AUSTIN HOUSE PRICE PREDICTION

Kesav Adithya Venkidusamy

Table of Contents

Introduction 2
Introduce the problem 2
Justify why it is important/useful to solve this problem 2
How would you pitch this problem to a group of stakeholders to gain buy-in to proceed? 2
Explain where you obtained your data 2
Organized and detailed summary of Milestones 1-3 3
EDA; include any visuals you think are important to your project 3
EDA 3
Visualization 4
Data preparation 8
Filter 8
Duplicate Check 9
Outlier Deduction 9
Dropping unwanted features 9
Missing data 10
Transformation 10
Model building and evaluation 11
Target Analysis 11
Correlation 11
Linearity check 13
Box Plot 14
Linear Regression 14
Interaction 14
Conclusion 14
What does the analysis/model building tell you? 14
Is this model ready to be deployed? 15
What are your recommendations? 15
What are some of the potential challenges or additional opportunities that still need to be explored? 15

Introduction

Introduce the problem

During covid-19 pandemic, the housing market has been increased exponentially across USA. Texas is one of the states which saw the house prices to skyrocket. Among various cities in Texas, I have chosen the house sale data available in and around Austin for this analysis. There are numerous factors contribute to increase in home price. We will explore what factors (such as location, square feet, school ratings) of the house affects the sale price. We will build and evaluate a regression model to predict features affecting the house prices in Austin, Texas location based on the available data.

Justify why it is important/useful to solve this problem

Predicting house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. It also helps to identify the features affecting the house price. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?

The main stakeholders for this project are marketing team of real estate companies, realtors who involved in selling and buying the house, sales team of home building companies who build and sell houses. In pitching the problem to these stake holders to gain buy-in, the main focus is to understand the reason for increase in house price. As everyone aware, pandemic has caused house price to skyrocket. This model could help the team to identify the features affecting the home price and focusing on those factors while buying and selling homes would eventually improve the revenue and net income of the company.

Explain where you obtained your data

In this project, we will work with a dataset on the Austin housing data. The data is downloaded from the Kaggle using the below link:

https://www.kaggle.com/datasets/ericpierce/austinhousingprices.

The dataset contains 47 columns and ~15000 rows. The dataset contains house sale price for 3 years (2018-2021) in and around Austin, Texas area. Basically, this is dataset extracted from Zillow website which tells the features of the house and sold price. "latestPrice" is the price that home has been sold and this will be used as target for our model. Some of continuous features available in the dataset are as follows which will use for model.

- Year built
- Lot size square ft
- Living area square ft
- Average school distance
- Average School rating
- Number of bathrooms
- Number of bedrooms
- Price per square foot
- Number of price changes
- Number of appliances

Below are some of the categorical variables present in the dataset.

- Has Association
- Has Cooling
- Has Heating
- Has Garage
- Patio porch
- Security
- Number of parking features
- Community
- Number of Primary
- Number elementary
- Number of middle and high schools

Organized and detailed summary of Milestones 1-3

EDA; include any visuals you think are important to your project

EDA

- The dataset has ~15000 rows and 47 columns and there is no missing data
- There is no duplicate row present across the dataset
- Data doesn’t have any null values present in any of the column

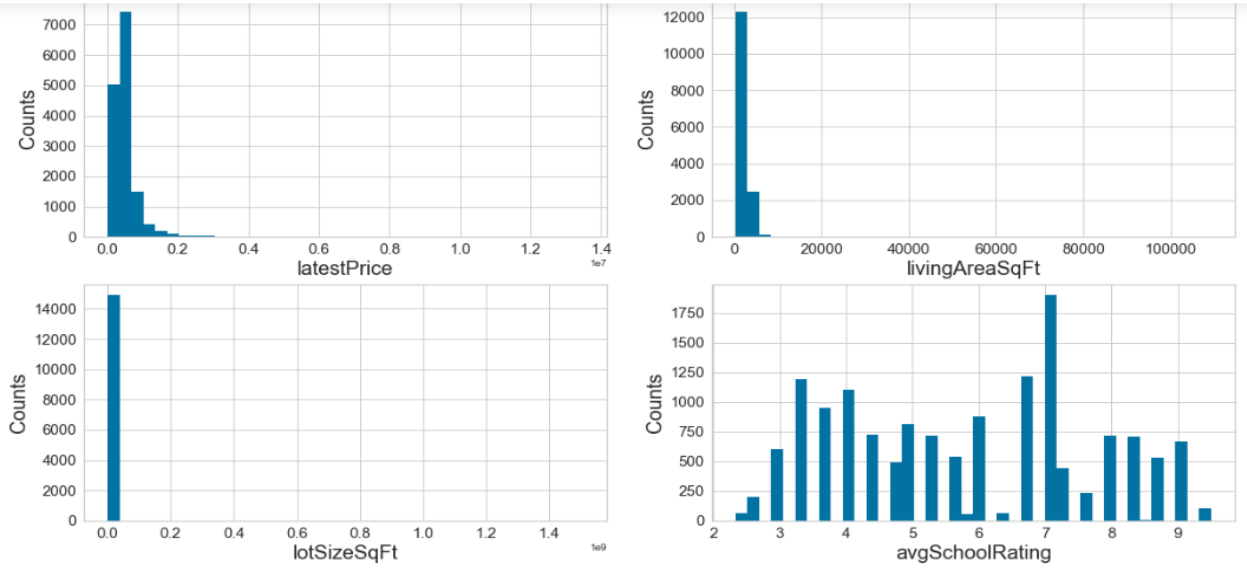
- Looking at the data, some of the columns/features present in the dataset may not be useful for the model. Those columns can be dropped from the dataset
- There are outliers present across most of the columns. Needs to be filtered out during data preparation step
- Only few columns present in the dataset are dimensions. Rest of the columns present in the dataset are measures with datatypes as int and float
- Both continuous and categorical variables are present in the dataset
- The target for the model will be “latestPrice” field present in the dataset

Visualization

Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. As a first step, histogram charts have been created for the following variables present in the dataset

1. latestprice - Latest price of the house
2. livingAreaSqFt - Living area in sq.ft
3. lotSizeSqFt - Total lot size in sq.ft
4. avgSchoolRating - Average school rating



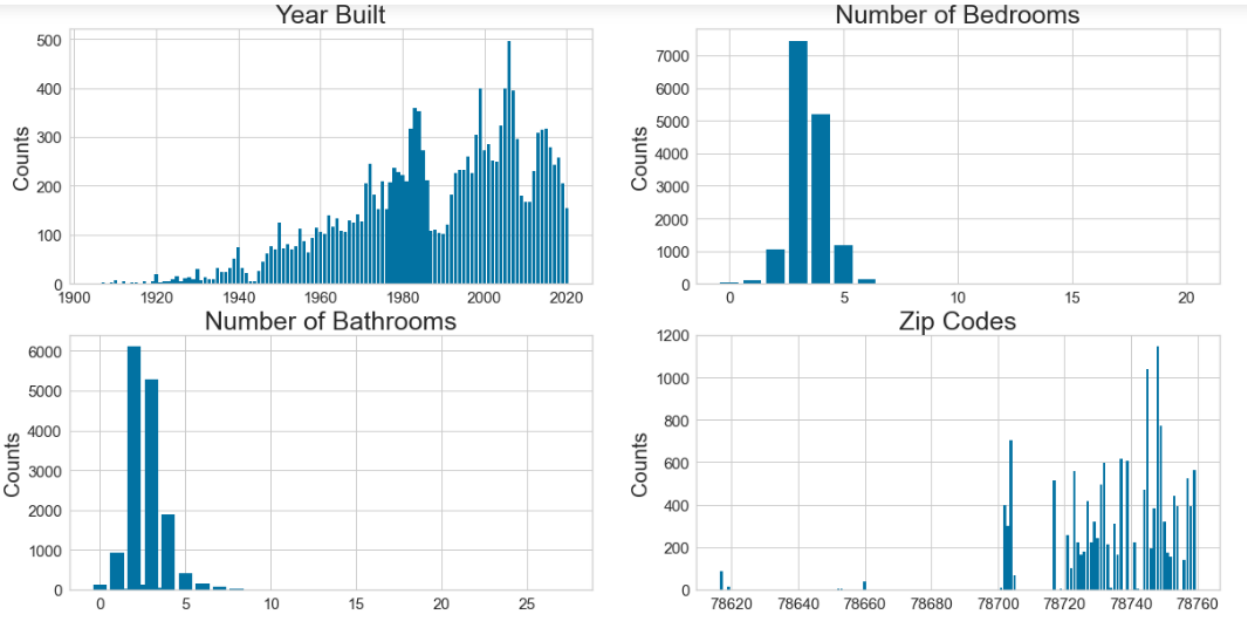
Observation

1. All the histogram charts depict the variables (home price, living square ft, total lot size square ft) except avgSchoolRating are skewed to the right which means the peak of graph lies to the left side of the center. On the right side of the graph, the frequencies of observations are lower than the frequencies of observations to the left side.
2. The first three histogram charts show the presence of outliers in those variables which caused the histograms to be right-skewed distribution.
3. The histogram for average school rating shows the type as multimode distribution as multiple peaks exist in the chart. The frequency of school rating is high at 7 and least at 2 which makes complete sense

Bar Chart

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. We will create bar chart for some of the below mentioned categorical values.

1. HomeType - Type of the home
2. yearBuilt - Year built
3. numOfBedrooms - Number of bed rooms present in the house
4. numOfBathrooms - Number of bathrooms present in the house
5. zipcode - Zip codes



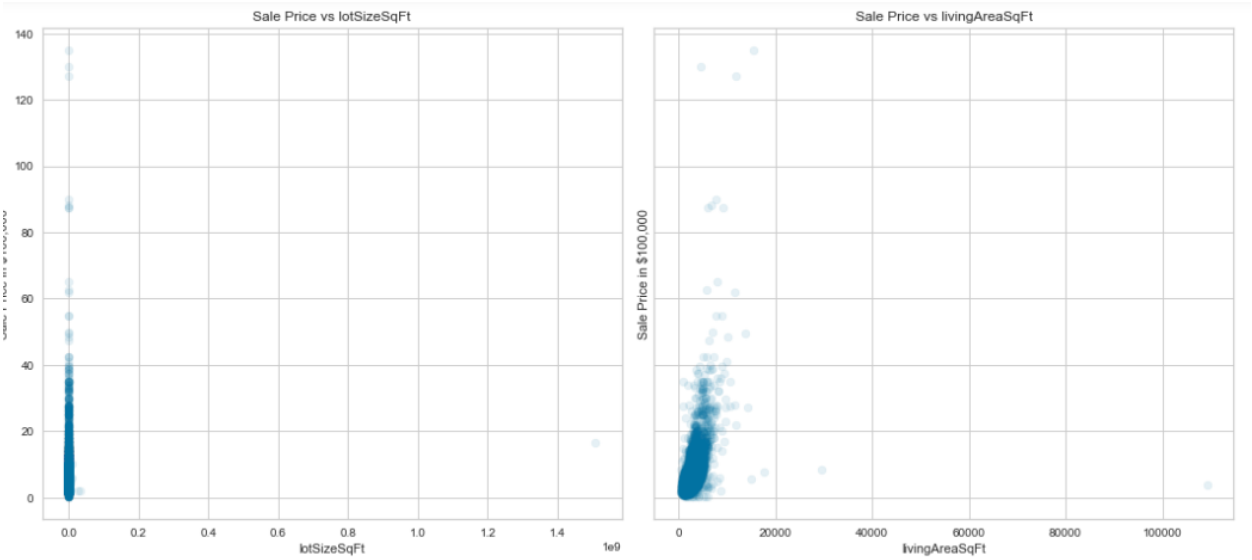
Observation

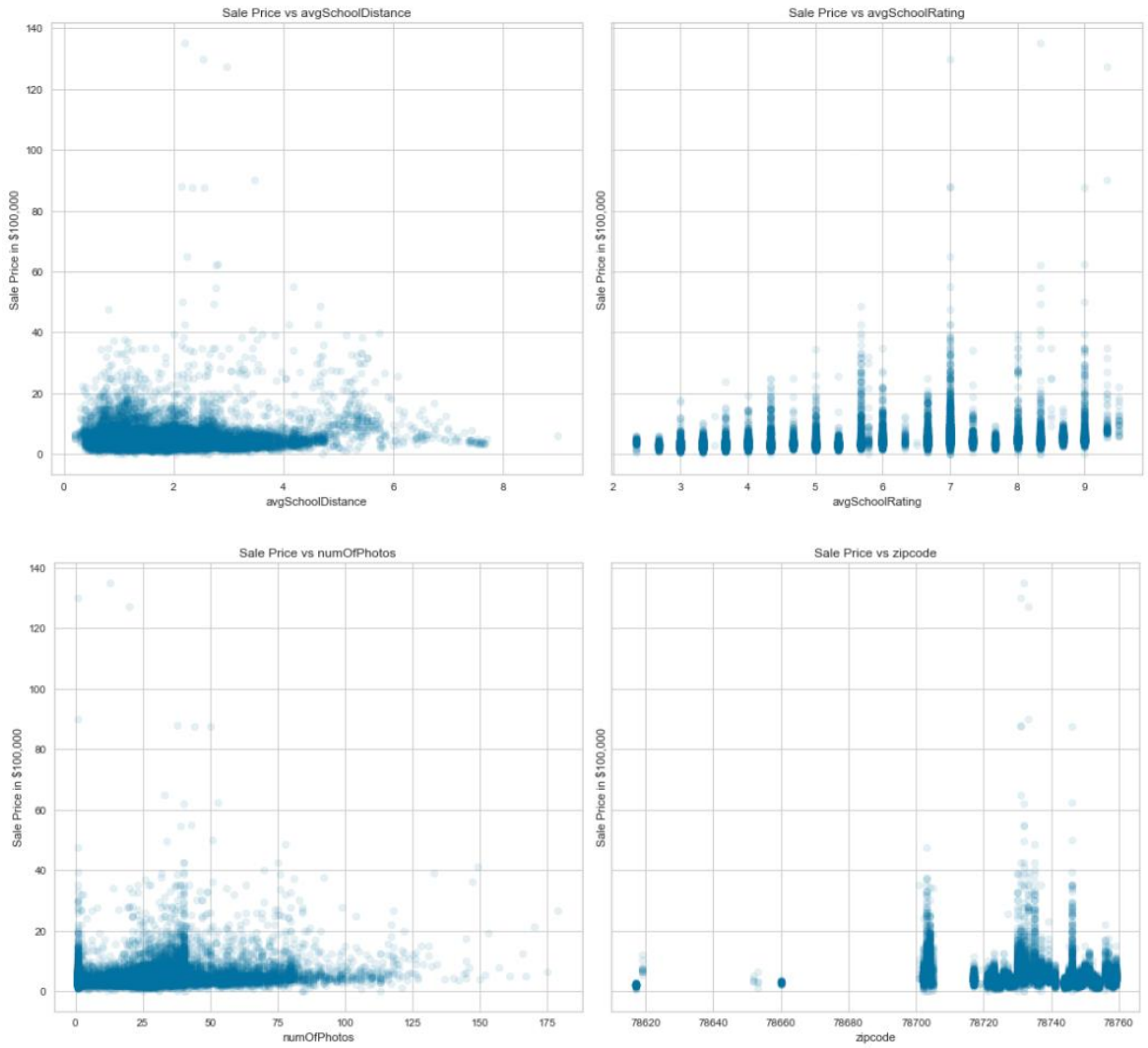
1. From Year Built bar chart, the houses built after 2000 are sold more compared to the houses built before 2000. The number of houses built around 80's also sold more. We will identify the reason while running the model.
2. The number of houses with bedrooms and bathrooms count as 2 and 3 were sold more compare to the houses with different numbers.
3. The houses present in zip codes greater than 78740 sold more than the houses present in other zip codes. There may be n number of reason. We will figure out during modeling training.

Scatter Plot

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. We will plot scatter chart for the below variable vs price

1. lotSizeSqFt
2. livingAreaSqFt
3. avgSchoolDistance
4. avgSchoolRating
5. numOfPhotos
6. zipcode



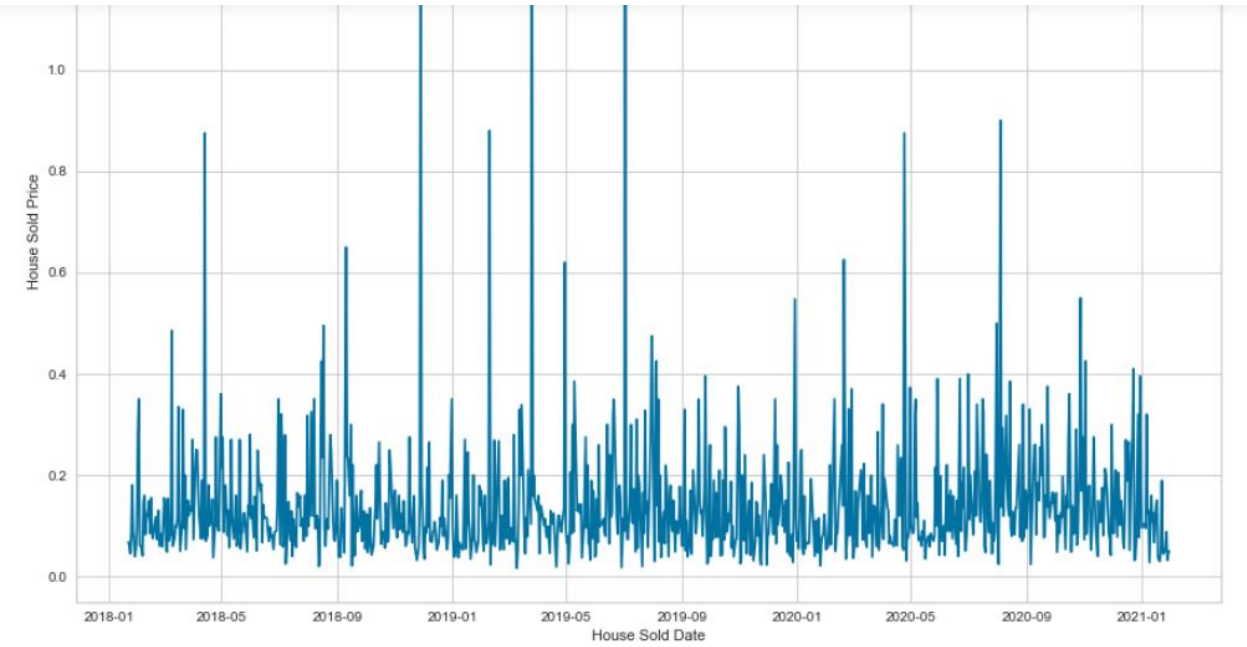


Observation

1. The first two charts which are plotted between the price of the house and total living and lot size in square ft tells increase in price of the house for increase in living and lot size in square ft. This makes complete sense as bigger house sold for high prices compared to small houses.
2. The second scatter chart which is plotted between the price of the house and school distance depicts the houses nearer to school are sold more compared to the houses which are far away from the house. However, the sale price is evenly distributed across all the distances.
3. The fourth chart which is plotted between price and avg school rating shows the houses having good school rating are sold more compared to the houses having poor ratings.
4. The fifth chart is to find if number of photos shown during the listing plays any role in the price of the house. Based on the chart, it is not significant feature in relation to the price.

5. Final chart which is plotted between the price and zip code shows the zip code plays a major role in price of the house. The houses present in some of the zip codes were sold for more price compared to te houses present in other zip codes

Time Series Chart



Observation

1. I tried to plot the time series chart between the sold price and date to understand the increase/decrease in house price over the time frame of 3 years.
2. However, the above chart shows there is a fluctuation in the house price over the time; As mentioned earlier, there are some outliers present in the data which might also be the reason for the fluctuation
3. The chart has been plotted by taking the maximum price sold for each date across various zip codes. If we filter the dataset for particular zip code and try to plot, then it would give clear picture of price flow over the time frame.

Data preparation

Filter

- Home Type: Looking at the percentage of number of records present for each home type, Single Family, Condo and Townhouse make up most of the data. So, other hometypes can be ignored from the dataset

```
Single Family            0.938699
Condo                    0.030980
Townhouse                0.011469
Multiple Occupancy       0.006328
Vacant Land              0.005471
Apartment                0.002439
Residential              0.002439
Mobile / Manufactured    0.001121
MultiFamily              0.000659
Other                    0.000395
Name: homeType, dtype: float64
```

- Lot size: There are records with total lot size less than 500. This doesn't make sense for Single Family, Condo and Townhouse home types. Considered only the houses with lot size greater than 500

Duplicate Check

The latitude and longitude fields are unique for each home. So, performed the duplicate checked based on these 2 fields and found no duplicates present in data.

Outlier Deduction

- Number of bedroom and bathrooms: There is couple of records with number of bedrooms as 20 and number of bathrooms as 27 which doesn't make sense. So, filter these records from dataset.
- Garage space: The data present for Garage space seems to be incorrect. So, the same has been force changes to numbers that make sense as below.
  Price < 1000000 then Garage Space = 3
  Price > 1000000 then Garage Space = 4
- Living Area and lot size square ft: For the square footage variables, I ultimately concluded that extremely large houses and lots are so seriously under-represented in the dataset that we won't be able to reliably predict on them anyway and they are better left off. So, I opt to remove via IQR on these items.

Dropping unwanted features

Out of 47 features present in the dataset, I feel below features doesn't make sense to be present in the dataframe and can be removed.

- zpid: Identifier field
- streetAddress: Address can be ignored as we considered zip code and city
- description: Description of the house; This doesn't make sense.
- hasSpa: I feel spa feature is least feature while considering the house purchase
- latest_salemonth: Sale date is present in the dataset; So, this can be ignored
- latest_saleyear: Sales date is present in the dataset; So, this can be ignored

- MedianStudentsPerTeacher: School rating is significant feature compared to MedianStudentsPerTeacher. So, this feature can be ignored.
- homeImage: This field contains image of the house. This can be ignored as this doesn't make sense.
- numOfWindowFeatures: I feel this is leaset feature while considering house
- numOfWaterfrontFeatures: I feel this is leaset feature while considering house
- numOfAccessibilityFeatures: I feel this is leaset feature while considering house
- latestPriceSource: I believe most of the people wouldn't look for the source of listing
- hasView: I believe this feature also contributes least to house price and can be ignored

Missing data

The dataset doesn't have any null values for any of the fields. So, there is no need for any transformation to populate the null values present in the dataset.

Transformation

Boolean Conversion

Boolean values (True and False) are present for the below list of features. So, these values need to be converted to values 1 and 0 for model training.

- hasAssociation
- hasCooling
- hasGarage
- hasHeating

Addition of feature

Price per square foot feature is not present in the dataset. I believe it would be useful information to include in the dataset to filter the outlier present in the dataset. This is calculated by latestPrice or sold price divided by livingAreaSqft (latestPrice/livingAreaSqFt)

Dummy variable creation

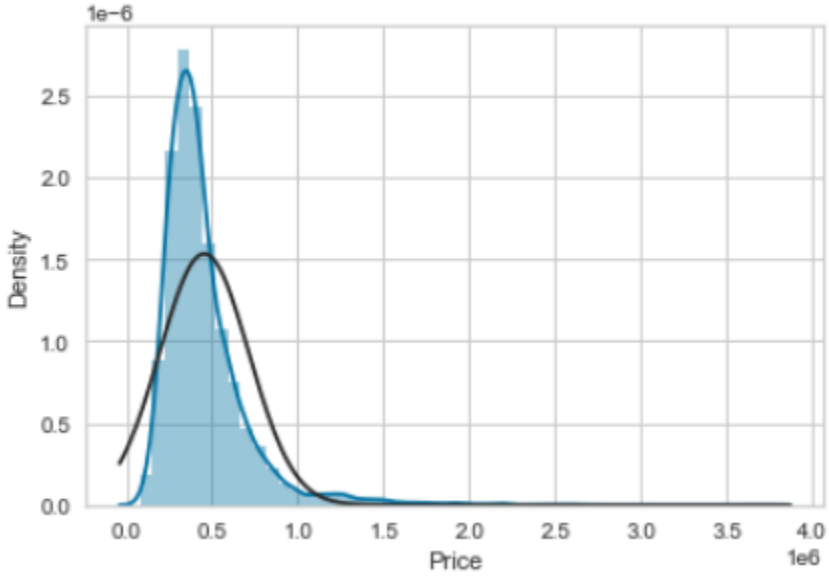
The below features have lot of values present in the dataset. Creating dummy variables for these features makes sense for model training.

- Zipcode
- City
- homeType

Model building and evaluation

Target Analysis

By plotting histogram and normal probability plot, I could see the target variable home price (**Price**) is highly right skewed.



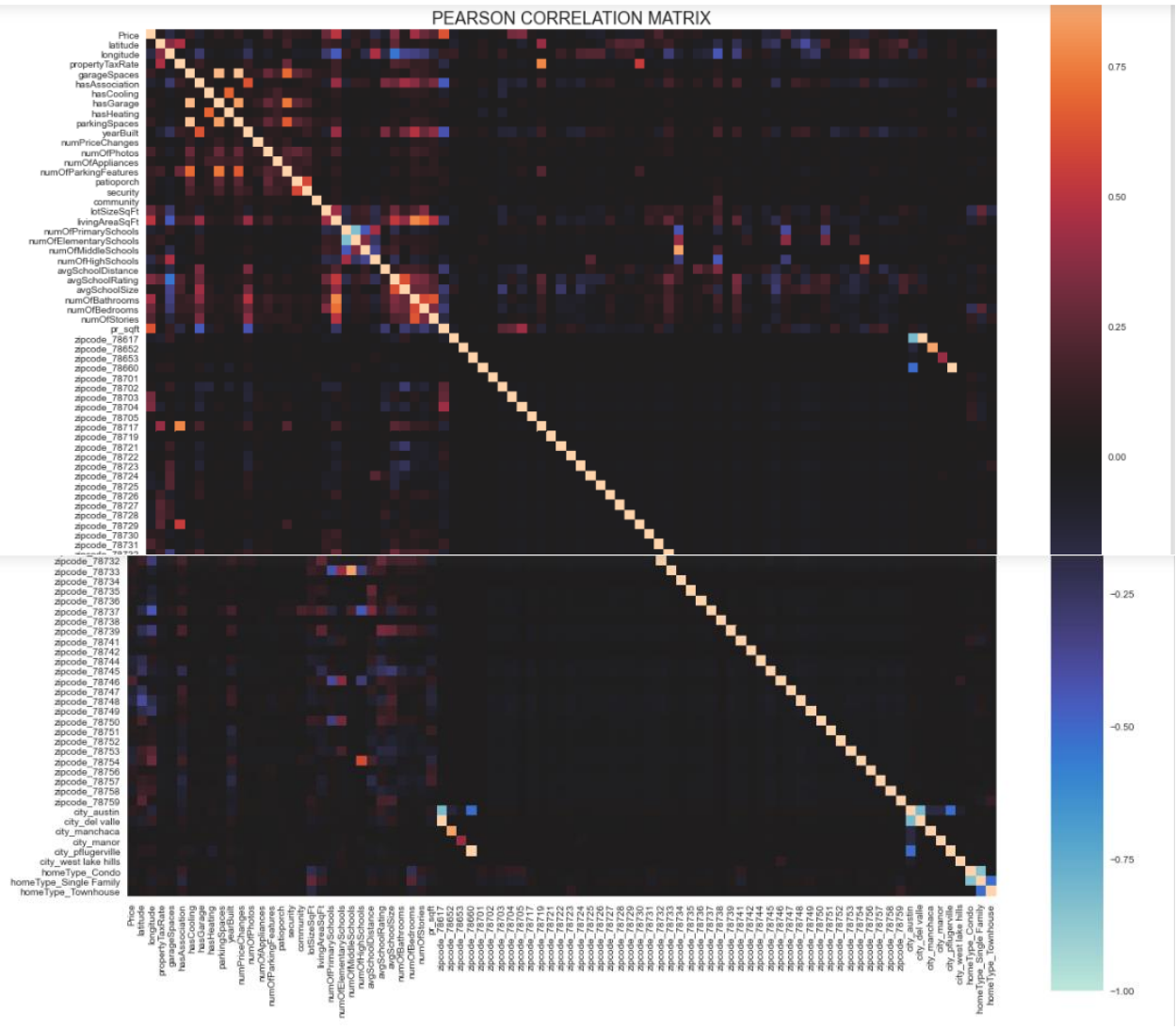
In addition, I calculated skewness and kurtosis for the target variable “Price” and the values are as follows.

```
Skewness: 3.042713
Kurtosis: 16.371716
```

The very positive kurtosis value indicates lots in the tails.

Correlation

Initially, Pearson correlation matrix has been calculated across all the features available in the dataset and the figure is as given below.



In addition, correlation matrix is calculated for the target variable “Price” against all other features available in the dataset and below is the observation.

Highly postively correlation: The below features are having high positive correlation with Sale Price. This means for increase in the value for below feature would increase the price of house.

1. Latitude
2. Avgerage School Rating
3. Number of Bathrooms
4. Number of Bedrooms
5. Price per square foot
6. Number of parking features
7. Living Area Suqare ft
8. Lot Size Square Ft

In addition, I have also noticed that among numerous zip codes present in Austin city, the below zip codes are having postive correlation with Price of the house. So, the price of the house is high in these zip codes compared to other zip codes present in the Austin City.

1. 78703
2. 78704

Among various cities present in the dataset, I see Austin is having high correlation with Price compared to other cities. This tells that price of house is somewhat high in Austin city compared to other cities.

Linearity check

Linearity check has been performed between sales price and the continuous variables and observed the following result.

1. **Price vs Year Built**: From the chart we could see chart looks like polynomial. For increase in year, there is a decrease in price of the house. The older houses sold for more price than the houses built in late 90s and beginning of 20s. This doesn't make sense. However, for the last few years, there is an increase in price for newer houses which makes sense.
2. **Price vs Lot Size Sq ft and Living Area Sq ft**: This chart looks somewhat linear. This makes complete senses as the price of house increases for increase in lot size and living area per square foot.
3. **Price vs Average School Distance**: The chart looks somewhat straight. The houses sold for more prices for the schools which are 5 to 6 miles away from the houses. However, the majority of the houses which are more than 6 miles away and less than 4 miles away from schools shows straight line, though few houses sold for more prices.
4. **Price vs average School rating**: This chart looks linear. This makes complete sense as the houses located in good school districts sold for more price compared to the houses which have poor school ratings.
5. **Price vs Number of Bedrooms and Bathrooms**: These charts are also look linear. The houses having a greater number of bedrooms and bathrooms sold for more price compared to the houses having a smaller number of bedrooms and bathrooms.
6. **Price vs Price per Sq foot**: This chart looks linear. This is somewhat confusing. Usually, the price per square foot will be less for bigger houses and more for smaller houses. However, this chart shows that price per square foot is high for the houses sold for more price compared to the houses sold for less price.
7. **Price vs Number of Price Changes**: This chart looks somewhat linear. There are multiple scenarios available for this comparison. There may be multiple reduction in house prices due to number of offers provided to the house. At the same time, the house price would have gotten increased due to number of offers to the house. We could see the same from the chart.
8. **Price vs Number of Appliance**: This chart will be mostly linear. For increase in number of appliances, the house price will also increase. This makes sense as price of the house and number of appliances have positive correlation.

Box Plot

Box plot is drawn for all the categorical variables. From all the box charts plotted, I could see the interquartile range values for all categorical variables lies around 500K of house price. This is expected as 250K to 500K house price would be the range for median price of the house sold and most of the variables lies in this range.

Linear Regression

Linear regression is commonly used for predictive analysis and modeling. Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Initially, I have chosen only continuous variables/features for my model and got the score as 86.86%. Below is the finding.

1. The low MAE value of 55452.67 indicates that the forecast value is nearer to the true values
2. The low RMSE of 8710.42 means that the predictions are not that far away from the values (mpg range from 77500-3750000), and the predictions are relatively accurate

Then, I run the linear regression model by including categorical variables.

Interaction

In regression, an interaction effect exists when the effect of an independent variable on a dependent variable change, depending on the value(s) of one or more other independent variables.

On running linear regression model on interaction (combination of every 2 features available in the dataset), there is an improvement in the overall score of the model. Though, score has been improved for most of the combination, the score has been reduced for some of the features like number of bedrooms and number of price changes and number of bathrooms and number of appliances.

Conclusion

What does the analysis/model building tell you?

All the continuous variables/features considered for the model building have considerable effect on home price. The R2 value of 0.8687 (86.87%) is somewhat high (0-1 scale) and means that the predictor mostly determines the observed value. These metrics are almost identical to testing set metrics in linear regression. By including categorical variables to the model, I see the score has been increased slightly to 90.5% which tells that all the categorical variables also have effect on home sale price. At the same time, mean absolute error (MAE: 48663.8169) and root mean squared error (RMSE: 78207.0110) have been reduced compared to previous dataset where continuous variables alone included.

Is this model ready to be deployed?

Though this model has several features available in the dataset, some of the external factors are missing in the dataset which are considered to be significant. Some of the features which I feel are significant to be considered are mentioned below.

- Interest rate
- Environment factors like pandemic
- Work from home making people to migrate to low cost of living area

So, model can be further improved by considering these external factors. So, this is not deployment ready.

What are your recommendations?

To build a most robust model, the dataset should include all the features affecting the house price. Since above mentioned external factors are not included, the model prediction might be incomplete. So, it would be nice to consider the dataset having all the features required for the modelling.

What are some of the potential challenges or additional opportunities that still need to be explored?

The major challenge I faced was with data transformation process and make the data ready for the model. There were lot of outliers present in the dataset which took considerable amount of time for the cleanup.

I would say the model can be further improved by adding home price prediction functionality. The “Sale Date” feature is present in the dataset which is not used in the model building. This feature can be used to predict the house price over the time which would be significant feature to this model.

References:

- Python Machine Learning Cookbook - Chris Albon
- Kaggle website