Predicting the Nightly Price of Denver's Airbnb in 2023:

Steps:

1. **Introduction and Problem Identification**
2. **Data Wrangling**
3. **Exploratory Data Analysis**
4. **Preprocessing And Training**
5. **Modeling**

1. Introduction and Problem Identification

Travel is one of the world's largest industries, and its approach has become commoditized. The travel industry has scaled by offering standardized accommodations in crowded hotel districts and frequently-visited landmarks and attractions. This one-size-fits-all approach has limited how much of the world a person can access, and as a result, guests are often left feeling like outsiders in the places they visit. Airbnb can help you with them.

In Denver, we are going to find a good way to predict the final price from our data like review scores, the number of bedrooms, and bathrooms. We should know how they affect to the final price.

Your property is never going to be expensive enough to satisfy your wishes, but it is never going to be cheap enough to satisfy a guest's wishes. The most common issue we hear is *'If I charge any more than $XX per night I won't get any bookings.'*

How can that data affect to the price? Or do they affect to the price at all?

We want to find the best model for predicting the price for Airbnb company.

- The purpose of the analysis: understanding the factors that influence Airbnb prices in Denver, or identifying patterns of all variables and Our analysis provides useful information for travelers and hosts in the city and also provides some best insights for Airbnb business.

- the factors that influence Airbnb prices in Denver, or identifying patterns of all variables and Our analysis provides useful information for travelers and hosts in the city and also provides some best insights for Airbnb business.

- This project involved exploring and cleaning a dataset to prepare it for analysis. The data exploration process involved identifying and understanding the characteristics of the data, such as the data types, missing values, and distributions of values. The data cleaning process involved identifying and addressing any issues or inconsistencies in the data, such as errors, missing values, or duplicate records and removing outliers.

- Through this process, we were able to identify and fix any issues with the data, and ensure that it was ready for further analysis. This is an important step in any data analysis project, as it allows us to work with high-quality data and avoid any potential biases or errors that could affect the results. The clean and prepared data can now be used to answer specific research.

- Once the data has been cleaned and prepared, now begin exploring and summarizing it with describing the data and creating visualizations, and identifying patterns and trends in the data. in exploring the data, may develop the relationships between different variables or the underlying causes of certain patterns or trends and other methods.

- using data visualization to explore and understand patterns in Airbnb data. We created various graphs and charts to visualize the data, and wrote observations and insights below each one to help us better understand the data and identify useful insights and patterns.

- Through this process, we were able to uncover trends and relationships in the data that would have been difficult to identify through raw data alone, for example factors affecting prices and availability. We found that minimum nights, number of reviews, and host listing count are important for determining prices, and that availability varies significantly across neighborhoods. Our analysis provides useful information for travelers and hosts in the city.

- The observations and insights we identified through this process will be useful for future analysis and decision-making related to Airbnb. And also, our analysis provides useful information for travelers and hosts in the city.

Source:

http://insideairbnb.com/get-the-data/

2. Data Wrangling Contents

3. Data Wrangling

In this section, I try to find the data that I need and I try to have the summary of the data. After that I clean and find the distributions of the columns that I want to select as features to predict the price. Finally, I will make the summary of the data. I won't bring all of visualization here. I select the effective of them here.

The details are listed below:

1. Find the Data that I Need Categorical Features from Raw Data:

   I select these columns bellow at first.
   'id', 'description', 'host id', 'area', 'latitude', 'longitude', 'property type', 'room type', 'bathrooms', 'bedrooms', 'number of reviews', 'last scraped', 'review scores rating', 'Review scores accuracy', 'review scores cleanliness', 'review scores check in', 'Review scores communication', 'review scores location', 'review scores value', 'Reviews per month', 'price'.

2. Number Of Missing Values by Column

   Half of the raw data approximately had missing values. I counted the number of missing values and percentages of them in each column and sorted them.
   The table below shows the table that I can find missing values.

	count	%
Reviews per month	757	15.228324
Review scores rating	756	15.208208
Review scores value	756	15.208208
Review score's location	756	15.208208
Review scores communication	756	15.208208
Review scores check in	756	15.208208

	count	%
Review scores cleanliness	756	15.208208
Review scores accuracy	756	15.208208
price	82	1.649567
bedrooms	65	1.307584
bathrooms	2	0.040233

3. Categorical Features
   Before going for categorical features, I would try to find the types of data for columns. description, area, property type, room type, bathrooms, and last scraped are object in the data frame. I will change those to the correct format. For example, the price column has $ sign and I will drop it and I select that as a number that I can use it later.
   After that I try to clean and organize the columns.

    1. Unique Home Type names

       I select the first word of description column, it's named 'Home Type'. When I compare the 'Home type' and 'property type', they show me similar info.

    2. Area

       The areas in Denver list below:

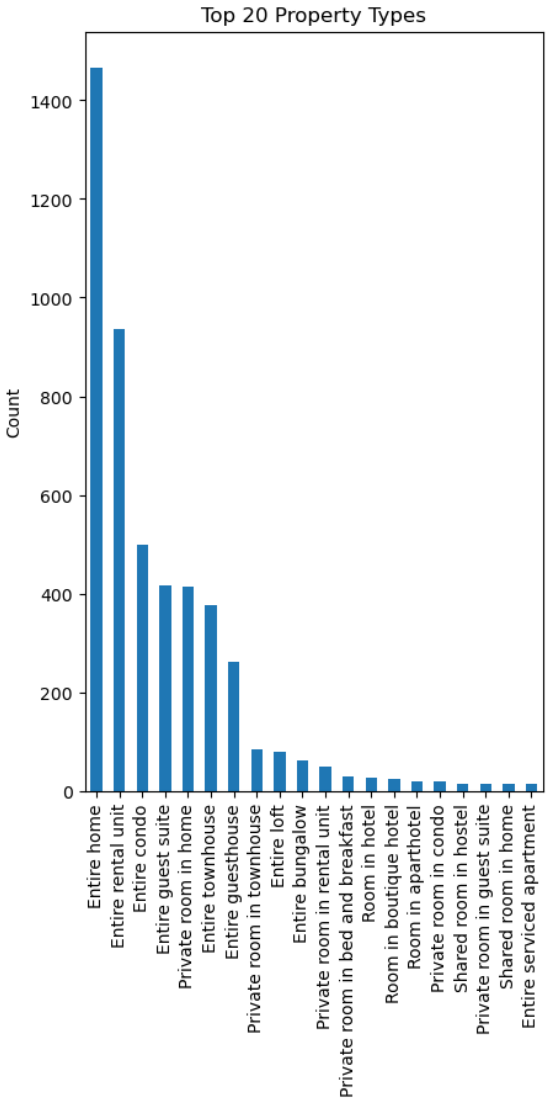
       'North Park Hill', 'Hale', 'Sloan Lake', 'Five Points', 'West Colfax', 'Sunnyside', 'Jefferson Park', 'Chaffee Park', 'West Highland', 'Highland', 'Union Station', 'Cole', 'City Park West', 'Lincoln Park', 'North Capitol Hill', 'CBD', 'East Colfax', 'Barnum', 'Gateway - Green Valley Ranch', 'Capitol Hill', 'Speer', 'Washington Park West', 'Hilltop', 'Whittier', 'Cherry Creek', 'Cheesman Park', 'University', 'Lowry Field', 'Stapleton', 'Cory - Merrill', 'City Park', 'Regis', 'Congress Park', 'Washington Virginia Vale', 'Skyland', 'Civic Center', 'College View - South Platte', 'Elyria Swansea', 'Clayton', 'Mar Lee', 'Ruby Hill', 'Northeast Park Hill', 'Athmar Park', 'Platt Park', 'Baker', 'Villa Park', 'University Hills', 'Montclair', 'Berkeley', 'Virginia Village', 'Montbello', 'Goldsmith', 'Overland', 'Belcarra', 'Country Club', 'Harvey Park', 'Hampden', 'Barnum West', 'Harvey Park South', 'Auraria', 'South Park Hill', 'Washington Park', 'University Park', 'Rosedale', 'Marston', 'Bear Valley', 'Windsor', 'Hampden South', 'DIA', 'Valverde', 'Globe Ville', 'Westwood', 'South moor Park', 'Willshire', 'Sun Valley', 'Indian Creek', 'Fort Logan'

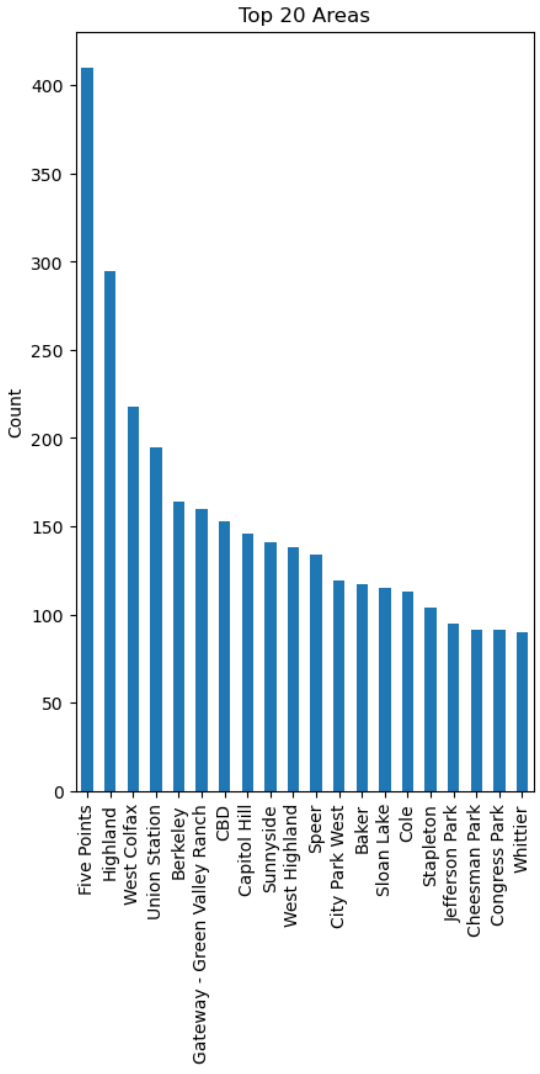
3. Number of distinct regions

   Here we have 77 distinct regions in my data

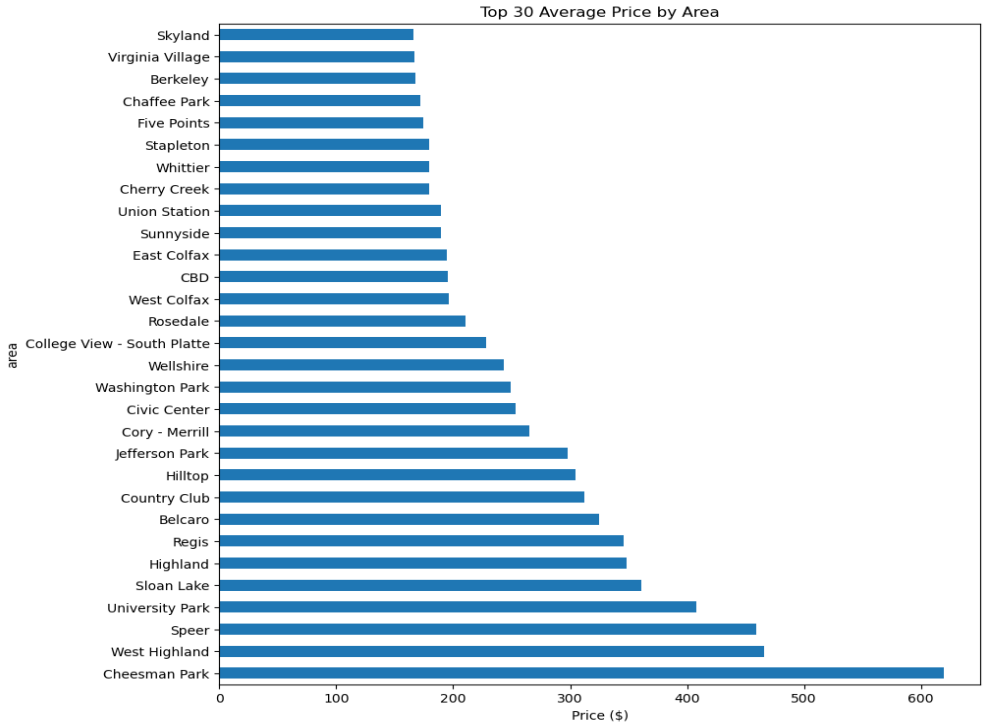
4. Distribution of by top 20 Property Type and Area

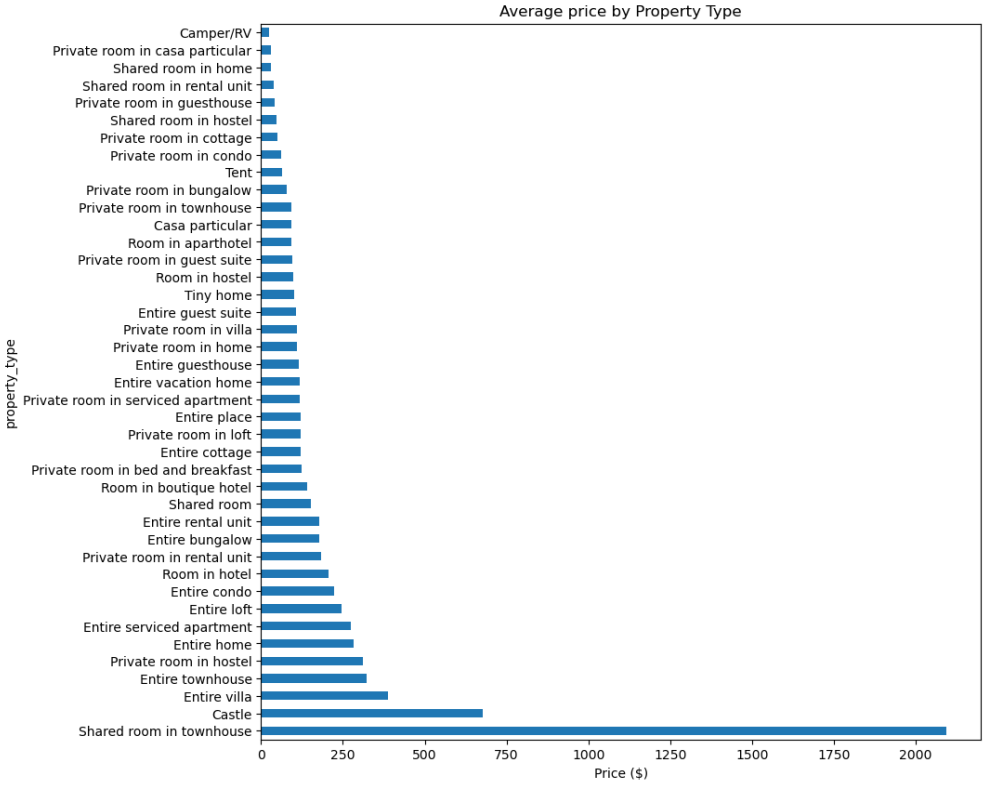
   I select top 20, because the visualization becomes better.



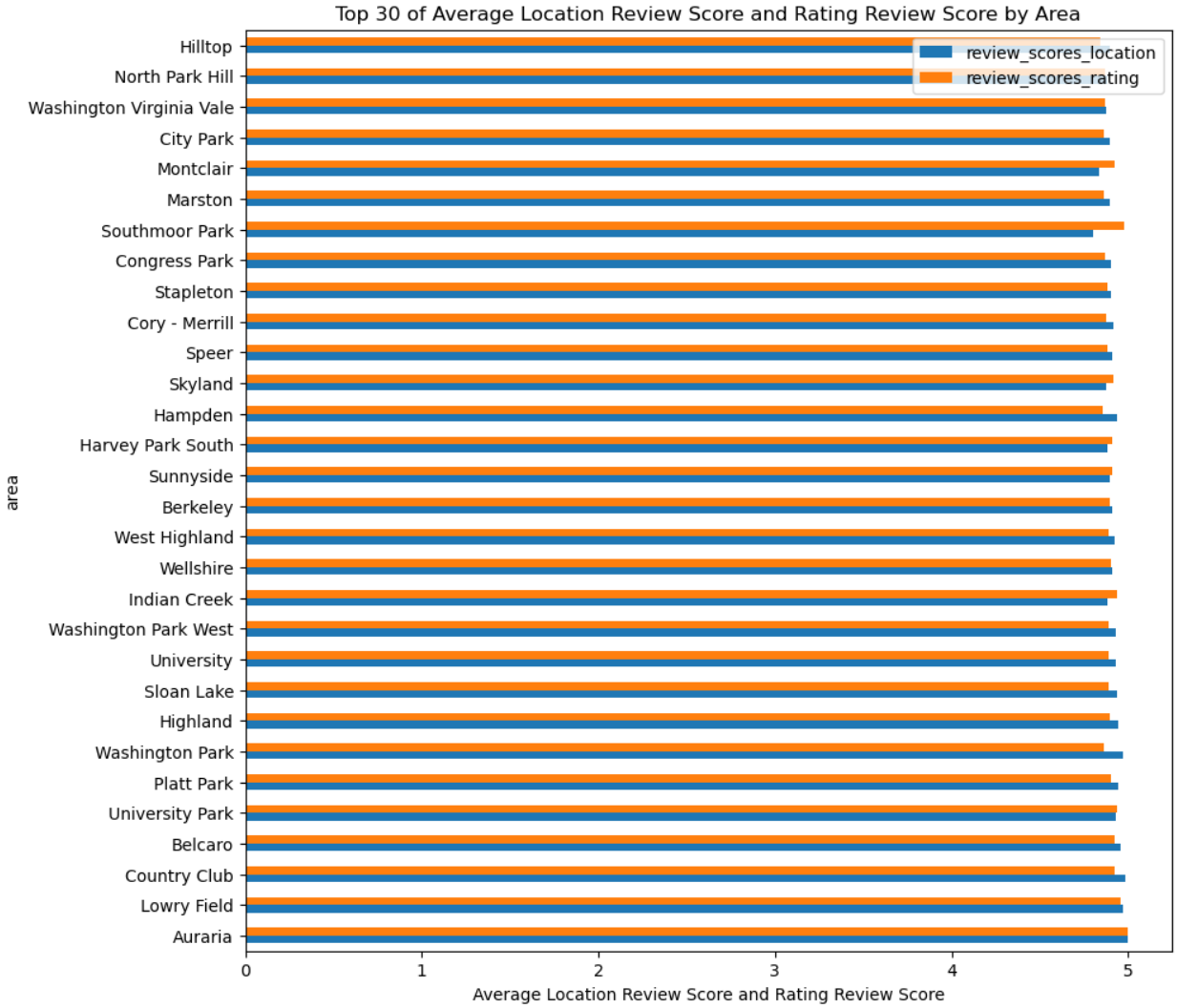


5. Distribution of Average of Price, Review Scores Rating and Review Score Location by Property Type and Area:
   a) Average of Price by Property Type and Area

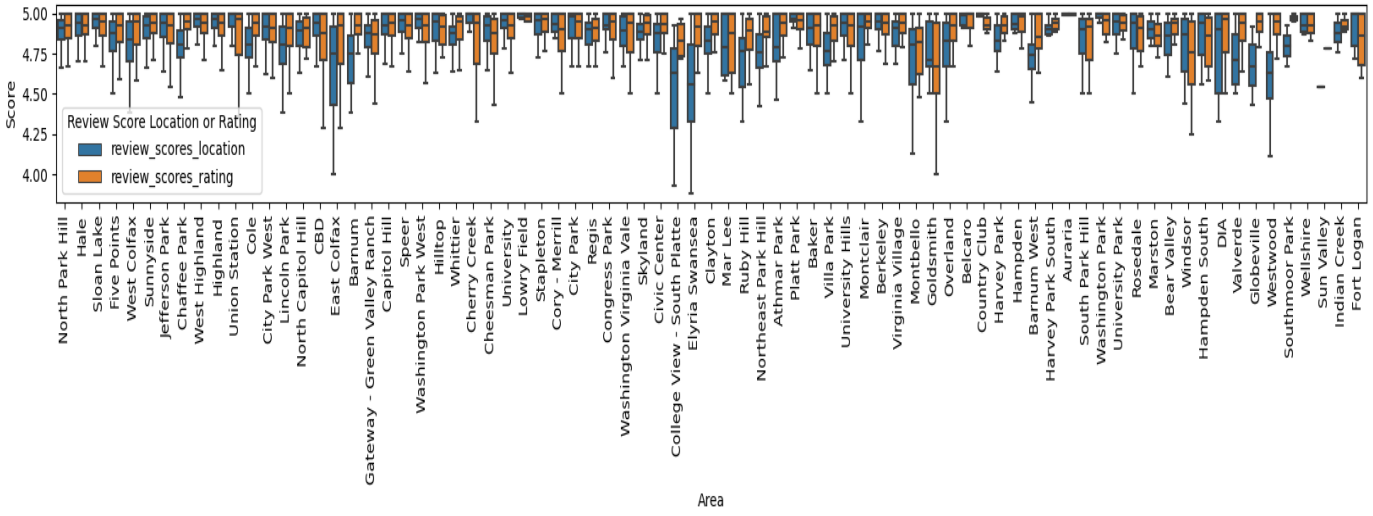

Top 30 Average Price by Area
Skyland
Virginia Village
Berkeley
Chaffee Park
Five Points
Stapleton
Whittier
Cherry Creek
Union Station
Sunnyside
East Colfax
CBD
West Colfax
Rosedale
College View - South Platte
Wellshire
Washington Park
Civic Center
Cory - Merrill
Jefferson Park
Hilltop
Country Club
Belcaro
Regis
Highland
Sloan Lake
University Park
Speer
West Highland
Cheesman Park
area
0
100
200
300
400
500
600
Price ($)


Average price by Property Type
Camper/RV
Private room in casa particular
Shared room in home
Shared room in rental unit
Private room in guesthouse
Shared room in hostel
Private room in cottage
Private room in condo
Tent
Private room in bungalow
Private room in townhouse
Casa particular
Room in aparthotel
Private room in guest suite
Room in hostel
Tiny home
Entire guest suite
Private room in villa
Private room in home
Entire guesthouse
Entire vacation home
Private room in serviced apartment
Entire place
Private room in loft
Entire cottage
Private room in bed and breakfast
Room in boutique hotel
Shared room
Entire rental unit
Entire bungalow
Private room in rental unit
Room in hotel
Entire condo
Entire loft
Entire serviced apartment
Entire home
Private room in hostel
Entire townhouse
Entire villa
Castle
Shared room in townhouse
property_type
0
250
500
750
1000
1250
1500
1750
2000
Price ($)

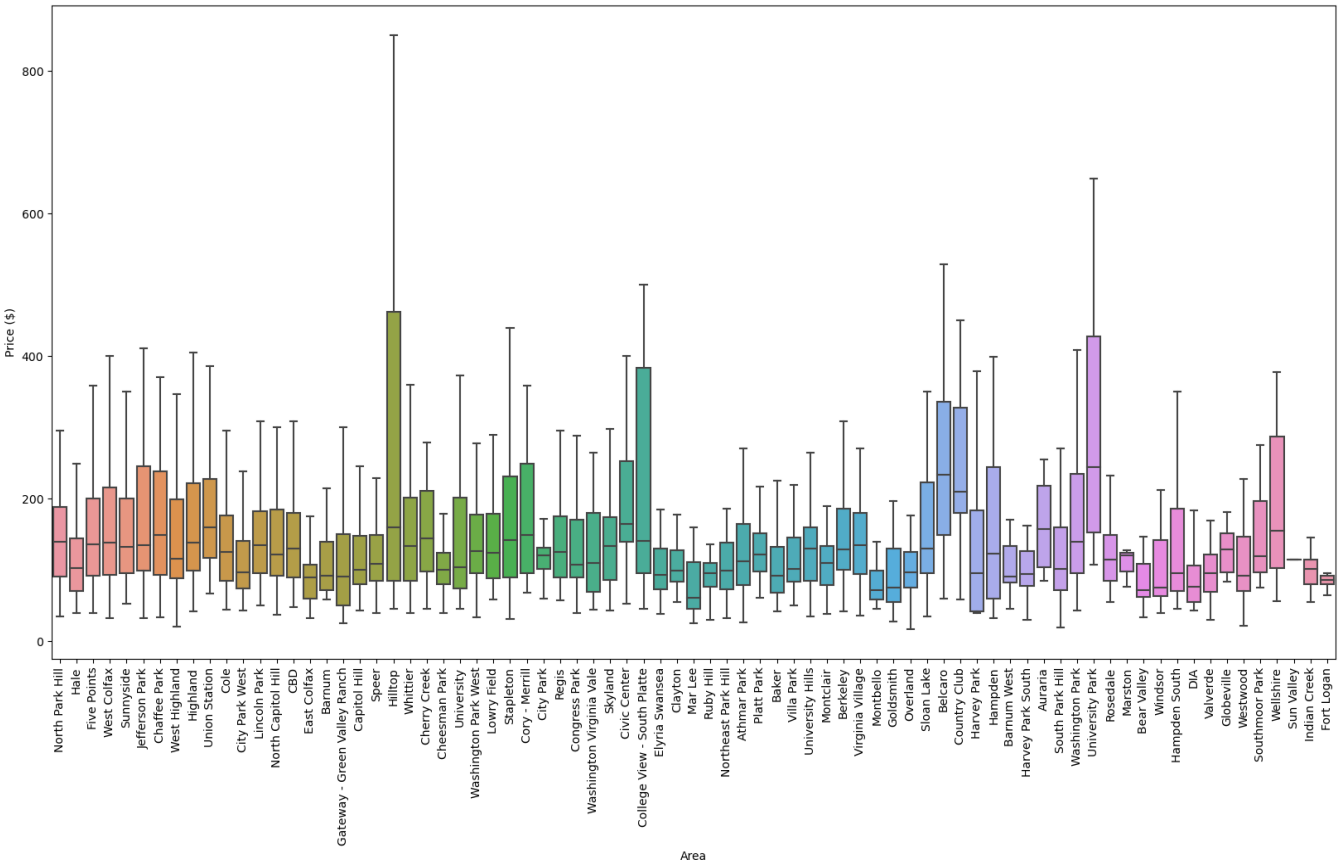
b) Average of Top 30 Review Scores Location and Review Score Rating by Area



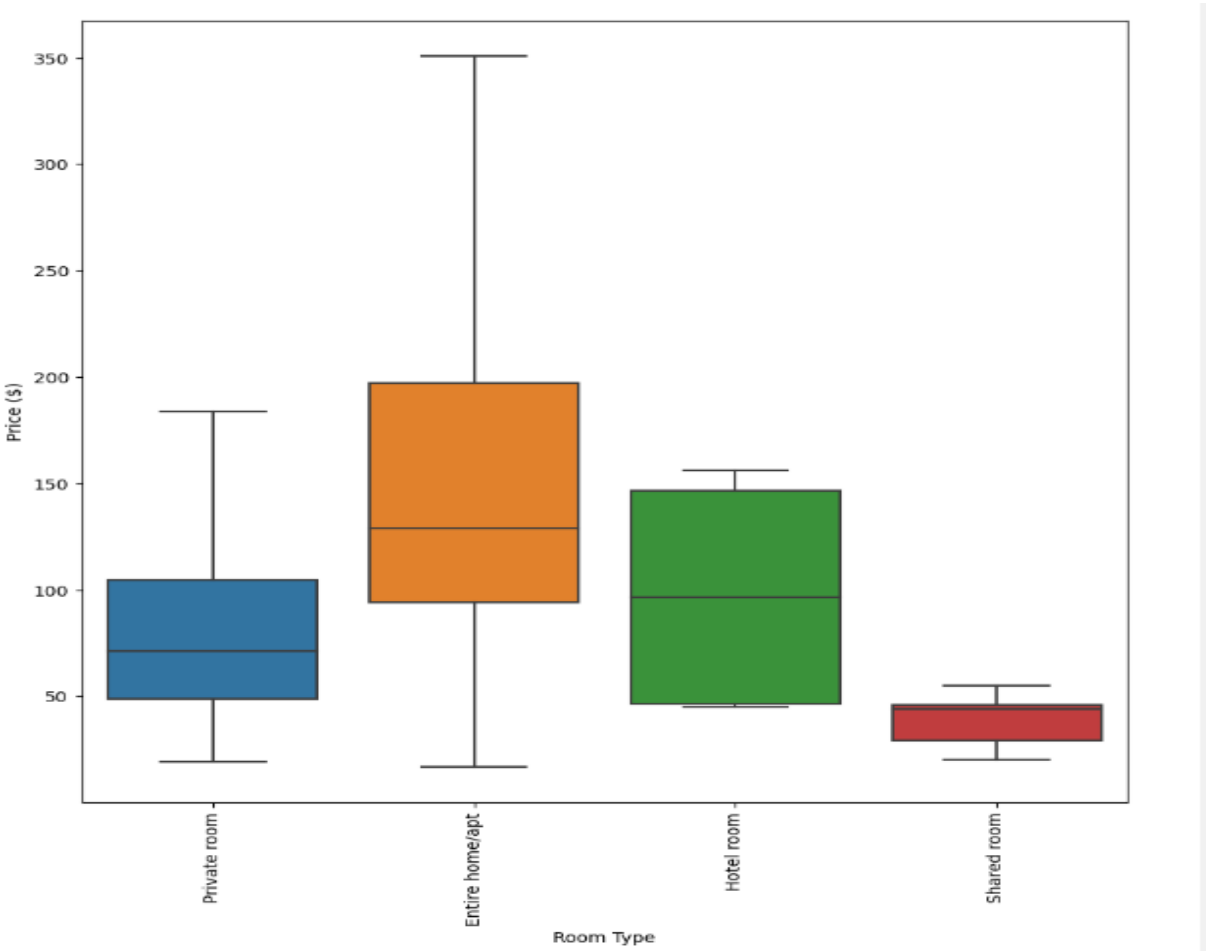
c) Distribution of Review Score Location, Review Score Rating and Price by Area



d) Distribution of Price by Area



e) Distribution of Room Type by Area

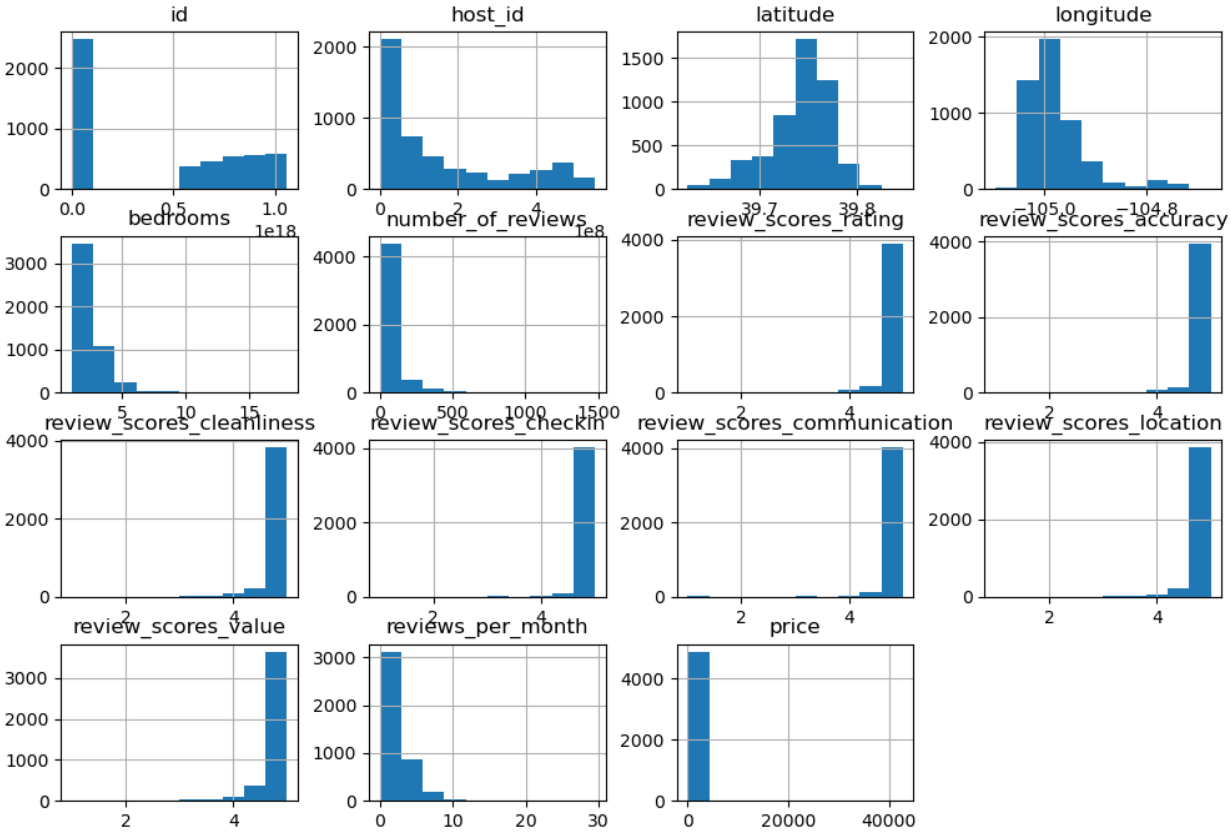


4. Numeric Features

1. Numeric Data Summary

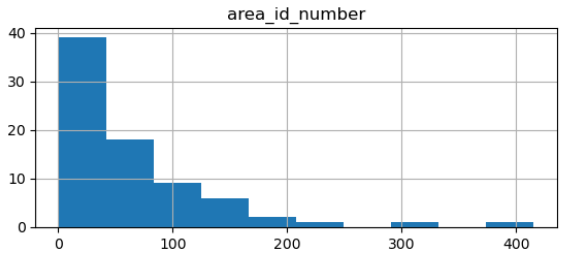
When I used “Describe code” to find out the aggregation of data. And also, I found out less than 2% of price column is missing. I decide to drop them.

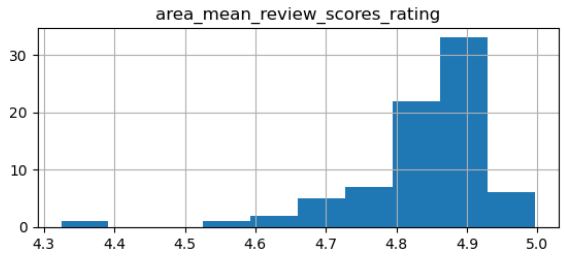
2. Distributions Of Feature Value

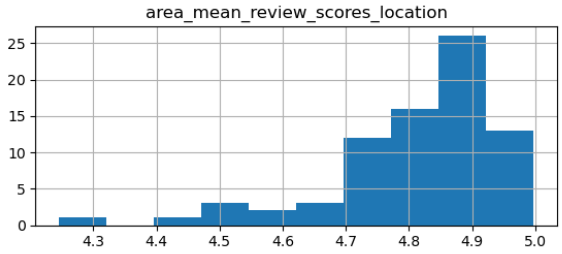


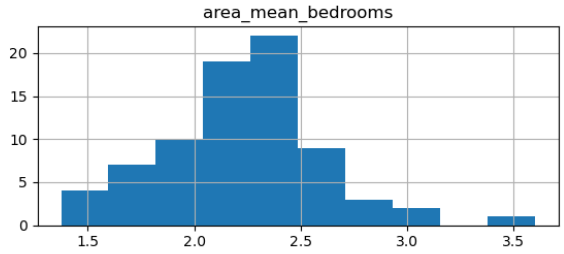
5. Drop Rows with no Price

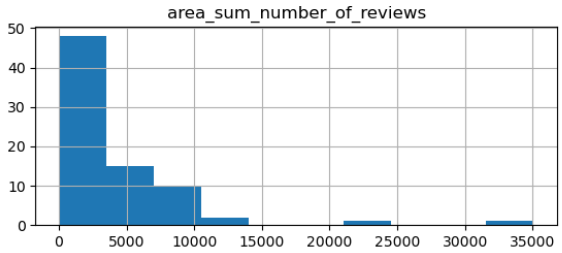
6. Review Summary Distribution

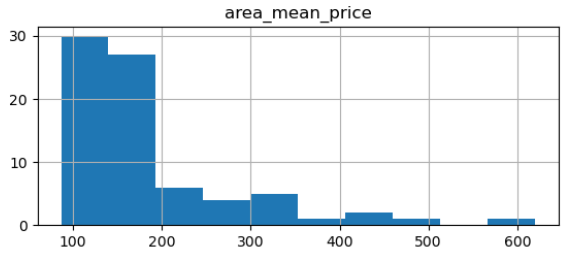












7. Target Feature (Price)
   I drop missing values of price.

Here our data is ready for EDA (Exploratory Data Analysis)

4. Exploratory Data Analysis

1. Top Areas by Order of Each of the Summary Statistics

   I set the area as index. I put mean of data in summary table.

   1) Top areas review score rating

      Auraria, South moor Park, Lowry Field, Globe Ville, University Park

   2) Top areas review score location

      Auraria, Country Club, Washington Park, Lowry Field, Belcarra

   3) Top amount of Id numbers

      Five Points, Highland, West Colfax, Union Station, Gateway / Green Valley Ranch

   4) Total population

      In the Data Wrangling part, I select the population of areas. I get this info from Wikipedia

Montebello	30348
Gateway / Green Valley Ranch	29201
Hampden	17547
Westwood	15486
Capitol Hill	14708

2. Visualizing High Dimensional Data

   One way to disentangle this interconnected web of relationships is via principle components analysis (PCA). This technique will find linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance they explain.

The basic steps in this process are:

1. scale the data (important here because our features are heterogenous)
2. fit the PCA transformation (learn the transformation from the data)
3. apply the transformation to the data to create the derived features
4. (optionally) use the derived features to look for patterns in the data and explore the coefficients

   1) Scale the data

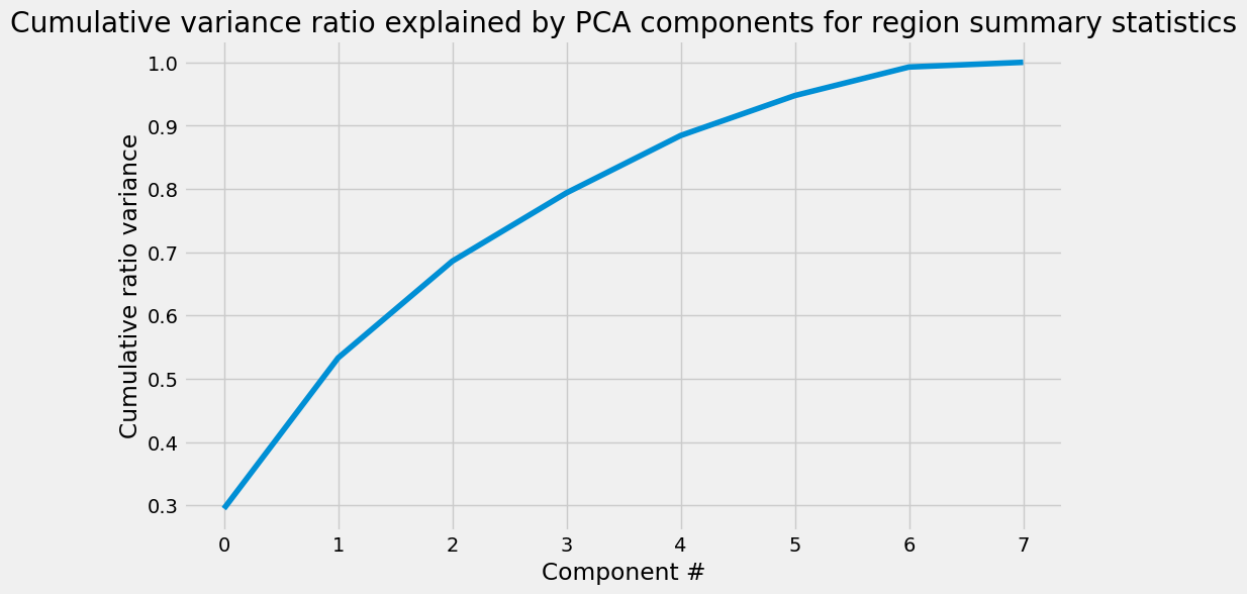
I only want numeric data here, although I don't want to lose track of the region labels, so it's convenient to set the area as the index.
scale () returns arrays. I want to visualize scaled data; I already copied the column names. Now I can construct a data frame from the arrays here and reintroduce the column names.

Verifying the scaling:
First of all, check the mean of the scaled features using panda's mean () Data Frame method.

Perform a similar check for the standard deviation using pandas' std () Data Frame method.

The numbers are close to 1. It's verified now.

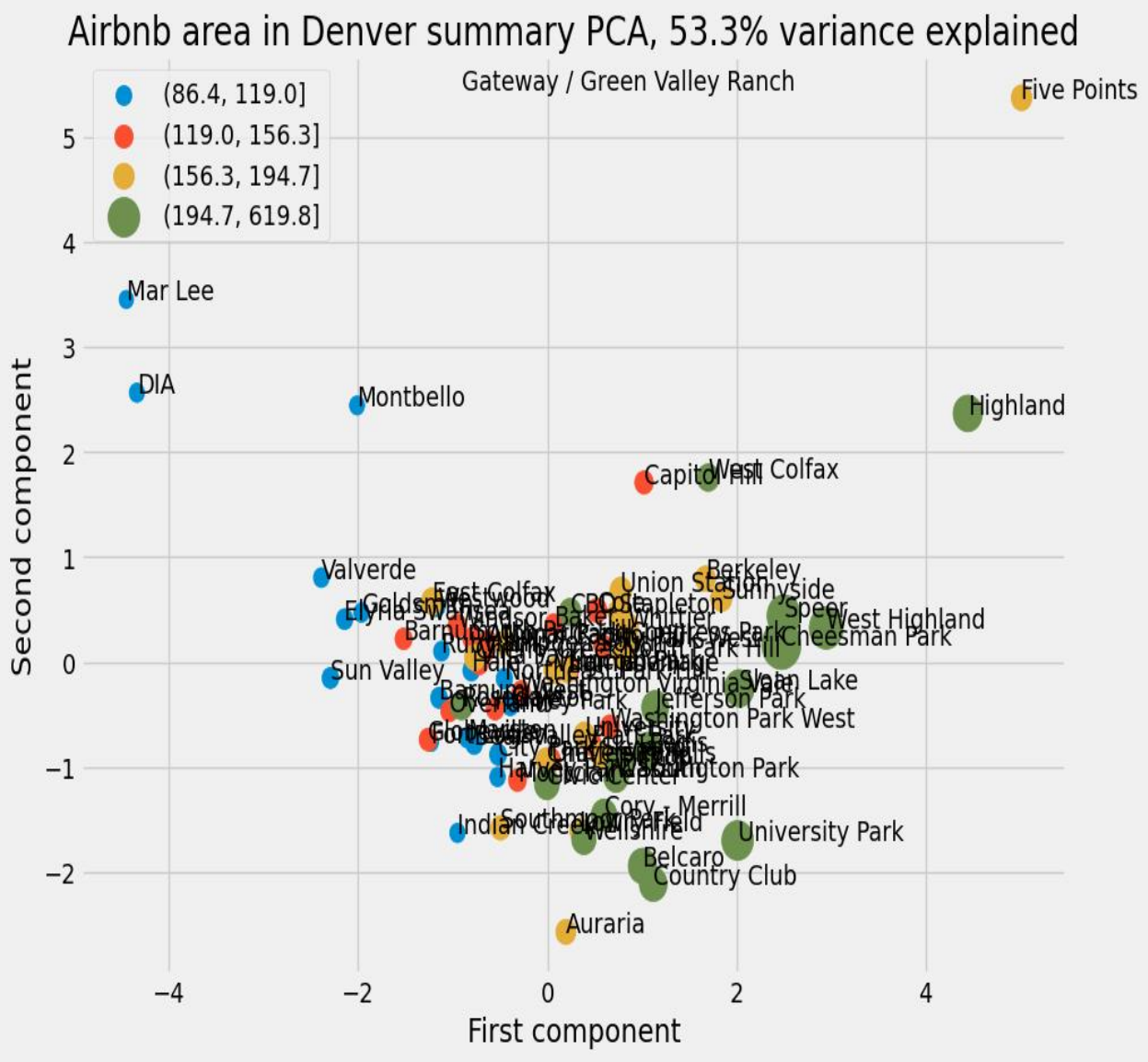
2) Calculate the PCA transformation
   Fit the PCA transformation using the scaled data.
   Plot the cumulative variance ratio with number of components.



The first two components seem to account for over 60% of the variance, and the first four for over 80%.

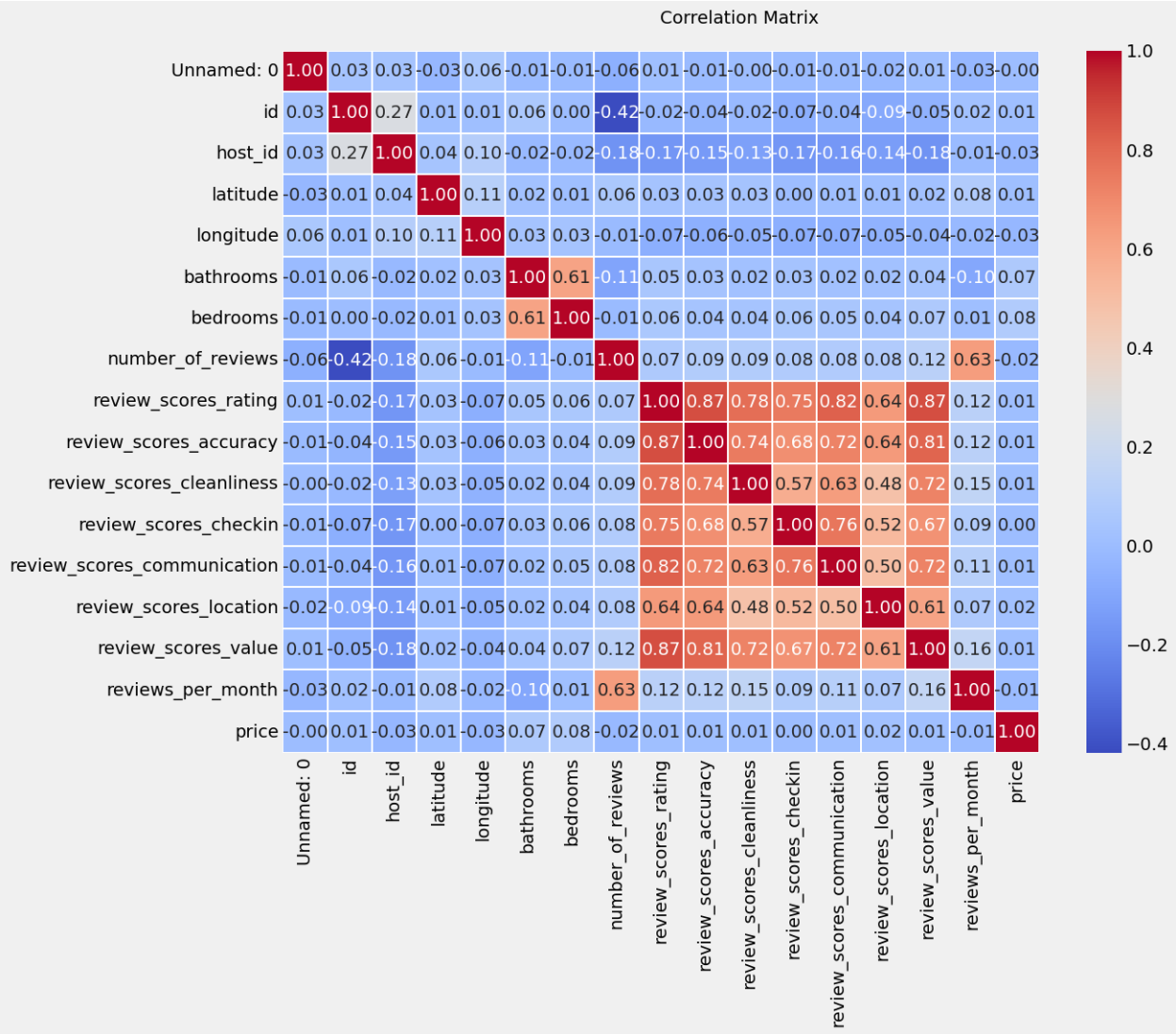
Apply the transformation to the data to obtain the derived features.

a) Adding Average Price and Quartile to Scatter Plot
The green points representing the upper quartile of price can be seen to the down, the right. There's also a spread of the other quartiles as well.



3) Feature Correlation Heatmap

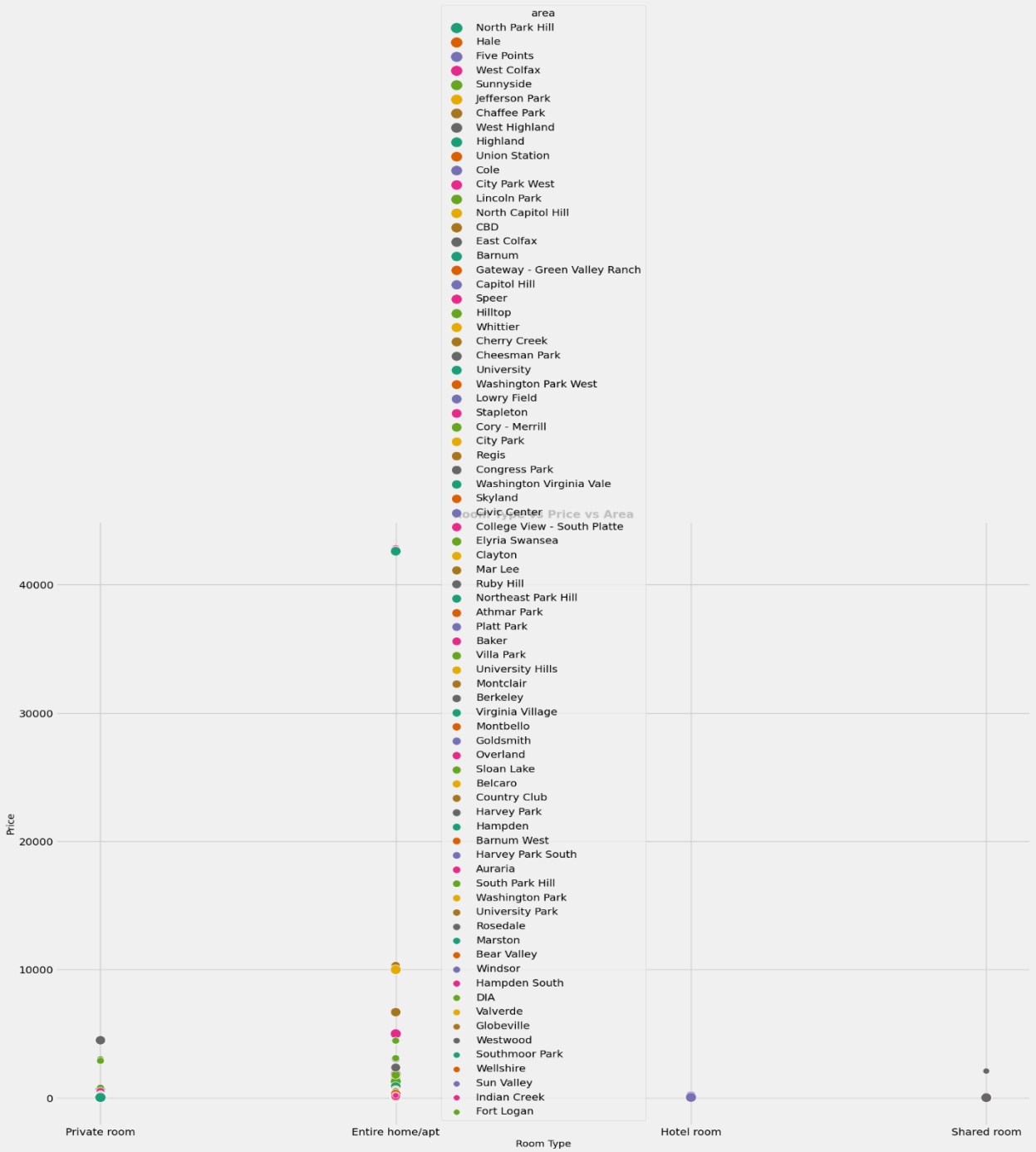
A great way to gain a high-level view of relationships amongst the features.



As I can see review scores are strongly related to each other's. Totally I decided to compare price with other features.

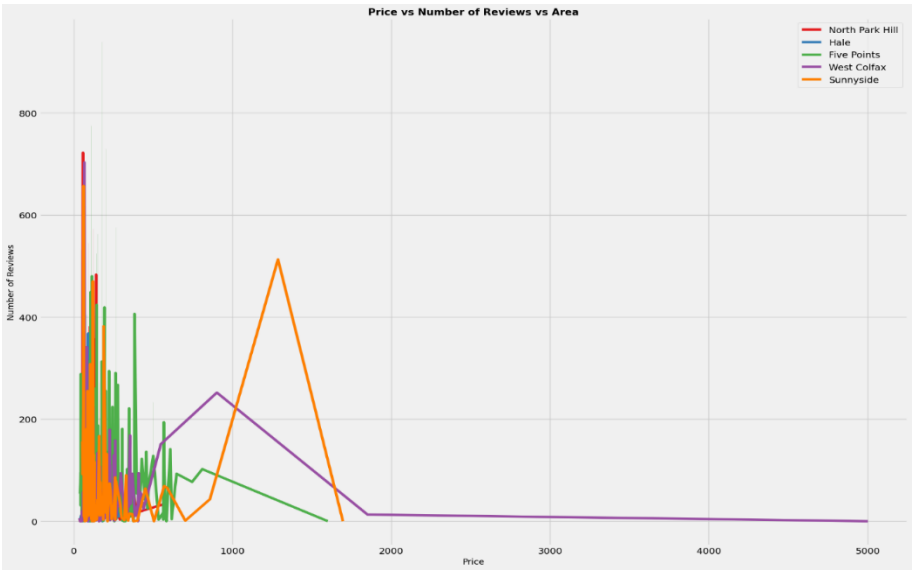
3. Explore Price with other Features

    1) Explore Price and Room Type by Area
    I group them by regions.



It shows us the most five Expensive areas are 'North Park Hill', 'Hale', 'Five Points', 'West Colfax', and 'Sunnyside'. So, I visualize them below.

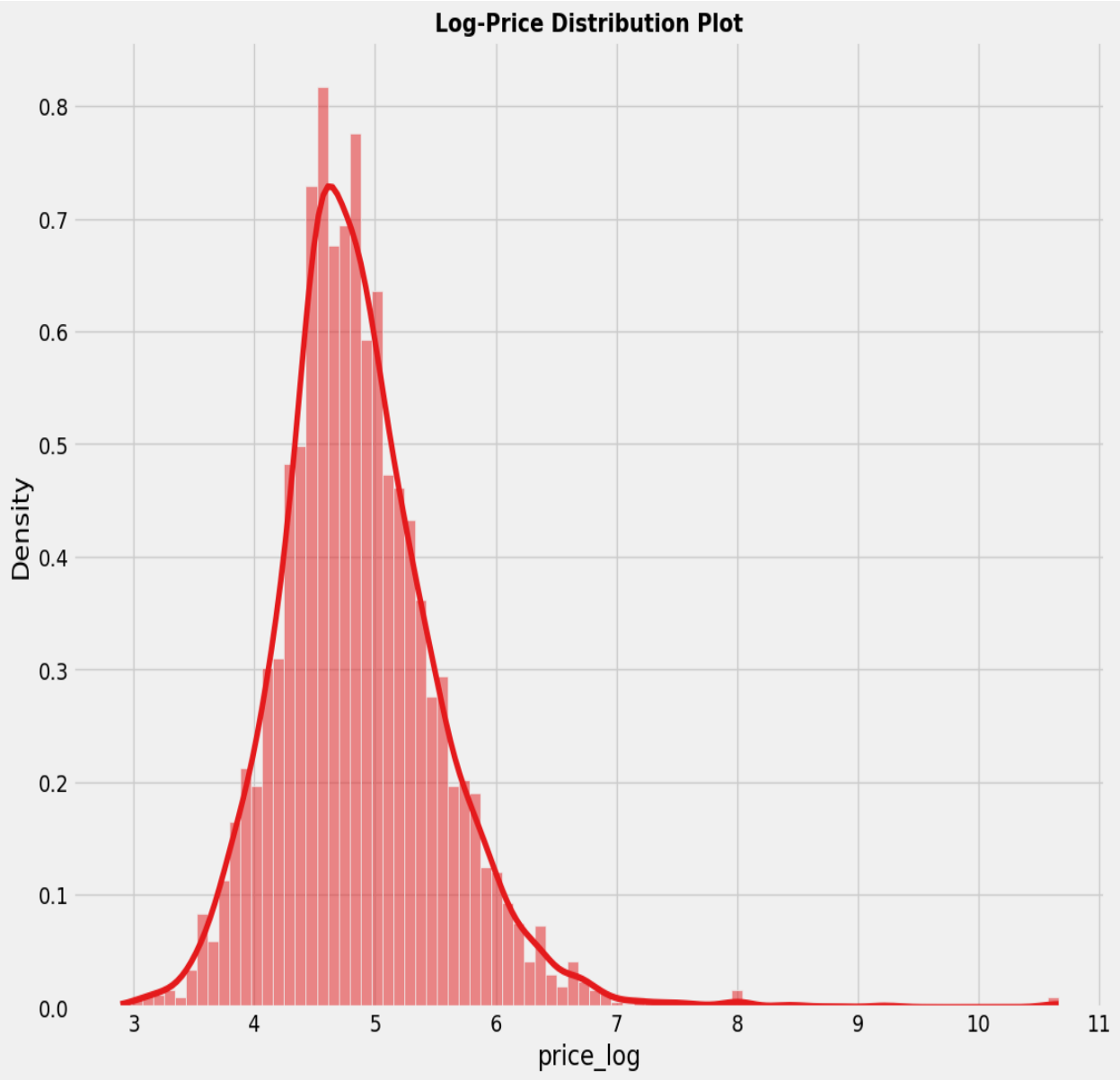
2) Explore Price and Number of Reviews by Highest Price Area



3) Log Price Distribution Plot

The above distribution graph shows that there is a right-skewed distribution on price. This means there is a positive skewness. Log transformation will be used to make this feature less skewed. This will help to make easier interpretation and better statistical analysis.
Since division by zero is a problem, log+1 transformation would be better.

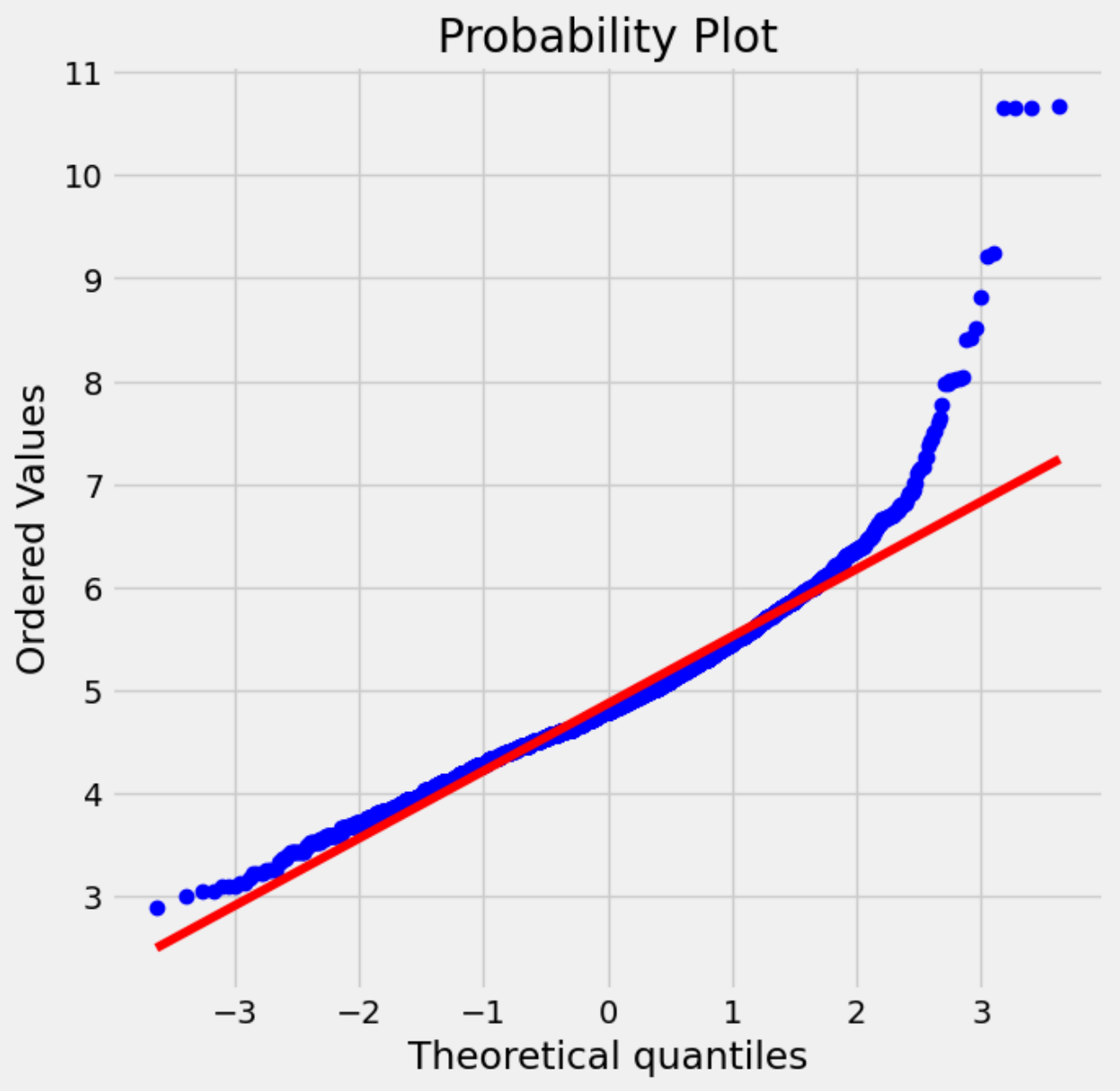
Log price gives us a better visualization of price distribution.



4) Probability Plot

The purpose of a probability plot (Q-Q plot) is to visually assess whether the data follows a particular distribution, in this case, a normal distribution. If the data points fall approximately along a straight line, it suggests that the data is normally distributed. Departures from the straight line indicate departures from normality.

The points curve upward. There are more values at the tails of the distribution, which suggests that the data has a higher frequency of extreme values or outliers.



5. Preprocessing And Training

Preprocessing is a critical step in the data science pipeline, involving various techniques to clean, transform, and prepare raw data for analysis. This stage ensures that the data is in a suitable format and quality for modeling and insights extraction. Typically, preprocessing includes steps such as handling missing values, removing duplicates, encoding categorical variables, scaling numerical features, and handling outliers.

1. Create dummy features for room types

   To create dummy features for room types, you can use one-hot encoding. This technique converts categorical variables into binary vectors, where each category becomes a separate feature with a value of 0 or 1.
   I did this step for room type.

2. Standardize numeric features using a scaler

   First, Standard Scaler technique will be used to normalize the data set. Thus, each feature has 0 mean and 1 standard deviation.

3. Train/Test Split

   I would have 70/30 train/test split.

6. Modeling

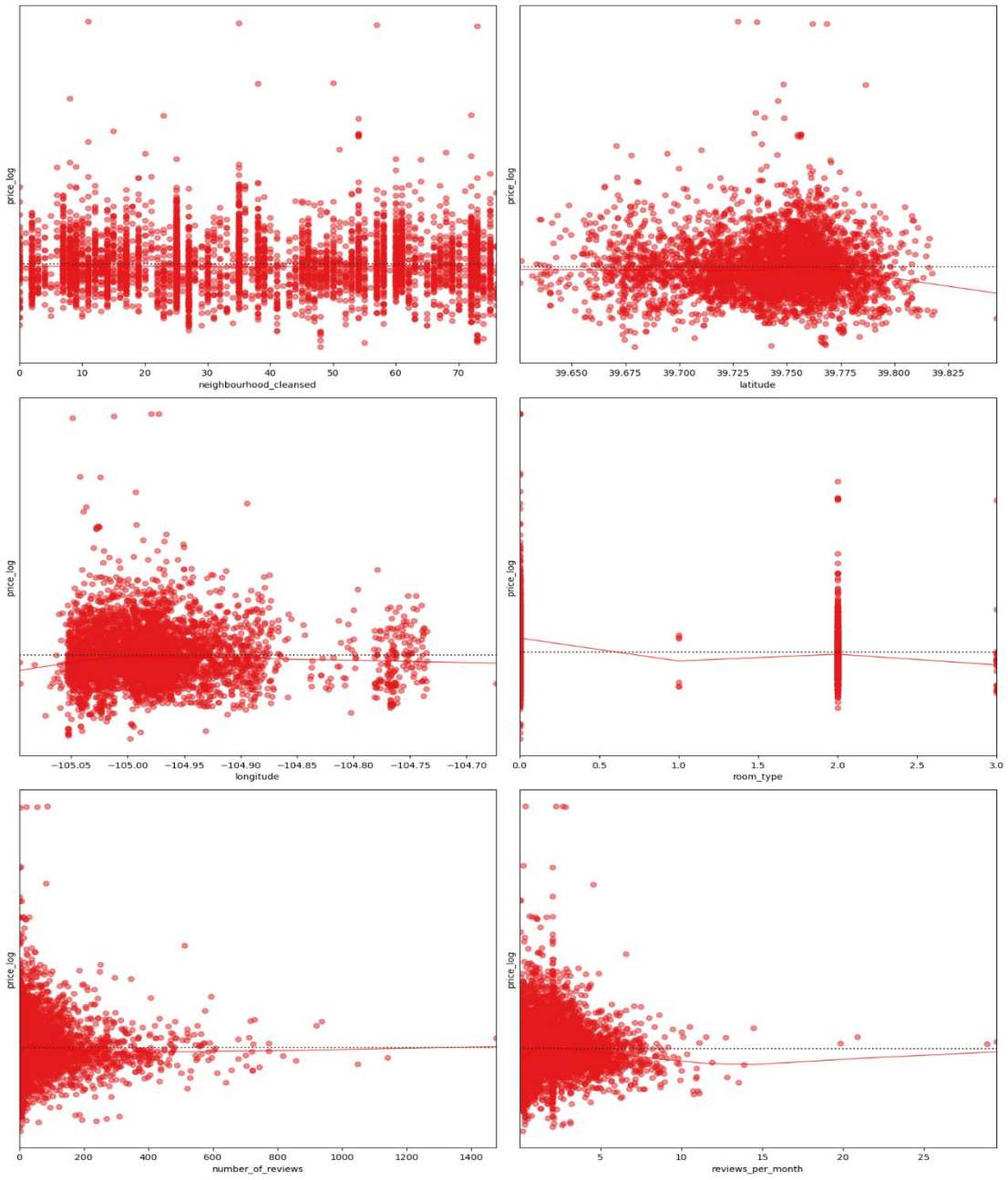
1) Residual Plots
2) Feature Selection and Grid Search
3) Model Scenarios
    A. Scenario 1(With All Features)
        a) K-Fold Cross Validation
        b) Polynomial Transformation
        c) Model Prediction
    B. Scenario 2(Without All Features)
        a) K-Fold Cross Validation
        b) Polynomial Transformation
        c) Model Prediction
4) Model Comparison
5) Variance Comparison
6) Conclusion
7) Summary

I select these columns below:
neighborhood cleansed, latitude, longitude, room type, number of reviews, reviews per month, log of price.
The models that I will create are linear regression, Ridge, Lasso, and Random Forest.

1) **Residual Plots**

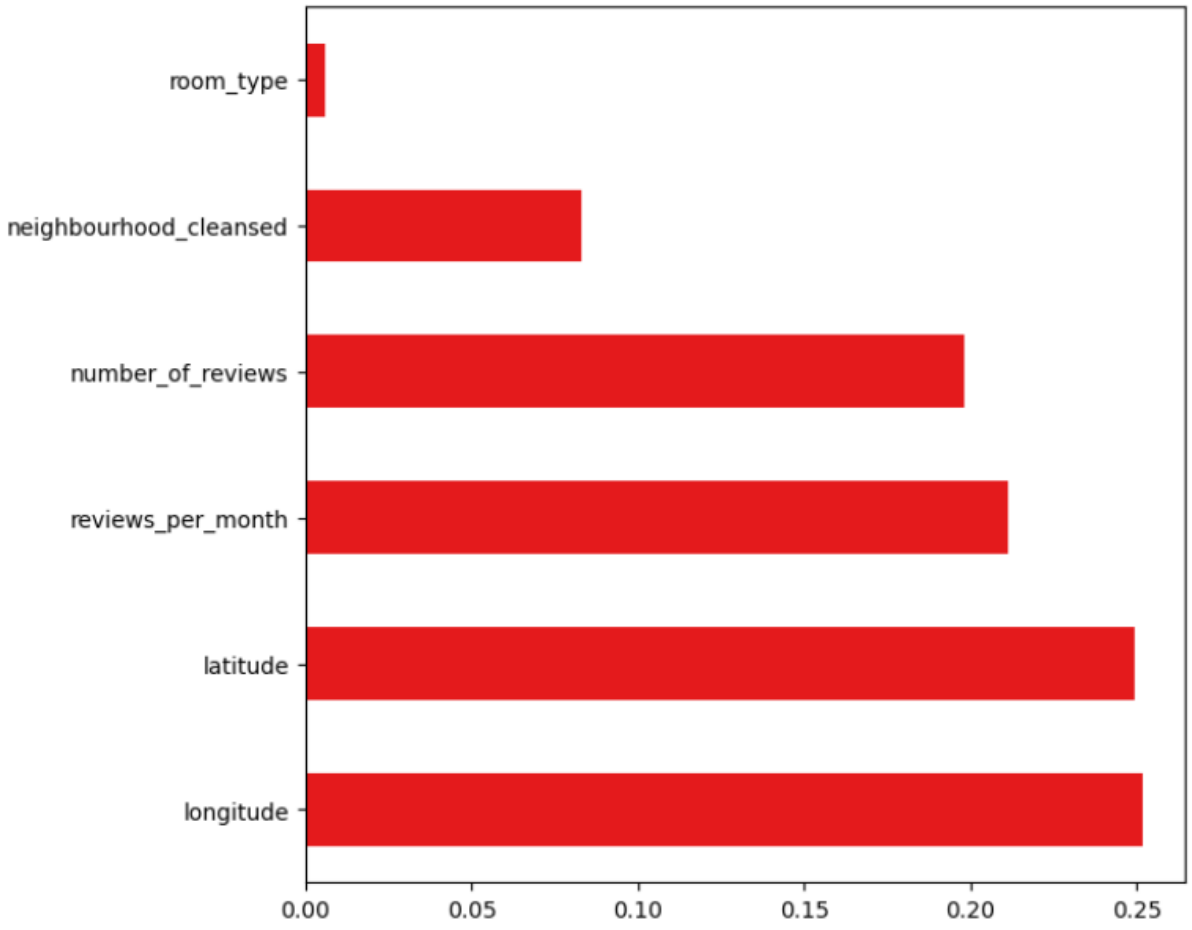
Residual Plot is strong method to detect outliers, non-linear data and detecting data for regression models. The below charts show the residual plots for each feature with the price.

An ideal Residual Plot, the red line would be horizontal. Based on the below charts, most features are non-linear. On the other hand, there are not many outliers in each feature. This result led to underfitting. Underfitting can occur when input features do not have a strong relationship to target variables or over-regularized. For avoiding underfitting new data features can be added or regularization weight could be reduced.



2) Feature Selection

Now it is time to build a feature importance graph. I use Extra Trees Classifier to create this visualization below.



The above graph shows the feature importance of dataset. According to that, neighborhood cleansed and room type have the lowest importance on the model. So, I decide to create two scenarios. First scenario is with all features. Second one is without room type and neighborhood cleansed.

3) Model Scenarios

A. Scenario 1(With All Features)

Correlation matrix, Residual Plots and Multicollinearity results show that underfitting occurs on the model and there is no multicollinearity on the independent variables. Avoiding underfitting will be made with Polynomial Transformation since no new features cannot be added or replaced with the existing ones.

In model building section, Linear Regression, Ridge Regression, Lasso, and Random Forest Regression models will be built. These models will be used to avoiding plain Linear Regression and show the results with a little of regularization.

First, ‘GridSearchCV’ method will be used to find the best parameters and tuning hyperparameters for each model. In this algorithm 5-Fold Cross Validation and Mean Squared Error Regression Loss metrics will be used.

a) K-Fold Cross Validation

   Before model building, 5-Fold Cross Validation will be implemented for validation.

b) Polynomial Transformation

   The polynomial transformation will be made with a second degree which adding the square of each feature.

c) Model Prediction

B. Scenario 2(Without room type and neighborhood cleansed)

a) K-Fold Cross Validation

   Before model building, 5-Fold Cross Validation will be implemented for validation.

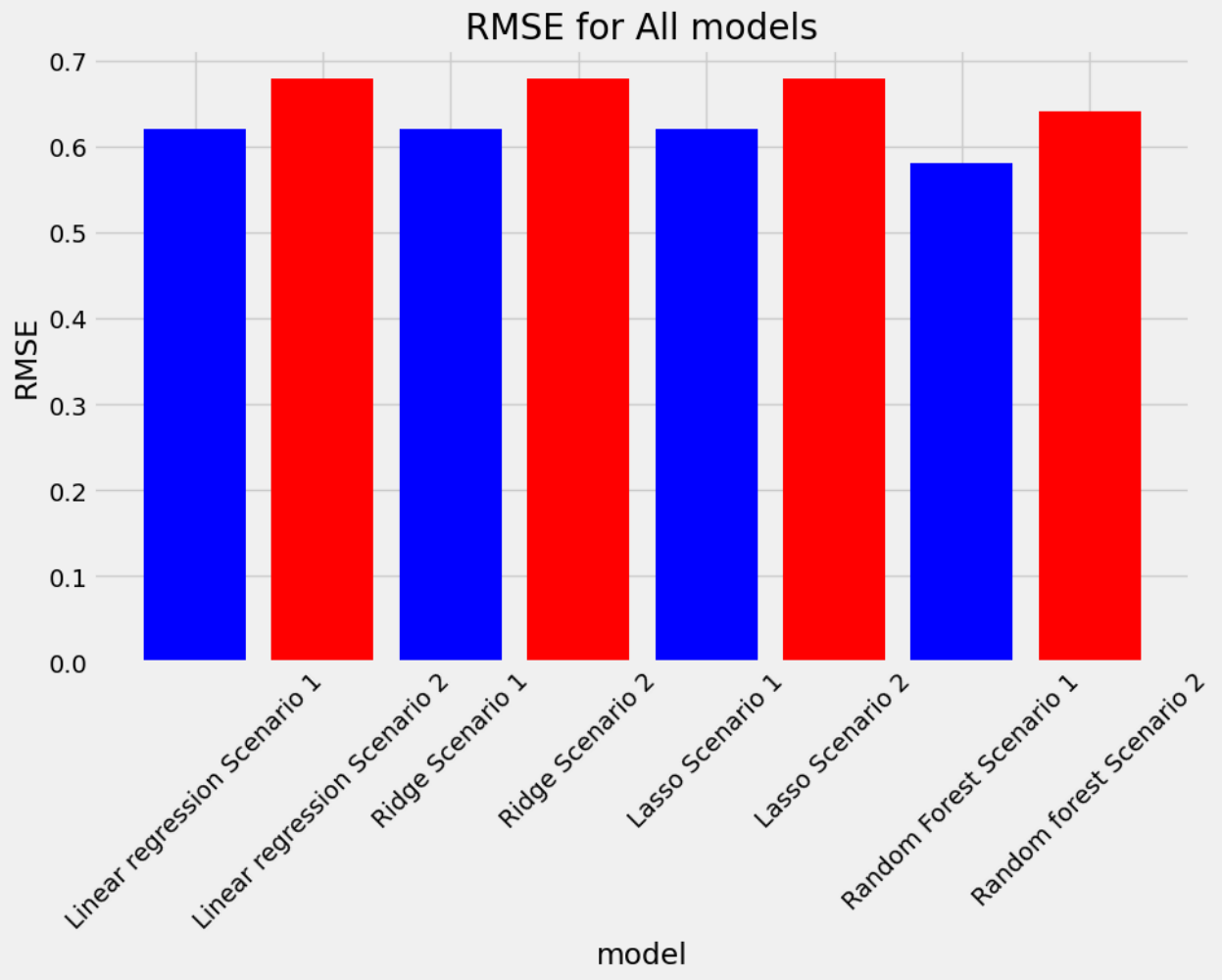
b) Polynomial Transformation
c) Model Prediction

4) Model Comparison

Root Mean Square Error (RMSE) shows how accurately the model predicts the response.

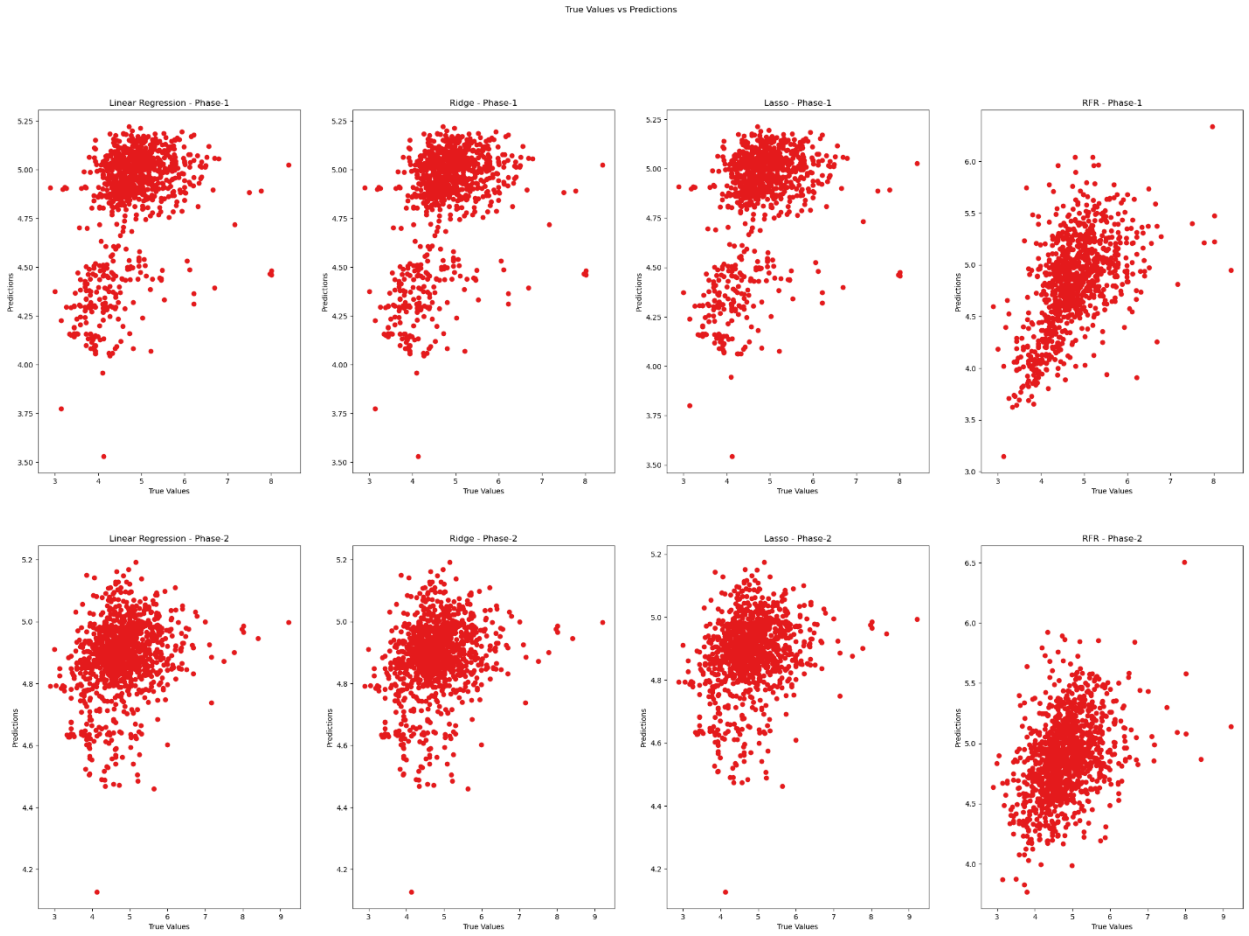
RMSE = sqrt [(Σ (Pi – Oi) ²) / n]

I visualize the RMSE for all models.



A lower RMSE indicates better accuracy of the predictive model. As we can see here scenario 2 has higher RMSE. The lowest RMSE is Random Forest in scenario 1. It means the best model is Random Forest with all feature that I select first.

5) True Values vs Predictions for all models



We can notice to these tips below:

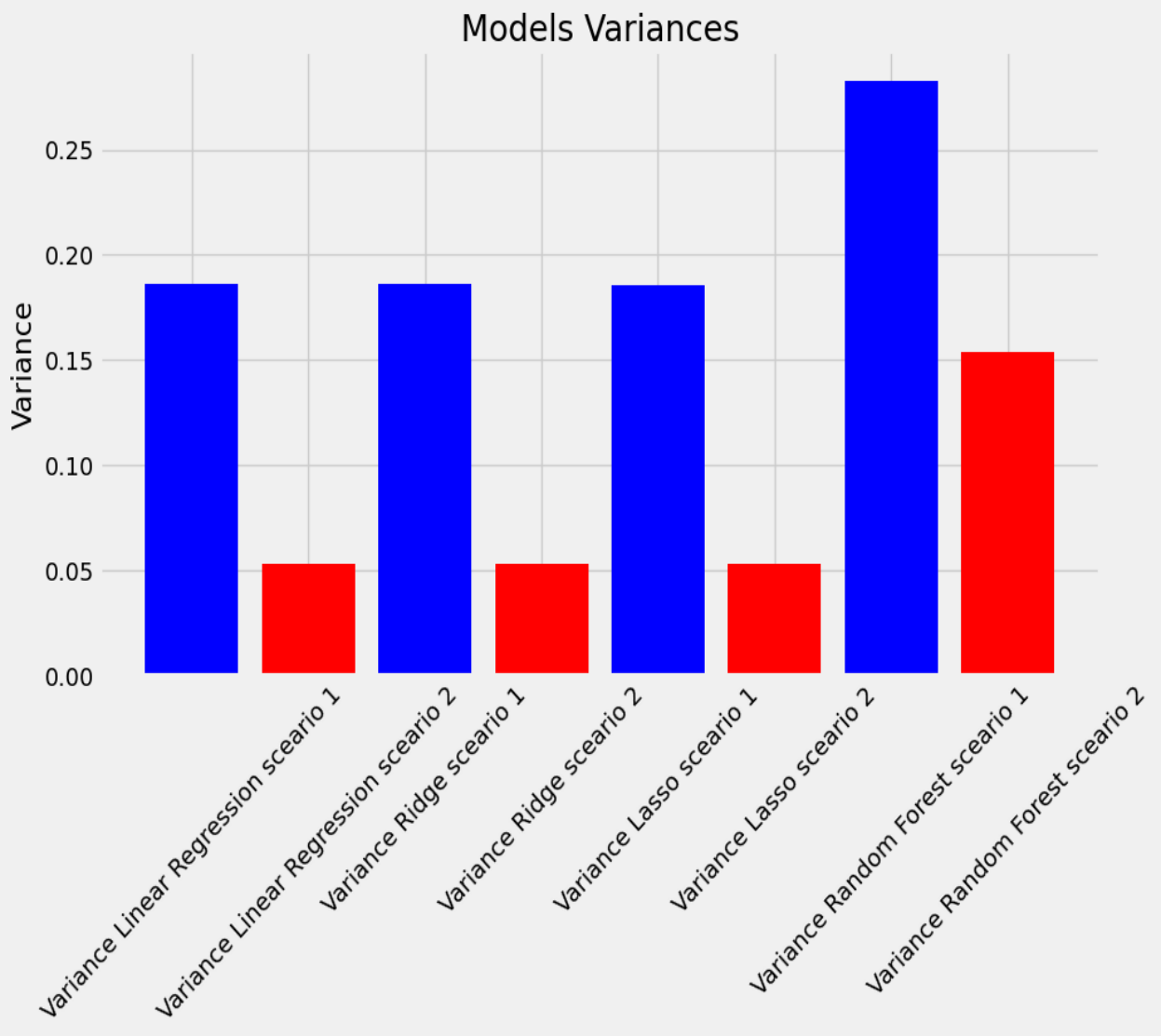
Model Accuracy: You can visually assess how well the model predictions match the actual values. A tight cluster of points along a diagonal line (with slope 1) indicates a strong correlation between predicted and actual values, suggesting good model accuracy.

Outliers: Outliers, or points far from the diagonal line, might indicate areas where the model performs poorly or where there are extreme values in the data.

Model Fit: The overall shape of the scatter plot can provide insights into the overall fit of the model. A non-linear relationship between actual and predicted values might suggest that the model is not capturing the underlying patterns in the data.

As we can see Random Forest with scenario 1 has tight cluster, less outliers. And this model has the better linear relationship between actual and predicted values.

6) Models Variances



Variance measures how much the model's predictions deviate from the mean of the actual values. A higher variance score indicates that the model can explain more of the variance in the data, meaning it captures the underlying patterns better.

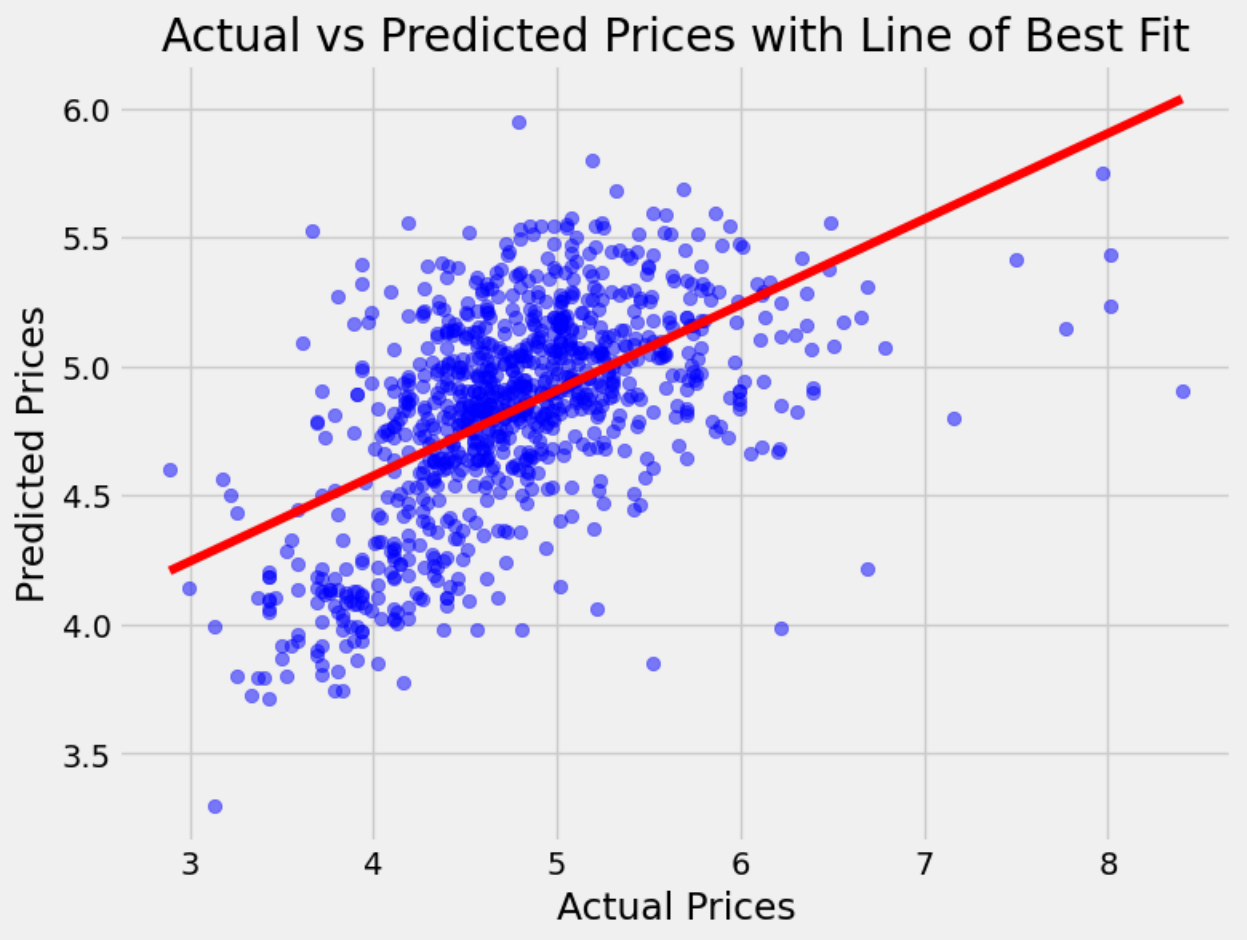
As we can see Variance Random Forest for scenario 1 has the highest Variance.

7) Create The Best Model

I try to use the random_forest_reg function that I defined before. It uses the GridSearchCV that gives us the best parameter for our random forest model that we defined. Here we have five parameters below:
'n_estimators': [10, 50, 100],
'max_depth': [None, 10, 20, 30],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [1, 2, 4],
'bootstrap': [True, False]

After that it give us the best parameters of the best model.
I will use the exact these parameters to find the best model.

This the actual and predicted of the best model Random Forest from scenario 1 with all features.



8) Conclusions

- Scenario 1 is better, so, we need to keep all features.

- Random forest is the best model in scenario 1 and 2.

- Finally, we can find the best model with the best parameters to predict the nightly Airbnb price.

9) Summary

- Four models were created here:

Linear Regression, Ridge, Lasso, and Random Forest.

- Two scenarios were existed:

1. Scenario with all deatures

2. Scenario without two features room type and neighborhood cleansed that were less important.

- Totally Random Forest with scenario 1 has the highest variance and lowest RMSE that it shows it was the best model for prediction of log price of denver airbnb.

- I use GridSearchCv to find the best paramaters for Random Forest scenario 1. The results is shown below:

The best parameters:

'bootstrap': True,

'max_depth': 20,

'min_samples_leaf': 4,

'min_samples_split': 5,

'n_estimators': 100

- I used these parameters to create the best model for my data, and it was saved.