FINAL PROJECT REPORT

Group Number: C1

STAT 306
Group Project

Name	Student Number	Preferred E-mail Address
Panle Li	41464579	panlerichard@gmail.com
Kashish Joshipura	27745629	kashishjoshipura@gmail.com
Ankur Bhardwaj	83640458	ankurb75@gmail.com
Aryan Ballani	61663514	aryanballani@gmail.com

Introduction

The dataset is sourced from Airbnb and is publicly available. The original dataset can be found on Kaggle (https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data).

This dataset focuses on Airbnb listing activity in New York City (NYC) for the year 2019. Airbnb has been used since 2008, providing a platform for guests and hosts to enhance traveling experiences by offering unique and personalized accommodations.

It contains comprehensive information about hosts, geographical availability, and key metrics necessary for making predictions and drawing conclusions. It encompasses details that contribute to understanding the dynamics of Airbnb listings in NYC.

Our question of interest from this dataset is :

- **What is the association between various properties, including room type, price, and availability, and the reviews per month of Airbnb listings in New York City for the year 2019?**
- Through regression analysis, this study aims to explore and quantify the relationships between these properties and the reviews per month, recognizing that correlation does not imply causation. The goal is to identify statistically significant associations and provide valuable insights into factors influencing the popularity of Airbnb listings

To address the research question regarding the influence of various properties on reviews per month for Airbnb listings in New York City in 2019, a systematic plan will be executed. Beginning with data cleaning, we will handle missing values, outliers, and ensure uniform data formatting. Multicollinearity checks will involve examining correlation matrices and calculating variance inflation factors. Model selection will employ exhaustive methods, evaluating models using Mallow's Cp, AIC, and Adjusted R-squared. Cross-validation techniques will be implemented to assess model performance. Visualizations, including residual plots and scatter plots, will aid in understanding the relationships between predictor variables and reviews per month. The interpretation of coefficients will guide insights into the practical implications of room type, price, and availability on listing popularity, culminating in a comprehensive report with visual summaries of the analysis.

The dataset includes the following variables, each providing valuable information for analysis:

Variable Name	Variable Description	Variable Type
`id`	Listing ID	integer
`name`	Name of the Listing	character
`host_id`	Owner's ID	integer
`host_name`	Name of the Owner	character

neighbourhood_group	Location (cities eg. - Manhattan, brooklyn, etc)	Factor: Manhattan, Brooklyn, Bronx, Staten Island, Queens
neighborhood	Area	character
latitude	Latitude Coordinate	double
longitude	Longitude Coordinate	double
room_type	Type of room	Factor: Shared, Private, and, Entire Home/Apartment
price	Price per night (in $)	integer
minimum_nights	Minimum number of nights one has to book the room	integer
number_of_reviews	Number of reviews	integer
last_review	Date of last review provided	character
calculated_host_listings_count	Amount of listings per host	integer
availability_365	Number of days the place is available in a year	integer
review_per_month (Response Variable)	Number of reviews per month	double

The dataset is rich in information, providing a comprehensive view of Airbnb listings in NYC for the specified year. This information will be leveraged to address research questions and make predictions about factors influencing reviews per month. The dataset is rich and has 48895 observations.

Analysis

Exploratory Data Analysis:

We initiated the data preprocessing by loading the dataset and eliminating irrelevant explanatory variables that were deemed to lack real-world significance. The selected variables for further analysis were as follows:

(i) `review_per_month`
(ii) `room_type`
(iii) `latitude`
(iv) `longitude`
(v) `price`
(vi) `neighbourhood_group`
(vii) `minimum_nights`
(viii) `availability_365`
(ix) `calculated_host_listings_count`

As part of the data transformation process, we introduced a new variable named `minimum_price` by multiplying the `price` and `minimum_nights` features. This derived variable is particularly meaningful in a real-world context, as it provides valuable insights into the minimum amount (in dollars) required for an individual to stay at a specific AirBnB, considering the interdependence of these features. In doing so we also ensured that we tackle any type of multicollinearity that might fester later on in our project.

Subsequently, we conducted an examination for any missing values across our variables and identified that our response variable, `review_per_month`, exhibited a deficit of 10,052 values. Faced with the decision of either imputing these missing values or opting for deletion of the corresponding observations, we chose the latter. This decision was driven by the consideration that imputing the missing values in the response variable might introduce bias to the model. Even with the removal of these 10,052 observations, we retained a substantial dataset comprising 38,843 observations, providing an ample dataset for training our model.

Continuing, we aimed to prevent multicollinearity in our dataset. Using the GGally library, we visualized and calculated the correlation between each continuous variable, ensuring a thorough assessment of interrelationships and effective mitigation of potential multicollinearity issues in regression analysis.

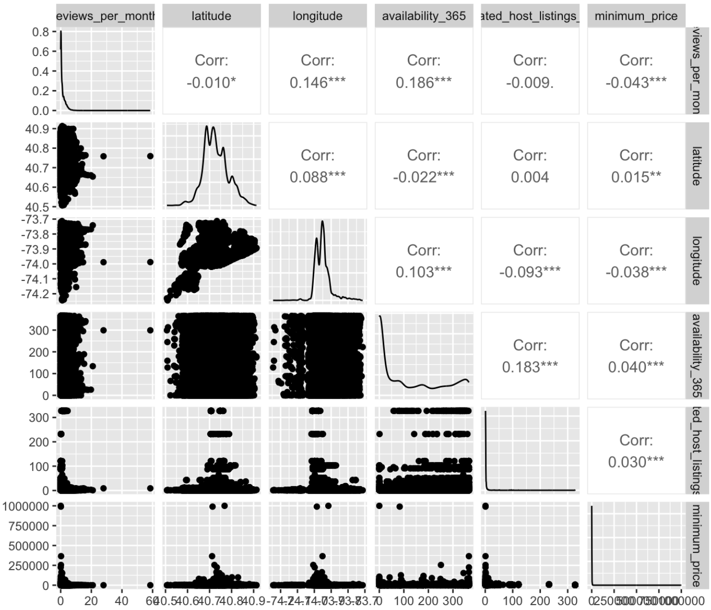


Fig1: *ggplot* showing correlation of every continuous variable with the rest

Looking at the results of **Fig1**, we noticed that not all variables were normal or had constant variance, so we decided to plot a histogram to examine the variables and apply the necessary transformations to ensure a good fit. The initial histograms are attached below:

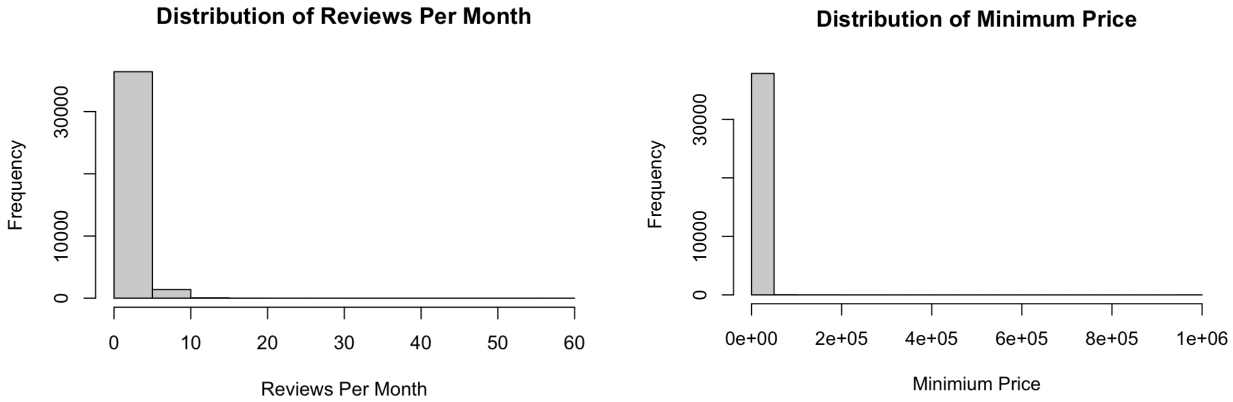
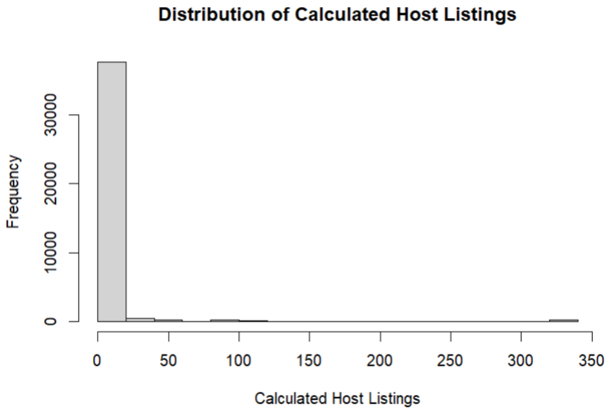


Fig2: Pre-transformation Histograms



Upon plotting these histograms, we also observed that `calculated_host_listings_count` had some outliers, so we removed these outliers before trying to transform the variables. After trying multiple different transformations on these variables, we came up with the final transformations which made our variables approximately normal.

`review_per_month → log(review_per_month)`

`minimum_price → log(minimum_price )^(⅓)`

`calculated_host_listings_count → log(calculated_host_listings_count)^(½)`

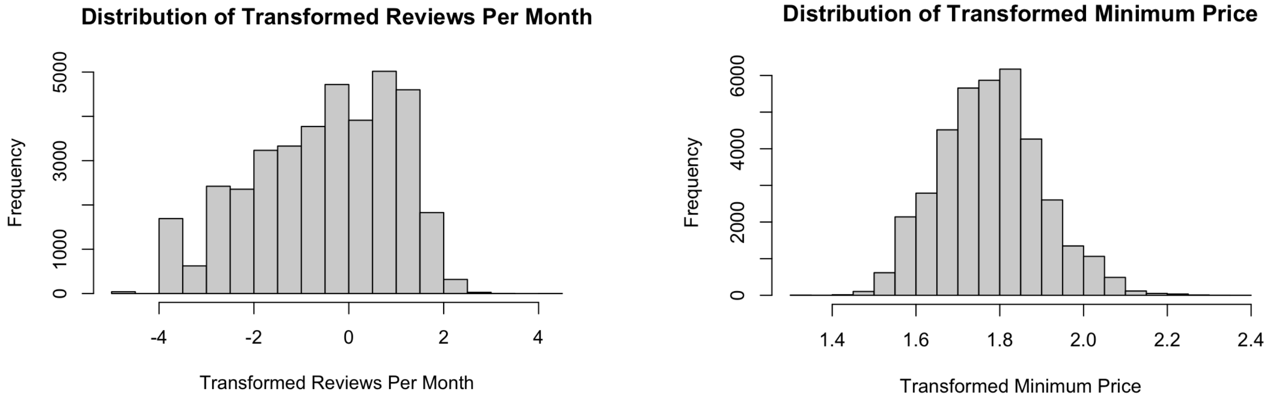
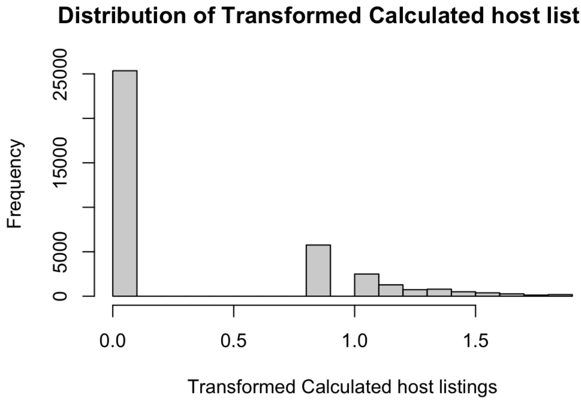


Fig3: Post-transformation Histograms



After applying all the necessary transformations, we plotted each continuous variable against our transformed response variable to check if the distribution is evenly spread out. We also examined the ***qqplot*** for the transformed response variable to check if the transformed response was now normally distributed. All the plots have been attached below.

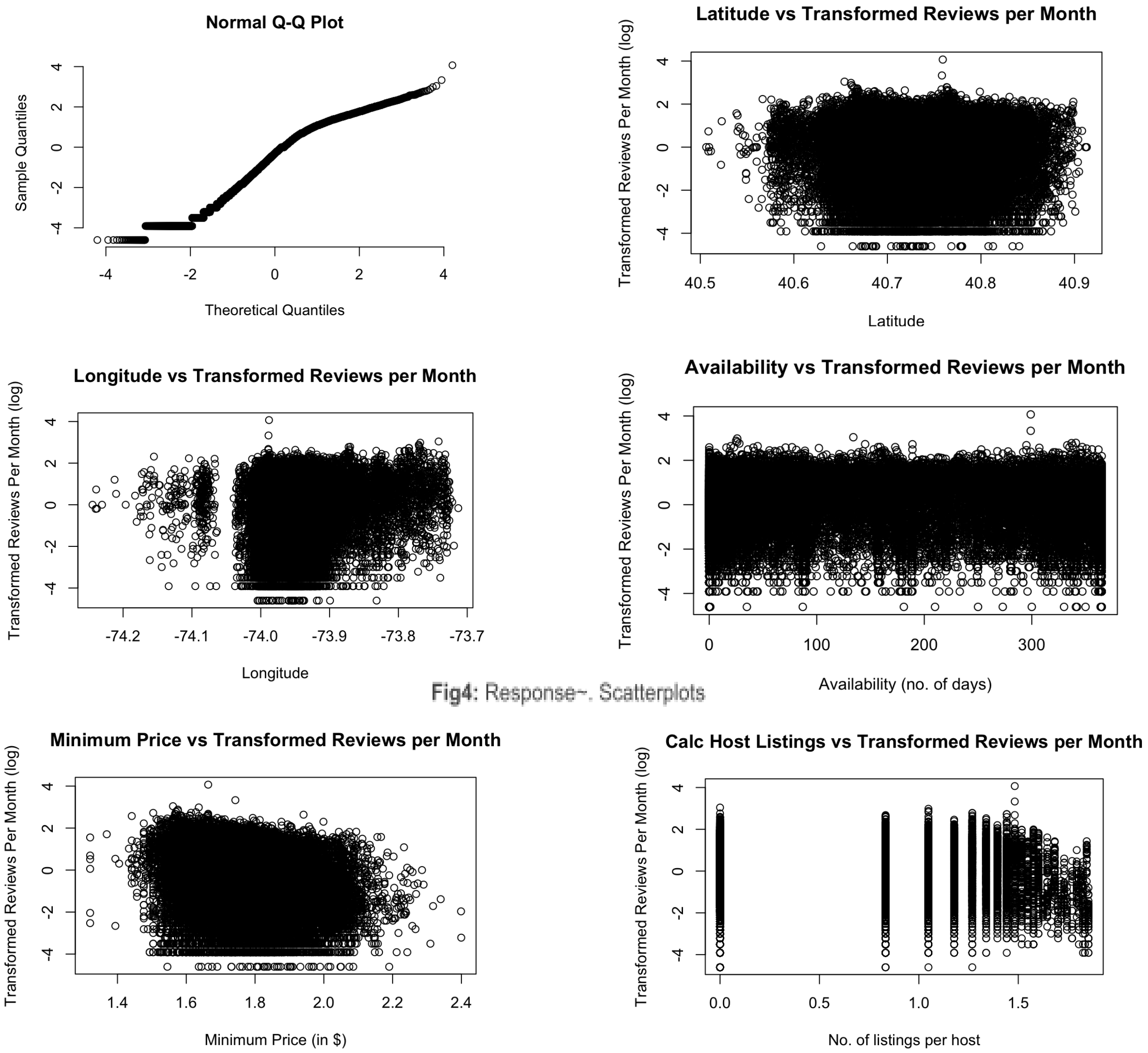


Fig4: Response~. Scatterplots

We were contempt with all the transformations and distributions so we moved on to the model training phase.

Model Training:

I. Full model without interactions:

Full Model:

$$\hat{Y} = \exp\left(0.0199 + 1.87x_1 - 1.35x_2 + 0.00337x_3 + 0.273x_4 - 3.51x_5 - 0.312x_6 - 0.0533x_7 - 0.19x_8 + 0.052x_9 - 0.482x_{10} - 0.815x_{11}\right)$$

where:

$\hat{Y}$ = reviews_per_month

x_1 = longitude

x_2 = latitude

x_3 = availability_365

$x_4 = \sqrt{\log(\text{calc\_host\_listings-count})}$

$x_5 = \sqrt[3]{\log(\text{minimum\_price})}$

$$x_{10} = \begin{cases} 1, & \text{room\_type = "Private Room"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{11} = \begin{cases} 1, & \text{room\_type = "Shared Room"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_6 = \begin{cases} 1, & \text{neighbourhood\_group = "Brooklyn"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_7 = \begin{cases} 1, & \text{neighbourhood\_group = "Mahatten"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_8 = \begin{cases} 1, & \text{neighbourhood\_group = "Queens"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_9 = \begin{cases} 1, & \text{neighbourhood\_group = "Staten Island"} \\ 0, & \text{otherwise.} \end{cases}$$

Baseline for room_type = "Entire Home/Apt"

Baseline for neighbourhood_group = "Bronx"

Using the selected variables along with the necessary transformations, we train a full regression model without any interactions. Afterwards, we use the exhaustive method through the regsubsets function to further study our model in depth. We took into consideration adj R^2, C_p and AIC statistics to help us validate our model and select the best one. We then created a C_p vs p plot.

Our best model indicates dropping Staten Island from `neighbourhood_group` however we shouldn't simply drop coefficients for the levels of a categorical variable even if they are not statistically significant. We can either drop the categorical variable as a whole or not drop it at all because dropping a single coefficient from a categorical variable might lead us to create a biased model. So, we proceeded to not drop the categorical variable and worked with the full model instead.

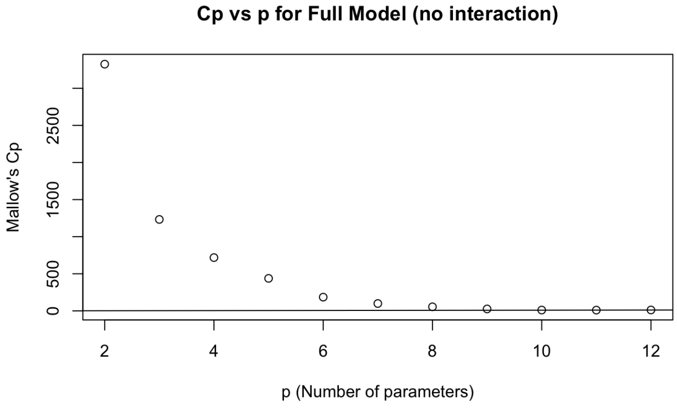


Fig5: C_p vs p plot for model with no interactions

We notice that our best C_p value (10.2198) is for the model with p = 11 because the ideal C_p value should be as close to p as possible. Furthermore, our adj R^2 value is 0.4632 and AIC value is 31847.6 for the selected best model.

Now we would like to study how interactions between different explanatory variables could further improve our full model.

II. Full model with `longitude`*`latitude` interaction :

Full Model :

$$\hat{Y} = exp\big(0.00015 - 0.036\,x_1 - 0.815\,x_2 + 0.00337\,x_3 + 0.273\,x_4 - 3.51\,x_5 - 0.344.x_6 - 0.102\,x_7 - 0.219\,x_8 + 0.089.x_9 - 0.482\,x_{10} - 0.815\,x_{11} - 4.96\,x_{12}\big)$$

where :

$\hat{Y}$ = reviews_per_month

x_1 = longitude

x_2 = latitude

x_3 = availability_365

$$x_4 = \sqrt{\log\,(\text{calc\_host\_listings-count})}$$

$$x_5 = \sqrt[3]{\log\,(\text{minimum\_price})}$$

$$x_{10} = \begin{cases} 1, & \text{room\_type} = \text{"Private Room"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{11} = \begin{cases} 1, & \text{room\_type} = \text{"Shared Room"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_6 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Brooklyn"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_7 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Mahatten"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_8 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Queens"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_9 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Staten Island"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{12} = (\text{latitude} * \text{longitude})$$

Baseline for room_type = " Entire Home/Apt"

Baseline for neighbourhood_group = " Bronx"

First interaction we will study is between `latitude` and `longitude`. We are interested in their interaction as we believe geographic location can be an important factor influencing the number of `review_per_month`. Hence, we train a full regression model with interaction between `latitude` and `longitude`. Afterwards, we use the exhaustive method through the regsubsets function to further study our model in depth. We took into consideration adj R^2, C_p and AIC statistics to help us validate our model and select the best one. We then created a C_p vs p plot.

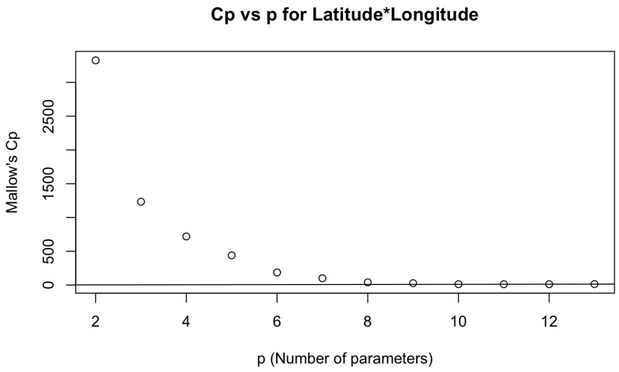


Fig6: C_p vs p plot for model with interaction between latitude and longitude

Our best model indicates dropping Staten Island from `neighbourhood_group` however we shouldn't simply drop coefficients for the levels of a categorical variable even if they are not statistically significant. We can either drop the categorical variable as a whole or not drop it at all because dropping a single coefficient from a categorical variable might lead us to create a biased model. So, we proceeded to not drop the categorical variable and worked with the full model instead.

We notice that our best C_p value (11.6132) is for the model with p = 12 because the ideal C_p value should be as close to p as possible . Furthermore, our adj R^2 value is 0.4632 and AIC value is 31847.4 for the selected best model. Since our adj R^2 and AIC values do not improve noticeably, we will further explore different interactions.

III. Full model with `minimum_price` * `neighbourhood_group` interaction:

Full Model :

$$\hat{Y} = \begin{cases} \exp\left(0.0195 - 1.41x_1 - 0.826x_2 + 1.775x_3 + 0.273x_4 - 4.011x_5 + 0.0195 x_6 - 0.479x_7\right), & \text{neighbourhood\_group} = \text{"Bronx"} \\ \exp\left(-1.8305 - 1.41x_1 - 0.826x_2 + 1.775x_3 + 0.273x_4 - 3.127x_5 + 0.0195 x_6 - 0.479x_7\right), & \text{neighbourhood\_group} = \text{"Brooklyn"} \\ \exp\left(-0.688 - 1.41x_1 - 0.826x_2 + 1.775x_3 + 0.273x_4 - 3.6263x_5 + 0.0195 x_6 - 0.479x_7\right), & \text{neighbourhood\_group} = \text{"Queens"} \\ \exp\left(0.4805 - 1.41x_1 - 0.826x_2 + 1.775x_3 + 0.273x_4 - 4.389x_5 + 0.0195 x_6 - 0.479x_7\right), & \text{neighbourhood\_group} = \text{"Manhatten"} \\ \exp\left(-0.457 - 1.41x_1 - 0.826x_2 + 1.775x_3 + 0.273x_4 - 3.7123x_5 + 0.0195 x_6 - 0.479x_7\right), & \text{neighbourhood\_group} = \text{"Staten Island"} \end{cases}$$

where :

$$\hat{Y} = \text{reviews\_per\_month}$$
$$x_1 = \text{longitude}$$
$$x_2 = \text{latitude}$$
$$x_3 = \text{availability\_365}$$
$$x_4 = \sqrt{\log(\text{calc\_host\_listings-count})}$$

$$x_5 = \sqrt[3]{\log(\text{minimum\_price})}$$
$$x_6 = \begin{cases} 1, & \text{room\_type} = \text{"Private Room"} \\ 0, & \text{otherwise.} \end{cases}$$
$$x_7 = \begin{cases} 1, & \text{room\_type} = \text{"Shared Room"} \\ 0, & \text{otherwise.} \end{cases}$$

Second interaction we will study is between transformed `minimum_price` and `neighbourhood_group` . We are interested in their interaction as we believe the prices might differ depending on the location of the neighbourhood and therefore can be an important factor influencing the number of `review_per_month`. Hence, we train a full regression model with interaction between transformed minimum price and neighbourhood group. Afterwards, we use the exhaustive method through the regsubsets function to further study our model in depth. We took into consideration adj R^2, C_p and AIC statistics to help us validate our model and select the best one. We then created a C_p vs p plot.

Our best model indicates dropping Staten Island from `neighbourhood_group` however we shouldn't simply drop coefficients for the levels of a categorical variable even if they are not statistically significant. We can either drop the categorical variable as a whole or not drop it at all because dropping a single coefficient from a categorical variable might lead us to create a biased model. So, we proceeded to not drop the categorical variable and worked with the full model instead.

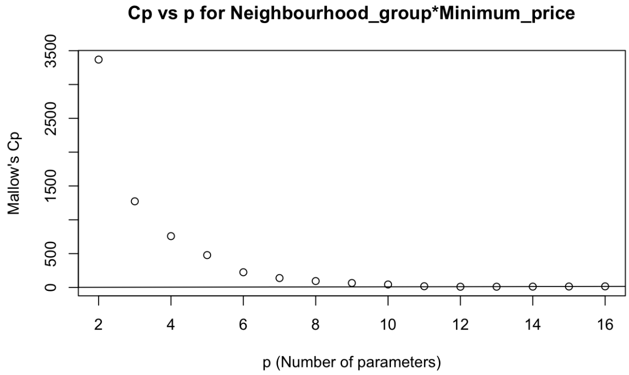


Fig7: C_p vs p plot for model with interaction between neighbourhood_group and minimum_price

We notice that our best C_p value (14.1035) is for the model with p = 15 because the ideal C_p value should be as close to p as possible . Furthermore, our adj R^2 value is 0.4641 and AIC value is 31811.8 for the selected best model. Our adj R^2 and AIC values improve very slightly, however we will further explore different interactions to improve our model.

IV. Full model with `minimum_price` * `calculated_host_listings_count` interaction:

Full Model:

$$\hat{Y} = exp\big(0.0019 + 1.923\, x_1 - 1.11\, x_2 + 0.00337 x_3 + 4.487\, x_4 - 2.39\, x_5 - 0.28\, x_6 - 0.038\, x_7 - 0.194\, x_8 + 0.083. x_9 - 0.468. x_{10} - 0.901\, x_{11} - 0.098\, x_{12}\big)$$

where :

$\hat{Y}$ = reviews_per_month

x_1 = longitude

x_2 = latitude

x_3 = availability_365

$$x_4 = \sqrt{\log\left(\text{calc\_host\_listings-count}\right)}$$

$$x_5 = \sqrt[3]{\log\left(\text{minimum\_price}\right)}$$

$$x_{10} = \begin{cases} 1, & \text{room\_type} = \text{"Private Room"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{11} = \begin{cases} 1, & \text{room\_type} = \text{"Shared Room"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_6 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Brooklyn"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_7 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Mahatten"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_8 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Queens"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_9 = \begin{cases} 1, & \text{neighbourhood\_group} = \text{"Staten Island"} \\ 0, & \text{otherwise.} \end{cases}$$

$$x_{12} = \left(\sqrt{\log\left(\text{calc\_host\_listings-count}\right)} \ast \sqrt[3]{\log\left(\text{minimum\_price}\right)}\right)$$

Baseline for room_type = " Entire Home/Apt"

Baseline for neighbourhood_group = " Bronx"

Third interaction we will study is between transformed minimum_price and transformed calculated_host_listings_count . We are interested in their interaction as we believe the prices might differ depending on how many properties a host owns. If a host owns multiple properties, they could potentially afford to rent out their properties for a cheaper price as compared to single property owners and therefore can be an important factor influencing the number of review_per_month. Hence, we train a full regression model with interaction between transformed minimum_price and transformed calculated_host_listings_count. Afterwards, we use the exhaustive method through the regsubsets function to further study our model in depth. We took into consideration adj R^2, C_p and AIC statistics to help us validate our model and select the best one. We then created a C_p vs p plot.

Our best model indicates dropping Manhattan from neighbourhood_group however we shouldn't simply drop coefficients for the levels of a categorical variable even if they are not statistically significant. We can either drop the categorical variable as a whole or not drop it at all because dropping a single coefficient from a categorical variable might lead us to create a biased model. So, we proceeded to not drop the categorical variable and worked with the full model instead.

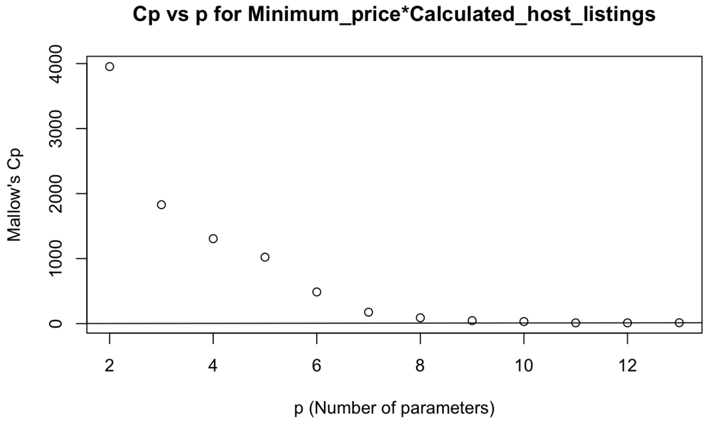


Fig8: C_p vs p plot for model with interaction between calculated_host_listings_count and minimum_price

We notice that our best C_p value (11.5183) is for the model with p = 11 because the ideal C_p value should be as close to p as possible . Furthermore, our adj R^2 value is 0.5757 and AIC value is 26275.2 for the selected best model. Our adj R^2 and AIC values improve noticeably and hence we will consider this model as our final model.

V. Final Model:

Based on the model selection process we went through, we decided to select the full model with interaction between transformed minimum_price and transformed calculated_host_listings_count. We supported our decision using the Mallow's Cp, adj R^2 and AIC values. Amongst all the trained models, we noticed that the model with interaction between transformed minimum_price and transformed calculated_host_listings_count has the best adj R^2 and AIC values and hence decided to select it as the best model.

Model Validation:

After selecting our final model, we conducted a two-fold cross-validation to assess its performance on previously unseen data. Both models trained on the folds demonstrated strong performance, yielding a total error of 2.6365. This level of accuracy is noteworthy for real-life predictions. The cross-validation results suggest that our model can predict the number of reviews per month within a margin of ±2.6365.

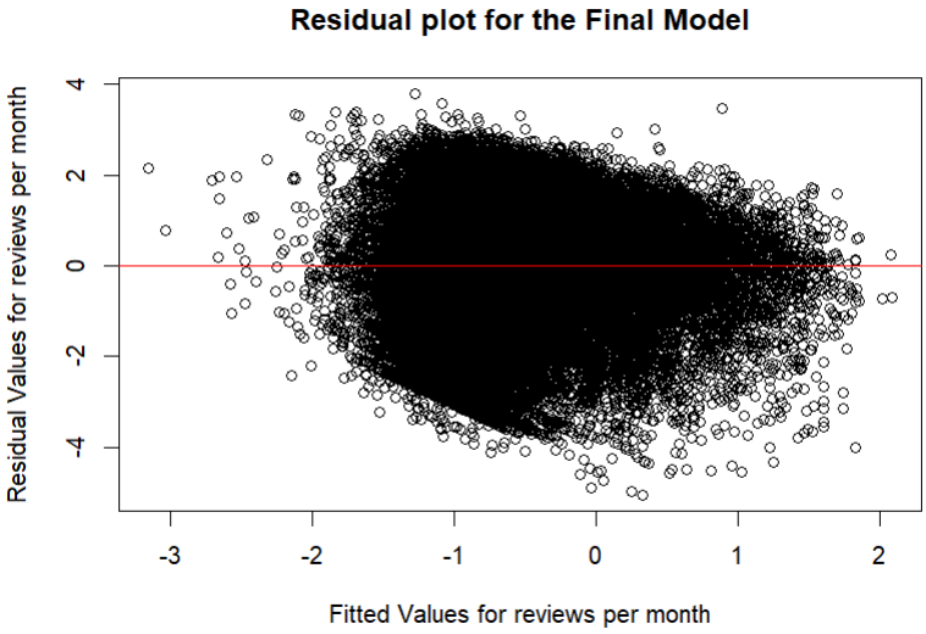


Fig9: Residual plot for final model

Additionally, we generated a residual plot for our final fitted model by plotting the fitted values against the predicted values. Upon inspection, we observed a lack of any discernible pattern in the residuals; they were randomly distributed with constant variance across all data points. This observation reinforces our confidence in the final fitted model as a good fit for the data.

Conclusion

Through a comprehensive analysis of the New York City Airbnb dataset for 2019, this study aimed to unravel the intricate associations between various property features and the reviews per month received by Airbnb listings. The exploration focused on understanding the factors influencing the popularity of these listings and sought to address the initial research question.

This analysis commenced with exploratory data examination and preprocessing steps, where relevant variables were selected, and transformations were applied to ensure the suitability of the dataset for regression analysis. Model training involved several iterations, starting with the full regression model and progressing to explore interactions between variables. The final model incorporates an interaction between transformed `minimum_price` and transformed `calculated_host_listings_count`, demonstrating the most favorable performance based on adjusted R-squared, Mallow's Cp, and AIC values.

The final model emphasized an association between host property ownership and the number of reviews per month. Interestingly, the model underscored the importance of considering host behavior and pricing strategies concerning the number of properties they manage, indicating potential implications for pricing policies and property ownership strategies on Airbnb.

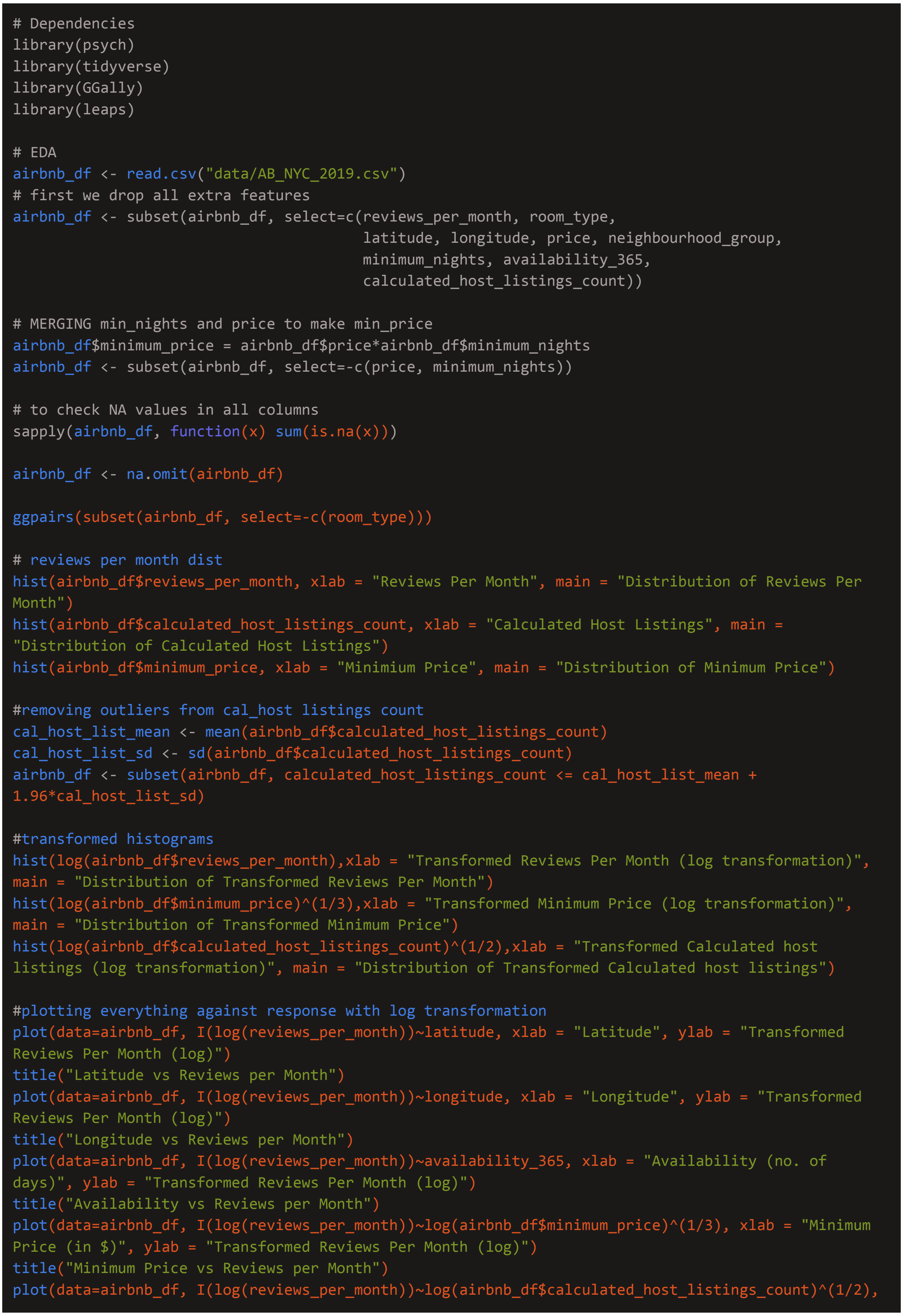
In conclusion, the analysis reveals the complex dynamics that influence the popularity of Airbnb listings in New York City. While the model provides insights, it is important to note that causality cannot be inferred from correlation alone. Future research work may explore additional factors in depth or refine the model to capture more of the nuances that influence the popularity of Airbnb listings to further deepen the understanding of consumer behavior in the context of short-term rentals.

The findings of this study provide a platform for policymakers and Airbnb hosts to make informed decisions about property management, pricing strategies, and enhancing the guest experience. All in all, the goal to identify statistically significant associations and provide valuable insights into factors influencing the popularity of Airbnb listings was achieved

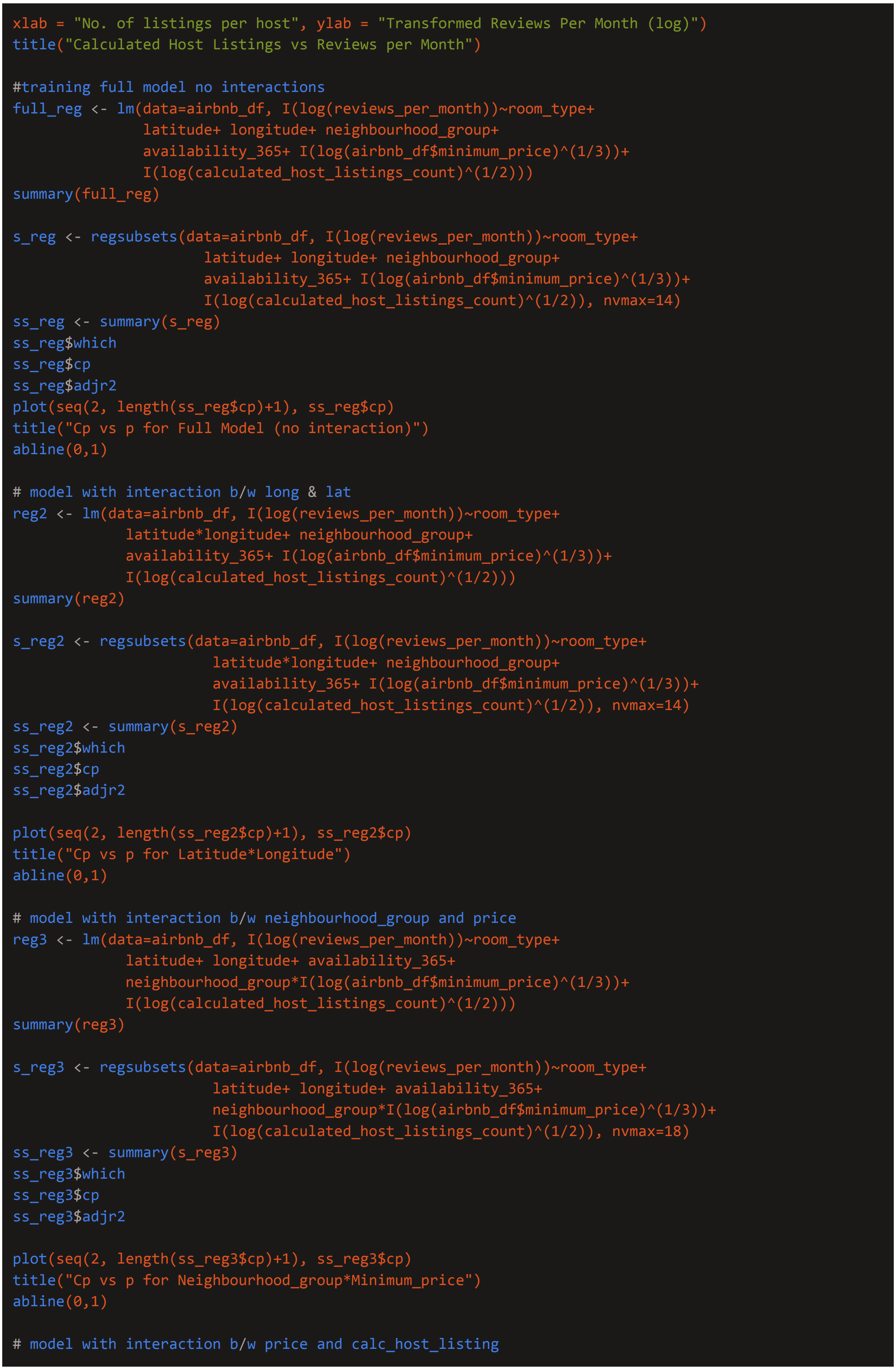
References

Aleksandr Blekh (1961, April 1). *Is it advisable to drop certain levels of a categorical variable?*. Cross Validated. https://stats.stackexchange.com/questions/141063/is-it-advisable-to-drop-certain-levels-of-a-categorical-variable

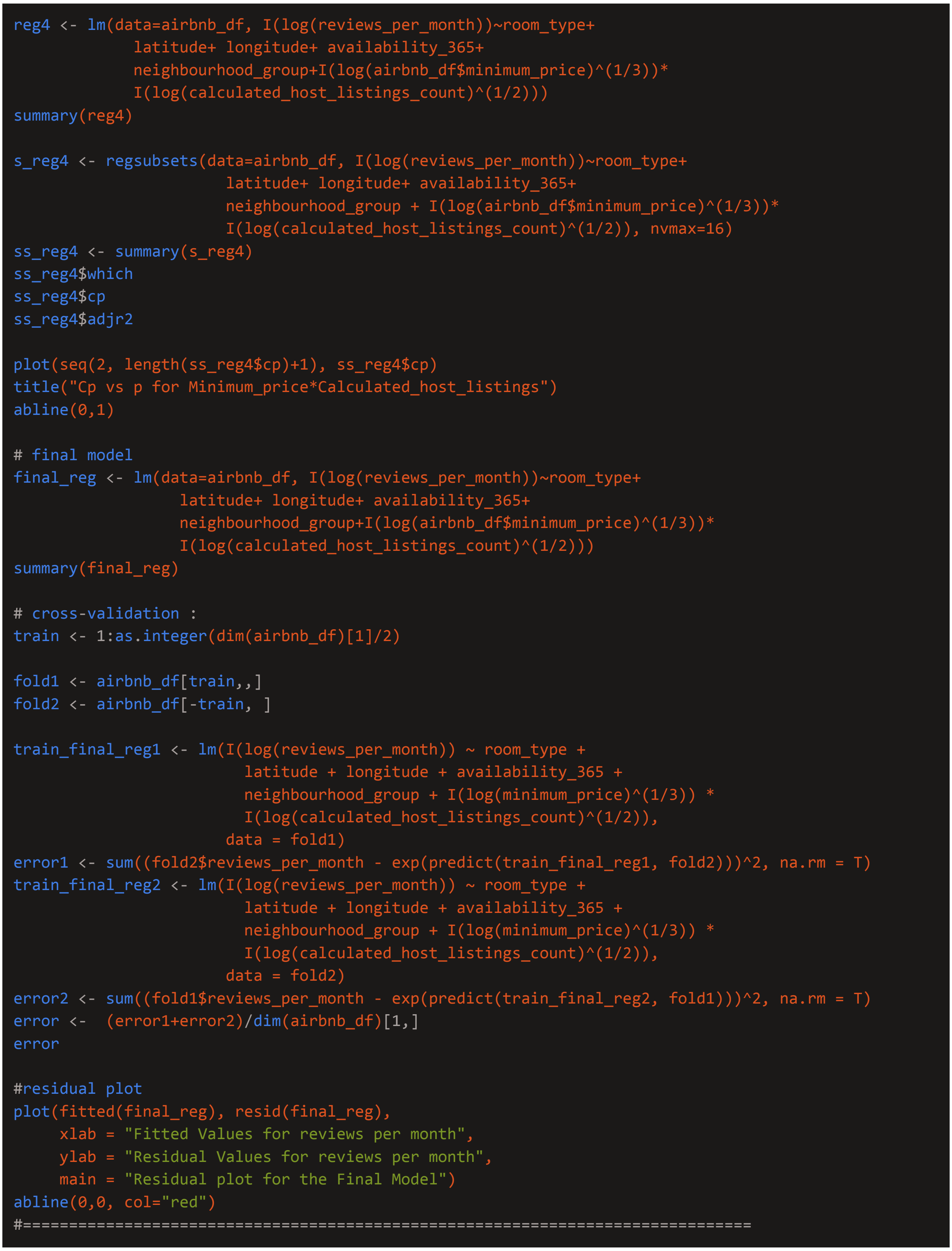
Apendix



```
# Dependencies
library(psych)
library(tidyverse)
library(GGally)
library(leaps)

# EDA
airbnb_df <- read.csv("data/AB_NYC_2019.csv")
# first we drop all extra features
airbnb_df <- subset(airbnb_df, select=c(reviews_per_month, room_type,
                                        latitude, longitude, price, neighbourhood_group,
                                        minimum_nights, availability_365,
                                        calculated_host_listings_count))

# MERGING min_nights and price to make min_price
airbnb_df$minimum_price = airbnb_df$price*airbnb_df$minimum_nights
airbnb_df <- subset(airbnb_df, select=-c(price, minimum_nights))

# to check NA values in all columns
sapply(airbnb_df, function(x) sum(is.na(x)))

airbnb_df <- na.omit(airbnb_df)

ggpairs(subset(airbnb_df, select=-c(room_type)))

# reviews per month dist
hist(airbnb_df$reviews_per_month, xlab = "Reviews Per Month", main = "Distribution of Reviews Per
Month")
hist(airbnb_df$calculated_host_listings_count, xlab = "Calculated Host Listings", main =
"Distribution of Calculated Host Listings")
hist(airbnb_df$minimum_price, xlab = "Minimium Price", main = "Distribution of Minimum Price")

#removing outliers from cal_host listings count
cal_host_list_mean <- mean(airbnb_df$calculated_host_listings_count)
cal_host_list_sd <- sd(airbnb_df$calculated_host_listings_count)
airbnb_df <- subset(airbnb_df, calculated_host_listings_count <= cal_host_list_mean +
1.96*cal_host_list_sd)

#transformed histograms
hist(log(airbnb_df$reviews_per_month),xlab = "Transformed Reviews Per Month (log transformation)",
main = "Distribution of Transformed Reviews Per Month")
hist(log(airbnb_df$minimum_price)^(1/3),xlab = "Transformed Minimum Price (log transformation)",
main = "Distribution of Transformed Minimum Price")
hist(log(airbnb_df$calculated_host_listings_count)^(1/2),xlab = "Transformed Calculated host
listings (log transformation)", main = "Distribution of Transformed Calculated host listings")

#plotting everything against response with log transformation
plot(data=airbnb_df, I(log(reviews_per_month))~latitude, xlab = "Latitude", ylab = "Transformed
Reviews Per Month (log)")
title("Latitude vs Reviews per Month")
plot(data=airbnb_df, I(log(reviews_per_month))~longitude, xlab = "Longitude", ylab = "Transformed
Reviews Per Month (log)")
title("Longitude vs Reviews per Month")
plot(data=airbnb_df, I(log(reviews_per_month))~availability_365, xlab = "Availability (no. of
days)", ylab = "Transformed Reviews Per Month (log)")
title("Availability vs Reviews per Month")
plot(data=airbnb_df, I(log(reviews_per_month))~log(airbnb_df$minimum_price)^(1/3), xlab = "Minimum
Price (in $)", ylab = "Transformed Reviews Per Month (log)")
title("Minimum Price vs Reviews per Month")
plot(data=airbnb_df, I(log(reviews_per_month))~log(airbnb_df$calculated_host_listings_count)^(1/2),
```



```
xlab = "No. of listings per host", ylab = "Transformed Reviews Per Month (log)")
title("Calculated Host Listings vs Reviews per Month")

#training full model no interactions
full_reg <- lm(data=airbnb_df, I(log(reviews_per_month))~room_type+
               latitude+ longitude+ neighbourhood_group+
               availability_365+ I(log(airbnb_df$minimum_price)^(1/3))+
               I(log(calculated_host_listings_count)^(1/2)))
summary(full_reg)

s_reg <- regsubsets(data=airbnb_df, I(log(reviews_per_month))~room_type+
                      latitude+ longitude+ neighbourhood_group+
                      availability_365+ I(log(airbnb_df$minimum_price)^(1/3))+
                      I(log(calculated_host_listings_count)^(1/2)), nvmax=14)
ss_reg <- summary(s_reg)
ss_reg$which
ss_reg$cp
ss_reg$adjr2
plot(seq(2, length(ss_reg$cp)+1), ss_reg$cp)
title("Cp vs p for Full Model (no interaction)")
abline(0,1)

# model with interaction b/w long & lat
reg2 <- lm(data=airbnb_df, I(log(reviews_per_month))~room_type+
             latitude*longitude+ neighbourhood_group+
             availability_365+ I(log(airbnb_df$minimum_price)^(1/3))+
             I(log(calculated_host_listings_count)^(1/2)))
summary(reg2)

s_reg2 <- regsubsets(data=airbnb_df, I(log(reviews_per_month))~room_type+
                       latitude*longitude+ neighbourhood_group+
                       availability_365+ I(log(airbnb_df$minimum_price)^(1/3))+
                       I(log(calculated_host_listings_count)^(1/2)), nvmax=14)
ss_reg2 <- summary(s_reg2)
ss_reg2$which
ss_reg2$cp
ss_reg2$adjr2

plot(seq(2, length(ss_reg2$cp)+1), ss_reg2$cp)
title("Cp vs p for Latitude*Longitude")
abline(0,1)

# model with interaction b/w neighbourhood_group and price
reg3 <- lm(data=airbnb_df, I(log(reviews_per_month))~room_type+
             latitude+ longitude+ availability_365+
             neighbourhood_group*I(log(airbnb_df$minimum_price)^(1/3))+
             I(log(calculated_host_listings_count)^(1/2)))
summary(reg3)

s_reg3 <- regsubsets(data=airbnb_df, I(log(reviews_per_month))~room_type+
                       latitude+ longitude+ availability_365+
                       neighbourhood_group*I(log(airbnb_df$minimum_price)^(1/3))+
                       I(log(calculated_host_listings_count)^(1/2)), nvmax=18)
ss_reg3 <- summary(s_reg3)
ss_reg3$which
ss_reg3$cp
ss_reg3$adjr2

plot(seq(2, length(ss_reg3$cp)+1), ss_reg3$cp)
title("Cp vs p for Neighbourhood_group*Minimum_price")
abline(0,1)

# model with interaction b/w price and calc_host_listing
```



```
reg4 <- lm(data=airbnb_df, I(log(reviews_per_month))~room_type+
             latitude+ longitude+ availability_365+
             neighbourhood_group+I(log(airbnb_df$minimum_price)^(1/3))*
             I(log(calculated_host_listings_count)^(1/2)))
summary(reg4)

s_reg4 <- regsubsets(data=airbnb_df, I(log(reviews_per_month))~room_type+
                       latitude+ longitude+ availability_365+
                       neighbourhood_group + I(log(airbnb_df$minimum_price)^(1/3))*
                       I(log(calculated_host_listings_count)^(1/2)), nvmax=16)
ss_reg4 <- summary(s_reg4)
ss_reg4$which
ss_reg4$cp
ss_reg4$adjr2

plot(seq(2, length(ss_reg4$cp)+1), ss_reg4$cp)
title("Cp vs p for Minimum_price*Calculated_host_listings")
abline(0,1)

# final model
final_reg <- lm(data=airbnb_df, I(log(reviews_per_month))~room_type+
                  latitude+ longitude+ availability_365+
                  neighbourhood_group+I(log(airbnb_df$minimum_price)^(1/3))*
                  I(log(calculated_host_listings_count)^(1/2)))
summary(final_reg)

# cross-validation :
train <- 1:as.integer(dim(airbnb_df)[1]/2)

fold1 <- airbnb_df[train,,]
fold2 <- airbnb_df[-train, ]

train_final_reg1 <- lm(I(log(reviews_per_month)) ~ room_type +
                         latitude + longitude + availability_365 +
                         neighbourhood_group + I(log(minimum_price)^(1/3)) *
                         I(log(calculated_host_listings_count)^(1/2)),
                       data = fold1)
error1 <- sum((fold2$reviews_per_month - exp(predict(train_final_reg1, fold2)))^2, na.rm = T)
train_final_reg2 <- lm(I(log(reviews_per_month)) ~ room_type +
                         latitude + longitude + availability_365 +
                         neighbourhood_group + I(log(minimum_price)^(1/3)) *
                         I(log(calculated_host_listings_count)^(1/2)),
                       data = fold2)
error2 <- sum((fold1$reviews_per_month - exp(predict(train_final_reg2, fold1)))^2, na.rm = T)
error <-  (error1+error2)/dim(airbnb_df)[1,]
error

#residual plot
plot(fitted(final_reg), resid(final_reg),
     xlab = "Fitted Values for reviews per month",
     ylab = "Residual Values for reviews per month",
     main = "Residual plot for the Final Model")
abline(0,0, col="red")
#================================================================================
```