IDS 570 - Project Report – Presented By

Alexandrina Almazan

Ashwin Kumar Mathi

Priyadharshini Govindarajan

Sahana Shrivats

Vigneshwaran Giri Velumani

Yokesh Vishwanathan Suresh Kumar

Table of Contents

Introduction

Airbnb is a peer-to-peer online marketplace and homestay network that enables people to list or rent short-term lodging in residential properties, with the cost of accommodation set by the property owner. The company receives percentage service fees from both guests and hosts in conjunction with every booking. It has over 3,000,000 listings in 65,000 cities and 191 countries. Airbnb charges their hosts 6-10% commission of each of the bookings done via the platform and 3% service charge from the guests.

Airbnb was launched in Australia in August 2012 and Melbourne has been one of the early adopters of Airbnb. Melbourne is one of the top 10 cities for global travelers on Airbnb. Ever since it launches, it has provided an opportunity for the hosts to make extra income by listing their properties for Airbnb.

Tourism has been an important contributor for Australia's GDP and Airbnb, Melbourne has contributed roughly $400m a year to the Victorian economy. Melbourne attracts tourists from around the world. It hosts some of Australia's most prestigious events throughout the year such as the Spring Racing Carnival, the Australian Open Tennis, Melbourne Food & Wine Festival, the Formula 1™ Australian Grand Prix, Melbourne International Comedy Festival, international jazz, film and writers' festivals, and Melbourne Festival in October. Melbourne embraces cultures from around the world is a warm and friendly city. Tourists typically enjoy the coastline, sports, food and wine, the cafe culture.

The Melbournians have always been open to the idea of Airbnb and with our analysis, we want to help the Airbnb hosts in Melbourne to decide which factors to consider before finalizing their listing's price.

How does Airbnb work?

Airbnb is a community-based, two-sided online platform that facilitates the process of booking private living spaces for travelers.

The steps below explain how Airbnb exactly works:

1. Hosts list out their property details on Airbnb along with other factors like pricing, amenities provided etc.
2. Travelers search for a property in the city (Melbourne in our case) where they wish to stay and browse available options according to price, amenities etc.
3. Booking is made through Airbnb where traveler pays the amount mentioned by host and some additional money as transaction charges.
4. Host approves the booking. Traveler stays there and finally Airbnb pays the amount to the host after deducting their commission.

The host and the traveler can rate each other and can write reviews based on the experience.

Approach to the Data

The Airbnb data from an online source: tomslee.net/airbnb-data. Professor Tom Slee currently works for SAP SE and writes about technology and society. The original data has 15 variables with 7280 observations. Out of the 15 variables there were 12 numeric and 3 factor levels.

While working with the dataset, some of the variables like latitude, longitude and collected have been removed as they are outside the scope of our analysis. To understand the impact of external factors on the price of an Airbnb listing, we have merged additional datasets to out Airbnb data based on the postal code of the listing. Columns such as number of violent crimes and property crimes in the property's neighborhood were gathered from the Victorian Burglary Statistics. Victoria offers a grant of up to $20,000 for first home buyers buying a new home in regional Victoria, or up to $10,000 for homes in cities. The maximum purchase price of eligible new homes is $750,000. Apart from the crime data, we also want to check whether the amount and number of First Home Owner Grant(FHOG) encourages the Melbournians to host on Airbnb. The FHOG is a one-off payment to encourage and assist first home buyers to buy or build a residential property for use as their principal place of residence. The number of grants and the value of the grants was added from the FHOG data.

The following are preliminary dataset parameters –

• room_id	• bedrooms
• host_id	• bathrooms
• room_type	• price
• city	• minstay
• neighborhood	• latitude
• reviews	• longitude
• overall_satisfaction	• collected
• accommodates	

We have removed the variables highlighted in red during data cleaning as they do not have any significance to our analysis.

Our analysis is primarily based on the variables highlighted in blue.

The additional dataset parameters are as follows –

Postal Code Data	**Crime Data**	**Grants Data**
• postal_code	• postal_code	• postal_code
• city	• violent_crimes	• no_of_grants
	• property_crimes	• grant_value

The below table mentions the final dataset after cleaning up of data –

Variable	**Type**	**Description**
room_id	Integer	Unique ID assigned by Airbnb to identify a room
host_id	Integer	Unique ID assigned by Airbnb to identify the host

room_type	Factor	The accommodation type - Private Room, Shared Room, Entire Home/Apt
city	Factor	The city where the accommodation is located
reviews	Numeric	The total number of reviews given for a particular accommodation
overall_satisfaction	Factor	The customer's rating given for the accommodation
accommodates	Factor	The number of people staying at the accommodation
bedrooms	Factor	The total number of bedrooms in the accommodation
bathrooms	Factor	The total number of bathrooms in the accommodation
price	Numeric	The price per for the accommodation
minstay	Factor	The minimum number of days the accommodation rented for
cityCheck	Factor	Differentiates cities within and outside of scope of the analysis
cityZone	Factor	Categorizes cities as Zone 1: Within 5Km or Zone 2: Btw 6to10Km
hostCount	Numeric	Count of the number of listings for each host in the two zones
hostCategory	Factor	Host count category- low, medium and high
price_cat	Factor	Price Category- High or Low
minstayCategory	Factor	Minstay category- one, two and three or above
postal_code	Factor	Postal code of the city

violent_crimes	Numeric	Total number of Violent crimes
property_crimes	Numeric	Total number of Property crimes
no.of.grants	Numeric	Total number of grants given for the particular postal code
grant.value	Numeric	Total number of grant value given for the particular postal code
total_crimes	Numeric	Sum of Violent Crimes and Property Crimes
overallSatisfactionCategory	Factor	Overall satisfaction category- low, medium and high
totalCrimesCategory	Factor	Total crimes category- low, medium and high

Assumptions

The analysis depends on important assumptions such as –

- The city limits are less than 5 km from Melbourne CBD, and these cities are a part of Zone 1 for our analysis. Majority of the tourist attractions are within the 5-km vicinity including shopping centers, events and multi-cuisine restaurants, hence 5 km was chosen as the city limits for the analysis.
- Cities at 6km-10km from Melbourne CBD are a part of Zone 2 for our analysis
- Cities beyond 10km from Melbourne CBD are not considered for our analysis.
- First Home Owner Grant(FHOG) restricts first time home buyers from renting out their unit for 1 year from the date of occupation.
- Utilization of Airbnb properties in both city zones is the same.

Challenges

The following are the list of challenges that had to be addressed –

- Our analysis compares the factors influencing the Airbnb prices in city and non-city zones. Deciding and coming up with the city limits of Melbourne from our data was our primary challenge.
- The data related to the crimes and FHOG was not available with our original data, nor was it available explicitly. This data needed to be pulled up manually for each postal code. As postal code was not a column in the Airbnb data, the postal code for each city was found out manually.
- The dataset contained a lot of "dirty" data. There were spelling mistakes, null values, and incorrect values in the fields which needed to be cleaned up.
- The unavailability of utilization and time-series data restricts the scope of our analysis.

Research Question

The research question that we are going to solve as part of this project is as follow –

"What factors help Airbnb's hosts (Melbourne) determine its pricing model for its city zones?"

Hypotheses

- The price of an Airbnb listing in Melbourne for the year 2016 is higher in a low crime zone
- Minimum stay for an Airbnb listing affects the price fixed by the hosts in Melbourne for the year 2016

Analysis

Dependent Variable

The dependent variable for our analysis is price of the listing. This price is not determined by Airbnb and it is the host who has full control over deciding the price for each listing. Our analysis will help to improve the pricing pattern for city zones in and around Melbourne. And based on the analysis, recommendations to improve revenue have been suggested to Airbnb. For businesses like Airbnb, price is an important indicator which determines how well the business performs in the longer run, and price is the primary factor that every customer considers before finalizing their booking.

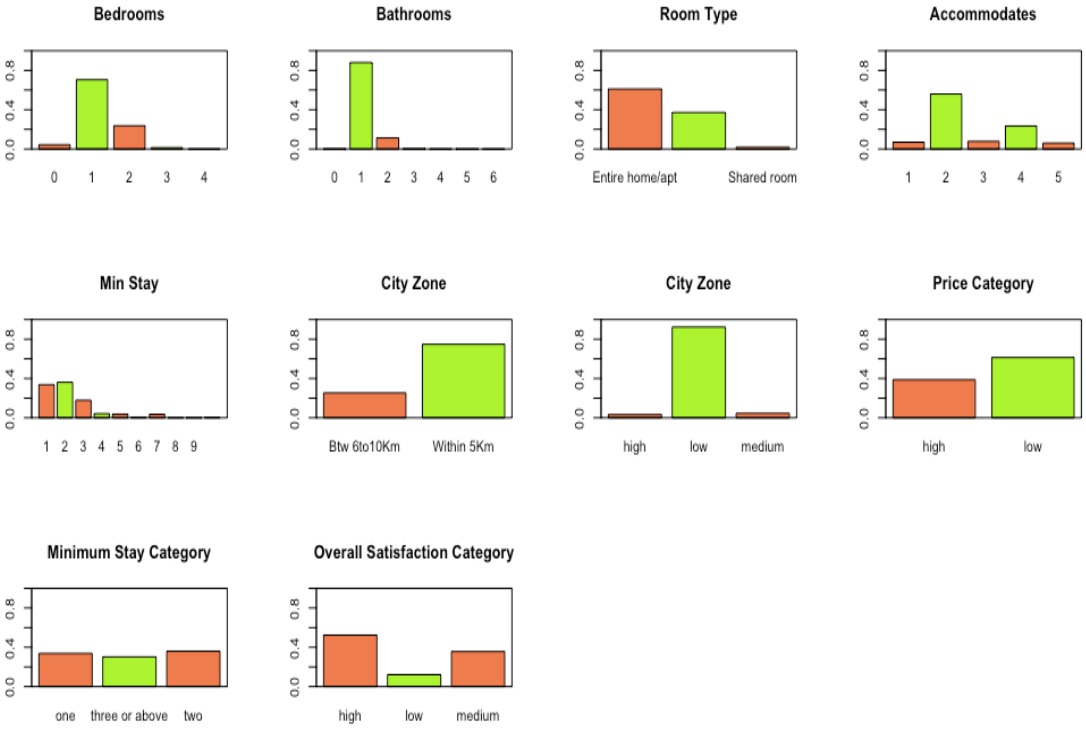
Independent Variable

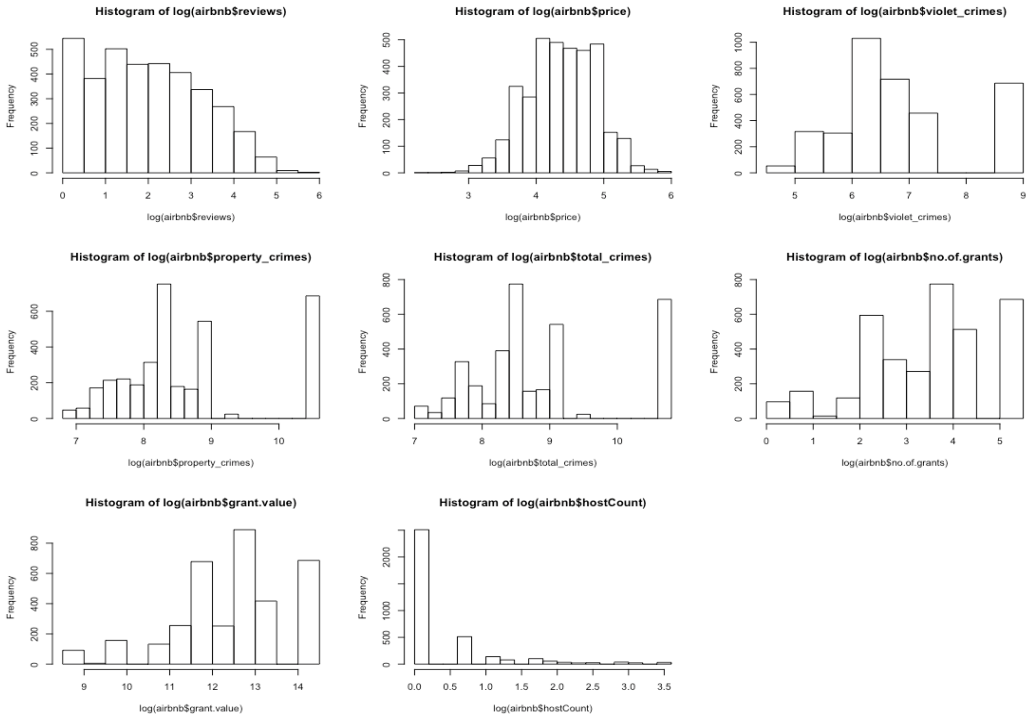
Accommodates, minstay, reviews, total crimes are considered as our independent variables.

Data Cleaning- Merging Data, Removing NAs, Adding Derived Columns and Removing Outliers

Columns which were not required for our analysis were removed. Intensive data cleaning was required as it involved merging 3 datasets- postal code, crime and FHOG Grant along with the original Airbnb dataset. Many null value fields were introduced, which needed to be removed. Columns like price_cat, cityCheck, cityZone, hostCount, hostCategory, total_crimes, minstayCategory, overallSatisfactionCategory and totalCrimesCategory were added. The final part of the data cleaning process was to remove the outliers from the data. This process was done once we obtained the results from the univariate analysis.

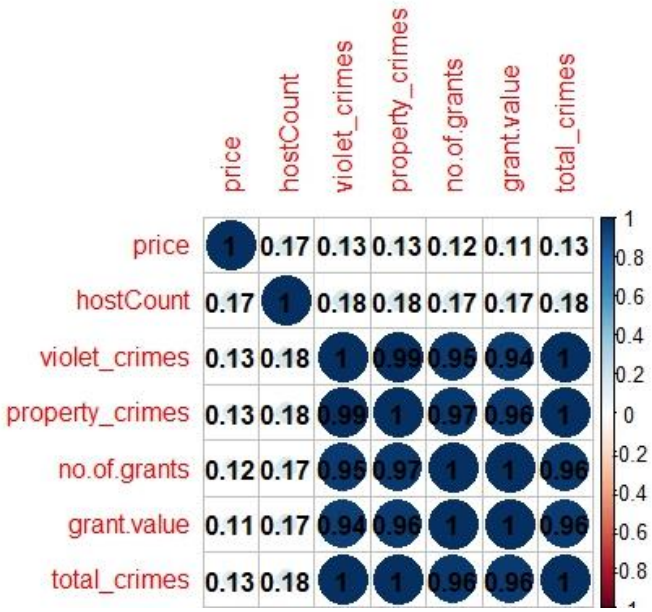
Univariate Analysis



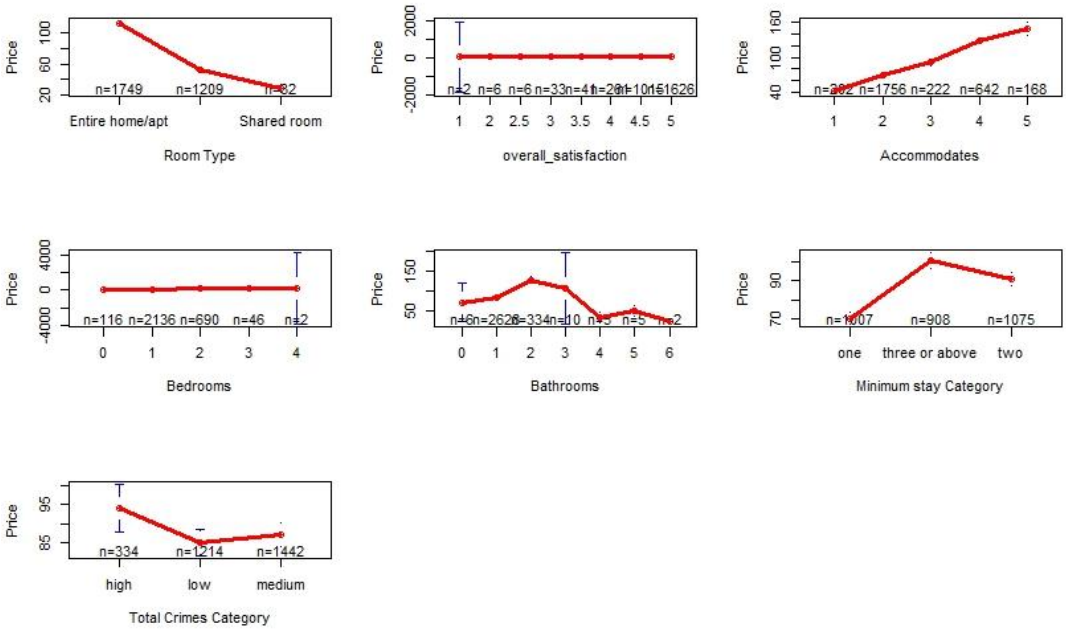


Bivariate Analysis

Price against other numeric attributes

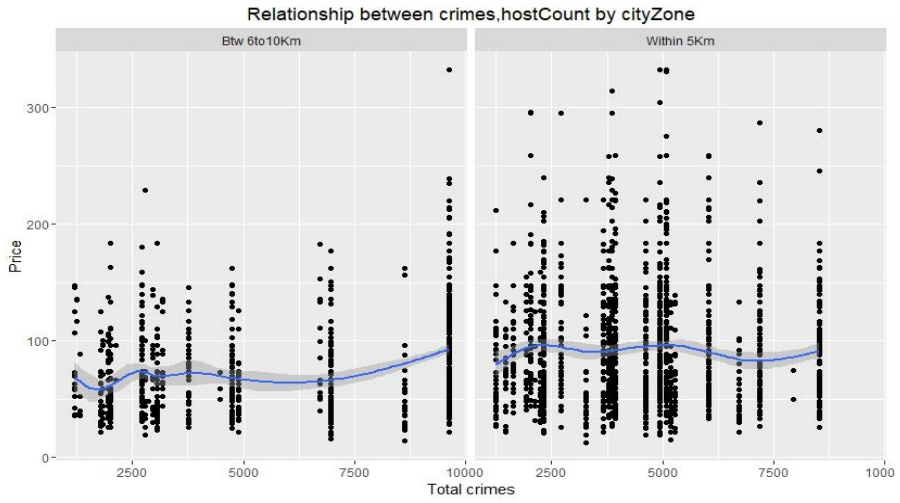


Price against other factors

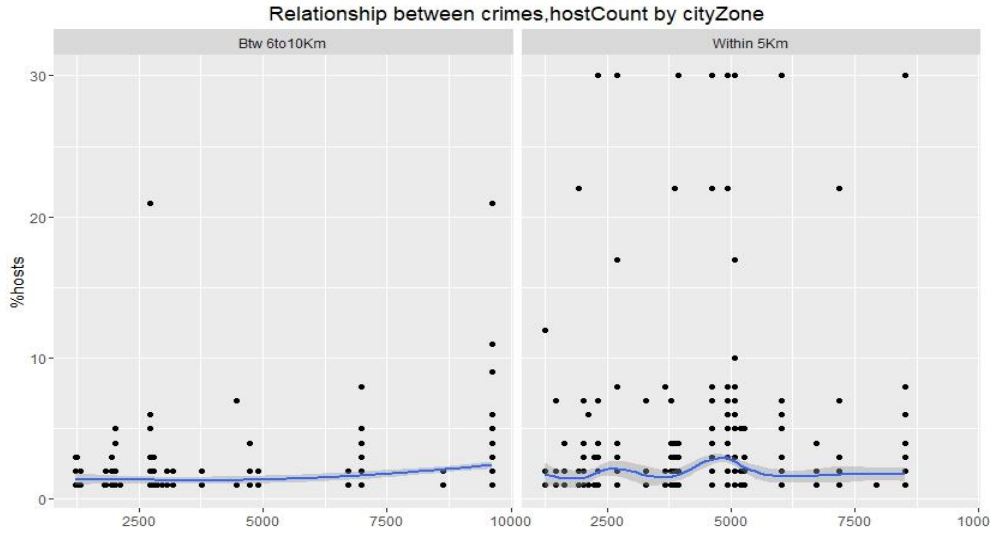


Hypothesis Testing

The qplot shows that the total_crimes in zone 1(within 5 kms) is higher where as in zone 2 (6 to 10 kms) it is comparatively lesser.

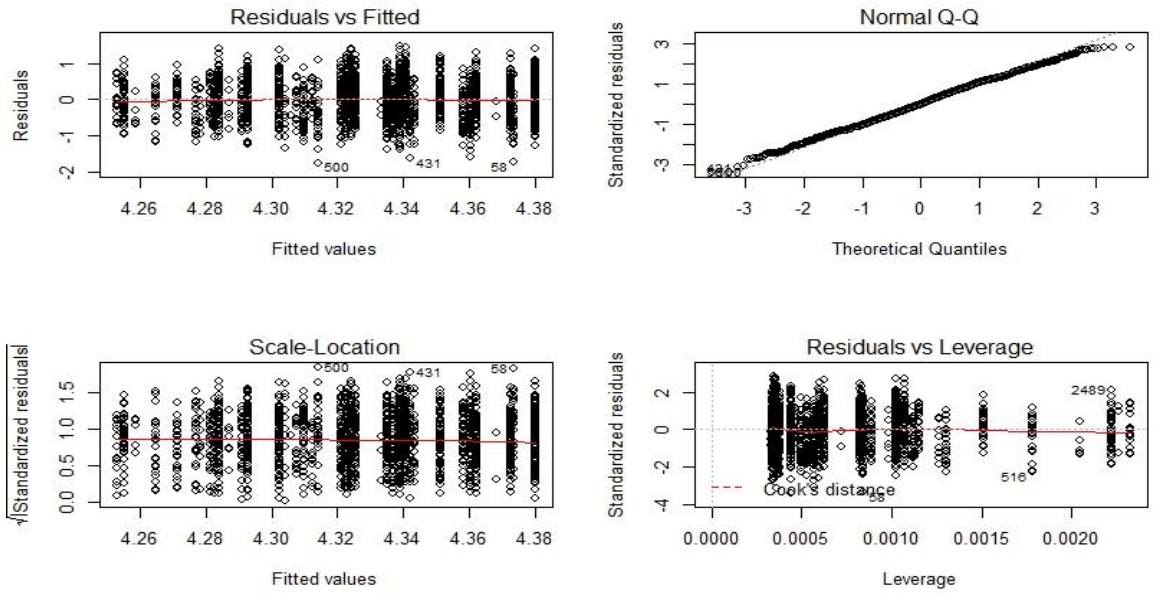


To further support the analysis the total crimes is plotted against the host count. As evident from the plot the host count is more in zone1 which confirms that crime is more within the 5-km range.



Linear Regression

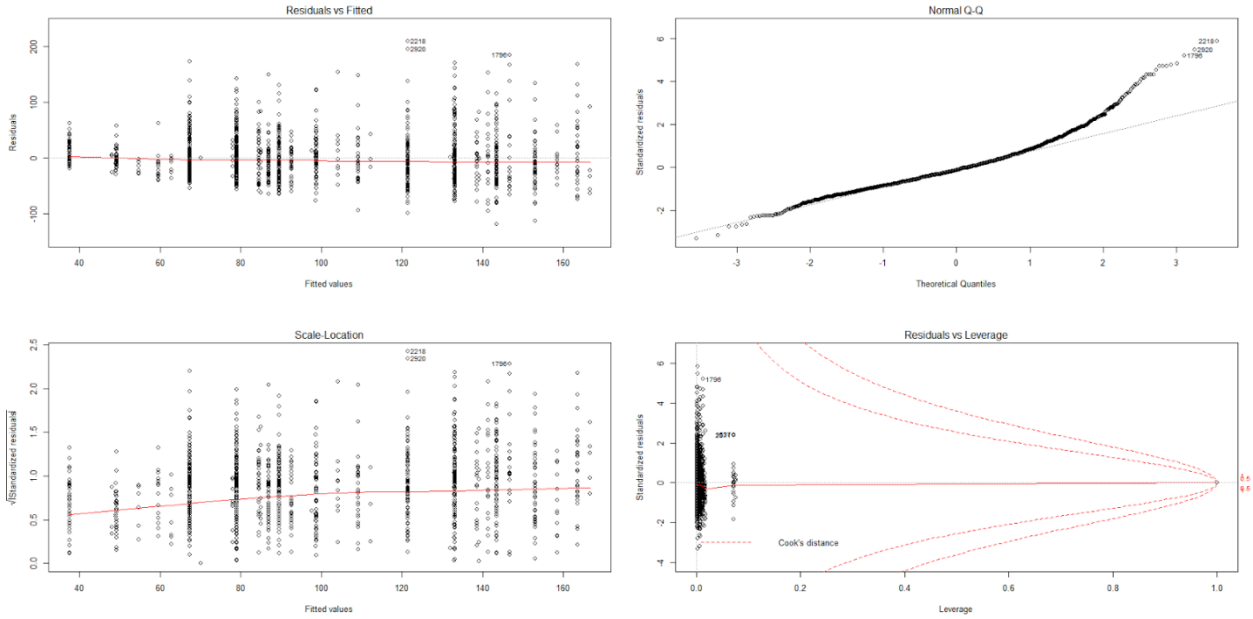
Model 1: Price Vs Total crimes for both the zones




```
Call:
lm(formula = (price) ~ (accommodates) + (minstay), data = five)

Residuals:
     Min       1Q   Median       3Q      Max
-103.883  -22.430   -5.424   15.480  205.863

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)    37.559      3.290  11.417 < 0.0000000000000002 ***
accommodates2  28.871      3.346   8.627 < 0.0000000000000002 ***
accommodates3  49.416      4.351  11.357 < 0.0000000000000002 ***
accommodates4  87.578      3.646  24.018 < 0.0000000000000002 ***
accommodates5 112.325      4.764  23.576 < 0.0000000000000002 ***
minstay2       10.090      1.982   5.090   0.0000003922899621 ***
minstay3       18.504      2.433   7.605   0.0000000000000437 ***
minstay4       11.287      4.093   2.758              0.00588 **
minstay5       18.855      4.694   4.017   0.0000611432471756 ***
minstay6       19.501     10.499   1.857              0.06340 .
minstay7       11.300      4.268   2.648              0.00816 **
minstay9        3.570     36.022   0.099              0.92106
minstay10      10.882      9.730   1.118              0.26356
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.99 on 1964 degrees of freedom
Multiple R-squared:  0.4532,    Adjusted R-squared:  0.4499
F-statistic: 135.7 on 12 and 1964 DF,  p-value: < 0.00000000000000022
```

We first added total_crimes in the linear regression, our Adjusted R squared value came to be very less of about 2.43%. So, we decided that total crime is not a good predictor of price for both the zones and moved on to next model.

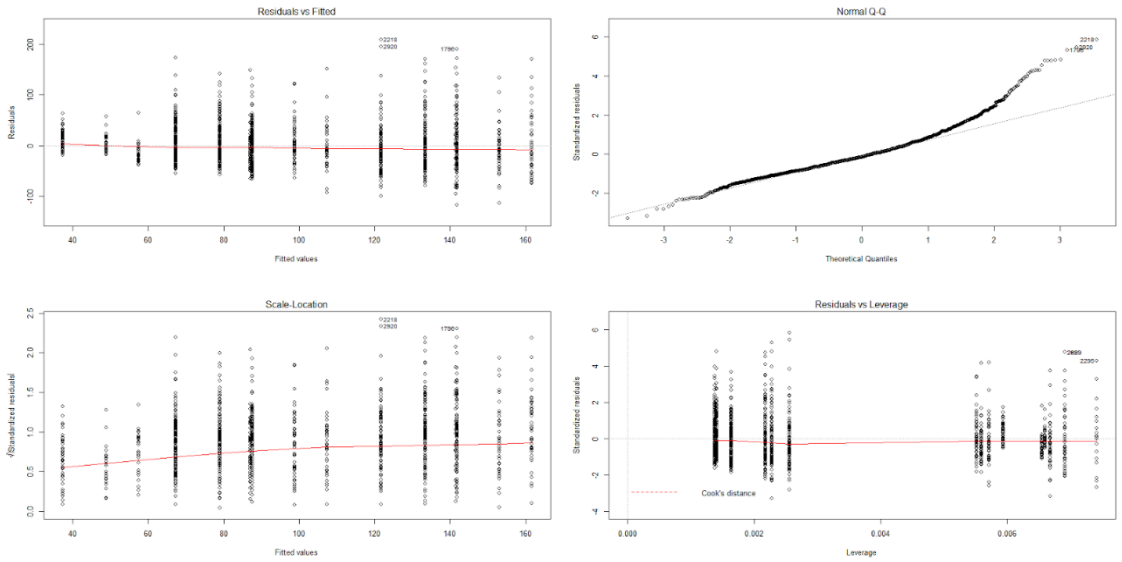
Model 2: Price with Accommodates and Minstay in Zone 1 (Within 5 Kms)




```
Call:
lm(formula = (price) ~ (accommodates) + (minstay), data = five)

Residuals:
     Min       1Q   Median       3Q      Max
-118.524  -23.073   -5.045   16.955  209.609

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)    37.463      2.761  13.570 < 0.0000000000000002 ***
accommodates2  29.900      2.836  10.543 < 0.0000000000000002 ***
accommodates3  49.533      3.657  13.543 < 0.0000000000000002 ***
accommodates4  83.928      3.048  27.533 < 0.0000000000000002 ***
accommodates5 103.956      3.918  26.535 < 0.0000000000000002 ***
minstay2       11.682      1.695   6.894     0.00000000000678 ***
minstay3       22.133      2.057  10.762 < 0.0000000000000002 ***
minstay4       17.139      3.474   4.933     0.00000086042291 ***
minstay5       25.239      4.048   6.235     0.00000000052429 ***
minstay6       17.627      9.648   1.827              0.06780 .
minstay7       11.573      3.982   2.907              0.00368 **
minstay9        2.637     35.801   0.074              0.94128
minstay10      10.569      9.644   1.096              0.27322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.78 on 2625 degrees of freedom
Multiple R-squared:  0.4391,    Adjusted R-squared:  0.4365
F-statistic: 171.2 on 12 and 2625 DF,  p-value: < 0.00000000000000022
```

As we can see from the significant codes of Accommodates and minstay are high, adding them in the model fitness our Adjusted R squared value went up to 43.65%.

Model 3: Price with Accommodates and Minstay category in Zone 1 (Within 5 Kms)



There are some nonlinear relations between accommodates and minstay which is pulling the Residual vs Fitted line a little bit below the horizontal line.

The residuals are not completely normally distributed. We can remove the tails of these outliers which have high variance.

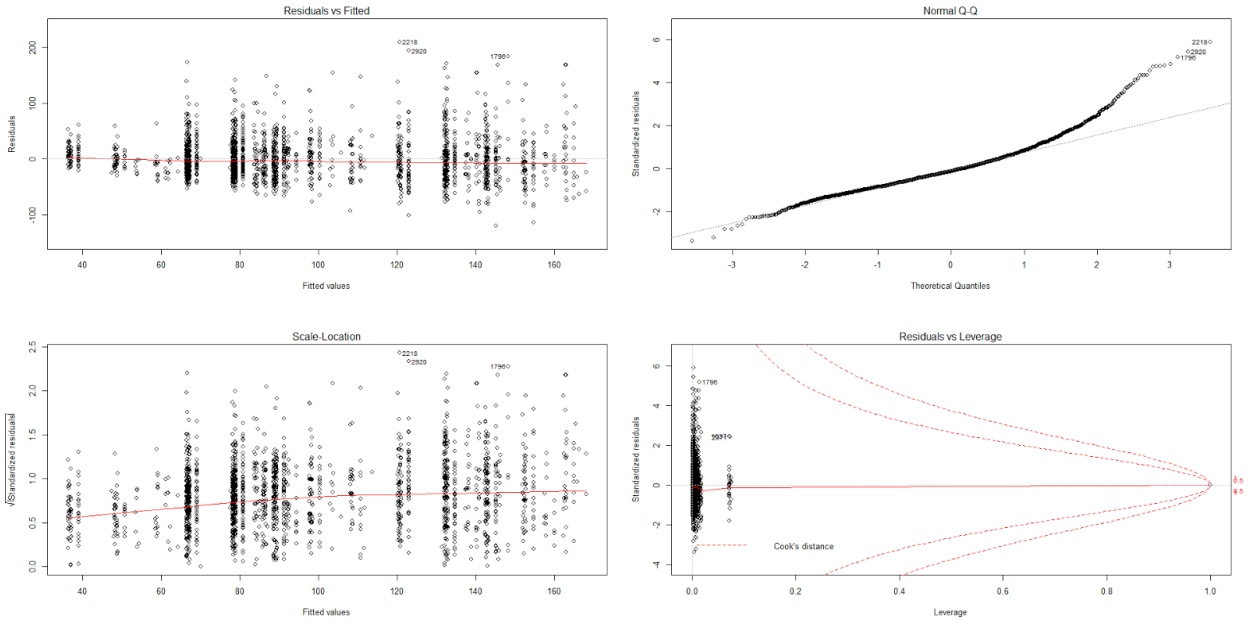
Our prediction may vary a little bit as the scale location plot is not perfectly horizontal.

As we have removed a lot of outliers, there is no point beyond Cook's distance we still think that there are some rows which are affecting the regression model, which can be further removed if we do more analysis.

```
Call:
lm(formula = (price) ~ (accommodates) + (minstayCategory), data = five)

Residuals:
     Min       1Q   Median       3Q      Max
-116.883  -23.199   -4.929   16.519  209.308

Coefficients:
                            Estimate Std. Error t value             Pr(>|t|)
(Intercept)                   37.283      2.759  13.513 < 0.0000000000000002 ***
accommodates2                 30.006      2.835  10.584 < 0.0000000000000002 ***
accommodates3                 49.817      3.654  13.632 < 0.0000000000000002 ***
accommodates4                 84.408      3.041  27.759 < 0.0000000000000002 ***
accommodates5                104.163      3.911  26.631 < 0.0000000000000002 ***
minstayCategorythree or above 20.191      1.767  11.430 < 0.0000000000000002 ***
minstayCategorytwo            11.640      1.696   6.864      0.00000000000834 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.81 on 2631 degrees of freedom
Multiple R-squared:  0.4368,    Adjusted R-squared:  0.4355
F-statistic: 340.1 on 6 and 2631 DF,  p-value: < 0.00000000000000022
```

Instead of adding Minstay we tried adding Minstay category in the model, but even though it is significant, this variable brought the Adjusted R squared value down from 43.65% to 43.55%. So, we tried to come up with more better models.

Model 4: Price with Accommodates, Minstay category & Grant.Value in Zone1 (Within 5 Kms)



```
Call:
lm(formula = (price) ~ (accommodates) + (minstay) + (grant.value),
    data = five)

Residuals:
     Min       1Q   Median       3Q      Max
-120.054  -22.670   -4.954   16.891  210.349

Coefficients:
                    Estimate    Std. Error t value             Pr(>|t|)
(Intercept)    36.315909723   2.869968444  12.654 < 0.0000000000000002 ***
accommodates2  30.081101165   2.838186007  10.599 < 0.0000000000000002 ***
accommodates3  49.501094074   3.656724279  13.537 < 0.0000000000000002 ***
accommodates4  83.877881114   3.047827065  27.521 < 0.0000000000000002 ***
accommodates5 103.807447753   3.918177991  26.494 < 0.0000000000000002 ***
minstay2       11.745177134   1.694814510   6.930     0.00000000000527 ***
minstay3       22.145029713   2.056234855  10.770 < 0.0000000000000002 ***
minstay4       17.169522437   3.473760451   4.943     0.00000081928356 ***
minstay5       25.229229248   4.046924377   6.234     0.00000000052755 ***
minstay6       17.896977585   9.647366038   1.855              0.06369 .
minstay7       11.778558432   3.983313261   2.957              0.00313 **
minstay9        2.940673293  35.793605857   0.082              0.93453
minstay10      11.013228265   9.646361595   1.142              0.25368
grant.value     0.000001694   0.000001161   1.459              0.14477
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.77 on 2624 degrees of freedom
Multiple R-squared:  0.4395,    Adjusted R-squared:  0.4367
F-statistic: 158.3 on 13 and 2624 DF,  p-value: < 0.00000000000000022
```

After including grant.value in linear model our model fitness increased by 0.02%but since grant.value does not contribute significantly to our model fitness, we decided not to include grant.value in our model and stick to Model 1, which is the best model for Zone 1(Within 5 Kms)

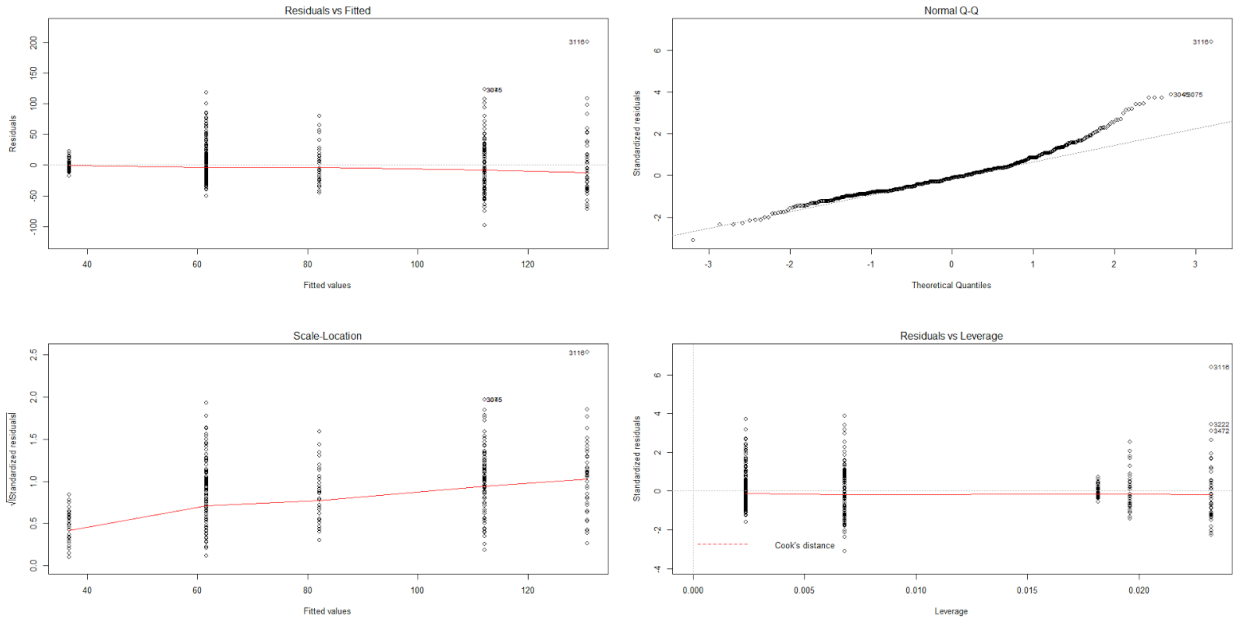
Model 5: Price with Accommodates in Zone2 (Between 6-10 Kms)

There is a bit of nonlinear relation between accommodates which is pulling the Residual vs Fitted line a little bit below the horizontal line towards the right.

The residuals are not completely normally distributed. We can remove the tails of these outliers which have high variance.

Our prediction may vary a little bit as the scale location plot is not horizontal in the

As we have removed a lot of outliers, there are no points beyond Cook's distance and, so we think that this model is good and there are no rows which are affecting this regression model.




```
Call:
lm(formula = (price) ~ (accommodates), data = more)

Residuals:
    Min      1Q  Median      3Q     Max
-98.068 -21.582  -3.673  12.418 201.209

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)    36.673      4.292   8.544 < 0.0000000000000002 ***
accommodates2  24.909      4.563   5.459    0.000000066109553 ***
accommodates3  45.465      6.188   7.347    0.000000000000555 ***
accommodates4  75.395      5.032  14.984 < 0.0000000000000002 ***
accommodates5  94.118      6.480  14.524 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.83 on 714 degrees of freedom
Multiple R-squared:  0.4063,    Adjusted R-squared:  0.403
F-statistic: 122.2 on 4 and 714 DF,  p-value: < 0.00000000000000022
```

For Zone 2, to predict the price well we initially added accommodates and minstay both, but the Adjusted R squared value was less than 40.3%. So, the above model with only including accommodates in the linear model is considered as the best model for Zone 2.

Findings and Results

- In Zone 1, the accommodation (number of bedrooms, bathrooms, etc.) specific to the property and its minimum stay requirement together determine the price of the listings fixed by the hosts.
- There is a mild correlation between the property's price and its minimum stay requirement
- In Zone 2, Airbnb hosts set the price of the listings only based on the size of the accommodation.
- Utilization factor and locality do not determine the price fixed by the hosts. Only the size of the listings is the deciding factor.
- Crime rate is comparable across the two city zones.
- There is no significant relationship between price and crime rates.
- Average price of listings in Zone 1(Less than 5kms from Melbourne CBD) is higher than that in Zone 2 (between 6 to 10kms from Melbourne CBD)

Conclusion

From our analysis, we find that accommodates and minstay helps to determine the pricing of the listing in Zone 1 with an adjusted r-squared value of 43.5%. Similarly, for Zone 2, accommodates helps to determine the pricing of listings with an adjusted r-squared value of 40.3%.

We recommend Airbnb to gather utilization data and the hosts should price the listings based on its utilization rather than just the unit size or minimum stay values in Zone 1 and Zone 2.

References

- Airbnb <https://www.airbnb.com/about/about-us>
- About Victoria <http://www.visitmelbourne.com/Information/About-Victoria>
- Tom Slee, Airbnb Data Collection: City Maps <http://tomslee.net/airbnb-data>
- Royal Automobile Club of Victoria (RACV), Victorian Burglary Statistics, https://www.racv.com.au/in-your-home/home-advice/burglary-statistics.html
- Burglary Statistics, <https://www.racv.com.au/in-your-home/home-advice/burglary-statistics.html>
- First Home Owner Grant <http://www.firsthome.gov.au>
- State Revenue Office, Victoria, First Home Owner Statistics <http://www.e-business.sro.vic.gov.au/corporate/statistics/SummaryServlet?postcodeInput=3000>
- State Revenue Office, Victoria, First Home Owner <http://www.sro.vic.gov.au/first-home-owner>
- Melbourne Hosing Market, Priya Ananthram <https://www.kaggle.com/priyaananthram/melbourne-property-analysis/data>
- How Airbnb Works | Insights into Business & Revenue Model<http://nextjuggernaut.com>
- Jacqui Alexander, 'How Airbnb is reshaping our cities'<http://theconversation.com/how-airbnb-is-reshaping-our-cities-63932>
- Herald Sun <http://www.heraldsun.com.au/news/victoria/airbnb-melbourne-400m-a-year-injected-into-victorian-economy/news-story/c97afe034dae2344f8f204876eb03284>