**Term Project**

**Airbnb Data Analysis**

**Course: SCS 3250 Foundation of Data Science**

**Instructor: Carl Jackson**

**By: Gemma, Manpreet, Senthan, Shady, Sixing, Steven**

**Table of Contents Page**

**Executive Summary……………………………………………………………………………….. i**

**Introduction**

**Project background………………………………………………………………………...i**

**Case Background and Objectives……………………………………………………….ii**

**Preparing the Data sources**

**Data source…………………………………………………………………………………..ii**

**Description of Data………………………………………………………………………....ii**

**Imports and Settings……………………………………………………………………….ii**

**Definitions (User-Defined functions)..........................................................………....ii**

**Extract, Load and Transform (ELT) Data**

**Picking columns of interest……………………………………………………...ii**

**Key feature engineering………………………………………………………….ii**

**Data type conversions…………………………………………………………...ii**

**Cleaning up the data……………………………………………………………...ii**

**Executive summary**

DS3250 consulting is here to provide detailed, relevant market research analysis to Uber Hotel to allow them to make proper decision on their investment in the travel and tourism industry. Our team is here to give UberHotel proper insight on their main competitor and target market to allow them to make proper informed decisions on their investment.

**Introduction**

**Project background**

Over time, the travel industry is being seen as one of the most lucrative markets around the world. As travel becomes easier and cheaper, more people take trips to get away from their everyday lifestyle. Just in Canada alone, the travel and tourism industry amounts to approximately 35.37 billion dollars (CAD). In total, the whole industry on a global scale has a valor sitting at approximately 7.6 trillion dollars. The contributions are not only beneficial in adding to the country’s GDP but also helps with generating jobs and additional income for citizens. Many countries are dependent on having visitors from all over the world to enjoy their cities. A huge part of this industry that has profited off the travel sector has been the hotel business. Throughout history, tourists were always known to stay in a motel or hotel on their getaways. Starting in 2008, a new way to lodge in a city came about called Airbnb and it started gaining traction very quickly. It was a means started gaining traction for people who wanted to rent out their homes to have an additional source of income. It is a platform where people, acting the same way as a hotel would, put up their private properties up for rent for those needing a quick place to crash when traveling overseas. Hosts determine their own prices they want to set based on the amenities they are offering, the neighborhood and the number of people who are staying over. Having started in 2008, Airbnb boasts revenue of at least 2.6 billion dollars. As of now, Airbnb find themselves in over 81,000 cities and roughly 191 countries. The competition in this market is present, but it is mainly for booking good hotel deals. There are a few websites that allow people to rent private properties such as tripping.com, FlipKey, VRBO and Windu. Presently, these are not as well known and don’t share the same size of brand name as Airbnb, but will be a force to be reckoned with as time goes by. Also, a few of these competitors have a niche target market that they are focused on acquiring. There are other problems with this sector than having competition. Companies such as Airbnb are facing multiple lawsuits for not following standard travel regulations and rules. Being their main competitor, the hotel industry is cracking down on cases of illegal housing which will result in strict fines from authorities. In our analytics, we want to see how profitable the industry is and wanted to determine how profitable a company will be in setting up a similar system as Airbnb in a big city like Toronto. In the province of Ontario, there are 15,000 hosts and more than half of them, 8600 listings, are situated in the capital city of Ontario. With a city that hosted over 43.7 million people in 2017, it is worth seeing how well a company will do in the room booking industry.

**Case background and Objectives**

Being a profitable industry, many people are trying to grab market share in the lodging industry. One of these companies is Egypt based [company Orascom Holdings](http://orascom.com/), one of the largest investment firms in the Middle East. After successfully launching Freedom Mobile as Canada's 4th largest mobile operator, Orasom's interest in Canada has increased. Orascom holdings decided to have their [Hotel Management Division](http://orascomhm.com/) join forces with Dubai Based Ikama Real Estate in order to establish UberHotels with an ambitious strategy to take market share from Airbnb.

Orascom Holding’s new adventure, UberHotels has a novel strategy of buying or leasing units in strategic areas in Toronto and marketing them at competitive rates to Airbnb. UberHotels aims at offering high value short term rentals for competitive rates. In addition, to help gain an edge over their competition, Orascom Holdings wants to roll out a mobile 'room service' team.

UberHotel has already secured funding from Orascom and is in the early stages of conducting feasibility studies to assess the viability of the project. That is where our team at DS3250 come in to assist in Orascom’s early planning phases. Based on the outcome of this study and findings, UberHotels intends to invest close to $153 million in phase 1. We are here to analyze the landscape of tourism and lodge renters in Toronto. That means taking a thorough look at the different neighborhoods in the surrounding areas, the different prices people are renting at, the size of the property, etc.

Before making any final decisions, Orascom Holdings is expecting the following deliverables to be given to them by DS3250 Consulting:

1) Insights about the consumer behaviour and industry trends. This allows for observations to be conducted in the point of view of the consumer. This includes which neighborhoods are often most chosen, the effects of local crime rate and the price ranges that are most commonly affordable for consumers

2) Data driven Pricing Guidelines and Manual that would be used as an objective tool for sales decisions. This is to help determine how a lodge would be priced in accordance to different variables by the property owner. This includes local attractions, accessibility of transport and the size of the lodge amongst other reasons.

3) Heat map for current Airbnb listings. This will provide a more visual explanation of which areas of the city have Airbnb spaces for rent. In turn, Orascom Holdings would see where to try and target their markets if a decision is made to proceed with the project.

**Preparing the Data Sources**

**Data Source**

All our data was procured from an online source titled[**http://insideairbnb.com/**](http://insideairbnb.com/). This link allows access to a plethora of data from the Airbnb site. From the available data sheets, we were able to extract, refine and filter important data to perform an exploratory analysis of the Airbnb dataset. This allows for us to understand the rental landscape in Toronto through all kinds of interactive and statistical visualisations. In addition to this, we also imported a date set file from the Toronto Police service- Public Safety Data, <http://data.torontopolice.on.ca/pages/major-crime-indicators> portal to further analyze the crime rate in each neighborhood.

**Description of Data**

From the website [**http://insideairbnb.com/**](http://insideairbnb.com/), we pulled the data from three CSV files. They are titled:

* *Listings.csv* - In this file, they had a summary information and metrics for listings in Toronto. This included details such as all current listings, the host names and their respective ID, and the price of the lodge.
* *Reviews.csv* - This file is made up of reviews customers had posted after their stay in the lodge. It includes reviews as well as the corresponding listings.
* *Calendar.csv* - This file has all the dates that the listings are available or booked in Toronto.
* *MCI\_2014\_to\_2018.csv* - This document has information on Major crimes in Toronto neighborhoods from the years 2014-2018.

We used these four data sets as they had all the relevant information we needed to follow through with our analysis.

**1.3. Imports & Settings**

For importing the date that we needed to analyze, we used both Python and Jupyter notebook. To better serve our purposes of analysing the data, we needed to call on additional packages that included pandas, numpy, math, statsmodels.api, matplotlib.pyplot, pandas.plotting, scatter\_matrix, warinings, seaborn, etc.

**Definitions (user-defined functions)**

In this part, we use def statements to create our own functions for further analysis. Firstly, we declare the function with the keyword def followed by the function name. Next, we write the arguments of the function that will be followed by the program statement to be executed.

Here are some examples of our def statement:

def printIn # puts a line under the text

def shape(df) # returns number of rows and columns in df

def unique # show the unique values in every column

def hist(df) # assume 1 column only

def analyze(df, col, QuestionAnswer): # prints a summary table of unique items, their count, as well as chart it

df super\_hosts(df, col,function, msg=’ ’): # compares super vs regular host on a given criteria

**Extract, Load, and Transform (ELT) Data**

**Picking columns of interest**

This stage is for choosing which columns are essential to answering our questions. Not all columns will pertain or be considered relevant as it is a giant data set. Many are missing proper values, others have values which repeat and there are columns who just have no value to add to the final computations. To facilitate inspection, certain rows were also transposed and added with the other relevant columns.

**Key feature engineering**

We inspected unique values in every column by using the *unique* function to check that they are valid and see if columns could then be classified into bins. The result was 26 out of 42 columns had less than 100 unique values. Furthermore, inconsistencies needed to be fixed in the city column as some neighborhoods went by different names even though it was the same location. Therefore, we tried to cut down on issues related to redundancy.

**Data type conversions**

A few columns needed to have type conversions in order to allow us to properly manipulate the date. The host\_since column contained data in integer format for which we then converted to year instead of date. The price column contains data in string format with a comma separator and currency symbol ‘$’ which we needed to convert to numeric values.

**Cleaning up the data**

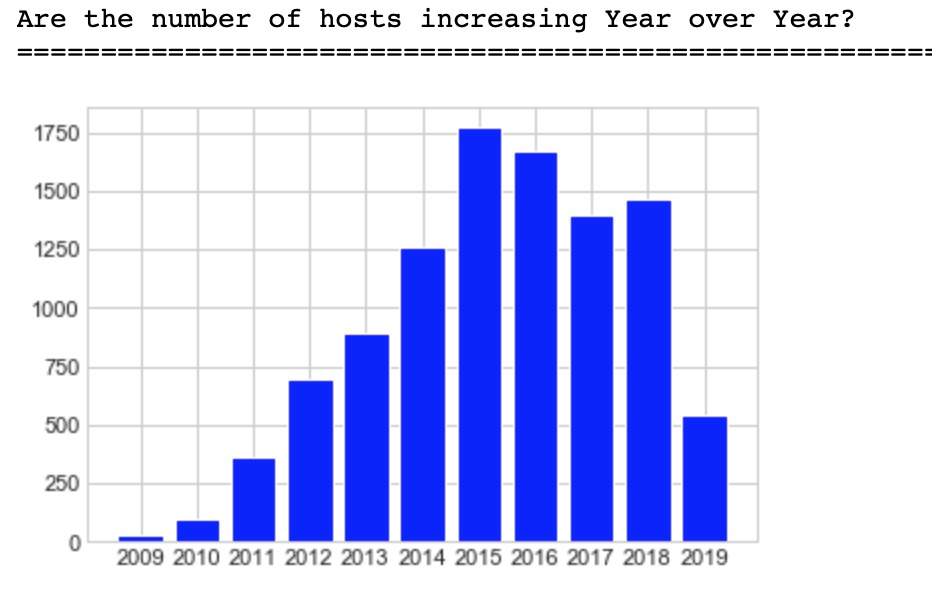
From a first glance of the Data, there were a lot of columns containing NaN values. This translated to having the data manipulated so that it was usable for proper calculation and tabulation of the date in question. Since there are lots of columns with NaN values, we have to perform a few imputations and transformations on our dataset to create the desired analysis. Some of these columns needed to be deleted such as city, host\_location, and host\_neighborhood. Also, data cleaning and preprocessing has been done to prepare the data for the next step of data analysis. We set df0 as the master final clean dataset. A quick glance at the dataset df0 shows that there are 10171 rows and 36 columns, which presents 10171 unique listings in Toronto in total and 36 attributes describing the detailed information of each listing.

**Data Analysis**

Insights about industry trends and customer behavior

**Growth of Airbnb in Toronto**

As we can see from the graph below, the number of hosts over the years shows an increase in the number of hosts over the years. This is a positive indication of the lodging market, showing a proper demand for it. However, the growth rate becomes slower starting from 2016 (See appendix 5.1,5.2). These reasons are what we need to further analyze as a team and figure out why there is a slow stagnation in growth over the last few years. As of now, after having cleaned up all the data, there are currently 10171 listings.



**Price statistics**

On average, it costs approximately 146$ a night to stay in a listing based in Toronto. Most listings are situated between 0-200$, and the most listings are approximately 100$. The highest valued ising is 10,000$ while the lowest is listed at 0. Further below, we will see how different factors come into effect with the pricing of each lodging.

**Crime statistics in Toronto and the effects on the number of Listings**

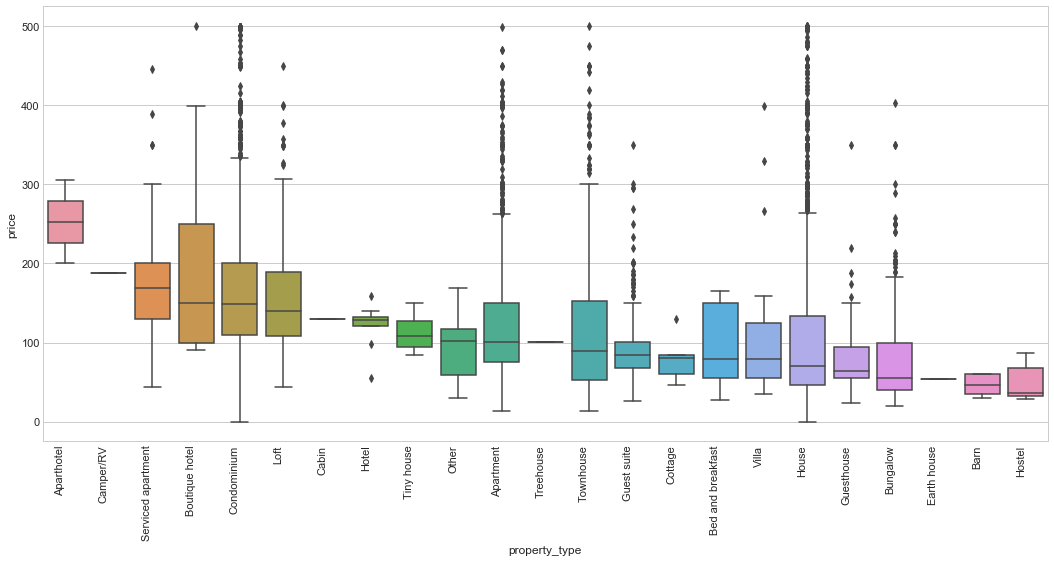
With this statistic, it was to be evaluated if crime has any sort of relevance with the number of listings in that neighborhood. If we look at the bottom five neighborhoods in table (See Appendix 10101)If we look at the last five cities on this chart we can see:

|  |  |  |
| --- | --- | --- |
| Neighborhood | Number of listings | Crime Rate |
| Waterfront Communities- The Island | 2222 | 5674 |
| Niagara | 436 | ~1300 |
| Annex | 412 | ~2800 |
| Church-Yonge Corridor | 337 | 6301 |
| Kensington-Chinatown | 332 | ~3300 |

With the exception of the Waterfront Communities-The island, a general trend can be observed that the fewer crimes in the neighborhood then the greater number of properties are placed as a listing. Upon a quick glance, the graph is indicative of this pattern.

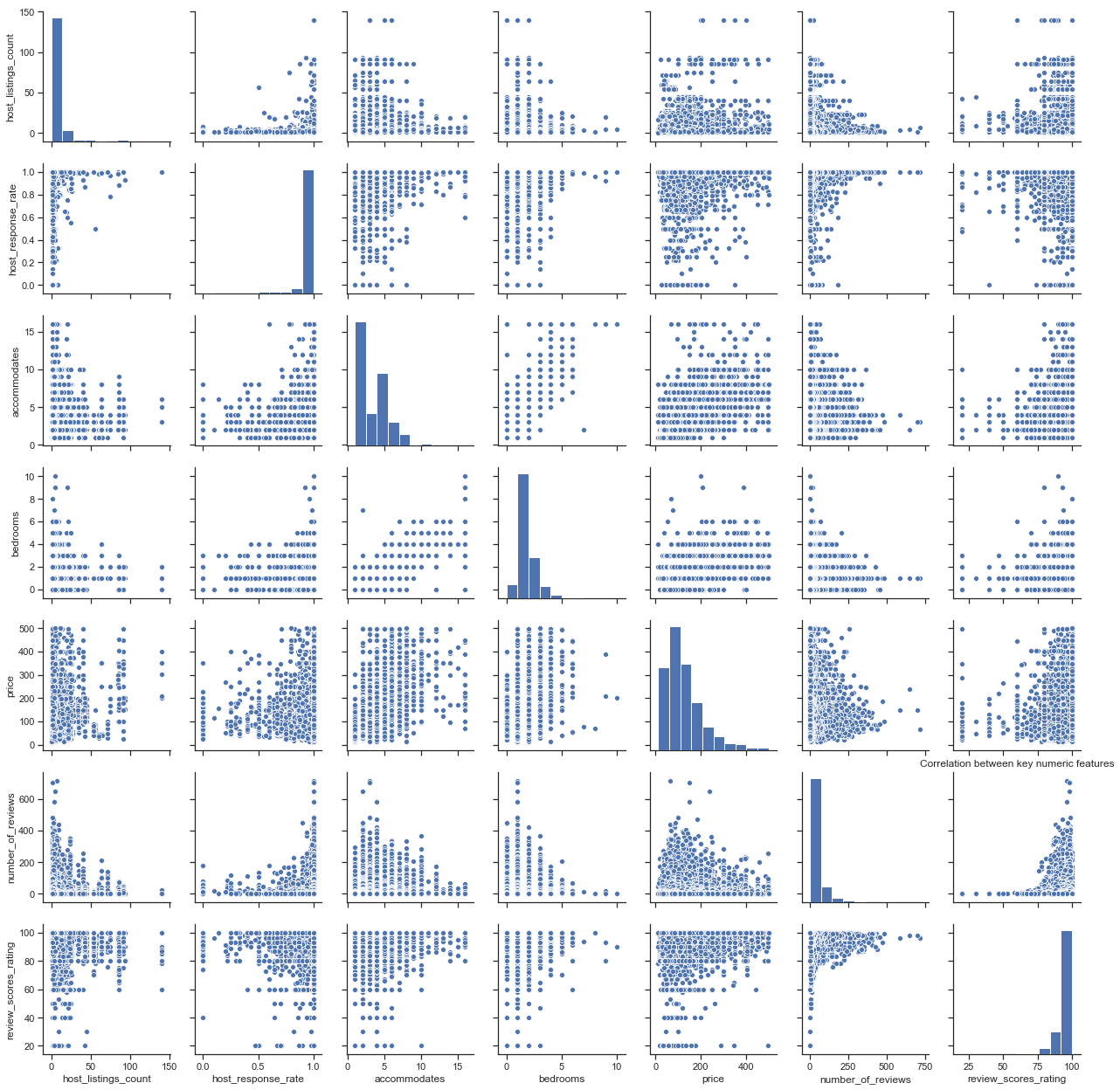
**Type of property vs Price**

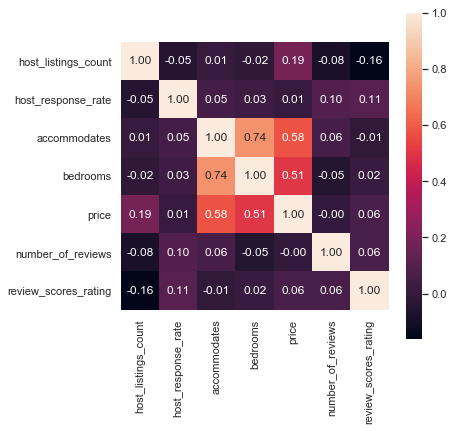
The prices vary depending on what kind of lodging property was rented. According to our box plot graph below, the more expensive properties to rent are Aparthotel, boutique hotels, loft, and condominiums. All other properties have most of their listings around the calculated average price of 146$

****

**3.1.2. Correlation of numeric features.**

In order to get a better understanding of how the attributes are correlated in Listings, we select several numeric features to draw pairplot and heatmap. The correlation matrix demonstrates that that many attributes are well-correlated, such as: accommodates, beds and bedrooms, we will apply this result to the price analysis.





**3.1.4. Analysis on SuperHost vs RegularHost**

In 2016, Airbnb launched Superhost program to evaluate the performance of hosts. The An Airbnb Superhost must meet the following four criteria to be qualified: host a minimum of 10 reservations in a year; respond to 90% of all guests within 24 hours; have at least an overall rating of 80% or higher; no cancelations. In this section, we try to figure out the exact distinctions of superhosts from regular hosts. (Table 6,6.1)

More than half of the hosts are superhosts, the ratio between Superhosts and regular host is 61.7% vs 38.3%.

How does the response time look like?

Hosts are very responsive, more than 91% hosts response within an hour.

Are super hosts more responsive than regular hosts? (Table 7)

In 83% of the categories, super hosts min host\_response\_rate is more than regular hosts.

Are super hosts more expensive than regular hosts? (Table 8)

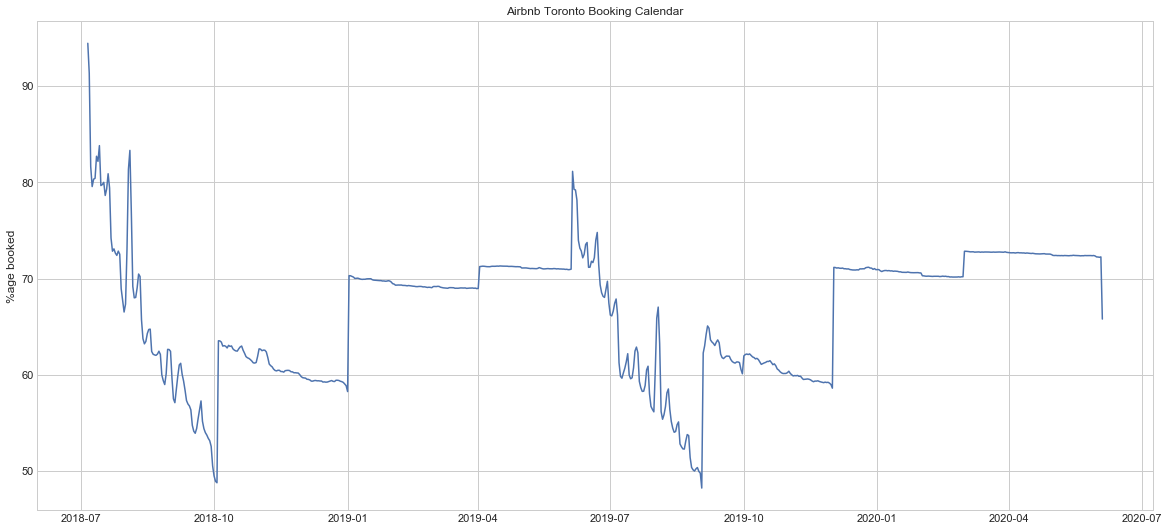
In 77% of the categories, super hosts median price is more than regular hosts.

Avg. Rating?

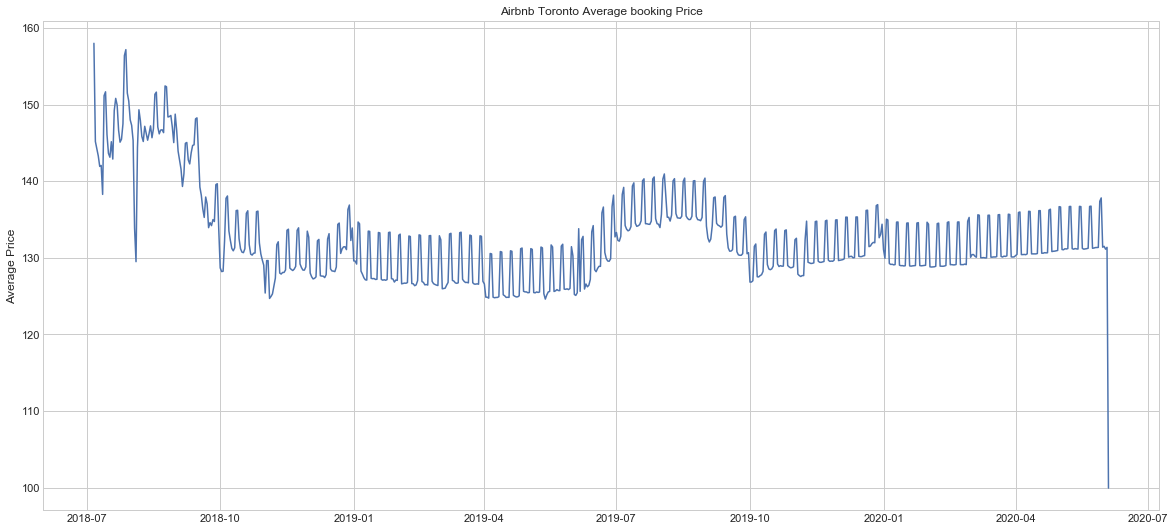
The ….. regular hosts did not qualified due to low rating.

**3.1.5. Analysis of seasonality.**

In this part, we perform time series analysis based on *calendar.csv* dataset and try to figure out what time of the year Airbnb are most popular in Toronto. The dataset covers price and availability for each day between 2018 and 2019 time frame. After calculating seasonality in demand, we can find that summer is the highest booked season and more than half of all the Airbnb listings are always booked.



What’s more, the Airbnb rental price is also affected by seasonality. The following plot shows how price changes over the year by month. It is obvious that the daily average prices of the listings in Summer is more expensive than the rest of the year.



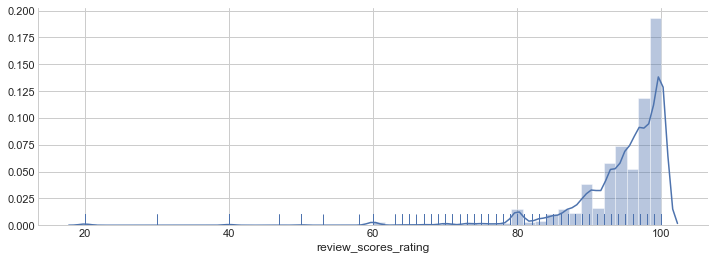
**3.1.6. Analysis of customer reviews**

The customer reviews are insightful and they can tell us a lot of the customer experience, feedback and how well those amenities met their expectations. The textual data mining of customer reviews requires a lot of data cleaning, for example, the words need to be stemmed and stop words need to be removed, etc. Since there are over a million reviews to analyze, we take a random sample - Top50K reviews in our case.

*wordcount* & *wordcloud*

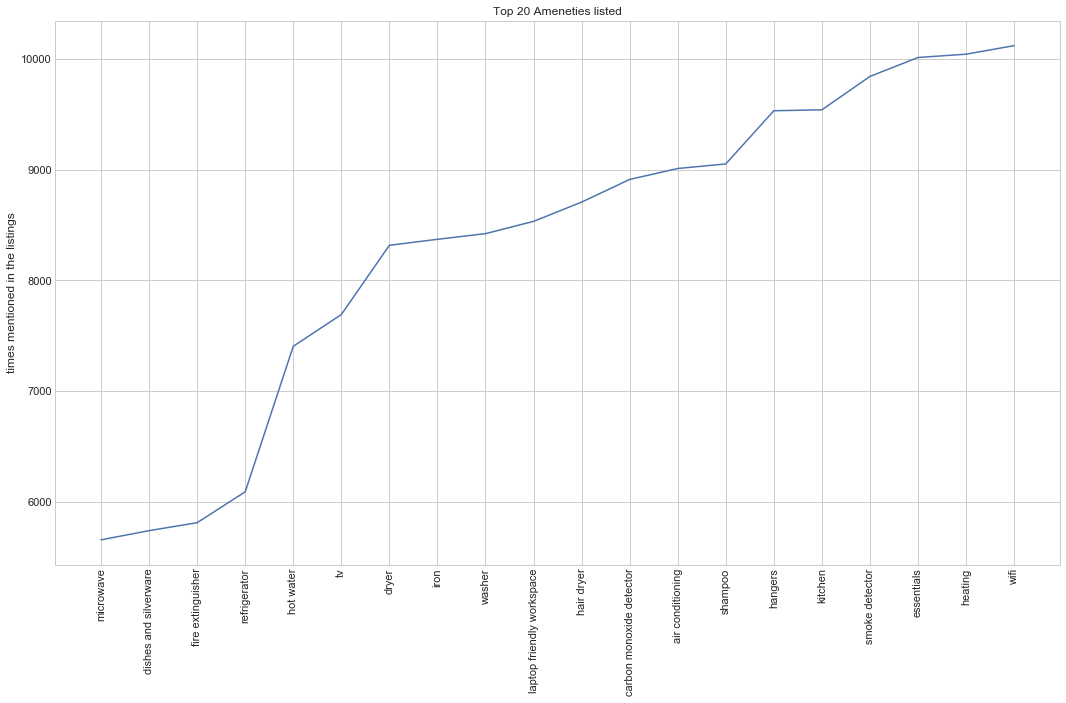
parking/ friendly/ restaurants/ comfortable/ quiet/ clean/ subway

****

There are a total of 10171 Airbnb listings in Toronto, the mean of total review scores rating is 94.494. To sum up, most reviewers leave high scores in Toronto Airbnb rental market.

**3.1.7. Amenities**

In this step, we firstly try to clean the amenities text field and replace insignificant values. In order to print most common word, *wordcount* statement is applied. The output shows that the top 20 most common words are as follows: wifi, heating, essentials, smoke detector, kitchen, hangers, shampoo, air conditioning, carbon monoxide detector, hair dryer, laptop friendly workspace, washer, iron, dryer, tv, hot water, refrigerators, fire extinguishers, dishes & silverware and microwave.

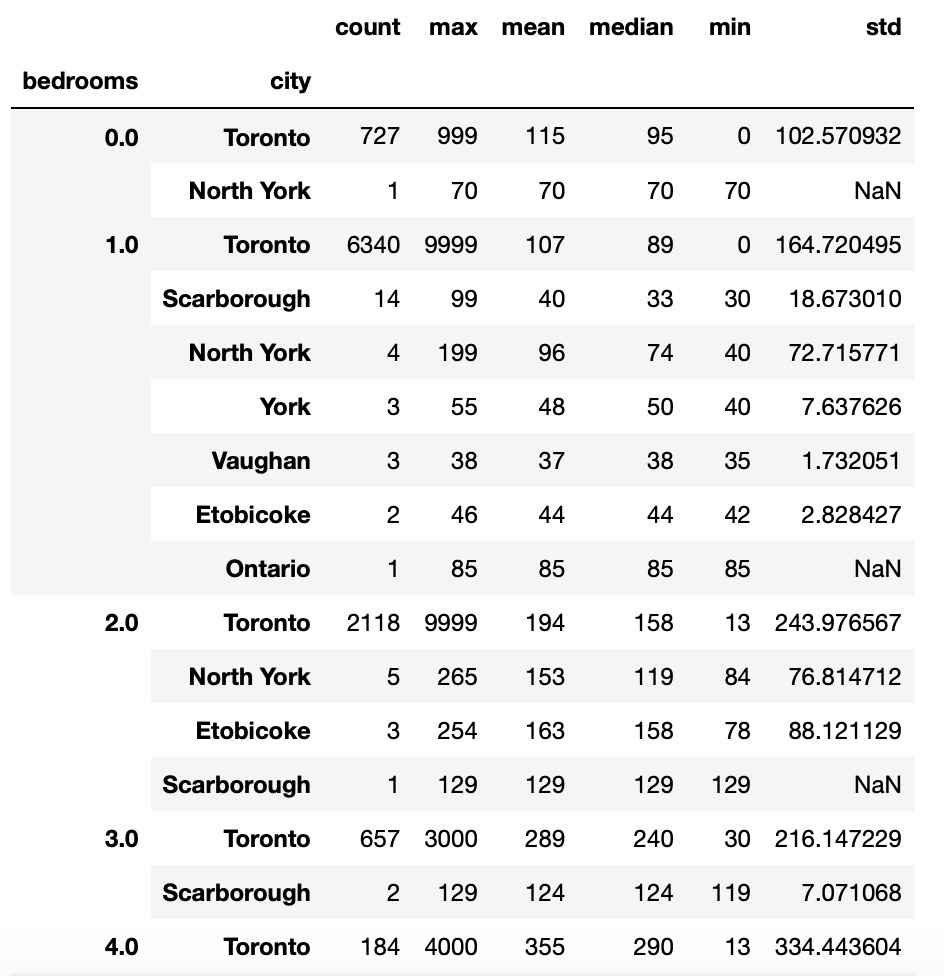
****

**3.2. Analysis of price**

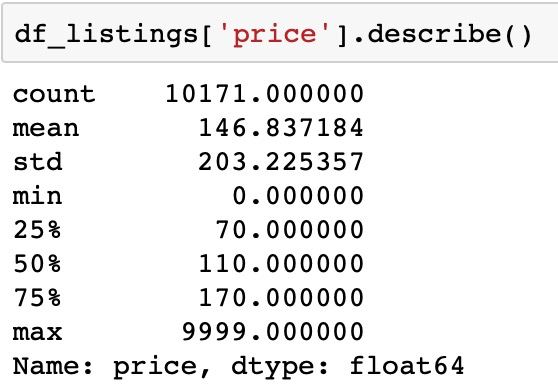
In this part, we explore the price of Airbnb rental market in Toronto from various aspects.

**3.2.1. Pricing Guidelines**

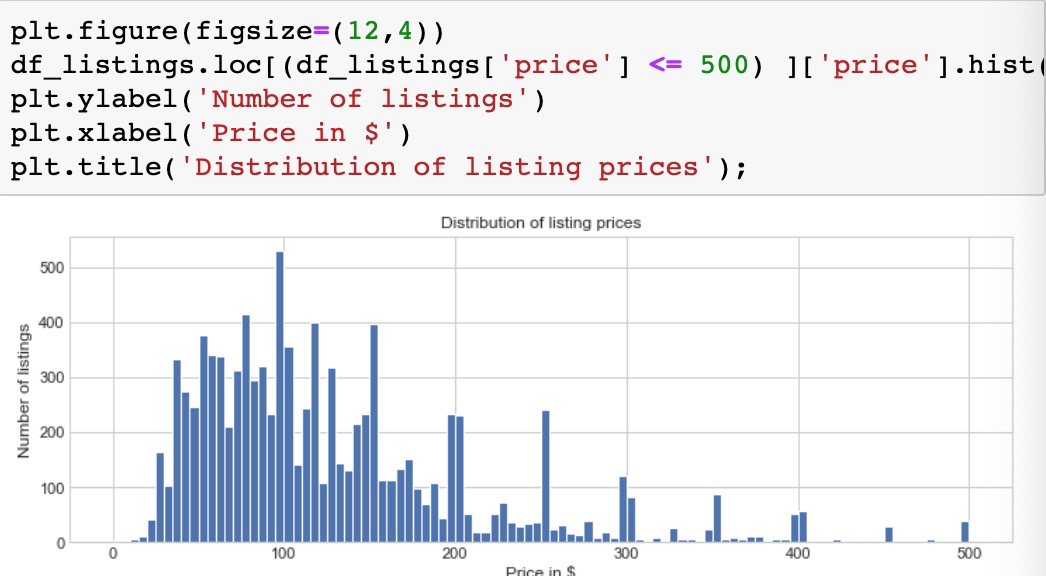
The following analysis gives us an outline of bedroom pricing in different cities of Toronto. It is obvious that the rental price of Toronto is nearly twice as North York, Scarborough, York, Vaughan and Etobicoke no matter how many bedrooms the unit is.

****

On average, the price to rent Airbnb in Toronto is $146.00. The most expensive Airbnb listings in Toronto is $9999 per night.

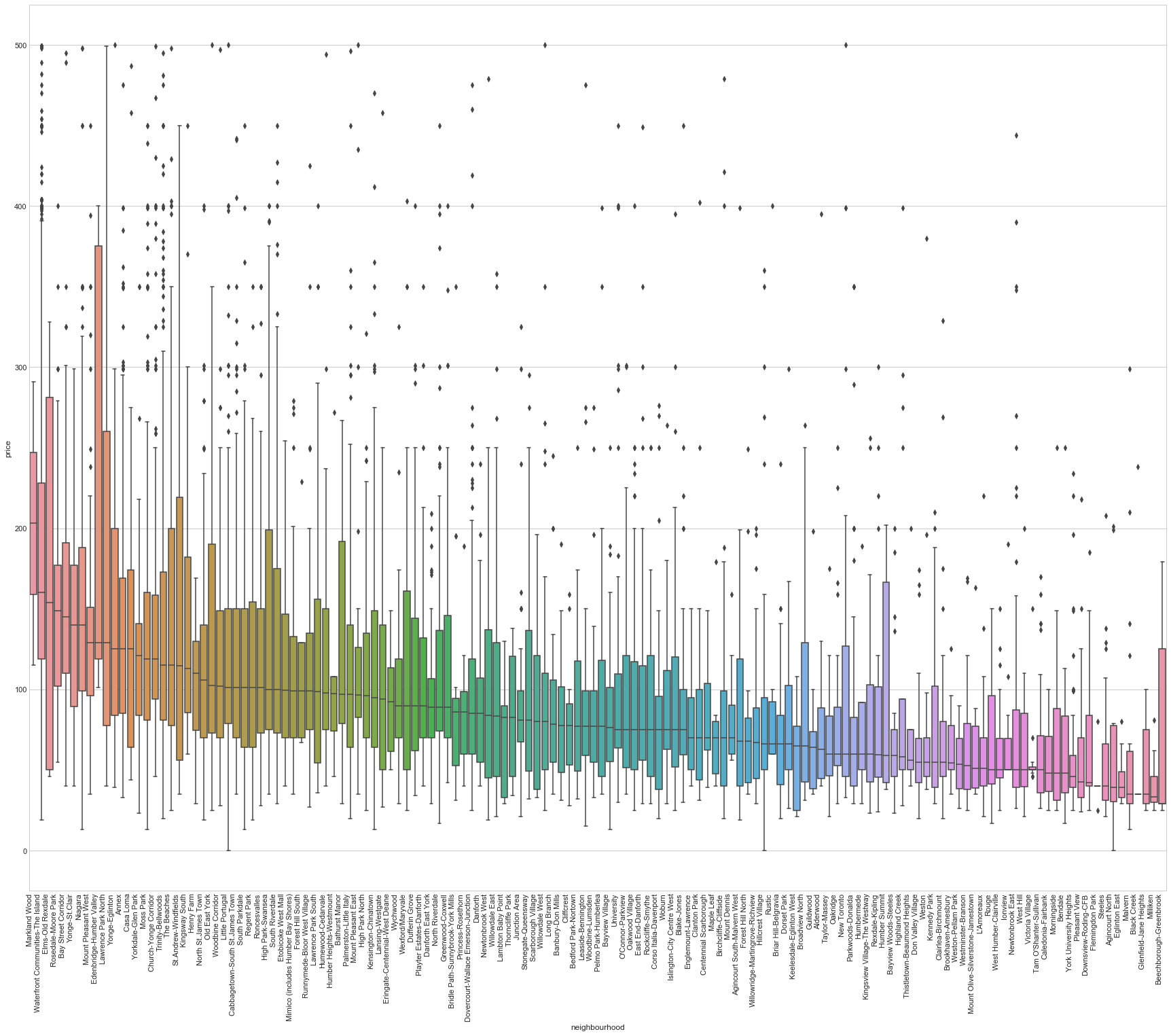


Then, we plot a histogram of listing price to analyze the price distribution. In order not to be affected by the extreme cases, we remove outlier which the listing price is more than $500 per night. The result indicates that most of the Airbnb rental prices are between 0 to $200.



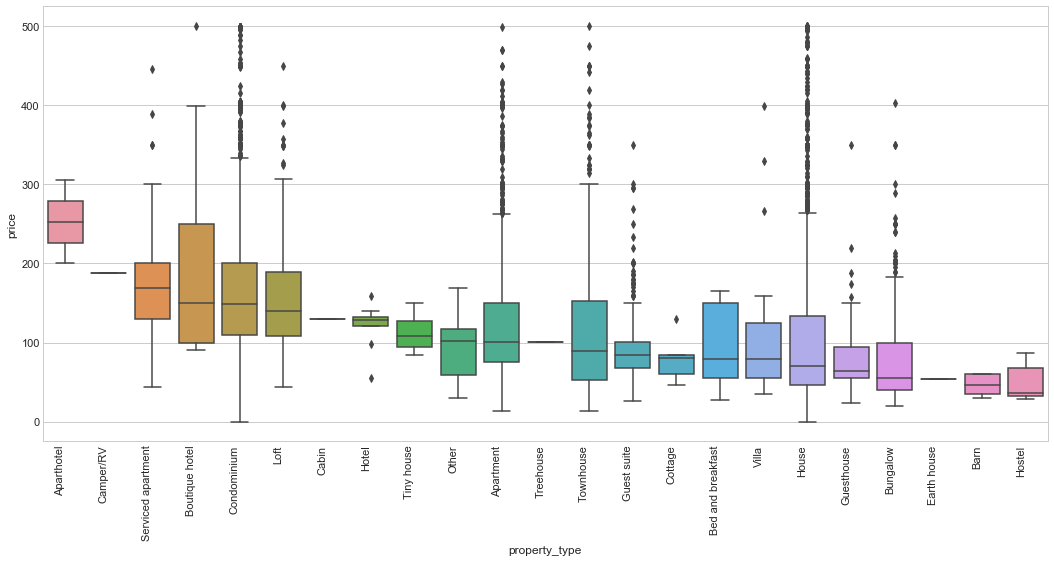
**3.2.2. Neighbourhood vs Price**

In this step, we use *sns.boxplot* statement to visualize the comparison of listing price between each neighbourhood. As we can see, the highest indicator with pink-orange color is Lawrence Park North which presents the most expensive neighbourhood in Airbnb rental market. When we look at the median price for each neighbourhood, Markland Wood is the highest one.

****

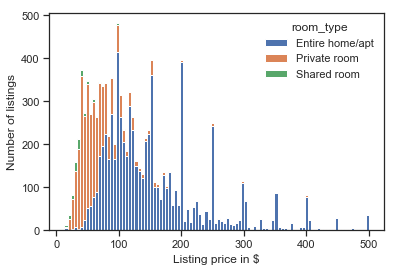
**3.2.3. Property type vs Price**

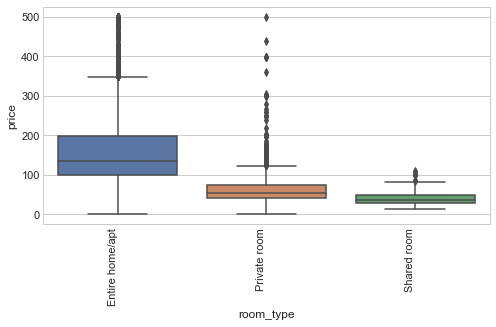
Since the box plot indicates the minimum and maximum price of all of the property type, apparently, we can conclude that the more expensive property types are Aparthotel, Serviced apartment, Boutique hotel, Condominium and loft.

****

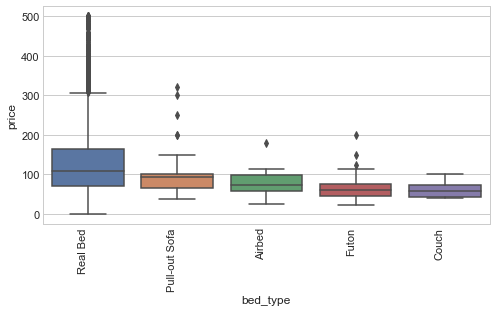
**3.2.3. Room type VS Price**

As we can see from the box plot, there is no doubt that the type of entire room/apt has the highest median price compared to private room and shared room. What’s more, entire home/apt also has the most number of listings.



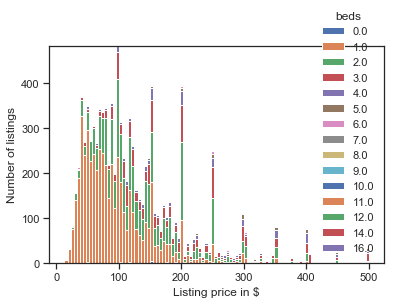
****

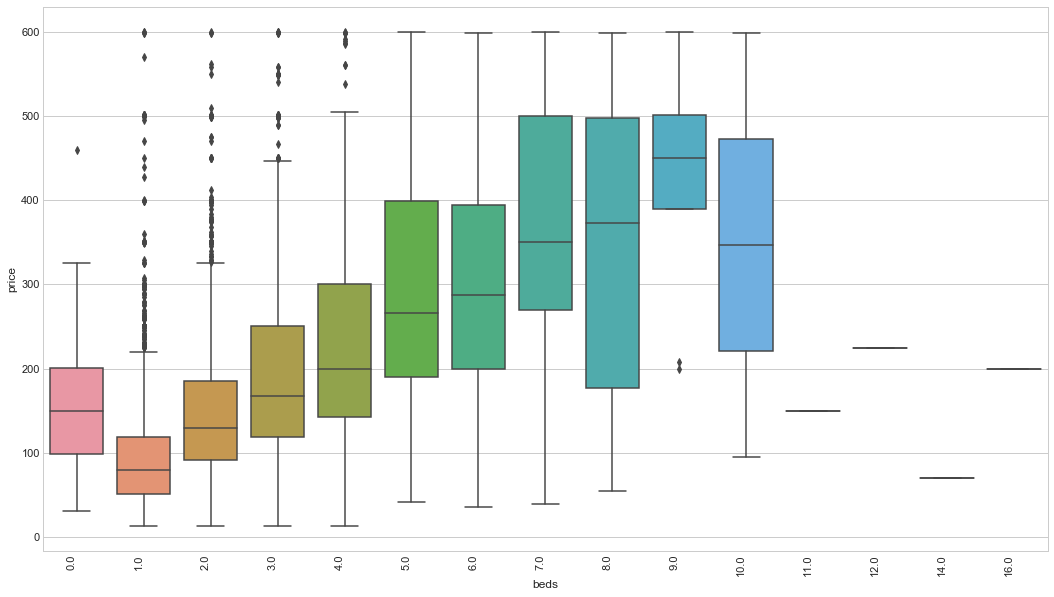
**3.2.4. Bed type VS Price**

****

**3.2.5. Number of beds vs Price**

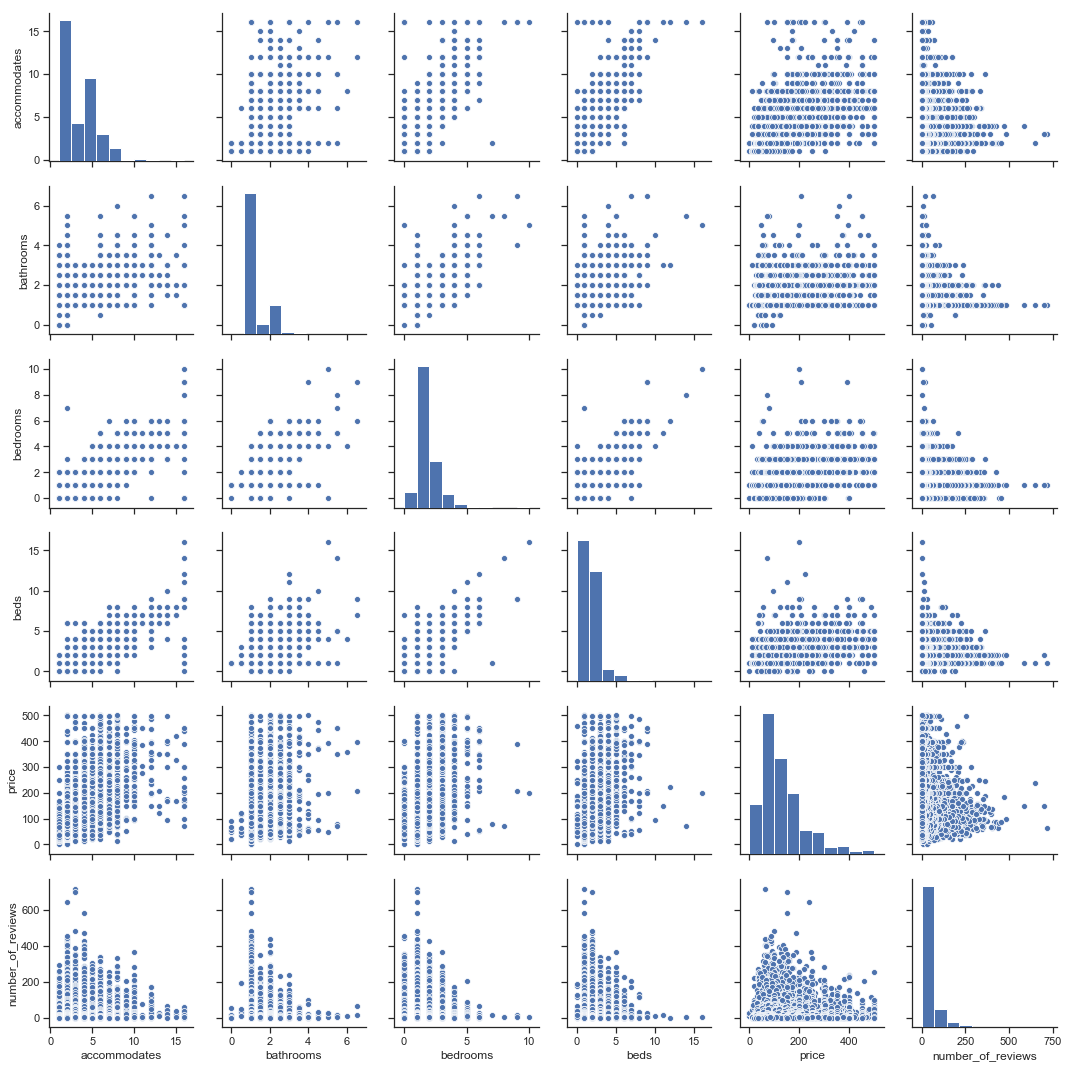
The one-bed listing constitute a significant component of the whole listings and it has a very wide range of prices.

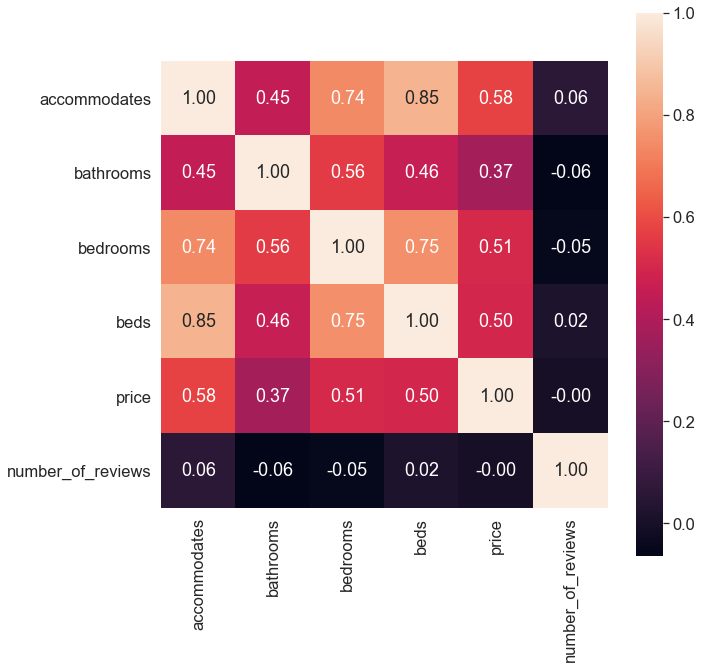
****

****

**3.2.6. Determine correlation between price and various attributes**

SNS/pair plot . What factors affect the daily renting price the most





Number of rooms vs Pricing:

As we can see from this model (Pricing Model vs rooms Graph), as the number of rooms increases then so does the price. A bigger loft can accommodate a huge group of people which means there would be more work for the host in terms of cleaning and preparing. We can also see there are some lodges with more rooms but at the same price as a lodge with fewer rooms. We are assuming the price discrepancy can be attributed to the neighborhood it is in or other factors such as crime rate to explain why they are priced similarly.

**3.2.7. Prediction using Logistic Regression and K-Nearest Neighbors Model.**

Model and Train

Divide the data into X & Y, along with a Train-Test Split

Then train it on a model and predict values on it

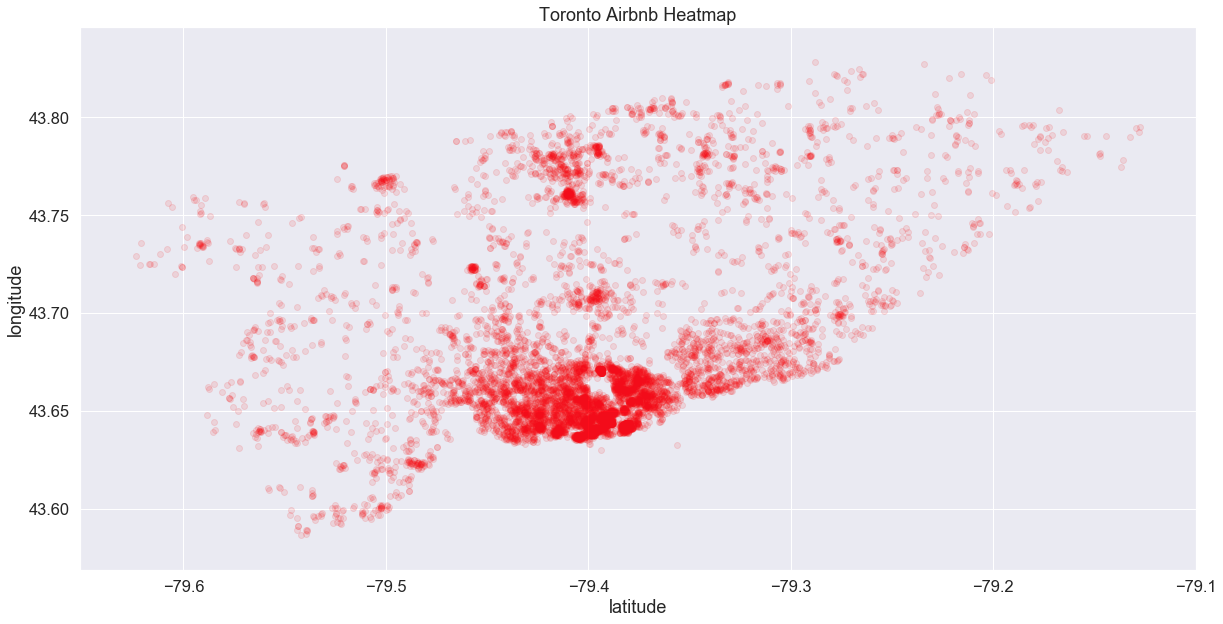
For regression, use Random Forest Regressor

Prediction of price is very difficult as we were only able to achieve about 6% accuracy using two different models (Logistic Regression and kNN). More work/data or different strategy is needed to train and test our model.

**3.3. Heat Map for Airbnb listings**

**3.3.1. Toronto Airbnb Heatmap**

In this section, we try to understand Toronto’s real estate scene by clustering the longitude and latitude of each listing. From the Heatmap, Airbnb listings tend to be closer to downtown and waterfront area, a small part of them gather together in North York region.

****

**3.3.2. Analysis of listings in each neighborhood**

Which neighborhoods are the most popular place in Toronto?

Number of listings/Neighborhood + Map (Table 9)

In Total, we have 10171 listings in each neighbourhood. The Waterfront Communities- The Island neighbourhood is the most popular place in Toronto with the listings number of 2222. In the meanwhile, the neighbourhood of Niagara, Annex, Church - Yonge Corridor and Kensington - Chinatown are quite popular compared to other areas.

Number of reviews per listing/ Neighborhood

**3.3.2. Affordability. Which area is the most expensive?** (Avg. listing price per neighborhood/ Neighborhood + Map)

We can

**3.3.3. What localities in Toronto are rated highly by guests?** (Avg. rating per neighborhood/Neighborhood + Map)

review\_scores\_rating column from listing.csv

**3.3.4. Analysis of crime in each neighborhood.**

To better understand the residential environment and determine where has the highest crime rate in Toronto, we also try to add analysis of crime in each neighborhood. The dataset called *MCI\_2014\_to\_2018.csv* which indicates the summary of Toronto major crime in 2014 to 2018 was procured from Toronto Police Service - Public Safety Data Portal.

In the coding process, we import *re package* for using regular expressions and *re.research* function to clean the city names. Then, we use *groupby* function to determine which neighborhood has the highest number of crimes. As we can see from the output, the neighbourhood of Church-Yonge Corridor, Waterfront Communities-The Island, West Humber-Clairville, Moss Park and Bay Street Corridor are the most dangerous communities. (Table 10)

In order to clarify the correlation between current Airbnb listings and crimes, we import *matplotlib.pyplot* function to visualize our data.

Table 101010

**1.2. Project Objectives**

**1.2.1. Insights about the consumer behaviour and industry trends**

**1.2.2. Data driven Pricing Guidelines**

**1.2.3. Heat map for Airbnb listings**

. **Conclusions:** What did you learn about your data set?

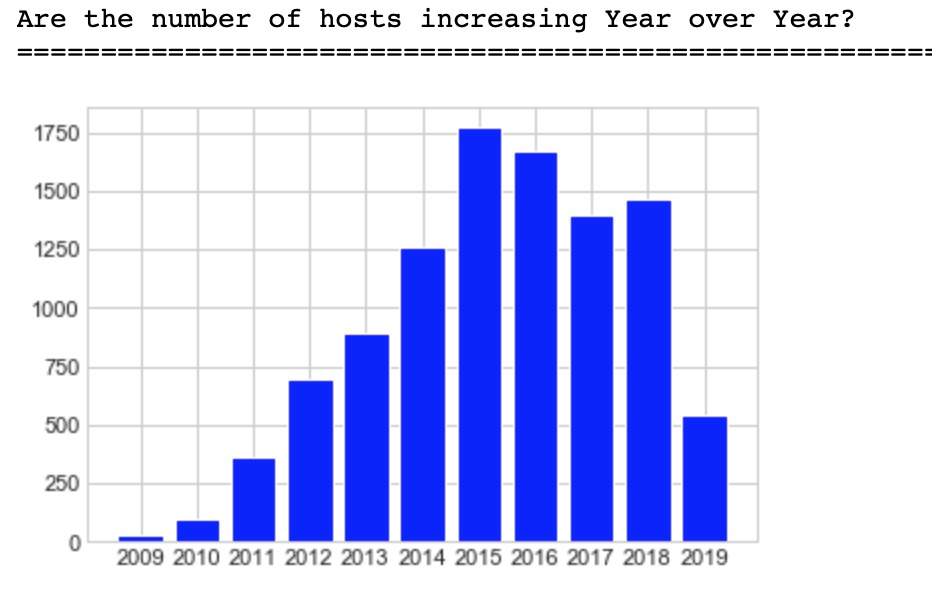
Using these heat maps, we were able to recommend some neighbors to explore for our company

If you’re doing a custom project the outline should include Objectives, Analysis, and Conclusions. The report should be about 15 pages in length. If you would like to include any code and samples of data, in addition to the Jupyter notebook, these should be included as an appendix, not in the main body of the report. Data samples should be limited to less than a page.

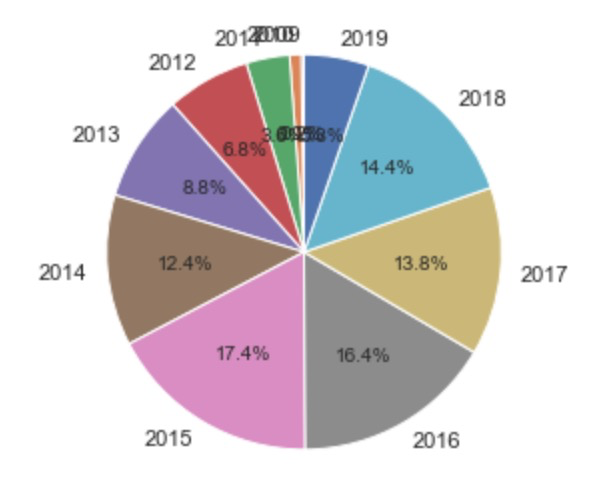
If you refer to any external sources in your report (articles, blog posts, the Statistics Canada website, etc.), you must include the references (link, book title) to the sources of the information that you reference. This includes a URL for the webpage where your dataset came from if from the Internet.

**Appendix**

**(Table 5)**



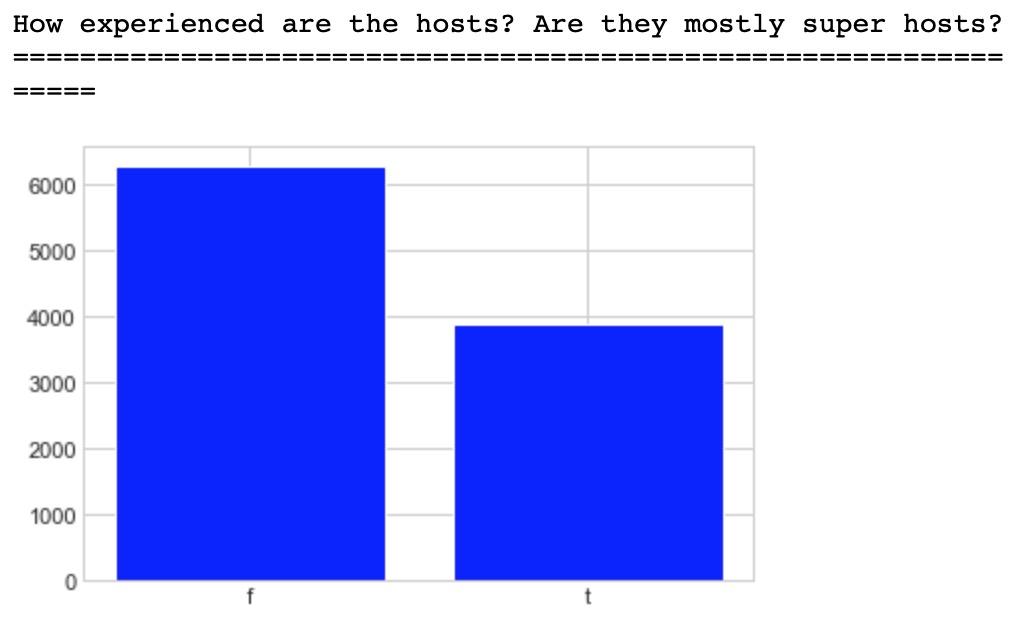
(Table 5.1)



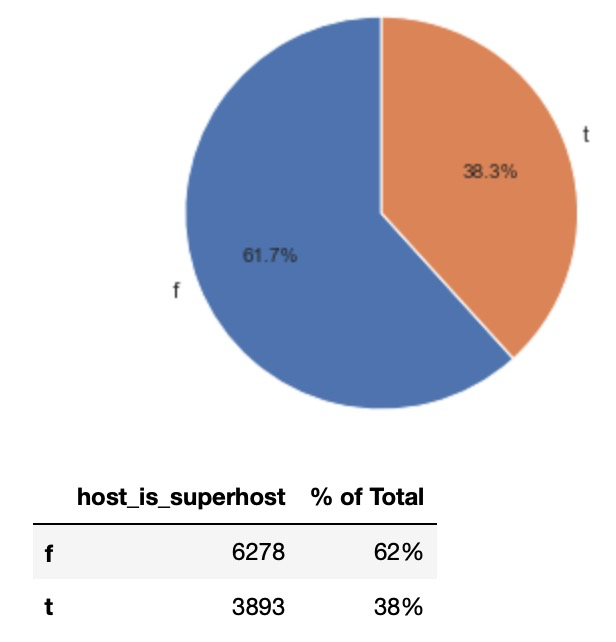
(Table 5.2)



**(Table 6)**

****

**(Table 6.1)**

****

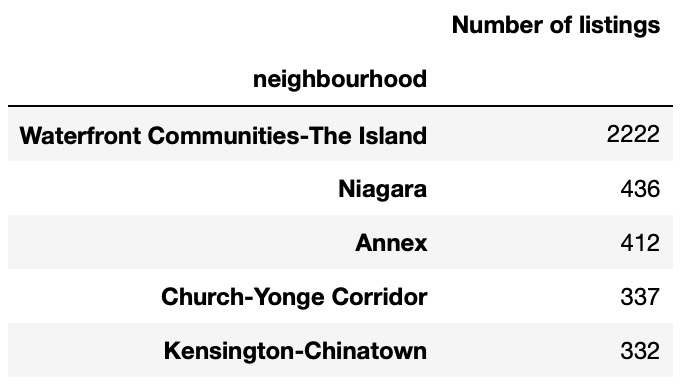
**(Table 7)**



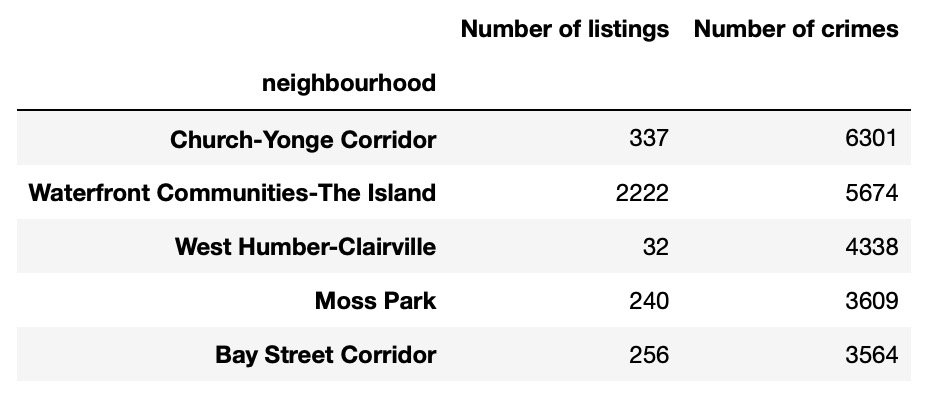
**(Table 8)**



**(Table 9)**



**(Table 10)**



**Bibliography**