MAT 360 – Generalized Linear Model
Final Project

Price Prediction on Airbnb Rentals

Kavit Patel & Eddie Cojulun

March 8, 2023

Abstract

This study uses multiple regression techniques to analyze the relationship of renting prices of Airbnb houses in consideration with some factors that may or may not affect it. The dataset, called "Airbnb Prices in European Cities", is obtained from Kaggle.com and includes a total of 20 independent datasets. It consists of 10 cities; those cities are divided into weekdays data and weekends data that adds upto 20 independent datasets. All the datasets have 20 rows including price, room, guest, lat and long, attraction index, metro index, bedroom, satisfaction rating and more to explain further. For the analysis and sake of the project we used the city Amsterdam. This means there are 2 datasets for the city. Having the same column names, we merged both the datasets. We mainly have a goal to compare the results of gamma model. The results summarized looks something like we made some new discoveries of unexpected significant factors.

Table of Contents

Introduction

Airbnb, Inc. is and American-based company operating an online marketplace for short-term homestays and experiences. The company acts as a middleman and charges a commission from each booking. It is a service that allows for property owners to rent out their spaces or entire property to travelers. The rising popularity to Airbnb can most likely be tributed to the fact that not everyone can afford to stay in a hotel and sometimes is could be very difficult to find a hotel room in a busy area. In 2021 alone, 300 million bookings were made on Airbnb.

In order to book a stay on Airbnb, a person needs to create an account and from there browse places to rent based on what location they want to stay at, when they want to check in, check out, and how many people will be staying at the place. Today there are listings from over 200 countries and 40,000 different cities. It is also possible to get a narrower list of options by specifying additional information such as neighborhood location, price, type of property, and more. In order to be host and put of a piece of property up for booking, a person needs to have an Airbnb account and must describe their property as accurately, along with listing all its amenities as accurately as they can so that they could be matched to the right guest. Hosts can set their own price and with some research, they could ensure their price is competitive with other listings. Adding photos is a great way to make their listing more attractive to potential guests. They are responsible for managing their listing by responding to inquiries/ booking requests and they have the final say on whether or not they want to book to someone.

Airbnb's potential for the future is significant. The different types and amount of data collected by Airbnb could prove to be incredibly useful in terms of modeling or estimating relationships that are affected by geographic locations in a technique called spatial econometric analysis.

The dataset we used for our study is named "". This dataset can be used to gain insight on the cost of Airbnb listings in some of the most popular European cities. It contains information on a variety of attributes such as room type, cleanliness rating, guest satisfaction, distance from the city center, and more. In addition to exploring general trends in prices across Europe, this data set can be used for deeper spatial econometric analysis. The data set consists of hundreds, if not thousands of different accounts (observations) for many cities throughout Europe, but the city we are focusing on in this project is Amsterdam while combing the two data files from weekday and weekend Airbnb bookings. For each booking, the following variables were given:

- **realSum:** a continuous variable which records the full price of accommodation for two people and two nights in EUR. (Numerical)
- **room_type:** a nominal variable which describes the type of the accommodation. ("Private room", "Entire home/apt", "Shared room")

- **room_shared:** a nominal variable which describes whether the room is shared or not. (Boolean)
- **room_private:** a nominal variable which describes whether or not the room is private or not. (Boolean)
- **person_capacity:** a continuous variable which records the maximum number of guests.
- **host_is_superhost:** a nominal variable which describes whether the host is a superhost or not. (Boolean)
- **multi:** a nominal variable which describes if the listing belongs to hosts with 2-4 offers (Boolean)
- **biz:** a nominal variable which describes if the listing belongs to hosts with more than 4 offers. (Boolean)
- **cleanliness_rating:** a continuous variable which records the cleanliness rating of the listing (numeric).
- **guest_satisfaction_overall:** a continuous variable which records the overall satisfaction rating of the listing. (numeric)
- **bedrooms:** a continuous variable which records the number of bedrooms in the listing. (0 for studios) (Numeric)
- **dist:** a continuous variable which records the distance from the city center. (Numeric)
- **metro_dist:** a continuous variable which records the distance from the nearest metro station. (Numeric)
- **attr_index:** a continuous variable which records the attraction index of the listing location (numeric).
- **attr_index_norm:** a continuous variable which records a normalized attraction index (0-100) (numeric).
- **rest_index:** a continuous variable which records the restaurant index of the listing location (numeric).
- **rest_index_norm:** a continuous variable which records the normalized restaurant index (0-100) (numeric).
- **lng:** a continuous variable which records the longitude of the listing location. (numeric)
- **lat:** a continuous variable which records the latitude of the listing location. (numeric)

The goal of our analysis is precisely to study the prices of Airbnb bookings in the city of Amsterdam, both on weekdays and weekends. In particular, we wanted to test whether a relationship exists between listings' amenities, descriptive data, and especially spatial metrics such as distance from the city's center as well as longitude and latitude. For this, we will use a regression model and gamma model since the level of measurement of rental price of Airbnb listings is numerically continuous and somewhat skewed.

Literature Review

"A Spatial Econometric Analysis of Regional Income Convergence in Mexico" (1999) by Miguel Flores and Sergio Rey, is a report that uses spatial econometric techniques to analyze the patterns of regional income convergence in Mexico. The study begins by providing an overview of the concept of income convergence and the different methods that are used to measure it. The author then presents a spatial econometric model that considers the spatial dependencies between regions in Mexico. Using data on per capita income for the period 1990-2010, the author finds evidence of conditional convergence, which means that poorer regions tend to grow faster than richer regions, but only up to a certain point. Beyond that point, there is no evidence of convergence. The author also finds evidence of spatial dependence, which means that the growth rates of neighboring regions are positively correlated. Overall, the study provides important insights into the patterns of regional income convergence in Mexico and factors that influence it.

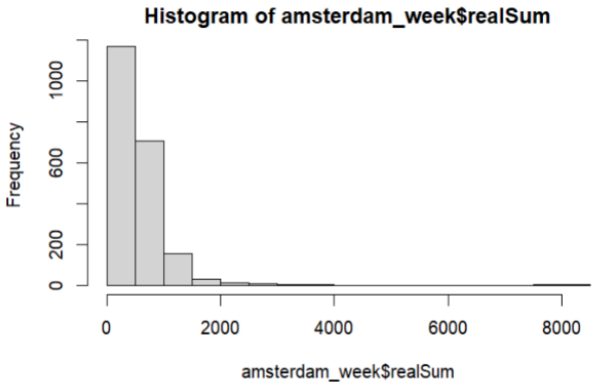
"Determinants of Airbnb prices in European cities: A spatial econometrics approach" by Kristóf Gyódi is a research paper that examines the factors that determine the prices of Airbnb listings in European cities. The paper uses a spatial econometrics approach to analyze the spatial patterns and determinants of Airbnb prices. Using data on more than 300,000 Airbnb listings in 36 European cities, the author finds that the location of the listing is the most important determinant of the price of an Airbnb accommodation. Listings located in central or highly desirable areas command higher prices than those located in less central or less desirable areas. Overall, the study provides valuable insights into the factors that determine the prices of Airbnb accommodation in European cities. The study highlights the importance of location, as well as the quality and size of the listing, in determining the price of an Airbnb accommodation. The use of spatial econometrics provides a deeper understanding of the spatial patterns and determinants of Airbnb prices in European cities.

Exploratory Data Analysis

Before we began any sort of modeling, we employed some EDA methods such as visualization with histograms for our numerical variables and bar plots for some of our nominal predictors, as well as the generation of a correlation matrix to see what numeric variables correlated with some of the other ones. The insights gained from our exploratory analysis phase informed subsequent analyses, such as hypothesis testing or machine learning modeling. By understanding the key features of our data, we were able to make more informed decisions about the appropriate analytical techniques to use and the potential limitations or biases of the data.

The first discovery that we took note of was the distribution of the realSum variable which is a recording of the total price for the Airbnb listing. It will be our response variable as our aims of this analysis is to accurately predict pricings for Airbnb bookings in the city of Amsterdam based on various attributes associated with each listing. The minimum value was $128.88 and a maximum of $8130.67 with a median of 460.2 and a mean of 573.1. The relatively higher mean

and the very high max compared to the median was something that stood out to us. After we are taking a look at the histogram we determined that our response variable was skewed and then utilized the Shapiro-Wilk test on realSum. The Shapiro-Wilk test is a statistical test used to assess whether a given dataset is normally distributed or not. It is a popular test for normality and it measures the deviation of the observed sample data from a normal distribution, taking into account the mean and the variance of the data The null hypothesis of the test is that the data follows a normal distribution. If the p-value of the test is less than the chosen significance level (e.g., 0.05), the null hypothesis is rejected, indicating that the data is unlikely to be normally distributed.



$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

```
        Shapiro-Wilk normality test

data:  amsterdam_week$realSum
W = 0.62982, p-value < 2.2e-16
```

Because the p-value was 2.2e-16 the null hypothesis was rejected and we determined that the varaible realSum was not normally distributed and either a transformation would have to be applied to our response varaible or another GLM would have to be used and not a regular linear regression model. Another fact we observed was that there are only two cases where the price of the listing was over $4000. Because of this observation we decided to create a separate data set almost exactly the same as the original this subset would not include those two cases real realSum was over 4000. We would use these two data sets (the original and subset) to compare the models created from each set.

Apart from the principal analysis we conducted on our response variable, we also engaged in some univariate analysis, creating histograms for all the other numerical explanatory variables and bar plots for our nominal and ordinal variables. Some of the explanatory variables like cleanliness_rating and guest_satisfaction_overall had somewhat skewed distributions after looking at their histograms and can be seen in Appendix A. Appendix A, Table A.2 also has some of the bar plots that we generated to see how some of the nominal variables and their different categories compared to each other in terms of pricing, our response variable.

Methods and Analysis

Before proceeding with out regression analysis, we generated a correlation matrix of all numerical variables in both the datasets to give us an idea whether we would be dealing with the possibility of multicollinearity (see Appendix A, Table A.3) There are only five cases of correlation above 0.6 (all statistically significant) and just one of those is larger than 0.8

(correlation between attr_index_norm and rest_index_norm). This scenario of relatively low correlation doesn't seem to suggest the possibility of multicollinearity may arise.

We studied the relationship between the price of listings and listing characteristics. Since the pricing of a booking is a continuous variable, the multiple regression model represents the natural choice to model the response variable. However, as mentioned before the assumption of normality is not satisfied but a linear regression model was fitted for the purpose of comparing it to a GLM later.

Y represents the price of a listing (realSum) and the exploratory variables employed are room type, whether the room was shared or not, whether the room was private or not, the capacity of people, whether not the host is a super host, whether or not the listing belongs to host with 2-4 offers (multi), whether or not the listing belongs to a host with more than 4 offers (biz), the rating of how clean the listing is, the overall rating of the listing, the number of bedrooms, the distance from the city center in km, the distance from the nearest metro station in km, the attraction index of the listing location, the restaurant index of the listing location, the longitude of the location and the latitude of the listing location.

We fitted a linear regression for both the original data set and the subset of the original without the two cases where realSum was above $4000. The variables room_type, host_is_superhost, multi, and biz are categorical variables (see Tables A.4-A.7 in Appendix A). The estimates for the original data set are reported in Table B.1 in Appendix B. And the estimates for the subset are reported in Table B.2 in Appendix B.

The very first thing that we noted from the model summary was that two variables were not defined because of singularities. This message occurs when you fit some model in R and two or more predictors have an exact linear relationship between the – known as perfect multicollinearity. This is because we used a categorical variable in a regression model and used dummy variables that took on values of True and False. One of our categorical variables was room_type and we also had variables room_shared and room_private which are categories of room_type so that is why these variables do not have coefficient estimates in the model. For our original data set, the parameter estimates for room_type Private room, room_type Shared room, person_capacity, bedrooms, dist, and lng are all statistically significant at least at the 5% level. When taking a look at the summary for the original data set, the parameter estimates seem to be incredibly small and barely impactful. The reason behind these very small estimate values is most likely because in the original data set, there are two observations of listings that are about $3000 more than $4000, while every other listing is below $4000. Because of the lack of listings with prices in the range of 4000 and 8000, the model can be influenced heavily by those observations in a suboptimal way.

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

But if we look at the model fitted on our subset, that excludes those two bookings with prices over 7k, we find some very important and meaningful results. Both room_type Private room and Shared room have negative coefficient estimates over 100 (-175 and -270 respectively) most likely describing how private rooms and shared rooms have a lower price on average compared to those of room_type base case, that is Entire home/apt. The variables person_capacity, bedrooms, and longitude have positive estimates greater 100, meaning as the values of these variables goes up, so does the price of the Airbnb listing. Finally, some smaller estimates like dist, which measures the distance from the listing location from the city's center and biz, which tells whether or not the host of the listing belongs to a host with more than 4 offers have negative values of -23.5 and -41.3 respectively, holding all other predictors constant. While cleanliness_rating was not statistically significant, guest_satisfaction_overall was positive with a parameter estimate of 2.4 which confirms our hypothesis that more highly rated listings correlate with more expensive bookings. Overall, these findings are consistent with our initial expectations and with the prevailing literature.

Compared to the model fitted on original data set, the model for the altered subset had strong signs of a better model such as an Adjusted R-Squared of 0.5423 which means a little more than half of the variance we see in the pricings for the Airbnb listings can be explained by the model fitted on the altered sub data set. The R-Squared for the original data set was only 0.4373. The standard error for the model fitted on the original data set was 324.4 while the model for the subset was only 246.8. The log likelihood, AIC, and BIC for the original data set was -14968.8, 29975.65, and 30082.81 respectively while the log likelihood, AIC, and BIC for the subset was -14386.3, 28810.58, 28917.72 where we generally want AIC and BIC to be as close to zero. These values are far from 0 but that is to be somewhat expected since our response variable is not normally distributed, so our model does not satisfy the assumption of normality.

The next technique we tried was again regular multi-linear regression but with a transformed response variable. From the histogram of realSum shown earlier, it is obvious that the pricings for all the listings is not exactly normally distributed. With this in mine, one common way to deal with right-skewed data is to transform the response variable using a mathematical function such as the logarithm, square root, or reciprocal. This can help to make the distribution more symmetrical and improve the performance of statistical models. However, it's important to note that some transformations can change the interpretation of the variable, so we chose a transformation that makes sense for our research question and data, therefore we transformed realSum with the logarithm function. The histograms of the realSum variable from both the original data set and the data set that excludes the 2 observations with pricings over $4000 can be seen in Table B.3 in Appendix B. By looking at the histograms and The Shapiro-Wilk test statistic "w" of 0.98 for both data sets, we can conclude the transformed response variable is normally distributed because in general, values of "w" above 0.95 are considered to provide strong evidence

that the data are normally distributed, while values below 0.90 are considered to indicate non-normality.

The estimates for these transformed-response-variable normal regression models can be seen in Table B.4 in Appendix B. There are a number of things to take note of such as the increased R and Adjusted R Squared with a value of 0.6895 for the original data set and 0.6947 for the model fitted on the subset. The log likelihood, AIC, and BIC are -372.69, 783.38, and 890.5 for the model fitted on the subset that excludes the two cases of luxury pricings which is a great improvement from the original linear models. For that same subset, the variables room_type Private room, room_type Shared room, person_capacity, multi, guest_satisfaction_overall, bedrooms, dist, rest_index_norm, lng, and lat all are significant parameter estimates. However, because we transformed our response variable using a log function the interpretation is altered. The coefficients represent the expected change in the log of the response variable for a one-unit increase in the corresponding predictor variable, holding all other predictor variables constant. In other words, the coefficient represents the percentage change in the expected value of the response variable for a one-unit increase in the predictor variable. With an estimate coefficient of -3.708e-01, we can interpret it as: if the room type is private then the percent change in listing price is decreased by 31% compared to listings where the room type is equal to "Entire home/apt". We take the exp() of the estimate coefficients and subtract 1 and then finally multiply it by 100 to get the percent change for a one-unit increase in that variable. For every 1 increase in number of bedrooms, the price goes up by exp(exp(1.427e-01) - 1) * (100) or 15% all other variables held constant. The mathematical interpretations of all the significant variables in the subset model can be seen in Table B.5 in Appendix B. Just like the untransformed response variable, the variable dist has a negative coefficient meaning the further away the listing location is from the city, the cheaper the booking costs.

$$g(\mu) = log(\mu) = \sum_{1}^{p} X_j \beta_j$$

Lastly, we fit a gamma generalized linear model because as a GLM gamma regression allows for us to model the relationship between the response variable and one or more predictor variables while accounting for the non-normal distribution of the response variable. Because our response variable is skewed to the right and not normally distributed it can be well described by a gamma model.

Where μ is the expected value of the listing price and the explanatory variables are room type, whether the room was shared or not, whether the room was private or not, the capacity of people, whether not the host is a super host, whether or not the listing belongs to host with 2-4 offers (multi), whether or not the listing belongs to a host with more than 4 offers (biz), the rating of how clean the listing is, the overall rating of the listing, the number of bedrooms, the distance

from the city center in km, the distance from the nearest metro station in km, the attraction index of the listing location, the restaurant index of the listing location, the longitude of the location and the latitude of the listing location. The estimates are reported in Table B.6 and B.7 in Appendix B.

The parameter estimates can be interpreted the same way as they are in our normal linear regression model with the transformed response variables because we used a log link function for our gamma regression. The estimates, specifically their signs are consistent with the estimates' signs of the other models and our pre-modeling intuitions. The variables room_type Private room, room_type Shared room, person_capacity, multi, bedrooms, dist, rest_index_norm, and lat all are significant parameter estimates. Unlike the other models, guest_satisfaction_overall and longitude are not statistically significant variables. The mathematical interpretations of all the significant variables in the subset model can be seen in Table B.8 in Appendix B.

The summaries of the fitted gamma models also display the AIC values of both the original data set and the subset that excludes the two cases of pricings over $4000. Contrary to our initial theories, the gamma regression model has much larger AIC and BIC values compared to the normal linear model where the response variable was transformed. It should be noted that this gamma regression model deals with the assumptions much better than both the regular linear models and regression models that have a transformed response variable. The residual vs fitted plot allows us to see how the assumption of constant variance is much better handled with the gamma regression as seen in Table B.9-B.11 in Appendix B.

Conclusion

Our analysis shows that many factors have a substantial impact on the pricing of Airbnb listings, according to the regression analysis that was done on the Airbnb data. The number of beds, person capacity, and longitude of the listing site were found to have a substantial favorable impact on pricing. On the other side, it was discovered that factors like location, whether or not there are more than four offers, and the style of accommodation (private or shared) have a large negative impact on pricing. The results are in line with both our initial predictions and the literature running a lot of tests on original model or the subset models also had different factor significance.

Prediction of prices using Gamma led us to somewhere closer to the results of this model. It is significant to notice that the model fit on the modified subset of data outperformed the model fit on the original dataset. The subset model's Adjusted R-squared value (0.6947) was greater than original (0.6895), suggesting that the model could account for more of the pricing volatility. The Moreover, the subset model's standard error was smaller (0.291) than the original standard error (0.297). The log likelihood, the AIC, and BIC for the model fitted on the original data set are -13438.79, 26915.57, and 27022.73 respectively, while the log likelihood, AIC, and BIC for the model fitted on the subset data set are -13346.11, 26730.22, and 26837.36 respectively. Coming

the number of significant factors. The original model had room_typePrivate, room_typeShared, person_capacity, guest_satisfaction_overall, bedrooms, dist, rest_index_norm, lng, lat as their high significant factors. There are 9 significant variables in this model. In other words, their variables had p-value <0.05. Now we compared this to the subset model. It turns out that it had another significant factor in addition to the ones in the original model. The factor was multi which makes it a total of 10 significant variables.

Overall, the regression analysis results offer useful information about the factors that affect the pricing of Airbnb listings in Amsterdam, which individuals may or may not use to improve pricing strategy and on the other hand customers can use to make educated decisions while booking Airbnb listings.

References

"Why Do People Stay in Airbnb?" *Why Do People Stay in Airbnb?*, hospitalityinsights.ehl.edu/travelers-airbnb-study.

Sergio J. Rey & Brett D. Montouri (1999) US Regional Income Convergence: A Spatial Econometric Perspective, Regional Studies, 33:2, 143-156, DOI: 10.1080/00343409950122945

Gyódi, Kristóf, and Łukasz Nawaro. "Determinants of Airbnb Prices in European Cities: A Spatial Econometrics Approach." *Tourism Management*, vol. 86, 2021, p. 104319., doi:10.1016/j.tourman.2021.104319.

Appendix A

Table A.1: Histograms of two ordinal variables (cleanliness_rating and guest_satisfaction_overall)

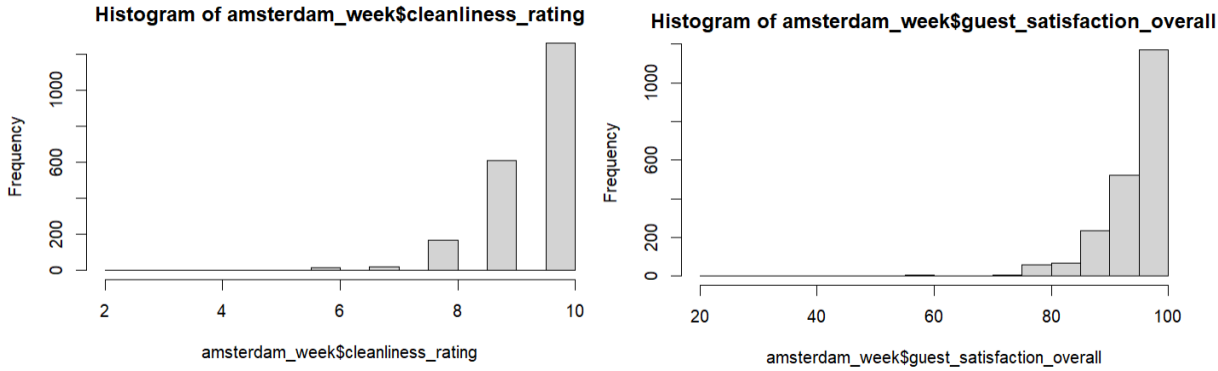
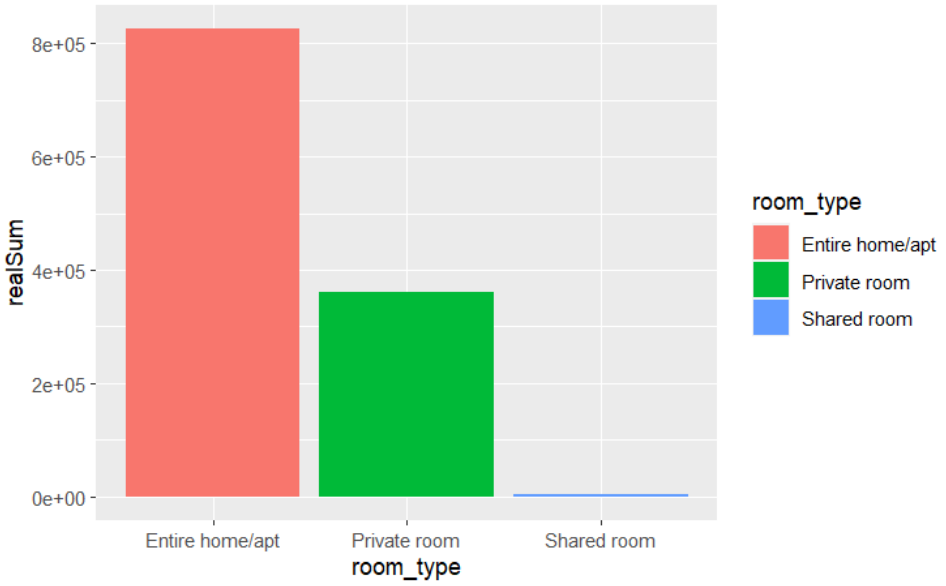


Table A.2: Bar plots for nominal variables against realSum (listing prices)



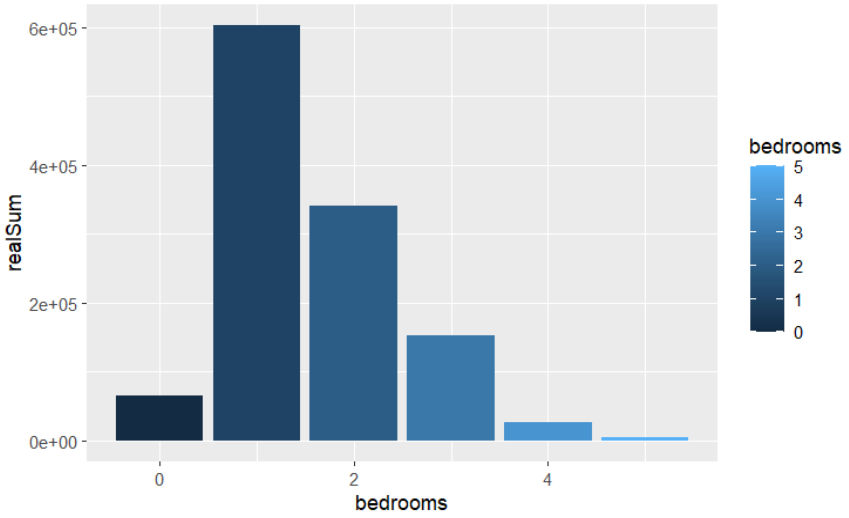


Table A.3: Correlation Matrix

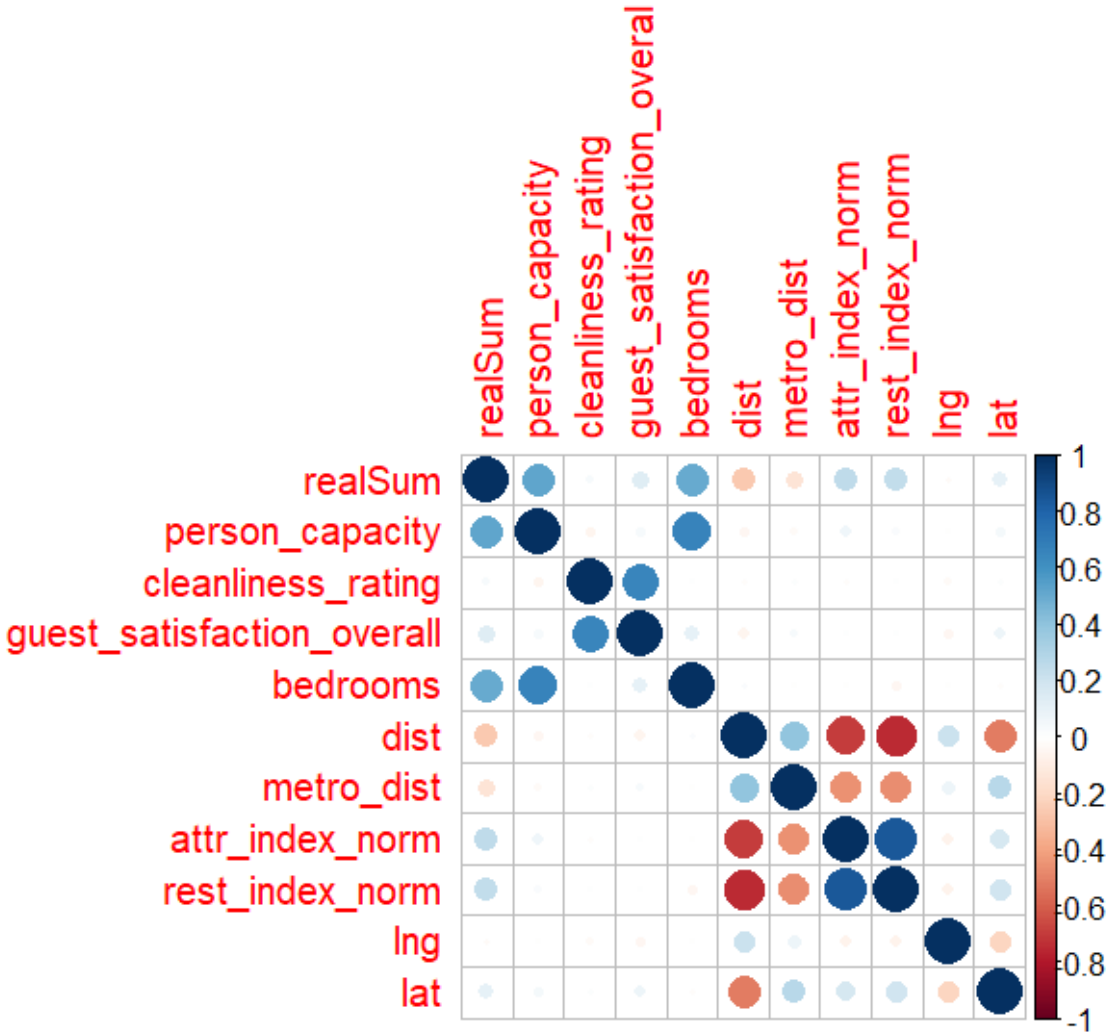


Table A.4: room_type bar plot

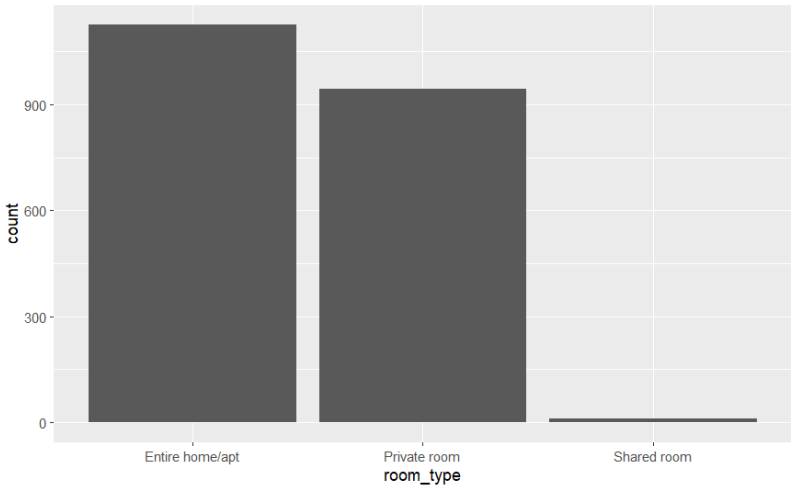


Table A.5: host_is_superhost bar plot

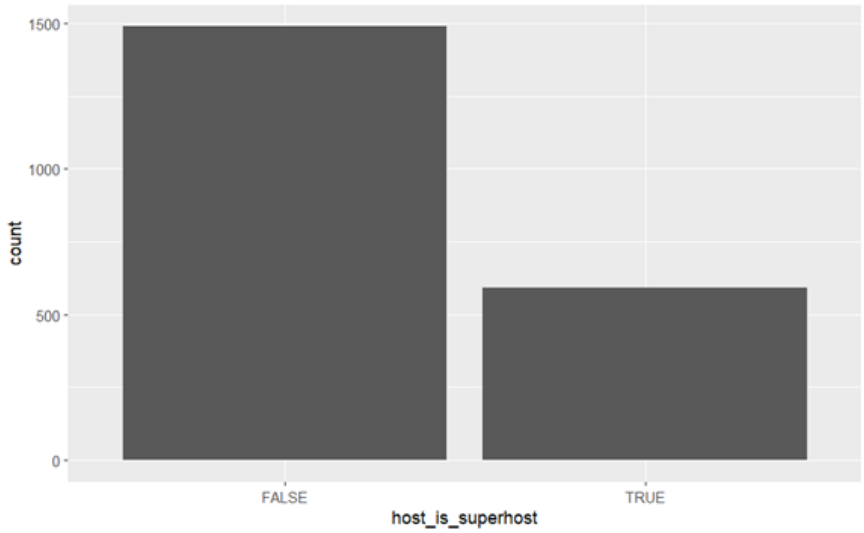


Table A.6: multi bar plot

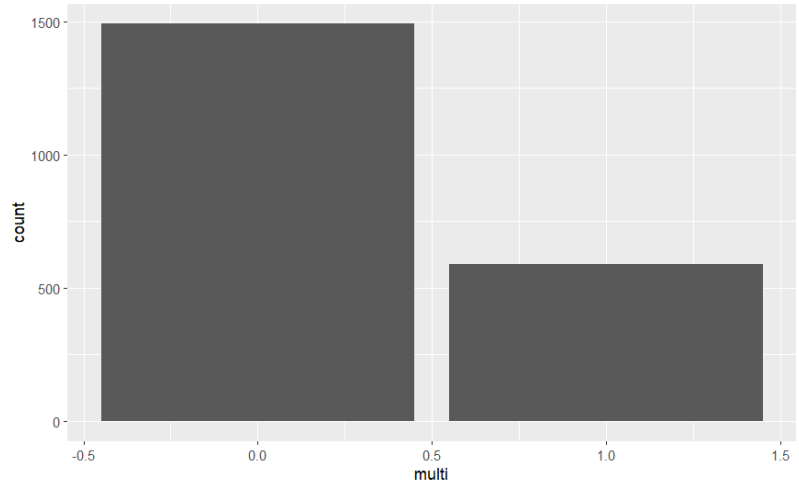
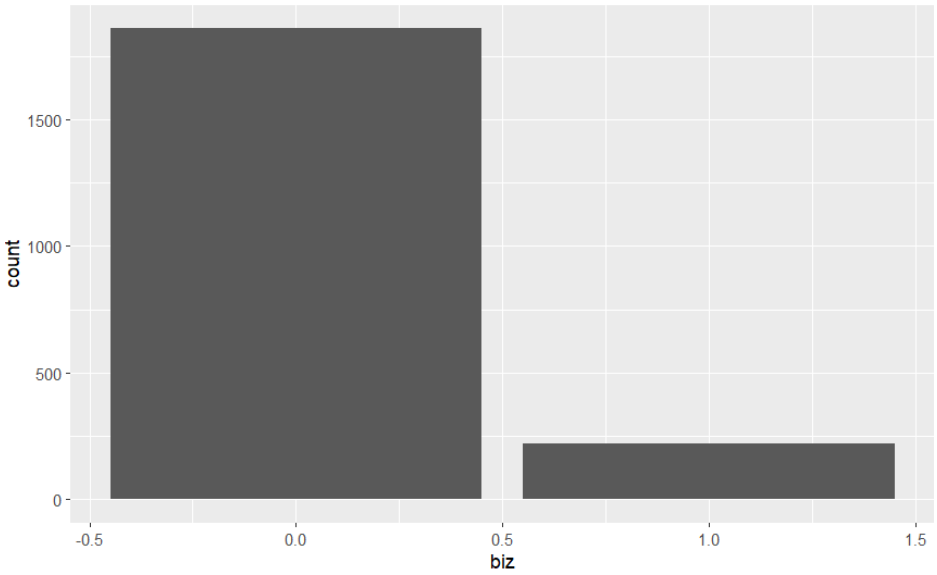


Table A.7: biz bar plot



Appendix B

Table B.1: Multi-Linear Regression Output (Original Data set)

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.540e+04	3.022e+04	-0.510	0.61036	
room_typePrivate room	-1.765e+02	1.697e+01	-10.402	< 2e-16	***
room_typeShared room	-2.879e+02	1.040e+02	-2.767	0.00571	**
room_sharedTRUE	NA	NA	NA	NA	
room_privateTRUE	NA	NA	NA	NA	
person_capacity	1.217e+02	9.497e+00	12.819	< 2e-16	***
host_is_superhostTRUE	-7.491e-01	1.703e+01	-0.044	0.96492	
multi	1.312e+01	1.702e+01	0.771	0.44097	
biz	-4.433e+01	2.486e+01	-1.783	0.07467	.
cleanliness_rating	9.650e+00	1.192e+01	0.810	0.41815	
guest_satisfaction_overall	2.841e+00	1.554e+00	1.828	0.06765	.
bedrooms	1.383e+02	1.329e+01	10.409	< 2e-16	***
dist	-2.047e+01	7.191e+00	-2.847	0.00446	**
metro_dist	-5.766e+00	1.192e+01	-0.484	0.62852	
attr_index	-7.023e+02	1.418e+03	-0.495	0.62043	
attr_index_norm	1.327e+04	2.678e+04	0.495	0.62032	
rest_index	-5.896e-01	5.162e-01	-1.142	0.25356	
rest_index_norm	9.439e+00	6.620e+00	1.426	0.15406	
lng	3.895e+02	1.924e+02	2.024	0.04308	*
lat	2.526e+02	5.746e+02	0.440	0.66021	

Table B.2: Multi-Linear Regression Output (Altered Subset)

```
Coefficients: (2 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              31979.1040 23020.6625   1.389 0.164938
room_typePrivate room     -175.0775    12.9093 -13.562  < 2e-16 ***
room_typeShared room      -270.6970    79.1607  -3.420 0.000639 ***
room_sharedTRUE                 NA         NA      NA       NA
room_privateTRUE                NA         NA      NA       NA
person_capacity            102.6504     7.2418  14.175  < 2e-16 ***
host_is_superhostTRUE       13.6530    12.9616   1.053 0.292308
multi                      -18.5122    12.9757  -1.427 0.153824
biz                        -41.2883    18.9103  -2.183 0.029121 *
cleanliness_rating           3.1813     9.0673   0.351 0.725732
guest_satisfaction_overall   2.4221     1.1824   2.048 0.040650 *
bedrooms                   134.7365    10.1124  13.324  < 2e-16 ***
dist                       -23.4881     5.4717  -4.293 1.85e-05 ***
metro_dist                  -3.9343     9.0652  -0.434 0.664335
attr_index                -784.7398  1078.7690  -0.727 0.467039
attr_index_norm          14825.3917 20374.0653   0.728 0.466904
rest_index                  -0.5958     0.3928  -1.517 0.129405
rest_index_norm              9.6727     5.0366   1.920 0.054935 .
lng                        310.5438   146.3969   2.121 0.034020 *
lat                       -641.8335   437.7238  -1.466 0.142720
```

Table B.3: Histograms of the transformed response variable of the original data set and subset

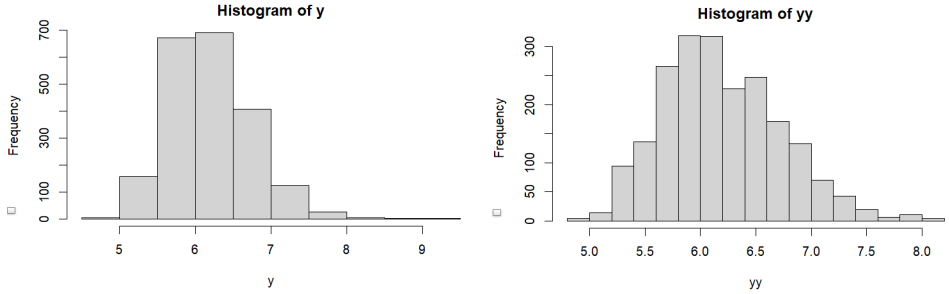


Table B.4: Gamma Model Summary for both the Original dataset and the subset

```
Coefficients: (2 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.144e+02  2.767e+01   4.133 3.72e-05 ***
room_typePrivate room     -3.712e-01  1.554e-02 -23.891  < 2e-16 ***
room_typeShared room      -6.141e-01  9.527e-02  -6.446 1.43e-10 ***
room_sharedTRUE                   NA         NA      NA       NA
room_privateTRUE                  NA         NA      NA       NA
person_capacity            1.792e-01  8.696e-03  20.608  < 2e-16 ***
host_is_superhostTRUE      1.119e-02  1.559e-02   0.718   0.4731
multi                     -2.985e-02  1.559e-02  -1.915   0.0556 .
biz                       -1.582e-02  2.276e-02  -0.695   0.4870
cleanliness_rating         9.811e-03  1.091e-02   0.899   0.3686
guest_satisfaction_overall 3.666e-03  1.423e-03   2.576   0.0101 *
bedrooms                   1.438e-01  1.217e-02  11.813  < 2e-16 ***
dist                      -6.155e-02  6.585e-03  -9.348  < 2e-16 ***
metro_dist                 3.440e-03  1.091e-02   0.315   0.7526
attr_index                -1.162e+00  1.298e+00  -0.895   0.3711
attr_index_norm            2.194e+01  2.452e+01   0.895   0.3710
rest_index                -8.752e-04  4.727e-04  -1.851   0.0643 .
rest_index_norm            1.500e-02  6.062e-03   2.474   0.0134 *
lng                        3.614e-01  1.762e-01   2.051   0.0403 *
lat                       -2.117e+00  5.261e-01  -4.024 5.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.297 on 2062 degrees of freedom
Multiple R-squared:  0.692,     Adjusted R-squared:  0.6895
F-statistic: 272.6 on 17 and 2062 DF,  p-value: < 2.2e-16
```

```
Coefficients: (2 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.282e+02  2.712e+01   4.725 2.46e-06 ***
room_typePrivate room    -3.708e-01  1.521e-02 -24.378  < 2e-16 ***
room_typeShared room     -6.091e-01  9.327e-02  -6.531 8.21e-11 ***
room_sharedTRUE                  NA         NA      NA       NA
room_privateTRUE                 NA         NA      NA       NA
person_capacity           1.736e-01  8.532e-03  20.351  < 2e-16 ***
host_is_superhostTRUE     1.538e-02  1.527e-02   1.007   0.3138
multi                    -3.906e-02  1.529e-02  -2.555   0.0107 *
biz                      -1.494e-02  2.228e-02  -0.671   0.5026
cleanliness_rating        7.928e-03  1.068e-02   0.742   0.4581
guest_satisfaction_overall 3.544e-03 1.393e-03   2.544   0.0110 *
bedrooms                  1.427e-01  1.191e-02  11.979  < 2e-16 ***
dist                     -6.243e-02  6.447e-03  -9.684  < 2e-16 ***
metro_dist                3.973e-03  1.068e-02   0.372   0.7099
attr_index               -1.186e+00  1.271e+00  -0.933   0.3508
attr_index_norm           2.241e+01  2.400e+01   0.933   0.3507
rest_index               -8.775e-04  4.627e-04  -1.896   0.0580 .
rest_index_norm           1.507e-02  5.934e-03   2.540   0.0112 *
lng                       3.384e-01  1.725e-01   1.962   0.0499 *
lat                      -2.378e+00  5.157e-01  -4.610 4.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2908 on 2060 degrees of freedom
Multiple R-squared:  0.6972,    Adjusted R-squared:  0.6947
F-statistic:   279 on 17 and 2060 DF,  p-value: < 2.2e-16
```

Table B.5: Estimate interpretations for significant variables of the realSum transformed linear regression model fitted on the subset.

```
> #Percent Change if room_type = Private room
> (exp(-3.708e-01) - 1) * (100)
[1] -30.9818
> #Percent Change if room_type = Shared room
> (exp(-6.091e-01) - 1) * (100)
[1] -45.61599
> #Percent Change for every increase in person capacity
> (exp( 1.736e-01) - 1) * (100)
[1] 18.95796
> #Percent Change if the listing belongs to a host with 2-4 offers
> (exp(-3.906e-02) - 1) * (100)
[1] -3.830699
> #Percent Change for every unit increase in guest satisfaction
> (exp(3.544e-03) - 1) * (100)
[1] 0.3550287
> #Percent Change for every one increase in bedrooms
> (exp(1.427e-01) - 1) * (100)
[1] 15.33837
> #Percent Change for every km increase in distance from City's center
> (exp(-6.243e-02) - 1) * (100)
[1] -6.052118
> #Percent Change for every unit increase in restaurant index of location
> (exp(1.507e-02) - 1) * (100)
[1] 1.518413
> #Percent Change for every unit increase in longitude
> (exp(3.384e-01) - 1) * (100)
[1] 40.27015
> #Percent Change for every unit increase in latitude
> (exp(-2.378e+00) - 1) * (100)
[1] -90.72641
```

Table B.6: Gamma Regression Model Summary (Original Dataset)

```
Coefficients: (2 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               65.3140477 35.8499865   1.822   0.0686 .
room_typePrivate room     -0.3804218  0.0201320 -18.896  < 2e-16 ***
room_typeShared room      -0.6238600  0.1234489  -5.054 4.72e-07 ***
room_sharedTRUE                   NA         NA      NA       NA
room_privateTRUE                  NA         NA      NA       NA
person_capacity            0.1863007  0.0112674  16.534  < 2e-16 ***
host_is_superhostTRUE     -0.0113373  0.0202053  -0.561   0.5748
multi                     -0.0180722  0.0201955  -0.895   0.3710
biz                       -0.0341612  0.0294903  -1.158   0.2468
cleanliness_rating         0.0148852  0.0141381   1.053   0.2925
guest_satisfaction_overall 0.0032477  0.0018439   1.761   0.0783 .
bedrooms                   0.1547364  0.0157696   9.812  < 2e-16 ***
dist                      -0.0606680  0.0085322  -7.110 1.59e-12 ***
metro_dist                 0.0055228  0.0141369   0.391   0.6961
attr_index                -1.0433290  1.6823289  -0.620   0.5352
attr_index_norm           19.7083757 31.7731391   0.620   0.5351
rest_index                -0.0008049  0.0006125  -1.314   0.1890
rest_index_norm            0.0141808  0.0078544   1.805   0.0711 .
lng                        0.4207595  0.2282834   1.843   0.0655 .
lat                       -1.1858558  0.6816792  -1.740   0.0821 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.1481279)

    Null deviance: 667.12  on 2079  degrees of freedom
Residual deviance: 205.66  on 2062  degrees of freedom
AIC: 26916
```

Table B.7: Gamma Regression Model Summary (Subset Dataset)

```
Coefficients: (2 not defined because of singularities)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.028e+02  3.300e+01   3.114  0.00187 **
room_typePrivate room     -3.777e-01  1.851e-02 -20.410  < 2e-16 ***
room_typeShared room      -6.126e-01  1.135e-01  -5.398 7.53e-08 ***
room_sharedTRUE                   NA         NA      NA       NA
room_privateTRUE                  NA         NA      NA       NA
person_capacity            1.715e-01  1.038e-02  16.520  < 2e-16 ***
host_is_superhostTRUE      1.125e-04  1.858e-02   0.006  0.99517
multi                     -4.204e-02  1.860e-02  -2.260  0.02394 *
biz                       -3.261e-02  2.711e-02  -1.203  0.22914
cleanliness_rating         9.683e-03  1.300e-02   0.745  0.45641
guest_satisfaction_overall 3.128e-03  1.695e-03   1.845  0.06518 .
bedrooms                   1.526e-01  1.450e-02  10.524  < 2e-16 ***
dist                      -6.330e-02  7.845e-03  -8.070 1.18e-15 ***
metro_dist                 6.860e-03  1.300e-02   0.528  0.59767
attr_index                -1.117e+00  1.547e+00  -0.722  0.47023
attr_index_norm            2.110e+01  2.921e+01   0.722  0.47015
rest_index                -8.167e-04  5.631e-04  -1.450  0.14709
rest_index_norm            1.443e-02  7.221e-03   1.999  0.04575 *
lng                        3.602e-01  2.099e-01   1.716  0.08624 .
lat                       -1.893e+00  6.275e-01  -3.017  0.00258 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.1251844)

    Null deviance: 625.78  on 2077  degrees of freedom
Residual deviance: 191.52  on 2060  degrees of freedom
AIC: 26730
```

Table B.8: Estimate interpretations for significant variables of the gamma regression model fitted on the subset.

```
> #Percent Change if room_type = Private room
> (exp(-3.777e-01) - 1) * (100)
[1] -31.45639
> #Percent Change if room_type = Shared room
> (exp(-6.126e-01) - 1) * (100)
[1] -45.806
> #Percent Change for every increase in person capacity
> (exp(1.715e-01) - 1) * (100)
[1] 18.70841
> #Percent Change if the listing belongs to a host with 2-4 offers
> (exp(-4.204e-02) - 1) * (100)
[1] -4.116857
> #Percent Change for every one increase in bedrooms
> (exp(1.526e-01) - 1) * (100)
[1] 16.48589
> #Percent Change for every km increase in distance from City's center
> (exp(-6.330e-02) - 1) * (100)
[1] -6.133817
> #Percent Change for every unit increase in restaurant index of location
> (exp(1.443e-02) - 1) * (100)
[1] 1.453462
> #Percent Change for every unit increase in latitude
> (exp(-1.893e+00) - 1) * (100)
[1] -84.93807
```

Table B.9: Residuals vs Fitted Plot for the regular regression model fitted on the subset.

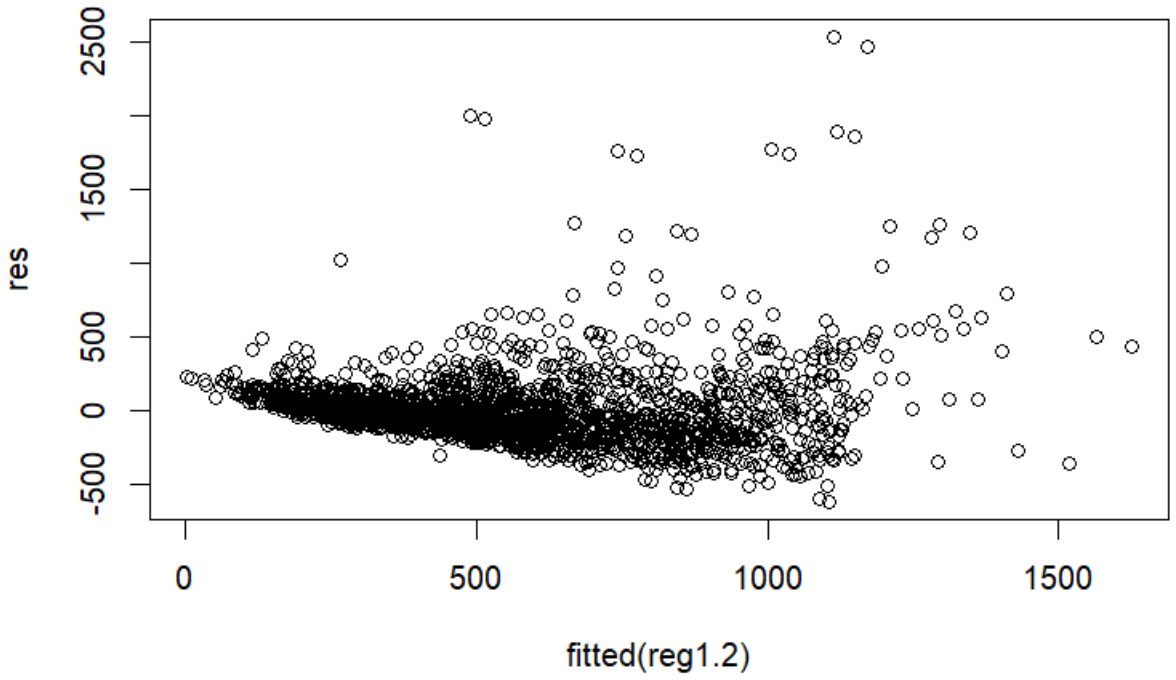


Table B.10: Residuals vs Fitted Plot for the regression model with the log() transformed response variable fitted on the subset.

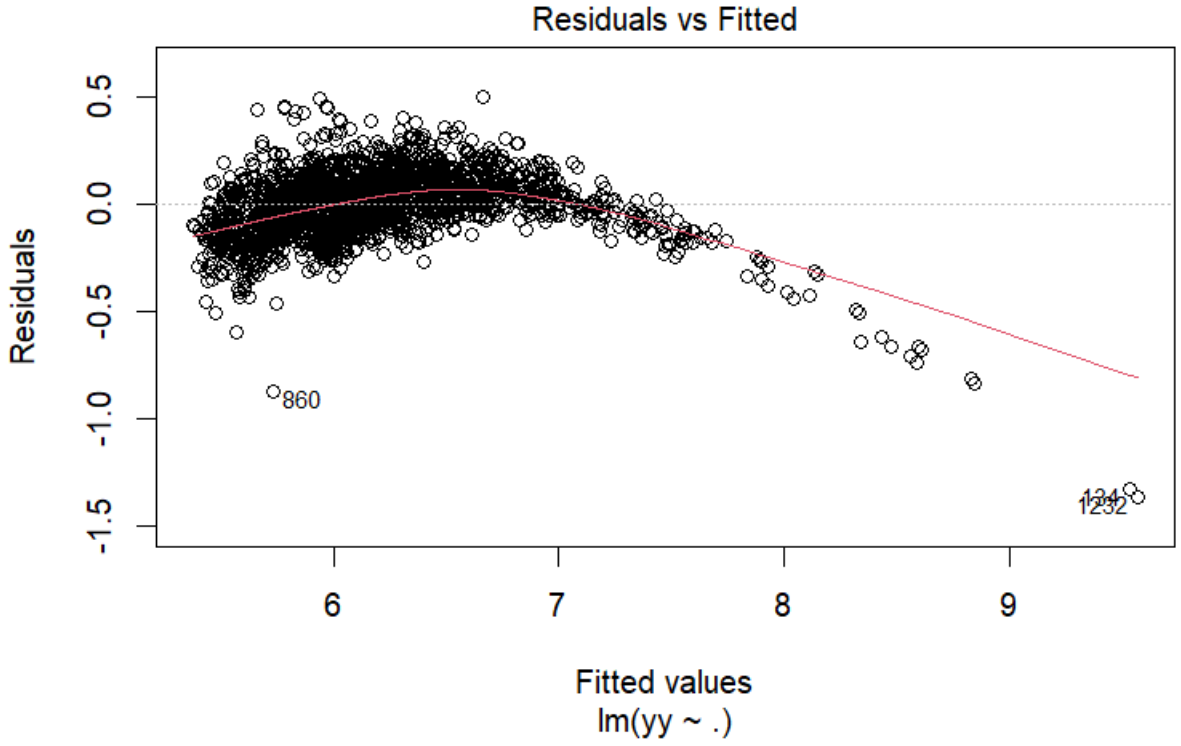
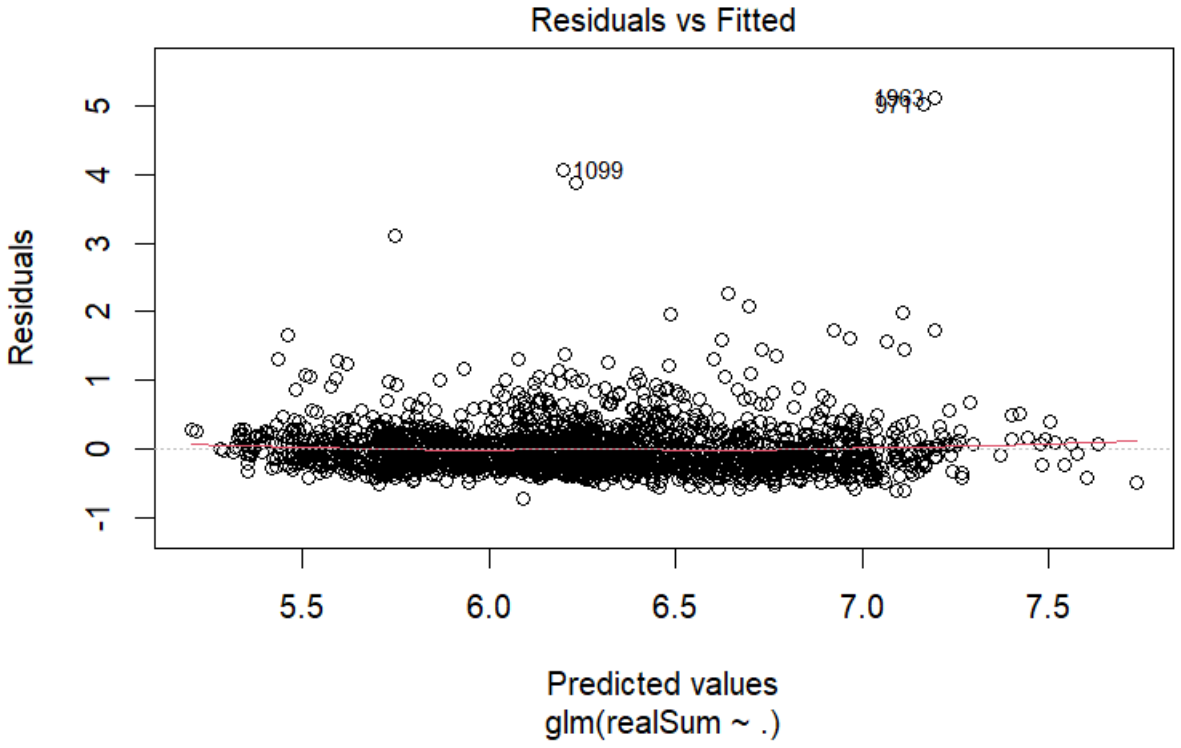


Table B.11: Residuals vs Fitted Plot for the gamma regression model fitted on the subset.



Appendix C

Code C.1 – Importing Libraries and Exploration

```
library(readr)
library(explore)
library(tidyverse)
library(corrplot)

amsterdam_weekdays <- read_csv("~/OneDrive - DePaul
University/FinalProject/MAT360/amsterdam_weekdays.csv")

amsterdam_weekends <- read_csv("~/OneDrive - DePaul
University/FinalProject/MAT360/amsterdam_weekends.csv")

dim(amsterdam_weekdays)
dim(amsterdam_weekends)


amsterdam_weekdays <- amsterdam_weekdays[,-1]
head(amsterdam_weekdays)

amsterdam_weekends <- amsterdam_weekends[,-1]
head(amsterdam_weekends)

amsterdam_week <- rbind(amsterdam_weekdays, amsterdam_weekends)
write_csv(amsterdam_week, "~/OneDrive - DePaul
University/FinalProject/MAT360/amsterdam_fullweek.csv")

dim(amsterdam_week)

amsterdam_week

names(amsterdam_week)

summary(amsterdam_week)
#describe(amsterdam_week)

# Some Exploratory Analysis

print(shapiro.test(amsterdam_week$realSum))
print(summary(amsterdam_week$realSum))
hist(amsterdam_week$realSum)
```

```
print(mean(amsterdam_week$realSum))
print(var(amsterdam_week$realSum))

subSet = amsterdam_week %>%
  filter(realSum <= 4000)
print(summary(subSet$realSum))
hist(subSet$realSum)

hist(amsterdam_week$person_capacity)
hist(amsterdam_week$cleanliness_rating)
hist(amsterdam_week$guest_satisfaction_overall)
hist(amsterdam_week$bedrooms)
hist(amsterdam_week$dist)
hist(amsterdam_week$metro_dist)
hist(amsterdam_week$attr_index_norm)
hist(amsterdam_week$rest_index_norm)
hist(amsterdam_week$lng)
hist(amsterdam_week$lat)

ggplot(amsterdam_week, aes(x=room_type)) +  geom_bar()
ggplot(amsterdam_week, aes(x=room_shared)) + geom_bar()
ggplot(amsterdam_week, aes(x=room_private)) + geom_bar()
ggplot(amsterdam_week, aes(x=host_is_superhost)) + geom_bar()
ggplot(amsterdam_week, aes(x=multi)) + geom_bar()
ggplot(amsterdam_week, aes(x=biz)) + geom_bar()
ggplot(amsterdam_week, aes(x=bedrooms)) + geom_bar()

ggplot(data=amsterdam_week, aes(x=room_type, y=realSum, fill = room_type)) +
geom_bar(stat="identity")
ggplot(data=amsterdam_week, aes(x=bedrooms, y=realSum, fill = bedrooms)) +
geom_bar(stat="identity")
ggplot(data=amsterdam_week, aes(x=multi, y=realSum, fill = host_is_superhost)) +
geom_bar(stat="identity")

names(amsterdam_week)

#Creating a correlation matrix with just the numerical variables and only the normalized versions
of attraction and resturaunt indexes
numerics <- amsterdam_week[, c(1, 5, 9:13, 15, 17:19)]
```

```
matrix.corr = cor(numerics)

print(matrix.corr)
corrplot(matrix.corr)
```

Code C.2 Analysis

```
#Linear Regression

#Taking
print(dim(amsterdam_week))
print(dim(subSet))

reg1.1 <- lm(realSum ~ ., data = amsterdam_week)
reg1.2 <- lm(realSum ~ ., data = subSet)

summary(reg1.1)

summary(reg1.2)

print(logLik(reg1.1))
AIC(reg1.1)
BIC(reg1.1)
print(logLik(reg1.2))
AIC(reg1.2)
BIC(reg1.2)

res <- resid(reg1.2)
plot(fitted(reg1.2), res)


#Multi-Linear Regress (Transforming sumReal)

y = log(amsterdam_week$realSum)
yy = log(subSet$realSum)

print(shapiro.test(y))
print(shapiro.test(yy))
hist(y)
```

```
hist(yy)

reg2.1 <- lm(y ~ .-realSum, data = amsterdam_week)
reg2.2 <- lm(yy ~ .-realSum, data = subSet)

summary(reg2.1)
summary(reg2.2)

print(logLik(reg2.1))
AIC(reg2.1)
BIC(reg2.1)
print(logLik(reg2.2))
AIC(reg2.2)
BIC(reg2.2)

#Percent Change if room_type = Private room
(exp(-3.708e-01) - 1) * (100)
#Percent Change if room_type = Shared room
(exp(-6.091e-01) - 1) * (100)
#Percent Change for every increase in person capacity
(exp( 1.736e-01) - 1) * (100)
#Percent Change if the listing belongs to a host with 2-4 offers
(exp(-3.906e-02) - 1) * (100)
#Percent Change for every unit increase in guest satisfaction
(exp(3.544e-03) - 1) * (100)
#Percent Change for every one increase in bedrooms
(exp(1.427e-01) - 1) * (100)
#Percent Change for every km increase in distance from City's center
(exp(-6.243e-02) - 1) * (100)
#Percent Change for every unit increase in restaurant index of location
(exp(1.507e-02) - 1) * (100)
#Percent Change for every unit increase in longitude
(exp(3.384e-01) - 1) * (100)
#Percent Change for every unit increase in latitude
(exp(-2.378e+00) - 1) * (100)

plot(reg2.1)
plot(reg2.2)
```

```
#Gamma GML

#Taking
print(dim(amsterdam_week))
print(dim(subSet))

gamma1.1 <- glm(realSum ~ .,family=Gamma(link="log"), data=amsterdam_week)
gamma1.2 <- glm(realSum ~ .,family=Gamma(link="log"), data=subSet)

summary(gamma1.1)

summary(gamma1.2)

#Percent Change if room_type = Private room
(exp(-3.777e-01) - 1) * (100)
#Percent Change if room_type = Shared room
(exp(-6.126e-01) - 1) * (100)
#Percent Change for every increase in person capacity
(exp(1.715e-01) - 1) * (100)
#Percent Change if the listing belongs to a host with 2-4 offers
(exp(-4.204e-02) - 1) * (100)
#Percent Change for every one increase in bedrooms
(exp(1.526e-01) - 1) * (100)
#Percent Change for every km increase in distance from City's center
(exp(-6.330e-02) - 1) * (100)
#Percent Change for every unit increase in restaurant index of location
(exp(1.443e-02) - 1) * (100)
#Percent Change for every unit increase in latitude
(exp(-1.893e+00) - 1) * (100)

print(logLik(gamma1.1))
AIC(gamma1.1)
BIC(gamma1.1)
print(logLik(gamma1.2))
AIC(gamma1.2)
BIC(gamma1.2)

plot(gamma1.1)
plot(gamma1.2)
```