Airbnb predictive pricing tool for tourists coming to Canada

Introduction

According to Statistics Canada, a recording breaking **22.1 million international tourists from abroad visited Canada in 2019.** Hotels have always been the mainstay for accommodations but the prices per night can be unaffordable for visitors looking to stay long-term for tourism or work. Airbnb was founded in 2008 and has since been proven to be a successful online platform to match hosts with unused space with international or local guests looking for an affordable place to lodge. Although it is often more affordable than hotels, **Airbnb does not have a direct effect on the listing prices and ultimately leaves the hosts to decide the listing prices**. In this analysis, we want to investigate which factors, ranging from if the host is a superhost or not to the number of bathrooms, are most likely influencing the price of Airbnb listings for cities in Canada. This predicitive tool may potentially help travellers better understand the reasoning behind the listed price of certain Canadian Airbnb listings.

Research Question

In this analysis, we aim to investigate the influence of various factors on the price of Airbnb listings across various Canadian cities to see which ones are most likely to impact the listed price. Which factors, ranging from property type to number of bedrooms, are most likely playing a role in determining the price of Canadian Airbnb listings?

Data Description

The Data folder contains raw Airbnb listings data for Montreal, New Brunswick, Ottawa, Quebec, Toronto, Vancouver, and Victoria. The datasets were obtained from the Inside Airbnb project, conceived and compiled by Murray Cox and John Morrix in 2019. Each row represents a single listing with detailed information such as location, price, and rating score. The cleaned dataset can be accessed here.

Variable	Type	Description
host_is_superhost	String	whether the host is a super host (TRUE or FALSE).
city	String	City of the listing belongs to. One exception: New Brunswick is a Province.
property_type	String	Property type of the listing.
room_type	String	Room type: Entire Room, Hotel room, Private room, or Shared room.
accommodates	Int	The number of people that can be accommodated in the unit.
bathrooms	Int	The number of bathroom in the unit.
bedrooms	Int	The number of bedrooms in the unit.
beds	Int	The number of beds in the unit.
cancellation_policy	String	Strictness of the cancellation policy.
price	Int	Price per night.

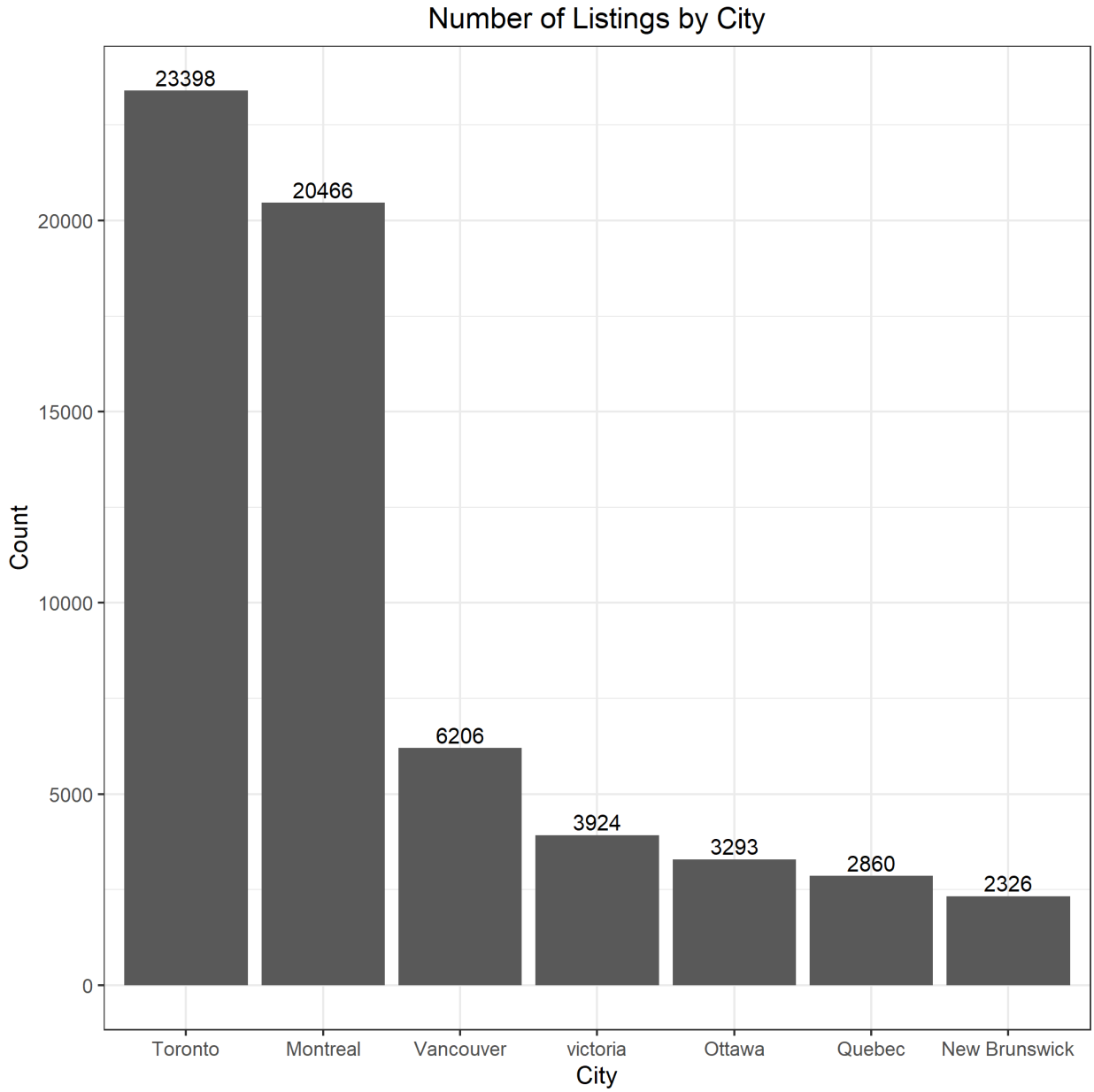
Exploratory Data Analysis

In this section, four plots are produced to give an insight into how does the dataset distribute. First, we read in the dataset.

```
data <- readr::read_csv(here("Data", "cleaned_data.csv"))
```

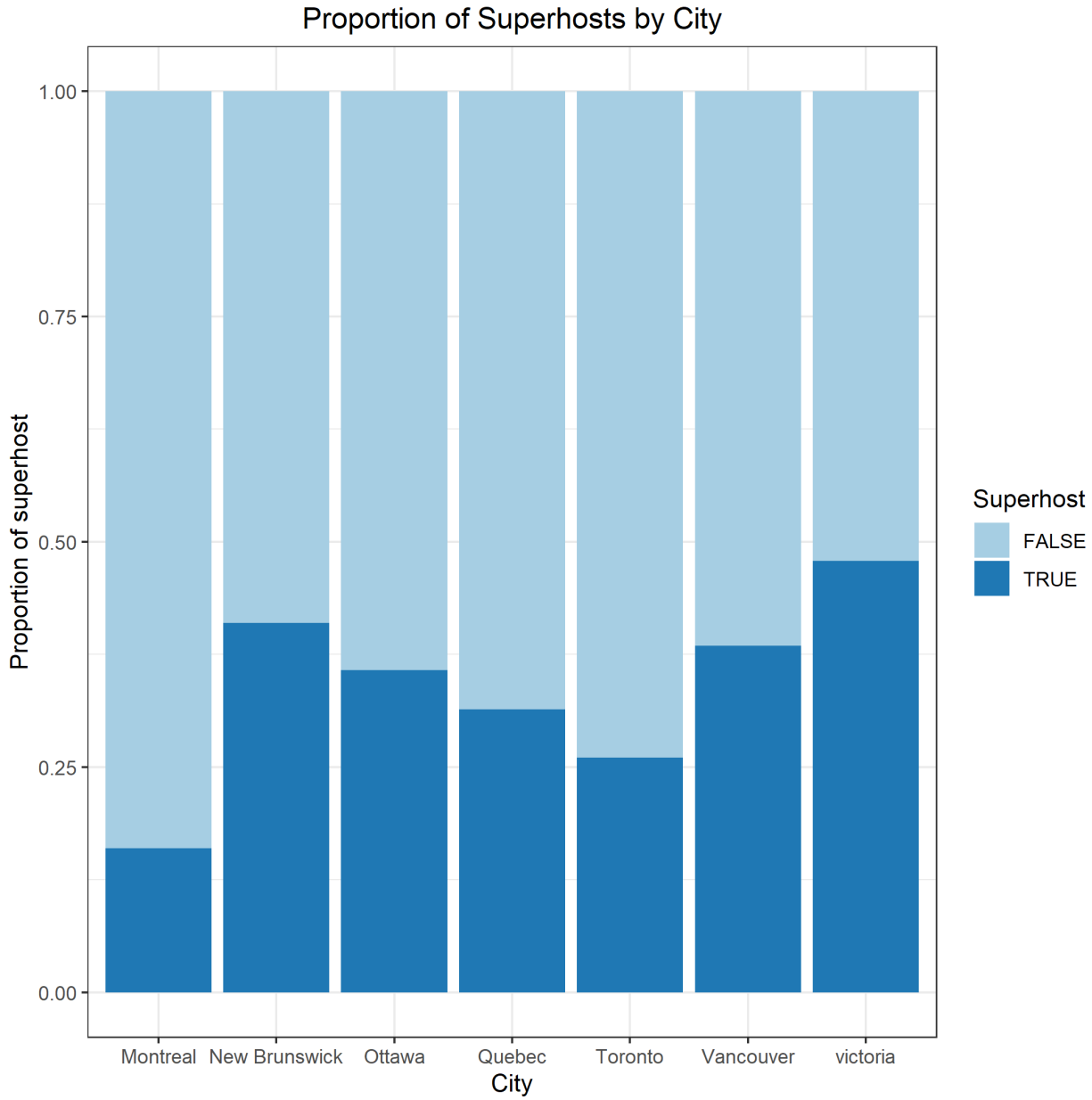
1. What is the number of Airbnb listings in different Canadian cities?

The barplot below shows the number of listings in different cities. From the plot, we can see that Toronto has the most number of listings (23398) and New Brunswick has the least number of listings (2326).



2. How many Airbnb superhosts are there in different Canadian cities?

The proportional bar chart below shows the percentage of superhosts in different cities. From the plot, Victoria seems to have the largest percentage of superhosts (48.0886850152905%), while Montreal seems to have the smallest percentage of superhosts (19.3784813837584%).

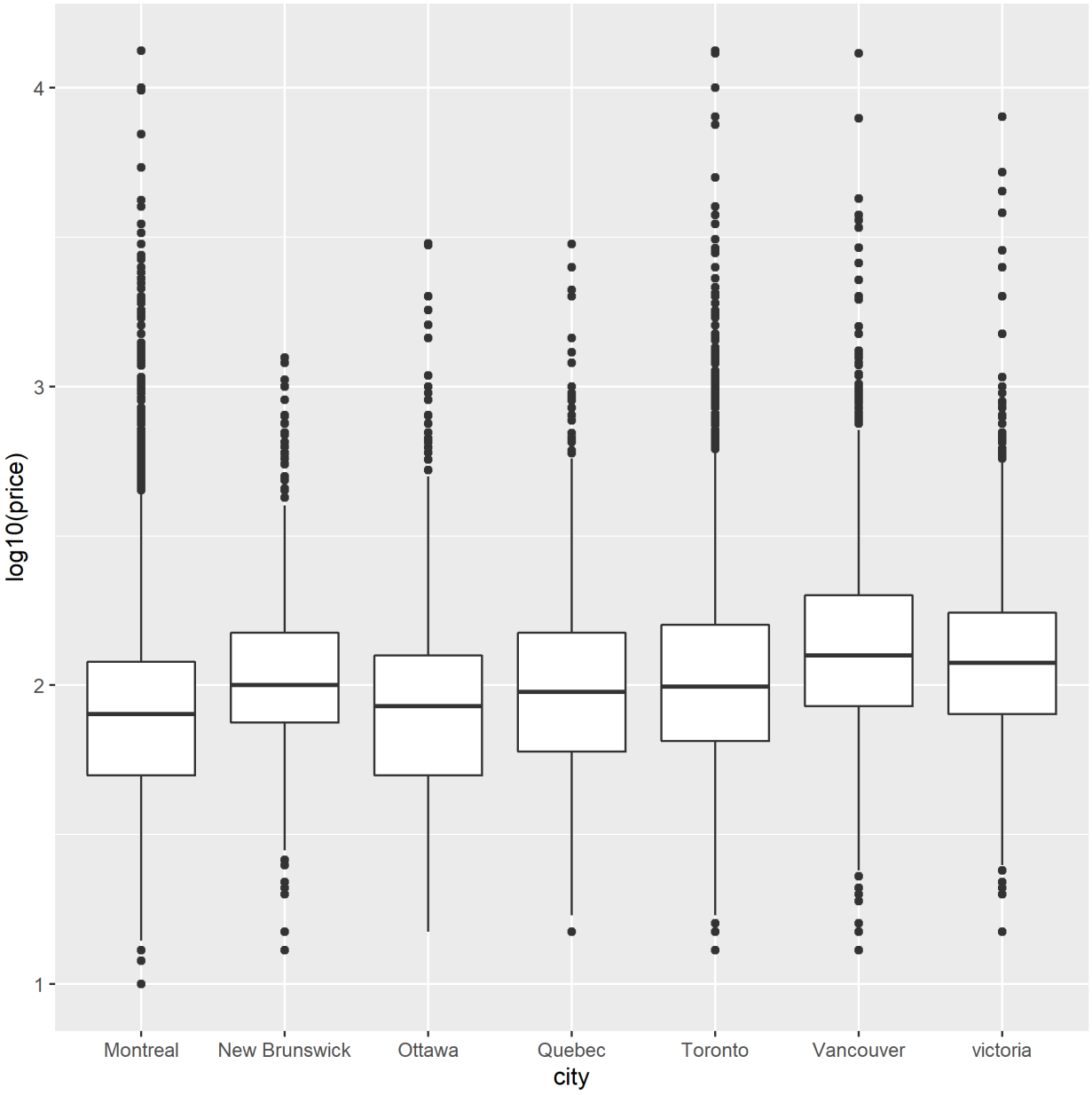


3. Is there a relationship between the number of accommodates and other features of the listing?

From the correllogram below, there is a strong relationship between the number of accommodates and the number of beds in the unit. This may cause a problem when performing linear regression analysis because some predictors are collinear. To solve the collinearity problem, we may decide to not use all of the variables (accommodates, bathrooms, bedrooms, and beds) as predictors in the linear regression model.

4. What is the distribution of the price per night in different Canadian cities?

The side-by-side boxplots shows the price per night (after log10 transformation) distribution in different cities. From the plots, we can see that there are some extremely high prices in the dataset. Further analysis will be required to figure out the reason for the extreme prices. Otherwise, we may need to consider them as outliers.



Analysis Methods

In this section, we perform some analysis to figure out which factors are most likely influencing the price of Airbnb listings for cities in Canada by following 3 steps. In the first step, we remove all the outliers. In the second step, fit a linear regression model. In the last step, do model diagnostics to check the assumptions and the model performance.

Step 1: Remove Outliers

We identify an observation as an outlier if it falls below `Q1-1.5*IQR` or above `Q3+1.5*IQR`, where Q1 and Q3 means the first and third quantile of all the observations in the corresponding city; IQR means interquartile range (`Q3-Q1`).

Step 2: Fit a linear regression model

Firstly, we build a full model for `price` by including all the potential predictors mentioned in the Data Description section. Then, do variable selection. Since `property type` has too many levels (44 levels), it is not appropriate to include in the linear regression model. Also, `bed` is removed because it is highly correlated with `accommodates`, `bathrooms`, and `bedrooms`.

Step 3: Model Diagnostics

Four plots are produced for model checking. The first plot is residual vs. fitted values plot. It is uswful for checking the assumption of linearity. The second plot is QQ plot, which is used to check the normality assumption of the residuals. The third plot is scale-location plot, which is useful for checking the assumption of homoscedasticity. The fourth plot is Cook's distance plot, which shows the measure of the influence of each observation on the regression coefficients.

Results

Let's look at the results of the linear regression model.

```
lm <- readRDS(here::here("RDS","step_lm.RDS"))
knitr::kable(tidy(lm))
```

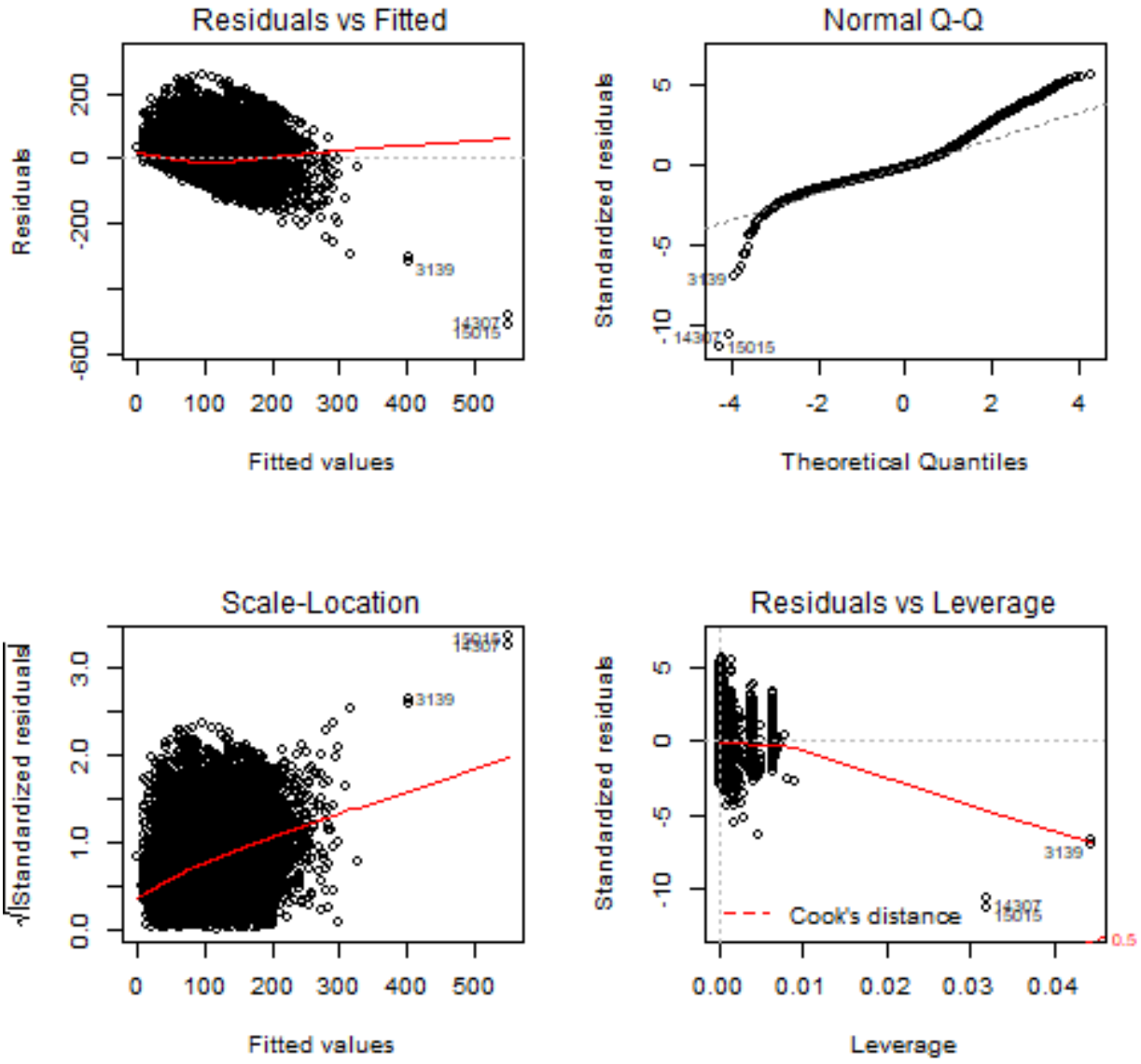
term	estimate	std.error	statistic	p.value
(Intercept)	49.389185	0.6988047	70.676663	0.0000000
host_is_superhostTRUE	2.603737	0.4339549	6.000017	0.0000000
cityNew Brunswick	14.695022	1.0447213	14.065973	0.0000000
cityOttawa	8.380514	0.9020077	9.290956	0.0000000
cityQuebec	11.040100	0.9611550	11.486284	0.0000000
cityToronto	32.873827	0.4621202	71.136957	0.0000000
cityVancouver	50.329410	0.7032114	71.570815	0.0000000
cityvictoria	33.765144	0.8410008	40.148764	0.0000000
room_typeHotel room	8.141642	2.8297850	2.877124	0.0040146
room_typePrivate room	-45.298859	0.4728658	-95.796437	0.0000000
room_typeShared room	-64.408283	1.7562995	-36.672722	0.0000000
accommodates	6.997359	0.1526387	45.842622	0.0000000
bathrooms	10.315506	0.4515773	22.843279	0.0000000
bedrooms	8.956279	0.3183921	28.129718	0.0000000
cancellation_policymoderate	-2.397571	0.4991631	-4.803182	0.0000016
cancellation_policystrict	4.076218	0.4789944	8.509951	0.0000000
cancellation_policysuper_strict	33.610818	3.6772103	9.140303	0.0000000

- The coefficient of `superhost` is positive means that if the host is a superhost, the Airbnb price is higher.
- Montreal is baseline `city`. The Airbnb price in Montreal is the lowest compared to other cities, while the Airbnb price in Vancouver is the highest. The price in Vancouver is $50 higher than Montreal in average by controlling other factors.
- For `room type`, apartment is the baseline. If the room type is hotel room, the price is higher; if the room type is private room, the price is lower; if the room type is shared room, the price is even lower for about $64.
- The coefficient of `accommodates` is 7 means that for each extra accommodate, the Airbnb price increases

by $7 in average.

- The coefficient of `bathroom` is 10 means that for each extra bathroom, the Airbnb price increases by $10 in average.
- The coefficient of `bedroom` is 9 means that for each extra bathroom, the Airbnb price increases by $9 in average.
- For `cancellation policy`, flexible cancellation is the baseline. Airbnb with moderate cancellation policy has lower price while the Airbnb with strict or super strict cancellation policy has higher price.

Then take a look at the model diagnostic:



There is an increasing trend in scale-location plot, so the assumption of homoscedasticity may not hold. From Cook's distance plot, there are several influential points, which need to remove in the future analysis.

Conclusions/Discussion

Superhost, city, room type, number of accommodates, number of bathrooms, number of bedrooms, and cancellation policy are all have significance effects on the Airbnb price. However, the model seems not fit very well based on model diagnostics. A possible reason is that there are too many categorical variables included in the linear regression model. For future analysis, we can remove the influential points and perform ANOVA instead of linear regression.

References

Statistics Canada. (2020, February 21). Travel between Canada and other countries, December 2019. Retrieved from https://www150.statcan.gc.ca/n1/daily-quotidien/200221/dq200221b-eng.htm?indid=3635-2&indgeo=0