**Final Project**

**Projecting AirBnb Prices in New York**

**Group 4**

**Balram Prasad, Matthew Campbell, Claire Zhu, Rafael Silva**

**Table of Contents**

Executive Summary………..…………………………………………………………...…………2

Introduction ……………………………………………………………………………...………..2

Methods

Data Collection………………………………………………………………..…………..3

Data Summary……………………………………………………………………..……...3

Data Preparation……………………………………………………………...………...…4

Modeling……………………………………………………………………..…………....4

Results…………..…………………………………………………………………..……………..9

Discussion………………………………………………………………………………………..11

Recommendations ………………………………………………………………..……………...12

Limitations and Potential Improvements……..………………………………..………………...12

Appendix ………………………………………………………………………………...……....13

Works Cited……………………………………………………..……………………………….23

**Executive Summary:**

This project will examine and analyze New York City AirBnb rental data from the year 2017 in order to better understand and accurately predict one-night AirBnb prices. We intend to present our findings to AirBnb, so that they can use them to help new AirBnb property owners better price their rental units. This will involve the examination of categorical predictors, like location and specific features of the listing (ex. Type of AirBnb, Room type within AirBnb, etc), as well as numeric predictors (number of bedrooms and bathrooms). Our variable of interest, also described later as our “response variable,” is price, and its nature and relation to other predictors will be analyzed throughout this report. Our prediction consists of two models: 1) a Quadratic Discriminant Analysis model serving as a preliminary model to roughly predict the price range of a prospective listing. 2) a Multivariate Linear Regression model serving as a model that produces a more specific point estimation.

First, correlations between predictors and the response variable will be examined with a Color Correlation Matrix, a table of correlation values, and series of scatter plots and bar graphs. Next, after selecting the appropriate predictors, we will create a model that incorporates them and illustrates their relationships with the response variable price. This will consist of linear regression (including standard MLR, forward selection, backward selection, and hybrid-stepwise selection), and logistic regression. Ideally, our model will help decrease price disparity among AirBnb rental units in New York City, and will provide new AirBnb owners with a useful means to properly price their units.

**Introduction:**

AirBnb is a privately held company headquartered in San Francisco. The company operates an online platform that works to connect property owners with those looking to rent property for a certain amount of time. The company essentially allows property owners to use their property as a hotel, and gives owners the freedom to set their own price.

While having the ability to set the price gives property owners a set of freedom, it does also present an issue. AirBnb’s platform operates as a competitive market place. When listing a property the owner wants to be sure that their listed price is competitive. Because AirBnb provides revenue to the property owner, this would classify as a business problem.

Our team assumed that our client is a property owner in New York. Our team was tasked with developing a model using AirBnb data to predict the price range of the given property.

The business objective of this project is to give our client a comprehensive model that will help them determine what price they should list their property at. For example, if a prospective host is planning to list a new property on Airbnb; however, due to the lack of experience, his/her pricing of the property is not consistent with the market price. Such mispricing can result in either a loss of potential clients and profits loss for the host or overpricing and a loss of client welfare. In either way, such mispricing causes inefficiency in the Airbnb housing market and hampers the sustainable growth of the company in the long run. To avoid this inefficiency, our model will help Airbnb improve the pricing efficiency for its housing market.

One potential cost for Airbnb to ensure the most efficient pricing is the cost associated with data collection and model updating. Because the economic condition of the region may change and pricing is dynamic, our model works the best if the training data incorporates the latest market information and reflects the most recent market sentiments. Therefore, to maintain the predictive power of the model, Airbnb should release its listings in a timely manner and incorporate new listings in the training model.

**Methods:**

***Data Collection:***

The data was received from AirBnb.com[[1]](#footnote-0). AirBnb did not officially disclose how they compiled the data, but given the data is from their own platform one would assume that that they drew from their internal databases.

***Data Summary:***

|  |  |  |  |
| --- | --- | --- | --- |
|  | type | statistics | descriptions |
| Name | qualitative |  | name of the property listed |
| Experiences\_offered | qualitative |  | whether the host has previous experience in Airbnb hosting |
| Host\_name | qualitative |  | name of the host |
| Host\_since | qualitative |  | date when the host made the listing (in the date-format) |
| Host\_total\_listings\_count | numeric | range: 0 - 2271  (highly right-skewed) | total number of listings that the host has |
| Host\_identity\_verified | categorical | 2 levels | whether the listed host is a verified user or not  25482 verified, 25183 non-verified |
| Neighbourhood\_cleansed | categorical | 225 levels | the neighborhood the property is located in |
| Neighbourhood\_group\_cleansed | categorical | 5 levels | which borough (in New York City) the property is located in |
| Room\_type | categorical | 3 levels | the type of room offered  26349 are entire home/apartment, 23163 are private room, 1153 are shared room |
| Property\_type | categorical | 33 levels | 40871 out of 50665 are apartments |
| Accommodates | numeric | range: 1 - 16  (right-skewed) | the capacity of the listed property |
| Bathrooms | numeric | range: 0 - 16.5  (highly right-skewed) | number of bathrooms |
| Bedrooms | numeric | range: 0 - 11  (highly right-skewed) | number of bedrooms |
| Beds | numeric | range: 0 - 21  (highly right-skewed) | number of beds |
| Bed\_type | categorical | 5 levels | 49683 out of 50665 are real beds |
| Price | numeric | range: 0 - 999  (highly right-skewed) | listing price |
| Extra\_People | numeric | range: 0 - 300  (highly right-skewed) | The cost associated with accommodating one additional guest |

***Data Preparation:***

Before running our model, we had to make sure that our data was formatted properly. There were two actions that we had to take to make our data workable. First, there were a few columns in our data that had NA’s. In order to run our regressions model, we had to remove these, and replaced all NA’s with 0. Secondly, the price was formatted with a $ at the start and was in decimal format. Because R recognizes values with symbols as strings, we had to remove the decimal and the $ sign. After doing this we then converted the column into a numeric value.

**Modeling:**

***Quadratic Discriminant Analysis:***

Predicting an absolute price number might be unrealistic, so one way to approach a pricing model is to utilize a classification technique and predict the price range a listing might fall into.

We first group the prices into 4 categories based on the quantiles such that the 4 groups are 1) the minimum to the first quantile, 2) the first quantile to the median, 3) the median to the third quantile, and 4) the third quantile to the maximum. This serves as the response variables we try to predict. Because the response variables are available for our dataset, exploratory techniques such as clustering are not appropriate. Instead, classifications such as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) serve our purpose better.

On one hand, LDA is better when there are relatively few observations so that reduced variance is critical in the analysis. Furthermore, LDA classifier results from assuming that the observations within each class are normally distributed with a class-specific mean vector and a common covariance-variance matrix. On the other hand, QDA is better when the training dataset is large and the assumptions for LDA are not met.

After some exploratory analysis (appendix), we conclude that QDA is a more appropriate model for our dataset for the following reasons: 1) our dataset has high dimensionality, 2) the correlation matrix within each class is significantly different from each other. As a result, QDA is used as our classification technique.

First, we allocate 80% observations to the training dataset and 20% to the testing dataset, which is used later to assess the predictive power of this model. Second, we use all the predictors (explanatory variables) to build the model based on the training dataset and use the model to predict the price range for the testing dataset. Third, we evaluate the result and performance of the QDA model by a confusion matrix and key parameters such as accuracy, specificity, and sensitivity. (under Results)

***Linear Regression Analysis:***

The data set initially contained 50665 observations of 20 variables. Price was used as our response variable; for any observation, price gives the dollar cost associated with renting that unit for one night. As our goal is to assist a new AirBnb owner with accurately pricing his or her rental unit, we removed the variables host\_total\_listings count, host\_since, and days\_since, which describe the number of times the host has previously listed the unit for rental, and how long the host has been affiliated with the AirBnb rental system (host\_since provides a specific date and days\_since provides a number of days since affiliation). Additionally, the variables experiences\_offered and host\_name were removed, because all entries for experiences\_offered were “none,” and because the name of the host should not logically impact the pricing of the unit. Neighbourhood\_Group\_Cleansed was used as a kind of location variable, and room\_type was used to describe the different kinds of rental units observed in the data set. The variables Neighbourhood\_Cleansed and Property\_Type were thus removed in favor of using the simpler variables Neighbourhood\_Group\_Cleansed and Room\_Type. And, as this report does not include text analysis, the text descriptions (given by the variable “name”) which describe the rental units are also left out of the modelling and analysis.

Regarding exploratory analysis of variable importance, scatterplots were used to examine the relationships between price and the numeric predictors (accommodates, bedrooms, bathrooms, beds, extra\_people) and bar graphs were used to examine the relationships between price and the average values of the categorical predictors (host\_identity\_verified, room\_type, neighbourhood\_group\_cleansed, and bed\_type). These visualizations are given in the appendix section of the report. For the scatterplots, corresponding regression equations are given in the plot titles. These equations were derived from the coefficient estimates from the lm() function for each plot, which are displayed in the appendix. Upon examination of each of the five individual simple linear regression models, it is clear that all five of the numeric predictors (Accommodates, Bathrooms, Bedrooms, Beds, and Extra People) are have a positive linear relationship with the response variable Price. This is sufficient for our exploratory data analysis, as we can conclude that these five numeric predictors play a notable role in the Price of AirBnb units in New York City, and we can plan on examining them more formally later on during the modelling process. Regarding the four selected categorical predictors, Host Identity Verified, Room Type, Bed Type, and Neighbourhood Group Cleansed, we can see that the variable Host Identity Verified has very little impact on price; that is, for both the True and False subsets of the data (being subsets in which the Host’s Identity was verified prior to rental, and those in which it was not, respectively), the mean value of price is almost identical. In other words, there does not seem to be a notable difference in price between these two categories, so the variable Host Identity Verified will not be seriously considered in modelling. Additionally, there seems to be some variation in price for the different levels of the variable Bed\_Type: units with the level “Real Bed” seem to have a substantially higher price than the four other kinds of units. This variable will be considered in modelling, but does not clearly appear to be extremely significant. Lastly, the variables Neighbourhood Group Cleansed and Room Type almost certainly have an impact on the response variable Price, as their different levels have different (mean) prices. So, in summary, we will initially consider the numeric predictors Accommodates, Bathrooms, Bedrooms, Beds, and Extra People, and the categorical predictors Neighbourhood Group Cleansed, Room Type, and Bed Type for modelling.

Next, the predictors must be examined for any signs of multicollinearity. Regarding the three numeric variables Accommodates, Bedrooms, and Beds, we see that there exists some serious multicollinearity (see Correlation Color Matrix and Correlation Values given in Appendix). Examining the individual correlations in the table provided in the appendix, we see that the correlation values between these variables are well above the desired range. Rationally, all three of these variables describe a similar characteristic of the an AirBnb unit: how many people can stay there. So, considering both the multicollinearity seen in the bolded red entries in the correlation table and the relevant dark red squares in the Color Correlation Matrix, and the individual correlations between predictors and the response variable, we decide to remove both Bedrooms and Beds from our modelling. Next, despite having a low-moderate correlation with price (0.269803), the variable Bathrooms seems to be another good variable to examine when analyzing price; this is because of the strength of the initial relationship between this variable and price given in the scatter plots, and because it logically follows that units with more bathrooms should tend to be more expensive, both because they are larger and more accommodating. This variable has a somewhat red square describing its relationship with price in the Color Correlation Matrix, indicating a moderate (positive) relationship between Lastly, the variables Extra People and Days Since both have very low correlations with the response variable Price, although they do not appear to have any serious issues with multicollinearity; this is seen in their correlations of 0.095683 and 0.011429, respectively, with price. So, after some analysis, we choose to consider only the numeric variables Accommodates and Bathrooms, and the categorical variables Room Type and Neighbourhood Group Cleaned, in our attempt to better understand AirBnb pricing in New York City.

In R, we model the relationship between the response variable Price and the predictors. For our initial model, we used the formula Price = Accommodates + Bathrooms + Neighbourhood Group Cleansed + Room Type. This was checked using forward, backward, and hybrid-stepwise selection methods; all three of these methods gave us the same model (Price = Accommodates + Bathrooms + Neighbourhood Group Cleansed + Room Type + Bedrooms + Beds). However, due to the previous examination of multicollinearity among predictors, we remove both Bedrooms and Beds from the model and instead just use Accommodates to represent the number of guests that can realistically stay at an AirBnb. The summary statistics for the model are given in the appendix. Also, the results from an F-Test for the model (with the full model being price = accommodates + bathrooms + room type + neighbourhood group cleansed, and the null model being price = 1) indicate that the four predictors we have selected are very important to the model. The F-Test statistic is 5170, with a p-value that is less than 2e-16, which is clearly very significant. Thus we do not reject the null hypothesis that at least one of these predictors is equal to zero, and instead choose to keep the four predictors in the model.

Finally, the residuals must be examined to verify the use of a linear regression model. Looking at a plot of the residuals against fitted values (see appendix), we see that the residual mean remains generally constant around zero, with slight fluctuation as price increases. For the purposes of our analysis, this is not terribly concerning. This is because most of the AirBnbs in the data set offer only a few bedrooms and bathrooms (e.g. 1-4), only accommodate a small number of guests (e.g. 1-4), and tend to have prices below or around $250. We also see that nearly all of the residuals are densely packed toward the left, indicating that most of the AirBnbs have a relatively low value of price (there are few AirBnbs that are priced above $250, shown by the lack of points on the right hand side of the residual plot). With this in mind, the residuals for the majority of AirBnbs (being the smaller, cheaper ones, given by the points toward the left of the plot) seem to have a more balanced relationship with the line y=0, suggesting the assumption that the residuals have a constant mean of zero is generally met. Going further, the QQ plot shows that the residuals are essentially normally distributed except toward the right side of the plot. As mentioned before, some sporadic variation toward the expensive, larger end of AirBnbs is expected, and the fact that a few of these units have large, positive residuals should not derail the model for the remaining 50,000 (approx.) observations. Also, the plot of Residuals against Leverage Values (also in the appendix section) shows no clear outliers in the data set, suggesting that we do not need to remove any observations. As previously described, the model was examined for multicollinearity, and the variables beds and bedrooms were already removed. So, the linear regression model assumptions generally hold true, and the linear model is a valid process for analyzing the data. Logistic regression was initially considered for modelling, but was not used as the response variable is not categorical. A final data visualization of the predicted values of price plotted against the true values of price is given in the appendix. As the line cuts through the middle of the data, and most of the points are densely centered around the line (especially towards the bottom right), we conclude the linear regression model is a good fit for the data and can be used to better understand and predict the value of the response variable price.

**Results:**

***Quadratic Discriminant Analysis:***

Our starting hypothesis is that the Quadratic Discriminant Analysis (QDA) serve as a classification technique that, given all the explanatory variables, classifies the price range of a listing with a high level of accuracy. At the very least, its classification accuracy should be higher than a random guess of 25%. After running the model, we show that the QDA model overall exhibits an accuracy of 49%, supporting our initial hypothesis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Predictions | | | | |
| Actual observations |  | $0 - $69 | $69 - $106 | $106 - $175 | $175 + |
| $0 - $69 | 803 | 1538 | 109 | 72 |
| $69 - $106 | 349 | 1542 | 333 | 345 |
| $106 - $175 | 81 | 685 | 537 | 1145 |
| $175 + | 38 | 219 | 239 | 2101 |

Table 1. Confusion matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | $0 - $69 | $69 - $106 | $106 - $175 | $175 + |
| sensitivity | 0.3183981 | 0.6002336 | 0.2193627 | 0.8090104 |
| specificity | 0.5489887 | 0.4547377 | 0.5783039 | 0.3822788 |
| accuracy | 0.491614 | | | |

Table 2. Evaluation based on sensitivity, specificity, and accuracy.

Sensitivity is calculated as the probability of the QDA model predicting that a property falls into a specific price range while the observed price is indeed within that range. From Table 2, we conclude that the model is best at correctly predicting the price range for the $175+ group and has the worst performance for the $106 - $175 group.

Specificity is calculated as the probability of the QDA model predicting that a property does not belong to a specific price range while the observed price is indeed not within that range. Also from Table 2, we conclude that the model is best at avoiding false prediction for the $106 - $175 group, even though the specificity is similar for each category.

Accuracy is calculated as the proportion of true predictions and measures how often the classifier is correct. From Table 2, we see that the QDA model makes accurate predictions 49% of the time.

***Linear Regression***

We started this project with a general hypothesis that Capacity (accommodates), Bathrooms, Neighbourhood Group and Room Type are all significant variables when predicting the price of a AirBnB listings.

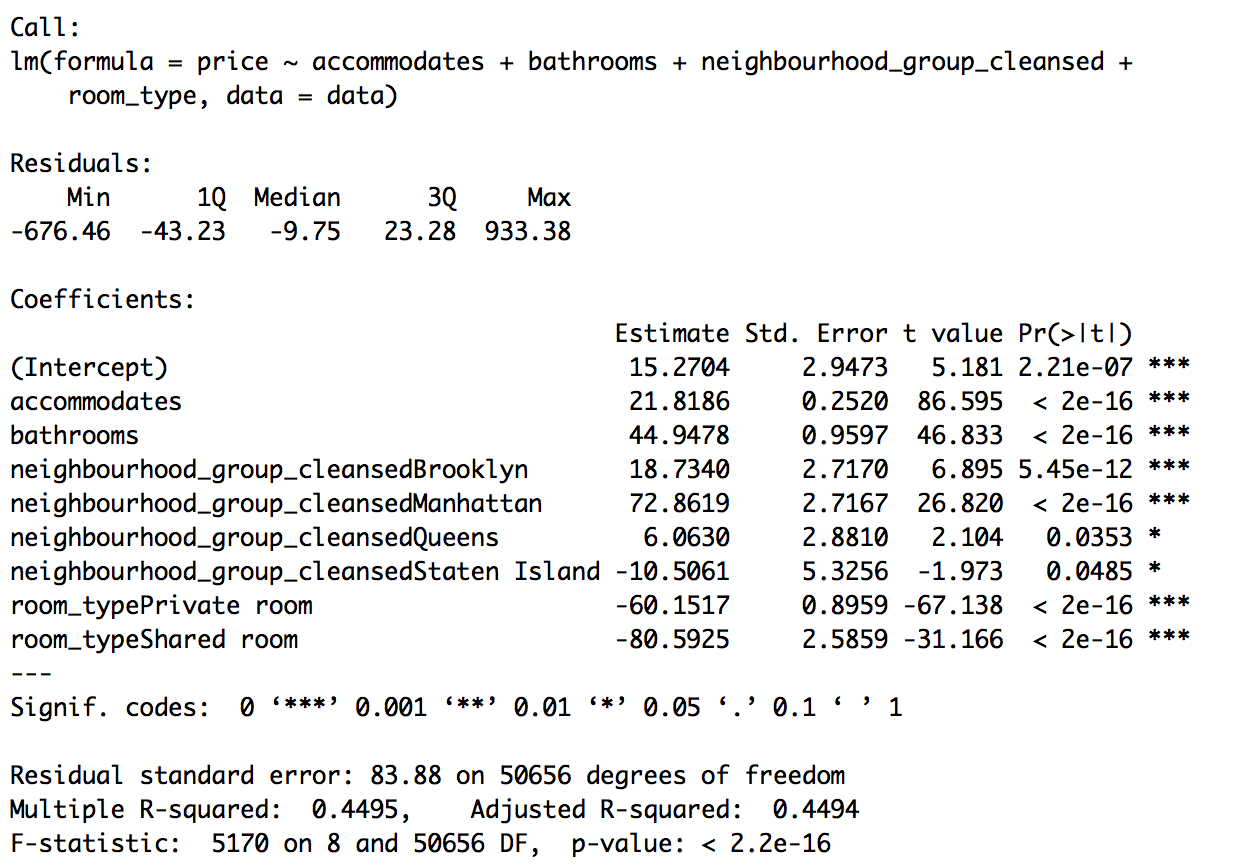


Table 3. Statistical Summary from Linear Regression Model with Price as Response Variable.

After running our Linear Model we observe that all mentioned variables have a good level of significance. Capacity, Bathrooms and the dummy variables Manhattan, Private Room and Shared Room are the most significant while the dummy variables Queens and Staten Island are only significant at the 95% confidence level. The resulting R-squared of 0.4495 is low. However, all individual p-values and the model p-values are small enough to give this model the confidence to be significant. The y-intercept of 15.27 of accounts for the base level price for all listings, serving as the benchmark of a default property (that rents out the entire house/apartment in Bronx).

***Discussion:***

***Linear Regression Analysis***

When pricing listings in AirBnb we will start from a baseline of $15. Listing in Manhattan will be price significantly higher than a similar listing located in another Borough. The coefficient 72.86 shows that on average a Manhattan location will bump the price by $72.86. On average a Brooklyn location will bump the price by $18.73. We can feel very confident about those two impacts in the total price of the listings. We can feel less confident about the impact of a Queens or Staten Island location on the total price of the listing, but they suggest a bump of $6 and a decrease of $10 respectively.

If the listing is just a private room within a residence, the price is decreased $60 on average and if it is only a share of that private room, the price is decreased by $80 on average.

The capacity and number of bathrooms are clearly important when predicting price. On average price is increased by around $22 for every person the listing can host. For the number of bathrooms, a $44 increase should be expected for each unit in the listing.

***Quadratic Discriminant Analysis***

The sensitivity numbers for the $69 - $106 group and the $175+ group are relatively high, (0.6 and 0.8, respectively) suggesting that if a property is predicted to belong to one of those two groups, the owner/host should be relatively confident that the intrinsic value of the property indeed belongs to the predicted group.

On the other hand, the sensitivity numbers for the $0 - $69 group and the $106 - $175 group are relatively low, (0.3 and 0.2, respectively) suggesting that if a property is predicted to belong to one of those two groups, it is reasonable for the owner/host to be suspicious about whether the QDA model is actually reflecting the market value of the property.

The accuracy for the overall QDA model is 0.49, a number that seems relatively small at the first sight, but because we have 4 price ranges, a random guess should give an accuracy of 25%. In contrast, the QDA model is two times more likely to accurately categorize a listing than a random guess, suggesting that the predictors are significant in terms of estimating the actual market price of an Airbnb listing.

As a result, our QDA can serve as a preliminary model to roughly predict the price range of a prospective listing. And it is especially effective for the $69 - $106 and $175+ group. If a prospective host would like a more specific point estimation, the host is encouraged to use our linear regression model.

**Recommendation:**

Our recommendation is that AirBnb adopt our linear regression model to recommend pricing to their customers. While our R2 is very low (~0.50), we believe that this model would still provide value to AirBnb, as they currently have no system in place. Most importantly, our model would only serve as a recommendation. Given the context we believe that our model is strong enough to warrant adoption, and believe that it would provide value to AirBnb consumers.

**Limitations and Potential Improvements:**

The effectiveness of our model relies on the dataset accurately reflecting the current market price. Therefore, it needs to be constantly updated and trained by the most recent listing data if our client adopts our pricing model to make recommendations to prospective hosts.

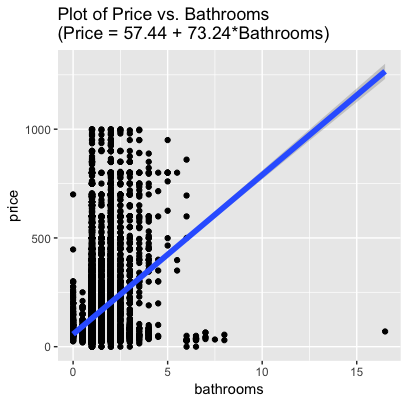
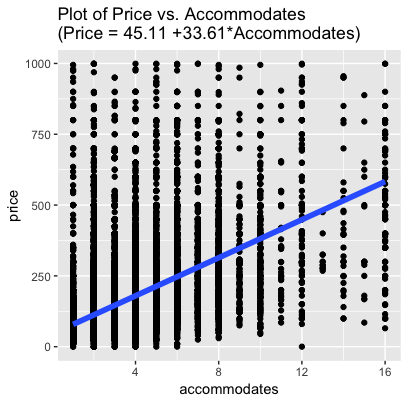
Another major limitation in our model is that it does not account for seasonality. It would be expected that this would play an important factor in prices. During holidays and summer it would be expected that there would be a greater demand for AirBnb’s, thereby pushing the price up.

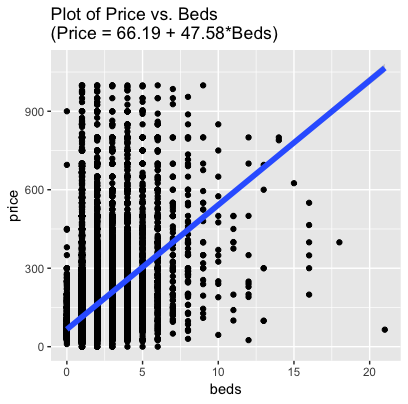
As discussed in the modelling section, the model does not perform as well for more expensive, larger AirBnBs with more or better features. Thus, our model is effective in predicting the one-night rental price for AirBnB units with a reasonable number of bedrooms and bathrooms (1-4), and inside a reasonable price level (below $250), but should not be extrapolated to predict the rental price of, perhaps, an AirBnb with 10 bedrooms and bathrooms. And, most importantly, as our model R-squared was 0.4495, we must conclude that over half (50%) of the variation in price is not explained by our set of predictors. So, future users of our model might consider incorporating additional parameters to describe and understand price. These could be variables that describe parking (parking spaces available, cost to park at the AirBnb), current economic conditions (GDP, inflation, Consumer Price Index), or other unforeseen factors.

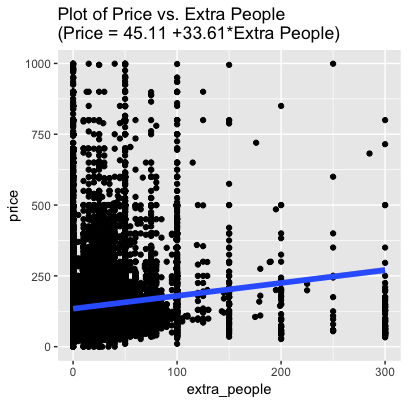
**Appendix:**

Section One: Exploratory Data Analysis: Scatter Plots and Bar Graphs

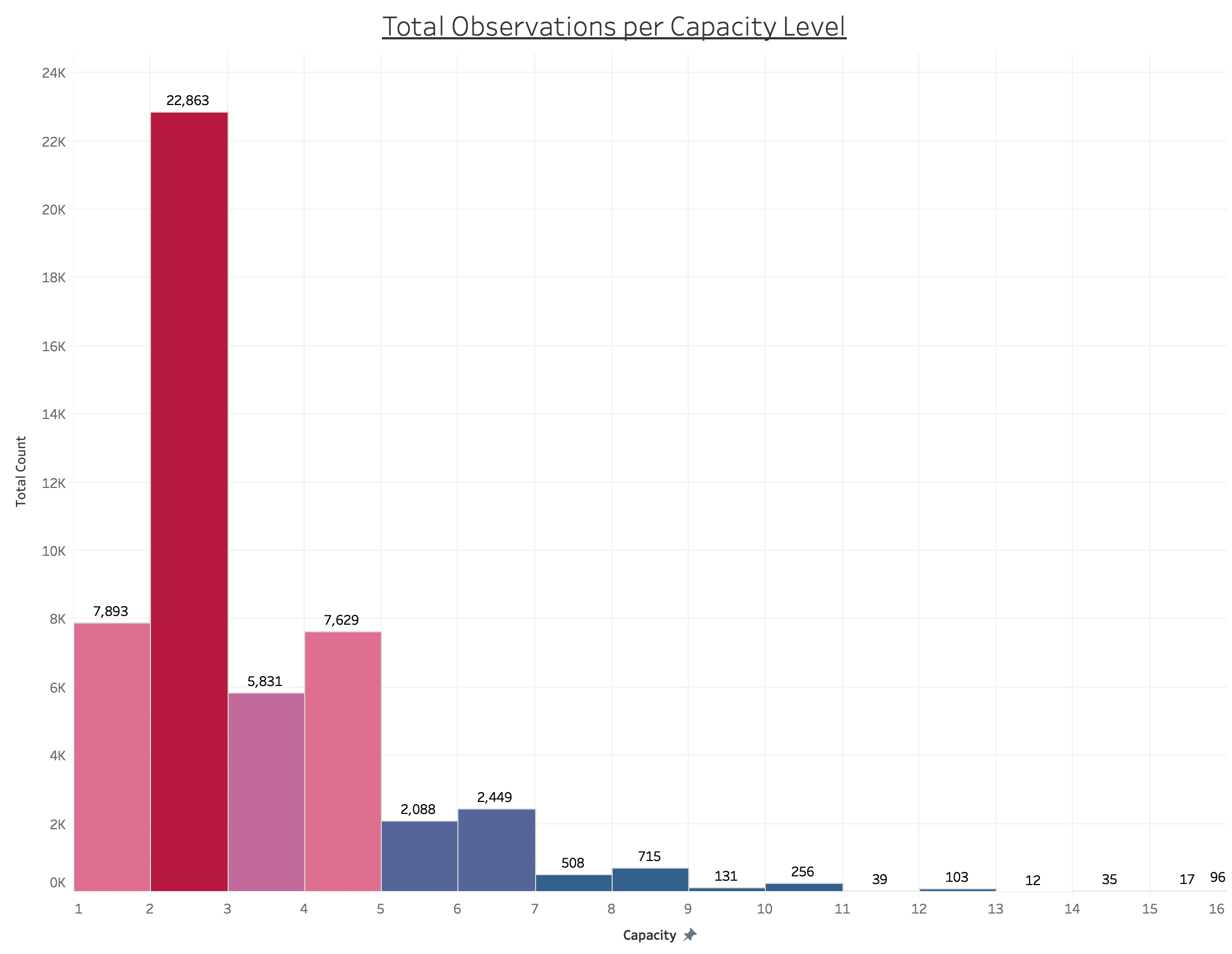
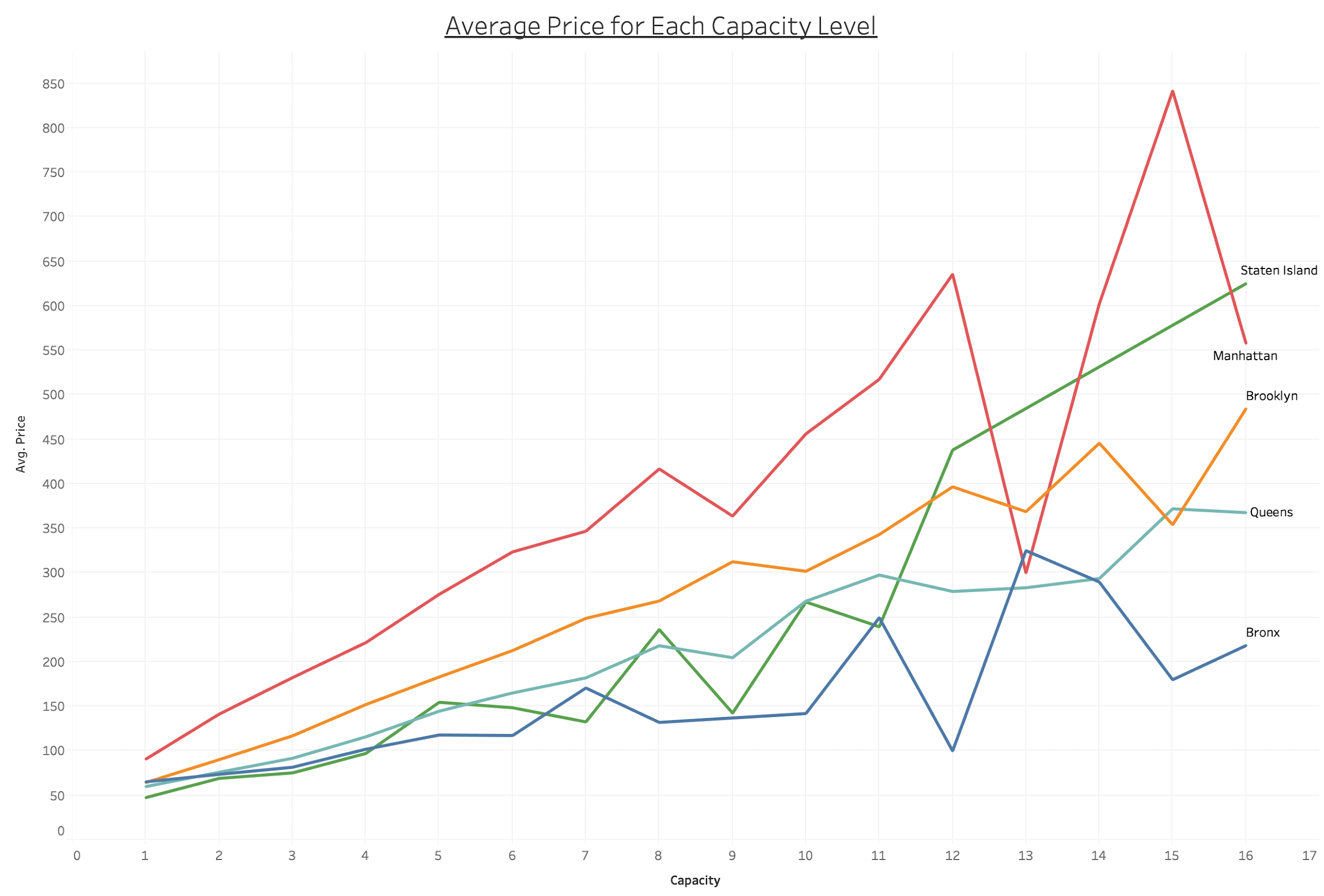
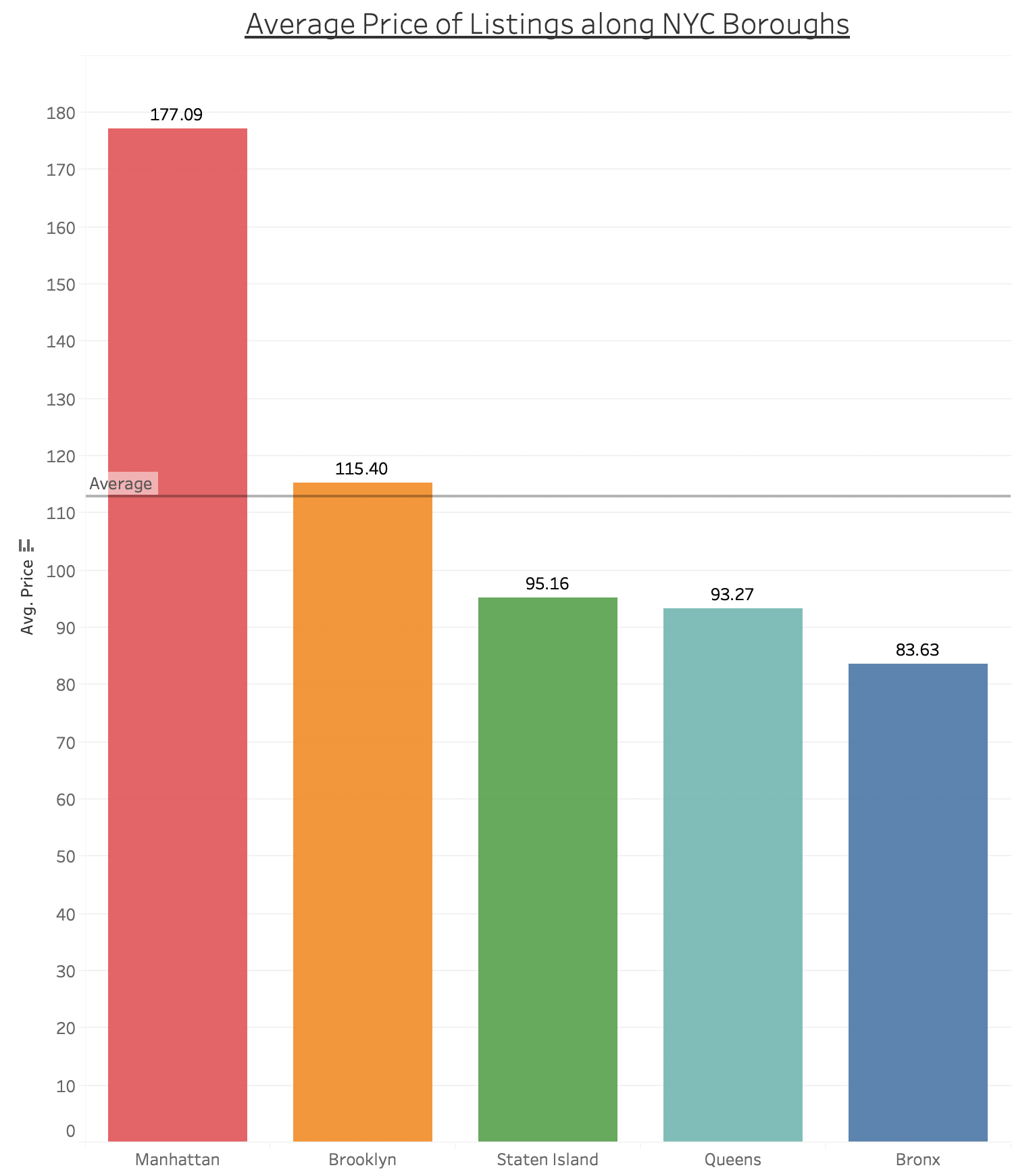
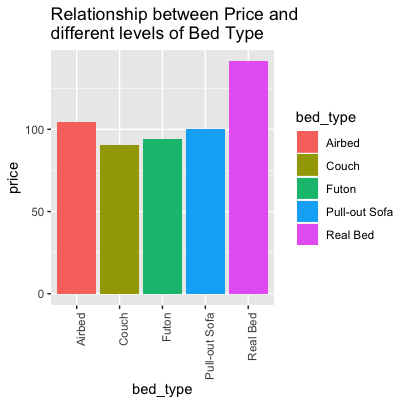
\*\*\*Bar Graphs give the mean price on the y-axis for each categorical variable\*\*\*

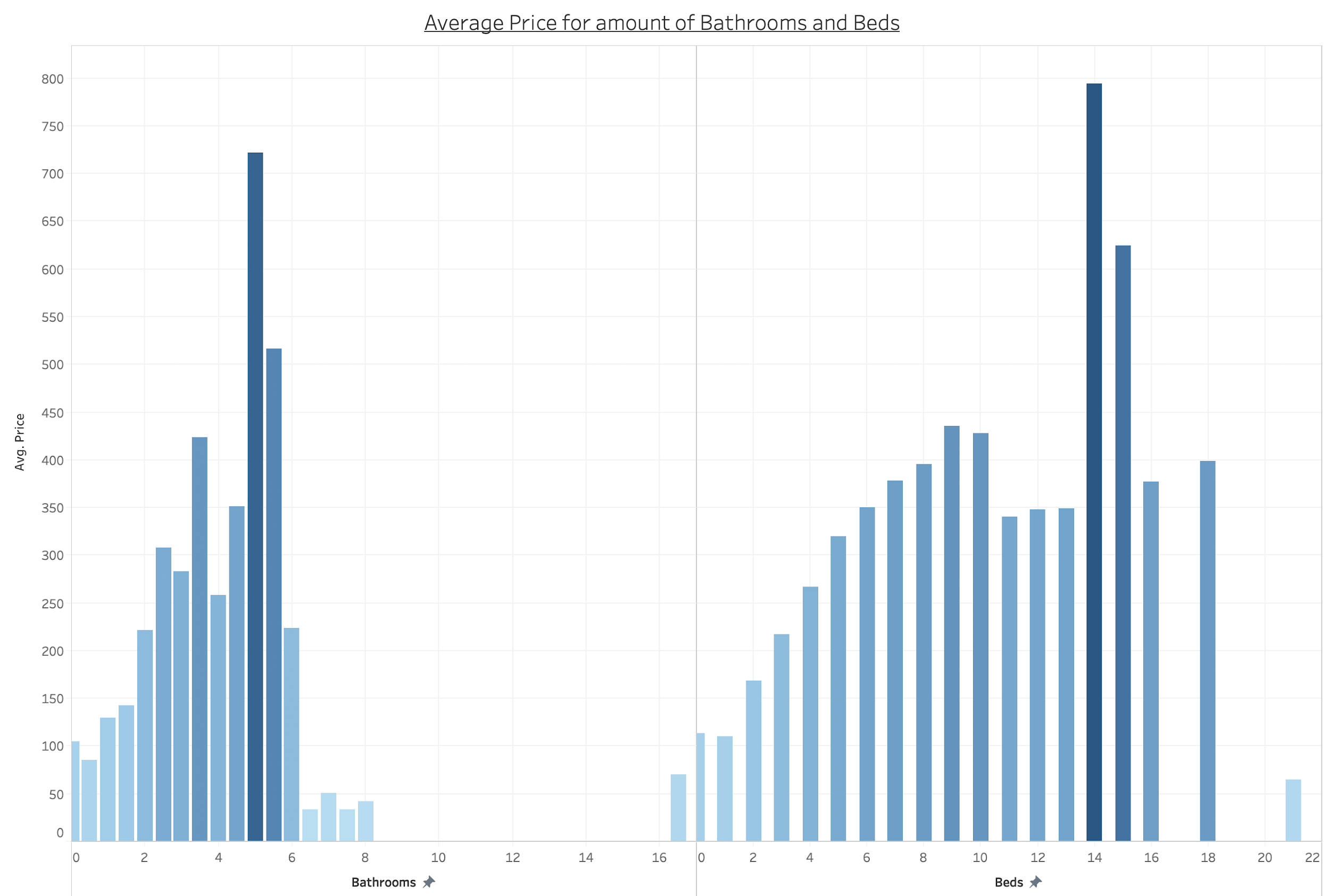
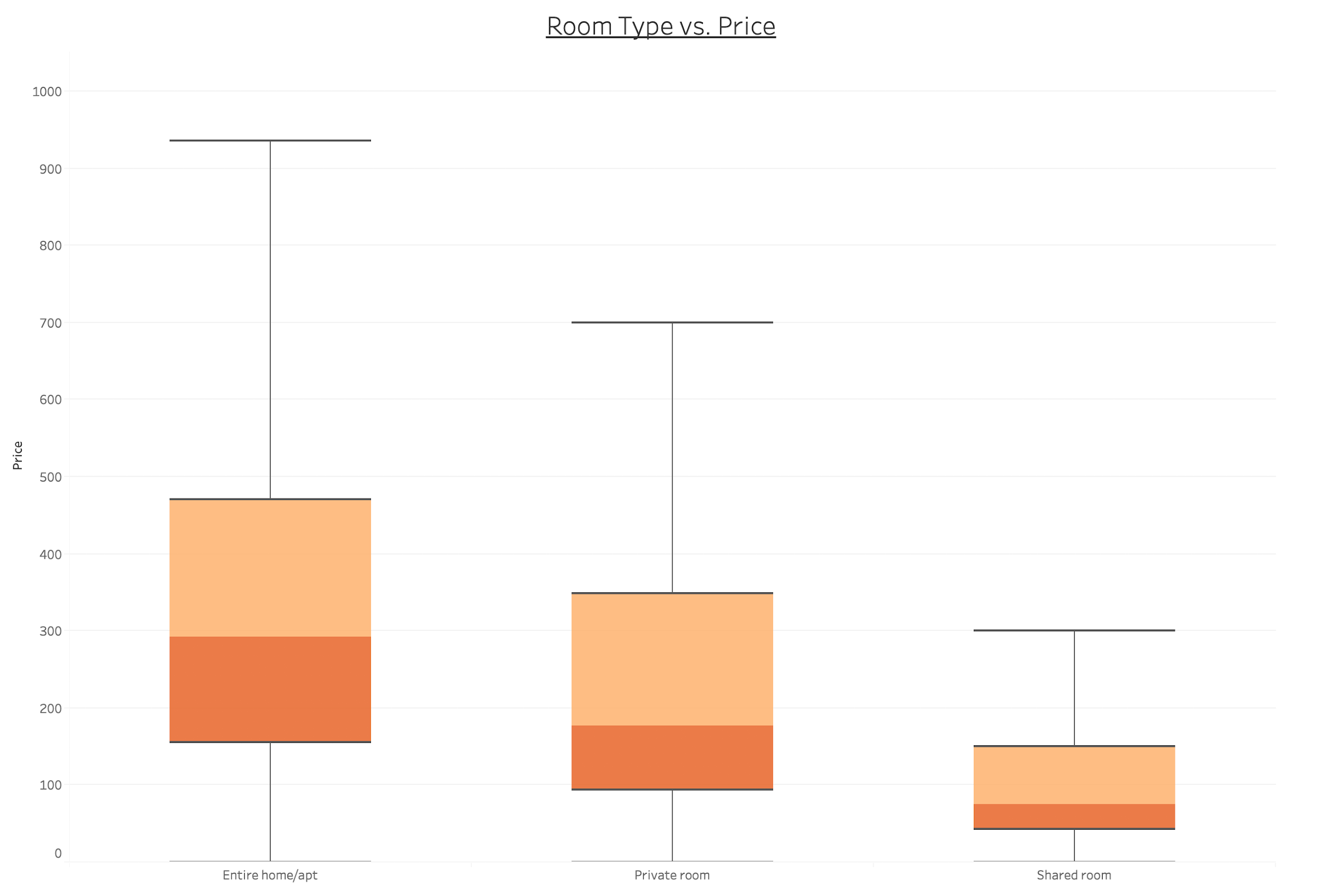










****

**Coefficient Estimates for the Five Linear Relationships above (for each scatterplot):**

\*\*\*These are for the 5 individual simple linear regression models given in the above scatterplots--this is not a comprehensive model summary, or a summary of one model containing all five of the below predictors. This information was obtained from the lm() and summary() functions in R.\*\*\*

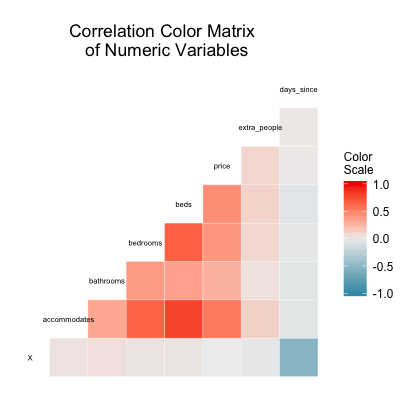
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Intercept Estimate | P-value (of Intercept) | Coefficient Estimate | P-value (of Coefficient) |
| Accommodates | 45.11 | <2e-16 | 33.61 | <2e-16 |
| Bathrooms | 57.44 | <2e-16 | 73.24 | <2e-16 |
| Bedrooms | 68.96 | <2e-16 | 61.27 | <2e-16 |
| Beds | 66.19 | <2e-16 | 47.58 | <2e-16 |
| Extra People | 134.1475 | <2e-16 | 0.4539 | <2e-16 |

**Individual Variable Correlations:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accommodates | Bath-  rooms | Bed- rooms | Beds | Price | Extra people | Days  since |

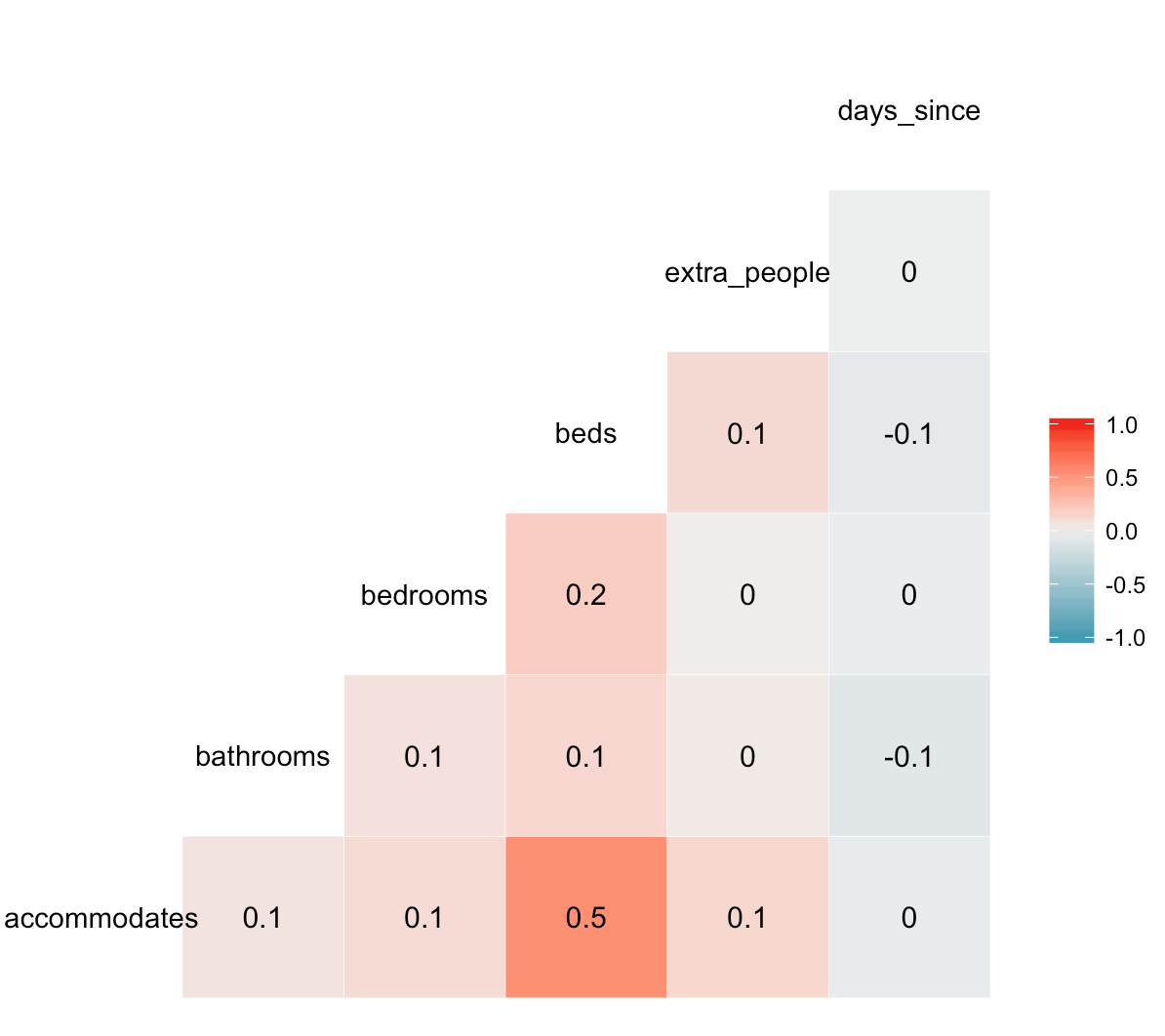
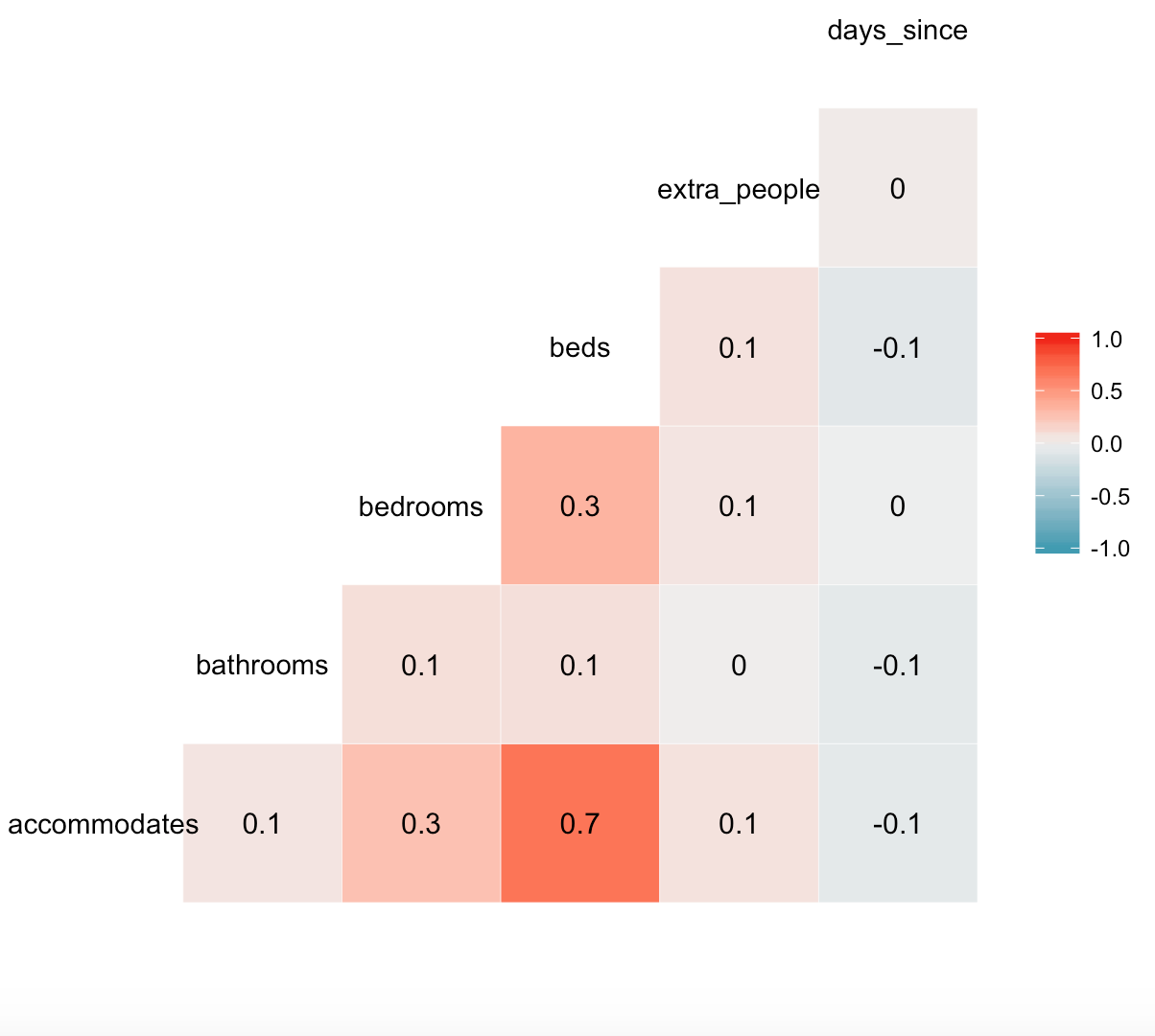
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Accommo-dates | 1.00000000 | 0.31727 | 0.64227 | 0.79773 | **0.543663** | 0.11659 | -0.04545 |
| Bathrooms | 0.31727 | 1.00000 | 0.37742 | 0.34864 | 0.269803 | 0.04612 | -0.04934 |
| Bedrooms | **0.64227** | 0.37742 | 1.00000 | 0.66339 | **0.399078** | 0.08948 | -0.02817 |
| Beds | **0.79773** | 0.34864 | **0.66339** | 1.00000 | **0.442743** | 0.11234 | -0.05265 |
| Price | 0.54366 | 0.26980 | 0.39907 | 0.44274 | 1.000000 | 0.095683 | 0.01142 |
| Extra people | 0.11659 | 0.04612 | 0.08948 | 0.11234 | 0.095683 | 1.00000 | 0.02025 |
| Days since | -0.04545 | -0.04934 | -0.02817 | -0.05265 | 0.011429 | 0.02025 | 1.00000 |

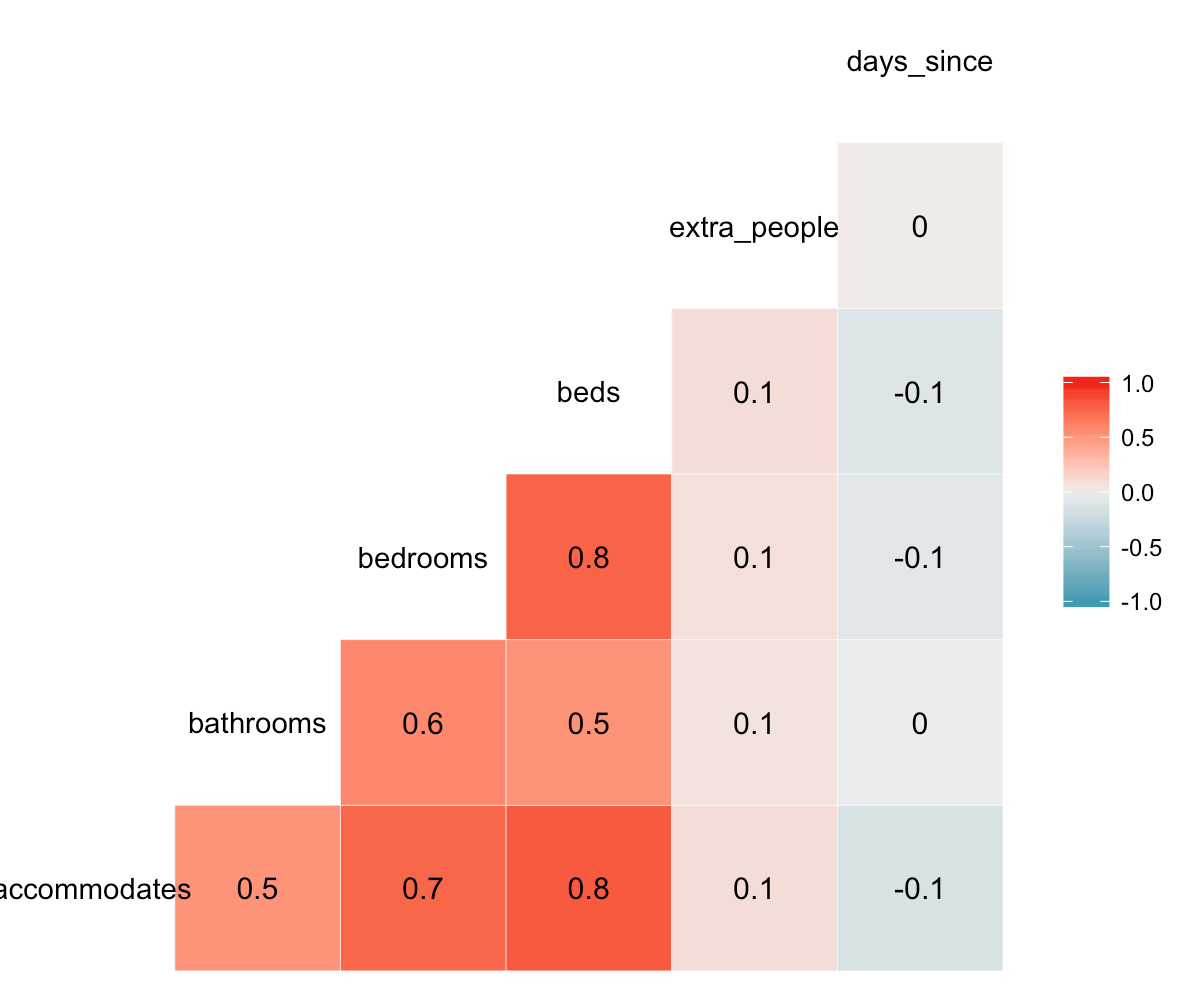
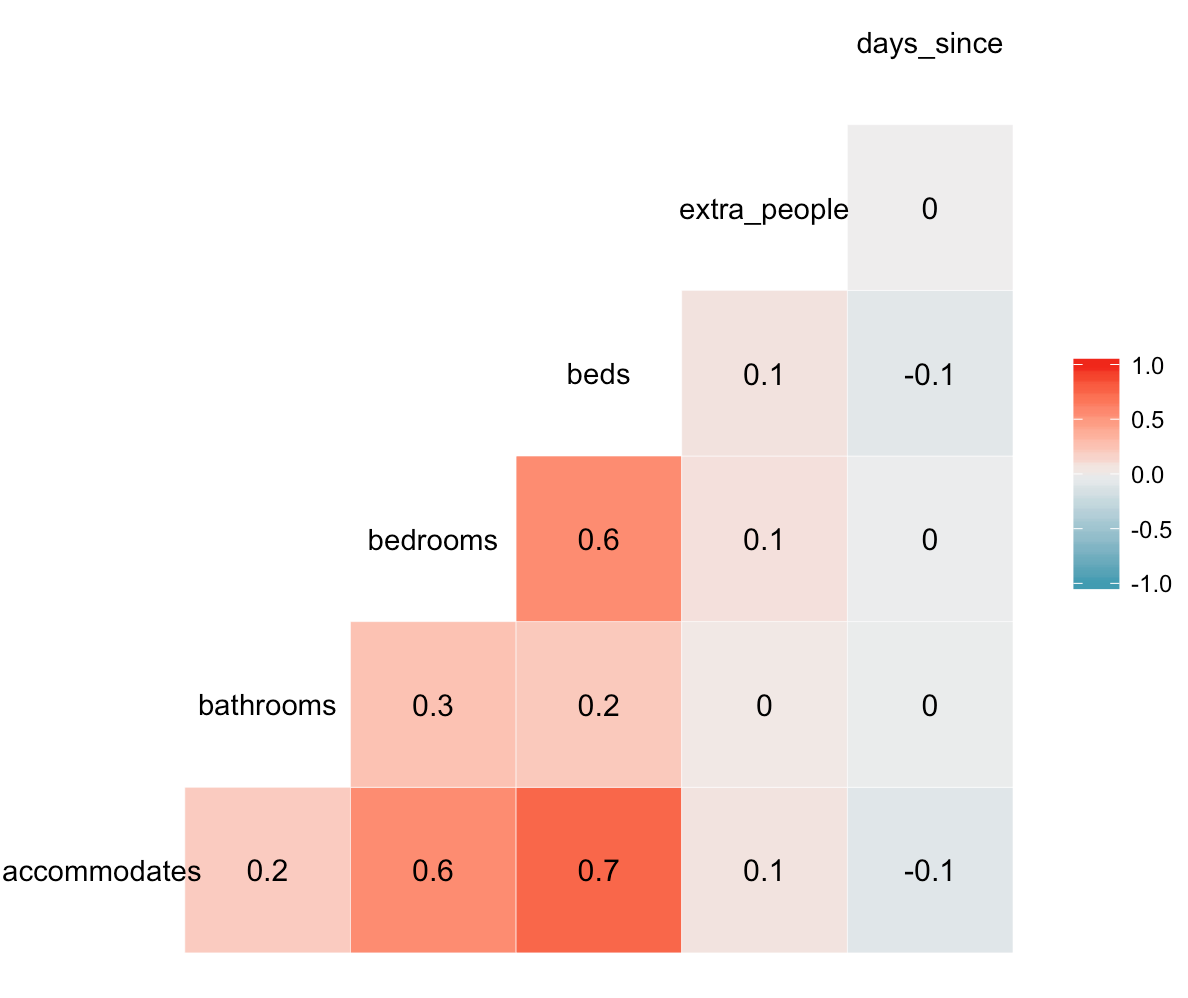
**Correlation Color Matrix:**



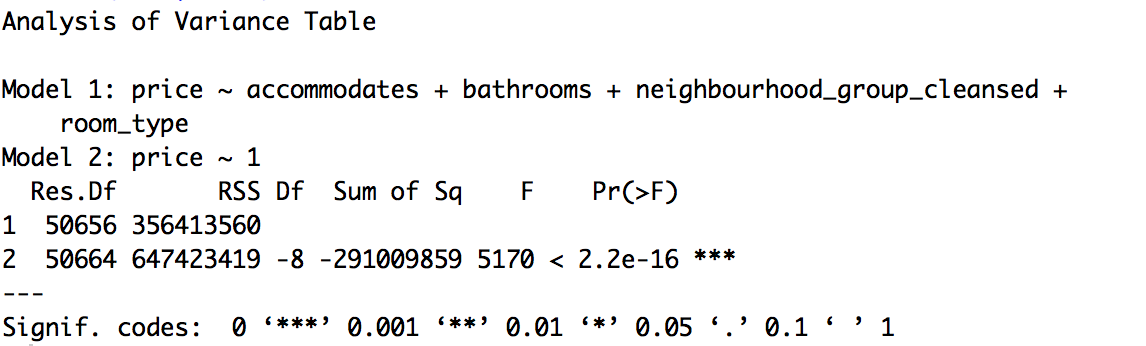
**LDA Assumptions Check - covariance matrix for each class (price range)**

Tables from left to right, top to bottom are the covariance matrices for the $0-$69, $69 - $106, $106 - $175, and $175+ price groups, respectively.

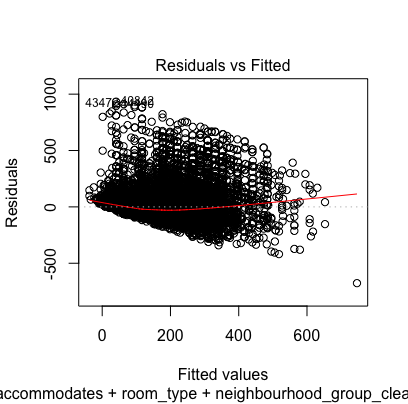




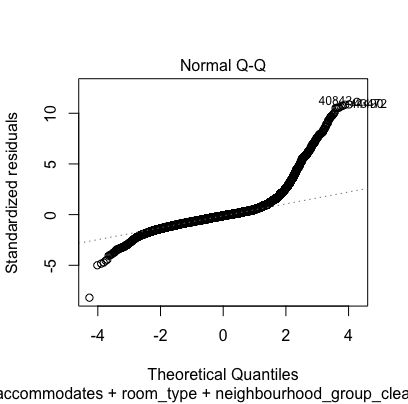
**Analysis of Variance Table:**



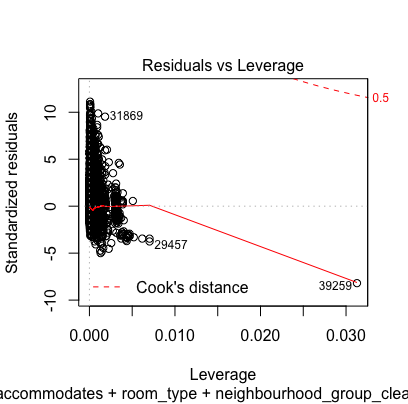
**Plot of Residuals Against Fitted Values:**

****

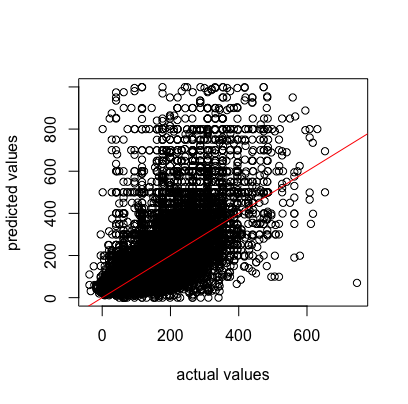
**QQ plot of Residuals:**

****

**Plot of Residuals vs. Leverage Values**

****

**Final Data Visualization of MLR Model:**

****

**Works Cited**

* “Inside AirBnB: Adding Data to the Debate.” AirBnB, 03 June, http://insideairbnb.com/new-york-city/. Web.

1. Inside Airbnb Website. http://insideairbnb.com/new-york-city/ [↑](#footnote-ref-0)