MA615 Final project

Shangchen Han

12/15/2019

I. Abstract:

Travel becomes more popular than before, because of releasing pressure during daily life. But traveler is likely to consider the price and quality of the accommodation as their first two concern. The company Airbnb provides information about the accommodations and it is easy for travelers to choose positions by themselves. For this project, its focus on analyzing the price of accommodations based on different variables, such as room types, number of reviews, neighborhoods, etc. Furthermore, the project focuses on the relationships between crime numbers and price of accomodations. To be specific, in order to analyze, the whole project is about EDA. For EDA, it is about finding the relationships between variables, and it also shows the changes in one specific variable based on different years or different conditions. Thus, the trend of changes is visible.
II. Introduction:

2.1 Background:

Airbnb is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. After founded, Airbnb became famous and popular because of the traveling trend. For travelers, they tend to consider price, satisfaction, review and safety as their concerns. Thus, I would like to do analyses such as EDA to find out the relationships between price and other factors and whether the place is safe of not. It will helpful for travelers to predict the price by themselves.

2.2 Data Sources:

The datasets I used for performing the analysis,'Airbnb Data Collection: Get the Data','Crime', which are particular in Vancouver, are obtained from the Tomslee website and Kaggle website. But, actually, the dataset is not integrated, which means airbnb only includes years from 2015 to 2017. And for the data in 2015 and 2017 have 4 months, so I cannot compare these 3 years directly.

2.3 Previous Work: Data combining and cleaning

The whole airbnb data has 20 files, so after imported I need to combine them together and choose available varibles, and then omit NAs both in airbnb and crime datasets.

```
dt <- bind_rows(dt1,dt2,dt3,dt4,dt5,dt6,dt7,dt8,dt9,dt10,dt11,dt12,dt13,dt14,dt15,dt16,dt17,dt1
8,dt19,dt20)
dt <- dt[,c(1:3,5:10,12:14)]
Van_dt <- na.omit(dt)
```

```
crime <- na.omit(crime)
```

I would not like to choose that price and review equal to zero, because these data are not representative. And then select the year and months, which are overlapped, so that I could compare the factorc in these date, and see whether there are some changes or not. They are April, Augest, November, December, respectively.

```r
Van_dt <- Van_dt %>%
  filter(Van_dt$price>0) %>%
  filter(Van_dt$reviews>0)
```

```r
## Try to split last_modified data in Van_dt
Van_dt <- separate(Van_dt,last_modified,into = c("date","hour"),sep = " ")
Van_dt <- separate(Van_dt,date,into = c("year","month","day"))

Van_dt1 <- Van_dt
Van_dt1 <- Van_dt1 %>% filter(year=="2015"| year=="2016")
Van_dt1 <- Van_dt1 %>% filter(month=="04"| month=="08"| month=="11"| month=="12")
```

```
##   room_type                     neighborhood      reviews
##  Length:67092       Downtown          :14979   Min.   :  1.00
##  Class :character   West End          : 8016   1st Qu.:  4.00
##  Mode  :character   Kitsilano         : 7552   Median : 10.00
##                     Mount Pleasant    : 6140   Mean   : 21.14
##                     Grandview-Woodland: 4310   3rd Qu.: 26.00
##                     Downtown Eastside : 3629   Max.   :489.00
##                     (Other)           :22466
##  overall_satisfaction  accommodates       bedrooms         price
##  Min.   :0.000        Min.   : 1.000   Min.   :0.000   Min.   :  10.0
##  1st Qu.:4.500        1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:  58.0
##  Median :5.000        Median : 2.000   Median :1.000   Median :  79.0
##  Mean   :4.407        Mean   : 2.991   Mean   :1.238   Mean   : 101.5
##  3rd Qu.:5.000        3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 116.0
##  Max.   :5.000        Max.   :16.000   Max.   :9.000   Max.   :8033.0
##
```

TYPE	YEAR	NEIGHBOURHOOD	Latitude	Longitude
Other Theft	2003	Strathcona	49.2698	-123.0838
Other Theft	2003	Strathcona	49.2698	-123.0838
Other Theft	2003	Strathcona	49.2698	-123.0838
Other Theft	2003	Strathcona	49.2698	-123.0838
Other Theft	2003	Strathcona	49.2698	-123.0838
Other Theft	2003	Strathcona	49.2698	-123.0838

```
##                                  TYPE             YEAR      NEIGHBOURHOOD
##  Theft from Vehicle                :172700   Min.   :2003   Length:476290
##  Mischief                          : 70413   1st Qu.:2005   Class :character
##  Break and Enter Residential/Other: 60862   Median :2009   Mode  :character
##  Other Theft                       : 52167   Mean   :2009
##  Theft of Vehicle                  : 38418   3rd Qu.:2013
##  Break and Enter Commercial        : 33845   Max.   :2017
##  (Other)                           : 47885
##     Latitude       Longitude
##  Min.   :49.07   Min.   :-124.5
##  1st Qu.:49.25   1st Qu.:-123.1
##  Median :49.27   Median :-123.1
##  Mean   :49.26   Mean   :-123.1
##  3rd Qu.:49.28   3rd Qu.:-123.1
```

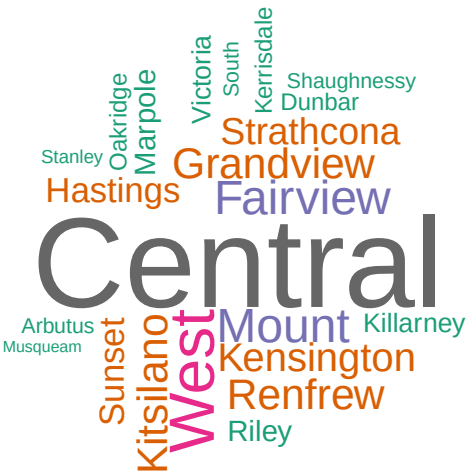
```
##  Max.   :49.76   Max.   :-122.8
##
```

They are some varibles in the data, which might be included in my datasets. And from this summary, we could see the features of these variables.

III. EDA part:

3.1 Text Analysis:

As can be seen, most accommondations are in Downtown, West End and Mount Pleasant from Airbnb text analysis.



From crime dataset's text analysis, most of crimes are in Central place, which is the same result as the text analysis for Airbnb. Furthermore, West End and Fairview are in top crime rate about survey. These two text analysis show the same result, because the more flourishing the more opportunities for crime.

3.2 Plots:

Fig.1 Distribution of neighborhood

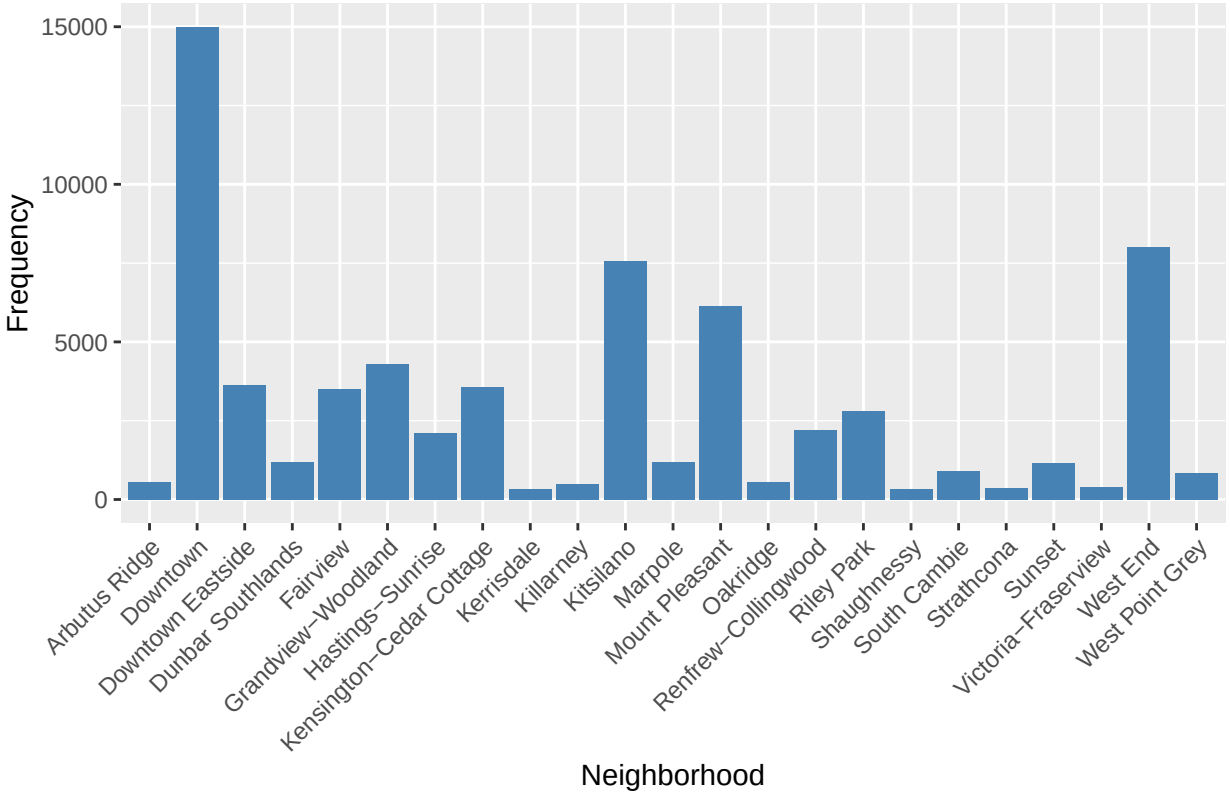


Fig.2 Crime frequency in neighborhood

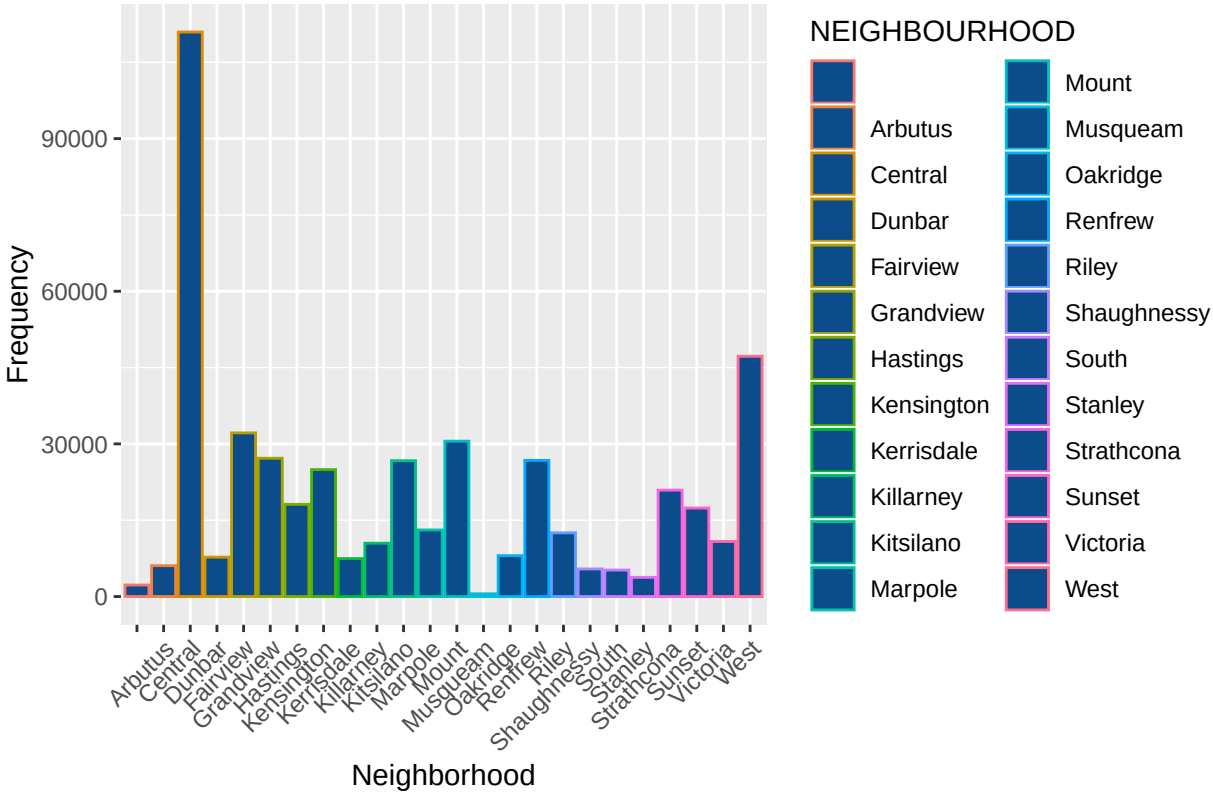
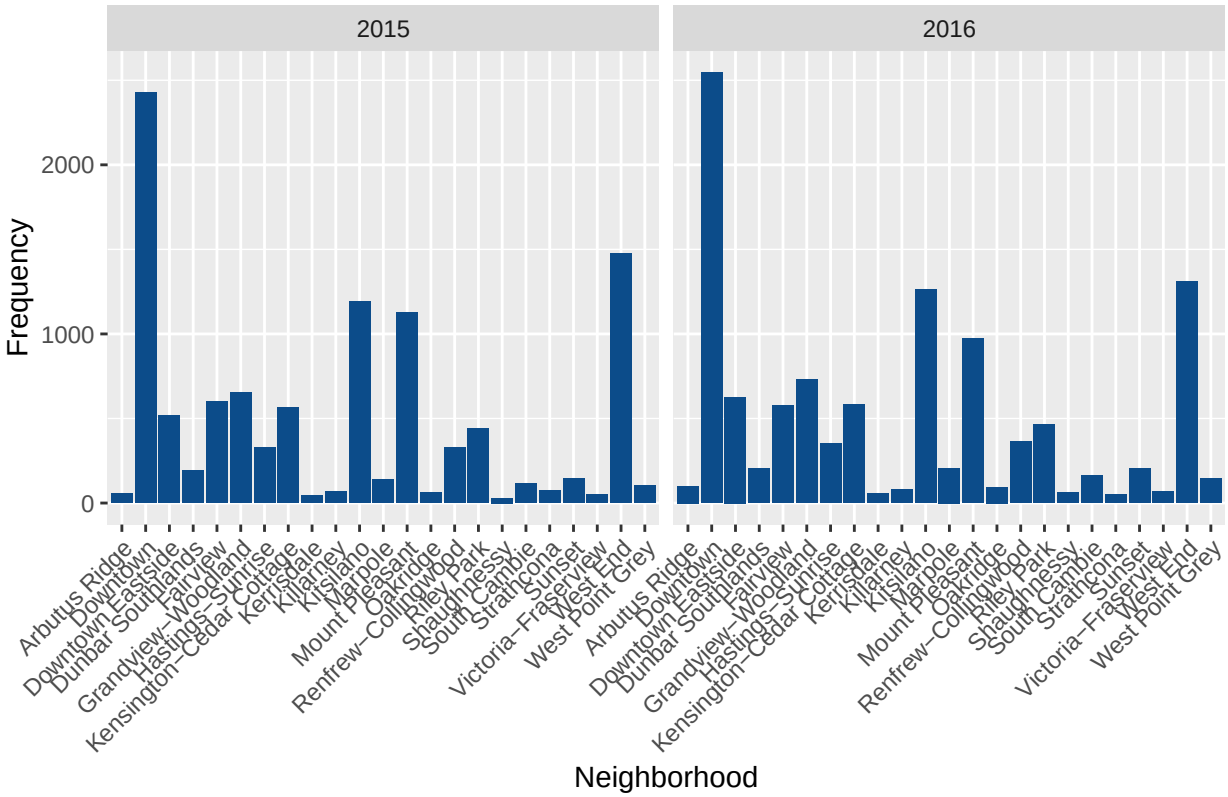


Fig.1 shows the distribution of neighborhood in Airbnb dataset, as can be seen, Downtown, West End, Kitsilano,Mount Pleasant and Grandview-Woodland are in top five places for travelers to choose. Fig.2 shows the distribution of neighborhood in Crime dataset, they are top five places: Central, West End, Fairview, Mount Pleasant and Grandview-Woodland. Compared these two datasets, it could be concluded that if a place has high check-in rate, then it will be together with high crime rate.

Fig.3 Distribution of neighborhood in 2015 & 2016 among 4 months



From Fig.3 and plot of text analysis, as we can see the most of the neighborhoods are in the Downtown region, and then West End and Kitsilano are also popular through the whole dataset. This was the same trend in 2015 and 2016, although there were slight differences between these two years among 4 months. Thus, we can guess that in the future, there will not change too much. From the result as Fig.1 and Fig.2, we can also guess, these five districts will have high crime rate.

Fig.4 Crime Type Survey

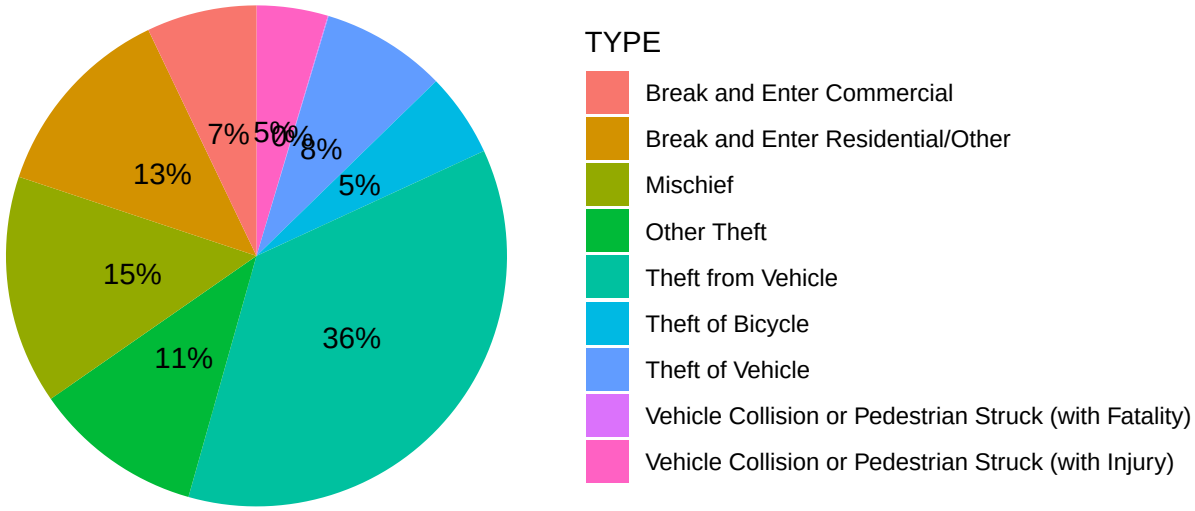
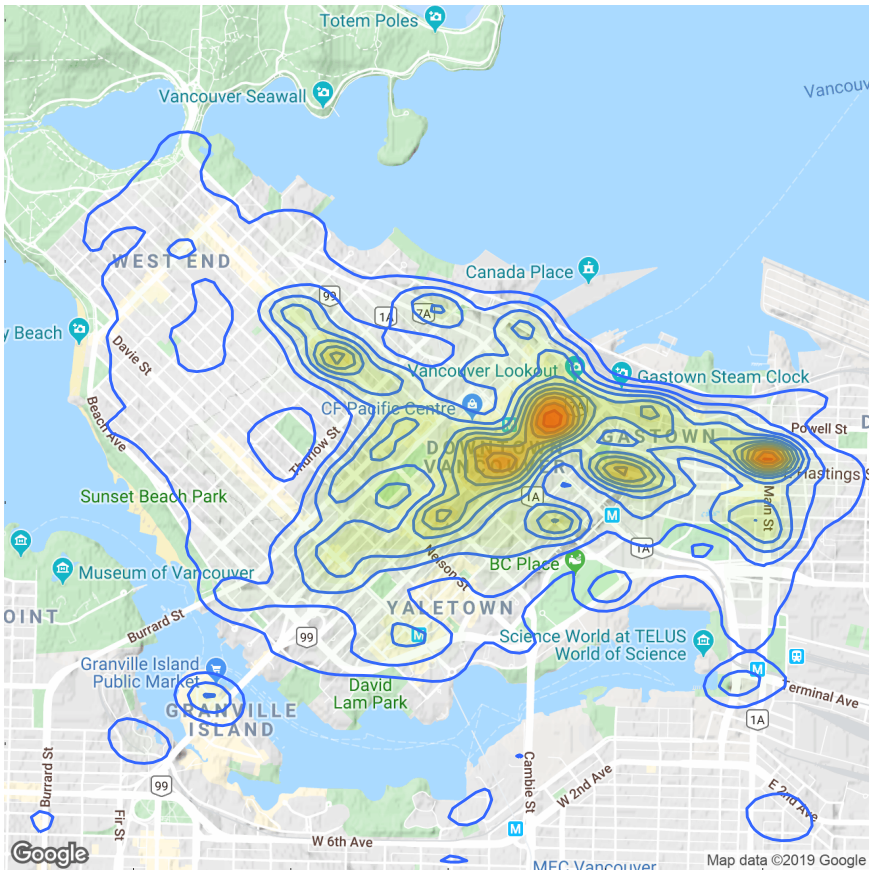
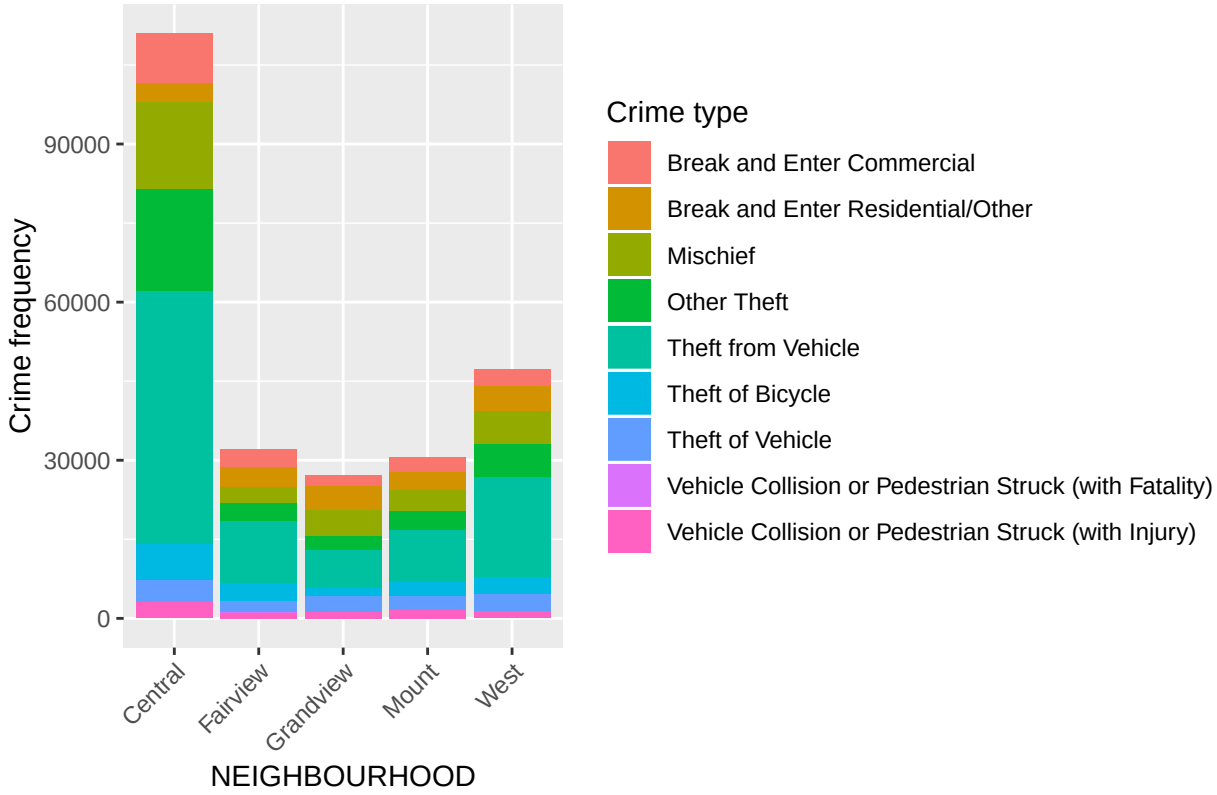


Fig.4 shows the type of crime. As can be seen, most of crime is "Theft from Vehicle", which is about 36 percent. The least crime type is "Vehicle Collision or Pedestrian Struck with Fatality", which is about 0 percent, but it does not mean it did not happen because it is a great dataset.



Because the most crime is "Theft from Vehicle", then from this map it can be seen, most of this crime happen in Downtown. Though West End is popular for travelers to choose, there are not too much "Theft from Vehicle" according to the total "Theft from Vehicle" number.

Fig.5 Crime survey in top 5 neighbourhoods



I selected top five neighborboods with high crime rate. Because "Theft from Vehicle" has the most number among crime dataset, so it happens the most. After that, "Other Theft" is at the second rank. In conclusion, although the total numbers of crime are different, in different districts have the same distribution of crime types.

Fig.6 Distribution of reviews

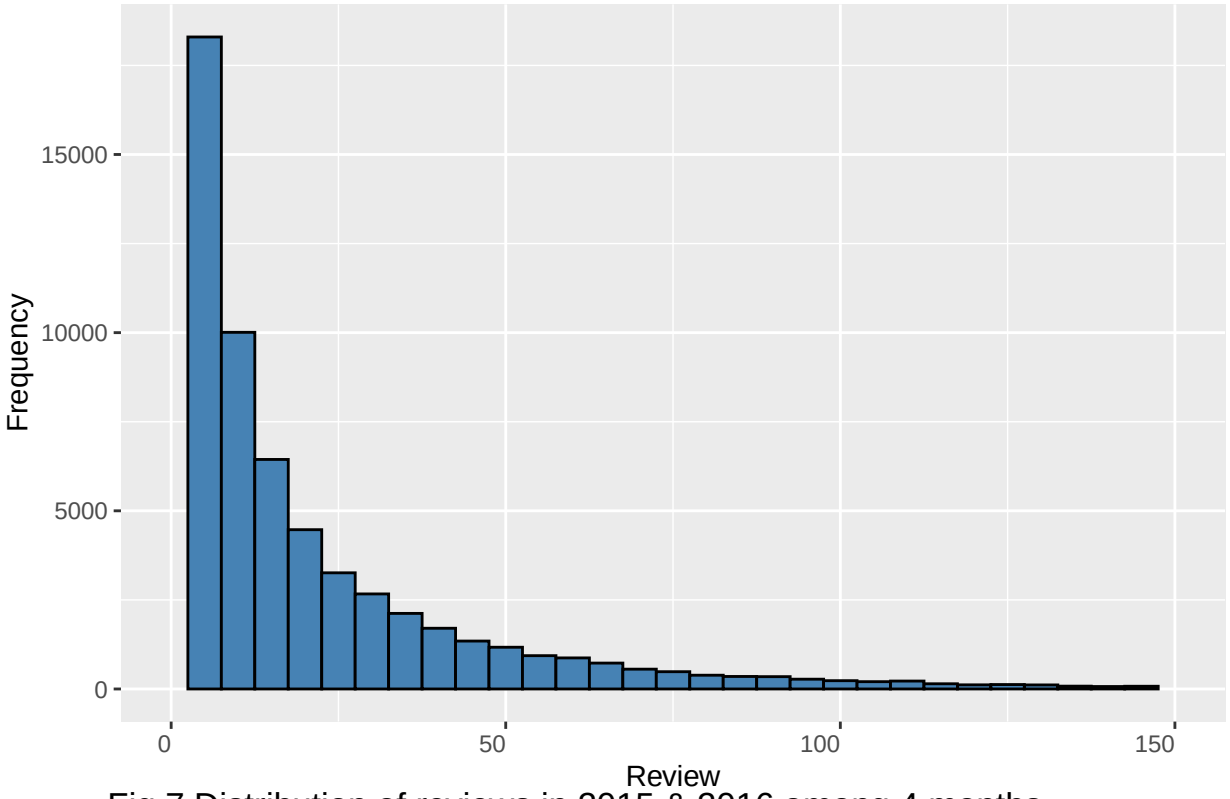
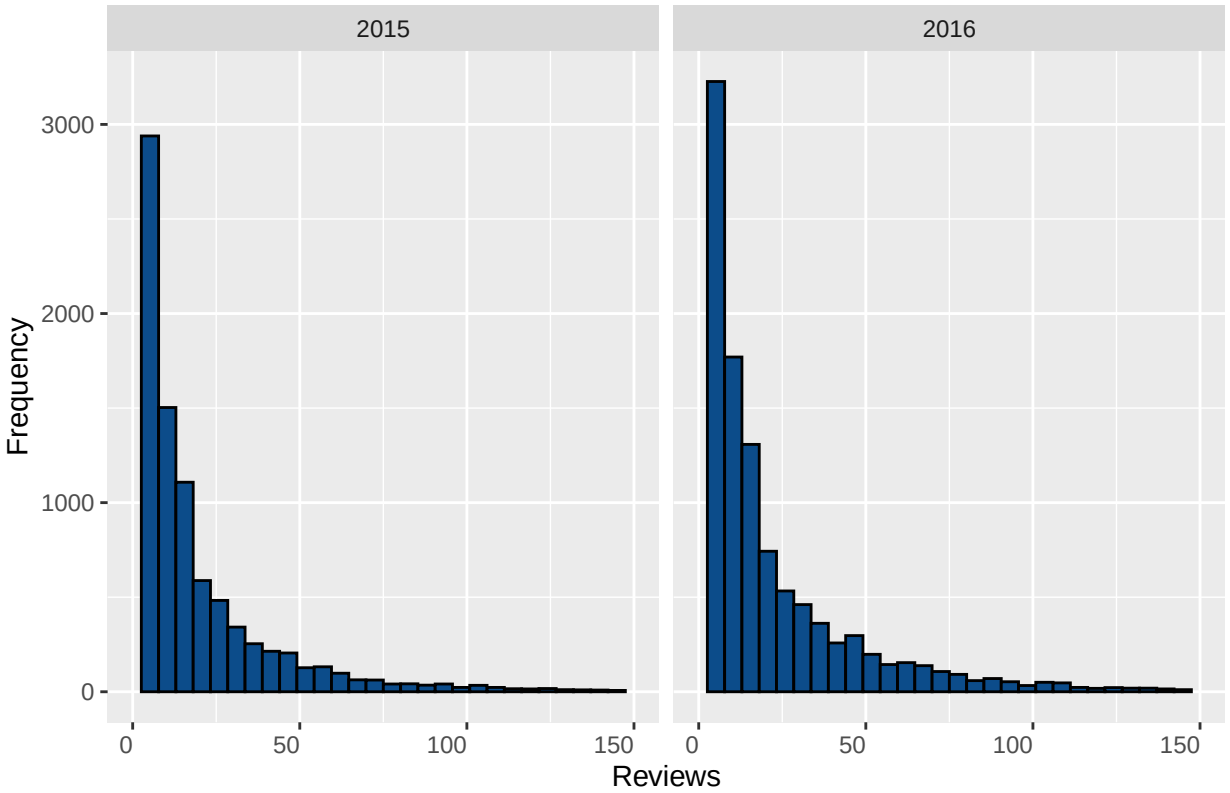


Fig.7 Distribution of reviews in 2015 & 2016 among 4 months



From the plots of the distribution of reviews, as can be seen, most of the Airbnb hosts have no reviews, and

this trend does not change, because in 2015 and 2016, they are in the same shape of the distribution.

Fig.8 Distribution of overall satisfaction

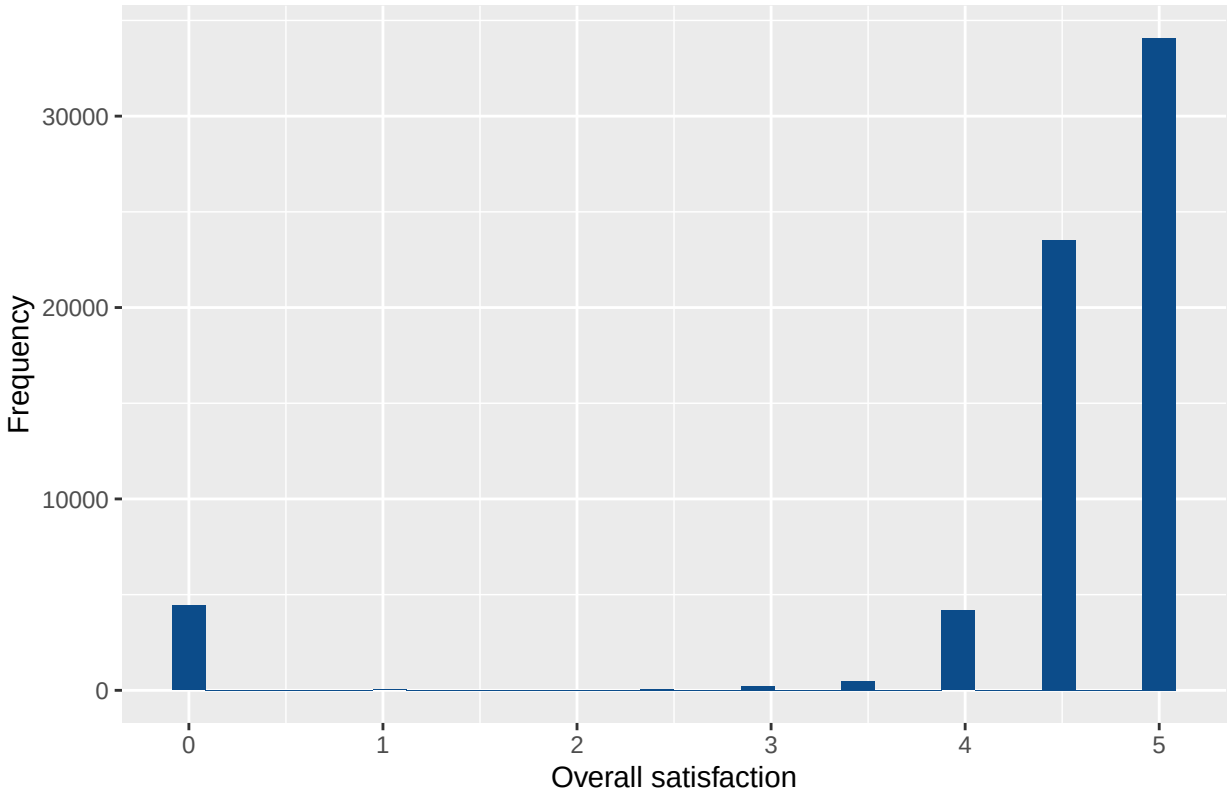
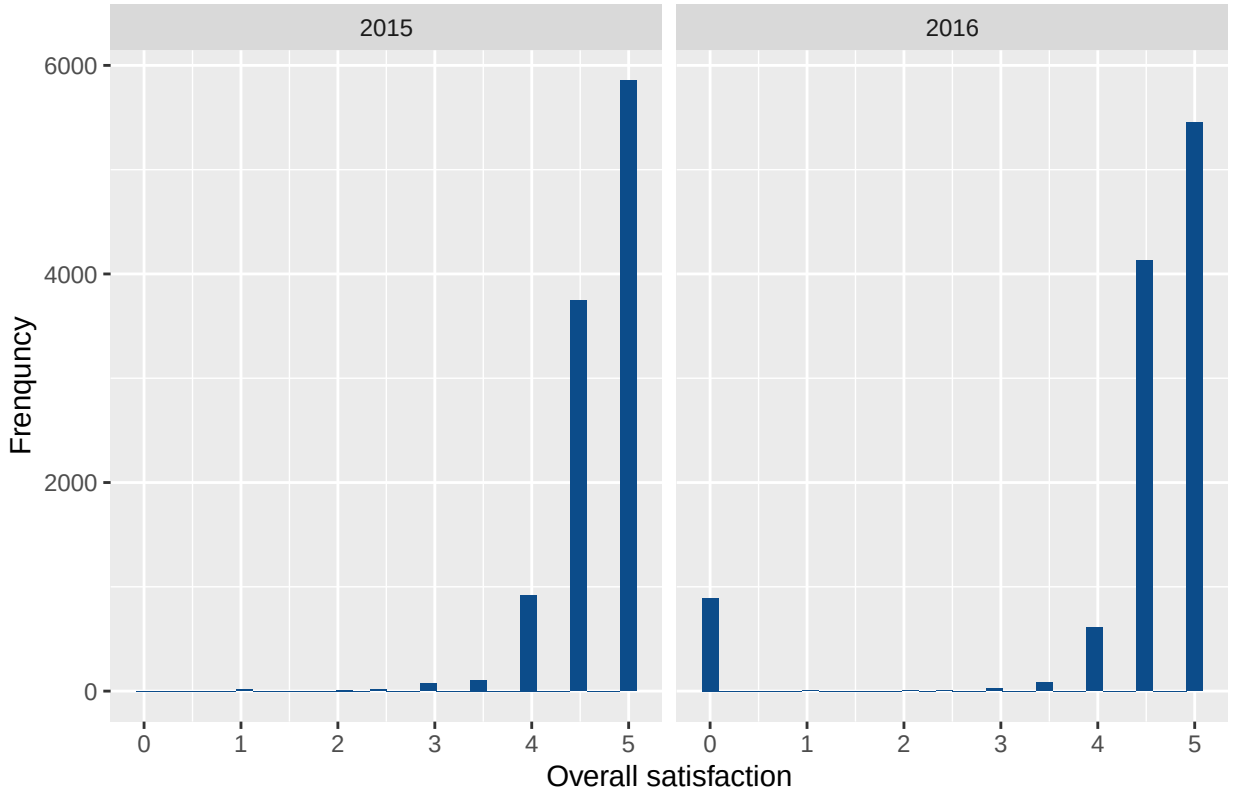


Fig.9 Distribution of overall satisfaction in 2015 & 2016 among 4 months



Most of the overall satisfaction rating is 4.5 and 5 points. To be specific, in 2015 and 2016 among 4 months, there are not too many changes. In 2016, there are more 0 points of rating Airbnb accommodations than in 2015.

Fig.10 Satisfaction in different room types

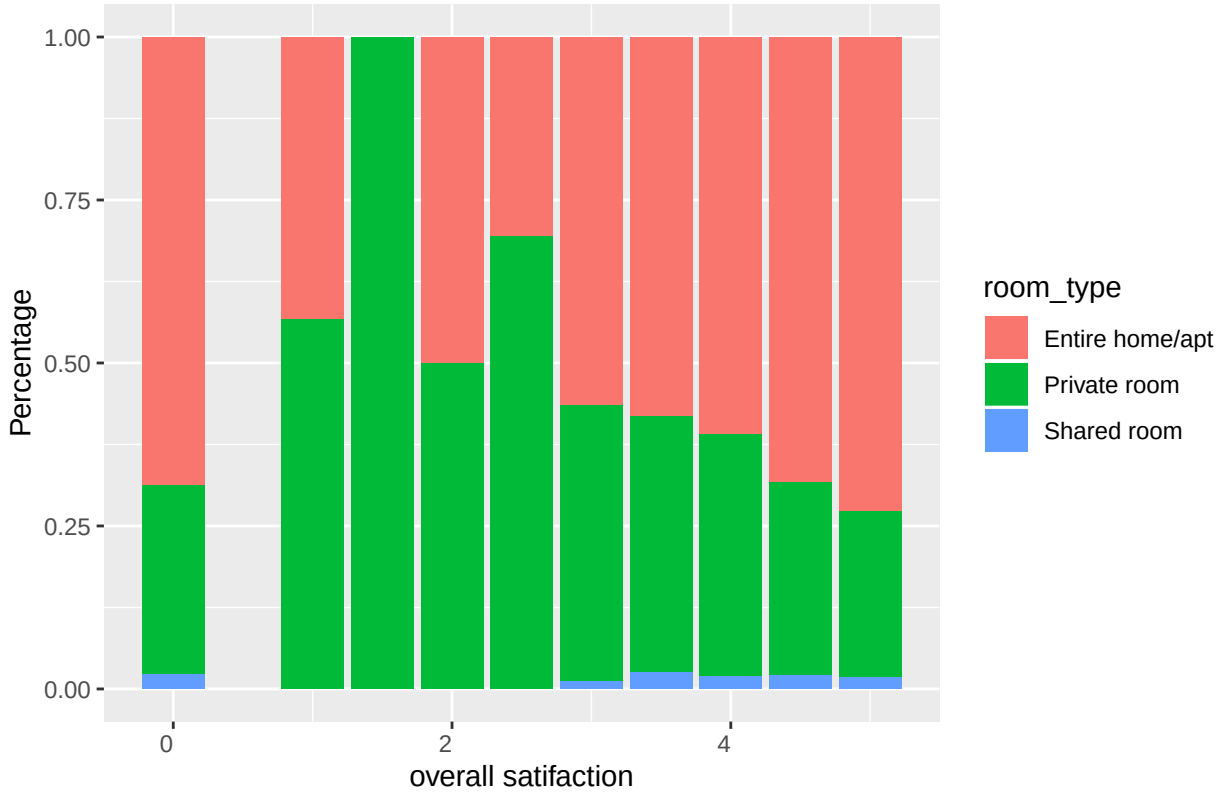
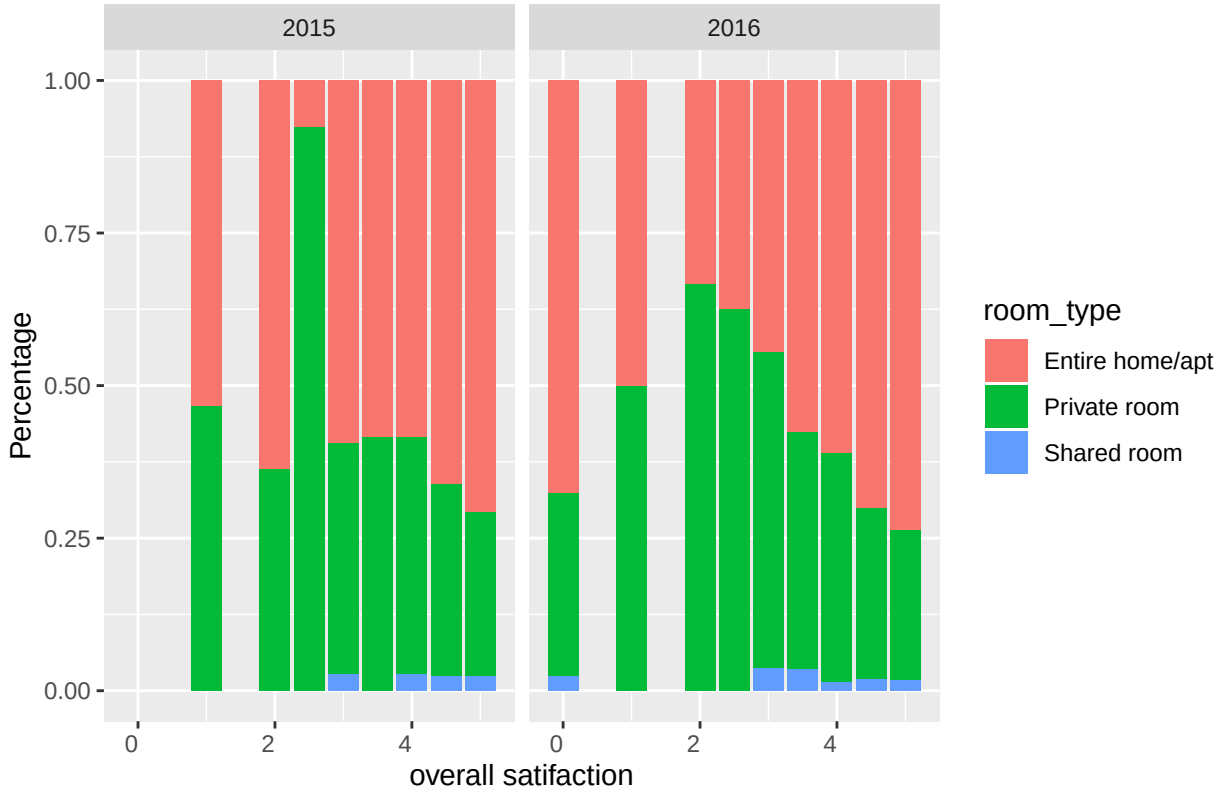


Fig.11 Satisfaction in different room types



From Fig.10 we could see the satisfaction for different room types. At first sight, the entire home/apt has the

largest percentage through most of the range of ratings. But in 1.5 point field, it is only for private room. From Fig.11, in 2015, it has the shape as Fig.10. In contrast, the entire home/apt and private room are complementary, and the rate of entire home/apt increases from 2 points to 5 points.

Fig.12 Room type survey

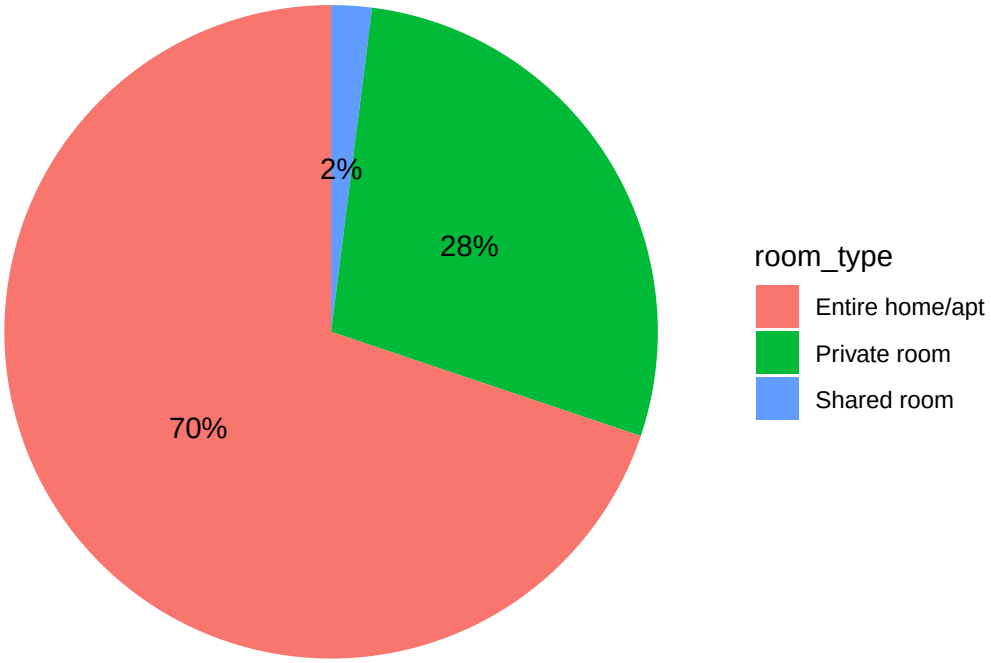


Fig.12 shows the percentage of room type, the entire home/apt occupy around 70% of whole dataset, and private room is in the second ranking around 28%. And shared room has the smallest percent(2%).

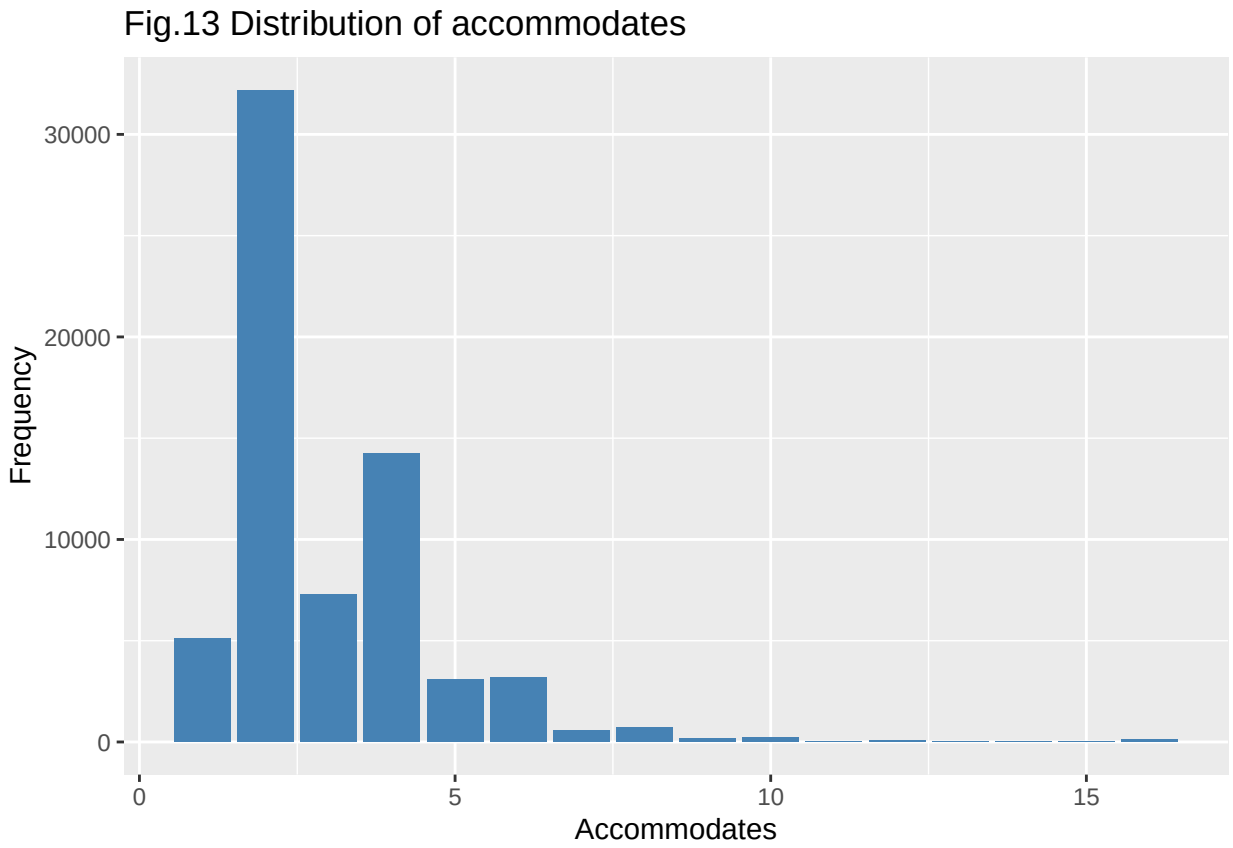
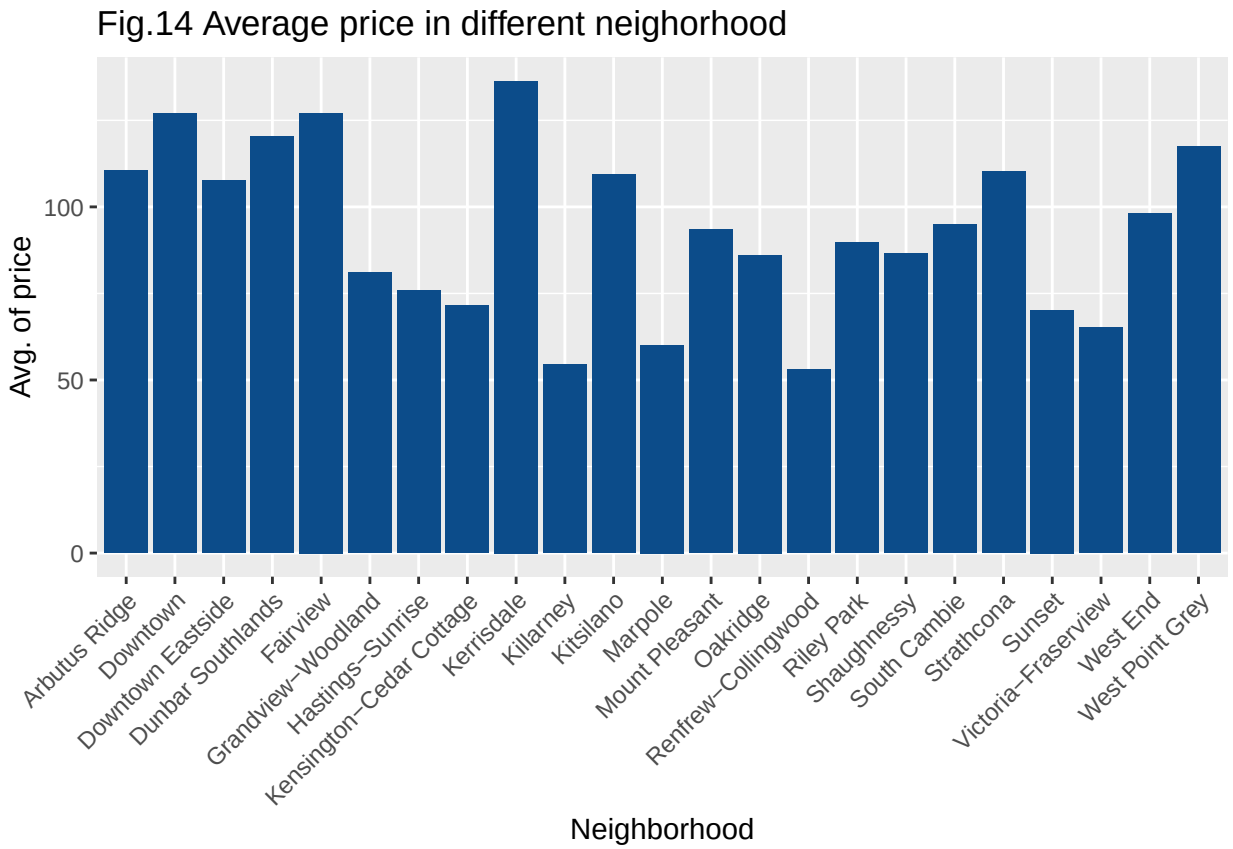
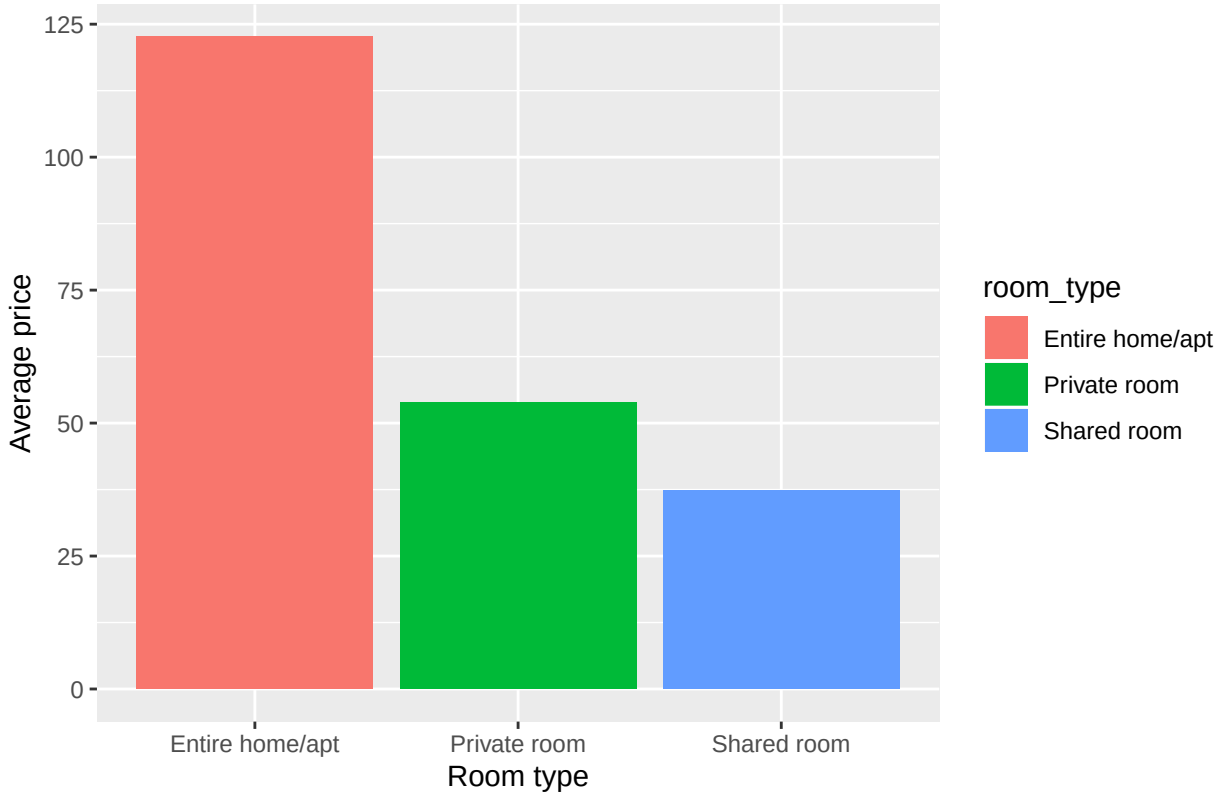


Fig.13 shows the distribution of accommodates. The most common range of accommodates is from 1 to 6, and the top three accommodates are 2,4,3.

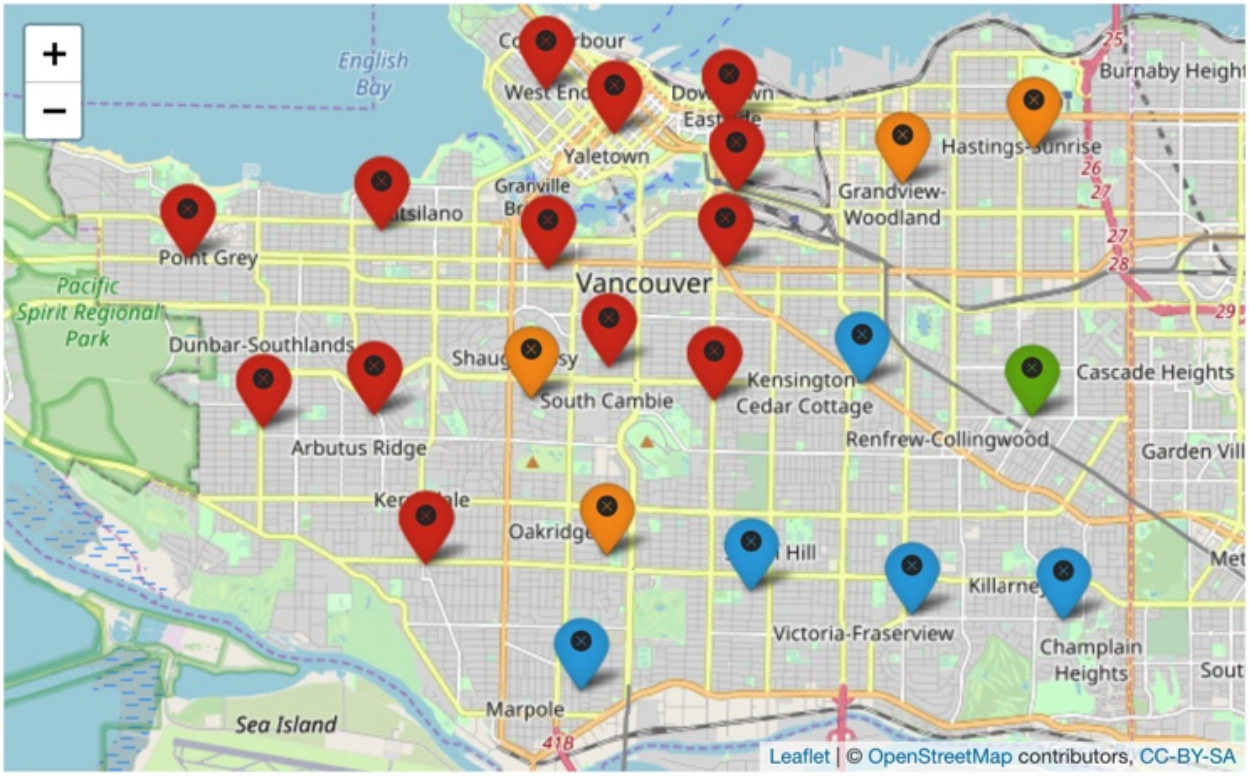


From Fig.14 we could see that the most expensive Airbnb position is in Kerrisdale around 175 per night, and Downtown, Fairview are also expensive. In contrast, Killarney and Renfrew-Collingwood have the cheapest position for living around 50 dollars.

Fig.15 Average price among different room types



From Fig.15, the average price for different room types could be seen. The entire home/apt has the highest price of around 125 per night, and the price of the private room is half of the entire home/apt. The price of the shared room is lowest at about 37.5 per night.



This map shows the different prices in different districts. I used median, mean, 3rd quartile to classify the regions. From green, blue, orange to red represents the average price is about 54, 72, 89 and above 89, respectively. Relatively costly positions are in the top left. Compared with result, the places in the top left are not only expensive, but are also in high crime rate.

Fig.16 Crime frequency between 2003 to 2017

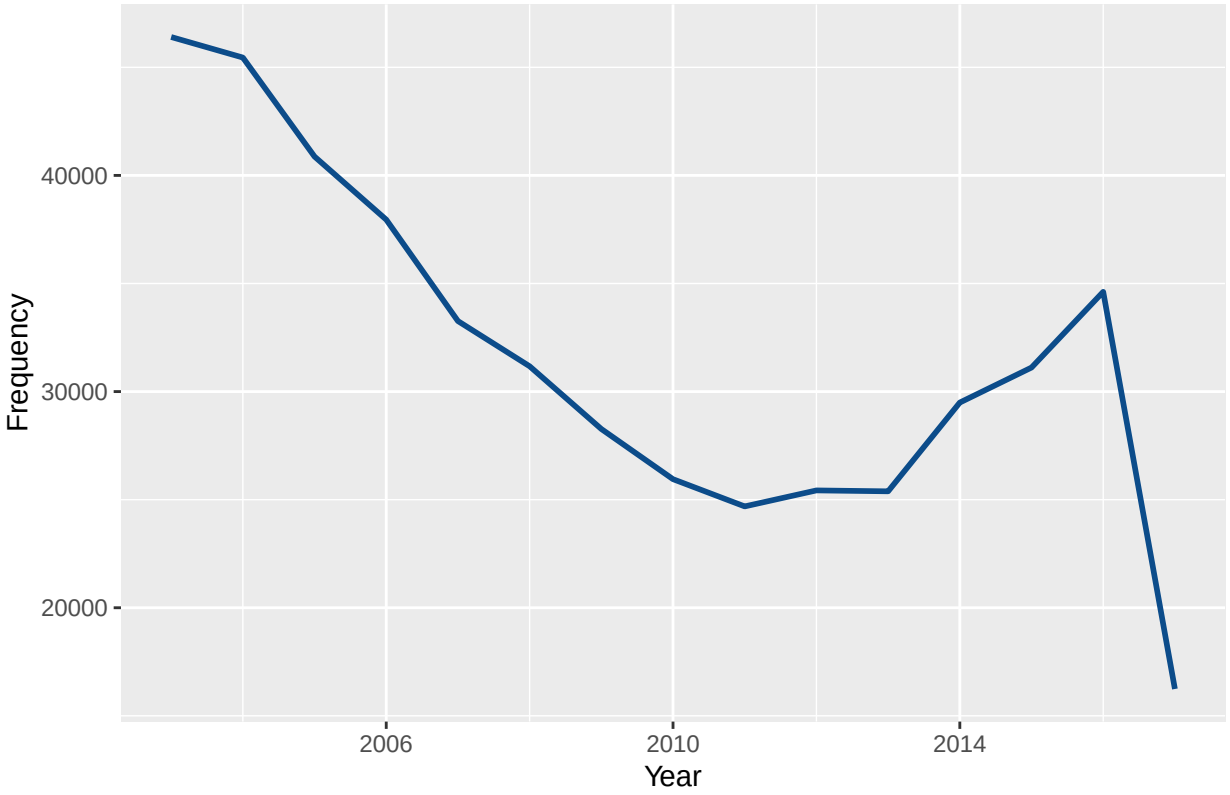


Fig.16 shows the total frequency of crimes from 2003 to 2017. 2003 had the highest crime frequency around 51000, and then the crime rate decreased significantly until 2011 around 25000. Furthermore, it increased dramatically, and it peaked at 2016 around 35000. From 2016 to 2017, there were sharp decrease to below 20000.

Fig.17 Changes of different crime type among years

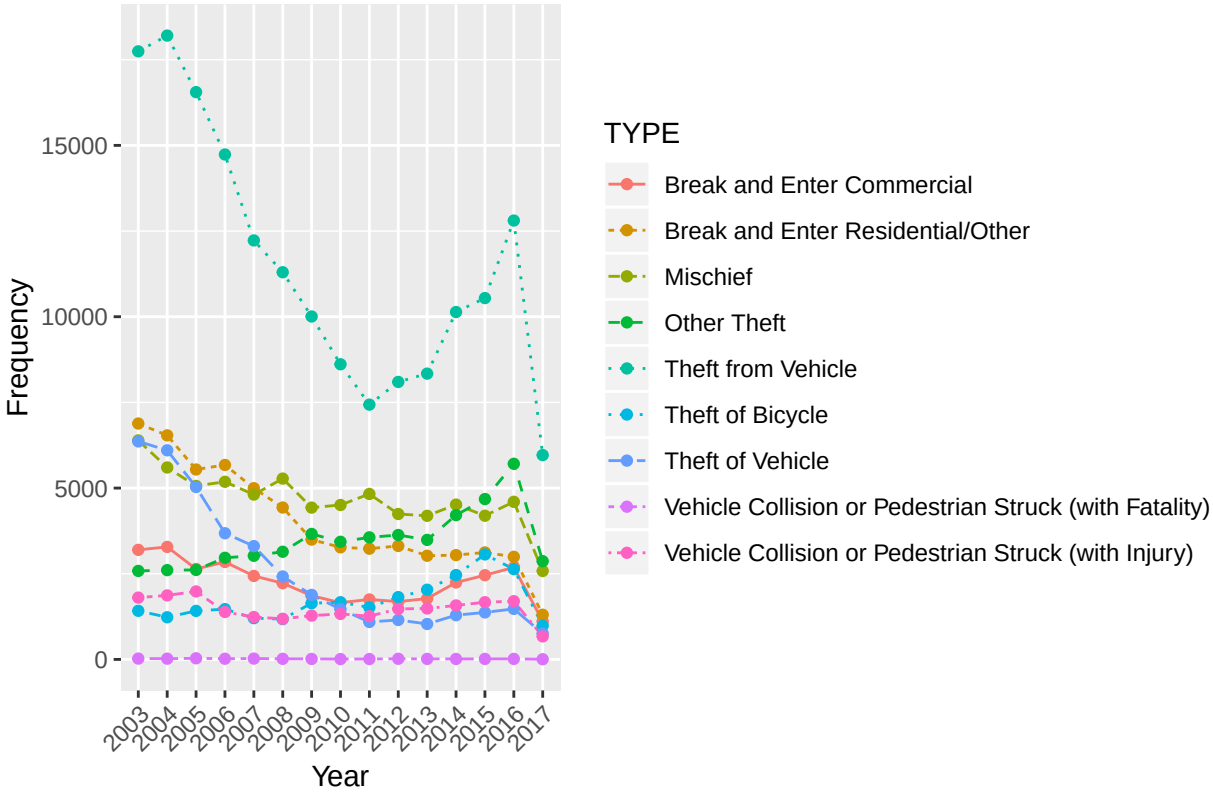


Fig.17 shows the changes of different crime types from 2003 to 2017. As can be seen, compared to 2003, in 2007 the number of "Theft from Vehicle" decreased dramatically. And the crime type of "Vehicle Collision or Pedestrian Struck with Fatality" is infrequency among this long range. We can predict that there will less crime in the future.

V. Discussion:

Implication

From this analysis of Airbnb in Vancouver, it can be concluded that the room type is the most influential factor in terms of price. And entire home/apt is the most expensive type in room type choices. And also, when travelers are choosing Airbnb, they should concern the number of bedrooms and the neighborhoods. From my survey, the number of bedrooms has a significant influence on the price of Airbnb. In terms of neighborhoods, downtown has the most number of accommodations, but the price is in the top three. In addition, downtown has the most number of crime. This is because travelers are likely to live in downtow, and this mean more chances for crime. But according to the whole crime trendy, there will be less crime in the future. And then when people plan to travel and want to choose Airbnb, they need to consider the room type, number of bedrooms and the location of accommodations.

Limitation

My airbnb dataset is only from 2015 to 2017, and the data in 2015 and 2017 are not complete. So, I cannot compare these three years directly. Thus, the result may have deviations, and I am not sure whether it is useful for 2019 or not. Besides, my report is only about Vancouver. Thus, it may not be compatible with another region. For the prediction part, the result could not be precise, since predictors are limited.

Future direction

To improve the precision, I would like to search for another bigger dataset, which includes predictors like the number of facilities in accommodations, and the conditions of transportation near the locations and so on.

VI. Reference:

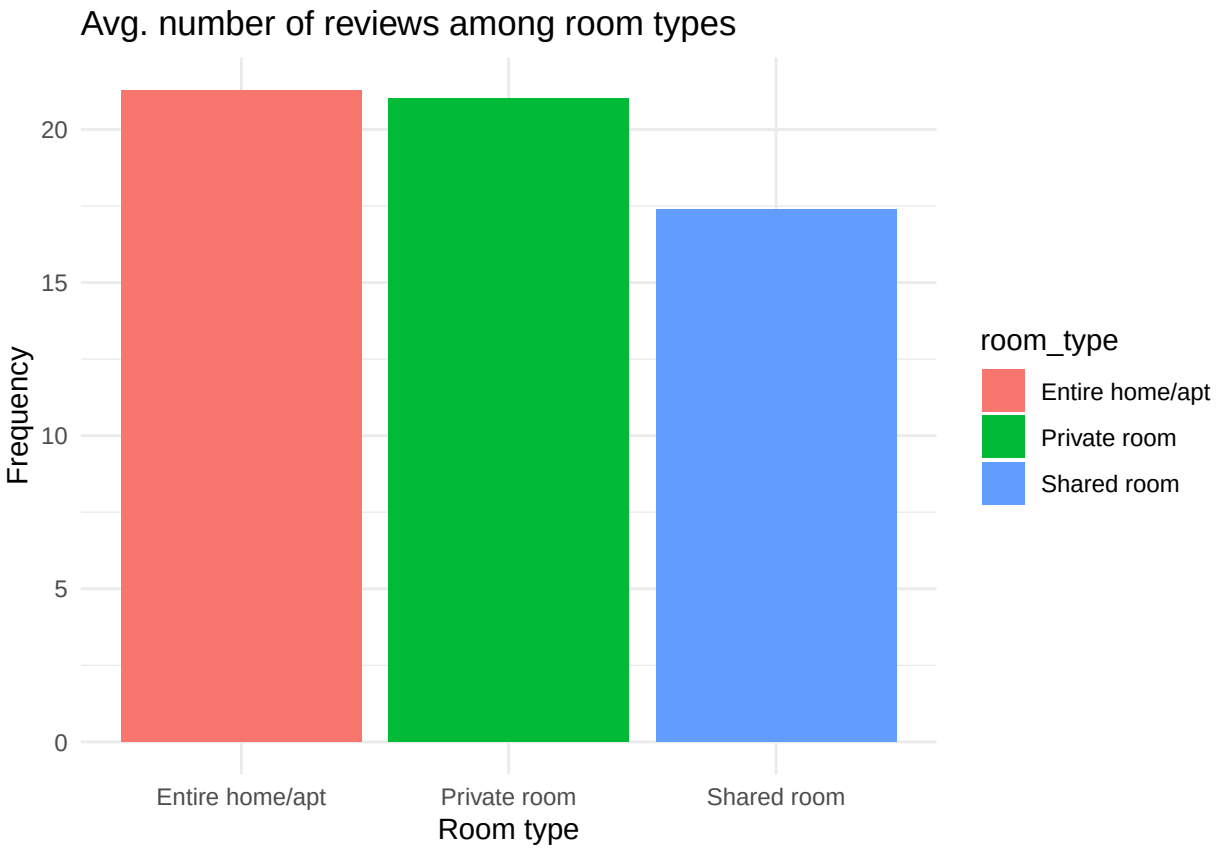
http://tomslee.net

https://en.wikipedia.org/wiki/Airbnb

http://kaggle.com/wosaku/crime-in-vancouver/data

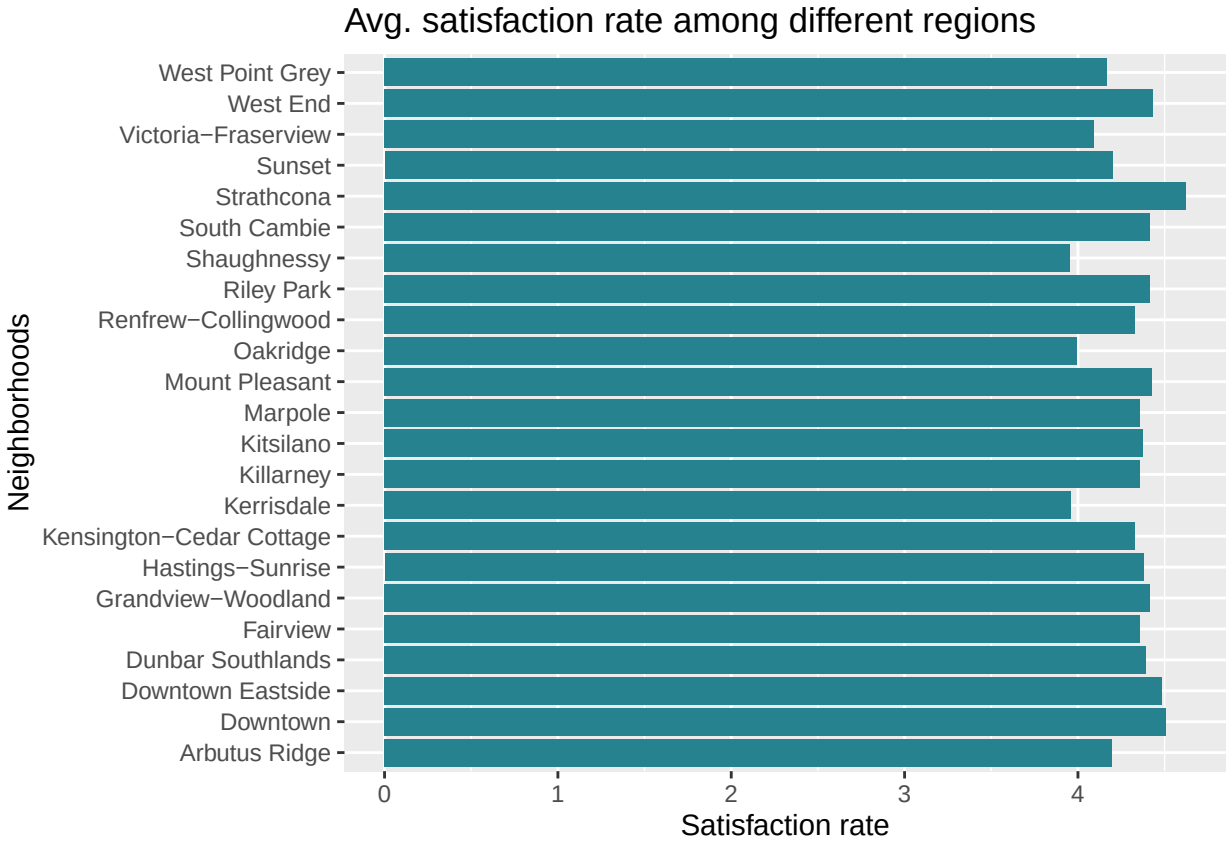
VII. Appendix:

Appendix I



The entire home/apt has the highest number of average reviews. Conversely, the type of shared room has the least information about reviews.

Appendix II



It could be found, Strathcona, Downtown, and Downtown Eastside have the top three high satisfaction rates.