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 obtain the summary statistics on the numerical features.From the summary table, there are signs of outliers which we will do further analysis to confirm. The two most notable ones are 42 beds and 1250 minimum amounts of nights.

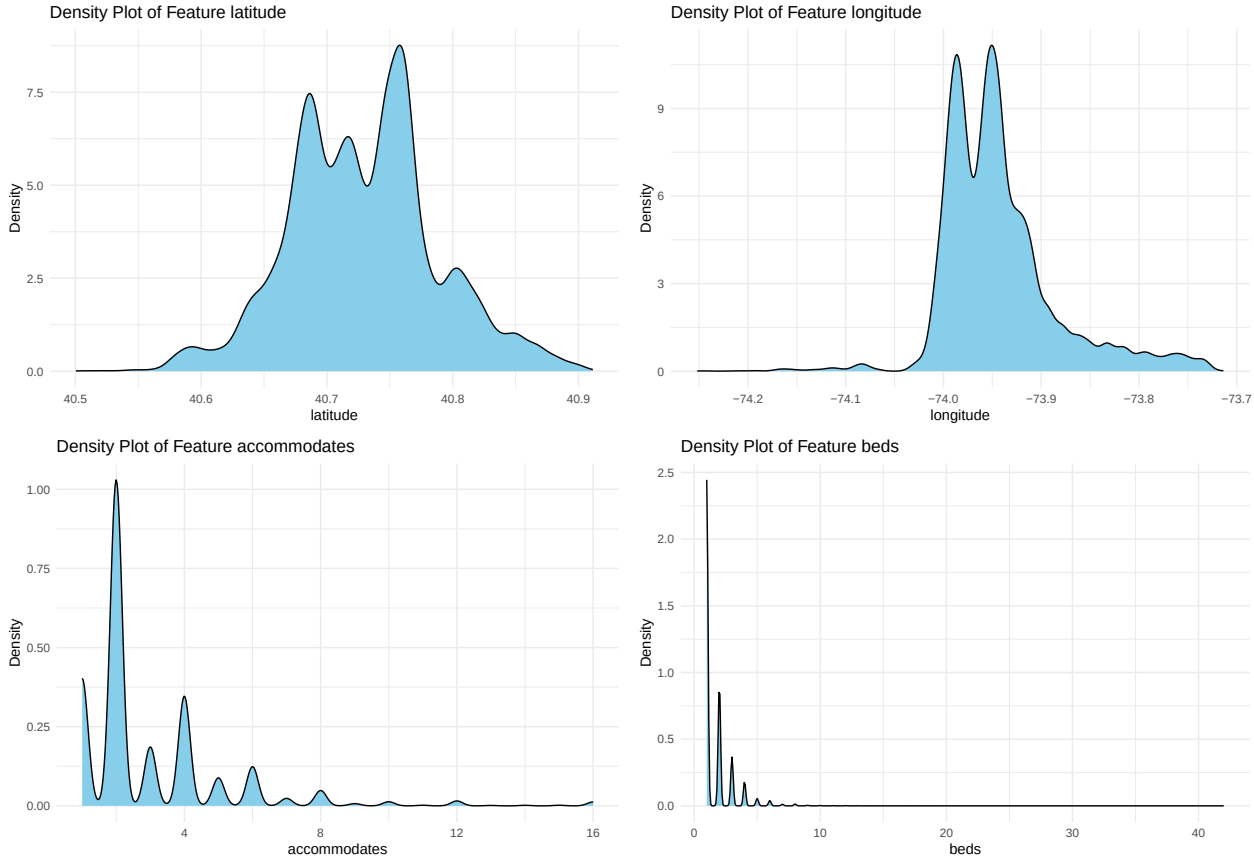
```
data <- read.csv('nydata.csv')

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

stargazer(num_and_price,type='latex', title = "Figure 1: Summary Stat for Numerical Feat
```

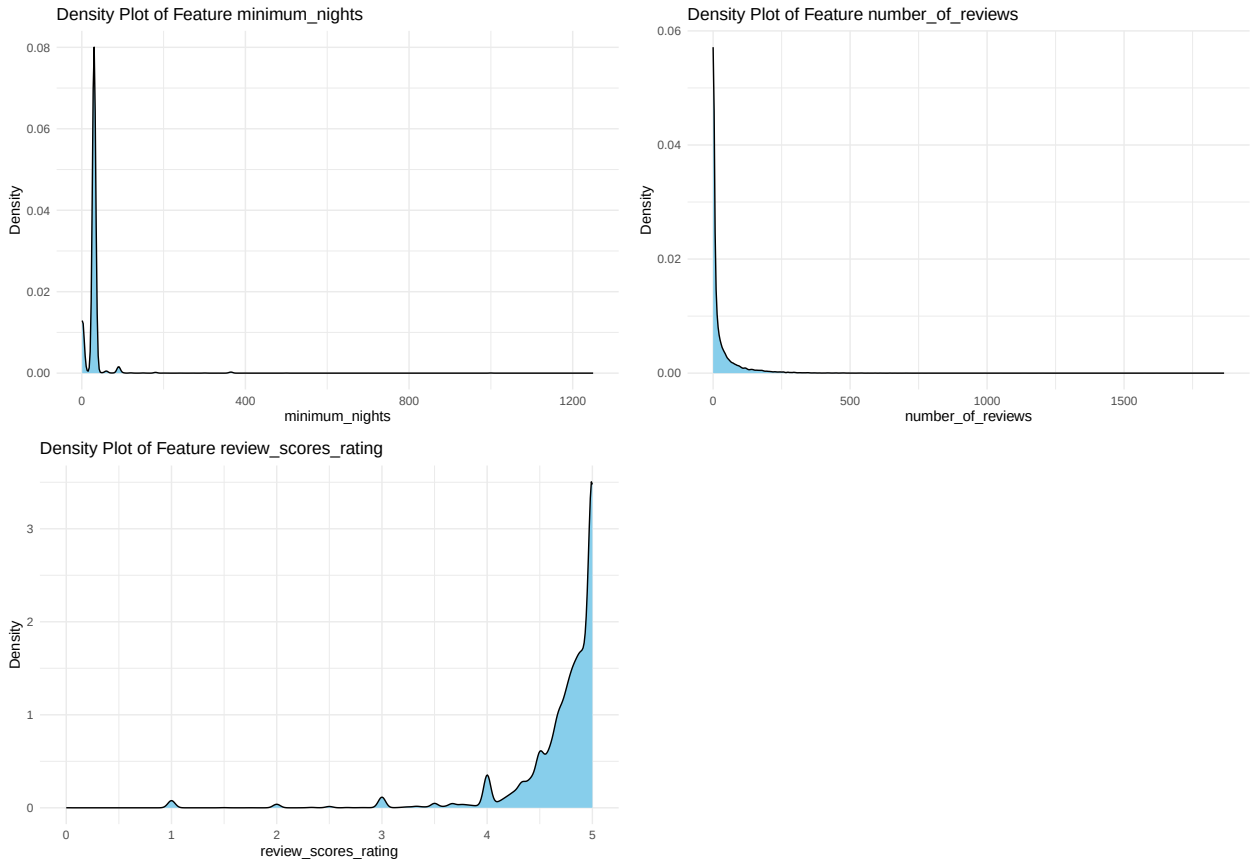
% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Feb 12, 2024 - 21:05:02

Table 1: Figure 1: Summary Stat for Numerical Features

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

Next, we plot the density for each of the feature to get a sense of the shape. From the plots of each feature, there are signs of skewness from each other the features. The features are also poorly distributed. This also confirms our claim that there are outliers. The main outlier exist with minimum amounts stay, price, and amount of beds. To remedy this, we will remove the outliers according to the summary statistics in figure 1.



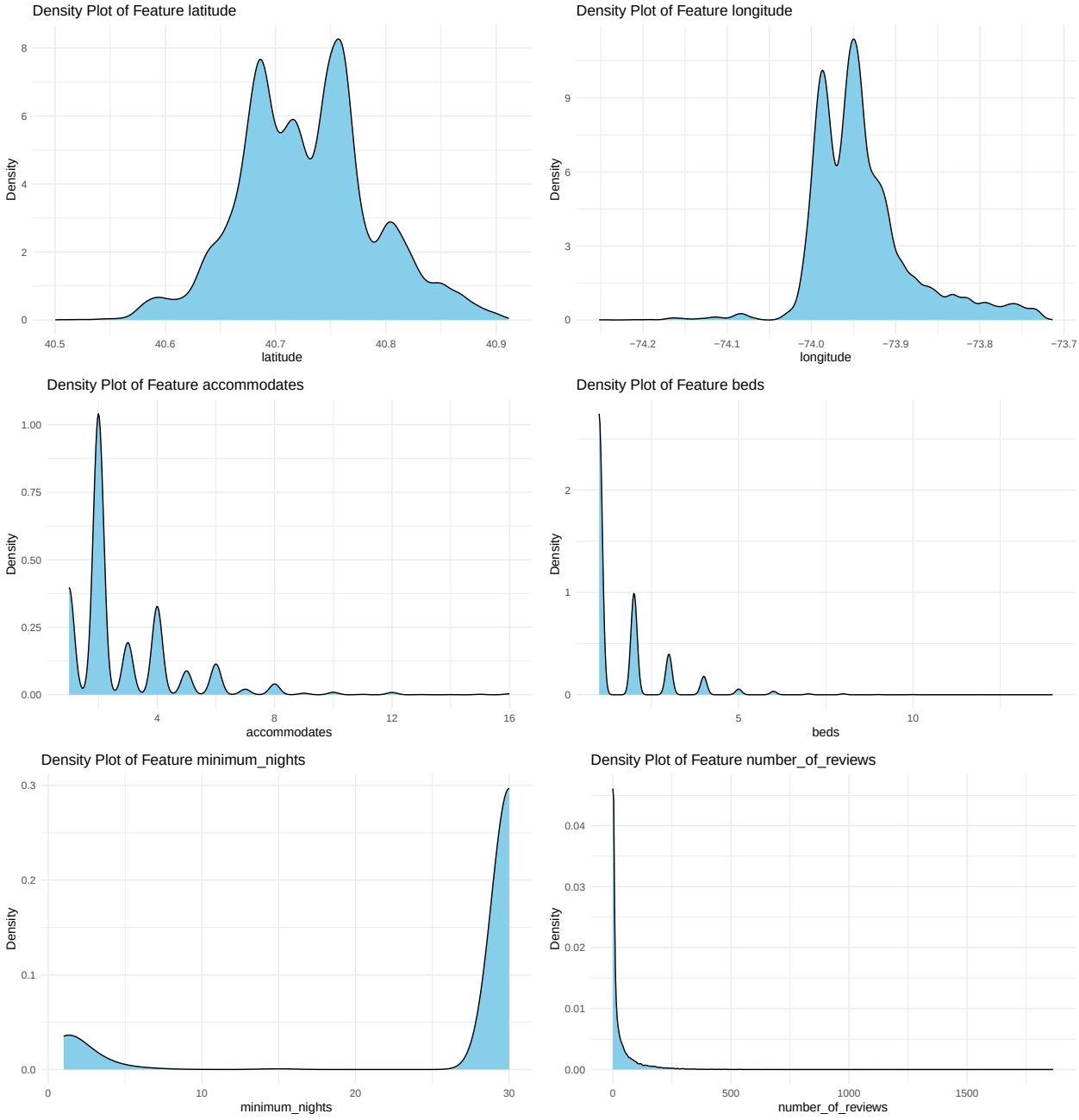


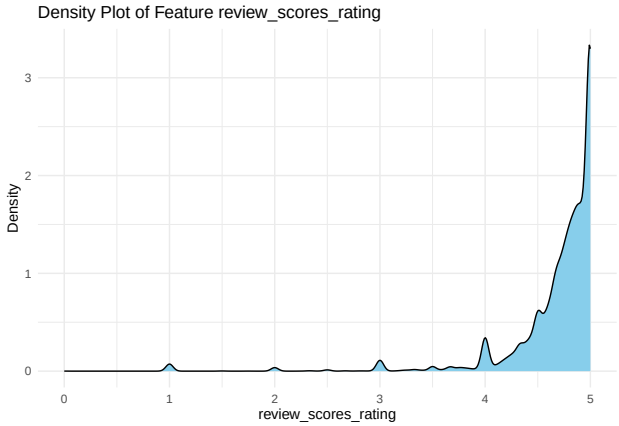
We clean the data by using two filters. The first one is to only look at listing that is under 500 dollars. The second one is that we only look at listings that have a minimum nights of under 30. We choose these two because of the distribution plots. Most data clustered around price less than or equal to 500 and minimum nights less than or equal to 30. Looking at figure 2, which is the summary stat for the cleaned data, and the density plots for the cleaned data, the skewness problem seems to be alleviated.

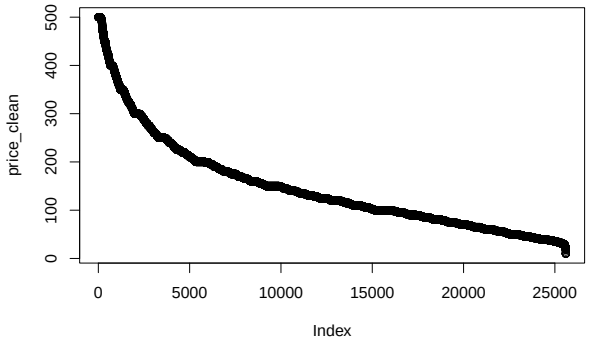
% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Feb 12, 2024 - 21:05:03

Table 2: Figure 2: Summary Stat for Cleaned Numerical Features

Statistic	N	Mean	St. Dev.	Min	Max
latitude	25,604	40.729	0.060	40.500	40.911
longitude	25,604	−73.940	0.060	−74.252	−73.714
accommodates	25,604	2.807	1.882	1	16
beds	24,947	1.648	1.079	1	14
minimum_nights	25,604	26.055	9.750	1	30
number_of_reviews	25,604	33.002	67.832	0	1,865
review_scores_rating	19,234	4.694	0.488	0.000	5.000
price_clean	25,604	146.213	96.199	10	500


Density Plot of Feature latitude
Density
latitude
Density Plot of Feature longitude
Density
longitude
Density Plot of Feature accommodates
Density
accommodates
Density Plot of Feature beds
Density
beds
Density Plot of Feature minimum_nights
Density
minimum_nights
Density Plot of Feature number_of_reviews
Density
number_of_reviews





Next, we check for correlation between the predictors. Because predictors such as number of reviews on a host might be correlated with other predictors that can reflect quality of housing , i.e number of beds and amenities, we want to make sure that our predictors does not show strong correlation. If our predictors are correlated, it can affect our model fit. To check, we created a correlation heat map of predictors. From figure 3, we see light correlation between some of the predictors. However, since no two predictor exhibit strong correlation, which is good. We will fit our model and analyze the model to see if our model is good fit.

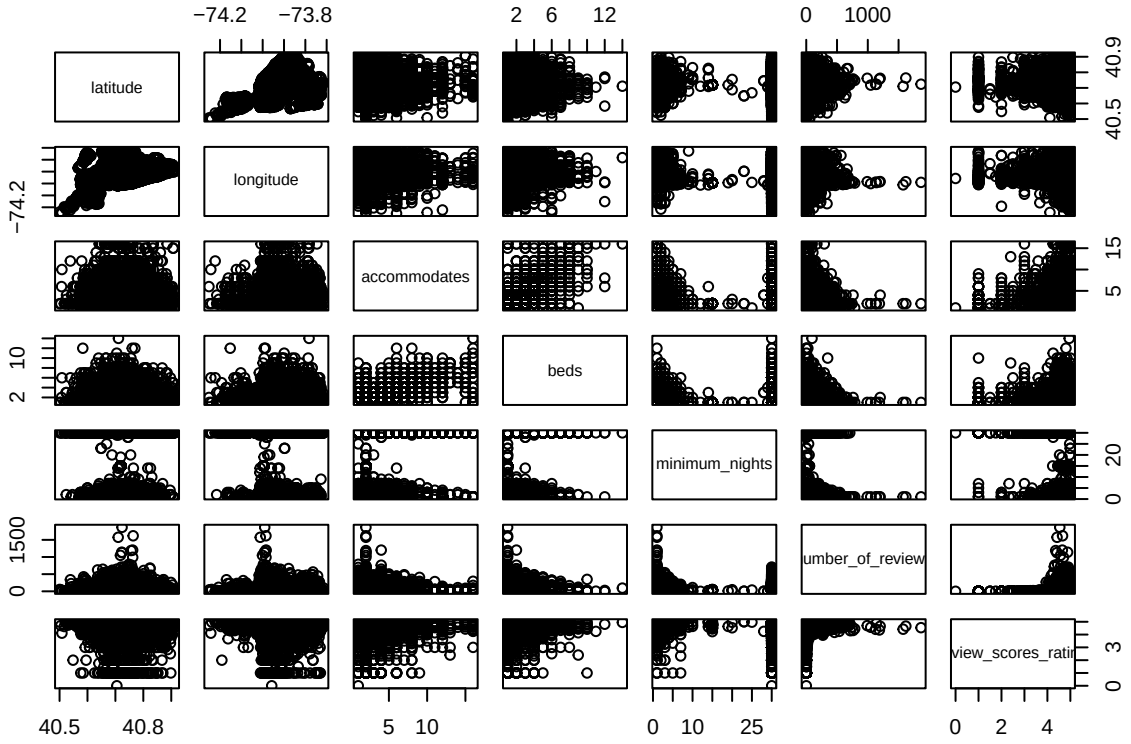
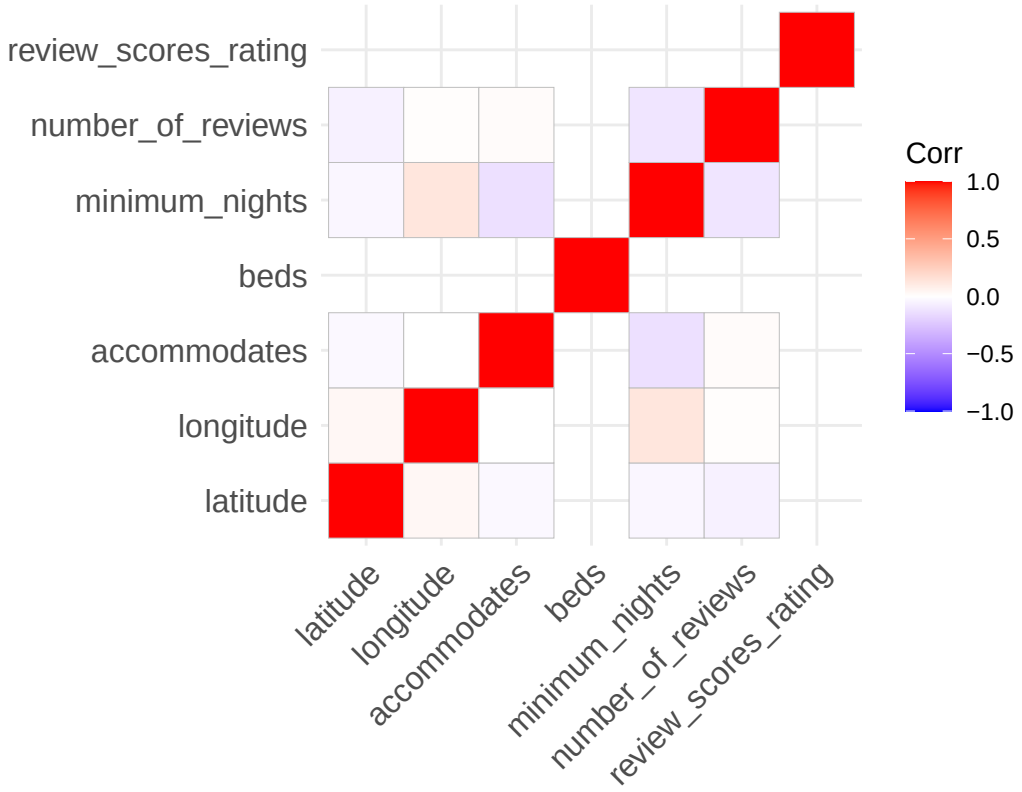
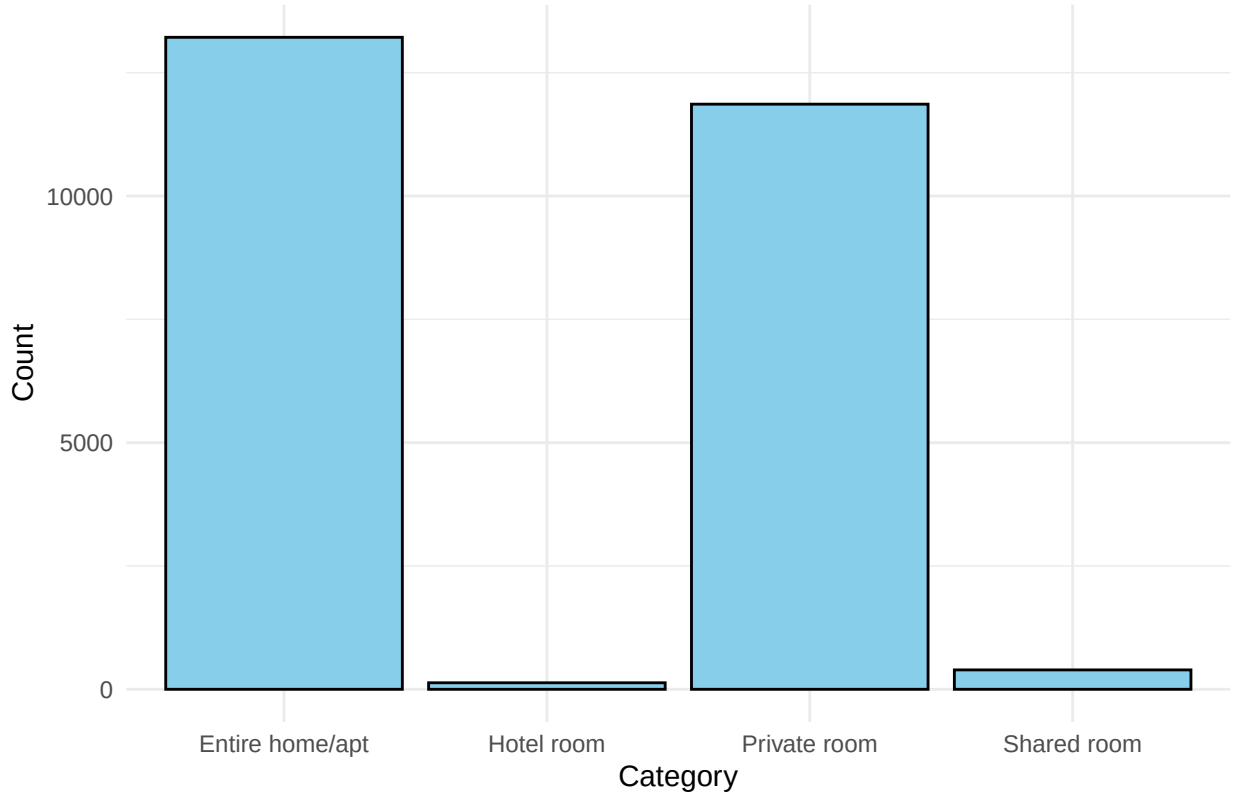


Figure 3: Correlation Heat Map



To conclude our exploratory data analysis, we plot the relationship between neighborhood and price. From figure 5, we break down the prices by the neighborhood. Note that the graph only shows 30 neighborhoods because there are around 200 total neighborhoods in New York. We will insert the total in the appendix.

Figure 4: Type of Housing Count Plot



% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Feb 12, 2024 - 21:05:23

3. Frequentist Analysis

3.1 Proposed Frequentist Model(s)

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ij} + \epsilon_i$$

- Y is the predicted rent price
- β_0 is the intercept
- $\beta_1, \ldots, \beta_p$ are the coefficients for the respective X_n independent variables
- ϵ_i is the error term

The variables that we included in our model were:

Table 3: Figure 5: Summary Stat by Neighborhood

	neighbourhood_cleansed	mean_price	median_price	min_price	max_price	total_listi
1	Allerton	117.825	94	31	500	40
2	Arden Heights	133.750	134.500	100	166	4
3	Arrochar	137.667	97.500	75	350	12
4	Arverne	134.603	108	33	375	63
5	Astoria	111.634	94.500	25	500	418
6	Bath Beach	147.667	110	40	399	24
7	Battery Park City	185.650	180.500	10	360	40
8	Bay Ridge	106.707	85	40	350	99
9	Bay Terrace	149.500	124.500	99	250	4
10	Bay Terrace, Staten Island	175	175	175	175	1
11	Baychester	107.424	98	40	300	33
12	Bayside	135.231	108	39	415	39
13	Bayswater	95.389	80	38	225	18
14	Bedford-Stuyvesant	123.060	100	10	500	1,943
15	Belle Harbor	221.286	225	92	350	7
16	Bellerose	136.200	107	50	300	15
17	Belmont	112.552	90	30	300	29
18	Bensonhurst	114.489	100	24	351	45
19	Bergen Beach	204.882	187	50	500	17
20	Boerum Hill	199.642	195	44	415	53
21	Borough Park	94.500	72.500	35	330	46
22	Breezy Point	150	150	150	150	4
23	Briarwood	140.111	82	37	499	36
24	Brighton Beach	134.375	119.500	50	385	48
25	Bronxdale	74.938	58.500	40	159	16
26	Brooklyn Heights	194.667	179	70	410	45
27	Brownsville	128.079	95	31	500	101
28	Bull's Head	124.833	101	60	200	6
29	Bushwick	109.358	85	18	500	1,008
30	Cambria Heights	124.159	106.500	43	350	44

- *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*, the number bathrooms in the Airbnb
- 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 start by fitting the following models.

- *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

The formula for the 3 models are as following:

- $Price = \beta_0 + \beta_1 \times area + \beta_2 \times accommodates + \beta_3 \times bathrooms + \beta_4 \times nights + \beta_5 \times room\,type + \beta_6 \times beds$
- $log(Price) = \beta_0 + \beta_1 \times area + \beta_2 \times accommodates + \beta_3 \times bathrooms + \beta_4 \times nights + \beta_5 \times room\,type + \beta_6 \times beds$
- $log(Price) = \beta_0 + \beta_1 \times area \times log(accommodates) + \ beta_2 \times area \times bathrooms + \beta_3 \times area \times log(min\ nights) + \beta_4 \times area \times room\ type + \beta_5 area \times bed$

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).

Figure 6: Area vs. Price

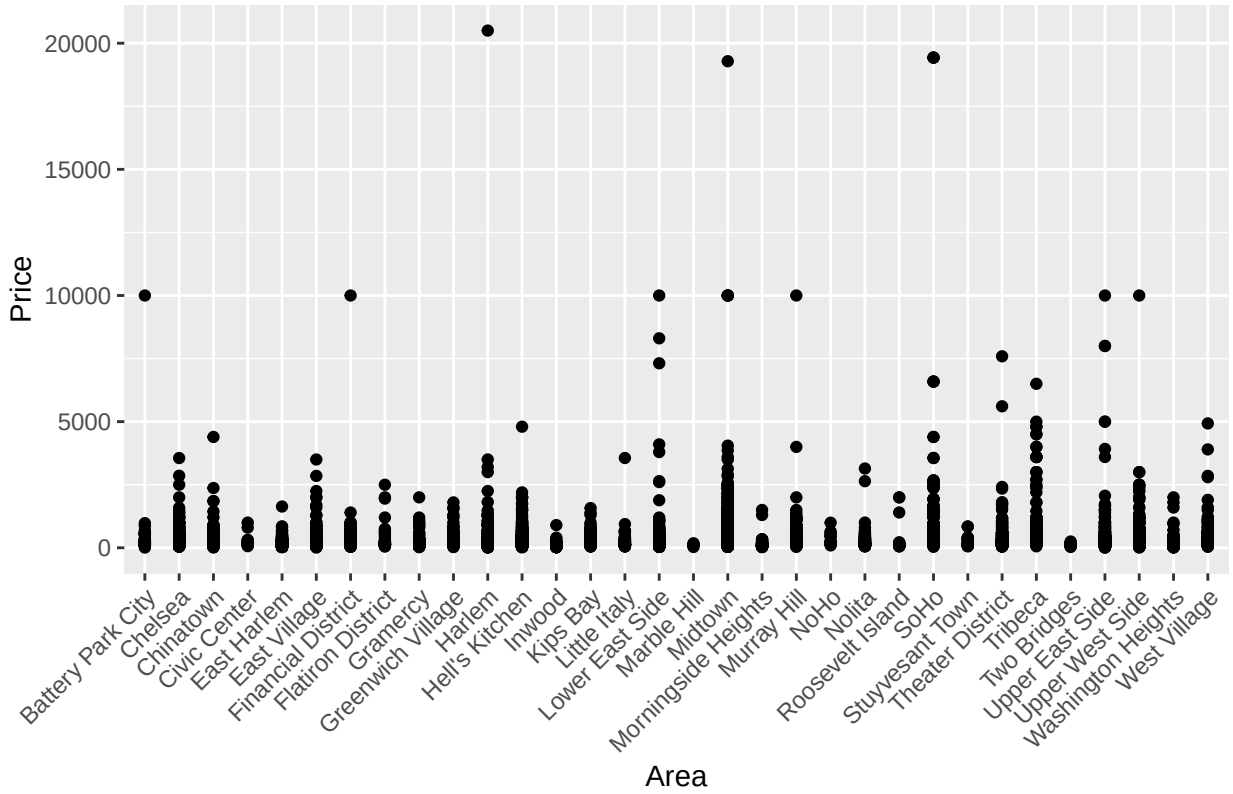
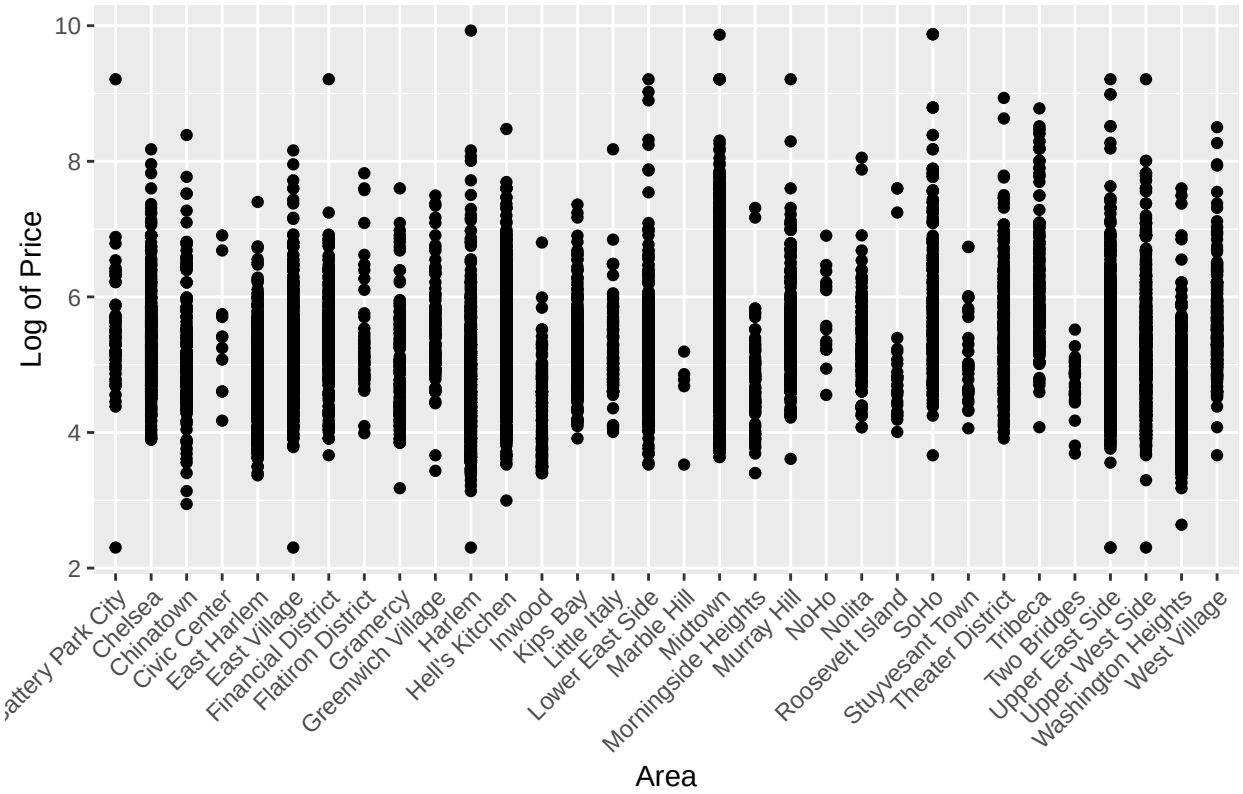


Figure 7: Area vs. Log(Price)



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.

Figure 9: Without Log Transformation

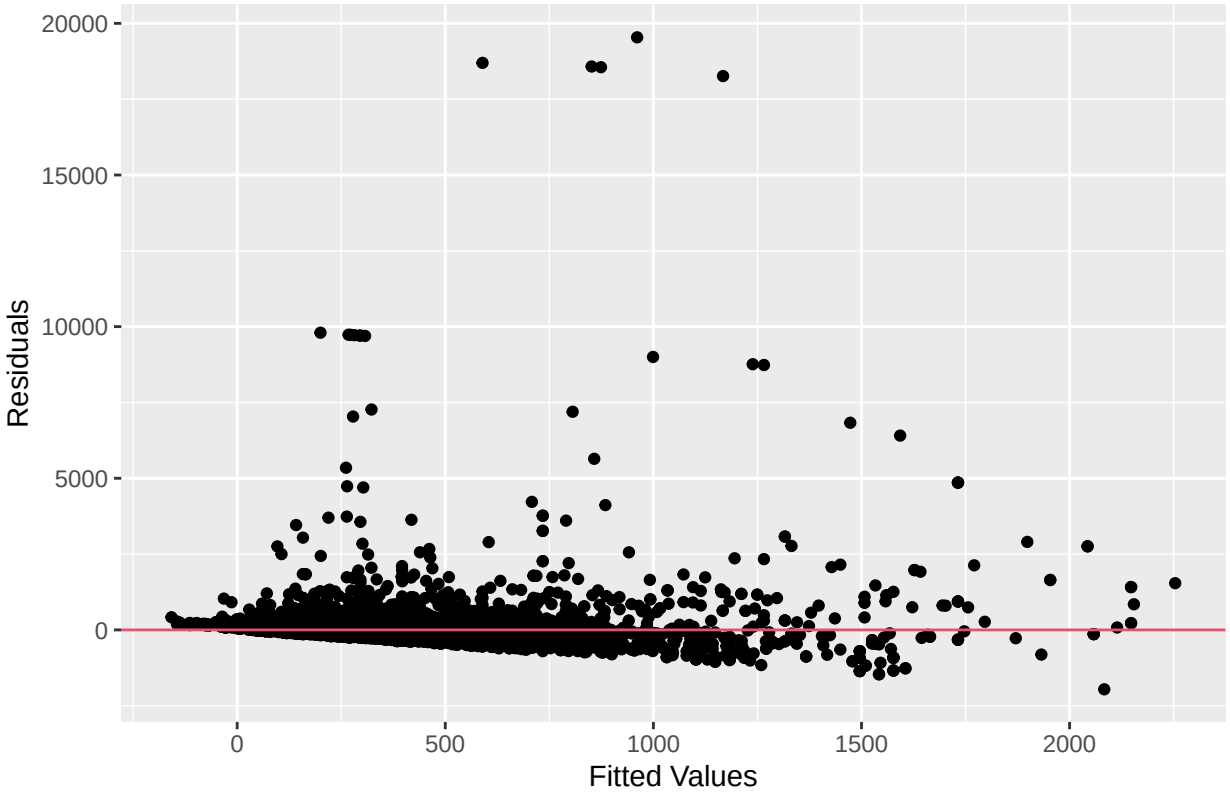
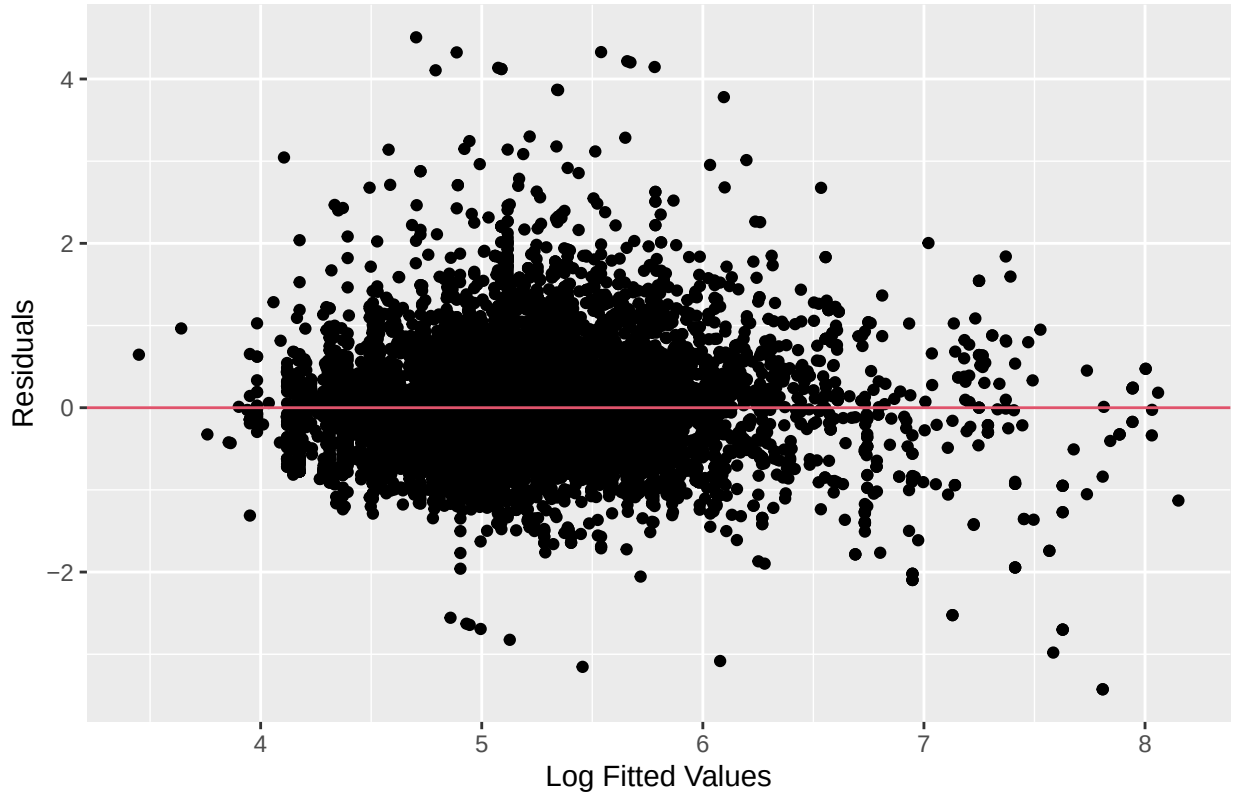


Figure 10: With Log Transformation



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*. From figure 11 and 12 which contains the Anova table, and AIC/BIC 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). From both figures, the model that includes interaction terms and log transform performs the best. Using our oringial data, we find that the mean

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Feb 12, 2024 - 21:05:27

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Feb 12, 2024 - 21:05:27

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

Table 4: Figure 11: Anova for 3 models

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
area	31	1,481.220	47.781	144.409	0
log(accommodates)	1	1,821.400	1,821.400	5,504.780	0
bathrooms	1	124.084	124.084	375.016	0
log(min_nights)	1	26.097	26.097	78.872	0
room_type	3	54.580	18.193	54.985	0
beds	1	0.022	0.022	0.067	0.795
area:log(accommodates)	30	94.135	3.138	9.483	0
area:bathrooms	30	105.990	3.533	10.678	0
area:log(min_nights)	31	96.530	3.114	9.411	0
area:room_type	60	153.788	2.563	7.747	0
area:beds	31	134.141	4.327	13.078	0
Residuals	12,060	3,990.350	0.331		

Table 5: Figure 12: AIC/BIC for 3 Models

	Name	AIC	BIC
1	Log Interaction Model	21,490.000	23,136.300
2	Log Transformed	23,197.000	23,493.600
3	Linear Model	191,358.000	191,655.000

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)

In the Bayesian context, we can use our Airbnb data in two ways. We can use a none-informative prior such as $1/\sigma^2$ that does follows $IG(0, 0)$. Alternatively, we can use this dataset as the prior and download the more recent 2024 as our data. For this proposal, we will use the reference prior, $1/\sigma^2$ due to the new data have not been published as the proposal due date.

We now formally define the notion for this section:

$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$

- $X_{n\times p}$ = data matrix
- $Y_{n\times 1}$ = price response vector
- $\beta_{p\times 1}$ = predictor vector
- $\epsilon_{n\times 1} \quad N(0, \sigma^2 I_{p\times p})$ = noise vector

4.2 Fitting the Bayesian model(s)

- Propose how you will fit the proposed Bayesian models.

Following the correlation heat map from exploratory data analysis, we see that some predictors are correlated. However, since the correlation are not over 0.5, we can assume that it will not a significant impact and each predictors are independent.

To start the Bayesian regression analysis, we will center data first by subtracting the mean from each column.

```
means = colMeans(numerical_features_clean, na.rm = TRUE)
numerical_features_clean_substracted <- numerical_features_clean
for (i in seq_along(df)) {
  numerical_features_clean_substracted[, i] <- ifelse(is.na(numerical_features_clean[, i
```

```
}

stargazer(numerical_features_clean_substracted,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: Mon, Feb 12, 2024 - 21:05:27

Table 6:

Statistic	N	Mean	St. Dev.	Min	Max
latitude	25,604	0.000	0.060	−0.228	0.183
longitude	25,604	−73.940	0.060	−74.252	−73.714
accommodates	25,604	2.807	1.882	1	16
beds	24,947	1.648	1.079	1	14
minimum_nights	25,604	26.055	9.750	1	30
number_of_reviews	25,604	33.002	67.832	0	1,865
review_scores_rating	19,234	4.694	0.488	0.000	5.000

Since this a multiple linear regression. We can define the distribution as the following:

- $\pi(\beta \mid \sigma^2, X, Y) \quad N(\hat{\beta}, \sigma^2(X'X)^{-1}$ where $\hat{\beta} = (X'X)^{-1}X'Y$
- $\pi(\sigma^2|X, Y) \ . \ IG((n-p)/2, SSE/2)$ where $SSE = \quad (Y - X\hat{\beta})'(Y - X\hat{\beta})$

Alternatively, we can fit a linear hierarchical model to better compensate for neighborhood data. To do this, we would use the brms library to do the following.

Stage 2

$\alpha|\beta, \Sigma_\alpha \sim \text{MVN}_2(\beta, \Sigma_\alpha)$

Stage 3

- $\beta|\mu, \Sigma_0 \sim \text{MVN}_2(\mu, \Sigma_0)$
- $\Sigma_\alpha \sim \text{Inv-Wishart}(R, \rho)$
- $\sigma_e^2 \sim \text{IG}(a_e, b_e)$

We'll be able to get a better understanding of to do this after 12/14/24 lectures and learning how to use the brms after the lab06.

To perform sensitivity test, we will make sure the following conditions are satisfied.

- $n > p$

- Making sure the predictors are independent
- $X'X$ is convertible.
- Checking collinearity
- $SSE > 0$

Another way that we can check for sensitivity is using data from prior years to do cross validation to make sure our model is correct.

For ensuring MCMC convergence, we will plot and monitor the following showing signs of convergence:

- trace plots
- histograms
- Gelman-Rubin convergence diagnosis
- Effective sample size

4.3 Prediction

After generating the posterior distribution, we can generate the predictive posterior distribution, and compute the predictive mean and credible intervals.

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. Another way we can improve our model is to use a informative prior derived from previous years' dataset. However, this might be difficult because of other variable impacting the Airbnb market such as the economy.

6. Contributions

For the first proposal, Bridgette and Jake split evenly. Bridgette did sections 3,4,and 5. Jake did selection 1 and 2. For the final propose, Jake did all the revision. We have spent 15+ hours for this proposal

References

Appendix

From Section 2:

```
summary_price <- as.data.frame(summary_price)

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
}
```

```
plot(price_clean)
```

```
pairs(numerical_features_clean)
```

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

```
#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()
```

```
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

```
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)
```