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

- Anit Mathew

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. However, while listing or booking a stay, many factors lead to attracting more customers or looking for cheaper places to stay. 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. In this project, we will analyze what factors lead to an increase or decrease in the prices of a listing.

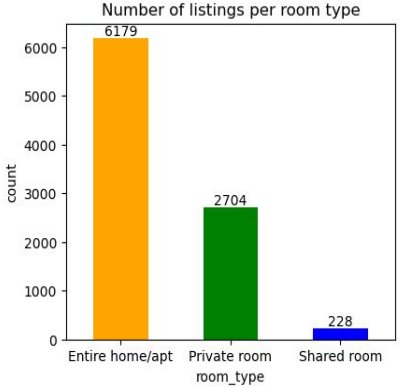
The queries that we would be researching will be as follows:

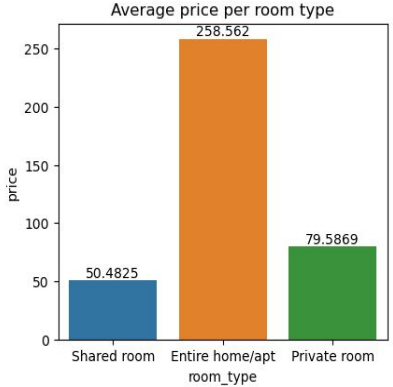
1. What are the major factors in the increase or decrease according to room type, bedroom, and accommodation in prices of any listing?
2. Does the neighborhood of San Diego affect the prices of the listings? Which neighborhood has the highest average price? Demonstrating it on a map and analyzing it.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

DATASET

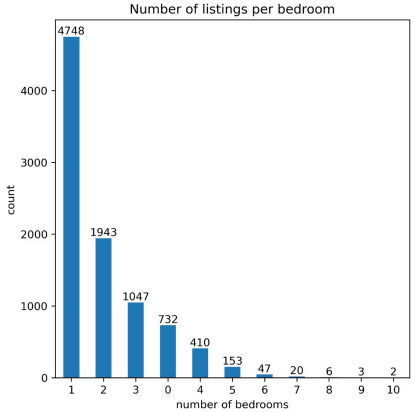
room_id	survey_id	host_id	room_type	country	city	borough	neighborhood	reviews	overall_satisfaction	ccommodates	bedrooms	bathroom	price	minstay	last_modified	latitude	longitude	location
11637213	1436	24705242	Shared room	NaN	San Diego	NaN	Pacific Beach	2	0	2	1	NaN	63	NaN	19:23.0	32.7952	-117.252712	0C0C26C020CCB65...
14351163	1436	87948847	Shared room	NaN	San Diego	NaN	Mountain View	0	0	1	1	NaN	80	NaN	19:23.0	32.6998	-117.106789	0C0BA826DC49359...
9327098	1436	31043523	Shared room	NaN	San Diego	NaN	Tierrasanta	0	0	2	1	NaN	75	NaN	19:23.0	32.8361	-117.084967	0C0620FED63056B...
17535919	1436	117987352	Shared room	NaN	San Diego	NaN	Pacific Beach	1	0	3	1	NaN	70	NaN	19:23.0	32.8012	-117.241454	0C05E30B8E68E66...
3688119	1436	13209607	Shared room	NaN	San Diego	NaN	Cortez Hill	1	0	2	1	NaN	78	NaN	19:23.0	32.7248	-117.165529	0C07651F4C0C75C...

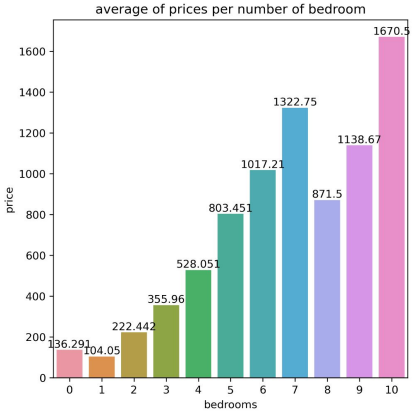
ANALYSIS: BARPLOTS

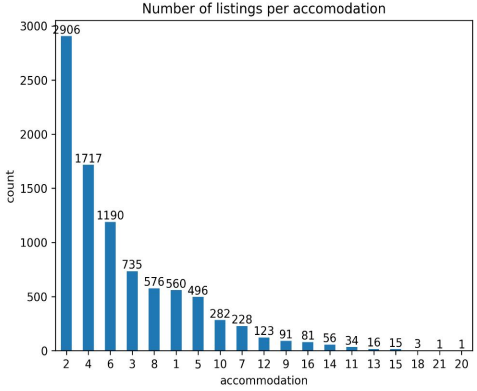


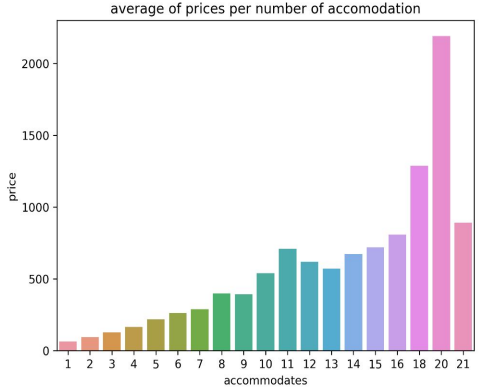


- Looking at the figures and plots, it is easily acceptable that Entire home listings which is in majority has the highest average price and shared room listings has the lowest average prices. We can analyze that here the price is in direct relation the number of listing. The more the listing, higher the average.
- By looking at the plot, we can analyze that average prices of the listings per bedroom increases with the increase in number of bedrooms. The price has direct relation with number of bedrooms.
- The interesting factor in the second plot is, no matter what the number of count of the listing is, the average price is increasing with the increase in number of accommodation. Though the count of listing with 20 people is lowest, the average price is the highest which is more than $2000.



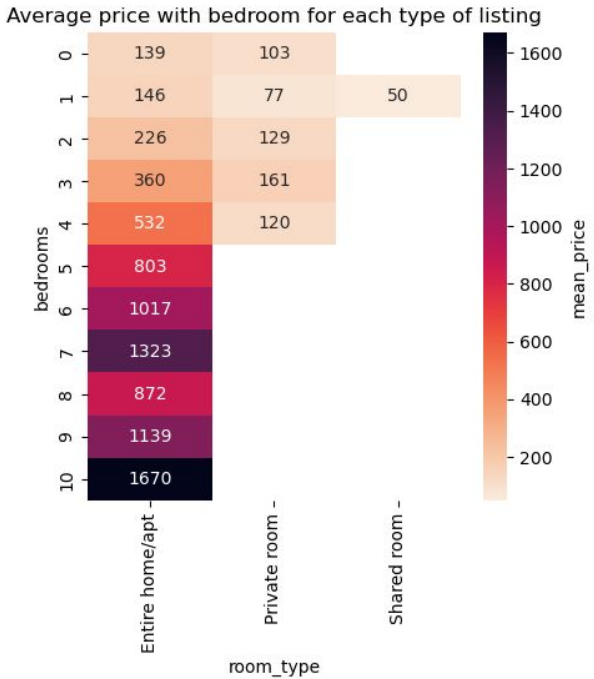


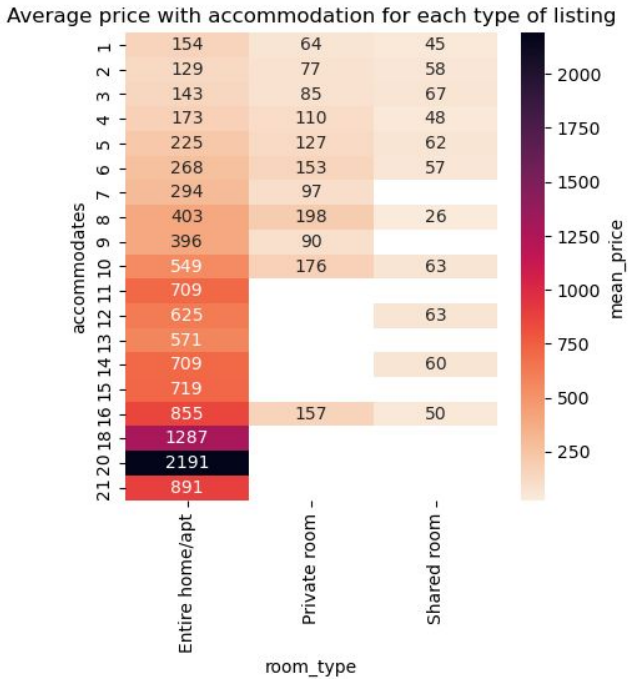




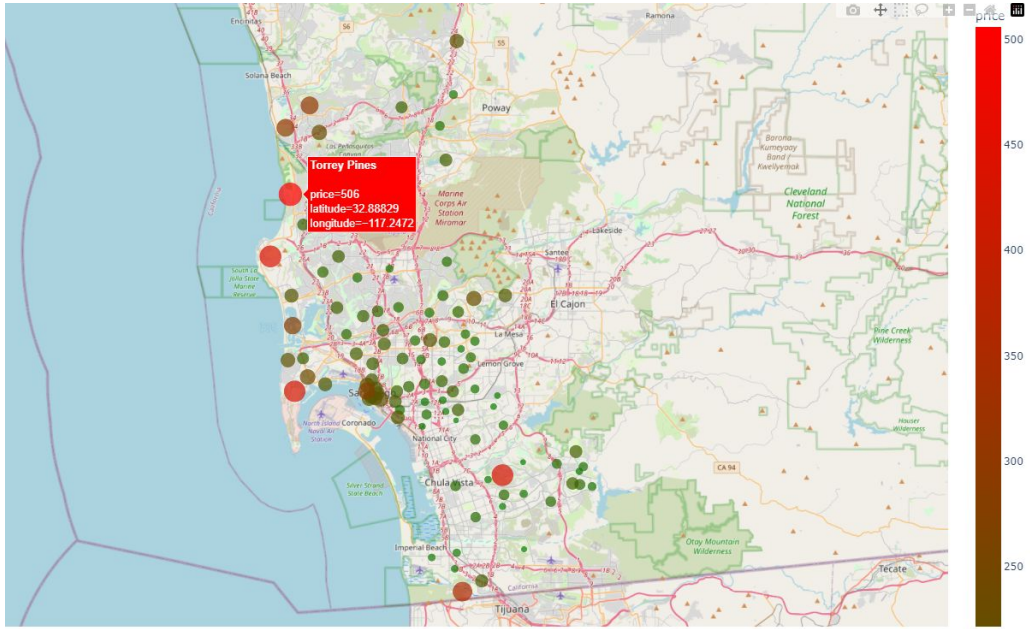
ANALYSIS: HEATMAP

Shared room has the lowest count in the city with only 1 bedroom and several accommodation. Its price is also at an average of $50 which is the cheapest option in the whole set. The relation which 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 bedroom and accommodation increases, average price also tends to increase





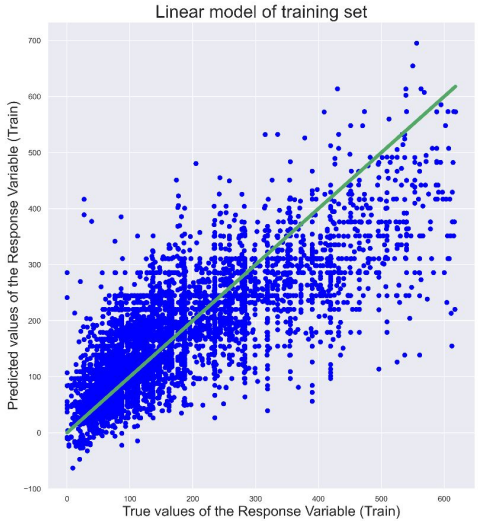
ANALYSIS: PRICE VS NEIGHBORHOOD

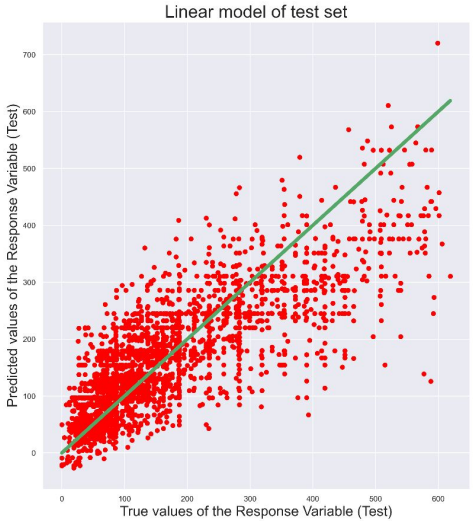


Torrey Pines which is near the beach has the highest average price which reaches upto $506. This provides us with a clear indication that the listings near the beach in San Diego is expensive compared to the ones which are in the city

MACHINE LEARNING: MULTIPLE LINEAR REGRESSION

Linear Regression performs the regression task to estimate a target variable value (price) based on provided independent variables (neighborhood, room type, accommodation, etc. It then tries to find a linear relationship between the variables and predicts the price based on the linear line.



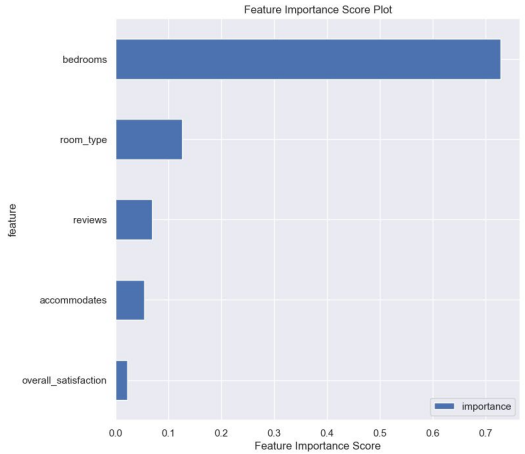


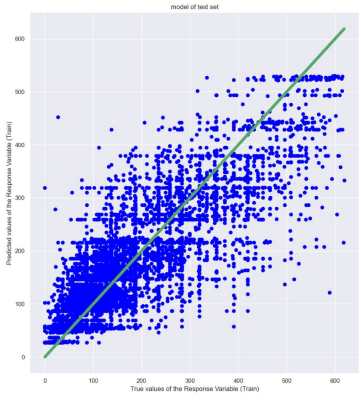
Points that lie on or near the diagonal line means that the values predicted by the Linear Regression model are highly accurate.

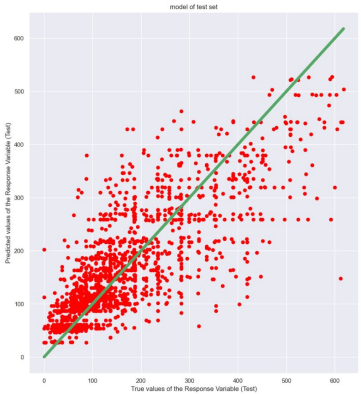
The adjusted R-squared is : 0.6417533558641315

RANDOM FOREST REGRESSION

We used the Random Forest Regressor to help predict the price while also finding out the most important variables (i.e features). Importance provides a score that indicates how useful or valuable each feature is.







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

The R squared value of the model : 0.6599691651942647

RESULTS

- Bedroom has the highest influence on price. With the increase in number of bedrooms, the price level also increases.
- San Diego has more number of entire-home 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 neighborhood.
- Reviews and overall satisfaction ratings does 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 feedback system cannot be relied on to predict the price.

LIMITATIONS AND FUTURE

The main limitation with this dataset is the less amount variables which makes it difficult to reach an accuracy of more than 70%. Both the Regression models provide a R-square 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 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, dataset, more features will be procured.


THANK
YOU