Northeastern University
College of Professional Studies

Airbnb Data Analysis in NYC, NY, USA (2011 - 2019)

Capstone Presentation

Name: Abhilash Kumar Dikshit

Student ID: 2702209

Course: MPS Analytics- ALY 6010

Campus: Vancouver, Canada

Airbnb New York Analysis (2011 - 2019)

Airbnb New York Analysis (2011 - 2019)
This dataset contains information about Airbnb host listings, neighbourhoods, price, minimum nights, review counts/rate and the availability throughout the year.

Limitations:
No data on customer stay duration, timeline, review score for each listing and no. of tourist attractions nearby.

Acknowledgements:
Dataset taken from Kaggle. This is a **public dataset of Airbnb**, and the original source can be found in their website.

Exploratory Data Analysis:

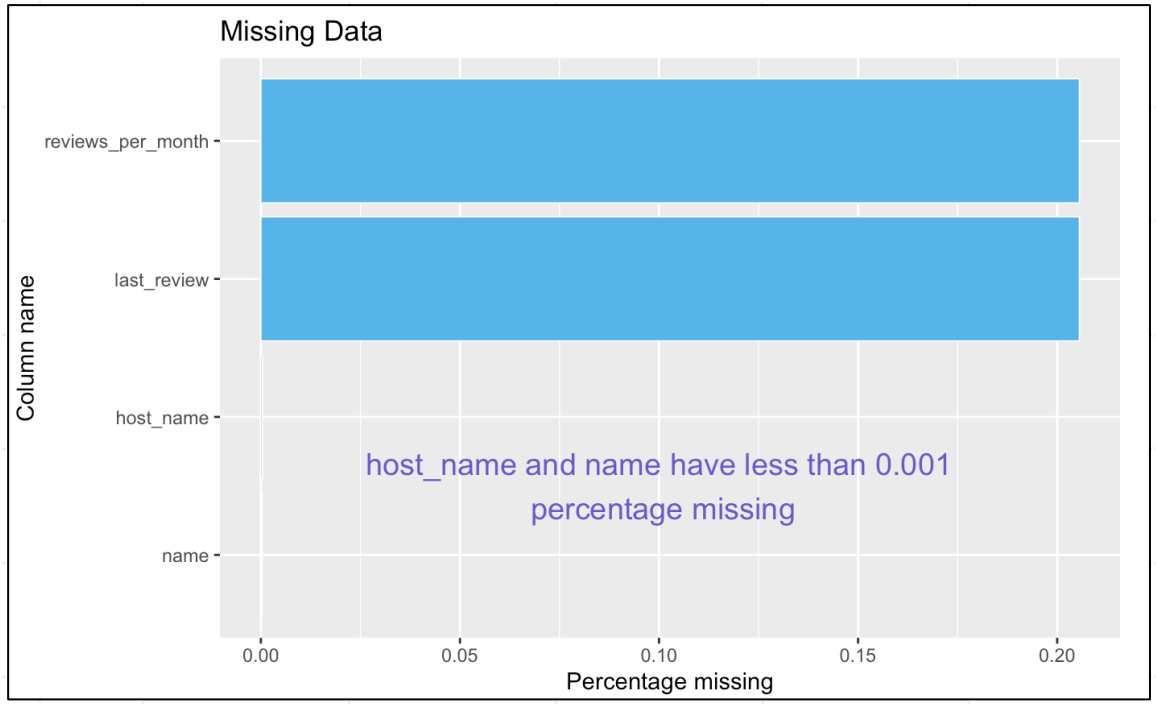
Data cleanup:

- Columns "id" and "host_id" is omitted.
- Column "last_review" has to be converted to Airbnb type using function ymd from lubridate package.
- Identified missing and NA values in the data.
- Respective columns "id" and "host_id" is omitted since they don't carry any useful information and hence wont' be used in predictive models.
- Column "last_review" has to be converted to Airbnb type using function ymd from lubridate package.

Exploratory Data Analysis

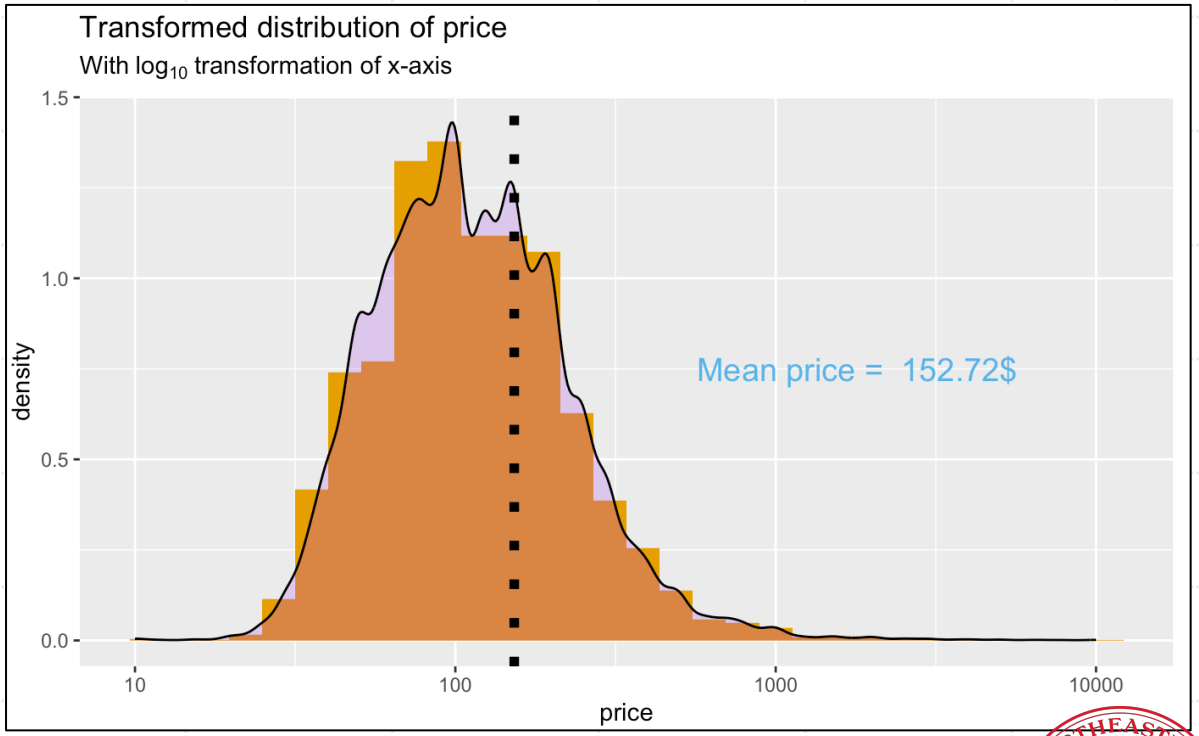
Data Storytelling:

- Data storytelling using **descriptive statistics and visualizations**



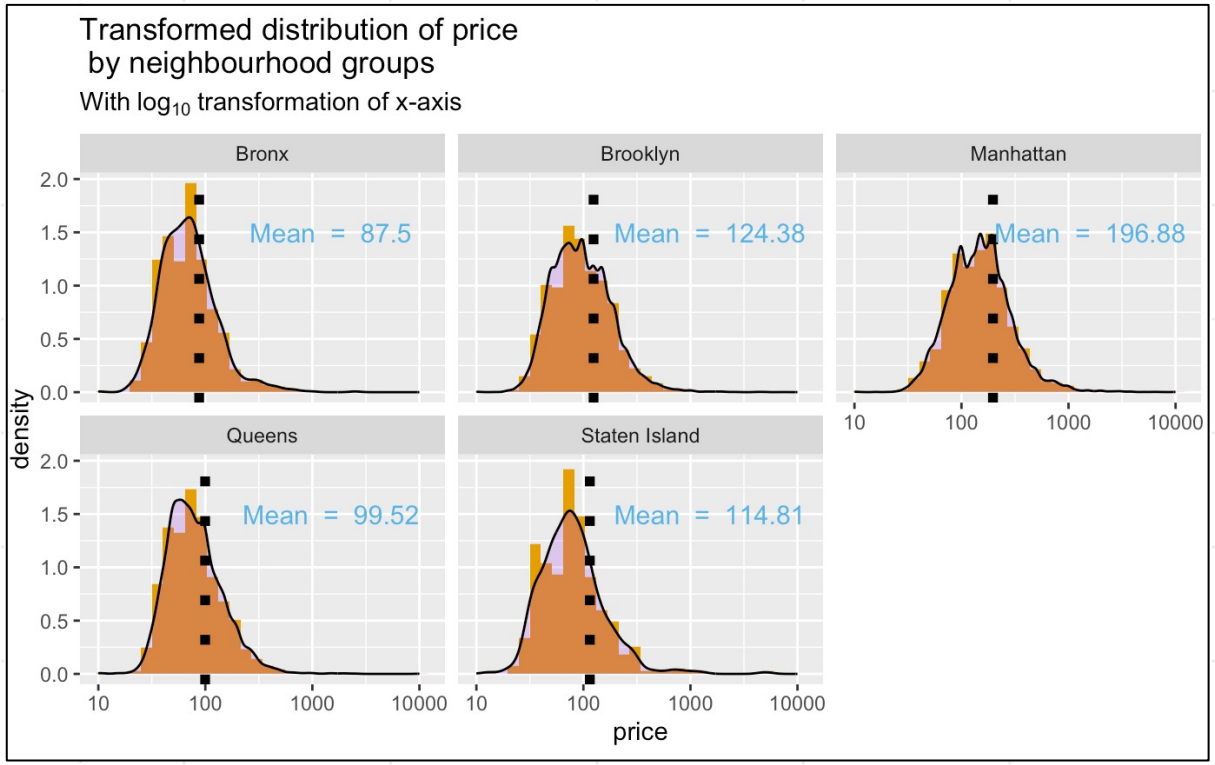
Histogram & Density with log10 transformation for price:

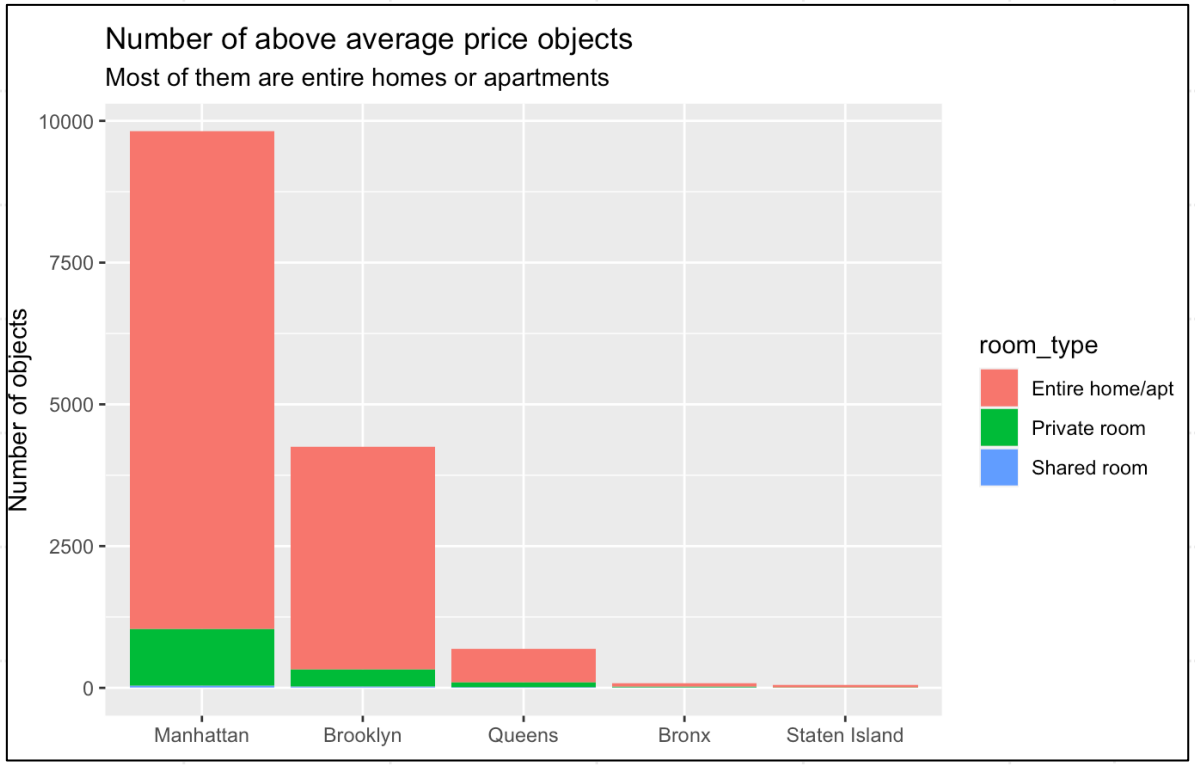
Original distribution is very skewed, logarithmic transformation can be used to gain better insight into data.



Exploratory Data Analysis

Histogram & Density with log10 transformation for Above Average Price Objects by Neighbourhood Areas: neighbourhood groups:

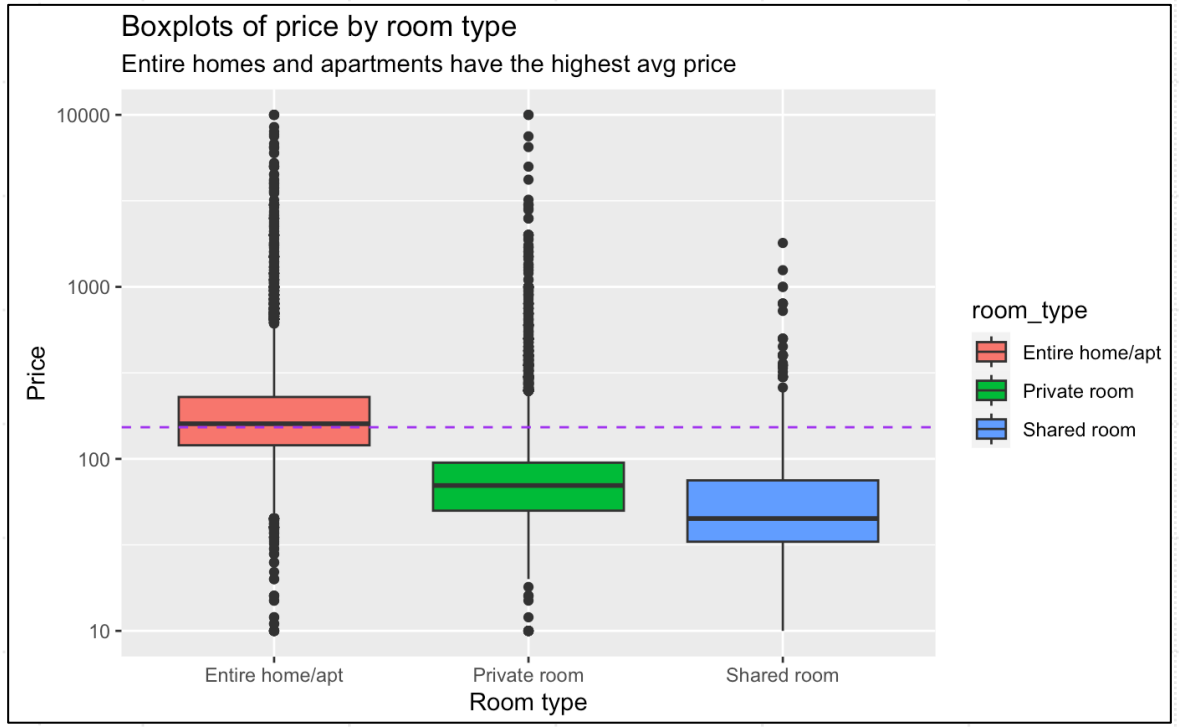




Exploratory Data Analysis

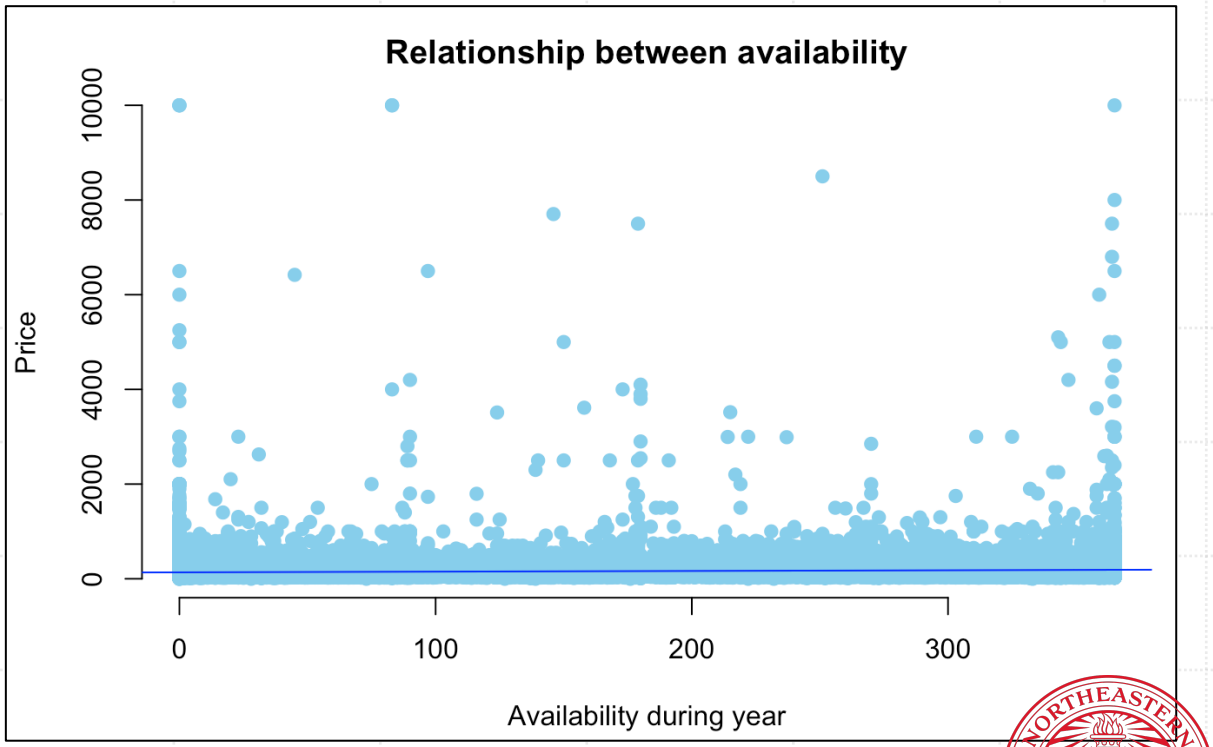
Boxplot of price by room type:

As expected, entire home or apartment type has the highest average price. It was also expected that shared rooms would have lower price than private rooms.



Scatter Plot for Price and Availability:

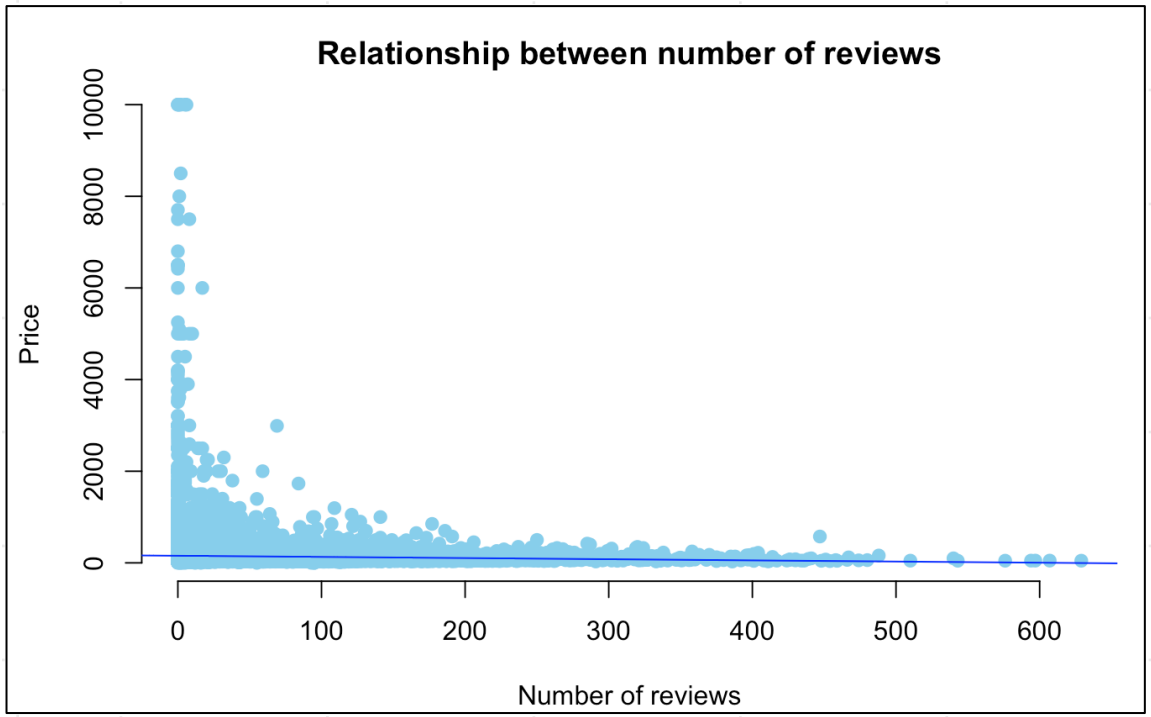
It's hard to see clear pattern, but there's a lot of expensive objects with few available days and many available days.



Exploratory Data Analysis

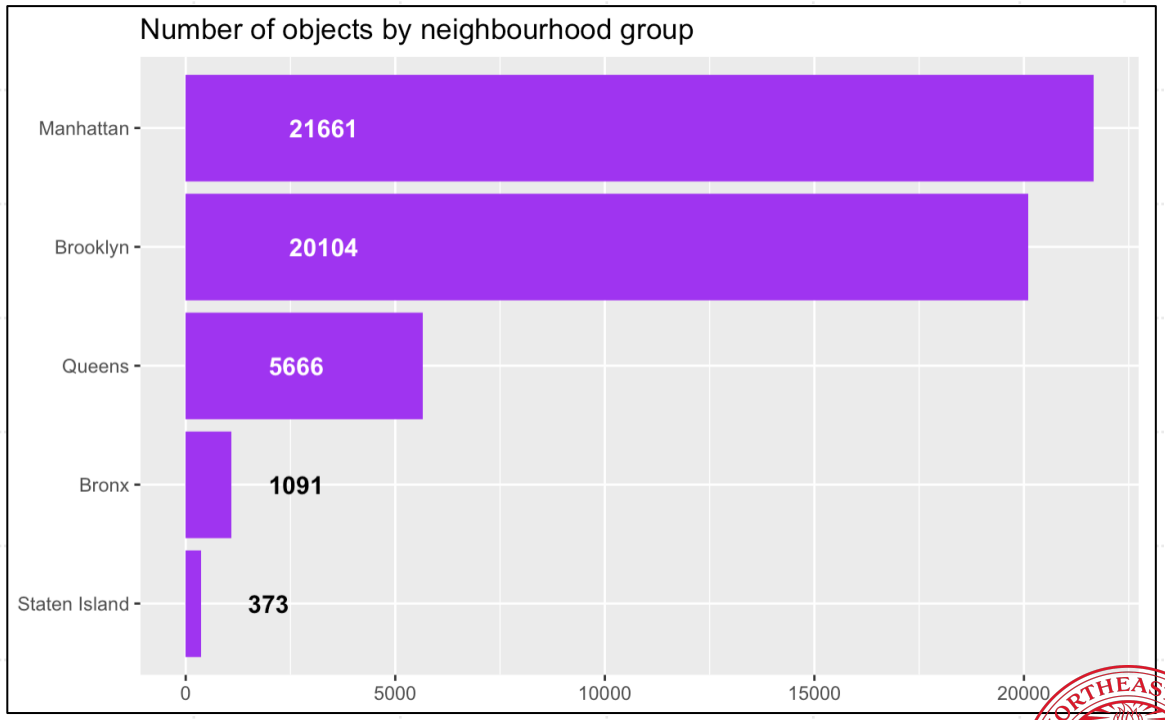
Scatter Plot for Price and Number of Reviews:

As expected, entire home or apartment type has the highest average price. It was also expected that shared rooms would have lower price than private rooms.



Number of objects by neighbourhood areas:

Manhattan has the highest number of objects while it's the smallest neighbourhood group by area. That can be explained by the fact that it's the most popular neighbourhood group with biggest GDP.



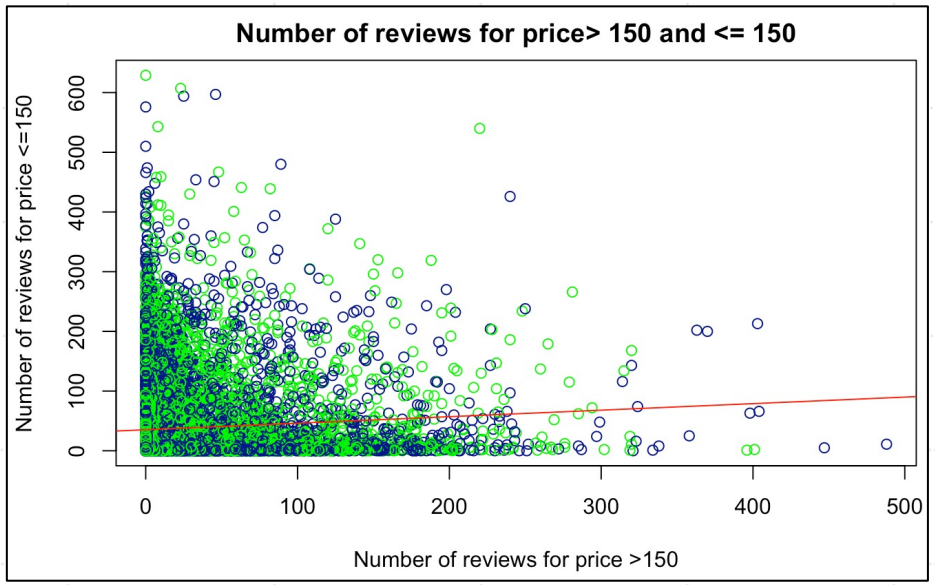
Hypothesis Testing

Does Airbnb prices affects the number of reviews?
Stating Null and Alternative hypothesis using two sample tests:

- *H0(claim)= Airbnb's with higher prices have higher reviews*
- *H1= Airbnb's with higher prices have lower reviews*

Results:

- We can easily interpret that p-value is less than 0.05 in our case, hence we can reject the null hypothesis.
- We can evidently conclude by the p-value that the true means are significantly different with the p-value.

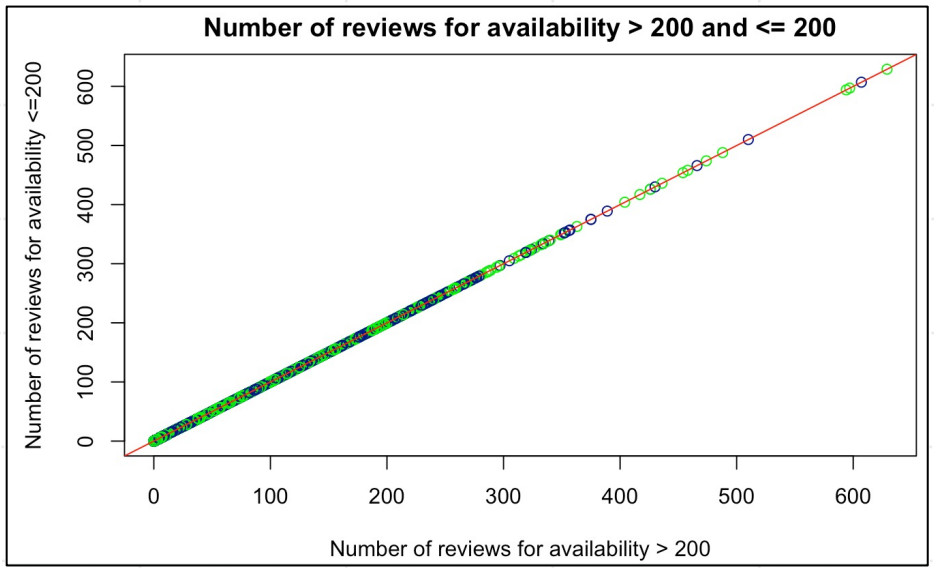


Does Airbnb availability for more than 200 days affects the number of reviews?

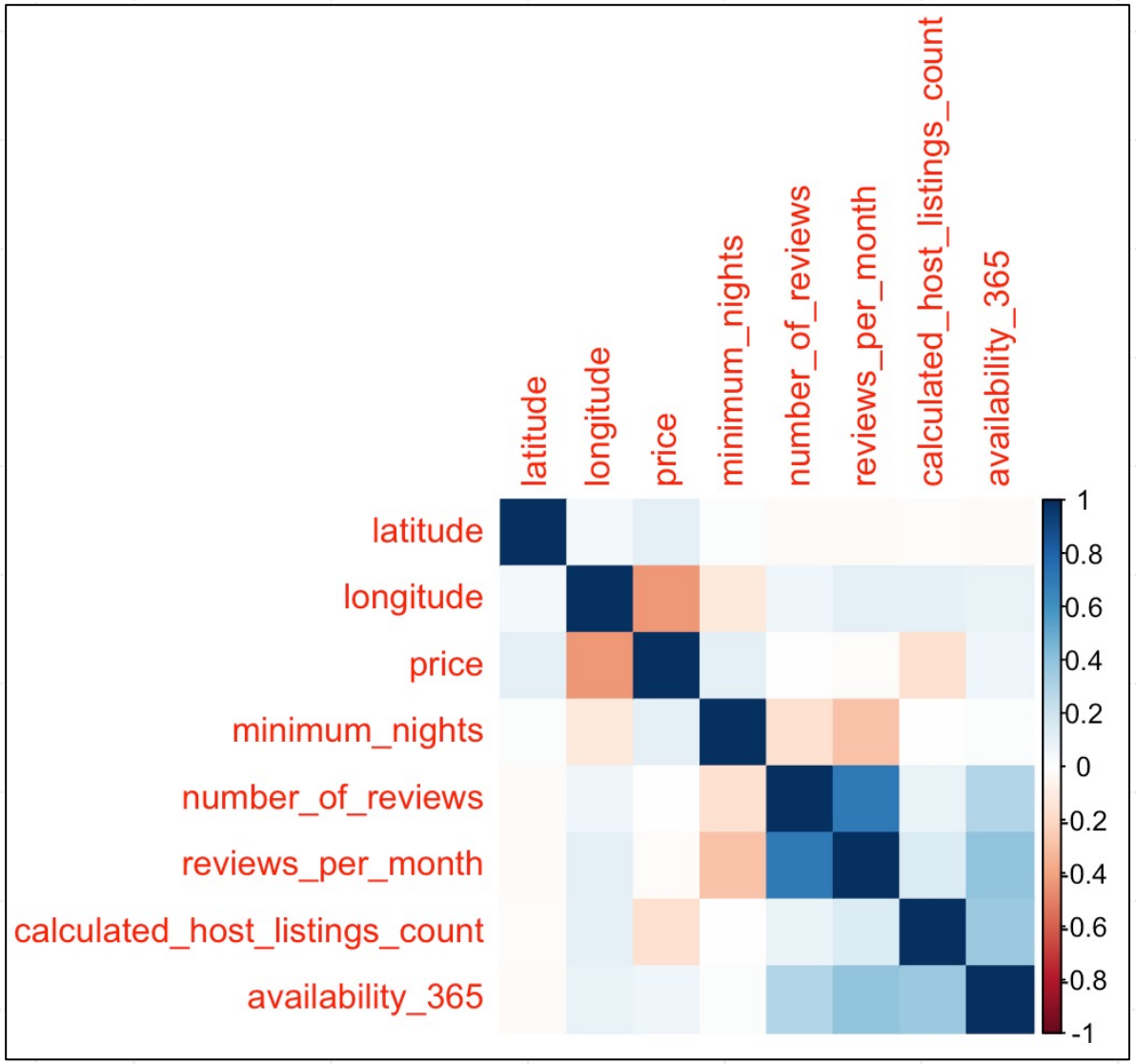
- *H0(claim)= Airbnb's with 365 days availability have higher reviews*
- *H1= Airbnb's with 365 days availability have lower reviews*

Results:

- We can easily interpret that p-value is less than 0.05 in our case, hence we can reject the null hypothesis.
- We can evidently conclude by the p-value that the true means are significantly different with the p-value.



Correlation matrix of Airbnb using Spearman correlation



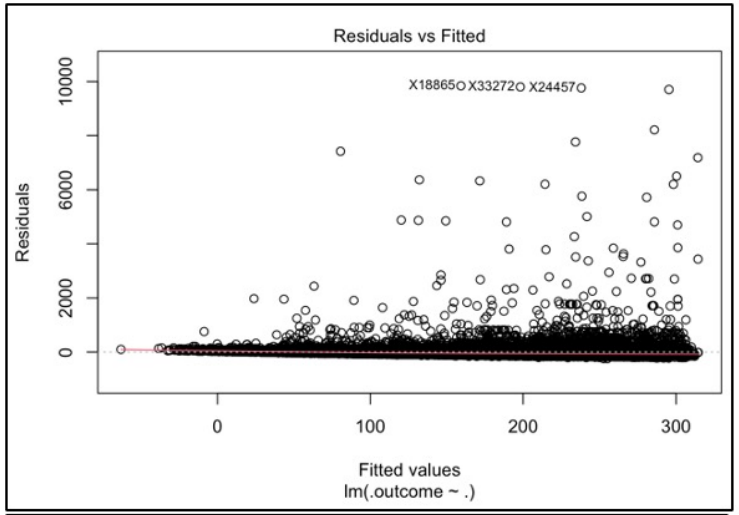
We used the "rcorr()" function from the "Hmisc" package in R to create a correlation matrix that shows the correlation coefficients between each variable in our data frame. The first matrix shows the correlation coefficients between the variables and the second matrix shows the corresponding p-values.

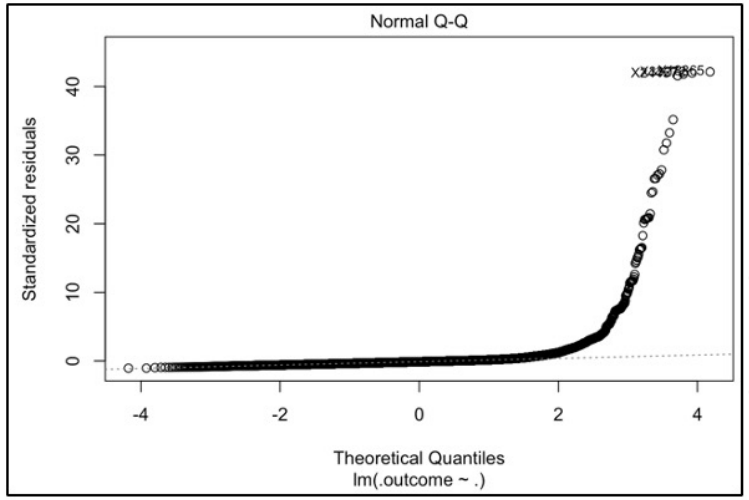
The correlation coefficient between price and number of reviews is -0.04 and the p-value for this correlation coefficient is 0.0000.

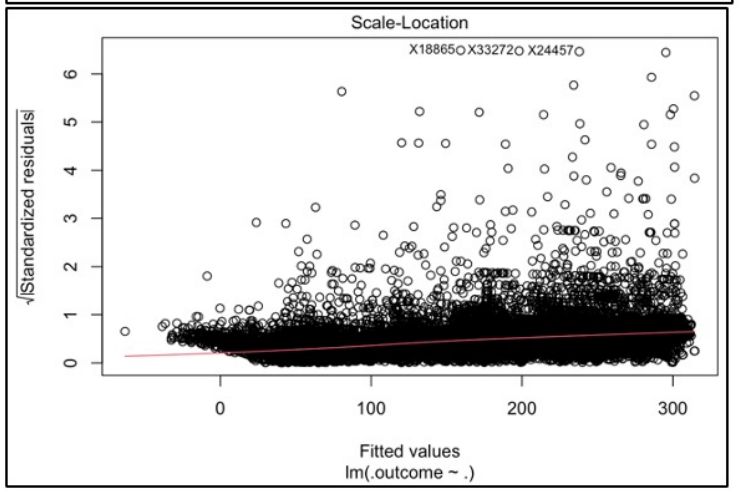
This tells us that the correlation between the two variables is negative but it's a statistically significant correlation since the p-value is less than 0.05.

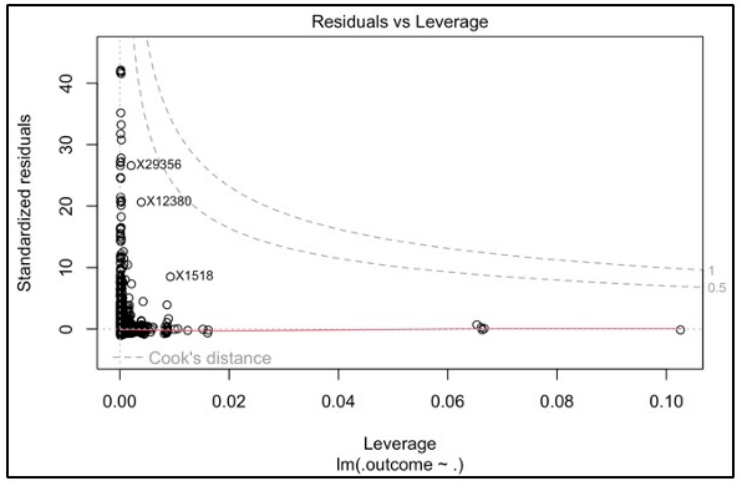
Linear Regression model 1

Comparing price with other variables like latitude, longitude, room type, minimum nights, availability in 365 days and neighbourhood group.







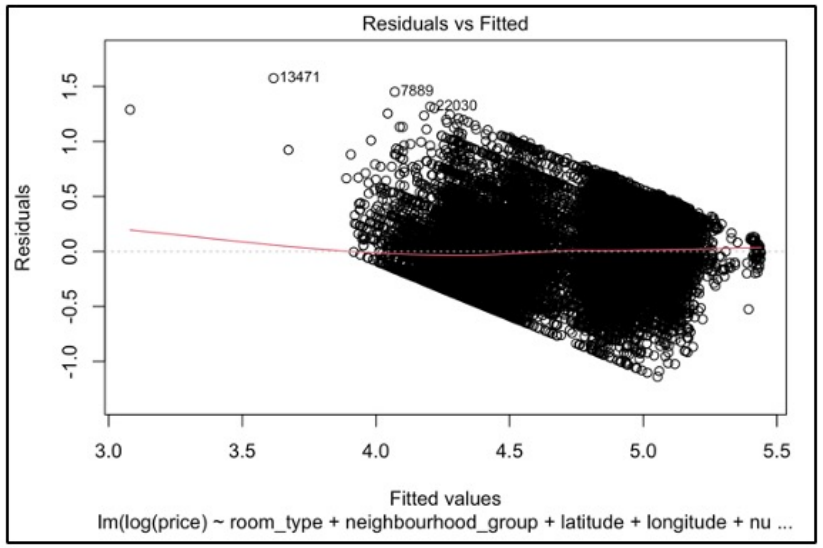


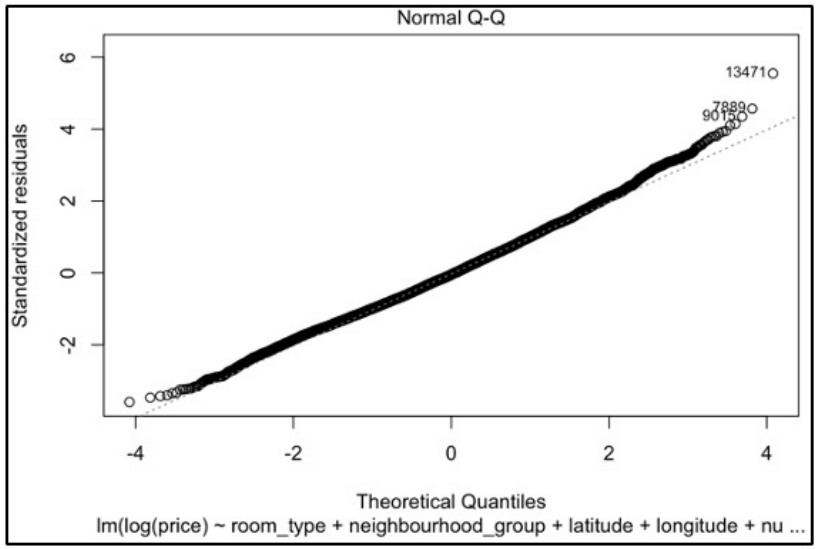
This model is not so good. Median residual error is -24.2, while it should be near 0. R2=0.1 is also not so good.

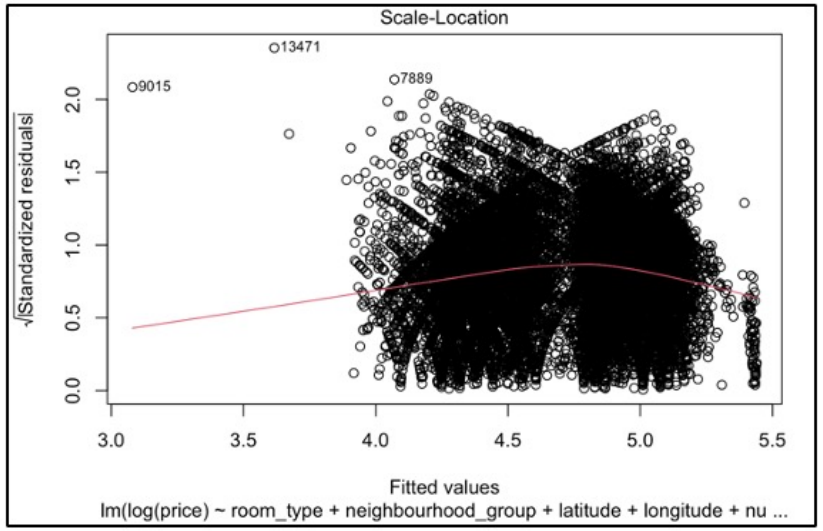
Normal Q-Q plot clearly shows that first linear model doesn't satisfy linear model assumptions (normal Q-Q plot should be straight line).
Since the model seems bad, it will not be used in predicting new prices.

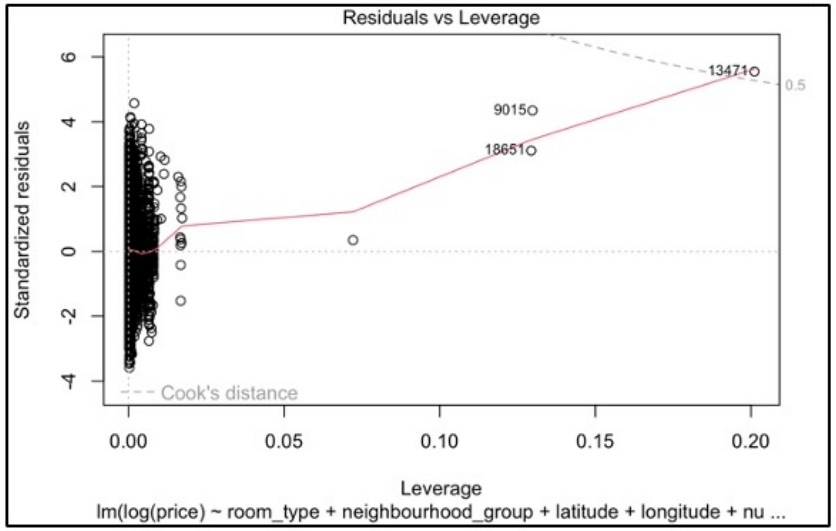
Linear Regression model 2

Second model will introduce logarithmic transformations. Also, training data set will be filtered by price, so outliers are removed.









Multiple Line Regression Model

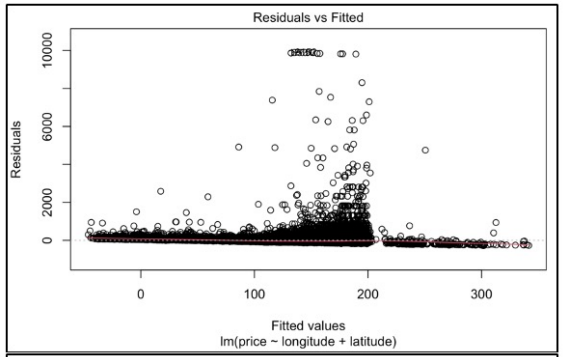
For price, longitude, and latitude

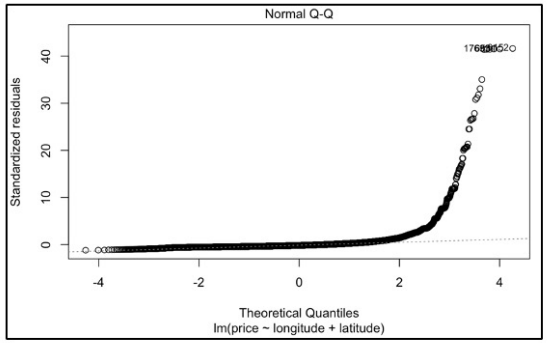
Residuals:

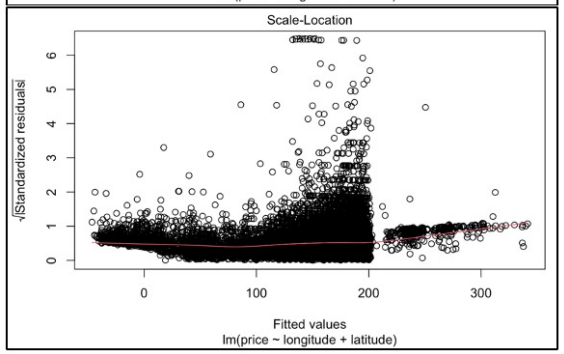
Min	1Q	Median	3Q	Max
-281.2	-78.0	-39.9	20.3	9867.7

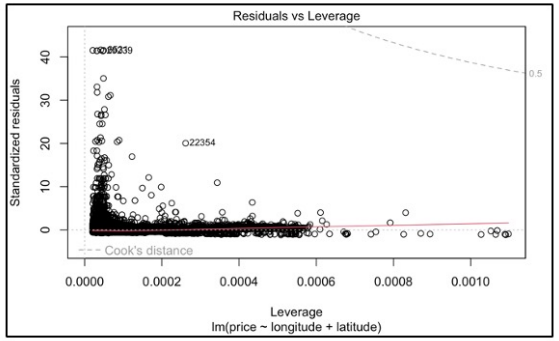
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-67533.95	1963.76	-34.39	<0.0000000000000002
longitude	-801.29	23.32	-34.36	<0.0000000000000002
latitude	206.98	19.74	10.48	<0.0000000000000002









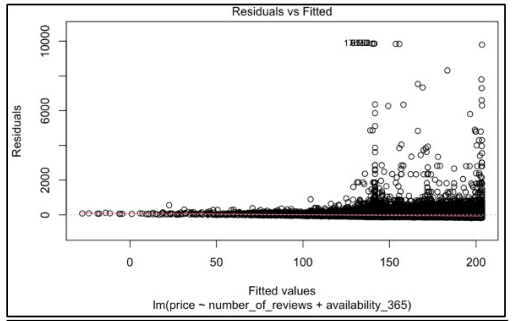
For price, number of reviews and availability for 365 days

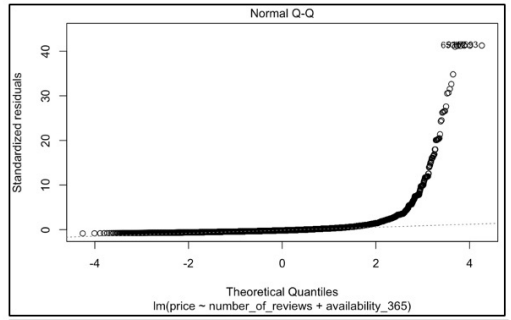
Residuals:

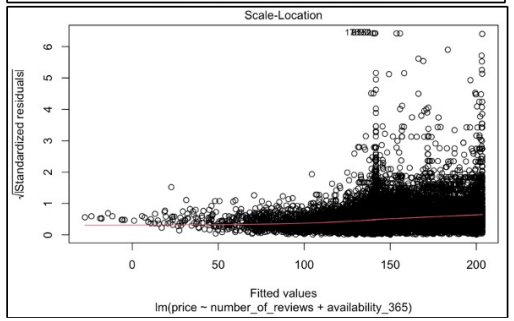
Min	1Q	Median	3Q	Max
-1.18692	-0.22518	-0.01606	0.20821	1.46626

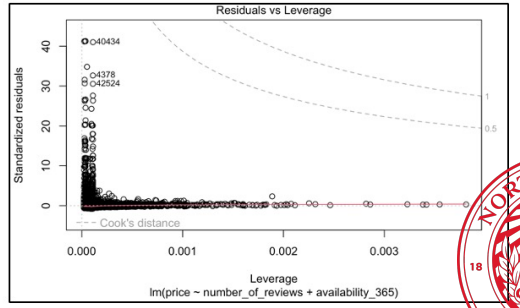
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	141.639282	1.480882	95.64	<0.0000000000000002
number_of_reviews	-0.344582	0.024616	-14	<0.0000000000000002
availability_365	0.169366	0.008332	20.33	<0.0000000000000002

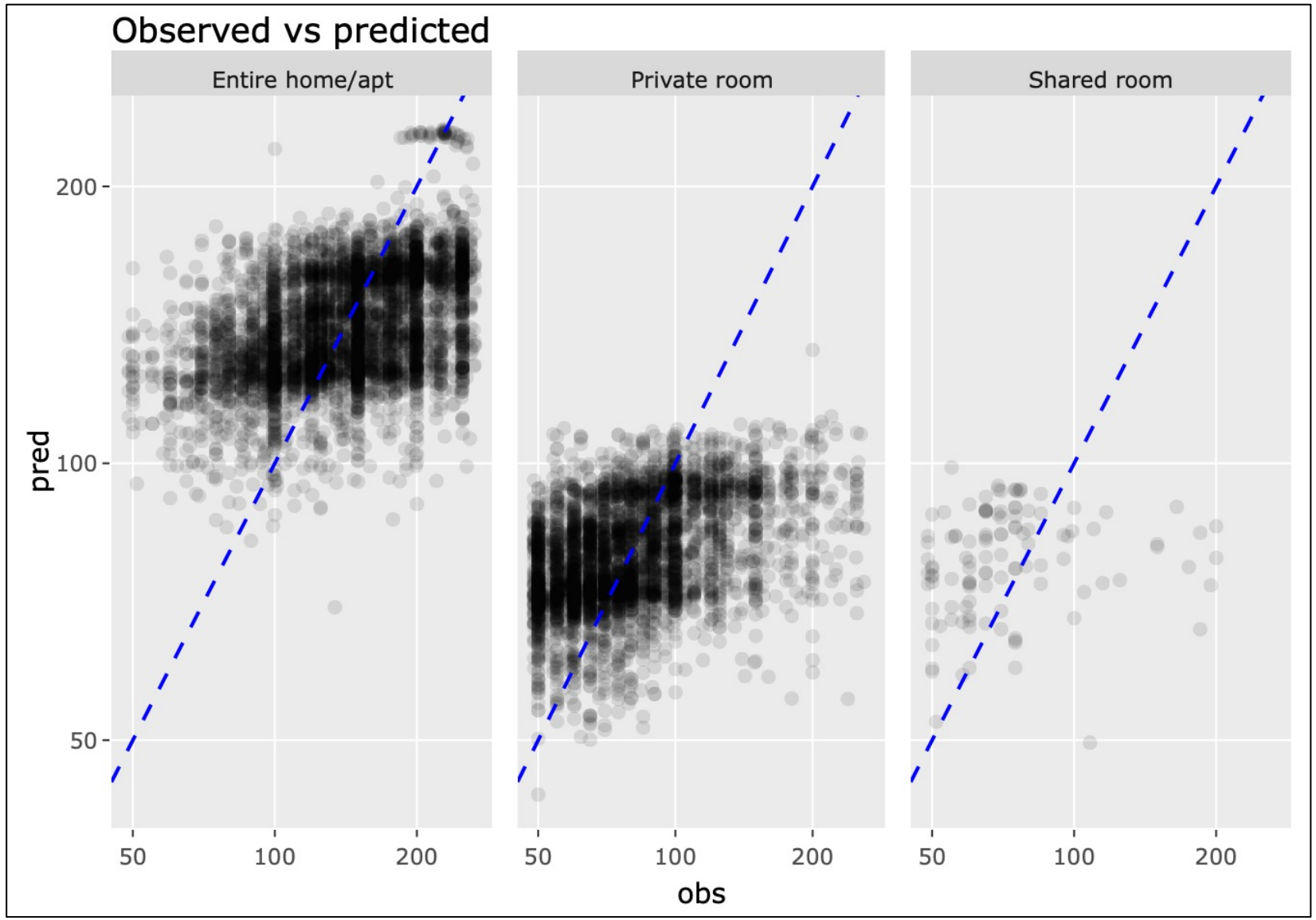








Observed vs predicted prices for training set using Linear regression model



Metrics for testing set: R2 = 0.43 and RMSE = 41.24

References

Bevans (2022, November 11);Datanovia (2019, December 26);*Linear Regression Example in r Using Lm() Function (n.d.);Domazet (2019, September 3);Investopedia (2022, August 31)*

Bevans, R. 2022, November 11. *Hypothesis Testing | a Step-by-Step Guide with Easy Examples*. https://www.scribbr.com/statistics/hypothesis-testing/.

Datanovia. 2019, December 26. *How to Do a t-Test in r: Calculation and Reporting*. https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/.

Investopedia. 2022, August 31. *What Is a Confidence Interval and How Do You Calculate It?*https://www.investopedia.com/terms/c/confidenceinterval.asp.

Linear Regression Example in r Using Lm() Function. n.d. https://www.learnbymarketing.com/tutorials/linear-regression-in-r/.

```
1  def gratitude():
2      print("Thank you.")
3
```

Any Questions?