QMB6304.002S20: - Analytical Methods for Business.

Airbnb – Austin Occupancy Rate Prediction

Submitted by: -

- Sathish Reddy

7th May 2020
Prof: Dr. Anol Bhattacherjee

Table of Contents

1.Executive summary

Airbnb has successfully disrupted the traditional hospitality industry as more and more travelers decide to use Airbnb as their primary accommodation provider. Since its inception in 2008, Airbnb has seen enormous growth, with the number of rentals listed on its website growing exponentially each year. After reading about the Airbnb occupancy rate in the USA, our interest was driven to find out what factors have a higher influence on occupancy rate, to investigate any such factors which can have some implementable outcomes to improve occupancy rate. This is a replicable problem for all new Airbnb hosts in Austin and relatively similar in other locations.

To pursue this analysis, we used the data provided by Inside Airbnb (http://insideairbnb.com/austin/), which is publicly available information about a city's Airbnb's listings that had been scraped and released for independent, non-commercial use. This includes details about the listing such as no. of rooms, guests included amenities, description, location as well as information about the occupancy rate such as price and the average number of reviews per month. A standard approach has been followed by performing the exploratory analysis, forming a hypothesis, and conducting the regression analysis on the data. The data has been modeled using 3 models and finding the one which predicts the data best and is practically suitable.

To build this model, the concept of occupancy rate is used as a proxy for potential future earnings. The Occupancy rate is defined as the number of days the listing is booked to the total number of days listing is available. The calculation of Occupancy Rate uses Inside Airbnb's The Occupancy Model (also known as 'San Francisco Model') (http://insideairbnb.com/about.html) based on the average length of stay and review rate. A review rate of 50% (http://insideairbnb.com/austin/#) to convert reviews to estimated bookings and the average length of stay in Austin is 3 days.

Occupancy Rate = Avg number of reviews * Avg length of stay /Review rate

Key findings from our analysis suggest that the Price of the listing, Room type, and contemporary amenities influence the Occupancy Rate of the listing. This report aims at making some actionable recommendations to new hosts and even the existing host who are facing low occupancy rate issues.

2.Problem Significance

Airbnb is a popular home-sharing platform, enabling people all over the world to lease or rent short-term lodgings. For potential hosts, this is a potentially lucrative option for their under-utilized vacation homes, spare rooms or even beds. However, it is difficult for new hosts to ascertain how much their property could earn them and whether it will be a worthy investment. So, the primary target client for this project is an Airbnb host.Using publicly available information, **this project will help in predicting the Occupancy Rate on prospective Airbnb listings in Austin.** So, that new hosts can concentrate on a few important factors and can get a high occupancy rate.

While there are available commercial tools such as Mashvisor (https://www.mashvisor.com), BiggerPockets (https://www.biggerpockets.com/airbnb-calculator) that predict Airbnb value, these models are generic with minimal features used and no indication of model accuracy. So, we are making this project considering various predictors variables that are practically reliable and building a predictive model using Regression Algorithms.

3.Prior literature

A listing's success, efficiency, and growth depend primarily on two kinds of factors or environments; the external and the internal. External factors include significant external problems such as markets, the effects of economic conditions and uncertainty, changes in government policies and regulations, information technology development, the demographics of a country or state, and the socio-cultural influences in the end. And various internal factors are leading to unforgettable hotel accommodation. The four key factors are; quality, place, facilities, and cleanliness, according to Lockyer (2005).

Customers who are happy with the quality of services would most likely return and may bring in new customers as well. Success or failure of a hotel is then largely determined on the satisfaction of the customers. Researcher claims that while quality customer service is not the only deciding factor, if pleased with the facilities, customers will certainly return to the same hotel and Price would ultimately spread good words to their friends and families. This will result in a higher occupancy rate, which would then translate into sales growth A positive emotion among customers that lead to positive mouth words, whereas negative emotions will result in complaining behavior. Gratification is the secret to business in the long term and to produce a further profit. Customer rates the listing essentially on experience.

4. Data Source and Preparation

4.1 Data Sourcing and Wrangling

To start with any Analysis, it is a known fact that the credibility and reliability of the data collected are of paramount importance. The entire data for this project has been sourced from the Inside Airbnb website, one of the leading providers of Airbnb listings data for independent, non-commercial use. The data in these files is quite overwhelming and contained thousands of observations for the Price, Average reviews per month, and information about the host. Since loading and analyzing these massive datasets at this level is quite challenging, we will use the detailed listings information for active listings from 1st March 2019 – 29th Feb 2020. An active listing is defined as a property that has been reviewed at least once during this time period.

Few of our predictor variables are character data types so we dropped them and important character variables like Amenities list, here we used regular expressions to clean and extracted meaningful insights.

4.2 Data Preparation

The data required for this project has been prepared in the following sequence and approach.

1. Airbnb Austin data file consisted of 7000 plus observations with more than 100 attributes or columns. So, we took only practically realistic predictor variables.
2. The variables with the values or categories 'Unknown' in most of the cells have been removed as they do not contribute to any meaningful interpretations.
3. All the continuous filtered are taken in a range following Box Plot condition and all the values falling out of range are removed.
4. only essential amenities like, Tv, Wi-Fi, Air conditioning, Free parking on-premise, and Heating are used in the model.
5. Occupancy Rate variable has been created using Average no of reviews, Review rate, and Average length of stay. This also forms our Target variable.

Variable Name	Definition
Occupancy_Rate	Occupancy rate of the listing
Listing_id	Unique Id of Listing
Accommodates	No. of people the listing can accommodate
Guests_included	No. of guests included in the base price
Bathrooms	No. of bathrooms
Bedrooms	No. of bedrooms
Beds	No. of beds
Price	Combination Price of listing and Cleaning fee
Extra_people	Price for extra guests not included in the base price
Minimum_nights	Minimum no. of nights required for booking
Maximum_nights	Maximum number of nights to book a room
Availability_365	No. of nights the listing is available for booking
Reviews_per_month	Average no. of reviews listing receives per month
Amenities	Types of Amenities included
Room_type	Type of room e.g. private, shared etc.
Host_is_superhost	Host is super host or not

5. Hypotheses

Now that all the data has been cleaning and the necessary variables have been obtained, the core hypotheses can be formulated to know exactly what we are going to test or infer in this project.

The following hypotheses have been laid out for this project:

1. H1a: βEntire_home_apt < βPrivate_room > βShared_room

   Most of the people, in general are interested in staying in a private room rather than a shared room. While an entire house is way costlier than a private room, we can say that private room type should be more significant than the entire house and shared room.

2. H2a: βPrice < 0

   Everyone tries to get a booking at the lowest possible rates and even when there are similar kinds of listings people tend to take the one at a low price. So, when the Price increases the chance of booking a listing should decrease.

3. H3a: βMinimum_nights < 0

   The Minimum number of nights needed for booking a listing acts as a major obstacle because there is a possibility that a person wants to book for a single night but there may be a minimum day to book. So, as the number of minimum nights to book increases the occupancy rate will decrease.

4. H4a: βGuest_included > 0

   The more the number of people can stay in a listing the more economical the listing is. So, as the number of guests included in a listing increases the occupancy rate should increase.

5. H5a: βsuper_host_true > 0

   Super host is those who are having good reviews for quite some time and their response time, availability is greater than average.

6. H6a: βBedrooms < 0

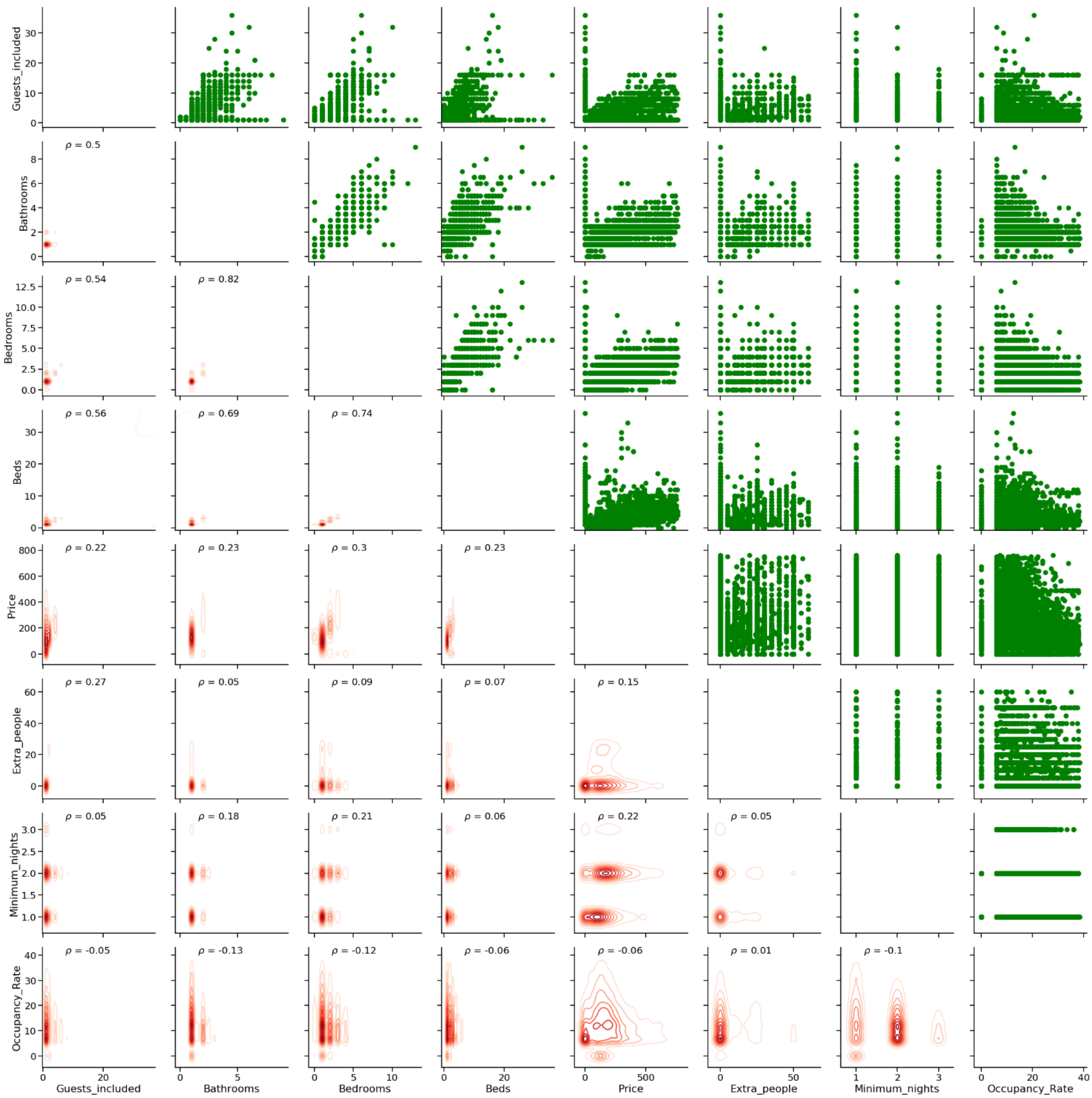
   If we can get a greater number of bedrooms consequently the price of the listing will be more than the relatively similar listings so, the occupancy rate should decrease as the number of bedrooms increases.

7. H7a: βBathrooms < 0

   Like the number of the bedroom if the number of bathrooms increases the price of the listing will also increase resulting in low occupancy rates.

6. Descriptive Analysis

To explore more about the data, we used a Scatter plot, kdeplot and Violin plot to see the trends in it. A Scatter plot has been plotted for all continuous variables or characteristics that are expected to have an impact on the Occupancy rate of the listings.

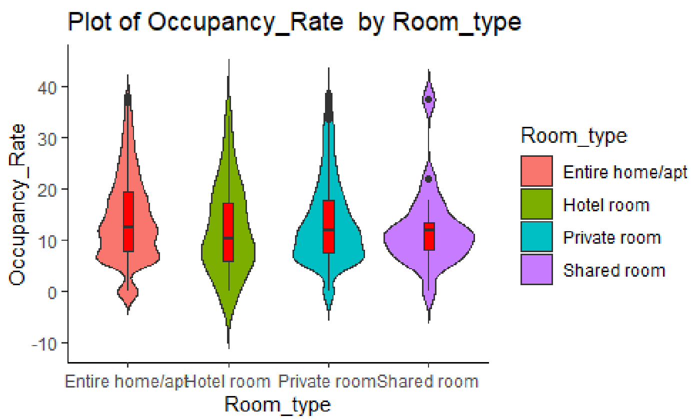


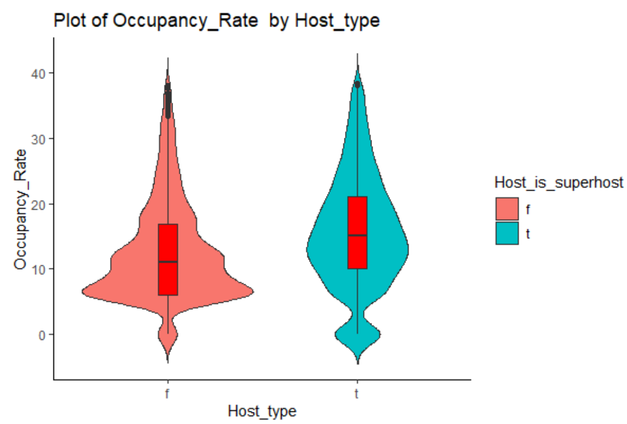
kdeplot have been plotted for all continuous variables or characteristics that are expected to have an impact on the Occupancy rate of the listings to check the density and on the top of each kdeplot, we mentioned the correlation coefficient.

From the correlation coefficient values, we can see that there is some correlation among the independent variables which is needed to be removed before using regression models and all the predictor variables are showing some sort of relation on the occupancy rate.

From the scatter plots we can say as occupancy rate of the listing decreases as the price, guests included, minimum nights to be booked increases.

Violin plots are drawn with an occupancy rate against the categorical columns.

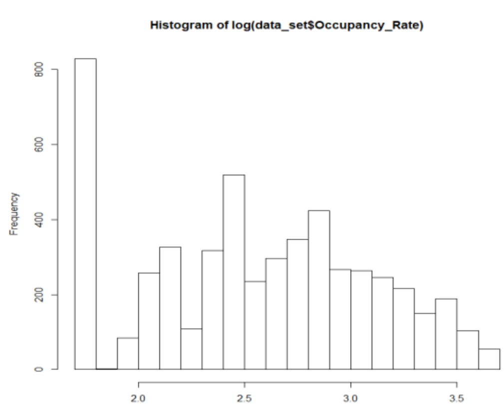


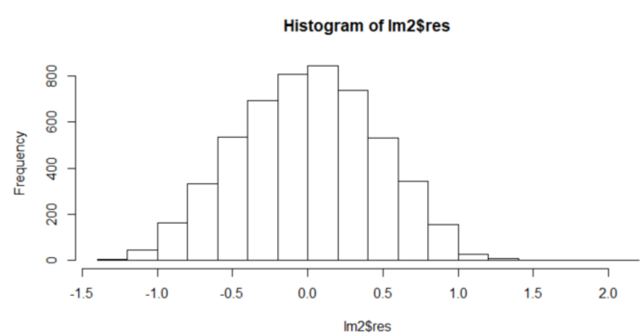


From the above violin plots of room type vs occupancy rate room type of Entire home or apartment and the private room has more occupancy rate when compared to the hotel room and shared room

In the violin plot of Host type vs occupancy rate, Host type False is standing more at occupancy rate less than 10 while those who are super are having high occupancy rates.

As the target variable Occupancy rate is right, we applied nonlinear transformation on it.





7.Modelling

Based on the above hypotheses and the descriptive analysis, the following models have been formulated: The target variable is the Occupancy rate, the models have been built using linear Regression and the analysis has been done in R.

Model 1

Occupancy_Rate = f (Accommodates, Guests included, Bathrooms, Bedrooms, Beds, Price, Extra people, Minimum nights, Availability_365, Tv, Wi-Fi, Air-conditioning, Free_parking_on_premises, Heating, Room_type, Host_is_superhost*).*

Model 2

Occupancy_Rate = f (Accommodates, Guests included, Bathrooms, Bedrooms, Beds, Price, Extra people, Minimum_nights, Availability_365, Tv, Wi-Fi, Air-conditioning, Free_parking_on_premises, Heating, Room_type, Host_is_superhost ,Price* Minimum_nights* Room_type).

Interaction among Price, Minimum_nights and Room_type.

Model 3

Occupancy_Rate = f (Accommodates, Guests included, Bathrooms, Bedrooms, Beds, Price, Extra people, Minimum_nights, Availability_365, Tv, Wi-Fi, Air-conditioning, Free_parking_on_premises, Heating, Room_type, Host_is_superhost , Price*Beds*Bedrooms).

Interaction among Price, Beds and Bedrooms.

7.1 Model Comparisons

The above models are linear regression models. Model 1 explains the effect of the independent variable on the response variable and there are no interactions between the independent variables. The model 2 includes all the predicted variables along with interactions among Price, Minimum_nights, and Room_type. This model explains the effect of interaction terms practically.it relates to the practical implementation and like the real-world scenario. A customer would like to book a room based on the price for a minimum number of nights. Model 3 is also the same, but interaction terms are different compared to both models. In this model, we are looking at interaction term price, Beds, and bedrooms. This model does not follow any realistic scenario. Basically, most of the customers would be interested in booking a room for a minimum price. And would not consider much type of beds and bedrooms. Model 2 is better for predicting the occupancy rate. There are certain predictors that are not much contributing to the occupancy rate in either manner. As from the above models Guests_included, Extra_people are not contributing much. Though they are not contributing much to the occupancy rate, Property owners are managers will get to know whether to invest money and time on the unfavorable predictors.

	Dependent variable:		
	Occupancy_Rate)		
	(1)	(2)	(3)
Accommodates	0.018*** (0.004)	0.019*** (0.004)	0.015*** (0.004)
Guests_included	0.0003 (0.003)	0.0003 (0.003)	0.0003 (0.003)
Bathrooms	-0.100*** (0.013)	-0.102*** (0.013)	-0.091*** (0.013)
Bedrooms	-0.073*** (0.010)	-0.073*** (0.010)	-0.123*** (0.014)
Beds	0.018*** (0.005)	0.017*** (0.005)	-0.018** (0.009)
Price	-0.0003*** (0.00004)	-0.001*** (0.0001)	-0.002*** (0.0001)
Extra_people	-0.0002 (0.0004)	-0.0001 (0.0004)	-0.0002 (0.0004)
Minimum_nights	-0.158*** (0.010)	-0.201*** (0.018)	-0.144*** (0.010)
Availability_365	-0.0003*** (0.0001)	-0.0003*** (0.0001)	-0.0003*** (0.0001)
Tv	0.027 (0.022)	0.032 (0.022)	0.037* (0.022)
Wifi	0.238*** (0.057)	0.230*** (0.057)	0.234*** (0.057)
Air_conditioning	0.215** (0.087)	0.242*** (0.087)	0.232*** (0.086)
Free_parking_on_premises	0.109*** (0.017)	0.100*** (0.017)	0.103*** (0.017)
Heating	0.181*** (0.047)	0.179*** (0.047)	0.166*** (0.047)
Room_typeHotel room	-0.171** (0.078)	0.021 (0.472)	-0.158** (0.078)
Room_typePrivate room	-0.220*** (0.019)	-0.023 (0.078)	-0.278*** (0.020)
Room_typeShared room	-0.157 (0.110)	-0.981 (0.679)	-0.218** (0.109)
Host_is_superhost	0.324*** (0.013)	0.327*** (0.013)	0.334*** (0.013)
Price:Minimum_nights		0.0002*** (0.0001)	
Price:Room_typeHotel room		-0.002 (0.002)	
Price:Room_typePrivate room		-0.003*** (0.001)	
Price:Room_typeShared room		0.004 (0.004)	
Minimum_nights:Room_typeHotel room		-0.026 (0.400)	
Minimum_nights:Room_typePrivate room		-0.042 (0.049)	
Minimum_nights:Room_typeShared room		0.898 (0.625)	
Price:Minimum_nights:Room_typeHotel room		0.001 (0.002)	
Price:Minimum_nights:Room_typePrivate room		0.001*** (0.0004)	
Price:Minimum_nights:Room_typeShared room		-0.005 (0.004)	
Beds:Price			0.0002*** (0.00003)
Bedrooms:Price			0.0004*** (0.00005)
Bedrooms:Beds			0.005*** (0.001)

Bedrooms:Beds:Price			-0.00004*** (0.00001)
Constant	2.271*** (0.095)	2.316*** (0.098)	2.458*** (0.096)
Observations	5,235	5,235	5,235
R2	0.218	0.227	0.237
Adjusted R2	0.215	0.223	0.234
Residual Std. Error	0.461 (df = 5216)	0.459 (df = 5206)	0.456 (df = 5212)
F Statistic	80.786*** (df = 18; 5216)	54.681*** (df = 28; 5206)	73.520*** (df = 22; 5212)
Note:			*p<0.1; **p<0.05; ***p<0.01

Beta coefficients for Model 2

Beta coefficients	Values
Accommodates	0.019***
Guests_included	0.0003
Bathrooms	-0.102***
Bedrooms	-0.073***
Beds	0.017***
Price	-0.001***
Extra_people	-0.0001
Minimum_nights	-0.201***
Availability_365	-0.0003***
Tv	0.032
Wifi	0.230***
Air_conditioning	0.242***
Free_parking_on_premises	0.100***
Heating	0.179***
Room_typeHotel room	0.021
Room_typePrivate room	-0.023
Room_typeShared room	-0.981
Host_is_superhostTRUE	0.327***
Price:Minimum_nights	0.0002***
Price:Room_typeHotel room	-0.002
Price:Room_typePrivate room	-0.003***
Price:Room_typeShared room	0.004
Minimum_nights:Room_typeHotel room	-0.026

Minimum_nights:Room_typePrivate room	-0.042
Minimum_nights:Room_typeShared room	0.898
Price:Minimum_nights:Room_typeHotel room	0.001
Price:Minimum_nights:Room_typePrivate room	0.001***
Price:Minimum_nights:Room_typeShared room	-0.005
Constant	2.316***

Accommodates, Bathrooms, Bedrooms, Beds, Price, Minimum_nights, Availability_365, Wifi, Air_conditioning, free_parking_on_premises, Heating, Host_is_superhostTRUE, Price:Minimum_nights, Price:Room_typePrivate room, Price:Minimum_nights:Room_typePrivate room, Constant are significant predictors. All these factors are influencing the Occupancy rate and we can predict occupancy rate proportional changes while we change the values by unit or some other values.

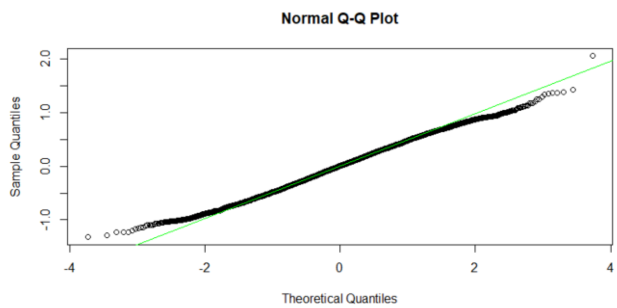
Guests_included, Tv, Room_typeHotel room, Room_typePrivate room, Room_typeShared room, Price:Room_typeHotel room, Price:Room_typeShared room, Minimum_nights:Room_typeHotel room, Minimum_nights:Room_typePrivate room, Minimum_nights:Room_typeShared room, Price:Minimum_nights:Room_typeHotel room, Price:Minimum_nights:Room_typeShared room
are not significant predictors. From these predictors we can suggest host to not to invest money or time on these predictors.

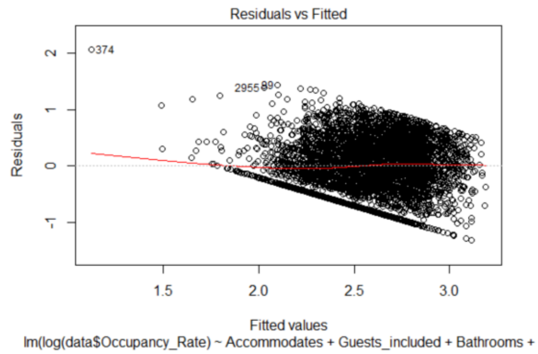
Through this model we can say about our hypothesis.

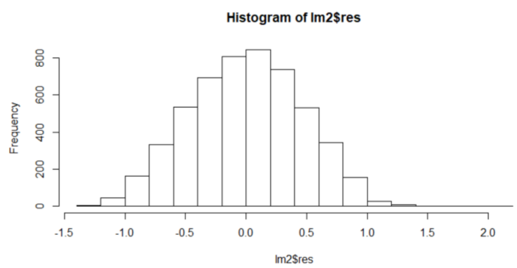
Hypothesis	Accepted/ Rejected	Significance	Inference
H1	Rejected	-	The chances of choosing a Private room when compared Hotel room is less.
H2	Accepted	Medium	As the price of listing increases the chances of booking are less resulting in low occupancy rate.
H3	Accepted	High	As the minimum number of nights to be booked increases the occupancy rate decreases.
H4	Rejected	-	Guests included in the base price does not seem to have any major effect on the occupancy rate.
H5	Accepted	High	Super host status increases occupancy rate at high scale and we can consider this as our top 5 priorities.
H6	Accepted	Low	As the number of bedrooms increases the occupancy rate of the listing will reduce.
H7	Accepted	Medium	As the number of bathrooms increases the occupancy rate of the listing will reduce.

7.Quality Checks

The model has been built using the logistic Regression approach, the following assumptions need to be satisfied to ensure the quality of the Analysis:







Linearity:
The model does not follow the linearity.
Homo Scedasticity:
The model error term is not the same across all independent variables.
Multivariate Normality:
The model residuals are not distributed normally
Multi-collinearity:
All the interaction variables in the model are dependent on each other. Hence, this assumption has also not satisfied.
Autocorrelation:
All the analysis has been performed on the data and the model has successfully passed the autocorrelation test. Hence there is no chance of Autocorrelation in this analysis. The below table shows the model evaluation methods.

Models	**Linearity Test**	**Multi Variate Normality Test**	**Homo Scedasticity Test**	**Multi collinearity Test**	**Auto correlation Test**
Model 1	Failed	Failed	Failed	Passed	Passed (1.85)
Model 2	Failed	Failed	Failed	Failed	Passed (1.85)
Model 3	Failed	Failed	Failed	Failed	Passed (1.824)

9.Recommendations

From our analysis, we can make few recommendations to the hosts which can be implemented to improve the Occupancy rate of the Listings.

- From the results of our analysis, we see that the price is indeed a crucial factor for the occupancy rate. As the price changes, the occupancy rate is also changing. A unit change in price happens, Occupancy rate decreased by 0.995. Based on this we can recommend hosts to maintain a minimum or optimal price depending on the scenario.
- From the statistical results, it is evident that the Amenities are playing a crucial role in generating the high occupancy rate. Maintain the most essential amenities in each property to generate good reviews and income.
- One of the well-known facts that are further established from our analysis is the Room types are important to Draw the attention of the customers. A unit change in the private room decreases the occupancy rate by 0.941. Customers would be interested in taking private rooms most of the time.
- One of the main factors is the minimum nights. Based on the model, it is significantly affecting the occupancy rate. A unit change in minimum nights would increase the occupancy rate by 1.87. It is a minimum requirement for a listing to keep Minimum_nights.
- From the observation, Super host is practically important in calculating the Occupancy rate(Super host means the host is trustworthy in the model) If, the host of the listing has the ability to maintain the good relationship between the customers, there is a favorable chance to gaining the more income from the same customers again and again. A unit change in the super host increases the occupancy rate by 1.38.
- In addition to the above-mentioned recommendations, it is equally important not to focus on unfavorable predictors like Guest included, extra fee per person.

These are some of the recommendations that our model predicts to improve the occupancy rate of listings.

10.References

Various sources helped us in building this project, to mention few of them are

Statistical tools for high-throughput data analysis: - http://www.sthda.com/english/.

Inside Airbnb: - http://insideairbnb.com/get-the-data.html

Occupancy Rate: - https://medium.com/mashvisor/what-is-airbnb-occupancy-rate-mashvisor-f73a7c9f1c8b

Andrew H. Van De Ven (October, 1976). On the Nature, Formation, and Maintenance of Relations among Organizations. Journal of The Academy of Management Review, Vol. 1, No. 4, pp. 24-36.

Atkinson, A. (1998). Answering the Internal Question: What Does the Customer Want?. The Cornell Hotel and Restaurant Administration Quarterly, 29(2), pp. 12-14.

Bell, R.A. and Morey, R.C. (1997). Are You in the Book? Hotel Attributes Bundles and Corporate Travel Department. Cornell Hotel & Restaurant Administration Quarterly, 55-61.

Carman, M.J. (1990). Consumer Perceptions of the Service Quality: Assesment of SERVQUAL Dimension. Journal of Retailing, 66, 33-55.

Cronin J. Joseph, Jr. and Steven A. Taylor (1992). Measuring Service Quality : A Re-examination and Extension. Journal of Marketing.

11.Appendix: Code

Since we used most of the data cleaning and wrangling techniques in Python so, the Complete code of this project which includes Data files, Data Wrangling, EDA and Visualization, Model building is available at https://github.com/sathishreddyjagan/Airbnb-Austin.

R code for linear regression and OLS assumptions:

```
data = read.csv ('C:/Users/ompra/OneDrive/Documents/QMB/project/Data_with_out-dummies.csv')
str(data)
hist(data$Occupancy_Rate)
hist(log(data$Occupancy_Rate))
str(data)
nrow(data)

lm1 = lm (log(data$Occupancy_Rate)~Accommodates+Guests_included+Bathrooms+
Bedrooms+Beds+Price+Extra_people+Minimum_nights+
Availability_365+Tv+Wifi+Air_conditioning
+Free_parking_on_premises+Heating+Room_type+
Host_is_superhost,data = data)

summary(lm1)
hist(lm1$res)

lm2 = lm (log(data$Occupancy_Rate)~Accommodates+Guests_included+Bathrooms+
Bedrooms+Beds+Price+Extra_people+Minimum_nights+
Availability_365+Tv+Wifi+Air_conditioning
+Free_parking_on_premises+Heating+Room_type+
Host_is_superhost+Price*Minimum_nights*Room_type,data = data)
summary(lm2)
hist(lm2$res)

lm3 = lm (log(data$Occupancy_Rate)~Accommodates+Guests_included+Bathrooms+
Bedrooms+Beds+Price+Extra_people+Minimum_nights+
```

```
Availability_365+Tv+Wifi+Air_conditioning+Free_parking_on_premises+Heating+Room_type+
Host_is_superhost+Price*Beds*Bedrooms,data = data)
summary(lm3)
hist(lm3$res)

library(stargazer)
stargazer(lm1,lm2,lm3,type = "text")

#'Assumptions
#'lm1,lm2,lm3

hist(log(data$Occupancy_Rate))

hist(lm1$res)
#Multi variate normality
qqnorm(lm1$res)
qqline(lm1$res, col="green")

shapiro.test(lm1$res)

norm  = rnorm(5235)
ks.test(norm, lm1$res)

#Homoscedasticity
library(car)
plot(lm1)
plot(lm1$fitted.values,log(data$Occupancy_Rate))
abline(0,1,col = 'red')
bartlett.test(list(lm1$res, lm1$fit))

#Multicoliniarity
vif(lm1)

#Auto correlation
library(lmtest)

dwtest(lm1)


#lm2 model
#Multi variate normality
qqnorm(lm2$res)
qqline(lm2$res, col="green")

shapiro.test(lm2$res)
```

```
norm  = rnorm(5235)
ks.test(norm, lm2$res)

#Homoscedasticity
library(car)
plot(lm2)
bartlett.test(list(lm2$res, lm2$fit))

#Multicoliniarity
vif(lm2)

#Auto correlation
library(lmtest)

dwtest(lm2)


#lm3 model
#Multi variate normality
qqnorm(lm3$res)
qqline(lm3$res, col="green")

shapiro.test(lm3$res)

norm  = rnorm(5235)
ks.test(norm, lm3$res)

#Homoscedasticity
library(car)
plot(lm3)
bartlett.test(list(lm3$res, lm3$fit))

#Multicoliniarity
vif(lm3)

#Auto correlation
library(lmtest)

dwtest(lm3)
```