Analysis of active Airbnb listings for San Diego as of August 2017.

Anit Mathew

Table of contents

Introduction

Airbnb has been a leading competitor in the travel industry. It has provided an integrated platform to the host and their customers to have a convenient experience of hosting and staying. The larger idea behind the business is that it allows residents to list a room in their house to travelers and earn additional revenue. However, while listing or booking a stay, many factors lead to attracting more customers or looking for cheaper places to stay. In this project, we will analyze what factors lead to an increase or decrease in the price of a listing and what major factor a customer looks into while booking the cheapest option. I travel a lot, and I browse a lot on Airbnb for cheaper options with certain amenities. I am curious to know if there is a relationship between amenities and prices and what major factors or amenities lead to price change through this dataset.

The queries that we would be researching will be as follows:

1. What are the major factors in the increase or decrease according to room type, bedroom, and accommodation in prices of any listing?
2. Does the neighborhood of San Diego affect the pricing of the listings? Which neighborhood has the highest average price? Demonstrating it on a map and analyzing it.
3. Is there any relation between prices and the amenities like 'bedrooms', 'room_type', 'reviews', 'overall_satisfaction',and 'accommodates'? How reliable is the relationship? Can we depend on them to make a prediction?

Dataset

The dataset has been collected over the course of 4 years by Tom Slee. Tom Slee has not just collected data for San Diego but for all major cities around the globe.

Link:https://tomslee.net/category/airbnb-data

1. room_id- A unique number that identifies an Airbnb listing.
2. host_id- A unique number that identifies an Airbnb host.
3. room_type- Divided into 3 types "Entire home/apt", "Private room", or "Shared room"
4. neighborhood- As with borough a subregion of the city or search area for which the survey is carried out. For cities that have both, a neighborhood is smaller than a borough. In some cities, there is no neighborhood information.
5. reviews- Total number of reviews that a listing has received.
6. overall_satisfaction- The average rating (out of five) a listing has received from its visitors.
7. accommodates- Total number of guests a listing can accommodate.
8. bedrooms- Number of bedrooms a listing has.

```
    room_id  survey_id    host_id    room_type  country       city  borough  \
0  11637213       1436   24705242  Shared room      NaN  San Diego      NaN
1  14351163       1436   87948847  Shared room      NaN  San Diego      NaN
2   9327098       1436   31043523  Shared room      NaN  San Diego      NaN
3  17535919       1436  117987352  Shared room      NaN  San Diego      NaN
4   3688119       1436   13209607  Shared room      NaN  San Diego      NaN

    neighborhood  reviews  overall_satisfaction  accommodates  bedrooms  \
0  Pacific Beach        2                   0.0             2         1
1  Mountain View        0                   0.0             1         1
2    Tierrasanta        0                   0.0             2         1
3  Pacific Beach        1                   0.0             3         1
4    Cortez Hill        1                   0.0             2         1

   bathroom  price  minstay last_modified   latitude   longitude  \
```

```
0       NaN     63      NaN       19:23.0  32.795259 -117.252712
1       NaN     80      NaN       19:23.0  32.699822 -117.106789
2       NaN     75      NaN       19:23.0  32.836102 -117.084967
3       NaN     70      NaN       19:23.0  32.801236 -117.241454
4       NaN     78      NaN       19:23.0  32.724846 -117.165529

                                                 location
0  0101000020E6100000A2D3F36E2C505DC0C26C020CCB65...
1  0101000020E6100000A5A487A1D5465DC0BA826DC49359...
2  0101000020E6100000508F6D1970455DC0620FED63056B...
3  0101000020E6100000425F7AFB734F5DC05E30B8E68E66...
4  0101000020E61000008962F206984A5DC07651F4C0C75C...
```

Methodology

The data collected by Tom Slee was easy to download from the website. Upon looking at the dataset, it was hard to understand.

- Data Cleaning: The data contained more than 5000 records which included columns which had no values. To start with the research, it was important to clean the data. So, removed all the columns like countries, bathrooms and, boroughs.

```
Data is cleaned and has no null values
```

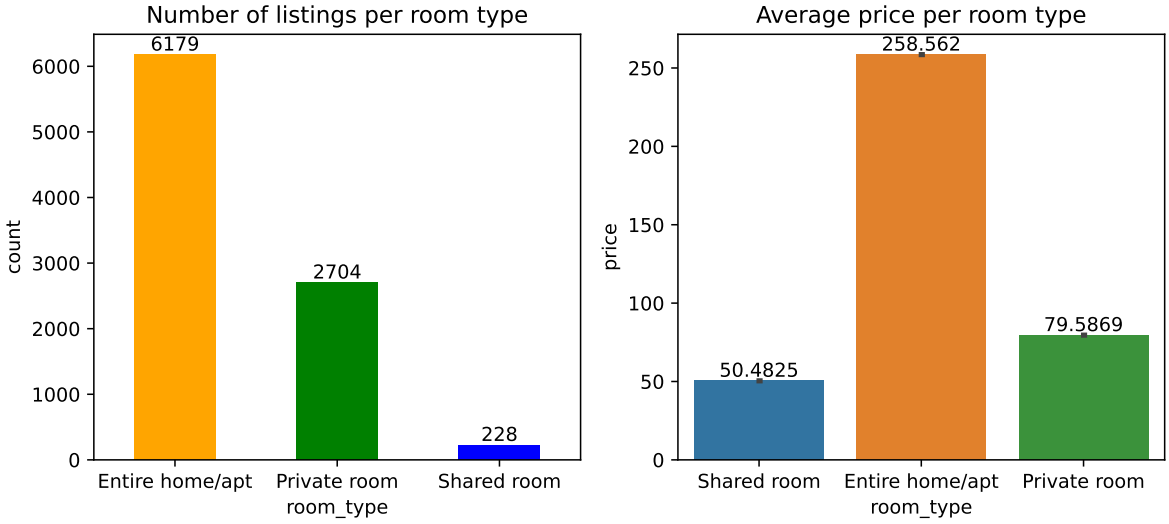
- For analyzing the prices in relation to variables like bedroom, room type, accommodation, we extracted the particular columns. First, we counted the number of listings according to room type, bedroom, and accommodation. It is very important to look into the count to have a better understanding of its relationship with the price. Thereafter, we calculated the average price in relation to every amenity which the listing has to offer. The analysis is done by plotting the results on bar charts to have a clear picture and understanding.
- If we need to rent an Airbnb, which place would be the best in the city - the central area or the beach area? Figuring out the answer might be difficult because every listing has a different price range. To better understand the pricing of each neighborhood, we grouped the dataset according to the neighborhood and calculated the average price for each neighborhood. To our surprise, we got 102 neighborhoods in the city of San Diego. It would be hard to plot the results over a bar chart, as the data is vast. So, we decided to plot the result over a map. The average pricing data was plotted over the map using latitude and longitude over the neighborhood. This gave us a clear picture to analyze which neighborhood is expensive and which is cheap.

- Do the amenities of the listing have any influence on the pricing? Which amenity has a larger influence on price? To answer these questions, we checked out whether there is any relationship between the amenities and pricing. We performed this analysis using 'Machine Learning'. We processed the data using label encoding so that labels could be converted into numeric form to make machine-readable. We performed Multiple Linear Regression and Random Forest Regression on the dataset to find how strong the relationship is and which feature is the most important.

Results

Visual analysis of the listing.

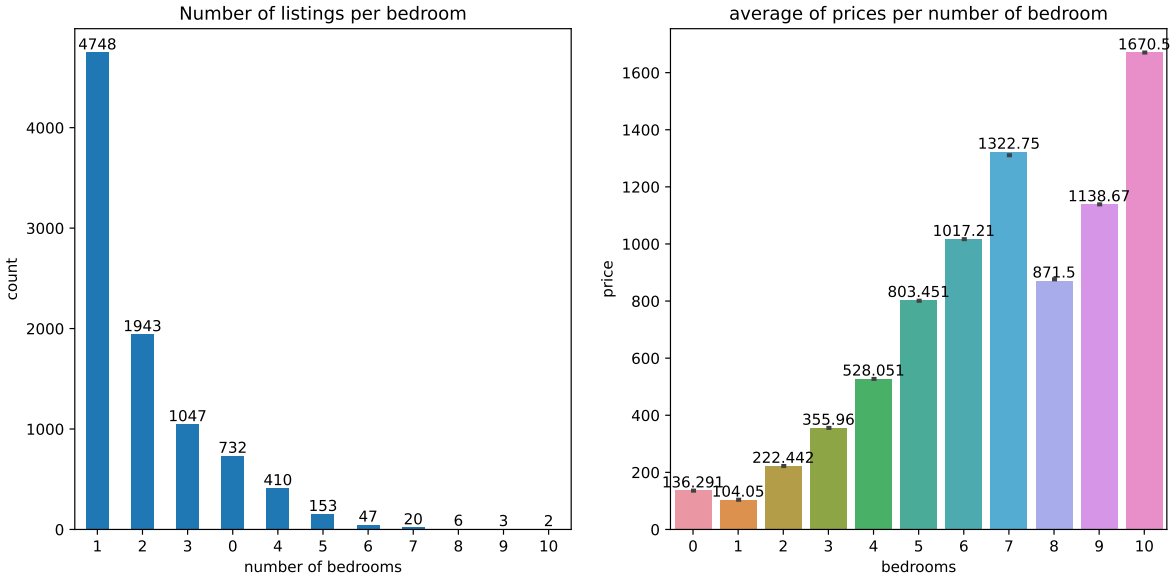
Price relation with room type



As per the first plot, entire home/apt has 6179, private room has 2704 and shared room has 228 listings throughout San Diego. Thereafter, while plotting average prices of listings as per room type. We found that entire home/apt has an average price of $258.562227, private room has an average price of $79.586908 and shared room has an average price of $50.482456.

Looking at the figures and plots, it is easily acceptable that entire home listings which are in majority have the highest average price, and shared room listings have the lowest average prices. We can analyze that here the price is in direct relation to the number of listings. The more the listing, the higher the average.

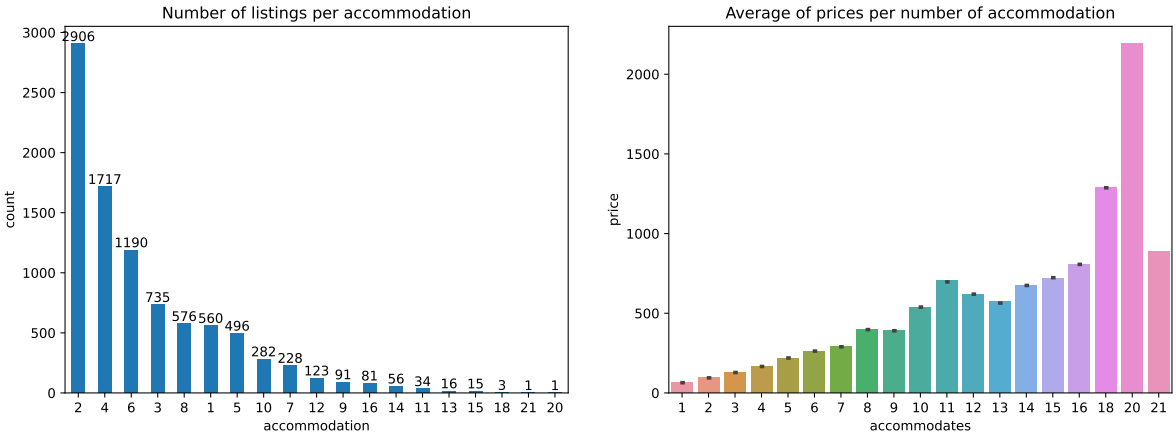
Price relation with number of bedrooms.



In the first plot, listings with 1 bedroom have the highest number with 4748, and listings with 10 bedrooms have the lowest count of 2 listings in San Diego.

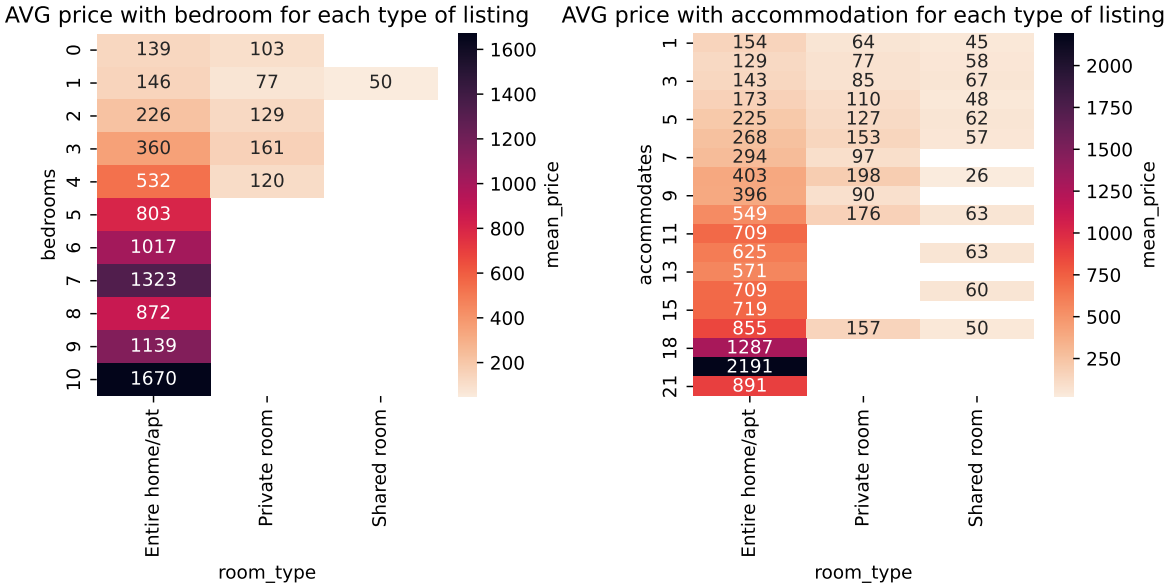
Further, while analyzing the second plot, listings which has the highest number of count, has the lowest average prices which is hard to believe. Along with that, the 2 listings which have 10 bedrooms have the highest average price in the city. By looking at the plot, we can analyze that the average prices of the listings increase with the increase in the number of bedrooms. The price has a direct relation with the number of bedrooms.

Price relation with number of accommodation.



In the first plot, listings that can accommodate 2 people have the highest number with 2906, and listings that can accommodate 21 and 22 people have the lowest count with just 1 listing. Looking at the plot the listings are pretty scattered, listings with even number of accommodations are more compared to the odd ones. But the interesting factor in the second plot is, no matter what the number of accommodations of the listing is, the average price is increasing with the increase in the number of accommodations. Though the count of listing with 20 people is the lowest, the average price is the highest which is more than $2000.

All the above listing gives us an overview of whether the price is related to the amenities provided by the listings. Looking at the plots above we can distinguish that prices are related to the space and amenities they provide. If we need a big sized apartment with more rooms and accommodation, we may need to pay more. To better understand the scenario, we tried to plot the average price with the number of bedrooms for each type of listing.



With the above heat map, we get an in-depth idea of average pricing. The shared room has the lowest count in the city with only 1 bedroom and several accommodations. Its price is also at an average of $50 which is the cheapest option in the whole set. The relation that we get here is that as the listings shift from shared room(right) to entire home/apt(left), the average price also tends to increase. And as the number of bedroom and accommodation increases, the average price also tends to increase.

So, we can conclude by analyzing the plots once again that if we need a big sized apartment with more rooms and accommodation, we may need to pay more.

Price relation with neighborhood

Neighborhood is a major factor when renting an Airbnb facility. Should I be getting a place in the city or by the beach? Which place would be the cheapest? Will a beach house be expensive? These are the major questions that arise in our thoughts. The question here is what is the price change according to the neighborhood?

With 102 neighborhoods, it is hard to understand which neighborhood has the highest price. So, we tried to plot listings on a map. Average prices for every neighborhood on the map provide a better understanding. By looking at the plot, we can understand that majority of neighborhoods have an average price below 250 but as we move to the coastal areas, the average price increases. Torrey Pines which is near the beach has the highest average price which reaches up to $506. This provides us with a clear indication that the listings near the beach in San Diego are expensive compared to the ones which are in the city.

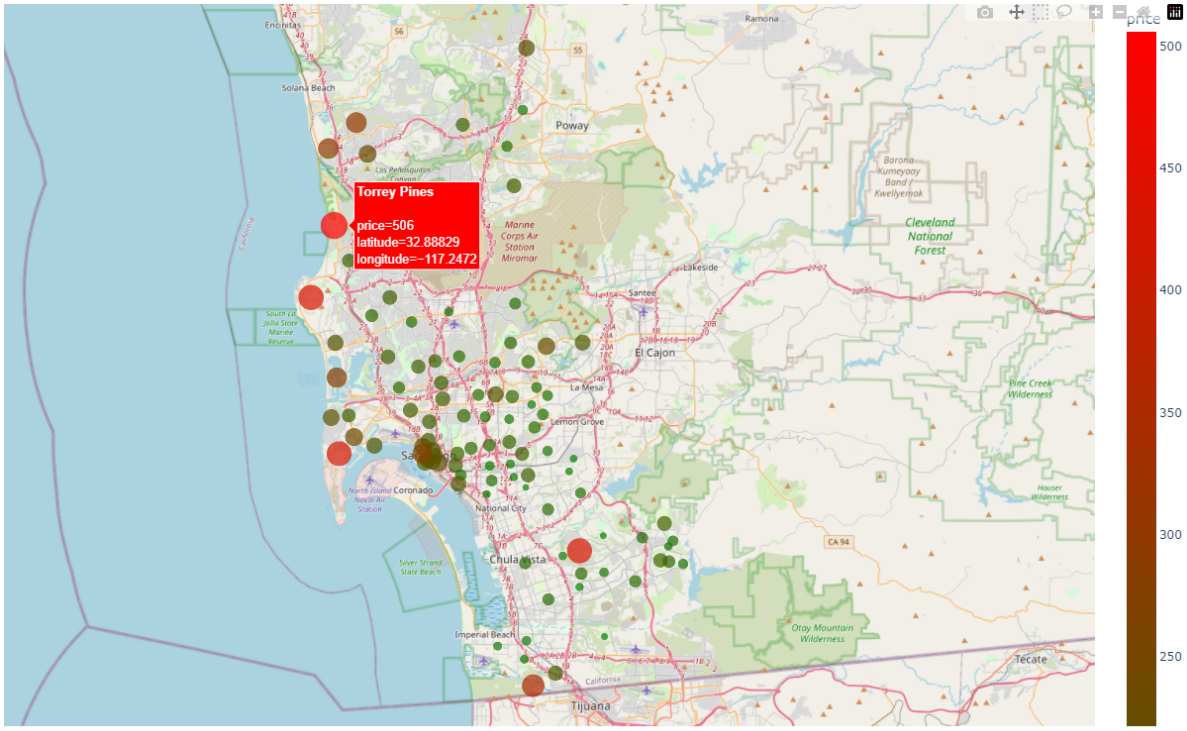
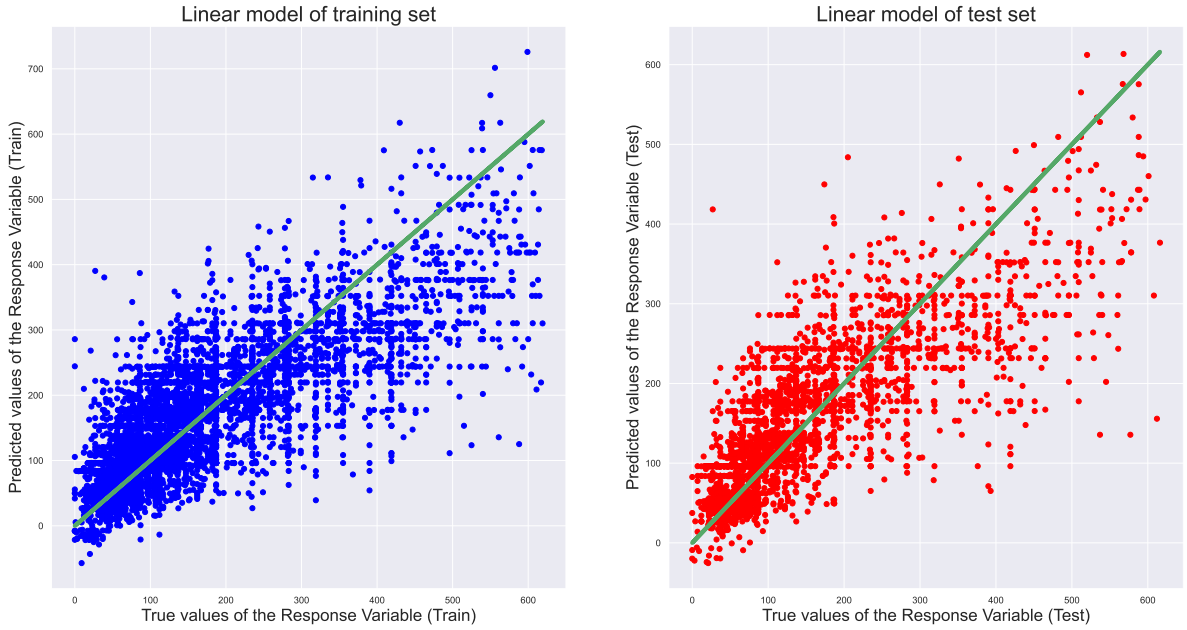


Figure 1: Average price per neighbourhood data

Machine Learning

Multiple Linear Regression model

Multiple Linear Regression model is a model that helps to determine a relationship between quantitative dependent variables with two or more independent variables in a straight line. In other words, we will be using this model to establish whether there is a relationship between price and amenities. We have fit the model as price = a * (Predictor variable) + b.



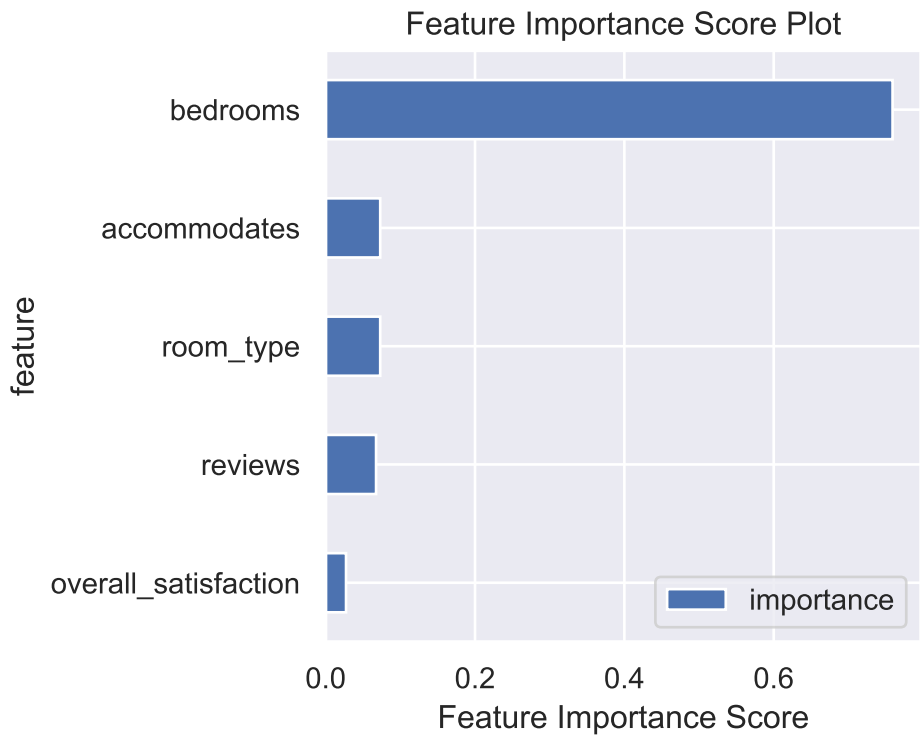
Points that lie on or near the diagonal line imply that the values predicted by the Linear Regression Model are highly accurate. The plot suggests that it has an R-squareed value of 0.63 which means it has an accuracy of 63%.

```
The adjusted R-squared is :  0.6417038502638214
```

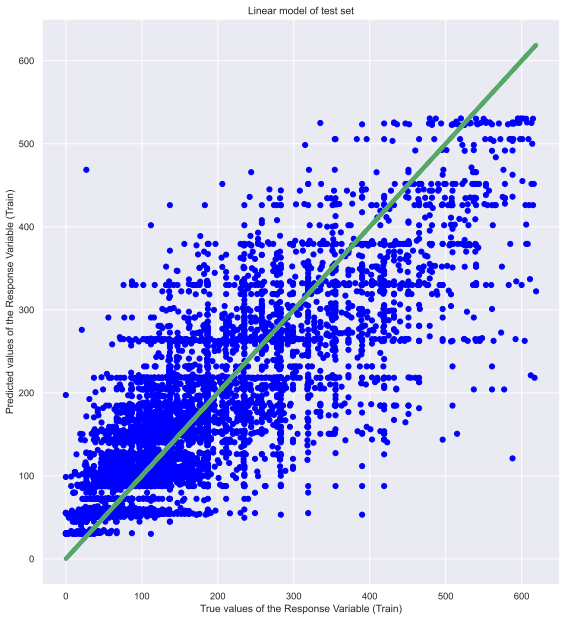
Random Forest Regression

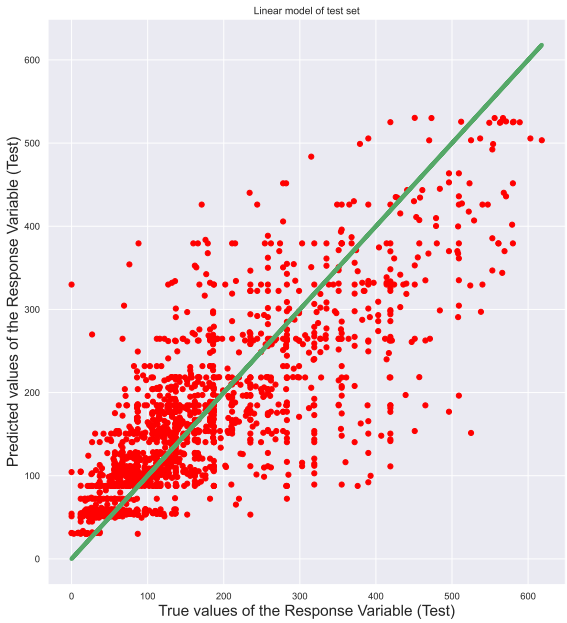
Random Forest Regression is a model that uses an ensemble learning method for regression. It is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. In this model, we will be predicting the price and also finding out the most important variable. The importance of a variable depicts how useful is that variable in building up the model accurately.

The above plot provides the information that the bedroom variable has the highest importance with more than 70% rate. The rest of the variables like room_type, reviews, accommodation, and overall satisfaction has an importance of 0 to 10%.




```
The R square value of the model :  0.6617507995844509
```





Points that lie on or near the diagonal line mean that the values predicted by the Linear Regression model are highly accurate. The plot suggests that it has an R-squared value of 0.65 which means it has an accuracy of 65% which is a little better than the linear model.

Therefore after analyzing both models, we can conclude that the predictor variables do have an influence on the price level and the bedroom has the highest influence compared to all other amenities.

Discussion

With all the insight received from the exploratory analysis and machine learning models, we can conclude that:

- Bedroom has the highest influence on price. With the increase in the number of bedrooms, the price level also increases.
- San Diego has more number of entire properties listed comparatively which have the highest price level.
- Listings near to the beach areas attract higher prices compared to listings in the city area.
- Torrey Pines has the highest average price compared to all 102 neighborhoods.
- Reviews and overall satisfaction ratings do not have much effect on price.
- There is a less but positive relationship between review and price system. It has less than 10% of importance, so the feedback system cannot be relied on to predict the price.

The main limitation of this dataset is the fewer amount variables which makes it difficult to reach an accuracy of more than 70%. Both the Regression models provide a R-squared value of around 65% which could have been better if it was more than 70%. One way to reach that goal is to add more data which will provide us with more insight into the data. There can be more factors to an Airbnb listing such as washrooms, air conditioner, refrigerator, hot water, parking facility, etc. which can enhance the strength of the model and helps to analyze the relationship with price in depth. If we need to work on this future, a dataset with more features will need to be procured.

Citiations

- Dataset : https://tomslee.net/category/airbnb-data
  - Author: Tom Slee
  - Year of Publication: October 9, 2017