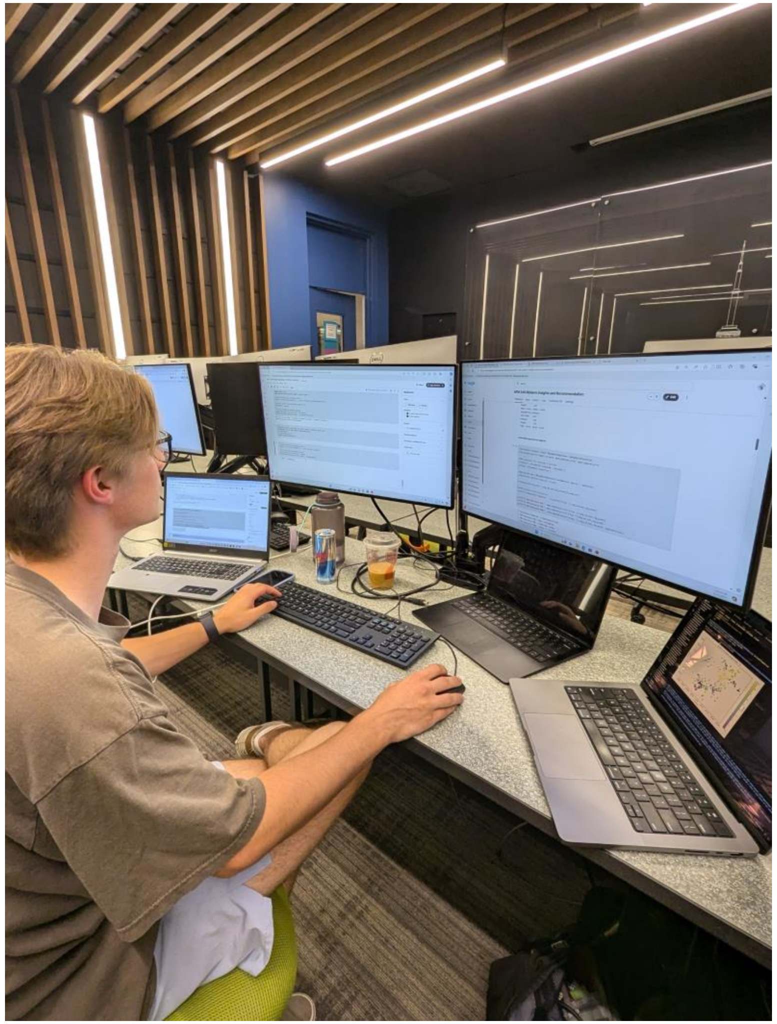
Nate’s Real Estate Advisory Firm

Austin Housing Market Analysis

Callum Stevenson, Jake Vanderweyst, Joel Palmer, Nathan Farquharson, Tim Sankey

Table of Contents

Key Insights and Recommendation



- **KNN Neighbours:** We found that house location significantly influences price, as properties within the same neighbourhood tended to be priced similarly.

- **Random Forest**: We used random forest regressions to predict prices revealed the key features that influence house value, highlighting key influencing factors.

- **House Value Appreciation**: We observed a general trend of houses appreciating over time within the data, emphasizing the value of time when making real estate investments.

Introducing our Client

- Name: Keehyung "Kee" Kim
- Recently hired at UT Austin as Data Science Professor
- Hired Nate's Real Estate Investment Advisory Group to find him a suitable investment property
- Wants to invest in a property that he can rent to families.
- Interested in flipping the property for profit in 5 years.

Quote from Kee

> I recently joined the University of Texas at Austin and am searching for a house in the city. I've hired Nate’s Group to find an investment property. I prefer family tenants. I want a property that can be flipped in a few years while generating monthly rental income. As a busy professor, I need a list of fewer than 100 properties that meet my financial and personal criteria.

- Keehyung Kim, 2024

Overview

Key Areas	Description
Objective	The primary goal is to develop predictive models to understand the factors influencing house prices in Austin, Texas, and to provide strategic insights and recommendations for real estate investment.
Data Analysis	We are analyzing house listings data, focusing on key variables like price, time, house characteristics, and location (longitude and latitude) to identify patterns and predictive relationships.
Model Development	This will be achieved through machine learning techniques such regression (linear, multi-variable, polynomial, logistic, and KNN), classification (binary, multi-class, multi-label), and random forest models to accurately predict house prices.
Insights and Implications	The analysis will help us understand how different factors affect house prices. We will gain knowledge of the data which can help investors make informed decisions and identify potential investment opportunities.
Strategic Recommendation	The insights from predictive models will be used to provide actionable recommendations for real estate investment strategies, helping maximize returns through undervalued properties and potential appreciation.

Filtering the Dataset

We want to filter the dataset to fit Kee's stated needs that brought him success in previous real estate investments. This revolves around finding homes tailored towards families

propertyTaxRate = 1.98

Filter the dataset for the lowest tax rate to have all houses be comparable. The lowest tax rate has been most attractive to families & implies a good living area

hasView = 1

He has found clients value having a view. It is also a differentiating factor in Austin as many houses do not have views

numOfBedrooms >= 2

Having over 2 bedrooms means family homes which is the demographic we are targeting

latestPrice <= 1,000,000

Setting a price cap allows us to remove any high-priced outliers

avgSchoolRating >= 4

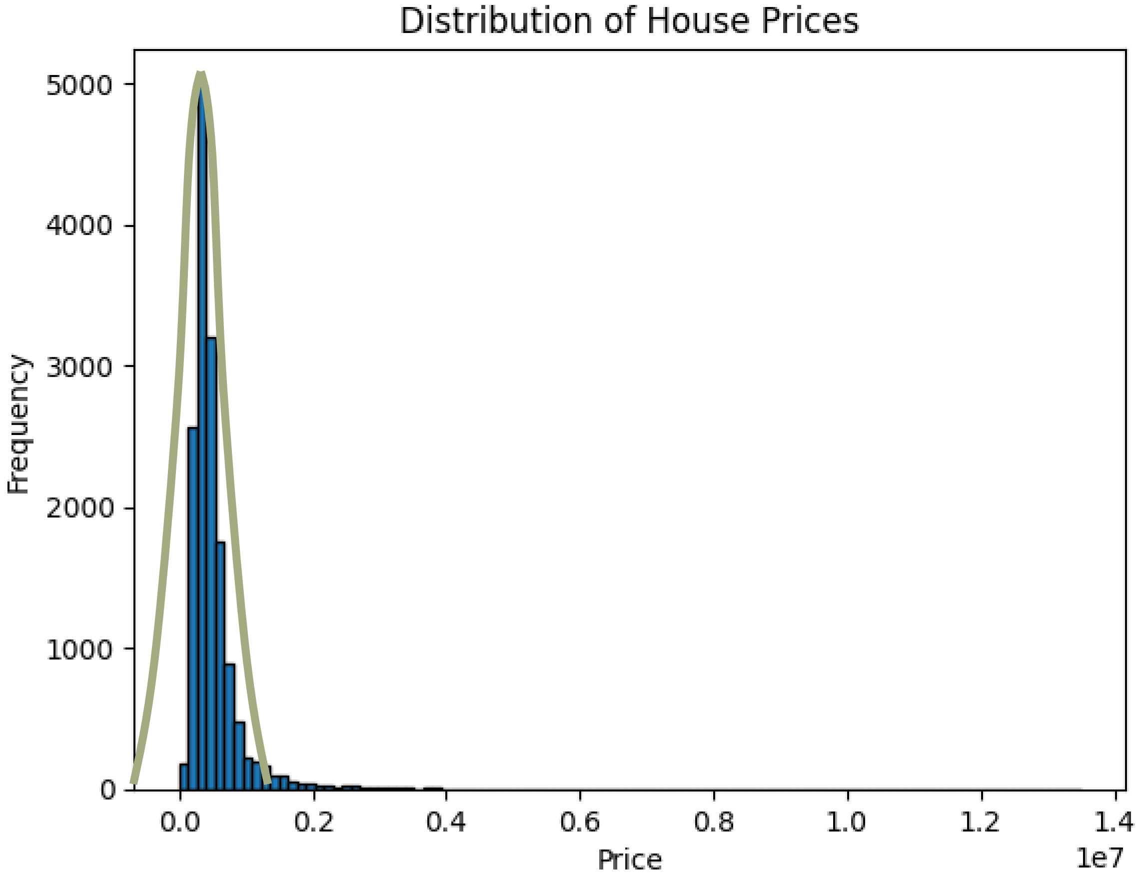
We find higher school rating areas to be more active and attractive in family home market.

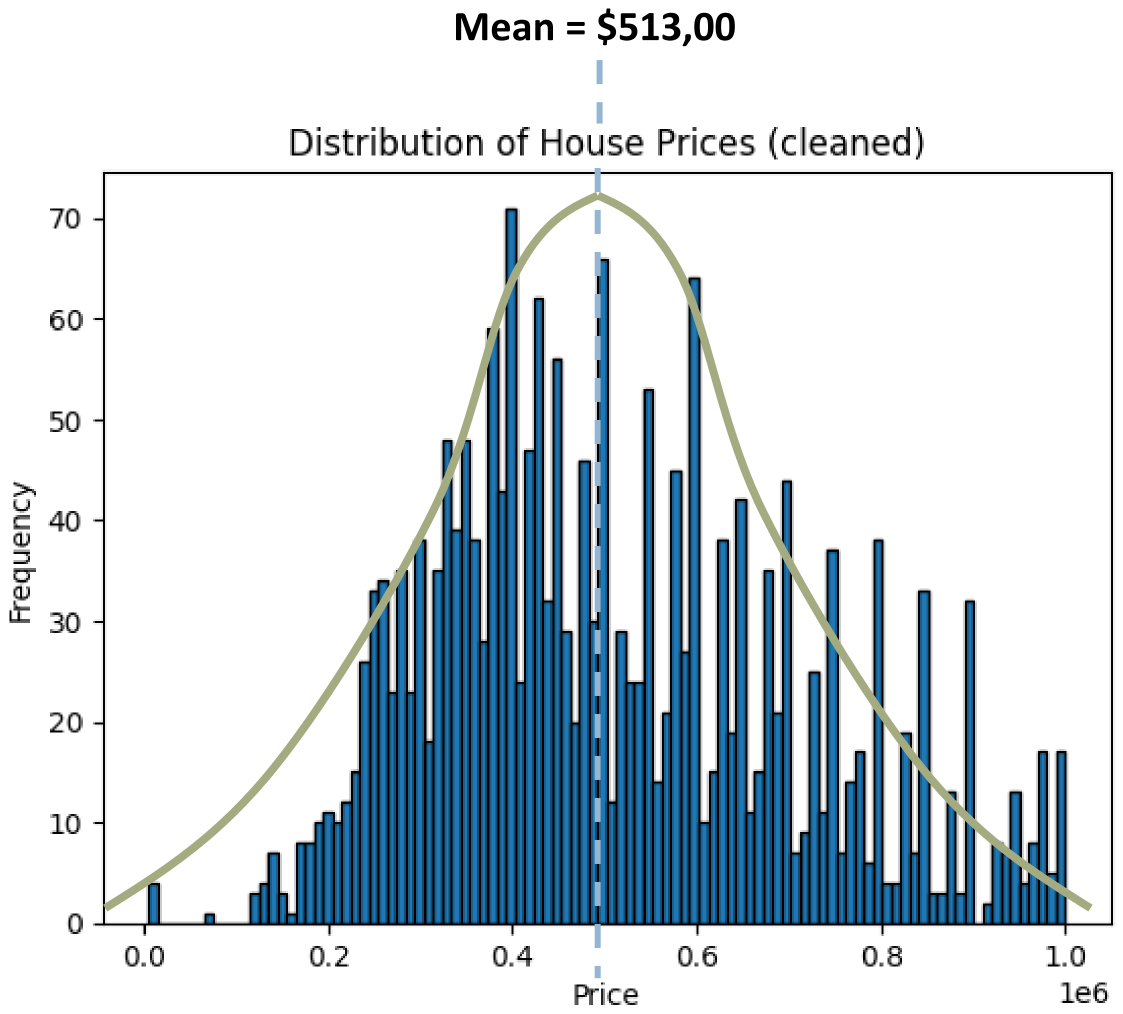
This brings the dataset from 15,171 homes to the 2,084 homes that fit these 5 conditions

1

Exploratory Data Analysis

Distribution of Price Analysis





The distribution of house prices looks very clustered on the left end with a very long tail stretching to $1.4M. These outliers at the upper end of the distribution make the data challenging to interpret.

When we remove the outliers, via a price cap, the distribution is much more spread and observable. This allows us to work with more comparable data, observe trends, and draw meaningful insights.

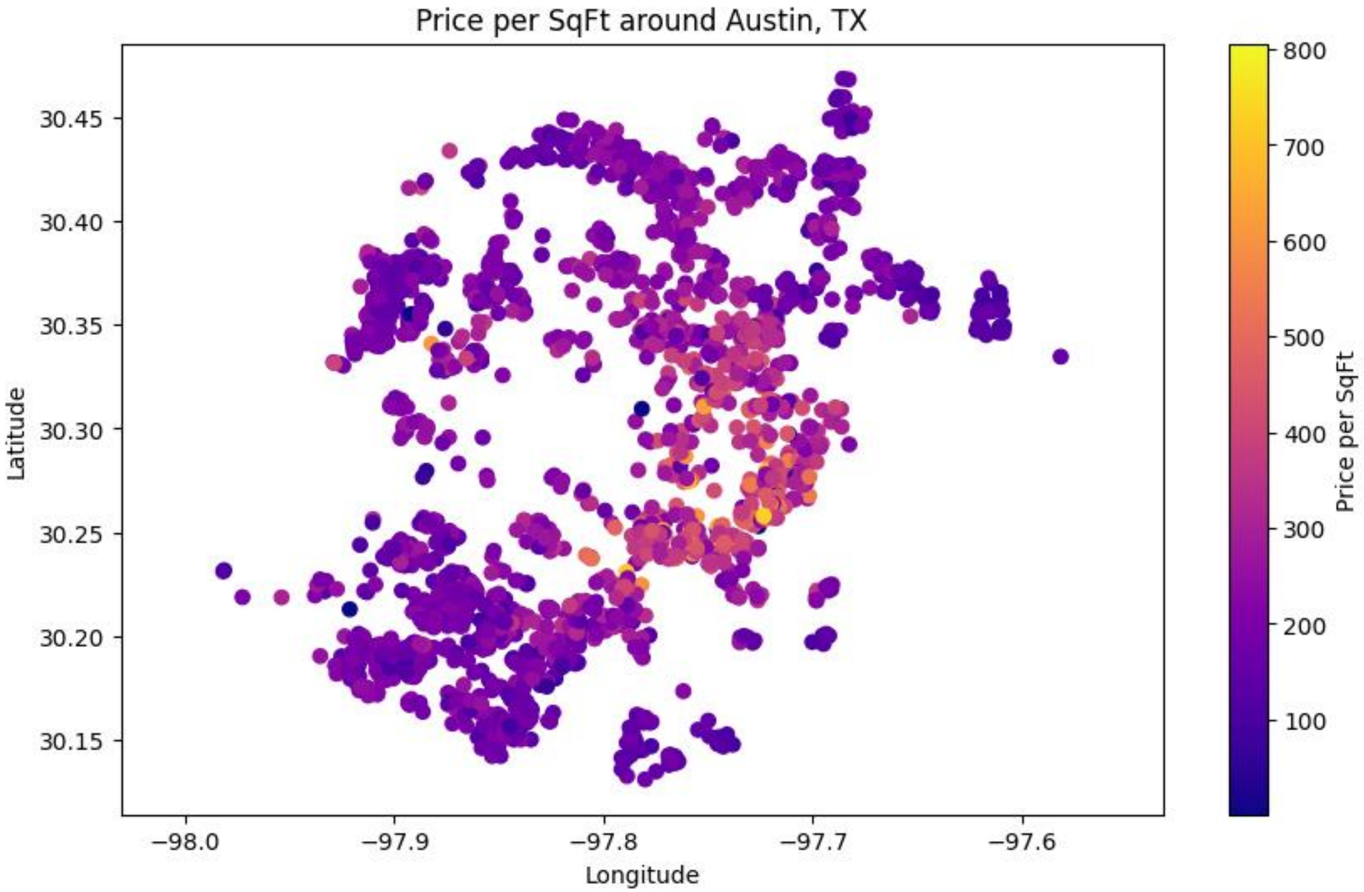
EDA Section 1 – Price per SqFt

Creation of "Price per SqFt"

- We know that livingAreaSqFt has the highest R-squared value with latestPrice
- A common metric in real-estate is price per square foot

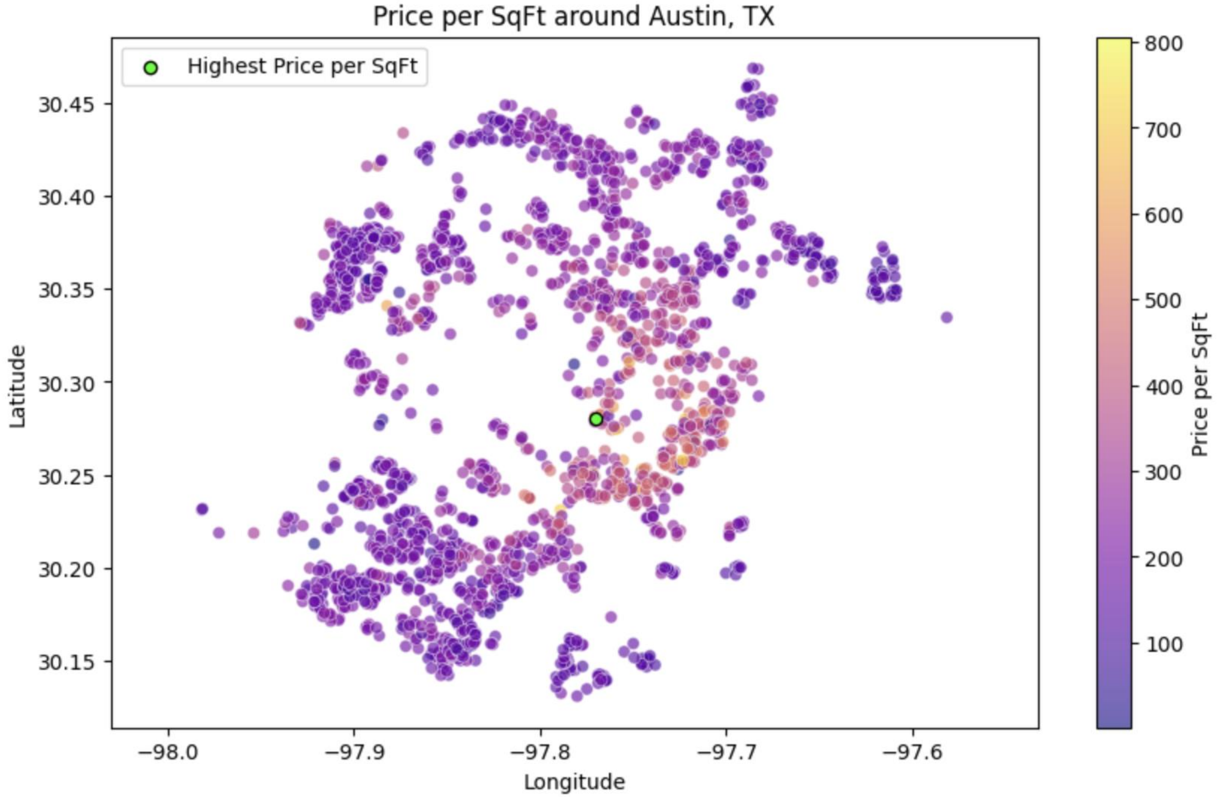
Plotting "Price per SqFt"

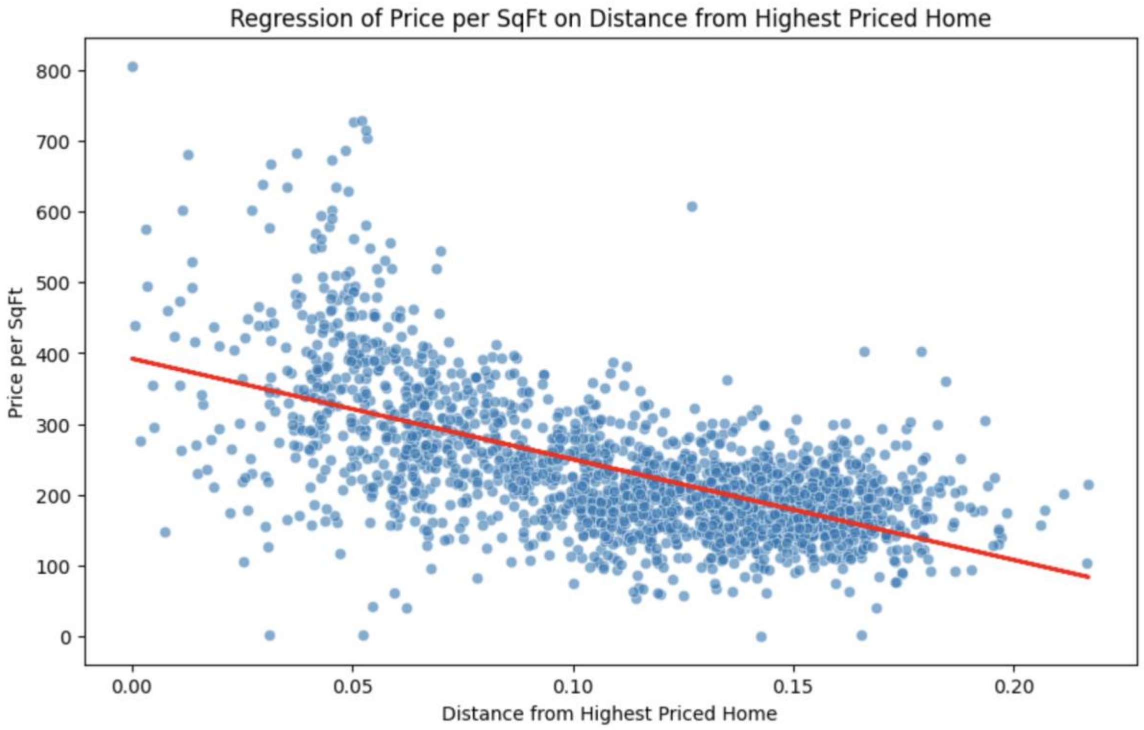
- Using latitude and longitude, we can create a scatterplot to show areas of Austin that have the highest price per square foot.
- Once plotted, the brighter points on the map show the areas with HIGH price per square foot.



*** As we move away from the red area, Price per SqFt DECREASES!**

pricePerSqFt and distanceFromMax



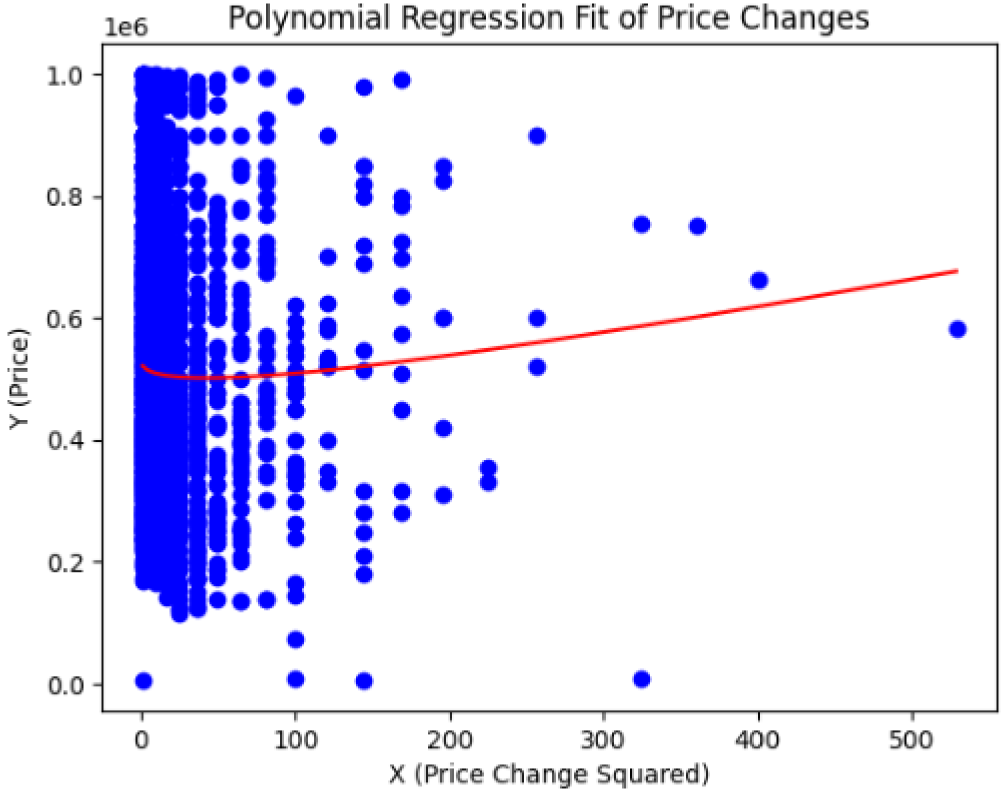


The green dot is the house with the highest pricePerSqFt. We created distanceFromMax to represent distance from this green dot.

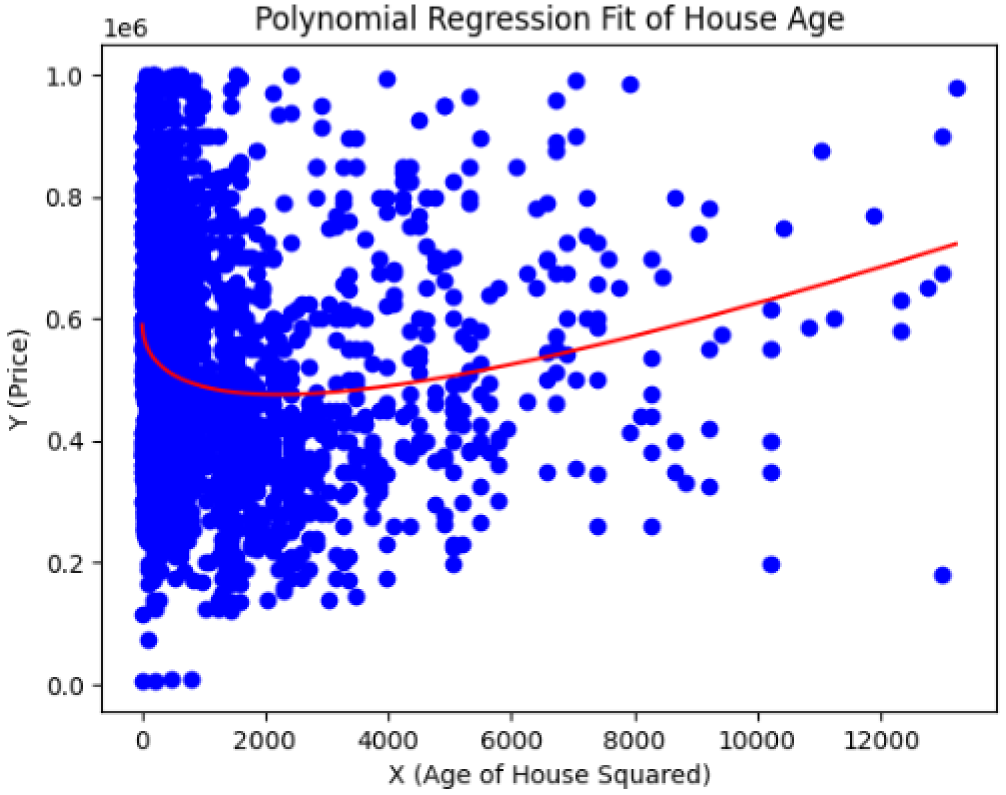
Running a linear regression model on distanceFromMax (X) and pricePerSqFt (Y), we found a notable relationship between the two. (R-squared value of 0.39)

EDA Section 2 - Polynominal Regression Analysis

We tested the number of price changes and house age (years since house was built) using a polynomial regression to assess each variables relationship to price and found a very weak predictive power for both variables.



Regression Type	R-Squared
Linear Regression	0.0121%
Polynomial Regression	0.2158%



Regression Type	R-Squared
Linear Regression	0.2569%
Polynomial Regression	2.8909%

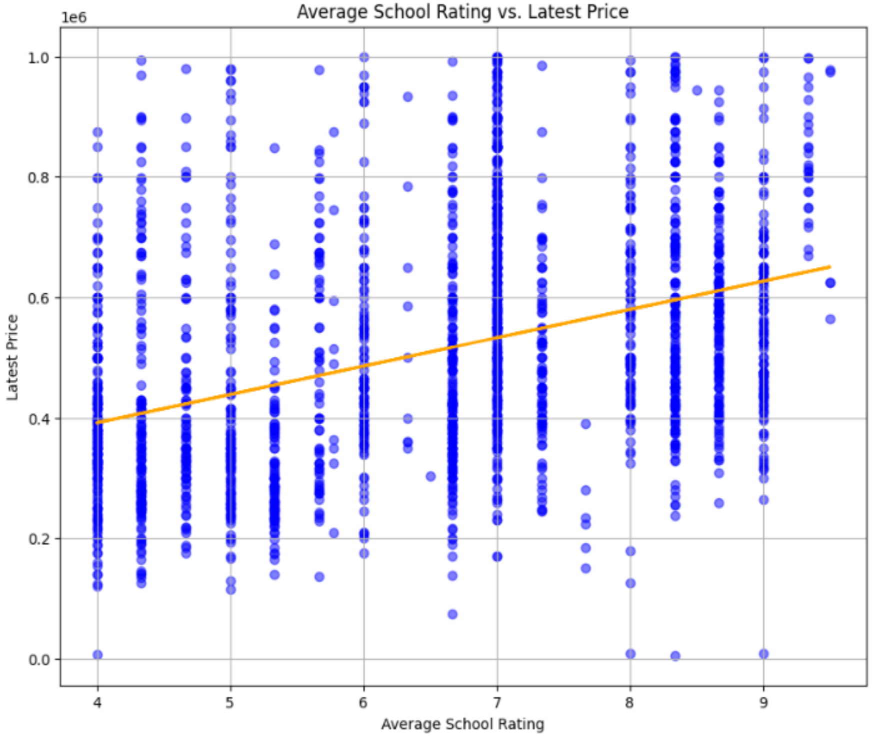
EDA Section 3 - General Analysis

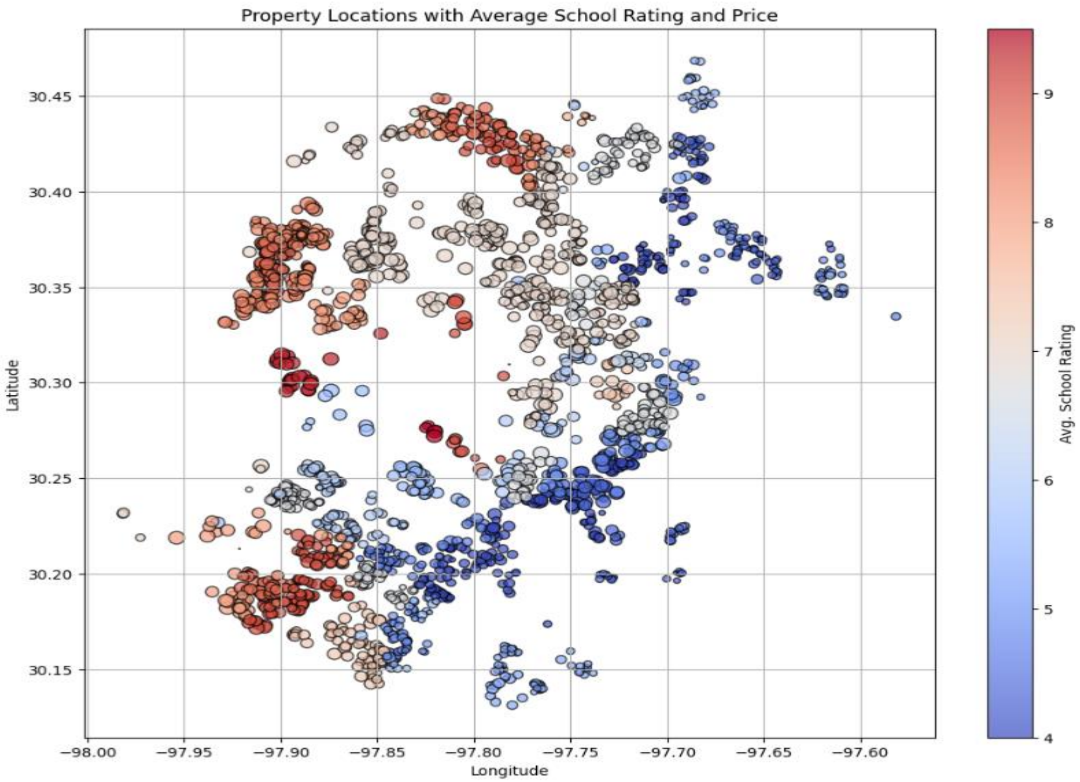
Highest Linear Regression R-square value	**Highest R-square in Multiple Linear Regression**
0.2952	**0.3369**
By livingAreaSqFt	**By livingAreaSqFt and numOfHighSchools**
Highest School Variable Correlation	**Highest Simple Linear School Variable**
0.3751	**0.1401**
By avgSchoolRating	**By avgSchoolRating**

avgSchoolRating Analysis

We identified avgSchoolRating as the most impactful variable within the school variables. Here we will explore it’s relationship with other variables that could prove impactful to our analysis

avgSchoolRating	
4.000000	186
4.333333	144
4.666667	88
5.000000	129
5.333333	97
5.666667	69
5.778694	9
6.000000	105
6.333333	9
6.500000	1
6.666667	205
7.000000	365
7.333333	86
7.666667	6
8.000000	105
8.333333	177
8.500000	1
8.666667	150
9.000000	122
9.333333	25
9.500000	5

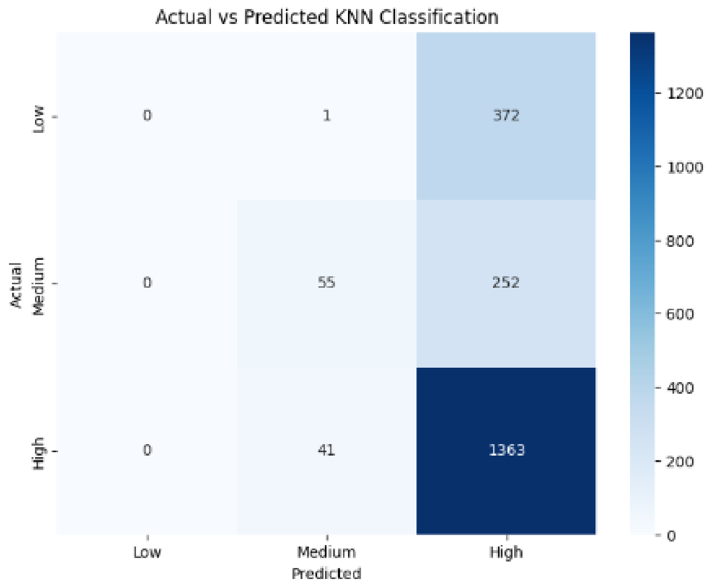




avgSchoolRating and latestPrice show positive correlation as previously mentioned where the higher the school rating the higher the price

We also explored average school ratings with location, noting that the better school areas (red) are in the west of Austin and lower ratings (blue) are on the east side

avgSchoolRating Classification Models



Low: 0-300k

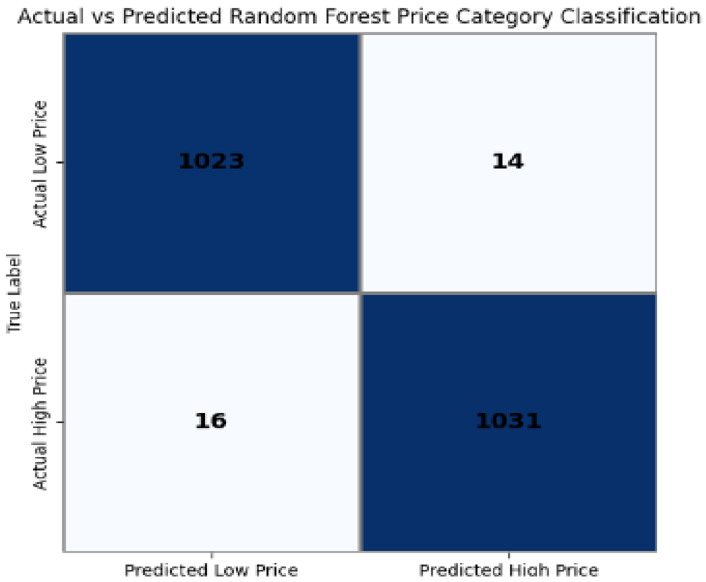
Med: 300k-700k

High: 700k-1M

68% Accuracy

This matrix shows the ability to predict housing price categories of low, medium and high in a KNN classification

The model was most accurate in predicting when house price were actually in the high-price category



Median: 479k

Low: 0-479k

High: 479k-1M

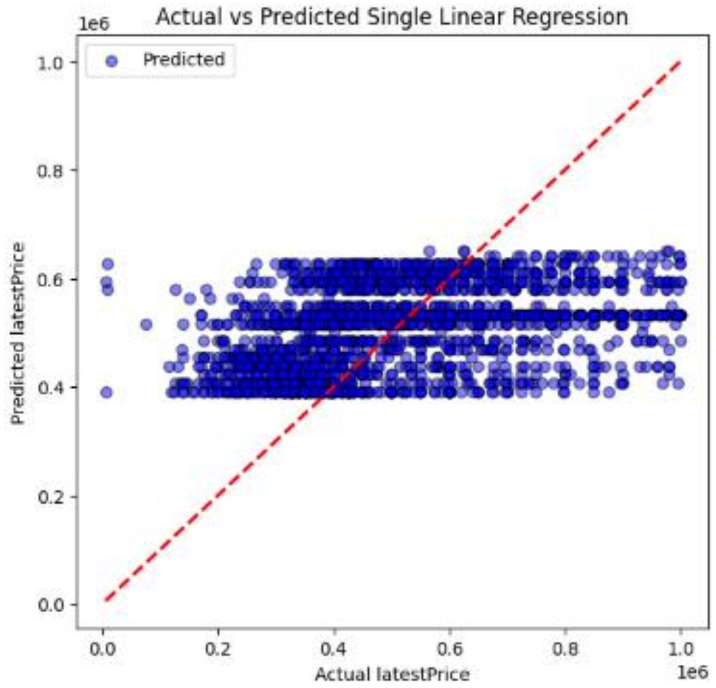
99% Accuracy

This matrix shows the ability to predict housing price categories of low and high around the median price in a random forest classification

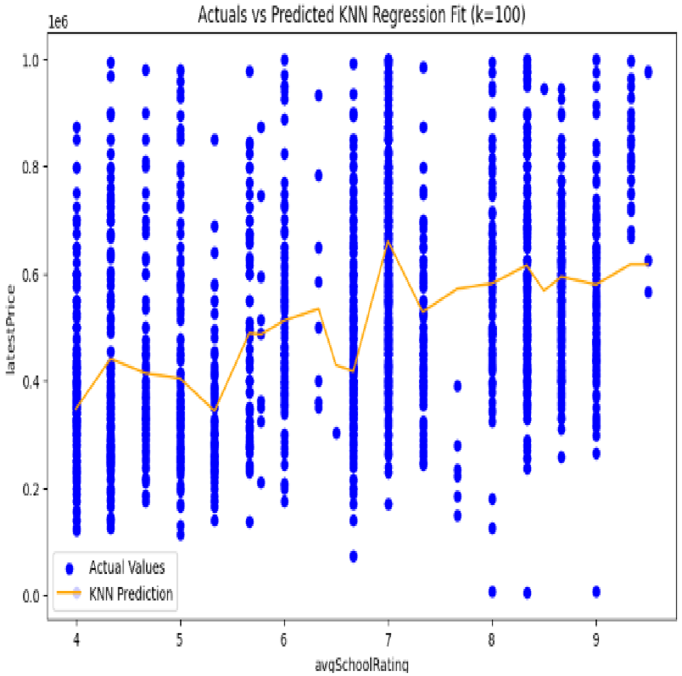
The model was very accurate in predicting both high and low prices compared to actual results

avgSchoolRating Regression Models

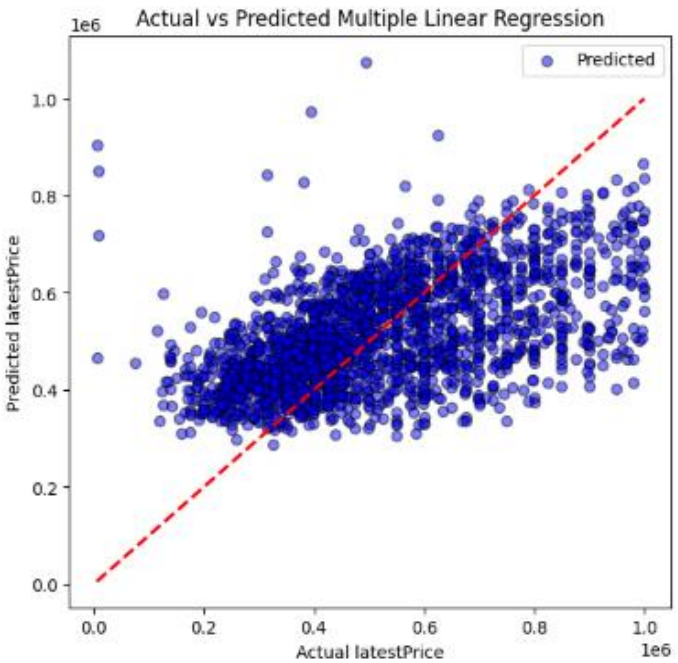
Using avgSchoolRating in regression models was not very accurate. The highest R square was in a multiple linear regression with livingAreaSqFt while the lowest was the single-linear regression



0.141 R-Square
49.95% MAPE



0.191 R-Square
47.61% MAPE



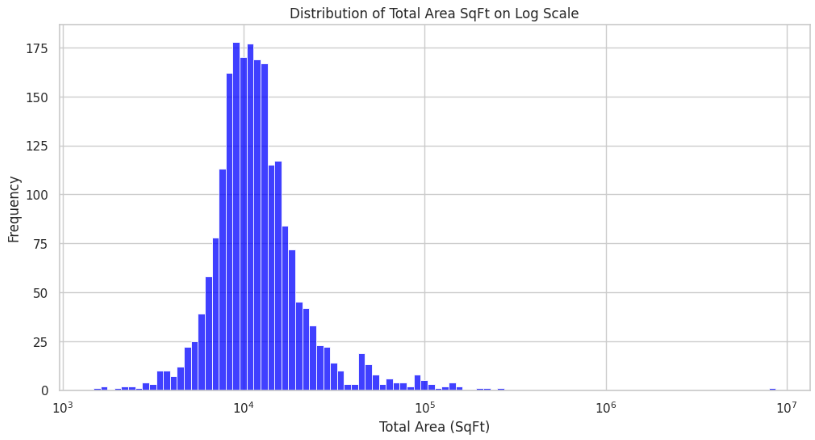
0.306 R-Square
50.84% MAPE
1.378 VIF

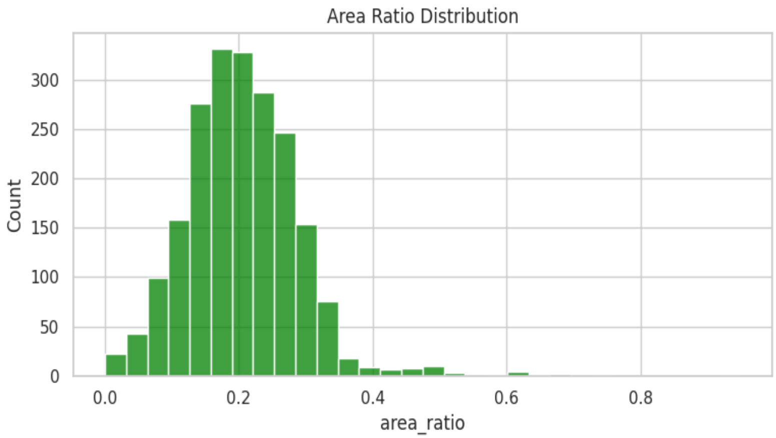
EDA Section 4 – Creating New Variables and Exploring SqFt

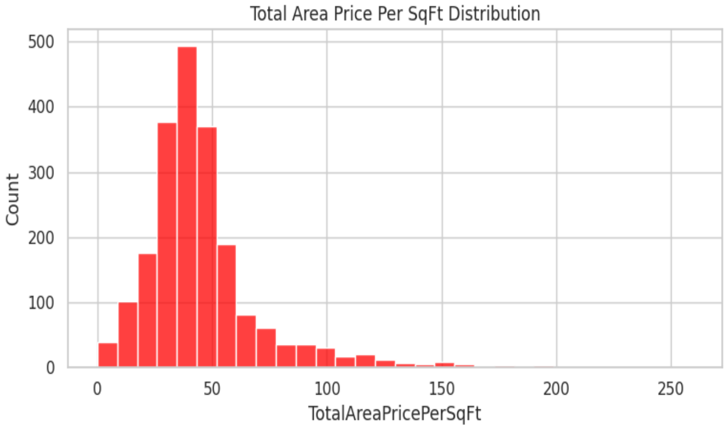
Key Variables	Living Area SqFt and Lot Area SqFt

Created Variables		
	Total Area	Living Area SqFt + Lot Area SqFt
	Living Area Ratio	Living Area SqFt / Lot Area SqFt
	Total Area Price / SqFt	Latest Price / Total Area SqFt

Exploring The New Variables







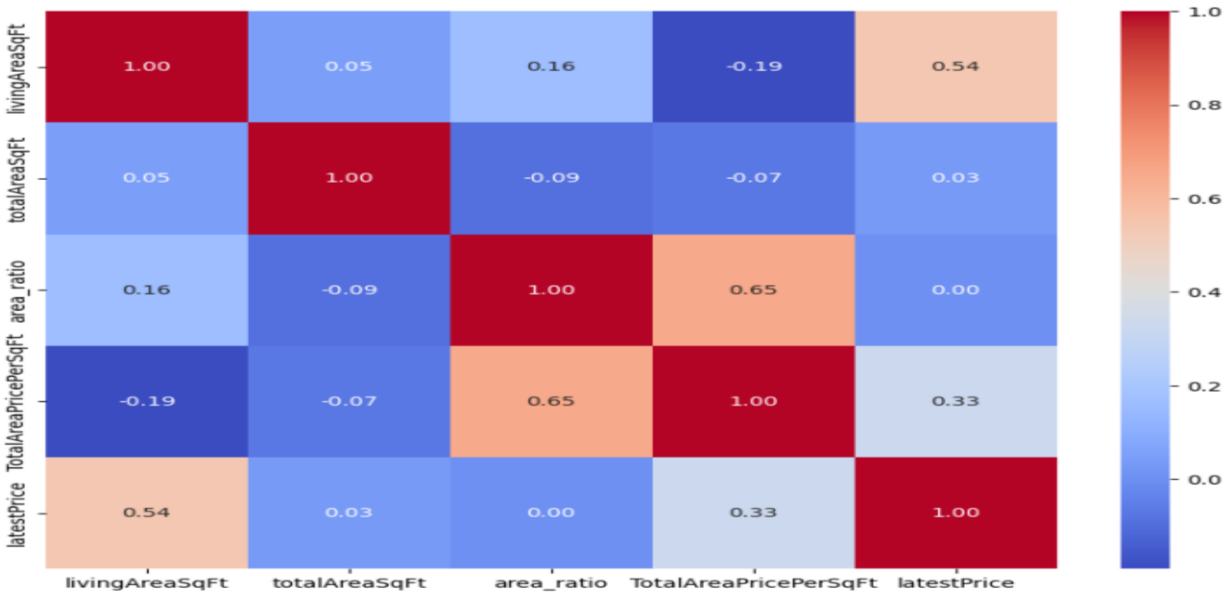
Total Area SqFt: Mean – 18999 SqFt

Area Ratio: Mean – 0.2052

Total Area Price / Sq: Mean - $45.93

EDA Section 4 – Understanding Correlation and Distribution

Correlation Matrix with Latest Price

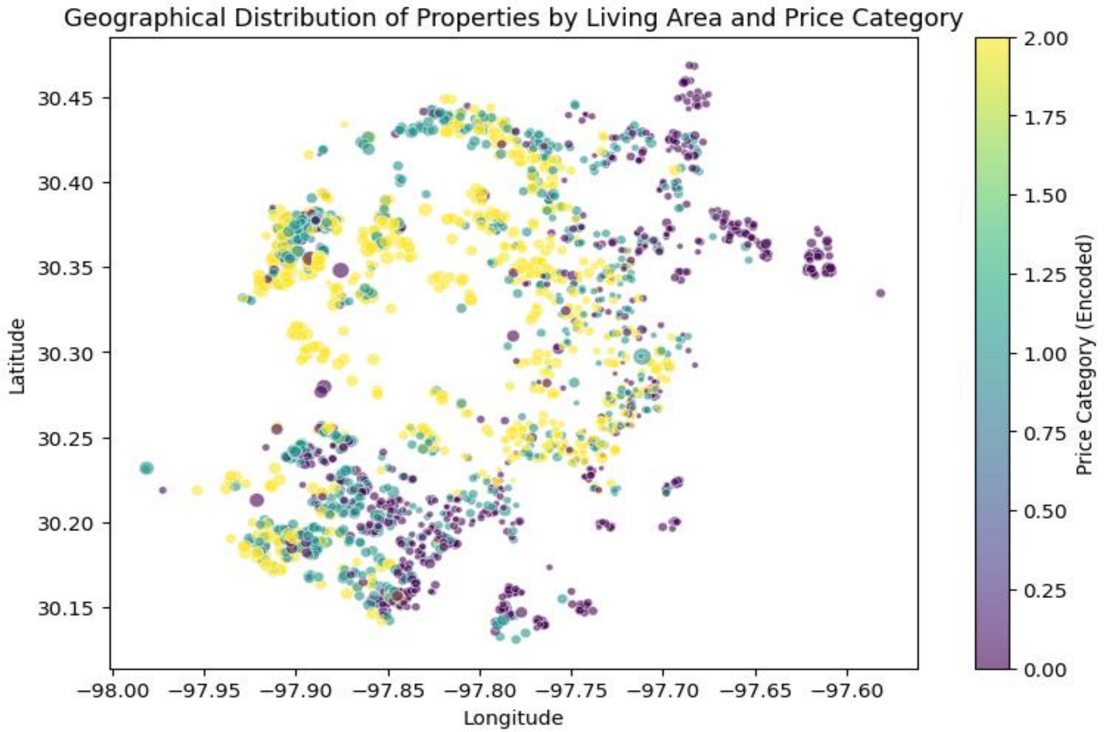


Living Area SqFt – (0.54) – Moderate Correlation

Total Area Price / SqFt – (0.33) – Low Correlation

Total Area SqFt – (0.03) – Poor Correlation

Geographical Distribution of Properties



Highlighted Trend

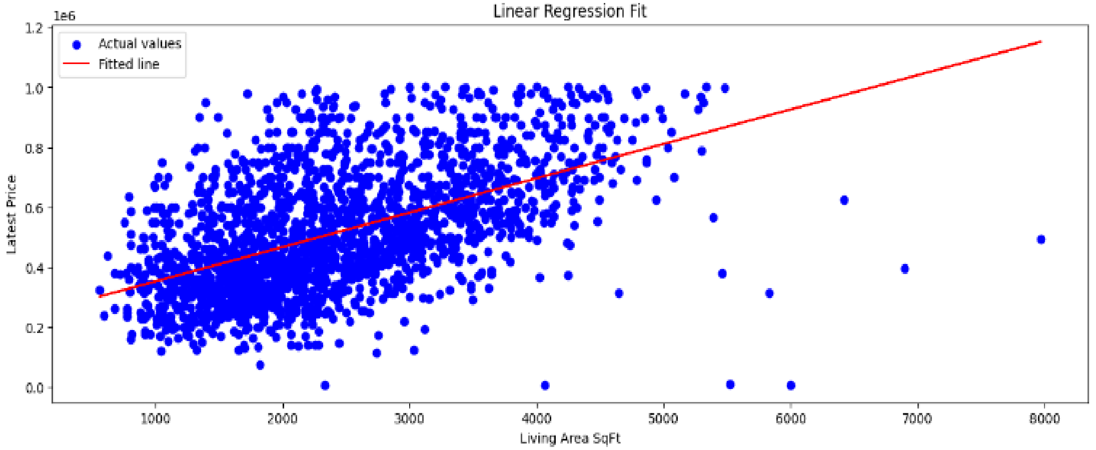
- Distribution aligns with noted trend that Price increase as the Size of the Living Area increases

Modelling Techniques – Linear Regression and KNN

Features → Living Area SqFt

Target → Latest Price

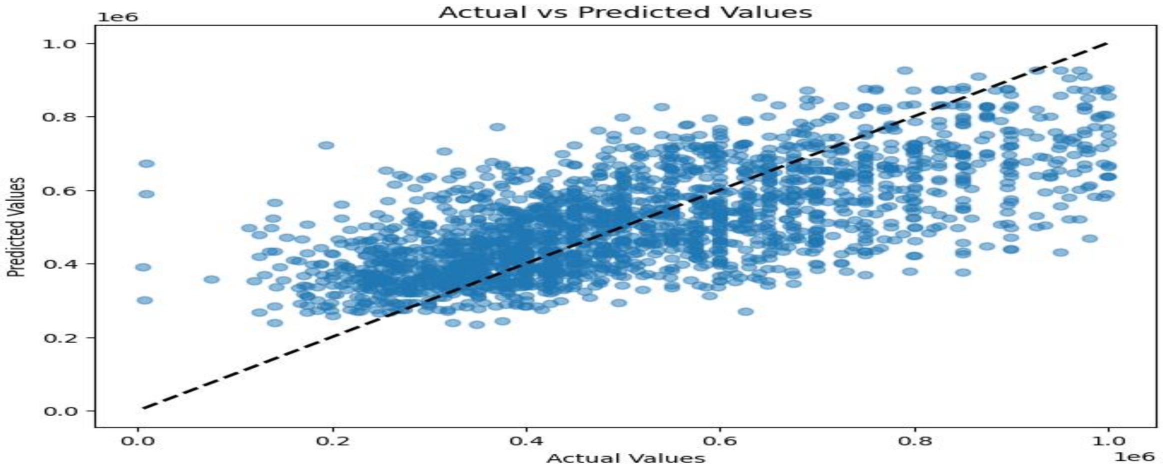
Linear Regression Model



Model Fit: 0.2952

MAE: $132,188

Best KNN Distance Model: Euclidean



R-Squared: 0.4517

MAE: $116,332

Predicting Price Using the KNN Distance Model: Euclidean

Predicted Price for New House

2500 SqFt → $485,165,194

Cross-Validation Results

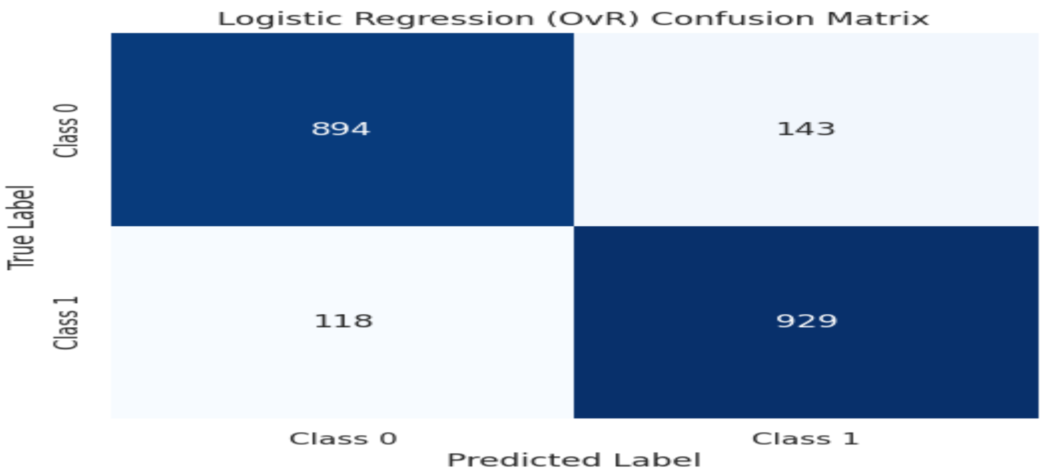
R-Squared: 0.1787
MAE: $141,959

Modelling Techniques – Logistical Regression and KNN

Features → Lot Size SqFt, Living Area SqFt, Total Area SqFt, Area Ratio, Total Area Price / SqFt

Target → Latest Price

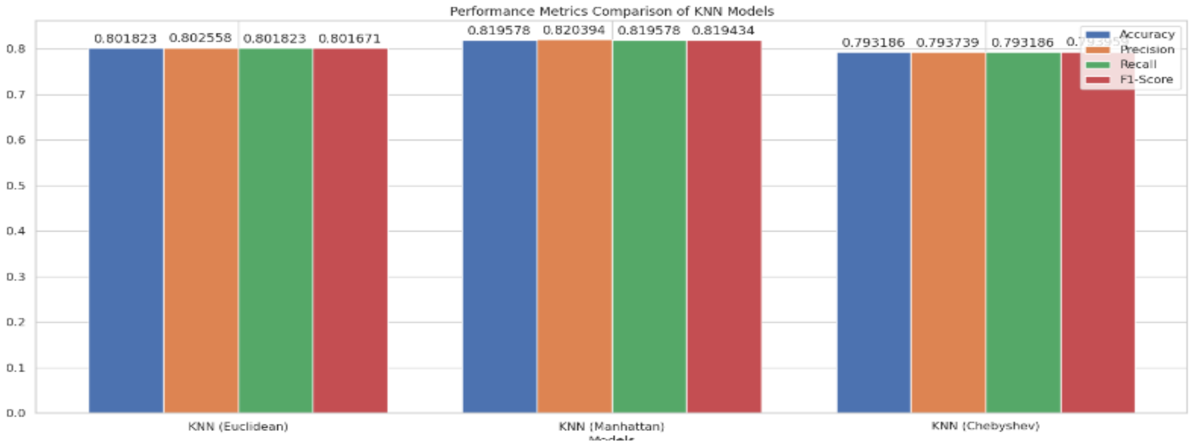
Logistical Regression Models (OvR and Multinominal)



Precision: 0.87

F1-Score: 0.87

Comparing KNN Distance Models



Top Performance: Manhatten

F1-Score: 0.82

Predicting Price Using Logistical Regression (OvR)

Lot: 7000 SqFt

Living: 3000 SqFt

→ Predicted Price for New House

$575,500,00

Cross-Validation Results

R-Squared: 0.1871

MAE: $139,207

EDA Section 5

Correlation Analysis

Strongest Correlation

Bathrooms – 0.4437

Moderate Correlation

Bedrooms - 0.3118

Stories - 0.2606

Photos - 0.1599

Weak Correlation

Window Features – 0.1091

Patio/Porch Features - 0.1062

Security Features – 0.0925

Parking Features – 0.0811

Negligible Correlations

Accessibility Features - 0.0186

Appliances - 0.0103

Waterfront Features – 0.0075

Bad Distribution Variables

Accessibility Features

Waterfront Features

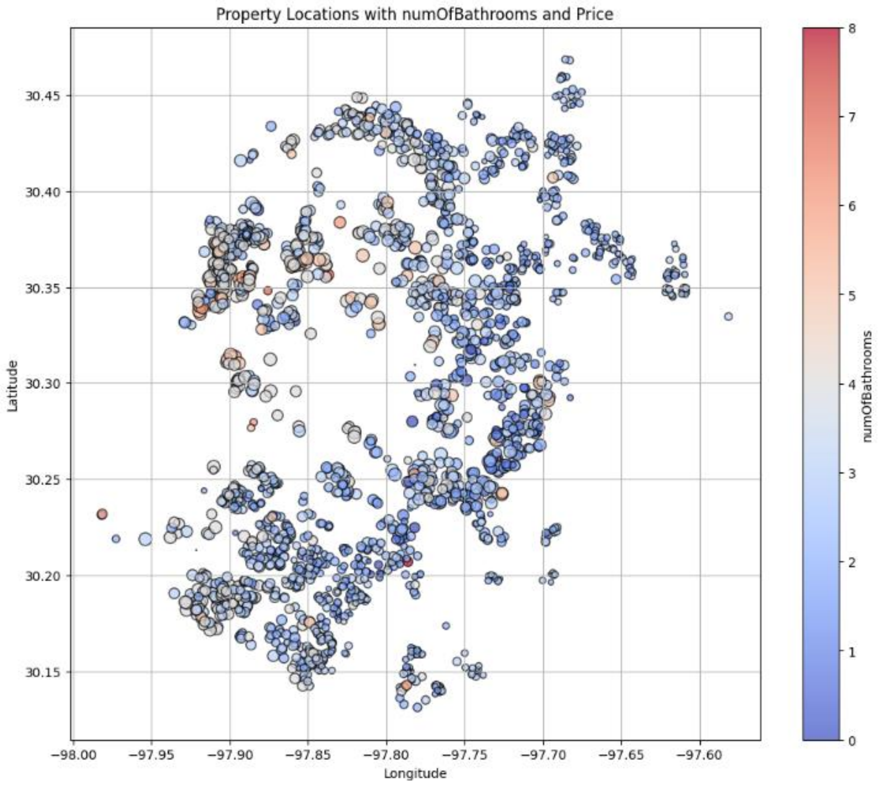
The vast majority of properties have no accessibility features or waterfront features, with only 19 properties having one or more accessibility features and 7 having one or more waterfront features.

Number of Bathrooms and Price Analysis

We identified the number of bathrooms is the most impactful variable within this cluster of variables. Here we will explore its relationship with other variables that could prove impactful to our analysis

R^2= 21.25%

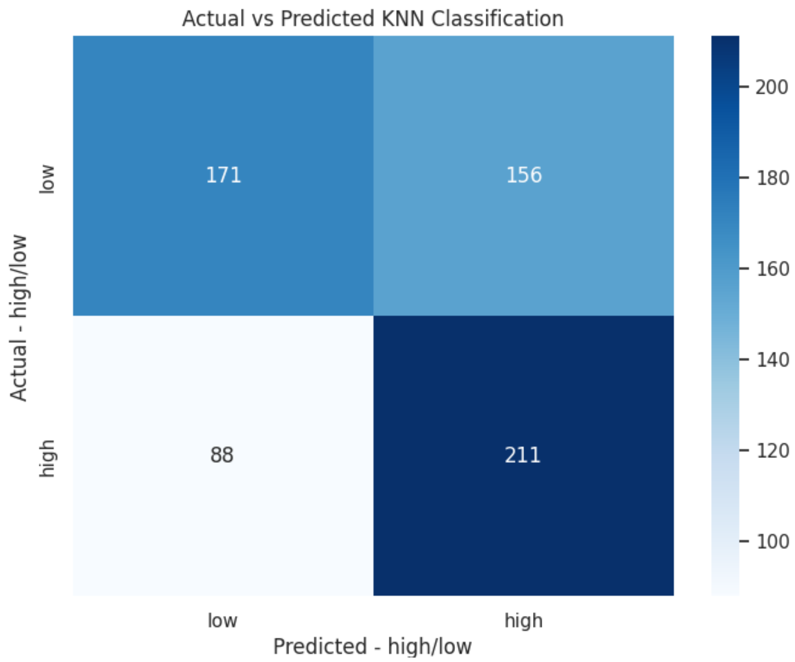
Number of bathrooms and property price has the strongest linear relationship in this group where a higher number of bathrooms generally means higher property price



We also explored number of bathrooms with location, noting that more bathrooms (red) are in the north-west of Austin and less (blue) are in other areas

Regression and Classification Models

We created a regression model and a KNN classification model predicting property price based on number of bedrooms, bathrooms and stories



Median: 479k

Low: 0-479k

High: 479k-1M

61% Accuracy

Number of bed, bath and stories has a 21.86% accuracy of predicting property price. This model has low predictive power and limitations.

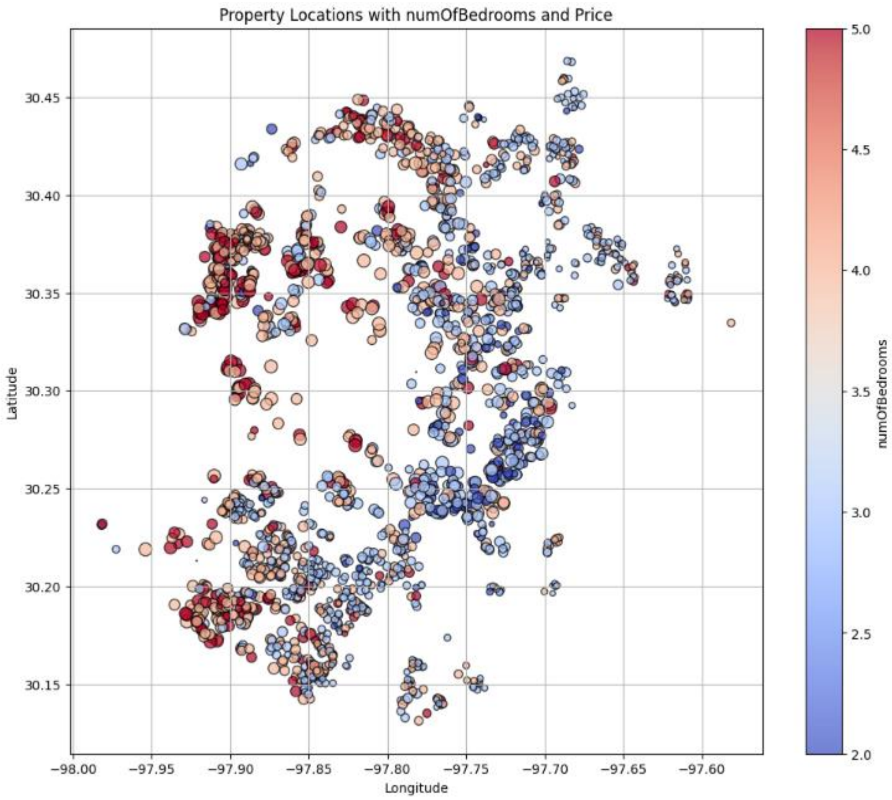
This matrix shows the ability to predict whether property price is low or high in a KNN classification, with anything over the median being high, and under being low

Number of Bedrooms and Price Analysis

We identified the number of bedrooms is one of the most impactful variables within this cluster of variables. Here we will explore its relationship with other variables that could prove impactful to our analysis

R^2= 11.98%

The number of bedrooms also displays a positive linear relationship with property price, reflecting that properties with more bedrooms are generally more expensive.



We also explored the number of bedrooms with location, noting that more bedrooms (red) are in the west of Austin and fewer bedrooms (blue) in other areas

2

Insights

Insight #1 – Pricing Based on Neighbours

Goal

- To determine the impact of home location and prior neighbouring sales on the price of a home

Rationale

- Many neighbourhoods have similar house features (size, bedrooms, bathrooms, etc.) and will share location characteristics (school rating, # of schools, tax rate, etc.)

Challenges

- Houses may not be exact matches to their neighbours, making a precise calculation challenging
- The model may over-generalize based on availability of house sales in the area

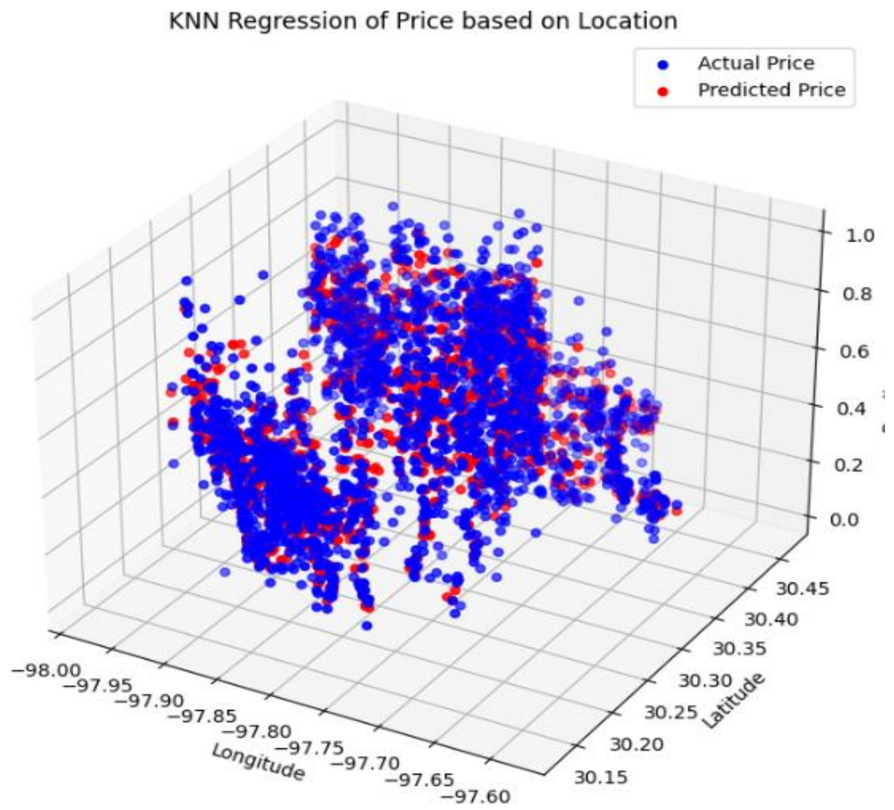
Solution

- Use a multi-variable K-Nearest Neighbours regression using longitude and latitude to approximate location
- Use the same KNN predictors to classify into price category (budget, mid-range, and premium) on a min-max scale

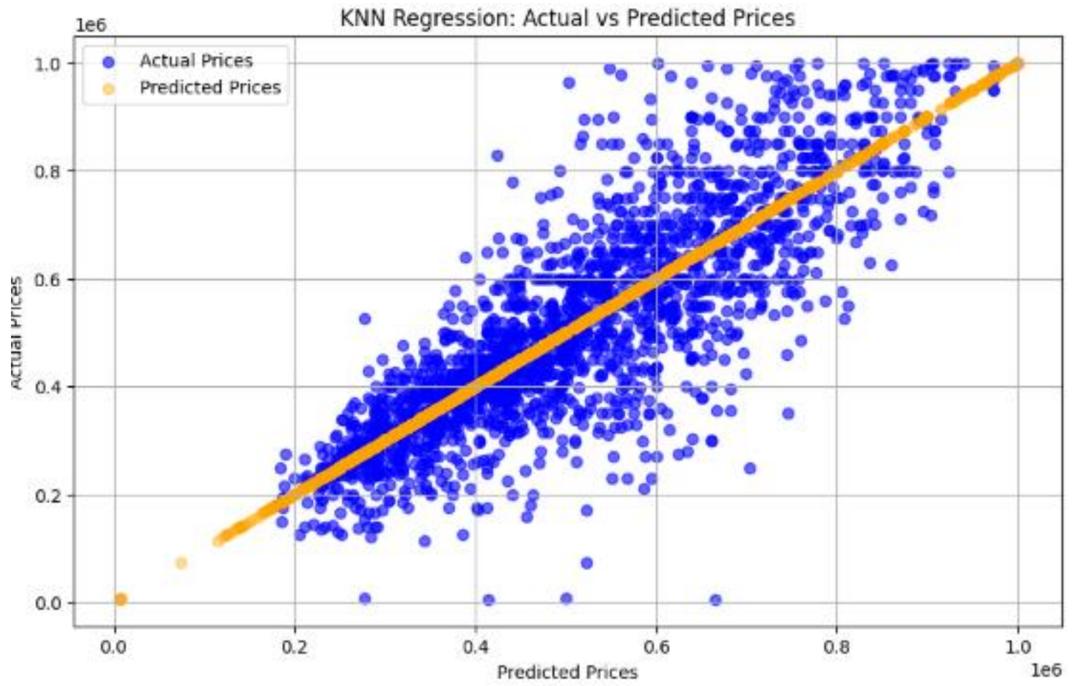
Neighbour Based KNN Regression

With a k value of 3 and distance calculated using the Manhattan method, the model was a strong predictor of price, showcasing its accuracy and suitability for ordinal data

Longitude and Latitude were used in a multivariable KNN regression model



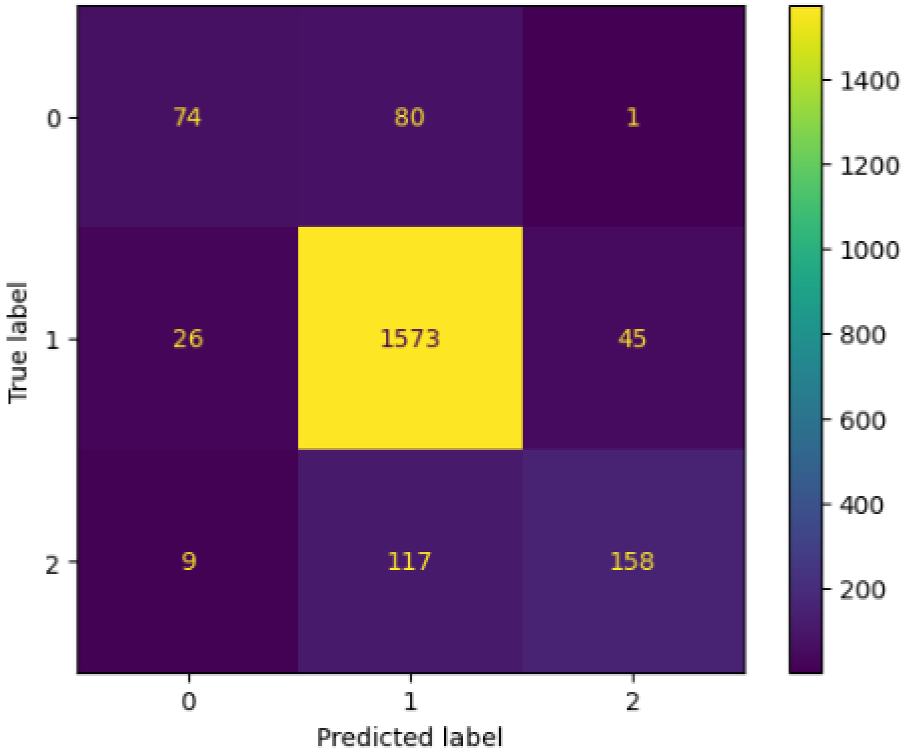
The model yielded an R-Squared of 71.3%, explaining the majority of price changes



Neighbour Based KNN Classification

With similar parameters to the KNN regression (k-value and distance method), the KNN classification model effectively split the predicted prices into the 3 defined price categories.

Confusion Matrix (below) visualizes the classification accuracy score of 86.7%



Categorized price using the Min-Max method and classified into the following categories

	Actuals	**Predicted**
Budget (Bottom 25%)	7%	5%
Mid-Range (Middle 50%)	79%	85%
Premium (Upper 25%)	14%	10%

Insight #2 – Predictive Pricing model

Goal

- Create a random forest model where we can predict the price of a house based on important variables to our client
- Look to uncover value with the latest price of houses by finding those with a higher predicted value than currently listed
- Narrow our dataset from 2084 observations to a list of 100 of the best-value homes to evaluate in our recommendation

Challenges

- It is difficult to ensure that the model does not have outliers showing value when in reality it is the model not fitting properly
- Should we focus on returning the highest dollar value return or the highest percentage return

Solution

- We identified 6 variables that provide us with a strong set of 1087 feasible homes fitting our client’s wish list for his investment
- Afterwards we filtered by its IQR to remove outliers and create a 100-observation dataset of the highest percentage returns

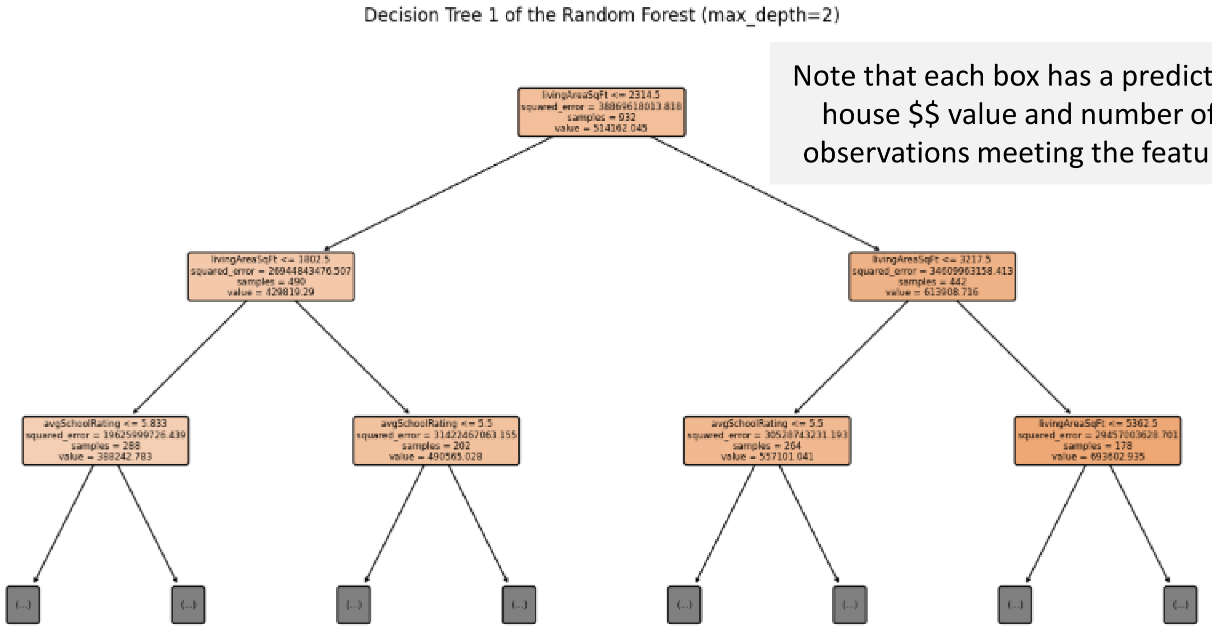
Random Forest Regression Model

Random Forest is a model that builds multiple decision trees based on training sets that output the mean prediction of the trees, thereby improves the predictive accuracy

How does it work?

1. Create multiple subsets of the training data through random sampling
2. For each subset, a decision tree is built
3. At each split on the tree, a random subset of features is considered for the best split
4. The predictions of the splits at the bottom are then averaged and used to arrive at a final prediction (in our case a predicted price)

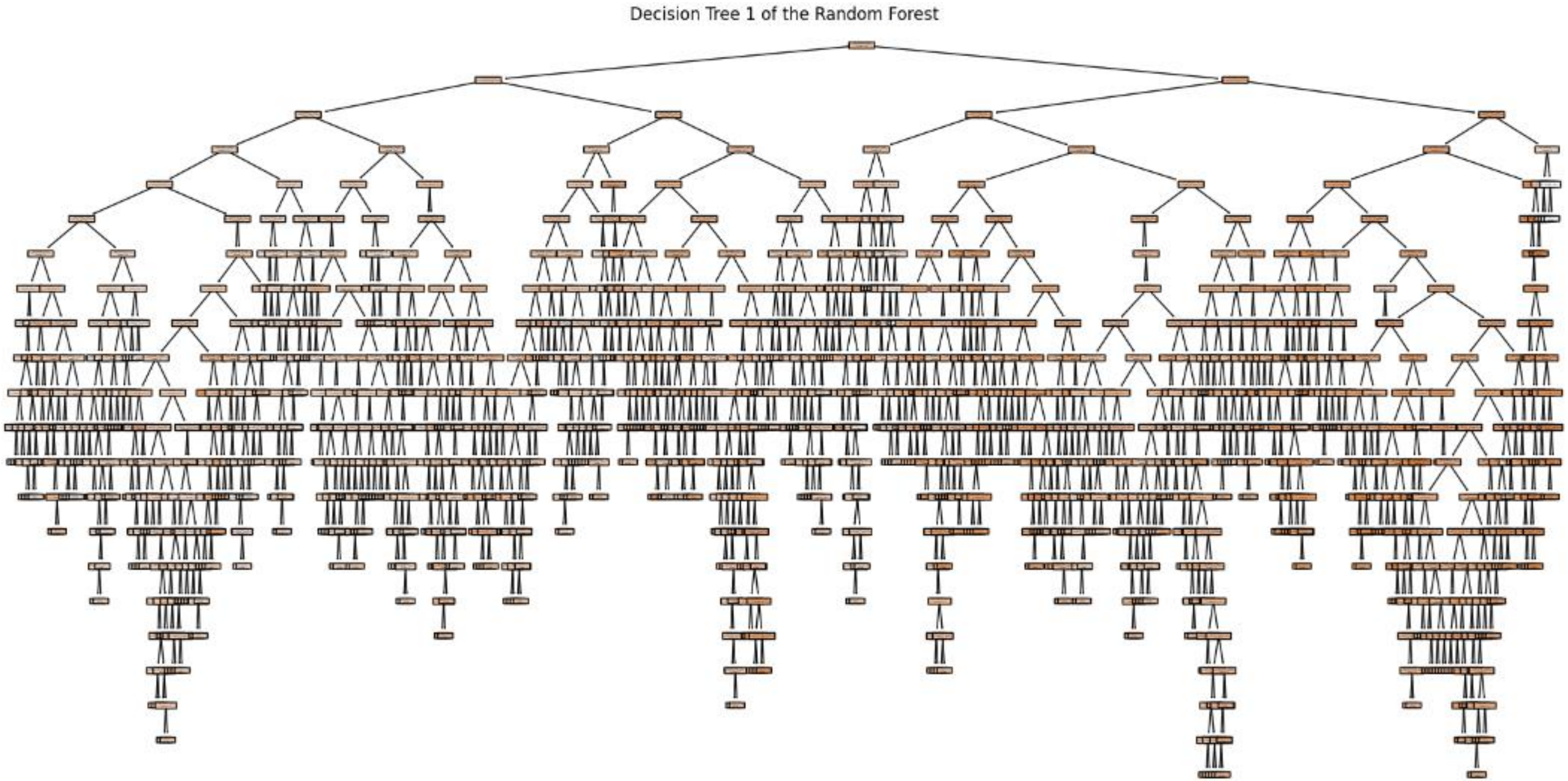
Visualization



Note that each box has a predicted house $$ value and number of observations meeting the feature

If the observation meets the features, you move down the left split of tree into the next feature. In our dataset it goes from measuring livingAreaSqFt to avgSchoolRating. If you do not meet the value then you go right where the limit is stretched higher

... And Now For a Single Decision Tree in Our Random Forest Model



Our model walks each of the observations through 100 trees…

Insight #2- Predictive Pricing Model

6 Filtered Variables

- avgSchoolRating >= 5
- livingAreaSqFt >= 1500
- numOfBedrooms >= 3
- numOfBathrooms >= 2
- hasGarage >= 1
- hasCooling >=1

These filtered variables aligns with our client's goal of finding ideal sized homes that are suitable to families in the Austin area

90.35%

Accuracy of model

$346,776

Predicted price of houses meeting these marks

Insight #2- Predictive Pricing Model

Here we show a sample of the top 10 percentage differences between predicted price and actual price, where predicted>actual to identify value buys before and after filtering out unrealistic outliers

House ID	Difference (%)
14580	3525.47
5796	2942.75
4823	2605.79
10996	221.48
4198	152.98
9732	146.48
5401	144.46
11026	133.68
13273	114.56
5326	106.47

House ID	Difference (%)
12370	40.00
5246	39.53
7156	39.30
14526	39.18
7425	38.84
5346	38.83
8995	38.45
13963	38.35
8615	38.00
8519	37.99

We now have a filtered predictive model to determine the price of a house, which has been subsettedd to the 100 best valued homes that will be used in our recommendation

Insight #3 - Austin Housing Market Appreciation

Goal

- To calculate the compound annual growth rate (CAGR) of the Austin housing market
- Determine whether the market is investable: positive CAGR == good investment, negative CAGR == bad investment.

Challenges

- We do not have historical housing data for each house. We are only given one price, representing the latest sale price.

Solution

- Using Sci-kit Learn library and K-Means clustering, we can group similar houses based on their attributes.
- After locating the largest cluster, we can reasonably assume that they appreciate at the same rate.
- We can use the latestPrice and latest_saleyear, along with the averages of those values among the cluster, to calculate a CAGR.

Insight #3 – Variable Creation

Cluster Identification – df_largest_cluster (DFLG for simplicity)

After finding the largest cluster of houses with similar features, we can begin to create variables and assign them to the new dataframe. The cluster contains 67 houses.

1. **average_latest_saleyear**
   This variable takes the mean of the latest saleyears from the cluster. *This value ended up being 2019.3*

2. **average_latest_saleprice**
   This variable takes the mean of the latest saleprice from the cluster. *This value ended up being $362,237*

3. **years_between_sales**
   This variable calculates the difference between the latest_saleyear and average_latest_saleyear for each of the 67 houses in the cluster.

4. **average_annual_appreciation**
   This variable calculates the CAGR using the variables above.

CAGR Formula =

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

Insight #3 – Applying CAGR Formula

Applying the CAGR formula to every single house in the cluster (DFLG), and taking the mean:

7.99%

Compound Annual Growth Rate of Austin Housing Market

A Decade of Impressive Growth in Austin Home Prices

- Over the past ten years, Austin has experienced exceptional real estate appreciation. Homes in the city have **seen a remarkable increase in value of 123.20%,** translating to an impressive average annual growth rate of 8.36% – **Neighborhoodscout**.

Source: https://www.noradarealestate.com/

7.99% CAGR is relatively close to the OVERALL Austin housing market annual growth rate. The variance could be due to the cluster that was chosen or the lack of historical price data.

3

Recommendation

Recommendation – Predict Viable Houses' 5-year Appreciation

Use 100 most underpriced houses from insight #2

The model from insight 2 was the most accurate when predicting house prices, and using this, we found the most underpriced houses

Apply the CAGR from insight #3 to 10 most recently sold houses on the list

To ensure that we are projecting into the future, and not a house with a latest_saleyear from OVER 5 years ago, we will ONLY project the 10 most RECENTLY sold houses

Present list of houses to our client

The list of 10 houses that is generated from this will be the final recommendation to Kee, our valued client

Recommendation List

House ID	latest_saleyear	latestPrice	projected_value	yearBuilt
4505	2020	258,000.00	378,944.55	2014
11008	2020	166,246.00	244,178.35	2012
2654	2020	550,000.00	807,827.51	1990
2156	2020	247,000.00	362,787.52	2011
12006	2020	280,000.00	411,257.65	2006
5807	2020	259,000.00	380,413.32	2003
649	2020	439,900.00	646,115.14	1998
7554	2020	245,000.00	359,850.44	1976
1731	2020	369,900.00	543,300.73	1997
2258	2020	419,000.00	615,417.69	1999

Houses in this list all appreciate by the CAGR, and are projected out 5 years

According to our model, the houses in this list should all be worth their projected_value in 5 years

The next steps for Kee to take would be to approach a real-estate agent to establish offers for the houses

Thank you

Questions?