Graphical user interface, text

Description automatically generated with medium confidence

Logo

Description automatically generated

ALY 6010

Capstone Report

Northeastern University, Vancouver, BC, Canada

Mohsen Soltanifar, PhD, AStat

Faculty Lecturer

December 2022



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

Abhilash Dikshit

Graduate Student at Northeastern University, Vancouver, BC, Canada

Master of Professional Studies in Analytics

**Abstract**

This data focuses on exploratory data analysis, but also includes two linear regression models and two multiple regression models along with predictive analysis with price as target variable. I have focused more on visualisation in the exploratory data analysis.

**Key values:**

Key Variables used: Price, Neighbourhood groups, Room Type, last review, number of reviews, availability\_365

Tests Performed: Hypothesis testing, Correlational testing, Regression testing

Plots used: Bar plot, Histogram, Box plot, Scatter plot

1. **Introduction About the Dataset**

### Context

Airbnb has expanded travel options and offered more distinctive, personalised ways for travellers to see the world. From 2011 through 2019, this dataset details listing activity and analytics in NYC, New York, USA.

### Content

This data file contains all the details required to learn more about hosts, their geographic accessibility, and the essential metrics to build forecasts and reach conclusions.

### Acknowledgements

Dataset taken from [Kaggle](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data). This is a public dataset of Airbnb, and the original source can be obtained on this [website](http://insideairbnb.com).

**2. Materials & Methods**

**2.1 Dataset**

Dataset taken from [Kaggle](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data). This is a public dataset of Airbnb, and the original source can be found in their website.

Number of Rows in the final dataset: 48895

Number of Columns in the final dataset: 14

Blank cells count in the final dataset: 10074

|  |  |
| --- | --- |
| **Variables** | **Class Type** |
| Name | Character |
| Hostname | Factor |
| Neighbourhood group | Factor |
| Neighbourhood | Factor |
| Latitude | Double |
| Longitude | Double |
| Room type | Factor |
| Price | Integer |
| Minimum Nights | Integer |
| Number of reviews | Integer |
| Last review | Date |
| Reviews per month | Double |
| Calculated host listing count | Integer |
| Availability 365 | Integer |

**Data Cleanup**

* Identified missing and NA values in the data.
* Showed the Dimension and summary of 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.
* Since I used stringsAsFactors = F in read.csv function, I have to transform following character columns to factor columns: host\_name, neighbourhood\_group, neighbourhood, room\_type.
* Column “last\_review” has to be converted to Airbnb type using function ymd from lubridate package.

**2.2 Statistical Analysis**

**2.2.1. Hypothesis tests**

Question1: Does Airbnb prices affects the number of reviews?

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

Question2: 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

**2.2.2. Correlation tests**

* Method: Spearman correlationfor Airbnb dataset
* Creating matrix of correlation coefficients and p-values for all the variables.

**2.2.3. Linear regression models**

* 1st Linear Regression model for price in comparison with latitude, longitude, room type, minimum nights, availability 365 days, neighbourhood group.
* 2nd Linear Regression model for price in comparison with room type, neighbourhood group, latitude, longitude, number of reviews, availability 365 days, reviews per month, calculated host listings count, minimum nights.
* Multiple Line Regression 1 for price, longitude, and latitude.
* Multiple Line Regression 2 for price, number of reviews and availability for 365 days.
* Confidence Intervals for Model Parameters.
* Predict prices for training set.

**2.2.4. Statistical software used**

|  |  |
| --- | --- |
| **Mac OS version: 13.1**  macOS is a Unix operating system developed and marketed by Apple Inc. | Table  Description automatically generated |
| **R Version: 4.2.2**  R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. | **Text  Description automatically generated** |
| **R studio version: 2022.12.0**  RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser. | **Graphical user interface, text, application  Description automatically generated** |

**3. Results**

3.1 Exploratory Data Analysis (EDA)

Columns “reviews\_per\_month” and “last\_review” have exactly the same value of percentage missing (~ 20.56 %) because if we don’t know when the last review was, we cannot calculate reviews per month).

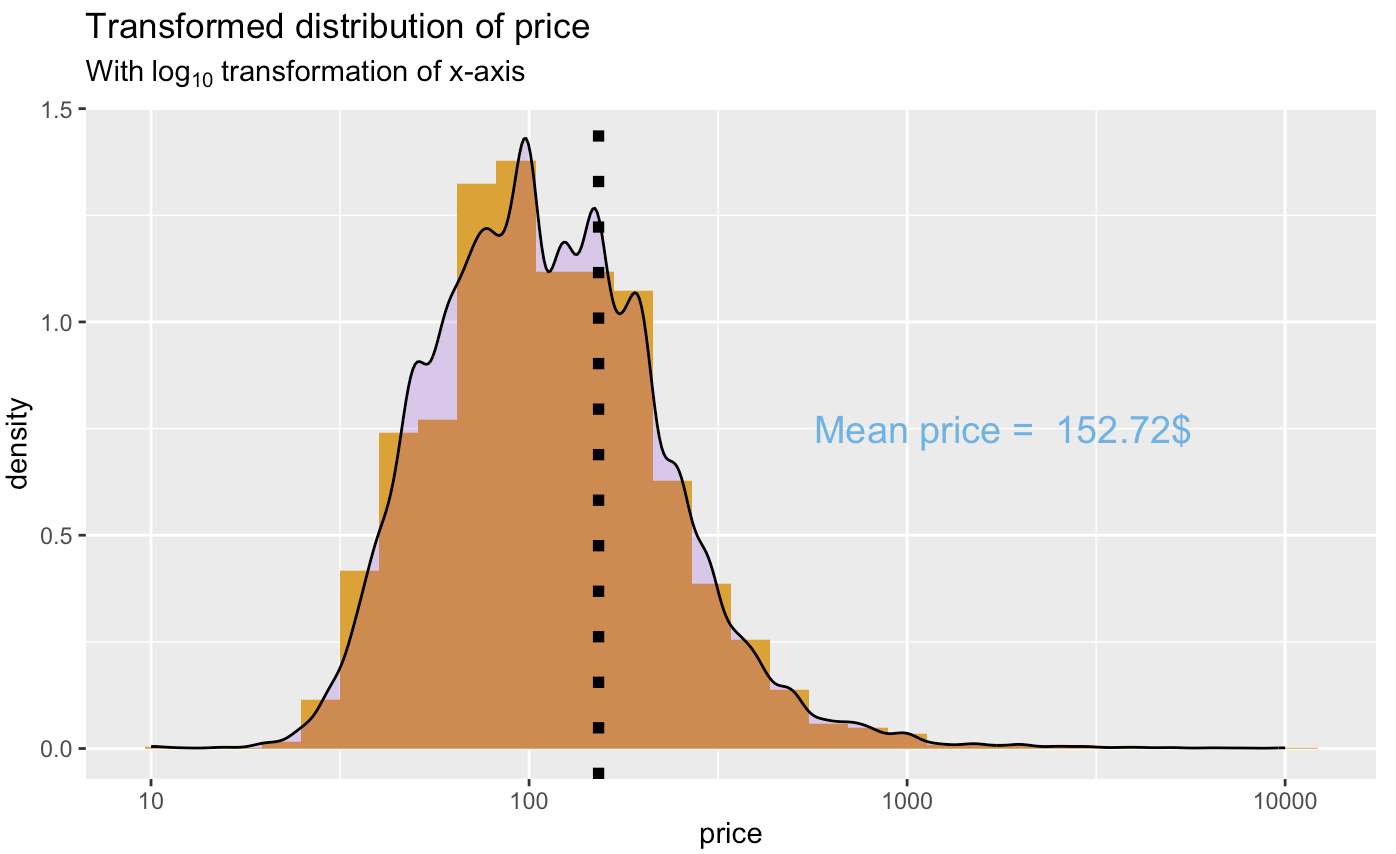
Timeline

Description automatically generated with low confidence

The most important (target) variable is price.

**Histogram & Density with log10 transformation for price:**

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



**Histogram & Density with log10 transformation for neighbourhood groups:**

**Chart, radar chart, surface chart

Description automatically generated**

**Above Average Price Objects by Neighbourhood Areas:**

**Chart, bar chart

Description automatically generated**

**Boxplot of price by room type:**

* Entire home or apartment
* Private Room
* Shared Room

|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated | 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:**

**Chart, scatter chart

Description automatically generated**

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

**Scatter Plot for Price and Number of Reviews:**

Chart

Description automatically generated

**Number of objects by neighbourhood areas:**

**Chart

Description automatically generated**

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.

3.2 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*

Created separate vectors *price\_high* and *price\_low* for Airbnb prices > $150 and <= $150.

Created scatter plot for x and y with same length as x.

Chart, scatter chart

Description automatically generated

Summed the number of reviews received on *price\_high* i.e., 296023 and for *price\_low* it is 841982.

Calculated the mean of higher number of reviews which is 19.81 and for lower number of reviews it is 24.79 and showed the summary of both the subsets.

Interpretation from the two-sample t-test:

* T is the test statistic value i.e., t = -12.145
* df is the degree of freedom i.e., df = 33308
* p-value is the significance level of t-test i.e., p-value < 0.00000000000000022
* Conf.int is the confidence interval of the mean at 95% i.e., conf.int = [-5.782239 -4.175280]
* A sample estimate is the mean value of the sample mean i.e., mean of x is 19.81678 and mean of y is 24.79554

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.

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

Stating Null and Alternative hypothesis using two sample tests:

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

Created separate vectors *price* *availability\_high* and *availability\_low* for Airbnb availability\_365> 200 days and <=200 days.

Created scatter plot for *availability\_high* and *availability\_low*.

**Chart, line chart

Description automatically generated**

Summed the number of reviews received on *availability\_high* i.e., 340612 and for *availability\_low,* it is 797393.

Calculated the mean of higher number of reviews which is 30.99 and for lower number of reviews it is 21.03 and showed the summary of both the subsets.

Interpretation for the two-sample t-test:

* T is the test statistic value i.e., t = 18.093
* df is the degree of freedom i.e., df = 15048
* p-value is the significance level of t-test i.e., p-value < 0.00000000000000022
* Conf.int is the confidence interval of the mean at 95% i.e., conf.int = [-8.88367 11.04240].
* A sample estimate is the mean value of the sample mean i.e., mean of x is 30.99854 and mean of y is 21.03551.

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.**

**Timeline

Description automatically generated with medium confidence**

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.

**Data Splitting**

Training set will be 70% percent of the original data. Objects with price equal to 0 will be omitted since price can’t be 0 (faulty records). They would make predictive models significantly weaker.

**1st Linear Regression model**

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

Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -287.7 | -62.7 | -24.5 | 15.5 | 9934.4 |

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | -28360.156 | 3974.79318 | -7.135 | 9.87E-13 |
| latitude | -208.34819 | 38.721078 | -5.381 | 7.4666E-08 |
| longitude | -500.79085 | 44.457645 | -11.264 | <2E-16 |
| `room\_typePrivate room` | -105.07581 | 2.663541 | -39.45 | <2E-16 |
| `room\_typeShared room` | -138.81411 | 8.536158 | -16.262 | <2E-16 |
| minimum\_nights | 0.135122 | 0.061115 | 2.211 | 0.027046 |
| availability\_365 | 0.163012 | 0.009931 | 16.414 | <2E-16 |
| neighbourhood\_groupBrooklyn | -35.83223 | 10.832431 | -3.308 | 0.000941 |
| neighbourhood\_groupManhattan | 28.650817 | 9.829601 | 2.915 | 0.003562 |
| neighbourhood\_groupQueens | -5.394891 | 10.490414 | -0.514 | 0.607067 |
| `neighbourhood\_groupStaten Island` | -152.35704 | 20.707774 | -7.357 | 1.92E-13 |

Residual standard error: 235.8 on 34207 degrees of freedom

Multiple R-squared: 0.09164, Adjusted R-squared: 0.09138

F-statistic: 345.1 on 10 and 34207 DF, p-value: < 0.00000000000000022

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.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | A picture containing chart  Description automatically generated |
| Chart, scatter chart  Description automatically generated | Chart, line chart  Description automatically generated |

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.

**2nd Linear Regression model**

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

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) | -127.0888 | 6.7128044 | -18.932 | <2E-16 |
| room\_typePrivate room | -0.542494 | 0.00445399 | -121.8 | <2E-16 |
| room\_typeShared room | -0.6637876 | 0.02026159 | -32.761 | <2E-16 |
| neighbourhood\_groupBrooklyn | -0.0432153 | 0.01940835 | -2.227 | 0.026 |
| neighbourhood\_groupManhattan | 0.15230698 | 0.01780571 | 8.554 | <2E-16 |
| neighbourhood\_groupQueens | 0.04003544 | 0.01884357 | 2.125 | 0.0336 |
| neighbourhood\_groupStaten Island | -0.606001 | 0.03642262 | -16.638 | <2E-16 |
| latitude | -0.5919605 | 0.06548045 | -9.04 | <2E-16 |
| longitude | -2.1102935 | 0.07558116 | -27.921 | <2E-16 |
| number\_of\_reviews | -0.0001121 | 0.00005345 | -2.098 | 0.0359 |
| availability\_365 | 0.00032225 | 0.00001818 | 17.722 | <2E-16 |
| reviews\_per\_month | -0.0003141 | 0.00156373 | -0.201 | 0.8408 |
| calculated\_host\_listings\_count | 0.00046581 | 0.00009215 | 5.055 | 4.34E-07 |
| minimum\_nights | -0.0011099 | 0.0001096 | -10.127 | <2E-16 |

Residual standard error: 0.3209 on 22015 degrees of freedom

Multiple R-squared: 0.4935, Adjusted R-squared: 0.4932

F-statistic: 1650 on 13 and 22015 DF, p-value: < 0.00000000000000022

This model is an improvement. Median residual error is now -0.0145, which is far better than -25.5 from the first model. R2=0.491 means that this model explains about 50% variance of target variable.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated | Chart, line chart  Description automatically generated |
| Chart, scatter chart  Description automatically generated | Chart, line chart  Description automatically generated |

**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 |

Residual standard error: 237.2 on 48892 degrees of freedom

Multiple R-squared: 0.0247, Adjusted R-squared: 0.02466

F-statistic: 619.1 on 2 and 48892 DF, p-value: < 0.00000000000000022

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated |  |
| Chart, scatter chart  Description automatically generated | Chart, scatter chart  Description automatically generated |

**Multiple Line Regression Model 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 |

Residual standard error: 238.9 on 48892 degrees of freedom

Multiple R-squared: 0.01066, Adjusted R-squared: 0.01062

F-statistic: 263.4 on 2 and 48892 DF, p-value: < 0.00000000000000022

|  |  |
| --- | --- |
| **Chart, scatter chart  Description automatically generated** | **Chart  Description automatically generated with medium confidence** |
| **Chart, scatter chart  Description automatically generated** |  |

**Confidence Intervals for Model Parameters**

Simple:Line Regression

|  |  |  |
| --- | --- | --- |
|  | 2.50% | 97.50% |
| (Intercept) | -140.24638 | -113.93122 |
| room\_typePrivate room | -0.5512241 | -0.5337639 |
| room\_typeShared room | -0.7035018 | -0.6240734 |
| neighbourhood\_groupBrooklyn | -0.0812571 | -0.0051735 |
| neighbourhood\_groupManhattan | 0.11740651 | 0.18720744 |
| neighbourhood\_groupQueens | 0.00310068 | 0.0769702 |
| neighbourhood\_groupStaten Island | -0.6773919 | -0.53461 |
| latitude | -0.7203069 | -0.4636142 |
| longitude | -2.258438 | -1.9621491 |
| number\_of\_reviews | -0.0002169 | -7.375E-06 |
| availability\_365 | 0.00028661 | 0.00035789 |
| reviews\_per\_month | -0.0033791 | 0.00275096 |
| calculated\_host\_listings\_count | 0.00028518 | 0.00064643 |
| minimum\_nights | -0.0013248 | -0.0008951 |

Multiple Line Regression for price, longitude, and latitude:

|  |  |  |
| --- | --- | --- |
|  | 2.50% | 97.50% |
| (Intercept) | -71382.941 | -63684.952 |
| longitude | -846.9973 | -755.5731 |
| latitude | 168.2826 | 245.6682 |

Multiple Line Regression for price, number of reviews and availability for 365 days:

|  |  |  |
| --- | --- | --- |
|  | 2.50% | 97.50% |
| (Intercept) | 138.736735 | 144.541828 |
| number\_of\_reviews | -0.3928289 | -0.2963344 |
| availability\_365 | 0.1530359 | 0.1856967 |

**4. Discussion**

**4.1 This work**

* Does price have any impact on the number of reviews received by customers?
* Does Airbnb availability have any impact on the number of reviews?
* What can we learn about different hosts and areas?
* What can we learn from predictions? (Ex: locations, prices, reviews, etc.)
* Which hosts are the busiest and why?
* Is there any noticeable difference of traffic among different areas and what could be the reason for it?

**4.2 Limitations**

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

**4.3 Future Work**

**Observed vs predicted prices for training set using Linear regression model**

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

Graphical user interface, chart, scatter chart

Description automatically generated

**5. References**

* Bevans ([2022, November 11](file:///Users/abidikshit/R_Projects/ALY6010/Final_Projects/FinalProject/Dikshit_Final_Project.html#ref-R-Career));Datanovia ([2019, December 26](file:///Users/abidikshit/R_Projects/ALY6010/Final_Projects/FinalProject/Dikshit_Final_Project.html#ref-R-Action));Linear Regression Example in r Using Lm() Function ([n.d.](file:///Users/abidikshit/R_Projects/ALY6010/Final_Projects/FinalProject/Dikshit_Final_Project.html#ref-R-Cran));Domazet ([2019, September 3](file:///Users/abidikshit/R_Projects/ALY6010/Final_Projects/FinalProject/Dikshit_Final_Project.html#ref-R-Material1));Investopedia ([2022, August 31](file:///Users/abidikshit/R_Projects/ALY6010/Final_Projects/FinalProject/Dikshit_Final_Project.html#ref-R-Material2))
* 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/>.
* Domazet, Josip. 2019, September 3. Mining NYC Airbnb Data. <https://www.kaggle.com/code/josipdomazet/mining-nyc-airbnb-data-using-r/report>.
* 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/>.

**6. Appendix**

my\_packages = c("tidyverse", "ggthemes", "ggExtra", "caret", "glmnet", "corrplot", "leaflet", "RColorBrewer", "plotly", "grid", "gridExtra", "dplyr", "ggplot2", "reshape2", "data.table", "knitr", "png", "ggimage", "GGally", "psych", "Hmisc")

lapply(my\_packages, require, character.only = T)

#install.packages(remotes)

#remotes::install\_github("r-lib/tidyselect")

airbnb = read.csv("/Users/abidikshit/R\_Projects/Data/AB\_NYC\_2019.csv", encoding="UTF-8", stringsAsFactors = F, na.strings = c(""))

cat("Number of Rows For Uncleaned Dataset:", nrow(airbnb), "\n") # Printing string and variable row count on the same line

cat("Number of Columns For Uncleaned Dataset:", ncol(airbnb), "\n")

cat("Blank cells count For Uncleaned Dataset:", sum(!complete.cases(airbnb))) # Displaying Blank Cells Count from the original data frame

colSums(is.na(airbnb))

#install.packages("DT", type = "binary")

library(DT)

# View the head and tail of the data

DT::datatable(headTail(airbnb, 3, 3, ellipsis = F), rownames = F)

options(scipen = 100)

cat("Dimension of the data","\n")

dim(airbnb)

cat("Summary of the data","\n")

summary(airbnb)

sapply(airbnb, class)

# Following columns can be omitted since they don’t carry any useful information and hence wont’ be used in predictive models:

# id

# host\_id

names\_to\_delete = c("id", "host\_id")

airbnb[names\_to\_delete] = NULL

# Since I used stringsAsFactors = F in read.csv function, I have to transform following character columns to factor columns:

# host\_name

# neighbourhood\_group

# neighbourhood

# room\_type

names\_to\_factor = c("host\_name", "neighbourhood\_group", "neighbourhood", "room\_type")

airbnb[names\_to\_factor] = map(airbnb[names\_to\_factor], as.factor)

# Column last\_review\_ has to be converted to airbnb type using function ymd from lubridate package

airbnb[c("last\_review")] = airbnb[c("last\_review")] %>% map(~lubridate::ymd(.x))

glimpse(airbnb)

th = theme(axis.title = element\_text(), axis.title.x = element\_text())

missing\_airbnb= airbnb %>% summarise\_all(~(sum(is.na(.))/n()))

missing\_airbnb= gather(missing\_airbnb, key = "variables", value = "percent\_missing")

missing\_airbnb= missing\_airbnb[missing\_airbnb$percent\_missing > 0.0, ]

ggplot(missing\_airbnb, aes(x = reorder(variables, percent\_missing), y = percent\_missing)) +

geom\_bar(stat = "identity", fill = "#56B4E9", aes(color = I('white')), size = 0.3)+

xlab('variables')+ coord\_flip() +

th +

labs(title = "Missing Data") +

xlab("Column name") + ylab("Percentage missing") +

annotate("text", x = 1.5, y = 0.1,

label = "host\_name and name have less than 0.001\n percentage missing",

color = "slateblue", size = 5)

# Columns reviews\_per\_month and last\_review have exactly the same value of percentage missing (~ 20.56 %) because if we don’t know when the last review was, we can not calculate reviews per month)

ggplot(airbnb, aes(price)) +

geom\_histogram(bins = 30, aes(y = ..density..), fill = "#E69F00") +

geom\_density(alpha = 0.2, fill = "purple") +

th +

ggtitle("Transformed distribution of price",

subtitle = expression("With" ~'log'[10] ~ "transformation of x-axis")) +

geom\_vline(xintercept = round(mean(airbnb$price), 2), size = 2, linetype = 3) +

scale\_x\_log10() +

annotate("text", x = 1800, y = 0.75,label = paste("Mean price = ", paste0(round(mean(airbnb$price), 2), "$")),

color = "#56B4E9", size = 5)

airbnb\_nh= airbnb %>%

group\_by(neighbourhood\_group) %>%

summarise(price = round(mean(price), 2))

ggplot(airbnb, aes(price)) +

geom\_histogram(bins = 30, aes(y = ..density..), fill = "#E69F00") +

geom\_density(alpha = 0.2, fill = "purple") +

th +

ggtitle("Transformed distribution of price\n by neighbourhood groups",

subtitle = expression("With" ~'log'[10] ~ "transformation of x-axis")) +

geom\_vline(data = airbnb\_nh, aes(xintercept = price), size = 2, linetype = 3) +

geom\_text(data = airbnb\_nh,y = 1.5, aes(x = price + 1400, label = paste("Mean = ",price)), color = "#56B4E9", size = 4) +

facet\_wrap(~neighbourhood\_group) +

scale\_x\_log10()

airbnb %>% filter(price >= mean(price)) %>% group\_by(neighbourhood\_group, room\_type) %>% tally %>%

ggplot(aes(reorder(neighbourhood\_group,desc(n)), n, fill = room\_type)) +

th +

xlab(NULL) +

ylab("Number of objects") +

ggtitle("Number of above average price objects",

subtitle = "Most of them are entire homes or apartments") +

geom\_bar(stat = "identity")

ggplot(airbnb, aes(x = room\_type, y = price)) +

geom\_boxplot(aes(fill = room\_type)) + scale\_y\_log10() +

th +

xlab("Room type") +

ylab("Price") +

ggtitle("Boxplots of price by room type",

subtitle = "Entire homes and apartments have the highest avg price") +

geom\_hline(yintercept = mean(airbnb$price), color = "purple", linetype = 2)

# 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.

x= airbnb$availability\_365

y= airbnb$price

plot(x, y, main = "Relationship between availability",

xlab = "Availability during year", ylab = "Price",

pch = 19, frame = F, col= "skyblue")

abline(lm(y ~ x, data = airbnb), col = "blue")

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

x= airbnb$number\_of\_reviews

y= airbnb$price

plot(x, y, main = "Relationship between number of reviews",

xlab = "Number of reviews", ylab = "Price",

pch = 19, frame = F, col= "skyblue")

abline(lm(y ~ x, data = airbnb), col = "blue")

airbnb %>% group\_by(neighbourhood\_group) %>% tally() %>%

ggplot(aes(x = reorder(neighbourhood\_group, n), n)) +

geom\_bar(stat = "identity", fill = "purple") +

ggtitle("Number of objects by neighbourhood group") +

geom\_text(aes(x = neighbourhood\_group, y = 1, label = paste0(n),

colour = ifelse(neighbourhood\_group %in%

c("Manhattan", "Brooklyn", "Queens"), '1', '2')),

hjust=-1.5, vjust=.5, size = 4,

fontface = 'bold') +

coord\_flip() +

scale\_color\_manual(values=c("white","black"), guide = F) +

labs(x = NULL, y = NULL)

# 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.

price\_high = subset(airbnb, subset=(airbnb$price>150), select= c(neighbourhood\_group:availability\_365))

price\_low = subset(airbnb, subset=(airbnb$price<=150), select= c(neighbourhood\_group:availability\_365))

price\_high

price\_low

length(price\_high$number\_of\_reviews) == length(price\_low$number\_of\_reviews)

x = price\_high$number\_of\_reviews

y = price\_low$number\_of\_reviews

plot(x, y[1:length(x)],main = "Number of reviews for price> 150 and <= 150", xlab="Number of reviews for price >150",ylab="Number of reviews for price <=150", col=c("darkblue","green"))

abline(lm(y[1:length(x)] ~ x), col = "red")

sum(as.numeric(price\_high$number\_of\_reviews, na.rm = TRUE))

sum(as.numeric(price\_low$number\_of\_reviews, na.rm = TRUE))

cat("Mean of higher number\_of\_reviews:", mean(price\_high$number\_of\_reviews), "\n")

cat("Mean of lower number\_of\_reviews:", mean(price\_low$number\_of\_reviews), "\n")

summary(price\_high$number\_of\_reviews)

summary(price\_low$number\_of\_reviews)

options(scipen = 100)

t.test(price\_high$number\_of\_reviews,price\_low$number\_of\_reviews, conf.level = 0.95)

availability\_high = subset(airbnb, subset=(airbnb$availability\_365>200), select= c(neighbourhood\_group:availability\_365))

availability\_low = subset(airbnb, subset=(airbnb$availability\_365<=200), select= c(neighbourhood\_group:availability\_365))

availability\_high

availability\_low

length(availability\_high$number\_of\_reviews) == length(availability\_high$number\_of\_reviews)

x = availability\_high$number\_of\_reviews

y = availability\_high$number\_of\_reviews

plot(x, y,main = "Number of reviews for availability > 200 and <= 200", xlab="Number of reviews for availability > 200",ylab="Number of reviews for availability <=200", , col=c("darkblue","green"))

abline(lm(y ~ x), col = "red")

sum(availability\_high$number\_of\_reviews, na.rm = TRUE)

sum(availability\_low$number\_of\_reviews, na.rm = TRUE)

cat("Mean of higher prices:", mean(availability\_high$number\_of\_reviews), "\n")

cat("Mean of lower prices:", mean(availability\_low$number\_of\_reviews), "\n")

summary(availability\_high$number\_of\_reviews)

summary(availability\_low$number\_of\_reviews)

options(scipen = 100)

t.test(availability\_high$number\_of\_reviews,availability\_low$number\_of\_reviews, conf.level = 0.95)

options(scipen = 100)

airbnb\_cor= airbnb[, sapply(airbnb, is.numeric)]

airbnb\_cor= airbnb\_cor[complete.cases(airbnb\_cor), ]

correlation\_matrix= cor(airbnb\_cor, method = "spearman", use = "complete.obs")

#correlation\_matrix

corrplot(correlation\_matrix, method = "color")

#create matrix of correlation coefficients and p-values

rcorr(as.matrix(airbnb\_cor))

airbnb= airbnb %>% mutate(id = row\_number())

airbnb\_train= airbnb %>% sample\_frac(.7) %>% filter(price > 0)

airbnb\_test= anti\_join(airbnb, airbnb\_train, by = 'id') %>% filter(price > 0)

# sanity check

nrow(airbnb\_train) + nrow(airbnb\_test) == nrow(airbnb %>% filter(price > 0))

options(scipen = 100)

first\_model= caret::train(price ~ latitude + longitude + room\_type + minimum\_nights + availability\_365 + neighbourhood\_group, data = airbnb\_train, method = "lm")

summary(first\_model)

# 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.

plot(first\_model$finalModel)

# 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.

learn= airbnb\_train %>% filter(price < quantile(airbnb\_train$price, 0.9) & price > quantile(airbnb\_train$price, 0.1)) %>% tidyr::drop\_na()

second\_model= lm(log(price) ~ room\_type + neighbourhood\_group + latitude + longitude

+ number\_of\_reviews + availability\_365

+ reviews\_per\_month +

calculated\_host\_listings\_count + minimum\_nights, data = learn)

# Summarize the results

summary(second\_model)

# This model is an improvement. Median residual error is now -0.0145, which is far better than -25.5 from the first model. R2=0.491 means that this model explains about 50% variance of target variable.

# Q-Q plot for this model looks much better than the previous one.

plot(second\_model)

airbnb\_test= airbnb\_test %>% filter(price <= quantile(airbnb\_train$price, 0.9) & price >= quantile(airbnb\_train$price, 0.1)) %>% tidyr::drop\_na()

pred\_regression= predict(second\_model, newdata = airbnb\_test)

pred\_regression= exp(pred\_regression)

RMSE\_regression= sqrt(mean( (airbnb\_test$price - pred\_regression)\*\*2 ))

SSE= sum((airbnb\_test$price - pred\_regression)\*\*2)

SSR= sum((pred\_regression - mean(airbnb\_test$price)) \*\* 2)

R2= 1 - SSE/(SSE + SSR)

regression\_results= tibble(

obs = airbnb\_test$price, pred = pred\_regression, diff = pred - obs,

abs\_diff = abs(pred - obs),

neighbourhood = airbnb\_test$neighbourhood,

name = airbnb\_test$name, group = airbnb\_test$neighbourhood\_group,

type = airbnb\_test$room\_type

)

regression\_plot= regression\_results %>%

ggplot(aes(obs, pred)) +

geom\_point(alpha = 0.1, aes(text = paste("Name:", name, "\nGroup:", group, "\nType:", type,

"\nPrice diff = ", diff))) +

th + scale\_x\_log10() + scale\_y\_log10() +

ggtitle("Observed vs predicted",

subtitle = "Linear regression model") + geom\_abline(slope = 1, intercept = 0, color = "blue", linetype = 2) + facet\_wrap(~type)

ggplotly(regression\_plot)

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

model\_mul1 = lm(price ~ longitude+latitude, data = airbnb)

summary(model\_mul1)

plot(model\_mul1)

model\_mul2 = lm(price ~ number\_of\_reviews+availability\_365, data = airbnb)summary(model\_mul2)

plot(model\_mul2)

cat("\n","Simple:Line Regression","\n")

confint(second\_model, level=0.95)

cat("\n","Multiple Line Regression for price, longitude and latitude:","\n")

confint(model\_mul1, level=0.95)

cat("\n","Multiple Line Regression for price, number of reviews and availability for 365 days:","\n")

confint(model\_mul2, level=0.95)

## NA