AFM 346 Midterm Report - Austin Housing Market Analysis

Group #1 – Callum Stevenson, Jake Vanderweyst, Joel Palmer, Nathan Farquharson, Tim Sankey

Introduction

As members of Nate's Real Estate Advisory Firm, we were tasked with exploring and identifying predictive relationships within the austinHousingData dataset. This dataset had 15,171 observations with 47 columns, each observation represented one unique house in the Austin, Texas area. The columns are various identification and feature factors ranging from location to house information to recent pricing information. Due to the size of this dataset, we will be subsetting to make it more manageable while also better meeting the needs/wants of our high-profile client, Kee. After preliminary discussions with him, we derived that he is looking for a list of feasible investment properties that he could rent to family tenants and hopefully sell for a profit in 5 years. This discussion allowed us to determine 5 variables to be the best way to focus our analysis:

- **PropertyTaxRate = 1.98:** To make the listings more comparable we decided to use the most common and lowest tax rate. By doing so this mostly eliminated the small outside cities around Austin as potential destinations and allowed our analysis to be better focused. In addition, lower tax rates typically mean better living area.
- **HasView = 1:** A house with a view is a distinguishable feature that can increase the price compared to an identical home without a view. This is also a feature that typically cannot be added later and is an effective identifier of feasible homes.
- **NumOfBedrooms >=2:** Typical family homes have a minimum of 2 bedrooms, so we wanted to eliminate any 1-bedroom options as it does not fit the scope of our analysis
- **LatestPrice <= 1,000,000:** This allowed us to filter any high outlier priced homes we identified in the dataset that would be unrealistic options for our client
- **AvgSchoolRating >= 4:** This relates to finding a property that is attractive to families as our ideal tenant typically has kids in school and would want a strong educational area.

With these filters, the dataset decreased from 15,171 to 2,084 homes that fit the general criteria our client is looking for. Notice that there are 2 NA's in the 'description' variable, this will have no impact on our analysis so we left these observations in.

To show our findings, we focused on three main insights. Firstly, using KNN Neighbours, we discovered that house location significantly impacts pricing, with properties in the same neighbourhood tending to be similarly priced. Secondly, the Random Forest analysis allowed us to identify the key features influencing house value. Lastly, our analysis of house value appreciation over time showed a general upward trend, highlighting the feasibility of real estate investments in the Austin area. These insights, combined with our detailed exploratory data analysis, regression models, and clustering techniques, contributed to a final recommendation to identify and invest in undervalued properties that have the potential for significant appreciation. The first insight that we will cover will explore how house location affects pricing using KNN Neighbours. The second insight will use Random Forest regressions to identify key features influencing house value. Finally, the third insight will analyze the trend of house value appreciation over time, highlighting the benefits of long-term real estate investments. To wrap up the report, we will

combine the findings from our insights to prepare a list of 10 investable properties for our valued client, Kee.

Data Exploration (EDA)

EDA Section 1

Price Analysis

First, we analyzed house prices and observed many outliers in the boxplot visualization of the raw dataset (exhibit 1). The histogram (exhibit 2) supported this observation and highlighted a significant positive skew with a high concentration of observations at the lower end of the distribution. The concentrated area in the histogram suggests a potential market grouping that warrants further investigation. Further, the outliers in the long tail make the data challenging to interpret. Thus, a price limit of $1 million, close to the $974,000 upper limit, was imposed to eliminate outliers. This price limit created a more evenly spread and observable distribution (exhibit 3) allowing us to work with more comparable data, observe clearer trends, and draw meaningful insights.

PricePerSqFt and DistanceFromMax

A common metric in real estate is price per square foot. In this dataset, we have the latestPrice and livingAreaSqFt for each property. To create the variable pricePerSqFt, we need to divide latestPrice by livingAreaSqFt for each house and assign the result to a new column.

With the pricePerSqFt calculated, we can plot this using a scatter plot, where the X-axis is longitude and the Y-axis is latitude, and the color of each point is determined by the value of pricePerSqFt (exhibit 4).

From analyzing this plot, it becomes evident that there is an area of Austin with higher values for pricePerSqFt than most other areas. As you move away from this area, the pricePerSqFt tends to decrease. To meet our objective of mathematically proving relationships between variables, we should perform linear regression on this variable.

We will approach this by identifying the point with the highest pricePerSqFt value and then assigning a value called distanceFromMax to every other point, representing the Euclidean distance from the point with the highest pricePerSqFt value (exhibit 5). Once this value is assigned to a new column in the dataset, we can perform linear regression with distanceFromMax as the independent variable (X) and pricePerSqFt as the dependent variable (Y). The model gives an R-squared value of 0.39, and once plotted, the relationship becomes very clear. Form this, we gather the conclusion that as you move away from the most expensive house in terms of pricePerSqFt, the value of your house in terms of pricePerSqFt declines (exhibit 6).

EDA Section 2

Number of Price Changes

Our analysis found a low correlation between the number of price changes and the latest price. A linear regression model had an R-squared value of 0.0002 and an RMSE of 200,928, indicating it explained very little of the price changes with large errors. The polynomial regression (exhibit 7) improved slightly, with

an R-squared of 0.0020 and an RMSE of 200,747 but remained inaccurate. Therefore, we did not proceed with using the number of price changes as a predictive variable.

Age of House

Similarly, our analysis of the age of houses also demonstrated a low correlation with price. A linear regression model yielded an R-squared value of 0.0030 and an RMSE of 200,647, indicating low predictive power of price with large errors. The polynomial regression model (exhibit 8) for house age showed some improvement, with an R-squared of 0.0294 and an RMSE of 197,976. Despite the improved predictive power in the age of house polynomial model we did not continue the analysis of house age as a predictive variable.

EDA Section 3

General Analysis

To get a better understanding of how each variable works with latestPrice we wanted to find the R-square value of each through a simple linear regression. livingAreaSqFt was the highest value at 0.295 showing that 29.5% of variation of latest pricing can be explained to livingAreaSqFt. In addition, we wanted to see the results of all possible pairs of multiple linear regression and found that livingAreaSqFt and numOfHighSchools with an R-squared value of 0.337. This makes sense as the size of a home is typically a strong factor in the homes pricing and the data supported this.

School Variables

We wanted to explore the subset of school variables offered in the dataset to see if there were any strong predictor variables to be used in our insight models (Exhibit 9). By creating a histogram of all the variables, we found that all the numOf... variables were typically 1 or 0 with middle schools and high schools reaching a maximum of 2. avgSchoolDistance has a right skewed distribution with many of the observations being <2. AvgSchoolRating has a more even distribution with most observations being around 7. avgSchoolSize and medianStudentsPerTeacher had similar left skewed distributions.

After getting a general picture of each variable, we calculated the correlation and R-squared between them and latestPrice. AvgSchoolRating had the highest values in both categories at 0.375 and 0.14 respectively, so we chose this as our variable of focus for this section.

To further enhance our visual of how avgSchoolRating is laid out we mapped it with longitude on the X axis and latitude on the Y axis with a blue-red colour scale (Exhibit 10). Noticing that on the west side of Austin there was primarily read colouring (high school rating) whereas the east was primarily blue signaling weak school ratings. This could be useful in determining preferred locations to purchase an investment home tailored to families.

For the exploratory analysis of avgSchoolRating we attempted to make predictive models from all the versions in class to see which would be most effective. As previously mentioned, the simple linear regression had an R-square of 0.14 and a MAPE of 0.49 (Exhibit 11). Note that because the variable takes on a value between 1-10, a MAPE value of 0.49 is moderate. Next we created a multiple linear regression with livingAreaSqFt which earlier was identified as our most effective single variable in the set, this combination resulted in an R-square of 0.306 and a MAPE of 0.5084 (Exhibit 12). It is interesting to note

that while the R-square increased meaning an improved model, the MAPE increased which shows a higher average error rate. In this model we also tested the VIF which was 1.378 meaning there is limited multicollinearity problems between variables. After this we created a KNN regression which produced a R-square of 0.19 and a MAPE of 0.476 (Exhibit 13). The takeaway from these regressions is that avgSchoolRating does a moderate job in predicting latestPrice however it certainly is not perfect which is expected as more than school rating determines the price of a house, however it is a factor.

To continue understanding avgSchoolRatings use we moved to classification methods beginning with a KNN classification. The model predicts if a house should have a low (0-300k), medium (300k-700k) or higher (700k+) price and compares it to the actual listed price (Exhibit 14). The accuracy of the model was 68% meaning houses are properly classified just over 2/3 of the time. From this we created a random forest classification model, which is a model we have not gone over in class but will be seen in this report. What this model does is derive a value based on the average result of a multitude of decision trees, or the "forest." This classification is measuring whether a house should have a low (0-478k) or high (478k+) price and compare it to its actual value. The model produced an accuracy of 99%, only misclassifying 30 of 2084 homes (Exhibit 15).

Overall, we can see that avgSchoolRating was much more effective as a classification variable, although this is expected as it is more likely to be accurate within categories rather than trying to derive exact values.

EDA Section 4

Exploring Living and Lot size

The next section of the EDA focused on critical variables, such as living area and lot area. To explore the intricacies between living and lot area, we created new variables, including the total area, the living area ratio, and the total area price per sqft. These newly created variables, albeit some more prominent outliers, were all evenly distributed, with the mean of each being 18,999 SqFt, 0.20, $45.93, respectively (Exhibit 16). To better understand the strength of relationships between the variables and our target variable of the latest price, a correlation matrix was created using living area, total area, area ratio, and total area price per sq ft against the latest price (Exhibit 17). The area ratio showed a 0 correlation to the latest price, and the total area sqft also had a low score of 0.03. However, the total area price per square foot had a much higher result with 0.33, and the highest correlation between the variables was from the living area sqft with 0.54, suggesting that size can be a major determinant of price. A geographical analysis was performed on the difference between the living area and price categories to further explore the living area sqft (Exhibit 18) The geographical visualization highlights the price distribution by region but also shows a similar distribution of home prices and home sizes throughout regions and price categories identified in the analysis. The expectation that areas and neighbourhoods are sized and priced similarly aided in creating the predictive models.

Connection to latestPrice

Although various predictive models were employed to analyze the data, including the K-Nearest Neighbours and a Linear Regression model, each model served a different purpose. While the linear regression showed the direct relationship between living area size and the latest price, the model was not

the best fit, with a r-squared value of 0.29, a lower-moderate score. However, the KNN distance metrics were modeled next, with the Euclidean distance metric showing the best result (Exhibit 19). The Euclidean distance measures the 'closeness' of instances considering the weighted average of the prices of the nearest neighbor properties. The model resulted in a r-squared value of 0.45, a moderate score, and expected as the geographical analysis suggested a distribution of price correlated with region and living area size. The Euclidean model was used to predict the price of a home using the single variable living area size. The input was 2500 sqft, resulting in a home price of $485 thousand, and cross-validation results with an r-squared of 0.17 and a mean average error of $141 thousand. The next model was approached differently, separating the target variable into high and low-price categories using logistical regression. The model used lot size, living area, total area, area ratio, and total area price per sq ft to help predict the price category. The model performed well with a precision and f1-score of 0.87, helping identify which properties fall into either category. The KKN distance model was also implanted and tested with the Manhattan metric, which performed the best with an f1-score of 0.82 (Exhibit 20). The Manhattan distance can be used to compare properties within the same neighborhood by measuring distances along actual city streets rather than through longitude and latitude lines. However, the logistical regression metrics performed better and were used to predict the price of a home based on a lot area of 7000 sq ft and a living area of 3000 sq ft for an expected price of $575 thousand with cross-validation results of r-squared equal to 0.18 and just under $140 thousand for the mean average error. The EDA provides the foundation for creating variables and models that will go on to generate insights and recommendations, enabling the client to capitalize on quality investment opportunities.

EDA Section 5

Property Features

This section reviews another group of variables to identify specific variables of interest and potential models to create new insights. In this section, we reviewed property features including the number of photos available for the property, accessibility features, appliances included, parking features, patio and porch features, security features, waterfront features, window features, bathrooms, bedrooms, and stories.

Correlation Analysis

The correlation analysis shows that the number of bathrooms stands out as the most impactful, with a strong positive correlation of 0.4437. This suggests that properties with more bathrooms generally result in higher prices. Bedrooms also show a significant positive correlation of 0.3118 with property prices. This suggests that an increase in the number of bedrooms is associated with an increase in property value. Other factors such as the number of stories (0.2606) and the number of photos (0.1599) display moderate positive correlations, indicating that multi-story properties and well-photographed listings tend to be valued higher.

However, some variables display weaker correlations with property prices. Features like windows, patios/porches, security, and parking have minor positive impacts, with correlation coefficients ranging from 0.0811 to 0.1091. Accessibility features, appliances, and waterfront features show negligible correlations, indicating they do not significantly influence property prices in the dataset. By looking into the negligible correlations, we found accessibility and waterfront features are poorly distributed in the

dataset, with only 19 properties having accessibility features and 7 having waterfront features. This scarcity makes them unreliable predictors for the model and not of interest going forward.

Variables of interest

- Number of Bathrooms: There is a strong positive correlation between the number of bathrooms and property prices, with an R^2 value of 21.25% (Exhibit 21). Properties with more bathrooms generally result in higher prices. These properties are mainly located in the northwest part of Austin (Exhibit 22).
- Number of Bedrooms: This variable also shows a positive correlation with property prices, though less strong, with an R^2 value of 11.98% (Exhibit 23). Properties with more bedrooms tend to be more expensive and are more commonly found in the western part of Austin (Exhibit 24).
- Number of Stories: The number of stories in a property has a slight positive correlation with property prices, with an R^2 value of 6.30% (Exhibit 25). Multi-story properties are more commonly found in central Austin (Exhibit 26).

Regression And Classification Model

Taking these 3 variables of interest in this section we created a regression model that had an accuracy of 21.86% at predicting property price (Exhibit 27). This model has low predictive power and limitations. We also made a KNN classification model predicting whether the property price is above or below the median (Exhibit 28). This model gave us an accuracy of 61%.

Insights

Insight 1 – Pricing Based on Neighbours

Our goal was to determine the impact of neighboring sales on home prices using a K-Nearest Neighbors (KNN) regression and classification model. Many neighborhoods have houses with similar features (size, bedrooms, bathrooms) and shared location characteristics (school rating, tax rate). However, not all houses in a neighborhood are exact matches, making precise predictions difficult based on available sales data.

We implemented a multi-variable KNN regression using longitude and latitude to approximate location, with a k-value of 3 and distance calculated using the Manhattan method (exhibit 31). This model predicted prices based on the nearest three house sales in the dataset and proved to be a strong predictor, with an R-squared value of 0.7043 and an RMSE of 109,227. It proved that neighbouring sales can effectively capture neighborhood price trends.

Next, we used the same predictors, and parameters, to classify houses into three price categories: budget, mid-range, and premium. We applied min-max standardization to express price as a percentage of the range (0 to 1), with the categories representing the bottom 25%, middle 50%, and upper 25% of the price range. This model (exhibit 32) achieved 86.7% accuracy, with a mean cross-validation accuracy of 66.2%. The classification highlighted how grouping homes into price categories can improve predictive power and is better tailored to the theory that neighbours have similar, but not exact, characteristics and prices.

For real estate investors, considering neighboring sales in pricing models is crucial given the influence of neighbourhood characteristics on home values. By leveraging KNN regression and classification,

investors can more accurately assess property values, identify investment opportunities, and anticipate market trends within neighborhoods.

Insight 2 – Create a Pricing Model

Our goal for this insight is to create a pricing model to predict exact price values using a random forest regression to narrow our list of 2084 homes to the 100 best valued homes for our recommendation. We define the "best value" as the homes with the highest percentage increase between their actual latestPrice and their predicted latestPrice that our model derives. We decided to use a random forest model as with the number of variables incorporated into predicting a price, we believe it captures their features the most accurately.

To narrow the focus of the model we created another sample dataset with filters on 6 key variables that were deemed essential to a well-designed family home who were also identified as useful predictors of latestPrice. The conditions for all homes to meet were >=5 avgSchoolRating, >=1500 livingAreaSqFt, >=3 numOfBedrooms, >=2 numOfBathrooms, >=1 hasGarage, >=1 hasCooling which were 1037 of our 2084 observations. For homes fitting the model we were able to produce an accuracy rating of 90.35% and a predicted price of $346,776. The predicted pricing model was then applied to all the fitting observations creating the new variable 'predicted_price.'

With this model we noticed there were a few outliers in the top 100 where the predicted value was unrealistically larger than their latest price. This was found using the new variable 'percentage_difference' and 'price_difference' where if the values were negative then the latestPrice was greater than the predicted_price value and vice versa. To combat the outlier issue, we decided to remove the observations using the IQR method. This removed unrealistic values, such as having a 3525% upside, from the highest valued home to a more realistic 40% difference. This filtering lowered our dataset from 1037 observation to 926. With the filtered dataset for outliers, we were able to create a new dataset by sorting the data in descending order based on 'percentage_difference' and taking the first 100 rows. This dataset will be used to find the top 10 optimal houses for our client to take the next steps in the purchasing process.

Insight 3 – Calculating the Appreciation Rate of the Austin Housing Market

The objective of this analysis is to determine the housing market appreciation rate in the Austin area to assess its invest ability. The main challenge encountered was the lack of historical pricing for each house, as the dataset only contains one price per house, labeled latestPrice. To address this, we clustered the dataset based on groups of very similar houses and selected the largest cluster with a range of latest_saleyear values. This approach assumes that houses within the same cluster appreciate at similar rates, allowing us to calculate the average growth rate using the different latestPrice values for the different latest_saleyear values.

To begin the analysis, we created a new data frame called df_similar, filtering down to a subset of houses that are nearly identical except for longitude and latitude. This narrowed the list down to 450 houses. We then used KMeans clustering to identify 10 clusters of houses similar in latitude and longitude. The analysis focused on the largest geographical cluster from this process, referred to as DFLG.

With the filtered and clustered data based on geographical location, we were confident that the houses in this subset would appreciate at the same rate. The next step was to apply the compound annual growth rate (CAGR) formula to each house in the dataset and then average the values.

$$\text{CAGR (\%)} = \left(\frac{\text{Ending Value}}{\text{Beginning Value}}\right)^{(1 \div t)} - 1$$

Source : https://www.wallstreetprep.com/knowledge/cagr-compound-annual-growth-rate/

To determine the ending value, we used the latestPrice for each house in the DFLG. For the beginning value, we used the average latestPrice from all the houses in the chosen cluster, assigned to the variable average_latest_price. This was feasible due to the dataset's similarity and the absence of historical pricing. Using the average latestPrice allows us to focus on relative appreciation. The t value in the CAGR formula was represented by a new variable, years_between_sales, calculated as the difference between the latest_saleyear for each house and the average_latest_saleyear, which is the mean latest_saleyear from the DFLG. With these variables, we created the annual_appreciation column and applied the CAGR formula to all the houses.

$$\left(\left(\frac{\text{DFLG['latestPrice']}}{\text{average\_latest\_price}}\right)^{\left(\frac{1}{\text{DFLG['years\_between\_sales']}}\right) - 1}\right) \times 100$$

After applying this formula to each house, we took the mean value, resulting in a compound annual growth rate (CAGR) of approximately 8%-10% after testing multiple clusters.

To check the accuracy of our calculation, we compared our results with online sources, which predicted similar appreciation rates for the Austin housing market.

With the calculated CAGR for the Austin housing market, we can now combine this with a list of 100 undervalued houses to create a comprehensive list for our valued client.

Recommendation

As previously mentioned, our recommendation will be using the predictive pricing model from insight 2 along with the appreciation information from insight 3 to create a top 10 list to present to our client, Kee. To do this we took the list of our 100 houses and identified the 10 most recent sale dates. Reason being if we were to look at a house from 2015 for example and appreciate it 5 years, we would still be looking at past pricing which would not be useful to our client today. With this we found 10 homes with the latest sale year in 2020. Using our predicted values, we appreciated the dollar amount by 5 years using the 7.99% CAGR from insight 3 to project the expected return Kee can expect from this investment based on capital gains alone. We recommend for Kee to continue renting his investment properties to families

throughout its investment life to generate an even greater return and offset any expenses incurred. The next steps we recommend for Kee is to contact the representatives of these 10 homes and make a choice of which he would like to invest into.

The key findings from our three compelling insights were:

1. Neighboring houses work as strong classification predictors of price for other homes due to comparable size, features, and location characteristics.
2. Identifying specific features within a home can be an effective way to obtain an accurate dollar value for the home.
3. When choosing real estate investments, it is important to consider macro economical factors such as the CAGR of the housing market.

All our insights related back to our role of being a real estate investment advisor and allowed us to go from a list of 15,000+ homes to 10 strong investment options.

Outside of simply taking our models values, we were able to derive various strategic insights that Kee should consider looking for this investment home and any in the future. He should analyze all variables before deciding which to base a decision on as each have different levels of prediction for the price. We also believe it is important to be specific in what you are searching for, finding an “investment house in Austin” is difficult, finding an “investment home in Austin meeting conditions X,Y,Z that can generate me an XX% return of X# of years” is much more defined and easier to identify value. We also recommend for Kee to continue focusing on family homes in good school areas as these were proven to generate strong capital gains return, appreciate well and are easy to rent during your holding period. With our list of 10 homes and strategic recommendations we believe our client is in an excellent position to succeed in the real estate investment space.

Exhibits:

Exhibit 1:

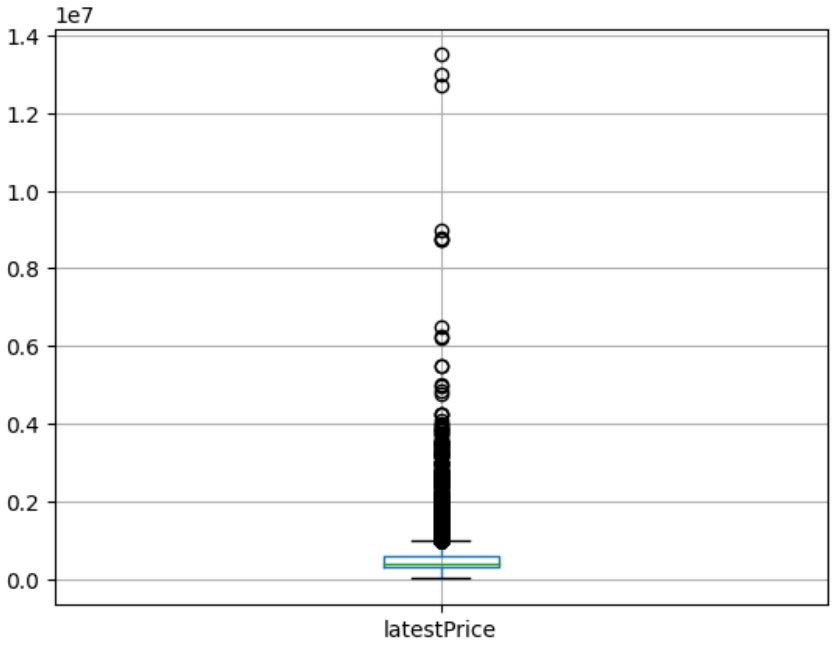


Exhibit 2:

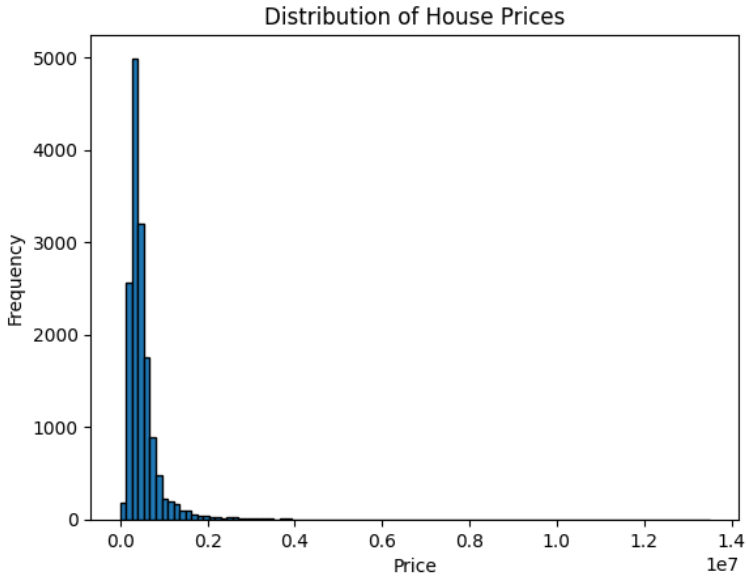


Exhibit 3:

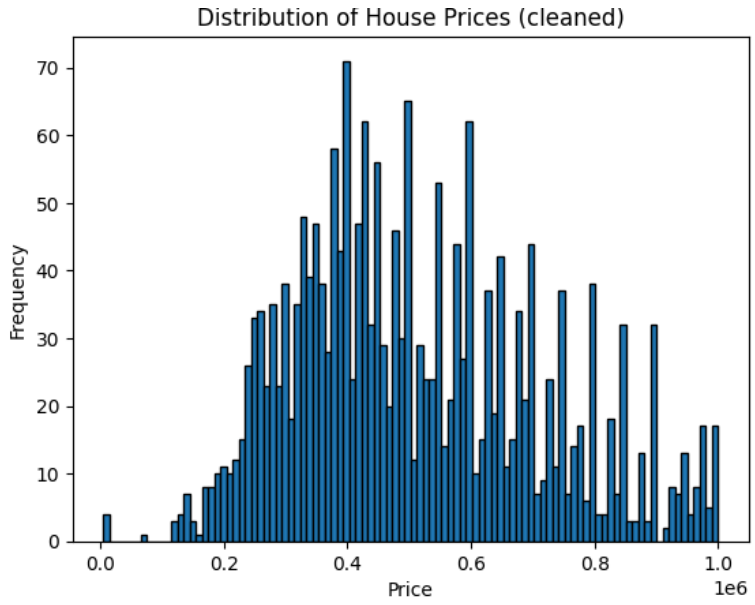


Exhibit 4:

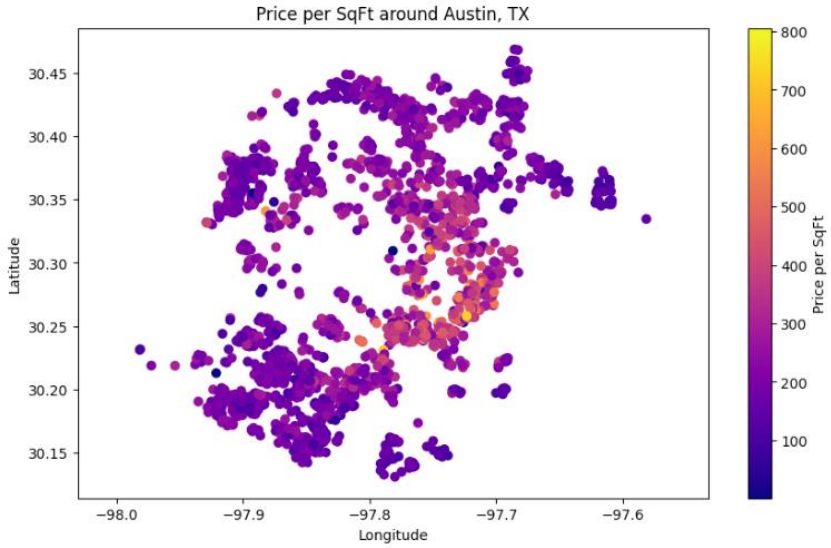


Exhibit 5:

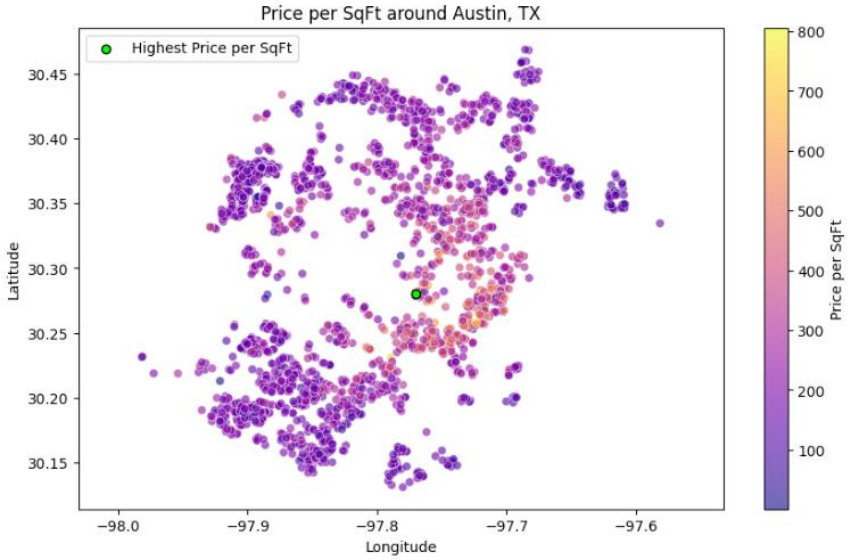


Exhibit 6:

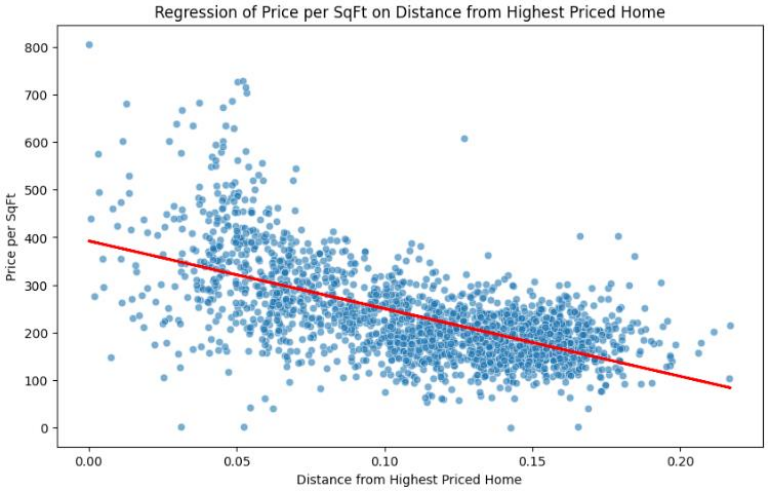


Exhibit 7:

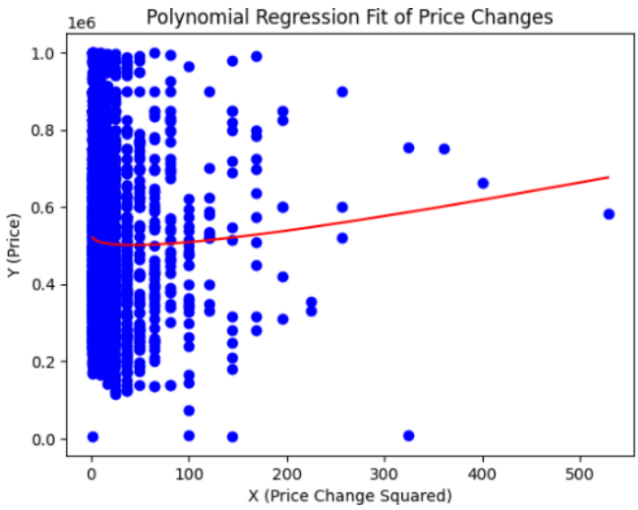


Exhibit 8:

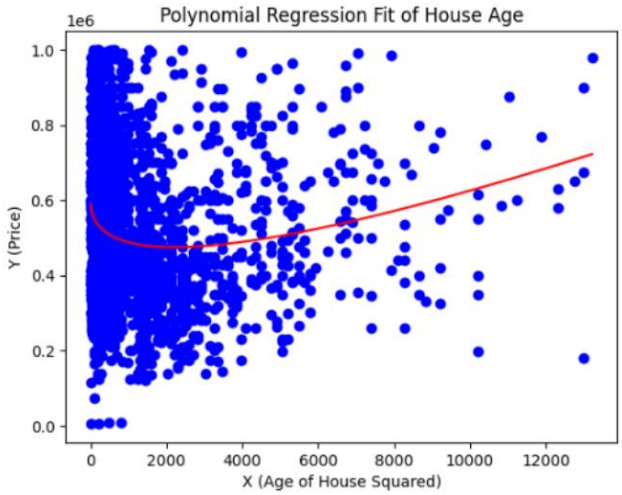


Exhibit 9

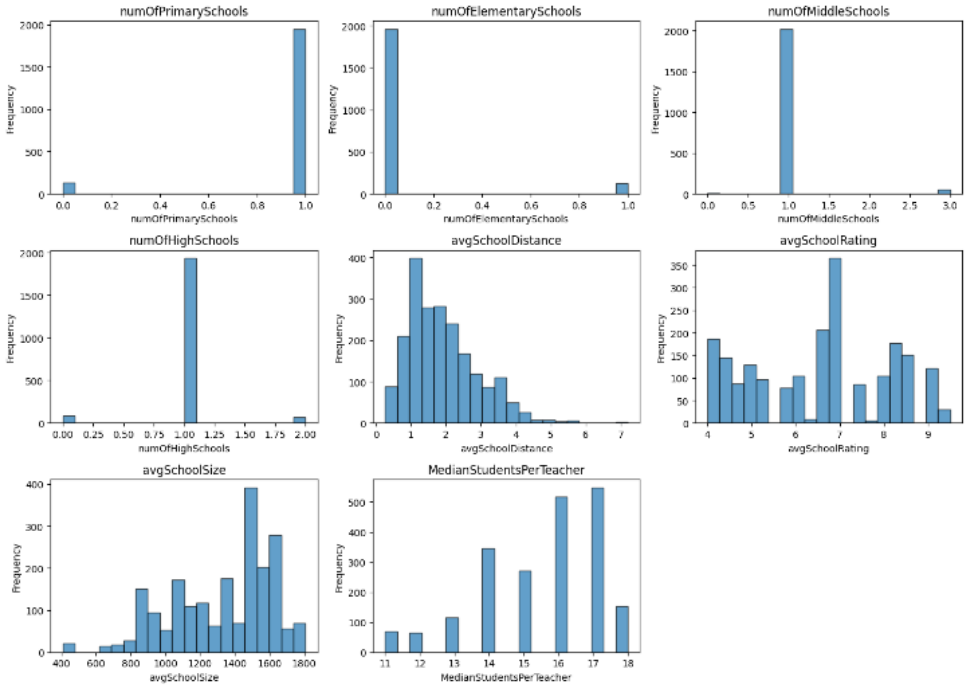


Exhibit 10

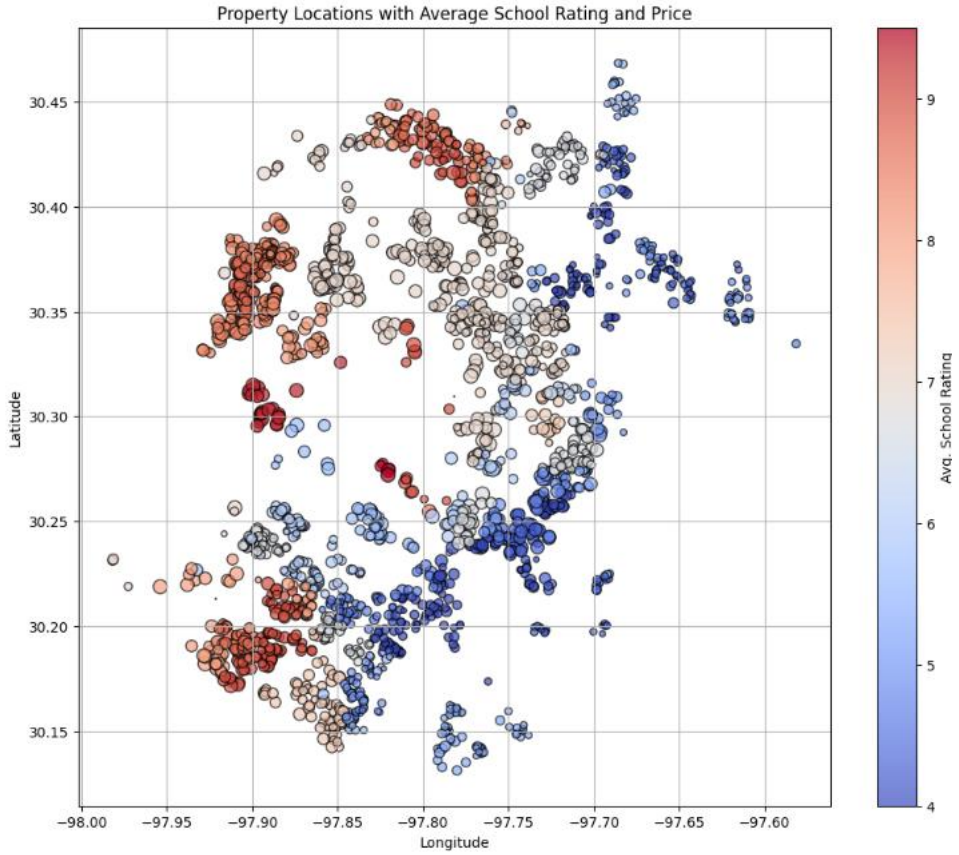


Exhibit 11

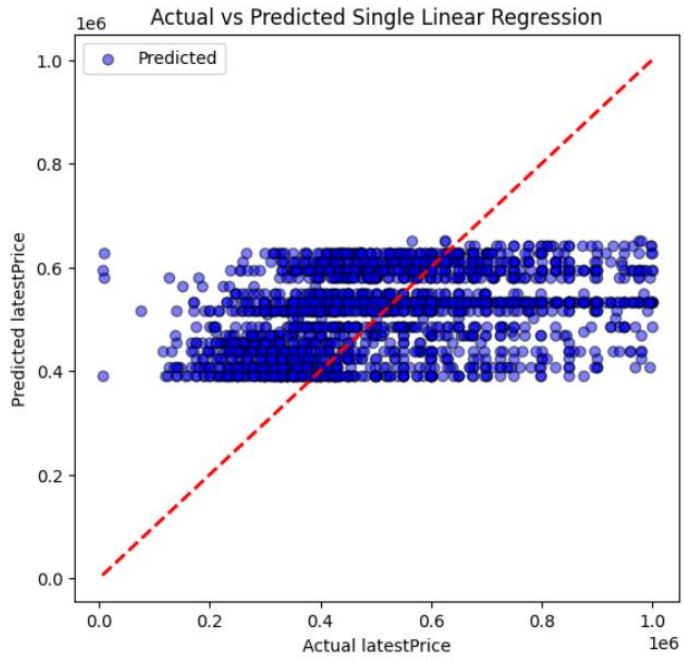


Exhibit 12

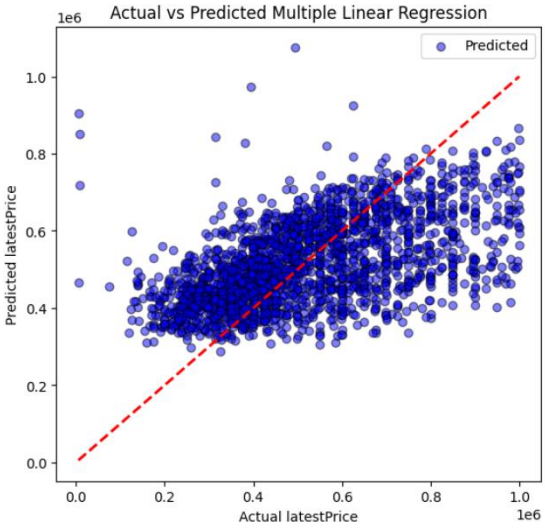


Exhibit 13

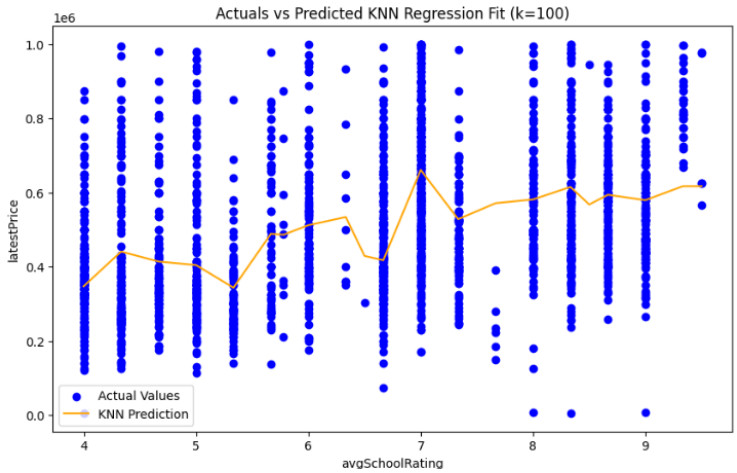


Exhibit 14

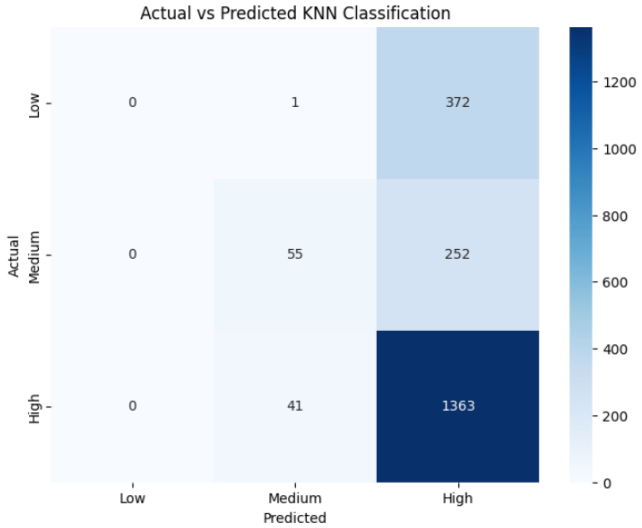


Exhibit 15

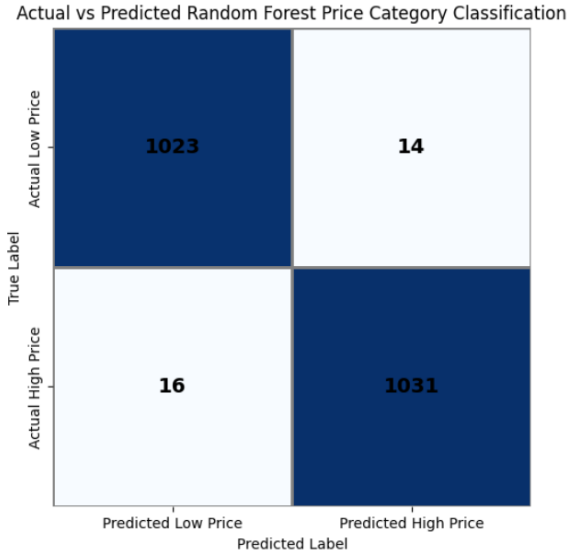


Exhibit 16

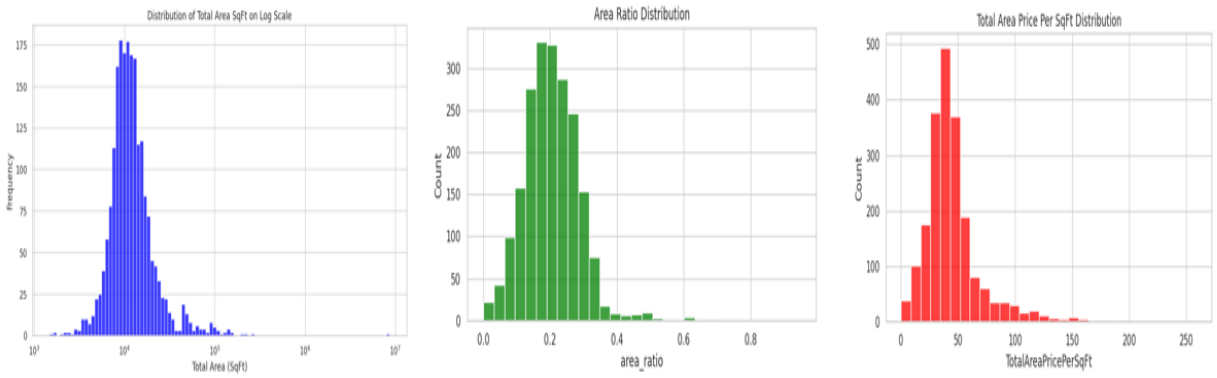


Exhibit 17

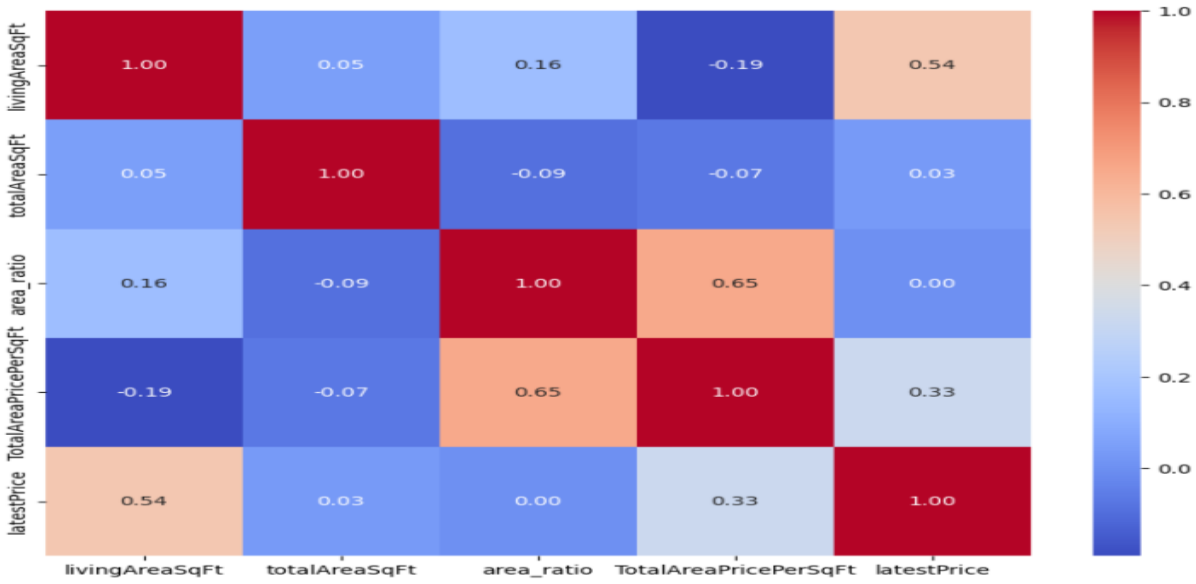


Exhibit 18

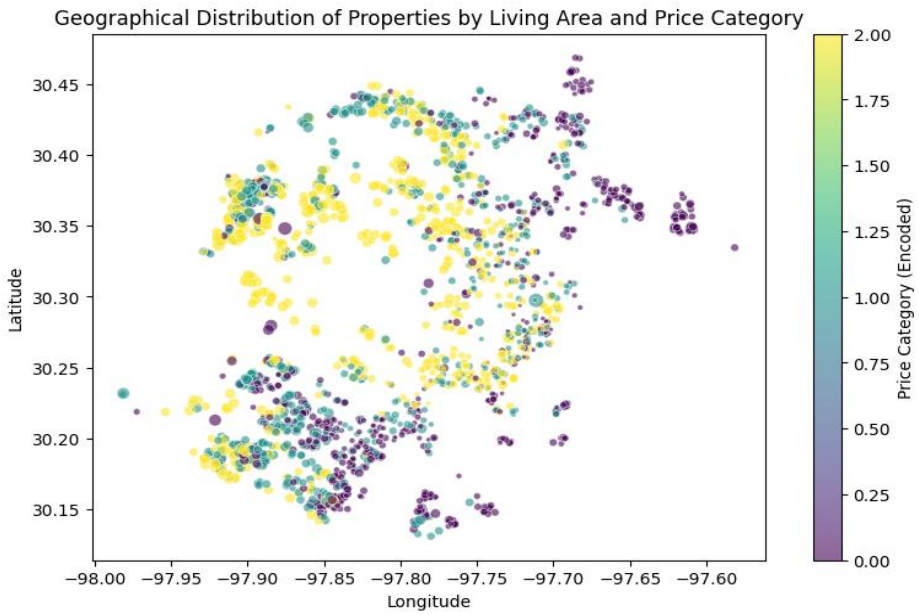


Exhibit 19

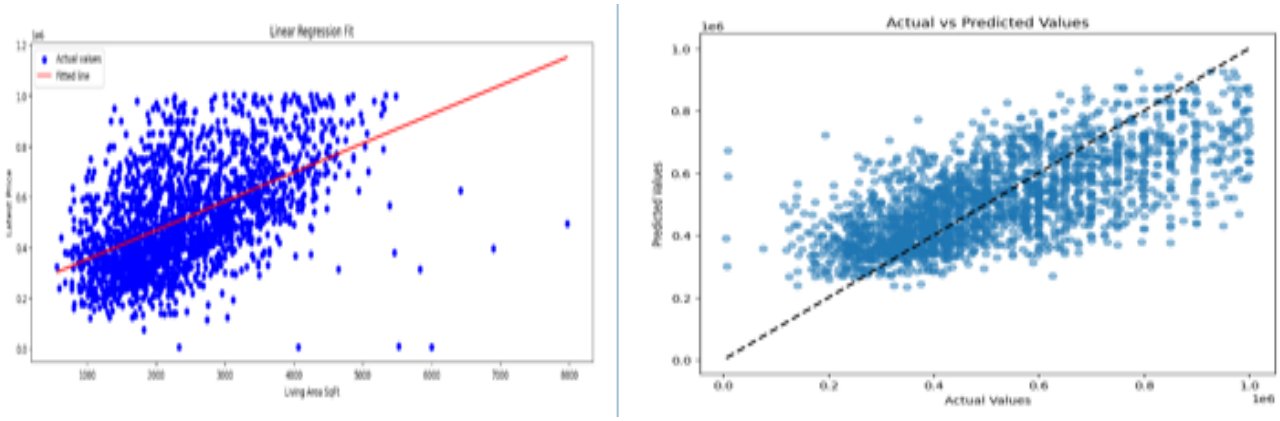
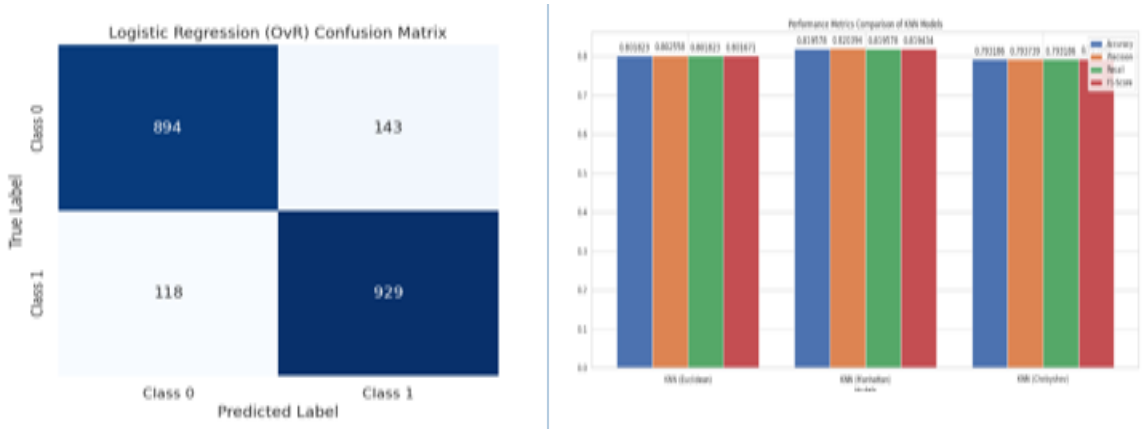


Exhibit 20



<u>Exhibit 21:</u>

Exhibit 22:

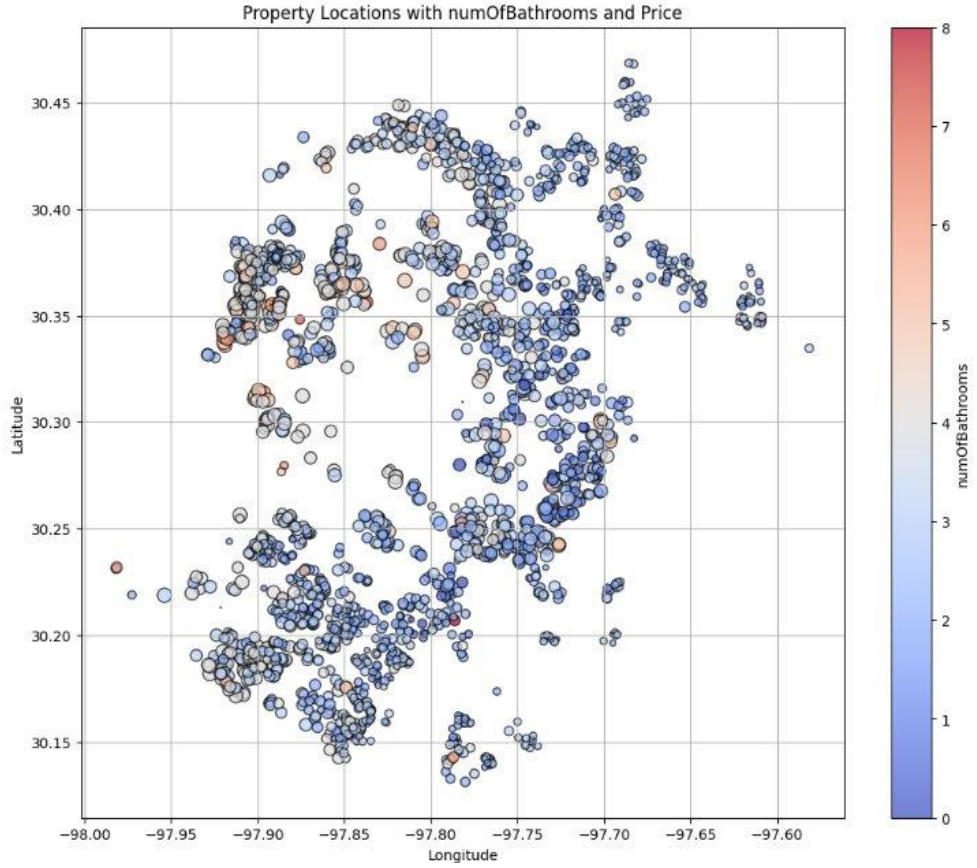


Exhibit 23:

Exhibit 24:

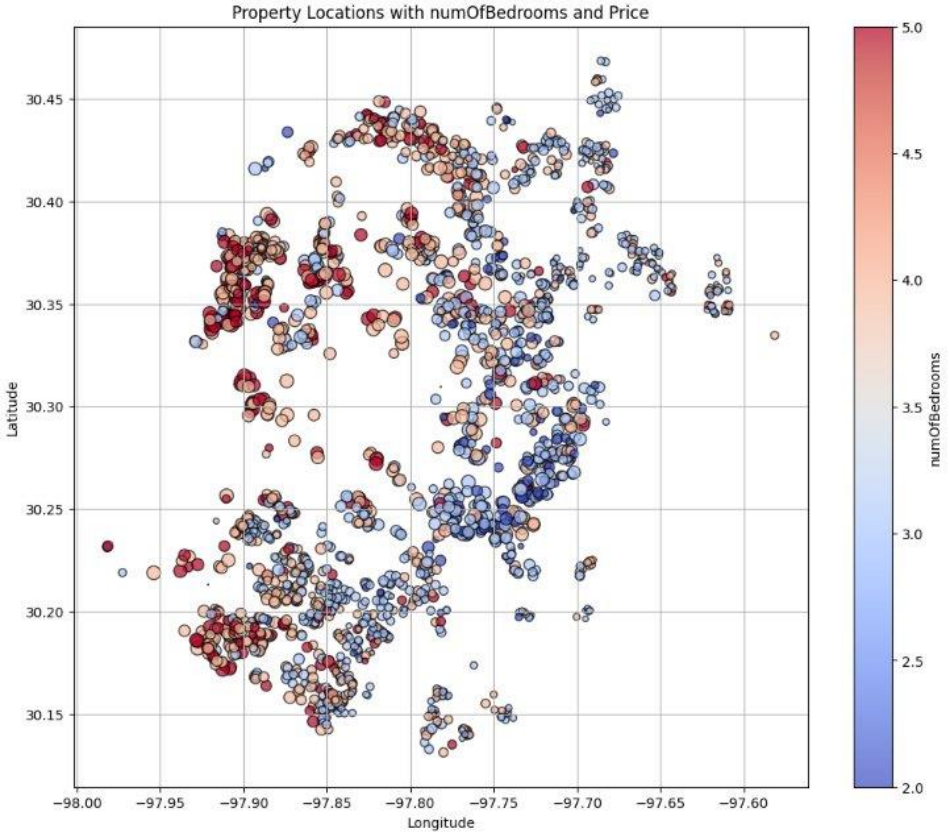


Exhibit 25:

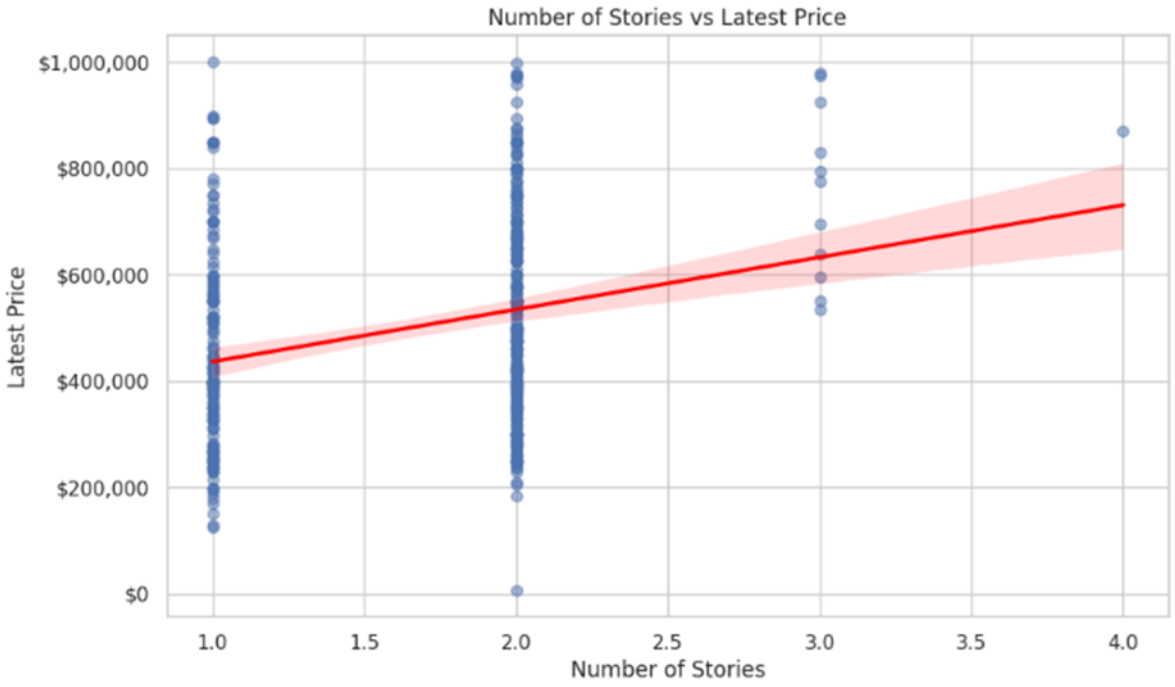


Exhibit 26:

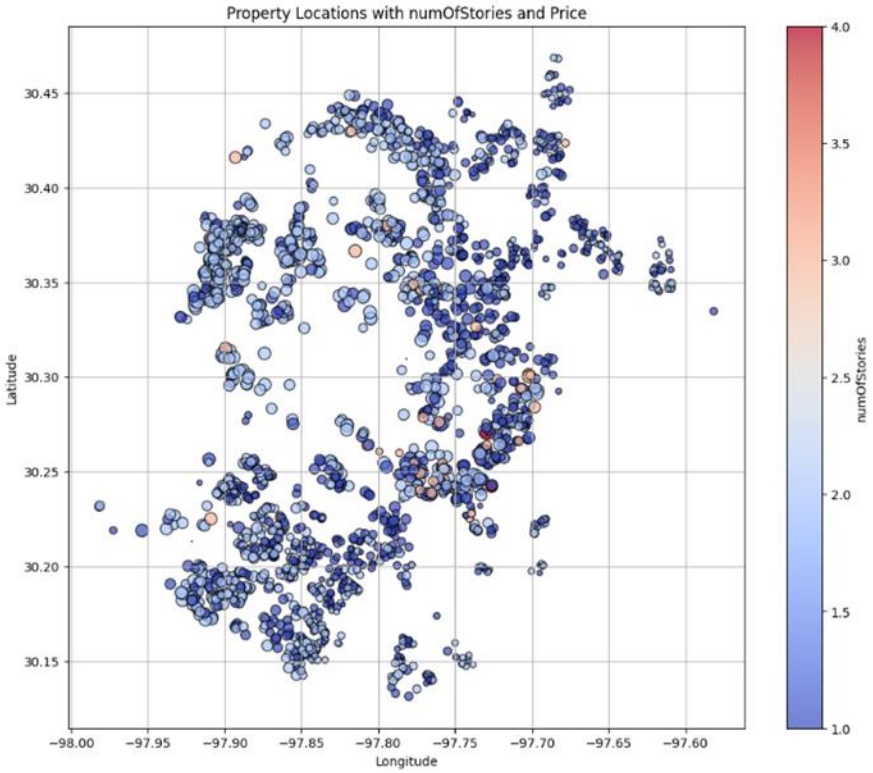


Exhibit 27:

Exhibit 28:

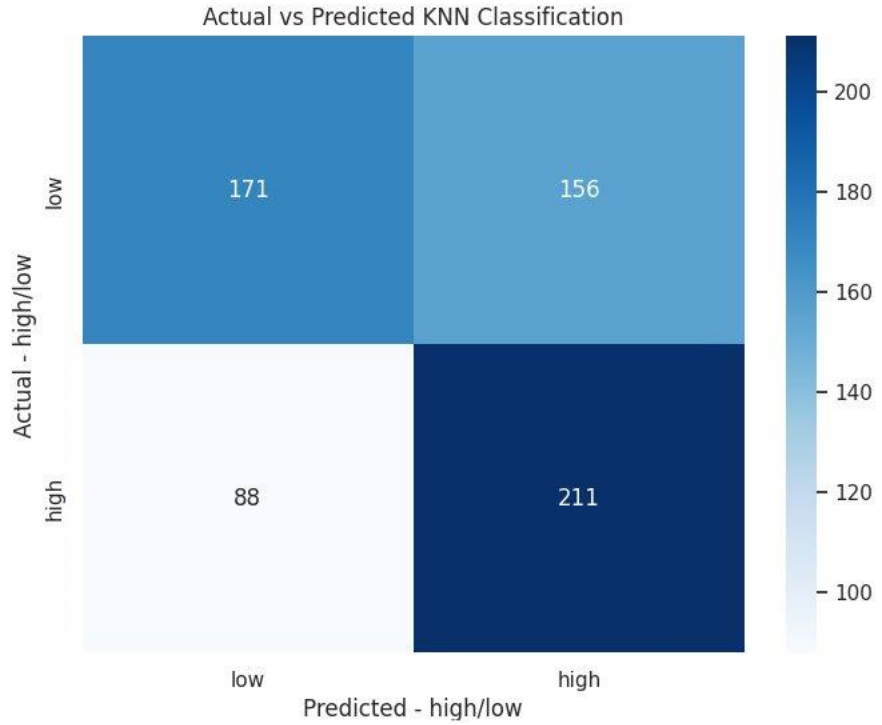


Exhibit 29:

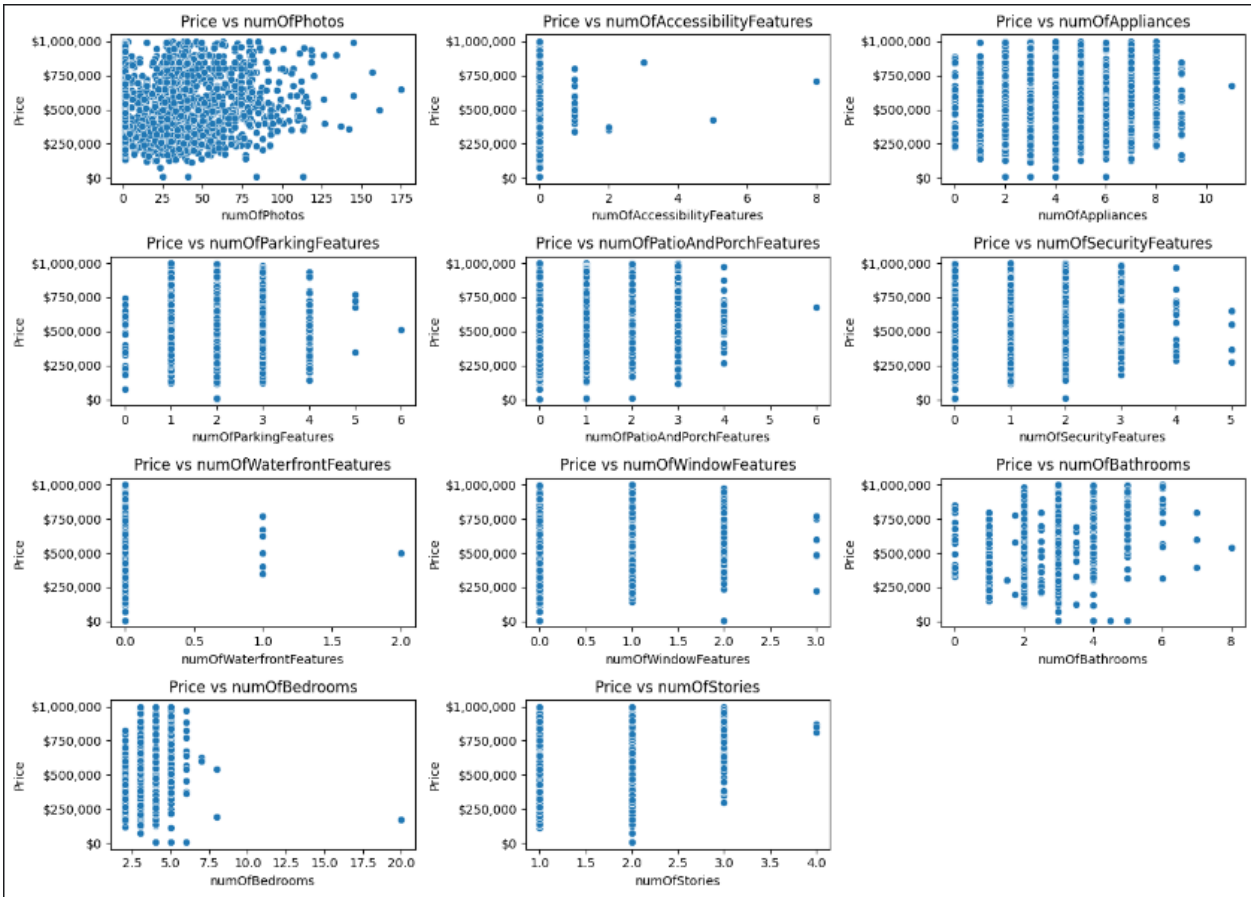

Price vs numOfPhotos
Price vs numOfAccessibilityFeatures
Price vs numOfAppliances
Price vs numOfParkingFeatures
Price vs numOfPatioAndPorchFeatures
Price vs numOfSecurityFeatures
Price vs numOfWaterfrontFeatures
Price vs numOfWindowFeatures
Price vs numOfBathrooms
Price vs numOfBedrooms
Price vs numOfStories
Price
numOfPhotos
numOfAccessibilityFeatures
numOfAppliances
numOfParkingFeatures
numOfPatioAndPorchFeatures
numOfSecurityFeatures
numOfWaterfrontFeatures
numOfWindowFeatures
numOfBathrooms
numOfBedrooms
numOfStories

Exhibit 30:


Average Latest Price by Number of Bathrooms
Average Latest Price
Number of Bathrooms
$0
$100,000
$200,000
$300,000
$400,000
$500,000
$600,000
$700,000
$800,000
0.0
1.0
1.5
1.75
2.0
2.5
3.0
3.5
4.0
4.5
5.0
6.0
7.0
8.0

Exhibit 31:

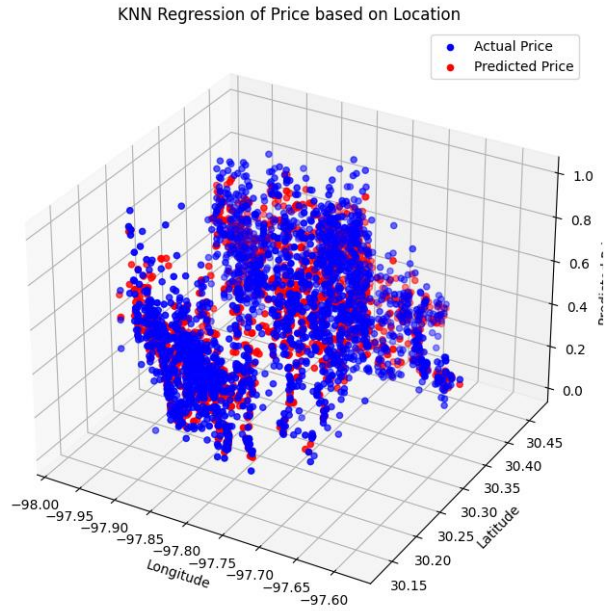


Exhibit 32: