Using Machine Learning to Predict Housing Prices in DC

Background

Housing plays a crucial role in urban development and the well-being of people. Access to stable housing is essential for maintaining a good quality of life and ensuring good health, education, and work opportunities. At an economic level, housing prices carry substantial weight in determining the overall social consumption level of the entire market (Kohler & Merwe, 2015). Therefore, high housing prices can significantly impact social development and the economy, creating economic instability and increasing social inequality (Iacoviello, 2015).

The housing market in the District of Columbia is a matter of concern due to the continued rise in prices. Although not as unaffordable as in San Francisco or New York City, the high prices in D.C. put a disproportionate strain on low- and middle-income families, who usually spend more than 30% of their income on housing (Lawrence, 2013). Washington's status as the political and economic hub of the country undoubtedly contributes to high prices. However, the housing market is influenced by various factors, including shifts in interest rates, general economic conditions, and demographic changes resulting from high levels of gentrification (Wilhelmsson et al., 2021). In the past decade, surges of well-paid individuals and developers have propelled housing development and values, resulting in significant growth in neighborhoods like Eckington and Brookland, with home prices in Trinidad soaring approximately 141% since 2009 (Alpert, 2015).

In order to address the persistent issue of population displacement caused by the housing burden, the District of Columbia government has implemented a range of affordable housing policies. To achieve this goal, the government has established affordable housing targets on a community-by-community basis, ensuring equitable distribution of housing opportunities and attracting new investments by constructing houses in areas with concentrated poverty. For instance, the Housing Production Trust Fund (HPTF) offers developers cost-efficient loans to build and maintain affordable housing (McCabe, 2021). Given the interdependence of housing prices with social development and the political economy, analyzing housing price trends can provide policymakers with comprehensive insights to develop targeted housing policies. It is important to tailor policies to the specific circumstances of each region to ensure their effectiveness in addressing local poverty and housing shortages.

Pardo and Pérez utilized GEE and GLMM methodologies to predict the average housing price in quarter periods, providing valuable insights and contributions to macro research on housing prices from an economic perspective (Pardo & Pérez, 2012). Besides, scholars have also employed Bayesian methods to analyze housing prices through the prediction of hedonic prices (Wheeler et al., 2013). However, traditional econometric methods and theories may no longer be sufficient. A more comprehensive and compatible framework is necessary to evaluate and identify the determinants of housing prices, enabling policymakers to formulate effective policies. This also means that housing price forecasts should account for the mutual influence between the local social environment, population distribution, and housing conditions.

Therefore, this study aims to explore the factors that influence the housing market in the District of Columbia. Specifically, it seeks to answer what determines housing prices. To achieve that, I will apply machine learning models to predict housing sales prices in the area. This approach will enable me to predict the price more accurately than traditional methods by using feature selection and reduction, and evaluating the importance of various features helps to identify the factors that shape the price. By gaining insight into the potential factors that shape the housing market dynamics, policymakers can enhance existing policies and better respond to housing demand.

Data

To build accurate models for predicting housing prices, I needed a dataset at the individual level. The primary dataset I utilized is at the city level and contains sale prices for all wards in DC, obtained from Kaggle. This dataset covers the period from 1947 to 2018, with individual housing units as the unit of analysis. To supplement this information, I incorporated two other datasets containing data on median income levels for families with children and racial proportions in each ward. These datasets were obtained from the Kids Count Data Center website and were originally from the U.S. Census Bureau's 2000 and 2010 Decennial Censuses. The generated dataset for this research consists of 41,649 observations and 49 features. Adding social factors to the original dataset serves a better policy implication purpose, as socioeconomic and demographic factors may also influence housing market trends.

Despite the comprehensiveness of the dataset, it has some limitations that should be acknowledged. Firstly, the race and income data only cover information for each ward from 2010 to 2020, which may limit the depth of analysis in studying the impact of social factors on housing prices. Secondly, the dataset exhibits a high degree of variability, with significant differences observed between the lowest and highest values. Furthermore, despite the vastness of the dataset, it still contains numerous missing values, which may affect the reliability and generalizability of the findings.

The following tables summarize continuous and categorical variables that will be applied to the models. In *Table 1*, I provided a brief explanation of the variables for each. The outcome variable, PRICE, has a mean value of approximately 607204, and a median value of about

534452, indicating that the income data is left-skewed. The minimum value of 1 suggests a possible error in the original data collection.

Table 1: Continuous Variables (N = 44695)

Variable	Description	Mean	Std. Dev.	Min	Median	Max
PRICE	Sale price for individual housing	607203.99	578903.03	1.00	534451.52	25100000.00
KITCHENS	Number of kitchens	1.27	0.68	0	1.00	44.00
GBA	Square footage of Gross Building Area, which includes all heated and cooled areas of a property, as well as any below-grade living space	1675.86	757.06	407.00	1575.93	15902.00
income($)	Income level for families with children	83663.68	59297.93	24096.00	60792.46	236711.00

Table 2: Categorical Variables (N = 44695)

Variable	Description	Observation	
HEAT	The type of heat	Forced Air	18527
		Hot Water Rad	15373
		Warm Cool	10133
		Ht Pump	622
CNDTN	Condition of the house	Good	21124
		Average	17524
		Very Good	5395
		Excellent	652
STRUCT	The structure type of the house	Row Inside	19270
		Single	10866
		Semi-Detached	6629
		Row End	5576
		Multi	2354
WARD	Wards in D.C.	Ward 6	8003
		Ward 4	7877
		Ward 5	7374
		Ward 7	6298
		Ward 3	5348
		Ward 1	3874
		Ward 8	3443
		Ward 2	2478

Table 2 shows the summary of building characteristics and ward distribution. Each variable contains observations for different distinct values. I found that some categorical variables had distinct values with a frequency of less than 500, which were too small to be

meaningful for prediction. Therefore, I decided to drop these values to improve the accuracy of my model.

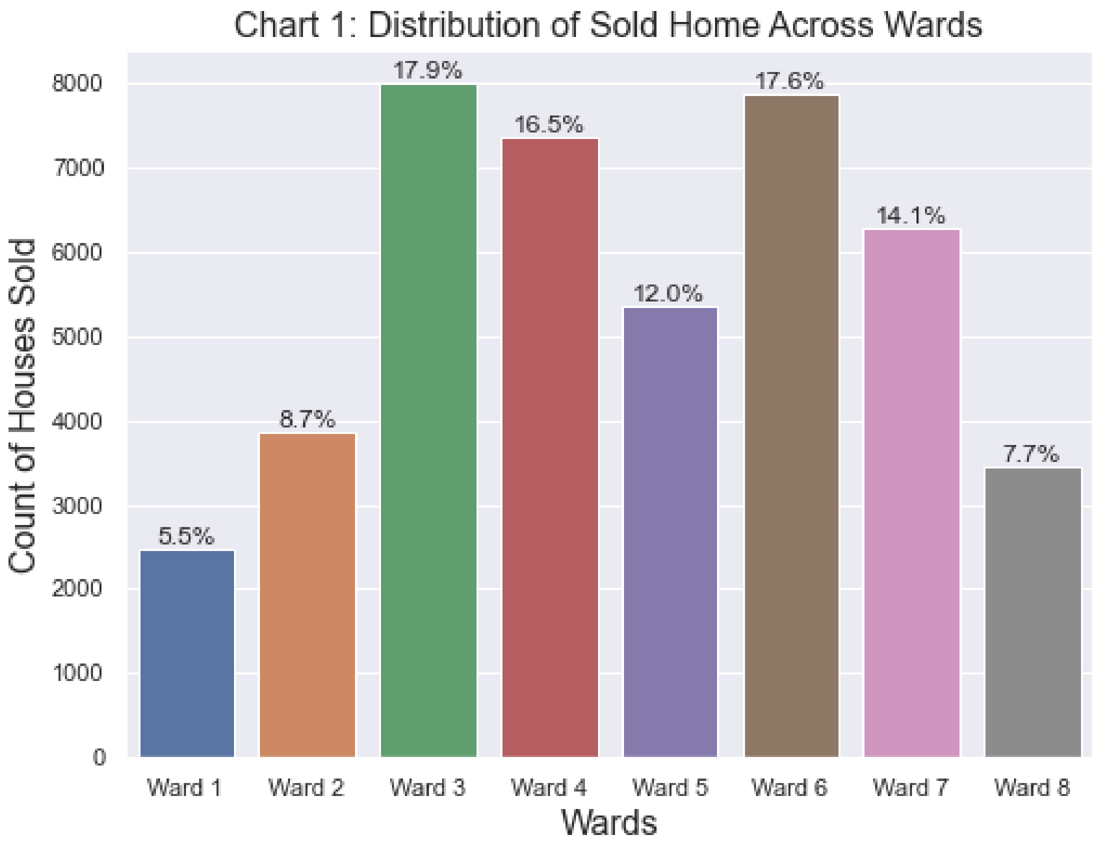


Chart 1 displays a histogram of the number of homes sold in different wards of Washington D.C. It is evident that Ward 3, Ward 4, and Ward 6 are the most popular living areas, whereas Ward 1 and Ward 8 had relatively fewer homes sold compared to other wards. *Chart 2* shows the trend of housing prices across different wards. The graph reveals a generally positive trend, indicating that housing prices have increased over time, except for a dip between 2007-2010, potentially due to the financial crisis. Notably, Ward 2 has the highest prices, followed by Ward 3, and they exhibited significantly higher prices compared to other wards, particularly after 2005.

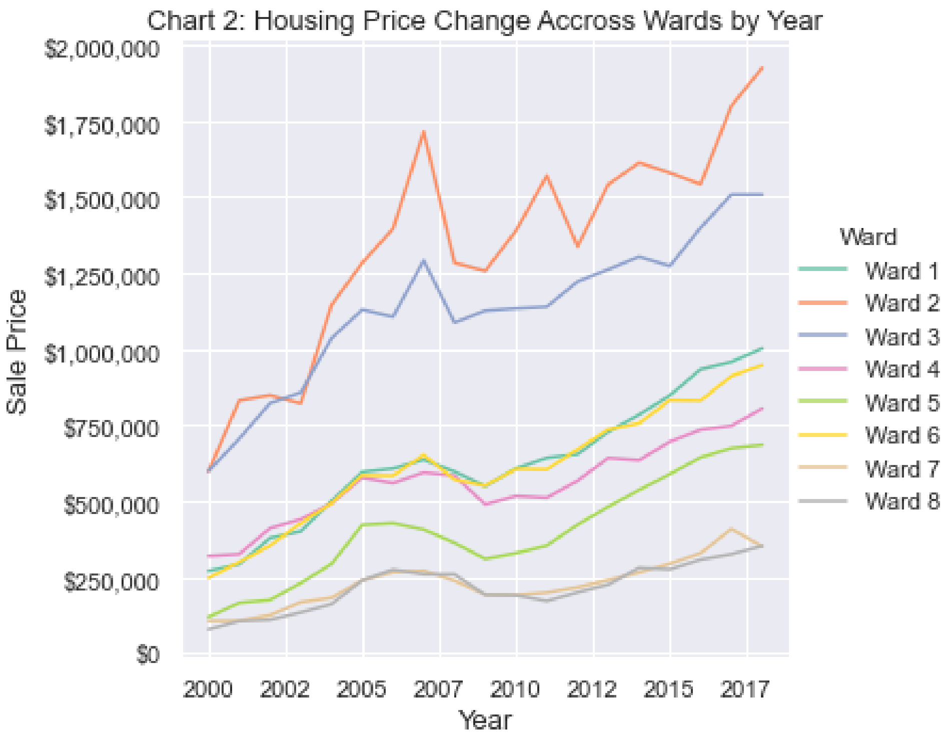


Chart 2: Housing Price Change Accross Wards by Year

In *Chart 3*, I examined the relationship between race proportions and housing price changes. The Asian and Hispanic populations in Washington D.C. are relatively small compared to the White and Black populations. Areas with a higher proportion of White residents tend to have higher housing prices, while areas with a higher proportion of Black residents tend to have lower housing prices.

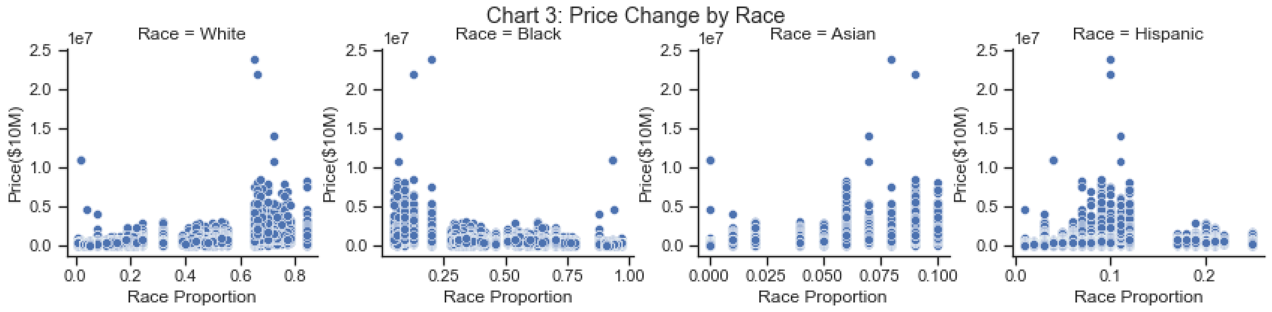


Chart 3: Price Change by Race

Race = White

Race = Black

Race = Asian

Race = Hispanic

Methodology

My research focuses on predictive analysis, specifically developing a model that can accurately predict housing prices and identify the key features that shape them. The target variable is the sale price, which is numerical. The feature matrix includes internal housing characteristics such as the number of bathrooms, bedrooms, kitchens, fireplaces, and square footage, as well as building characteristics such as the year of construction, number of stories, roof type, wall type, structure type, and condition. Additionally, social factors, including income level, race proportion, and date and wards, are also taken into account.

Prediction

My goal is to develop a high-accuracy model by comparing the performance of one parametric model and three non-parametric models. For the parametric model, I chose LASSO regression which helps prevent overfitting and performs feature selection through regularization. This is particularly useful since some features had very low values during the previous data processing. LASSO shrinks some coefficients to zero, removing them from the model by imposing a penalty on the impact of each feature. Scholars have used LASSO to predict stock prices (Roy, 2015). For example, they worked with Goldman Sachs Group to predict the future price of a chosen stock. The model outperformed Ridge as the RMSE and MAPE values were lower.

I selected three non-parametric models for my analysis, including Decision Tree, Random Forest, and XGBoost. The Decision Tree model was chosen due to its computational efficiency and intuitive decision process. The model structure consisted of a root node, internal nodes, and leaf nodes, with each layer making judgments recursively based on specific conditions (Roy, 2020). As an example, Decision Tree was successfully applied to predict airfare prices, with the

Random Forest regressor outperforming other algorithms in accuracy for predicting flight ticket prices between March and June 2019 (Shaw, 2020). I chose Random Forest due to its anti-overfitting properties and high accuracy. This model generated multiple regression trees, each trained on a different bootstrap sample of the training dataset using a random subset of input features. The final projection was obtained by averaging the output of all the trees (Galasso et al., 2022). For instance, it was used to predict the number of COVID-19 cases at the U.S. county level, and the model achieved a high R2 and low MAE. It could handle an expansive training feature set while also delivering good performance (Galasso, 2021).

I chose XGBoost due to its ability to handle large and complex datasets, allowing for high accuracy while avoiding overfitting. Unlike Random Forest, which builds decision trees independently and relies on randomness, XGBoost is a form of regularized Gradient Boosting that creates a sequence of decision trees, accumulates weak learners at each step, and optimizes the trees using gradient descent (Pan, 2018). For example, XGBoost has been used to predict hourly PM2.5 concentrations in air quality data, with scholars finding it to have the best performance among the different regression models they tested (Pan, 2018).

Testing Accuracy and Feature Importance

I split the dataset into training and test sets before constructing the models, and used several performance evaluation metrics including MAE, MSE, RMSE, and R-squared. R-squared measures how well the model fits the data, with higher values indicating better results. While non-parametric models typically perform better on complex and large datasets with non-linear relationships, parametric models such as Lasso may introduce more bias and underfit the training data. The performance of decision trees, random forests, and XGBoost models may improve progressively as they become more complex, but this can lead to overfitting and high variance in the training set. To improve accuracy and reduce overfitting, I used grid search technique to tune hyperparameters of the models. Grid search tests a range of hyperparameters and evaluates the performance of the model using cross-validation, allowing us to choose the optimal set of hyperparameters that will help interpret the data more effectively. In the end, I will evaluate the feature importance of my prediction model to identify factors that impact house price forecasts. Understanding the dynamics of the housing market and its various factors is crucial for policymakers to make informed decisions that promote fair and sustainable housing for all residents.

Findings

After evaluating the models, it became apparent that the performance improved gradually with each model. *Table 3* shows the performance of the four models. To obtain the optimal alpha that delivers the lowest cross-validation error and the ideal accuracy, I used LassoCV and divided the data into six folds for Lasso regression. The optimal alpha value of 0.005 produced an MSE of 0.16, an MAE of 0.27, and an R-squared of 0.72. These outcomes demonstrate a significant improvement in performance compared to the results obtained before applying LassoCV, where the accuracy rate was only 0.22. The substantial boost in performance when setting alpha to a small value suggests that the unpenalized model was overfitting the data. With alpha =0.005, the penalty added to the function is negligible, and most of the coefficients of housing interior features are not shrunk to zero, while the coefficients related to building characteristics are shrunk to zero. This indicates that the interior features of the housing data are more significant in determining the target variable than the building characteristics for Lasso regression.

After applying the decision tree model, I used a validation curve to determine the ideal max_depth, dividing the data into 6 folds. The resulting graph indicated that the model tended to overfit after reaching a maximum depth of 8. Therefore, I chose a max_depth of 7, which resulted in an MAE of 0.28 and an accuracy rate of approximately 68%. Subsequently, I performed parameter tuning using grid search, which involves trying different hyperparameters to find the combination that yields the best performance on the validation set. This optimized the parameters of max_leaf_nodes, min_sample_leaf, and max_depth to yield the best model with the best estimator. The obtained results demonstrated a marginal improvement over the baseline, with an MAE of 0.29 and an accuracy rate of 70%. However, I noticed that this led to overfitting in the training dataset, and the previous hyperparameters were already close to optimal.

In implementing the random forest model, I followed the same procedure as the decision tree model by setting up a six-fold cross-validation and creating a validation curve. Prior to determining the optimal max_depth, the model achieved an accuracy rate of 78%. After setting the max_depth to 8, however, the accuracy rate slightly decreased to 75%. This suggests that the original model may have been overfitting to the training data. To address this issue, I also used the grid search method. While the model performed well on the training set with the previous hyperparameters, there is a risk that it may not generalize well to new data.

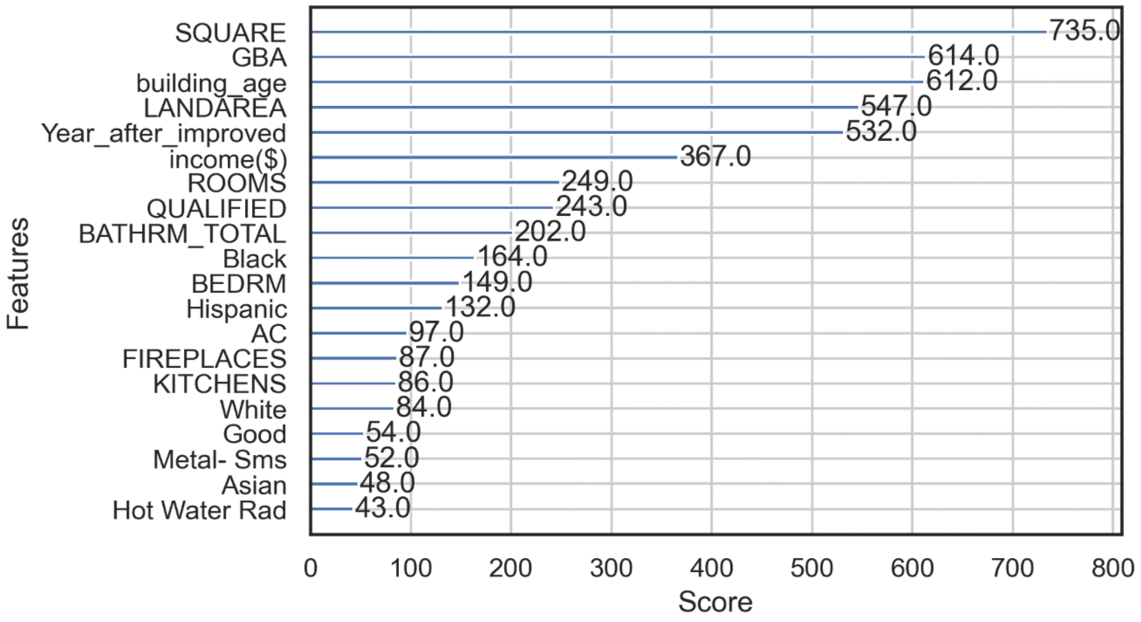
The XGBoost model, with a max depth of 5, achieved an 80% accuracy rate. However, analysis of the results indicated that the decision tree model performed the worst among the four models due to its high cross-validation score, while XGBoost achieved the highest accuracy rate. Nevertheless, the validation curve revealed severe overfitting in the data. Therefore, the random forest model is the optimal choice for this study, as it achieved a satisfactory accuracy rate while demonstrating relatively low overfitting. Furthermore, the analysis revealed that parameter tuning can enhance model performance, and selecting the optimal model depends on balancing accuracy and overfitting.

Table 3: Performance Comparison

Model	MAE	MSE	RMSE	R squared	Cross Val Score
LASSO	0.28	0.16	0.40	0.72	0.18
Decision Tree	0.29	0.17	0.41	0.71	0.22
Random Forest	0.26	0.14	0.38	0.75	0.18
XGBoost	0.21	0.11	0.33	0.81	0.17

The rank of feature importance in *Chart 4* indicates that housing footage, gross building area, building age, land area, building improvement, and income are the main factors for predicting housing prices.

Chart 4: Feature Importance



Conclusion

The findings of this study are of significant importance for the housing market in DC. Housing features such as square footage, building age, and interior attributes are major contributors to shaping housing prices. Policymakers can leverage these insights to develop effective housing policies that promote affordability and allocate resources optimally. It also allows real estate developers to make informed decisions regarding the construction of new properties and the renovation of existing ones. Similarly, potential homebuyers can use these findings to guide their purchase decisions and negotiate prices based on the key attributes of a property.

However, one of the major limitations is the computational expense involved in running the random forest algorithm and conducting grid search on a large dataset. For future study, poverty rate and interest rate can also be added as potential factors. Moreover, the use of standard scaler and log transformation, while useful for the algorithm, can also increase the difficulty for researchers to interpret the results.

The difference in performance between the parametric and non-parametric models was not significant, therefore, further research can use other metrics to determine which approach is best suited for predicting housing prices in DC. Other metrics could be used to determine which approach is best suited for this type of prediction task. Additionally, the research predicts the housing prices in DC in general, but each ward has its own unique situation that may require more focused analysis. Future research could delve deeper into the specific characteristics of different areas in DC to gain a more nuanced understanding of housing prices and the factors that affect them.

Bibliography

Alpert, D. (2015). *House prices are skyrocketing in central DC neighborhoods, but not in outlying ones*. Greater Greater Washington. Retrieved March 1, 2023, from https://ggwash.org/view/40203/house-prices-are-skyrocketing-in-central-dc-neighborhoods-but-not-in-outlying-ones

Annie E. Casey Foundation. (2021). POVERTY BY WARD IN DISTRICT OF COLUMBIA. Retrieved May 09, 2023, from https://datacenter.aecf.org/data/tables/9070-poverty-by-ward

Annie E. Casey Foundation. (2021). RACE/ETHNICITY OF TOTAL POPULATION BY WARD IN DISTRICT OF COLUMBIA. Retrieved May 09, 2023, from c

Correa, C. (2018). DC_Property_data. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/christophercorrea/dc-residential-properties

Galasso, Cao, D. M., & Hochberg, R. (2022). A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. *Chaos, Solitons and Fractals, 156*, 111779–111779. https://doi.org/10.1016/j.chaos.2021.111779

Iacoviello, M. (2015). Financial business cycles. *Review of Economic Dynamics*, 18(1), 140-163.

Marion Kohler, & Michelle van der Merwe. (2015). Long-run Trends in Housing Price Growth. *In Bulletin* (Issue September Quarter 2015). Reserve Bank of Australia.

McCabe, B. (2021). *The Housing Production Trust Fund, explained*. Greater Greater Washington. Retrieved March 1, 2023, from https://ggwash.org/view/80343/what-is-the-housing-production-trust-fund-anyway

Pan. (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series. Earth and Environmental Science*, 113(1), 12127–. https://doi.org/10.1088/1755-1315/113/1/012127

Pardo, & Pérez, T. (2013). Analysis of housing prices by GEE and GLMM methodologies: a longitudinal study. *Applied Stochastic Models in Business and Industry*, *29*(5), 552–563. https://doi.org/10.1002/asmb.1940

Roy, A. (2020). A dive into decision trees. Retrieved May 7, 2023, from https://towardsdatascience.com/a-dive-into-decision-trees-a128923c9298

Roy, Mittal, D., Basu, A., & Abraham, A. (n.d.). Stock Market Forecasting Using LASSO Linear Regression Model. In *Afro-European Conference for Industrial Advancement* (pp. 371–381). Springer International Publishing. https://doi.org/10.1007/978-3-319-13572-4_31

SHAW, V. (2020). Airfare Price Prediction. *Kaggle*. Retrieved from https://www.kaggle.com/code/vinayshaw/airfare- price- prediction

Wadud, I. K. M. M., Bashar, O. H., Ali Ahmed, H. J., & Dimovski, W. (2022). Property price dynamics and asymmetric effects of economic policy uncertainty: New evidence from the Australian Capital Cities. *Accounting & Finance*, *62*(4), 4359–4380. https://doi.org/10.1111/acfi.13014

Wilhelmsson, M., Ismail, M., & Warsame, A. (2021). Gentrification effects on housing prices in neighbouring areas. *International Journal of Housing Markets and Analysis*, *15*(4), 910–929. https://doi.org/10.1108/ijhma-04-2021-0049

Wheeler, Páez, A., Spinney, J., & Waller, L. A. (2014). A Bayesian approach to hedonic price analysis. *Papers in Regional Science*, 93(3), 663–683. https://doi.org/10.1111/pirs.12003

Appendix A

Data Preprocessing

This appendix describes data preprocessing steps that were applied to the dataset before training the models.

The income and race data in the primary dataset only covered information for each ward from 2010 to 2018, while the housing dataset had information from 1947 to 2018. In order to have more observations for housing market analysis, I merged the income and race data from 2000 with the housing data corresponding to each ward from 2000 to 2009. This means that for the years 2000 to 2009, the income and race data were the same as the values from 2000. This approach enabled me to increase the sample size and obtain more comprehensive data for the analysis without significantly affecting the results.

The distribution of the target variable, price, was very right-skewed. To help the models interpret the training data more effectively, a log transformation was applied to the target variable to convert it into a more systematic distribution.

Numerical variables, such as square footage and number of bedrooms, had a wide range of values. To ensure that each variable was given equal importance during training, standard scaling was applied to normalize the values.