**Analysis of the Housing Market in Two US Cities from 2020-2022**

| Nikhil Muthuvenkatesh  *Drexel University*  nm3297@drexel.edu | Rohan Ukkalam  *Drexel University*  rsu24@drexel.edu | Luqing Qi  *Drexel University*  lq53@drexel.edu | Seyi Oyesiku  *Drexel University*  *so536@drexel..edu* |
| --- | --- | --- | --- |

***Abstract* - The aim of this study is to develop a predictive model for real estate prices using multiple value datasets. The first dataset is from a realtor database that includes property information on property buyers and sellers in the United States, which was filtered down to New York and Philadelphia. Using data visualization, we explored the datasets to identify preprocessing requirements and select appropriate machine learning models to achieve our goals. Our study contributes to the real estate industry by providing a reliable model for predicting future property prices.**

**Keywords**: Real Estate, Pennsylvania, New York, Pricing Prediction

#### **Introduction**

Since the COVID-19 pandemic in 2020, real estate pricing has been quite volatile. The initial lockdowns led to a large drop in property values across the country, but shortly thereafter there was a huge spike that has continued up until now. This project will implement a model that is trained with real estate data over the last two years to predict the future prices for housing. The goal of this capstone project is to correctly predict real estate pricing for people considering when to purchase real estate in the cities of New York and Philadelphia.

Our project is driven by our shared interest in real estate investment as graduate students, and our goal is to provide valuable insights into the future of this market. We believe that our findings could benefit not only individual home buyers but also financial institutions such as banks, mortgage companies, and escrow companies. Moreover, we hope to offer actionable information to anyone considering purchasing a home in the cities we analyze. In the next couple of years, we will find this information useful. A model that can predict house pricing in these two cities within the next couple of months will include visualizations for a layman to better understand the trends for the past few years and include several areas of interest. We will be training the models with real estate data from both cities in the period of 2020-2022.

The nature of the project is that it is tackling what is known as a Time Series Analysis Problem. Time series analysis is “a statistical technique dealing in time series data, or trend analysis” (Tyagi, 2021). As the project is attempting to analyze market trends in order to predict future prices, it falls firmly under the purview of a time series analysis problem, and it is with this mindset that the project continued its next phases.

#### **Dataset Description**

We have selected several datasets for our project, with the first dataset originally sourced from [www.realtor.com](http://www.realtor.com) a real estate listing website operated by the News Corp subsidiary Move, Inc. and based in Santa Clara, California. However, as this dataset includes properties from across the United States, we created a filter to focus only on properties in Pennsylvania and New York, which are the two states of interest for our project.

We are making use of a version of this dataset which is available on Kaggle. Our goal is to predict the price of properties in these two states, with the ‘price’ column serving as the target variable and the other data feature serving as possible predictors. By focusing on these specific regions and leveraging this dataset, we believe we can gain valuable insights into the real estate markets of Pennsylvania and New York, and offer accurate predictions of future property prices.

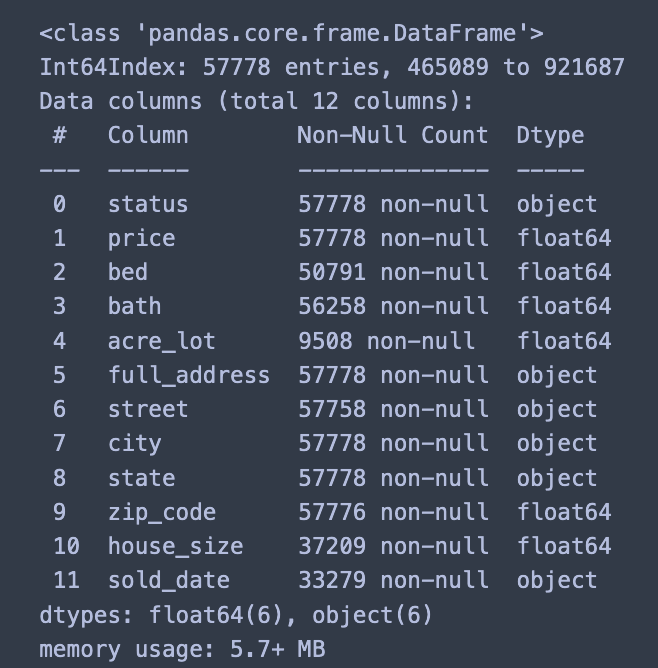


Fig. 1. Dataset Variables

The second database used is a public-use database by Fannie Mae and Freddie Mac. Loan-level Public-Use Databases (PUDBs) are released annually to meet FHFA’s requirement under 12 U.S.C 4543 and 4546(d) to publicly disclose data about the Enterprises’ single-family and multifamily mortgage acquisitions. The datasets supply mortgage lenders, planners, researchers, policymakers, and housing advocates with information concerning the flow of mortgage credit in America’s neighborhoods. The PUDB single-family datasets include loan-level records that include data elements on the income, race, and gender of each borrower as well as the census track location of the property, loan-to-value(LTV) ratio, age of mortgage note, and affordability of the mortgage.

The third database used is the Zillow Home Value Index. The Zillow Home Value Index measures monthly changes in median price at property-level in different cities in the United States. We use this dataset to study the housing price change in Philadelphia and New York from the years 2020 to 2022.

The fourth database used is the Rolling Sales Data at NYC. The Department of Finance's Rolling Sales file lists properties sold in the last twelve-month period in New York City for tax classes 1,2, and 4. These fields include the neighborhood, the building type, the square footage of the real estate, and the sale price.

#### **Exploring Data**

1. Realtor Data

1. Imbalanced Data

Imbalanced data refers to classification problems where we have unequal instances for different classes of the target variable. The only data in this dataset that seemed imbalanced were the zip codes, but it makes sense because they are assigned to each property, as opposed to being an actual feature of the property. The zip code is based on how the city organizes different parts of itself. Therefore, we had expected there to be some imbalance in this feature. The rest of the features we are using for our analysis are balanced already.

1. Null Values

The values in the dataset that can be considered null are the NaN values for the sets we are working with. These represent values that are missing from the columns in our datasets. Figure 1 shows all the columns in our realtor dataset that have null values. Figure 2 below shows the distribution of the null values in our dataset.

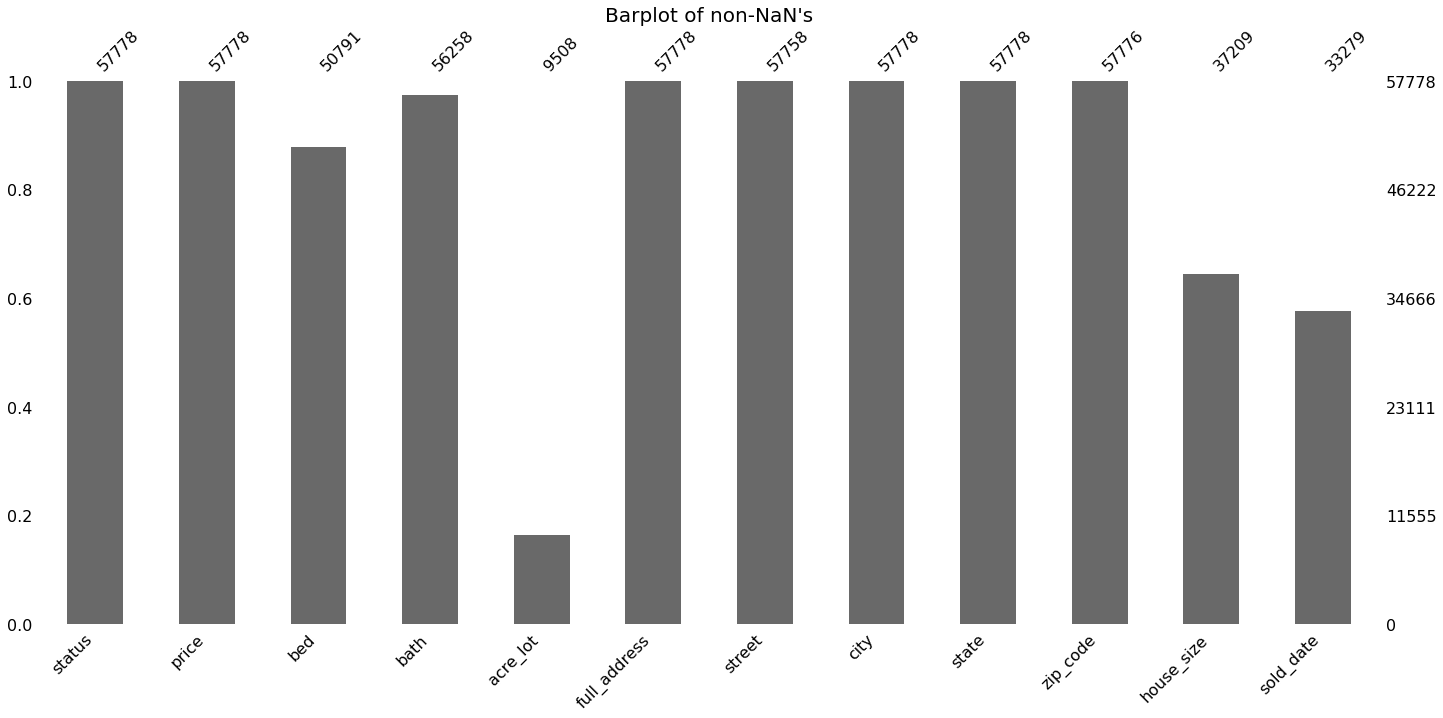


Fig. 2. NaN Values Visualization

1. Categorical Features

Categorical features are normally

non-numeric, represented by strings, variables that represent a limited number of values that assign qualitative property. They usually represent categories. The categorical features within this data are status, full\_address, street, city, and state. Due to this nature, for future machine learning models, we will have to preprocess these categorical features using various encoding techniques such as label encoding, One Hot encoding, or hash encoding. The following graph is a heat map of various Philadelphia zip codes and the average price of property within said zip code.

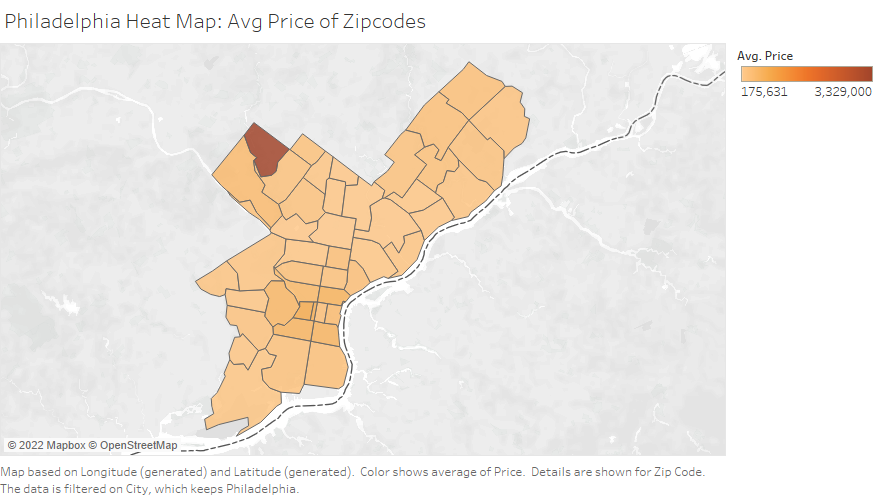


Fig. 3. Heat Map of Average Price

1. Continuous Features

Continuous features are normally numerical features that are quantified, but they have no limit in terms of the range of values that can be encompassed in them. Our continuous features in this study are price, acre lot size, and house size.

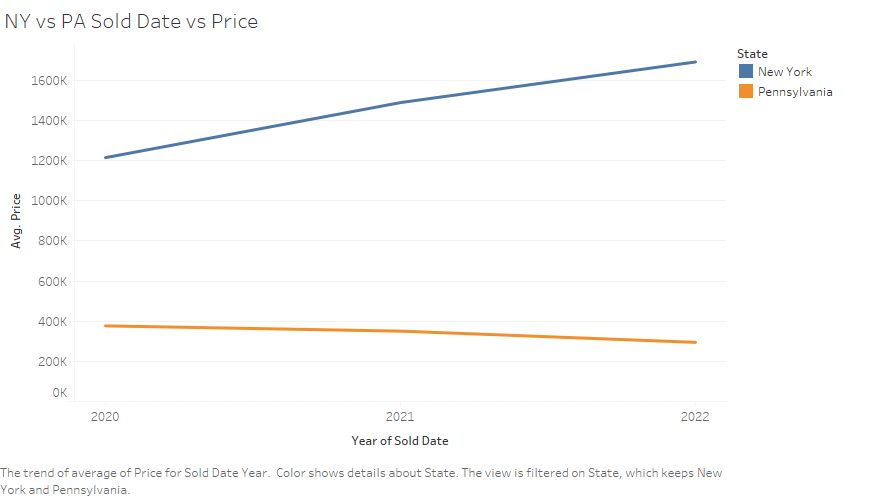


Fig. 4. Average Price over Time

1. Discrete Features

Within the data are several discrete data features, such as bed and bath, that are of interest to us. The number of bedrooms and bathrooms for a given property would indicate its potential price. The following figure shows the average amount of beds a property has per its location. .

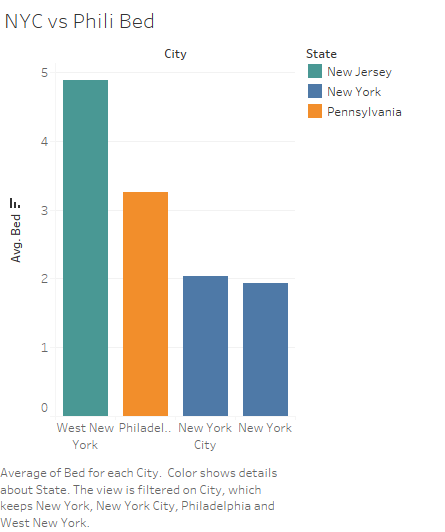


Fig. 5. Average beds in a property

2. Loan-level Public-Use Databases (PUDBs)

In the dataset, we noticed that the bank allowed us to have two different borrowers apply the loan for each mortgage loan. We didn’t find any NA value in our dataset and did some basic demographics analysis about borrowers.

The Average Total Monthly Income Amount is equal to $10872.35. Most mortgage borrowers have 360 months to pay back their loans. The Average Mortgage interest range is equivalent to 2.84. The average Mortgage amount is equal to $227,637.9.The average age for Borrower one is equal to 44, and the average age for Borrower two is equal to 68. We got 34 categorical features and 22 continuous features.

1. Data Features

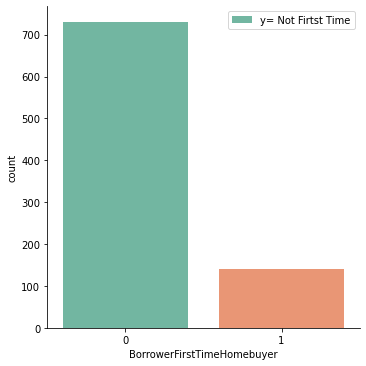


Fig. 6. Countplot of First-Time House Owner

In this graph, the number zero represents someone who is not buying a house for the first time. Number one represents someone who is a first-time home buyer. From this graph, we could tell, the majority of the borrowers are not buying a house for the first time.

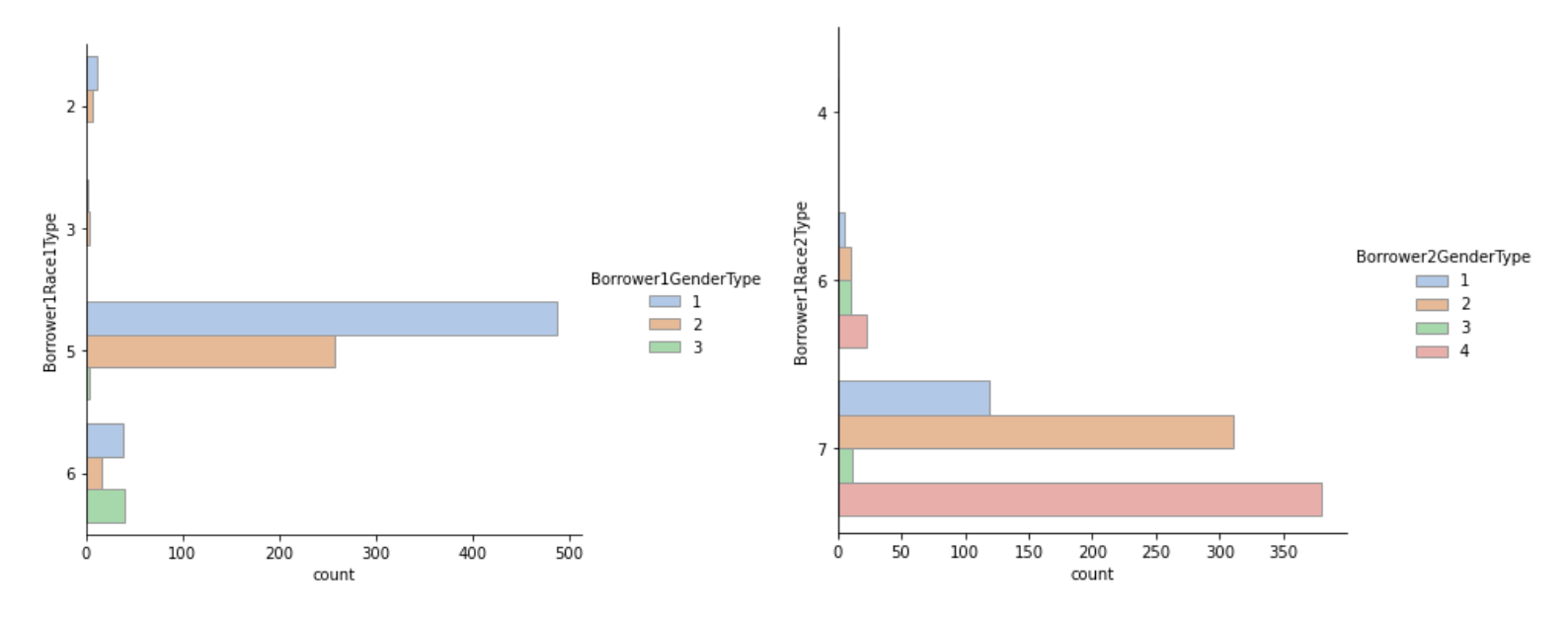


Fig. 7

From the upper graph, For the hue-axis, we can tell that the Numeric code indicates the sex of the first or primary Borrower. Number one represents males. Number two represents females. Number three represents information not provided by the Borrower. Number four represents No Co-Borrower, for the primary borrower is male. For co-borrower, we could find most loans don't have co-borrower. For the y-axis, we can tell the race of the Borrower. Numeric code indicates the Borrower's race. 2 = Asian, 3= Black or African American, 4 = Native Hawaiian or other Pacific Islander, 5 = White, 6= Information not provided by Borrower, 7= Not Applicable. From the left graph, we can tell that most of the borrowers are white males. From the right chart, we can find that a borrower is an institution, a corporation, or a partnership.

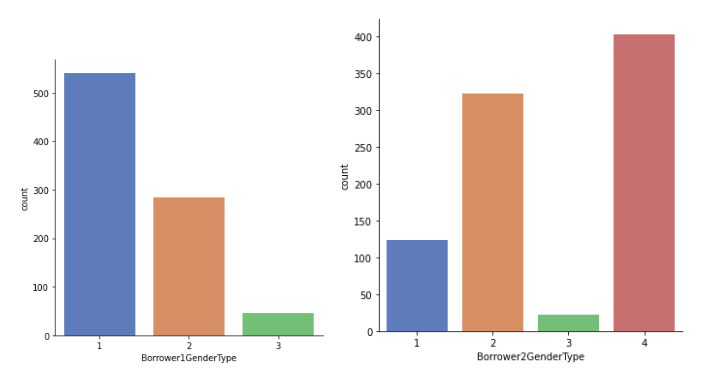


Fig. 8. Gender Distribution of Buyers

This graph shows the gender distribution for borrower one and borrower two. Numeric code indicating the sex of the co-borrower. 1= Male, 2 = Female, 3 = Information not provided by borrower, 4= No-borrower. Our primary borrowers have twice as many male borrowers as female borrowers. For our co-borrower, we could see more females than males. We could also find that a certain amount of loan just got one borrower.

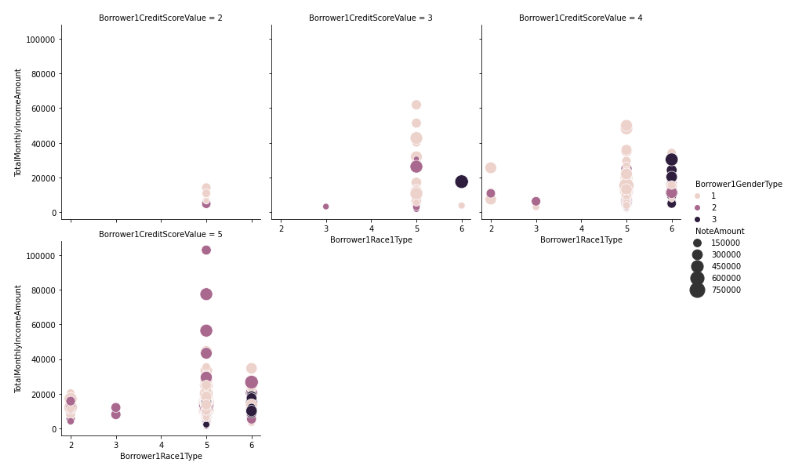


Fig. 9. Loan Distribution of Borrowers

These graphs show the number of loans under different credit ratings for primary borrowers of different races and monthly incomes. Numeric code indicating the race of the Borrower. 1= American Indian or Alaska Native, 2= Asian, 3= Black or African American, 4= Native Hawaiian or other Pacific Islander, 5= White, 6= Information not provided by Borrower. Numeric code indicating the sex of the first or primary borrower. 1= Male, 2=Female, 3= Information not provided by the borrower. From this graph, we could see that Caucasians have borrowed more than other groups with the same credit score. We also found that white women tended to borrow more with a credit rating of five.

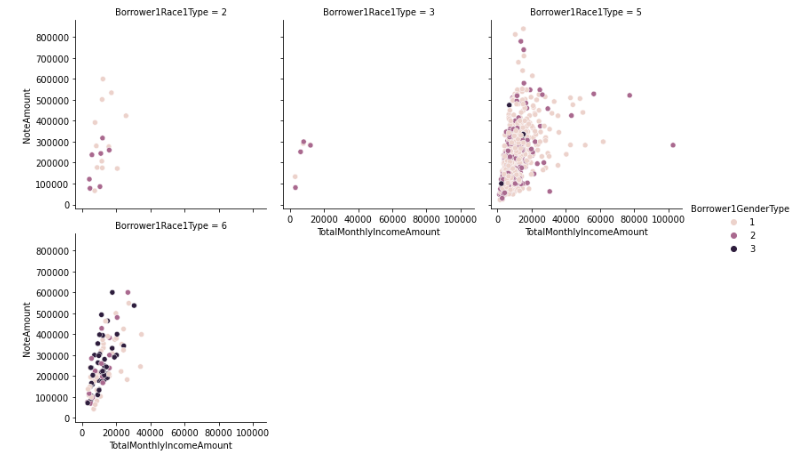


Fig. 10.

The numbers in this graph and the graph above represent the same race group and gender. We found a positive correlation between monthly income and the number of loans. At the same time, we found that whites have more income and loans than other ethnic groups.

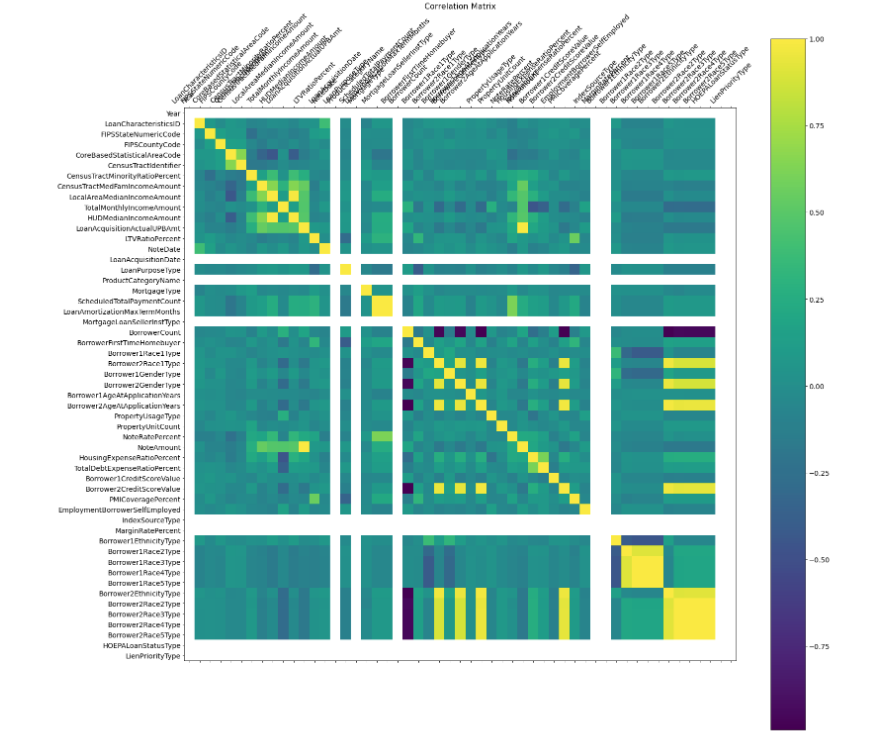


Fig. 11. Correlation Matrix

From this correlation matrix, we found that Loan Acquisition Actual UPB Amount, Census Tract Median Family Income amount, HUD Median Income Amount, Local Area Median Income Amount, and Total Monthly Income Amount. From this graph, we can conclude that three important factors can affect the number of loans you can take:

1. The number of loans you borrow from the bank.
2. The median income for the house purchased.
3. The average monthly payment of a borrower.

3. Zillow Home Value Index

We got 893 records and 278 dimensions in our dataset. We filtered data into Philadelphia and New York to study the housing marketing price change during the pandemic. Most of the variables are numeric variables. We picked up variables from 1/01/2020 to 9/30/2022. We did data wrangling through Python and made data visualization through Excel.

1. Data Features



Fig. 12. Housing Price over time

Before we started our work, we tried to understand the impact of the pandemic on housing prices in different regions. From this chart, we can see that the average home price in New York is twice that of Philadelphia, and property prices in both places have steadily increased during the pandemic.

4. Rolling Sales Data at NYC

We picked our dataset from 10/1/2021 to 9/30/2022. We got some missing value problems in the dataset. We dropped the variables that we don't need, such as EASEMENT and APARTMENT NUMBER. We used model imputation for our category variables. We used mean imputation for our numeric variables. We got 6 category variables and 13 numeric variables.

1. Data Features

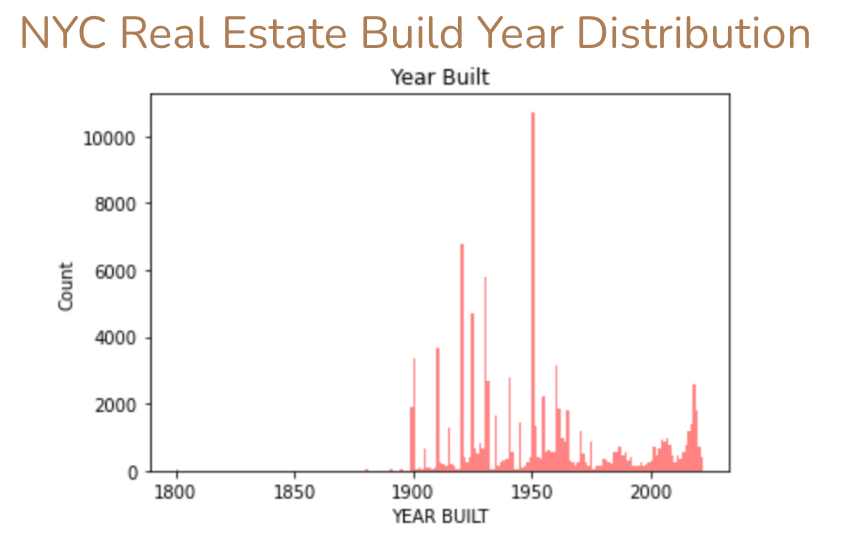


Fig. 13. NYC Build Year Distribution

From this chart, we can see that from 1900 to 1950, there was an upward trend in new Real Estate construction in New York City, reaching its peak in 1950.

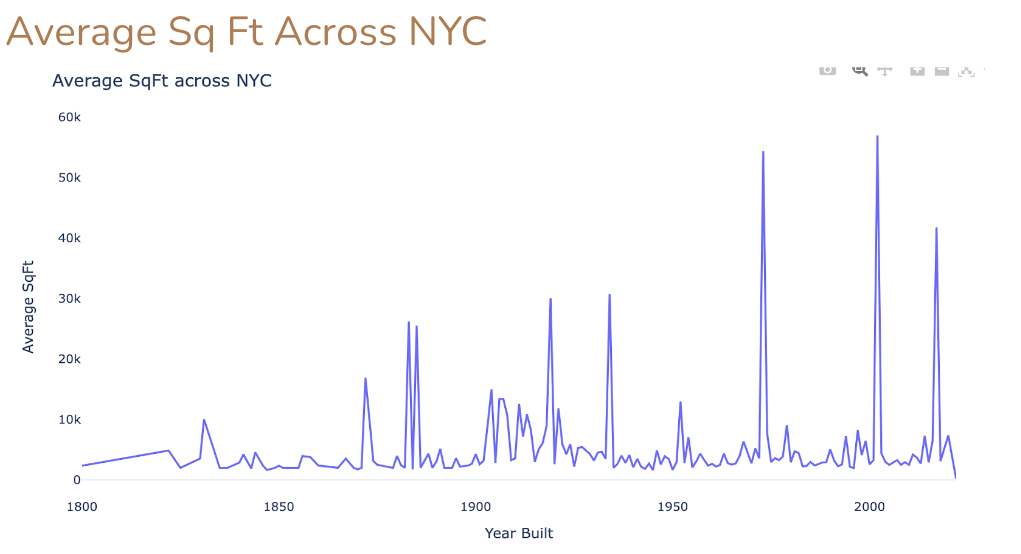


Fig. 14. NYC Average Square Footage

We can find that in the period between 1900 and 1950, there were two peaks. The first peak was in 1920, and the 20s were also the era of the great boom in the real estate market, and it became fashionable to build skyscrapers between 1920 to 1930. New York City built more skyscrapers in these ten years than at any other time in history. The Chrysler Building was the symbol of that era. The second peak may involve Roosevelt’s New Deal, which dramatically increased investment in infrastructure. The Lincoln Tunnel and the Empire State Building have become icons of this period.

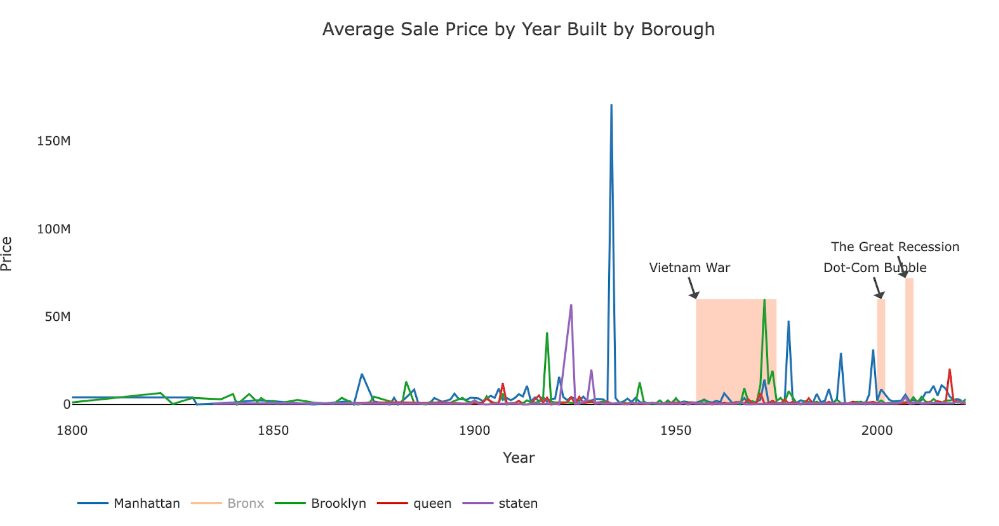


Fig. 15. Avg Price of NYC borough over time

We can see the same trend in this graph of average sales price change in different New York City areas. Manhattan had the highest average home price between 1900 to 1950. In the later stages of the Vietnam War, the average home price in the Brooklyn area was higher than in Manhattan. We found that the New York area strongly correlated with the U.S. economy. For example, the dot-com bubble and the 08 subprime mortgage crisis significantly impacted the real estate market.

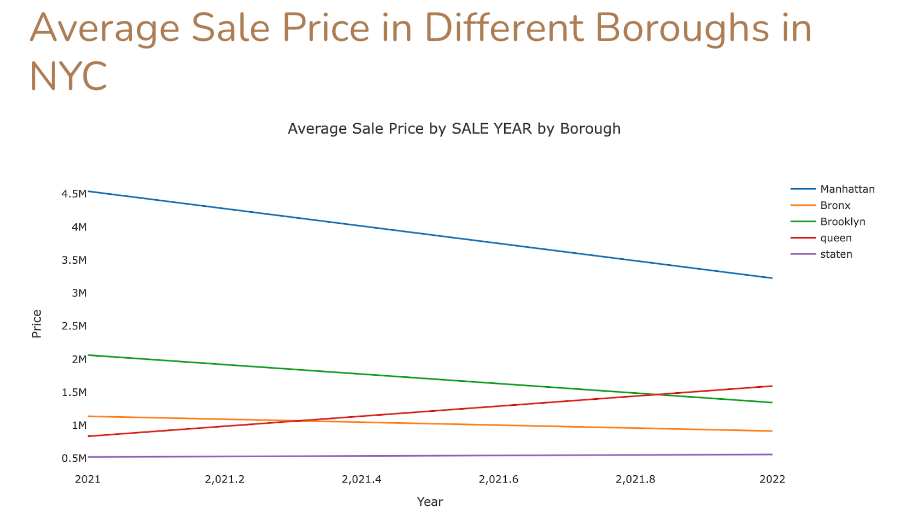


Fig. 16. NYC Sale Price over Time

We also explored average sale prices in different boroughs of New York City. From this graph, we can tell that the average sale price in Manhattan dropped from Oct 2021 to Set 2022. In contrast, we can see a clear trend improvement in Queens' average sale price.

#### **IV. Model Selection**

The project focuses on predicting a target variable - price- which necessitates the use of supervised machine learning models. We employed both regression and classification models in our analysis. Specifically, we used classification to group price predictions into specific ranges, and then regression models to predict the actual pricing values and compare them to the true values.

To begin, we used a multiple linear regression model to analyze the impact of multiple features on the price. The coefficients next to each feature in the multiple linear regression equation. In addition to multiple linear regression, we also utilized other regression models to further refine our predictions. By combining these models with classification techniques, we were able to develop a robust approach to predicting property prices in Philadelphia and New York.

*Y = ax0+bx + cx + dx…*

Eq. 1

The next regression model that was used was Extreme Gradient Boosting (XGBoost), which has become a very popular machine learning tool lately. It is an ensemble method that uses multiple decision trees that are able to correct the prediction errors made by the previous trees. The main benefit of XGBoost is its ability to apply gradient boosting quickly, efficiently, and effectively.

Decision trees can be very effective at both classification and regression tasks. They use the features that are fed into the model to make decisions based on binary (yes or no) questions. Starting at the root with the first question, the tree is able to build more nodes for itself based on the answers that the data from the features provides.

However, sometimes we may want to use multiple decision trees with our data to potentially get better results. This is where the Random Forest came into play. Like the decision tree, a random forest can be both a classifier or regressor. In this study, we used it as both. The Random Forest is also similar to the XGBoost method in that each decision tree that is built out next is able to learn from previous trees.

Finally, the logistic regression classification method was used for this study. Because they are easily interpreted, logistic regression models can also be used to describe the relationship between one independent variable with multiple dependent variables. They also do not require too much computational power and are relatively easily trained and implemented.

#### **V. Clustering Analysis for Public Use Database from Faannie Mae and Freddie Mac**

Clustering is a technique commonly used in exploratory data analysis to gain insights into the data structure. The primary objective of clustering is to identify subgroups in the data where data points in each subgroup are similar, while data points in different clusters are dissimilar. The process involves finding homogeneous subgroups in the data where the similarity between data points can be measured using metrics such Euclidean-based distance or correlation-based distance.

In our analysis of the Public Use Database from Fannie Mae and Freddie Mac, we performed K-means clustering with an optimized number of clusters set to four, determined using the elbow method. The features we chose include: Total Monthly Income Amount and Note Amount. The elbow method involves plotting the sum of squared distance (SSE) between data points and their assigned clusters’ centroid against the number of clusters. We selected the optimal number of clusters at the point where they start to flatten out and form an elbow. Our analysis revealed four distinct groups in the data.

Cluster 1: This cluster includes individuals with an average monthly income between $5,000 to $10,000, an average total monthly income of $6,149.6, and an average note amount of $127,487. This group mainly consists of first-time buyers and has the largest number of USDA loans. The main reason for the loan is cash-out refinance, and the average interest rate is 2.83. USDA loans are a type of mortgage loan that is offered by the United States Department of Agriculture (USDA) to help low – and moderate – income borrowers in rural areas become homeowners. So we called our cluster 1 farmer group.

Cluster 2: This cluster includes individuals with an average monthly income between $10,000 to $15,000, an average total monthly income of $10,338, and an average note amount of $259,292. The average interest rate is 2.87, and the average age of the borrower is 45. This group has the largest number of VA loans.VA loans are a type of mortgage loan that are available to military service members, veterans, and eligible surviving spouses. These loans are backed by the Department of Veterans Affairs (VA) and offer several benefits, including: first, no down payment required. Second, with a VA loan, borrowers don’t have to pay for private mortgage insurance. Third, VA loans typically offer lower interest rates than conventional loans, which can result in significant savings over time. So, we call our cluster 2 warrior group.

Cluster 3: This cluster includes individuals with an average monthly income above $15,000, an average total monthly income of $16,231, and an average note amount of $446,868. The average interest rate is 2.90, and the average credit score is 4.48. The main reasons for the loan are purchase, and no cash-out refinance.

Cluster 4: This cluster includes individuals with an average monthly income above $50,000, an average total monthly income of $53,704, and an average note amount of $404,492. The main reasons for the loan are purchase, and no cash-out refinance, and the average interest rate is 2.77. The average age of borrower one is 50.

K-means clustering has several advantages, including simplicity of implementation, scalability to large datasets, and generalizability. However, there are also some drawbacks to this method. Firstly, K-means clustering is sensitive to outliers in the data, which can significantly affect the resulting clusters. Secondly, choosing the optimal number of clusters (k) can be a challenging task, as it requires prior knowledge of the data or multiple iterations to find the best k value. Lastly, K-means clustering assumes that the clusters are spherical, with equal variances in all dimensions, and that each cluster has roughly the same number of observations. If any of theses assumptions are violated, K-means clustering may not produce accurate results. As described by Jim Brown (2021), “If any one of these three assumptions is violated, then k-means will fail.”

#### **VI. Predictive Model for Loan-level Public-Use Databases (PUDBs)**

We have a dataset with 34 categorical and 22 continuous features. We removed some variables that were not useful for our model, such as Year, LoanCharacteristicsID, Bank, FIPSStateNumerCode, and CoreBasedStatistical AreaCode. To reduce the number of highly correlated variables, we used the Variance Inflection Factor (VIF) method. Then, we used a regression model to identify the most important features for predicting Note Amount.

To prepare our data for modeling, we checked for skewness and kurtosis and applied Box Cox transformation to correct for highly skewed data. Additionally, we standardized our data to ensure that all features were on the same scale and equally important for our machine learning algorithm. Our ultimate goal is to use the multiple regression model to predict the Note Amount based on the input data.

One advantage of multiple regression is that it allows for the simultaneous examination of the relationship between a dependent variable and several independent variables. It also facilitates the identification of the strength and direction of these relationships, and it can be used to predict the value of the dependent variable based on the independent variables. Additionally, multiple regression can help to control for the effects of confounding variables.

However, there are several limitations to multiple regression. For instance, it assumes that the relationship between the dependent and independent variables is linear, which is not always the case. Furthermore, it assumes that there is no multicollinearity among the independent variables, which can lead to unreliable estimates of the regression coefficients. The assumption of normality and equal variance of errors may also not hold in some cases, and the analysis can be sensitive to outliers and influential observations, potentially affecting the results.

The initial version of our model showed a small eigenvalue of 9.65e-28, which suggests the presence of strong multicollinearity problems or singularity issues in the design matrix. In order to solve this issue, we employed stepwise regression, a method for selecting the most relevant predictor variables while excluding the less significant ones in linear regression models. By systematically identifying the crucial variables from a pool of potential predictors, stepwise regression aims to enhance the model’s performance.

We were able to produce an optimal MSE, MAE, and RMSE values of 0.0743,0.211,and 0.273 respectively, and additionally captured a higher R-squared of 0.761 compared to the original model R-squared value of 0.497. We also got the top 5 important variables in the regression model by looking at the absolute values of the coefficients. They are: Borrower Gender Type, MortgageloanSellerInstType, MortgageType, PropertyUnitOUNT, HOEPALoanstatusType.

In the context of PUDBs, MortgageloansellerInstType refers to the institution that sold the mortgage loan to the purchasing entity. This could be a depository institution like a bank or savings and loan association, a non-depository institution like a mortgage company or finance company that originates and sells mortgage loans, a private entity like an investment bank that purchases mortgage loans and pools them into mortgage-backed securities (MBS) to sell to investors, or a federal agency like Fannie Mae or Freddie Mac that purchases mortgage loans from originating lenders. The MortgageloanSellerInstType field is an important piece of information for understanding the mortgage market and analyzing loan performance.

PropertyUnitCount refers to the number of units that are part of a given mortgaged property. This field can have several possible values, such as: 1 unit, 2-4 units, 5-49 units, and 50 or more units. The propertyUnitCount field is an important variable in analyzing loan performance because the number of units in a mortgaged property can affect the borrower’s ability to repay the loan.

HOEPALoanstatusType refers to the status of a mortgage loan under the Homeownership and Equity Protection Act. HOPEA is a federal law that provides additional protection for certain types of high-cost mortgage loans. For example, borrowers with low credit scores or history of delinquencies and defaults may be considered higher risk by lenders.

We still got smallest eigenvalue of a matrix is very close to 0, it indicates that there maybe problem of multicollinearity. Inorder to address this issues, we may need to consider the following options: obtain more data, and using regularization methods such as Ridge or Lasso regression.

#### **VII. Results and Discussion**

There are a few different types of results that we received from our study. The first part of the results we saw were with the training and test data that we created from our realtor dataset with both the regression and classification models. The second part of the results comes from using our models to predict pricing with features from more recent real estate data to find out how well these models do with real time data. We do not train the model with this new data, we simply input the full dataset with the features we need, and then compare the results with the true pricing in that dataset.

Before we were able to create the models, we had to select the features we would use. We dropped many of the categorical features that we had because they were mainly used for us to filter out data initially. We also had a separate column with the categorical version of the price to be used with our classification models. The data we used for the regression models included the original price variable, and we used the categorical version for the classification models.

Because we have two different types of models, we had to create a training and test set to use for regression and training/test sets to use for the classification. Once both of those splits were completed, we applied a Robust Scaler to the training and test sets because we have a number of different ranges of values for each of our features, so we need to have some scale to use when applying our models.

The multiple linear regression model did the worst out of all models implemented with the training and test data, as expected. It performed with an accuracy of 0.645, and this was quite low compared to our other models. The mean absolute error was 589415.15, and the root mean square error was 843770.85. Given that our pricing ranges from $40,000 to about $9.9 million, these errors are reasonable. However, we will need a model that is able to predict housing with a higher accuracy.

The next model implemented was the XGBoost. Unlike the other models which the team has studied extensively through prior coursework, this was a new model to work with. However, because it works similarly to a random forest regressor, the team still had some familiarity with it. The errors were vastly lower than the multiple linear regressions. We saw a mean absolute error of 0.7230 and a root mean square error of 0.125. These values are normalized for the dataset, so it seems that this model performed fairly well with the training and test.

Compared to the multiple linear regression, the decision tree regressor fared very well with an accuracy value of 0.822 with a normalized mean absolute error of 0.0067 and a normalized root mean square error of 0.0473. This accuracy is much better than what we saw with the multiple linear regression, and the normalized errors are even lower than what we saw with the XGBoost. From what we have seen so far, the decision tree regressor seems to perform the best.

The decision tree classifier performed very well with classifying pricing in the correct ranges. The accuracy was 0.9752, and we can see the confusion matrix in Figure 17 below.

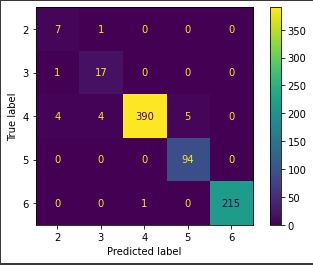


Fig. 17: Confusion Matrix for the Decision Tree Classifier

We see that this model was able to label the houses ranging upwards of $100,000 the best out of all of the other price ranges. The largest percentage of our dataset fell in class 3 of prices ranging from $100,000 to $150,000. As we can see, there were only a small number of properties that were classified incorrectly by this model. Predicting a price range as opposed to an exact price seems to be much easier with machine learning, given the accuracy we saw with this model compared to the regression models used previously.

Finally, after all the models were created, trained, and tested, we validated them on future data. To do this, the models had been exclusively trained on data from the last decade, i.e. 2010 to 2020, in both the training and testing face. Taking the three best models - Decision Tree Classifier, Decision Tree Regressor, and Random Forest, these models were then given data from 2020 onwards, and their predictions were compared to the actual prices. The results were less than stellar, with heavy hits to accuray in all models. The Decision Tree Classifier had an accuracy of 61% when predicting future prices, while the Random Forest had an accuracy of 58%. The RMSE score of the Decision Tree Regressor was around 0.2.

#### **VIII. Conclusion**

Real Estate pricing since the start of the pandemic has been extremely volatile. With the initial steep decline, there was also an extreme spike in the pricing over the last couple of years in housing pricing. This study focused on the two cities of Philadelphia and New York. The datasets used here were filtered to these two cities for analysis. The goal was to use continuous features along with the categorical features in our datasets to try to predict the future pricing of housing in these two major cities.

The first step was acquiring the datasets that were needed for this analysis. The realtor data was originally from [www.realtor.com](http://www.realtor.com), but we accessed it through Kaggle. It details the pricing of real estate across the United States, but we wanted to focus solely on Pennsylvania and New York. Therefore, a mask had to be created for the data to be filtered to just those two states.

The second dataset comes from a Loan-Level Public-Use Database. It shows information about all mortgage acquisitions of single-family and multi-family homes over the last few years we are using for analysis. It also provides details of loans and demographic

information about the loan holders.

The third dataset is a Zillow Home Value Index that provides us with the median price of homes in the United States.

The final dataset we used was rolling sales data from properties in NYC. The Department of Finance is where this data was found and it includes property information as well.

The next step was exploring the data that we had. We were able to identify the missing values and separate the categorical and continuous features in our datasets. We also had to identify imbalanced features.

After processing and exploring the data, we created several machine learning models trained on the preprocessed data, in order to see if they could accurately predict future prices. We did this by training them on data between the time period of 2010 to 2020, and then testing them on data from 2020 onwards. Of the models we created, Random Forest had done the best, followed by the Decision Tree Classifier and the Decision Tree Regressor.

Taking the three best models forward into the next phase, we had the models predict “future data”, i.e. data from after 2020, to see how well they did. All models had heavy decreases in accuracy.

It is clear that more work needs to be done on this project to have a truly accurate real estate price predictor. Some more turning on the models is required, and indeed it might be the case that entirely new models, better suited to the nature of the problem, might need to be created. Regardless, the work done in this project has laid a firm foundation for future groundwork.

#### **IX. References**

Santarelli, M. (2022, December 7). *Is it a good time to buy a house or should I wait until 2023-2024*. Norada Real Estate Investments. Retrieved December 8, 2022, from <https://www.noradarealestate.com/blog/is-it-a-good-time-to-buy-a-house/>

Santarelli, M. (2022, November 26).Housing market trends 2022: Will house prices fall ?. Norada Real Estate Investments. Retrieved December 8, 2022, from <https://www.noradarealestate.com/blog/housing-market-trends/>

Rothstein, R (2022, November 29).Housing Market predictions for 2023: Will home prices (Finally) Fall ? Forbes Advisor. Retrieved December 7, 2022, from <https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/>

Tyagi, N. (2021, March 4). *Introduction to time series analysis in Machine Learning*. Analytics Steps. Retrieved March 18, 2023, from <https://www.analyticssteps.com/blogs/introduction-time-series-analysis-time-series-forecasting-machine-learning-methods-models>

Ali, R. (2020, September 22). *Predicting your business's future*. Oracle NetSuite. Retrieved March 18, 2023, from https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml

Jim Brown (2021, May6) Why using K-means clustering? Retrieved <https://knowledgeburrow.com/why-using-k-means-clustering/>