Smith SCHOOL OF BUSINESS Queen's University | Global Master of Management Analytics

GMMA 860

The Acquisition and Management of Data

Professor Alex Scott

BUSINESS MEMO

Predictive Pricing Model & Dashboard for Airbnb Properties in Canada

Sunday, May 30, 2021

Team LA

Student Names
Amol Gupta
Gbenga Ilori
Hari Saripalli
Jessie Niles
Jovial Zhang
Patrick Linehan
Shirley Hu

Order of files:

Filename	Pages	Comments and/or Instructions
Team LA – 860 Group Project	26	
Airbnb Dashboard.twb	N/A	Tableau dashboard file Click to view dashboard online
Airbnbproject.R	N/A	R file containing codes
All Files.zip	N/A	Contains source files for all cities: 7 individual files for each city and 1 file with all cities combined

Contents

MANAGEMENT SUMMARY ***1***

TECHNICAL REPORT ***4***

Introduction **4**

Data Source 4

Data Dictionary 4

Section 1: Data Cleaning & Feature Engineering **5**

Merging and Joining 5

Data Cleaning 5

Missing and Incomplete Data 5

Feature Engineering - Dummy Variables 6

Section 2: Data Exploration **7**

Univariate Analysis 7

Bivariate Analysis 11

Section 3: Interactive Dashboard **17**

Section 4: Hypothesis Testing **19**

Model Development 19

Linear Regression 20

Prediction using Test Data 21

Model Limitation / Future Work 21

Conclusions **22**

Appendix A: Data Dictionary ***23***

MANAGEMENT SUMMARY

Overview

Airbnb is a vacation rental online marketplace company based in the US. The main business of the company is based on short-term rental of properties all around the world. It provides a perfect platform for communities of local hosts and short-term tenants to connect and offers unique living experiences to tourists. In this project, we took a deep dive into publicly available Airbnb data for seven sampled Canadian cities[1]. Our business case is two-fold: One is from a host perspective and the other from a guest perspective.

New hosts often struggle to properly price their property. They risk losing revenue if priced too low and may lower occupancy rates (resulting in low revenue) if priced too high. To address this issue for hosts, we analyzed Airbnb datasets and developed a predictive model to determine the price of a listing based on different features in the data set. Ultimately, a potential Airbnb host who wants to list a property on the website can use this model as a guide to price their listing and understand what drives the occupancy rate and revenue. Also, we provided a guide to help a potential host decide whether to rent property long-term or list it on Airbnb to maximize revenue from the property.

For guests visiting a new city, it can be difficult to know how much to expect to pay for accommodation. Also, when evaluating places to stay, it is hard to know if the property is listed at a fair price. To help guests with this, we developed a dashboard that provides information on all the listings in each city with real-time filters to select the listing of choice and respective pricing. Ultimately, we try to answer the question: 'How much should I expect to pay in my choice of neighborhood in each city?'

Summary of Key Findings

1. Price is most sensitive to:
    - Location factors: Listings in Toronto, Vancouver, and Victoria drive higher prices.
    - Room type: Price is lower for a shared room than a private room.
    - A higher number of bedrooms, bathrooms and larger sized listings will drive higher prices.

2. Revenue is more correlated with occupancy rates than with price. Superhosts tend to get a higher occupancy rate than a non-superhost. For a new host, occupancy rates are compared with non-superhosts for the purposes of analysis.

Future Work

This project provides hosts pricing guidance when listing their properties on Airbnb. This will offer insight into projected revenue but does not inform the host of potential profitability of their property. We propose future work to analyze the expenses (including purchase costs) hosts expect to incur in each city to provide a more wholistic and granular picture for their investment analysis.

[1] Seven Canadian cities sampled are: Toronto, Vancouver, Montreal, New Brunswick (province data only), Quebec City, Ottawa and Victoria. The data are sourced from both the insideAirbnb (http://insideAirbnb.com/) and the Canada Mortgage and Housing Corporation (https://www.cmhc-schl.gc.ca/) websites.

Location, Room Type, Number of Rooms, and Accommodation Capacity are strongest predictors of Price

Our price predictor model based on linear regression uses data from seven Canadian cities. The model identifies 10 variables that are statistically significant in predicting prices. A potential host who wants to list a property on the Airbnb website can use the model to price their listing to maximize the occupancy rates and revenues.

The relevant variables are listed below.

1. Number of people the listing can accommodate
2. Number of bedrooms
3. Number of bathrooms
4. Room type: Private or Shared
5. Location: Toronto, Victoria, or Vancouver
6. Reviews: Number and Score Rating

Total revenue is a function of both price and occupancy rates. To maximize revenue, hosts need to also consider occupancy rates in the location of choice.

Superhosts have higher occupancy rates and higher monthly revenue for their listings across all seven cities. They have more visibility on the Airbnb listings, higher earning potential, and access to exclusive rewards. This is highlighted in Figure 1.

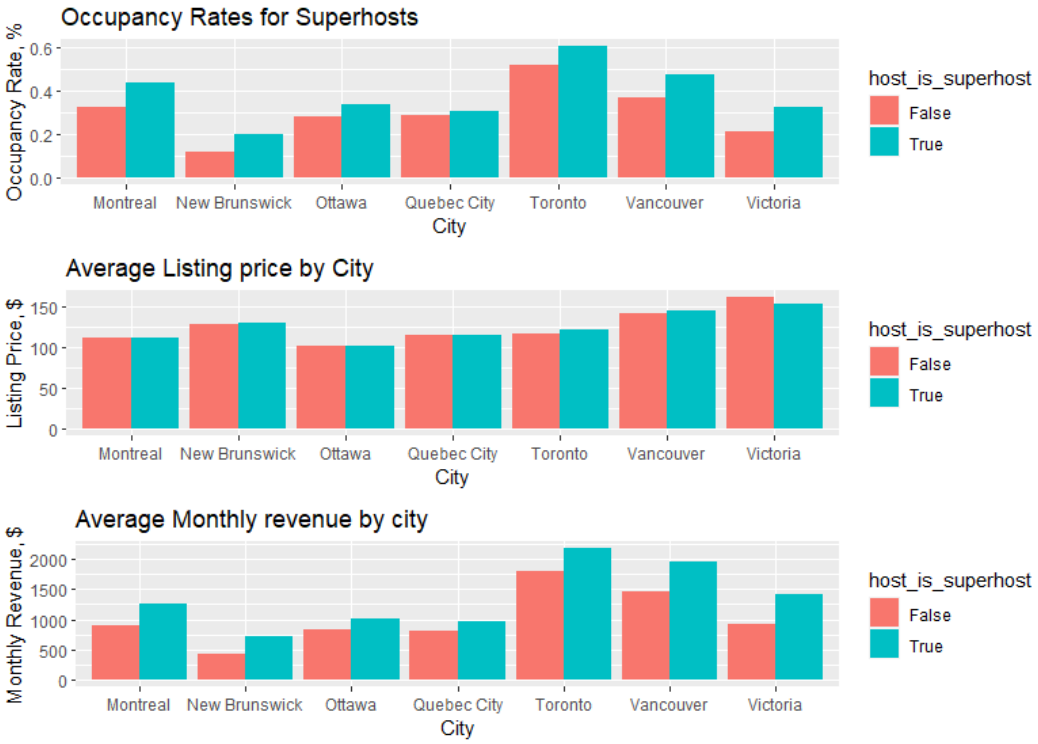


Figure 1. Listing Overview

Interactive Dashboard Provides Information to Aid Listing Selection and Investment Evaluation

A dashboard with filters was built for each of the seven Canadian cities and displays KPIs including:

- Average rental price for the entire city.
- The average number of beds for all the listings.
- The average number of days of availability for rental over the future thirty calendar days.
- The average minimum and maximum number of nights that can be rented.

Click to view dashboard

The dashboard also shows a map of the distribution of all the Airbnb rental properties by neighbourhood. The size of each dot represents the average rental price, among other features. A potential guest can drill down to the listing of interest or make comparison across different cities and/or elements as needed. A capture of a dashboard for one of the cities can be seen in Figure 2. Click the blue button above to view the entire dashboard online.

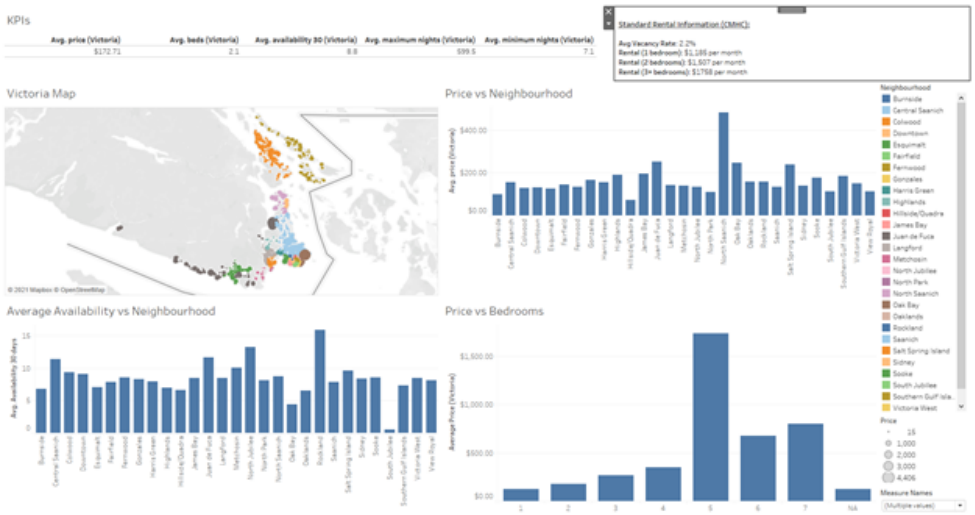


Figure 2. Dashboard snapshot for Victoria

The dashboard is also useful for potential investors who may be seeking to purchase a rental property in one of the seven cities. It incorporates traditional rental information from the Canada Mortgage and Housing Corporation (CMHC) that shows the average vacancy and rental rates. This can be used to compare rental rates and average vacancy in the location of interest against Airbnb occupancy and pricing information using the various filters on the dashboard.

TECHNICAL REPORT

Introduction

Data Source

Data for this project was obtained from *insideairbnb*[2] (http://insideairbnb.com/). Data from seven cities from across Canada was obtained. An overview of the datasets can be seen in Table 1 below.

Table 1. Summary of Observations

City	No. of Observations
Montreal	3,271
New Brunswick	1,945
Ottawa	2,587
Quebec	2,289
Toronto	15,542
Vancouver	4,299
Victoria	2,919
TOTAL	**32,853**

Data Dictionary

We modified an existing data dictionary from the *InsideAirbnb* website to clearly identify what each variable means. The modified dictionary reduces the number of fields and further populates the dictionary information. The dictionary can be found in *Appendix A*.

To take a closer insight from short-term rental industry in Canada, we have chosen major data sets from seven cities to analyze. The following factors might be related to the pricing strategy:

- Location (latitude and longitude) to illustrate area in each city
- Response rate: how fast host will response to customer request
- Neighbourhood: used for identifying cities and district of city
- Property type: private room, entire house, entire apartment and entire loft will be main types for analysis
- Room type: like property type to show the features of room
- Bathroom: the number of bathrooms in use
- Bedroom: the number of bedrooms in the renting area
- Price
- Minimum and maximum nights: requirements for the duration of renting time
- Availability: defined as availability of the listing X days in the future as determined by the calendar
- Reviews: number, scores, accuracy

A detailed description of each factor can be found in *Appendix A.*

[2] Inside Airbnb. (2020, August). Get the Data - Data Dictionary. Retrieved May 22, 2021, from Inside Airbnb: http://insideairbnb.com/get-the-data.html

Section 1: Data Cleaning & Feature Engineering

Merging and Joining

The dataset for each city was loaded into RStudio. Since all seven worksheets have identical headings, we used the *rbind()* function to merge seven worksheets into one single worksheet. To join the data from all seven cities, we started by adding the city name to a new column in each individual dataset. Next, we consolidated all seven tables into one master table.

Data Cleaning

We use *str()* command to show the structure of the data frame. To clean the data, we conducted the following:

- Filtered data to select 40 of 75 variables to be used in analysis (by using *select()* function)
- Converted "City" and "Room Type" to factors (by using *as.factor()* function)
- Converted strings to numeric as required (by using *as.numeric()* function)
- Converted characters to binary as required (by using *as.binary()* function)

Many of the text in the character-type variables, including "name" and neighborhood overview", had line break (
) characters that needed to be cleaned. We used *gsub()* to remove the many instances of "br" from the descriptions.

Missing and Incomplete Data

To begin investigating any missing and incomplete data in our dataset, we displayed all the missing data. We did this by using *md.pattern()* command.

After running the above command, the plot we generated can be observed in Figure 3. The rightmost column shows the number of missing variables in the missing pattern:

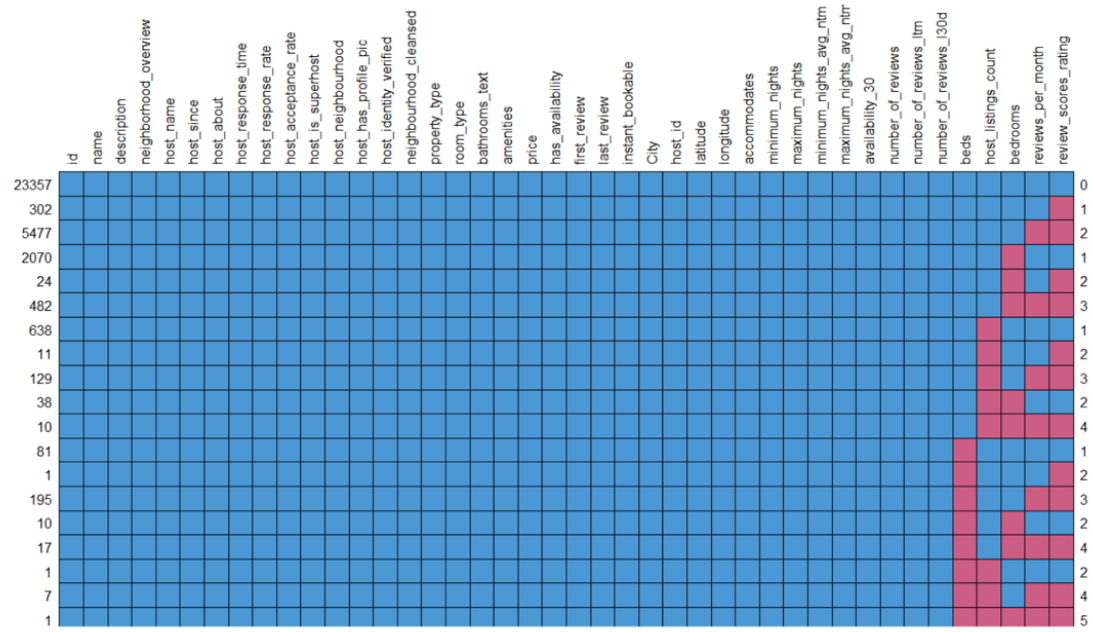


Figure 3. Missing Data Visualization

We ran the *dimensions*() command to check the total number of rows and columns. We reduced the data to 40 columns and 32,852 rows.

Following that, we removed all the missing values out from the dataset. We used the *na.omit()* command to do this.

We then checked the pattern again to see if there are any missing variables remaining. Figure 4 shows the plot we generated based on the command we ran:

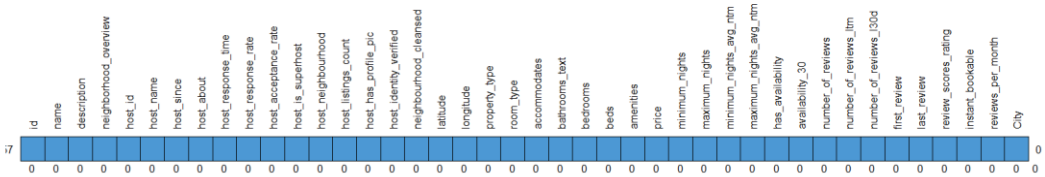


Figure 4. Data with missing values removed

We removed the missing variables but there were still some "N/A" values that remain in the dataset. We first deleted the N/A values in the character columns and then omitted them. We installed and loaded the *"naniar"* package to use and ran the *replace_with_na()* command. This replaced the N/A values with blank values. We then used the *na.omit()* command to delete the blank values in *host_response_rate* column. Complete rows were deleted in the *host_response_rate* column that has the blank values in it.

Feature Engineering - Dummy Variables

First, we converted the *City* and *Room Type* columns to a *factor* category as this represents a categorical variable. Next, using the *ifelse()* command, we created new dummy variables for each city. This allowed us to capture the differences in each city in our model.

Section 2: Data Exploration

Univariate Analysis

We investigated the summary statistics, such as the maximum, minimum, range, mean and mode for the numeric variables in our data. Then we visualized the distribution of these numeric variables (excluding indicator variables) on a series of histograms.

We divided the data into different subcategories such as host information, accommodation information, price, reviews, duration of the stay, availability, and the information about the reviews. Below is the summary of our findings.

Review of host data:

- 31.43% of hosts are Superhosts
- The average host response rate is 91% and the average host acceptance rate is 82.3%. 31.43% of hosts are Superhosts
- The average number of listings for a host are 9.315 listings
- 99.66% of hosts have a profile picture
- 80.95% of hosts have their identity verified

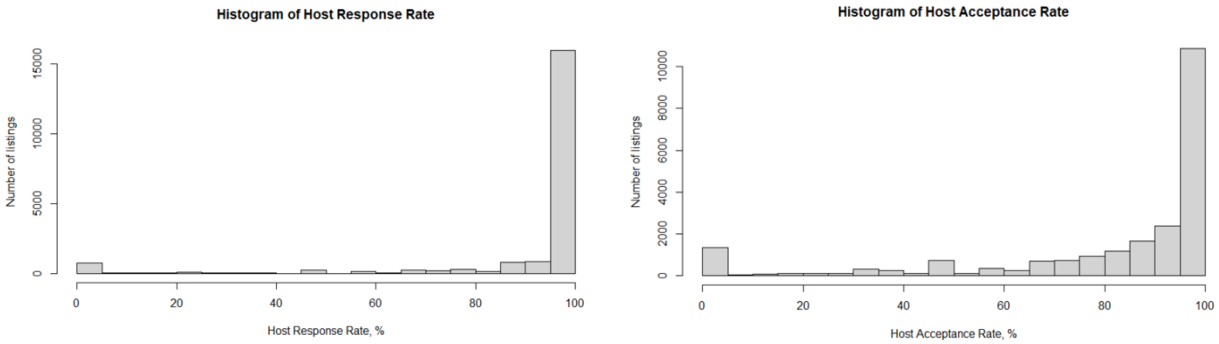


Figure 5. Visualization of host data

Review of accommodation information:

- The average number of people that can be accommodated per listing is 3.352
- The average number of bedrooms is 1.544
- The average number of beds in 1.8
- The average number of bathrooms is 1.206

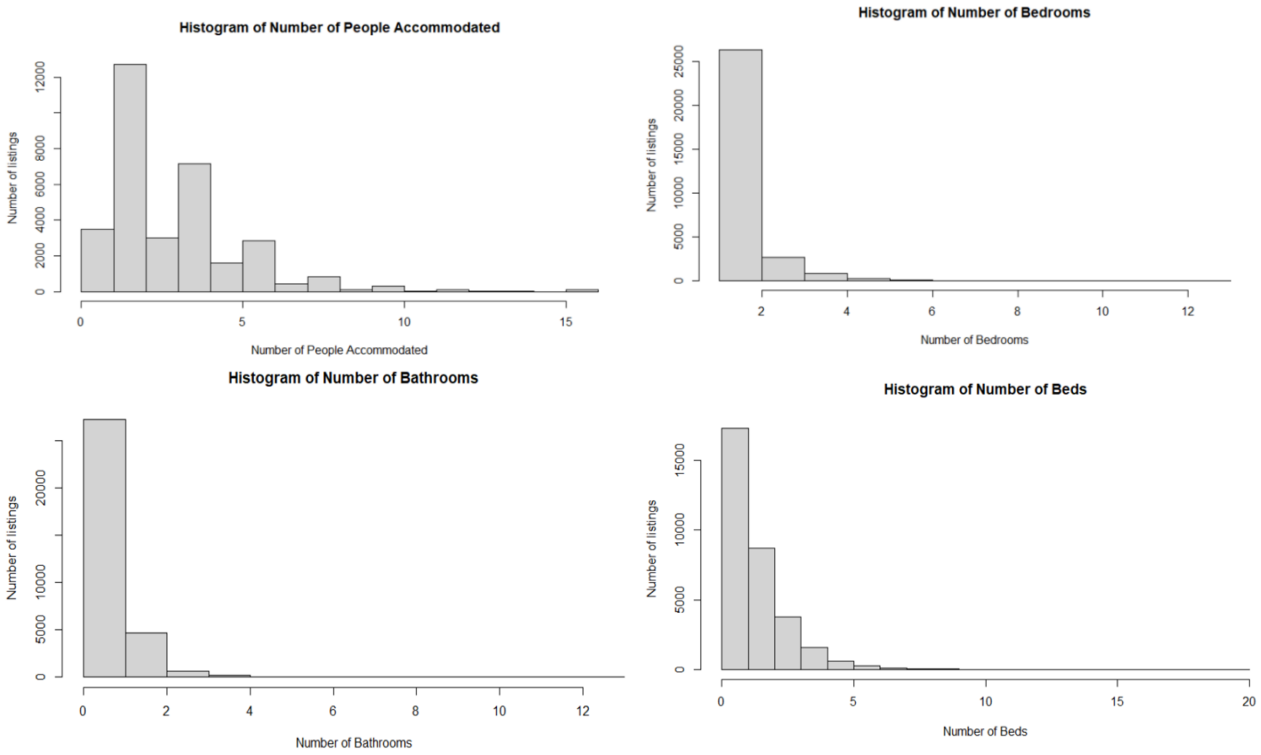


Figure 6. Visualization of property information

Review of pricing information:

- The average price is $125.20

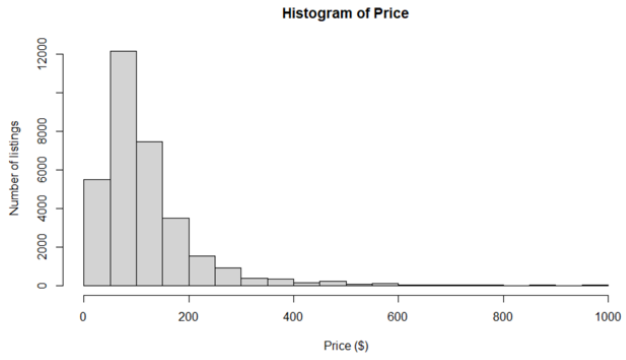


Figure 7. Visualization of pricing information

Review of stay duration information:

- The average minimum number of nights is 12
- The average maximum number of nights is average maximum number of nights is 31,182

Review of availability information:

- 97.23% of listings have availability within the next 30 days
- The average number of nights available within the next 30 days for all listings is 10.6
- 32.05% of listings are instantly bookable

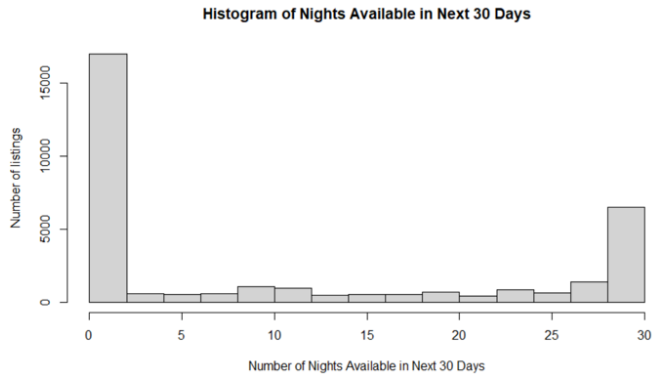


Figure 8. Visualization of listing availability

Review of review information:

- The average number of reviews that a listing received is approximately 33
- The average number of reviews that a listing received in the last month is approximately 4
- The average review scores rating for a host is 94.76%
- The average number of reviews per month that a listing receives is 1.187

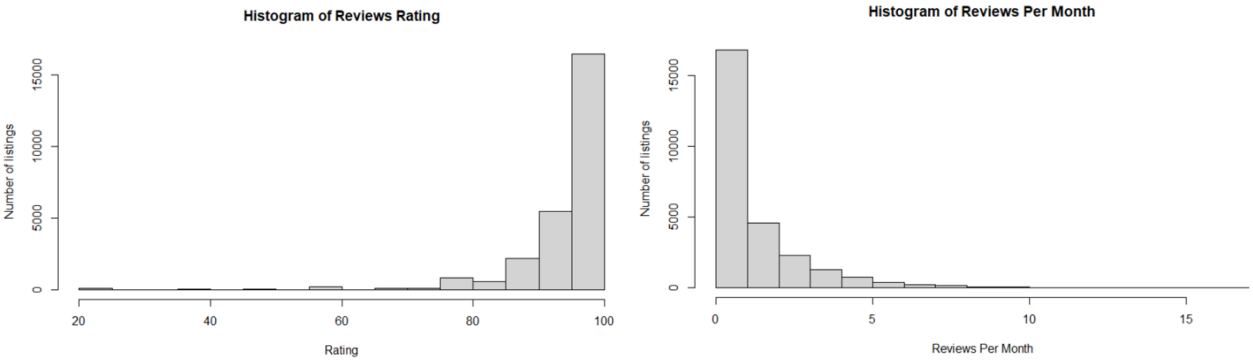


Figure 9. Visualization of review information

Top Keyword Analysis

We explored the most common words used in listing names, descriptions, and neighbourhood overview.

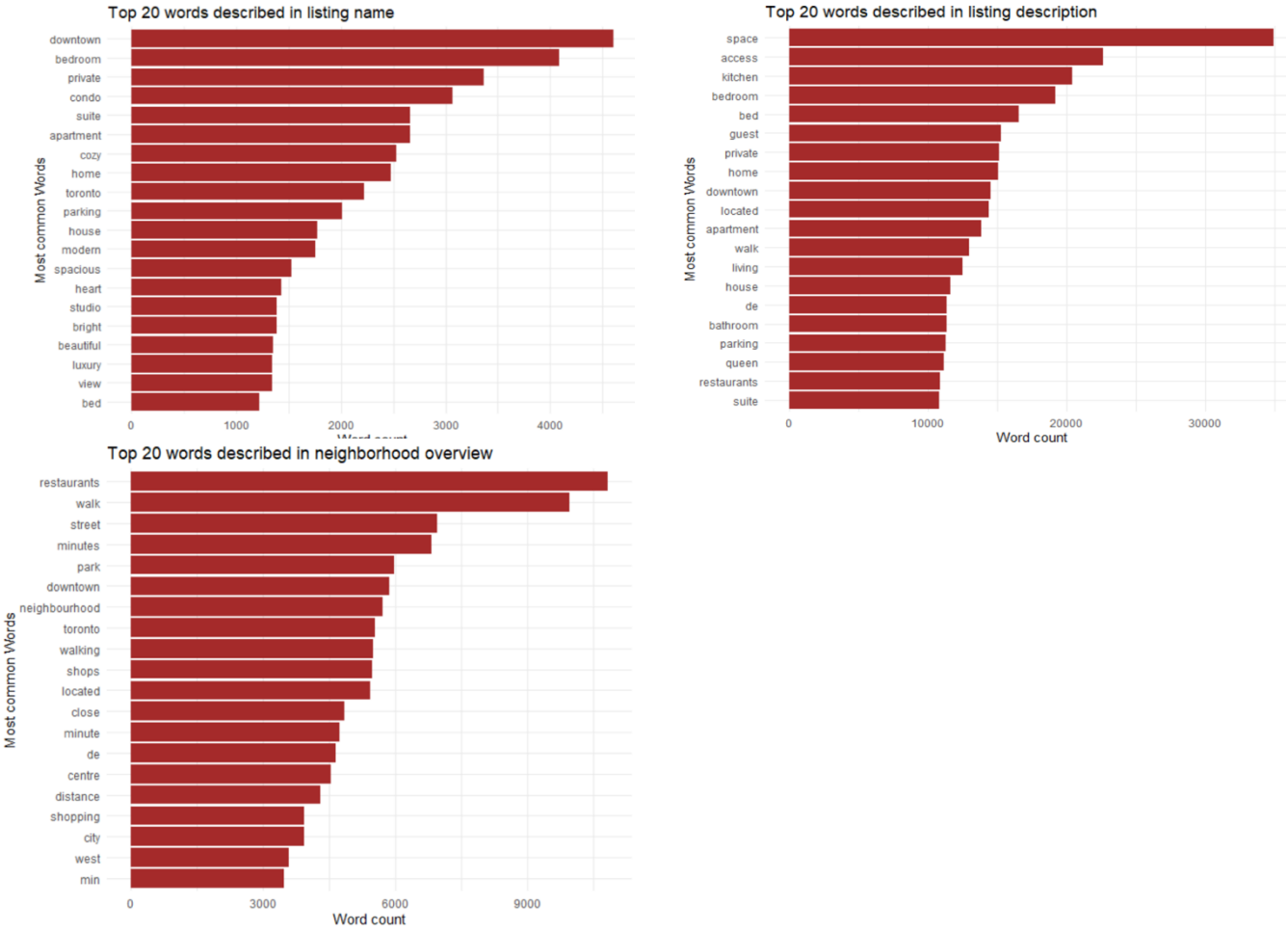


Figure 10. Top words used in listings

Figure 10 highlights the top words mentioned in all listings. The keywords that standout are “downtown”, “restaurants”, and “private”. This suggests that most listings like to highlight that either they are close to downtown, if not in downtown itself, that they are in a close proximity to good restaurants, and that they would like to emphasize privacy in their listings.

Bivariate Analysis

Correlation

We conducted a correlation analysis on all numeric variables. Some of the highlights of the results of the analysis are as follows:

- Price is moderately positively correlated with “accommodates”, “bedrooms”, “beds”, and “bathrooms”. This makes sense intuitively.
- Reviews per month is moderately positively correlated with “number of reviews”, “number of reviews last month” and “number of reviews last 130 days”. This makes sense intuitively.

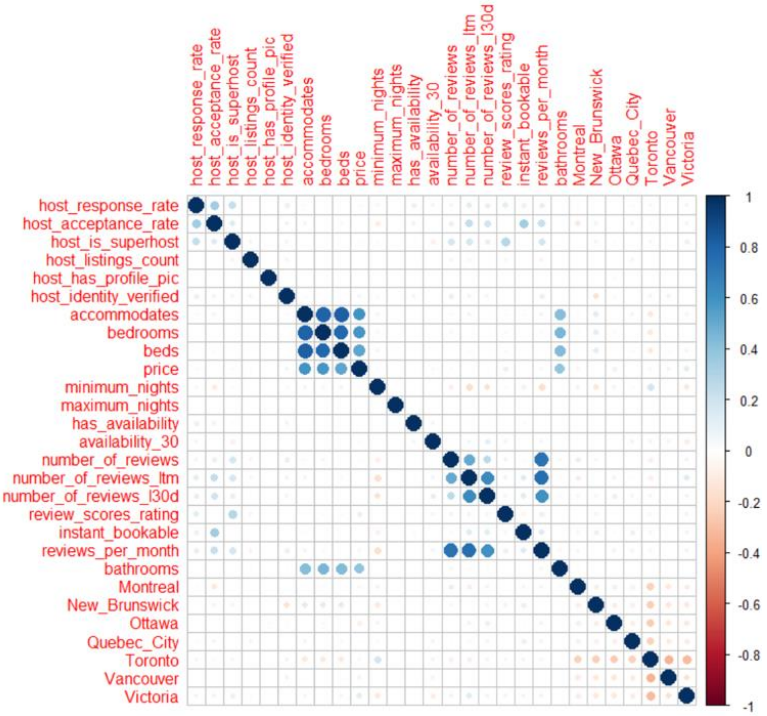


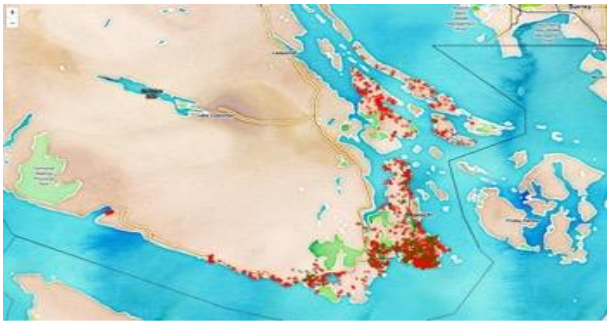
Figure 6. Correlation plot

Price vs. Location Analysis

We used the leaflet package to visualize the price and location of each listing. The results of this can be seen in the heat maps in Figure 12. Red circles on the map represent listings over $100 and green dots are listings below $100. In general, we can see the listings in and around the downtown areas of cities are more expensive than the rest.

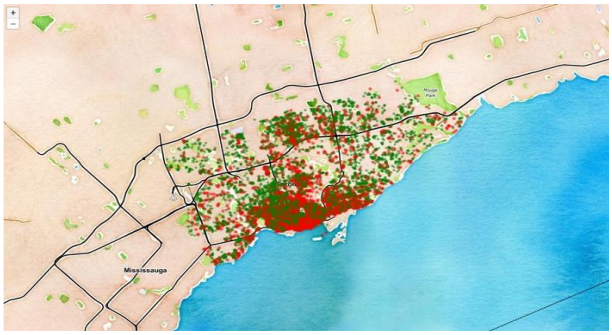
All Cities

Victoria

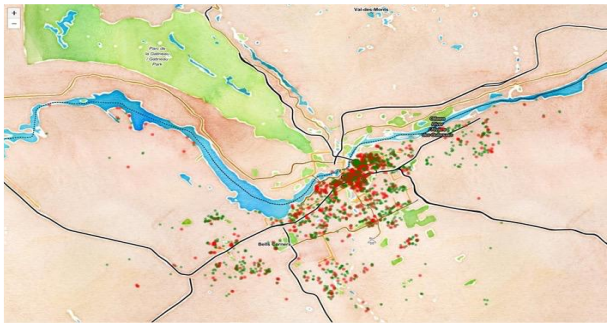


Vancouver

Toronto

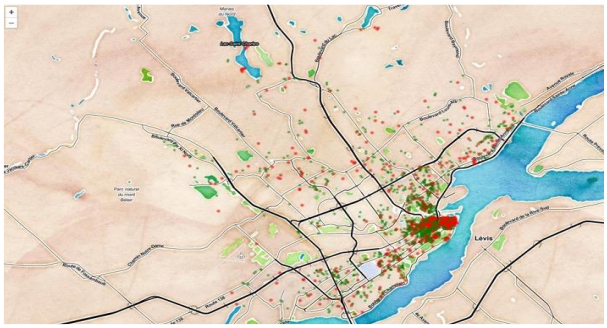


Ottawa



Montreal

Quebec City



New Brunswick

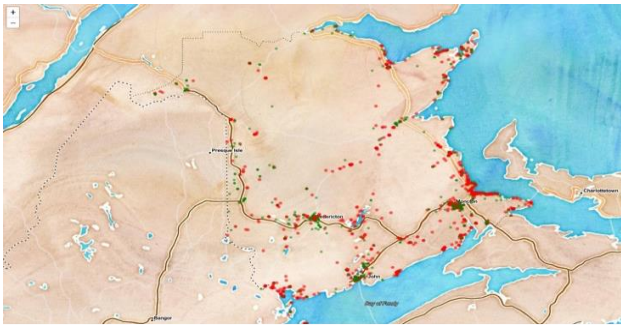


Figure 12. Snapshot of listing location by city

Price Comparison

We compared the average price in each city to five other variables: City, Number of Beds, Number of Bedrooms, Number of Bathrooms, and Number of People Accommodated.

Victoria has the highest average price per listing, followed by Vancouver and Toronto. Montreal has the lowest price per listing. Average price initially rises sharply as Number of Beds, Bedrooms, and Bathrooms rise but subsequently begins to decline. However, we suspect that there may not be enough data points for listings with more than 9 bedrooms and 8 bathrooms. We also observed that the price goes up if the place can accommodate more people. These findings can be observed in Figure 13.

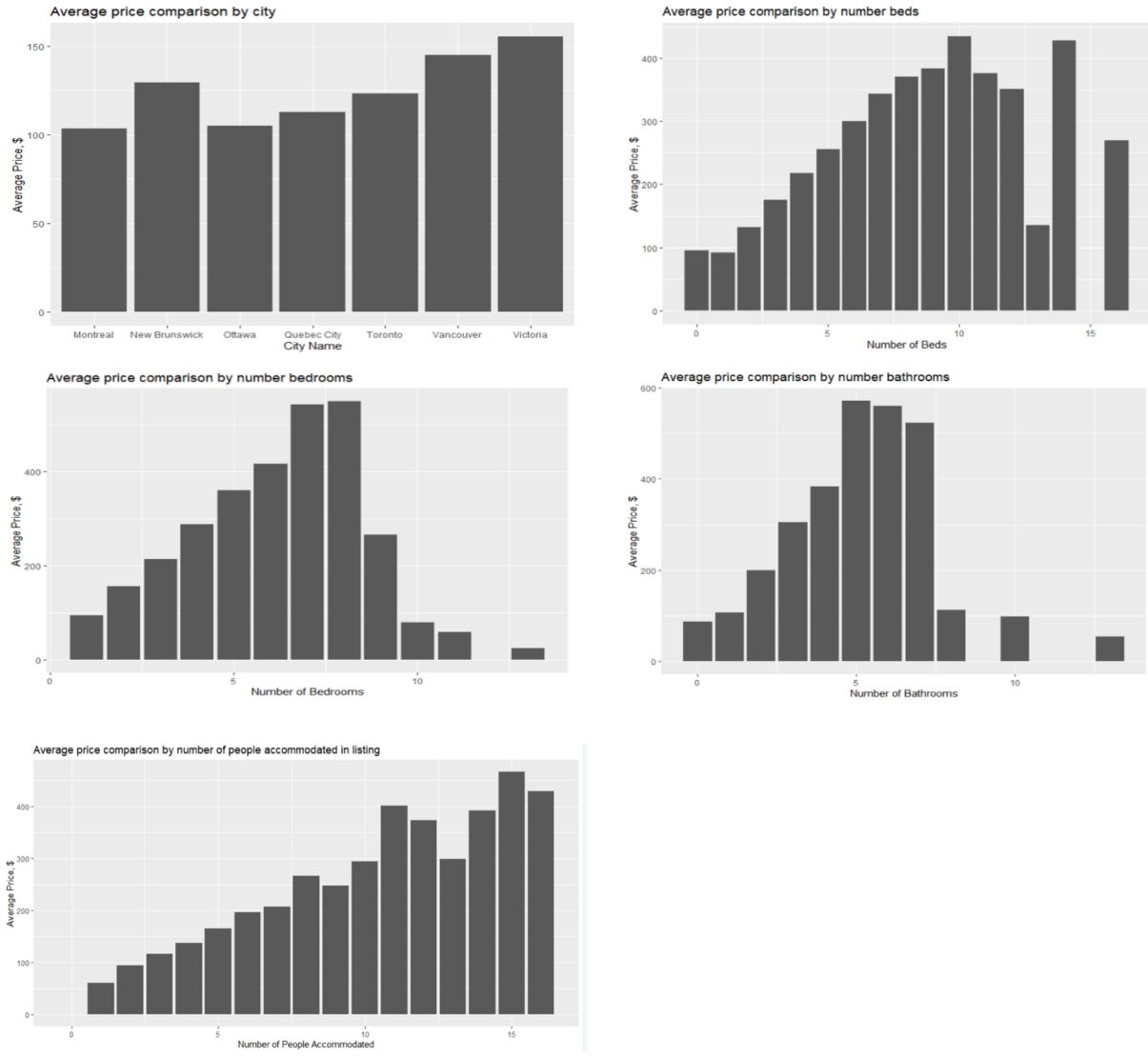


Figure 13. Price analysis

Occupancy Rates and Monthly Revenue

To build our occupancy model, we needed to make an assumption on the guest review rates. We have seen data suggesting review rates of 30% to 70%. We decided to choose a conservative estimate of a review rate of 50%. This means that only 50% of the guests leave a review after their stay. We used this in conjunction with the minimum nights for the booking to estimate the days of occupancy for a listing. Monthly revenue was then calculated using the price of a listing multiplied by the number of days the listing is occupied. A breakdown of revenue and occupancy rates by city can be seen in Figure 14.

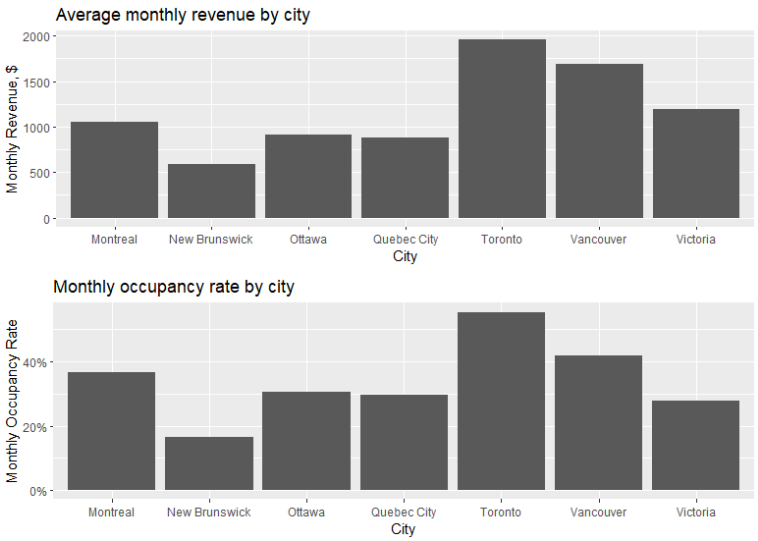


Figure 14. City overview

It is interesting to observe that the monthly revenue for a host is more correlated with the occupancy rate compared to the price of a listing. This can be seen in Figure 15.

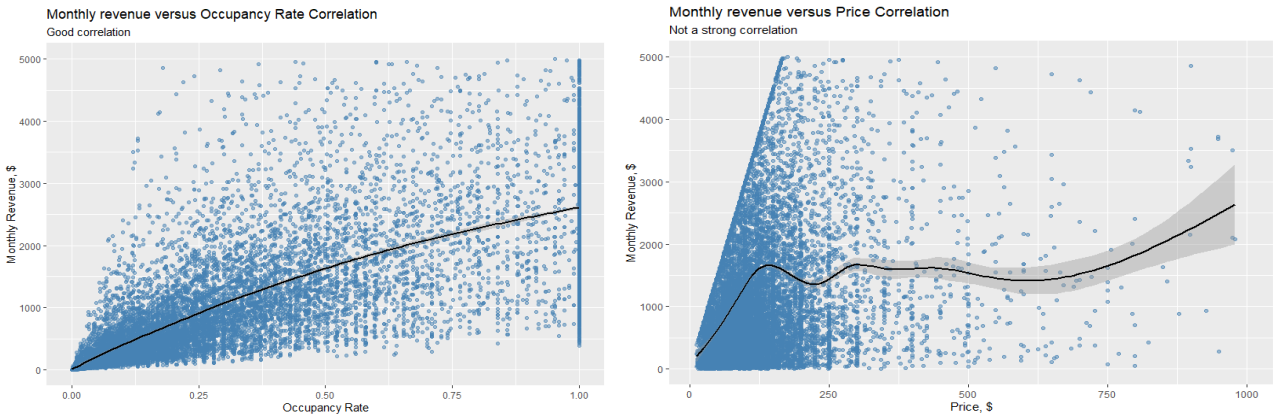


Figure 15. Revenue overview

We found that the highest monthly revenue for Airbnb hosts is in Toronto, followed by Vancouver and Victoria.

Superhosts

Superhosts are experienced hosts who provide an extraordinary experience for their guests. Superhosts have more visibility on the Airbnb listings, have higher earning potential and access to exclusive rewards.

We found that even though a superhost and a non-superhost, have the same listing price on average, superhosts have higher occupancy rates as well as higher monthly revenue for their listings across all seven cities. This can be seen in Figure 16.

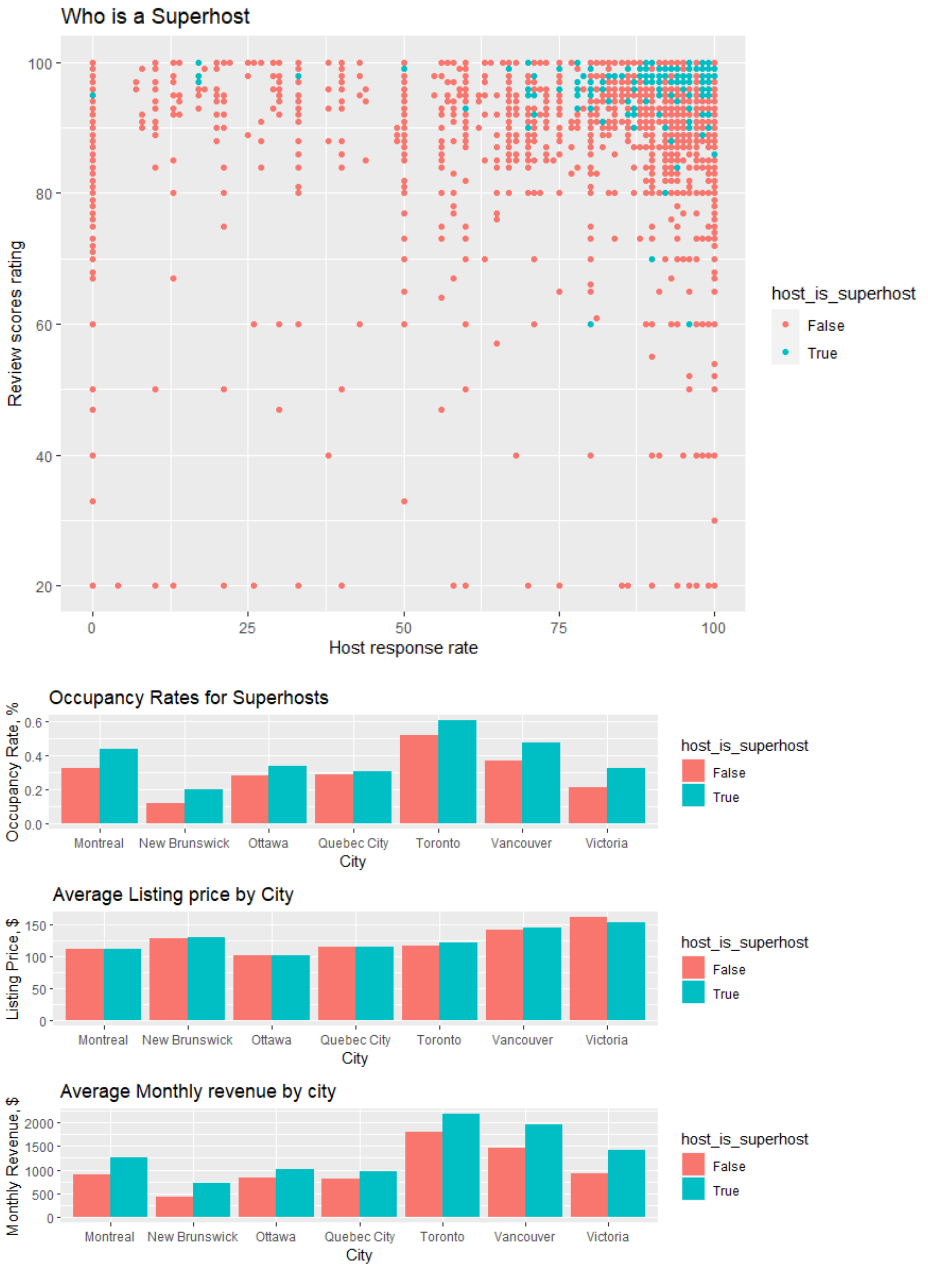


Figure 16. Superhost overview

We can see that superhosts generally maintain a response rate of more than 90% and a review score of more than 90. Though there are some outliers in the data, our observations are in line with Airbnb's

guidelines for being a superhost. Also, we can see that many Airbnb hosts lie in the high review score and high response rate region yet do not qualify for the superhost badge. Clearly, there are other factors involved in being a superhost, such as hosting more than 10 stays and maintaining a cancellation rate of less than 1%[3].

List or Rent

We gathered public data from CMHC website to see the average monthly rents across these seven Canadian cities. We compared them to the expected monthly revenue when a property is listed on Airbnb. Ultimately, we set out to answer the question, "to maximize my monthly revenue, should I rent my property for a long-term rental, or list it on Airbnb website?". We observed that if a host owns a property in Toronto, Vancouver, or Montreal, it would be wise to list it on Airbnb as it has the potential to provide more revenue than a long-term rental. For properties in New Brunswick, Quebec City, Ottawa, and Victoria, one would be better off by renting it out. Figure 17 shows the rental vs. Airbnb income for all seven cities.

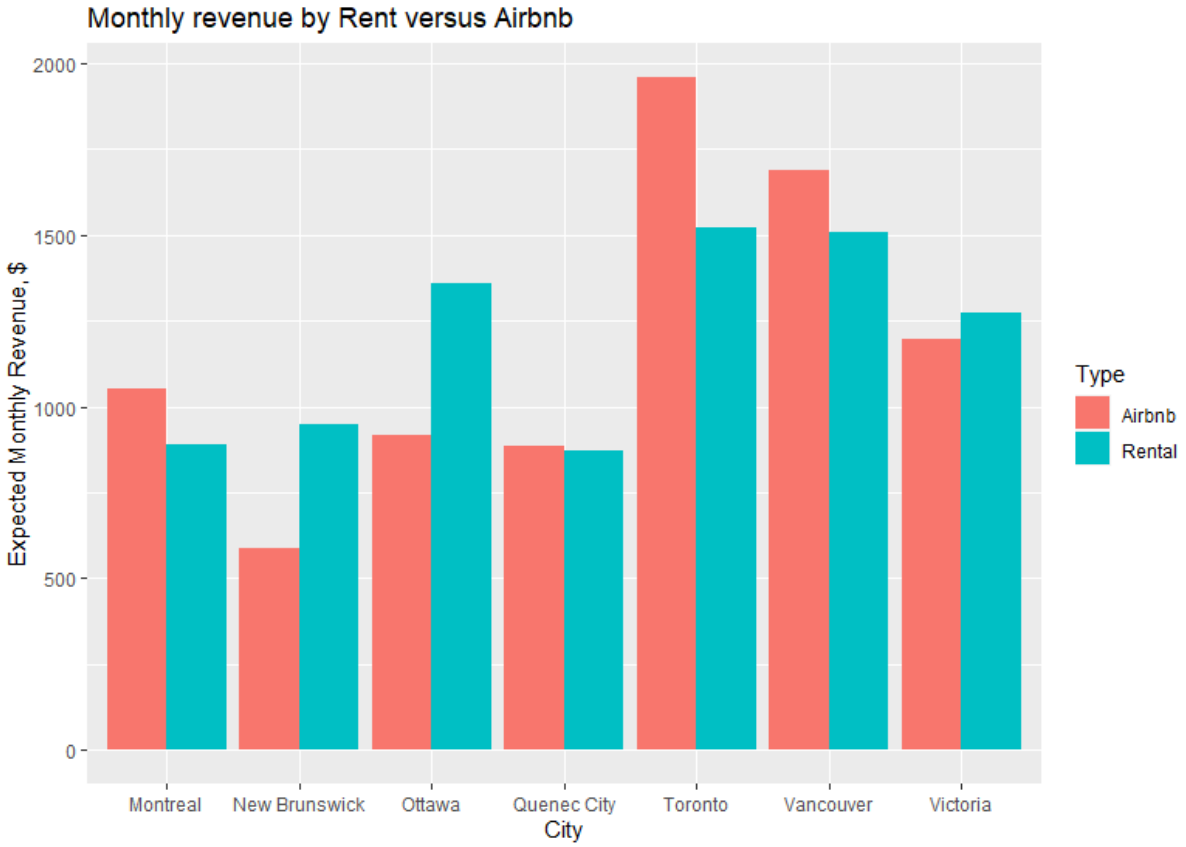


Figure 17. Monthly revenue - rent vs Airbnb

[3] Airbnb, (2021). Superhosts. https://www.Airbnb.ca/d/superhost

Section 3: Interactive Dashboard

Click to view dashboard

A series of eight dashboards were created in Tableau to help both guests and hosts observe and interact with the data. For visitors to one of these cities, it will help them to determine if a listing is fairly priced and aid in their decision to rent from the host. From the perspective of an investor who is seeking to purchase a rental property in one of the cities. An eighth dashboard shows all seven locations was also created.

Data sources for the dashboard were obtained from the InsideAirbnb website as .csv files. This data was cleaned and filtered using R with the modified .csv files being saved as eight separate files – one for each location and the last file containing all the data. The files were resaved as Excel files as it appears that Tableau works better with this file type. All the Excel files were then loaded into Tableau and linked to the file containing all the data with a many-to-many relationships joined by the "host id" field. This relationship is visualized in Figure 18.

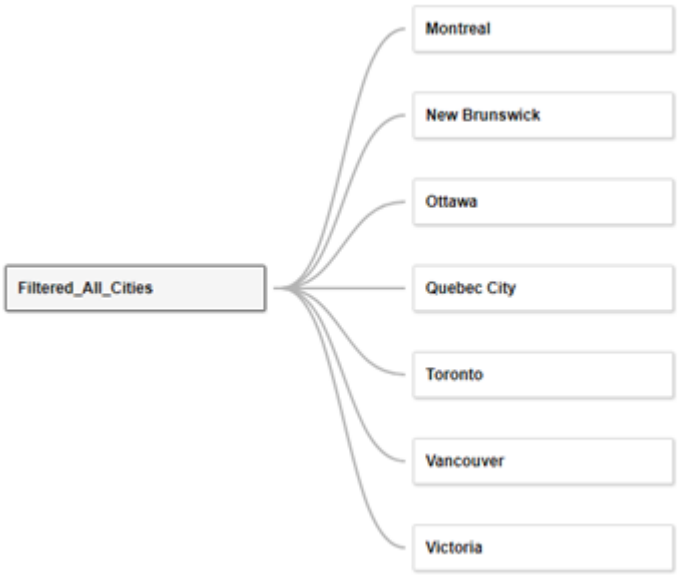


Figure 18. Tableau relationships

Each of the city dashboards contains six elements that displays the data gathered from the data sources.

1. A Key Performance Indicator (KPI) table showing:
    a. Average rental price for the entire city.
    b. The average number of beds for all the listings.
    c. The average number of days of availability for rental over the future thirty calendar days.
    d. The average minimum and maximum number of nights that can be rented.
2. Traditional rental information from the Canada Mortgage and Housing Corporation (CMHC) that shows the average vacancy and rental rates.
3. A map showing the distribution of all the Airbnb listing locations.

4. Airbnb rental properties are represented by a do and colour coded by neighbourhood. The size of each dot represents the average rental price – the higher the price, the larger the dot and the lower the price the smaller the dot.
5. A bar plot showing the average rental price for each neighbourhood.
6. A bar plot showing the availability for each neighbourhood over the next thirty days.
7. A plot showing the average rental price for properties with different numbers of bedrooms.

Except for the CMHC element, all of the other elements are enabled with filters that are linked with the other elements in the dashboard. These filters can be used to view more specific information. The Canada-wide dashboard replaces the neighbourhood data field with the city data field and the corresponding plots reflect this change. There is also no CMHC element for this dashboard as it was determined that this information was more valuable at the local level and not as important at the national level.

An example of these dashboards (for Victoria) is shown in Figure 19. Dashboards for the remaining six locations and for the national level can be found in the team Tableau file and on the web version of the dashboard here.

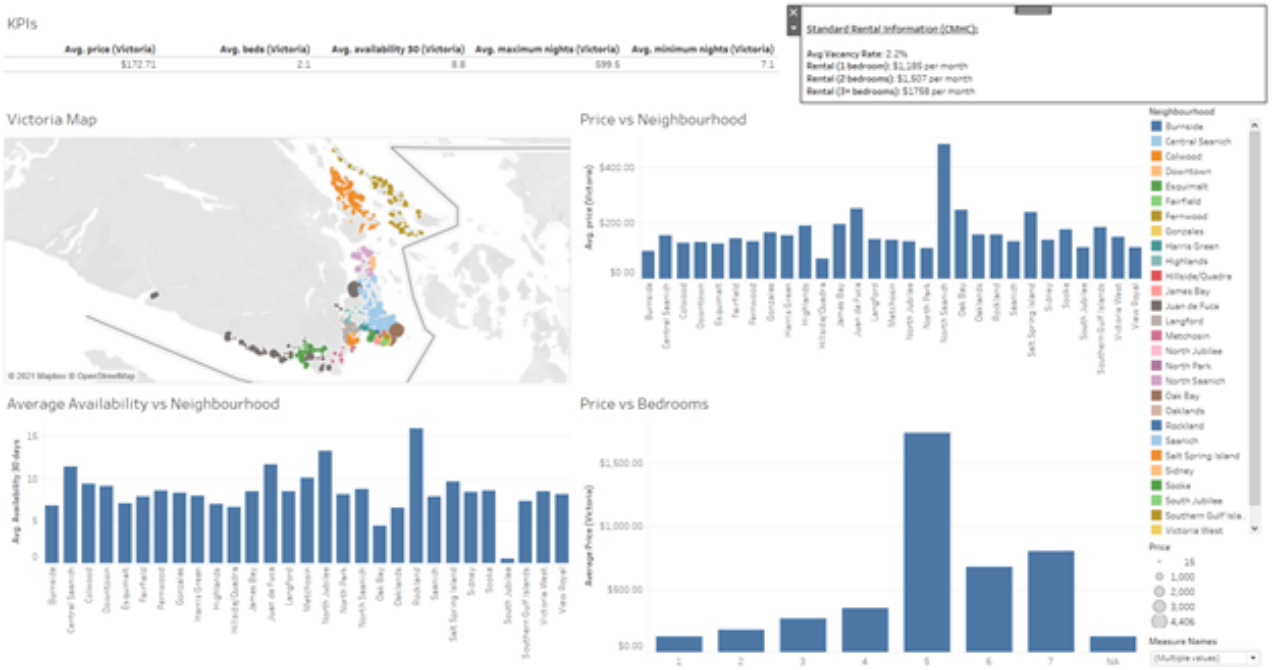


Figure 7. Tableau dashboard snapshot

Section 4: Hypothesis Testing

Model Development

After cleaning the data and conducting an exploratory data analysis, we built a linear regression to predict the price of a listing on a city-level. We created dummy variables for all seven cities, room type, and for instant bookability. These dummy variables gave us the flexibility to capture the incremental effects of each city and room type separately. We used random sampling to select 70% of the data to be used as a training set and the remaining 30% as testing set. Out of the many variables available in the data set, we used a TTT (test-test-test) approach and came up with the final list of variables to include in our model. Ultimately, we came up with a final model for price prediction as follows:

$$price = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \beta_5 X5 + \beta_6 X6 + \beta_7 X7 + \beta_8 X8 + \beta_9 X9 + \beta_{10} X10$$

Where;

Table 2. Price Prediction Model

Coefficients	Variables
$\beta_0 = -48.29$	Intercept term
$\beta_1 = 11.26$	X1 = No of people it can accommodate
$\beta_2 = 25.6$	X2 = Number of bedrooms
$\beta_3 = -0.55$	X3 = Number of reviews in the last 12 months
$\beta_4 = 0.57$	X4 = Review score rating (out of 100)
$\beta_5 = 27.62$	X5 = Number of bathrooms
$\beta_6 = -37.28$	X6 = '1' for private room, '0' otherwise
$\beta_7 = -74.65$	X7 = '1' for shared room, '0' otherwise
$\beta_8 = 42.03$	X8 = '1' if the listing is in Victoria, '0' otherwise
$\beta_9 = 38.49$	X9 = '1' if the listing is in Vancouver, '0' otherwise
$\beta_{10} = 25.94$	X10 = '1' if the listing is in Toronto, '0' otherwise

From the coefficients, we can see that the number of people a place can accommodate, number of bedrooms and bathrooms have an impact on the price of a listing. A listing also commands a premium price if the listing is in Victoria, Vancouver, or Toronto, compared to if it is in other cities.

Recognizing that correlation does not imply causation, to further illustrate the model, we ran a sample calculation with the following sample coefficient values:

Sample calculation coefficients:

X1 = property can accommodate 4 people

X2 = 1 bedroom

X3 = 12 reviews in last 12 months

X4 = review score rating of 91

X5 = 1 bathroom

X6 = 1 for private room

X7 = 0 for shared room

X8 = 1 for the listing being in Victoria

X9 = 0 for listing being in Vancouver

X10 = 0 for listing being in Toronto

Sample Calculation:

Predicted Price = -50.52 + (11.211)(4) + (26.78)(1) – (0.63)(12) + (0.63)(91) + (24.5)(1) – (37.3)(1) – (60.35)(0) + (42.94)(1) + (36.85)(0) + (26.04)(0)

Predicted Price = -50.52 + 44.844 + 26.78 – 7.56 + 57.33 + 24.5 – 37.3 + 0 + 42.94 + 0 + 0

Predicted Price = $101.01

Linear Regression

The regression results are presented below in Figure 20 and shows all statistically significant variables. The R-squared suggests that model is only able to explain approximately 42% of price variation. The significant F-statistic validates the usefulness of the model.

The regression plots for the error term below are generally good though the Normal Q-Q plot does not show normality of the error term for data above the 2nd quartile.

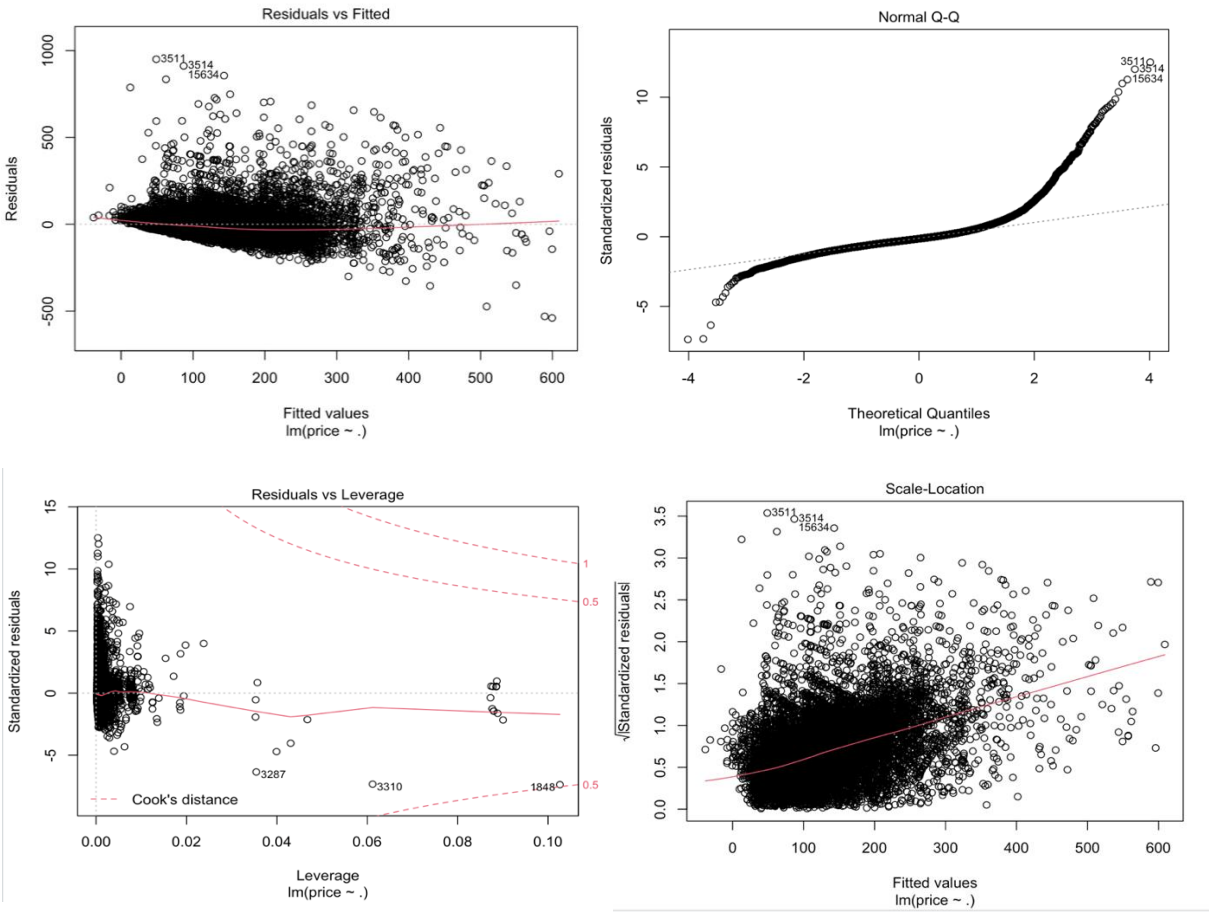


Figure 208. Regression plots

We also ran the *ncvTest* and it showed that the data is homoscedastic. The result is below.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 9739.095, Df = 1, p = < 2.22e-16
```

Prediction using Test Data

We used the linear regression model to predict the prices of listings for our testing set. After comparing the testing vs training set results, we believe that the model performs well. It is not overfitted to our data.

```
Residuals:
    Min      1Q  Median      3Q     Max
-483.14  -37.17  -11.66   19.40  906.25

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -48.29474    7.03912  -6.861 7.08e-12 ***
accommodates          11.26002    0.48876  23.038  < 2e-16 ***
bedrooms              25.60082    1.15338  22.196  < 2e-16 ***
number_of_reviews_ltm -0.54915    0.05288 -10.385  < 2e-16 ***
review_scores_rating   0.57324    0.07128   8.042 9.40e-16 ***
bathrooms             27.61717    1.18727  23.261  < 2e-16 ***
private_room         -37.27621    1.48595 -25.086  < 2e-16 ***
shared_room          -74.65131    7.04884 -10.591  < 2e-16 ***
Victoria              42.02683    2.22740  18.868  < 2e-16 ***
Vancouver             38.48740    1.92952  19.947  < 2e-16 ***
Toronto               25.93596    1.38898  18.673  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.43 on 16722 degrees of freedom
Multiple R-squared:  0.4119,    Adjusted R-squared:  0.4116
F-statistic:  1171 on 10 and 16722 DF,  p-value: < 2.2e-16
```

Figure 21. Linear regression output in R

The model passes all our quality control checks and it is the best model given the dataset. The positive prediction test result is displayed below.

```
         R2      RMSE       MAE
1 0.3821446 78.48699 47.52125
```

Model Limitation / Future Work

It should be noted that the model can successfully predict a listing price on a city-level. We understand that there are different neighbourhoods in each city with varying prices across these neighbourhoods. If such a level of granularity is needed, we need to build separate models for each city independently. This would take into consideration the listing by neighbourhood. The approach that we used to build our model can be extended to each city separately, however, that is beyond the scope of this project.

The model has a few limitations and provides direction for future work. These are:

1. The data not a time series but instead a snapshot in time. It does not capture any seasonality or price fluctuations.
2. The model does not capture the purchase cost or expenses associated with the property. The model only illustrates the revenue portion and not provide insight into the profitability of the property. We understand that listing on Airbnb involves services fees and other miscellaneous expenses that are not built into this model.
3. There are some qualitative factors that have not been captured in this model that can impact the pricing. For example, if the listing is unusually unique or if the host provides extra amenities such

as welcome packs etc. The model also does not capture the pricing for peak demand periods such as sports events or festivals.

4. The numbers of observations of data from Toronto is significantly higher than the rest of the cities. This may result in a data bias in our model. Future refinements to the model should address this issue.
5. The QQ plot tells us that the data is over dispersed where the data deviates from the straight line after 2nd quartiles. This suggests that the model is unable to predict the large values and cannot be trusted for extremely large values in the dataset.

Conclusions

The analysis presented here is introductory, high-level overview of Airbnb properties listed across seven cities in Canada. We did significant data cleaning and exploratory analysis to gain meaningful insights from the data. We built a linear regression model to determine the price of a listing based on the different parameters available. We came up with an occupancy model to discover how occupancy rates and monthly revenues vary across the cities. Ultimately, we provided a useful guide to answer the question: "Should we list a property on Airbnb or rent it comparing potential revenue?". Finally, we created a series of eight dashboards in Tableau to help a prospective investor or a guest to observe and interact with the data.

Though the model can be further refined and improved (as discussed in the limitations section of this report), we believe it provides a sound starting point to understand the pricing of Airbnb listings and the revenue associated with it.

Appendix A: Data Dictionary

Field	Data Type	Calculated (Yes/No)	Description
id	integer	No	Airbnb's unique identifier for the listing
name	text	No	Name of the listing
description	text	No	Detailed description of the listing
neighborhood_overview	text	No	Host's description of the neighbourhood
host_id	integer	No	Airbnb's unique identifier for the host/user
host_name	text	No	Name of the host. Usually just the first name(s).
host_since	date	No	The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest.
host_about	text	No	Description about the host
host_response_time	string	No	Text describing the response rate of the host.
host_response_rate	numeric (float)	No	A percentage value that shows how often the host responds.
host_acceptance_rate	numeric (float)	No	That rate in percentage at which a host accepts booking requests.
host_is_superhost	boolean [t=true; f=false]	No	A true / false flag indicating if the host is a superhost or not.
host_neighbourhood	text	No	The neighbourhood that the host lives in.
host_listings_count	text	No	The number of listings the host has (per Airbnb calculations)
host_has_profile_pic	boolean [t=true; f=false]	No	Indicates if the host has a profile picture or not.
host_identity_verified	boolean [t=true; f=false]	No	Indicates if the host identify has been verified.
neighbourhood_cleansed	text	Yes	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.

latitude	numeric (float)	No	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
longitude	numeric (float)	No	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
property_type	text	No	Self-selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field
room_type	text	No	[Entire home/apt|Private room|Shared room|Hotel] All homes are grouped into the following three room types: Entire place Private room Shared room Entire place Entire places are best if you're seeking a home away from home. With an entire place, you'll have the whole space to yourself. This usually includes a bedroom, a bathroom, a kitchen, and a separate, dedicated entrance. Hosts should note in the description if they'll be on the property or not (ex: "Host occupies first floor of the home") and provide further details on the listing. Private rooms Private rooms are great for when you prefer a little privacy, and still value a local connection. When you book a private room, you'll have your own private room for sleeping and may share some spaces with others. You might need to walk through indoor spaces that another host or guest may occupy to get to your room. Shared rooms Shared rooms are for when you don't mind sharing a space with others. When you book a shared room, you'll be sleeping in a space that is shared with others and share the entire space

			with other people. Shared rooms are popular among flexible travelers looking for new friends and budget-friendly stays.
accommodates	integer	No	The maximum capacity of the listing
bathrooms_text	string	No	The number of bathrooms in the listing. On the Airbnb web-site, the bathrooms field has evolved from a number to a textual description. For older scrapes, bathrooms is used.
bedrooms	integer	No	The number of bedrooms
beds	integer	No	The number of bed(s)
amenities	json	No	
price	currency	No	daily price in local currency
minimum_nights	integer	No	minimum number of night stay for the listing (calendar rules may be different)
maximum_nights	integer	No	maximum number of night stay for the listing (calendar rules may be different)
minimum_nights_avg_ntm	numeric (float)	Yes	the average minimum_night value from the calender (looking 365 nights in the future)
maximum_nights_avg_ntm	numeric (float)	Yes	the average maximum_night value from the calender (looking 365 nights in the future)
has_availability	boolean	Yes	[t=true; f=false]
availability_30	integer	Yes	avaliability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
number_of_reviews	integer	No	The number of reviews the listing has
number_of_reviews_ltm	integer	Yes	The number of reviews the listing has (in the last 12 months)
number_of_reviews_l30d	integer	Yes	The number of reviews the listing has (in the last 30 days)
first_review	date	Yes	The date of the first/oldest review
last_review	date	Yes	The date of the last/newest review
review_scores_rating	integer	No	The overall review score of the listing represented by a value of 0 to 100 with 0 being the lowest score and 100 being the highest score.

instant_bookable	boolean	No	[t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing.
reviews_per_month	numeric (float)	Yes	The number of reviews the listing has over the lifetime of the listing
City	string	No	The city associated with the listsing. There are seven used in the data: Victoria, Vancouver, Toronto, Ottawa, Montreal, Quebec City and New Brunswick (data for entire province).

Note:

Data dictionary has been modified from the version found on insideAirbnb.[4]

[4] Inside Airbnb. (2020, August). Get the Data - Data Dictionary. Retrieved May 22, 2021, from Inside Airbnb: http://insideairbnb.com/get-the-data.html