**Predictive Analysis of Price on Berlin Airbnb Listings**

Vivek Advani, Samrita Ghosh, Ruying Situ, Aniruddh Vashi, Nishit Patel

California State University, East Bay

**Abstract:** Airbnb, Inc.. , is an American online marketplace and hospitality service brokerage company based in San Francisco, California, United States. Members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. In this paper we used the data of over 22.6k listings of Berlin as of Nov 2018. In the process of analysis, we discarded many variables through preprocessing and exploring the dataset and we build a linear regression model to express the relationship between expected price and number of predictors for a Airbnb listing in Berlin. We also used regression trees to predict the price of the listings through which we got important rules for price prediction of the listings. After measuring our models in terms of error metrices we compared them according to the RMSE (Root Mean Square Error) values that we got by running the models on unseen data.

**Motivation:** This project explored the utility of Airbnb Berlin dataset for gaining insights of the pricing strategy and to remain competitive in the market. We found the dataset to be like what we worked on in class and we could use our course knowledge to leverage useful information from the dataset. This would also give us an opportunity to implement our course knowledge to answer specific questions regarding the listing pricing. The specific

questions which we tried to answer using our analysis were:

A) What are the key features that determine the pricing of a listing in Berlin?

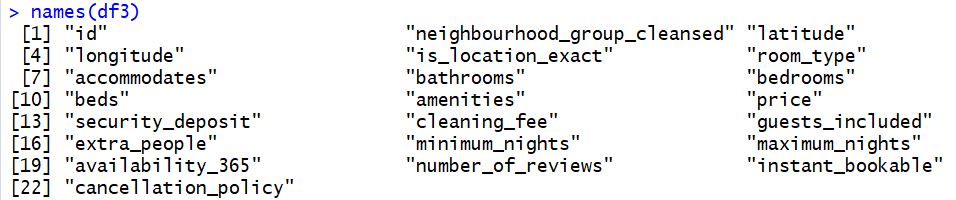
B) Is distance from the center of Berlin a factor in determining the price of a listing?

**Introduction:** Airbnb is an American enterprise which maintains a lodging plat form, enabling sets to rent and share their properties in terms of a reliable, review based and active community. Airbnb benefits both users and renters, while properties are rented to increase renter’s income, whereas renters get the opportunity to reside in cozy accommodations which are beneficial even from a value -for-money perspective. The best approach to understand a specific technique or a method is to recreate and practically try it. Having learned the concepts of data mining in depth this semester, we tried to widen the practical skills by performing this project. This project analyzes Airbnb listings in the city of Berlin to better understand how different attributes such as amenities, location, amongst others can be used to accurately predict the price of a new listing that is optimal in terms of the host’s profitability yet affordable to their guests. This model is intended to be helpful to the internal pricing tools that Airbnb provides to its hosts. The analysis begins with exploring and examining the data to make necessary transformations that can be conducive for a better understanding of the problem at large while helping us make hypotheses. We built and trained a model based on the descriptions and guidelines given in this report and interpret the results to understand what factors influence the pricing of the listings which will eventually help the hosts to decide the rent of their listings. The project then concludes with a discussion on the business implications, associated risks and future scope.

**Data Handling:**

Data Source : Kaggle (<https://www.kaggle.com/brittabettendorf/berlin-airbnb-data#listings.csv)-22552> rows and 96 columns.

Final columns selected:

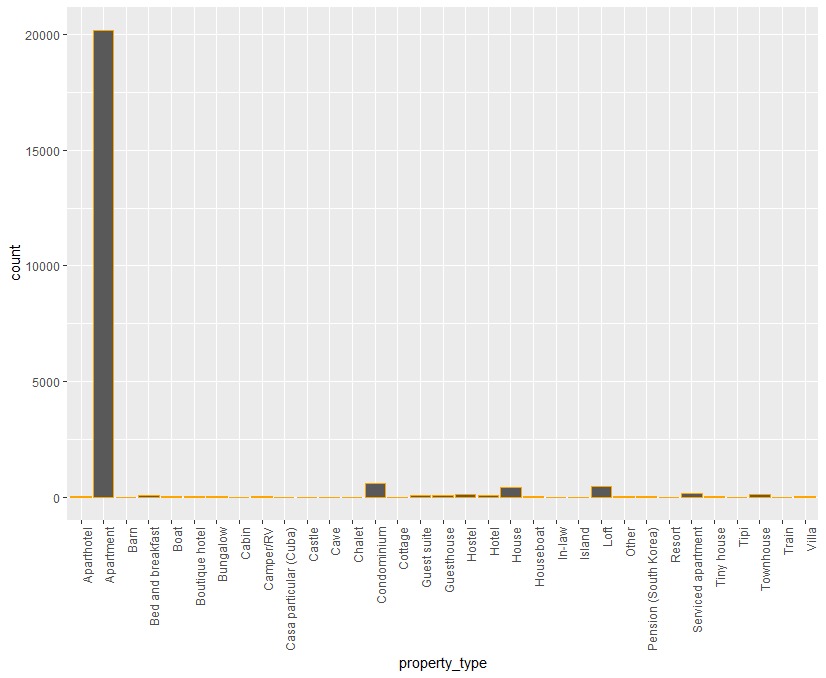


Text and unwanted columns: We removed 63 columns based on the data in the columns, mainly all text data and irrelevant data such as summary, name, access and columns with irrelevant information such as scrape id, last scraped, etc. We also removed columns with no data such as thumbnail \_URL and medium\_URL.

Missing Data: We removed 8 columns with missing data and NA data.

We also removed few columns based on the nature of the data in the columns.

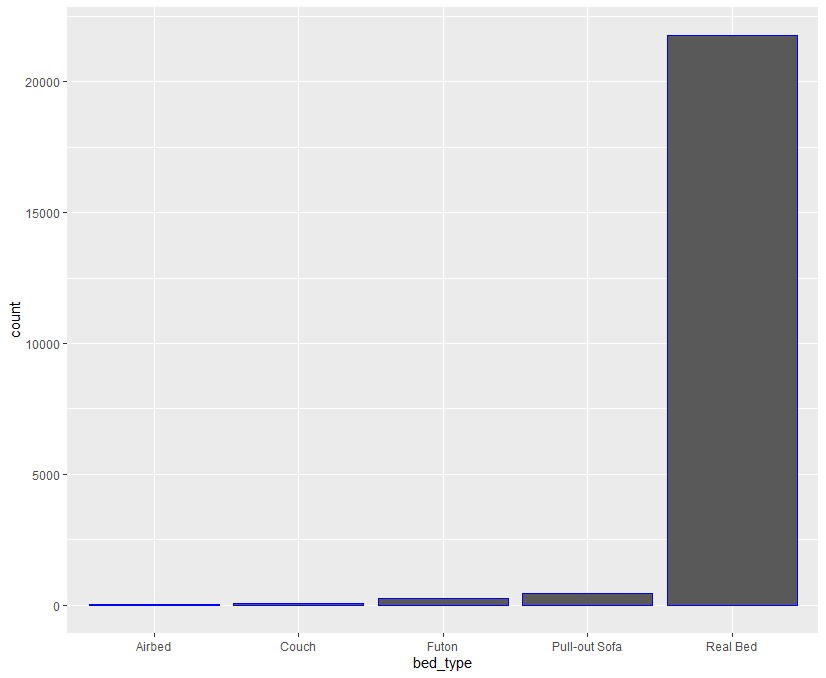
* At first, we plotted a bar graph for the property type column to get an idea of the different property types available in Berlin:



Here we can see more that 95% of the listings are apartments, hence we could say that this column won’t be contributing towards our analysis. Hence, we removed this column from out final dataset for regression.

* We plotted the graph for the column of bed type column:

It can be seen that more than 95% of the data is real bed, hence we removed this column from the dataset for regression.



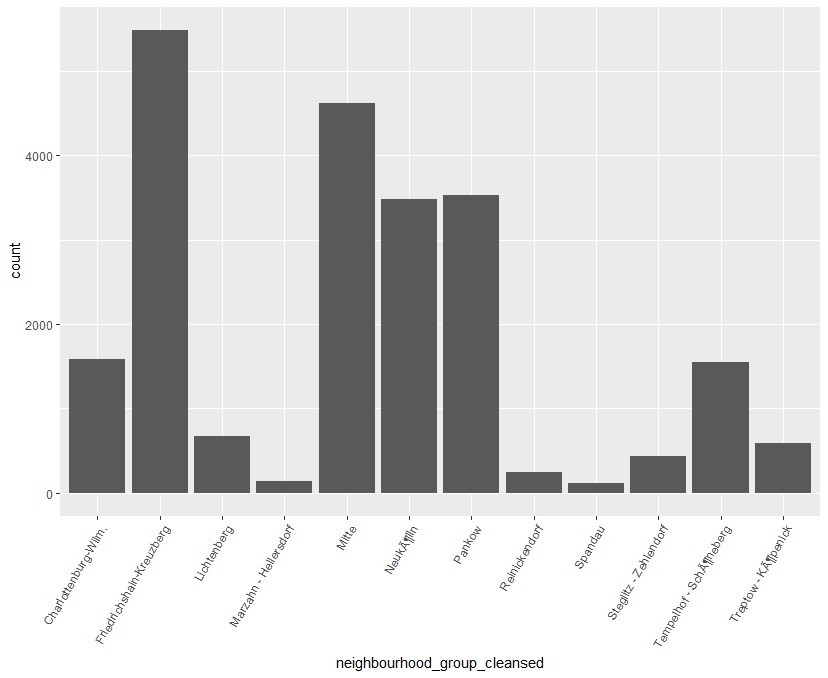
We tried to analyze, using a heatmap whether the availability\_30,availability\_60,availability\_90 and availability\_365 are correlated.

A screenshot of a cell phone

Description automatically generated

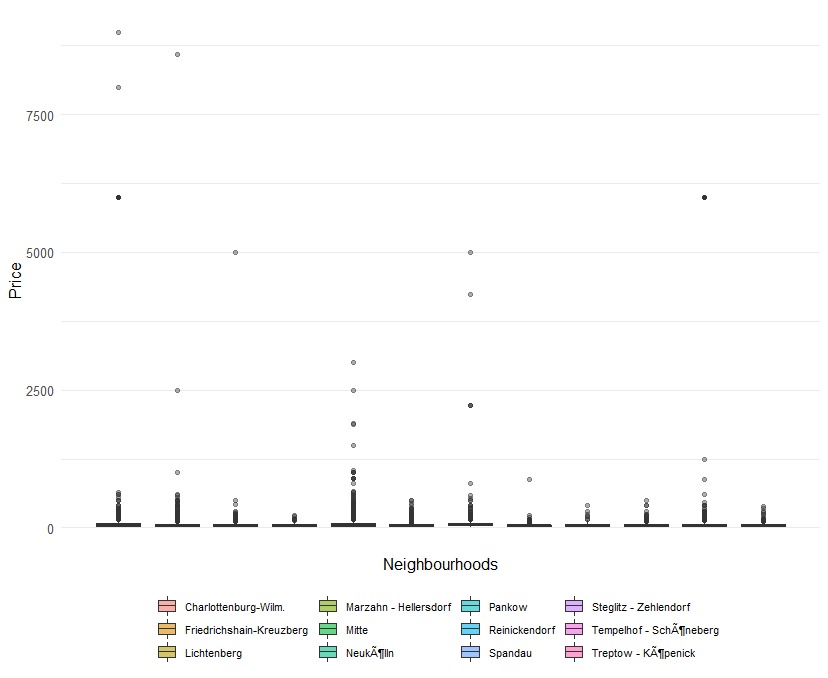
**Exploratory Data Analysis:**

We explored the number of listings there are in each neighborhood to know which neighborhoods more listings have as compared to others.



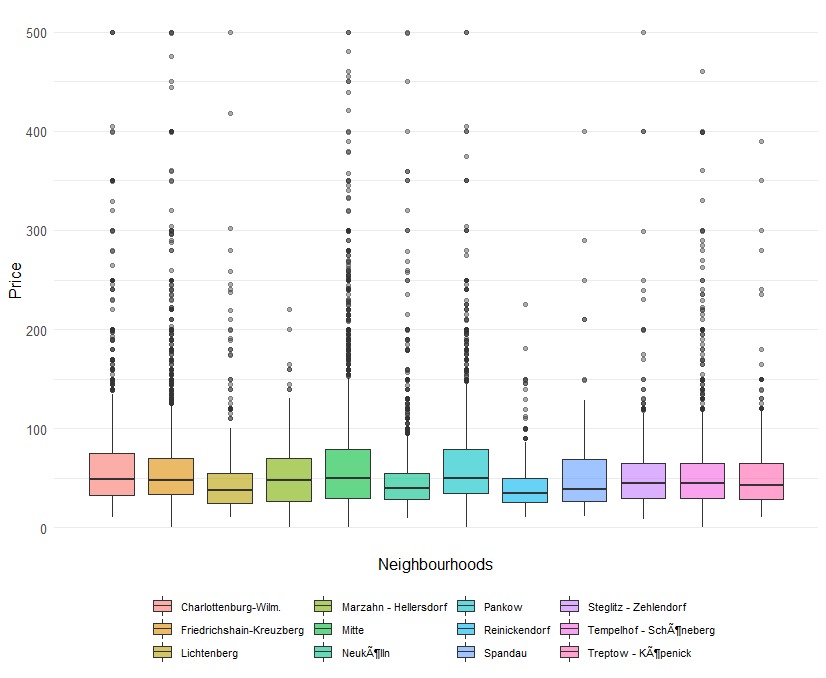
Here we can see that there are few neighborhoods with very high number of listings, this can help the hosts to know the competition in their neighborhood.

Pricing trend according to neighbors:

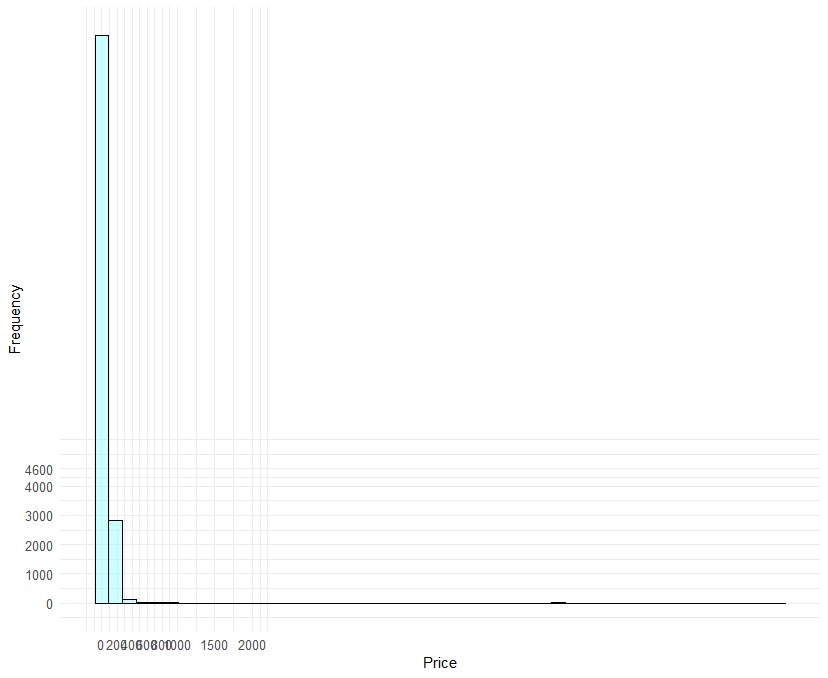


Here we can see that the price value is mostly between 0 to 500, hence in order to view our data better we zoom the graph between 0 to 500 price values.

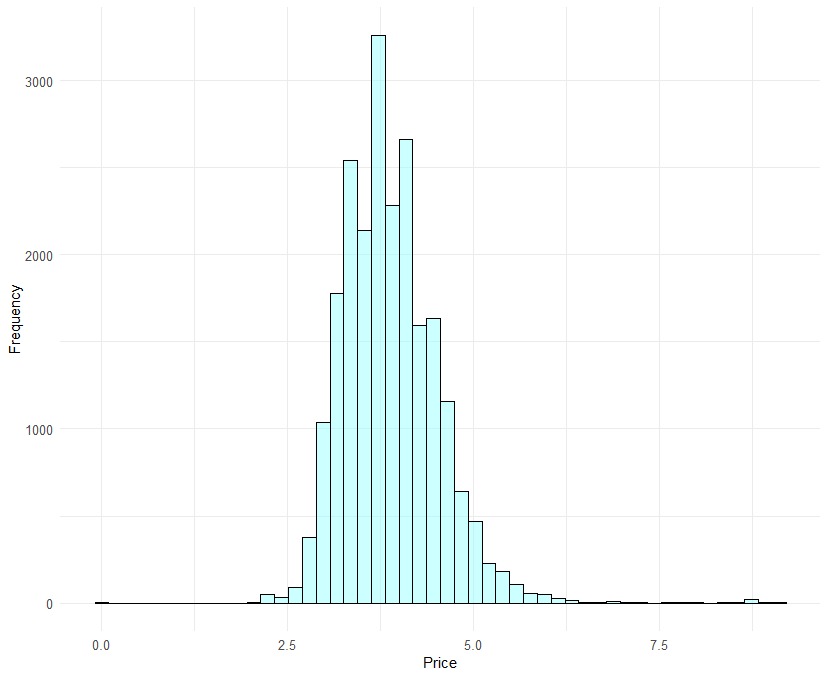
We can see from the graph below that the median price has a similar trend for all the neighborhoods whereas few neighborhoods have some listings which have higher price compared to the median. Those listings stand out from other listings in the same neighborhood in terms of price.



Plotting the frequency graph of price:

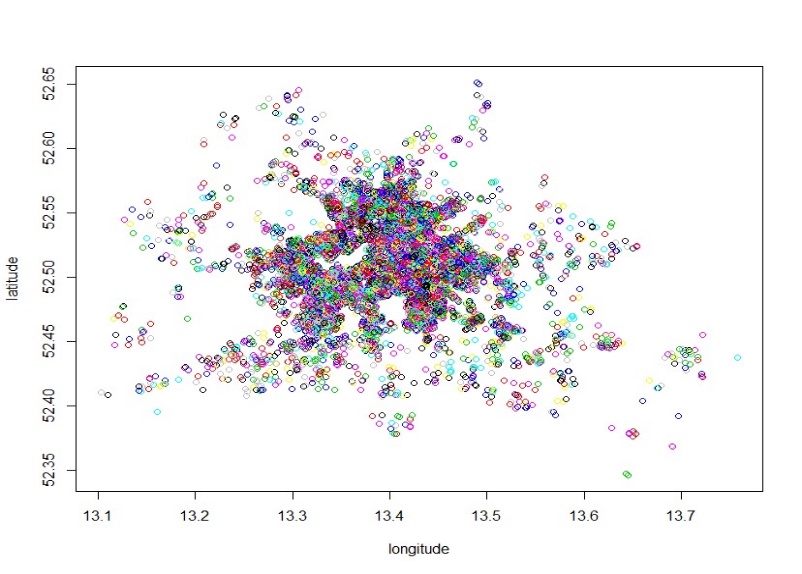


Here we can see that the price is highly right skewed. Hence we took the log of price to see the frequency.



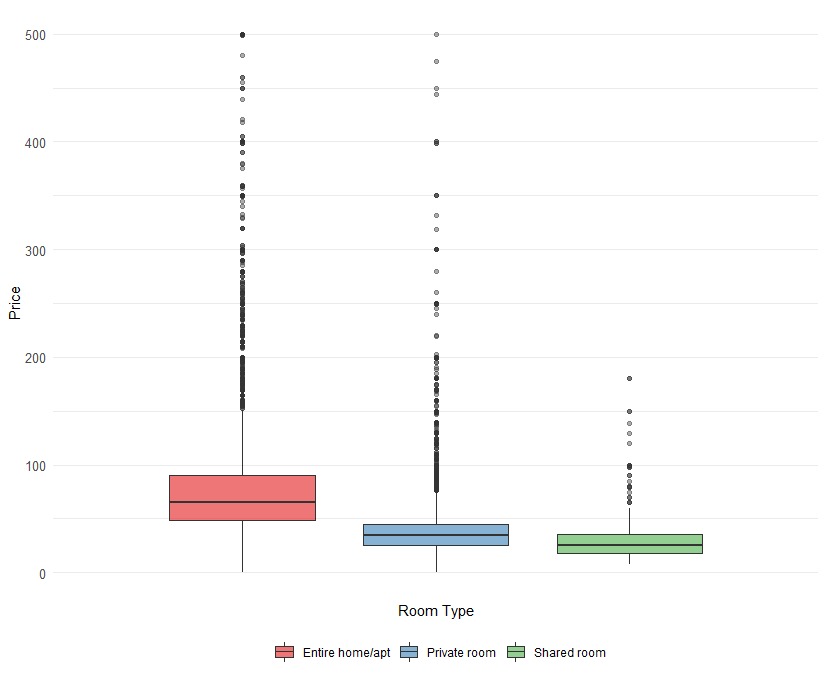
Here we can see there Are few outliers above 7.5, which we also noticed in the boxplot of Price VS Neighborhoods.

Price variation according to location:



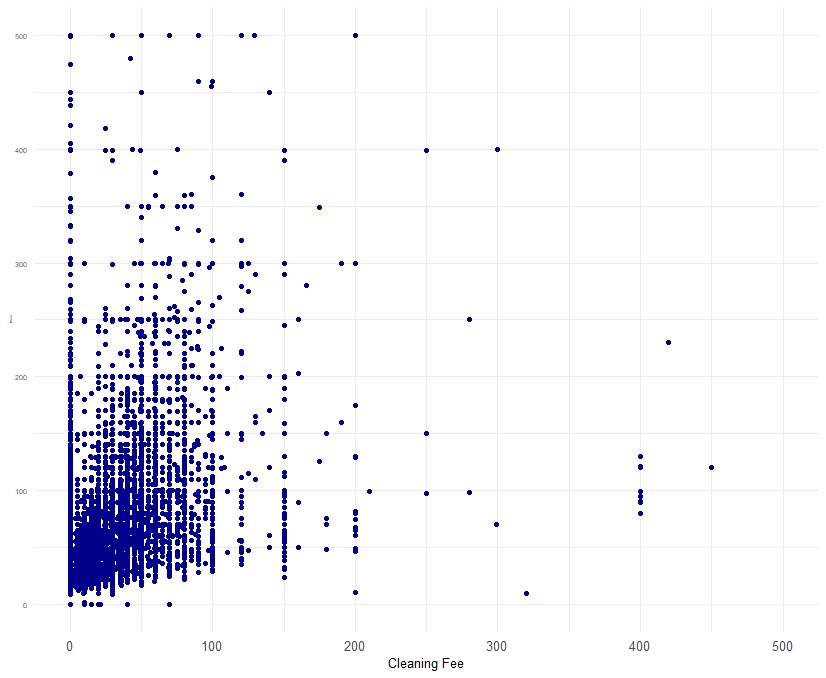
Here all the points in the plot represent each listing location and the color scheme shows us the price(darker the color, higher the price). We can see that most of the listings are located nearby each other and the price is higher of the listings falling in the middle of the graph.

We now tried to explore if the type of the room rented affects the pricing of the listing using the below graph:



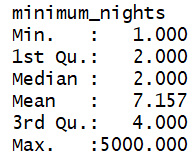
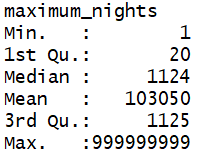
Here we can see that the price is highest for the entire apartment listings followed by private room listings followed by shared room .

We wanted to know whether the listings charging a cleaning fee have higher price, hence we plotted cleaning fee vs price graph. From the graph below we can see that the price of the listing is not affected by the cleaning price. Listings having same amount of cleaning fee have various prices.

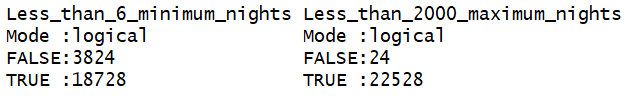


**Feature Engineering:**

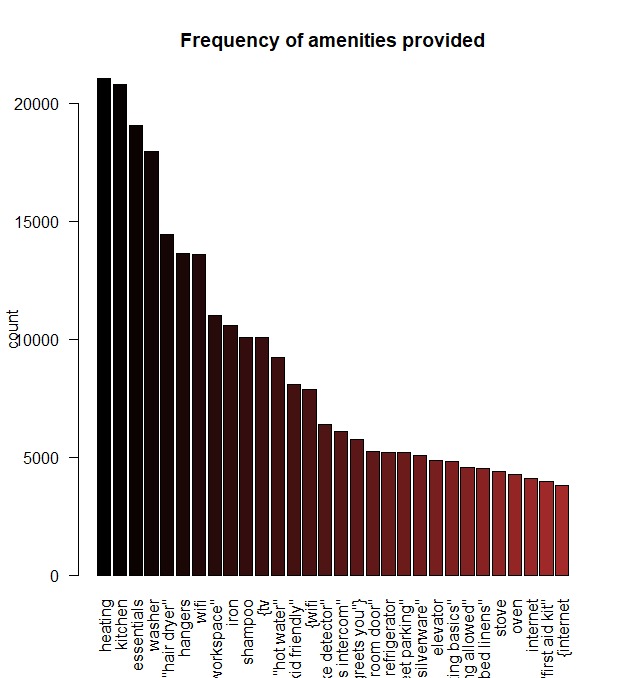
1. While preprocessing the data we saw that the columns of minimum\_nights and maximum\_nights has values such as 5000 and 999999999 which were clearly outliers:

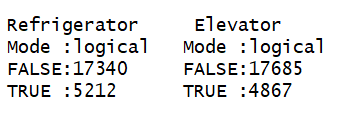
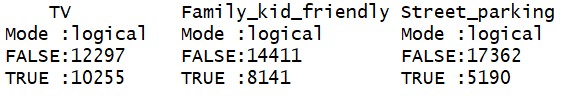
** **

We wanted to address this issue without blindly deleting the much higher values, hence we bucketed the data and created two columns named “less than 6 minimum nights” and “less than 2000 maximum nights”.

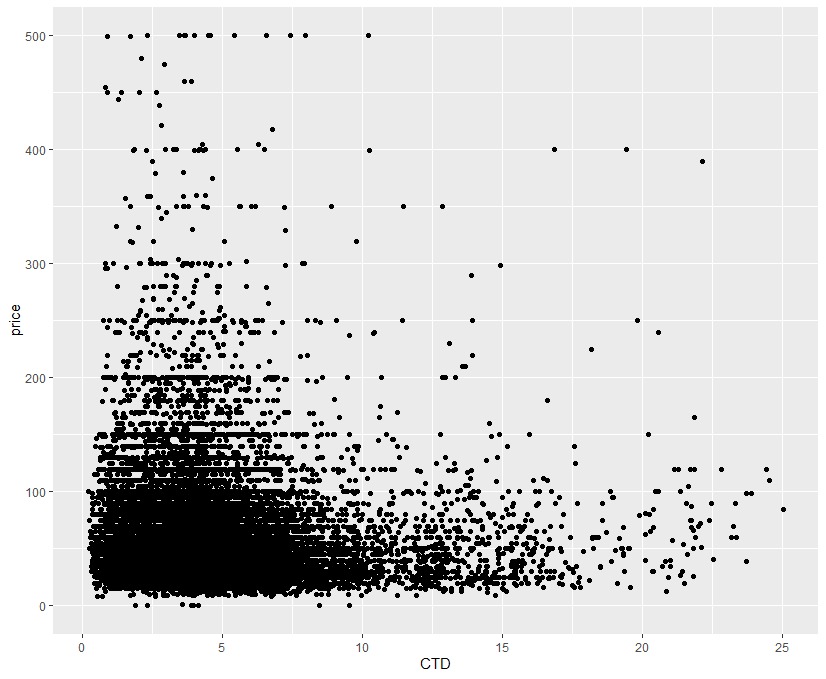


**2)**We further worked on the amenities column, so that we could know providing which amenities effects the price of the listing. We calculated the frequency of words in the amenities column , and then we created 5 new columns of the amenities which were significant but quiet unique and not provided by every host.





**3)**Location is an important factor when deciding the price of a listing. The listings in some neighborhood might be pricey just by the virtue of their location. We thought of calculating the distance of the listing location ,using the latitude and longitude of the listing from the centroid of Berlin, and then we plotted the price vs listing graph to know the variation in price due to it’s distance from the centroid of Berlin.

****

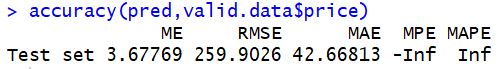
Here we can see that the prices of the listings are higher when they are closer to the center of Berlin. Hence, distance from center(CTD) is an important factor when deciding the price of a listing.

**Data Modelling:**

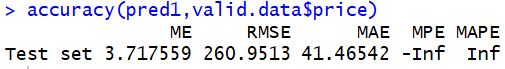
We partitioned our data into three sets namely Training data(60%), validation data(20%) and Test data(20%).

**Simple Linear regression**

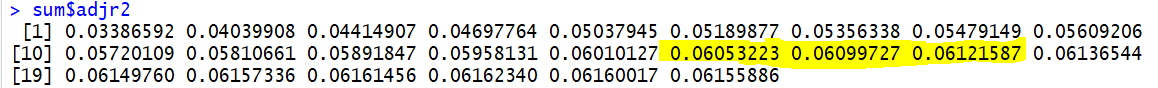
We then ran a simple linear regression model using all the selected predictors and got the following output:



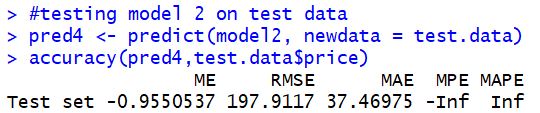
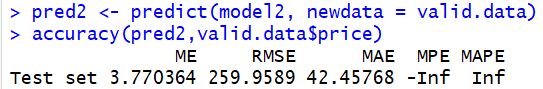
Next we selected the significant predictors using the p values we got from the previous model:



Both the models perform similar in terms of accuracy, hence we thought of running an exhaustive search to build a better model.

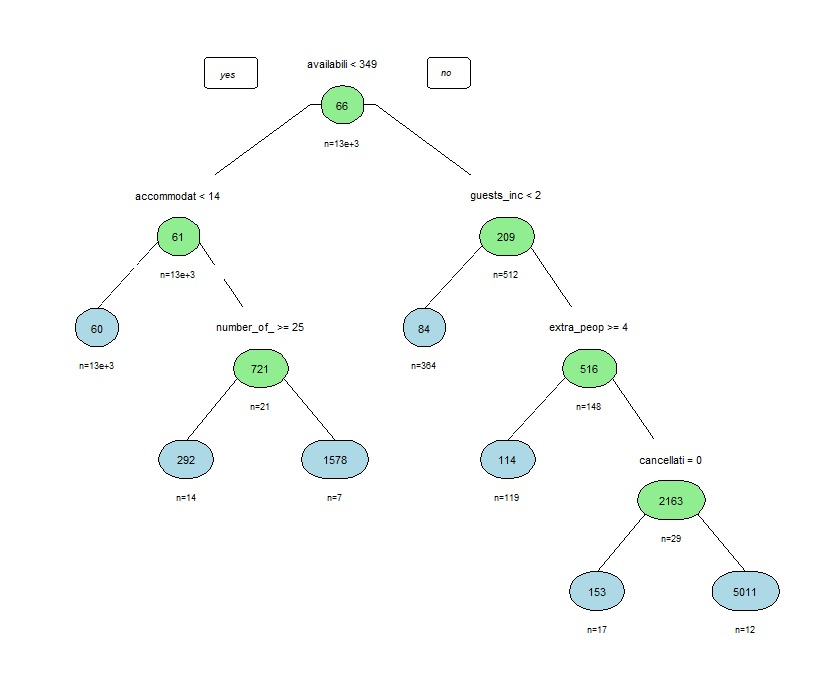


Based on this output we created three models using 15,16 and 17 variables and model 2 with 16 variables performed best among the three models with an RMSE of 259.9589 on valid data and 197.912 on test data

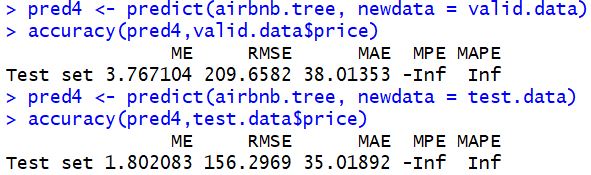


**Regression Tree:**

We ran a regression tree using all the predictors against price for the training data:



After running the tree on valid and test data we got the following outputs:



The regression tree gives us a better accuracy for predicting price on test data. Hence we would suggest regression tree as the best model that can be use to predict the price of a new listing using the predictors used in the regression tree.

**Business implication and risk mitigation:**

Using the Airbnb Berlin data about the Airbnb listings we have looked at approaches which can be used for predicting the price for a new listing. Hosts can make use of historical data to get an idea about how they can determine the price of their listing by using the data of how much other hosts are charging for similar listings. The model we developed has practical value to the client as they can not only get an idea of the prices, they should charge but also gives a chance to mark up/down the prices.

Airbnb might provide price tips based on multiple sources of information and multiple factors. Although, the hosts do not know if they bias these prices based on personal interests. Airbnb might want their listing prices varied so more customers can convert. So, the additional value this tool adds is an unbiased estimator.

**Conclusion:** In the framework of this paper, we analyzed Airbnb listings from Berlin. Our objective was to implement Linear models, to approximate prices of listings conditional to explanatory variables which would

provide the optimal equilibrium between performance and parsimony. We cleaned and formulated the data so they would be apt for data analysis. Next step was to plot the data; plotting enables us to distinguish relationships between response and predictors and obtain a general idea of the individual variables’ distributions and also find any multicollinearity between the variables. Subsequently, we used linear regression methods to generate models to predict price against various variables and compared their performance using error metrics such as RMSE.

**Future scope :**

We have features that were not used in this exercise because of time limitations. These are majorly text based features and would require natural language processing to convert into features. In continuation of the work, these features can be used:

1. Description (Text)

2. Reviews (Text)