**TEAM PROJECT FINAL REPORT**

*Section :* ***B*** *| Team :* ***15*** *| Members: Xue Chen -* ***xc183****; Shikha Dhurka -* ***sd430****; Xin Dong -* ***xd63****; Layla Lin -* ***yl810****; Akshat Samir Patel -* ***ap573****; Yuhan Zhu -* ***yz713***

**Business Understanding**

Airbnb, Inc is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, CA, the platform is accessible via a website and mobile app. Similar to Google, Facebook, and Uber, it works on a model called - Platform business model. The two facets of the Airbnb model are guests (customers) and hosts.

According to WSJ, Airbnb was able to solve the major problem that it faced during the COVID-19 outbreak, i.e., the lack of guests. As the travel restrictions are eased across the world, it saw a 50% growth in gross bookings value in the first quarter of 2021 as compared to the previous year, indicating that customers are willing to travel again. But, Airbnb’s relationship with the hosts became tenuous after it had to cancel the bookings because of the travel restrictions. Even though it offered millions of dollars to hosts in relief, it only covered a fraction of their lost earnings. In fact, Airbnb’s active listings grew less than 10% over the last 12 months. **Thus, Airbnb needs to provide lucrative incentives to the hosts to add more listings so that it can satisfy the likely increase in demand that is to occur. Along with this, Airbnb needs to take advantage of the anticipated travel boom by attracting millions of guests to book from them.**

Now that we have identified the problem, let us go through some causes that could **prevent** Airbnb from taking the advantage of the travel boom:

* **Causes for guests**: Price, Service Quality, Availability, Brand image/ brand disloyalty, Attractive pricing by competitors, New competitors entering the market.
* **Causes for hosts**: Fewer monetary incentives in running the listings, Refund for cancellation from Airbnb, Lack of incoming guests.

From the above causes, we would like to focus on the **“Pricing”** of the listings as solving it could help in tackling both the facets of the Airbnb business model.

1. Demand Side - It can help attract new customers, help retain existing customers via offering attractive prices as compared to competitors.
2. Supply Side - It can help in increasing the number of listings (by adding new hosts or retaining existing hosts) via offering lucrative incentives.

Secondly, we know that Airbnb is a customer-obsessed company, and **so we would like to understand how they feel about the brand, how they feel after staying at one of their listings**.

**Goal:** To take the advantage of the upcoming travel boom by attracting customers and making sure there is enough supply.

**Ideas to process:** Offer attractive prices; Understand the brand’s image in customers’ minds to serve them better.

For Pricing, we have used predictive algorithms like multiple linear regression and random forest to predict the prices of different listings in Boston. We have considered multiple factors such as the number of **bathrooms, bedrooms, beds,**  **reviews,** whether the host is a **superhost,** how many **guests** are **included,** the **availability in the next 30 days,** the **review score rating.** We also created several dummy variables for the categorical variables in our dataset. The categorical variables include **neighborhood, amenities, property type, room type, bed type, cancellation policy, and host response time.** To understand the brand’s image, we have performed **text mining** and **sentiment analysis** on 65k+ reviews data for listings in Boston city.

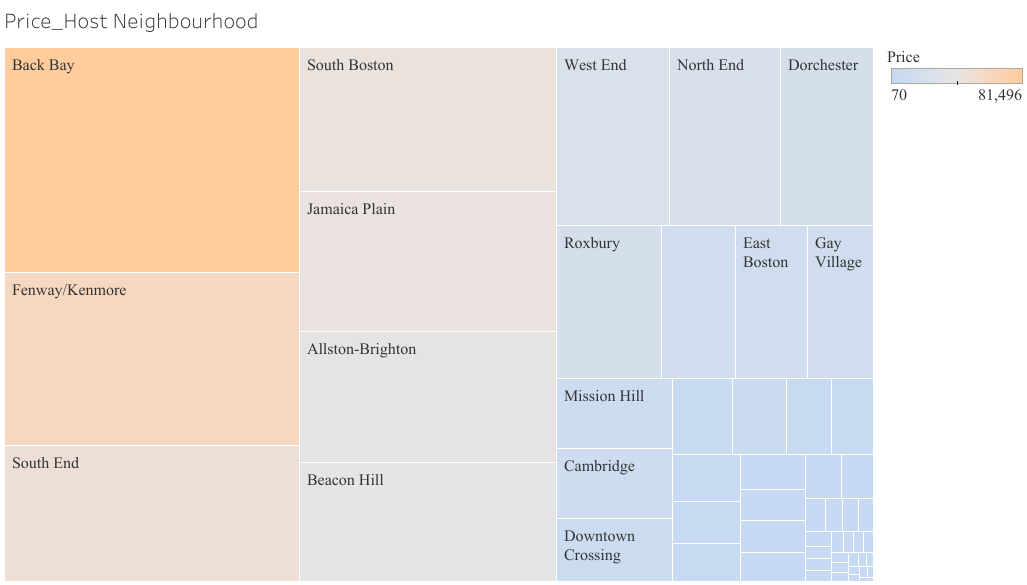
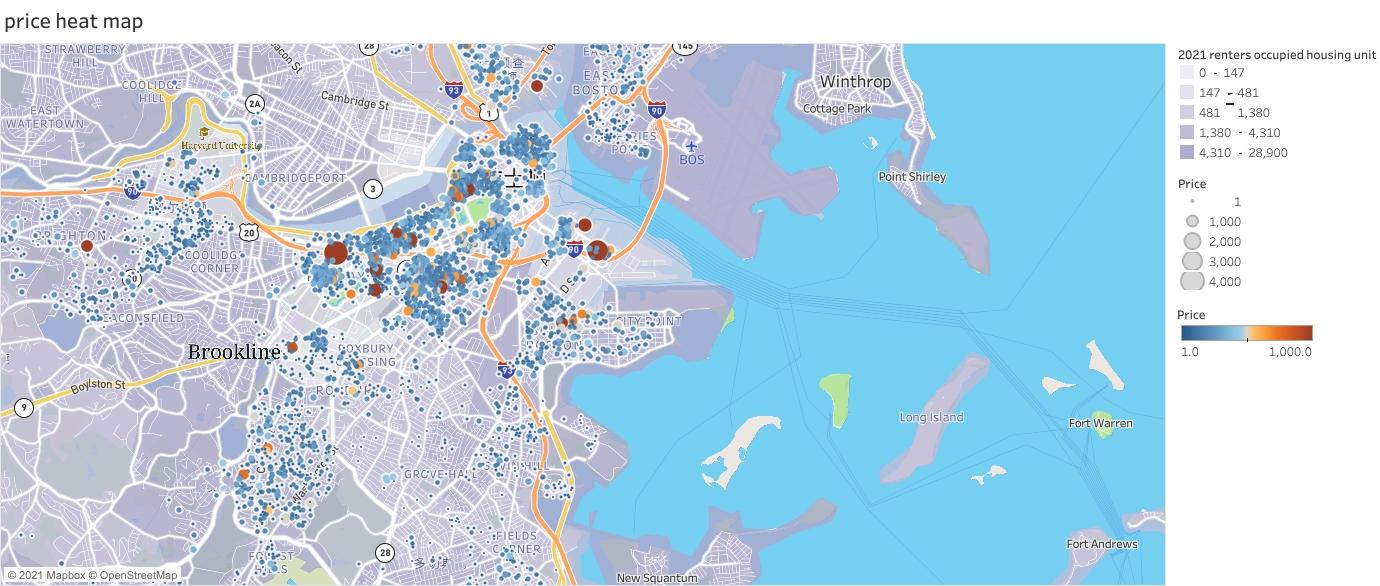
**Data Understanding**

We chose the Boston Airbnb open dataset from Kaggle. It is the official disclosed data from Airbnb, which can be found at <http://insideairbnb.com/get-the-data.html>. The dataset has 3 files out of which we have used 2 files - “listings.csv” and “reviews.csv”. The listings dataset consists of 94 columns/features and of 3585 observations. As our goal is to help the company make better pricing strategies for the listing housing, our target variable chosen is **price**. In general, among the 94 features, we divided them into demographic attributes(id, name), geographic attributes( market, latitude, longitude, zip code, street, neighborhood), and other attributes. We have also used another file “reviews.csv” which has 65k+ comments to analyze the sentiments.

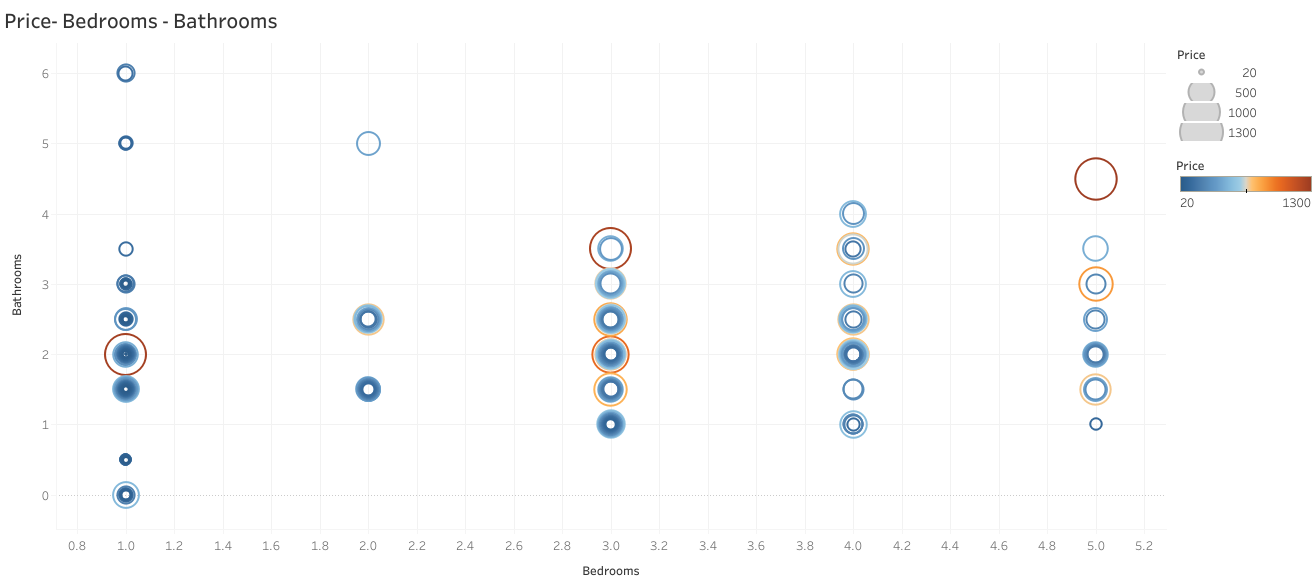
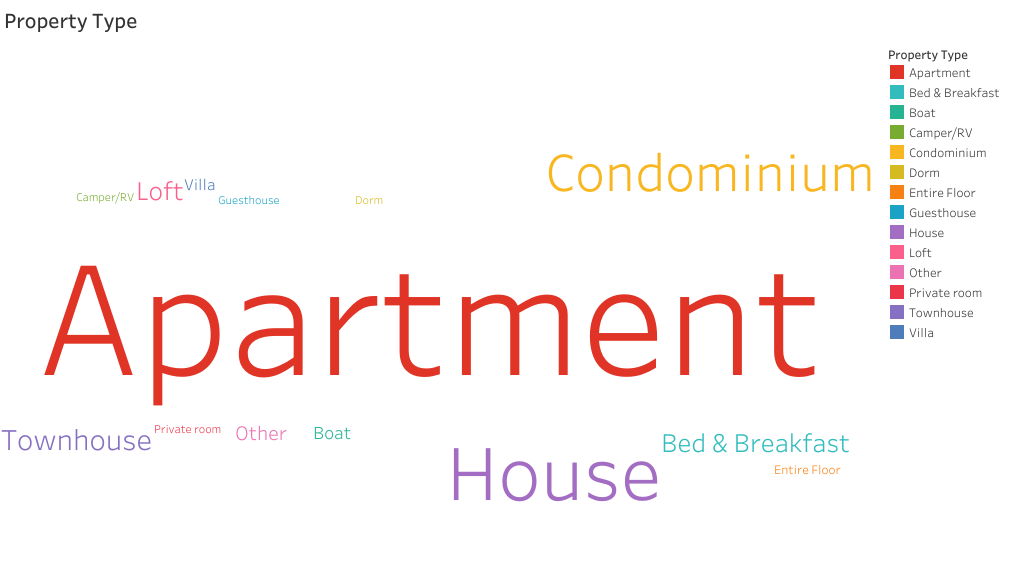
**Data visualization**:

In order to have a better understanding of the relationship between our target variable price and other features from the beginning, we did the following exploration series utilizing Tableau.

1. **Pricing Heatmap:** First, we did the price heat map with the dimension of latitude and longitude to explore the street or area that generates the highest revenue.
2. **Price and Host Neighborhood:** Secondly, we dived deeper into the neighborhood to see if there’s a correlation between listing price and neighborhood. The result was aligned with what we found out from visualization 1, the price heat map. Certain neighborhoods are way more vibrant than others and these areas are in accordance with the heating streets in visualization.



1. **Price and Property Type word cloud:** Then we focus on a unit scale, creating a word cloud to see which property type is more profitable in Boston. Apartment, house, and condominium turned out to be the top 3 profitable types of housing.
2. **Price and Bedrooms and Bathrooms:** On the unit scale, we then tried to find the relationship between price and the number of bedrooms and bathrooms. As can be seen from visualization4, the highest pricing listing tends to be close to the line where bedroom: bathroom approaches 1:1.



**Potential Bias**: Though our dataset is official and concise, there may exist some inevitable potential bias: (1) Our dataset only consists of information of one given year without a long time tracking, so there may exist some selection bias on the data. (2) Our dataset focuses more on the host level without a fair balance on customer demographics and so on. (3) The review data may contain positive reviews and very few negative reviews, thus making it difficult to provide a clear analysis.

**Data Preparation:**

The Airbnb Boston data consists of **3585 observations and 94 variables**. Our logic to remove irrelevant variables consists of two ways: 1. if the data has too many null values 2. if the columns are not relevant to the business issue. To begin with, as the number of unique listing ids equaled the number of observations, we could conclude that each row was a unique listing. We then **converted strings to numeric values** by removing the dollar sign in front of “price”, “security\_deposit”, “cleaning\_fee”, “extra\_people”, the percentage sign in front of “host\_response\_rate”, and the percentage sign in front of the “host\_acceptance\_rate”. We identified the **redundant columns** such as ‘listing\_url’, ‘scrap\_id’, ‘last\_scraped’, ‘host\_thumbnail\_url’ and removed them or columns with **only NA values** such as the ‘experiences\_offered’ column. We also changed certain **categorical variables to dummy variables**, including “amenities”, we **identified certain important amenities** through online research to understand which amenities guests look for. The top 9 that we selected were “wireless internet”, “pets”, “free parking”, “AC”, “heating”, “washer”, “pool”, “kitchen”, and “TV” and created dummy columns for the same. We also created **dummy variables** for 'neighbourhood\_cleansed', 'room\_type', 'bed\_type', 'host\_response\_time', 'property\_type', 'cancellation\_policy'. For the amenities column since each row had multiple amenity names in a row, we used the ‘**stringr**’ package to detect a particular string in the amenity column and then created dummy columns for them. We decided to **remove the various character columns** such as ‘name’, ‘description’, ‘summary’, ‘space’, ‘notes’, ‘interaction’, and ‘transit’ instead of creating dummy variables for keywords used in these columns. We did so because we wanted to **avoid overfitting** data by using too many variables in our model. We already created several dummy variables for more important factors (such as property type, room type, neighborhood, etc) that would contribute to the price of the house. We converted columns “host\_is\_superhost”, “host\_identity\_verified”, and “instant\_bookable” to **‘1’ for ‘t’ and ‘0’ for ‘f’** for better understanding and **factored those variables**. We checked for the correlation between each of the five availability columns and price to decide which column to select for our regression model. Since ‘availability\_30’ had the highest correlation we decided to include that in our model.

For the sentiment analysis, we created a corpus of the comments column. We did the cleaning of the corpus by converting the corpus **to lower case, removing punctuation, numbers, unhelpful words, and stemming**. Lastly, we created a **term-document matrix** to analyze the comments.

**Modeling**

**For sentiment analysis:**

We have used the get\_sentiment function to determine the sentiment score of the comments column. Further, we compared the results of three methods - syuzhet(), bing() and afinn() to ensure the consistency of sentiment scores. The sentiment scores can help us understand what the customers think about the brand. A positive sentiment score ensures that the customers are happy with the services and a negative sentiment score means that we need to understand the pain points of the customers to ensure they are happy with the provided services. Alternatively, topic modeling can be used to cluster the words based on the topics they fall into. But, because of the high number of rows (65k+) in the “comments” column, it becomes a performance-intensive task to achieve.

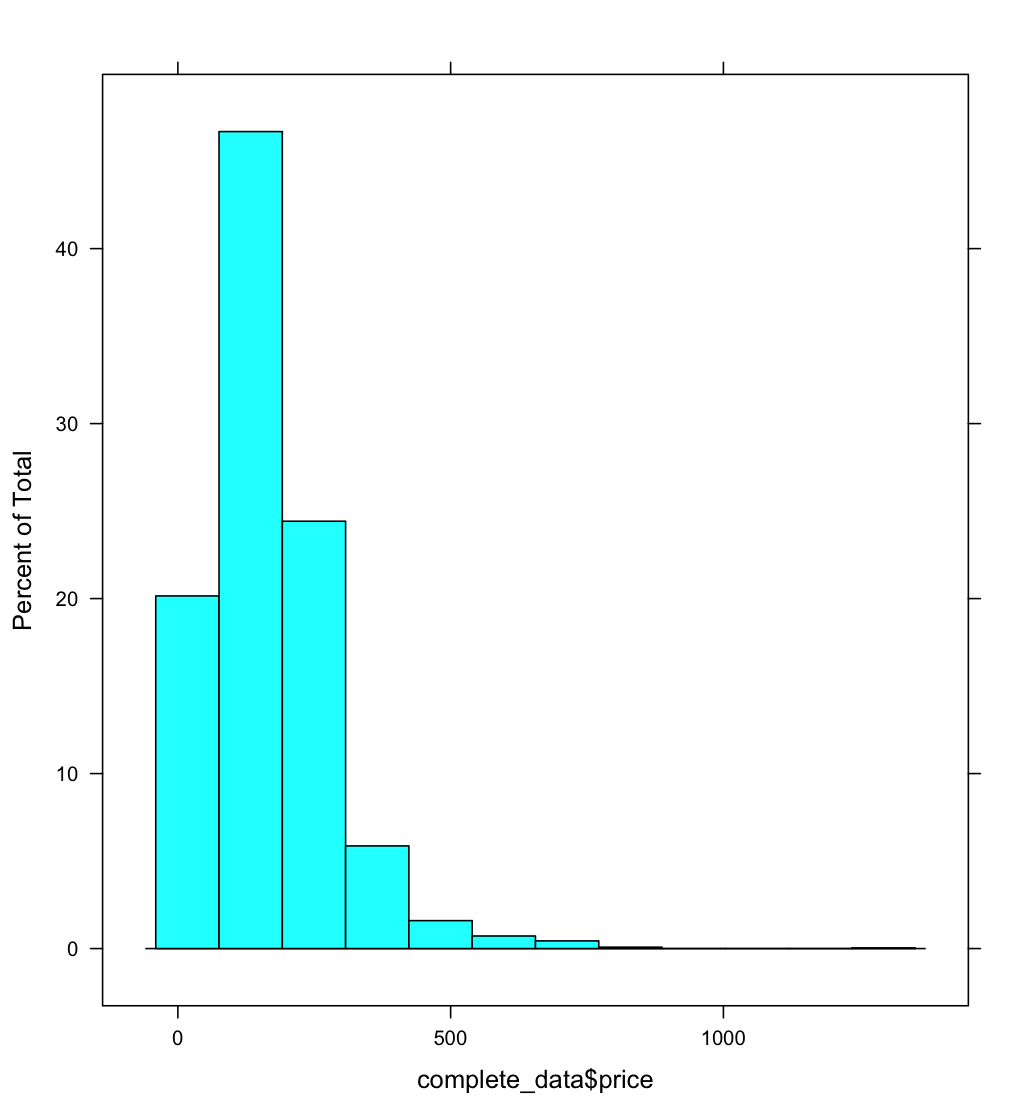
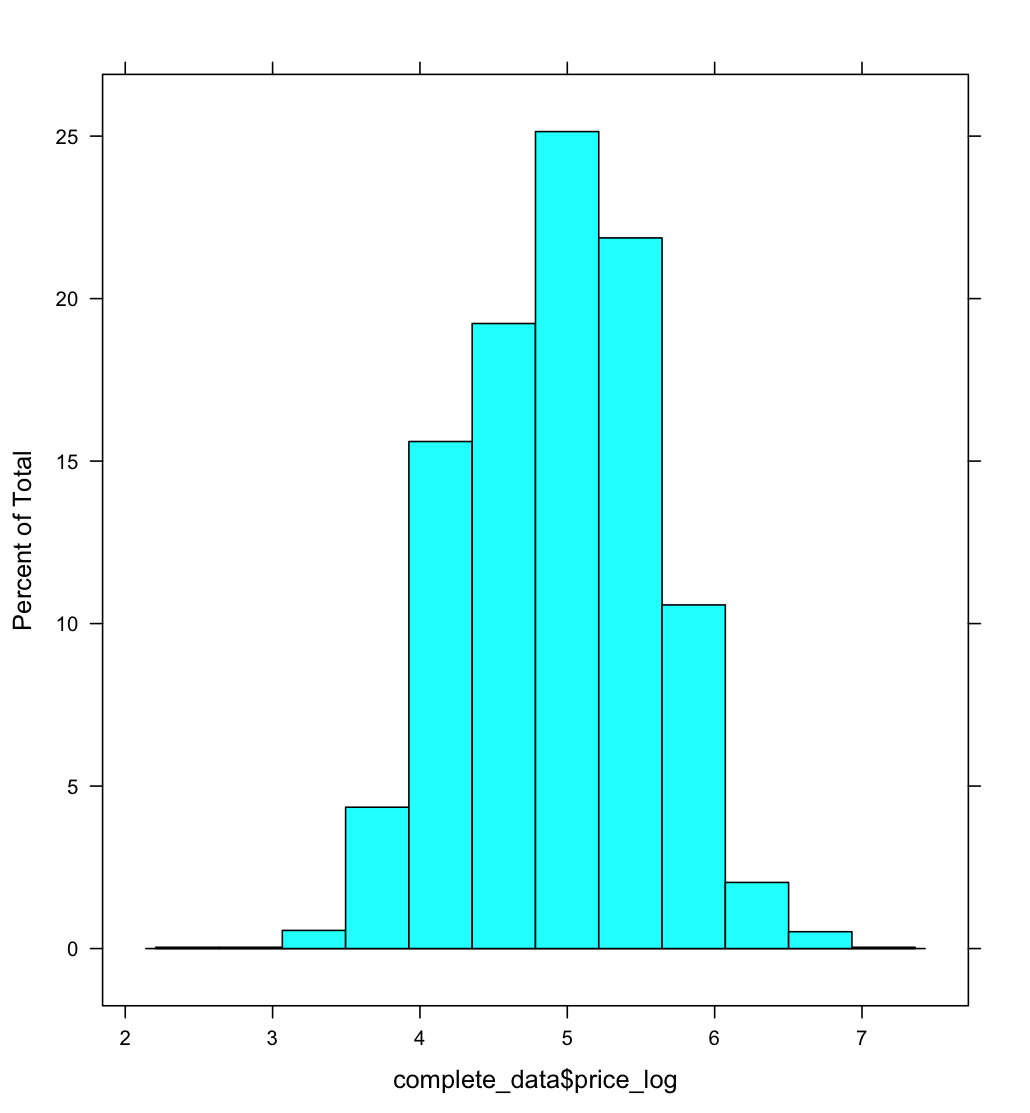
**Price Prediction:**

To predict the price of Airbnb listing we decided to use two predictive modelling techniques:

1. Multiple Linear Regression, supervised learning technique used to predict the value of one dependent variable based on the values of two or more independent variables.
2. Random Forest, unsupervised learning technique that combines many classifiers to provide solutions to complex problems. It consists of many decision trees.

We first partitioned the data into Training data and Testing data using the package ’caret’ and code ‘createDataPartition’. Since the dataset for pricing has a small number of rows (3585 observations) we decided to split the data into 80% training data and 20% testing data. Since ‘price’ is not normally distributed and skewed to the left side of the histogram, we decided to use the **logarithm of price** as the dependent variable instead of ‘price’ to transform the distribution of price into a more normally shaped bell curve as seen below.

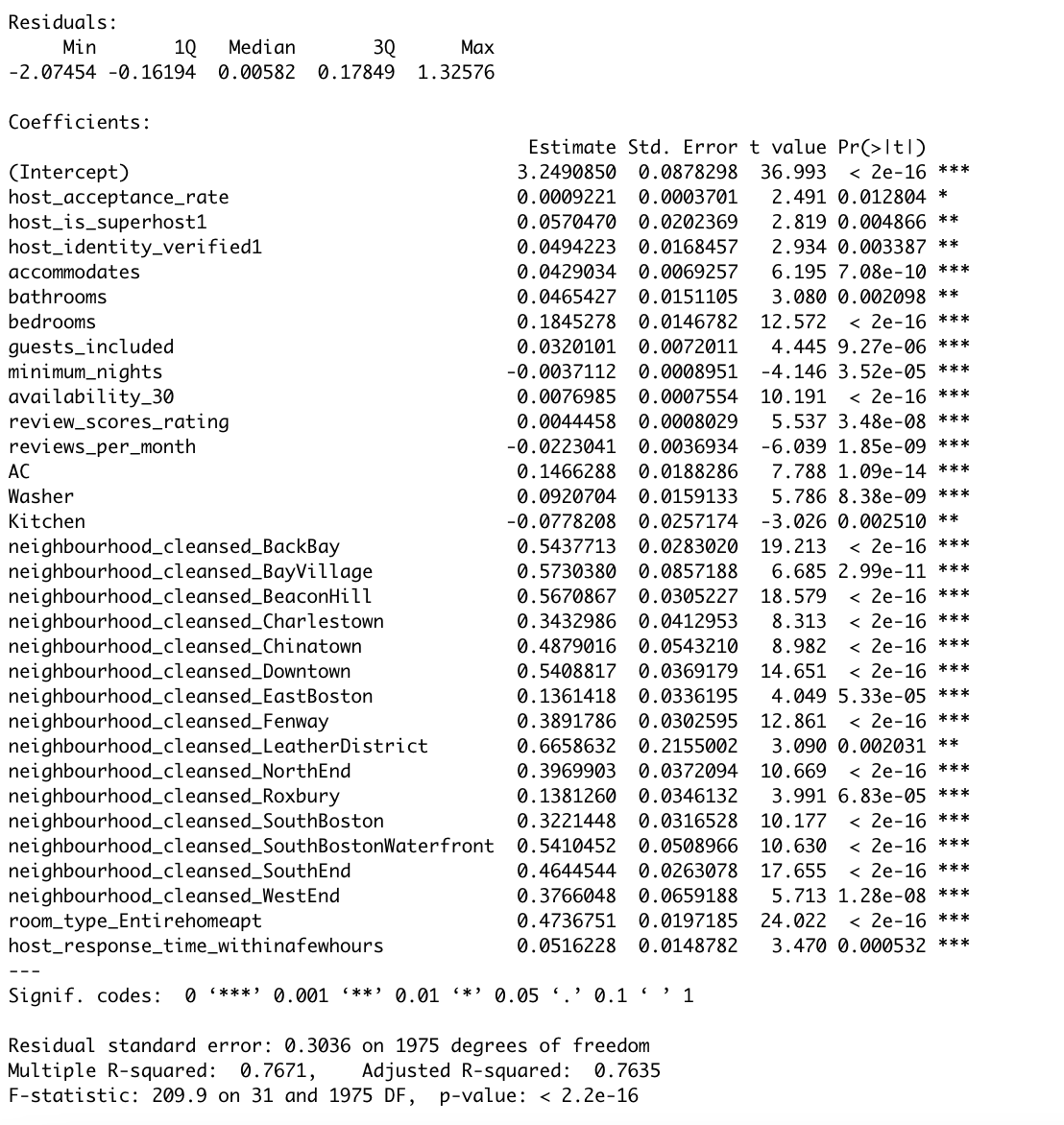
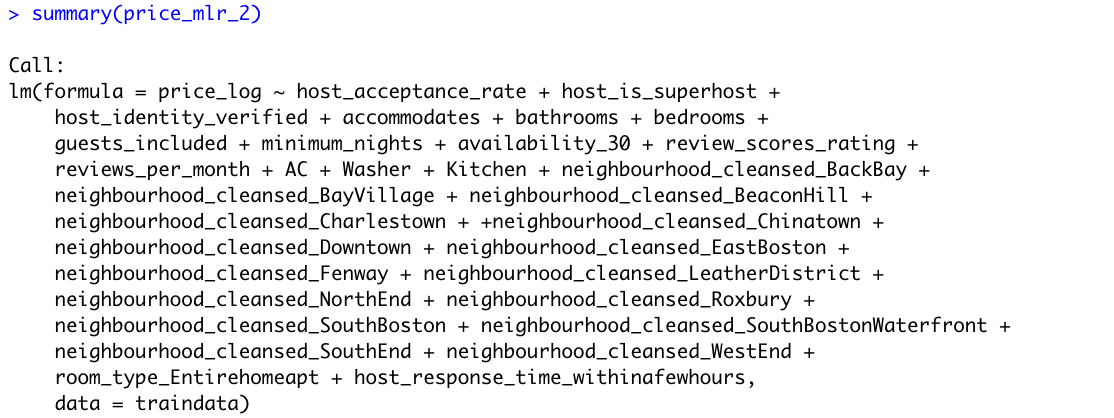
**Distribution of Price Distribution of log of Price**



Using ‘price\_log’ as the dependent variables and all the other selected variables as the independent variables, we ran a Multiple Linear Regression model. Based on the results of the model, we removed the insignificant variables with p-values greater than the 0.05 threshold. Our first model had an adjusted R square of 0.7821. We ran the linear regression model again with only the significant variables from model 1 and we achieved a satisfactory model with an adjusted r square of 0.7773. We then used the ‘car’ package to test for multicollinearity in the model. Using the function ‘vif’ we saw the ‘room\_type\_Entirehomeapt’ and ‘room\_type\_Privateroom’ had a value greater than 5 which suggests that those two variables have a high correlation. Therefore we removed ‘room\_type\_Privateroom’ from our second model since it had a p-value greater than ‘room\_type\_Entirehomeapt’. We also removed some neighborhood dummy columns with higher p-values to reduce the RMSE of our testing data by avoiding overfitting.



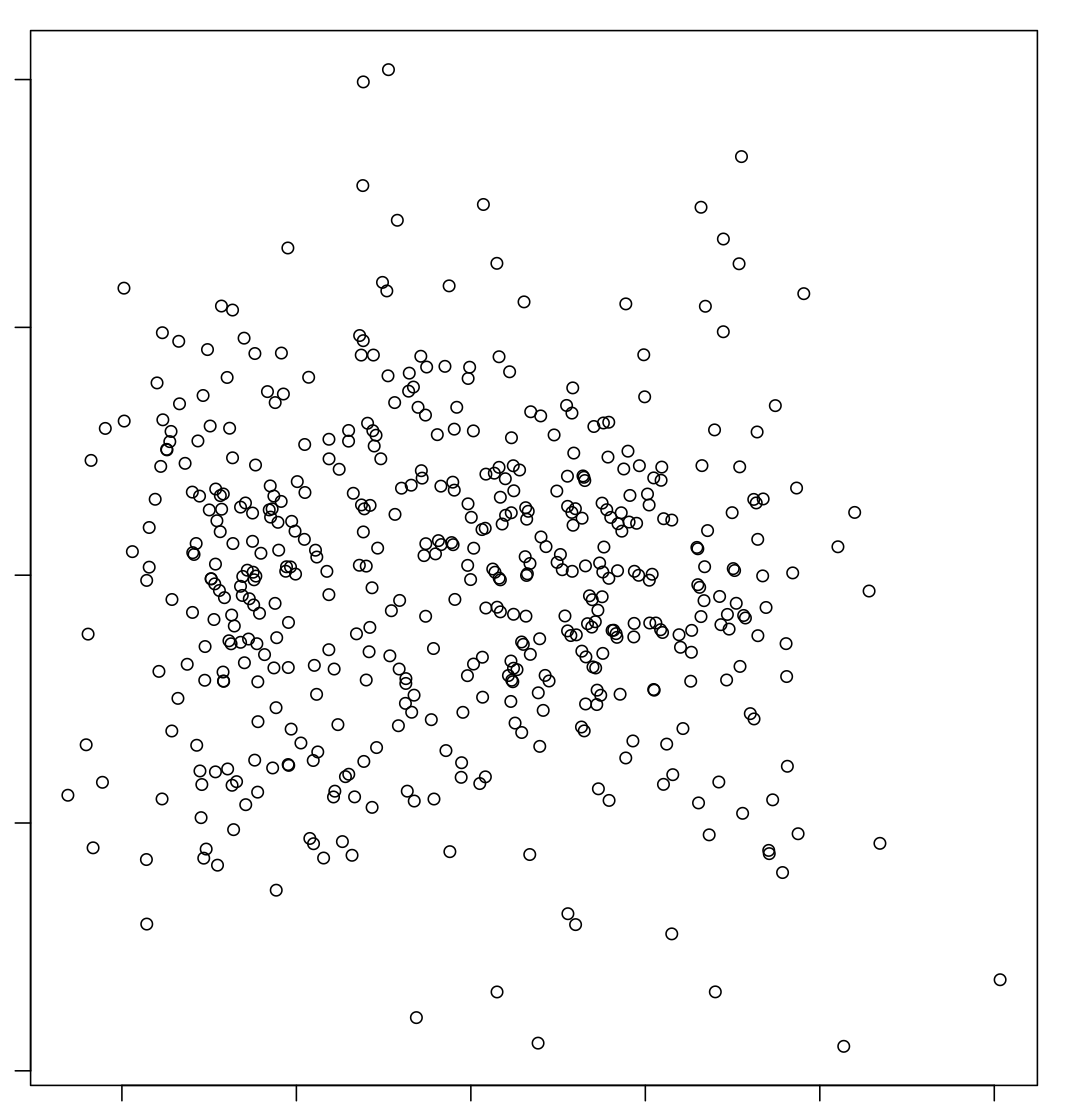
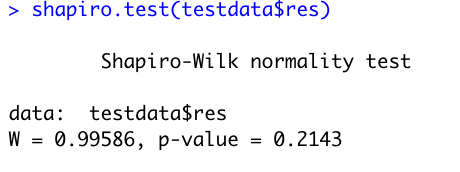
Our final model as seen below has an adjusted R square of 0.7635 which is relatively good considering the cleaned dataset consisted of only 2506 observations. All the variables have p-values less than 0.05.



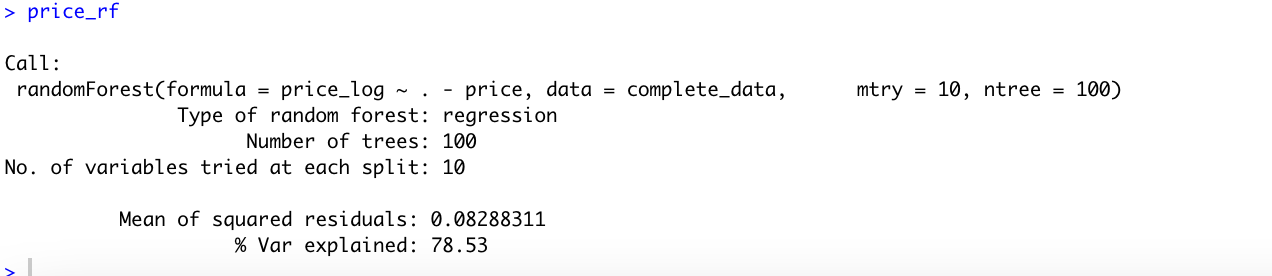
The **RMSE of our training data is 0.3011555** and of the **testing, data is 0.3092533**. Even though the RMSE of testing data is increasing it is only a slight increase which could be due to the large number of factors included in the model.

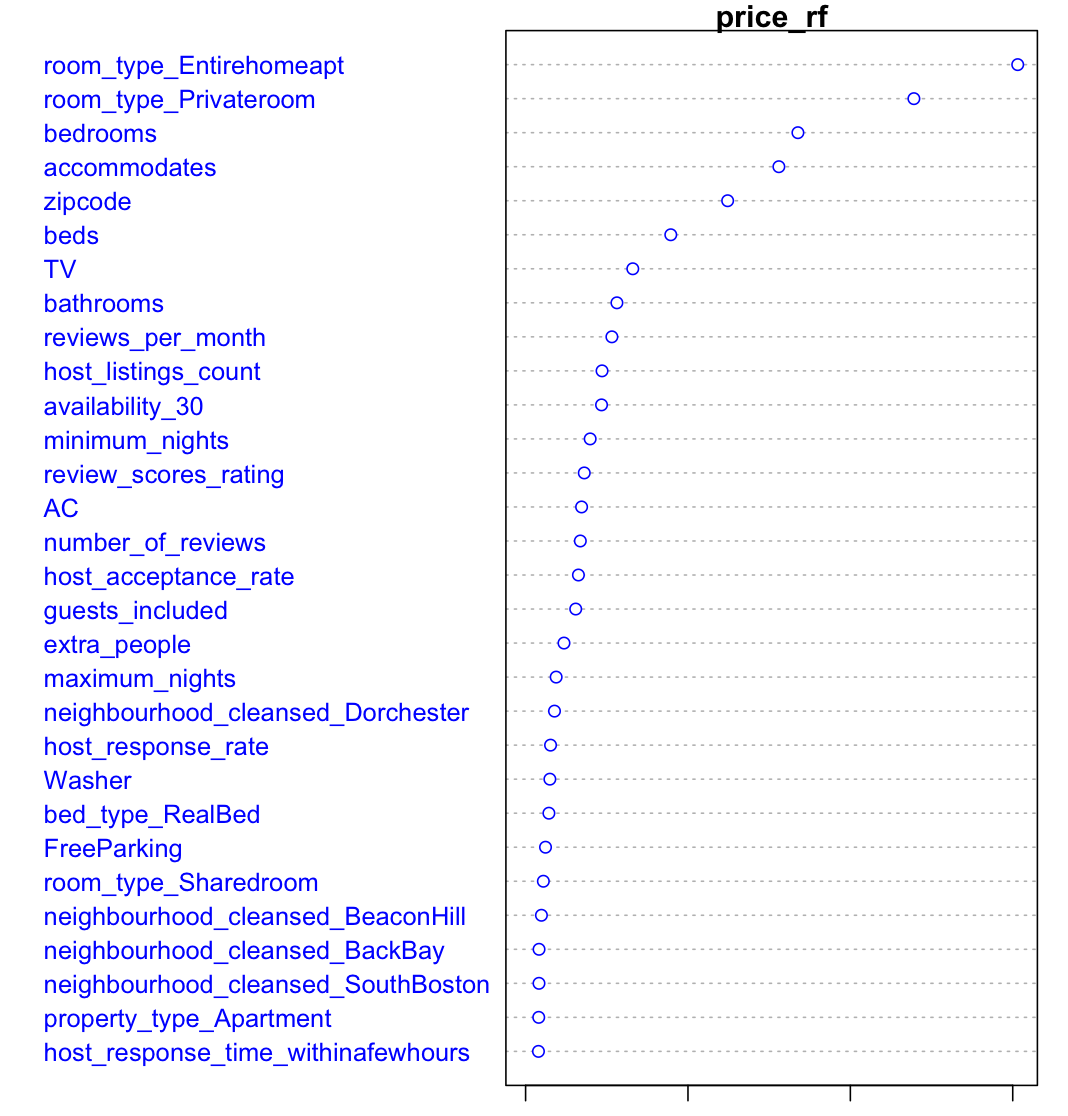
The random pattern of the plot between residual predicted values suggests that the model provides a good fit.

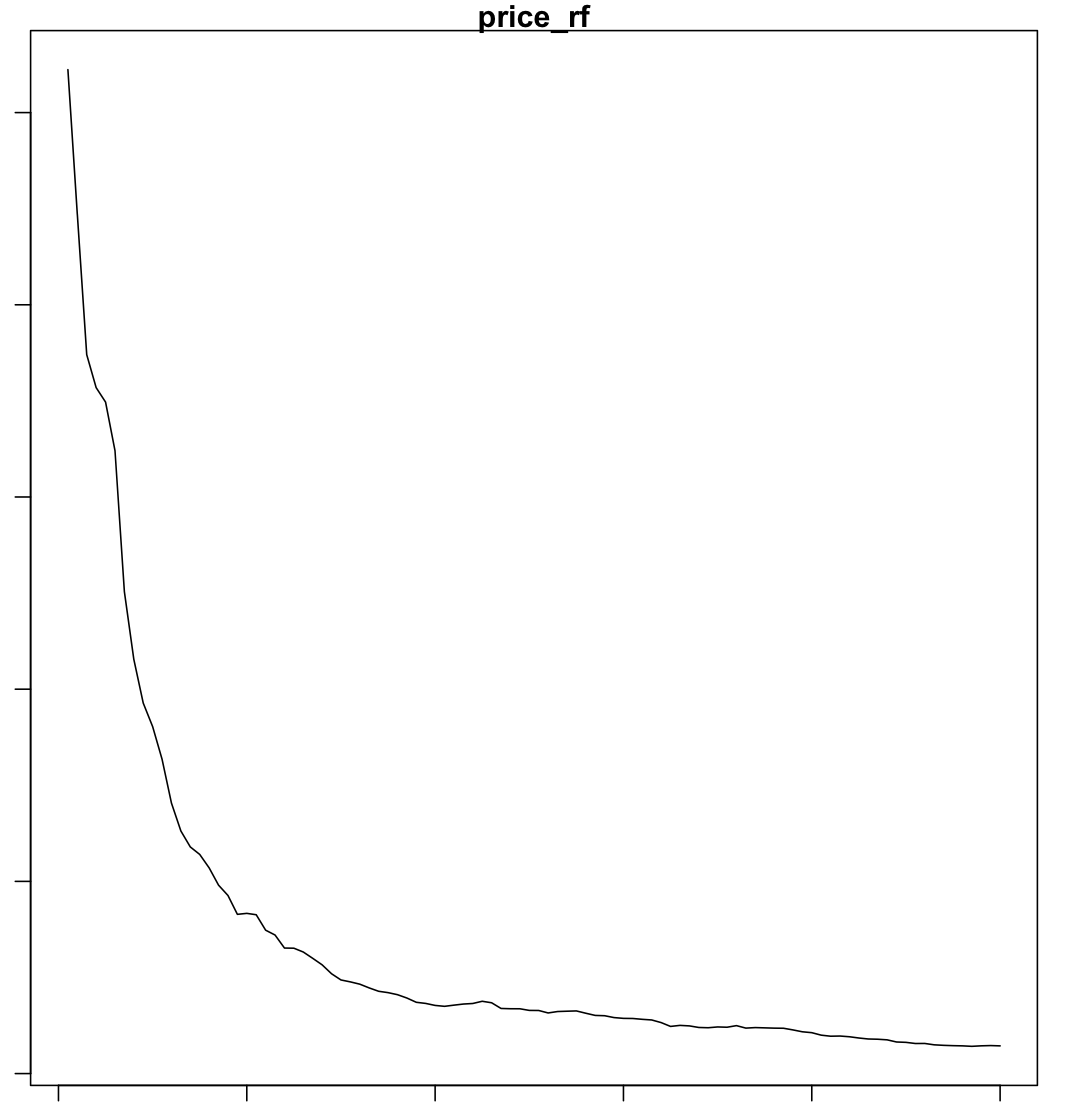
We performed the **Shapiro Wilk test** to check for the distribution of the residuals. Since the p-value of the Shapiro test is 0.2143 and is greater than 0.05, we can assume that errors follow a normal distribution.



**Random Forest:** We used Random Forest as an unsupervised predictive modeling technique. We chose the number of trees as 100 and the number of splits at each node as 10 and got a **mean of squared residuals of 0.08288** and the **percentage of variance explained by the model is 78.53**. We recommend using the random forest model as it has a lower RMSE compared to the Multiple Linear regression model and explains the variance better to predict the price of listings.



We then plotted the random forest model to see the error line. As seen from the plot below, the error reduces as the number of trees increases. We then checked to see the **important variables** in the model using the formula ‘price\_rf$importance’ and plotted them as seen below using the formula ‘varImpPlot’ :







As per the random forest model, the room\_type dummy variables, bedrooms and accommodates are the most significant variables. Followed by the number of guests the house accommodates, zip code, and the number of beds the house has.

**Evaluation**

**For Price Prediction:**

Evaluation metrics for regression models are quite different than the metrics used for classification

models because we are now predicting in a continuous range. We can first see the Root Mean

Squared Error (RMSE). For the lineage regression model 4, the RMSE is 0.3011555 using the

train data and 0.3092533 using the test data. There is a slight increase in the RMSE which could mean there is overfitting of data due to the number of variables. But these values alone are not intuitive. Thus, we take the R-Squared metrics, seeing how good our regression model is compared to a very simple model that just predicts the mean value of target from the train set as prediction, into consideration. And our regression shows 0.7635 in R-Squared, which is quite good. For the random forest, we have a mean of squared residual 0.082, and the percentage of variance explained is 78.53.

Airbnb can try the model, but pricing is a difficult thing to predict. They definitely have other variables

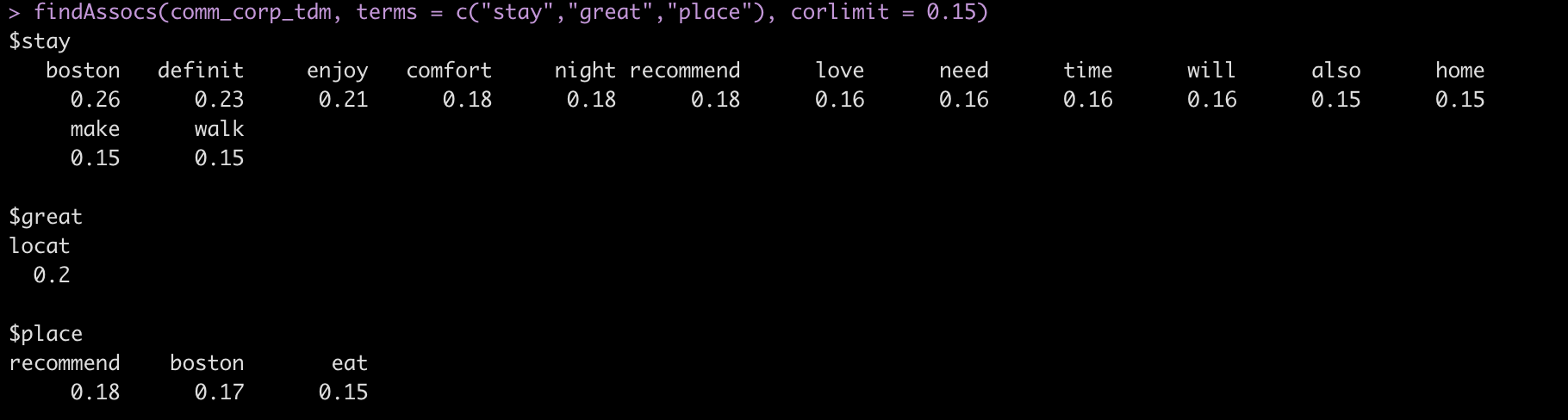
list in the model. We know that the pricing market is driven by supply and demand. However seemingly simple it may sound, there are literally an infinite amount of factors that come into play. We can technically create a model that factors in all the known factors that have moved the markets before. But the vast majority of market factors are literally impossible to know beforehand.

**For Sentiment Analysis:**

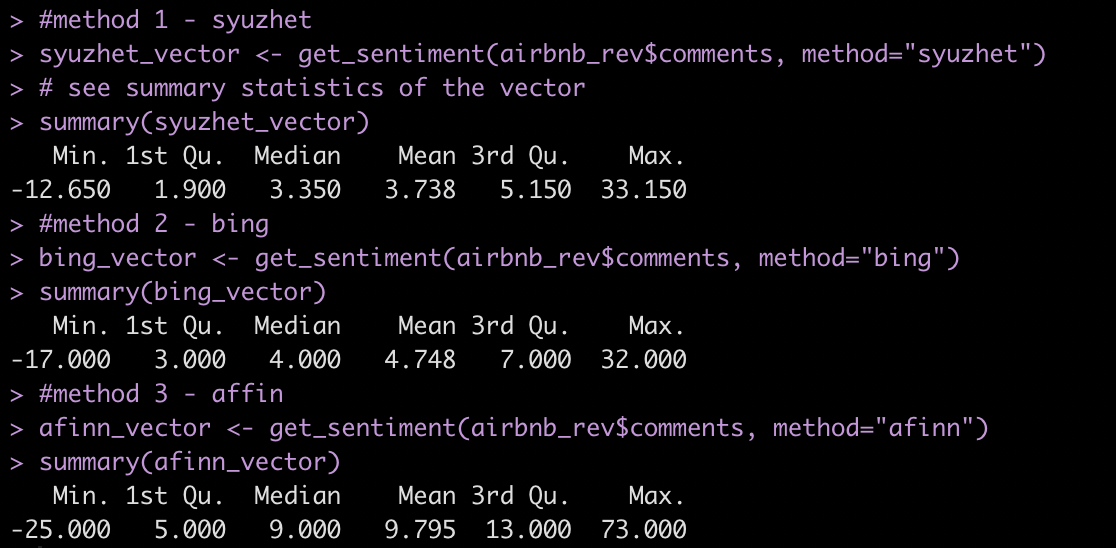
1. We calculated the 50 most frequently used words in the customer reviews to understand what the customers think about the brand and generated a word cloud for the same. All the top 50 words were either positive or neutral, with the top three being - ‘stay, great, and place’.



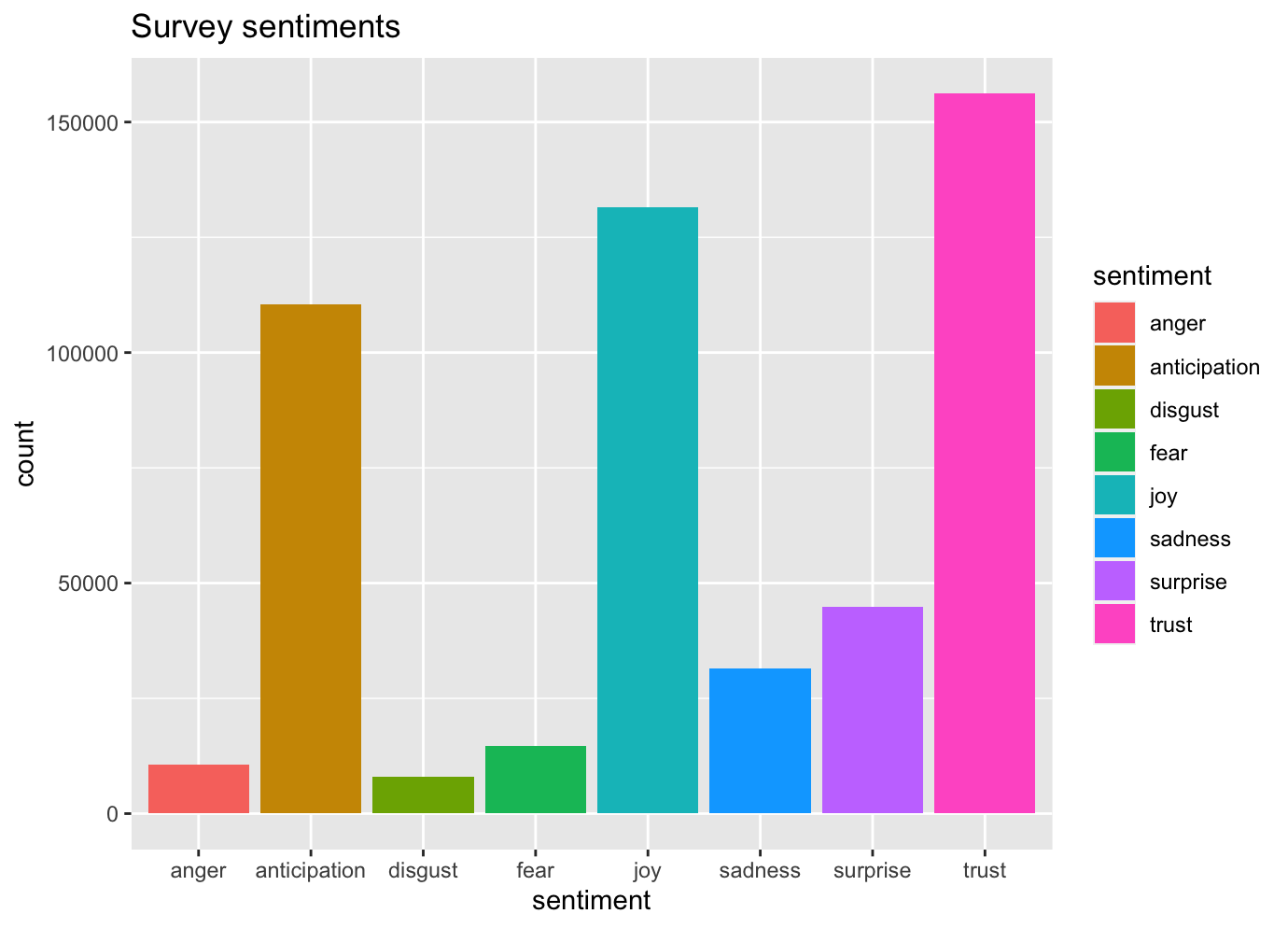
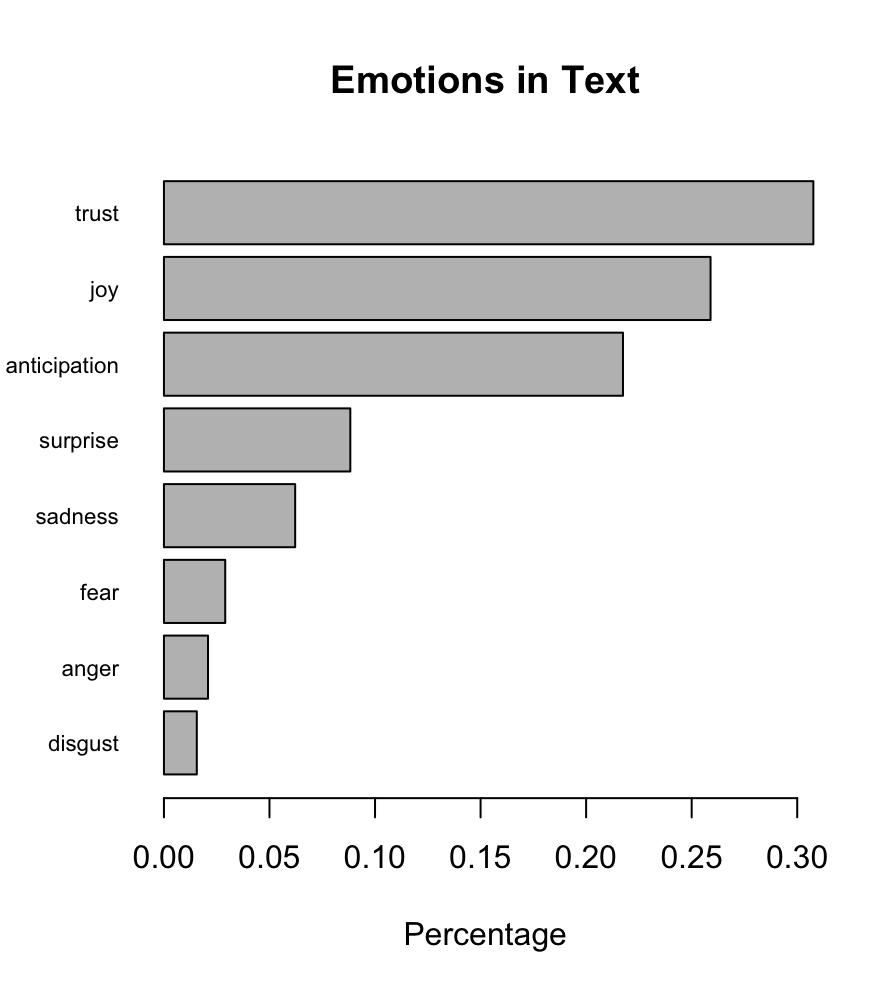
2. Next we calculated the associations between the top 3 words and also the words with a frequency greater than 15000 and a correlation of 0.15. We observed similar results where the top words and their associated words were either positive or neutral which is good for the brand.



3. Lastly, we calculated the sentiment scores using three methods and found similar results. From the 1st Quantile onwards all the words had positive sentiment scores. All three methods showed consistent results.



4. We analyzed the count of words associated with 8 different sentiments -



The results show that customers feel positive about the brand with words such as “trust, joy, anticipation, surprise” coming out as the top words.

**Deployment**

We assume that the price of a house in Airbnb is decided by the market supply, market demand and house providers are not precisely aware of the market. By applying this model, Airbnb can predict the estimated price for each listing using the data and reviews they collected. After prediction, the estimated price can act as a reference price for house providers, and it helps providers design an appropriate price, which leads the platform and house owners to earn more money. Also, from the result of the data mining, Airbnb can determine the key contributors to the price and use data to find out the customers’ preferences. By collecting the preference of customers, Airbnb can collect more house resources that have good performance and recommend the house with similar reviews.

Before deployment action, Airbnb should consider a few questions about the data mining model. The first issue is feasibility - the firm should carefully check each feature that is used in the model can be collected before doing prediction. If the feature can only be known when the purchase is completed, there is no need for Airbnb to do the prediction since the prediction price is out of date. Secondly, the firm should take into consideration the difficulty of data collection and the computational cost of model deployment. Since the review data changes every day and new customers will leave a new comment for listing, it is expensive for firms to run an online dynamic model and re-train the model using all data. Airbnb should fully consider the investment/cost and return of the deployment. Also, Airbnb should establish a data center to save all data for disaster recovery.

There are some ethical problems that need to be considered. For Airbnb, it owns the data from house providers and customers, which include demographic data, behavior data, and other data like order id and session id. Airbnb should keep private information secure and fully design the information access permission mechanism before collecting data. Moreover, extra data except review data should not be collected unless with customers’ permission. Also, since the comments facilitate price prediction, Airbnb has the obligation to check the authenticity of the comments and evaluate the predicting price to ensure it is not overestimated. Otherwise, it may ruin the platform’s credit and lose the trust of customers.

Finally, we should consider the risk behind the proposal. The first risk is that the data do not provide enough information for the business problem. It often happens when we cannot find differences when exploring data or the predicting model cannot provide good performance. To mitigate the risk, we fully consider the business problem, read related literature, find theory, and carefully select the features that should contribute to price prediction using the feature selection method. The second risk is the overfitting issue of the Random Forest model. To mitigate this risk, we split the dataset into a training set and testing set, and apply the model in the testing dataset to evaluate the model as well as avoid overfitting issues. Also, we tune the parameter in the Random Forest model to control the tree depth and the maximum number of iterations. We also introduce the ROC curve to examine the model performance. All efforts we make are aimed to control risk and ensure effective findings.

**Appendix**

**Individual Contribution:**

**Akshat Samir Patel (ap573):**

* **Cleaned the review.csv data file (comments data) and performed the sentiment analysis on the same to come up with multiple outputs(word cloud and sentiment charts)**
* **Wrote the business understanding, and sentiment analysis part in Modeling and Evaluation**

**Xin Dong (xd63):**

* **Wrote data preparation part of the report**
* **Helped with formulating the model**
* **Prepared slides and present**

**Xue Chen (xc183):**

* **Wrote deployment part**
* **Edited the format of the report**

**Shikha Dhurka (sd430):**

* **Data preparation and data mining - cleaned the listing.csv file**
* **Data modeling - MLR and random forest**
* **Wrote the data preparation and modeling part of the report**

**Yuhan Zhu (yz713):**

* **Data exploration and used Tableau for data visualization**
* **Wrote the data visualization part of report**

**Layla Lin (yl810):**

* **Wrote the evaluation part of report**

**Tableau Links:**

1. <https://public.tableau.com/app/profile/yu7490/viz/dsfinalproject-bostonairbnbmap-price/Sheet1#1>
2. <https://public.tableau.com/app/profile/yu7490/viz/dsfinalproject-bostonairbnbmap-price_hostneighbourhood/Sheet2?publish=yes>
3. <https://public.tableau.com/app/profile/yu7490/viz/dsfinalproject-bostonairbnb-propertytype/Sheet4?publish=yes>
4. <https://public.tableau.com/app/profile/yu7490/viz/dsfinalproject-bostonairbnbmap-price-bedrooms-bathrooms/Sheet3#1>
5. <https://public.tableau.com/app/profile/yu7490/viz/dsfinalproject-bostonairbnbmap-ofreviews/Sheet1#1>

**Research Links:**

1. <https://www.wsj.com/articles/airbnb-has-mo-money-mo-problems-11620944283>
2. <https://www.airbnb.com/resources/hosting-homes/a/the-best-amenities-to-offer-right-now-203>
3. <https://www.airbnb.com/resources/hosting-homes/a/how-to-earn-money-on-airbnb-282>
4. <https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-deep-learning-f46d44afb8a6>
5. <https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a>