STAT 27410 Final Project Proposal

Brigette Kon and Jake Wei

1. Introduction

Airbnb has become one of the most popular choices for people traveling and seeking housing. However, one problem exists. It is difficult for owners to come up with prices given the location and amenities. It is also hard to predict prices given seasons. Different features such as host ratings, season, location, and number of bedrooms make these two questions immensely complex. In this project, we try to provide insight into predicting prices given different neighborhoods in New York. The data set we picked comes from Inside Airbnb, which is an organization that periodically scrapes data from Airbnb listings. There are 12 columns in the data set. Four are descriptions of the host and 10 can serve as explanatory variables. The explanatory variables are: neighborhood, room_type, accommodation, bed, price, minimum nights, and number of reviews. The response variable is the rent price. Firstly, We will's start with exploratory analysis and data cleaning. Secondly, we'll use the frequentist approach to fit the data and analyze the models. Lastly, we'll describe how we are going to fit the model with respect to a Bayesian approach.

We used R-Markdown to prepare this document.

2. Exploratory Data Cleaning and Data Analysis

To start, we load our New York data from Inside Airbnb, and

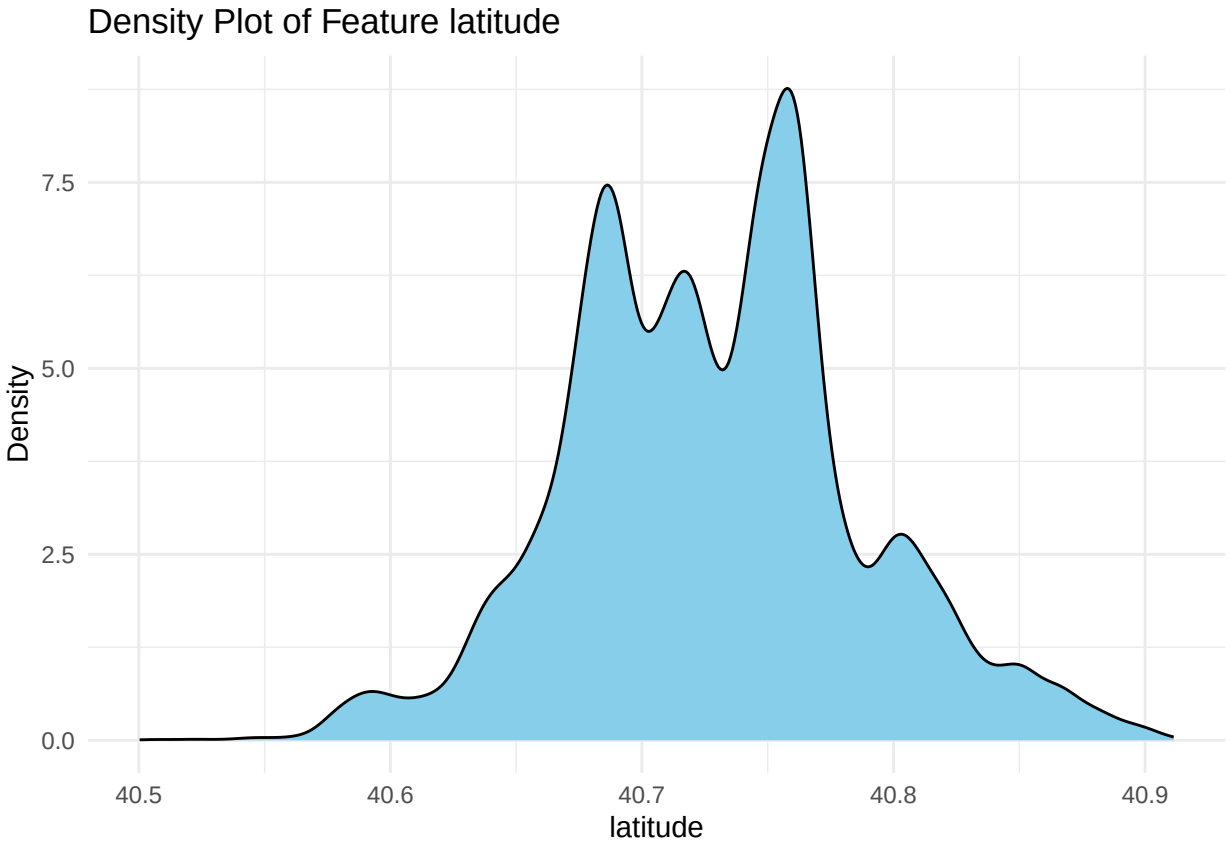
```
data <- read.csv('nydata.csv')

numerical_features <- data[c(3:9,12)]
categorical_features <- data[10:11]
price = summary(as.data.frame(data$price))

stargazer(numerical_features,type='latex')
```

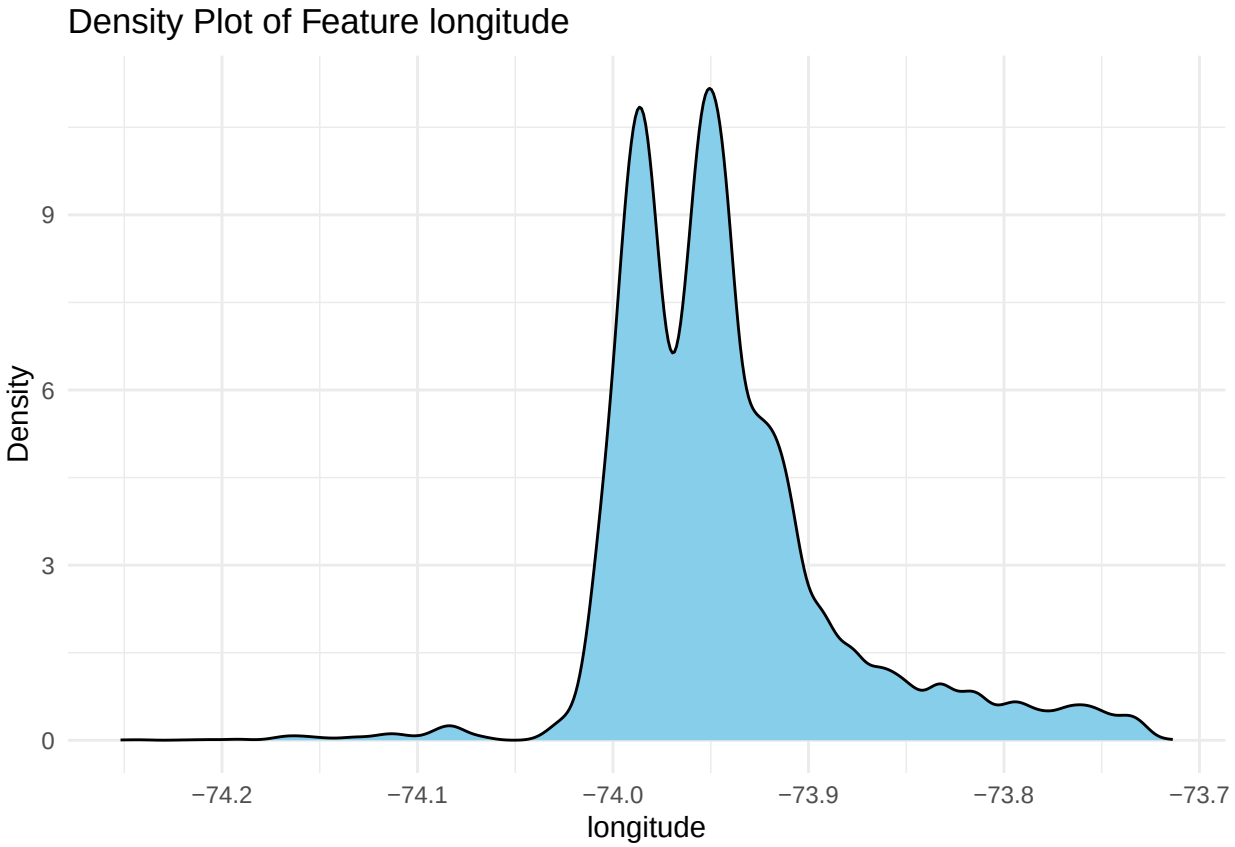
% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Feb 09, 2024 - 17:34:12

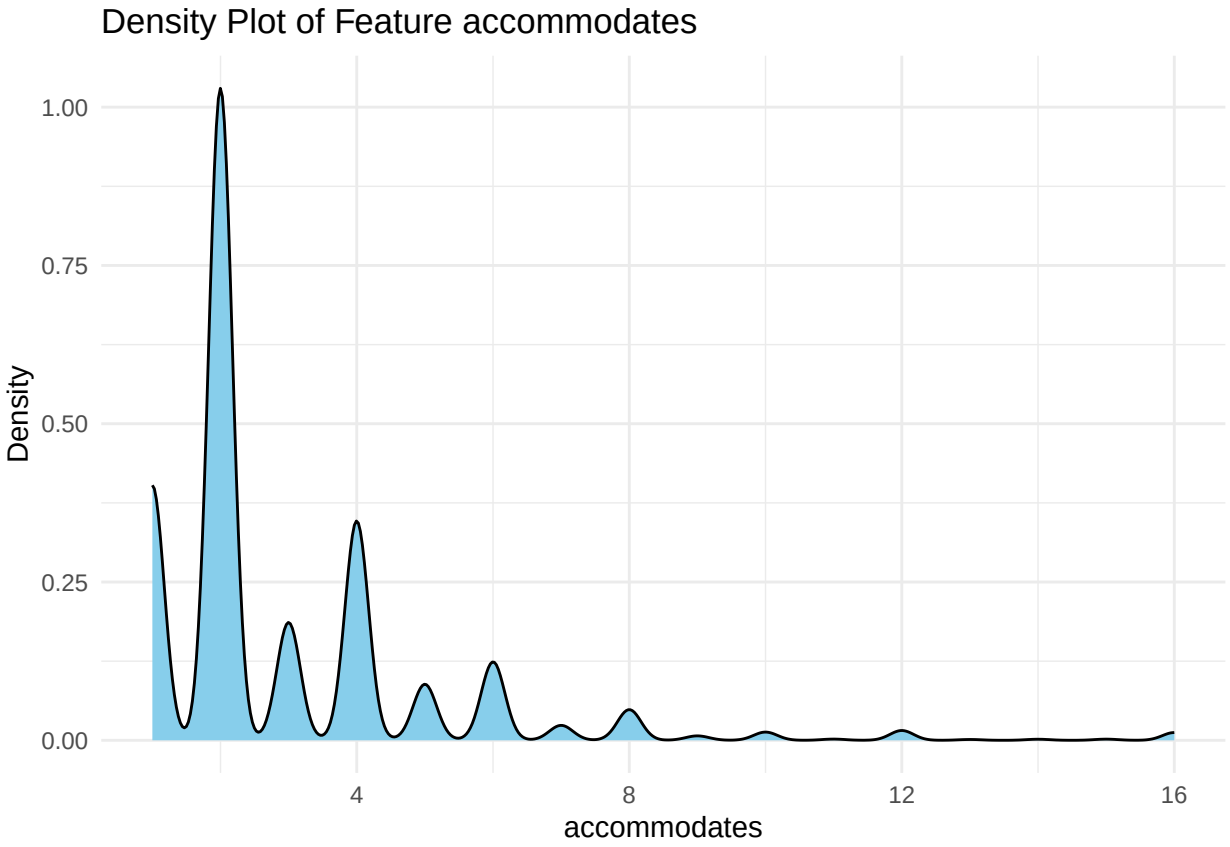
Table 1:

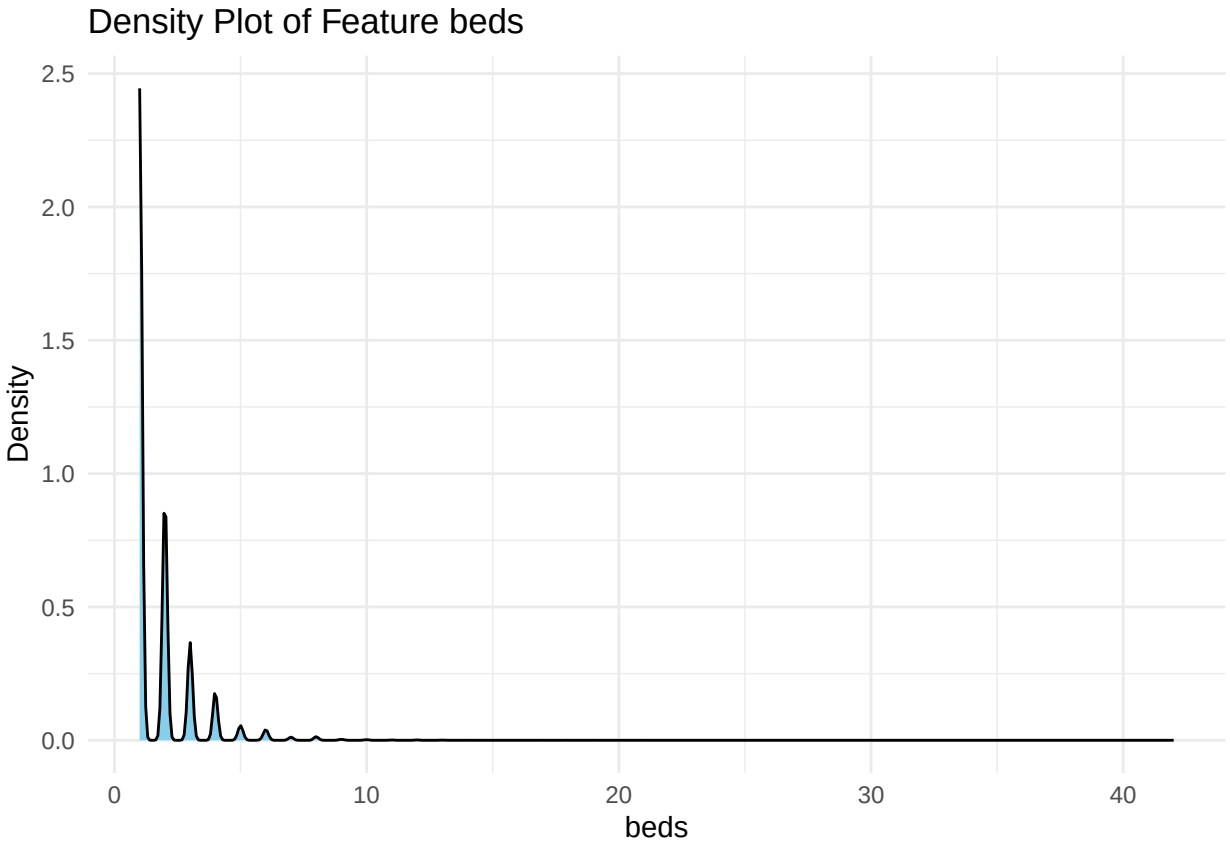
Statistic	N	Mean	St. Dev.	Min	Max
latitude	29,091	40.729	0.058	40.500	40.911
longitude	29,091	−73.943	0.059	−74.252	−73.714
accommodates	29,091	2.953	2.182	1	16
beds	28,363	1.710	1.237	1	42
minimum_nights	29,091	29.578	33.708	1	1,250
number_of_reviews	29,091	30.428	65.061	0	1,865
review_scores_rating	20,768	4.696	0.494	0.000	5.000
price	29,091	212.543	946.727	10	100,000

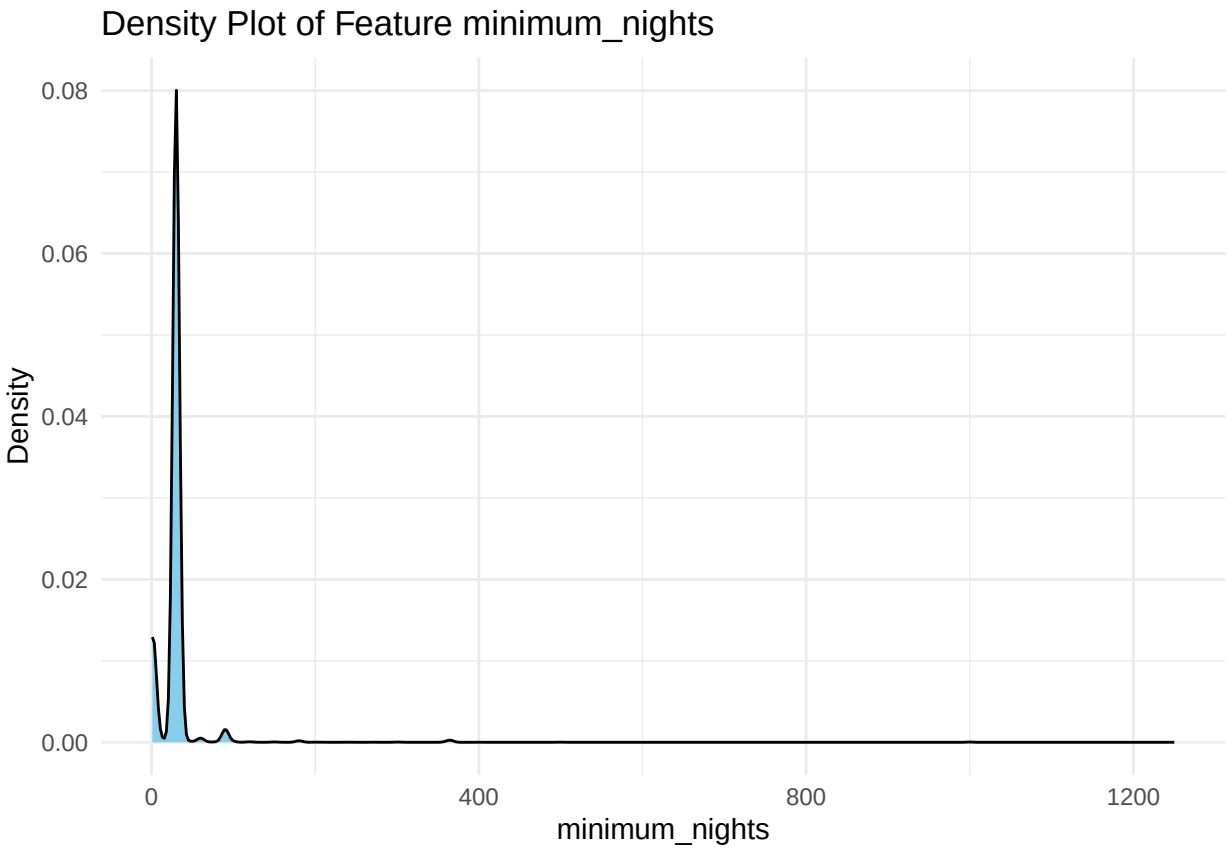
```
##     latitude      longitude      accommodates        beds
##  Min.   :40.5   Min.   :-74.3   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:40.7   1st Qu.:-74.0   1st Qu.: 2.00   1st Qu.: 1.00
##  Median :40.7   Median :-74.0   Median : 2.00   Median : 1.00
##  Mean   :40.7   Mean   :-73.9   Mean   : 2.95   Mean   : 1.71
##  3rd Qu.:40.8   3rd Qu.:-73.9   3rd Qu.: 4.00   3rd Qu.: 2.00
##  Max.   :40.9   Max.   :-73.7   Max.   :16.00   Max.   :42.00
##                                                 NA's   :728
##  minimum_nights   number_of_reviews review_scores_rating     price
##  Min.   :   1.0   Min.   :   0.0    Min.   :0.00         Min.   :    10
##  1st Qu.:  30.0   1st Qu.:   0.0    1st Qu.:4.63         1st Qu.:    79
##  Median :  30.0   Median :   5.0    Median :4.83         Median :   128
##  Mean   :  29.6   Mean   :  30.4    Mean   :4.70         Mean   :   212
##  3rd Qu.:  30.0   3rd Qu.:  30.0    3rd Qu.:5.00         3rd Qu.:   210
##  Max.   :1250.0   Max.   :1865.0    Max.   :5.00         Max.   :100000
##                                     NA's   :8323
```

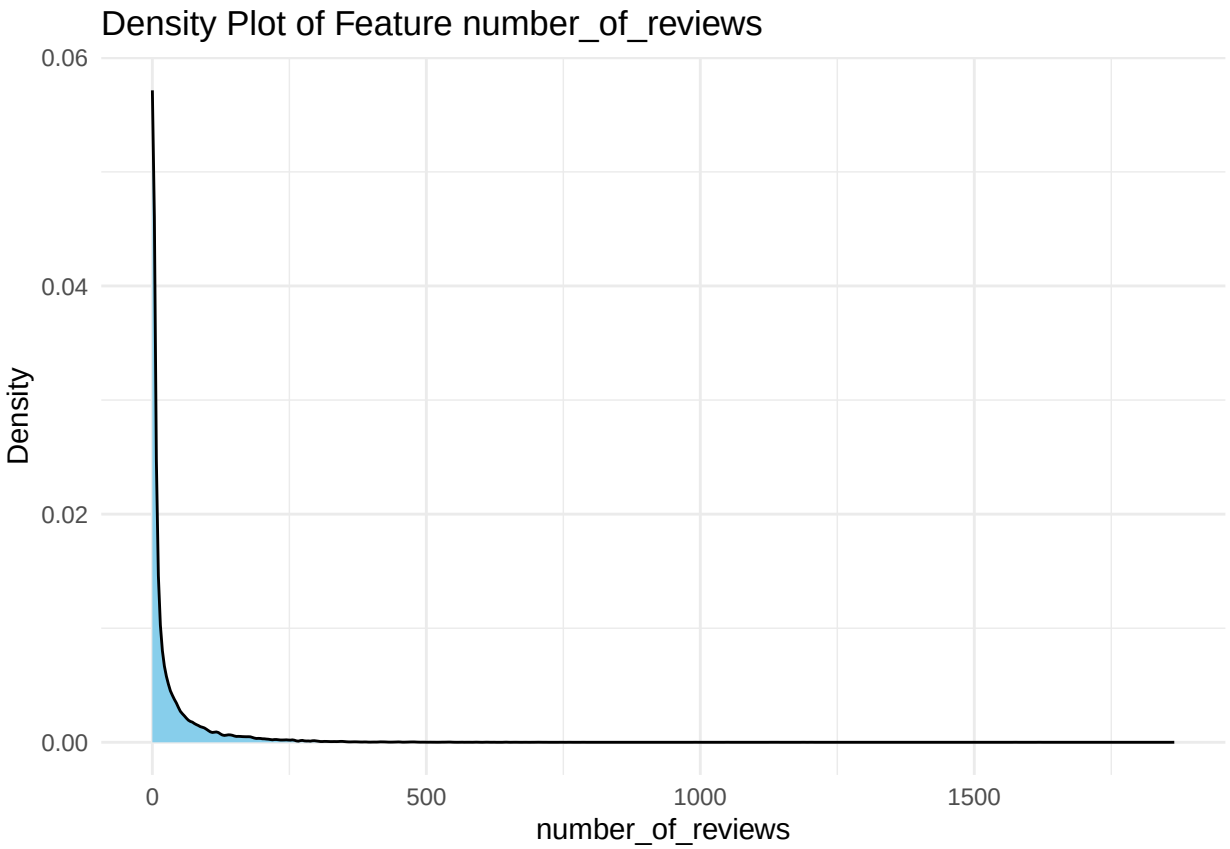

Density Plot of Feature latitude
Density
7.5
5.0
2.5
0.0
40.5
40.6
40.7
40.8
40.9
latitude

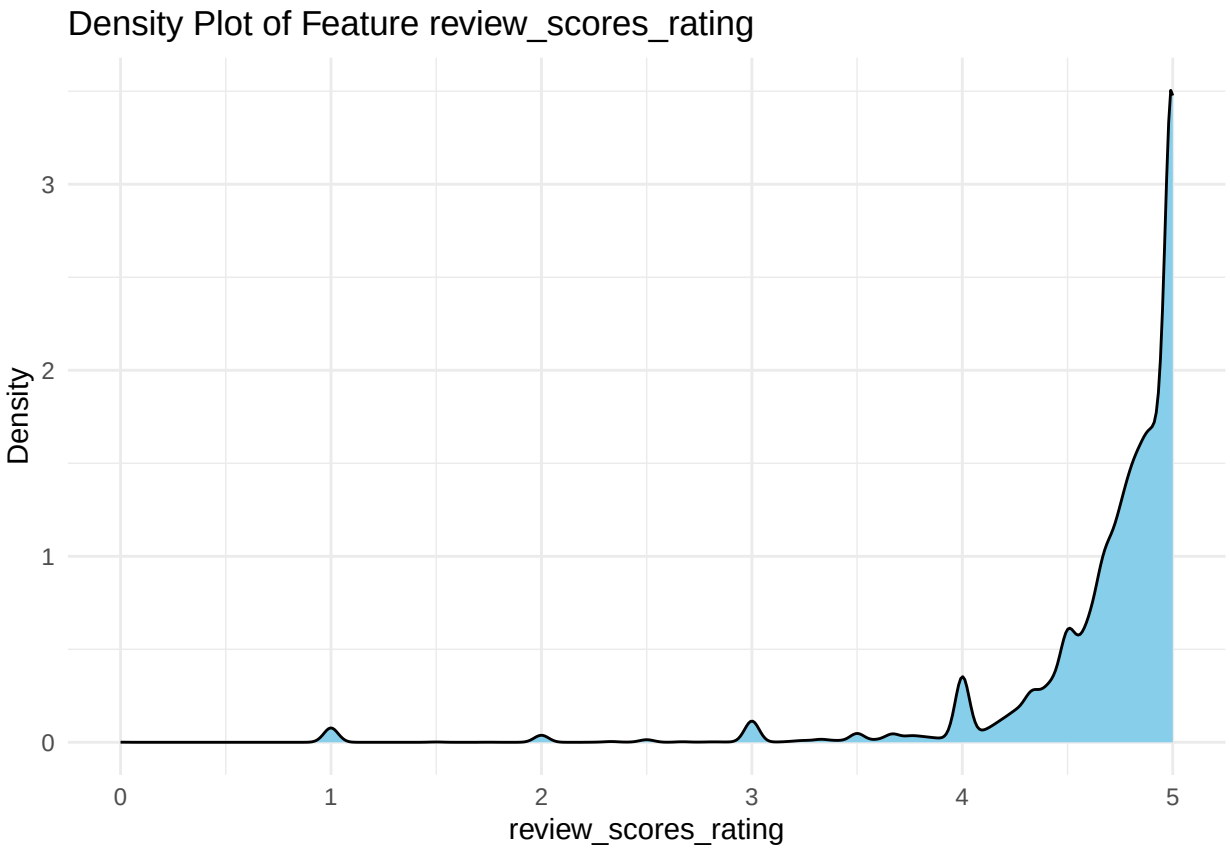

Density Plot of Feature longitude
Density
9
6
3
0
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
longitude

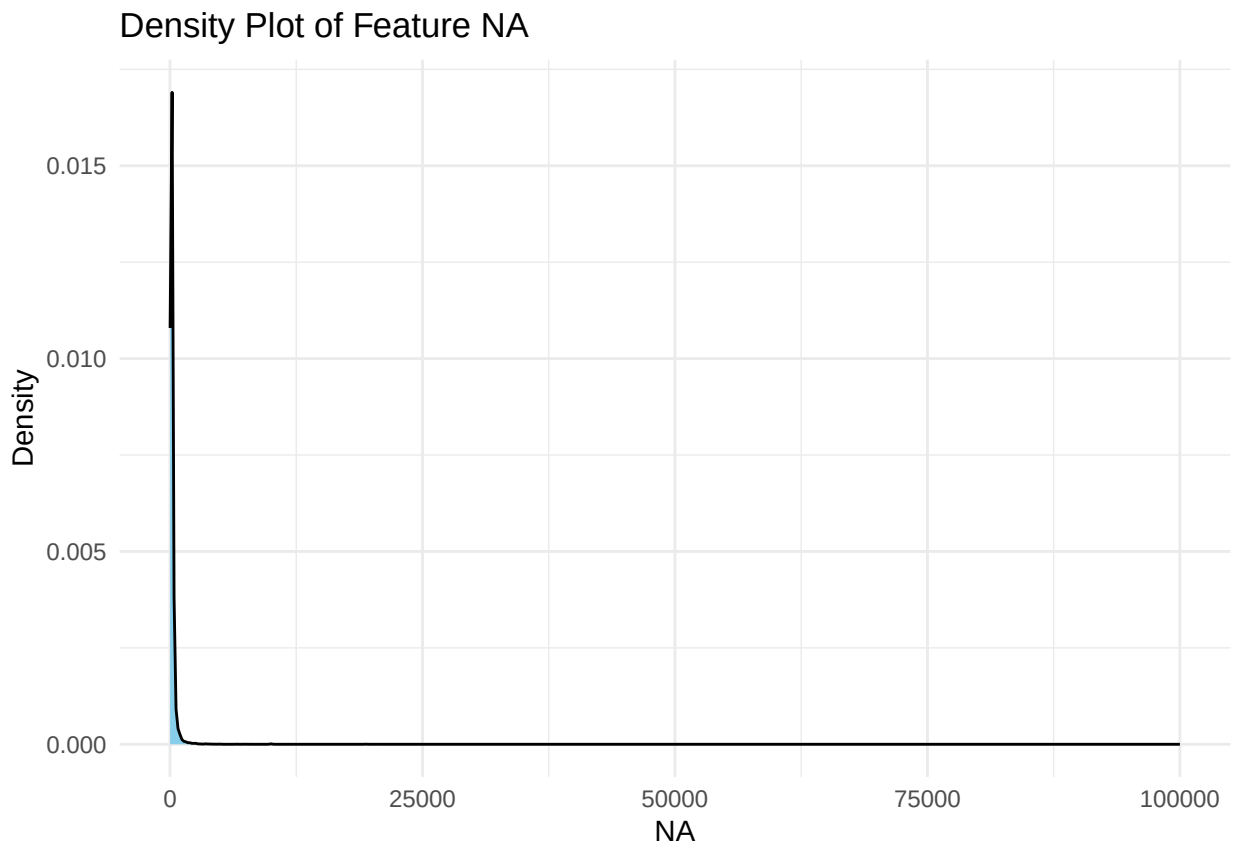

Density Plot of Feature accommodates
Density
1.00
0.75
0.50
0.25
0.00
4
8
12
16
accommodates


Density Plot of Feature beds
Density
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
beds


Density Plot of Feature minimum_nights
0.08
0.06
0.04
0.02
0.00
Density
0
400
800
1200
minimum_nights


Density Plot of Feature number_of_reviews
0.06
0.04
0.02
0.00
Density
0
500
1000
1500
number_of_reviews


Density Plot of Feature review_scores_rating
Density
0
1
2
3
0
1
2
3
4
5
review_scores_rating

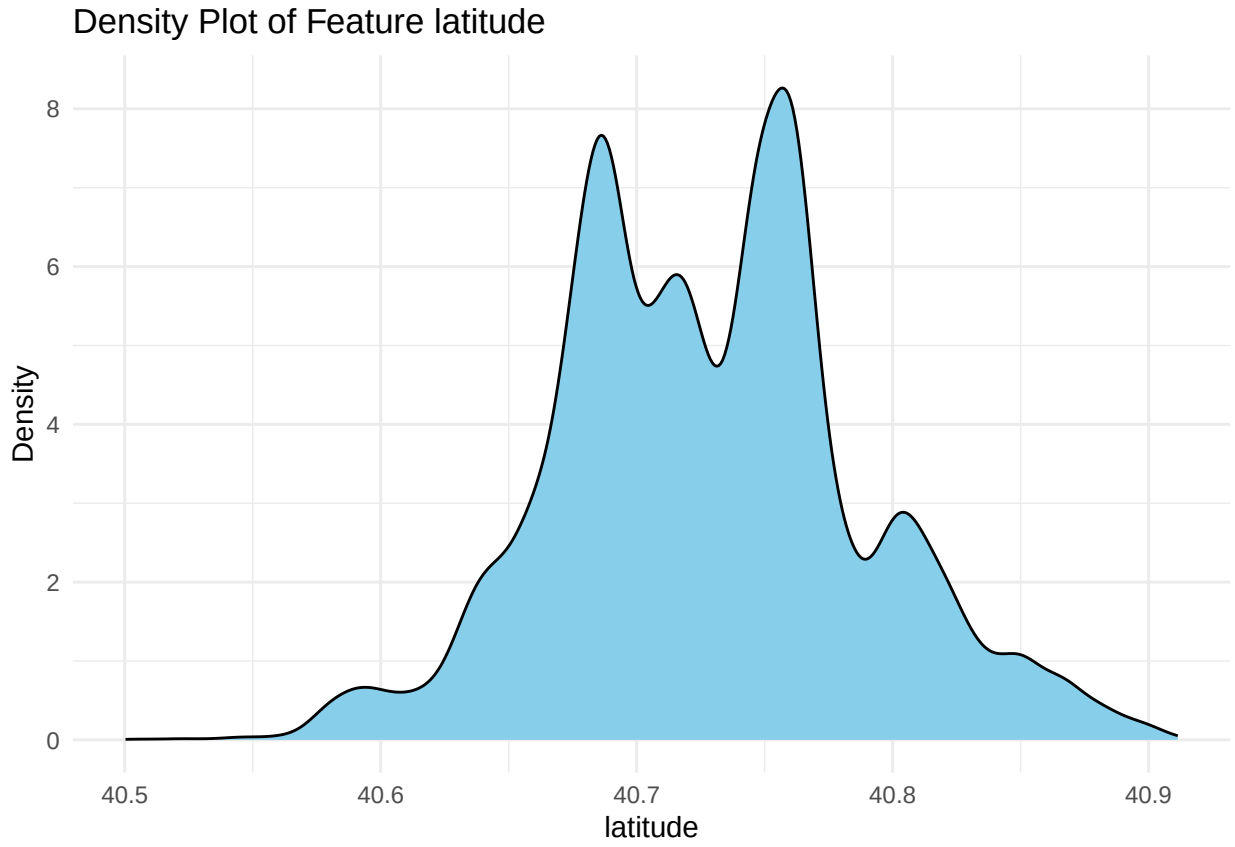


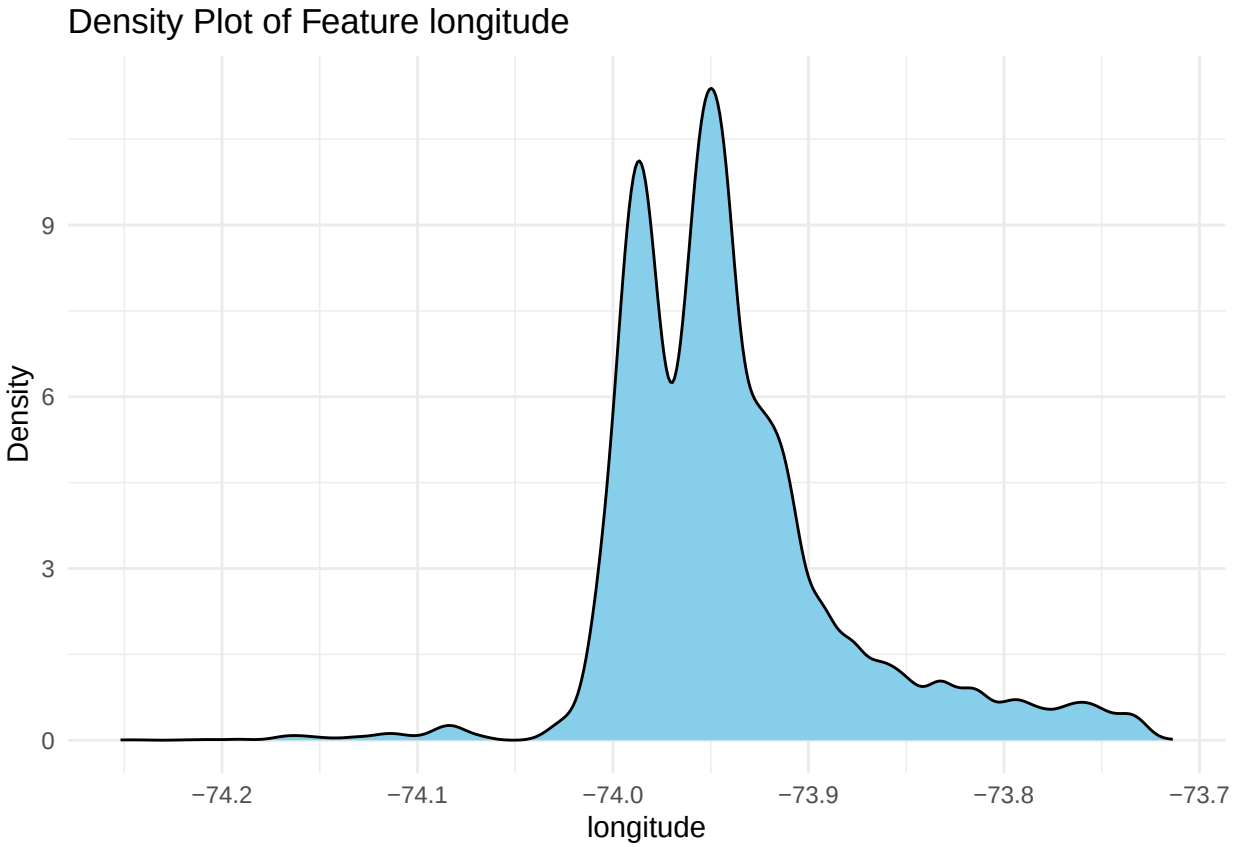
From this initial data plotting, we realized that there data points that don't make sense. Some of the variables are poorly distributed For example the max price of 10000. In the next step, we will attempt to remove these outliers. Judging from IQR of each variables, we can see that majority of airbnbs in New York is below $500 and less than 30 nights. We will create a new dataframe from this.

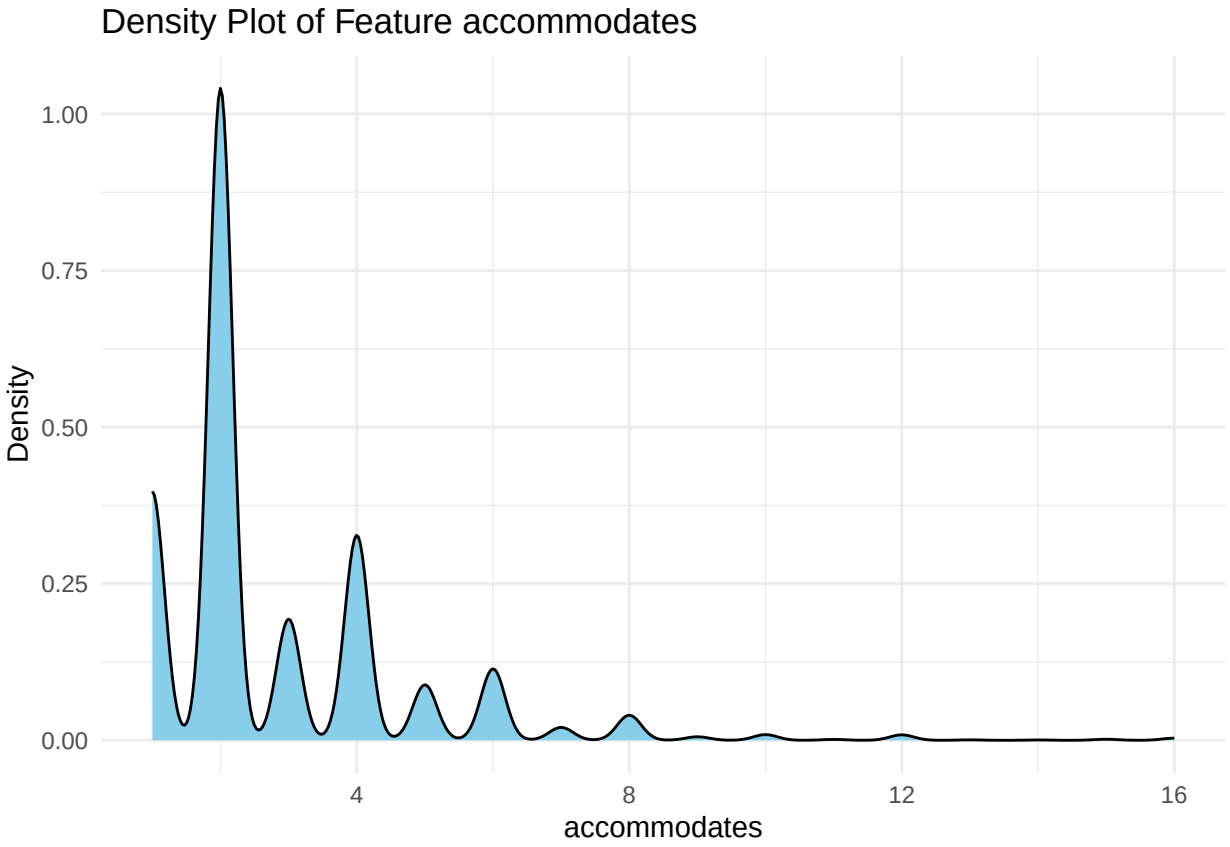
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10      75     120     146     190     500
```

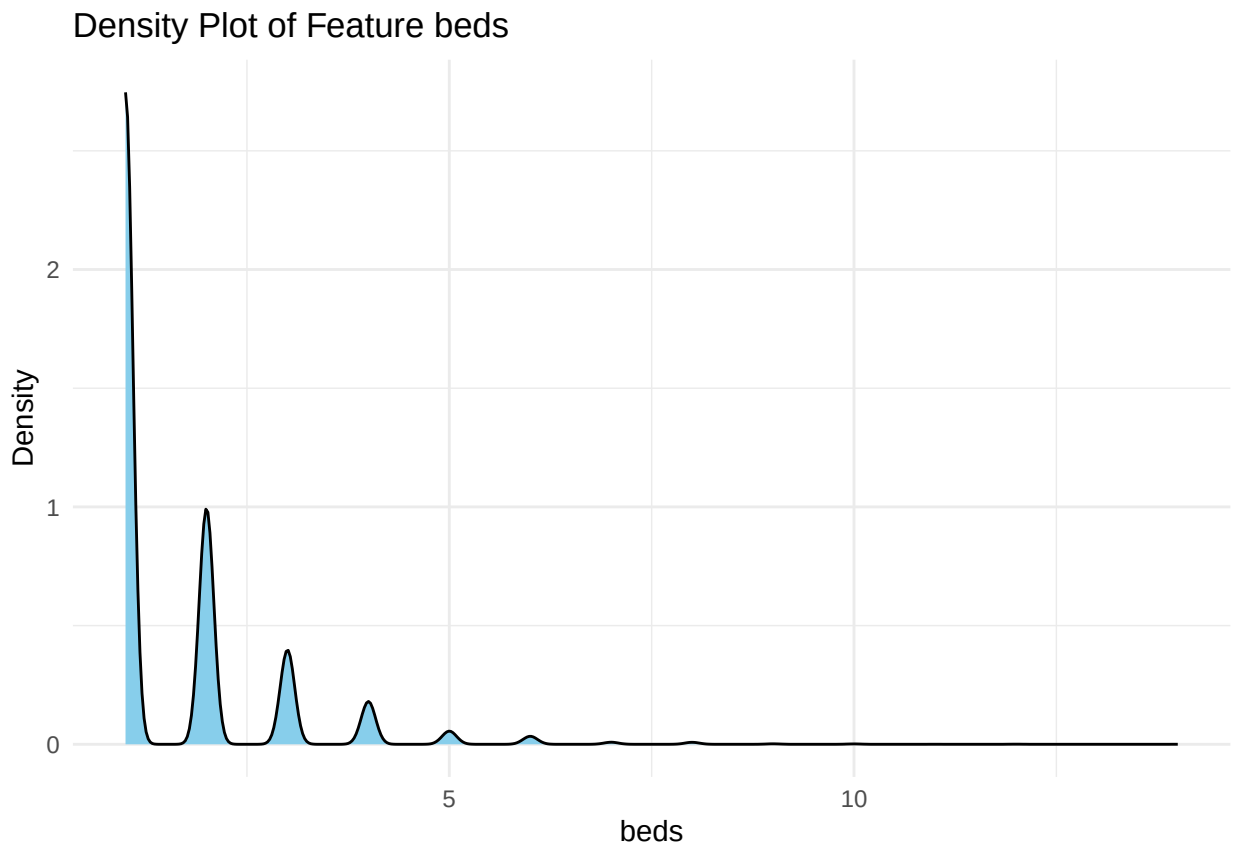
```
##     latitude      longitude      accommodates        beds       minimum_nights
##  Min.   :40.5   Min.   :-74.3   Min.   : 1.00   Min.   : 1.00   Min.   : 1.0
##  1st Qu.:40.7   1st Qu.:-74.0   1st Qu.: 2.00   1st Qu.: 1.00   1st Qu.:30.0
##  Median :40.7   Median :-73.9   Median : 2.00   Median : 1.00   Median :30.0
##  Mean   :40.7   Mean   :-73.9   Mean   : 2.81   Mean   : 1.65   Mean   :26.1
##  3rd Qu.:40.8   3rd Qu.:-73.9   3rd Qu.: 4.00   3rd Qu.: 2.00   3rd Qu.:30.0
##  Max.   :40.9   Max.   :-73.7   Max.   :16.00   Max.   :14.00   Max.   :30.0
##                                                 NA's   :657
##  number_of_reviews review_scores_rating
##  Min.   :   0      Min.   :0.00
##  1st Qu.:   1      1st Qu.:4.63
##  Median :   7      Median :4.83
```

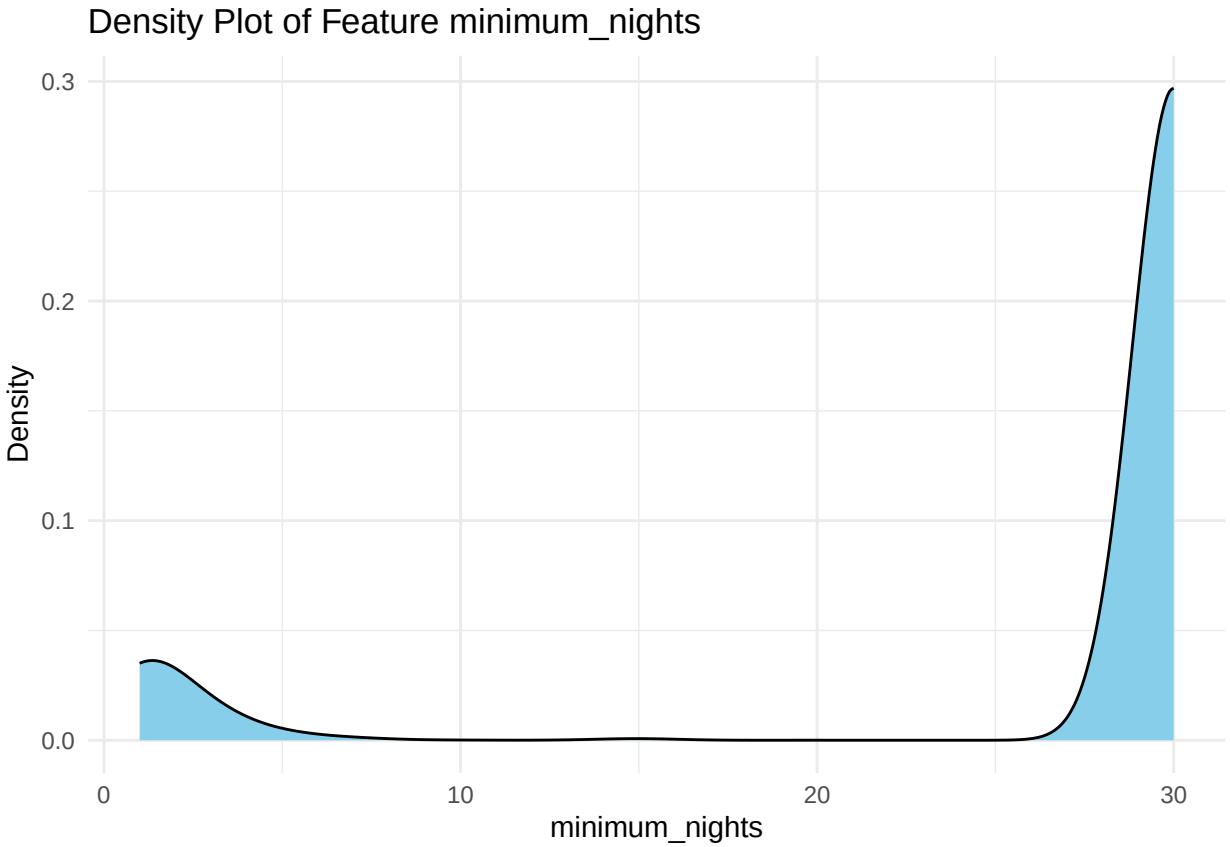
```
##  Mean   :  33      Mean   :4.69
##  3rd Qu.:  35      3rd Qu.:5.00
##  Max.   :1865      Max.   :5.00
##                    NA's   :6370
```

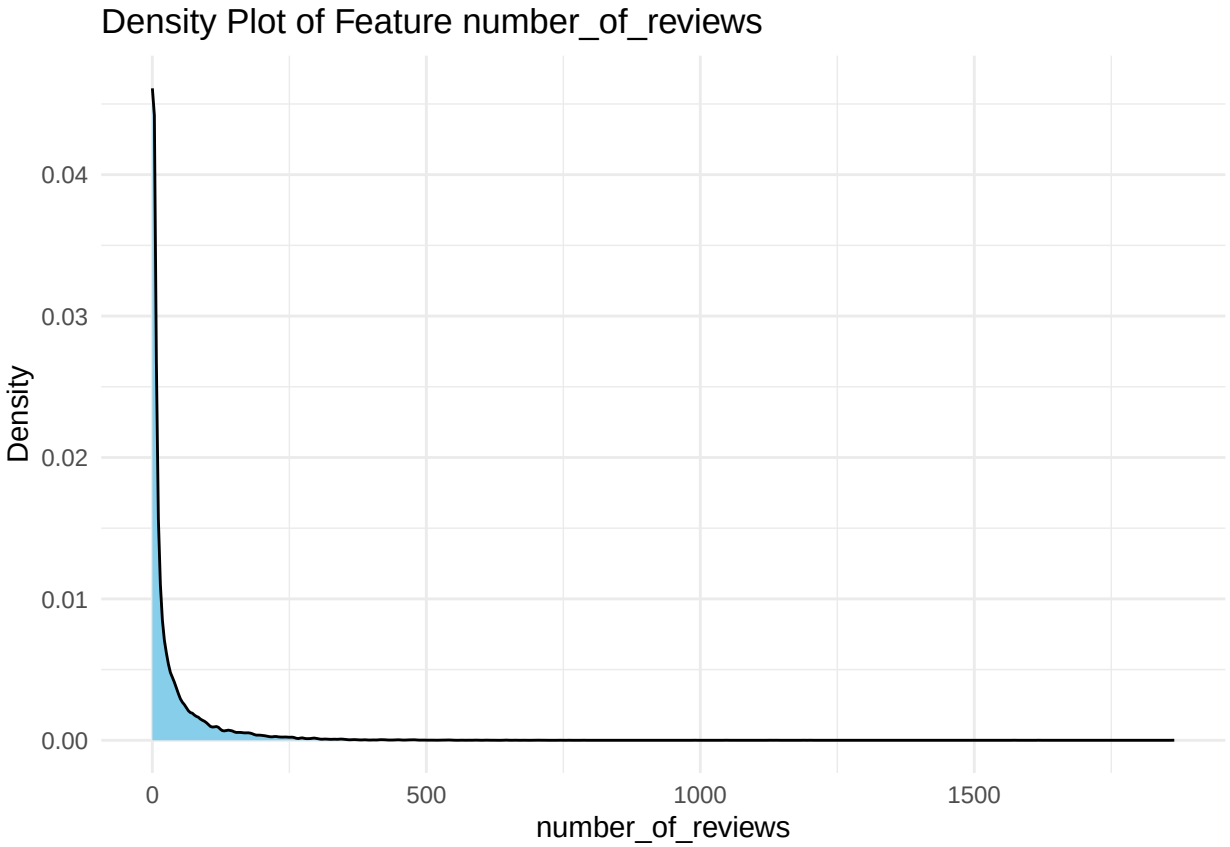


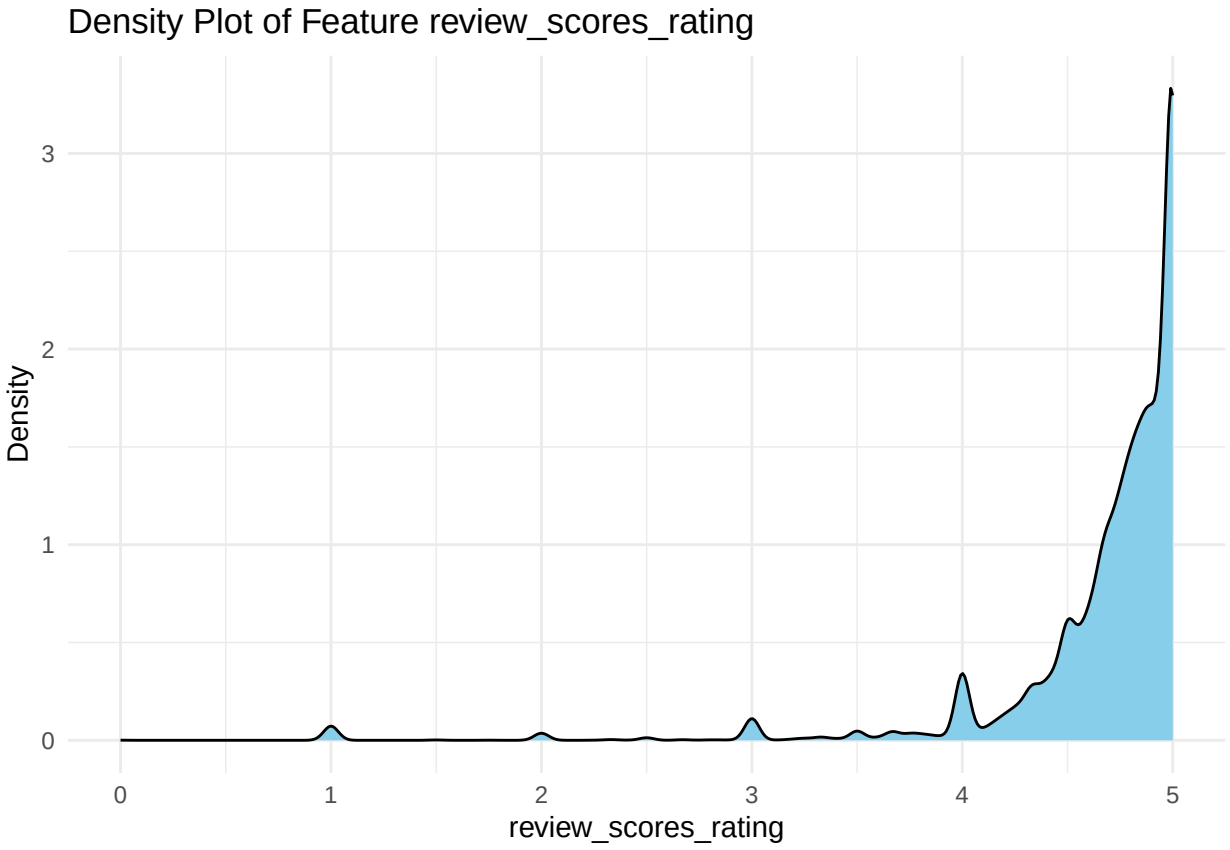

Density Plot of Feature longitude
Density
9
6
3
0
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
longitude

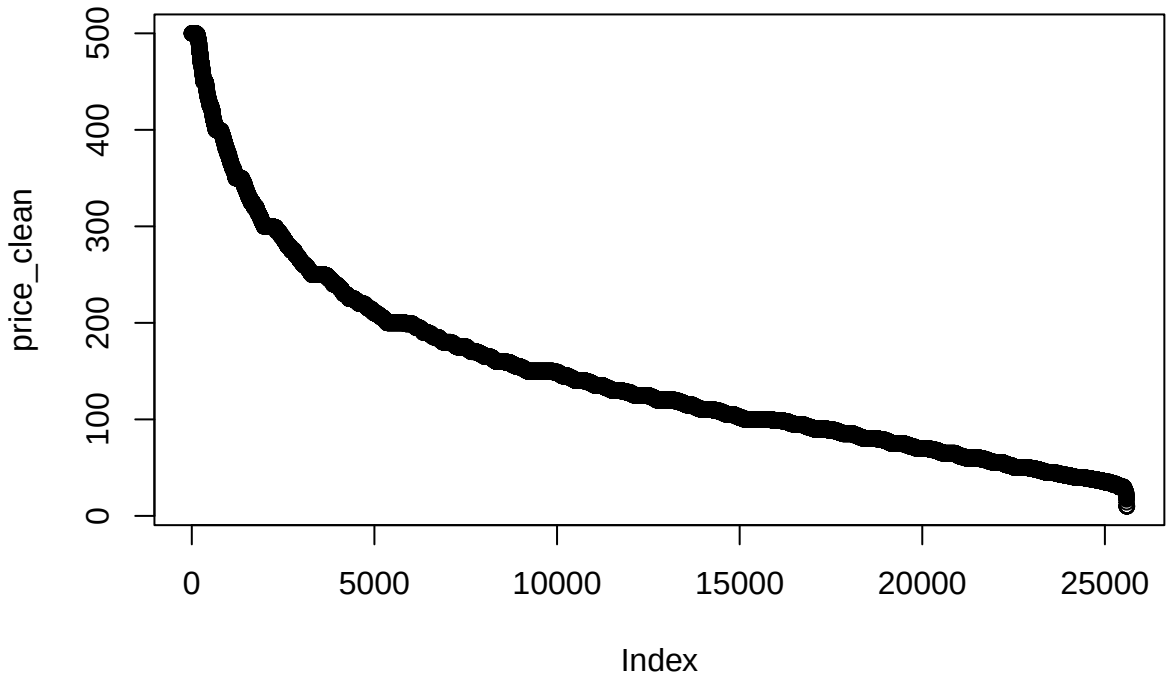

Density Plot of Feature accommodates
Density
1.00
0.75
0.50
0.25
0.00
4
8
12
16
accommodates


Density Plot of Feature beds
Density
2
1
0
5
10
beds

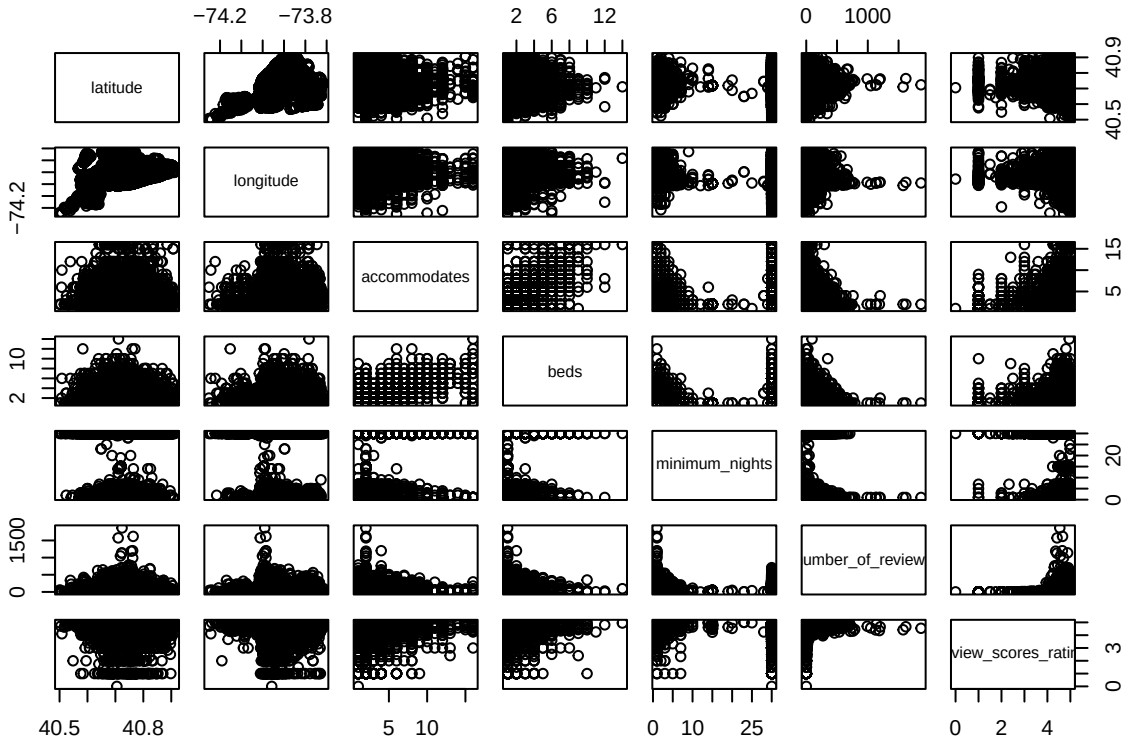

Density Plot of Feature minimum_nights
0.3
0.2
0.1
0.0
Density
0
10
20
30
minimum_nights


Density Plot of Feature number_of_reviews
Density
0.04
0.03
0.02
0.01
0.00
0
500
1000
1500
number_of_reviews

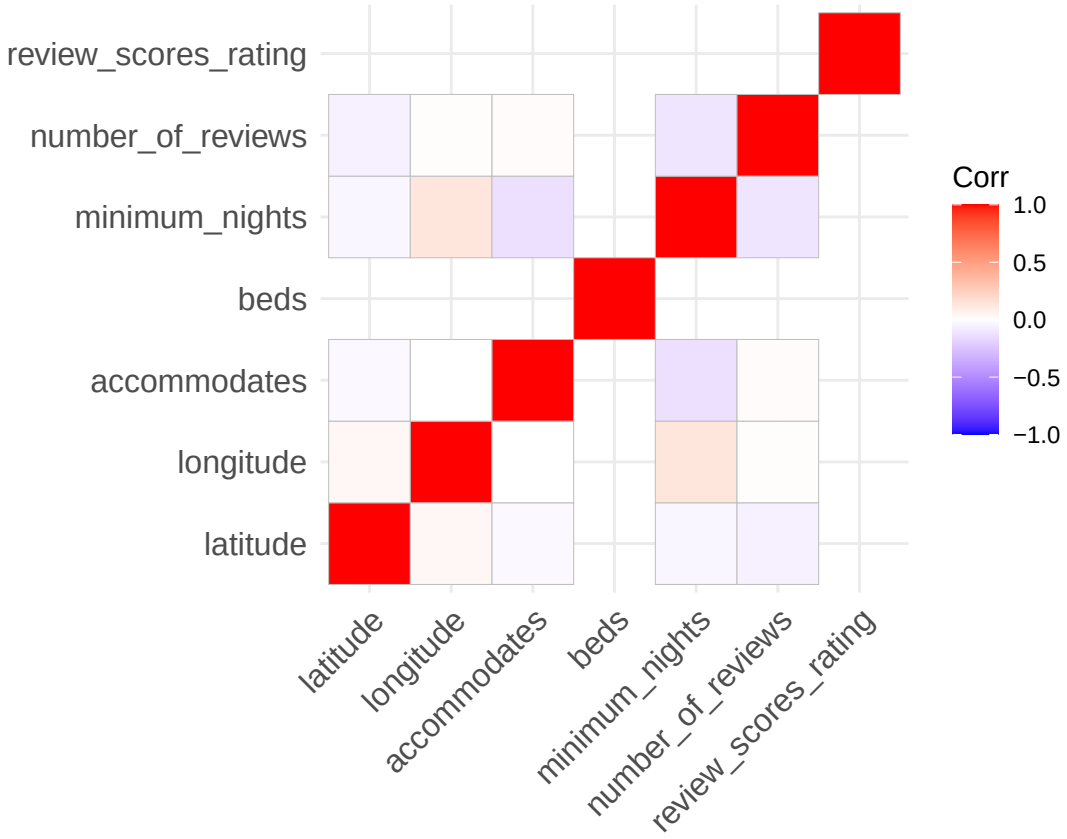

Density Plot of Feature review_scores_rating
Density
0
1
2
3
0
1
2
3
4
5
review_scores_rating



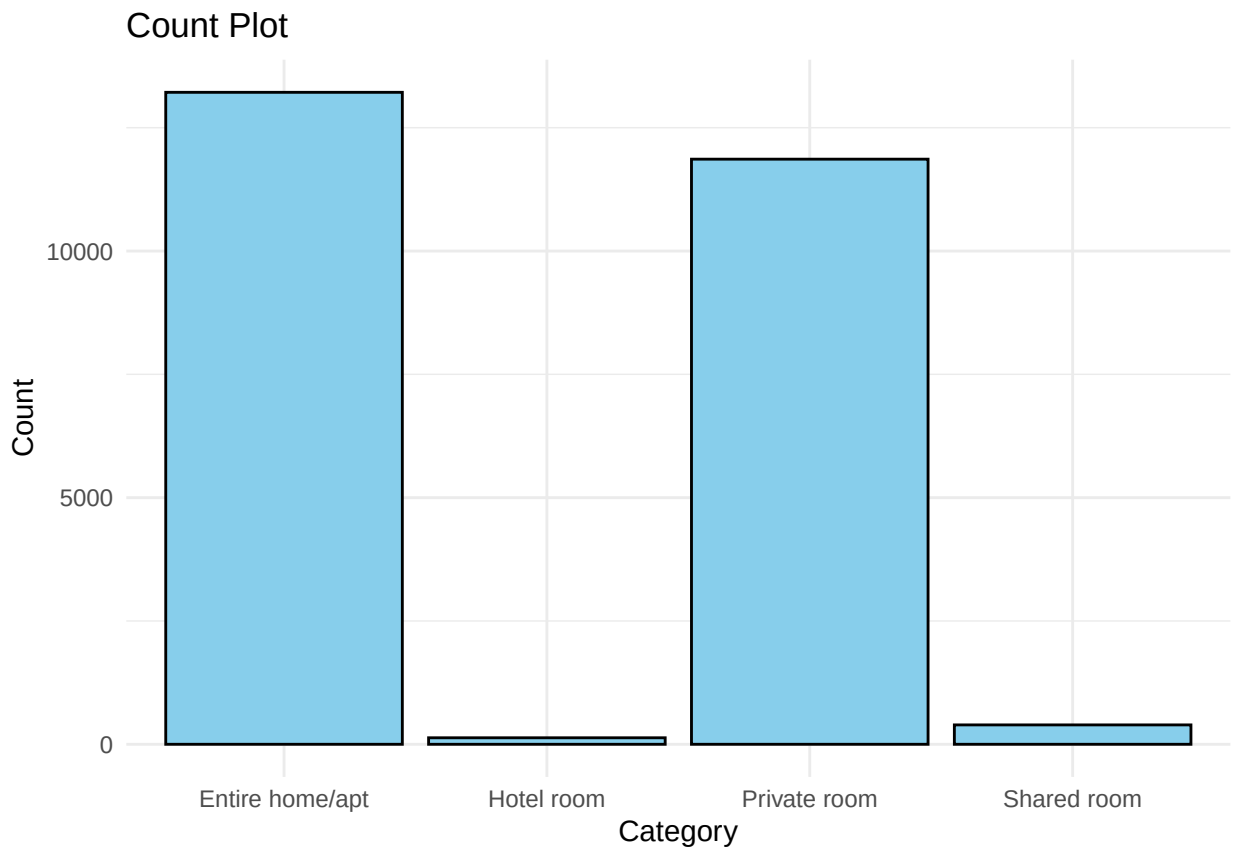
For the features now, it seems a lot more realistic after we removed outliers.




```
##                          latitude    longitude accommodates beds minimum_nights
## latitude               1.0000000  0.041978829 -0.030393082   NA     -0.0441245
## longitude              0.0419788  1.000000000 -0.000308472   NA      0.1301885
## accommodates          -0.0303931 -0.000308472  1.000000000   NA     -0.1255540
## beds                          NA           NA           NA    1             NA
## minimum_nights        -0.0441245  0.130188542 -0.125554031   NA      1.0000000
## number_of_reviews     -0.0591530  0.012256735  0.021937858   NA     -0.1091235
## review_scores_rating          NA           NA           NA   NA             NA
##                       number_of_reviews review_scores_rating
## latitude                      -0.0591530                   NA
## longitude                      0.0122567                   NA
## accommodates                   0.0219379                   NA
## beds                                  NA                   NA
## minimum_nights                -0.1091235                   NA
## number_of_reviews              1.0000000                   NA
## review_scores_rating                  NA                    1
```



we check for correlation between the variables. There seem to be relationship between number of reviews and minimum nights but because the correlation doesn't seem be a strong one.




```
## # A tibble: 221 x 6
##    neighbourhood_cleansed     mean_price median_price min_price max_price
##    <chr>                           <dbl>        <dbl>     <dbl>     <dbl>
##  1 Allerton                         118.         94          31       500
##  2 Arden Heights                    134.        134.        100       166
##  3 Arrochar                         138.         97.5        75       350
##  4 Arverne                          135.        108          33       375
##  5 Astoria                          112.         94.5        25       500
##  6 Bath Beach                       148.        110          40       399
##  7 Battery Park City                186.        180.         10       360
##  8 Bay Ridge                        107.         85          40       350
##  9 Bay Terrace                      150.        124.         99       250
## 10 Bay Terrace, Staten Island       175         175         175       175
## # i 211 more rows
## # i 1 more variable: total_listings <int>
```

3. Frequentist Analysis

3.1 Proposed Frequentist Model(s)

We will be using Multiple Linear Regression to analyze our dataset.

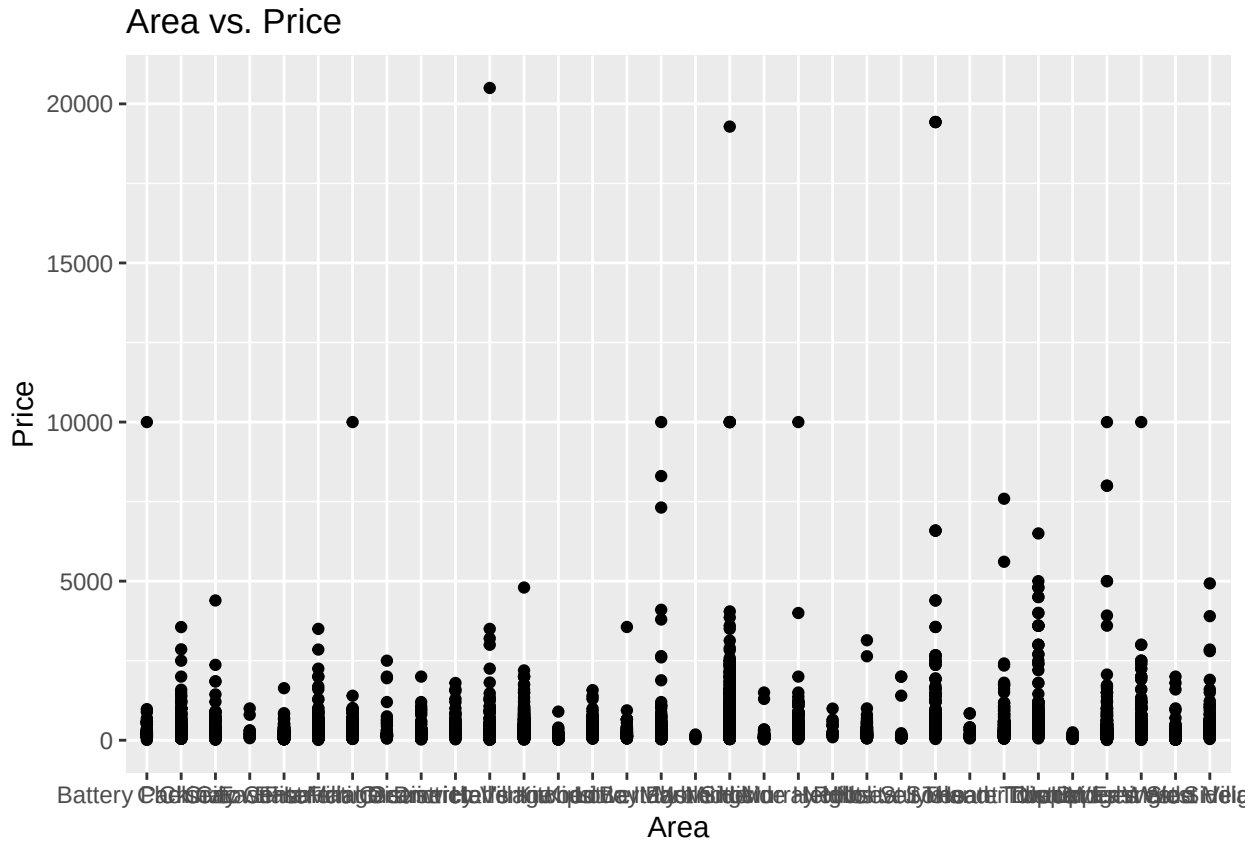
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon$$

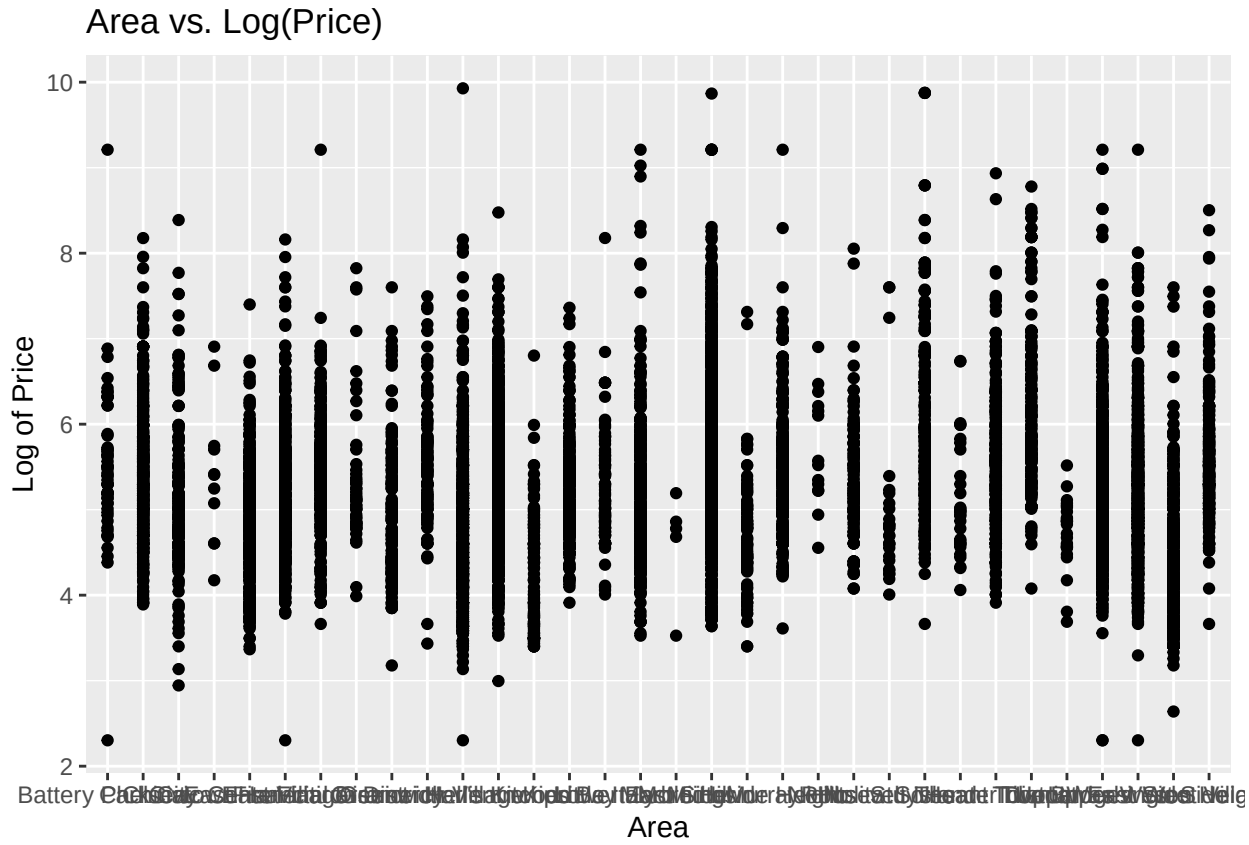
* Y is the predicted rent price * β_0 is the intercept * $\beta_1, \ldots, \beta_n$ are the coefficients for the respective X_n independent variables * ϵ is the error term

The variables that we included in our model were: * "area", the New York neighborhood of the Airbnb listing * "accommodates", the number of individuals the Airbnb can accommodate * "beds", the number of beds in the Airbnb * "bathrooms" * "min_nights", the minimum number of nights required to book the Airbnb * "room_type", whether the Airbnb listing was a private room, shared room, hotel room, or the entire home/apartment

3.2 Fitting the Frequentist Model(s)

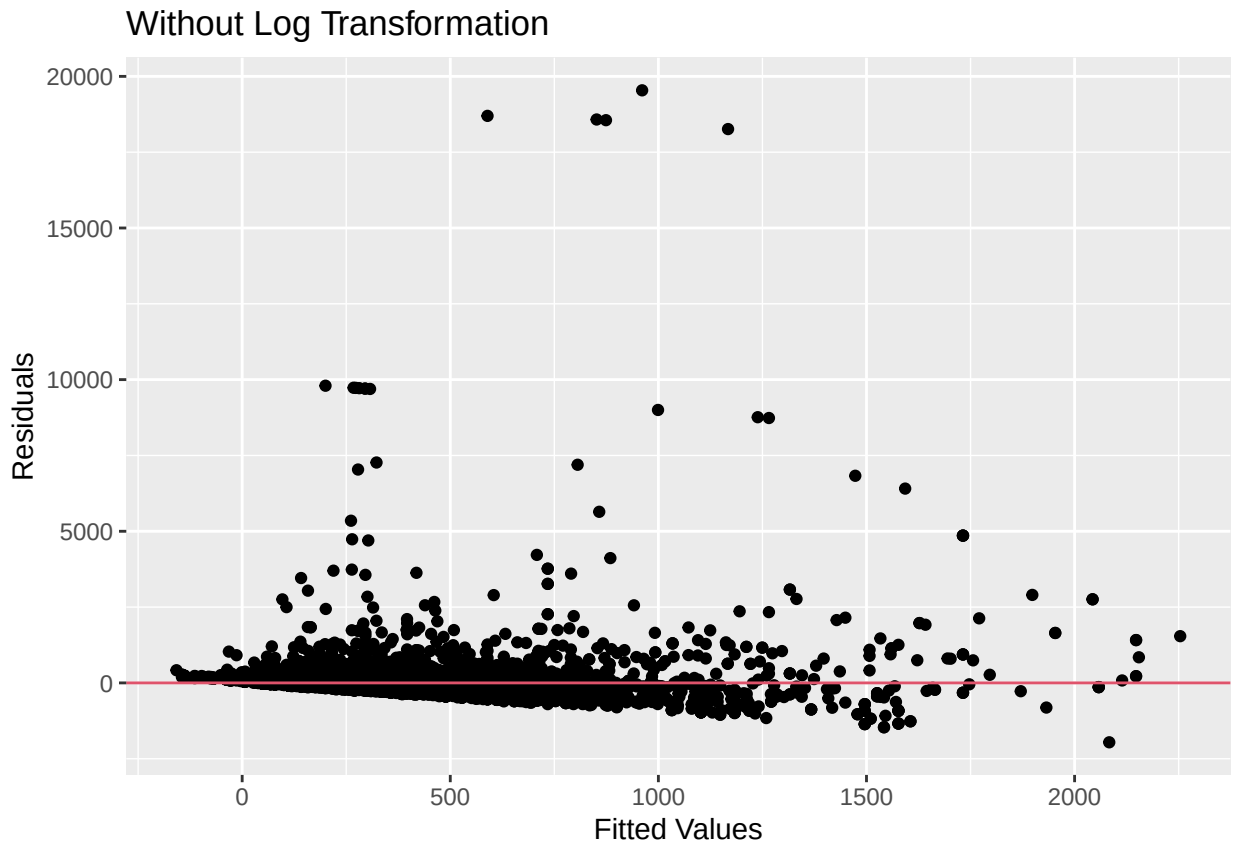
We fit the model by first testing if the model needed a log transformation. Thus, we graphed area vs. Price and area vs. log(Price).

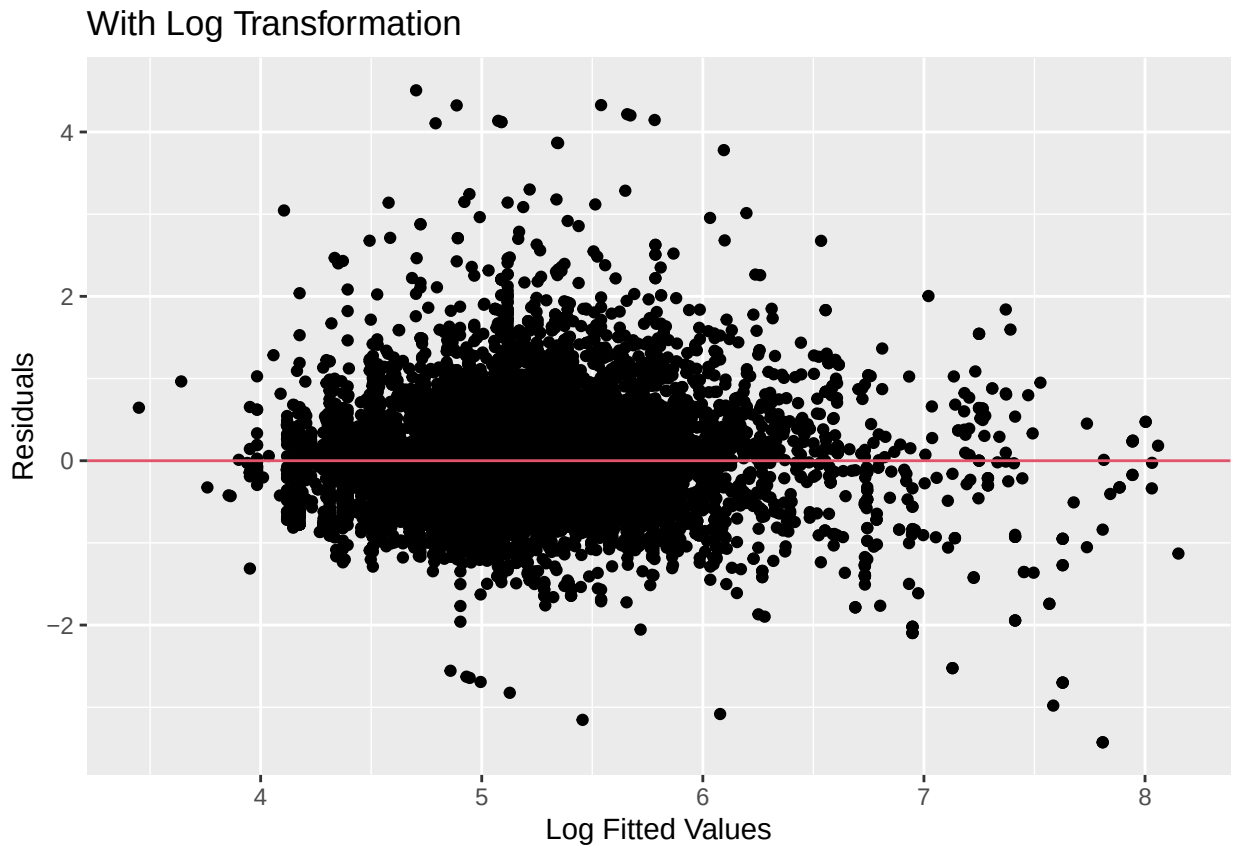




From the graphs above, we can see that after performing a log transformation on Price, the data becomes more linear and variance becomes more constant.

This is further demonstrated if we graph the fitted values against the residuals.


Without Log Transformation
20000
15000
10000
5000
0
Residuals
0
500
1000
1500
2000
Fitted Values



Thus, we decided to go with log(Price) in fitting our final model.

We included interaction terms between *area* and all other independent variables in the fitted model, and took the log of *accommodates* and *min_nights*. Looking at the R-squared values, we can see a general comparison between the model without a log transformation (lm24), the model with a log transformation (log24), and the final model we arrived at with a log transformation and interaction terms (lm1).

Model	R^2
lm24	0.148
log24	0.416
lm1	0.506

Given that there are 32 neighborhoods in our sample, we will not display the final equation of the fitted model as there are 288 coefficients, but we will briefly explain our results below.

In our final fitted model, the intercept $\beta_0 = 4.328$, which means if the Airbnb could accommodate 0 people, had 0 beds and bathrooms, required 0 minimum nights to book, was the entire home or apartment, and was located in Battery Park, the estimated rent price of the Airbnb is \$75.78. We also found that the rent price of the Airbnb increases when it can accommodate more people, has more bathrooms, or is a private room and decreases the

more minimum nights are required to book, is a hotel room or shared room, or has more beds.

As for the neighborhoods, we found that, compared to Battery Park, the neighborhoods that increased the rent price of an Airbnb the greatest were Civic Center, Inwood, Morningside Heights, and Roosevelt Island, and the neighborhoods that decreased the rent price the greatest were Flatiron District, Grammercy, and Stuyvesant Town.

4. Bayesian Analysis

4.1 Proposed Bayesian Model(s)

We do not know exactly what Bayesian model(s) we are going to use to analyze our dataset since we do not know how to handle multiple independent variables. As for obtaining the prior, we have data from 2023 that we can use in order to obtain a prior for our analysis.

4.2 Fitting the Bayesian model(s)

- Propose how you will fit the proposed Bayesian models.
- Propose how you will perform sensitivity analysis of the Bayesian models, i.e., how the posterior distribution is affected by the prior
- Propose how you will check the MCMC convergence.

4.3 Prediction

In this section, propose how you can make predictions using the Bayesian model.

5. Discussion

We can improve our model by narrowing down the neighborhoods where we do not have enough information in order to make reliable predictions and remove it from our model. While this means that we will have to exclude predictions for that area, that could potentially help in obtaining a better fitted model.

6. Contributions

Brigette and Jake contribtued to the project equally, with Jake doing sections 1 and 2, and Brigette doing sections 3, 4, and 5. We worked on the proposal for 10+ hours.

References

Appendix

From Section 2:

```
data <- read.csv('nydata.csv')

numerical_features <- data[3:9]
categorical_features <- data[10:11]
price <- data$price
summary(price)
summary(numerical_features)

features <- c("latitude", "longitude", "accommodates", "beds", "minimum_nights", "number

for (i in 1:ncol(numerical_features)) {
  p <- ggplot(numerical_features, aes(x = numerical_features[[i]])) +
    geom_density(fill = "skyblue", color = "black") +
    labs(title = paste("Density Plot of Feature", features[i]), x = features[i], y = "De
    theme_minimal()

  plot(p)  # Print each plot
}
```

```
#clean data
data_clean = subset(data,price <= 500 & minimum_nights <= 30)
numerical_features_clean <- data_clean[3:9]
categorical_features_clean <- data_clean[10:11]
price_clean <- data_clean$price

summary(price_clean)
summary(numerical_features_clean)

#graph again
for (i in 1:ncol(numerical_features_clean)) {
  p <- ggplot(numerical_features_clean, aes(x = numerical_features_clean[[i]])) +
    geom_density(fill = "skyblue", color = "black") +
    labs(title = paste("Density Plot of Feature", features[i]), x = features[i], y = "De
    theme_minimal()

  print(p)  # Print each plot
}
```

```r
plot(price_clean)
```

```r
pairs(numerical_features_clean)
```

```r
cor_matrix <- cor(numerical_features_clean)
cor_matrix
ggcorrplot(cor_matrix)
```

```r
#graphs with the categorical features
ggplot(categorical_features_clean, aes(x = categorical_features_clean$room_type)) +
  geom_bar(stat = "count", fill = "skyblue", color = "black") +
  labs(title = "Count Plot", x = "Category", y = "Count") +
  theme_minimal()
```

```r
neighborhood_Price <- data_clean[11:12]
summary_price <- neighborhood_Price %>%
  group_by(neighbourhood_cleansed) %>%
  summarise(mean_price = mean(price),
            median_price = median(price),
            min_price = min(price),
            max_price = max(price),
            total_listings = n())

summary_price
```

From Section 3.2

```r
airbnb24 <- read.csv('ny_ny.24.csv')
airbnb24$logP <- log(airbnb24$price)

lm24 = lm(price ~ area + accommodates + bathrooms + min_nights + room_type + beds, data=
log24 = lm(logP ~ area + accommodates + bathrooms + min_nights + room_type + beds, data=

lm1 = lm(logP ~ area*log(accommodates) + area*bathrooms + area*log(min_nights) + area*ro

## Residuals vs. Y
ggplot(airbnb24, aes(x=area, y=price)) + geom_point() +
  xlab("Area") + ylab("Price") + labs(title="Area vs. Price")
```

```r
ggplot(airbnb24, aes(x=area, y=logP)) + geom_point() +
  xlab("Area") + ylab("Log of Price") + labs(title="Area vs. Log(Price)")
```

```r
## Residuals vs. fitted values
ggplot(airbnb24, aes(x=lm24$fit, y=lm24$res)) + geom_point() +
  xlab("Fitted Values") + ylab("Residuals") + labs(title = "Without Log Transformation")
  geom_hline(yintercept = 0, col=2)
```

```r
ggplot(airbnb24, aes(x=log24$fit, y=log24$res)) + geom_point() +
  xlab("Log Fitted Values") + ylab("Residuals") + labs(title = "With Log Transformation"
  geom_hline(yintercept = 0, col=2)
```

```r
summary(lm1)$r.squared
summary(log24)$r.squared
summary(lm24)$r.squared

coef(lm1)[1]  #log(price)
tidy(lm1)
tidy(lmUES)
```