Office Space and Gentrification in King County: A Machine Learning Based Approach

Ethan Cano

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

Master of Architecture

University of Washington
2025

Committee:

Narjes Abbasabadi

Vince Wang

Tomás Mendez Echenagucia

Program Authorized to Offer Degree:

Department of Architecture, College of Built Environments

University of Washington

Abstract

Office Space and Gentrification in King County: A Machine Learning Based Approach

Ethan Cano

Chairs of the Supervisory Committee:

Narjes Abbasabadi
Department of Architecture

Ruoniu (Vince) Wang
Runstad Department of Real Estate

In recent years, King County has seen a surge in housing costs and unaffordability, leading to a housing crisis. Local newspapers and community groups have pointed to large companies and the highly paid employees they attract as responsible for unaffordability and gentrification in the area. Though intuitively this may appear to be the case, this thesis uses newly available Machine Learning (ML) technology to quantitatively investigate the importance of offices as they correlate with gentrification. Making use of data made available through the American Community Survey and King County GIS, patterns of gentrification examined from 2010 to 2019. Following this, demographic and aggregated office-related variables are established at the block group level, and new GPU-boosted methods of performing ML and SHAP analysis are used to investigate the level to which office-related variables, such as taxable land value, office age, etc., are correlated with a prediction of gentrification within a block group.

Table of Contents

List of Tables and Figures

Dedication

For Dev, who supported me through it all, I love you so much. Mom, Em, Ellie, Grandma, and the village that raised me: I love you all more than I can express, thank you for getting me here. To Ms. Jameson, who helped cultivate my desire to learn, thank you so much for your guidance in my younger years. Thank you to Valerie Greer and the rest of the faculty at the University of Utah who helped me get to this point, and to the friends who I made there. Thank you to Tomas, Narjes, and Vince for all of your help in making this dream a reality, and thank you to the rest of the faculty at the University of Washington who have helped me along in my journey. Finally, thank you to my lovely friends from my cohort who have made these last two years so much fun; I appreciate you as well.

Introduction

In the midst of a housing crisis where, in 2021, 76% of households in King County were paying more than 30% percent of their income[1] toward housing costs, it becomes imperative to investigate the underlying causes of this issue and work with policymakers to combat the problem. However, policymakers and non-profits have limited time and resources to contribute to combating these pressing issues[2]. As such, identifying areas of highest need for assistance in the fight to maintain affordability can help policymakers to best use their time and resources. The arrival of high-paying companies and their new highly skilled workers, specifically in tech, is often associated with neighborhood change and decreased affordability[3], though this relationship is often tenuously established with quantitative information.

The research conducted in this thesis is done in service of policymakers in the process of making decisions with relation to the protection and edification of affordable housing in King County. Specifically, using machine learning (ML) techniques, this thesis will interrogate the connection between office spaces and the gentrification of the surrounding area using a variety of variables to describe office spaces within a Census Block Group. In doing so, this thesis attempts to answer the question of the importance of office-related variables as they correlate with gentrification in King County from 2010 to 2019. With this information, the research will be used as a tool to describe the characteristics of block groups (and the offices within them) that indicate gentrification could be on the horizon.

Machine learning techniques present an exciting new opportunity for analysis of the conditions in King County. With large repositories of data from the King County Open GIS Portal[4], the King County Assessor's Office[5], and the American Community Survey (ACS)[6], a wealth of information is available for analysis down to geographical areas as small as a parcel for the King County Open GIS Portal and a census block group for the ACS. Though the emergence of these big data sources and machine learning technologies allow for more complex analysis of the problem of gentrification, up to this point, studies on the relationship between office space and gentrification have been limited and have not used machine learning as a method of analysis[7]. New Python packages have also allowed us to better see inside the black box of machine learning algorithms and further examine the contributions of individual features to ML model decisions[8]. Additionally, new technologies brought forth through the cuML Python package in the Rapids Python Environment[9] have allowed for GPU-boosted processing of

[1] Bick, "How to Tackle Lack of Regional Affordable Housing?"
[2] Kingston and Bolton, "New Approaches to Funding Not-for-profit Organisations."
[3] "Seattle City Council Introduces New Affordable Housing Policy Options – Puget Sound Sage."
[4] "King County GIS Open Data."
[5] "King County Department of Assessments: eReal Property."
[6] Manson et al., "IPUMS National Historical Geographic Information System: ACS 5-Year Estimate 2006-2010"; Manson et al., "IPUMS National Historical Geographic Information System: ACS 5-Year Estimate 2015-2019."
[7] Freyd, "Large Tech Office Openings and the Onset of Gentrification."
[8] Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.mol
[9] "Welcome to cuML's Documentation! — Cuml 25.04.00 Documentation."

machine learning tasks, feature importances, and feature interactions, making analysis of these features significantly faster.

While there have also been quantitative studies in Seattle looking into gentrification trends, these focused on gentrification as it relates to light rail[10] and did not make use of machine learning for analysis. Other studies have made use of ML to predict areas of gentrification using tweets[11], green infrastructure[12], or general census data[13]. However, they have not used ML with a focus on office space and gentrification, and they have not analyzed gentrification in King County. Therefore, with the lack of coverage on machine learning for gentrification analysis in King County, and the lack of machine learning analysis on the relationship between gentrification and office spaces, this thesis seeks to fill in this gap to help local policymakers.

A local NGO in Seattle, Puget Sound Sage, speaks on the arrival of new, high-paying, and highly skilled tech workers in the area, resulting in rapid increases in rents and housing costs.[14] The arrival of a new, highly skilled workforce resulting in rapid shifts in neighborhood costs and demographics begins to point toward the arrival of these corporations and their highly skilled laborers as a gentrifying force. This thesis will also investigate the relationship between tech companies and gentrification quantitatively, in order to better understand the basis for these claims.

Literature Review

Defining and Measuring Gentrification

In order to properly assess gentrification, it is important to have a robust background on how gentrification is defined and put into operation in other studies on the topic. In defining gentrification, it is important to understand that gentrification does not have a solidified definition. As such, the definition of gentrification varies from study to study. When looking at studies from the Drexel University Urban Health Collaborative, gentrification is defined across three categories: compositional changes in neighborhood demographics, increased property values, and physical signs of reinvestment[15]. Broader definitions of gentrification speak of the process by which higher income individuals displace people with lower incomes, and the repurposing of neighborhoods to make them more appealing to this new clientele[16]. Others define gentrification as the physical and social transformation of low-income neighborhoods as the result of in-migrating middle and upper class households[17] or speak of the visual improvements

[10] Hess, “Light-Rail Investment in Seattle.”
[11] Chapple et al., “Monitoring Streets through Tweets.”
[12] Assaad and Jezzini, “Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach.”assa
[13] Thackway et al., “Building a Predictive Machine Learning Model of Gentrification in Sydney.”
[14] “Seattle City Council Introduces New Affordable Housing Policy Options – Puget Sound Sage.”
[15] “A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US.”
[16] Palafox and Ortiz-Monasterio, “Predicting Gentrification in Mexico City Using Neural Networks.”
[17] Assaad and Jezzini, “Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach.”

to businesses or housing stock[18]. In these definitions, it is also important to note that gentrification is not a singular event, but a process[19]. Due to the gradual nature of gentrification, the quantitative study of gentrification should involve analysis that looks across years to examine the changes in a neighborhood.

When assessing these definitions of gentrification, one can extract variables to analyze quantitatively, and it is necessary to find easily quantifiable areas of interest in order to perform machine learning operations on large datasets. For example, variables such as the median income of an area or the average price of a house are easily quantifiable, while the changing visual composition of a neighborhood is not. Looking into the variables also requires the delineation of areas of study into categories such as “non-gentrifiable”, “gentrified”, or “gentrifying”. By separating these areas into categories, our study can more easily identify sites of risk.

The separation of areas categorically has precedent across other studies on related topics. When looking at the work of Palafox and Ortiz-Monasterio, areas of analysis in Mexico city are separated into the categories of “non-gentrifiable”, “non-gentrifying”, and “gentrifying”[20], however the measures that stratify areas of analysis into these categories are not shared. In the research carried out by Wang and Melton-Fant, areas of analysis are split into “gentrifiable” and “non-gentrifiable” categories, followed by a second division of “gentrifiable” tracts into “gentrified” or “non-gentrified” categories[21]. Their research also defines these categories more specifically. A gentrifiable tract is defined as a tract that was below the citywide median household income in the Census years of analysis, while a non-gentrifiable tract was not[22]. Further, a gentrified tract was defined as an area that saw a gross rent or median home value increase above the citywide median increase, and an increase in college educated residents above the citywide increase between the start and end of a study period[23]. The research also indicated the relative “age” of gentrification, breaking down gentrified areas into the categories of old (gentrified from 1990 to 2000, but not 2000-2019), recent (gentrified from 2000-2019, but not in the previous decade), and continued gentrification (tract underwent gentrification in every time period)[24].

Drexel University’s Urban Health Collaborative differs from previous examples in that their team broke out gentrified areas into the categories of “eligible, but did not gentrify”, “evidence of gentrification”, and “evidence of intense gentrification”[25]. This adds a further level of refinement to the category of gentrification, with intense gentrification being added to provide an even more robust glimpse into gentrified areas. In order to differentiate these categories, the team at Drexel utilized quartiles for analysis: ineligible tracts could not have <50 people or be in

---

[18] Reades, De Souza, and Hubbard, “Understanding Urban Gentrification through Machine Learning.”
[19] Chapple et al., “Monitoring Streets through Tweets.”
[20] Palafox and Ortiz-Monasterio, “Predicting Gentrification in Mexico City Using Neural Networks.”
[21] Wang and Melton-Fant, “Does Inclusionary Housing Alleviate the Negative Health Impacts of Gentrification?”
[22] Wang and Melton-Fant.
[23] Wang and Melton-Fant.
[24] Wang and Melton-Fant.
[25] “A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US.”

the top quartile of median household income, not gentrified tracts had below median increases in proportions of residents with college degrees or below median increases in home values and below median increases in gross rent, gentrifying tracts saw above median increases in the proportion of residents with college education and 50th to 75th percentile increases in gross rent or home value, and tracts experiencing intense gentrification saw above median increases in proportion of college education and 75th percentile or higher increases in gross rent or home value[26].

In the work of Chapple et al., researchers split gentrifiable areas out based on other factors. A tract eligible for gentrification to occur, by their definition, required affordable housing (rents or housing values under the area median) and at least two out of the four following demographic characteristics (relative to the median): low college education, and high amount of low-income households, renters, or non-white households[27]. Neighborhood change was then characterized using the terms "gentrification" (the influx and investment of people in low-income areas), "displacement" (the out-migration of low-income households in low-income areas without their replacement), and "exclusion" (displacement taking place in high income neighborhoods)[28]. Gentrification is measured using the aforementioned markers, while displacement and exclusion are measured using the absolute loss of low-income households between census years and decrease in in-migration[29].

Lastly, gentrification has also been measured using available indices from the government[30]. In Australia, the Socioeconomic Index for Advantage and Disadvantage (SEIFA) from the Australian Bureau of Statistics was used by Thackway et al. in their study of gentrification in Sydney[31]. This index combined a set of 21 variables into a robust measure of the economic situation of an area, allowing gentrification to be ascertained by looking at the standard deviation of change from the 2011 SEIFA rank to the 2016 SEIFA rank of an area[32]. If the SEIFA rank of an area increased by 1 standard deviation, the research team considered the area to be gentrifying, while less than 1 standard deviation was considered to be "noise".[33]

Office Space, the Tech Industry, and Gentrification

As previously mentioned, an influx of tech companies and workers in the Seattle area has led to concern from community organizations about gentrification associated with the arrival of these new, highly skilled workers[34]. When looking at the rapid arrival of approximately 15,000 (mostly high-salaried) employees to the city to support the operations of Amazon, these fears do

---

[26] "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US."
[27] Chapple et al., "Monitoring Streets through Tweets."
[28] Chapple et al.
[29] Chapple et al.
[30] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[31] Thackway et al.
[32] Thackway et al.
[33] Thackway et al.
[34] "Seattle City Council Introduces New Affordable Housing Policy Options – Puget Sound Sage."

not appear to be unfounded[35]. Amazon is also far from the only tech-based employer in the city of Seattle. Other companies, such as Microsoft, Tableau, Facebook, Google, Adobe, and Zillow, have all settled in Seattle and other parts of King County, bringing an army of workers, anti-union sentiment, and environmental issues to boot[36]. These companies bring not only workers with high salaries that could disrupt and gentrify areas, but also are often accused of not "fitting in" with the local culture or disrupting local social infrastructure[37].

To look into the social and cultural disruption caused by the arrival of tech workers to an area, the article from Atkins[38], as well as papers by Goldman[39] and Maharawal[40] turned to another tech hub, San Francisco, to analyze and unpack tech-based gentrification in the city. In "San Francisco's Dot Com Boom 2.0", Goldman speaks on the neoliberal economic policies targeted at the privatization and dismantling of public services that come through gentrification, resulting in roll-backs of services essential to the functioning of these lower-income communities[41]. Goldman quotes Smith[42], citing that gentrification is a neoliberal, "revanchist"[43] process by which poor people are dispossessed of their land. Maharawal calls the dispossession of poor communities from their land as a result of tech-based gentrification "tech colonialism", and details some of the issues that arise from the arrival of wealthier, whiter groups to these areas[44]. Maharawal's work is supported by the work of Hess, who details how white residents do not tend to integrate into the existing social fabrics of the neighborhoods that they "colonize"[45]. This can be seen in the way white residents tend to see gentrifying, predominantly low-income and minority neighborhoods as empty swaths of land to be commoditized instead of reckoning with the existing social networks and communities that have inhabited these "ancestral" lands[46]. An example of this is provided at San Francisco's Mission Playground. Maharawal details the Mission District's primarily Latin community, and the arrival of tech workers as a result of companies like Dropbox settling in the area[47]. Upon arrival, these tech workers began grabbing up use permits for soccer fields and attempting to kick local, BIPOC children off of the playfields[48]. These playfields were traditionally open and available for the children of the neighborhood to use without a permit, however, due to a lack of social integration into the existing network of the community, these predominantly white tech workers quarreled with the

35 Atkins, "The Backlash Against Tech in Seattle."
36 Atkins.
37 Atkins.
38 Atkins.
39 Goldman, "THE 'GOOGLE SHUTTLE EFFECT:' GENTRIFICATION AND SAN FRANCISCO'S DOT COM BOOM 2.0."
40 Maharawal, "Tech-Colonialism."
41 Goldman, "THE 'GOOGLE SHUTTLE EFFECT:' GENTRIFICATION AND SAN FRANCISCO'S DOT COM BOOM 2.0."gol
42 Smith, *The New Urban Frontier: Gentrification and the Revanchist City*.
43 Smith.
44 Maharawal, "Tech-Colonialism."
45 Hess, "Light-Rail Investment in Seattle."
46 Maharawal, "Tech-Colonialism."
47 Maharawal.
48 Maharawal.

youth over the issue of permitting, leading to the “Mission Playground Is Not For Sale” movement in San Francisco[49].

The findings of Freyd[50] and Goldman[51] demonstrate the economic issues associated with the arrival of big tech companies. Freyd details the disproportionate increase in home prices within 1 km of tech offices[52], while Goldman speaks of the disproportionate increase in the rent prices of 1 and 2 bedroom units within a 5 minute walkshed of the Google shuttle stops in San Francisco[53]. Delving further into issues caused by gentrification across the social and economic spectrum, Kohn details five primary harms of gentrification, namely: residential displacement, exclusion, transformation of public, social, and commercial space, polarization, and homogenization[54]. Kohn highlights residential displacement as the most harmful feature, and exclusion as harmful, but to a lesser extent[55]. Kohn describes how residential displacement as a result of gentrification is a form of “bad luck” for the existing residents in a neighborhood, and describes policies that could be implemented to help socialize this misfortune instead of forcing the lowest income residents in an area to bear the brunt of the burden[56].

However, gentrification can also provide benefits to the areas in which it occurs. Kohn touches on the points of transformation of public, social, and commercial space, polarization, and homogenization as factors that can benefit or harm an area depending on the preexisting conditions of the neighborhood[57]. For example, if robust anti-displacement frameworks are in place, upgrading the infrastructure of a neighborhood may allow existing residents to access better resources. Though these upgraded resources can be beneficial, Kohn also warns of neighborhoods becoming unwelcoming or alien to existing residents as these upgrade occur[58]. In 2019, a study from Brummet and Reed attempted to create the first causal demonstration of how gentrification benefits and harms adults and children who originally inhabit gentrifying neighborhoods[59]. Though Kohn speaks of the harms of residential displacement[60], Brummet and Reed explain that many original residents are able to remain in their neighborhoods throughout the process of gentrification and partake in the benefits that arise from it[61]. Specifically, children in these gentrifying neighborhoods are more likely to attend and complete college, and end up in

---

49 Maharawal.
50 Freyd, “Large Tech Office Openings and the Onset of Gentrification.”
51 Goldman, “THE ‘GOOGLE SHUTTLE EFFECT:’ GENTRIFICATION AND SAN FRANCISCO’S DOT COM BOOM 2.0.”
52 Freyd, “Large Tech Office Openings and the Onset of Gentrification.”
53 Goldman, “THE ‘GOOGLE SHUTTLE EFFECT:’ GENTRIFICATION AND SAN FRANCISCO’S DOT COM BOOM 2.0.”
54 Kohn, “What Is Wrong with Gentrification?”
55 Kohn.
56 Kohn.
57 Kohn.
58 Kohn.
59 Brummet and Reed, “The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children.”
60 Kohn, “What Is Wrong with Gentrification?”
61 Brummet and Reed, “The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children.”

neighborhoods with lower poverty rates, higher levels of college-educated residents, and more employed residents[62].

Upon reviewing literature on the topic of gentrification, it is apparent that the benefits and harms of gentrification are not universally agreed upon, but instead hotly contested. Though the issues are subjects of much debate, they provide us with a framework for analysis, demonstrating the ills of gentrification, including polarization[63], lack of social cohesion[64], and residential displacement[65], and its benefits to existing residents and their children[66]. The topics of debate demonstrate the need for analysis into the relationship between office spaces and gentrification, as they may provide us with key indicators of areas that are likely to gentrify, in turn allowing policymakers to have more time to plan ahead and prepare these areas for future bouts of gentrification.

Machine Learning as a Means of Analysis

When looking at the current landscape of gentrification prediction, Thackway speaks of indices such as the Displacement Risk Index in Seattle or the Index of Neighborhood Change in LA[67]. Both of these indices allow for the prediction of areas of risk in an area, however Thackway deems both to be largely reliant on rudimentary statistical evaluations and qualitative data, leading to inaccuracies and lack of trust from government decision makers[68]. Thackway also speaks of the use of linear regression as a method of analysis for these predictive tasks, but details the complexity of gentrification as a problem and the possibility of collinearity arising from related factors such as education, income, or occupation, indicating that simple linear regression does not provide a sufficiently robust form of statistical analysis[69]. Instead, the study suggests that machine learning (ML) be used to tackle predictive problems associated with gentrification, as ML methods allow researchers to parse long and wide datasets, leading to more comprehensive and accurate predictive results[70]. The work of Assaad and Jezzini provides further emphasis on the use of machine learning, highlighting the ability of machine learning techniques in handling complex, nonlinear patterns from large datasets[71]. The pair also underscore the scarcity of research using ML to predict green gentrification, and the pressing need for additional research using ML in order to improve urban resilience, equity, and access[72].

---

[62] Brummet and Reed.
[63] Kohn, "What Is Wrong with Gentrification?"
[64] Maharawal, "Tech-Colonialism."
[65] Kohn, "What Is Wrong with Gentrification?"
[66] Brummet and Reed, "The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children."
[67] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[68] Thackway et al.
[69] Thackway et al.
[70] Thackway et al.
[71] Assaad and Jezzini, "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach."
[72] Assaad and Jezzini.

The scarcity of research on ML and prediction of green gentrification extends to ML as a form of analysis for the relationship between gentrification and office spaces. Freyd's work on tech offices and home price increases begins to pick apart the idea of office space as it relates to gentrification, however, as previously mentioned, gentrification is not solely defined by increases in home prices[73]. Gentrification extends beyond only home price increases to increases in rent, changes in demographics, and displacement of existing community members. For this reason, extending the study to include more than just home price increase will provide a more robust interrogation of gentrification as it relates to office spaces. Additionally, Freyd's research analyzes home price increases from tech office arrivals through a difference-in-difference analysis instead of incorporating ML techniques[74]. The lack of ML in the study done by Freyd opens an opportunity for a new and scarcely researched direction to analyze the problem of gentrification as it relates to office buildings.

Other analyses, such as those done by Reades[75], Thackway[76], Chapple[77], Palafox[78], and Assaad[79] have all made use of ML techniques to analyze the problem of gentrification, however none of these projects have looked at gentrification through the lens of office buildings like Freyd[80]. Though these papers have not focused their attention on office buildings, the lessons learned about ML techniques are applicable to the research in this thesis. The analysis performed by Reades utilized Random Forest (RF) regression, a machine learning technique that arrives at predictions by sourcing predictive power from the fittest trees in a "forest" of decision trees[81]. Citing the work of Reades, Thackway speaks of the validity of RF analysis for predicting gentrification, but advocates instead for the emerging technology of Gradient Boosted Machines (GBMs) in ML-based analysis[82]. The method employed by GBMs is similar to that of RFs, however, in place of sourcing data randomly from the fittest trees, the best performing trees in GBM models are used to create the next generation of trees in the decision making process, allowing for the refinement of predictive metrics from generation to generation of prediction[83]. Apart from GBMs, Palafox and Ortiz-Monasterio have used Neural Networks to analyze gentrification in Mexico City[84], while Assaad and Jezzini analyzed gentrification in NYC using K-Nearest Neighbors, Decision Trees, Random Forests, Artificial Neural Networks (ANNs), Logistic Regression, and Naive Bayes algorithms, in order to test the effectiveness of each ML

[73] Freyd, "Large Tech Office Openings and the Onset of Gentrification."
[74] Freyd.
[75] Reades, De Souza, and Hubbard, "Understanding Urban Gentrification through Machine Learning."
[76] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[77] Chapple et al., "Monitoring Streets through Tweets."
[78] Palafox and Ortiz-Monasterio, "Predicting Gentrification in Mexico City Using Neural Networks."pala
[79] Assaad and Jezzini, "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach."
[80] Freyd, "Large Tech Office Openings and the Onset of Gentrification."
[81] Reades, De Souza, and Hubbard, "Understanding Urban Gentrification through Machine Learning."
[82] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[83] Thackway et al.
[84] Palafox and Ortiz-Monasterio, "Predicting Gentrification in Mexico City Using Neural Networks."

method[85]. In Assaad and Jezzini's study, the Artificial Neural Network appeared to produce the most accurate results[86], indicating that ANNs and GBMs appear to provide the most robust analysis with regard to predictive tasks involving gentrification. To further the accuracy of these ML methods, the works of Reades[87], Assaad[88], and Palafox[89] also indicate the need for K-Fold Cross-Validation in order to tune the hyperparameters of each of these models when performing predictive tasks.

Shap Values

The previously mentioned studies surrounding ML and gentrification have focused on using ML to predict areas of gentrification in the future. However, in this thesis, the research does not focus on future forecasting, but rather using ML on past data to examine which features were most important in pushing a block group toward a gentrification prediction. ML produces powerful predictive models, however they often appear to be "black boxes" taking in and outputting information without explanation as to why[90]. In order to interpret these models, Lundberg and Lee released the SHapley Additive exPlanations (SHAP) Python package in 2017[91]. This package uses Shapley values, which are based in game theory, to determine the average marginal contribution of each feature to a machine learning model[92]. By doing this, we can determine how important a feature within the model is, and how much it contributes to a gentrification prediction, which can serve as an analog for correlation. Using SHAP for feature importance analysis, and as an analog for correlation, has been done before in the field of materials science, where SHAP was used to analyze which features most important in determining compressive strength of an alkali-activated material[93]. Though not in the field of real estate and urban studies, this approach does have precedent and can easily be applied to a study of gentrification. SHAP values also prove useful beyond simple correlation analysis by providing insight into the interactions between features[94]. SHAP allows for a more robust understanding of where features interact with each other to produce higher or lower SHAP values[95]. Molnar demonstrates this through an example where the sex of a penguin is being predicted[96]. In this

[85] Assaad and Jezzini, "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach."
[86] Assaad and Jezzini.
[87] Reades, De Souza, and Hubbard, "Understanding Urban Gentrification through Machine Learning."
[88] Assaad and Jezzini, "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach."
[89] Palafox and Ortiz-Monasterio, "Predicting Gentrification in Mexico City Using Neural Networks."
[90] Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.
[91] Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions."
[92] Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.
[93] Zheng et al., "A Data-Driven Approach to Predict the Compressive Strength of Alkali-Activated Materials and Correlation of Influencing Parameters Using SHapley Additive exPlanations (SHAP) Analysis."
[94] Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.
[95] Molnar.
[96] Molnar.

example, SHAP interaction values are used to demonstrate that penguins with a lower body mass and longer bill tend to point to a prediction of male, while penguins with middling to higher masses and shorter bills tend to point toward a prediction of female[97].

Though SHAP values provide powerful insight into the inner workings of machine learning models, it is important to remember the difference between correlation and causation[98]. The SHAP package works well in revealing correlations within ML models, however this does not indicate that these correlations are guaranteed to continue into the future or that policymakers manipulating these features in the future will guarantee a change in future outcomes, as correlation does not equal causation[99]. It is also important to note that SHAP values are model-specific, meaning that SHAP values will change depending on the variables fed into an ML model[100]. This may also lead to confusion by overstating or understating the importance of a variable for a prediction if the model does not include a robust set of features[101]. Therefore, when analyzing an ML model using SHAP, it is important to include an extensive and robust set of features in order to ensure a more accurate representation of conditions. However, it must be noted that SHAP values still have limitations when assessing robust datasets.

Methodology

Operationalizing Gentrification

While all of the definitions of gentrification previously listed have merit, in this research, we sought to utilize the gentrification metric set forth by Drexel University's Urban Health Collaborative[102]. However, when putting this into practice, the multiclass approach created difficulties with computation due to imbalanced class data and the need for increased computational power to deal with the additional classes. To address this, the metrics created by Drexel's Urban Health Collaborative were collapsed down into a binary, with the ineligible and eligible, but did not gentrify[103] creating class "0" and the evidence of gentrification and evidence of intense gentrification categories[104] forming class "1". With the creation of two binary categories, computation was simplified, which eased issues associated with running ML models and computing SHAP and SHAP interaction values.

---

[97] Molnar.
[98] "Be Careful When Interpreting Predictive Models in Search of Causal Insights — SHAP Latest Documentation."
[99] "Be Careful When Interpreting Predictive Models in Search of Causal Insights — SHAP Latest Documentation."
[100] Hwang et al., "SHAP-Based Explanations Are Sensitive to Feature Representation."
[101] Hwang et al.
[102] "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US."
[103] "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US."
[104] "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US."

Data Collection - Block Groups

All of the studies performing statistical analysis or predictive tasks cited in this thesis have used data from the Census Bureau in their respective countries and areas of analysis. In turn, the analysis performed in this thesis will make use of data from the Census Bureau as well. In addition to this, a number of the cited studies, for example Thackway's study of Sydney[105] or Freyd's study of tech spaces[106], made use of data from other third-party sources, such as AirDNA or the Zillow ZTRAX repository. Though these external sources provide useful sets of data, the limited funding and time available for the research conducted in this thesis make the free and easily accessible data from the US Census and the American Community Survey (ACS) compelling. Therefore, in order to facilitate the process of analysis, data from the ACS will be used.

The Census data brings with it the issue of granularity. The works of Hess[107], Thackway[108], and Freyd[109] analyze Census data from the scale of the block group, while the research conducted by Reades[110], Chapple[111], Assaad[112], Wang[113], and Drexel's Urban Health Collaborative[114] analyzed data at the Census Tract level, indicating that both levels of granularity provide valid approaches to tackling the issue of granularity. The ACS readily provides data at both levels of granularity, however the increased level of granularity provided at the block group level proves beneficial for the research carried out in this thesis. The increased specificity of the block group allows for greater detail into smaller areas facing gentrification issues. In addition to this, ML models traditionally have thrived on larger datasets for training purposes[115], and the number of observations for analysis jumps from around 400 census tracts to over 1400 census block groups in King County, providing a more robust dataset for the model to work with.

As previously mentioned, definitions of gentrification describe change to an area over time, making gentrification a process, and not a singular event. As such, a baseline and analysis year must be chosen for the examination of gentrification. In this study, 2010 has been identified as the baseline year for examination. The year 2010 marked the first time a complete 5-year ACS estimate was made available to the public, which provides a more accurate and robust estimate of the conditions within a block group. The analysis year of 2019 was selected for the cutoff year in this study in order to avoid any possible noise derived from demographic and economic shifts

[105] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[106] Freyd, "Large Tech Office Openings and the Onset of Gentrification."
[107] Hess, "Light-Rail Investment in Seattle."
[108] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[109] Freyd, "Large Tech Office Openings and the Onset of Gentrification."
[110] Reades, De Souza, and Hubbard, "Understanding Urban Gentrification through Machine Learning."
[111] Chapple et al., "Monitoring Streets through Tweets."
[112] Assaad and Jezzini, "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach."
[113] Wang and Melton-Fant, "Does Inclusionary Housing Alleviate the Negative Health Impacts of Gentrification?"
[114] "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US."
[115] "Machine Learning Models Can Produce Reliable Results Even with Limited Training Data | University of Cambridge."

due to the COVID-19 pandemic. Selecting 2010 and 2019 for this study also proves useful, as both years use the 2010 Census block group boundaries, meaning the boundaries for block groups do not change from the baseline year to the year of analysis. Across the study, there are 1422 block groups ready for analysis within the boundary of King County.

Over 350 demographic and economic variables were collected from the 2010 and 2019 ACS via the NHGIS IPUMS database. A new variable was then formed to represent gentrification, with "0" representing an area ineligible to gentrify, "1" representing an area that was eligible, but did not gentrify, "2" representing an area that showed evidence of gentrification, and "3" representing an area that showed evidence of intense gentrification, as defined by the Drexel Urban Health Collaborative[116]. These were then collapsed down into binary categories, with the original class "0" and "1" categories constituting the new non-gentrified category ("0"), and the original class "2" and "3" categories constituting the new gentrified category ("1"). A codebook of the variables pulled from the 2010 ACS can be found in Appendix A[117], and the codebook for the 2019 ACS variables can be found in Appendix B[118]. Key variables will be discussed in later sections.

Data Collection - Offices

In order to analyze changes in these census block groups as they relate to office buildings, a comprehensive list of office buildings must be gathered for analysis. In order to compile this list, parcel data has been collected from the King County GIS department. The data received from King County GIS includes detailed parcel data from 2024, including the present use of the parcel with tags such as "condominium", "apartment", or "parking lot". Initially, the research in this thesis focused only on high tech spaces found with the tag "High Tech / High Flex," however, due to constraints with the data received from King County GIS, analysis was expanded to include labels containing the words "office" and "tech". This led research in the direction of not focusing on tech offices exclusively, but the largest offices in King County. Many of these offices are owned by large tech corporations, however the research focus was expanded to analyze office spaces in general instead of tech offices in specific. This selection process attempted to separate out a robust list of all office space from the list of all parcels within King County. Using the Playwright package in Python for automated web scraping, the properties were then fed into the King County Assessor's website, where information on the year built, taxable land value, taxable improvement value, and total taxable value was obtained. Though the list of offices came from 2024 parcel information, when scraping the Assessor's website, information was gathered on taxable land and improvement values from 2010. Taxable values were chosen in this instance to eliminate medical offices and non-profit institutions,

116 "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US."
117 Manson et al., "IPUMS National Historical Geographic Information System: ACS 5-Year Estimate 2006-2010."
118 Manson et al., "IPUMS National Historical Geographic Information System: ACS 5-Year Estimate 2015-2019."

which may not have the same effect as a more traditional office. After obtaining key information on all office spaces within King County, the parcel-level information was then aggregated to the block group level, creating the following key statistics:

"OFFICE_COUNT": The number of offices within a BG

"AGE_RANGE": Age of oldest office in BG minus the age of youngest office in BG
"MAX_AGE": Age of oldest office in BG
"MIN_AGE": Age of newest office in BG
"MEAN_AGE": Average age of office in BG
"MEDIAN_AGE": Median age of office in BG
"AGE_STD": Standard deviation of office ages in BG

"LAND_VALUE_RANGE": Office with highest taxable land value - Office with lowest taxable land value in a BG
"MAX_LAND_VALUE": Office with the highest taxable land value in the BG
"MIN_LAND_VALUE": Office with the lowest taxable land value in the BG
"MEAN_LAND_VALUE": Average office land value in the BG
"MEDIAN_LAND_VALUE": Median office land value in BG
"LAND_VALUE_STD": Standard deviation of office land values in the BG

"VALUE_RANGE": Office with highest taxable overall value - Office with lowest taxable overall value in a BG
"MAX_VALUE": Office with the highest taxable overall value in the BG
"MIN_VALUE": Office with the lowest taxable overall value in the BG
"MEAN_VALUE": Average overall office value in the BG
"MEDIAN_VALUE": Median overall office value in BG
"IMPR_STD": Standard deviation of overall office values in the BG

"IMPR_VALUE_RANGE": Office with highest taxable improvement value - Office with lowest taxable improvement value in a BG
"MAX_IMPR_VALUE": Office with the highest taxable improvement value in the BG
"MIN_IMPR_VALUE": Office with the lowest taxable improvement value in the BG
"MEAN_IMPR_VALUE": Average office improvement value in the BG
"MEDIAN_IMPR_VALUE": Median office improvement value in BG
"IMPR_VALUE_STD": Standard deviation of office improvement values in the BG

By aggregating the thousands of office points to block group-specific statistics, block groups and the offices within them were more easily compared with each other. However, another key limitation of the research conducted in this study lies in additions, expansions, or renovations to these properties. The expansion or addition of new properties to these sites often increase the capacity of employees an office space can host, which could lead to an influx of new talent. However, including new additions into the study leads to a level of complexity that creates difficulty when web scraping information from the assessor's web page. Due to the limited time frame of this project, only the initial year these office spaces were built is assessed, though a longer study in the future could investigate these additions and expansions more deeply.

Selection of ML Algorithm

Due to the limited resources available for this project, two tools were necessary to bridge the gap in computational power for machine learning and SHAP analysis. First, the RAPIDS Python environment allowed for processing of ML tasks with the GPU acceleration in place of the CPU, allowing for parallel processing and significantly faster compute times. Within RAPIDS, the cuML package specifically allowed for this quicker processing. cuML does not provide Gradient Boosted Machines (GBMs) or Neural Networks (NNs) which stood out as being more robust when used for gentrification-related tasks[119]. However, cuML does provide a Random Forest Classifier (RF), which has previously been used for ML tasks related to gentrification[120]. Though not necessarily as robust as other ML methods like GBMs or NNs, RFs still provide a valid approach to tackling the issue, as documented by previous studies. Therefore, when prioritizing shorter compute times for the abbreviated schedule of this project, cuML's Random Forest Classifier provided the best solution.

Tools Used

In addition to cuML and the RAPIDS environment, RunPod was utilized to remotely access stronger computer hardware for ML tasks. In the case of this project, RunPod allowed for remote access to an Nvidia RTX 5090, greatly improving the speed of ML tasks which previously would crash the kernel of locally available systems.

The RAPIDS environment comes loaded with various machine learning and data science tools for coding in the Python programming language. For handling of large datasets, the Pandas and Numpy packages were used to allow for better data management. The Playwright package has been used to handle web scraping tasks and collection of data, while the NHGIS IPUMS API allowed for the drawdown of ACS data from 2010 and 2019. GeoPandas was also used for geospatial data processing with relation to the selected census tracts and office spaces. SciKit Learn was incorporated alongside cuML for ML and interpretation tasks.

---

[119] Assaad and Jezzini, "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach"; Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."

[120] Reades, De Souza, and Hubbard, "Understanding Urban Gentrification through Machine Learning."

In their research, Palafox and Ortiz-Monasterio mention the need to help make complex sets of data more understandable and consumable to policymakers[121]. As the research in this thesis is being done in service of policymakers here in King County, selection of proper data visualization packages is important for the processing of the data extracted. In keeping with the work of Palafox, the SHAP, Seaborn, Matplotlib, and NetworkX packages were all employed to help produce graphics in order to make the technical information provided more digestible for policymakers.

Results and Discussion

After running a Random Forest Classifier with an X matrix consisting of independent variables and block group observations and a y matrix consisting of the dependent variable of binary gentrification classifications in 2019, the results of the Random Forest Classifier were validated using a Kendall's Tau[122] analysis. As this model was not being used for prediction, the train_test_split function[123] was not used, and the entire dataset was fed into the model[124], meaning evaluation metrics such as AUC or accuracy could not be used. In order to verify the model did not see wild swings in variable importance, the same Random Forest Classifier was run with multiple perturbations in random state seeds[125]. The SHAP results from these models were then compared using Kendall's Tau in order to judge the similarity in rankings of the variables and their SHAP values as a way to assess reliability[126]. In this analysis, 4 rounds of random seeding and Kendall's Tau consistently returned a value of 0.839 or higher, indicating that the model was returning consistent results. The variables identified in Table 1 were identified as the top contributors to a prediction of gentrification by their mean absolute SHAP values.

Table 1 - Top Overall SHAP Values (Mean Absolute Value)

Feature	Mean_Abs_SHAP
JOIE001 - Median Household Income	0.041886117
JSNE010- Renter occupied: no vehicle available	0.021109603
JTBE001 - Number of Renters	0.016174736
PERCENT_COLLEGE_ATTAINMENT - Percentage of people with college degrees	0.011207276
JLZE034 - Number of Females aged 22-24	0.009548083

121 Palafox and Ortiz-Monasterio, "Predicting Gentrification in Mexico City Using Neural Networks."
122 Kendall, "A New Measure of Rank Correlation."
123 "Train_test_split."
124 Shmueli, "To Explain or to Predict?"
125 Alvarez-Melis and Jaakkola, "On the Robustness of Interpretability Methods."
126 Alvarez-Melis and Jaakkola.

MEAN_LAND_VALUE	0.008476684
JTBE010 - Gross Rent as a percentage of income: 50% or more	0.006278855
MIN_LAND_VALUE	0.006241452
JSNE006 - Owner occupied, 3 vehicles	0.006120103
JTIE001 - Median Home Value	0.005699299

As would be expected, a few variables used to classify tracts as gentrified or not gentrified (household income, median home value, and percent college attainment) were in the top ten variables with the highest SHAP values. However, median gross rent fell to rank 26 in terms of SHAP values. Of note, the SHAP values of the top ten most important features varied by an order of magnitude from 1 to 10. Table 2 contains a list of the SHAP values for all office variables in the study, with a division between variables between rank 1 and 100, and variables between rank 101-398.

Table 2 - Top Office SHAP Values (Mean Absolute Value)

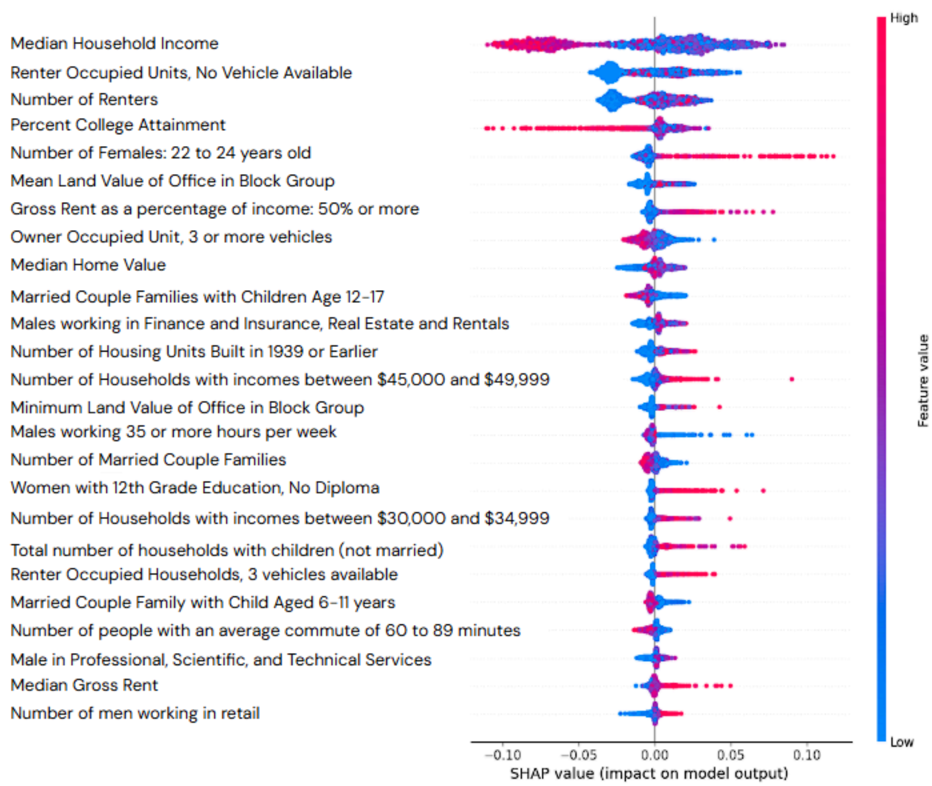
Rank	Feature	SHAP Value
6	MEAN_LAND_VALUE	0.008476684
8	MIN_LAND_VALUE	0.006241452
22	MIN_AGE	0.00289262
25	IMPR_VALUE_STD	0.002742971
39	MIN_VALUE	0.00213544
52	MIN_IMPR_VALUE	0.001778072
58	MEDIAN_LAND_VALUE	0.001689153
81	LAND_VALUE_STD	0.001418529
83	MAX_IMPR_VALUE	0.001394609
119	MAX_LAND_VALUE	0.001121822
148	MAX_AGE	0.000874964
153	AGE_STD	0.000860072
177	LAND_VALUE_RANGE	0.000783567
189	MEAN_VALUE	0.000737839
211	OFFICE_COUNT	0.000665506
237	IMPR_VALUE_RANGE	0.00059389
246	MEDIAN_VALUE	0.000575873
262	MEDIAN_AGE	0.000504665
289	MEAN_AGE	0.000437834

293	MEAN_IMPR_VALUE	0.000435001
305	MEDIAN_IMPR_VALUE	0.000401869
310	AGE_RANGE	0.00038594
320	MAX_VALUE	0.000355993
331	VALUE_RANGE	0.000323863
338	VALUE_STD	0.000311565

When looking at the top mean absolute SHAP values for office variables, we find 9 office variables within the top 100 most important variables, with 4 landing in the top 25, and 2 landing in the top 10. Though these mean absolute values are useful on their own, SHAP values provide even more detailed insight locally, helping to explain what factors contributed to the prediction of gentrification within a block group. Furthermore, with the above SHAP values being a mean absolute calculation of variable importance, local effects within specific ranges for a variable may be lost. SHAP again provides deeper insight by allowing us to see what local ranges contribute to the highest SHAP values for a variable.

Figure 1, a beeswarm plot for the top 25 most important variables by mean absolute SHAP, provides a window to begin looking into local effects of the variables fed into the ML model.

Figure 1 - Beeswarm Plot



A beeswarm plot displays the SHAP values for a variable based on all 1422 observations fed into the ML model. The plot in Figure 1 demonstrates an important characteristic of SHAP values in the ML model, where negative SHAP values indicate a push toward a prediction of not gentrified, while positive SHAP values demonstrate a push toward a gentrification prediction. In Figure 1, the variable with the highest importance based on mean absolute SHAP value, Median Household Income indicates that block groups with higher median incomes tend to produce negative SHAP values, while block groups with lower and middling median incomes tend to produce positive SHAP values. Intuitively, this makes sense, as areas with high median incomes are not eligible to gentrify per the aforementioned definition used to describe gentrifying and non-gentrifying areas. However, we can also look at variables outside of our definition, such as number of people with an average commute of 60-89 minutes, and see that block groups with higher numbers of people with a 60 to 89 minute commute tend to be less likely to receive a gentrification prediction from the model. This could be interpreted to indicate that people with longer commutes tend to live in wealthier suburbs further away from their jobs, or that people in these block groups tend to have more money to put into transportation to reach far away employment opportunities.

Beyond the beeswarm plot, SHAP also provides decision plots for individual observations (Block Groups) that give insight into why an observation leaned toward a prediction of gentrification or not. When looking at the top twenty most impactful variables for the BG in the observation plot below, the pink observation line for the BG is far to the right (ending at 0.70) of the baseline (at about 0.57) for an observation, indicated by the grey line. In this instance, the decision line sitting to the right of the baseline designates a higher likelihood of a prediction of gentrification, while a decision line to the left of the baseline designates a lower likelihood of a gentrification prediction. Among the top 20 variables contributing to a prediction of gentrification for this block group are 4 office-related variables, IMPR_VAL_STD, MIN_LAND_VALUE, MEAN_LAND_VALUE, and OFFICE_COUNT. From these office variables, we can see a high variance in office improvement values, with a standard deviation value of $87,260,200, indicating that the offices are not uniformly all high value or all low value, but have an imbalance, which could indicate a pricing gap, making the area ripe for gentrification. This idea of imbalance is further supported by the minimum and mean land values, where the minimum land value sits at $1,036,800, and the mean land value sits at $9,831,405, again indicating a higher value area where lower priced office space can be found. This BG also had an office count of 40, a relatively high amount when compared to other block groups, which could indicate that this area is a job hub and attracts new movement and investment into the area. Other notable demographic features from the ACS include a low median household income at $13,477 per year, a low percentage of people with a college degree (29.5%), and relatively inexpensive housing with a median home value of $350,000. Figure 3 highlights the BG from the observation plot, allowing for a view of where this observation sits within the county.

Figure 2 - Observation Plot

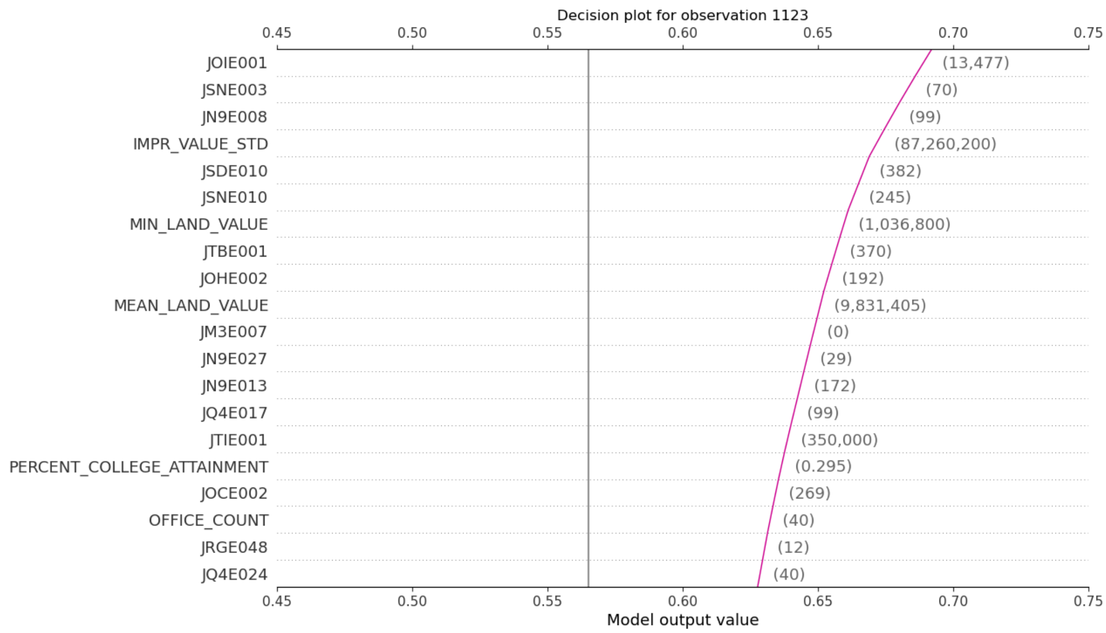
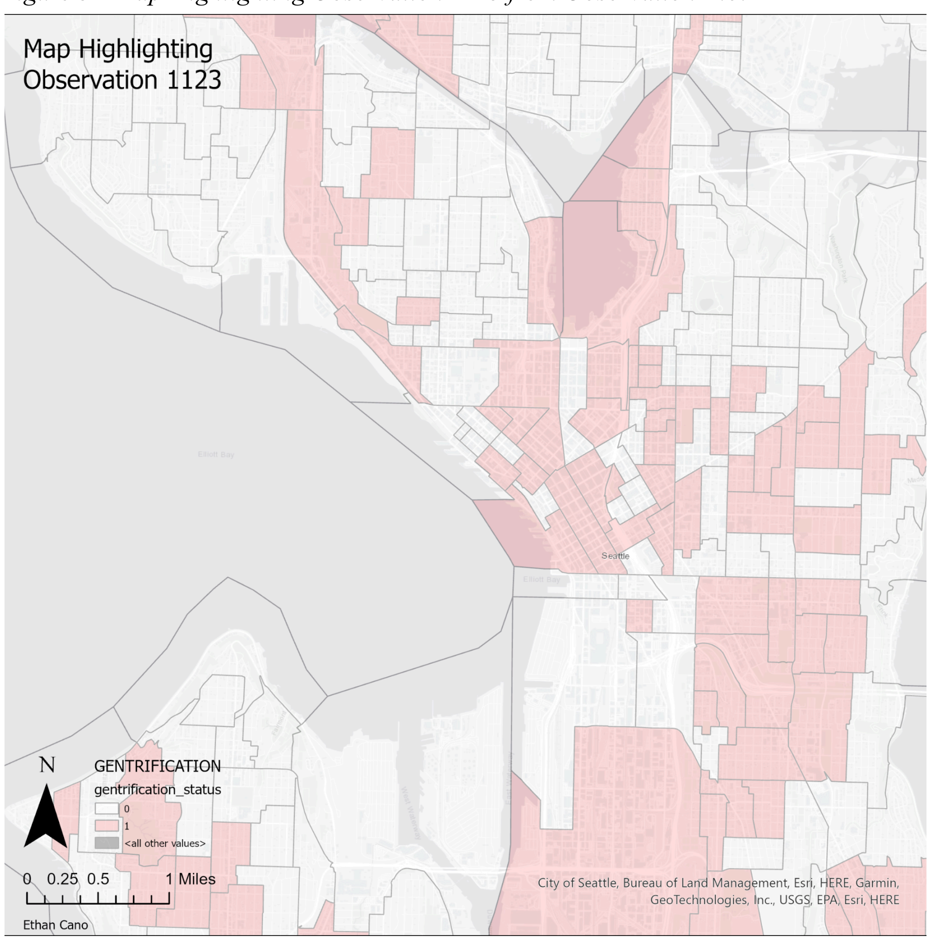
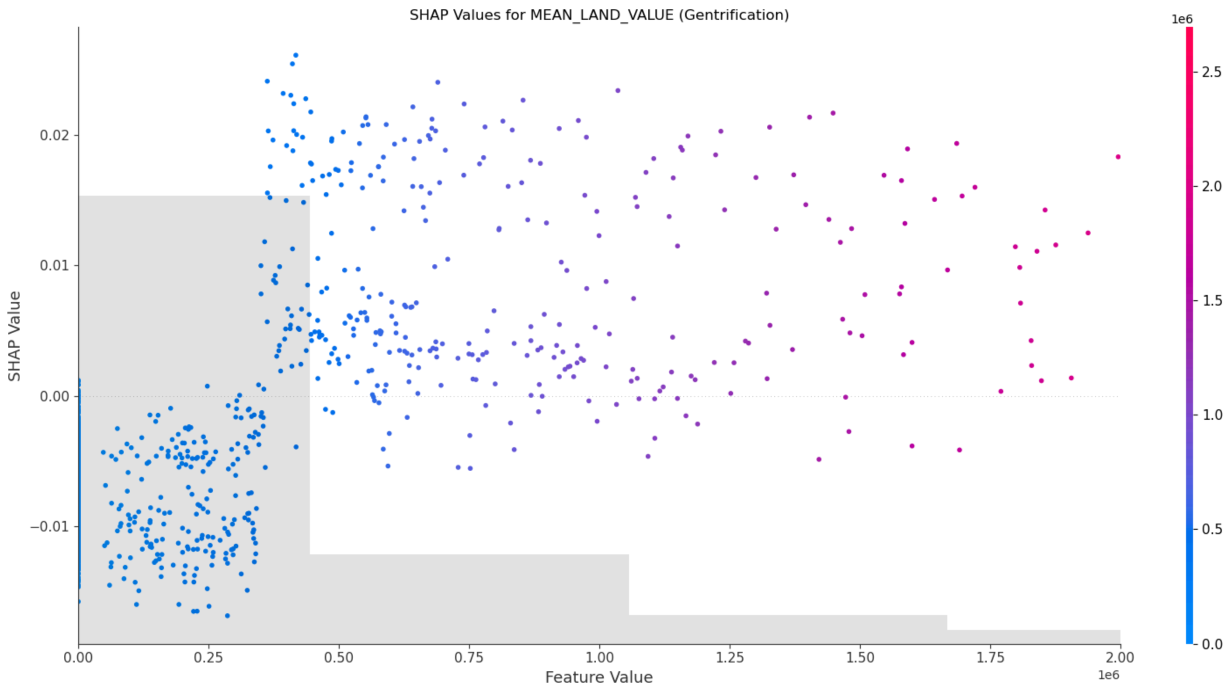


Figure 3 - Map Highlighting Observation 1123 from Observation Plot



The observation plot demonstrates certain ranges where variables tend to produce stronger pushes toward gentrification, as displayed by the particular values for office improvement standard deviation, office count, minimum land value, and mean land value. In order to investigate these particular values further, scatter plots can be generated to demonstrate the SHAP values for certain variables across different feature value ranges. The scatter plot in Figure 4 shows SHAP values on the y-axis and mean land values for offices within a BG on the x-axis.

Figure 4 - SHAP Scatter Plot for Mean Land Values of Offices within a BG



Upon examining the scatter plot, a shift becomes visible at approximately $350,000 in land value. Below a mean land value of $350,000 for offices within a block group, SHAP values tend to be negative, marking a push away from a prediction of gentrification, while above $350,000, SHAP values tend to be consistently positive, generally falling between 0 and 0.02. This may signify that areas with mean office land values below $350,000 have not quite reached the level of attractiveness to be considered undervalued by wealthier investors looking for value gaps.

Again, when examining the SHAP scatter in Figure 4, the magnitude of the values must be considered. In order to better understand the magnitudes displayed in scatter plots for office-related variables, 3 other scatter plots for demographic variables in the top 10 have been provided. Figure 5 displays the scatter plot for JOIE001, the top ranked variable by mean absolute SHAP. The scatter plot in Figure 6 provides information on JTBE001 (the number of renters in a BG), the 3rd ranked variable, and the scatter plot in Figure 7 represents the SHAP values for JTIE001 (median home value), the 10th ranked variable. These scatter plots related to demographic variables in the top 10 allow for a better understanding of the true significance of certain value ranges within additional SHAP scatters for office variables.When comparing results from the SHAP scatter for mean land values of offices within a BG, the values for this variable appear to generally have a smaller magnitude than that of a variable like median household income (Figure 5), which ranges from values of -0.1 to 0.075. The SHAP values for the mean land value of offices within a BG are, however, relatively comparable to the SHAP values for variables like number of renters in a BG (Figure 6), which ranges from approximately -0.04 to 0.04, or median home value (Figure 7), whose SHAP values range from -0.02 to 0.02.

Figure 5 - SHAP Scatter Plot for Median Household Income

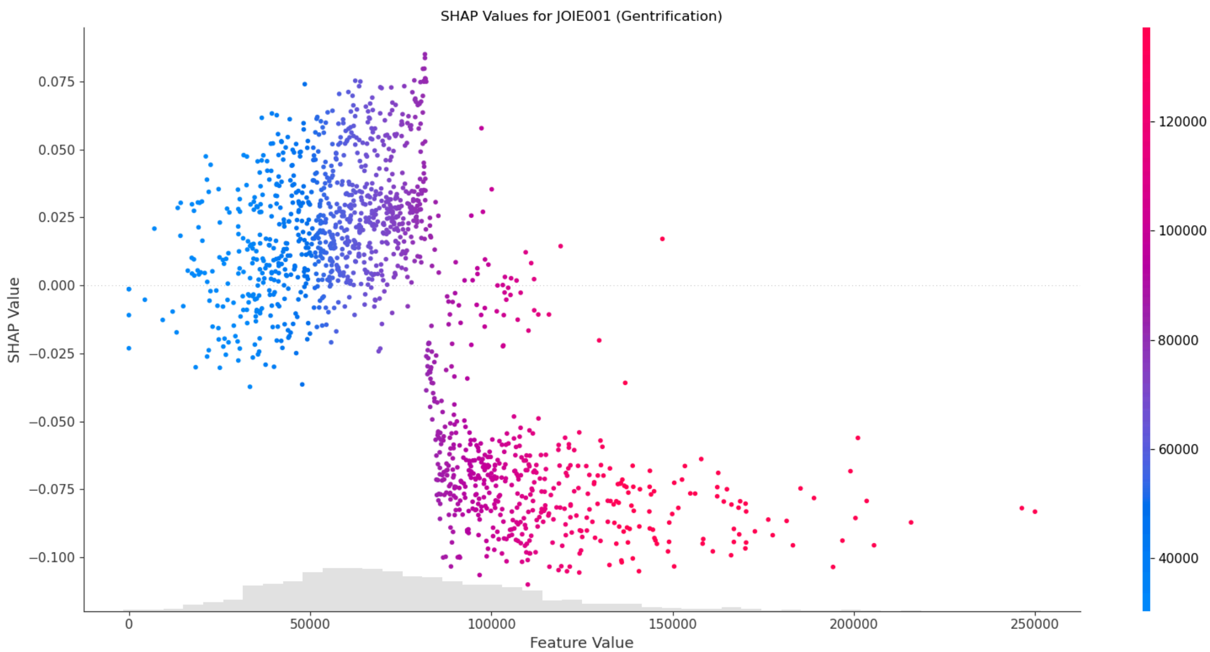
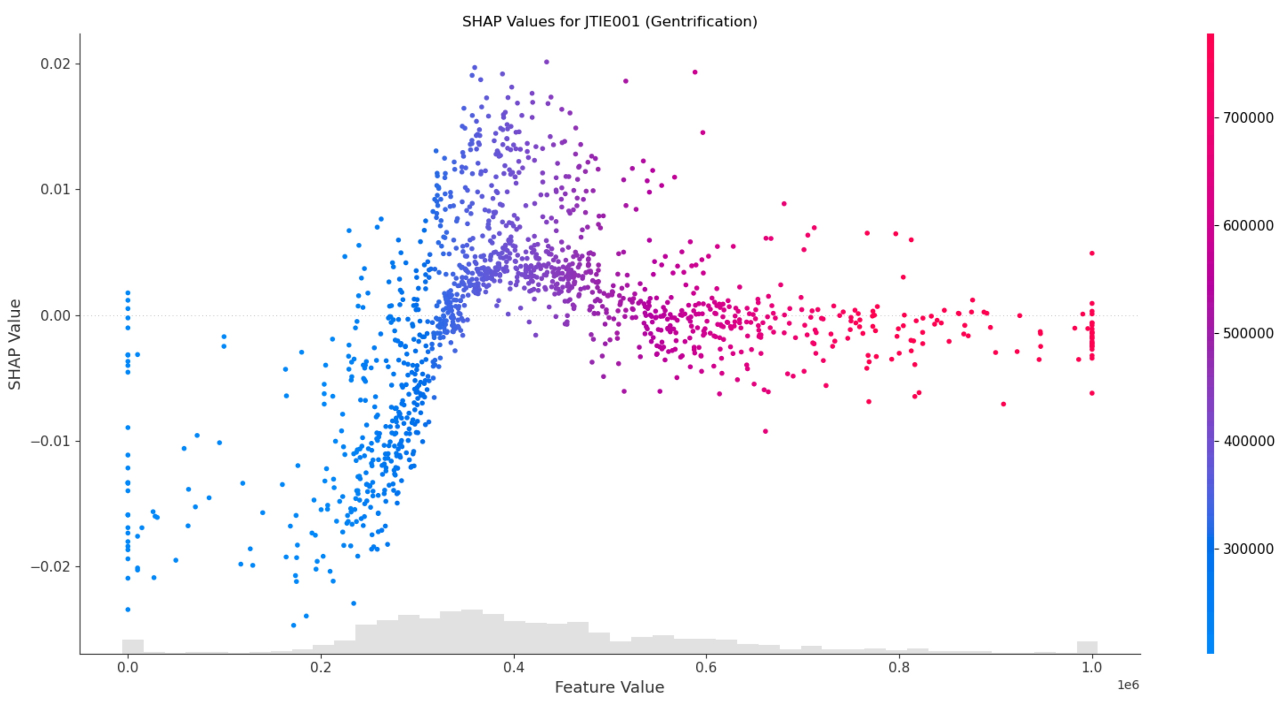


Figure 6 - SHAP Scatter Plot for Number of Renters in BG

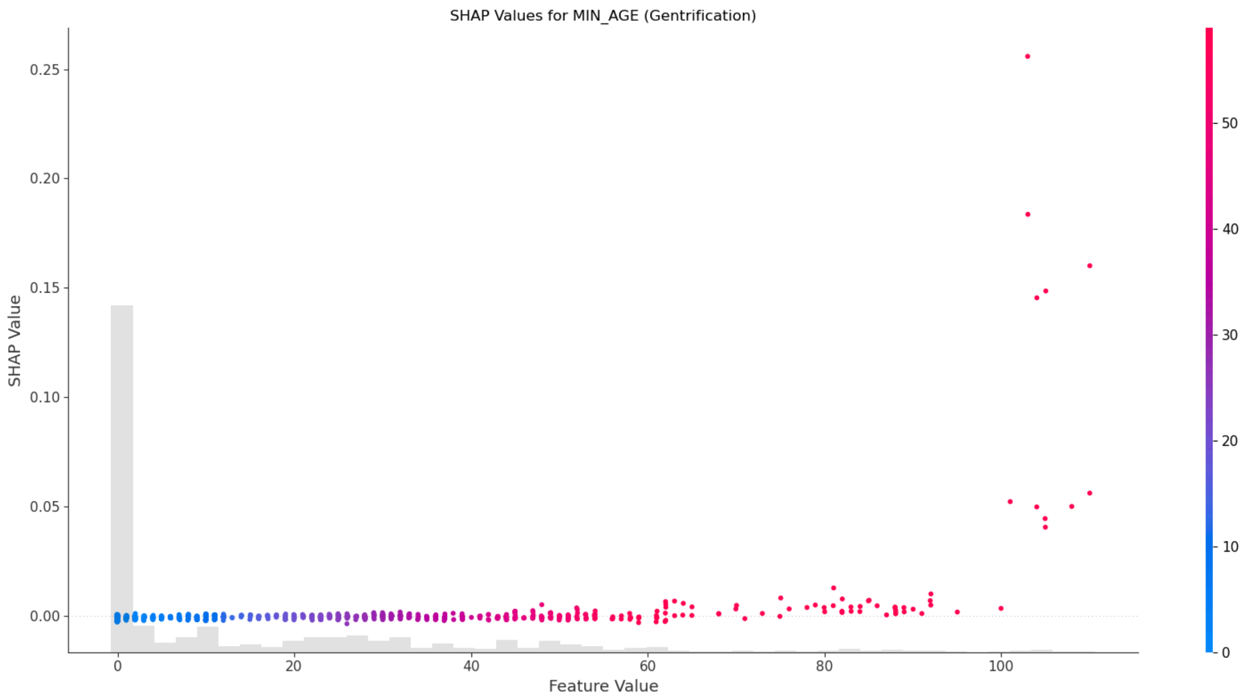
Figure 7 - SHAP Scatter Plot for Median Home Value



Investigating the minimum age of an office within a BG also provided notable results. In Figure 8, a scatter plot with SHAP values on the y-axis and minimum office ages in a block group on the x-axis, a strong surge in SHAP values appears when the minimum office age is over

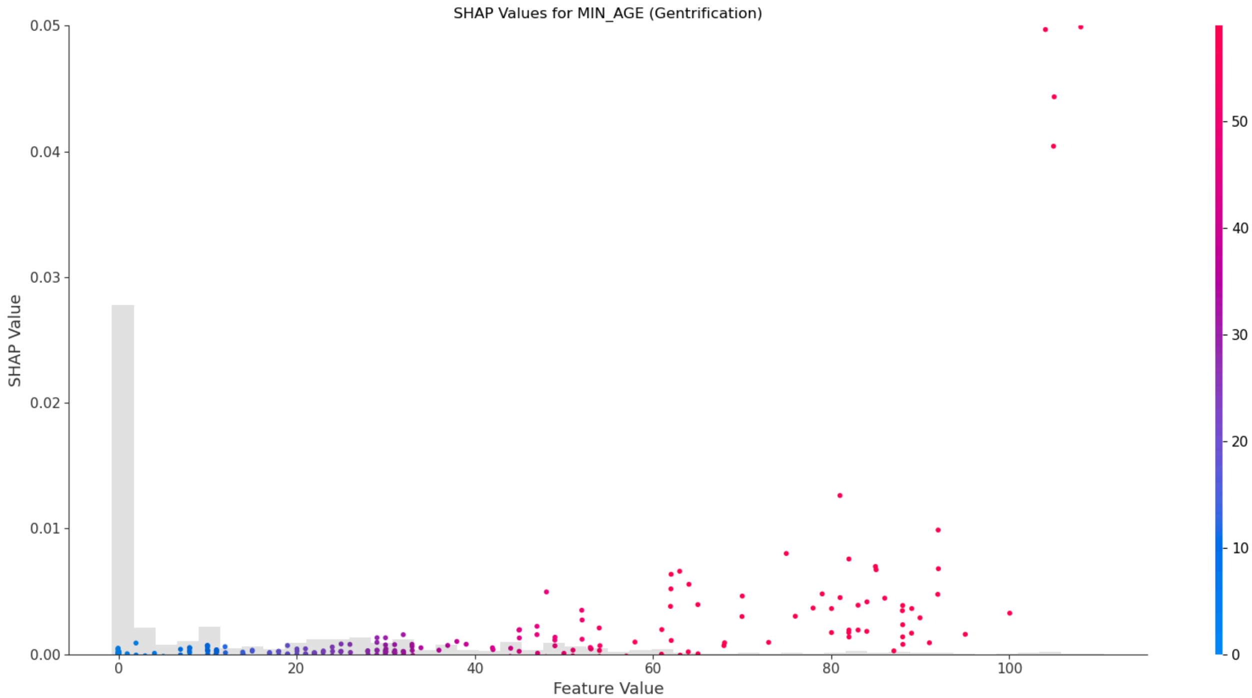
100 years old. This indicates that BGs containing offices all over 100 years old contribute highly to a prediction of gentrification within our model. The SHAP values for BGs with offices all over the age of 100 years old range in value from approximately 0.05 to 0.25, which were among the highest SHAP values from any variable analyzed in this model, even exceeding the high SHAP values seen from the median household income plot in Figure 5.

Figure 8 - SHAP Scatter Plot for Minimum Age of an Office within a BG



With the SHAP values in Figure 8 sitting as highly as they do, Figure 9 allows for a more zoomed in view and deeper investigation of the values before the minimum office age in a block group hits 100 years old.

Figure 9 - SHAP Scatter Plot for Minimum Age of an Office within a BG (Zoomed)



After zooming in on the scatter plot for Figure 9, the SHAP values for minimum age of an office within a BG demonstrate a positive, albeit weaker, trend from 40 to 100 years old. In this case, with SHAP Values generally ranging from 0-0.01, the minimum age of an office shows a push toward gentrification, though generally weaker in predictive power than other variables like mean land value or median household income.

Delving into the SHAP values for the IMPR_VALUE_STD variable also provides insight into the ranges in which the standard deviation of the improvement value of offices within a BG contributes toward a prediction of gentrification. In Figure 10, a zoomed in view with SHAP values on the y-axis and feature values from $0 to $10 million on the x-axis, SHAP values tend to be neutral with a slight negative tilt from values $0 to $6 million. However, after reaching $6 million SHAP values tend to be positive and in the 0 to 0.02 range. In Figure 11, the scatter plot is further zoomed out to encompass feature values from $0 to over $120 million. In this range, the trend exhibited past $6 million appears to continue, with SHAP values ranging from 0-0.02. These values again are not as strong of a predictor as a variable like median household income, though they do point to a push toward gentrification after surpassing $6 million. The increasingly positive SHAP values after $6 million in standard deviation of improvement values for offices within a BG again could indicate a mismatch of values within a block group, where certain offices are very expensive while others are undervalued, indicating the possibility of a lucrative investment.

Figure 10 - SHAP Scatter Plot for the Standard Deviation of Improvement Values for Offices within a BG (0-10M)

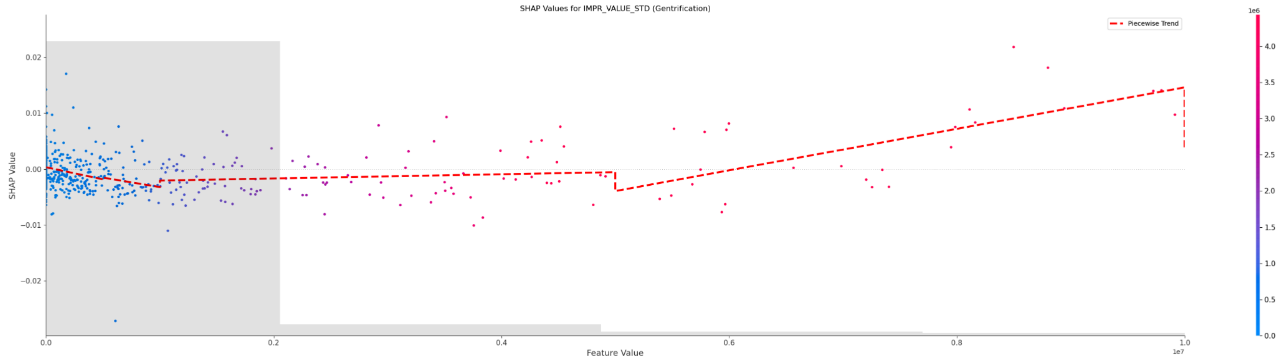
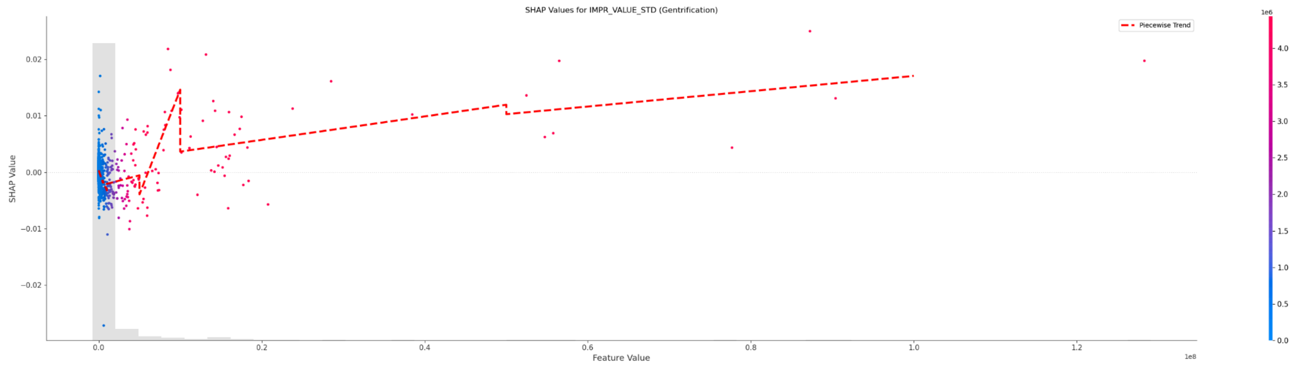


Figure 11 - SHAP Scatter Plot for the Standard Deviation of Improvement Values for Offices within a BG

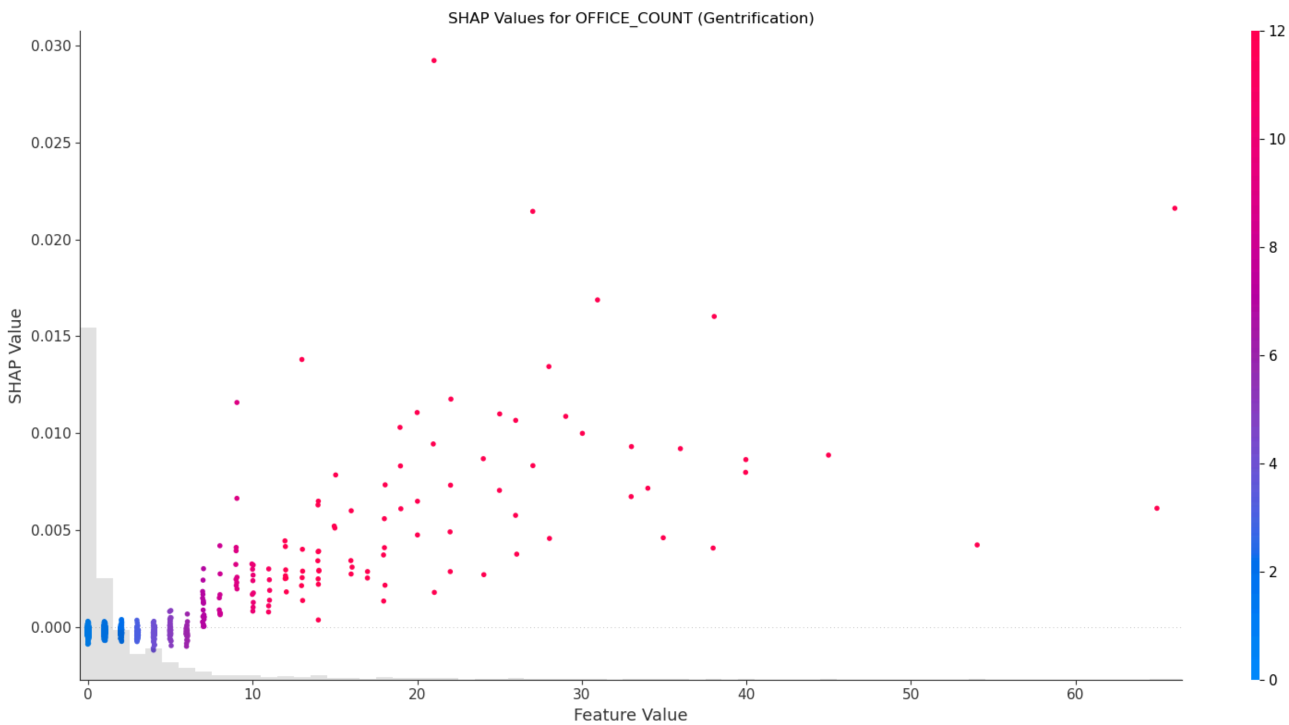


Of note in this analysis, the mean absolute SHAP values for office-related variables were calculated only by examining BGs where offices were present. In doing this, mean absolute SHAP values were not artificially deflated by including an excess of zeros for BGs without offices. However, the exception to this rule was the OFFCE_COUNT variable, where a zero value would be meaningful, indicating the lack of an office in a BG. Though it was meaningful to understand the difference in SHAP values produced by BGs without offices, Table 3 demonstrates the uneven weighting of office counts across BGs in King County, with the vast majority of BGs falling within the 0-10 office range. However, viewing the scatter plot of office counts allows for a better understanding of what occurs outside of the 0-10 count range. The scatter plot in Figure 12 illustrates that for the majority of block groups falling into the 0-10 offices range, SHAP values tend to fall between 0 and 0.005, which provides an explanation for the relatively low median absolute SHAP value of the OFFICE_COUNT variable. Looking further, when office counts extend past the value of 10, SHAP Values become larger in magnitude and move in a more positive direction, typically falling between 0.0025 and 0.015. With the SHAP values increasing as office counts increase, this seems to demonstrate a correlation between number of offices within a BG and likelihood of a prediction of gentrification. More offices in a BG could indicate that the BG is a job center and is therefore attracting rapid growth in incomes and education levels, leading to higher prices for rent and homeownership.

Table 3 - Bin Count for the OFFICE_COUNT variable

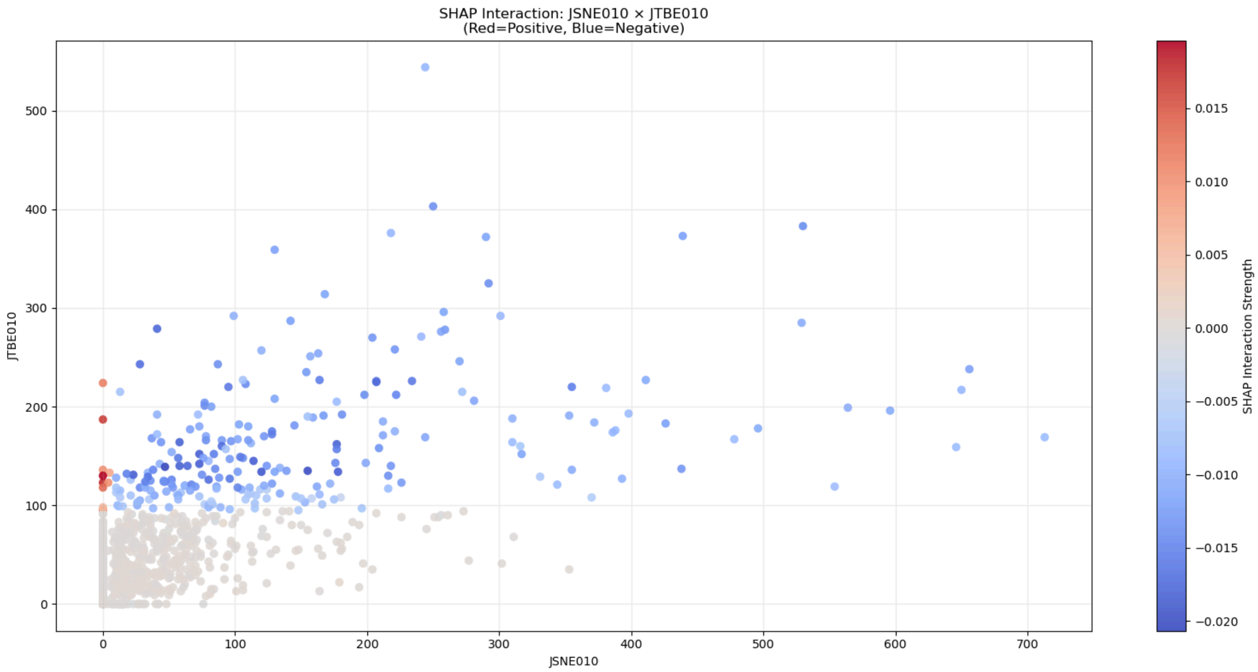
0-10	1328
10-20	57
20-30	22
30-40	9
40-50	3
50-60	1
60-70	2

Figure 12 - Scatter Plot of Office Counts within a BG



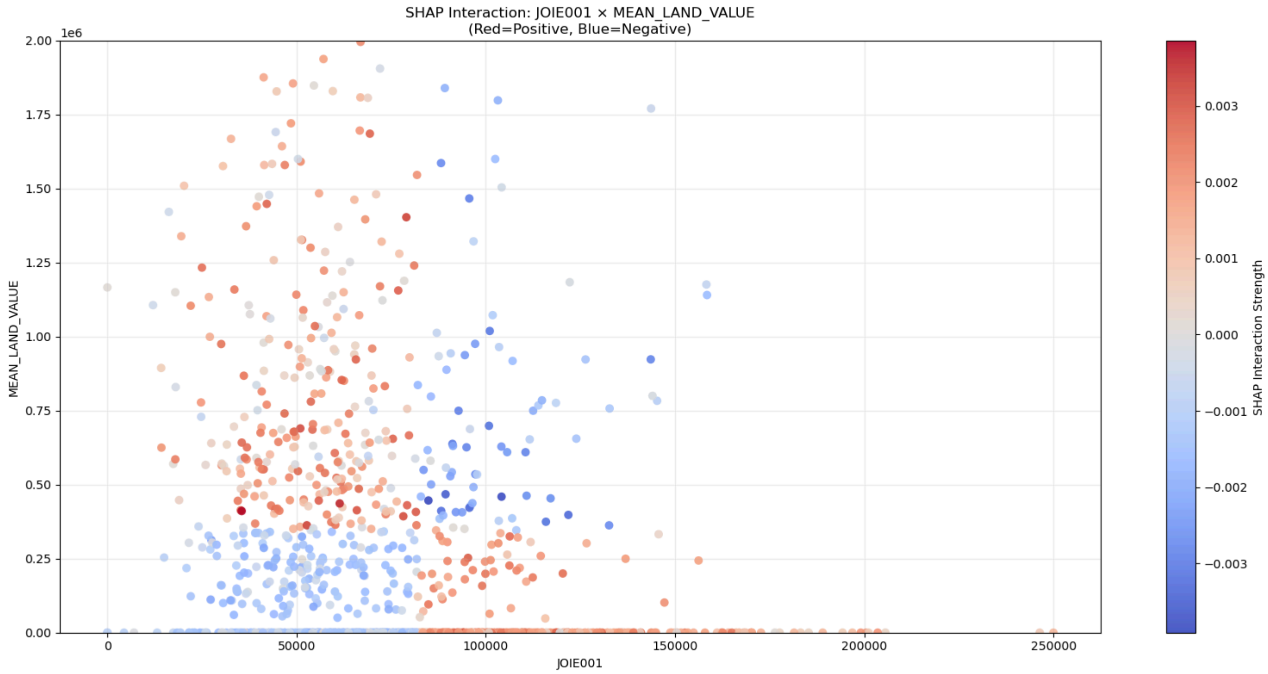
Beyond pure SHAP values, the approach of using SHAP also allows for us to examine interactions between variables. SHAP interaction strength demonstrates how two variables interact and affect each other across different thresholds. An example and baseline can be seen in Figure 13, which demonstrates the strongest SHAP interaction partnership across all variables, number of renter occupied units with no vehicles available in a BG (JSNE010) and total number of renter occupied units in a BG (JTBE001). In this interaction plot, BGs with less than 100 renters almost uniformly have SHAP interaction values around zero, however, past 100 renters, SHAP interaction values appear to be positive when there are few renter occupied households without a vehicle available and negative when there are more than about 10 renter occupied households with no vehicle available. It is important to note in this plot that SHAP interaction values range from -0.020 to 0.015.

Figure 13 - SHAP Interaction Plot For Renter Occupied Households with No Vehicle Available and the Total Number of Renter Occupied Units



With a baseline established, a SHAP interaction plot can be examined for the strongest interaction featuring an office-related variable. The strongest interaction found was between mean land value for offices in a BG and median household income (JOIE001). The results in Figure 14 demonstrate a connection between mean office land value and median household income in a BG, wherein SHAP interaction scores are generally positive where mean office land value is above $350,000 and median household income is below approximately $80,000, or where office land value is under $350,000 and median household income is above $80,000. This could indicate imbalances within block groups, where higher land values and lower median incomes represent rent gaps. Furthermore, areas with higher median incomes and lower land values could represent areas of opportunity for investment, where land values are cheap and they can get the most bang for their buck. However, in this instance it is again important to take note of the magnitude of the SHAP interaction values, which are significantly smaller than SHAP values from the strongest pairing. In this case, SHAP values range from about -0.003 to 0.003, indicating a much weaker feature interaction than that of the variables displayed in Figure 13.

Figure 14 - SHAP Interaction Plot For Mean Land Value of Offices Within a BG and Median Household Income



In addition to the above analysis, an analysis was also conducted on office buildings constructed in the interim period between 2010 and 2019. This analysis was particularly interesting as it related to the arrival of offices linked to big tech companies, which Puget Sound Sage[127] described as being a major contributor to gentrification. When examining construction in the interim period, 86 BGs saw the construction of a new office building between 2010 and 2019. Of these 86 BGs, 11 BGs contained office buildings with tech company ties (Amazon, Microsoft, Google, Meta, etc.).

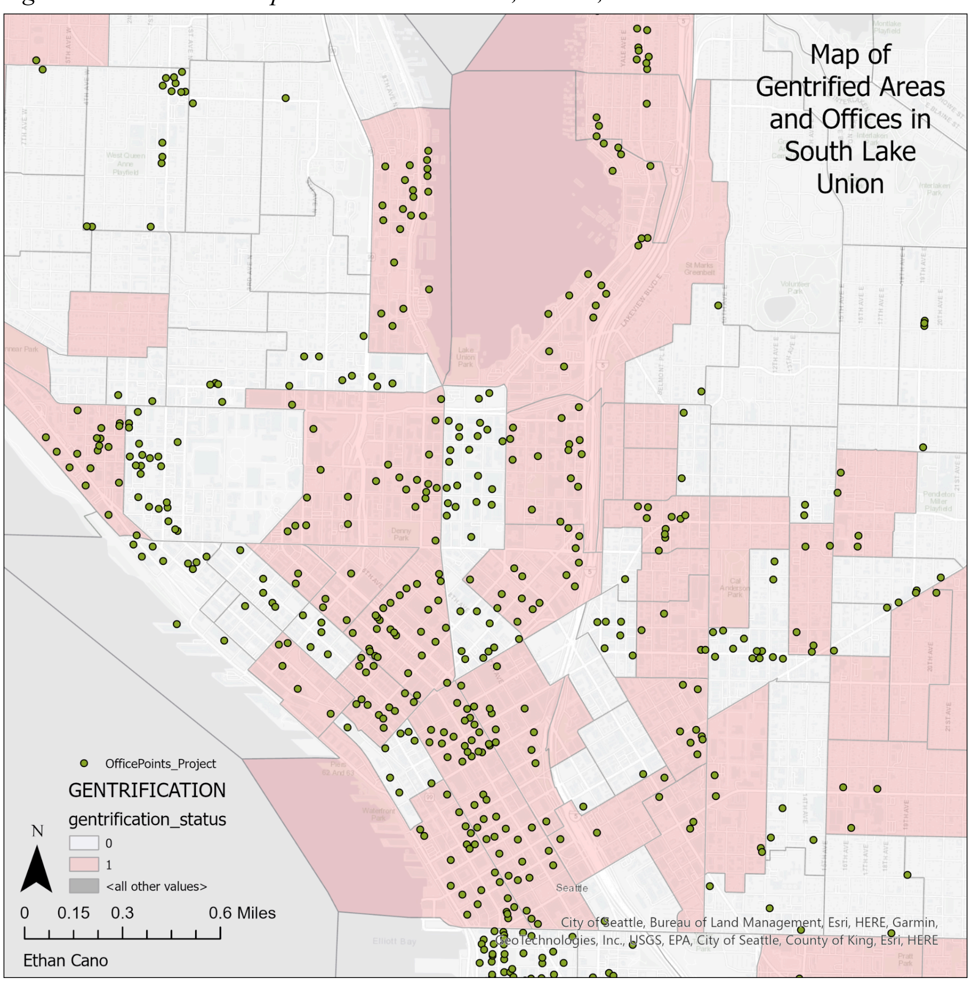
Of these 11 BGs, 7 BGs gentrified while 4 did not. However, of the 4 BGs that did not gentrify, 2 BGs were possibly contributors to spillover effects. To further investigate these spillover effects, the highlighted BG in Figure 15 can be examined. In Figure 15, the highlighted BG sits within South Lake Union, a neighborhood in Seattle often described as an area gentrified by tech companies[128]. The highlighted BG in Figure 15 demarcates the Block Group that saw the highest level of office construction between 2010 and 2019 (indicated by the highlighted blue dots), while also seeing a large amount of that construction come in the form of new offices for Amazon, Meta, and other tech giants. However, the BG directly to its right also provides additional information on the role of tech companies in the process of gentrification. The BG to the right is not classified as gentrified, but was one of the BGs that saw tech company office construction between 2010 and 2019. This BG may have not received a gentrification classification because its median household income in 2010 sat above $96,000, a relatively high level of earning. However, this BG is completely encircled by gentrified areas, all of which have

[127] "Seattle City Council Introduces New Affordable Housing Policy Options – Puget Sound Sage."
[128] Wallick, "Gentrification in Seattle: Amazon Overpowers the City Council."

median household incomes sitting around $30,000 to $40,000. This may indicate that this BG with a relatively high level of income is serving as an anchor from which gentrification is radiating outward to surrounding block groups. In other words, this non-gentrified BG may be creating spillover effects where it is not gentrified itself due to its pre-existing level of affluence, but is leading to gentrification in the surrounding areas. Another BG with a median household income of $93,359 saw the development of a Microsoft office, and while it did not gentrify itself, an adjacent BG with a median household income of around $71,000 did gentrify.

The remaining two BGs contained offices for the company Tableau. Neither one of these BGs gentrified, however one BG contained an office built in 2016, while the other contained an office built in 2018, which may have not had enough time to see gentrification effects by the 2019 analysis year. While none of the above information is causal evidence directly placing blame upon tech companies for the gentrification that occurred in these 11 Block Groups, the results are nonetheless important and merit further research and discussion into the issue of tech companies' roles in the gentrification of an area.

Figure 15 - A Block Group in South Lake Union, Seattle, WA



Conclusion

The issue of gentrification proves salient in Seattle due to increasing levels of unaffordability[129] and the socioeconomic issues that arise as a result of gentrification, including social disruption[130], residential displacement[131], and exclusion[132]. Fortunately, ML models provide a new method of investigating complex problems like gentrification[133] and SHAP values allow for a better understanding of the contributing factors leading to a gentrification prediction in an ML model[134]. By better understanding the contributing factors leading to gentrification, policymakers can better help protect communities against the damaging effects of gentrification, while also allowing for community members to take advantage of the opportunities presented in gentrifying neighborhoods[135].

Using a Random Forest (RF) Classification model in the cuML Python package, gentrification from 2010 to 2019 in King County has been analyzed. Approximately 400 independent variables were fed into this model and their feature importance values were extracted in order to better understand the black box behind the RF model. In this instance, the SHAP package was used for analysis, allowing for interpretation of the importance of these variables and their correlations with a prediction of gentrification.

With multiple office-related variables in the top 10 in terms of median absolute SHAP values and other office-related variables returning relatively strong SHAP values within certain thresholds, office-related variables provide an interesting and relevant basis for analysis of gentrification in a Block Group.

In this study BGs with mean office land values above $350,000, minimum office ages above 100 years old, standard deviations of office improvement values above $6 million, and office counts above 10 all seemed to correlate with a prediction of gentrification. Of these key variables, minimum office age of over 100 produced the strongest SHAP values, reaching levels above median household income, the variable with the highest median absolute SHAP value. In addition to this, weak, but significant connections between mean office land value and median household income could be seen through SHAP interaction plots.

Furthermore, though not causally linked, the interactions between tech companies and gentrification do not appear to be occurring randomly or by accident, and merit future investigation. This investigation could be carried out on the scale of multiple major MSAs, thereby providing a more robust dataset for analysis with ML.

[129] "Seattle City Council Introduces New Affordable Housing Policy Options – Puget Sound Sage."
[130] Maharawal, "Tech-Colonialism."
[131] Kohn, "What Is Wrong with Gentrification?"
[132] Kohn.
[133] Thackway et al., "Building a Predictive Machine Learning Model of Gentrification in Sydney."
[134] Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.
[135] Brummet and Reed, "The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children."

This research builds on the previous work contributed by Freyd demonstrating that technology offices have a positive and significant impact on house prices[136] by examining office-related variables with a more robust definition of gentrification and looking into specific office-related variables including office age, land value, and improvement value. Furthermore, the research conducted in this thesis has allowed for a deeper understanding of office-related variables through the use of ML and SHAP, giving deeper insight into the value ranges for offices where a prediction of gentrification is likely to take place.

With regard to policy, the findings in this thesis indicate that action could be taken to protect renters in areas with a mix of offices with low and high improvement values, older buildings, high office counts, and office land values over $350,000. Such actions could include rent caps or freezes, increased dispersal of Housing Choice Vouchers, or organizing of tenant unions in these areas to help ensure strong anti-displacement measures are in place for the future. However, these suggestions merit more thought and time from policymakers themselves.

In terms of other areas in which future investigation could occur, the aforementioned spillover effects could be further quantitatively or qualitatively studied to provide a more concrete answer to whether or not they exist or are linked with the gentrification that appears to be occurring in the surrounding area. Additionally, this study made use of data from 2010 to 2019. With large changes occurring post-COVID, another study could investigate the changing role of offices with the rise of remote and hybrid work. Beyond this, a study with a larger and more robust dataset could allow for predictive tasks to occur, where demographic and office-related variables are used to better predict future areas of gentrification within major metropolitan areas. This study was limited in time and to the geographic region of King County. By extending the dataset to multiple MSAs, a future study could bear out trends at a national level on office-related variables.

[136] Freyd, "Large Tech Office Openings and the Onset of Gentrification."

Appendix A

Variable Codebook for ACS 2006-2010 5-year estimate from NHGIS IPUMS. These variables were fed into the ML model in addition to the aforementioned selected office variables.

Table 1: Sex by Age
Universe: Total population
Source code: B01001
NHGIS code: JLZ

JLZE001: Total
JLZE002: Male
JLZE003: Male: Under 5 years
JLZE004: Male: 5 to 9 years
JLZE005: Male: 10 to 14 years
JLZE006: Male: 15 to 17 years
JLZE007: Male: 18 and 19 years
JLZE008: Male: 20 years
JLZE009: Male: 21 years
JLZE010: Male: 22 to 24 years
JLZE011: Male: 25 to 29 years
JLZE012: Male: 30 to 34 years
JLZE013: Male: 35 to 39 years
JLZE014: Male: 40 to 44 years
JLZE015: Male: 45 to 49 years
JLZE016: Male: 50 to 54 years
JLZE017: Male: 55 to 59 years
JLZE018: Male: 60 and 61 years
JLZE019: Male: 62 to 64 years
JLZE020: Male: 65 and 66 years
JLZE021: Male: 67 to 69 years
JLZE022: Male: 70 to 74 years
JLZE023: Male: 75 to 79 years
JLZE024: Male: 80 to 84 years
JLZE025: Male: 85 years and over
JLZE026: Female
JLZE027: Female: Under 5 years
JLZE028: Female: 5 to 9 years
JLZE029: Female: 10 to 14 years
JLZE030: Female: 15 to 17 years
JLZE031: Female: 18 and 19 years

JLZE032: Female: 20 years
JLZE033: Female: 21 years
JLZE034: Female: 22 to 24 years
JLZE035: Female: 25 to 29 years
JLZE036: Female: 30 to 34 years
JLZE037: Female: 35 to 39 years
JLZE038: Female: 40 to 44 years
JLZE039: Female: 45 to 49 years
JLZE040: Female: 50 to 54 years
JLZE041: Female: 55 to 59 years
JLZE042: Female: 60 and 61 years
JLZE043: Female: 62 to 64 years
JLZE044: Female: 65 and 66 years
JLZE045: Female: 67 to 69 years
JLZE046: Female: 70 to 74 years
JLZE047: Female: 75 to 79 years
JLZE048: Female: 80 to 84 years
JLZE049: Female: 85 years and over

Table 2: Total Population
Universe: Total population
Source code: B01003
NHGIS code: JMA
JMAE001: Total

Table 3: Race
Universe: Total population
Source code: B02001
NHGIS code: JMB
JMBE001: Total
JMBE002: White alone
JMBE003: Black or African American alone
JMBE004: American Indian and Alaska Native alone
JMBE005: Asian alone
JMBE006: Native Hawaiian and Other Pacific Islander alone
JMBE007: Some other race alone
JMBE008: Two or more races
JMBE009: Two or more races: Two races including Some other race
JMBE010: Two or more races: Two races excluding Some other race, and three or more races

Table 4: Hispanic or Latino Origin by Race
Universe: Total population
Source code: B03002
NHGIS code: JMJ
JMJE001: Total
JMJE002: Not Hispanic or Latino
JMJE003: Not Hispanic or Latino: White alone
JMJE004: Not Hispanic or Latino: Black or African American alone
JMJE005: Not Hispanic or Latino: American Indian and Alaska Native alone
JMJE006: Not Hispanic or Latino: Asian alone
JMJE007: Not Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
JMJE008: Not Hispanic or Latino: Some other race alone
JMJE009: Not Hispanic or Latino: Two or more races
JMJE010: Not Hispanic or Latino: Two or more races: Two races including Some other race
JMJE011: Not Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races
JMJE012: Hispanic or Latino
JMJE013: Hispanic or Latino: White alone
JMJE014: Hispanic or Latino: Black or African American alone
JMJE015: Hispanic or Latino: American Indian and Alaska Native alone
JMJE016: Hispanic or Latino: Asian alone
JMJE017: Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
JMJE018: Hispanic or Latino: Some other race alone
JMJE019: Hispanic or Latino: Two or more races
JMJE020: Hispanic or Latino: Two or more races: Two races including Some other race
JMJE021: Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races

Table 5: Means of Transportation to Work
Universe: Workers 16 years and over
Source code: B08301
NHGIS code: JM0
JM0E001: Total
JM0E002: Car, truck, or van
JM0E003: Car, truck, or van: Drove alone
JM0E004: Car, truck, or van: Carpooled
JM0E005: Car, truck, or van: Carpooled: In 2-person carpool
JM0E006: Car, truck, or van: Carpooled: In 3-person carpool

JM0E007: Car, truck, or van: Carpooled: In 4-person carpool
JM0E008: Car, truck, or van: Carpooled: In 5- or 6-person carpool
JM0E009: Car, truck, or van: Carpooled: In 7-or-more-person carpool
JM0E010: Public transportation (excluding taxicab)
JM0E011: Public transportation (excluding taxicab): Bus or trolley bus
JM0E012: Public transportation (excluding taxicab): Streetcar or trolley car (carro publico in Puerto Rico)
JM0E013: Public transportation (excluding taxicab): Subway or elevated
JM0E014: Public transportation (excluding taxicab): Railroad
JM0E015: Public transportation (excluding taxicab): Ferryboat
JM0E016: Taxicab
JM0E017: Motorcycle
JM0E018: Bicycle
JM0E019: Walked
JM0E020: Other means
JM0E021: Worked at home

Table 6: Travel Time to Work
Universe: Workers 16 years and over who did not work at home
Source code: B08303
NHGIS code: JM2
JM2E001: Total
JM2E002: Less than 5 minutes
JM2E003: 5 to 9 minutes
JM2E004: 10 to 14 minutes
JM2E005: 15 to 19 minutes
JM2E006: 20 to 24 minutes
JM2E007: 25 to 29 minutes
JM2E008: 30 to 34 minutes
JM2E009: 35 to 39 minutes
JM2E010: 40 to 44 minutes
JM2E011: 45 to 59 minutes
JM2E012: 60 to 89 minutes
JM2E013: 90 or more minutes

Table 7: Own Children Under 18 Years by Family Type and Age
Universe: Own children under 18 years
Source code: B09002
NHGIS code: JM3
JM3E001: Total

JM3E002: In married-couple families
JM3E003: In married-couple families: Under 3 years
JM3E004: In married-couple families: 3 and 4 years
JM3E005: In married-couple families: 5 years
JM3E006: In married-couple families: 6 to 11 years
JM3E007: In married-couple families: 12 to 17 years
JM3E008: In other families
JM3E009: In other families: Male householder, no wife present
JM3E010: In other families: Male householder, no wife present: Under 3 years
JM3E011: In other families: Male householder, no wife present: 3 and 4 years
JM3E012: In other families: Male householder, no wife present: 5 years
JM3E013: In other families: Male householder, no wife present: 6 to 11 years
JM3E014: In other families: Male householder, no wife present: 12 to 17 years
JM3E015: In other families: Female householder, no husband present
JM3E016: In other families: Female householder, no husband present: Under 3 years
JM3E017: In other families: Female householder, no husband present: 3 and 4 years
JM3E018: In other families: Female householder, no husband present: 5 years
JM3E019: In other families: Female householder, no husband present: 6 to 11 years
JM3E020: In other families: Female householder, no husband present: 12 to 17 years

Table 8: Household Type (Including Living Alone)
Universe: Households
Source code: B11001
NHGIS code: JM5
JM5E001: Total
JM5E002: Family households
JM5E003: Family households: Married-couple family
JM5E004: Family households: Other family
JM5E005: Family households: Other family: Male householder, no wife present
JM5E006: Family households: Other family: Female householder, no husband present
JM5E007: Nonfamily households
JM5E008: Nonfamily households: Householder living alone
JM5E009: Nonfamily households: Householder not living alone

Table 9: Household Type by Household Size
Universe: Households
Source code: B11016
NHGIS code: JNW
JNWE001: Total
JNWE002: Family households

JNWE003: Family households: 2-person household
JNWE004: Family households: 3-person household
JNWE005: Family households: 4-person household
JNWE006: Family households: 5-person household
JNWE007: Family households: 6-person household
JNWE008: Family households: 7-or-more person household
JNWE009: Nonfamily households
JNWE010: Nonfamily households: 1-person household
JNWE011: Nonfamily households: 2-person household
JNWE012: Nonfamily households: 3-person household
JNWE013: Nonfamily households: 4-person household
JNWE014: Nonfamily households: 5-person household
JNWE015: Nonfamily households: 6-person household
JNWE016: Nonfamily households: 7-or-more person household

Table 10: Sex by Educational Attainment for the Population 25 Years and Over
Universe: Population 25 years and over
Source code: B15002
NHGIS code: JN9
JN9E001: Total
JN9E002: Male
JN9E003: Male: No schooling completed
JN9E004: Male: Nursery to 4th grade
JN9E005: Male: 5th and 6th grade
JN9E006: Male: 7th and 8th grade
JN9E007: Male: 9th grade
JN9E008: Male: 10th grade
JN9E009: Male: 11th grade
JN9E010: Male: 12th grade, no diploma
JN9E011: Male: High school graduate, GED, or alternative
JN9E012: Male: Some college, less than 1 year
JN9E013: Male: Some college, 1 or more years, no degree
JN9E014: Male: Associate's degree
JN9E015: Male: Bachelor's degree
JN9E016: Male: Master's degree
JN9E017: Male: Professional school degree
JN9E018: Male: Doctorate degree
JN9E019: Female
JN9E020: Female: No schooling completed
JN9E021: Female: Nursery to 4th grade

JN9E022: Female: 5th and 6th grade
JN9E023: Female: 7th and 8th grade
JN9E024: Female: 9th grade
JN9E025: Female: 10th grade
JN9E026: Female: 11th grade
JN9E027: Female: 12th grade, no diploma
JN9E028: Female: High school graduate, GED, or alternative
JN9E029: Female: Some college, less than 1 year
JN9E030: Female: Some college, 1 or more years, no degree
JN9E031: Female: Associate's degree
JN9E032: Female: Bachelor's degree
JN9E033: Female: Master's degree
JN9E034: Female: Professional school degree
JN9E035: Female: Doctorate degree

Table 11: Ratio of Income to Poverty Level in the Past 12 Months
Universe: Population for whom poverty status is determined
Source code: C17002
NHGIS code: JOC
JOCE001: Total
JOCE002: Under .50
JOCE003: .50 to .99
JOCE004: 1.00 to 1.24
JOCE005: 1.25 to 1.49
JOCE006: 1.50 to 1.84
JOCE007: 1.85 to 1.99
JOCE008: 2.00 and over

Table 12: Household Income in the Past 12 Months (in 2010 Inflation-Adjusted Dollars)
Universe: Households
Source code: B19001
NHGIS code: JOH
JOHE001: Total
JOHE002: Less than $10,000
JOHE003: $10,000 to $14,999
JOHE004: $15,000 to $19,999
JOHE005: $20,000 to $24,999
JOHE006: $25,000 to $29,999
JOHE007: $30,000 to $34,999
JOHE008: $35,000 to $39,999

JOHE009: $40,000 to $44,999
JOHE010: $45,000 to $49,999
JOHE011: $50,000 to $59,999
JOHE012: $60,000 to $74,999
JOHE013: $75,000 to $99,999
JOHE014: $100,000 to $124,999
JOHE015: $125,000 to $149,999
JOHE016: $150,000 to $199,999
JOHE017: $200,000 or more

Table 13: Median Household Income in the Past 12 Months (in 2010 Inflation-Adjusted Dollars)
Universe: Households
Source code: B19013
NHGIS code: JOI
JOIE001: Median household income in the past 12 months (in 2010 inflation-adjusted dollars)

Table 14: Public Assistance Income in the Past 12 Months for Households
Universe: Households
Source code: B19057
NHGIS code: JPB
JPBE001: Total
JPBE002: With public assistance income
JPBE003: No public assistance income

Table 15: Sex by Work Status in the Past 12 Months by Usual Hours Worked per Week in the Past 12 Months by Weeks Worked in the Past 12 Months for the Population 16 to 64 Years
Universe: Population 16 to 64 years
Source code: B23022
NHGIS code: JQ4
JQ4E001: Total
JQ4E002: Male
JQ4E003: Male: Worked in the past 12 months
JQ4E004: Male: Worked in the past 12 months: Usually worked 35 or more hours per week
JQ4E005: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 50 to 52 weeks
JQ4E006: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 48 and 49 weeks

JQ4E007: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 40 to 47 weeks

JQ4E008: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 27 to 39 weeks

JQ4E009: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 14 to 26 weeks

JQ4E010: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 1 to 13 weeks

JQ4E011: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week

JQ4E012: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 50 to 52 weeks

JQ4E013: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 48 and 49 weeks

JQ4E014: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 40 to 47 weeks

JQ4E015: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 27 to 39 weeks

JQ4E016: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 14 to 26 weeks

JQ4E017: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 1 to 13 weeks

JQ4E018: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week

JQ4E019: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 50 to 52 weeks

JQ4E020: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 48 and 49 weeks

JQ4E021: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 40 to 47 weeks

JQ4E022: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 27 to 39 weeks

JQ4E023: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 14 to 26 weeks

JQ4E024: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 1 to 13 weeks

JQ4E025: Male: Did not work in the past 12 months

JQ4E026: Female

JQ4E027: Female: Worked in the past 12 months

JQ4E028: Female: Worked in the past 12 months: Usually worked 35 or more hours per week

JQ4E029: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 50 to 52 weeks
JQ4E030: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 48 and 49 weeks
JQ4E031: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 40 to 47 weeks
JQ4E032: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 27 to 39 weeks
JQ4E033: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 14 to 26 weeks
JQ4E034: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 1 to 13 weeks
JQ4E035: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week
JQ4E036: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 50 to 52 weeks
JQ4E037: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 48 and 49 weeks
JQ4E038: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 40 to 47 weeks
JQ4E039: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 27 to 39 weeks
JQ4E040: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 14 to 26 weeks
JQ4E041: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 1 to 13 weeks
JQ4E042: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week
JQ4E043: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 50 to 52 weeks
JQ4E044: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 48 and 49 weeks
JQ4E045: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 40 to 47 weeks
JQ4E046: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 27 to 39 weeks
JQ4E047: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 14 to 26 weeks
JQ4E048: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 1 to 13 weeks
JQ4E049: Female: Did not work in the past 12 months

Table 16: Sex by Industry for the Civilian Employed Population 16 Years and Over
Universe: Civilian employed population 16 years and over
Source code: C24030
NHGIS code: JRG

JRGE001: Total
JRGE002: Male
JRGE003: Male: Agriculture, forestry, fishing and hunting, and mining
JRGE004: Male: Agriculture, forestry, fishing and hunting, and mining: Agriculture, forestry, fishing and hunting
JRGE005: Male: Agriculture, forestry, fishing and hunting, and mining: Mining, quarrying, and oil and gas extraction
JRGE006: Male: Construction
JRGE007: Male: Manufacturing
JRGE008: Male: Wholesale trade
JRGE009: Male: Retail trade
JRGE010: Male: Transportation and warehousing, and utilities
JRGE011: Male: Transportation and warehousing, and utilities: Transportation and warehousing
JRGE012: Male: Transportation and warehousing, and utilities: Utilities
JRGE013: Male: Information
JRGE014: Male: Finance and insurance, and real estate and rental and leasing
JRGE015: Male: Finance and insurance, and real estate and rental and leasing: Finance and insurance
JRGE016: Male: Finance and insurance, and real estate and rental and leasing: Real estate and rental and leasing
JRGE017: Male: Professional, scientific, and management, and administrative and waste management services
JRGE018: Male: Professional, scientific, and management, and administrative and waste management services: Professional, scientific, and technical services
JRGE019: Male: Professional, scientific, and management, and administrative and waste management services: Management of companies and enterprises
JRGE020: Male: Professional, scientific, and management, and administrative and waste management services: Administrative and support and waste management services
JRGE021: Male: Educational services, and health care and social assistance
JRGE022: Male: Educational services, and health care and social assistance: Educational services
JRGE023: Male: Educational services, and health care and social assistance: Health care and social assistance

JRGE024: Male: Arts, entertainment, and recreation, and accommodation and food services

JRGE025: Male: Arts, entertainment, and recreation, and accommodation and food services: Arts, entertainment, and recreation

JRGE026: Male: Arts, entertainment, and recreation, and accommodation and food services: Accommodation and food services

JRGE027: Male: Other services, except public administration

JRGE028: Male: Public administration

JRGE029: Female

JRGE030: Female: Agriculture, forestry, fishing and hunting, and mining

JRGE031: Female: Agriculture, forestry, fishing and hunting, and mining: Agriculture, forestry, fishing and hunting

JRGE032: Female: Agriculture, forestry, fishing and hunting, and mining: Mining, quarrying, and oil and gas extraction

JRGE033: Female: Construction

JRGE034: Female: Manufacturing

JRGE035: Female: Wholesale trade

JRGE036: Female: Retail trade

JRGE037: Female: Transportation and warehousing, and utilities

JRGE038: Female: Transportation and warehousing, and utilities: Transportation and warehousing

JRGE039: Female: Transportation and warehousing, and utilities: Utilities

JRGE040: Female: Information

JRGE041: Female: Finance and insurance, and real estate and rental and leasing

JRGE042: Female: Finance and insurance, and real estate and rental and leasing: Finance and insurance

JRGE043: Female: Finance and insurance, and real estate and rental and leasing: Real estate and rental and leasing

JRGE044: Female: Professional, scientific, and management, and administrative and waste management services

JRGE045: Female: Professional, scientific, and management, and administrative and waste management services: Professional, scientific, and technical services

JRGE046: Female: Professional, scientific, and management, and administrative and waste management services: Management of companies and enterprises

JRGE047: Female: Professional, scientific, and management, and administrative and waste management services: Administrative and support and waste management services

JRGE048: Female: Educational services, and health care and social assistance

JRGE049: Female: Educational services, and health care and social assistance: Educational services

JRGE050: Female: Educational services, and health care and social assistance: Health care and social assistance

JRGE051: Female: Arts, entertainment, and recreation, and accommodation and food services

JRGE052: Female: Arts, entertainment, and recreation, and accommodation and food services: Arts, entertainment, and recreation

JRGE053: Female: Arts, entertainment, and recreation, and accommodation and food services: Accommodation and food services

JRGE054: Female: Other services, except public administration

JRGE055: Female: Public administration

Table 17: Occupancy Status
Universe: Housing units
Source code: B25002
NHGIS code: JRJ
JRJE001: Total
JRJE002: Occupied
JRJE003: Vacant

Table 18: Tenure
Universe: Occupied housing units
Source code: B25003
NHGIS code: JRK
JRKE001: Total
JRKE002: Owner occupied
JRKE003: Renter occupied

Table 19: Year Structure Built
Universe: Housing units
Source code: B25034
NHGIS code: JSD
JSDE001: Total
JSDE002: Built 2005 or later
JSDE003: Built 2000 to 2004
JSDE004: Built 1990 to 1999
JSDE005: Built 1980 to 1989
JSDE006: Built 1970 to 1979
JSDE007: Built 1960 to 1969
JSDE008: Built 1950 to 1959
JSDE009: Built 1940 to 1949

JSDE010: Built 1939 or earlier

Table 20: Tenure by Vehicles Available
Universe: Occupied housing units
Source code: B25044
NHGIS code: JSN
JSNE001: Total
JSNE002: Owner occupied
JSNE003: Owner occupied: No vehicle available
JSNE004: Owner occupied: 1 vehicle available
JSNE005: Owner occupied: 2 vehicles available
JSNE006: Owner occupied: 3 vehicles available
JSNE007: Owner occupied: 4 vehicles available
JSNE008: Owner occupied: 5 or more vehicles available
JSNE009: Renter occupied
JSNE010: Renter occupied: No vehicle available
JSNE011: Renter occupied: 1 vehicle available
JSNE012: Renter occupied: 2 vehicles available
JSNE013: Renter occupied: 3 vehicles available
JSNE014: Renter occupied: 4 vehicles available
JSNE015: Renter occupied: 5 or more vehicles available

Table 21: Median Gross Rent (Dollars)
Universe: Renter-occupied housing units paying cash rent
Source code: B25064
NHGIS code: JS5
JS5E001: Median gross rent

Table 22: Gross Rent as a Percentage of Household Income in the Past 12 Months
Universe: Renter-occupied housing units
Source code: B25070
NHGIS code: JTB
JTBE001: Total
JTBE002: Less than 10.0 percent
JTBE003: 10.0 to 14.9 percent
JTBE004: 15.0 to 19.9 percent
JTBE005: 20.0 to 24.9 percent
JTBE006: 25.0 to 29.9 percent
JTBE007: 30.0 to 34.9 percent
JTBE008: 35.0 to 39.9 percent

JTBE009: 40.0 to 49.9 percent
JTBE010: 50.0 percent or more
JTBE011: Not computed

Table 23: Median Value (Dollars)
Universe: Owner-occupied housing units
Source code: B25077
NHGIS code: JTI
JTIE001: Median value (dollars)

Appendix B

Variable Codebook for ACS 2015-2019 5-year estimate from NHGIS IPUMS. These variables were not fed into the ML model, but select variables from this dataset were used for gentrification classification per the definition used in this thesis.

Table 1: Sex by Age
Universe: Total population
Source code: B01001
NHGIS code: ALT0
ALT0E001: Total
ALT0E002: Male
ALT0E003: Male: Under 5 years
ALT0E004: Male: 5 to 9 years
ALT0E005: Male: 10 to 14 years
ALT0E006: Male: 15 to 17 years
ALT0E007: Male: 18 and 19 years
ALT0E008: Male: 20 years
ALT0E009: Male: 21 years
ALT0E010: Male: 22 to 24 years
ALT0E011: Male: 25 to 29 years
ALT0E012: Male: 30 to 34 years
ALT0E013: Male: 35 to 39 years
ALT0E014: Male: 40 to 44 years
ALT0E015: Male: 45 to 49 years
ALT0E016: Male: 50 to 54 years
ALT0E017: Male: 55 to 59 years
ALT0E018: Male: 60 and 61 years
ALT0E019: Male: 62 to 64 years
ALT0E020: Male: 65 and 66 years
ALT0E021: Male: 67 to 69 years
ALT0E022: Male: 70 to 74 years
ALT0E023: Male: 75 to 79 years
ALT0E024: Male: 80 to 84 years
ALT0E025: Male: 85 years and over
ALT0E026: Female
ALT0E027: Female: Under 5 years
ALT0E028: Female: 5 to 9 years
ALT0E029: Female: 10 to 14 years
ALT0E030: Female: 15 to 17 years

ALT0E031: Female: 18 and 19 years
ALT0E032: Female: 20 years
ALT0E033: Female: 21 years
ALT0E034: Female: 22 to 24 years
ALT0E035: Female: 25 to 29 years
ALT0E036: Female: 30 to 34 years
ALT0E037: Female: 35 to 39 years
ALT0E038: Female: 40 to 44 years
ALT0E039: Female: 45 to 49 years
ALT0E040: Female: 50 to 54 years
ALT0E041: Female: 55 to 59 years
ALT0E042: Female: 60 and 61 years
ALT0E043: Female: 62 to 64 years
ALT0E044: Female: 65 and 66 years
ALT0E045: Female: 67 to 69 years
ALT0E046: Female: 70 to 74 years
ALT0E047: Female: 75 to 79 years
ALT0E048: Female: 80 to 84 years
ALT0E049: Female: 85 years and over

Table 2: Total Population
Universe: Total population
Source code: B01003
NHGIS code: ALUB
ALUBE001: Total

Table 3: Race
Universe: Total population
Source code: B02001
NHGIS code: ALUC
ALUCE001: Total
ALUCE002: White alone
ALUCE003: Black or African American alone
ALUCE004: American Indian and Alaska Native alone
ALUCE005: Asian alone
ALUCE006: Native Hawaiian and Other Pacific Islander alone
ALUCE007: Some other race alone
ALUCE008: Two or more races
ALUCE009: Two or more races: Two races including Some other race

ALUCE010: Two or more races: Two races excluding Some other race, and three or more races

Table 4: Hispanic or Latino Origin by Race
Universe: Total population
Source code: B03002
NHGIS code: ALUK
ALUKE001: Total
ALUKE002: Not Hispanic or Latino
ALUKE003: Not Hispanic or Latino: White alone
ALUKE004: Not Hispanic or Latino: Black or African American alone
ALUKE005: Not Hispanic or Latino: American Indian and Alaska Native alone
ALUKE006: Not Hispanic or Latino: Asian alone
ALUKE007: Not Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
ALUKE008: Not Hispanic or Latino: Some other race alone
ALUKE009: Not Hispanic or Latino: Two or more races
ALUKE010: Not Hispanic or Latino: Two or more races: Two races including Some other race
ALUKE011: Not Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races
ALUKE012: Hispanic or Latino
ALUKE013: Hispanic or Latino: White alone
ALUKE014: Hispanic or Latino: Black or African American alone
ALUKE015: Hispanic or Latino: American Indian and Alaska Native alone
ALUKE016: Hispanic or Latino: Asian alone
ALUKE017: Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
ALUKE018: Hispanic or Latino: Some other race alone
ALUKE019: Hispanic or Latino: Two or more races
ALUKE020: Hispanic or Latino: Two or more races: Two races including Some other race
ALUKE021: Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races

Table 5: Means of Transportation to Work
Universe: Workers 16 years and over
Source code: B08301
NHGIS code: ALU1
ALU1E001: Total
ALU1E002: Car, truck, or van
ALU1E003: Car, truck, or van: Drove alone

ALU1E004: Car, truck, or van: Carpooled
ALU1E005: Car, truck, or van: Carpooled: In 2-person carpool
ALU1E006: Car, truck, or van: Carpooled: In 3-person carpool
ALU1E007: Car, truck, or van: Carpooled: In 4-person carpool
ALU1E008: Car, truck, or van: Carpooled: In 5- or 6-person carpool
ALU1E009: Car, truck, or van: Carpooled: In 7-or-more-person carpool
ALU1E010: Public transportation (excluding taxicab)
ALU1E011: Public transportation (excluding taxicab): Bus
ALU1E012: Public transportation (excluding taxicab): Subway or elevated rail
ALU1E013: Public transportation (excluding taxicab): Long-distance train or commuter rail
ALU1E014: Public transportation (excluding taxicab): Light rail, streetcar or trolley (carro publico in Puerto Rico)
ALU1E015: Public transportation (excluding taxicab): Ferryboat
ALU1E016: Taxicab
ALU1E017: Motorcycle
ALU1E018: Bicycle
ALU1E019: Walked
ALU1E020: Other means
ALU1E021: Worked from home

Table 6: Travel Time to Work
Universe: Workers 16 years and over who did not work from home
Source code: B08303
NHGIS code: ALU3
ALU3E001: Total
ALU3E002: Less than 5 minutes
ALU3E003: 5 to 9 minutes
ALU3E004: 10 to 14 minutes
ALU3E005: 15 to 19 minutes
ALU3E006: 20 to 24 minutes
ALU3E007: 25 to 29 minutes
ALU3E008: 30 to 34 minutes
ALU3E009: 35 to 39 minutes
ALU3E010: 40 to 44 minutes
ALU3E011: 45 to 59 minutes
ALU3E012: 60 to 89 minutes
ALU3E013: 90 or more minutes

Table 7: Own Children Under 18 Years by Family Type and Age

Universe: Own children under 18 years
Source code: B09002
NHGIS code: ALU4
ALU4E001: Total
ALU4E002: In married-couple families
ALU4E003: In married-couple families: Under 3 years
ALU4E004: In married-couple families: 3 and 4 years
ALU4E005: In married-couple families: 5 years
ALU4E006: In married-couple families: 6 to 11 years
ALU4E007: In married-couple families: 12 to 17 years
ALU4E008: In other families
ALU4E009: In other families: Male householder, no spouse present
ALU4E010: In other families: Male householder, no spouse present: Under 3 years
ALU4E011: In other families: Male householder, no spouse present: 3 and 4 years
ALU4E012: In other families: Male householder, no spouse present: 5 years
ALU4E013: In other families: Male householder, no spouse present: 6 to 11 years
ALU4E014: In other families: Male householder, no spouse present: 12 to 17 years
ALU4E015: In other families: Female householder, no spouse present
ALU4E016: In other families: Female householder, no spouse present: Under 3 years
ALU4E017: In other families: Female householder, no spouse present: 3 and 4 years
ALU4E018: In other families: Female householder, no spouse present: 5 years
ALU4E019: In other families: Female householder, no spouse present: 6 to 11 years
ALU4E020: In other families: Female householder, no spouse present: 12 to 17 years

Table 8: Household Type (Including Living Alone)
Universe: Households
Source code: B11001
NHGIS code: ALU9
ALU9E001: Total
ALU9E002: Family households
ALU9E003: Family households: Married-couple family
ALU9E004: Family households: Other family
ALU9E005: Family households: Other family: Male householder, no spouse present
ALU9E006: Family households: Other family: Female householder, no spouse present
ALU9E007: Nonfamily households
ALU9E008: Nonfamily households: Householder living alone
ALU9E009: Nonfamily households: Householder not living alone

Table 9: Household Type by Household Size
Universe: Households

Source code: B11016
NHGIS code: ALV1
ALV1E001: Total
ALV1E002: Family households
ALV1E003: Family households: 2-person household
ALV1E004: Family households: 3-person household
ALV1E005: Family households: 4-person household
ALV1E006: Family households: 5-person household
ALV1E007: Family households: 6-person household
ALV1E008: Family households: 7-or-more person household
ALV1E009: Nonfamily households
ALV1E010: Nonfamily households: 1-person household
ALV1E011: Nonfamily households: 2-person household
ALV1E012: Nonfamily households: 3-person household
ALV1E013: Nonfamily households: 4-person household
ALV1E014: Nonfamily households: 5-person household
ALV1E015: Nonfamily households: 6-person household
ALV1E016: Nonfamily households: 7-or-more person household

Table 10: Sex by Educational Attainment for the Population 25 Years and Over
Universe: Population 25 years and over
Source code: B15002
NHGIS code: ALWF
ALWFE001: Total
ALWFE002: Male
ALWFE003: Male: No schooling completed
ALWFE004: Male: Nursery to 4th grade
ALWFE005: Male: 5th and 6th grade
ALWFE006: Male: 7th and 8th grade
ALWFE007: Male: 9th grade
ALWFE008: Male: 10th grade
ALWFE009: Male: 11th grade
ALWFE010: Male: 12th grade, no diploma
ALWFE011: Male: High school graduate (includes equivalency)
ALWFE012: Male: Some college, less than 1 year
ALWFE013: Male: Some college, 1 or more years, no degree
ALWFE014: Male: Associate's degree
ALWFE015: Male: Bachelor's degree
ALWFE016: Male: Master's degree
ALWFE017: Male: Professional school degree

ALWFE018: Male: Doctorate degree
ALWFE019: Female
ALWFE020: Female: No schooling completed
ALWFE021: Female: Nursery to 4th grade
ALWFE022: Female: 5th and 6th grade
ALWFE023: Female: 7th and 8th grade
ALWFE024: Female: 9th grade
ALWFE025: Female: 10th grade
ALWFE026: Female: 11th grade
ALWFE027: Female: 12th grade, no diploma
ALWFE028: Female: High school graduate (includes equivalency)
ALWFE029: Female: Some college, less than 1 year
ALWFE030: Female: Some college, 1 or more years, no degree
ALWFE031: Female: Associate's degree
ALWFE032: Female: Bachelor's degree
ALWFE033: Female: Master's degree
ALWFE034: Female: Professional school degree
ALWFE035: Female: Doctorate degree

Table 11: Ratio of Income to Poverty Level in the Past 12 Months
Universe: Population for whom poverty status is determined
Source code: C17002
NHGIS code: ALWV
ALWVE001: Total
ALWVE002: Under .50
ALWVE003: .50 to .99
ALWVE004: 1.00 to 1.24
ALWVE005: 1.25 to 1.49
ALWVE006: 1.50 to 1.84
ALWVE007: 1.85 to 1.99
ALWVE008: 2.00 and over

Table 12: Household Income in the Past 12 Months (in 2019 Inflation-Adjusted Dollars)
Universe: Households
Source code: B19001
NHGIS code: ALW0
ALW0E001: Total
ALW0E002: Less than $10,000
ALW0E003: $10,000 to $14,999
ALW0E004: $15,000 to $19,999

ALW0E005: $20,000 to $24,999
ALW0E006: $25,000 to $29,999
ALW0E007: $30,000 to $34,999
ALW0E008: $35,000 to $39,999
ALW0E009: $40,000 to $44,999
ALW0E010: $45,000 to $49,999
ALW0E011: $50,000 to $59,999
ALW0E012: $60,000 to $74,999
ALW0E013: $75,000 to $99,999
ALW0E014: $100,000 to $124,999
ALW0E015: $125,000 to $149,999
ALW0E016: $150,000 to $199,999
ALW0E017: $200,000 or more

Table 13: Median Household Income in the Past 12 Months (in 2019 Inflation-Adjusted Dollars)
Universe: Households
Source code: B19013
NHGIS code: ALW1
ALW1E001: Median household income in the past 12 months (in 2019 inflation-adjusted dollars)

Table 14: Public Assistance Income in the Past 12 Months for Households
Universe: Households
Source code: B19057
NHGIS code: ALXL
ALXLE001: Total
ALXLE002: With public assistance income
ALXLE003: No public assistance income

Table 15: Sex by Work Status in the Past 12 Months by Usual Hours Worked per Week in the Past 12 Months by Weeks Worked in the Past 12 Months for the Population 16 to 64 Years
Universe: Population 16 to 64 years
Source code: B23022
NHGIS code: ALY1
ALY1E001: Total
ALY1E002: Male
ALY1E003: Male: Worked in the past 12 months
ALY1E004: Male: Worked in the past 12 months: Usually worked 35 or more hours per week

ALY1E005: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 50 to 52 weeks

ALY1E006: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 48 and 49 weeks

ALY1E007: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 40 to 47 weeks

ALY1E008: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 27 to 39 weeks

ALY1E009: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 14 to 26 weeks

ALY1E010: Male: Worked in the past 12 months: Usually worked 35 or more hours per week: 1 to 13 weeks

ALY1E011: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week

ALY1E012: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 50 to 52 weeks

ALY1E013: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 48 and 49 weeks

ALY1E014: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 40 to 47 weeks

ALY1E015: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 27 to 39 weeks

ALY1E016: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 14 to 26 weeks

ALY1E017: Male: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 1 to 13 weeks

ALY1E018: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week

ALY1E019: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 50 to 52 weeks

ALY1E020: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 48 and 49 weeks

ALY1E021: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 40 to 47 weeks

ALY1E022: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 27 to 39 weeks

ALY1E023: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 14 to 26 weeks

ALY1E024: Male: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 1 to 13 weeks

ALY1E025: Male: Did not work in the past 12 months

ALY1E026: Female
ALY1E027: Female: Worked in the past 12 months
ALY1E028: Female: Worked in the past 12 months: Usually worked 35 or more hours per week
ALY1E029: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 50 to 52 weeks
ALY1E030: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 48 and 49 weeks
ALY1E031: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 40 to 47 weeks
ALY1E032: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 27 to 39 weeks
ALY1E033: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 14 to 26 weeks
ALY1E034: Female: Worked in the past 12 months: Usually worked 35 or more hours per week: 1 to 13 weeks
ALY1E035: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week
ALY1E036: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 50 to 52 weeks
ALY1E037: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 48 and 49 weeks
ALY1E038: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 40 to 47 weeks
ALY1E039: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 27 to 39 weeks
ALY1E040: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 14 to 26 weeks
ALY1E041: Female: Worked in the past 12 months: Usually worked 15 to 34 hours per week: 1 to 13 weeks
ALY1E042: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week
ALY1E043: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 50 to 52 weeks
ALY1E044: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 48 and 49 weeks
ALY1E045: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 40 to 47 weeks
ALY1E046: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 27 to 39 weeks

ALY1E047: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 14 to 26 weeks

ALY1E048: Female: Worked in the past 12 months: Usually worked 1 to 14 hours per week: 1 to 13 weeks

ALY1E049: Female: Did not work in the past 12 months

Table 16: Sex by Industry for the Civilian Employed Population 16 Years and Over

Universe: Civilian employed population 16 years and over

Source code: C24030

NHGIS code: ALZH

ALZHE001: Total

ALZHE002: Male

ALZHE003: Male: Agriculture, forestry, fishing and hunting, and mining

ALZHE004: Male: Agriculture, forestry, fishing and hunting, and mining: Agriculture, forestry, fishing and hunting

ALZHE005: Male: Agriculture, forestry, fishing and hunting, and mining: Mining, quarrying, and oil and gas extraction

ALZHE006: Male: Construction

ALZHE007: Male: Manufacturing

ALZHE008: Male: Wholesale trade

ALZHE009: Male: Retail trade

ALZHE010: Male: Transportation and warehousing, and utilities

ALZHE011: Male: Transportation and warehousing, and utilities: Transportation and warehousing

ALZHE012: Male: Transportation and warehousing, and utilities: Utilities

ALZHE013: Male: Information

ALZHE014: Male: Finance and insurance, and real estate, and rental and leasing

ALZHE015: Male: Finance and insurance, and real estate, and rental and leasing: Finance and insurance

ALZHE016: Male: Finance and insurance, and real estate, and rental and leasing: Real estate and rental and leasing

ALZHE017: Male: Professional, scientific, and management, and administrative, and waste management services

ALZHE018: Male: Professional, scientific, and management, and administrative, and waste management services: Professional, scientific, and technical services

ALZHE019: Male: Professional, scientific, and management, and administrative, and waste management services: Management of companies and enterprises

ALZHE020: Male: Professional, scientific, and management, and administrative, and waste management services: Administrative and support and waste management services

ALZHE021: Male: Educational services, and health care and social assistance

ALZHE022: Male: Educational services, and health care and social assistance: Educational services

ALZHE023: Male: Educational services, and health care and social assistance: Health care and social assistance

ALZHE024: Male: Arts, entertainment, and recreation, and accommodation and food services

ALZHE025: Male: Arts, entertainment, and recreation, and accommodation and food services: Arts, entertainment, and recreation

ALZHE026: Male: Arts, entertainment, and recreation, and accommodation and food services: Accommodation and food services

ALZHE027: Male: Other services, except public administration

ALZHE028: Male: Public administration

ALZHE029: Female

ALZHE030: Female: Agriculture, forestry, fishing and hunting, and mining

ALZHE031: Female: Agriculture, forestry, fishing and hunting, and mining: Agriculture, forestry, fishing and hunting

ALZHE032: Female: Agriculture, forestry, fishing and hunting, and mining: Mining, quarrying, and oil and gas extraction

ALZHE033: Female: Construction

ALZHE034: Female: Manufacturing

ALZHE035: Female: Wholesale trade

ALZHE036: Female: Retail trade

ALZHE037: Female: Transportation and warehousing, and utilities

ALZHE038: Female: Transportation and warehousing, and utilities: Transportation and warehousing

ALZHE039: Female: Transportation and warehousing, and utilities: Utilities

ALZHE040: Female: Information

ALZHE041: Female: Finance and insurance, and real estate, and rental and leasing

ALZHE042: Female: Finance and insurance, and real estate, and rental and leasing: Finance and insurance

ALZHE043: Female: Finance and insurance, and real estate, and rental and leasing: Real estate and rental and leasing

ALZHE044: Female: Professional, scientific, and management, and administrative, and waste management services

ALZHE045: Female: Professional, scientific, and management, and administrative, and waste management services: Professional, scientific, and technical services

ALZHE046: Female: Professional, scientific, and management, and administrative, and waste management services: Management of companies and enterprises

ALZHE047: Female: Professional, scientific, and management, and administrative, and waste management services: Administrative and support and waste management services

ALZHE048: Female: Educational services, and health care and social assistance

ALZHE049: Female: Educational services, and health care and social assistance: Educational services

ALZHE050: Female: Educational services, and health care and social assistance: Health care and social assistance

ALZHE051: Female: Arts, entertainment, and recreation, and accommodation and food services

ALZHE052: Female: Arts, entertainment, and recreation, and accommodation and food services: Arts, entertainment, and recreation

ALZHE053: Female: Arts, entertainment, and recreation, and accommodation and food services: Accommodation and food services

ALZHE054: Female: Other services, except public administration

ALZHE055: Female: Public administration

Table 17: Occupancy Status
Universe: Housing units
Source code: B25002
NHGIS code: ALZK

ALZKE001: Total
ALZKE002: Occupied
ALZKE003: Vacant

Table 18: Tenure
Universe: Occupied housing units
Source code: B25003
NHGIS code: ALZL

ALZLE001: Total
ALZLE002: Owner occupied
ALZLE003: Renter occupied

Table 19: Year Structure Built
Universe: Housing units
Source code: B25034
NHGIS code: AL0D

AL0DE001: Total
AL0DE002: Built 2014 or later
AL0DE003: Built 2010 to 2013
AL0DE004: Built 2000 to 2009
AL0DE005: Built 1990 to 1999
AL0DE006: Built 1980 to 1989

AL0DE007: Built 1970 to 1979
AL0DE008: Built 1960 to 1969
AL0DE009: Built 1950 to 1959
AL0DE010: Built 1940 to 1949
AL0DE011: Built 1939 or earlier

Table 20: Tenure by Vehicles Available
Universe: Occupied housing units
Source code: B25044
NHGIS code: AL0N
AL0NE001: Total
AL0NE002: Owner occupied
AL0NE003: Owner occupied: No vehicle available
AL0NE004: Owner occupied: 1 vehicle available
AL0NE005: Owner occupied: 2 vehicles available
AL0NE006: Owner occupied: 3 vehicles available
AL0NE007: Owner occupied: 4 vehicles available
AL0NE008: Owner occupied: 5 or more vehicles available
AL0NE009: Renter occupied
AL0NE010: Renter occupied: No vehicle available
AL0NE011: Renter occupied: 1 vehicle available
AL0NE012: Renter occupied: 2 vehicles available
AL0NE013: Renter occupied: 3 vehicles available
AL0NE014: Renter occupied: 4 vehicles available
AL0NE015: Renter occupied: 5 or more vehicles available

Table 21: Median Gross Rent (Dollars)
Universe: Renter-occupied housing units paying cash rent
Source code: B25064
NHGIS code: AL05
AL05E001: Median gross rent

Table 22: Gross Rent as a Percentage of Household Income in the Past 12 Months
Universe: Renter-occupied housing units
Source code: B25070
NHGIS code: AL1B
AL1BE001: Total
AL1BE002: Less than 10.0 percent
AL1BE003: 10.0 to 14.9 percent
AL1BE004: 15.0 to 19.9 percent

AL1BE005: 20.0 to 24.9 percent
AL1BE006: 25.0 to 29.9 percent
AL1BE007: 30.0 to 34.9 percent
AL1BE008: 35.0 to 39.9 percent
AL1BE009: 40.0 to 49.9 percent
AL1BE010: 50.0 percent or more
AL1BE011: Not computed

Table 23: Median Value (Dollars)
Universe: Owner-occupied housing units
Source code: B25077
NHGIS code: AL1H
AL1HE001: Median value (dollars)

Bibliography

Alvarez-Melis, David, and Tommi S. Jaakkola. "On the Robustness of Interpretability Methods." arXiv, June 21, 2018. https://doi.org/10.48550/arXiv.1806.08049.

Assaad, Rayan H., and Yasser Jezzini. "Assessing and Predicting Green Gentrification Susceptibility Using an Integrated Machine Learning Approach." *Local Environment* 29, no. 8 (August 2, 2024): 1099–1127. https://doi.org/10.1080/13549839.2024.2353058.

Atkins, Drew. "The Backlash Against Tech in Seattle: What We Can Learn from San Francisco." Seattle magazine, May 27, 2014. https://seattlemag.com/seattle-living/backlash-against-tech-seattle-what-we-can-learn-san-francisco/.

Baker, Dwayne Marshall, and Bumsoo Lee. "How Does Light Rail Transit (LRT) Impact Gentrification? Evidence from Fourteen US Urbanized Areas." *Journal of Planning Education and Research* 39, no. 1 (March 1, 2019): 35–49. https://doi.org/10.1177/0739456X17713619.

"Be Careful When Interpreting Predictive Models in Search of Causal Insights — SHAP Latest Documentation." Accessed June 5, 2025. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html.

Bick, Carolyn. "How to Tackle Lack of Regional Affordable Housing? King County Calls In Experts." South Seattle Emerald, July 25, 2024. https://southseattleemerald.org/news/2024/07/25/how-to-tackle-lack-of-regional-affordable-housing-king-county-calls-in-experts.

Blog, Gawain Chin-DS. "Correlation vs SHAP: Understanding Feature Importance in ML Models." *Medium* (blog), February 18, 2024. https://medium.com/@gawainchin/correlation-vs-shap-understanding-feature-importance-in-ml-models-d6b52b1fba28.

Brummet, Quentin, and Davin Reed. "The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, July 16, 2019. https://doi.org/10.21799/frbp.wp.2019.30.

Chapple, Karen, Ate Poorthuis, Matthew Zook, and Eva Phillips. "Monitoring Streets through Tweets: Using User-Generated Geographic Information to Predict Gentrification and Displacement." *Environment and Planning B: Urban Analytics and City Science* 49, no. 2 (February 1, 2022): 704–21. https://doi.org/10.1177/23998083211025309.

DEV Community. "The Story of XGBoost: A Machine Learning Revolution," November 23, 2024. https://dev.to/aashwinkumar/the-story-of-xgboost-a-machine-learning-revolution-bib.

"Disappearing Affordable Housing in a High-Tech Town – The Seattle Globalist," February 24, 2020. https://seattleglobalist.com/2020/02/24/affordable-housing-seattle-disappearing-amazon-lihi-tech4housing/89275.

Freyd, Benjamin. "Large Tech Office Openings and the Onset of Gentrification," 2022. https://www.semanticscholar.org/paper/Large-Tech-Office-Openings-and-the-Onset-of-Freyd/5b28d93b85802db83029ff54aed07a2c66f12b39.

Gebreyesus, Yibrah, Damian Dalton, Sebastian Nixon, Davide De Chiara, and Marta Chinnici. "Machine Learning for Data Center Optimizations: Feature Selection Using Shapley Additive exPlanation (SHAP)." *Future Internet* 15, no. 3 (March 2023): 88. https://doi.org/10.3390/fi15030088.

Goldman, Alexandra. "THE 'GOOGLE SHUTTLE EFFECT:' GENTRIFICATION AND SAN FRANCISCO'S DOT COM BOOM 2.0." University of California, Berkeley, Spring 2013. https://svenworld.com/wp-content/uploads/2014/01/Goldman_PRFinal.pdf.

Hess, Chris L. "Light-Rail Investment in Seattle: Gentrification Pressures and Trends in Neighborhood Ethnoracial Composition." *Urban Affairs Review* 56, no. 1 (January 1, 2020): 154–87. https://doi.org/10.1177/1078087418758959.

Hwang, Hyunseung, Andrew Bell, Joao Fonseca, Venetia Pliatsika, Julia Stoyanovich, and Steven Euijong Whang. "SHAP-Based Explanations Are Sensitive to Feature Representation." arXiv, May 13, 2025. https://doi.org/10.48550/arXiv.2505.08345.

Kendall, M. G. "A New Measure of Rank Correlation." *Biometrika* 30, no. 1/2 (1938): 81–93. https://doi.org/10.2307/2332226.

"King County Department of Assessments: eReal Property." Accessed June 11, 2025. https://blue.kingcounty.com/Assessor/eRealProperty/default.aspx.

"King County GIS Open Data." Accessed June 11, 2025. https://gis-kingcounty.opendata.arcgis.com/.

Kingston, John, and Margaret Bolton. "New Approaches to Funding Not‐for‐profit Organisations." *International Journal of Nonprofit and Voluntary Sector Marketing* 9, no. 2 (January 25, 2004): 112–21.

Kohn, Margaret. "What Is Wrong with Gentrification?" *Urban Research & Practice* 6, no. 3 (November 1, 2013): 297–310. https://doi.org/10.1080/17535069.2013.846006.

Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." arXiv, November 25, 2017. https://doi.org/10.48550/arXiv.1705.07874.

"Machine Learning Models Can Produce Reliable Results Even with Limited Training Data | University of Cambridge," September 19, 2023. https://www.cam.ac.uk/research/news/machine-learning-models-can-produce-reliable-results-even-with-limited-training-data.

Maharawal, Manissa M. "Tech-Colonialism: Gentrification, Resistance, and Belonging in San Francisco's Colonial Present." *Anthropological Quarterly* 95, no. 4 (2022): 785–813.

Manson, Steve, Jonathan Schroeder, David Van Riper, Katherine Knowles, Tracy Kugler, Finn Roberts, and Steven Ruggles. "IPUMS National Historical Geographic Information System: ACS 5-Year Estimate 2006-2010." Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D050.V19.0.

———. "IPUMS National Historical Geographic Information System: ACS 5-Year Estimate 2015-2019." Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D050.V19.0.

Molnar, Christopher. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3rd ed., 2025. https://christophm.github.io/interpretable-ml-book/.

Palafox, Leon, and Pedro Ortiz-Monasterio. "Predicting Gentrification in Mexico City Using Neural Networks." In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–5, 2020. https://doi.org/10.1109/IJCNN48605.2020.9207685.

Pegler, Claudia, Hankan Li, and Dorina Pojani. "Gentrification in Australia's Largest Cities: A Bird's-Eye View." *Australian Planner* 56, no. 3 (July 3, 2020): 191–205. https://doi.org/10.1080/07293682.2020.1775666.

Reades, Jonathan, Jordan De Souza, and Phil Hubbard. "Understanding Urban Gentrification through Machine Learning." *Urban Studies* 56, no. 5 (April 1, 2019): 922–42. https://doi.org/10.1177/0042098018789054.

scikit-learn. "Train_test_split." Accessed June 4, 2025. https://scikit-learn/stable/modules/generated/sklearn.model_selection.train_test_split.html.

"Seattle City Council Introduces New Affordable Housing Policy Options – Puget Sound Sage." Accessed February 28, 2025. https://pugetsoundsage.org/new-affordable-housing-policy-options/.

Shmueli, Galit. "To Explain or to Predict?" *Statistical Science* 25, no. 3 (2010): 289–310. https://doi.org/10.1214/10-STS330.

Smith, Neil. *The New Urban Frontier: Gentrification and the Revanchist City*. New York: Routledge, 1996.

Tate, Robert F. "Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation." *The Annals of Mathematical Statistics* 25, no. 3 (1954): 603–7.

Thackway, William, Matthew Ng, Chyi-Lin Lee, and Christopher Pettit. "Building a Predictive Machine Learning Model of Gentrification in Sydney." *Cities* 134 (March 1, 2023): 104192. https://doi.org/10.1016/j.cities.2023.104192.

Urban Health Collaborative. "A Measure of Gentrification for Use in Longitudinal Public Health Studies in the US," March 10, 2022. https://drexel.edu/uhc/resources/briefs/Measure-of-Gentrification-for-Use-in-Longitudinal-Public-Health-Studies-in-the-US/.

Wallick, Kiki. "Gentrification in Seattle: Amazon Overpowers the City Council." Gettysburg College, May 4, 2020. https://cupola.gettysburg.edu/cgi/viewcontent.cgi?article=1945&context=student_scholarship.

Wang, Ruoniu, and Courtnee Melton-Fant. “Does Inclusionary Housing Alleviate the Negative Health Impacts of Gentrification?” *Housing Policy Debate* 33, no. 1 (January 2, 2023): 72–84. https://doi.org/10.1080/10511482.2022.2099933.

“Welcome to cuML’s Documentation! — Cuml 25.04.00 Documentation.” Accessed May 25, 2025. https://docs.rapids.ai/api/cuml/stable/.

Zheng, Xinliang, Yi Xie, Xujiao Yang, Muhammad Nasir Amin, Sohaib Nazar, Suleman Ayub Khan, Fadi Althoey, and Ahmed Farouk Deifalla. “A Data-Driven Approach to Predict the Compressive Strength of Alkali-Activated Materials and Correlation of Influencing Parameters Using SHapley Additive exPlanations (SHAP) Analysis.” *Journal of Materials Research and Technology* 25 (July 1, 2023): 4074–93. https://doi.org/10.1016/j.jmrt.2023.06.207.