airbnb

Chicago Listing

Sumeet Kukreja

TABLE OF CONTENT

Executive Summary

This report elucidates the statistical analysis for the data of Airbnb hosts listings in Chicago. The project's goal is to determine the **important factors** that affects the **price of an Airbnb listing in Chicago region**.

Our methodology starts with obtaining the data from **"Inside Airbnb"**, an independent, non-commercial entity that provides data for Airbnb listings for major cities around the world. Our team has explored the data set of Chicago listings provided by Inside Airbnb. We then moved to cleaning the data. The original data set had 5207 observations and 95 variables which after removing the rows where a lot of variables had NA values was reduced to **4018 observations and 28 variables**. The cleansed data includes additional metrics like number of crimes reported, gender and Walkscore which either we obtained from the internet or did feature engineering to extract some features out of the data taken from Inside Airbnb.

The subsequent section defines our research question and hypotheses. We have made an attempt to find the important factors that affect the price of a listing in Chicago. We formulated three hypotheses based upon our previous analysis.

Our analysis begins with determining the dependent and independent variables that will help us in finding results to our research question and hypotheses. The dependent variable is the price of a particular listing and independent variables include number of reviews, number of accommodates, bedrooms, bathrooms, room type and cancellation policy which we decided by doing the univariate, bivariate analysis. The control variable is Gender and Chicago region. Chicago region is divided into 4 parts-north, east, west and south- based on the neighbourhood.

In the next section, we have shown the statistical tests results that were performed for checking our hypotheses. The tests include t.test, chi-square test and analysis of variance (ANOVA) test. Finally, regression models were built where we ensured that we take care of the assumptions of the regression and take out the outliers to explain the maximum variance for the model built.

The final section describes the results based on our findings through the various statistical tests performed. Based on that we have made conclusions to our findings and answering our hypotheses.

Introduction

About Airbnb and our data

Airbnb is an online marketplace that enables customers to rent short-term stays. The company does not own any property and cost of the accommodation is set by the property owner. The company receives percentage service fees from both guests and hosts for every booking. It has over 2,000,000 listings in 34,000 cities and 191 countries. We have obtained the data set from Inside Airbnb. This data is sourced from publicly available information from the Airbnb website.

We don't have a particular time frame for our dataset. Originally the dataset contained 95 variables and 5207 observations (This dataset can be found in the **Rdata file** that we have provided alongside the report). Since most of the data was text, we didn't use it. Analysing that would be out of the scope of this project as it would require text mining techniques. After cleaning and adding external data, we came up with the below described dataset which will be used for Statistical analysis. This dataset has **28 variables and 4018 observations**.

Variable	Type	Description
Id	Integer	Unique identifier for the observations
Host_Id	Integer	ID of the Host
Host_name	Character	Name of the host
Gender	Factor	Gender of the Host
Host_total_listings_count	Numeric	Total listings per host
Neighbourhood_cleansed	Factor	Neighbourhood where the listing is posted
Latitude	Numeric	Latitude of the listings
Longitude	Numeric	Longitude of the listings
Property_type	Factor	Type of the property
Room_type	Factor	Type of room
Accommodates	Factor	Number of occupants available for a listing

Bathrooms	Factor	Number of bathrooms
Bedrooms	Factor	Number of bedrooms
Beds	Factor	Number of beds
Bed_type	Factor	Type of the bed
Price	Numeric	Price of the Listing
Minimum_night	Factor	Minimum nights available for a particular listing
Availability_30	Numeric	Availability of Listings in 30 days
Availability_60	Numeric	Availability of Listings in 60 days
Availability_90	Numeric	Availability of Listings in 90 days
Availability_365	Numeric	Availability of Listings around the year
Number_of_reviews	Numeric	Total number of reviews for a particular listing
Cancellation_policy	Numeric	Level of cancellation policy of listings
Chicago_region	Factor	Division of neighbourhoods in East, West, North and South regions
Zipcode	Factor	Zip code of the different listings
Multiple_Host	Factor	Value is 1 if the listing has multiple hosts, else 0
No_Crime_Reported	Numeric	Number of crimes reported for a particular zip code
Walkscore	Numeric	Walkability of any area based on zip code

Scope and Limitations of the dataset:

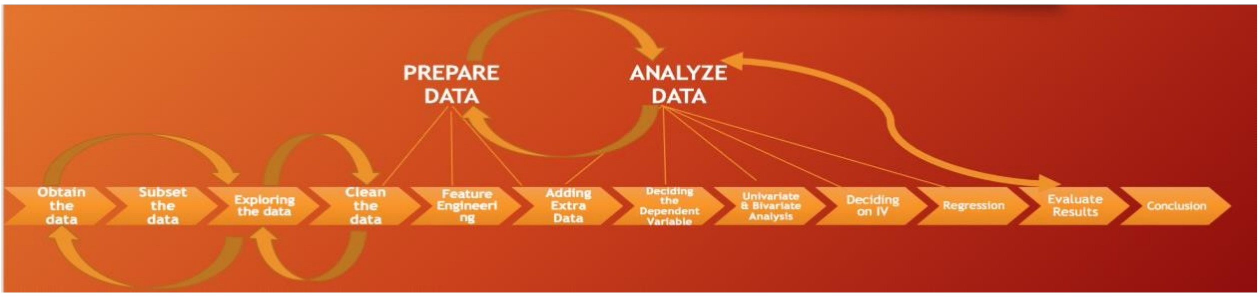
***Scope*: In this report, we have tried to incorporate the techniques like Univariate analysis, Bivariate analysis, Hypothesis formulation, ANOVA, Correlation test and Regression which covers the whole spectrum of this project and defines the scope of our work.

***Limitations*:**

1. **Time of year:** Seasonality, e.g. Christmas holidays, affects the price of a particular listing which is not provided in the dataset. Since listing price is seasonal in nature, it could be predicted more accurately if the time of the year was also present but unfortunately, we don't have any variable which tells us about that.
2. **Sentiment of the review:** Our dataset has information on the number of reviews for each of the listings. There is no information regarding the sentiments of the review. It would help to model the price more accurately if the positivity/negativity of the review could be measured via text or ratings of the review.
3. **Time Frame of the data:** We do not have a time period for the data as to when it was collected. It could have helped us in understanding the variations in the price over a period of time. As of now, we will be assuming the price that was taken at a single point of time.

Workflow

The workflow of our project has followed the below mentioned methodology:



Data cleaning

The original data set contained a lot of missing values for price, neighbourhood, zip code, bathrooms and bedrooms. The rows containing NA values for these variables were removed. Hence reaching 4018 rows from original 5207 rows.

In addition, there were columns that had numeric values but we converted them to factors as they had only limited number of values for a lot of factors, hence converting them to factors with 2 or 3 levels made sense. The following variables were converted into their corresponding types as required for our analysis:

- **Gender** – Character to factor
- **Price** – character to numeric after removing '$' sign. Also, did a log transformation after this
- **Minimum Nights** – numeric to factor. Here, a lot of levels had very few values, so, we made 2 levels after converting minimum_nights to a factor. The 2 levels were: '1' and 'More than 1'
- **Room type** – numeric to factor
- **Accommodates** – numeric to factor. Again, combined a lot of levels and finally, accommodates factor variable had 2 levels. 'Atmost 2', 'More than 2'
- **Bed type** – character to factor
- **Cancellation policy** – character to factor (level 'super_strict_30' or 'strict' were combined to one level, 'super_strict' as 'super_strict_30' were very few in number
- **Zip code** – numeric to zip code

Feature Engineering

We have performed feature engineering in our dataset by creating **Gender, Chicago Region, Multiple_host** variables. We used the **package gender** to determine the gender of a host by his/her name. In case, the listing is put up by more than individual, then we have assigned 'Both' as the 3rd level for the factor variable gender. Based on the similar idea, a listing posted by more than one individual is assigned a value of 1 and the one posted by an individual a value of 1 for the variable multiple_host. Also, we wanted to do our analyses on Chicago region rather than neighbourhoods because there were multiple neighbourhoods, so, we have divided the Chicago region into North, South, East, West.

Adding External Data

We have added the **'Walkscore'** and **'Number of crimes reported'** variables which were obtained on the basis of zip code and neighbourhood respectively. Walkscore measures the **walkability of any address**, it also measures the **pedestrian friendliness** by analysing **population density and road metrics** such as block length. We got this score from www.walkscore.com. We felt that this is a crucial factor whenever someone is looking to book an Airbnb listing and might impact price of that listing too. **Number of crimes** reported reveals the average number of crimes reported in a month in a particular **zip code**. Again, we made use of the various crime data posted on the internet for Chicago region and mapped it with the respective zip codes. This was done because crime rate is also an important factor which customers look into when making any booking.

Research Question and Hypotheses

***Research question**:* Which are the important factors that affect the price of an Airbnb listing in a Chicago region?

Hypotheses:

Hypothesis 1: The average number of reviews across the Chicago region is same.

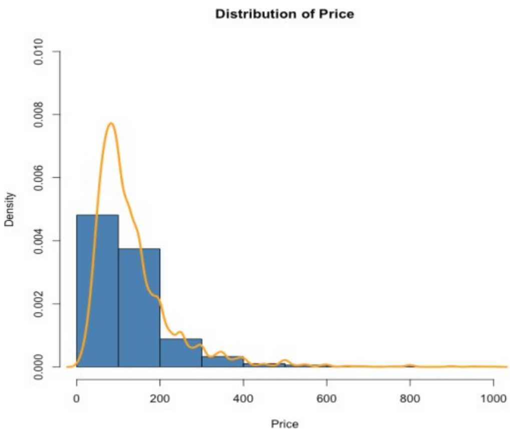
Hypothesis 2: The cancellation policy is uniform across the gender of the hosts.

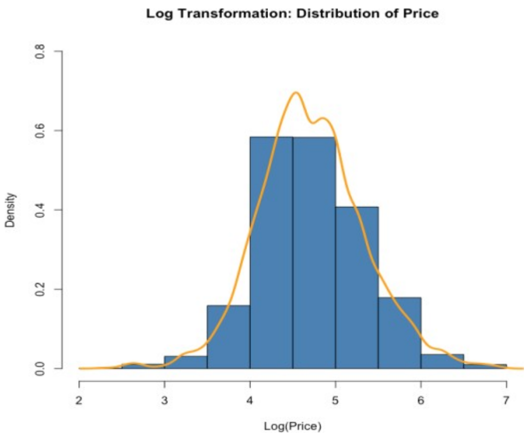
Hypothesis 3: The price of a listing in a Chicago region is higher when the number of reviews are more.

Univariate and Bivariate Analysis

Univariate:

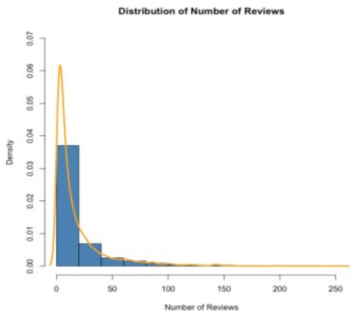
Univariate Analysis: Price Distribution

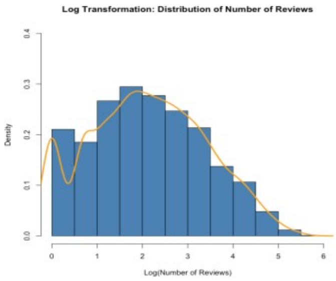




Price is **right skewed** therefore we used log transformation to make it **normally distributed**. **Price** is our **Dependent Variable (DV)** in the regression model and one of the main assumptions of Regression is for the dependent variable to be normally distributed, so, we have made sure that the distribution of price is normal.

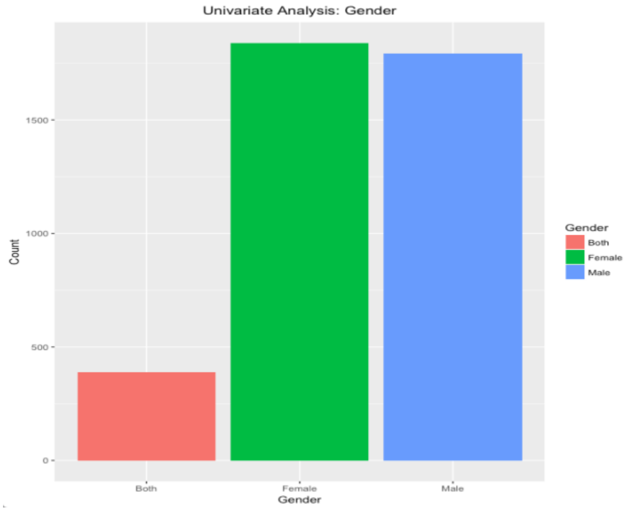
Univariate Analysis: Number of Reviews





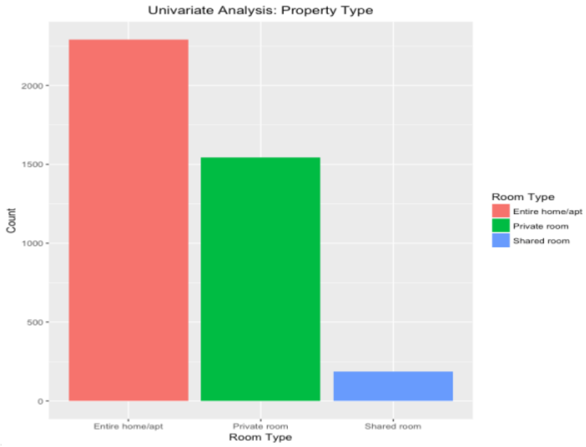
Number of reviews is **right skewed** and therefore we have performed a log transformation to make it into a normal distribution. Second plot displays the normal distribution of number of reviews.

Univariate Analysis: Gender



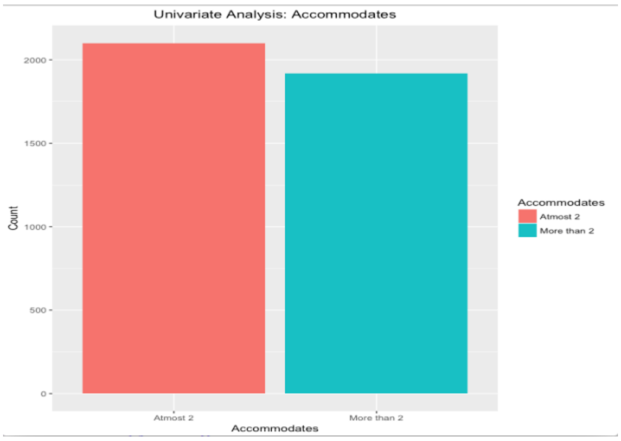
Number of listings by male host and female host is almost same across all Airbnb listings in Chicago.

Univariate Analysis: Property Type



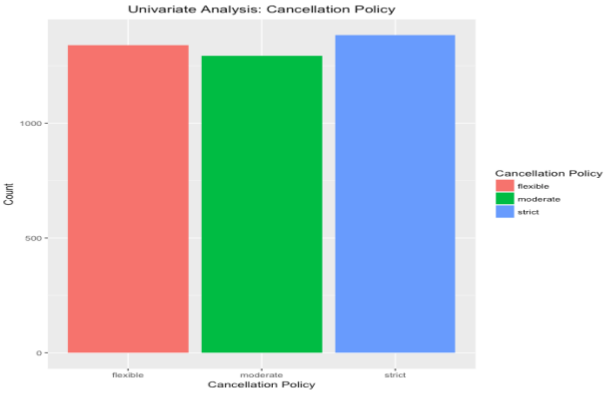
This plot demonstrates that ‘Entire home/apt’ category has maximum count across Airbnb listings in Chicago.

Univariate Analysis: Accommodates



Accommodates variable have been divided into 2 categories**: ‘Atmost 2’ and ‘More than 2’**. Accommodates had values like 3, 4 and 5 which were too less in number and similar case was there for values like 0,1,2. As a result to obtain a substantial percentage for some levels and to use this variable in the regression, we made 2 levels for this variable.

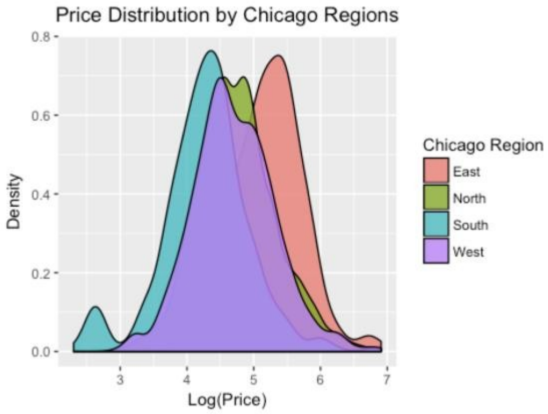
Univariate Analysis: Cancellation Policy



The cancellation policy is almost same among three categories: **Flexible, Moderate and Strict**.

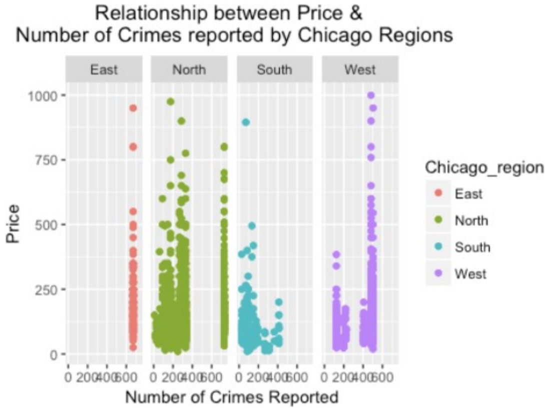
Bivariate Analysis

Price distribution by Chicago Region



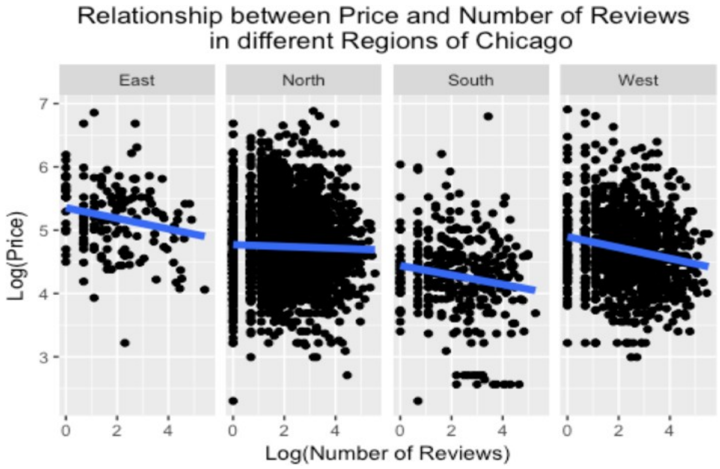
The distribution of log(price) for all the regions of Chicago is normally distributed.

Price & Number of Crimes Reported by Chicago Region



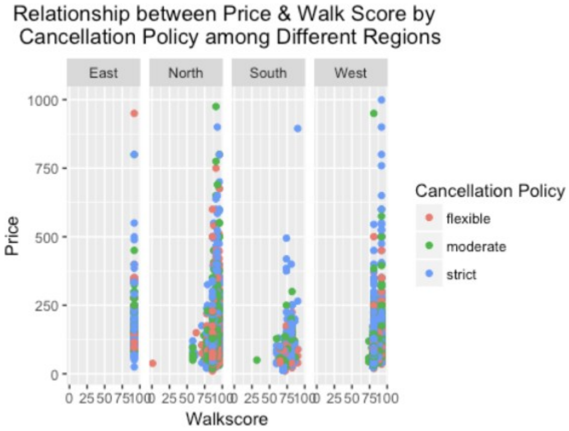
This shows the relationship between price and number of crimes in the various Chicago region. The number of the crimes reported in North are less. Also, as the number of crimes reported are increasing, the price is increasing as well but that's not something which is consistent across all the Chicago regions.

Price & Number of Reviews by various regions of Chicago



Running a bi-variate analysis between the number of reviews and price across all the regions of Chicago shows that there is a negative relationship between the price and reviews. This means that the increase in number of reviews decreases the price of a listing, which also suggests that probably the sentiment of the reviews is mostly on the negative side.

Price Vs Walk Score



Walk Score is a metric of walkability access in a particular neighbourhood. The bi-variate analysis of price and walk score for different regions of Chicago shows that the price increases with regions of higher walk score suggesting that price of a listing seems to increase with the convenience a visitor will have near the listing which makes complete sense.

Statistical Test

Analysis of Variance (ANOVA)

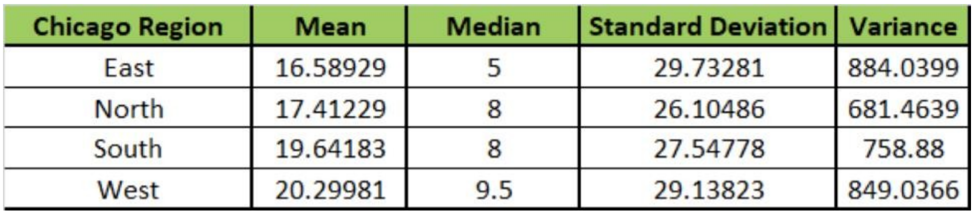
To test our **first hypothesis**, we did an ANOVA test on logarithmic of number of reviews(numeric) and Chicago region(factor).

Null Hypothesis: The average number of reviews across different Chicago regions is same

Alternate Hypothesis: The average number of reviews across different Chicago regions is not same

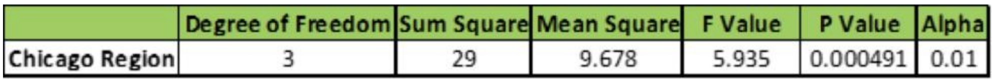
Before conducting this test, we made sure the following requirements were satisfied:

- Population distribution is approximately normal
- Variance in groups is about equal
- Sample sizes are adequate
- Samples sizes are independent



Chicago Region	Mean	Median	Standard Deviation	Variance
East	16.58929	5	29.73281	884.0399
North	17.41229	8	26.10486	681.4639
South	19.64183	8	27.54778	758.88
West	20.29981	9.5	29.13823	849.0366

The results of the ANOVA test are shown below:



	Degree of Freedom	Sum Square	Mean Square	F Value	P Value	Alpha
Chicago Region	3	29	9.678	5.935	0.000491	0.01

The p-value of 0.00049 is low and less than alpha (0.01). Since **p-value < alpha**, we will **reject our null hypothesis**.

This means that the average number of reviews across different Chicago regions is not same. To further understand where this difference is most significant, we do a post ANOVA test i.e. a Tukey HSD test. The result of this test is shown below:

Chicago Region	Difference	Lower	Upper	P-value
North-East	0.24481335	-0.072420101	0.5620468	0.0764511
South-East	0.30625909	-0.06729153	0.6798097	0.0521689

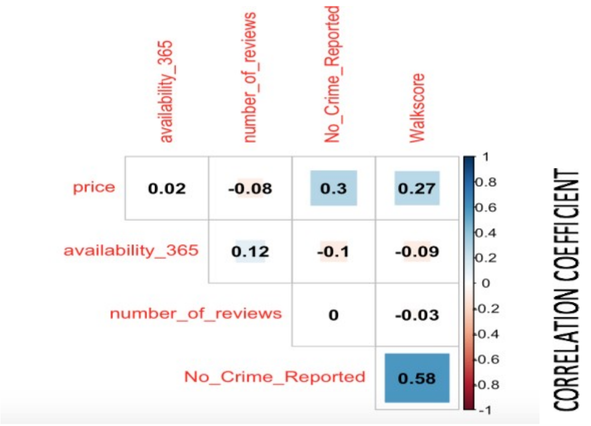
West-East	0.38986777	0.059180125	0.7205554	0.0013881
South-North	0.06144574	-0.166116444	0.2890079	0.8348028
West-North	0.14505441	-0.001910876	0.2920197	0.0113964
West-South	0.08360868	-0.162362456	0.3295798	0.714585

The results show that there's a **significant difference in the number of reviews** between **West-East, West-North**

region. This is a good finding, so it made a lot of sense to use the post-Tuckey test.

Correlation

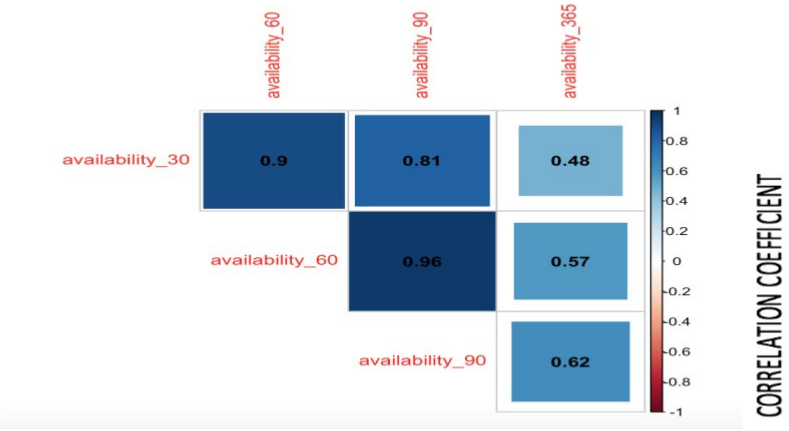
To test our **third hypothesis**, we construct a correlation matrix:



We have intentionally removed all the diagonals as the correlation of any variable with itself is always going to be 1. Although, the correlation between price and number of reviews is not much but it is still -0.08, so, atleast it gives us the idea that as the number of reviews increases, the price goes down, but suggests that probably the reviews were mostly negative regarding any listing.

If we do a correlation test on price and number of reviews, we get a p-value of 0.00000008 which is less than alpha(0.5) which confirms the relationship between price and number of reviews summarized in the previous paragraph.

Also, we constructed a correlation matrix of the variables availability_30, availability_60, availability_90, availability_365 as we thought that they might have a strong correlation and that is evident when we look at the correlation matrix.



Chi-Square Test

To test our **second hypothesis**, of checking where there is a relationship between cancellation policy and the gender of the host, we use a chi-square test.

After running the chi-square test, we got a p-value of 0.000003701 which again is lesser than the alpha (0.05), which shows that there is an association between the cancellation policy and gender of the host.

Multiple Regression Model

A multiple regression model describes a continuous response variable as a function of one or more predictor variables. It's a very useful method used for predictive analysis and helps in understanding the relationship between a dependent and one or more independent variables. A linear model essentially attempts to fit a single line through a scatter plot. The simplest form of multiple regression model is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Where:

Y -> Dependent variable

b0 -> Intercept when all independent variables = 0

b1, b2… -> Coefficient values of the various independent variables

X1, X2… -> Independent variables

After doing univariate and bivariate analysis, we selected accomodates, number of reviews, cancellation policy, chicago region, gender, bathrooms, number of crime reported, walkscore.

Since, number of reviews was right skewed, we made a log transformation to make the distribution of number of reviews normal.

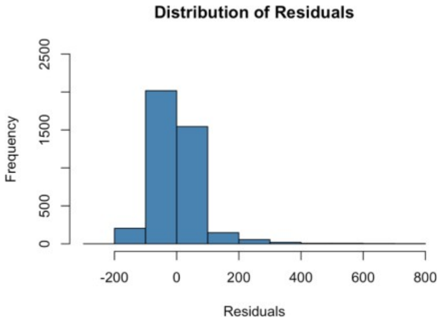
We have built 3 regression models where the 1st model is built with Price as our dependent variable, 2nd model is built with log(price) being our dependent variable and 3rd model is when we have removed outliers from our 2nd model. In all the 3 models, we have kept our independent variables to be the same.

Checking for Multicollinearity:

Also, before running any of our models, we checked for **Multicollinearity**. The **Variance Inflation Factor** for all of our independent variables taken in the regression model was **below 2** and as a result, we have taken care of the **no multicollinearity** condition for our models.

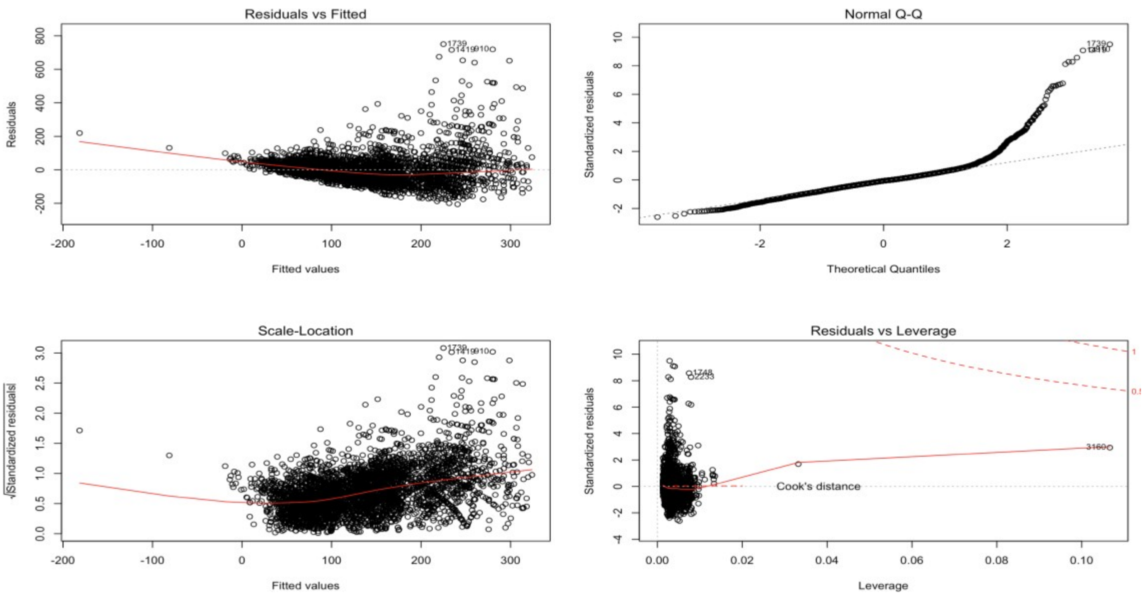
MODEL 1:

In Model-1, we got the Adjusted R-squared value of **40.22%** with most of the independent variables being highly significant. But, we cannot base any of our analysis based on this result because firstly, our dependent variable in this model is right skewed and thereby, not normally distributed which is one of the main conditions of regression.



Even the distribution plot of the residuals is not normally distributed in this one.

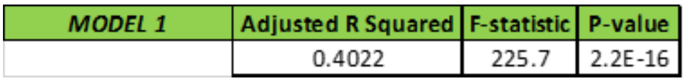
Diagnostic Plots of Model-1:



We don't see the random pattern of the residuals around the horizontal line in plot1 nor do we see a linear relationship from plot2 when we examine the Normal Q-Q plot.

So, this model as expected doesn't tell us anything on which we can base our conclusion on.

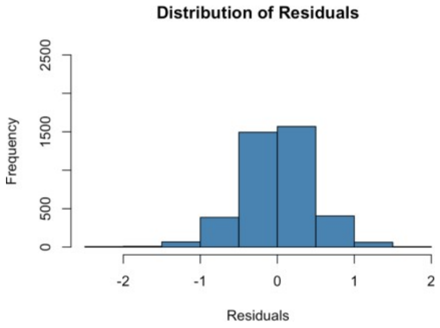
Result:



MODEL 1	Adjusted R Squared	F-statistic	P-value
	0.4022	225.7	2.2E-16

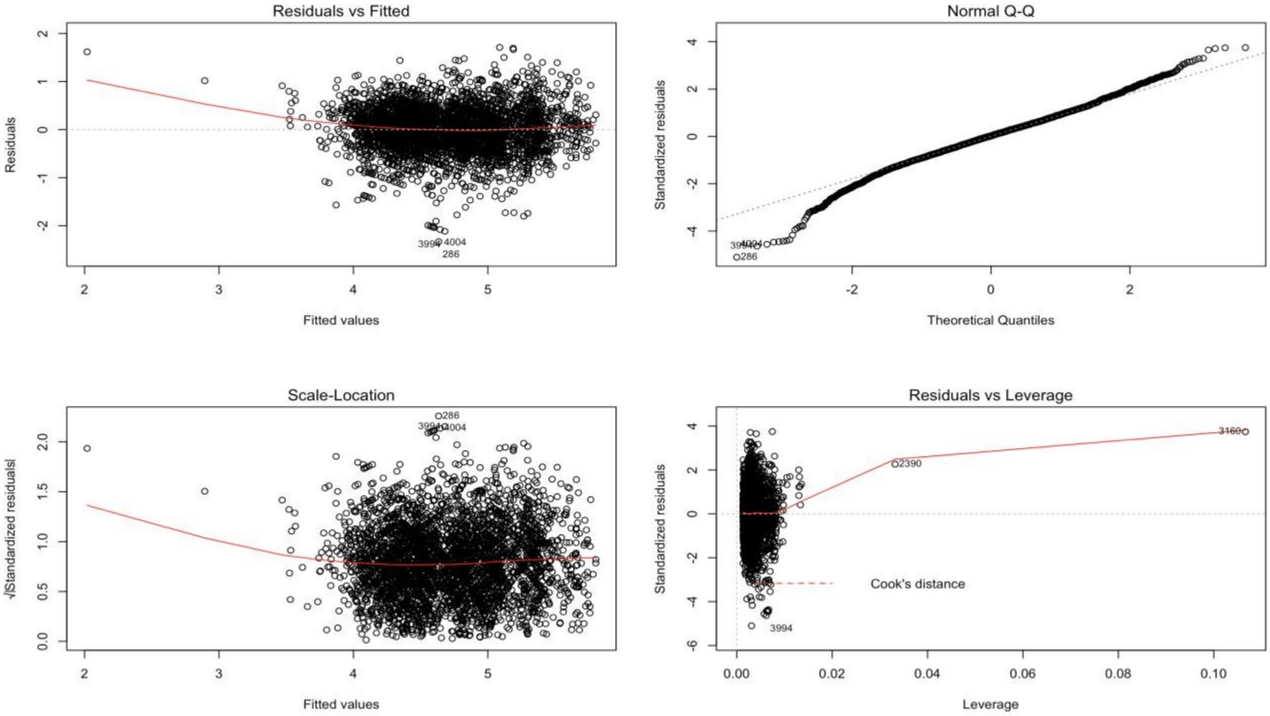
MODEL 2:

In Model-2, now, our dependent variable is log(price) which follows normal distribution. As expected, the adjusted R squared increases to 46.11% which is a good increase.



Now, the distribution of the residuals also becomes normal.

<u>Diagnostic Plots of Model-2:</u>



Compared to the Model-1, the Normal Q-Q plot looks much better and that's expected because we have already seen the distribution of the residuals which looks normal now. Even the spread of the residuals now looks random which was missing in Model-1.

But, from the Residuals vs Leverage plot, it is clear that if we eliminate our outliers, then our model will probably be even better and that would be the right model where we'll be in a state to analyze the different independent variables and their effect on price.

Result:

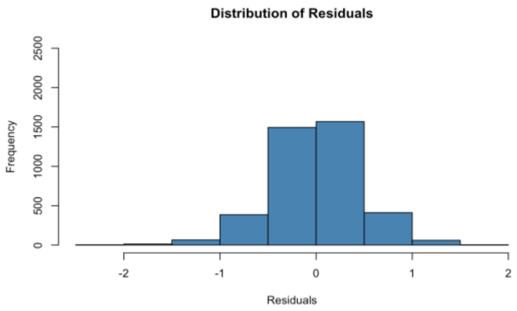
MODEL 2	Adjusted R Squared	F-statistic	P-value
	0.4611	286.7	2.2E-16

<u>MODEL 3:</u>

We then looked at the plot of the hatvalues of model-2 and picked the 2 most influential outliers. The reason why we just picked 2 influential outliers because thes hat values of the 3rd most influential outlier wasn't that high and we infact tried removing that outlier from the model as well but did not see any difference. So, we stuck at removing only 2 outliers.

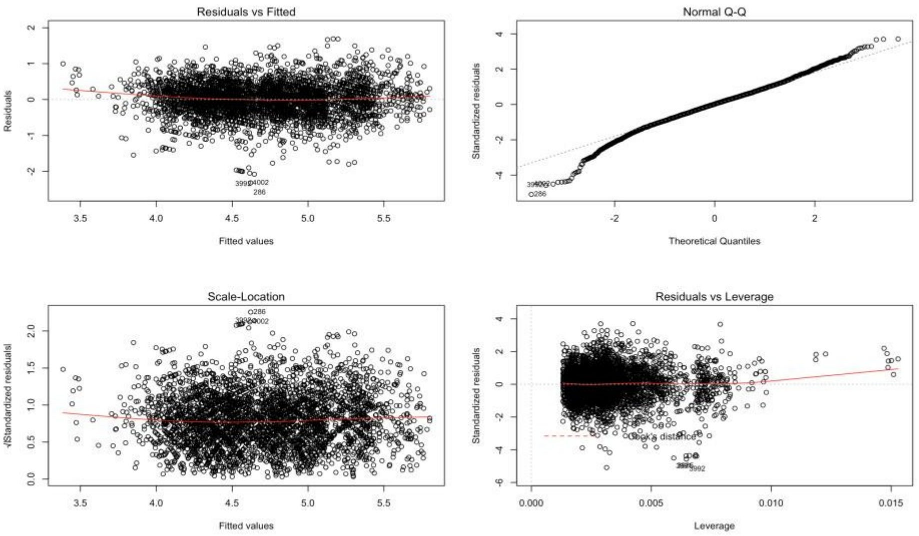
We further analyzed those two outliers to understand why they were different from the other data points and the reason was walkscore. One of them had a walkscore of 1 and the other one had a walkscore of 32 which is quite low as compared to the other listings which had a much higher walkscore.

Now, after removing the outliers, the adjusted R squared value came out to be 46.32% which is not much of an increase from Model-2 but surely bring some improvement which shows the importance of removing outliers from our regression model.



The distribution of the residuals is normal for Model-3.

Diagnostic Plots of Model-3:

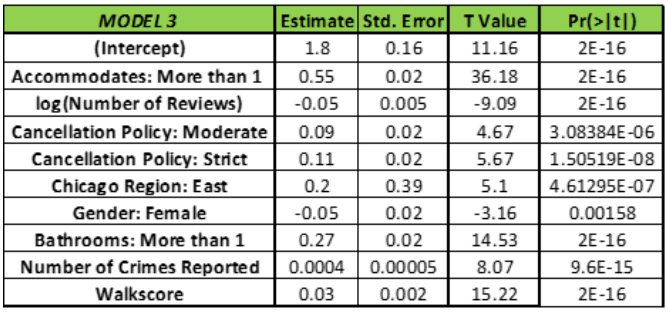


This diagnostic plot looks the best amongst all the models. The Normal Q-Q plot shows normally distributed residuals. We now see equal spread residuals about a horizontal line indicating almost equal variance.

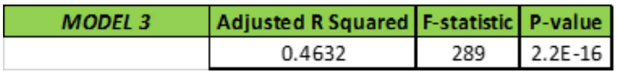
Now, we go about understanding the effect of different independent variables on log(price). Walkscore has a coefficient of 0.0285, now since, price has gone through log transformation, this means that for a unit increase in walkscore, we would expect to see about 2.8% increase in price, since e^0.0285 = 1.028 keeping every other independent variable constant.

The most interesting results come from factors like gender, accomodates, cancellation policy, chicago region etc.

Results:



MODEL 3	Estimate	Std. Error	T Value	Pr(>|t|)
(Intercept)	1.8	0.16	11.16	2E-16
Accommodates: More than 1	0.55	0.02	36.18	2E-16
log(Number of Reviews)	-0.05	0.005	-9.09	2E-16
Cancellation Policy: Moderate	0.09	0.02	4.67	3.08384E-06
Cancellation Policy: Strict	0.11	0.02	5.67	1.50519E-08
Chicago Region: East	0.2	0.39	5.1	4.61295E-07
Gender: Female	-0.05	0.02	-3.16	0.00158
Bathrooms: More than 1	0.27	0.02	14.53	2E-16
Number of Crimes Reported	0.0004	0.00005	8.07	9.6E-15
Walkscore	0.03	0.002	15.22	2E-16



MODEL 3	Adjusted R Squared	F-statistic	P-value
	0.4632	289	2.2E-16

Adjusted R squared: This measure is a predictor of how well the model is fitting with the data points. It explains the % of variance explained by the independent variables used in the model. We got an adjusted R squared value of **46.32%** which is a good number.

Gender: We would expect to see a decrease of about 4.69% in price if the listing is posted by a female as compared to the male, since the coefficient value is e^-0.048 = 0.9531. ((0.9531-1)/1)

Accommodates: We would expect to see an increase of about 72.96% in price when the number of accommodates changes from at the most 2 to more than 2.

Cancellation Policy: We would expect to see an increase of about 11.07% in price when the cancellation policy changes from flexible to strict.

Chicago Region: We would expect to see an increase of about 22% in price when the listing is posted in East Chicago region as compared to the North Chicago region.

Number of Bathrooms: We would expect to see an increase of about 30% in price when the number of bathrooms changes from 1 to More than 1.

So, the most important factors which affect the price of the Chicago listings are: Accommodates, Cancellation Policy, Walkscore, Number of Bathrooms.

Conclusion & Business Implementation

The model which we built can really help the Data Science team of Airbnb to provide recommendations to the hosts whenever they are putting up any listing because now we are the different factors that are impacting the price of listings in Chicago region. Of course, we cannot generalize this model and make a statement about different states because the situation might be totally different here but atleast for Chicago regions, we can back our claims based on the analysis we have done. Also, some of the interesting results which we got by performing some of the tests which were not stated as hypotheses by us were as follows:

- There is no difference in the average number of reviews if a listing is posted by a male as compared to if it's posted by a female.

- There is an association between the minimum number of nights & accommodates. A larger group never preferred to book a listing offering only 1 minimum of nights.

To conclude about the research question which we defined at the start of our project:

Which are the important factors that affect the price of an Airbnb listing in a Chicago region?

The important factors are: accommodates, number of bathrooms, cancellation policy, walkscore, gender etc.

Lessons Learnt

- The number of accommodates a listing offers is really important when it comes to affecting the price of an Airbnb listing in Chicago region

- The number of bathrooms is very important
- Accessibility near the area where you are booking your Airbnb listing is important