Airbnb Data Analysis

Sweta Patel - 100915164
Peter El Khoury - 100861813
Agnes Li - 100673905

Professor- Amir Rastpour

Outline

- Business Case
- Dataset
- Preliminary Data Analytics
- Exploratory Data Analytics
- Emotional Analytics
- Feature Importance
- Model Building
- Findings and Summary

Business Case

- The question to investigate is;
  - What are the factors affecting the pricing of the listing?
    - This is essential for the hosts.
      - Setting the price correctly is crucial for Airbnb because an excessively expensive listing will result in fewer bookings, and an excessively low listing will result in a loss of profit for Airbnb. Regression analysis will aid in making some degree of accurate price prediction.
      - Based on sentimental analysis of the customer reviews, hosts will be able to understand the emotions of the customers and make necessary improvements as when required.

Dataset

- The dataset is collected from insideairbnb.com which is mission driven project that provides research and advocacy on the impact of Airbnb on residential areas.
- For our use case, we are using the Airbnb dataset for Toronto region.
- The dataset contains the information about listings and the customer reviews for every listing starting from the year 2009 to 2023.
- Listings dataset contains information about the hosts, properties and different review scores.
- It contains 18921 observations with 75 features.
- Reviews dataset contains information about the customer reviews for every listing.
- It contains 4,82,000 observations with 6 features.

id	property_type
name	room_type
description	accommodates
host_id	bedrooms
host_name	bathrooms_text
host_since	beds
host_location	amenities
host_response_rate	price
host_acceptance_rate	has_availability
host_response_time	availability_30
host_is_superhost	number_of_reviews
host_identity_verified	review_scores_rating
host_verifications	review_scores_cleanliness
neighbourhood_cleansed	review_scores_communication
latitude	review_scores_checkin
longitude	review_scores_location

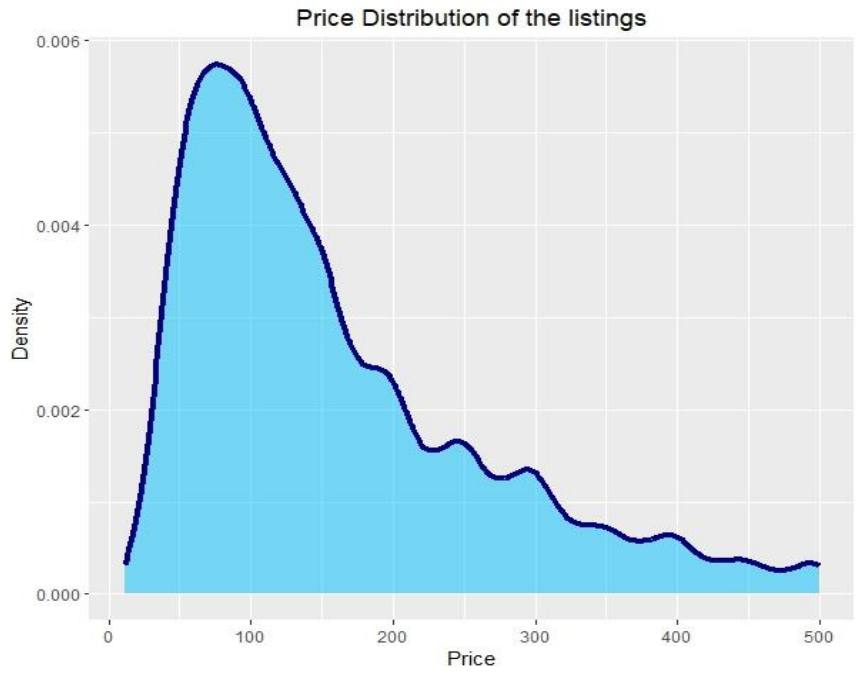
Listings Table

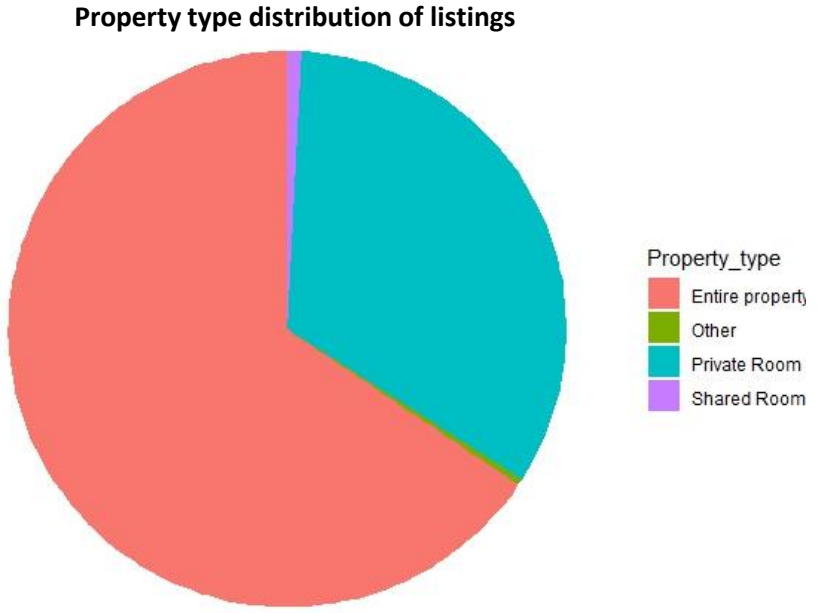
id
date
listing_id
reviewer_id
reviewer_name
comments

Reviews Table

Preliminary Data Analytics

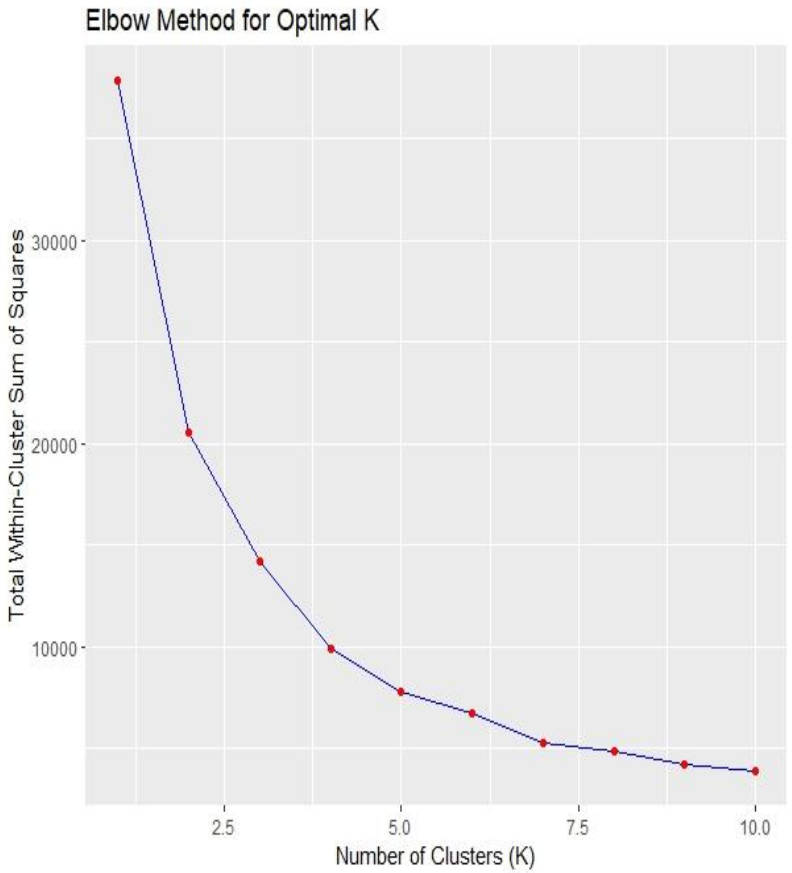
- **Missing or null values-** Getting sum of null values for all variables and the ones with all null values like bathrooms were eliminated from the dataset. For some important features like review_scores_rating, null values were replaced by 0.
- **Removal of unimportant features-** Many variables like listing_url, name, description, host_url,host_name etc. were removed.
- **Removing outliers-** We analysed that about 95% of our data have the prices in the range of 0 to $500 and hence we removed the rest.
- **Releveling of data-** Columns like property_type has lot of levels, so they are reduced to three levels namely shared room, private room, hotel room and other depending upon the input.
- Similarly with amenities, new columns like basics, facilities, parking, bath_essentials, kitchen_appliances , safety_measures and long_term_stays_allowed and the original column is dropped.

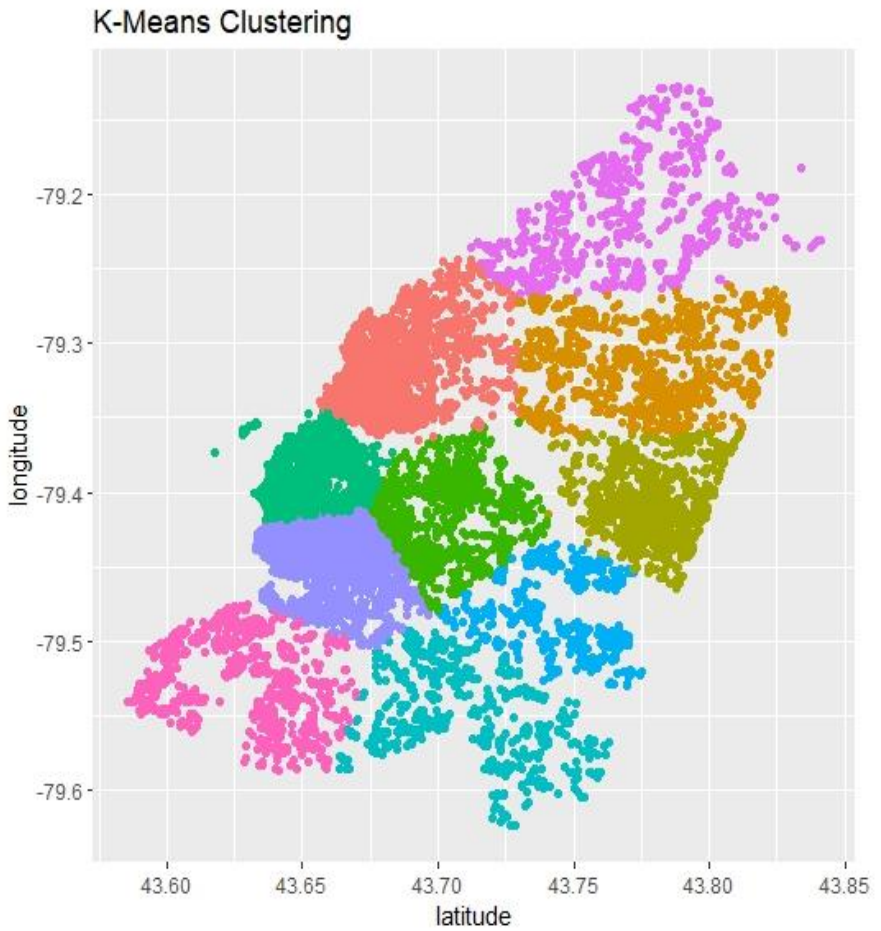


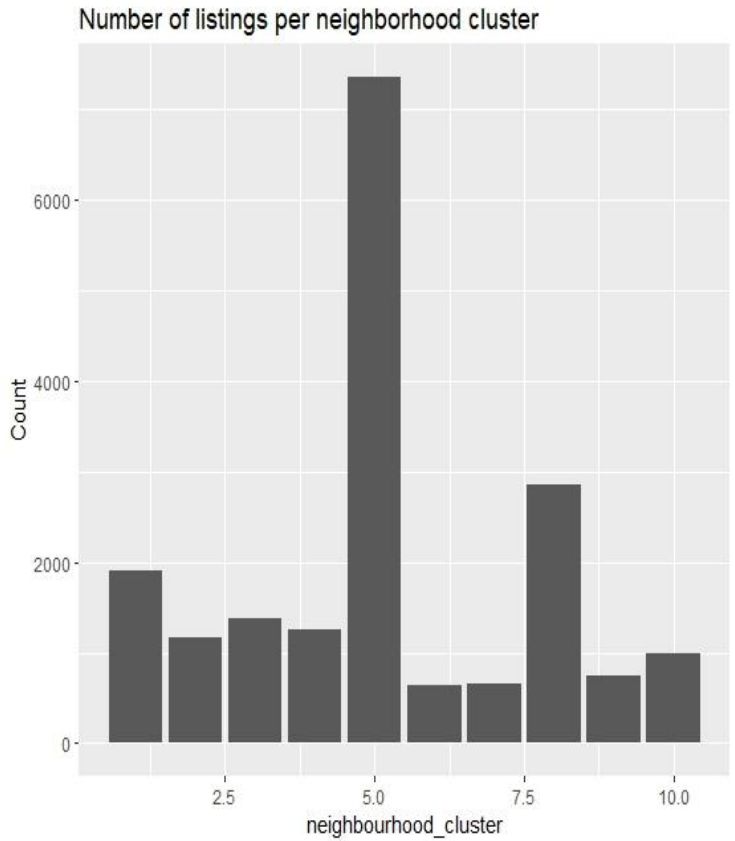


Preliminary Data Analytics

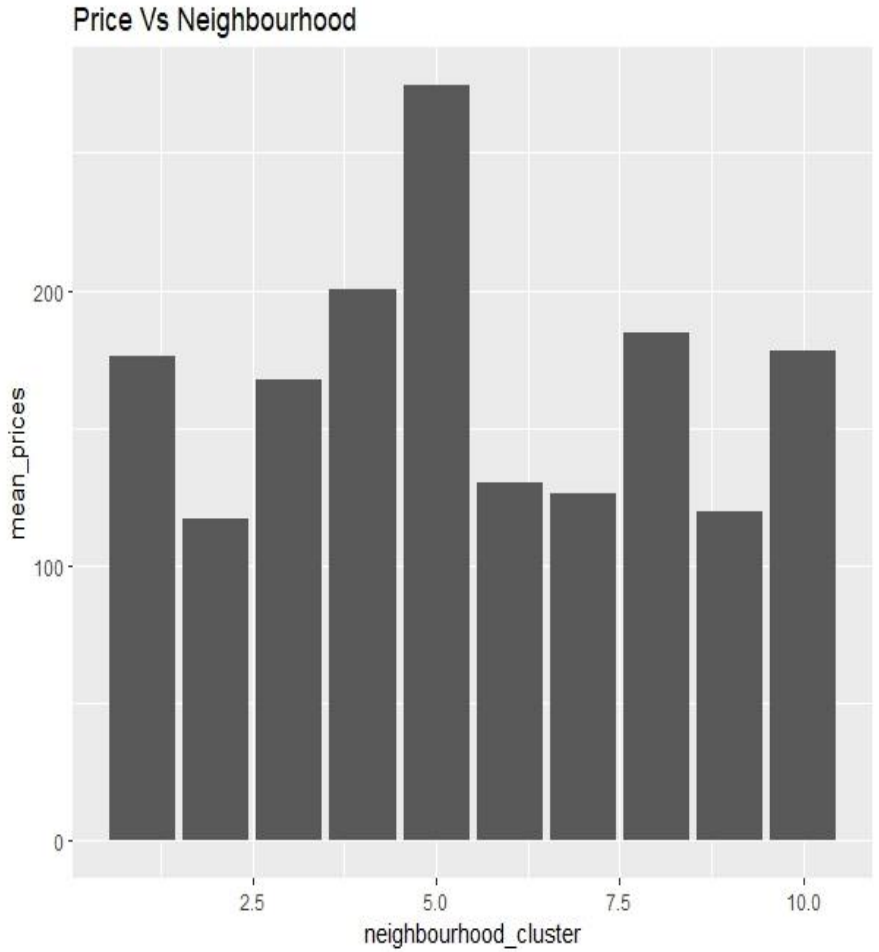
- Columns like latitude and longitude of listing were used to group listings in clusters using KNN algorithm.

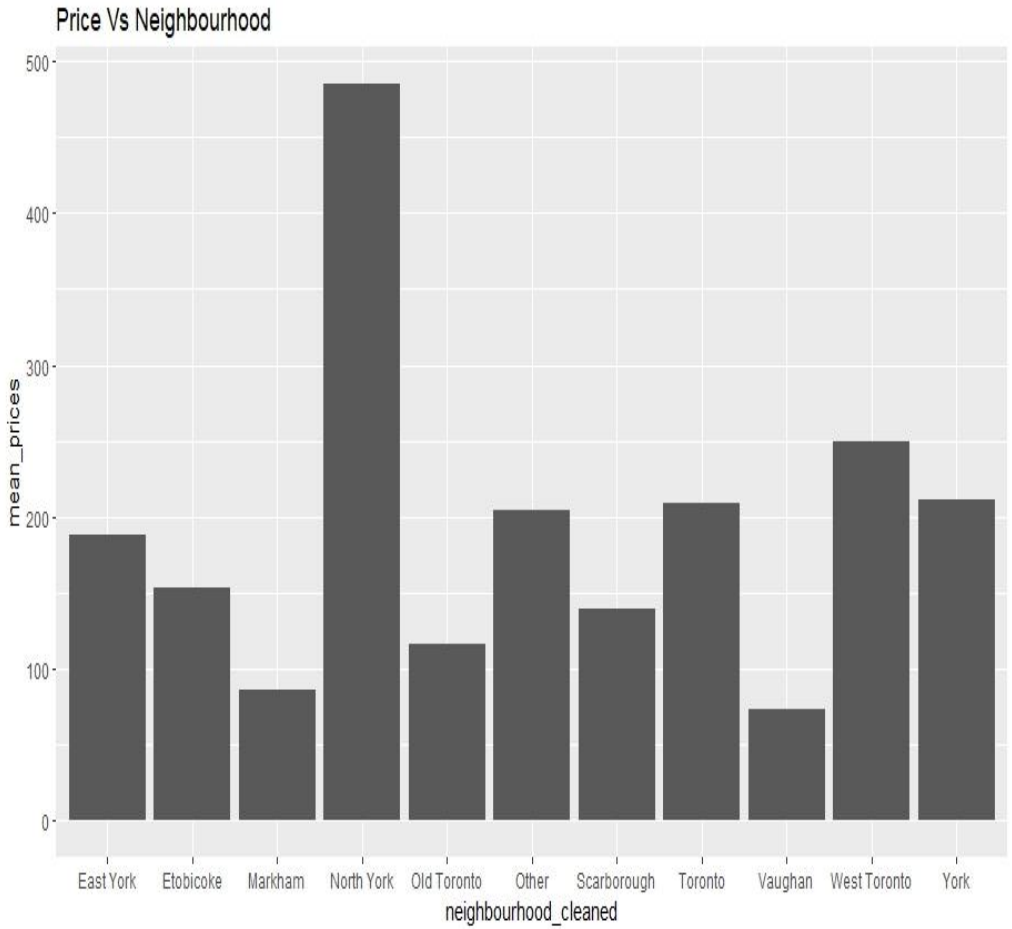




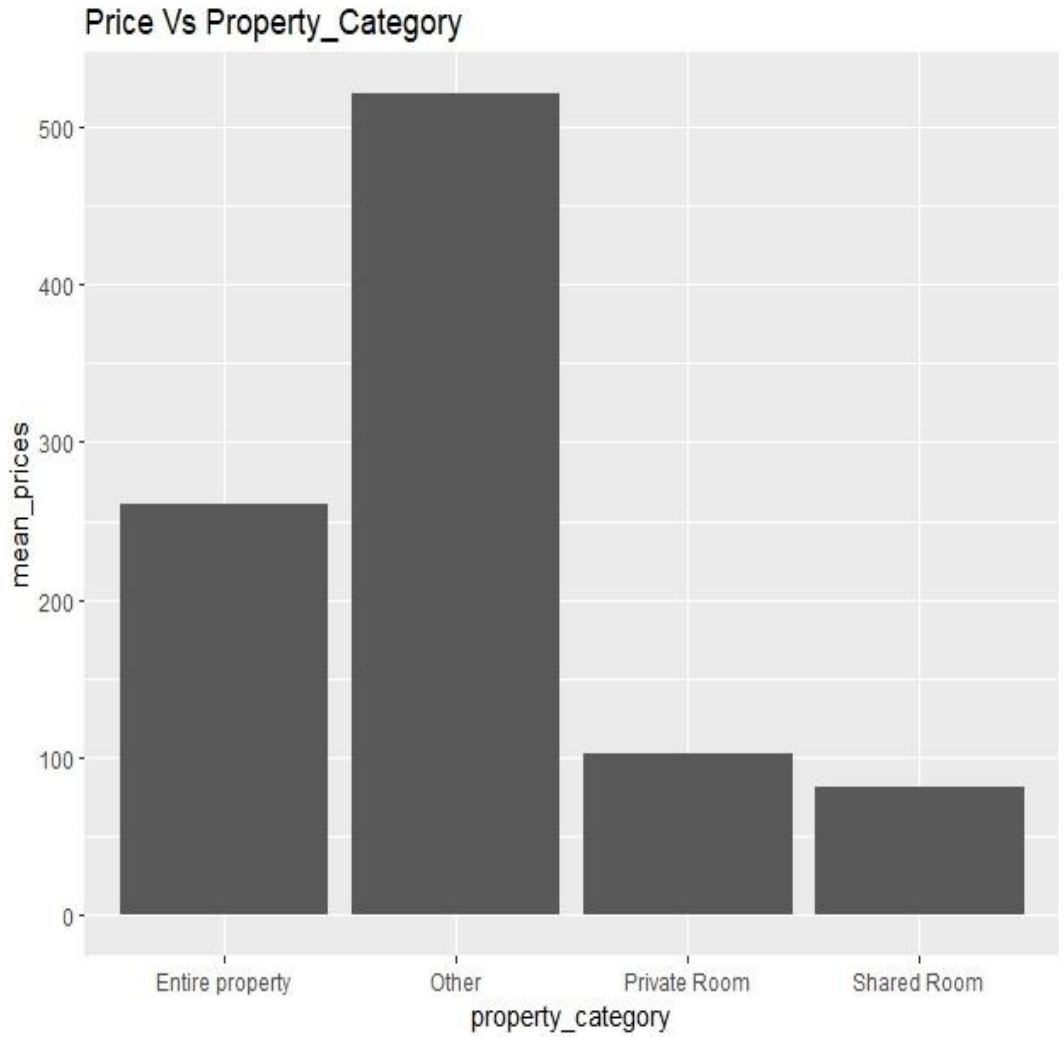


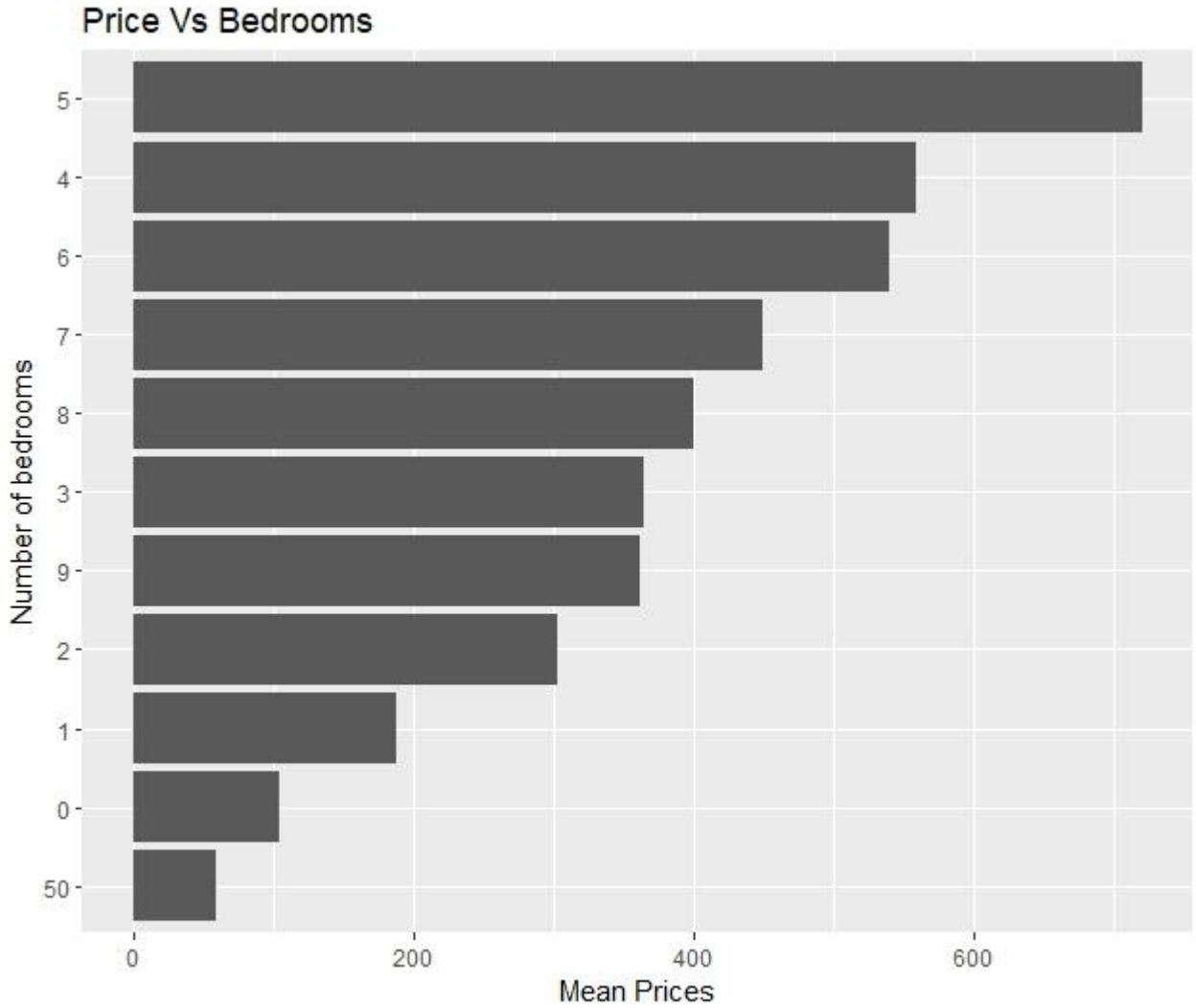
Exploratory Data Analysis



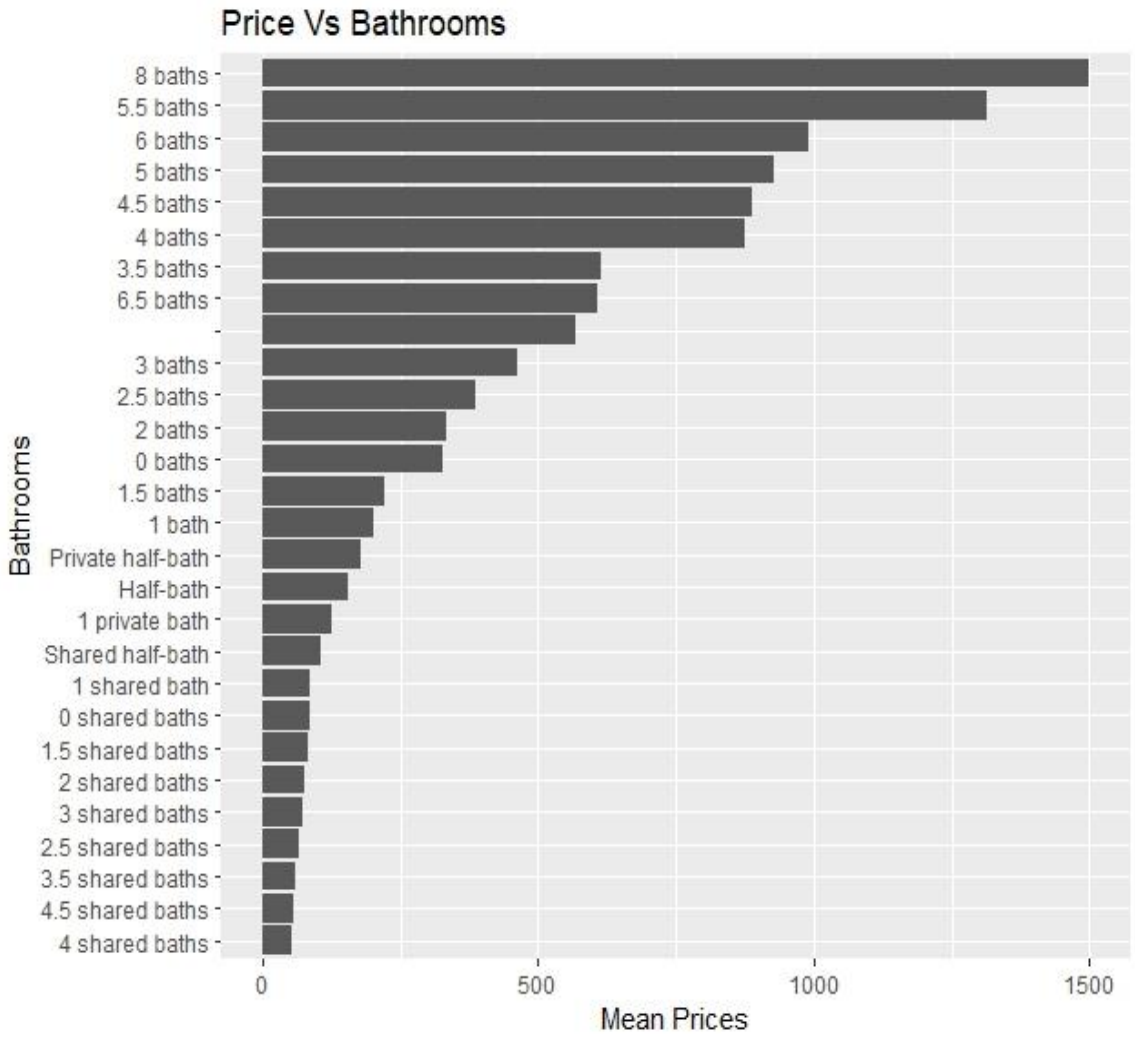


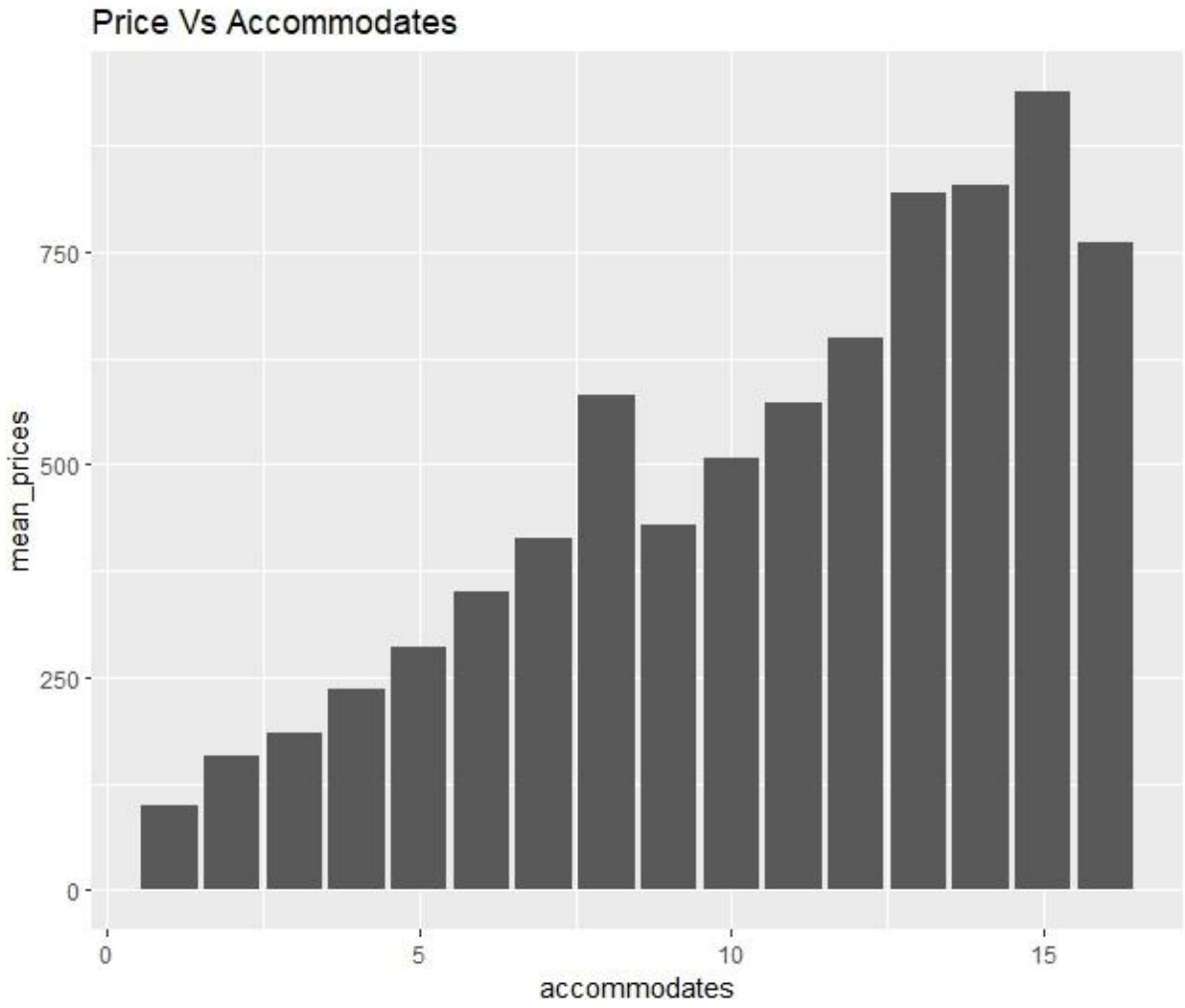
Exploratory Data Analysis





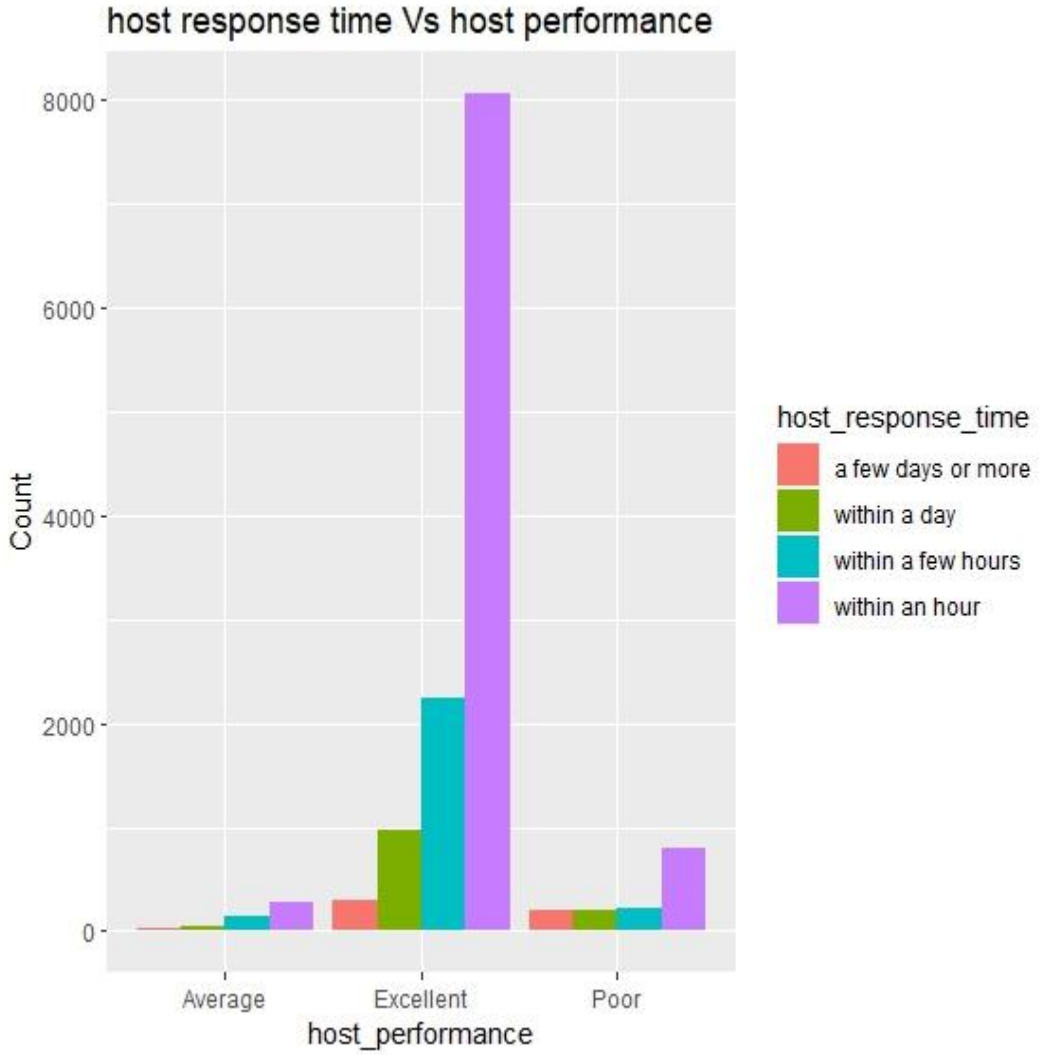
Exploratory Data Analysis

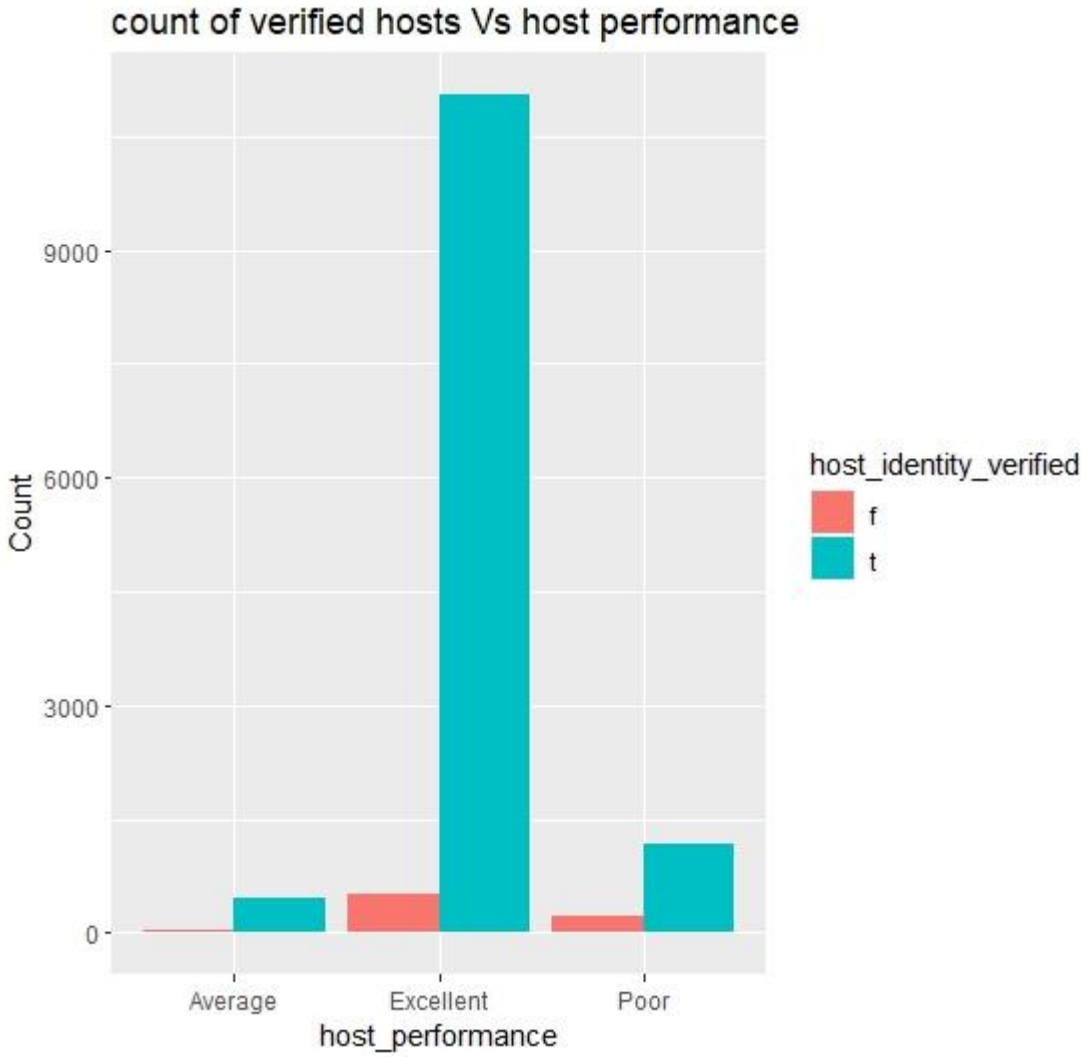




Exploratory Data Analysis

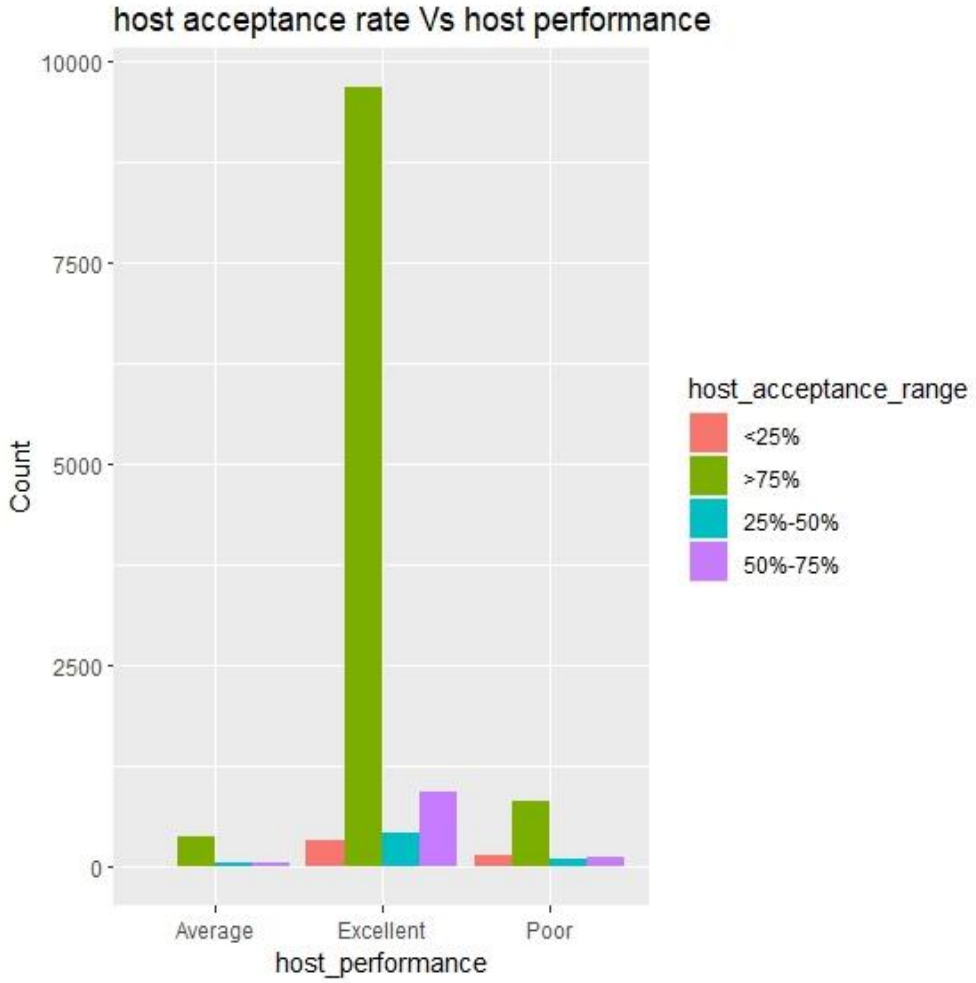
- Based on customer review scores, performance of host is categorized as average, excellent and poor and then the traits of host is analysed.

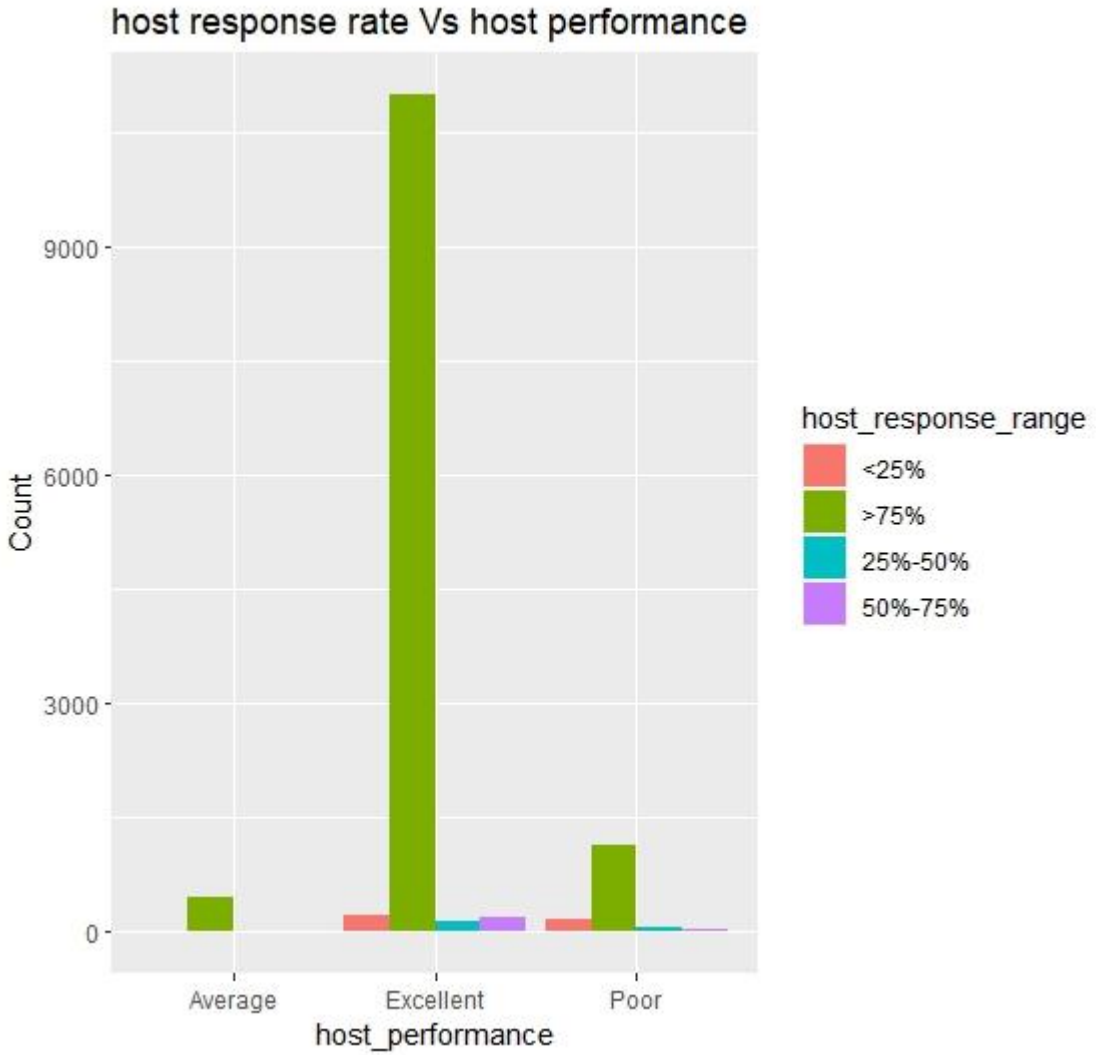




Exploratory Data Analysis

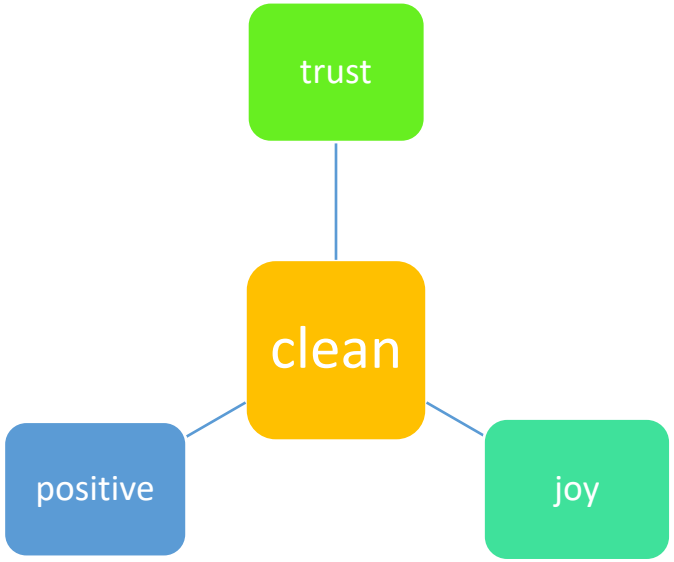
- Based on customer review scores, performance of host is categorized as average, excellent and poor and then traits of host is analysed.

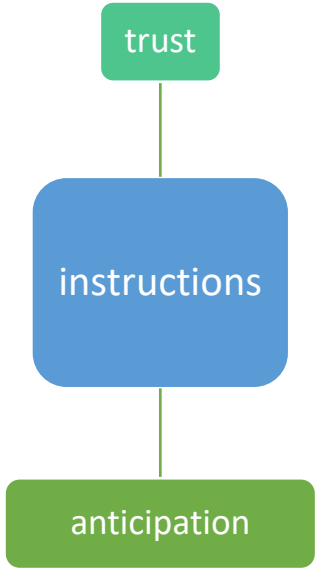


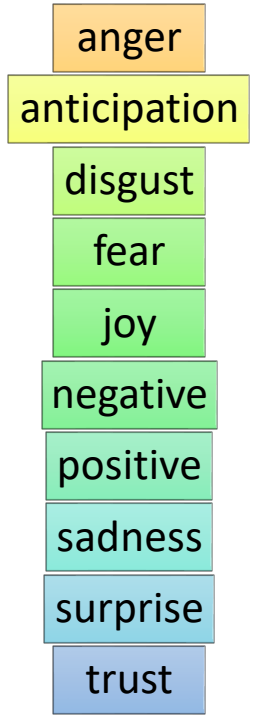


Review Analysis / Emotional Analysis

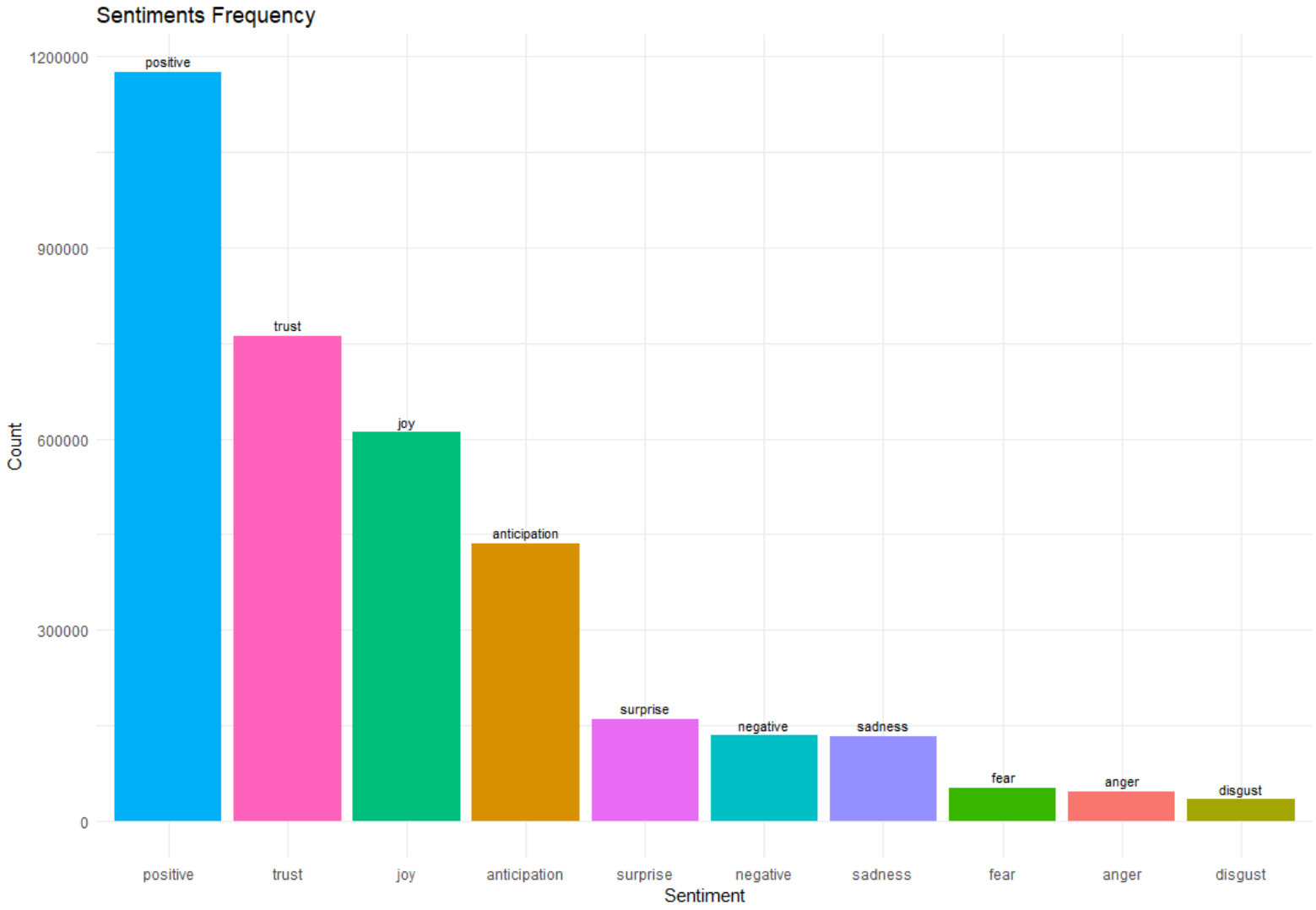
- An emotional lexicon called the NRC Lexicon (National Research Council) attributes emotions to English words.
- The comments on all the listings were analyzed.

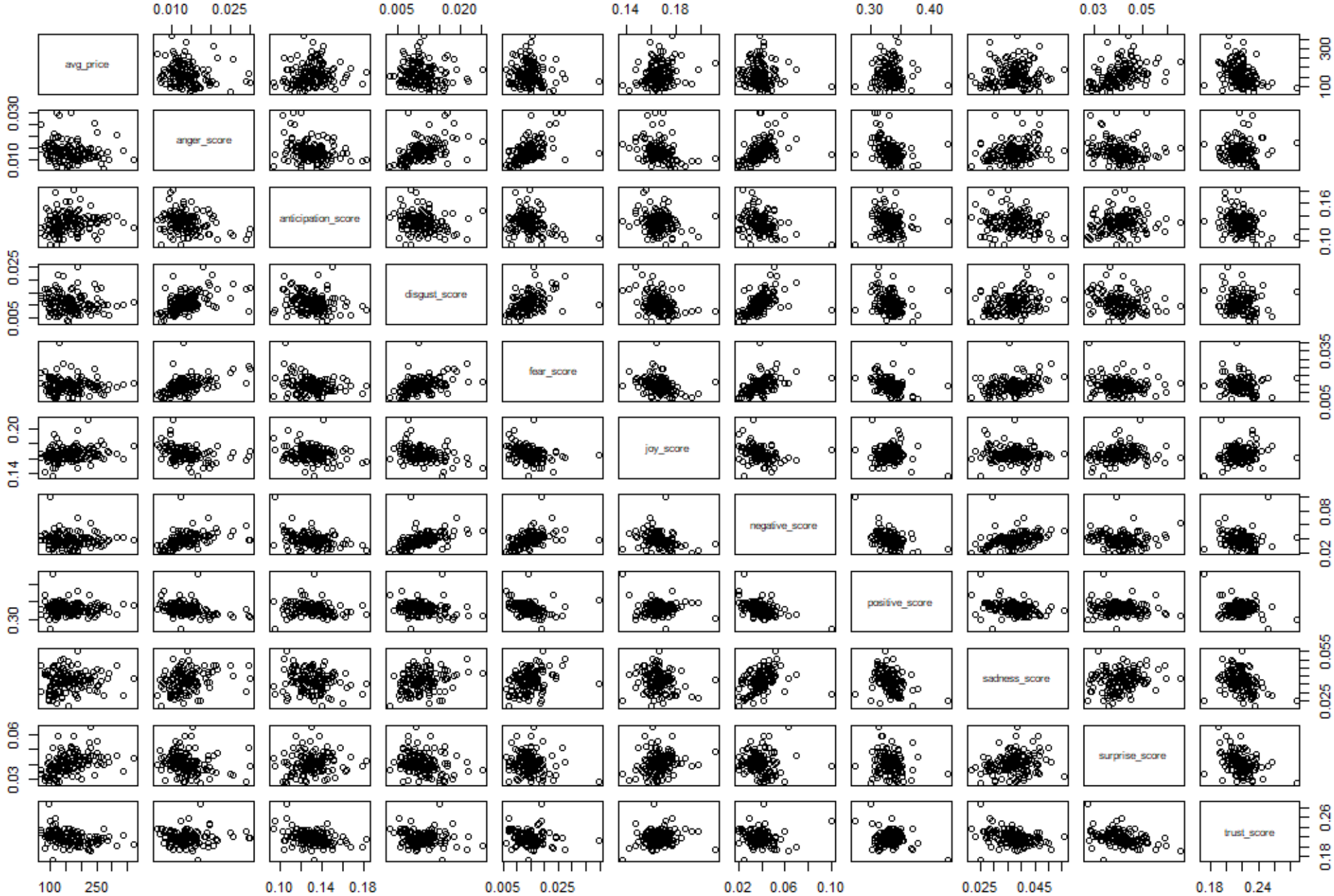


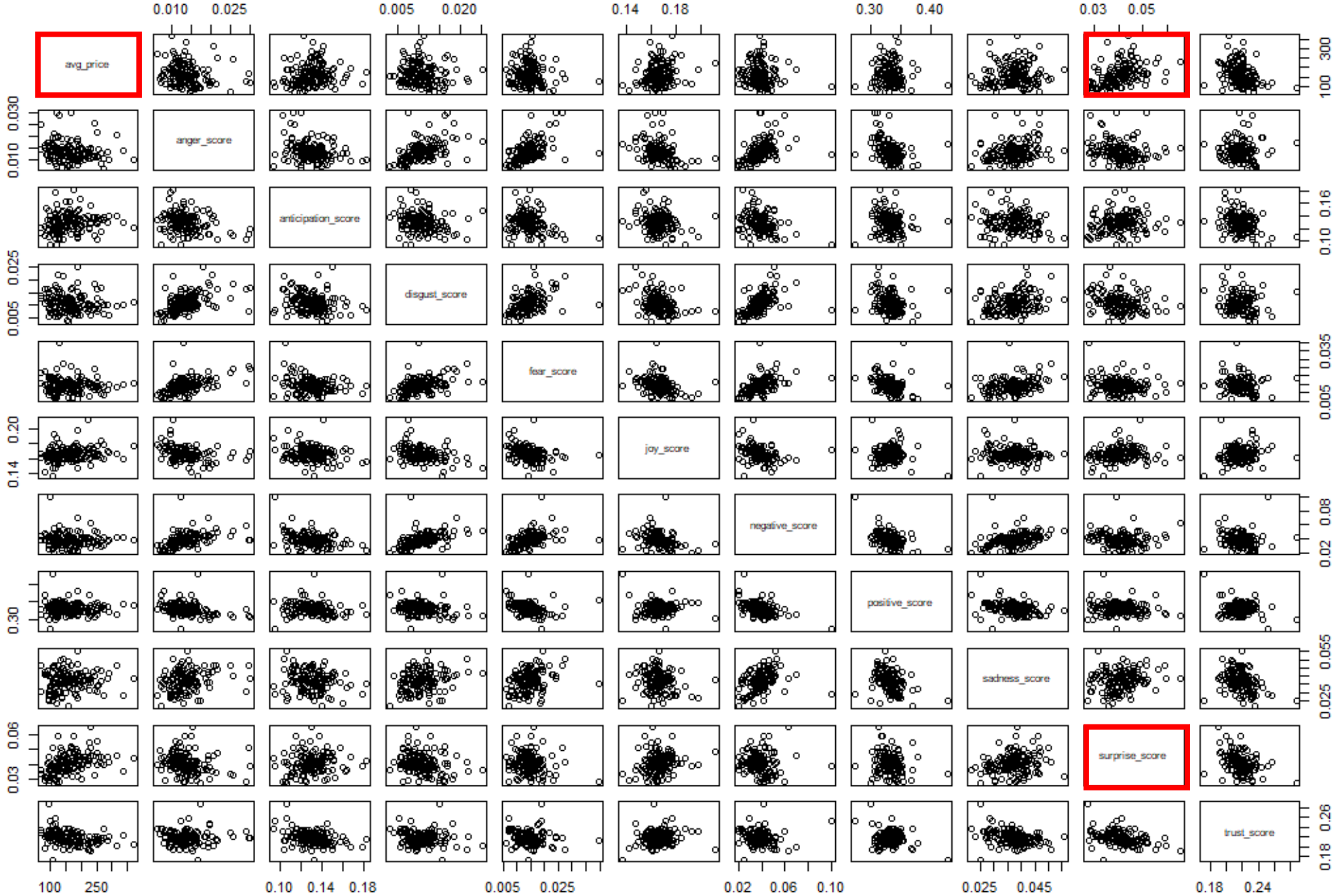




Sentiments Frequencies



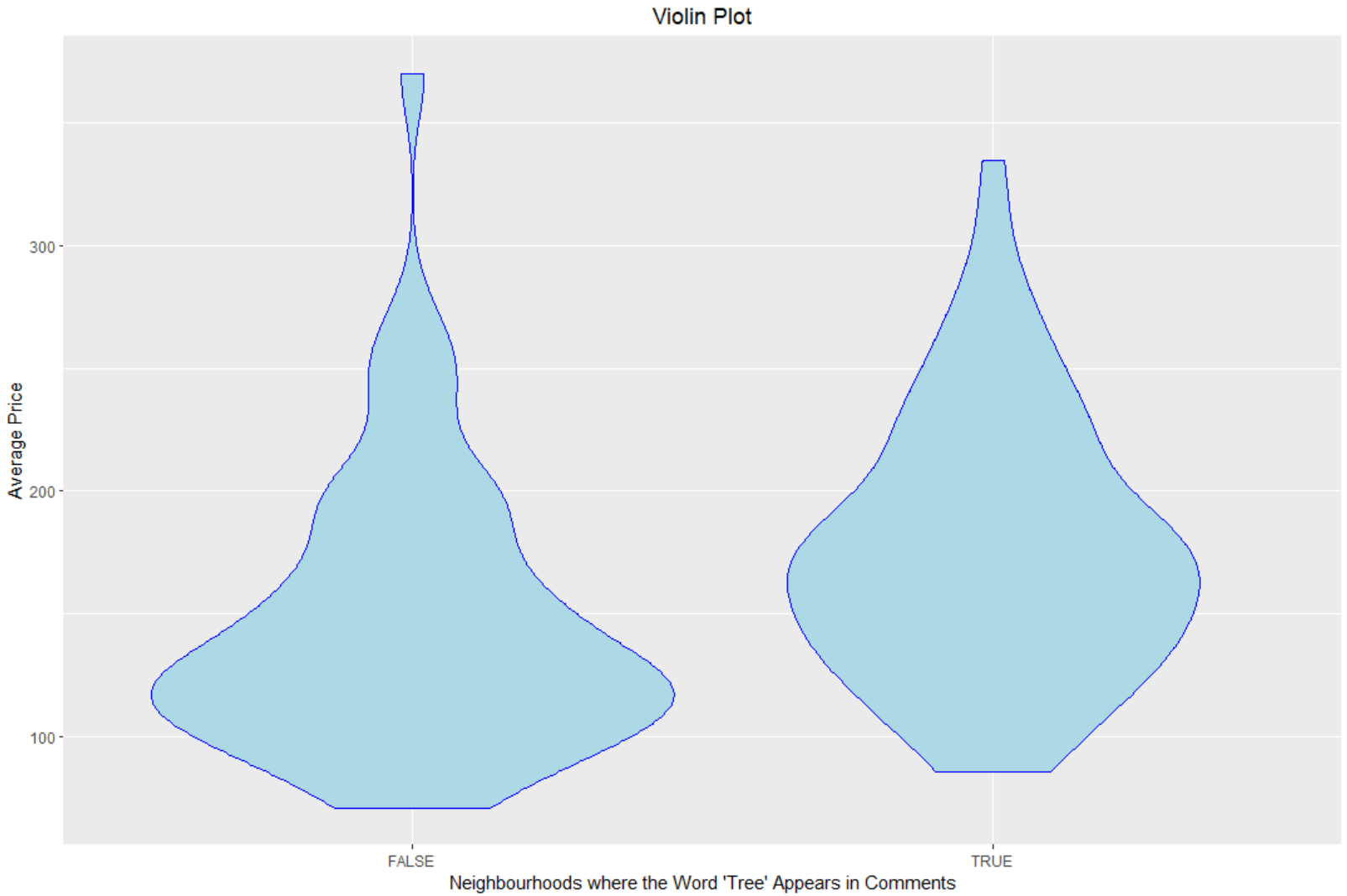

avg_price
anger_score
anticipation_score
disgust_score
fear_score
joy_score
negative_score
positive_score
sadness_score
surprise_score
trust_score


avg_price
anger_score
anticipation_score
disgust_score
fear_score
joy_score
negative_score
positive_score
sadness_score
surprise_score
trust_score

Top Words for Surprise

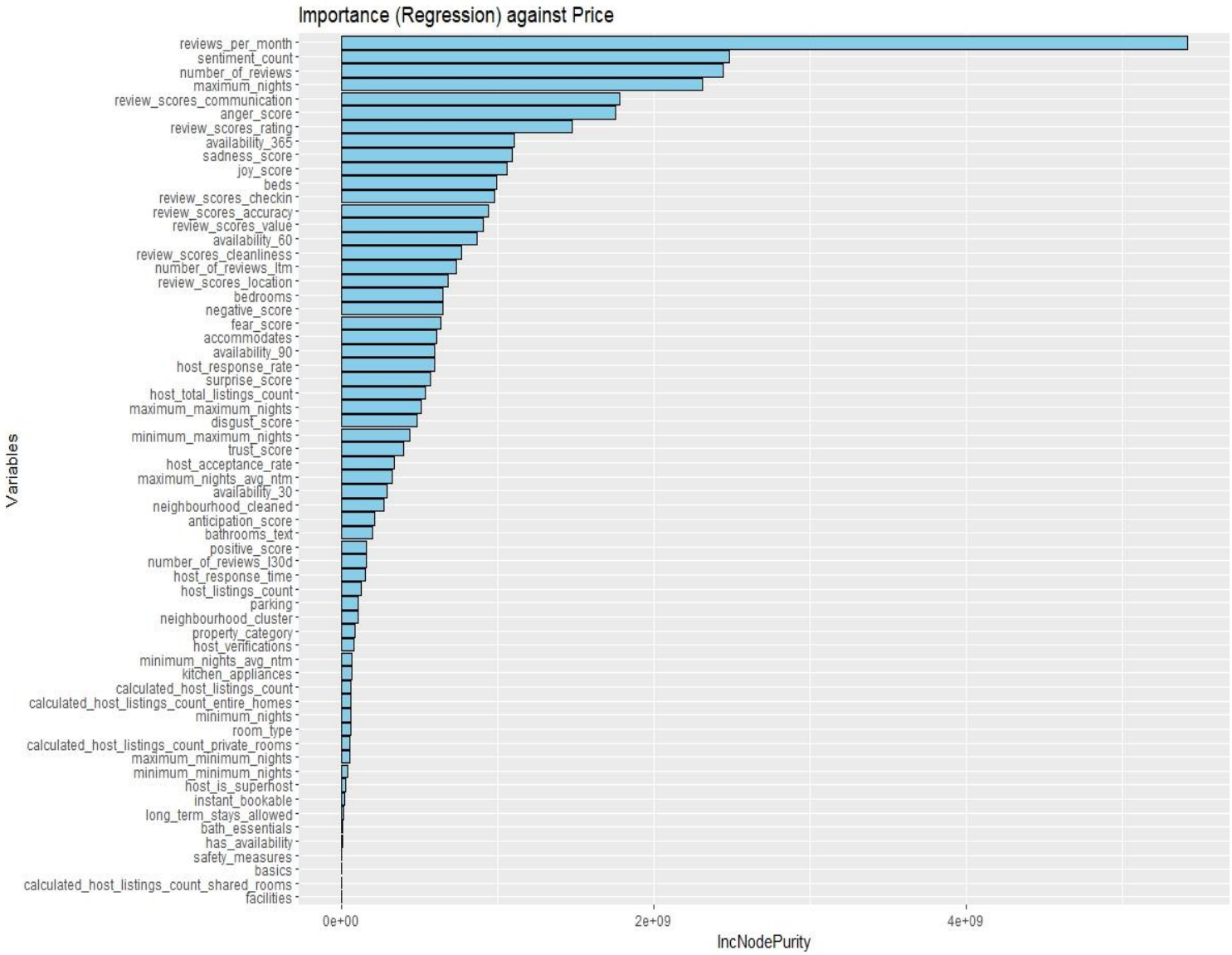
- The word tree is of interest
- Let's see how it affects the avg. price per neighborhood.

Top Words for Surprise



Feature Importance

- Used RandomForest Algorithm to find out the significance of all the variables on our dependent variable which is price.
- IncNodePurity- The amount that the model error rises when a specific variable is randomly permuted or shuffled.



Random Forest for Sentiment Scores

- ***Anticipation*** had the highest impact on price.

- Top words conveying anticipation corroborate previous findings.

Building the Models

- Two models were built:
  - Linear Regression Model
  - Random Forest

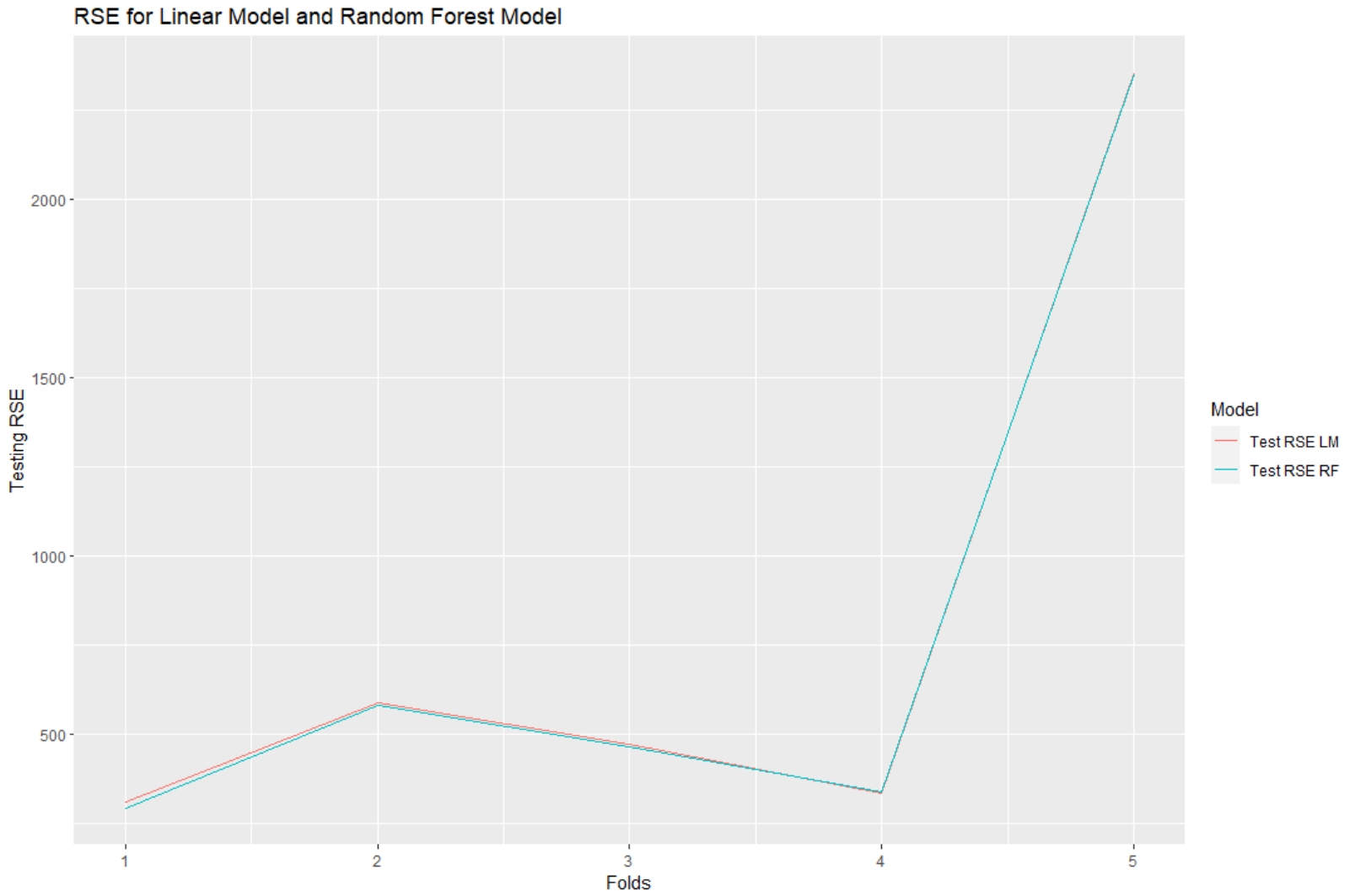
- The top 50 variables from the IncNodePurity bar chart were used.

- We compared the two models used 5-Fold cross-validation and then calculated the average RSEs for the test set.

Building the Models

- Linear Model:
  - $k = 5$
  - Mean RSE = 812
- Random Forest:
  - $m = \sqrt{p} = \sqrt{51} = 7$
  - Number of trees = 100
  - Mean RSE = 805

Building the Models



Findings and Summary

- Through our analysis we conclude the following findings.

1) Assist hosts in determining the factors that influence the pricing of a listing.
2) Before posting a listing on Airbnb, our model provides hosts with a price proposal based on all relevant factors such as the listing's location, listed properties, available amenities, and so on.
3) Assist hosts in becoming aware of the characteristics they must maintain, such as response rate and acceptance rate, in order to better serve clients.
4) The model could help users understand the reviews of a listing using the emotion metric.

Questions