Leveraging Machine Learning to Uncover Integrated Impacts of Urban Environment Features on Human Health

Jiafei Zhang

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2025

Committee:
Narjes Abbasabadi
Tomás Méndez Echenagucia
Edmund Y. W. Seto

Program Authorized to Offer Degree:

College of Built Environments

Abstract

Leveraging Machine Learning to Uncover Integrated Impacts of Urban Environment Features on Human Health

Jiafei Zhang

Chair of the Supervisory Committee:
Narjes Abbasabadi
College of Built Environments

While human health is significant to sustainable development as cities keep increasing and are estimated to contain 70% of the population by 2050, the quantitative relationship of how the urban environments influence human health remains so poorly understudied due to the convoluted connection and co-effect among numerous urban factors, as well as the large-scale data integration difficulty. Previous researchers studying urban health mainly focus on finding specific urban environment feature's influence or try to analyze the mechanisms of features co-effect on a coarse scale larger than census tract level or just with qualitative approach. Thus, this research set and reached the objectives to quantify the essential, specific and integrated relationship between urban environment and human health, and specify urban indicators' co-effect at census tract level mainly from the urban environment's perspective. With the development of data science, various statistical methods including Machine Learning (ML), SHAP analysis, and Pearson correlation analysis offer promising opportunities for analyzing such complex relationships, facilitated by the increasing availability of urban data by modern tools and the advancements in computational efficiency. By utilizing these methods, this research analyzes 6044 census tracts in 10 US metropolitans (New York City, Los Angeles, Chicago, Houston, Phoenix, Jacksonville, San Francisco Area, Seattle Area, Washington DC Area, Boston Area) with data from 2015 to 2024, builds a framework for census tract level urban health index developed from literature review and WHO urban health indicators framework (2014), quantifies 27 urban environment features' influence on human health indicators especially on life expectancy, and builds a high performance urban health prediction model for future intervention suggestion for policy makers, making significant meaning towards the urban planning practice.

Key Words: urban health index, machine learning, integrated urban feature importance, quantitative correlation, urban planning intervention, census tract

Give My Most Sincere Gratitude and Honor

Towards:

My thesis supervisor who leads me all along the way,

My research committee members who give me and this research great support,

My advisor who is always so warm hearted,

My classmates who bring lots of joy,

My friends who company me throughout the process.

For

This research cannot be done without all your great help and sponsor.

THANKS

CONTENT

ABSTRACT

Leveraging Machine Learning to Uncover Integrated Impacts of Urban Environment Features on Human Health

Jiafei Zhang

Chair of the Supervisory Committee:
Narjes Abbasabadi
College of Built Environments

While human health is significant to sustainable development as cities keep increasing and are estimated to contain 70% of the population by 2050, the quantitative relationship of how the urban environments influence human health remains so poorly understudied due to the convoluted connection and co-effect among numerous urban factors, as well as the large-scale data integration difficulty. Previous researchers studying urban health mainly focus on finding specific urban environment feature's influence or try to analyze the mechanisms of features co-effect on a coarse scale larger than census tract level or just with qualitative approach. Thus, this research set and reached the objectives to quantify the essential, specific and integrated relationship between urban environment and human health, and specify urban indicators' co-effect at census tract level mainly from the urban environment's perspective. With the development of data science, various statistical methods including Machine Learning (ML), SHAP analysis, and Pearson correlation analysis offer promising opportunities for analyzing such complex relationships, facilitated by the increasing availability of urban data by modern tools and the advancements in computational efficiency. By utilizing these methods, this research analyzes 6044 census tracts in 10 US metropolitans (New York City, Los Angeles, Chicago, Houston, Phoenix, Jacksonville, San Francisco Area, Seattle Area, Washington DC Area, Boston Area) with data from 2015 to 2024, builds a framework for census tract level urban health index developed from literature review and WHO urban health indicators framework (2014), quantifies 27 urban environment features' influence on human health indicators especially on life expectancy, and builds a high performance urban health prediction model for future intervention suggestion for policy makers, making significant meaning towards the urban planning practice.

KEY WORDS: urban health index, urban environment feature importance, machine learning, quantitative and systematic correlation analysis, census tract, urban planning intervention

1. Introduction and Background

- 1.1 Global Background and Research Gaps

Modern cities are dynamically facing widely recognized challenges of sustainability, which need to ensure economic growth, maintain social justice, and protect the environment [1]. Among them, human health and well-being are forever the most important themes and targets of what developments are for. As cities are estimated to contain 68% of the population by 2050 [2], most of the environmental influence on health would happen in the urban area. Researchers studying these issues are trying to find what urban features are important and to what extent and how they can influence or are correlated with human health, giving valid reference to policy making with the models trained and tools developed based on that, and giving essential insights in the direction of meaningful quantitative research in the future.

Cities and human health conditions, as complex systems, are often hard to make an integration of comprehensive but still specific recognition, the gap between macro view and micro view always exists larger than the granularity people expect. As Jane Jacob mentioned in *The Death and Life of Great American Cities*, such a phenomenon is like blind men who felt the elephant and pooled their findings [3]. While most research may have a deep dive into a specific problem to have sharp insights, the bigger picture is hardly tangible, being less explored thus researchers are not able to know how a specific point would contribute towards the system; research that focus on a macro view of the city may also bring several specific instances but still hard to have a proper granularized subdivision system to frame and match with micro approaches [4].

At the same time, with the trend of digitization, availability of emerging large datasets, optimizing of efficiency of computation, data science methods including Machine Learning (ML) for big data are promising to fill in such research gap, revolutionize the way we analyze and plan urban areas, building systematic and quantitative analysis framework for the integration of macro and micro views, and opening new opportunities for the sustainable city agendas [1], [5]. Thus, this research studies the correlations between urban form indicators and human health mainly with data science methods, focusing particularly on utilizing ML algorithms to uncover the integrated impacts of built environment features on human health.

- 1.2 Key Definitions

Health: The World Health Organization (WHO) defines health as a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. For mental health definition, it is defined as a state of wellbeing where the individuals realize their potential, cope with the normal stresses of life, can work productively and fruitfully, and can make contributions to the community [6]. Barton et al. (2009) defines the sphere of direct planning influence towards human settlements is by the built environment, which is the physical form and management of places such as the buildings, spaces, streets, and networks [7]. And the definition of urban physical environment is physical factors that comprise settlement area, facilities, utilities, green area, and transportation in the urban area [8], [9], [10], [11].

Health Indicators: 5 original categories including 35 features of urban health index are defined by WHO in 2014 (*The Urban Health Index: A handbook for its calculation and use*) [12], this study’s census tract level index is developed from it, by removing some indicators not applicable for census tract level, and adding some human activity indicators as well as comprehensive urban environment features like economic diversity and walkability from recent literature reviews. The indicators selected are listed below, which are generally applicable and important for neighborhood level health analysis.

Indicator	*Category*	*Developed From*
Life Expectancy	Health Indicators	WHO2014 – Health - Condition
General Health Condition		WHO2014 – Health - Condition
Mental Poor health Ratio		WHO2014 – Health - Condition
Physical Poor health Ratio		WHO2014 – Health - Condition
Physical Inactivity Rate	Human Activity	Lit Review
Economic Diversity		Lit Review (Jaine Jacobs, Kevin Lynch, Reid Ewing…)
Health Insurance Coverage		WHO2014 – Economics - Service
Unemployment Rate		WHO2014 – Economics – Unemployment
Population Density		WHO2014 – Geography – Population Density
Building Statistics	Built Environment	WHO2014 – Environment – Built Environment
Spatial Development Pattern		WHO2014 – Geography – Land Use
Air Quality	Natural Environment	WHO2014 – Environment – Air Quality
Plant Coverage		WHO2014 – Environment – Built Environment
Open Water Ratio		WHO2014 – Environment – Built Environment
Street Network Density	(Built Environment) Street and Traffic	WHO2014 – Geography – Spatial Patterns
Walkability		Lit Review (Jaine Jacobs, Kevin Lynch, Reid Ewing…)
Public Transport Accessibility		WHO2014 – Environment – Transportation
Hospital/Pharmacy Accessibility		WHO2014 – Health – Health care access

Table 1.2-1: urban health indicators used in this research

Technical terminology and abbreviations:

Abbreviation	*Full terminology*
ML	Machine Learning
DL	Deep Learning
SVR	Support Vector Regression
DT	Decision Tree
RF	Random Forest
SLFN	Single Hidden Layer Feedforward Neural Network
MLP	Multi-Layer Perceptron
CNN	Convoluted Neural Network
GAN	Generative Adversarial Network
ResNet	Residual Neural Network
SHAP	SHapley Additive exPlanations
API	Application Programming Interface
POI	Point Of Interest
GMP	Google Map Platform
OSM	Open Street Map
CDC	Centers for Disease Control and Prevention
ACS	American Community Survey
USALEEP	U.S. Small-area Life Expectancy Estimates Project
PLACES	Population Level Analysis and Community EStimates
NLCD	National Land Cover Database
EPA	U.S. Environmental Protection Agency
SLD	Smart Location Database

Table 1.2-2: Tech abbreviation

- 1.3 Research Objectives and Contributions

Though a great amount of research regarding urban health has been done, barely any one of them focused on all-factors integrated quantitative mechanisms especially on the fine grained (census tract) level, and few of them utilized the great potential of currently emerging data science methods and tools in integration research, and those using ML is still far less than exploiting its potential for both the tool and the field. Most researchers studying these issues start with the attempt to find specific urban environmental features' influence on human health quantitatively or try to analyze the mechanisms of how these features co-effect human health qualitatively, usually on a larger scale than neighborhood; or trying to research on individual's perceptual level to study certain features, but less related to the big picture of an urban scale. Thus, this research would fill in this gap using systematic data driven methods, analyzing 6044 census tract data from US metropolitans with ML and building an urban health consultant tool as an approach for policy makers to make queries and get recommendations.

To be specific, this research's objectives include:

1. Build the essential quantitative index of the urban environment for human health developed from literature review.

2. Train high performance models for health conditions prediction based on urban environmental features. The ideal expectation of the research contribution is to build the framework that covers most essential features of the urban environment at the census tract level, and that model should reach some prediction accuracy (e.g. 0.5-0.85).

3. Quantifying the effect of 10+ most important urban features for human health for US metropolitans. Quantifying essential and specific relationship between urban built environment and human health, specifically census tract environment features with life expectancy and mental disorder rate. Expected contribution would conceal some insightful findings and analyses that haven't been found previously.

4. Give Intervention analysis and suggestions. Also discuss interventions and possible important futural practice and research directions based on this research.

(**5. Optional - Build a tool to query the census tract human health conditions and give suggestions.** Expected result is a tool that, user can makes the query for both city level and neighborhood level; if it's their neighborhood's level, the tool analyzes census tract's health condition and give intervention advice based on correlation found by this research with analysis visualization (ideally with slidable bars that user can play with the indicators change and see the health performance improvement for their census tract), If it's the city level, show analysis of the whole city and which neighborhood should be improved in which features, thus help urban planning decision makers with quantitative evaluation reference.)

2. LITERATURE REVIEW AND RESEARCH FRAMEWORK DEFINITION

- 2.1 LITERATURE REVIEWS AND URBAN-HUMAN HEALTH CORRELATION FRAMEWORK BUILDUP

From literature review, research macro recognition frameworks, workflows, methodologies, and specific technical and statistical implementation and analysis methods are found to be great references of this research. From macro understanding of the urban system, the indexes selection, to the database build-up references, and statistical and machine learning analysis workflow and technical instructions, important literature related closely to this research are listed below with their main content and descriptions.

Lit Type	*Scope*	*Main Content or Contribution*
Urban health Framework	[12] The Urban Health Index: A handbook for its calculation and use. Kobe, Japan: World Health Organization; 2014	The UHI provides a flexible approach to selection, amalgamation, and presentation of health data. Its purpose is to furnish visual, graphical, and statistical insight into various health indicators and health determinants within particular geographic boundaries and health disparities with a focus on capturing intra-urban health disparities. The UHI may be used by public health workers, evaluators, statisticians, program managers, academic researchers, and decision makers to examine the current status of urban areas, to assess change and the effect of program interventions, and to plan for urban improvements.
Database	[13] Healthy Cities, a comprehensive dataset for environmental determinants of health in England cities. 2023	A comprehensive urban health dataset buildup method and result
Background	[3] The Death and Life of Great American Cities. Jane Jacobs, 1961.	A critique of 1950s urban planning policy, which gives insights and observations of how urban features function together and make influence comprehensively.
	[14] Delirious New York: A Retroactive Manifesto for Manhattan. Rem Koolhaas, 1978.	foundational text in architectural theory and urbanism. It explores the development of Manhattan's unique urban condition, which Koolhaas terms "Manhattanism", arguing that the city's extreme density, verticality, and inherent chaos fostered a "Culture of Congestion" and shaped its architectural identity.
	[10] Measuring the Unmeasurable: Urban Design Qualities Related to Walkability. 2009	Raised urban environment qualities that are essential to design that links the physical environment to human perception and human activity.
Urban Health Literature Review	[4] Neighbourhood community life and health: A systematic review of reviews. 2020	A review on what kind of urban environment indicators may specifically influence what health conditions.
	[15] Living environment and its relationship to depressive mood: A systematic review. 2018	How urban features influence depression (mostly on a large scale than census tract)
	[16] A Bidirectional Associations between Urban Physical Environment and Mental Health: A theoretical framework. 2020	Build the framework for analyzing how urban influence health: built environment and natural environment influence human activity, and these together influence human health.

	[17] Community design, street networks, and public health. 2014	Community level summary and findings for environment indicators and health
ML Urban Application Literature Review	[5] Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges. 2022	Listed ML methods and application for urban problem analysis
	[1] Machine learning for spatial analyses in urban areas: a scoping review. 2022	ML and urban analysis lit review
ML & Urban Research Methodology Reference	[18] Urban energy use modeling methods and tools: A review and an outlook. 2019	ML and urban energy modeling methodology, especially for numerical data
	[19] An integrated data-driven framework for urban energy use modeling (UEUM). 2019	Take Chicago as an example, built the data-driven framework for urban energy use modeling
Specific ML and Urban health analysis	[20] Deep learning in urban analysis for health. 2022	Methodology for DL in certain urban health analysis with similar scale to this research
	[11] Identifying correlations between depression and urban morphology through generative deep learning. 2023	Methodology for CycleGAN in mental health analysis: analysis of urban features from satellite image.
	[21] Use of Deep Learning to Examine the Association of the Built Environment with Prevalence of Neighborhood Adult Obesity. 2022	Methodology for DL in satellite urban image feature and obesity analysis
	[22] A big data evaluation of urban street walkability using deep learning and environmental sensors. 2020	ML methods for getting walkability feature
	[23] Subjective or objective measures of street environment, which are more effective in explaining housing prices. 2022	Methodology can be referenced (it explores the relationship between street environment features and housing price using ML)
	[24] A novel walkability index using google street view and deep learning. 2023	Methodology can be referenced for creating a novel index

Table 2.1: Literature review

A framework developed from these Lit reviews for analyzing urban features and health correlation is shown in [Fig.2.1-1], which considers urban-built-environment, natural environment, and human activity as a whole that affects human health, and within them urban-built-environment, natural environment again influences human activities, and specific urban and human features shown in [Fig.2.1-2] can found their proper positions within this qualitative integrated collection – such framework is a more amplified and meaningful one compared with the WHO's initial urban health index contents [Fig.2.1-3].

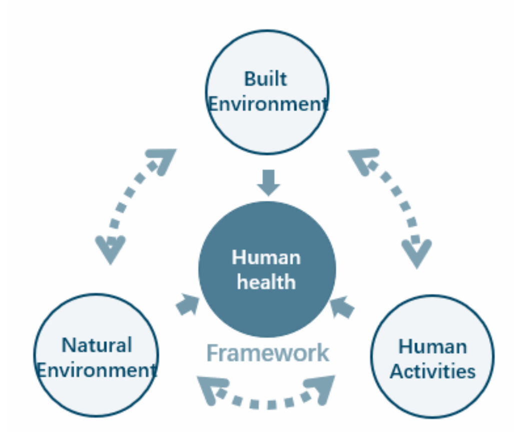


Fig. 2.1-1: Framework of city and human health relations, adapted from literature [16]

Below, integrating specific info into this “built-environment, natural-environment, human-activity – human-health” framework, we could see it comprises comprehensive urban features like Land use, Layout, Building Statistics, Infrastructure, Density, and Landscape, and can be further disassembled as more sub-categories like social-economics (population statistics, employment, social resource accessibility…), density (compactness, complexity, centrality…), mobility (walkability, various transport accessibility…), pollution (air quality, water quality, plant cover, food/earth/other-resources pollution, industrial land use…), climate (weather, plant cover type…), energy (mobility, industrial, commercial…), etc., which can be represented in this graph [Fig.2.1-3].

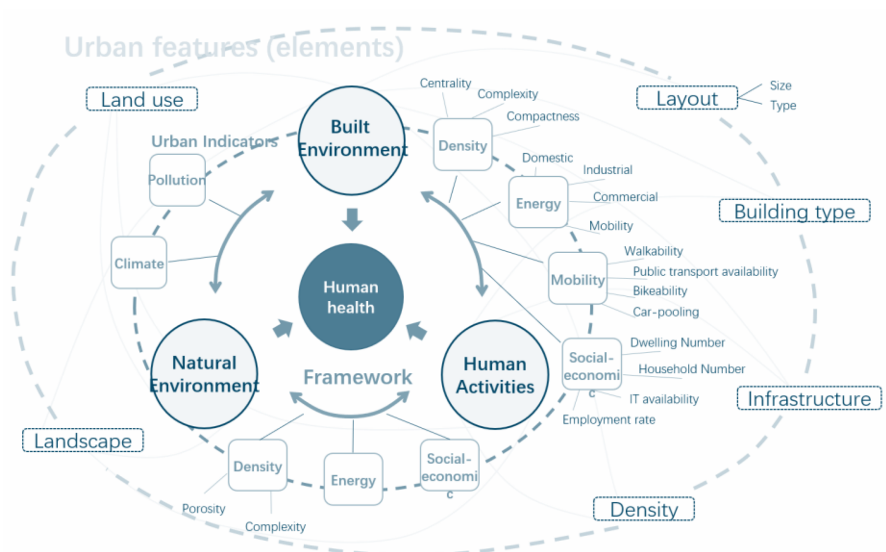


Fig. 2.1-2: How do city affect human health – formed based on previous lit review

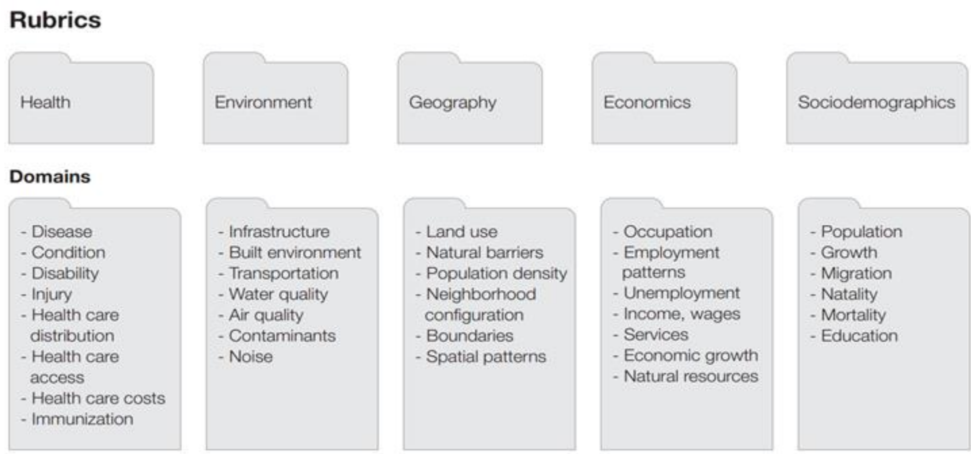


Fig. 2.1-3: Framework for Classifying Measures of Urban Health by WHO 2014 [12].

An evolving framework for a clearer, prioritizable and tangible quantitative research approach is shown below, where we can realize urban features (Physical environment and social environment measures) influence human activities then human health.

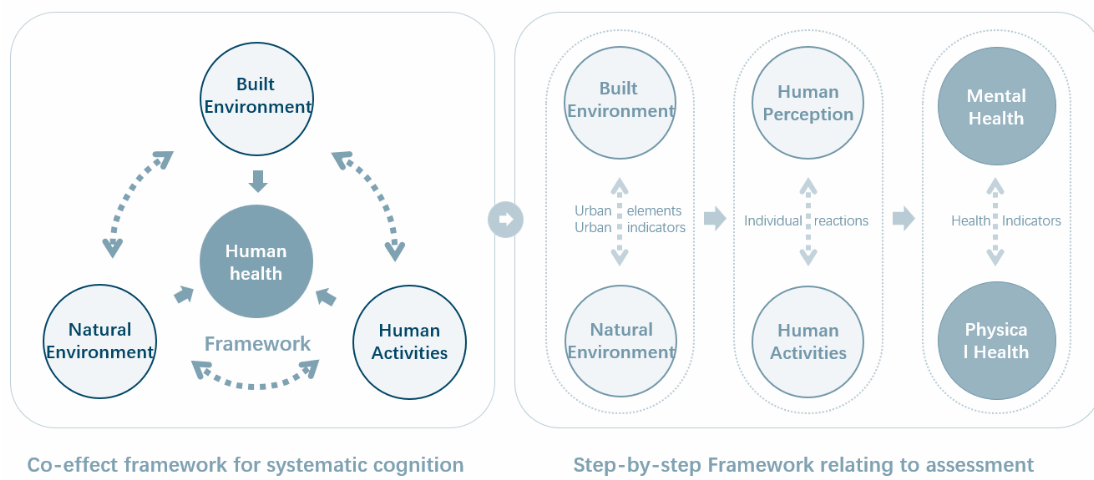


Fig. 2.1-4: How do cities affect human health – evolved framework for analysis (Edited by author).

Among all the features, what and how to choose, or how to define the features we want to use to form the urban health index framework better, an important concept can be introduced to help us here: urban-environmental-activity-friendliness – which literally integrates the core content of the framework, and can found its relationship with planner's familiar concepts like urban design quality, human perceptual qualities, walkability, etc. in previous urban design and planning historical discussions [Fig.2.1-5], so we could take such understanding and take relevant urban design features discussed as a helpful reference for the selection or definition of our features; another feature selection method relies on the data availability, and this research incorporate both these two methods to settled the features selected above. Thus, as is also mentioned, in the WHO 2014 features, the category of built environment is defined very coarsely; through works made in this research, we sub-define it with these essential factors: building density, building height, building floor area ratio, building type, tree canopy coverage ratio, shrub coverage ratio, open space ratio, developed low intensity ratio, developed medium intensity ratio, and developed high intensity ratio. There's no mention of walkability and economic diversity among WHO indexes, but since they are significant factors that link the communities built environment to human activity and neighborhood perception friendliness, as the graph from “Measuring the Unmeasurable: Urban Design Qualities Related to Walkability” shows, so they are added to the urban indicators’ list. Through these processes the selected indicators system could be considered well defined and valid to later analysis and model training processes.

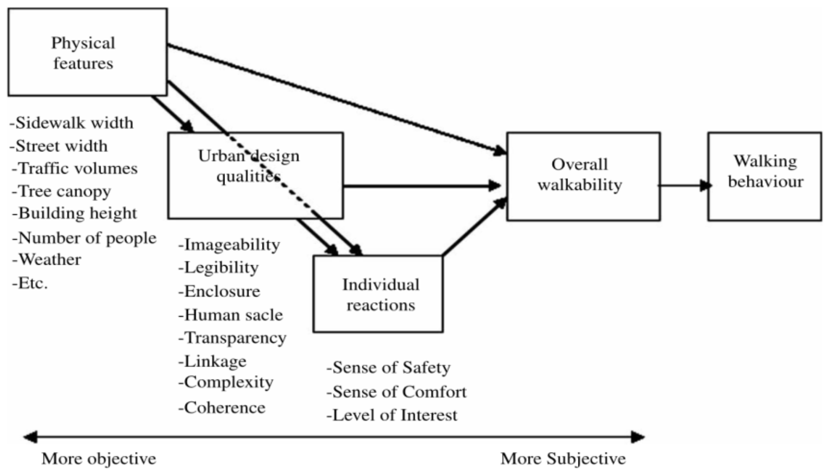


Fig. 2.1-5: Walkability conceptual framework from Reid Ewing & Susan Handy [10]

- 2.2 LITERATURE REVIEWS WITH ML APPLICATION IN URBAN STUDY

Literature has shown good examples of utilizing emerging statistical methods including ML to do such urban research. Generally, there are two approaches for ML application in urban research: Train model for statistical prediction, which aims at getting the high accuracy ML model to predict futural or unseen conditions based on known dataset or urban features, which may utilize all kinds of ML algorithms. Another approach is to train AI models that can do some direct design, among them most researchers are doing 2D image modification and design, baseline algorithms are mainly deep learning algorithms, including CNN, GAN, etc.

General Data driven approach for urban research includes traditional statistical model and artificial intelligence-based model, where ML is the central part of AI; ML algorithms include supervised learning, unsupervised learning, and reinforcement learning. From algorithms comparison mentioned by literature reviews regarding urban study [1], [5], this research decides to mainly use supervised learning regression algorithms and compares several of them for better learning and predicting urban health performance for US urban census tracts based on the input features. This may include predicting urban areas that are similar to the selected census tracts for current health conditions without direct health data, or predicting future health conditions given some certain intervention on some indicators – Since the urban built and natural environment hardly changes, so gaining as many indicators as possible regardless of their minor variance in time and spatial resolution is meaningful, finally the built database includes almost all features whose data is available at some time period in the last 10 years, while for each indicator the newest convincing data is used – such database can be considered useful for next decades, so one possible usage for models trained later can be predicting future health conditions given some certain intervention on indicators.

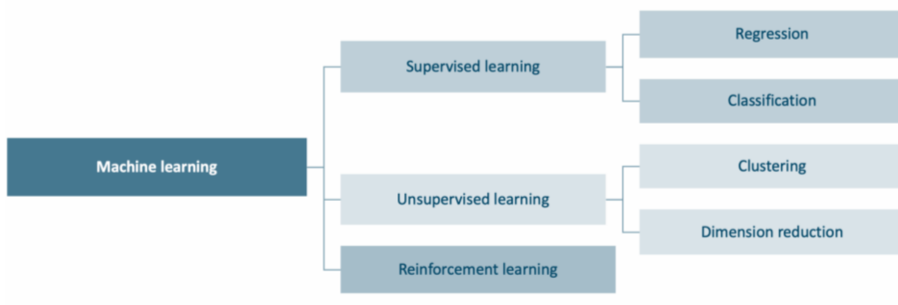
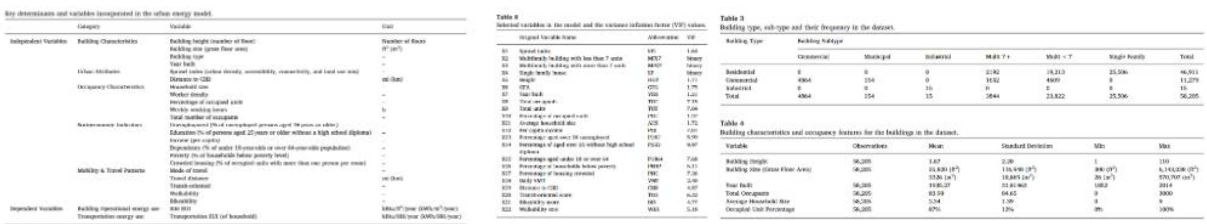
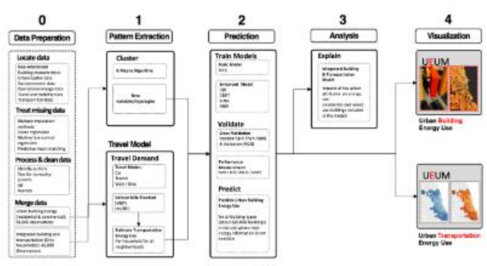


Fig. 2.2-2: Machine learning categories

- 2.3 METHODOLOGY AND WORKFLOW FROM LITERATURE REVIEWS





Reported data-driven error range as an accuracy metric for multiple previous studies of urban building energy use.

Source	Model	Accuracy Metrics		
		MAPE (%)	RMSE	MAD/MAE
[98]	MLR (OLS)	–	–	0.41–1.48
	RDF	–	–	0.40–1.32
	SVM	–	–	0.40–1.26
[99]	ANN (MLP)	20.6–22%	–	–
[97]	k-NN	1.81–2.38%	–	–

Fig. 2.3: Reference of methodology – Integrated data driven method for urban energy use modeling by Narjes Abbasabadi [19]

Literature like "An integrated data-driven framework for urban energy use modeling (UEUM)" introduced a method pipeline shown in [Fig. 2.3] that can be applied to this research, similarly it includes the steps of data integration, ML model training, tuning and comparisons, and result analysis and instructions. "Healthy Cities, a comprehensive dataset for environmental determinants of health in England cities" takes the UK as an example to build a comprehensive urban health dataset, whose method is also applied to this research to integrate various original datasets for the 10 US metropolitan areas. ML models for numerical urban data, according to ML and urban research methodology literatures [1], [5], [19], always include the comparison among the baseline linear regression, polynomial regression, decision tree and random forest, support vector model, and ANN models; the models' training and tuning methods are all utilized in this research.

Based on these literatures, this research's context, meaning, general scope, specific direction and methods applicable all emerge clearly. In the following section implementation methods of this research are elaborately described.

3. Methodology and Technical Implementation

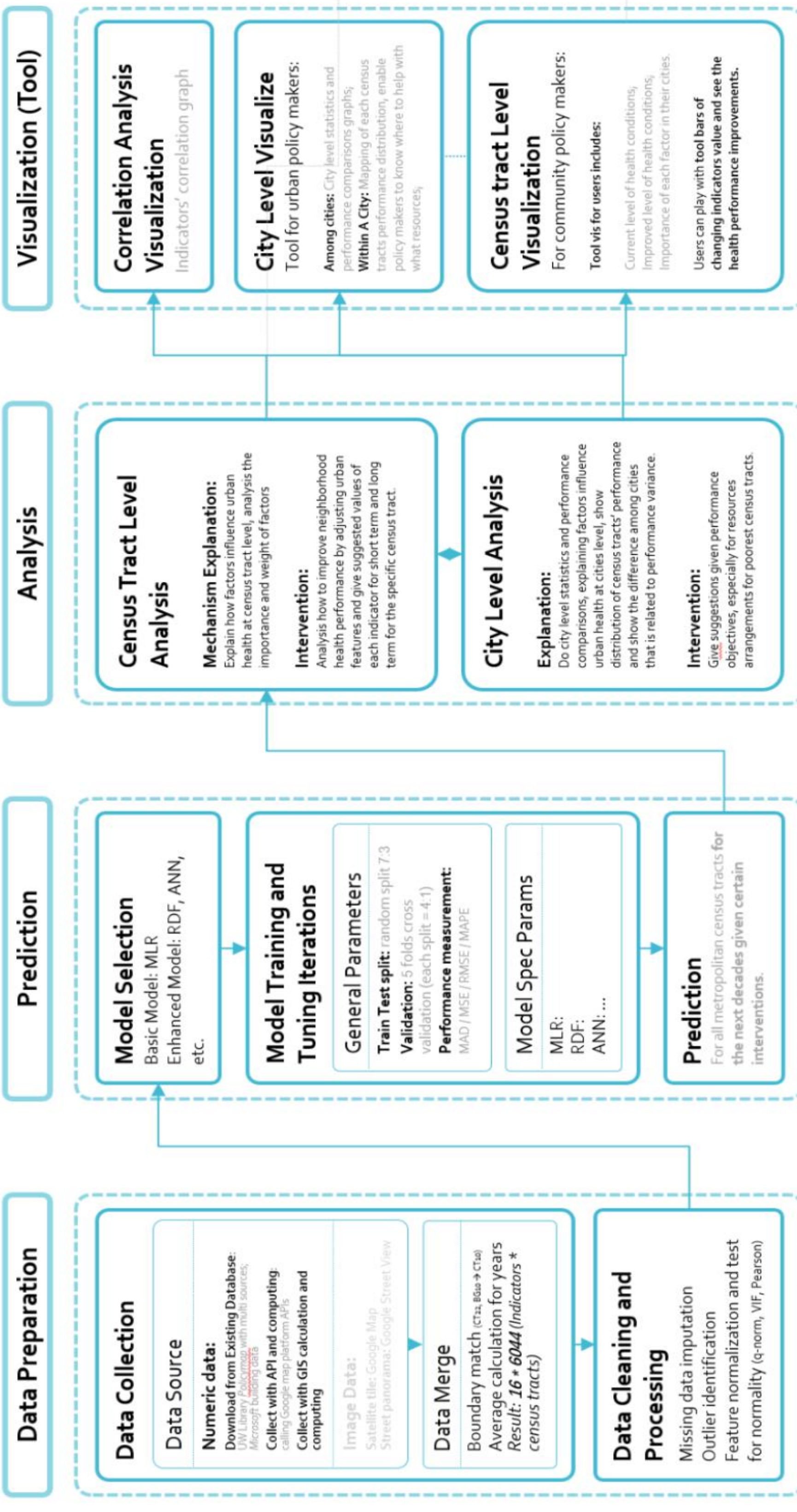


Fig.3.1: Research Implementation Workflow

3.1 General Workflow

The overall workflow of this study shown in [Fig.3.1] includes these major processes: data preparation, ML predictive model training, result analysis and visualization tool development. From the chosen urban features through literature review mentioned above, the urban sub-category indicators data are collected to make an original dataset, which mainly incorporates numerical data at present; the data cleaning and integration process then formatted all data to be matched with the census tract 2010 boundaries and normalized it, making it ready for ML model training. The buildup of the prediction model takes specific statistical steps including distribution analysis, ML model selection, feature engineering, ML model training, performance evaluation and parameter tuning, and analysis for feature insights during this process; thus, this process finally finds some best fitting models which are used for thorough analysis of indicator relationships and for further predictions for retrofitting strategies given some indicators adjustments. Based on the model, further analysis steps not only include the overall features analysis and intervention analysis that shows what kind of improvements for what features should be applied, but also developed an algorithm for census tract level analysis and retrofitting strategies recommendation. Such results are then converted to a web-based tool as a final demonstration, making it easy and friendly for the public and the policy makers to consult about city or neighborhood conditions within these metropolitan areas.

3.2 Data Processing

3.2.1 Data Sources and Integration Summary (Data Collection)

Dataset build up process utilizes several methods, including several statistical methods for the well sampling and integration of various data, and computational data API calling and calculation by python, together with geographical processing in ArcGIS; finally, all urban features are aggregated by census tract geographical boundaries. This section would give specific descriptions and explanations of the process.

The general city selection method is: 10 US cities, 6044 census tracts in total are selected, which is enough for comprehensive ML model training and analysis while keeping relative time efficiency. The 10 cities are all modern ones in sound economic development conditions, 9 of which are of the top 15 largest population metropolitan area and they scattered across the US, so they can be representative of typical US developed modern cities, thus the census tracts data are in the same large urban context but having enough variance, making the original data more universal and ML model trained based on it more generalizable. These cities are: New York City, Los Angeles, Chicago, Houston, Phoenix, Jacksonville, San Francisco Area, Seattle Area, Washington DC Area, Boston Area.

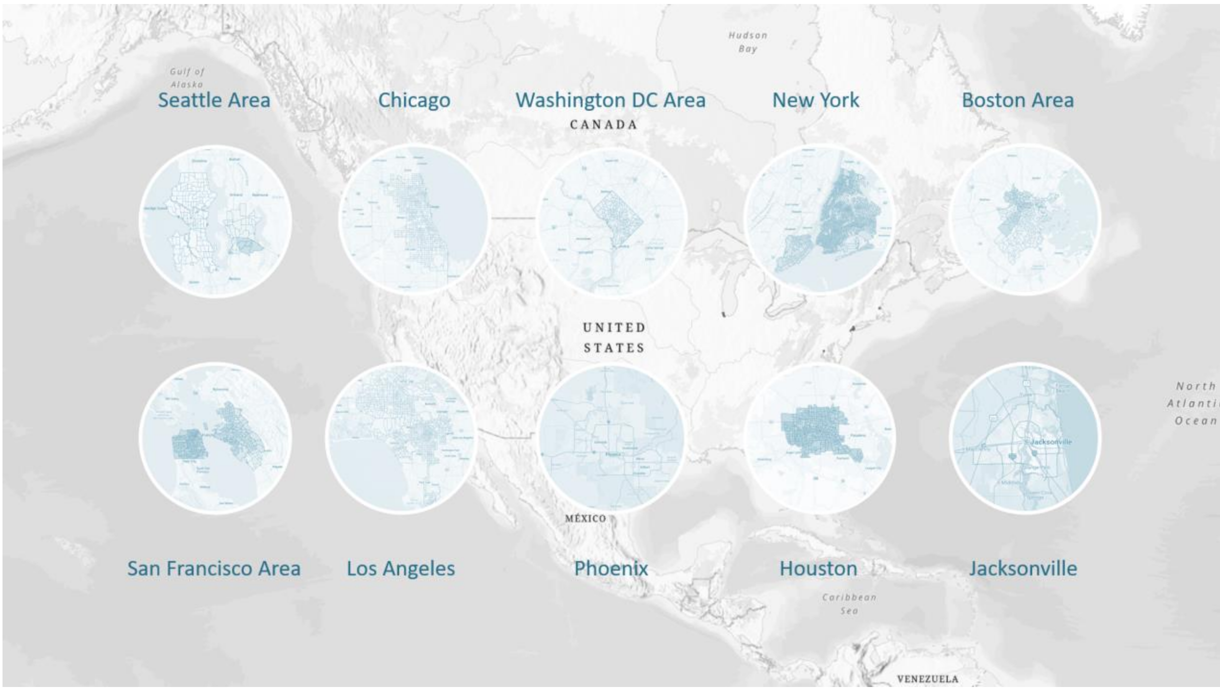


Fig.3.2.1: 10 cities selected

The integrated dataset shown in [table 3.2.1] consists of 5 large categories, below is the simplified name of each: Health Metrics – Life Expectancy, Poor Mental Health, Poor Physical health, Poor Health, Good Health; Human Related Features – Population Density, Physical Inactivity, Economic Diversity, Unemployment, Health Insurance Coverage; Natural Environment Features – Air Quality Index, Tree Canopy Cover, Shrub Cover, Herb Cover, Haypast Cover; Built Environment Statistics – Building Height, Building Density, Plot Ratio, Building Footprint Area Variance, Building Area Variance, Open Space Ratio, Low Development Intensity Ratio, Medium Development Intensity Ratio, High Development Intensity Ratio; Street and Traffic Features – Street Network Density, Walkability, Nearest Distance to Transit, Nearest Hospital Distance, Nearest Hospital Driving Time, Nearest Pharmacy Distance, Nearest Pharmacy Driving Time. Below is an introduction of each of these original datasets.

Health Metrics – from CDC USALEEP and PLACES.

Life Expectancy: The average number of years a person born in this tract would be expected to live, as of 2010 to 2015, from CDC USALEEP.

Poor / Good Health: Crude percent of [fair or poor] / [good or excellent] self-rated health status among adults in 2019-2021 each year and average. A multilevel regression and post-stratification approach was applied to BRFSS and ACS data to compute a detailed probability among adults who report their general health status as ["fair" or "poor"] or ["good" or "excellent"]. The probability was then applied to the detailed population estimates at the appropriate geographic level to generate the prevalence. This and below health metrics are all from CDC PLACES.

Poor Mental Health: Crude percent of frequent mental distress among adults in 2019-2021 each year and average. A detailed probability among adults aged >= 18 years who report that their mental health (including stress, depression, and problems with emotions) was not good for 14 or more days during the past 30 days.

Poor Physical Health: Crude percent of frequent physical distress among adults in 2019-2021 each year and average. a detailed probability among adults who reported 14 or more days, during the past 30 days, that their physical health (including physical illness and injury) was not good.

Built Environment Features – originally from multiple databases like Microsoft dataset and each city's data portal, open street map, etc., with the calculations in GIS and python.

Building Density: The sum of building footprint area per unit of land area.
Building Height: Average Building Height of a census tract.
Plot ratio: The sum of all building areas per unit of land area.
Building Area Variance: Statistical Variance of all buildings total area in a census tract.
Building Footprint Area Variance: Statistical Variance of all building's footprint area in a census tract.

Open Space Ratio: Areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover. These areas most commonly include large-lot single-family housing units, parks, golf courses, and vegetation planted in developed settings for recreation, erosion control, or aesthetic purposes.

Low Development Intensity Ratio: Areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 20% to 49% percent of total cover. These areas most commonly include single-family housing units.

Medium Development Intensity Ratio: areas with a mixture of constructed materials and vegetation. Impervious surfaces account for 50% to 79% of the total cover. These areas most commonly include single-family housing units.

High Development Intensity Ratio: highly developed areas where people reside or work in high numbers. Examples include apartment complexes, row houses and commercial / industrial. Impervious surfaces account for 80% to 100% of the total cover.

Human Related Features – from ACS and CDC PLACES.

Population Density: Estimated number of people per square mile, between 2019-2023.

Physical Inactivity: Crude percent of no leisure-time physical activity among adults in 2019-2021 each year and average. A multilevel regression and post-stratification approach was applied to BRFSS and ACS data to compute a detailed probability of having no leisure-time physical activity (reporting 'No' to the question: "During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?"). The probability was then applied to the detailed population estimates at the appropriate geographic level to generate the prevalence.

Economic Diversity: Economic diversity in 2021 is calculated as diversity of employment using 8 major sectors: retail, office, service, industry, entertainment, education, healthcare, and public administration. Diversity of these employment sectors is calculated as "entropy," or the mixture or evenness of distributions across employment types. EPA-SLD summarizes for every Census block group.

Unemployment Ratio: Estimated percent of civilian people age 16 years or older in the labor force who were unemployed, between 2018-2022.

Health Insurance Non-coverage: Estimated percent of the population without health insurance coverage, between 2019-2023. Percentage calculations were suppressed in cases where the denominator of the calculation was less than 10 of the unit that is being described (e.g., households, people, householders, etc.).

Natural Environment Features – from NLCD and google API.

Average Air Quality Index: Universal AQI collected from google air quality, which takes 9 times average AQI for a census tract during a month in 2024.10. The higher the value is, the better the air quality is.

Tree Canopy Cover: Tree canopy cover is measured as the percentage of ground covered by tree projections. Originally at a resolution of 30-meters.

Shrub Cover: Areas dominated by shrubs; less than 5 meters tall with shrub canopy typically greater than 20% of total vegetation. This class includes true shrubs, young trees in an early successional stage or trees stunted from environmental conditions.

Herb Cover: Areas dominated by graminoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling but can be utilized for grazing.

Haypast Cover: Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops, typically on a perennial cycle. Pasture/hay vegetation accounts for greater than 20% of total vegetation.

Street and Traffic Features – from EPA-SLD and google maps

Street Network Density: Total road network density in 2021. This was calculated by dividing the number of roadway links from intersection to intersection by the land area of the block group. The Environmental EPA-SLD summarizes for every Census block group. The same as walkability index and distance to the nearest transit stop.

Walkability Index: The National Walkability Index in 2021 identifies areas with mixtures of land use and transportation infrastructure that may encourage walking as a mode of transportation. This index comprises four ranked measures: intersection density (+), distance to the nearest transit stop (-), employment diversity (+), and employment and housing diversity (+). More walkable areas rate higher on intersection density, have lower distances to the nearest transit stop, and have higher employment and employment plus housing diversity.

Distance to Nearest Transit Stop: Distance (2021) is calculated from the population-weighted centroid of the block group to the nearest transit stop.

Distance to Nearest Hospital / Pharmacy: Straight distance from a census tract center towards nearest hospital or pharmacy.

Driving Time to Nearest Hospital / Pharmacy: Google map navigation driving duration from the census tract central to nearest hospital or pharmacy.

Types	*Feature*	*Dataset Name*	*Year*	*Data Unit*	*Original Space Unit*	*Source*
Human-Health	*Life Expectancy*	Life Expectancy	2010-2015	year	Census Tract 2010	CDC National Center for Health Statistics, Small-area Life Expectancy Estimates. [25]
	Poor Health	Fair or poor health rate	2019-2022	percent	Census Tract 2022	CDC: Population Level Analysis and Community Estimates (PLACES). [26]
	Good Health	Excellent health rate	2019-2022	percent		
	Physical Health	Frequent physical distress rate	2019-2022	percent		
	Mental Health	Frequent mental distress rate	2019-2022	percent		
Human-Related-Factors	*Population Density*	Estimated number of people per square mile	2018-2022	number per square miles	Census Tract 2022	Census: Decennial Census and American Community Survey (ACS) [27]
	Health Insurance Coverage	No health insurance coverage rate	2018-2022	ratio		
	Employment Density	Estimated percentage of people aged 16 years or older who were unemployed	2018-2022	percent		
	Physical Activity	Physical Inactivity Rate	2022	percent		CDC PLACES [26]
	Social Economic Activity	Economic Diversity Index	2021	index (entropy 0-1)	Block Group 2010	
Natural-Environment Features	*Air Quality*	Air Quality Index (AQI), Dominant Pollutant, Average Air Quality Level, Worst Air Quality Level)	2024	number (uaqi)	Census Tract 2010	Google Maps API [28] [29]
	Plant Coverage	Tree Canopy Coverage	2016	percent		National Land Cover Database at Census Tract Level (2016, mainland US) [30]
		Shrub, Herb, Haypast Cover	2016	percent		
Built-Environment-Related-Features	*Building Statistics*	Building Height (Footprint)	2014-2024	meter	Building	Microsoft [31], OSM [32] [33], Figshare [34], and each city's Data Portal (E.g. some especially useful ones like for LA [35], SF [36])
		Building Density	2014-2024	square meter /	Building	GIS + Python Calculation

	Building Floor Area Ratio	2014-2024	square meter	Building	
Spatial Pattern	Percent of Open Space (developed ct)	2016	percent	Census Tract 2010	National Land Cover Database at Census Tract Level (2016, mainland US) [30]
	Percent of Low Intensity (developed ct)	2016	percent	Census Tract 2010	
	Percent of Median Intensity (developed ct)	2016	percent	Census Tract 2010	
	Percent of High Intensity (developed ct)	2016	percent	Census Tract 2010	
Street Network	Street Network Density	2021	number per square miles	Block Group 2010	EPA Smart Location Database [37]
Transit Stop Distance	Distance to nearest transit stop	2021	meter	Block Group 2010	
Walkability	National Walkability Index	2021	index (score 1-20, 20 is high)	Block Group 2010	
Health resource accessibility	Shortest distance and time to a hospital;	2024	meters & seconds	Census Tract 2010	Google Maps API [28] [38] [39]
	Shortest distance and time to a pharmacy)				

Note: Some data are directly from the Policy Map platform, which gathered various original source data [40].

Table.3.2.1: Database information

During the process, how to define the space and time resolution, how to make proper sampling of the data, and then how to match all data to be in the same census tract boundaries are the main challenges. Considering that most urban features are not changing drastically in recent years since these cities are already in a relatively highly developed condition, so including more features available regardless of their minor variance in a recent time period or space can make the system more amplified, and the exploration and possible results of this research could be meaningful for decades later, therefore this research uses a broad time resolution of 10-years, and an earlier common base of spatial resolution defined by 2010 census tract boundaries.

The 10 years' time resolution means all data collected are within 2015-2024, while most are collected at a time near 2020, like many features including 4 health metrics are having the records of each year from 2019-2021 and this research takes the average of them as that feature's data, to enhance the data's generalizability as much as possible; some records like building height data is always in very slow completing process but even the most recent version still have some historical blank, so this research prioritized the most recent height data available while incorporating multiple sources; the air quality and hospital/pharmacy accessibility data are collected in 2024, considering air quality's fluctuation in time a 9 time-periods' sampling are made for each census tract to reduce its inaccuracy and randomness; the most recent life expectancy (at birth) available is gained by survey from 2010-2015, a bit earlier than 2020, but since it is such an essential and comprehensive feature that is gained through a long time period and can still reflect decades of urban environment influence, so we still take it as the most important health metric.

As for the spatial resolution, data of each feature have several original boundaries, like census block group in 2010, census tract in 2022, and census tract in 2010; the former two in most cases are subdivisions of or overlap with census tract 2010 boundaries, so computation methods are utilized to match the spatial boundaries to be all as Census Tract 2010, which would be described in next paragraphs.

Overcoming some other challenges like the acquisition of the height, air quality, and hospital/pharmacy accessibility data utilized some technical methods, which would also be described soon in the following sections.

- 3.2.2 Census Tract Data Matching

Mapping of the spatial boundary includes some sticky data processing, but important steps include: Coding to find those CT22 who is only within one CT10's boundary range or with boundary overlapping (within certain tolerance), by checking the minimum and maximum range of the boundary coordinates; For those within 2 or more 2010 CT XY boundary ranges, check distance of center point and difference of the shape length etc., and change tolerance to stricter values to ensure partial extremes of the CT22's reflection to CT10; for those few still undecidable, check manually. Very little data during census tracts mapping with too large a difference of boundaries would be removed. If a census tract comprises several subdivision blocks, then relevant features would be re-calculated by the weight of the block's population (or land area), to make an average value as their representation for that census tract.

Space Unit	*ID Digits*	*ID Format* *(E.g. UW)*	*Boundary*
Block Group 2010	12	530330053022 530330053023	Subdivisions of census tract 2010
Census Tract 2010	11	53033005302	

Census Tract 2022	11	53033005303	Mostly subdivisions of census tract 2010

Table.3.2.2: Info for mapping different space units

Fig.3.2.2: Census tract boundary difference of year 2022 and 2010

- 3.2.3 Google API Calling for Air Quality and Health Facility Accessibility Data Collection

Google Maps API is used for collecting air quality data and the health facility accessibility.

The air quality API allows for getting the universal air quality index (a score range from 0 to 100, higher means better air quality) at a certain coordinate at a specific 1-hour time period from the last 30 days when the API call (an http request) is sent; 9 timestamps air quality index score is get for an average, recorded as follows:

'2024-10-10T16:00:00Z' – 2024-10-10T09:00:00 (9:00 AM PDT),
'2024-10-10T21:00:00Z' – 2024-10-10T14:00:00 (2:00 PM PDT),
'2024-10-11T02:00:00Z' – 2024-10-10T19:00:00 (7:00 PM PDT),
'2024-10-20T16:00:00Z' – 2024-10-20T09:00:00 (9:00 AM PDT),
'2024-10-20T21:00:00Z' – 2024-10-20T14:00:00 (2:00 PM PDT),
'2024-10-21T02:00:00Z' – 2024-10-10T19:00:00 (7:00 PM PDT),
'2024-10-30T16:00:00Z' – 2024-10-20T09:00:00 (9:00 AM PDT),
'2024-10-30T21:00:00Z' – 2024-10-20T14:00:00 (2:00 PM PDT),
'2024-10-31T02:00:00Z' – 2024-10-10T19:00:00 (7:00 PM PDT).

It also enables the user to get the dominant pollutant, pollution level, etc., which can be used for further analysis, though non-numerical data are not used when training the ML models in this research.

The healthcare accessibility utilized the “nearbySearch” and “computeRoutes” API. Nearby health facilities are searched at the centroid coordinate of a census tract in a radius of 5km, and straight-line distance is calculated for the nearest hospital and pharmacy, and the “computeRoutes” calculated the driving time for getting the nearest facilities. Utilizing such http requests, these crucial information are gained to make the urban-environment-health-evaluation system more amplified and later results more convincing.

Feature Name	*Google API Name*	*Extra Computing*	*Result – Collected Data*
Health Care Accessibility	nearbySearch computeRoutes	Compute the nearest place to the current census tract center and then	Shortest distance and time to a hospital,

		call the computeRoutes API for that specific nearest place	Shortest distance and time to a pharmacy
Air Quality	Air quality gets hourly history		AQI, dominant pollutant, average air quality level, worst air quality level

Table.3.2.3: Google maps API for air quality and health facility accessibility data collection

- 3.2.4 Building Height Data Merge Process

Building height data sources used are: Microsoft Global ML Building Footprints; US Building height; Open Street Map; Data Portal of each city. For most cities, a basic dataset of OSM building footprint is matched with the height data gained from the Microsoft Global ML Building Footprints dataset, with the patches for large blocks missing data of height from other sources. Below is a simple process of how the base dataset is obtained – Download relevant grids' data from Microsoft database and then cut them by census tract boundaries, make patches for missing data, and match the height data with the original OSM's building footprints according to the overlapping of the geometries (or the nearest centroid). Some cities have direct well-completed 3D data from their data portal, like LA, so we utilized it directly. It is worth mentioning that all sources have lots of defects, like height inaccuracy, footprint error, data bias (overestimation or underestimation of most data or some certain types within it), outdated data or too much missing buildings, which makes the integrated dataset have its intrinsic limitations. In fact, all datasets of each feature may have kinds of limitations, which would be discussed in chapter 5, but these datasets can still be very important and useful in building up the system, learnable and interpretable by later machine learning training, tuning and analyzing processes.

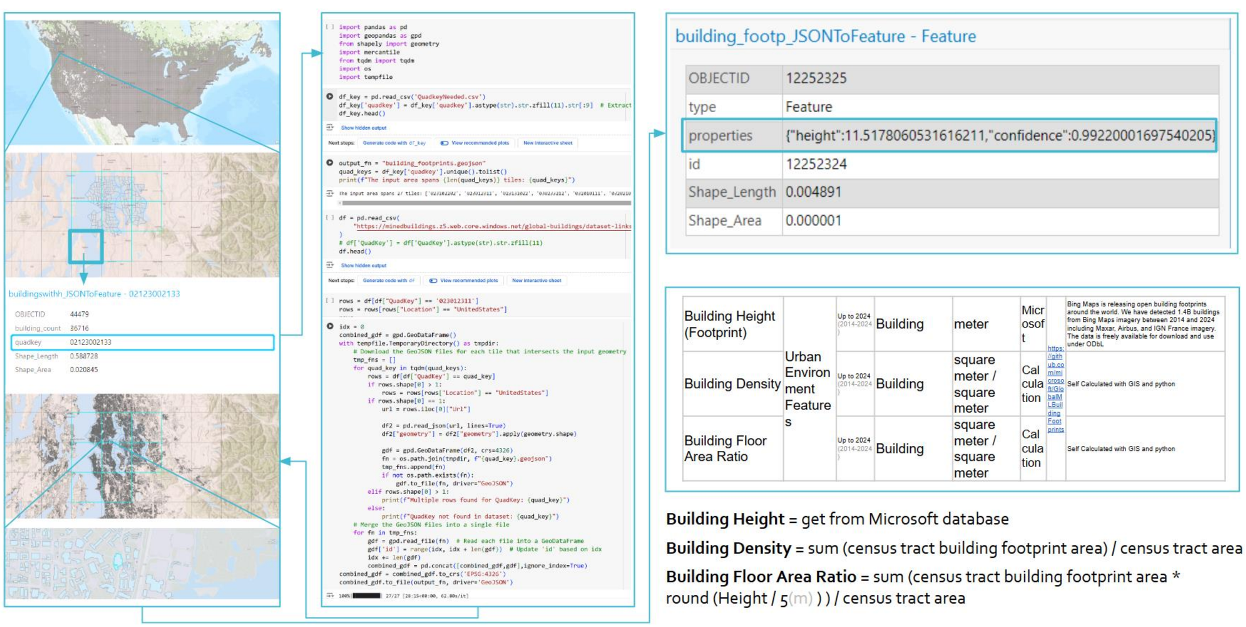


Fig.3.2.4: Building height data collection from Microsoft Global ML Building Footprints

- 3.3 Machine Learning Training and Tuning Methods

After building the database and processes of cleaning, several statistical and machine learning algorithms are applied to learn the relationships – from data distribution analysis, Pearson Correlation, to ML models comparisons and the best model MLP's deep interpretation

– important results would be discussed along with the statistical methods and ML models in the result chapter (chapter 4), while this section would describe the specific methods used for machine learning model training and tuning, like parameters grid search method, before we jump to results.

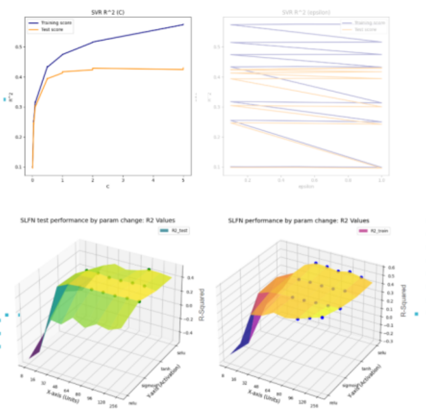
For linear regression models, the baseline model OLS and upgrade models like Bayesian are tested, several penalty algorithms like LASSO or Ridge are applied to them, and different parameters of k for k-folds cross validation are also tested.

For Polynomial Regression, degrees (the highest power of the independent variable (x) in the polynomial equation) of 1 to 4 are tested so it's found that a degree of 2 performs best for life expectancy prediction.

Grid search is used for multi-hyperparameters model tuning. For Support Vector Regression, a grid search is done for hyperparameter "C (Regularization Parameter)" and "Epsilon (ε - margin of tolerance around the predicted values)"; for ANN models, multi-dimensional grid search is applied to the choice of each layer's units and activation function, which shrinks the optimal hyperparameters range and helps the tuning to best models.

ML Models for Regression:

Model Type	Model Name
Linear Regression	OLS Baseline
	BayesianRidge with CV
Polynomial Regression	Polynomial ElasticNet with CV
Support Vector Regression	SVR with RBF Kernel
Tree	Decision Tree
	Random Forest
ANN	SLFN
	2 Hidden Layers



	Activation	Performance	8	16	32	48	64	80	96	128	256
Train	relu	RMSE	4.59	4.23	3.37	3.30	3.09	3.15	3.07	2.92	2.78
		R2	-0.21	-0.03	0.35	0.38	0.45	0.43	0.46	0.51	0.55
	sigmoid	RMSE	4.69	3.27	3.14	3.11	3.09	3.08	3.05	3.04	3.02
		R2	-0.26	0.38	0.43	0.45	0.45	0.45	0.46	0.47	0.47
	tanh	RMSE	4.11	3.22	3.12	3.08	3.08	3.09	3.09	3.05	3.15
		R2	0.03	0.40	0.44	0.45	0.46	0.45	0.45	0.46	0.43
	selu	RMSE	4.02	3.51	3.16	3.00	3.02	3.08	2.99	3.07	3.15
		R2	0.07	0.29	0.43	0.48	0.47	0.45	0.48	0.46	0.43
Test	relu	RMSE	5.07	4.80	3.77	3.81	3.48	3.70	3.46	3.54	3.36
		R2	-0.49	-0.33	0.18	0.16	0.30	0.21	0.31	0.27	0.35
	sigmoid	RMSE	4.67	3.27	3.06	3.02	3.01	3.00	3.00	2.97	2.96
		R2	-0.26	0.38	0.46	0.47	0.47	0.48	0.48	0.49	0.49
	tanh	RMSE	4.03	3.31	3.08	3.05	3.11	3.11	3.09	3.08	3.15
		R2	0.06	0.36	0.45	0.46	0.44	0.44	0.45	0.45	0.42
	selu	RMSE	4.41	3.96	3.40	3.24	3.35	3.26	3.12	3.13	3.19
		R2	-0.13	0.09	0.33	0.39	0.35	0.39	0.44	0.43	0.41

Fig.3.3: Grid search method for ML model tuning

3.4 Tool Development Methods and Implementations

Tool development consists of steps of client needs definition, frontend UI design, system components design, tech stack choice, and integration tests. Major users for this tool are urban policy makers who want to know the urban current conditions of each indicators, the general health condition, and how to do the interventions for each feature and how big a difference can be made, so the preliminary demo version developed at present makes both the city's page and the census tract's page, where features distribution can be glanced and compared at city level, and intervention suggestions and effects are also available at census tract level, with the Javascript and React as the frontend tech stack and python Flask as the backend framework.

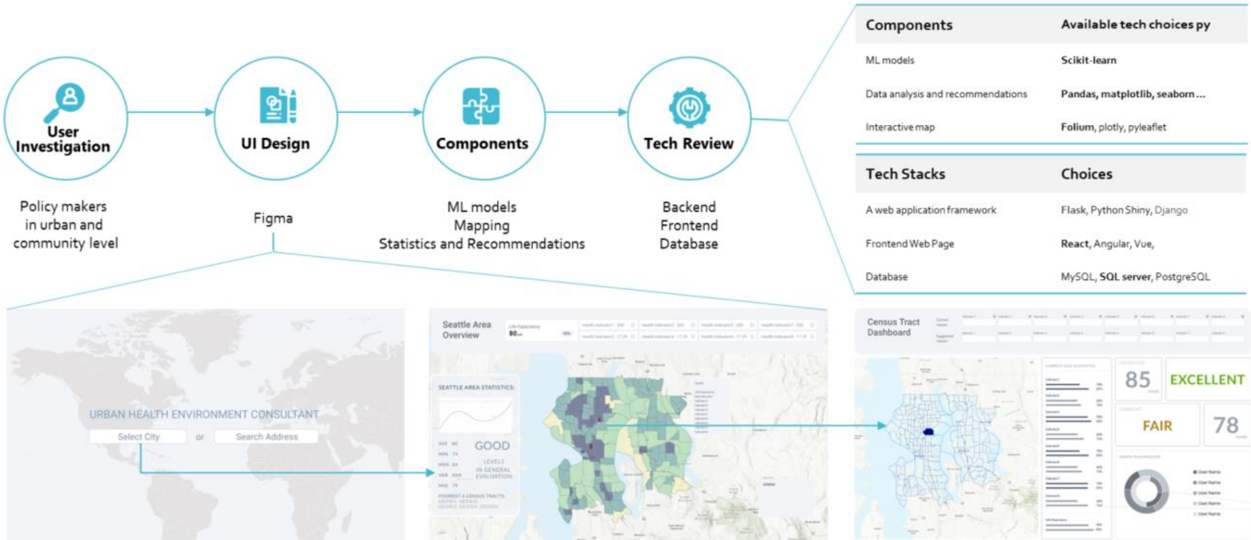


Fig.3.4: Web tool development methods

After the description of important methods like preparing the database, training models and developing tools, an elaborate analysis would be made for the most important part – result discussion.

4. Result Analysis

Aiming to know urban features and health performance relationship, we'd elucidate these following questions in this part: 1. how **features** are relevant, what the most important features are and whether irrelevant features exist; 2. how well can **models** learn the relationship and predict and what's the most meaningful model; 3. how the most important features exerts influence, what are meaningful **interrelationships**, or what's the possible important causation within these relationships – this section gives in-depth analysis of these problems by comparing various cities, models and indicators, starting from single parameter analysis, correlation analysis, to model performance analysis and prediction, and finally explores best intervention strategies and their effects, with the validation, discussion, or thorough elaboration of some city districts' examples.

To start with, the following tables give an overall glance of different models performance for distinct health indicators so that we know all our health measures are shown to have a good or strong relationship with urban features if good models are applied; the tables also show whether a certain important parameter(physical inactivity rate) can exert all the influence on each health measurements. With the first table below, it's obvious that Poor General Health Ratio, Good General Health Ratio, and Poor Physical Health Ratio within census tract are having pretty strong correlation with urban environment and human indicators reaching a R^2 of 0.83 and more, while Poor Mental Health Ratio and Life Expectancy are not as extremely strong but still in good correlation with urban and human indicators; further, urban and human indicators are found to perform influence together instead of impacting health by mainly one or two important ones among them, as the most important single parameter Physical Health Inactivity Rate cannot reaches as high the prediction performance as including all features (Table …), with a R^2 reduction of 0.09-0.30 for the single X in comparison to all features included. Moreover, the Life Expectancy and Poor Mental Health Rate, though not build up the highest performance model in R^2 compared to other health measurements, are or more reality importance, since Life Expectancy is a typically more comprehensive parameter, and the mechanism of improving them both are especially vague and harder to be quantified; while single parameter like Physical Inactivity Rate can be found in strong correlation to physical or general health condition(with a R^2 gap of less than 0.1 compared to full parameters model), it is not as important for life expectancy and poor mental health(with a R^2 gap of 0.3 and 0.17 each compared to full parameters model). Thus, to decode the complex relationships between urban, human, and health with better reality meaning, we'd deep dive analyzing all these health measurements, especially in life expectancy and mental health rate.

Model Type	*Model Name*	*Metric*	*Life Expectancy*	*Mental Health*	*Physical Health*	*Poor Health*	*Good Health*
Linear Regression	OLS Baseline	R^2	0.3675	0.5494	0.8059	0.8374	0.7421
		RMSE	3.305	1.9946	1.5416	3.1812	4.0164
	Bayesian Ridge with CV	R^2	0.3873	0.5641	0.8085	0.8403	0.7471
		RMSE	3.253	0.9617	1.5314	3.1530	3.9776
Polynomial Regression	Polynomial ElasticNet with CV	R^2	0.4402	0.5488	0.8095	0.8424	0.7562
		RMSE	101.5	67.834	51.913	106.44	132.72
Support Vector Regression	SVR with RBF Kernel	R^2	**0.5025**	**0.6494**	**0.8209**	**0.8572**	**0.8091**
		RMSE	2.931	1.7595	1.4811	2.9812	3.4556
Tree	Decision Tree	R^2	0.4093	0.5633	0.7997	0.8440	0.7644
		RMSE	3.194	1.9636	1.5660	3.1161	3.8391
	Random Forest	R^2	**0.5449**	**0.7084**	**0.8502**	**0.8847**	**0.8373**
		RMSE	2.804	1.6044	1.3545	2.6781	3.1908

ANN	SLFN	R^2	**0.5161**	**0.6467**	**0.8369**	**0.8717**	**0.8247**
		RMSE	2.895	1.7662	1.4205	2.8258	3.3114
	2 Hidden Layers	R^2	**0.5208**	**0.6631**	**0.8451**	**0.8799**	**0.8347**
		RMSE	2.877	1.7246	1.3773	2.7334	3.2155

Table.4-1: ML Model Results, Performance and Parameters

Linear Model Performance (R-squared, Amplified Data)									
Y	X - Physical Inactivity	Physical Inactivity	X	Physical + Mental Health Statistics	X + Physical & Mental Health Statistics	Good + Poor Health Statistics	X + Good & Poor Health Statistics	All Health Statistics	X + All Health Statistics
Life Exp	0.3270	0.2275	**0.3675**	0.4025	**0.466**	0.364	0.4357	0.4125	**0.4755**
Mental Health	0.4480	0.4965	0.5494	--	--	0.7891	0.8056	--	--
Physical Health	0.5350	**0.7567**	0.8059	--	--	0.9632	0.9748	--	--
Good Health	0.5644	0.6609	0.7422	0.6742	0.7966	--	--	--	--
Poor Health	0.6025	**0.7854**	0.8374	0.9589	0.9746	--	--	--	--
ANN Model Performance (R-Squared)									
Y			X						X + All Health Statistics
Life Exp			0.5208						**0.5788**
Mental Health			0.6631						--
Physical Health			0.8451						--
Good Health			0.8347						--
Poor Health			0.8799						--

Table.4-2: Feature engineering demos

- 4.1 Traditional Statistical Single Parameters Analysis and Cities Analysis

This section utilizes traditional statistical methods as the first approach to analysis parameters, cities, and relationships. Specific statistical methods including various data distribution checking, Pearson correlation analysis and single parameter's dive for its effect on life expectancy and poor mental health. Data distribution and single important parameter analysis can help us to know the base of the system; by doing city comparisons of different parameters ranking, we can have some initial important observations such as if the trend of each large parameter category is in accordance with health performances, etc; visualizing the Pearson correlation analysis of any two parameters, we can see more systematically yet directly how single factor and factor category have relation with each other. Thus, by understanding single ones thoroughly, it would be much easier for further understanding and uncovering of more complex interrelationships with more complex models.

- 4.1.1 Distribution Analysis

Distribution analysis not only detects apparent abnormalities from the original data, helps with asserting the data's generalizability after the data processing, and also helps with interpreting the results in this research.

Distribution of the health metrics, human related features, and street and traffic related features are generally unimodal and slightly skewed, however they are bumpy or show some variance in local span of their density curves, meaning the relationship behind them are by no means simple. Physical inactivity and average building height are the only bimodal features, air quality is the only trimodal feature, and most of the built environment and natural environment features have a spike at 0 with a long right tail which shows a difference with human-related features and partially explains why these single feature of built or natural environment is not having a high correlation score in later analysis – their combination pattern as a whole is possibly more important when considering them. Among all these features, building height is having multiple special symptoms: it's both bimodal, 0 spiking, and have a large span of range, a long right tail, which can help us to understand why it is one of the most important feature learned by all kinds of different data models in later analysis for life expectancy. Many of other 0 spiking and long tail features like shrub or grass cover ratio, distance to nearest hospital or pharmacy, turns out not to be essential features influencing health prediction, while building area and footprint area variance, though have some importance in later analysis, are showing an extreme zero-Inflation, which makes the area variance parameters intrinsically less interpretable.

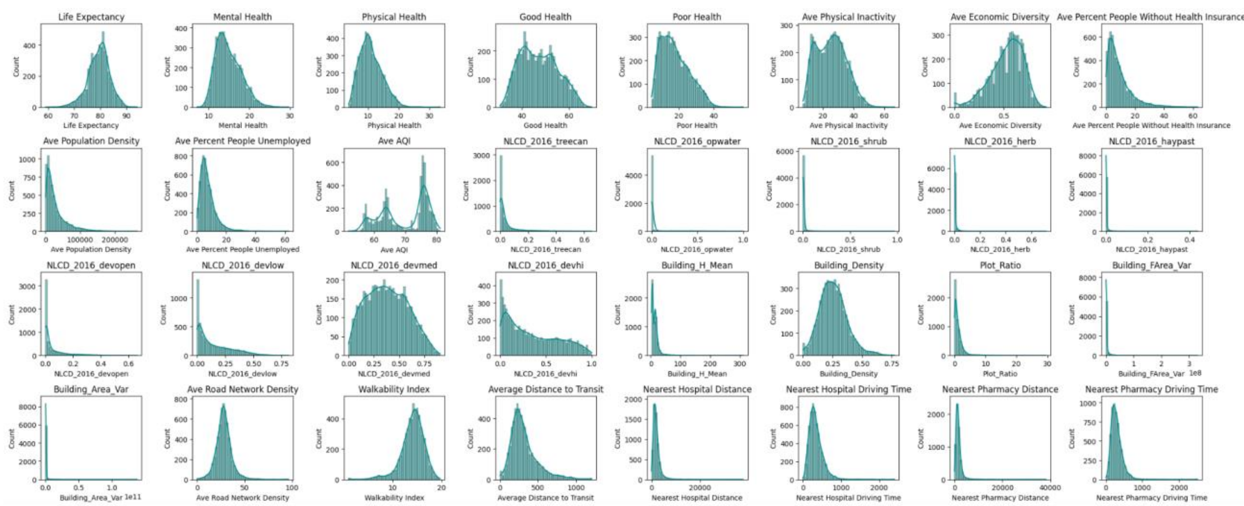


Fig.4.1.1-1: Distribution of the original data

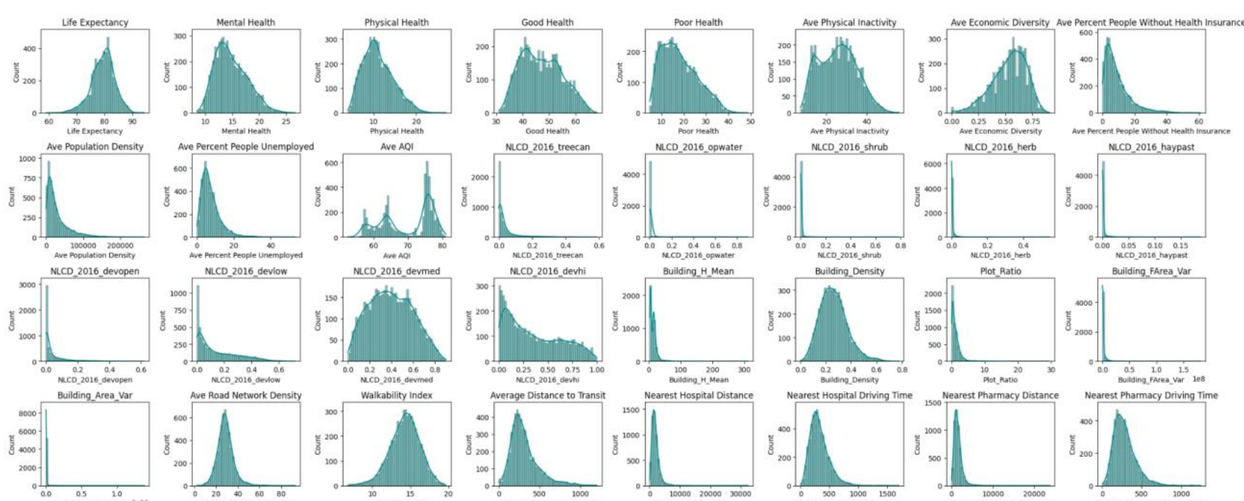


Fig.4.1.1-2: Distribution of the cleaned data

Comparison of distribution of all features show subtle difference before and after data cleaning (excluding all rows with null value), most data stay the same distribution but are having less extreme values, implying an improvement in the data's generalizability. The null value mainly exists in the life expectancy's original data, corresponding urban census tracts are large landscape or education areas that are having few people for a long term stable statistics, and these area's spatial and development pattern are also different from living areas, so excluding null values means excluding these extreme data that can increase later model's generalizability, which by common sense in accordance with the distribution comparison results.

Feature Names	Null	Feature Names	Null
GEOID	0	NLCD_2016_haypast	0
Sits in State	0	NLCD_2016_devopen	0
Life Expectancy	561	NLCD_2016_devlow	0
Mental Health	73	NLCD_2016_devmed	0
Physical Health	73	NLCD_2016_devhi	0
Good Health	64	Building_H_Mean	10
Poor Health	73	Building_Density	10
Ave Physical Inactivity	125	Plot_Ratio	10
Ave Economic Diversity	1	Building_FArea_Var	10
Ave Percent People Without Health Insurance	117	Building_Area_Var	10
Ave Population Density	4	Ave Road Network Density	1
Ave Percent People Unemployed	121	Walkability Index	1
Ave AQI	1	Average Distance to Transit	156
NLCD_2016_treecan	0	Nearest Hospital Distance	2
NLCD_2016_opwater	0	Nearest Hospital Driving Time	2
NLCD_2016_shrub	0	Nearest Pharmacy Distance	2
NLCD_2016_herb	0	Nearest Pharmacy Driving Time	2

Table.4.1.1: Null values in the original data

- ### 4.1.2 City Parameter Ranking Comparison

City ranking of features average value gives us macro inspection of the data and results. Accordance is shown between various health metrics, and between them with urban features, and therefore within general basic expectations, though no any single environmental feature can have perfectly close ranking with health metrics; such observation further validate the data's generalizability and verified that some data's variance in original time and space granularity is relatively trifle in successfully finding essential features in macro perspective, or a ten year's data collection span is able to give insights in such research, not hard to understand considering the urban environment features not always change drastically in decades.

For health metrics, cities of higher average life expectancy ranking are also of a quite close high ranking of temporary physical and mental health conditions.

For built and natural environmental indicators, it can be observed that very low average height, density, plot ratio, high-development-ratio and poorer air quality (represents by low ranks) cities – like Phoenix and Jacksonville – are also showing a low average life expectancy rank, but high average life expectancy cities like Seattle, San Francisco and Boston do not definitely have city-level's high height, density, or high-development-ratio, which asserts the complexity of the urban system and also strengthened the importance of doing such analysis in such a finer grain in this research – census tract level. Tree canopy cover ratio, as we would typically suppose that more trees means better, in this urban land and resource trade-off competition doesn't show such a direct pattern – highest tree canopy coverage cities like Jacksonville and Seattle show two poles of average life expectancy ranking, and low tree canopy coverage city San Francisco shows a relatively high life expectancy mean, which implies that tree canopy cover may not be in top priority features in the system, and these features are possibly working together to make a larger influence.

Street and transportation features do not show a significant relationship with health metrics at this city's level.

For human related features, some apparent accordance can easily be found, like high economic diversity, low physical inactivity and high population density (higher ranking) cities are generally with high average life expectancy ranking, and they've shown to possibly make this effect together since low physical inactivity alone (with low economic diversity together like Chicago and Washington DC) would not result in a high average life expectancy ranking, so do high economic diversity with a high physical inactivity (like Jacksonville).

These city level observations and comparisons reveal some opportunities in finding what are truly essential, strengthen the assertion of these data's generalizability, yet also show how far a macro interpretation can get so far, paving the way for detailed inspection in later sections.

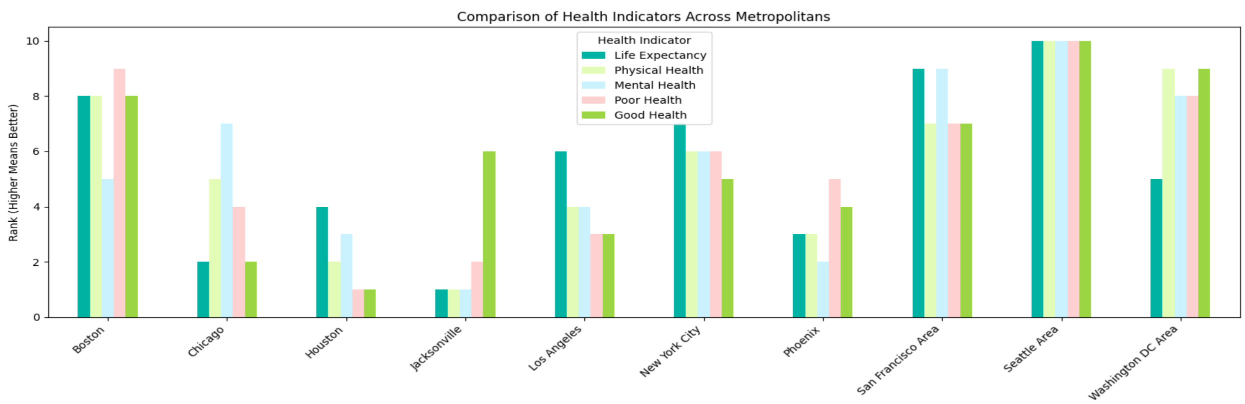


Fig.4.1.2-1: City ranking of health metrics

Feature Names	Order	Feature Names	Order
Life Expectancy	+	NLCD_2016_devopen	+
Mental Health	–	NLCD_2016_devlow	+
Physical Health	–	NLCD_2016_devmed	+
Good Health	+	NLCD_2016_devhi	+
Poor Health	–	Building_H_Mean	+
Ave Physical Inactivity	–	Building_Density	+
Ave Economic Diversity	+	Plot_Ratio	+
Ave Percent People Without Health Insurance	–	Building_FArea_Var	+
Ave Population Density	+	Building_Area_Var	+
Ave Percent People Unemployed	–	Ave Road Network Density	+
Ave AQI	+	Walkability Index	+
NLCD_2016_treecan	+	Average Distance to Transit	–
NLCD_2016_opwater	+	Nearest Hospital Distance	–
NLCD_2016_shrub	+	Nearest Hospital Driving Time	–
NLCD_2016_herb	+	Nearest Pharmacy Distance	–
NLCD_2016_haypast	+	Nearest Pharmacy Driving Time	–

Order: + means a positive ranking mechanism applied towards this parameter (larger ranking number like 10 means higher value in this parameter), - means a reverse ranking mechanism is applied (larger ranking number means lower value in this parameter); the target is to generate graph comparison of common sense that higher value means better for health for each indicator, to make the comparison easier to understand.

Table.4.1.2: Ranking mechanism engineering – positive ordered or reverse ordered

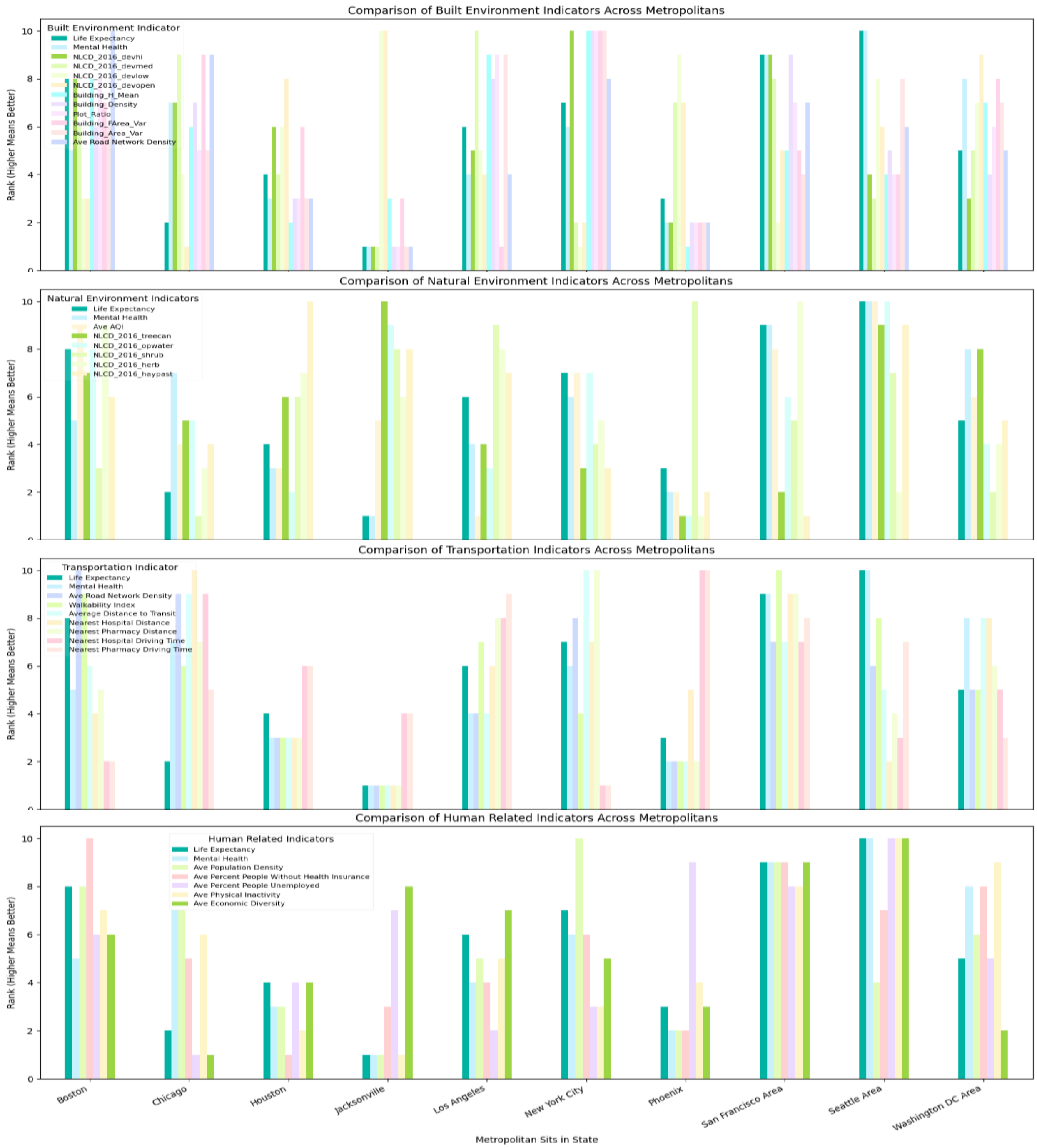


Fig.4.1.2-2: City ranking of urban environment metrics

- 4.1.3 Single Parameters Correlation

The table indicates complex and integrating relationships within the system that no single feature or category can make up the majority influence, though some are more important or work in an apparent way. Human activity, including physical and economic activity both relate to health apparently; the built and natural environment exerts influence in an intricate way, but some clearly readable relationship shows their relation on population density and human activity, thus in an integrated system process they together influence human health. Traffic and healthcare accessibility are not shown to have significant or direct relation towards health, but they are also important for maintaining the system and providing guarantee especially when health issues arise. Specific meaningful relationships observations and findings are written below.

a. Relationship within each feature category

A majority of health measurements are significantly correlated. From the Pearson correlation analysis chart below [Fig.4.1.3]: general health (poor or good health ratio) is found to most closely correspond to physical health condition, while life expectancy is found especially relevant to poor mental health ratio, still poor physical health condition is also of significant relevance to life expectancy. This can be partially understood, for health indicators except life expectancy are more of short term reflection parameters (say within 5 years) gained by self-perceptual survey, while life expectancy reflects decades of health condition; poor physical health and poor general health are found to have a Pearson value of 0.98, which implies they can almost be treated as the same feature, but when health condition turns better, like for good health, the physical health condition is not as important; mental health, while correlating with temporary health in a relatively mild way (compared to physical health, but still strong in statistics with Pearson value of 0.85 for poor and 0.61 for good health), have a longer important effect on health, thus this important correlation is reflected by another long-term parameter - life expectancy, so we'd emphasize on life expectancy and poor mental health performance more in later analysis.

Human activity indicators (population density, physical inactivity rate, economic diversity, unemployment, no health insurance rate) themselves are found to have weak yet meaningful correlation among some of them. Physical inactivity relates to low health insurance coverage, high unemployment, and a relatively low economic diversity; unemployment and economic diversity is showing a weak negative correlation, while short of health insurance coverage is not found to have apparent correlation with them. These human activities, though many are correlated, cannot be neglected singly, or substitute for each other.

Natural environment indicators (air quality index, all kinds of plant coverage, and open water ratio) within the group of themselves are not showing convincing correlation, partially due to the urban environment is still mainly influenced by human and built environment activities, thus these features are showing more relation to other urban and human features which would be discussed later.

Built environmental features also have strong inner correlations, which is also in accordance with common sense. Higher building density, height, plot ratio, and high development ratio, as well as higher building area variance and footprint area variance are found to have at least mild positive correlation among them each, but not all parameters are found to have strong correlation. Building Height correlates to plot ratio, they mildly influence building area and footprint area variance (which serves as reflections for built environment diversity). Building density's direct correlation with building height is not as apparent as with high development ratio and plot ratio, and building height's direct correlation with high development ratio is even weaker (with a Pearson value of only 0.35). Higher development ratio and density are also found in some weak accordance with higher street network density and walkability. Similarly, low development and spare space's ratio correlates, and both are in negative correlation with high density, height, variance and walkability features.

Street network, traffic, and spatial distance features also have some relationship. Nearest hospital and pharmacy distance or driving time are in some accordance, and they are all shown to have some weak relationship with street network density and walkability, which also reflects their relationship to a diverse of urban activities: relatively the higher street network density is, the more active a area is, which probably symbolizes a relatively developed community, thus with higher economic activities and development level, easier accessibility to healthcare resources like shorter distance to hospital or pharmacy. A question arises here for us: given the observations so far, however, can we say the inverse thing - can we of course build a healthier neighborhood by starting with higher network density, shorter distance to public transit, hospital, and pharmacy? The answer by best models built later in this article shows no, or street network is not of great importance found in this article given it is within a reasonable range in most cities in the US metropolitans, and the setup of health facilities are more like to deal with results, they serve as infrastructural guarantee that solves existing health problems, but can do less with improving life expectancy, improving long term health measurements like mental health condition, especially in improving people's health of medium level, say, not excellent or poor.

b. Relationship between each feature category group

Human activity vitality relates with good health in direct ways with highest Pearson scores compared to other groups, which means higher economic diversity, higher employment rate, higher health insurance coverage, and higher physical activity level relates apparently with better health measurements. Some economic features and population density also have many correlations with natural and built environment indicators. Among all these relationships, **physical inactivity** rate shows very strong synchronization with self-perceptual health conditions (good/poor health, physical or mental health), but its correlation with life expectancy is mild. Such cases are similar for health insurance dis-coverage, implying that no single feature can explain most of the system. Employment and Economic diversity are found to mildly influence life expectancy. **An interesting finding is population density**, it doesn't relate to any health statistics or any human physical or economical activities directly, but it has mild or strong relationships with all the built environment factors, and has mild relationships towards natural features (air quality and tree canopy coverage), as well as public transportation accessibility; many of these other features that population density relates to are showing some kind of correlation with life expectancy, so such phenomenon probably implies more that it's firstly better urban built and natural environment that can afford more people (reproduction) or allures more people to move in and then shapes better societal life and raises health condition, instead of population density raises first and then the built environment as well as the societal environment become better (this deduction is based on the city's currently already built condition and the finding are more meaningful for futural steps, instead of explaining thoroughly how the history forms from the very beginning).

Among Natural environment features, only air quality and tree canopy coverage has some weak relationship with health statistics, however they are showing stronger (yet still mild) correlation towards many built environment density related features. **It's interesting to find that high development area's ratio is in positive relationship with good air quality, but medium development area's ratio is in negative relationship to good air quality, and even more negative than low development or open area,** but this doesn't mean we should tear down medium development area's buildings for better air quality, instead it means comparing to build from open space to medium development place, building medium development areas into high development areas first would make more sense towards both the natural environment and human health. **Good air quality** also appears with higher population density, higher insurance coverage, better accessibility to public transportation, and slightly more water and plant coverage area. Another important topic is how can we build an excellent community given all these relationships with health, here we can discuss it a bit regarding the natural environment. Typically we would think an excellent community should be excellent in every aspects, however, not all good features can be collected as much as possible, we have to choose what are more important ones, and **natural environment features always make a concession** during the process, that's where we must set a threshold, even just from the health's perspective. **Tree canopy coverage and open water area are found in reverse trend to development density** - a higher ratio of medium to high development area appears to have lower plant and water coverage; since it's shown that tree canopy cover and open water area is having a weak but positive influence towards health from the chart, so there's still long journey for us to balance the ratio of each natural environment indicators and built environment indicators during urban development, and finding better quantified balancing pattern may need more research attention.

Built Environment features, though only building height, density and plot ratio show directly weak correlations with health, are having a variety of relationships with all other feature groups, such as their strong relationship with population density, air quality, plant coverage, open water area, public transportation and healthcare accessibility. Most of these observations and relationships are discussed above, yet another interesting point here is **considering health influence from development density with healthcare accessibility**. A higher medium-development-ratio appears to be in accordance with shorter distance and driving time towards hospital and pharmacy, but higher high-development-ratio appears to have longer driving time towards hospital and pharmacy, though the distance is no longer; such a phenomenon probably means either traffic congestion exists for high-development-area, or geographical block exists beside them (like a large central park, lake, etc.). Higher high-development-ratio, medium-development-ratio, or low-development-ratio doesn't have apparent correlation towards health, so as hospital and pharmacy availability, which means that health resource availability is not the cause of better or poor health, but more of a infrastructure gathered along with the development of urban area, and although people need them, building new hospitals or pharmacies should not be firstly considered for increasing all neighborhood residents' health condition. Traffic congestion is demonstrated as a problem by this case when development density rises, such problems also remind us higher development density is not always good for all living aspects, yet it's still showing a relatively positive relation with health.

Another parameter worth a clarification is the **walkability index** here, which is directly calculated from the average of economic diversity index, unemployment rate index, street network density index, and public transportation accessibility index, so it doesn't have much independent value or can serve as the true "walkability" from the urban designer or researchers perspective, though it shows some relationship with other groups, it's the basic 4 factors those form it that are truly making the correlations, and the correlations are usually very weak, which can be seen as a bad example to build such integrated index from some brute-force way.

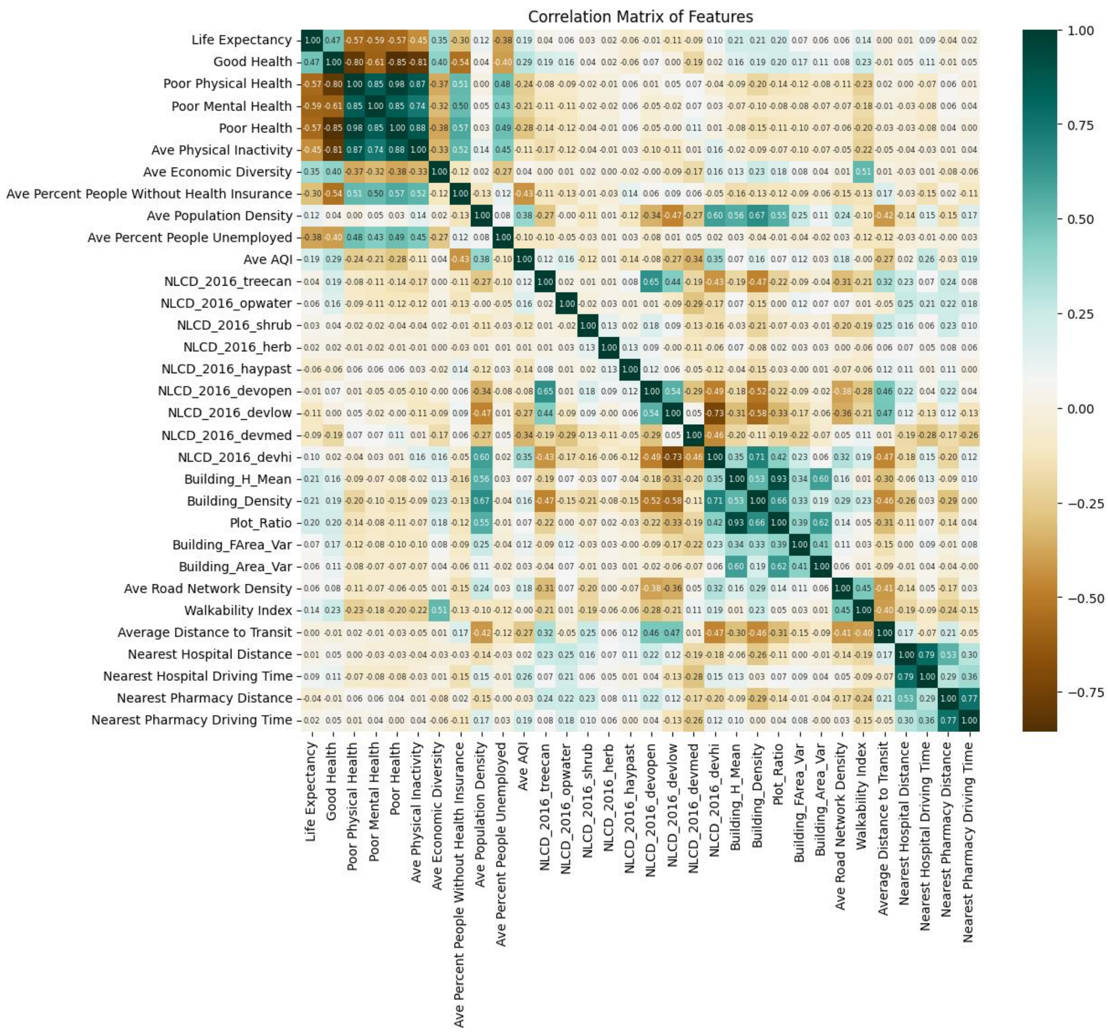
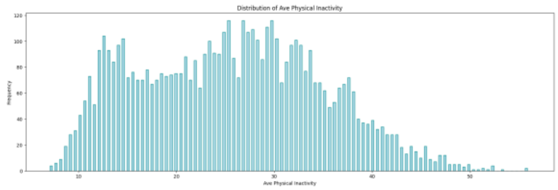


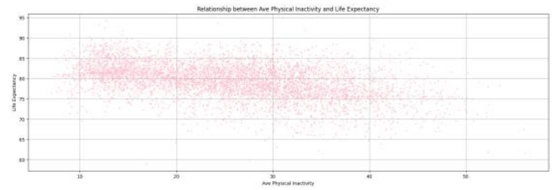
Fig.4.1.3: Pearson correlation analysis

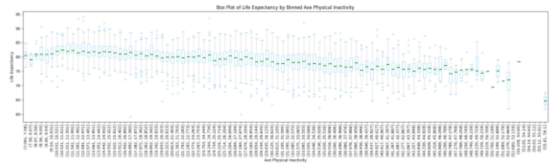
The Pearson correlation score gives us loads of significant info and also suggests a list of important single parameters that needs elaborate inspection before we dive into model analysis. The following part would give these 11 parameters more basic quantitative analysis: building height, physical inactivity, plot ratio, building density, economic diversity, tree canopy coverage, building footprint area variance, unemployment, population density, high development ratio, road network density. Through scatter plot and box plot visualizing them with life expectancy and poor mental health rate, together with their distribution graph, we would clearly know some meaningful trends.

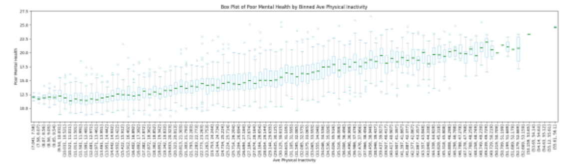
- 4.1.3.1 Physical Inactivity

A simple clear trend between physical inactivity and health measurements shows that, when physical inactivity rate increases, in combination comes the decrement of life expectancy and increment of poor mental health ratio. The only exception is when physical inactivity rate very low, say no leisure-time physical activity adults' percent is below 12 percent, then increment in physical inactivity doesn't relate to worse health.



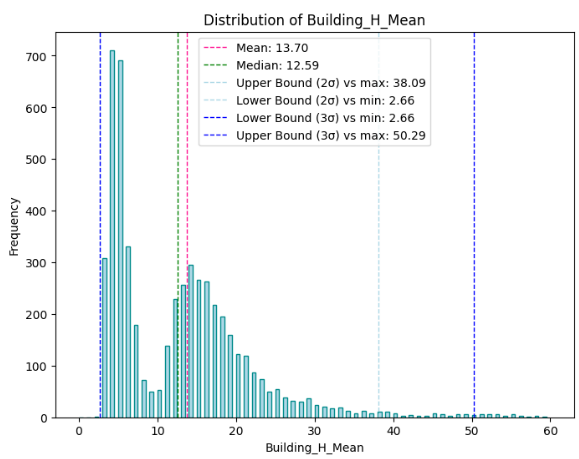


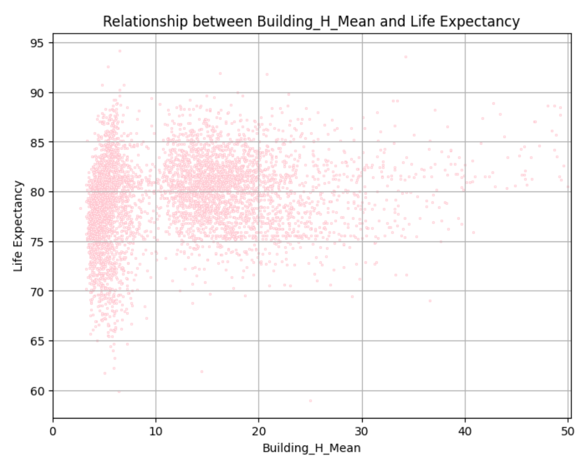


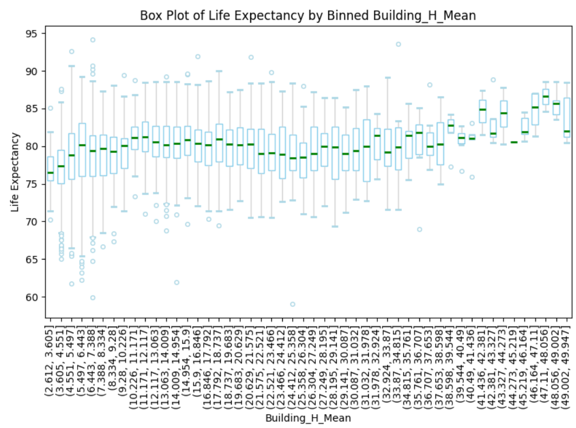


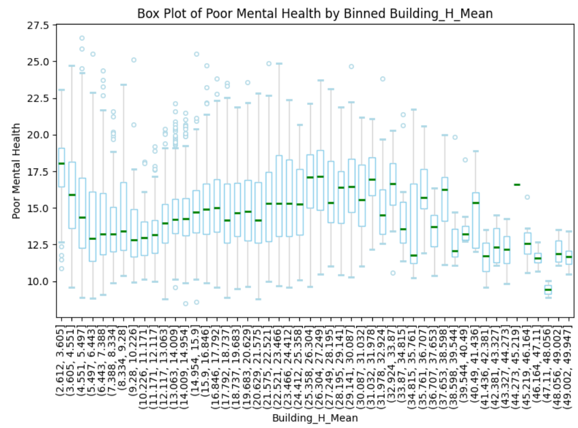
- 4.1.3.2 Building Height

Building height has 2 peaks in its distribution, one is at 2 layers height, symbolizing a majority of small residences; one is at 5 layers height, symbolizing a majority of medium sized apartments or office buildings. Through the box plots, the trend of building height varying with life expectancy and poor mental health rate is clear: for range near the first building height peak, higher average height means higher life expectancy and better mental health; then the health status keeps steadily sliding down a bit when average height keeps increasing, but mechanism for these parts are not very clear with just these analysis; when the average height reaches 25m, it arrives at a health performance valley, then the health status goes steadily up with higher average height, where the height can represent city center area with typically higher height, density, and more economic activity and diversity, thus as low life expectancy cases and poor mental health cases reduce, the health performance keeps to enhance again.



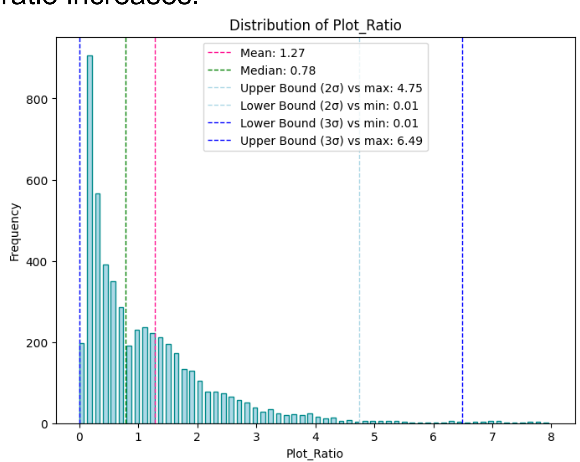


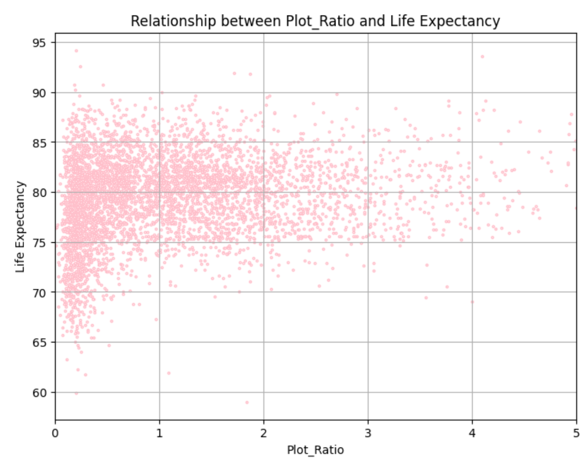


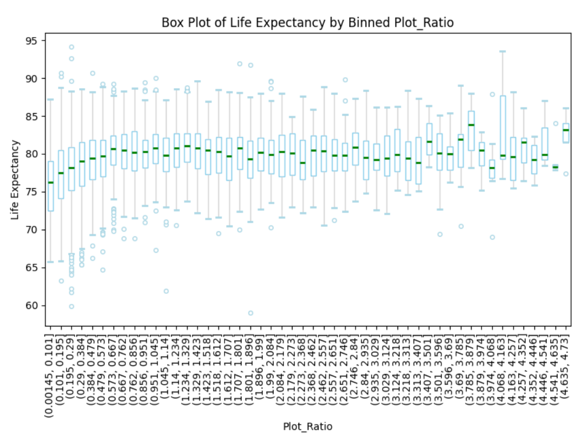


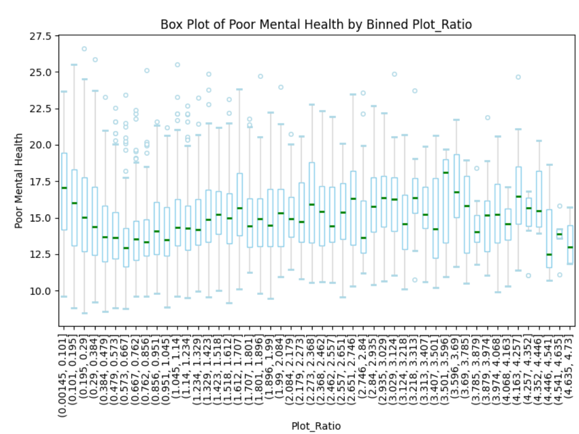
- 4.1.3.3 Plot Ratio

Plot ratio graph suggests similar trends with building height, partially because it is calculated from building height and footprint area. When plot ratio is below 0.78 (the range near the first plot ratio distribution peak), health performance keeps enhance as plot ratio increase; while greater than 0.78, life expectancy turns to keep the same value (or fluctuate in a very small range) and mental health performance goes worse in a mildly decreasing trend as plot ratio increases.



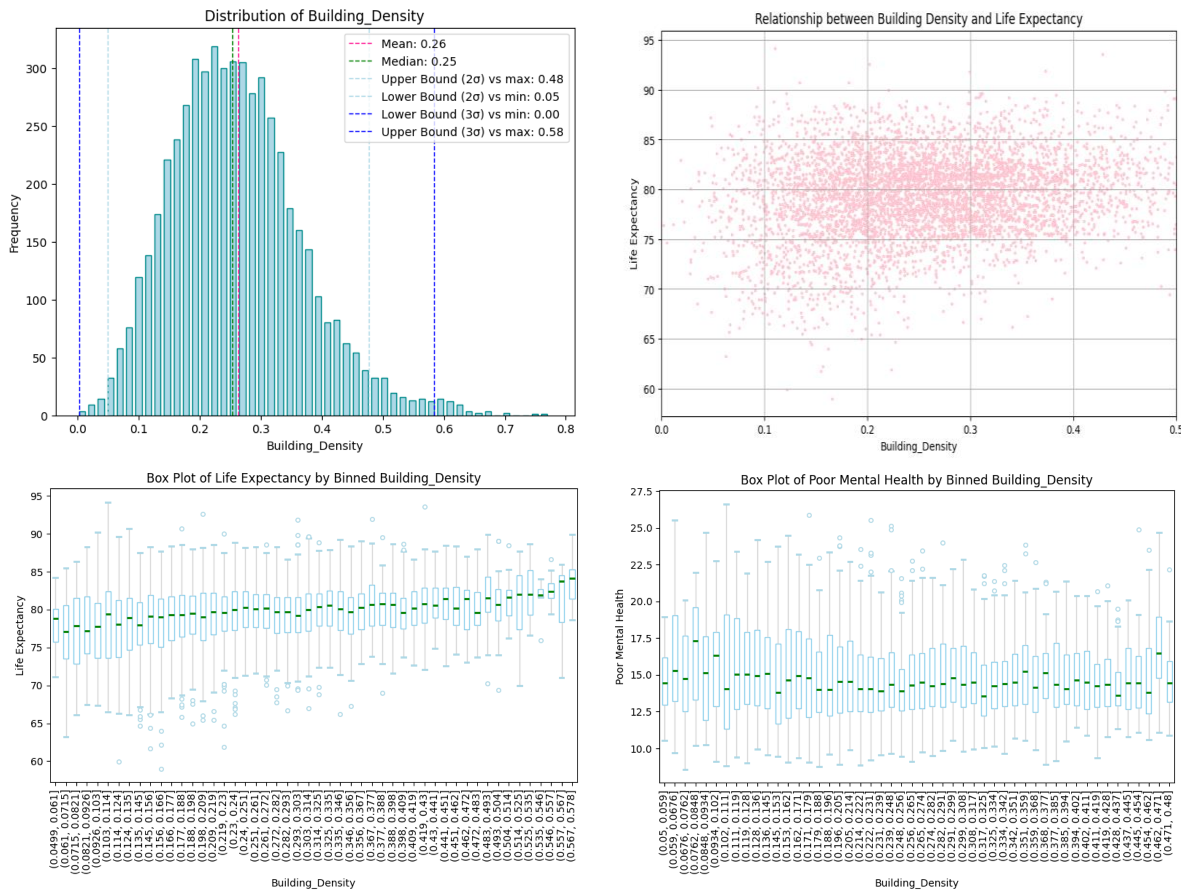






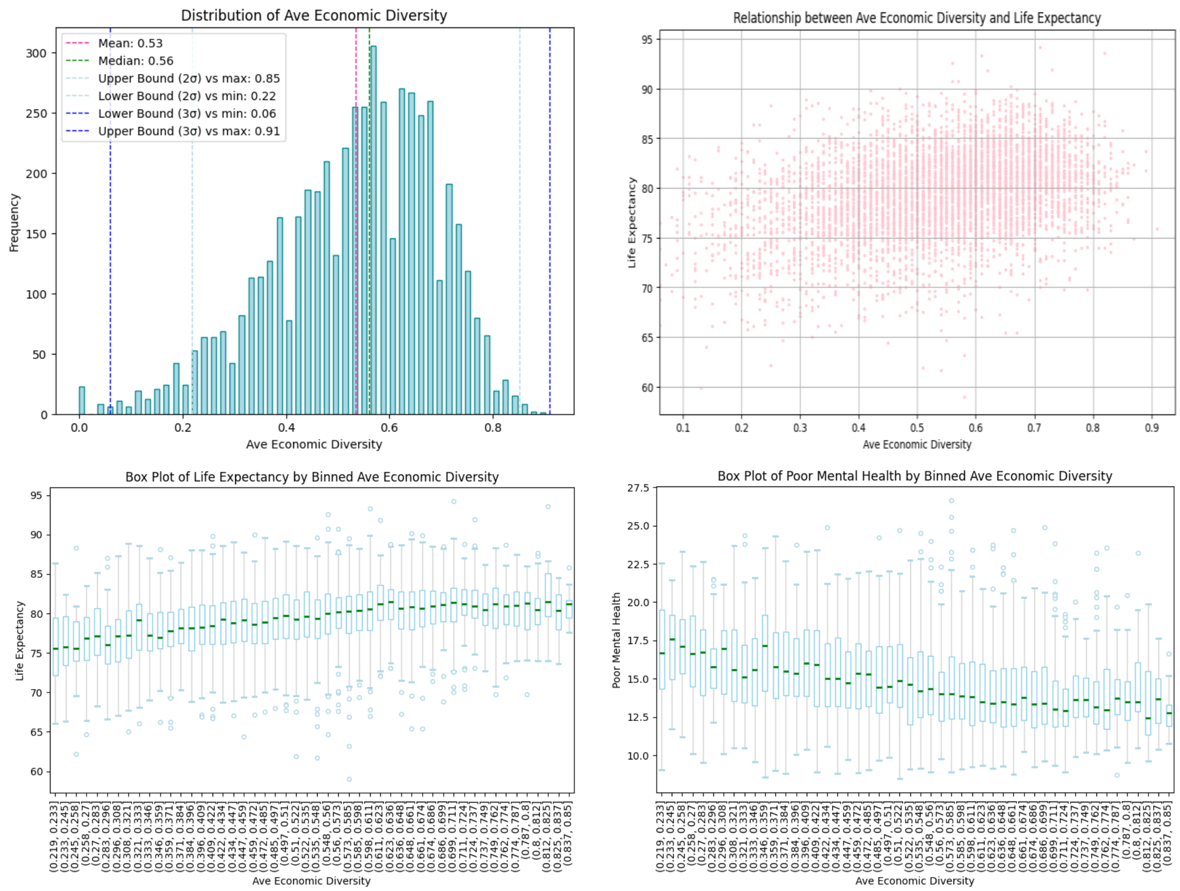
- 4.1.3.4 Building Density

Building density demonstrates only 1 peak in distribution, and its relation towards life expectancy is clear: always having a positive correlation trend, which means the density increment is always in combination with the life expectancy enhancement trend, but the overall fluctuating range is not as large as with building height. Moreover, mental health condition doesn't change apparently as building density changes, which implies that building height may be more important and can reflect more complex information regarding health, although the Pearson correlation score for height-health and density-health is almost the same.



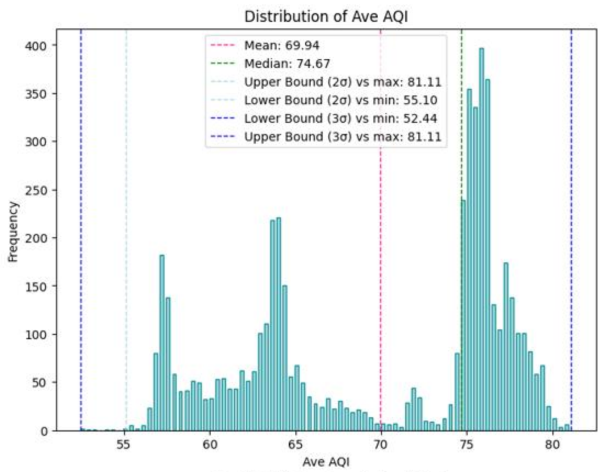
- 4.1.3.5 Average Economic Diversity

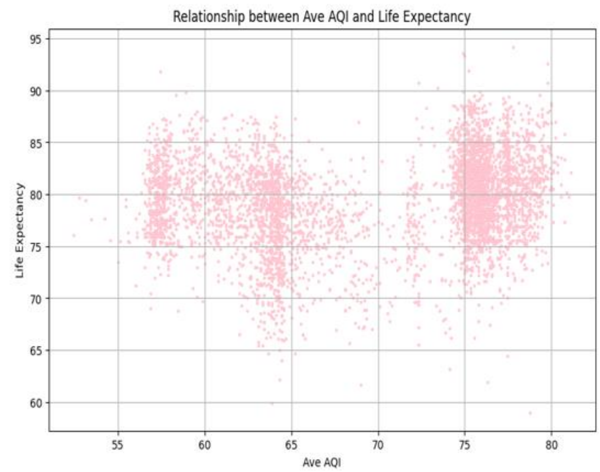
The economic diversity value is an index calculated and normalized to fit between 0 to 1, higher score means higher diversity, it only has 1 peak in distribution. Within reasonable sampling range (2 sigma's range), the correlation trend with health is also very clear. When the economic index score is below 0.64, health performance enhances mildly as the index score increases; once after the score reaches this threshold of 0.64, higher economic diversity doesn't significantly relate to health performance enhancement, but the variance of health performance also becomes smaller, which shows a more stable good health condition. At the same time, the stable peak of life expectancy reaches around 81 years old as the trend ends, which is not as high as the trend peak in building height and life expectancy, or building density and life expectancy, which also implies that economic diversity are making difference only in a limited range where economic diversity is not very high, and the influence is although significant but still limited.

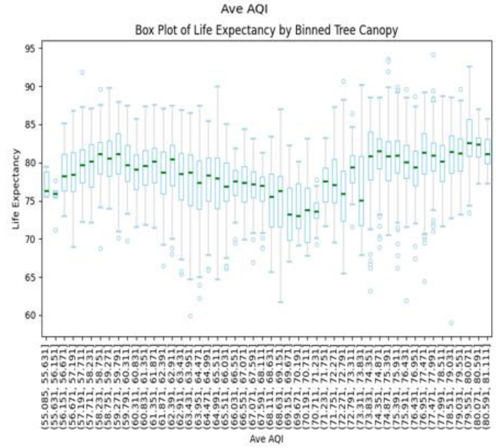


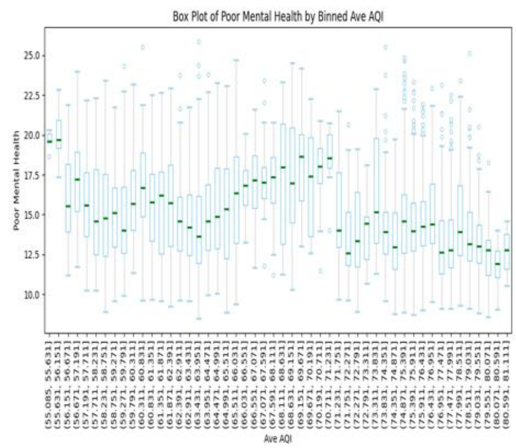
- 4.1.3.6 Air Quality

Air quality index shows a multi-modal distribution, it's apparent correlation with life expectancy and mental health is also interesting and hard to understand at first glance, like why after the beginning ascending trend there's a negative correlation with AQI and life expectancy – only after the joint analysis with other methods and models (in 4.1.1 Pearson correlation and 4.2 Partial and SHAP dependency tables and importance analysis) can this be understood: while high AQI score would have a positive effect to push up the life expectancy prediction, low AQI may generally have a dragging effect, and when AQI is lower than a threshold, a raise in its value correlates with a decrease in average building height (which can symbolize the development level and social economic vitality) thus a general weaken in many other good-for-health indicators, so the presented life expectancy at that local range drops. Such cases show many of these simple analyses can only make more sense if put into the integrated system and analyzed with more powerful tools like the MLP model and its interpretation tools.



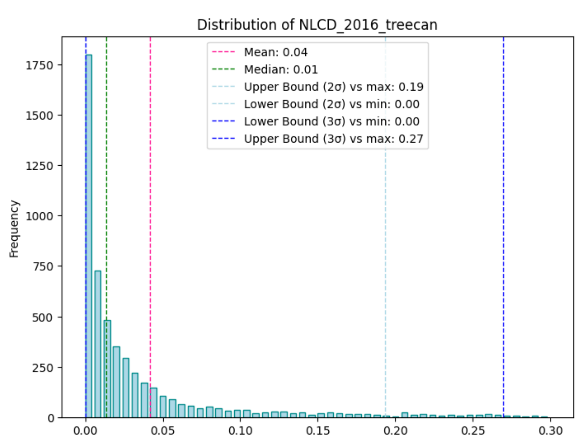


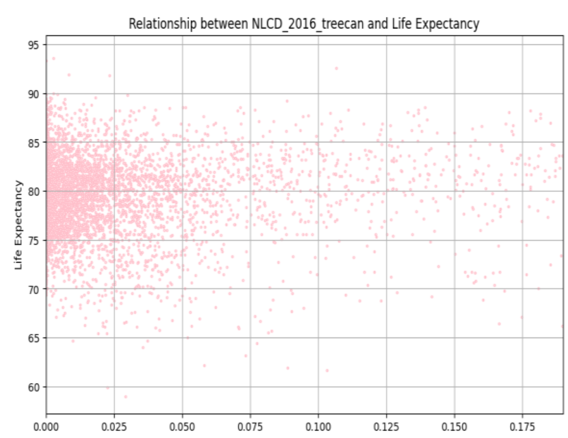


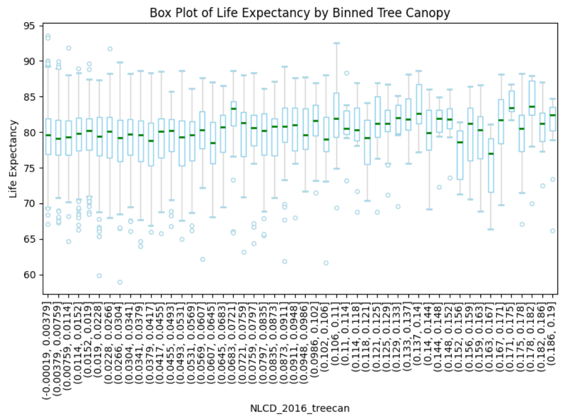


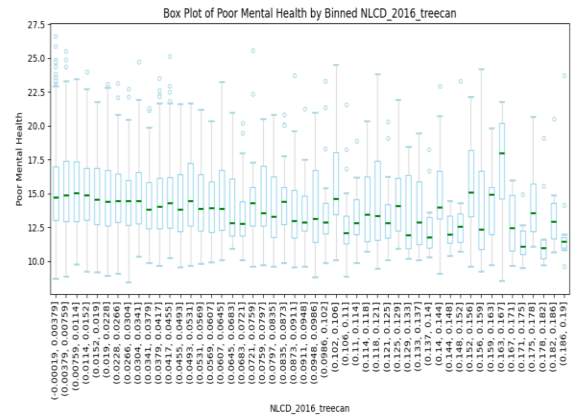
- 4.1.3.7 Tree Canopy Cover

Tree canopy's relationship trend with health indicators are more vague and having bigger variance, but through the graph below we can still see a weak positive trend between tree coverage and health condition - when tree coverage increases, life expectancy also goes sup slightly in the trend, and poor mental health rate goes down mildly in the trend.



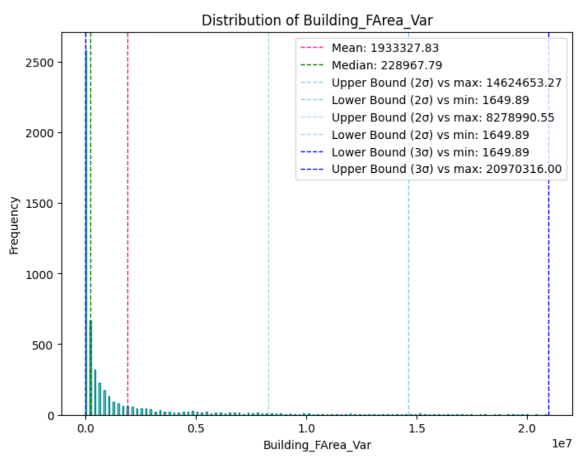


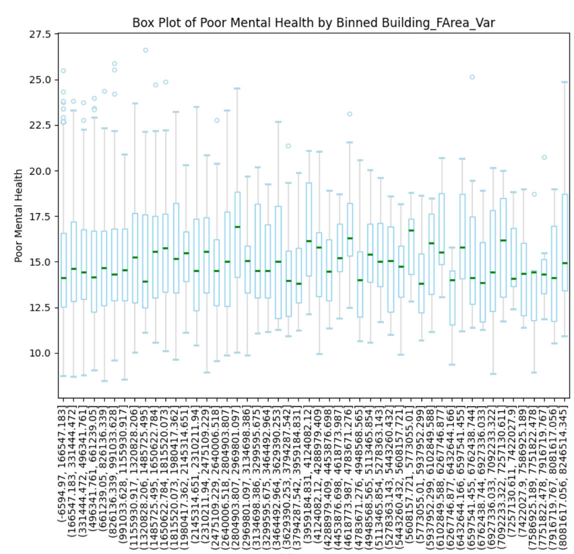




- 4.1.3.8 Building Footprint Area Variance

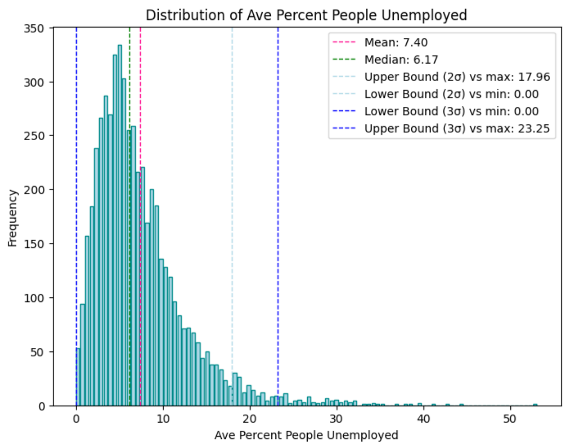
Building footprint area variance is one of the parameters that reflects the community built environments' diversity range, it shows some weak positive relationship towards population density and good health ratio, but as shown in the graph below, for life expectancy and mental health, there's too much variance within the trend and the general trend is also steady, not showing apparent relationship, so this factor may contribute in a more vague and implicit way.

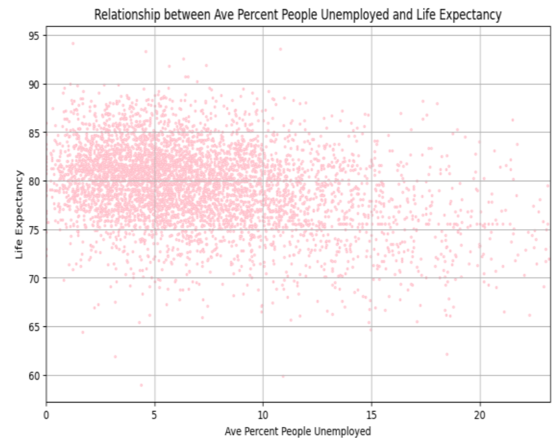




- 4.1.3.9 Unemployment Rate

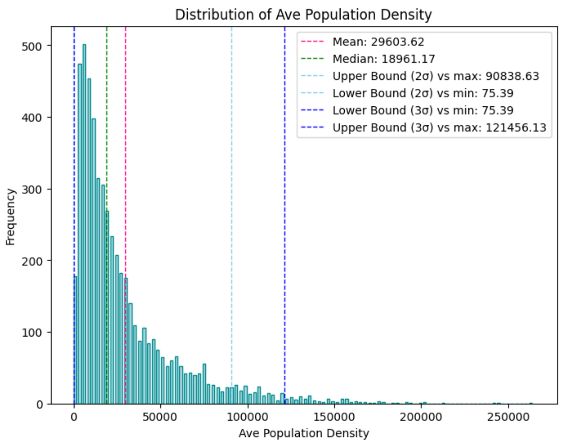
Unemployment rate increment obviously has some bad influence on health, only when unemployment rate is very low (below 2.3%) can it make little difference to communities average health condition, and typically an unemployment rate of 18% is considered a very poor condition that is worse than 95% of census tracts, which should be given enough attention early for city management staff or policy makers.

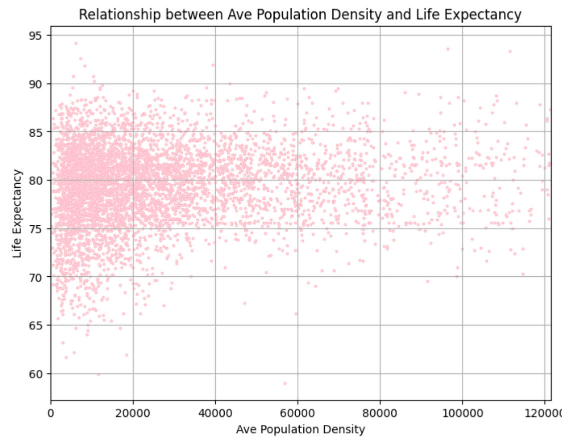


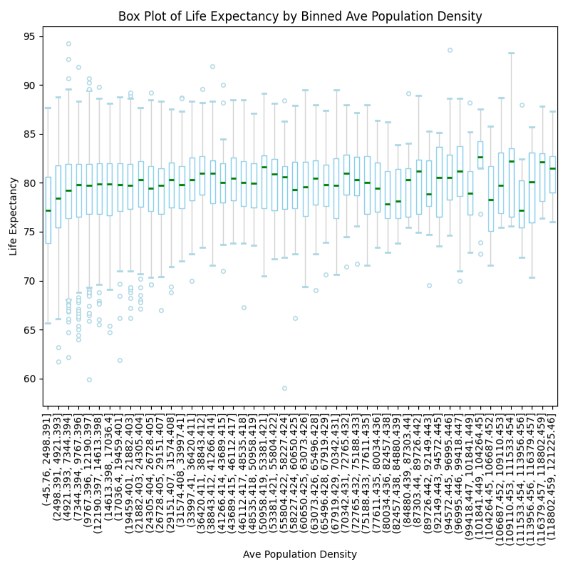


- 4.1.3.10 Population Density

Population density, as we also found in the above Pearson correlation analysis, doesn't apparently influence health; however, some points' findings also worth attention: by the graph below, when the density is very low (below 10000 people per square mile), an increment in density would significantly result in better community health condition, since the average life expectancy increase from 76 to 80 simply with the increasing of population density from 0 to 10000; when density is below 46000 people per square mile, increment in density can greatly improve the lower bound of life expectancy of that density, as is the case - from 65 to 74 years old; such finding implies that, when density is below 46000, we could encourage immigration safely (together with measures like building up the affordance of the community), simply from human health's perspective.

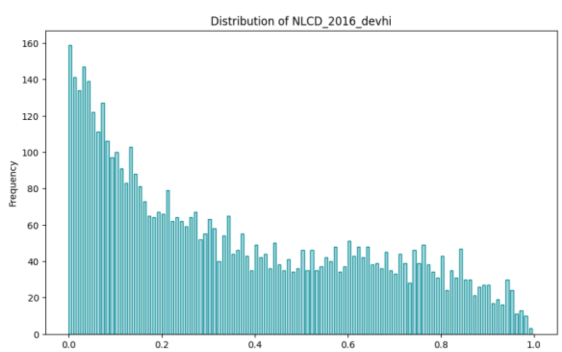


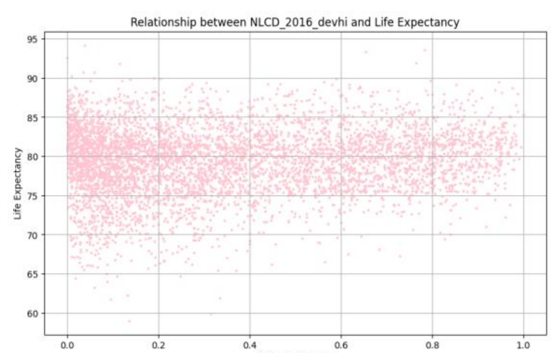


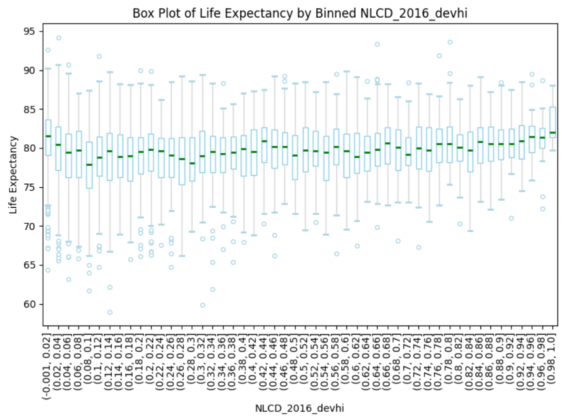


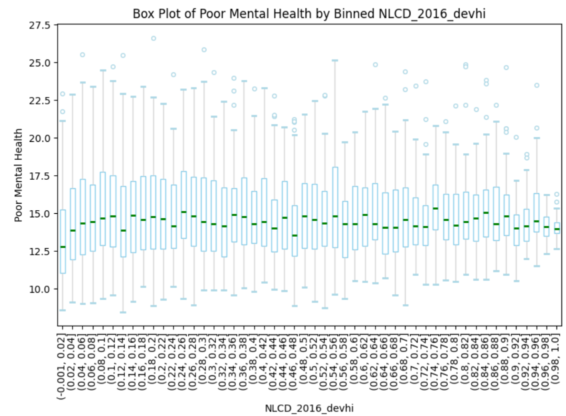
- 4.1.3.11 High Development Area Ratio

High development area ratio correlates with health mainly with the lower bounds – when high-development-ratio increases from 0.1 to 1.0, the lower bound value of life expectancy of that high-development-ratio also increases from approximately 68 to 76 years old – a great enhancement, while the upper bound value keeps the same at about 87 years.



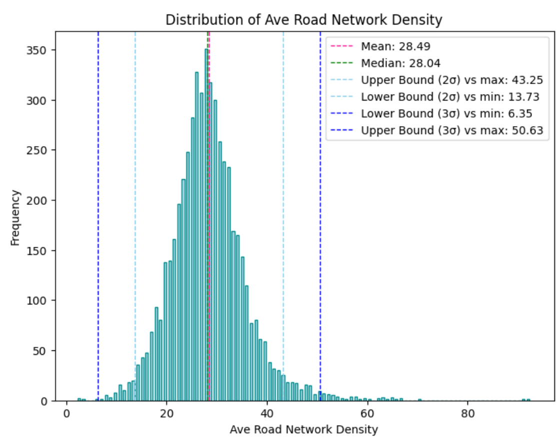


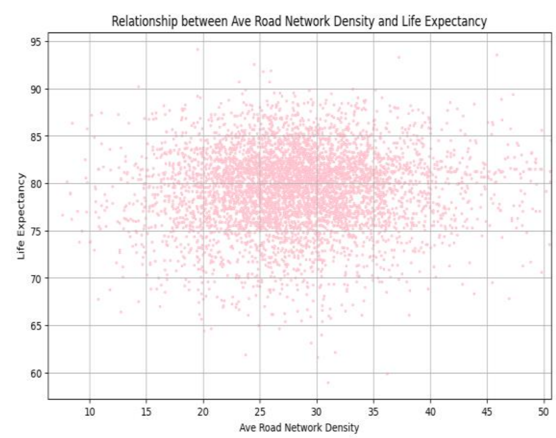


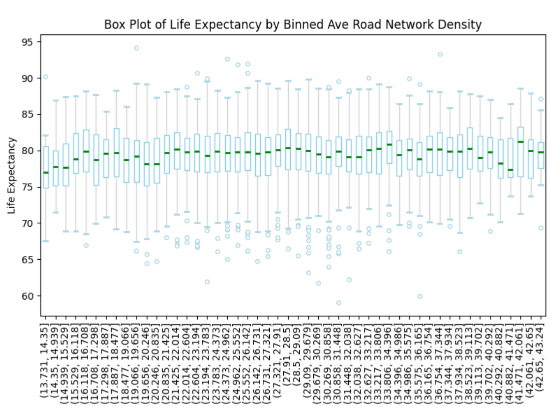


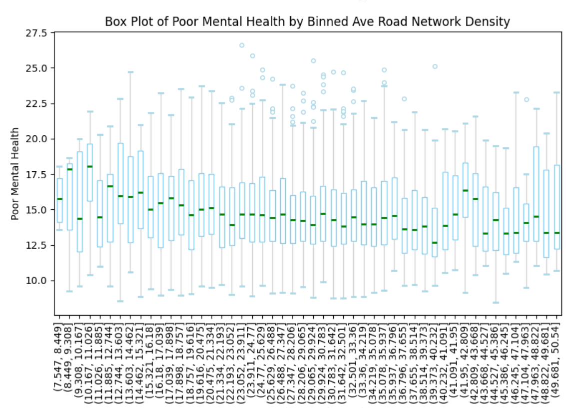
- 4.1.3.12 Road Network Density

Road network density shows a complex pattern in influencing health. For life expectancy, when network density is very low – lower than 16.7, higher density relates higher life expectancy; when density is greater than 16.7, life expectancy doe   sn't show enhancing or decreasing tendency as network density changes. As for poor mental health ratio, higher network density is shown to have a clear linear negative relationship when density is below 40, i.e. in most network density ranges (approximately a 95% range from 6.35 to 40.00), higher network density always relates to better mental health condition.









Thoroughly exploring these individual parameters shows their extremely basic and important necessity. Having known all these results, the next steps model training and comparison, parameters analysis and complex interpretation analysis as well as conclusions could be much more solid.

- **4.2 ML ANALYSIS AND INDICATOR INSIGHTS**

In the above single parameter analysis, though many parameters are showing significant correlation and interpretable trend with health indicators, each trend is also with high local variance, and all findings regarding them each also demonstrate that they only work in a limited range, as well as the majority of low Pearson values meaning single parameter's weakness in explanation the whole system, thus suggesting the integration of other parameters can exploit the data better and leading to more integrated results of systematic explanation meaning.

This section tries to find out how well machine learning models can do for systematic analysis and explanation, find the best performance models, and get realistic interpretations on indicators importance, interrelationships, influence on health and exerting influence mechanisms, and possible causation relationship inferences. A bunch of machine learning models are applied to explore the hidden mechanisms within the data, models include OLS Baseline Linear Regression, Bayesian Ridge Linear Regression with CV (k-folds cross validation), Polynomial ElasticNet Regression with CV, SVR with RBF kernel, Decision Tree, Random Forest, SLFN, Multiple Hidden Layer ANN. Among them, the Random Forest model sometimes performs best, but always has the problem of overfitting and the model tends to intrinsically overestimated few single parameter's importance; 2 Hidden Layer ANN model is an excellent choice with both high performance (and sometimes best performance) and good interpretability without the problem of overfitting, thus it's interpretation might be more meaningful and closer to reality; other models like Polynomial Regression can also give some insightful explanation supplementations, and Linear Regression models, as a baseline model, can demonstrate how well the trained machine learning models can do compare to ordinary statistical methods, also sets up the context of how easy the result analysis can reach a high performance, making all models mentioned here worth comparing.

- 4.2.1 All Model Results and General Comparisons

In social science, a value of R^2 above 0.5 could be considered acceptable, given the model is theoretically sound and results are statistically significant. In our model comparison, if a machine learning model is of R^2 above 0.5 that can hardly enhance again, and shows a significant advantage compared to the baseline linear model, we could say it has exploited the potential of the training method and the dataset and is therefore meaningful.

In the previously showed overall comparisons table [4-1], some very high performance are achieved by some model, while the models that are considered with best interpretability and reality meanings are models for predicting Life expectancy and Mental health since they show a great enhancement from the baseline linear model and make monumental statistical breakthrough to integrate the urban and human related features effects into them as the R^2 performances reach their competitive peak.

Model Type	*Model Name*	*Metric*	*Life Expectancy*	*Mental Health*	*Physical Health*	*Poor Health*	*Good Health*
Linear Regression	OLS Baseline	R^2	0.3675	0.5494	0.8059	0.8374	0.7421
		RMSE	3.305	1.9946	1.5416	3.1812	4.0164
	Bayesian Ridge with CV	R^2	0.3873	0.5641	0.8085	0.8403	0.7471
		RMSE	3.253	0.9617	1.5314	3.1530	3.9776

Polynomial Regression	Polynomial ElasticNet with CV	R²	0.4402	0.5488	0.8095	0.8424	0.7562
		RMSE	101.5	67.834	51.913	106.44	132.72
Support Vector Regression	SVR with RBF Kernel	R²	**0.5025**	**0.6494**	**0.8209**	**0.8572**	**0.8091**
		RMSE	2.931	1.7595	1.4811	2.9812	3.4556
Tree	Decision Tree	R²	0.4093	0.5633	0.7997	0.8440	0.7644
		RMSE	3.194	1.9636	1.5660	3.1161	3.8391
	Random Forest	R²	**0.5449**	**0.7084**	**0.8502**	**0.8847**	**0.8373**
		RMSE	2.804	1.6044	1.3545	2.6781	3.1908
ANN	SLFN	R²	**0.5161**	**0.6467**	**0.8369**	**0.8717**	**0.8247**
		RMSE	2.895	1.7662	1.4205	2.8258	3.3114
	2 Hidden Layers	R²	**0.5208**	**0.6631**	**0.8451**	**0.8799**	**0.8347**
		RMSE	2.877	1.7246	1.3773	2.7334	3.2155

Table.4-1: ML Model Results, Performance and Parameters

For life expectancy, through the model training process, test R^2 performance improved from 0.3675 (Linear baseline model) to 0.5449 (Random Forest Model); the best model - ANN model reaches a test R^2 of 0.5208, and SVR model reaches a test R^2 of 0.5025, both of them achieves an acceptable performance, and 2nd degree polynomial model with test R^2 of 0.4402 (significant higher than linear model) here is also strong in demonstrating multifactor interweaved relationship, all of them can help us to understand more thoroughly how the urban features work on human health.

The best test R^2 performance model Random Forest shows an intrinsic problem - overfitting, with a training R^2 of 0.9028 and a test R^2 of only 0.5449. This overfitting and its intrinsic training mechanism together impair its interpretation's reality meaning. Because it utilizes a batch of decision tree models, each decision tree model is having high variance and tends to emphasize singly one or two most important features and neglect others, therefore the final random forest model also tends to have some high variance on few features and evenly treat many other less important features since they are impaired to 0 in trees (so weight calculation kind of losing efficacy for them during the RF model training), making the importance analysis goes less real. Such a case would be explained with more graphs later in the following section.

ML Model Results, Performance and Parameters (For Life Expectancy)

Model Type	Model Name	Metric	Train	Test	Hyperparameters
Linear Regression	OLS Baseline	R²	0.3576	0.3675	All by default
		RMSE	3.341	3.305	
	Bayesian Ridge with CV	R²	0.3451	0.3873	cv = 10, max_iter=1000, tol=0.001, alpha_1=1e-6, alpha_2=1e+6, lambda_1=1e-6, lambda_2=1e+6
		RMSE	3.359	3.253	
Polynomial Regression	Polynomial ElasticNet with CV	R²	0.4122	0.4402	PolynomialFeatures: degree = 2 ElasticNetCV: l1_ratio=0.04, cv=6, alphas = np.logspace(-10, 4, 15), tol=1, max_iter=10000
		RMSE	208.6	101.5	
Support Vector Regression	SVR with RBF Kernel	R²	0.5711	0.5025	kernel = 'rbf', C = 5, epsilon = 1, CV = 5, gamma = 'scale'
		RMSE	2.730	2.931	
Tree	Decision Tree	R²	0.4410	0.4093	max_depth = 5
		RMSE	3.116	3.194	
	Random Forest	R²	0.9026	0.5449	max_depth=15, max_features=1.0, max_leaf_nodes=1024, n_estimators=1000
		RMSE	1.301	2.804	

ANN	SLFN	R^2	0.5203	0.5161	Activation = 'sigmoid', Units = '1024', model.compile(optimizer='adam', loss='mse', metrics=['mae']) epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping] Early stopping: monitor='val_loss', patience=100
		RMSE	2.840	2.895	
	2 Hidden Layers	R^2	0.5506	0.5208	layer1: units=128, activation='sigmoid'; layer2: units=96, activation='tanh'; optimizer='adam', loss='mse', metrics=['mae'] early_stopping: monitor='val_loss', patience=200 epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping]
		RMSE	2.794	2.877	

Table 4.2.1-1

For mental health, through the model training process, test R^2 performance improved from 0.5494 (Linear baseline model) to 0.7084 (Random Forest Model); 2 hidden layers ANN models is also considered excellent model reaching a test R^2 of 0.6631, higher than SVR (with R^2 of 0.6494) and also with lower overfitting error. Polynomial models, sadly, show a worse performance than baseline model, thus cannot be referred for interweaving relationship interpretation, and on the other hand show the robustness of ANN and Random forest models.

ML Model Results, Performance and Parameters (For Poor Mental Health)

Model Type	Model Name	Metric	Train	Test	Hyperparameters
Linear Regression	OLS Baseline	R^2	0.5785	0.5494	All by default
		RMSE	1.9895	1.9946	
	Bayesian Ridge with CV	R^2	0.5751	0.5641	cv = 10, max_iter=1000, tol=0.001, alpha_1=1e-6, alpha_2=1e+6, lambda_1=1e-6, lambda_2=1e+6
		RMSE	1.9855	0.9617	
Polynomial Regression	Polynomial ElasticNet with CV	R^2	0.6122	0.5488	PolynomialFeatures: degree = 2 ElasticNetCV: l1_ratio=0.03, cv=5, alphas = np.logspace(-10, 4, 15), tol=1, max_iter=10000
		RMSE	129.70	67.834	
Support Vector Regression	SVR with RBF Kernel	R^2	0.7999	0.6494	kernel = 'rbf', C = 10, epsilon = 1, CV = 5, gamma = 'scale'
		RMSE	1.3708	1.7595	
Tree	Decision Tree	R^2	0.6388	0.5633	max_depth = 5
		RMSE	1.8416	1.9636	
	Random Forest	R^2	0.9482	0.7084	max_depth=15, max_features=1.0, max_leaf_nodes=1024, n_estimators=1000
		RMSE	0.6973	1.6044	
ANN	SLFN	R^2	0.7425	0.6467	Activation = 'sigmoid', Units = '1024', model.compile(optimizer='adam', loss='mse', metrics=['mae']) epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping] Early stopping: monitor='val_loss', patience=100
		RMSE	1.5549	1.7662	
	2 Hidden Layers	R^2	0.7601	0.6631	layer1: units=1024, activation='sigmoid'; layer2: units=128, activation='sigmoid'; optimizer='adam', loss='mse', metrics=['mae']
		RMSE	1.5007	1.7246	

					early_stopping: monitor='val_loss', patience=1000 epochs=5000, batch_size=200, validation_split=0.2, callbacks=[early_stopping] early_stopping: monitor='val_loss', patience=200 epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping]

Table 4.2.1-2

Another result detail needing clarification is, for Good Health, Poor Health, and Poor Physical Health model fitting, the OLS Baseline Linear Regression model already shows some pretty high performance of R^2 above 0.74, but with joint reference to table [4.2.1-5, 4-1, 4-2], we know this is because a single parameter Physical Inactivity Ratio itself contributes to this high performance, yet with ANN models the performance can still be enhance a little bit - for good health prediction R^2 improved from 0.7421 to 0.8347, for poor physical health prediction R^2 improved from 0.8059 to 0.8451, for poor health prediction R^2 improved from 0.8374 to 8799, which could be pretty hard especially the baseline R^2 value is already very high, so it still demonstrates the great meaning of all other urban and human related features.

ML Model Results, Performance and Parameters (For Poor Physical Health)

Model Type	Model Name	Metric	Train	Test	Hyperparameters
Linear Regression	OLS Baseline	R^2	0.8120	0.8059	All by default
		RMSE	1.5831	1.5416	
	Bayesian Ridge with CV	R^2	0.8114	0.8085	cv = 10, max_iter=1000, tol=0.001, alpha_1=1e-6, alpha_2=1e+6, lambda_1=1e-6, lambda_2=1e+6
		RMSE	1.5730	1.5314	
Polynomial Regression	Polynomial ElasticNet with CV	R^2	0.8219	0.8095	PolynomialFeatures: degree = 2 ElasticNetCV: l1_ratio=0.04, cv=6, alphas = np.logspace(-10, 4, 15), tol=1, max_iter=10000
		RMSE	104.74	51.913	
Support Vector Regression	SVR with RBF Kernel	R^2	0.8903	0.8209	kernel = 'rbf', C = 5, epsilon = 0.5, CV = 5, gamma = 'scale'
		RMSE	1.2091	1.4811	
Tree	Decision Tree	R^2	0.8240	0.7997	max_depth = 5
		RMSE	1.5316	1.5660	
	Random Forest	R^2	0.9730	0.8502	max_depth=15, max_features=1.0, max_leaf_nodes=1024, n_estimators=1000
		RMSE	0.5998	1.3545	
ANN	SLFN	R^2	0.8687	0.8369	Activation = 'sigmoid', Units = '1024', model.compile(optimizer='adam', loss='mse', metrics=['mae']) epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping] Early stopping: monitor='val_loss', patience=100
		RMSE	1.2875	1.4205	
	2 Hidden Layers	R^2	0.8714	0.8451	layer1: units=1024, activation='sigmoid'; layer2: units=96, activation=relu'; optimizer='adam', loss='mse', metrics=['mae'] early_stopping: monitor='val_loss', patience=200
		RMSE	1.3091	1.3773	

					epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping]

Table 4.2.1-3

ML Model Results, Performance and Parameters (For Good Health)

Model Type	Model Name	Metric	Train	Test	Hyperparameters
Linear Regression	OLS Baseline	R^2	0.7416	0.7421	All by default
		RMSE	3.9788	4.0164	
	Bayesian Ridge with CV	R^2	0.7427	0.7471	cv = 10, max_iter=1000, tol=0.001, alpha_1=1e-6, alpha_2=1e+6, lambda_1=1e-6, lambda_2=1e+6
		RMSE	3.9799	3.9776	
Polynomial Regression	Polynomial ElasticNet with CV	R^2	0.7664	0.7562	PolynomialFeatures: degree = 2 ElasticNetCV: l1_ratio=0.04, cv=6, alphas = np.logspace(-10, 4, 15), tol=1, max_iter=10000
		RMSE	257.10	132.72	
Support Vector Regression	SVR with RBF Kernel	R^2	0.8977	0.8091	kernel = 'rbf', C = 20, epsilon = 1, CV = 5, gamma = 'scale'
		RMSE	2.5029	3.4556	
Tree	Decision Tree	R^2	0.7906	0.7644	max_depth = 5
		RMSE	3.5817	3.8391	
	Random Forest	R^2	0.9727	0.8373	max_depth=15, max_features=1.0, max_leaf_nodes=1024, n_estimators=1000
		RMSE	1.2927	3.1908	
ANN	SLFN	R^2	0.8686	0.8247	Activation = 'sigmoid', Units = '1024', model.compile(optimizer='adam', loss='mse', metrics=['mae']) epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping] Early stopping: monitor='val_loss', patience=500
		RMSE	2.8364	3.3114	
	2 Hidden Layers	R^2	0.8907	0.8347	layer1: units=1024, activation='sigmoid'; layer2: units=128, activation='sigmoid'; optimizer=tf.keras.optimizers.Adam(learning_rate=0.001), loss='mse', metrics=['mae'] early_stopping: monitor='val_loss', patience=1000 epochs=5000, batch_size=200, validation_split=0.2, callbacks=[early_stopping]
		RMSE	2.5867	3.2155	

Table 4.2.1-4

ML Model Results, Performance and Parameters (For Poor Health)

Model Type	Model Name	Metric	Train	Test	Hyperparameters
Linear Regression	OLS Baseline	R^2	0.8455	0.8374	All by default
		RMSE	3.1888	3.1812	
	Bayesian Ridge with CV	R^2	0.8445	0.8403	cv = 10, max_iter=1000, tol=0.001, alpha_1=1e-6, alpha_2=1e+6, lambda_1=1e-6, lambda_2=1e+6
		RMSE	3.1824	3.1530	
Polynomial Regression	Polynomial ElasticNet with CV	R^2	0.8518	0.8424	PolynomialFeatures: degree = 2 ElasticNetCV: l1_ratio=0.04, cv=6, alphas = np.logspace(-10, 4, 15), tol=1, max_iter=10000
		RMSE	212.29	106.44	

Support Vector Regression	SVR with RBF Kernel	R^2	0.9293	0.8572	kernel = 'rbf', C = 10, epsilon = 1, CV = 5, gamma = 'scale'
		RMSE	2.1570	2.9812	
Tree	Decision Tree	R^2	0.8602	0.8440	max_depth = 5
		RMSE	3.0331	3.1161	
	Random Forest	R^2	0.9806	0.8847	max_depth=15, max_features=1.0, max_leaf_nodes=1024, n_estimators=1000
		RMSE	1.1278	2.6781	
ANN	SLFN	R^2	0.9110	0.8717	Activation = 'sigmoid', Units = '1024', model.compile(optimizer='adam', loss='mse', metrics=['mae']) epochs=1000, batch_size=200, validation_split=0.2, callbacks=[early_stopping] Early stopping: monitor='val_loss', patience=500
		RMSE	2.4195	2.8258	
	2 Hidden Layers	R^2	0.9074	0.8799	layer1: units=1024, activation='sigmoid'; layer2-10: units=32, activation=LeakyReLU(alpha=0.1); optimizer=tf.keras.optimizers.Adam(learning_rate=0.0005, beta_1=0.9, beta_2=0.999, epsilon=1e-06), loss='mse', metrics=['mae'] early_stopping: monitor='val_loss', patience=100 epochs=5000, batch_size=200, validation_split=0.2, callbacks=[early_stopping]
		RMSE	2.4685	2.7334	

Table 4.2.1-5

Linear Model Performance (R-squared, Amplified Data)									
Y	X - Physical Inactivity	Physical Inactivity	X	Physical + Mental Health Statistics	X + Physical & Mental Health Statistics	Good + Poor Health Statistics	X + Good & Poor Health Statistics	All Health Statistics	X + All Health Statistics
Life Exp	0.3270	0.2275	**0.3675**	0.4025	**0.466**	0.364	0.4357	0.4125	**0.4755**
Mental Health	0.4480	0.4965	0.5494	--	--	0.7891	0.8056	--	--
Physical Health	0.5350	**0.7567**	0.8059	--	--	0.9632	0.9748	--	--
Good Health	0.5644	0.6609	0.7422	0.6742	0.7966	--	--	--	--
Poor Health	0.6025	**0.7854**	0.8374	0.9589	0.9746	--	--	--	--

ANN Model Performance (R-Squared)		
Y	X	X + All Health Statistics
Life Exp	0.5208	**0.5788**
Mental Health	0.6631	--
Physical Health	0.8451	--
Good Health	0.8347	--
Poor Health	0.8799	--

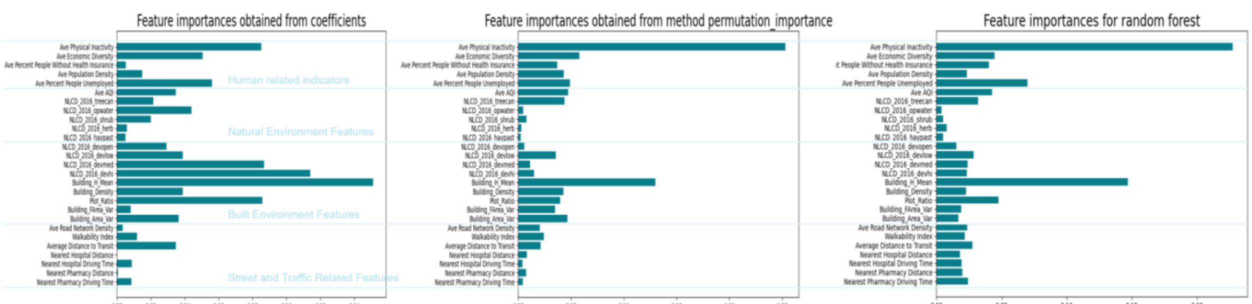
Table.4-2: Feature engineering demos (previously showed)

- 4.2.2 Specific Model Result Explanation

The following tables and models are all for life expectancy’s interpretation.

Before interpreting the feature’s importance in detail, let’s find out what model is most worthy to interpret. Below is the feature importance graph by typical models - Linear Regression baseline model, MLP, and RF. Linear model attributes importance improperly, putting too much weight on some built environment features like “Building Height”, being insensitive to some important features like ”physical inactivity”, and attributing those with some collinearity improperly, exemplified by “High Development Ratio”, “Medium Development Ratio”, and “Low Development Ratio”. The other two charts (MLP and RF feature importance) with high R^2 look similar, so both these right two importance lists are valuable.

Both MLP and RF feature importance find most important features like “physical inactivity”, “building height”, and they both find some subtle natural environment features as least important, such as “open water ratio”, “shrub ratio”. The RF model attributes more importance towards human and economic related factors in general, which is in the seemingly accordance with our previous Reason correlation analysis, however it attributes much importance on some second important factors that are not found by MLP and Pearson analysis like “unemployment” and “plot ratio”, and treats all other built environment, street and traffic parameters evenly, which might be far from the reality; Moreover, RF’s high overfitting problem also reduces the model’s robustness and generalization ability, we tend to believe it cannot generalize better than MLP with only a slightly higher R^2 value, and some important relationships or insights are muted by RF; MLP’s (2 Hidden Layers ANN model) feature importance emphasize both human related features and built environment features, and can catch accurately some second important features like Economic diversity while attributing enough importance on some implicit but important features like population density and building area variance, therefore MLP’s interpretation might be closer to reality and more meaningful.



Baseline linear model – Under Fitting MLP (2 Hidden Layers ANN) Random Forest - Over Fitting

Fig.4.2.2-1: Typical Model Feature Importance Comparison

MLP, the best performance model, for life expectancy prediction it’s found that 2 hidden layers with the below structure beats best. The first layer is 128 units using sigmoid activation, the second layer is 96 units using tanh activation function. Blue lines represent positive weights, red lines are negative weights. Such complexity shown visually helps us to understand why simply statistical models are not doing a good job regarding complex and non-linear relationships like exploring the convoluted urban attributes effect on human health.

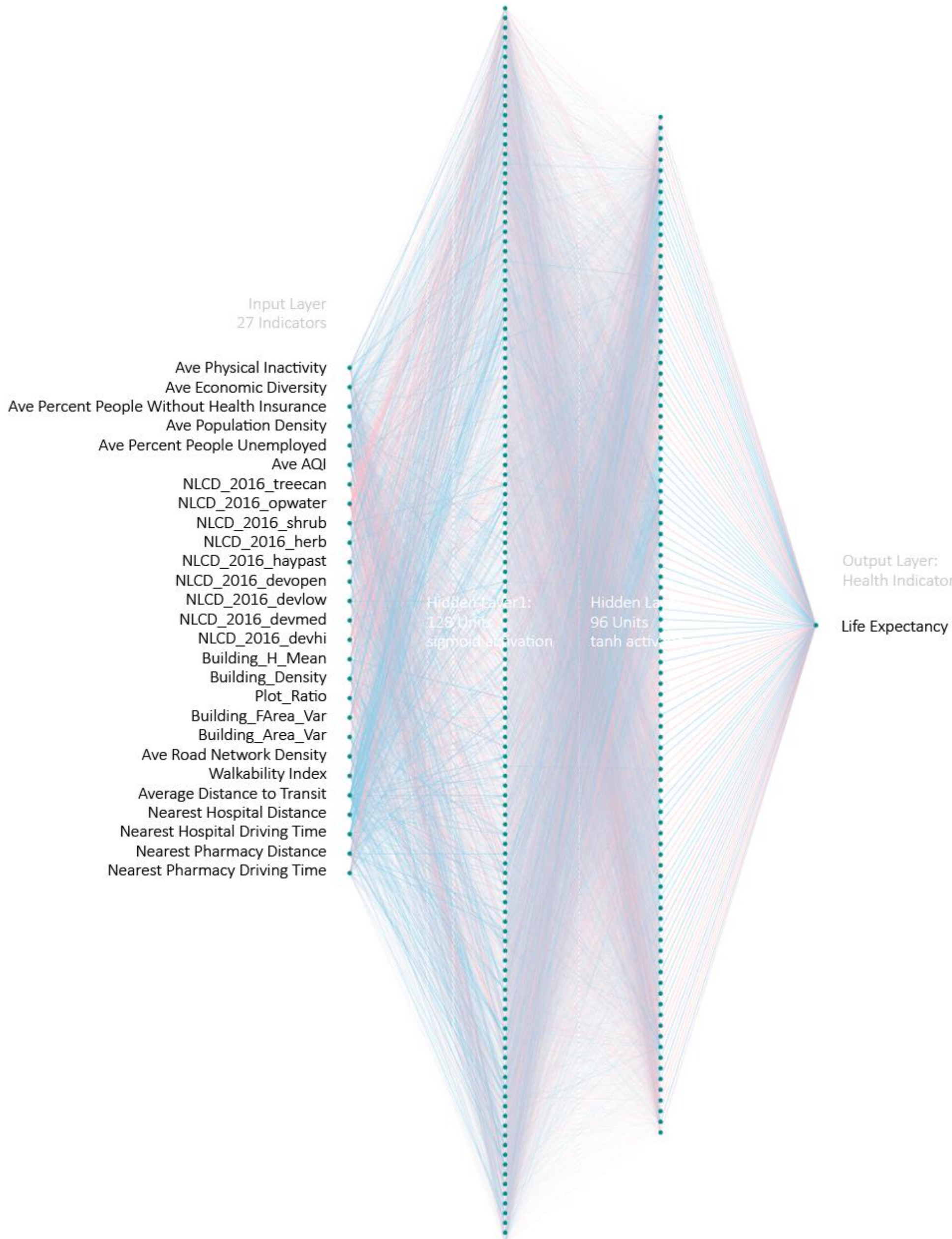


Fig.4.2.2-2: Best Performance ANN Model for Life Expectancy

To uncover the black box, two methods of analyzing importance are used - SHAP importance analysis, and Permutation importance analysis. Importance visualization charts are shown for comparison [Fig.4.2.2-3, Fig.4.2.2-4].

SHAP (SHapley Additive exPlanations): SHAP values are based on the principle of Shapley values, a concept from cooperative game theory, which provides a mathematically rigorous framework for attributing contributions. It is a powerful method for explaining the output of any machine learning model. SHAP calculates the impact of every feature to the target variable (called shap_value) using combinatorial calculus and retraining the model over all the combination of features that contains the one we are considering. The average absolute value of the impact of a feature against a target variable can be used as a measure of its importance.

It can reflect both the global and local importance well - By averaging the absolute SHAP values across all instances, we can get a global feature importance ranking, showing what features, on average, have the biggest impact; SHAP values also provide local explanations, revealing how a specific feature affects the prediction for a particular data point.

Permutation Importance: Permutation importance helps determine the influence of every feature on a model's precision or performance. By permuting a feature, it loses its initial connection with the output and shows how much the model relies upon that particular characteristic for making precise forecasts, therefore it’s also a good approach for globally evaluating features importance.

The two methods gives some similar global findings: most important 3 factors are physical inactivity (negative), building height (positive), economic diversity (positive); unemployment rate, air quality index, building density, population density, tree coverage, plot ratio and building footprint area are also found of significant importance following the first 3 by both models, but the sequence are different; building area variance and health insurance non-coverage are two main factors that doesn’t appear as very important in SHAP analysis, different from the permutation importance approach. They both also agree with some least important features: the subtle natural environment features like shrub or grass coverage, and open water ratio - they give people good feeling in the communities but are proved of less importance, though it might because of detection methods limitation like smaller planter or grass’s overlapping with tree canopy is more possible calculated as tree coverage; but from macroscope, these features turns to be more microscope and less important. Hospital and Drug stores accessibility, however, turns out with less doubt as not important, which corresponds to our previous correlation analysis, verifying them as a result (poor health issues) fixing role instead of long-life expectancy causing role in nowadays developed urban environment.

Many informative differences are also shown by two methods, especially because SHAP is better at inspecting local findings and permutation importance can give a more quantitative global sense. Below is a thorough interpretation of each unique finding.

Permutation importance found that temporary physical inactivity rate are far more important than all other features, even when complex urban mechanisms are found, and then building height’s importance is far more important than all other features; Building height’s much importance is not realized by previous findings, compare to other built environment features like density, plot ratio and area variance, building height can symbolize more of the cities all aspects - desire, honor, rights, power, density, etc., it’s a symbolization of the gathering of all human’s societal activities, and all other explicit activities importance, like economic diversity, unemployment ratio, is not as important compared to building height.

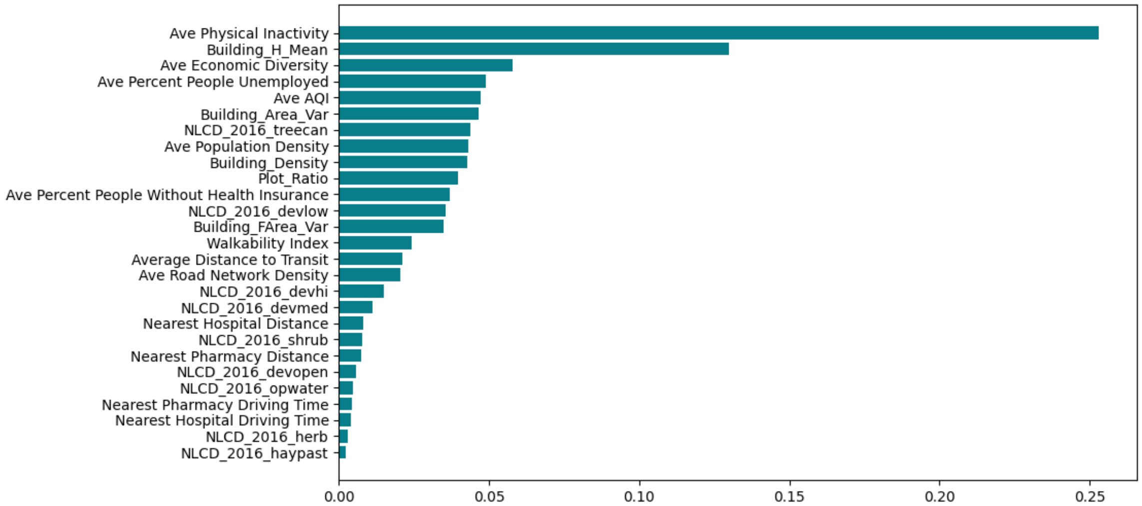


Fig.4.2.2-3: Feature importance obtained from permutation importance method

The SHAP model tends to rank the most distinct and essential features higher, which makes the result more meaningful towards reality intervention, since the basic variables are easier to do changes with.

SHAP is very insightful in showing up air quality and population density relative higher importance, which implies the MLP model has learned some essential implicit complexity within the system, since as we previously mentioned, population density is not directly related to health and human activities, it relates highly with the built environment feature and works in an implicit way. SHAP also ranks unemployment lower than air quality, population and building density and tree coverage, which implies that unemployment is a more temporary attribute whose change can be caused by other more essential features.

Another distinct difference between these two methods is with the built environment diversity related features - building area variance is considered 6th important by permutation importance, but only 15th importance by SHAP, the building footprint area variance is considered of 10th importance by SHAP but neglect by permutation importance. We'd consider building area variance's importance detection in permutation analysis is not essential, since it is a factor whose can be approximated by building height and footprint area, and in this research we used the calculated value instead of direct data from reality, thus bringing more collinearity for this feature with height, footprint area and density, so it shows a higher ranking in permutation importance which has a disadvantage in recognizing collinearity and therefore attributes wrongly the part of footprint-area variance towards building area variance. Same thing happens for the plot ratio, whose importance for health mainly relies on building height and density. From this model, it implies that other than building height and building density, building footprint area variance as a building diversity reflection, is another essential factor.

As for the natural environment, SHAP also shows its significance in detecting what are truly important - air quality and tree coverage. Therefore, SHAP model has proved to have greatly exploited the dark box, making the ANN model much more transparently explained.

Yet here we have to admit that, although the MLP model has learned many complex relationship, it still has limitation in interpretation, and it's possible it still omits some deeper implicit relationship, like for air quality, natural climate and open water ratio and plant coverage are both natural factors influencing greatly towards it, while higher development density might also have the problem of heavy traffic and more industrial gas exhaustion which impair the air quality, but in this ranking not all hidden insights for air quality are not easily interpretable. Such finer grained exploration directions are shown to be in great need for further exploration in future works.

Fig.4.2.2-4: SHAP value (impact on model output) - Top 20 variables

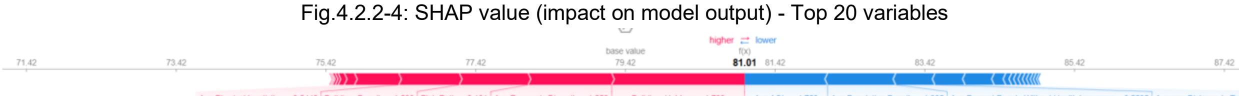


Fig.4.2.2-5: SHAP value (impact on model output for single instance)

SHAP has its practical meaning also in its high interpretation ability for single instance prediction, such as the above [Fig.4.2.2-5] is point prediction (symbolize 1 census tract's life expectancy prediction), SHAP shows how the MLP model makes the prediction, that is how each attribute is calculated for their contribution to push the value towards the final predicted value. This good tool can be used for visualization interpretation for policy makers.

The partial dependency table shows better how each variable makes its contribution towards the result. When joined with another variable it also interprets some hidden non-linear relationship better. For example, the graph of [air quality, building height, SHAP], it shows that high building height areas have two extreme groups, one is of poorest air quality which could probably because of higher traffic and industrial activity density that brings too much unhealthy emission, another one is with good air quality, thus height doesn't always boost a better health in every perspective, more research attention should also be put here to quantify what further development patterns regarding the balance of building height, density and air quality (with some other possible important factors like traffic or industrial density) are truly better.

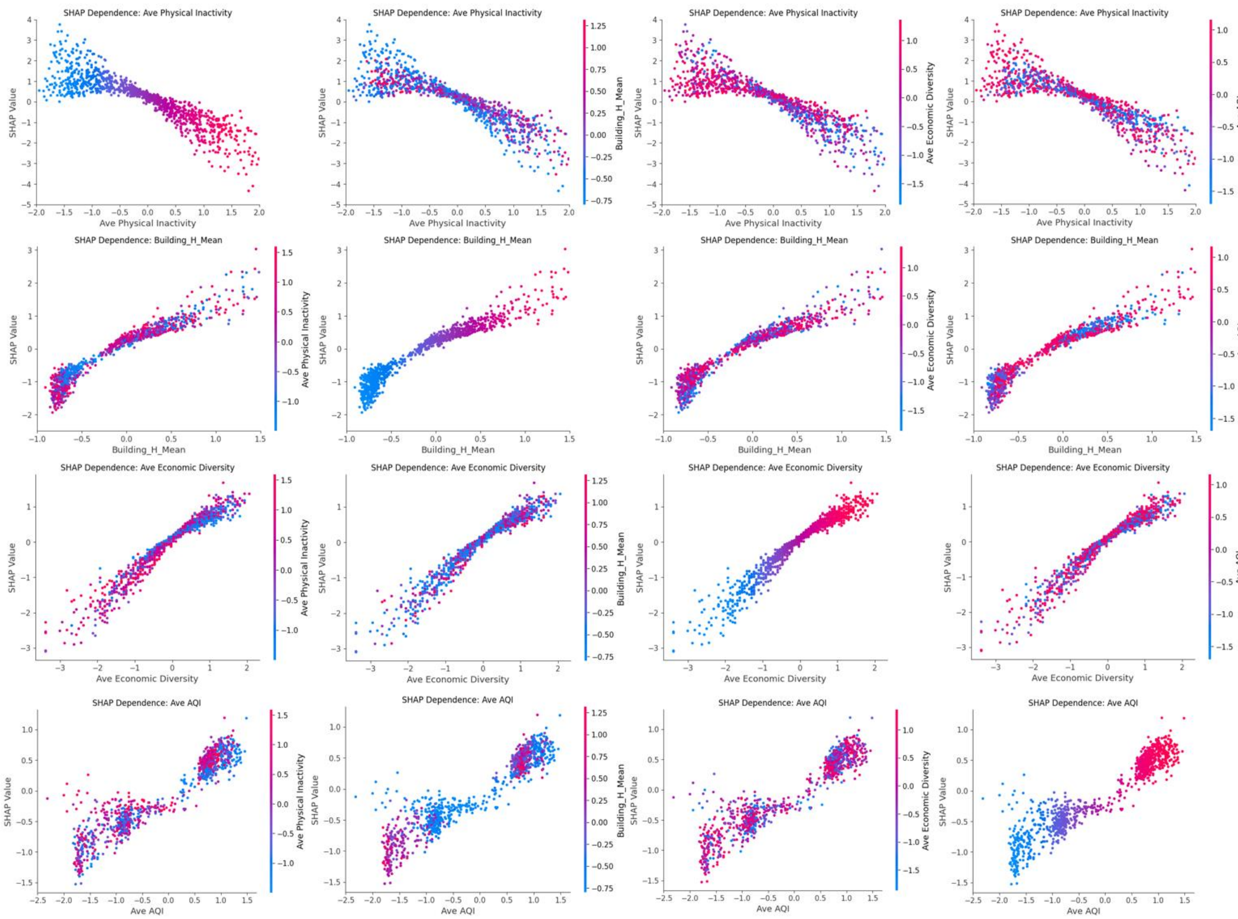


Fig.4.2.2-6: SHAP Dependency Values - Top 4 important features dependence matrix

The SLFN model has some supplementary findings. The SLFN with a train R^2 of 0.5359 and test R^2 0.5147 shows some other analysis perspective: the building area variance here, as a reflection of building diversity, might also be an important feature; in this importance list, building height and unemployment relative importance improves greatly compared to MLP interpretation; since this model R^2 performance is not significantly lower than MLP model, it suggests that some variables are still under-exploitation, which might introduce some interesting topics for further research, like for urban development, how important is building type diversity and when do we need to consider it, when do we need to balance the buildup efficiency versus building diversity, etc.

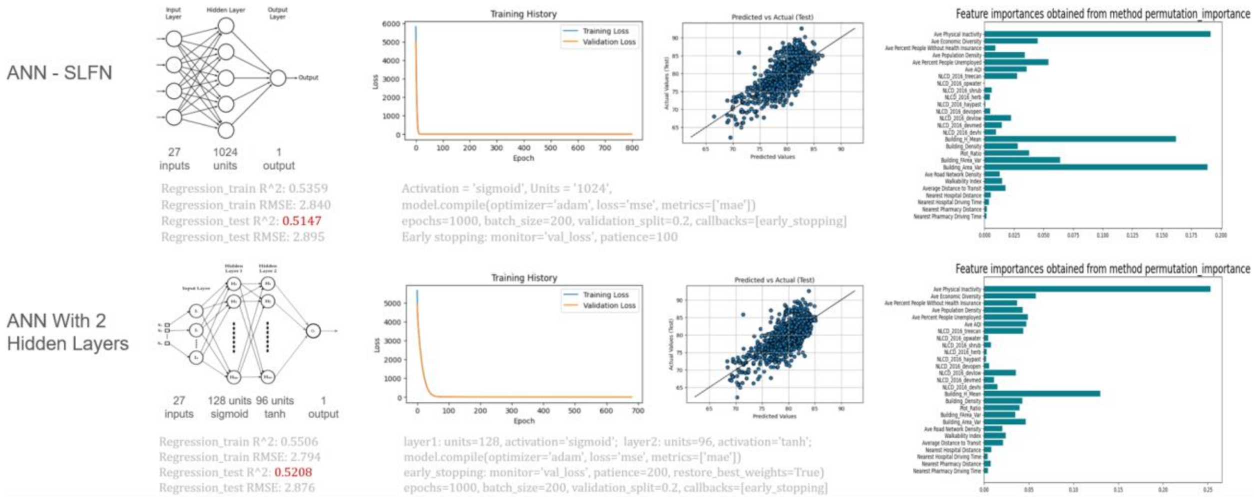


Fig.4.2.2-7: ANN Model Performance and Interpretation Charts

Tree Models give similar essential insights, here we'd only list some supplementary findings: DT prefers sparse features, so as RF. Both them show that among density features, a census tract's low-development-ratio is also an important factor, and RF also emphasized a little bit more on the public transport accessibility factor (average distance to transit) which is also an easy improvable factor for relieving air pollution issues, which can be meaningful; however the RF model treats all other built environment and street or traffic features as evenly not important, neglecting their interrelationship with building height, economic diversity, and many complexities, thus it cannot give convincing generalizations, and they also cannot attributes collinearity well due to the problem of too much underfitting (decision tree) or overfitting (random forest), therefore their results are just of moderate reference value.

Fig.4.2.2-8: DT and RF Model Performance and Interpretation Charts

Polynomial regression, on its own, can do better than DT models, reaching a test R^2 of 0.44 because of its partial learning of the complexity in the system, and learns some interweaving relationships. 2-degree polynomial model gives the importance rank as follows: A multiplication of high-development-ratio and tree coverage stands on the top; Building density * tree coverage follows, then is the economic diversity itself, next follows dev open * building density, and then comes several density relevant features' combination with themselves or with natural features. Shrub coverage, as a single parameter, also shows some importance. This analysis gives us a new interpretable perspective in realizing how urban features relate; MLP models are pretty much black boxes while polynomial models are clear in intersection relationships. From the result we can understand that, most density related features work most effectively in a way together with natural features, and spatial density patterns is a complex theme that includes smaller area (community or block, for example) local spatial patterns and their combination patterns, yet with diversity or spontaneity evolving or decaying modes in relation to economics and migrations patterns, their relationship analysis should be defined better and multiple methods including numerical and graphical (computer vision) statistics should be done to push forward the spatial pattern analysis with other urban features, which introduce another top priority interesting work for the future, basing on this research.

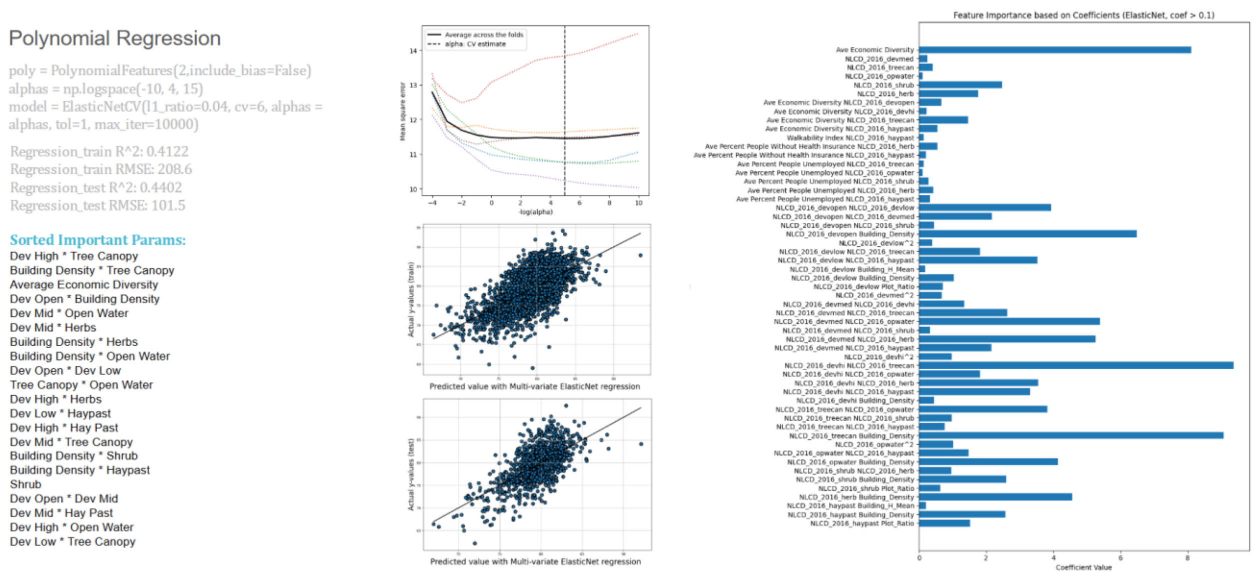


Fig.4.2.2-9: Polynomial Model Performance and Interpretation Charts

Through the exploration of model training, tuning, comparison and interpretation, best models with the most important features as well as essential affecting mechanisms or implicit relationships are elaborately demonstrated and discussed from a systematic and scientific perspective with modern quantitative methods. Such results not only set up the base framework for this big question, but also provide a context for discussing such problems scientifically and systematically, which can further point out some most important directions and provide valid suggestions for intervention strategies for future development. The following section would dive into intervention strategies evaluation and suggestion.

- 4.3 Intervention Evaluation and Suggestion

In this section the interventions are given to some features of all census tracts as a very rough futural development situation mimic to see how life expectancy would enhance using the MLP model, thus providing some reference for policy makers.

By only improving one large category's features to a reasonable extent, here we use the increment or decrement by the initially 25% central span’s corresponding value (half of the middle 50%’s span’s value, in normal distribution is 0.67.4 standard deviation), and setting others as default (same as their original values), we can compare the life expectancies improvement across feature groups and cities. Apparent improvements can be revealed by average value increment and by a greater number of performance improved census tracts than life expectancy decreased census tracts. All intervention strategies applying (adjust all feature groups by meaningful features) will result in a best result - an improvement in average life expectancy of 3.2 years (from 79.43 to 82.63), while improving 4419 census tracts life expectancy versus only 907 decreases; If only do single parameter change, effect rank is : built environment features > human related features > natural environment features > street and traffic features.

Single changes of built environment features can raise life expectancy from 79.43 to 80.8, for reference the change extents are recorded below:

Building height	+6.1418m
Building density	+0.0697
Plot ratio	+0.6523
High-dev-ratio	+24.17%
Low-dev-ratio	-10.87%
Medium-dev-ratio	-15.42%
Building footprint area variance	-5.16e+5
Building area variance	-1.56e+7

Single changes of human related features can raise life expectancy from 79.43 to 80.67, the changes are recorded below:

Physical inactivity rate	-7.29%
Economic diversity index (score)	+0.105
Health insurance non-coverage	-4.37%
Unemployment	-2.84%

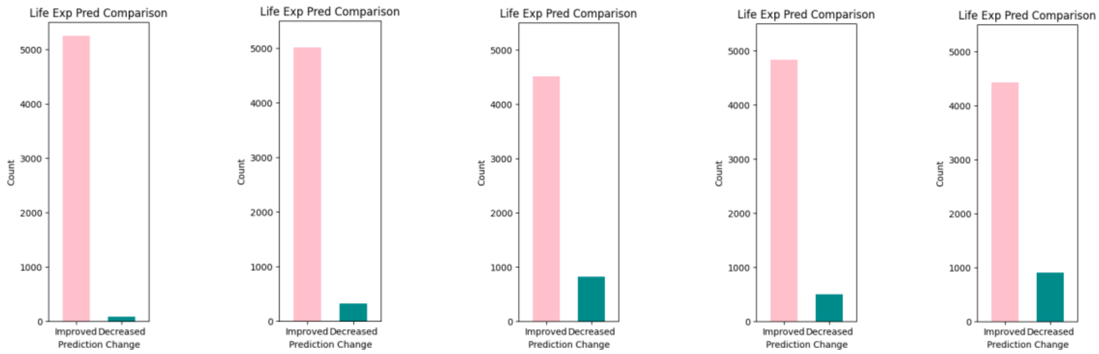
Single changes of natural features can raise life expectancy from 79.43 to 79.87, the changes are recorded below:

Air quality index (score)	+6.27
Tree canopy ratio	+1.87%

Single changes of natural features can raise life expectancy from 79.43 to 79.77, some of them are abnormal or counter-intuitive, the changes are recorded below:

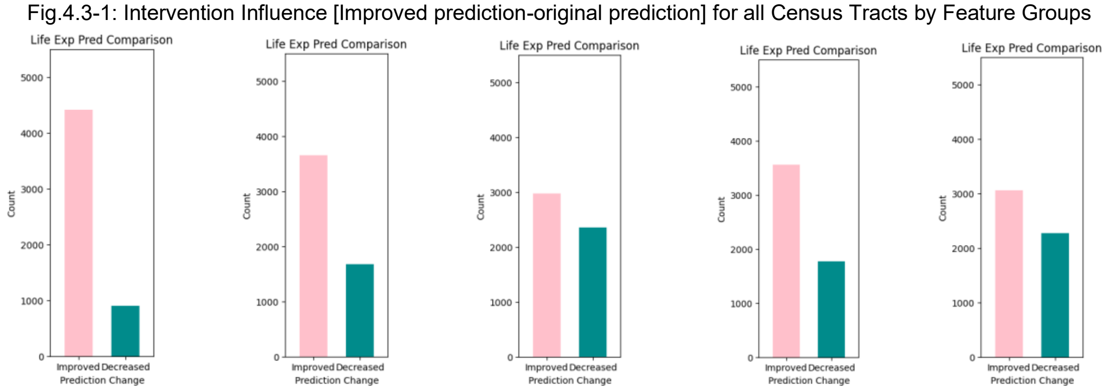
Road network density	+4.1187
Nearest hospital distance	+635.50m
Nearest hospital driving time	-101s
Nearest pharmacy distance	-516.75m
Nearest pharmacy driving time	-84.38s

Taking all these feature adjustments above reaches the life expectancy enhancement of 3.2 years. In reality, single type parameter change is usually not possible since the urban is a system where many attributes are interrelated to be changed together, like building more in low dev-density area typically will also boost the economic activities and influence plat coverage, however since we cannot do experiments in real world, this analysis is still meaningful in giving us intuitives on intervention strategies, this reference could be considered with more local verification research to help to reach final strategies.



All Features | Built Environment Features | Street Network Features | Human Related Features | Natural Environment Features

Fig.4.3-1: Intervention Influence [Improved prediction-original prediction] for all Census Tracts by Feature Groups



All Features | Built Environment Features | Street Network Features | Human Related Features | Natural Environment Features

Fig.4.3-2: Intervention Influence [Improved prediction-original true value] for all Census Tracts by Feature Groups

City inspections below show the model can in general fits for all cities, but it also gives a local abnormal finding - when simply improving the natural environment attributes, Houston and Phoenix are showing a decrease in life expectancy comparing to its original value because of the air quality feature; after checking their distribution as well as the partial dependency table for air quality [Fig.4.3-7], it can be understood since a large part of their spots fit into the decreasing part of the partial dependency table; such a case means the model still has some limitations, not all features of all cities can be learned perfectly by it.

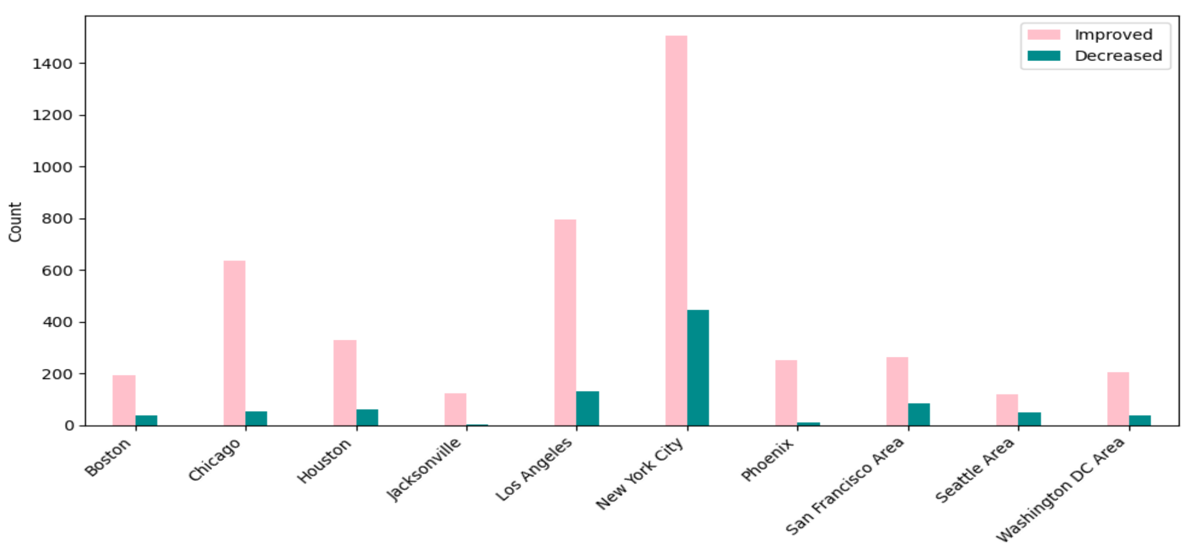
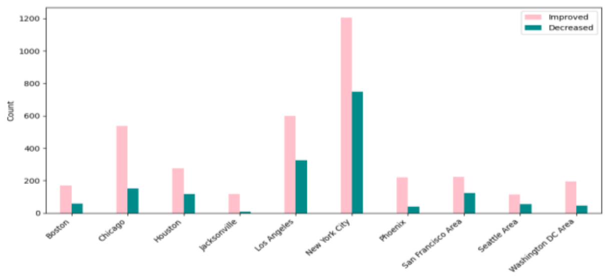
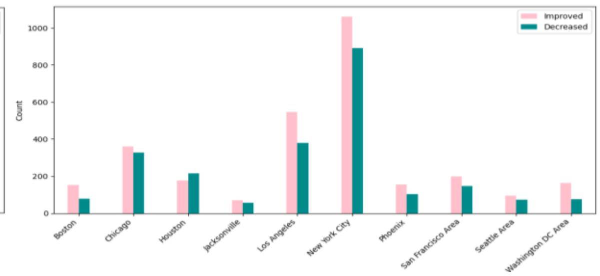


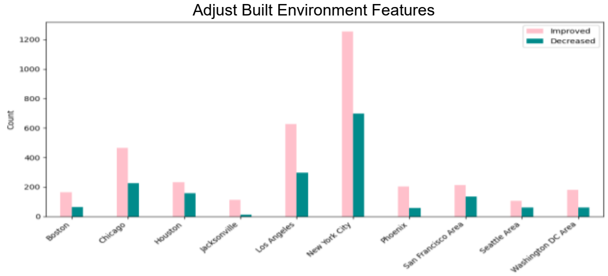
Fig.4.3-3: Single feature group adjustment for city comparison [Improved prediction-original prediction]



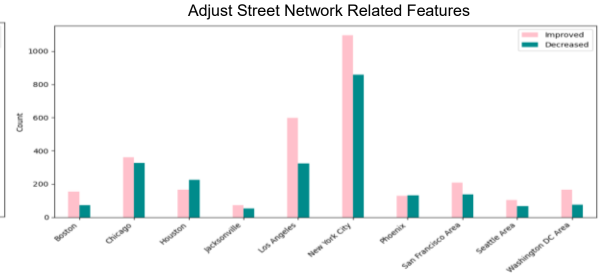
Adjust Built Environment Features



Adjust Street Network Related Features



Adjust Human Related Features



Adjust Natural Environment Features

Fig.4.3-4: Single feature group adjustment for city comparison [Improved prediction-original true value]

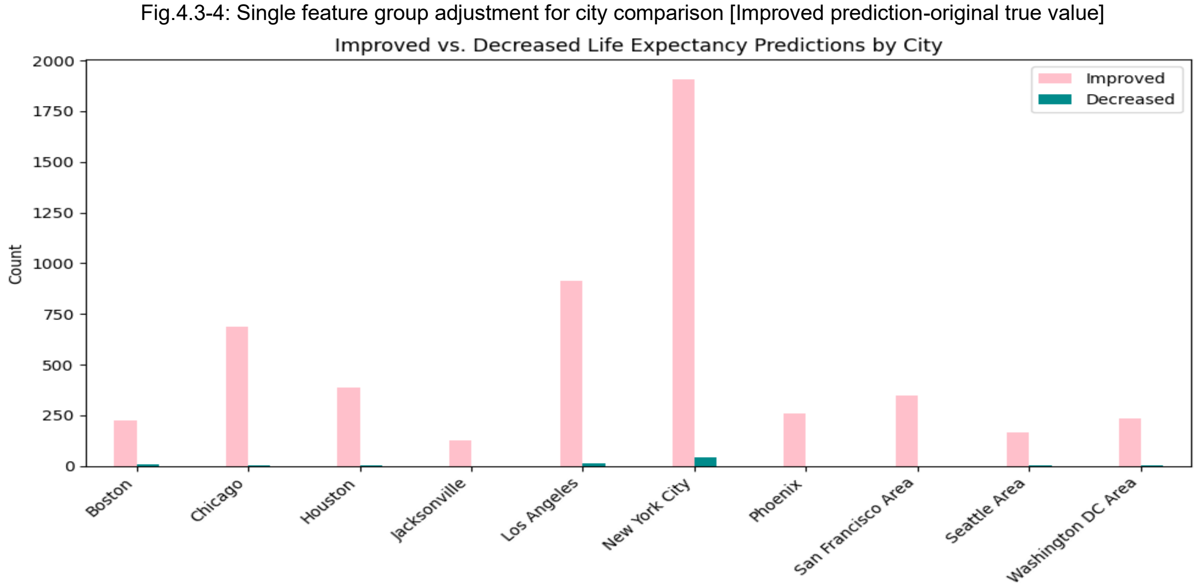
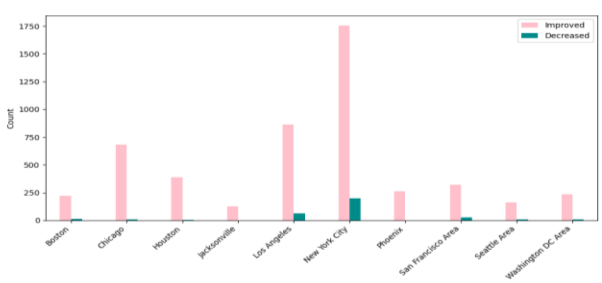
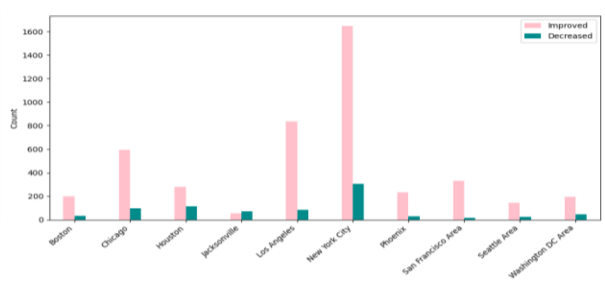


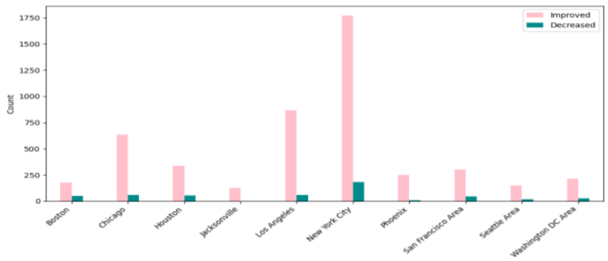
Fig.4.3-5: Adjust all features for city comparison [Improved prediction-original prediction]



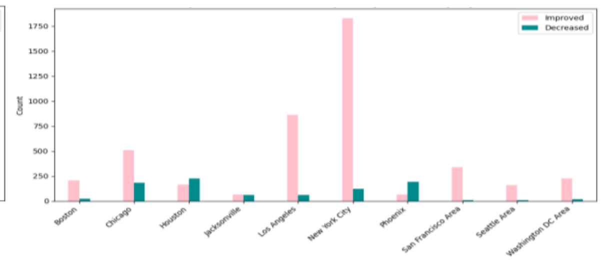
Adjust Built Environment Features



Adjust Street Network Related Features

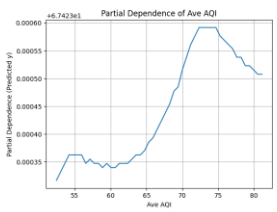


Adjust Human Related Features

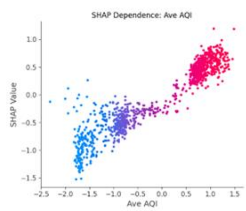


Adjust Natural Environment Features

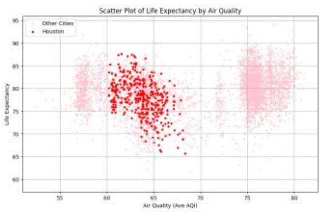
Fig.4.3-6: Single feature group adjustment for city comparison [Improved prediction-original prediction]



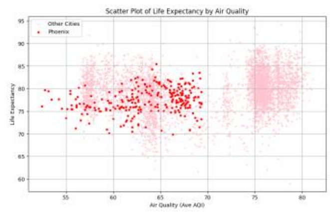
Partial Dependency Table



SHAP value for AQI



Houston



Phoenix

Fig.4.3-7: Abnormal checking - distribution checking of Houston and Phoenix Air Quality Data

- 4.4 TOOL DEVELOPMENT DEMONSTRATION

A preliminary version, as a way of the research result transformation, demonstrates the city level and census tract level analysis and intervention suggestions; while this only shows a few of the urban health indexes with only very few essential functions, it provides a prototype for expanding the web application soon. An amplified version with all features by category and their influence analyses with various ML models, and better intervention suggestions with higher user interactivity can be a great future development direction.

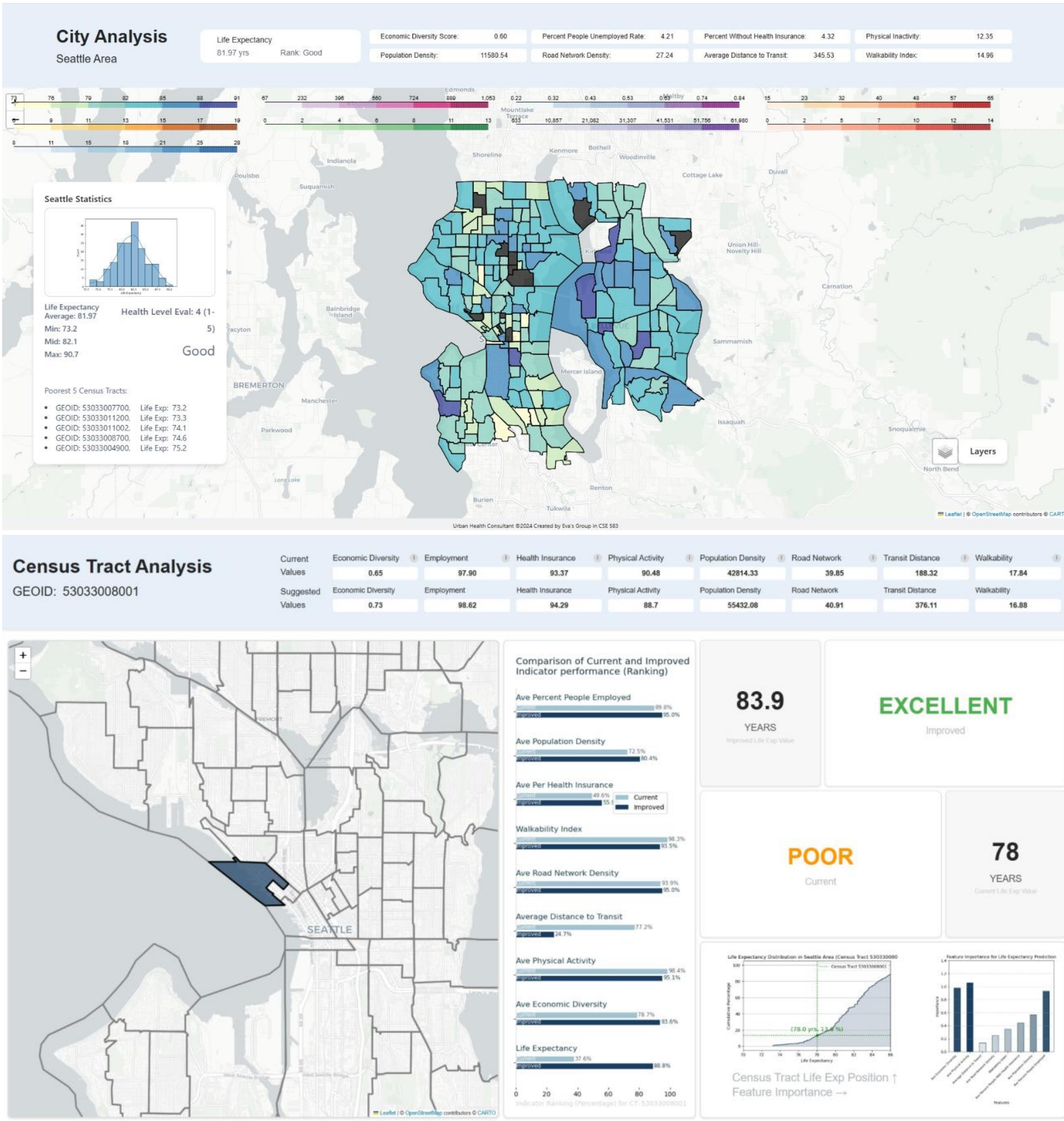


Fig.4.4: Web tool development demonstration of a prototype version

5. Discussion, Explanation and Limitations

- 5.1 Interesting Details, Cases and Abnormal Examples Analysis

Discussions in this section mainly focus on interesting topics or details and within them the competing features or abnormal findings' interpretation.

As beginning, it should be clarified here again how cities are selected to ensure the validity of this research: For machine learning, more data always means better, but to acquire data given the limited time and resources, only the most representative and necessary data should be included, so this research finally included 10 metropolitans (of 6044 census tracts, enough for machine learning), 9 of which are within top 15 largest population cities, and geographically distributed evenly in the US, to ensure the them to be representative for general urban environment in large US cities.

Secondly, how well can the best trained models learn the convoluted relationship well worth discussing. Many cases in previous sections have shown that the Pearson Correlation score is very valuable in showing many surface relationships but not revealing why or more fundamental correlations, while ANN models (MLP) have learned the features well and the SHAP importance analysis has given deeper insights. Here more discussions and typical examples are worth mentioning again for this assertion. By Pearson table we can see some superficial contradictions, like important natural features air quality and tree canopy cover have weak positive correlation, but building density and high-development-area has positive correlation with air quality versus the negative correlation with tree canopy cover; these contradictions can be understood considering the SHAP importance analysis for MLP model – both air quality and tree canopy cover are having a higher up pushing effect for prediction when their value improves greatly, but when their values are low they have a negative push against the prediction result, such interpretation is a decomposing of their correlation phenomenon; and they as well as density are shown to be more important and essential factors as opposed to development-ratio features, which in reality are also decomposing factors of density – the MLP model learns all the interrelationship, inner-relationship (sub-inclusion) and the complexity well. These cases also verified the MLP model's generalization ability of learning the important trends while not being too sensitive towards each feature data's local defects, exemplified by the partial dependence and SHAP dependence charts. SHAP dependence charts is a local result meaning each points contribution towards prediction of a specific feature, while partial dependence table is a more global reflection of how a specific feature would influence prediction since each local span on the curve is calculated by a bunch of points on average. The SHAP dependency charts and the local unevenness and bumps on partial dependency curves means the MLP model learns the complex relationships well for being sensitive to local attributes and abnormalities, while the clear general trends by both charts without too much local variance means the macro relationships are also well learned, they together strengthened the asserts of the models well generalizability.

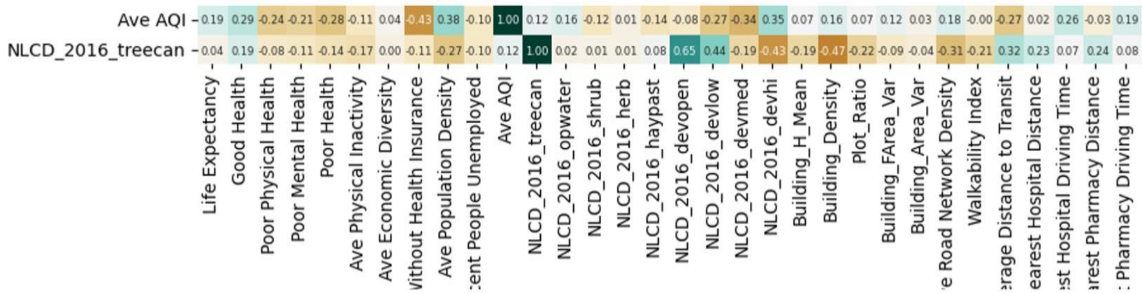


Fig.5.1-1: Pearson Correlation

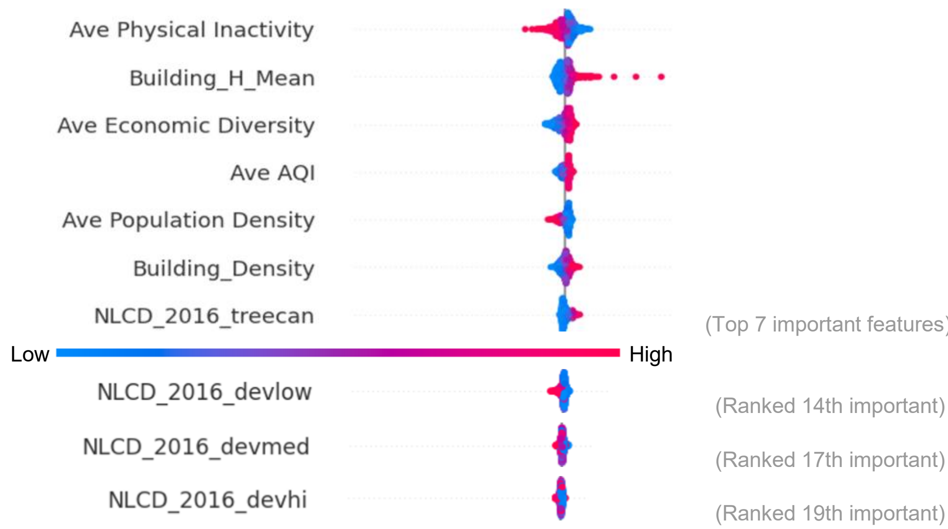


Fig.5.1-2: SHAP values

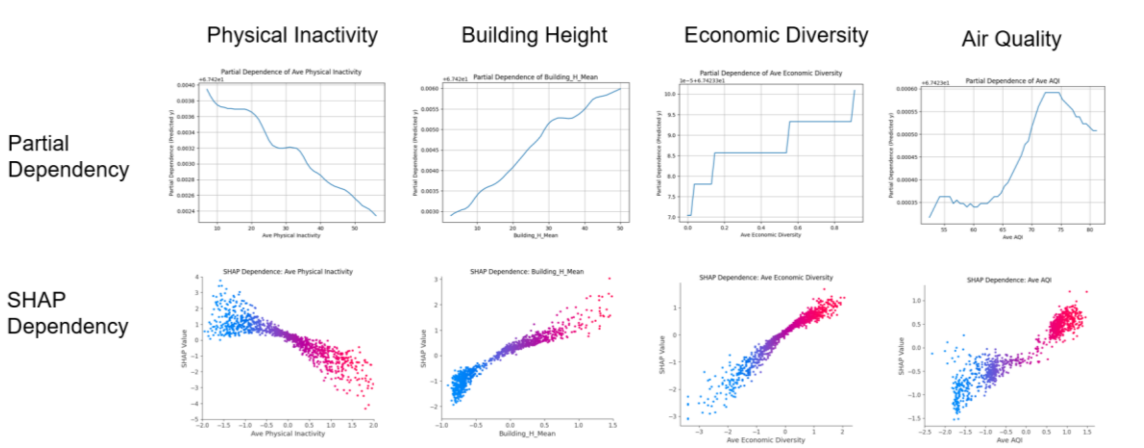


Fig.5.1-3: Partial v.s. SHAP dependence – global v.s. local effect of feature towards prediction

Based on the research analysis, one interesting topic further is how to bring such a synthetic system's learning result towards practical implementation reference in urban planner's daily works, in other words, apart from using the model's prediction is there other meaningful reference for intervention – Walkability can be a good understanding perspective or cut-in-point. Physical inactivity and walkability, while both are human activity related features, the former one is more of each individual's human physical evaluation that cannot be directly changed by urban planners, but walkability in a broader sense is an index built as the bridge of urban environment features and human physical activity, which would be better understood as human's environmental activity friendliness and can be changed directly by urban planners, in fact all features of this research are all greatly overlapping with or originally from walkability and urban design quality relevant indexes according to literature reviews, so models trained in this research can also be understood as an index of physical inactivity plus environmental activity friendliness, and so better definition and adjustment of such “walkability” can make a great effect for environmental health evaluation in urban planner’s typical works. Since the current unchangeable physical inactivity is strongly correlated to health, one easy way of judging whether the walkability we define is good is through its relationship with physical inactivity, strong correlation with physical inactivity and not very strong correlation with features defining the walkability would be an ideal case.

Current walkability as one of the single features within the 27 indexes is a very narrow-understanding and crude one that has great improving space in its more meaningful re-definition in future works. In this research it is only defined by 4 features simple add-up (intersection density (+), distance to the nearest transit stop (-), employment diversity (+), and employment and housing diversity (+)), which does not synthesize well; the initial block groups walkability is aggregated and averaged to map towards the census tract level, which causes some loss of sensitivity of it’s relationship towards other urban features; moreover, since some important features like a central park may be out of the census tract boundary but still can contribute much towards the walkability, so current calculation within the certain census tract boundary is far too less than synthesizing well, better algorithms should be applied based on various features’ proper spatial granularity.

Moreover, how such importance ranking could be understood is also interesting and needs discussion. Census tract average building height, out of expectation, is proved such a significant feature; among the top ten most important features, the built environment features takes up 4 positions (building height, density, plot ratio, footprint area variance), social-economic features takes up 3 position, 2 are natural environment features and 1 is human’s physical activity, such results would raise interesting discussions like how urban form, spatial pattern and architecture forms respond to cultural, economic, and human’s mental and physical activities, which has been an important theme in several architectural and planning masterpieces, like in ***Delirious New York*** *(1978)*, Rem Koolhaas points out: Height is not just about maximizing real estate; it's about symbolic power, liberation from the ground, and the creation of parallel worlds stacked above each other, and this vertical layering mirrors the diversity and chaotic vitality of urban life; skyscraper as reaching the height’s extreme, is not merely as a functional or economic solution, but as an architectural embodiment of human fantasies, ambition, and excess; skyscrapers as representations of human desire, power, economic forces, and cultural psychology, they are not just buildings, they are manifestations of the collective urban unconscious – this research findings perfectly respond to such historical inspection, and with this joint and interdisciplinary perspective we can finally understand why building height feature is so important and out of expectation, and work so organically with other environmental features to make such synthetic influence. Although much of the urban pattern’s specifics remain to be discovered in future works, this research surely has made a big quantitative step forward.

- 5.2 LIMITATIONS

First of all, it should be mentioned that many limitations do exist in this research, but they are inevitable at this stage since the research gap is so broad and the very first and essential step in furthering any specific features (more meaningful and quantitative exploration, iteration or discussion) later is to build such a context first, what is done throughout this research. Now here, the limitations are detailly listed as followings – primarily data limits, system limits, and approach limits – aiming to build futural better system through overcoming such obstacles.

Data limitations are the most essential ones for this research's implementation. Data accuracy and generalizability throughout various time periods and space ranges (of various granularity), data collection methods' defects, data amount's inadequate, etc., these are all contributing towards this system and models' defects. Through this research some data are gained basically highly relying on some computer science approaches and algorithms, and it's very common to see such industries iteration over time with the explosion development of technology and efforts put into it; the current approaches and their concatenations are by no means permanent and data are by no means perfect, much can be done in the future to improve the data accuracy.

Building height dataset, for example, is especially in need of improving accuracy as it's proved such an essential feature but poor in data collection's consistency – the current height data are from multiple sources yet not any source can cover all census tracts in these 10 metropolitans as needed with good accuracy, and different data sources are having different bias thus making the concatenation with even more deviation from the reality, and given the theme and time of this research, inspections into each data sources specific accuracy and tests on how they would combine to be best are not proved as a major thing previously thus not taken into much consideration, but they still can be a limitation of this research and should be put much emphasis on in furthering this research.

Plant coverage dataset also has some methods' intrinsic limits, since this research utilized the satellite data which is doing recognition based on 2D features, so it's in-sensitive towards the actual overlapping plants cover of all kinds, such approach makes tree canopy cover detection more accurate, but micro or human view perspective features like shrub, grass, etc. are easily underestimated if overlapped, given that urban environment small plants are already few, these underestimate may cause large inaccuracy.

Most other features data are collected near 2020 though large spatial and population surveys or with multiple years average, therefore are considered enough for this research, however there's still limits and some features can always do better in data collection. Air quality index data is an extreme example of possibly not very adequate construction of the dataset in time and space granularity – the current data collected are from October 2024, a nearly ten years span opposed to life expectancy data period – 2015. Air quality data was collected by the only accessible approach for all these census tracts – through google map API, but still only very recent air quality data (30 days) are remaining accessibility and no other source is found to have enough historical data, so while current AQI data have a moderate generalizability in this research, it still can be improved by collecting more data in a longer time span, especially as it's proved to be of top 5 important features.

System limits exist as there would always be out-of-system features this system is incapable of, and there would always be more features to be included into. Genetic features are essential to human health but we can do nothing with it by this system; humanity features, social economic features, mobility features, as well as natural environment features like weather, etc. are only covered with the most essential ones in this research, as they turned out to be correlated and can train models with good performance, it's also shown that some features are under-exploit and waiting for more features joining to deeperly uncover the black box.

Moreover, analysis approaches taken in this research are mainly statistical or data science approaches, however to give more insights, joint approach of urban planning like case studies can also be applied – like analyzing 2D plots of typical or interesting census tracts to give some insights from the urban planner's perspective – it could be a good supplement to make this research a better interdisciplinary one.

Apart from all limits above, some minor but general restrictions like time and computation resources accessibility are making some obstacles too, more insights can be made and possibly even better models can be trained; however, as it's said, all models are wrong, some are useful, both the model and formation of the dataset could this quotation be applied, leading to the overcoming of these obstacles found for the completion of the system in the future.

6. CONCLUSION

- 6.1 FUTURE WORKS

Future works remain not only for overcoming this research's obstacles and limitations mentioned above, like improving data's granularity and accuracy especially, but also by expanding this research's insights on futural important directions.

As interdisciplinary research, deeper diving with image data analysis as well as joint approaches from both data science and traditional architecture and urban planning realm should be made for unboxing the urban development patterns and targeting at some practical platform or tools development that can significantly and directly help urban planners to make planning decisions. Computer vision methods for analyzing census tract satellite data and street view data are crucial for further uncovering the development pattern of the urban neighborhoods: the combination of built and natural environment features like street network connectivity with building density, height, development density, and plant coverage, are usually forming the planning of the spatial pattern together, so they need to be analyzed using innovative and customized algorithms during computer vision analysis. Environmental human view factors mainly rely on analysis of street view images, and it can also be joined with some human perception acquisition through some survey methods, with building, streets, plants, air quality, etc. their combination effect on human view and perception perspective, to better define the environmental friendliness for human activity.

With these spatial patterns and human activity friendliness further explored, more interesting and interdisciplinary research can be done for the thorough unboxing of urban environmental influence on health and therefore make urban and neighborhood planning never be unfounded again, for instance interesting directions can be but not limited to the following:

1. Build and test the synthetic walkability index.
2. Do health and height – spatial pattern – economic vitality analysis.
3. Do health and development – diversity related analysis.
4. Do health and spatial pattern – Air quality joint analysis.
5. Do health and spatial pattern – plant coverage joint analysis.
6. Do mental health and height – human perception related analysis.

- 6.2 CONCLUSION AND CONTRIBUTION

To give it an end, conclusion and contributions are summarized here. This research built the scientific, quantitative and systematic framework base or context that covers essential important features of urban environment for human health; this research trained good performance model for health prediction, which could serve as a valid reference for urban development policy; and this research explained the relative importance rank of essential urban features and pointed out important directions of this field's future. In one word, this research has made significant progress in the exploration and unraveling of the integrated relationship of urban features and human health, with the help of advanced machine learning and computation analysis methods diving into the 5 large dimensional urban and human related indexes and metrics' exploration, analysis and evaluation, across the most representative US metropolitans neighborhoods, greatly bridging the research gap and making significant meaning towards the urban planning intervention practice and the scientificalization of the realm's development.

APPENDIX - REFERENCES

[1] Y. Casali, N. Y. Aydin, and T. Comes, “Machine learning for spatial analyses in urban areas: a scoping review,” *Sustain. Cities Soc.*, vol. 85, p. 104050, Oct. 2022, doi: 10.1016/j.scs.2022.104050.

[2] United Nations Department of Economic and Social Affairs, *World Urbanization Prospects: The 2018 Revision*. Erscheinungsort nicht ermittelbar: United Nations, 2019.

[3] J. Jacobs, *The death and life of great American cities*, Vintage books ed. New York: Vintage Books, 1992.

[4] E. Pérez *et al.*, “Neighbourhood community life and health: A systematic review of reviews,” *Health Place*, vol. 61, p. 102238, Jan. 2020, doi: 10.1016/j.healthplace.2019.102238.

[5] S. C. Koumetio Tekouabou, E. B. Diop, R. Azmi, R. Jaligot, and J. Chenal, “Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5943–5967, Sep. 2022, doi: 10.1016/j.jksuci.2021.08.007.

[6] “World Health Organization website.” Accessed: Jun. 09, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[7] J. Barton, R. Hine, and J. Pretty, “The health benefits of walking in greenspaces of high natural and heritage value,” *J. Integr. Environ. Sci.*, vol. 6, no. 4, pp. 261–278, Dec. 2009, doi: 10.1080/19438150903378425.

[8] H. F. Guite, C. Clark, and G. Ackrill, “The impact of the physical and urban environment on mental well-being,” *Public Health*, vol. 120, no. 12, pp. 1117–1126, Dec. 2006, doi: 10.1016/j.puhe.2006.10.005.

[9] S. L. Handy, M. G. Boarnet, R. Ewing, and R. E. Killingsworth, “How the built environment affects physical activity: views from urban planning,” *Am. J. Prev. Med.*, vol. 23, no. 2 Suppl, pp. 64–73, Aug. 2002, doi: 10.1016/s0749-3797(02)00475-0.

[10] R. Ewing and S. Handy, “Measuring the Unmeasurable: Urban Design Qualities Related to Walkability,” *J. Urban Des.*, vol. 14, no. 1, pp. 65–84, Feb. 2009, doi: 10.1080/13574800802451155.

[11] D. W. Newton, “Identifying correlations between depression and urban morphology through generative deep learning,” *Int. J. Archit. Comput.*, vol. 21, no. 1, pp. 136–157, Mar. 2023, doi: 10.1177/14780771221089885.

[12] J. WHO Centre for Health Development (Kobe, S. Weaver, D. Dai, C. Stauber, R. Luo, and R. Rothenberg, “The urban health index: a handbook for its calculation and use,” World Health Organization, Geneva, 9789241507806, 2014. Accessed: Dec. 01, 2024. [Online]. Available: https://iris.who.int/handle/10665/136839

[13] Z. Han, T. Xia, Y. Xi, and Y. Li, “Healthy Cities, A comprehensive dataset for environmental determinants of health in England cities,” *Sci. Data*, vol. 10, no. 1, p. 165, Mar. 2023, doi: 10.1038/s41597-023-02060-y.

[14] R. Koolhaas, *Delirious New York: A Retroactive Manifesto for Manhattan*. New York: Monacelli Press, Incorporated, 2014.

[15] N. Rautio, S. Filatova, H. Lehtiniemi, and J. Miettunen, “Living environment and its relationship to depressive mood: A systematic review,” *Int. J. Soc. Psychiatry*.

[16] M. A. Marzukhi, N. M. Ghazali, O. L. H. Leh, and H. Yakob, “A Bidirectional Associations between Urban Physical Environment and Mental Health: A theoretical framework,” *Environ.-Behav. Proc. J.*, vol. 5, no. 13, p. 167, Mar. 2020, doi: 10.21834/e-bpj.v5i13.2048.

[17] W. E. Marshall, D. P. Piatkowski, and N. W. Garrick, “Community design, street networks, and public health,” *J. Transp. Health*, vol. 1, no. 4, pp. 326–340, Dec. 2014, doi: 10.1016/j.jth.2014.06.002.

[18] N. Abbasabadi and M. Ashayeri, "Urban energy use modeling methods and tools: A review and an outlook," *Build. Environ.*, vol. 161, p. 106270, Aug. 2019, doi: 10.1016/j.buildenv.2019.106270.

[19] N. Abbasabadi, M. Ashayeri, R. Azari, B. Stephens, and M. Heidarinejad, "An integrated data-driven framework for urban energy use modeling (UEUM)," *Appl. Energy*, vol. 253, p. 113550, Nov. 2019, doi: 10.1016/j.apenergy.2019.113550.

[20] D. W. Newton, "Deep learning in urban analysis for health," in *Artificial Intelligence in Urban Planning and Design*, Elsevier, 2022, pp. 121–138. doi: 10.1016/B978-0-12-823941-4.00018-4.

[21] A. Maharana and E. O. Nsoesie, "Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity," *JAMA Netw. Open*, vol. 1, no. 4, p. e181535, Aug. 2018, doi: 10.1001/jamanetworkopen.2018.1535.

[22] Y. Li, N. Yabuki, T. Fukuda, and J. Zhang, "A big data evaluation of urban street walkability using deep learning and environmental sensors".

[23] W. Qiu *et al.*, "Subjective or objective measures of street environment, which are more effective in explaining housing prices?," *Landsc. Urban Plan.*, vol. 221, p. 104358, May 2022, doi: 10.1016/j.landurbplan.2022.104358.

[24] D. Ki, Z. Chen, S. Lee, and S. Lieu, "A novel walkability index using google street view and deep learning," *Sustain. Cities Soc.*, vol. 99, p. 104896, Dec. 2023, doi: 10.1016/j.scs.2023.104896.

[25] "NVSS - United States Small-Area Life Expectancy Estimates Project." Accessed: Jun. 13, 2025. [Online]. Available: https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html

[26] CDC, "About PLACES: Local Data for Better Health," PLACES: Local Data for Better Health. Accessed: Jun. 13, 2025. [Online]. Available: https://www.cdc.gov/places/about/index.html

[27] "Census Bureau Data." Accessed: Jun. 13, 2025. [Online]. Available: https://data.census.gov/

[28] "Google Maps Platform," Google for Developers. Accessed: Jun. 13, 2025. [Online]. Available: https://developers.google.com/maps

[29] "History | Air Quality API," Google for Developers. Accessed: Jun. 13, 2025. [Online]. Available: https://developers.google.com/maps/documentation/air-quality/history

[30] K. Schertz, O. Kardan, M. Rosenberg, and M. Berman, "National Land Cover Database at Census Tract Level (2016, mainland US)," Oct. 2021, doi: 10.17605/OSF.IO/HXMWJ.

[31] *microsoft/GlobalMLBuildingFootprints*. (Jun. 13, 2025). Python. Microsoft. Accessed: Jun. 13, 2025. [Online]. Available: https://github.com/microsoft/GlobalMLBuildingFootprints

[32] "Geofabrik Download Server." Accessed: Jun. 13, 2025. [Online]. Available: https://download.geofabrik.de/north-america/us.html

[33] "Microsoft Buildings Footprint Training Data with Heights - Overview." Accessed: Jun. 13, 2025. [Online]. Available: https://www.arcgis.com/home/item.html?id=3b0b8cf27ffb49e2a2c8370f9806f267

[34] "US Building height." figshare. Accessed: Jun. 13, 2025. [Online]. Available: https://figshare.com/s/e412c045834cf7d4c211

[35] "County of Los Angeles Open Data." Accessed: Jun. 13, 2025. [Online]. Available: https://data.lacounty.gov/

[36] "Building Footprints | DataSF." Accessed: Jun. 13, 2025. [Online]. Available: https://data.sfgov.org/Geographic-Locations-and-Boundaries/Building-Footprints/ynuv-fyni/about_data

[37] O. US EPA, "Smart Location Mapping." Accessed: Jun. 13, 2025. [Online]. Available: https://www.epa.gov/smartgrowth/smart-location-mapping

[38] "Nearby Search (New) | Places API," Google for Developers. Accessed: Jun. 13, 2025. [Online]. Available: https://developers.google.com/maps/documentation/places/web-service/nearby-search

[39]“Get a route | Routes API,” Google for Developers. Accessed: Jun. 13, 2025. [Online]. Available: https://developers.google.com/maps/documentation/routes/compute_route_directions
[40]“PolicyMap - Dig Deeper.” Accessed: Jun. 13, 2025. [Online]. Available: https://www.policymap.com/newmaps/e/washington

APPENDIX – TABLE INDEXES

APPENDIX – CHART INDEXES