Data Selection Report

Dataset

I have selected a dataset of AirBnb from "Data.World". This dataset gives us information about the Netherlands and particularly of Amsterdam. It has several dimensions that will allow to implement tools and techniques to get the required results.

Dimensions: 31 (Numeric and Non-Numeric)

Observations: 7285

Why this dataset?

Airbnb is a company which provide facility of accommodation to the people who travel from one place to another for a specific amount of time on their vacations, holidays, study visit, business visit and visiting to their family members. Airbnb provide vacation rentals, cabins, beach houses, unique house, apartments, private accommodation and much more to their customers for an initial push at a new place.

I got interest in this dataset because I am much passionate about travelling and visiting new places and explore the best places of world. So it's quite a task to choose and select a place for your living before planning to travel a place where you don't know much about it. Secondly, I am an international student and travelled to UK for my higher studies, so in my recent experience of travelling I suffered bit about living. This enables my interest to work on such a dataset and understand the analytics about Airbnb.

Aim and Goal

Aim of using dataset is to make a quick and correct decision while selecting your accommodation. Every person has his/her own choice and preferences and we must consider those. In this regard, some preferences and requirements of customers will be taken and analyze using data analytic techniques to classify some good options for the customer, moreover, based on number of accommodates, bedrooms, beds, etc a price will be predict to give an idea to the customer and make it easy to make decision.

Data Visualization

In order to select a dataset, a slight analysis is necessary to implement the different algorithms from various categories that include regression, classification and clustering. Basically we are two categories supervised and unsupervised learning. There are few algorithms which I will be using in this assessment.

Supervised Learning

1. Regression
   a. Linear Regression
   b. Logistic Regression
   c. Lasso Regression
   d. Multivariate Regression

   In this regression model we need to choose two algorithms. My first choice is Linear Regression and second it not yet decided but will be selected from the above options

2. Classification
   a. KNN
   b. SVM
   c. Decision Tree
   d. Random Forest

   Not yet decided about any algorithm.

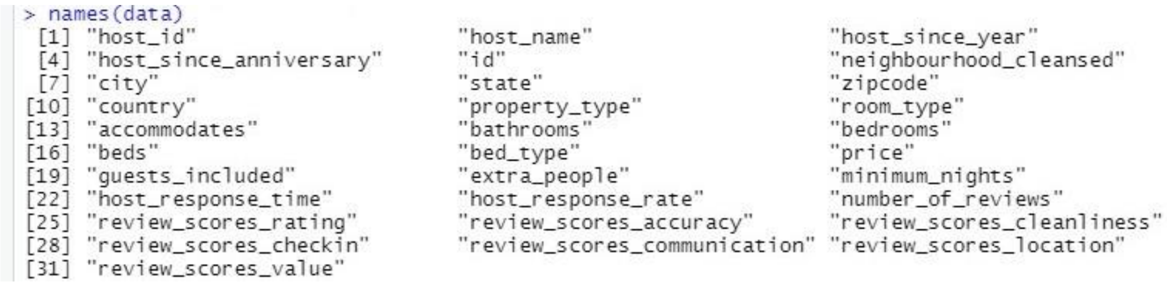
Unsupervised Learning

1. Clustering
   a. K-means clustering
   b. Clustering Dataset
   c. OPTICS
   d. Affinity propagation

Keeping in view the above algorithms, dataset is analyzed whether it suits the algorithms and can these algorithms apply on a selected dataset. Some visualizations are listed below.

Dimensions

Below are the names of dimension of selected dataset



```
> names(data)
 [1] "host_id"                   "host_name"                     "host_since_year"
 [4] "host_since_anniversary"    "id"                            "neighbourhood_cleansed"
 [7] "city"                      "state"                         "zipcode"
[10] "country"                   "property_type"                 "room_type"
[13] "accommodates"              "bathrooms"                     "bedrooms"
[16] "beds"                      "bed_type"                      "price"
[19] "guests_included"           "extra_people"                  "minimum_nights"
[22] "host_response_time"        "host_response_rate"            "number_of_reviews"
[25] "review_scores_rating"      "review_scores_accuracy"        "review_scores_cleanliness"
[28] "review_scores_checkin"     "review_scores_communication"   "review_scores_location"
[31] "review_scores_value"
```

Scatterplots and Boxplot of some dimensions

Boxplots

1. Price

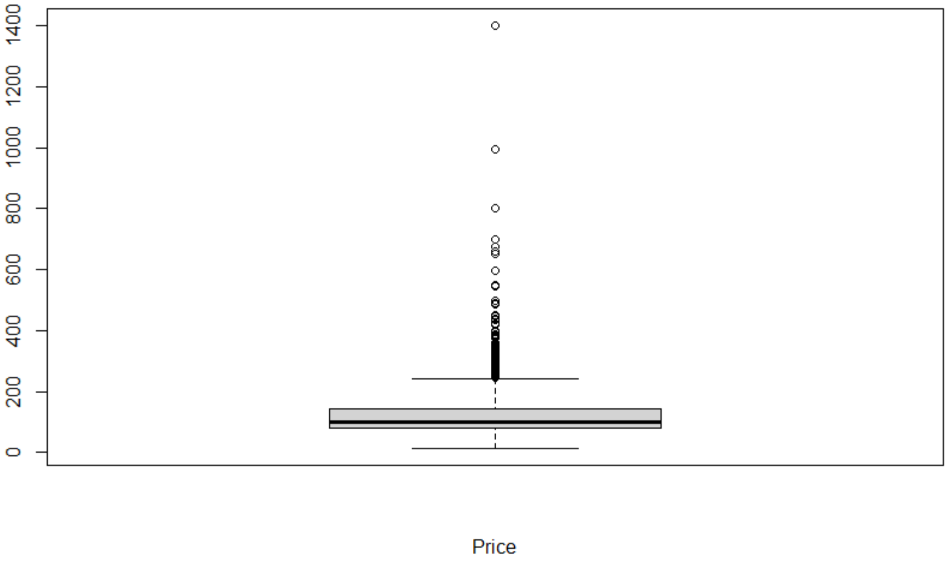


Figure: Before Data Cleansing

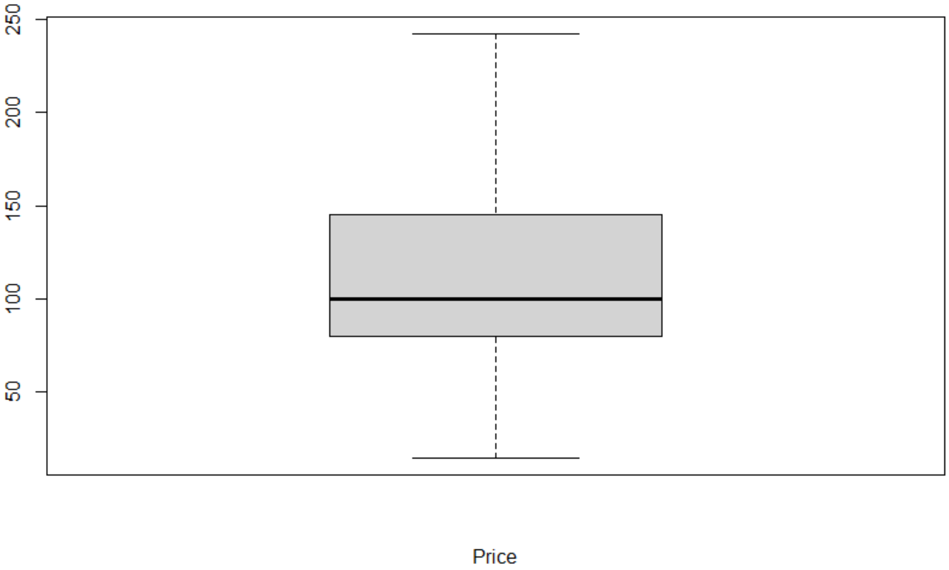
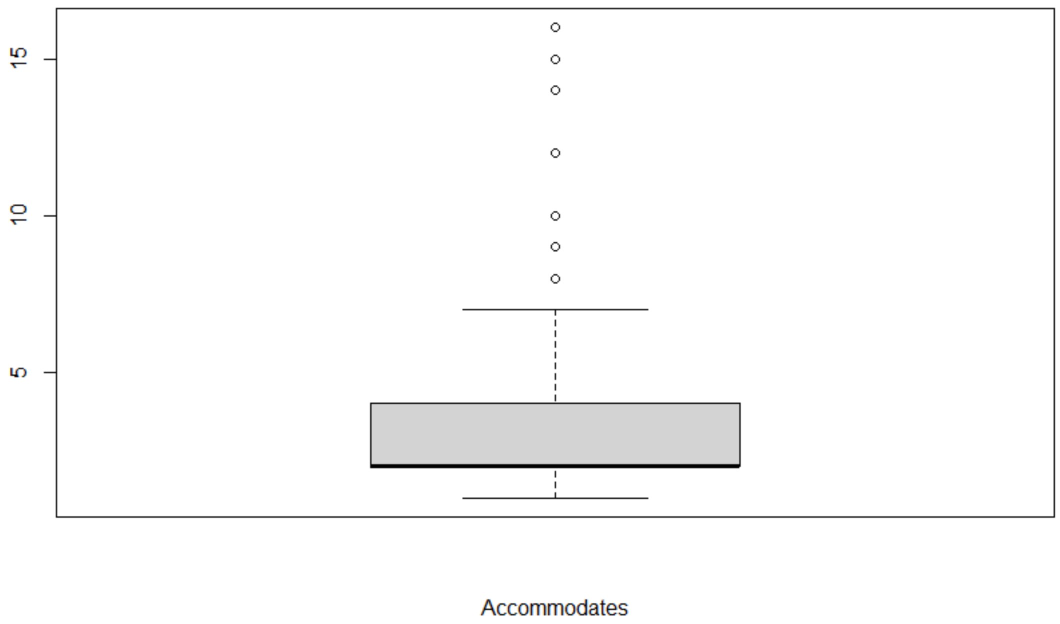
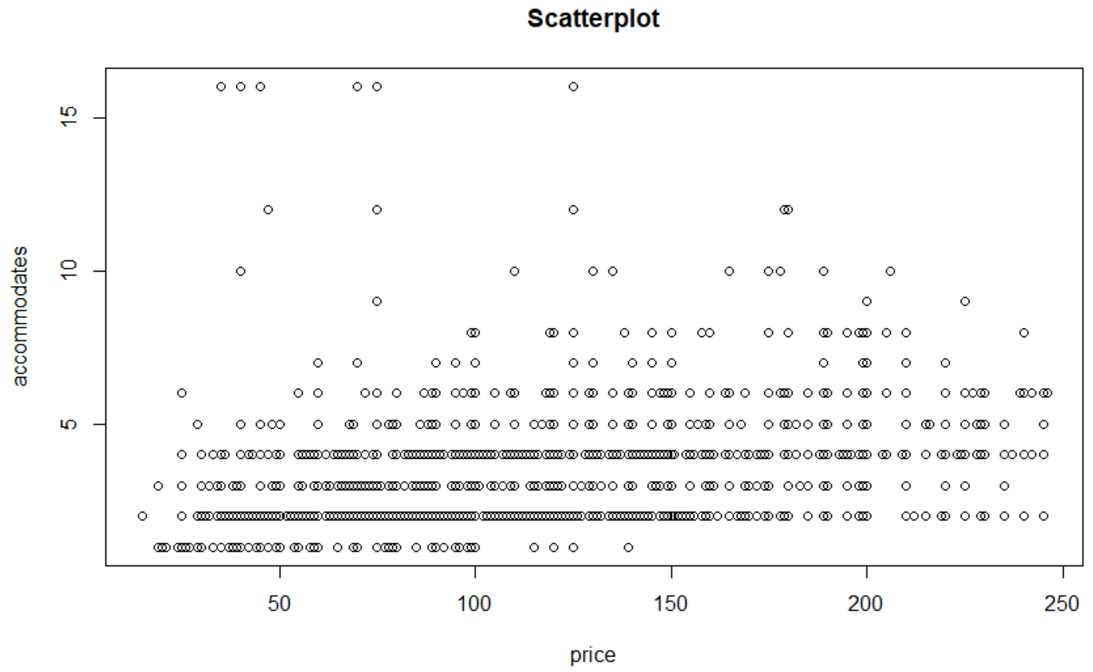


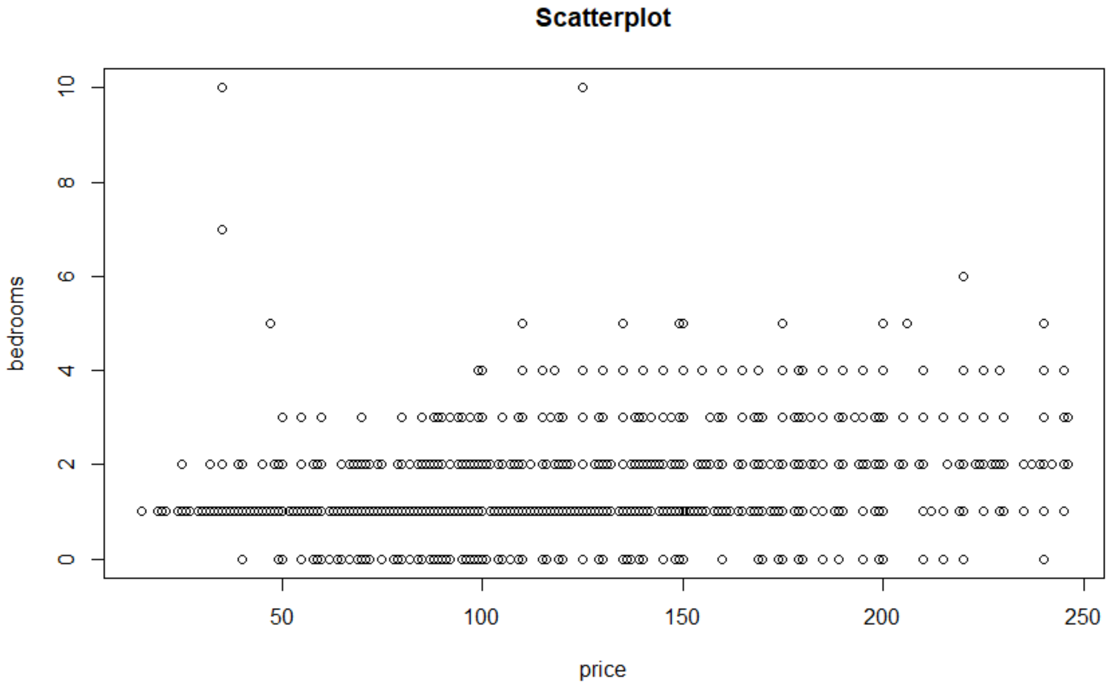
Figure: After Data Cleansing

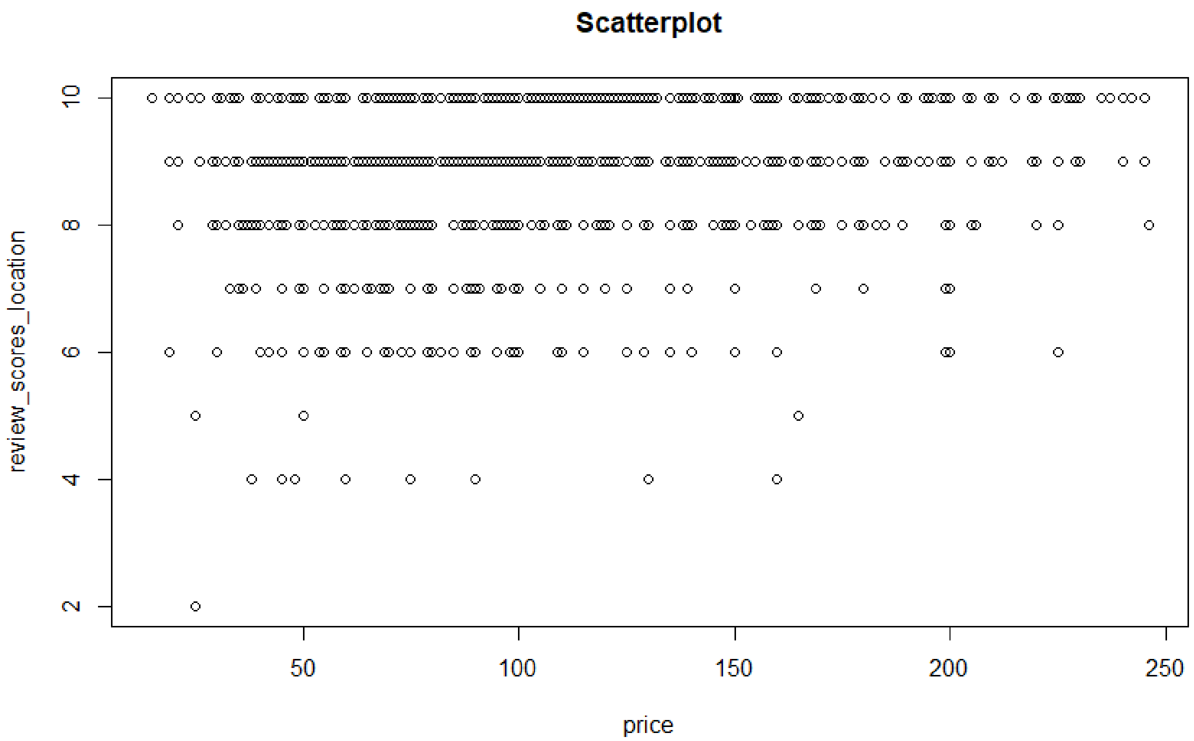
2. Accommodates

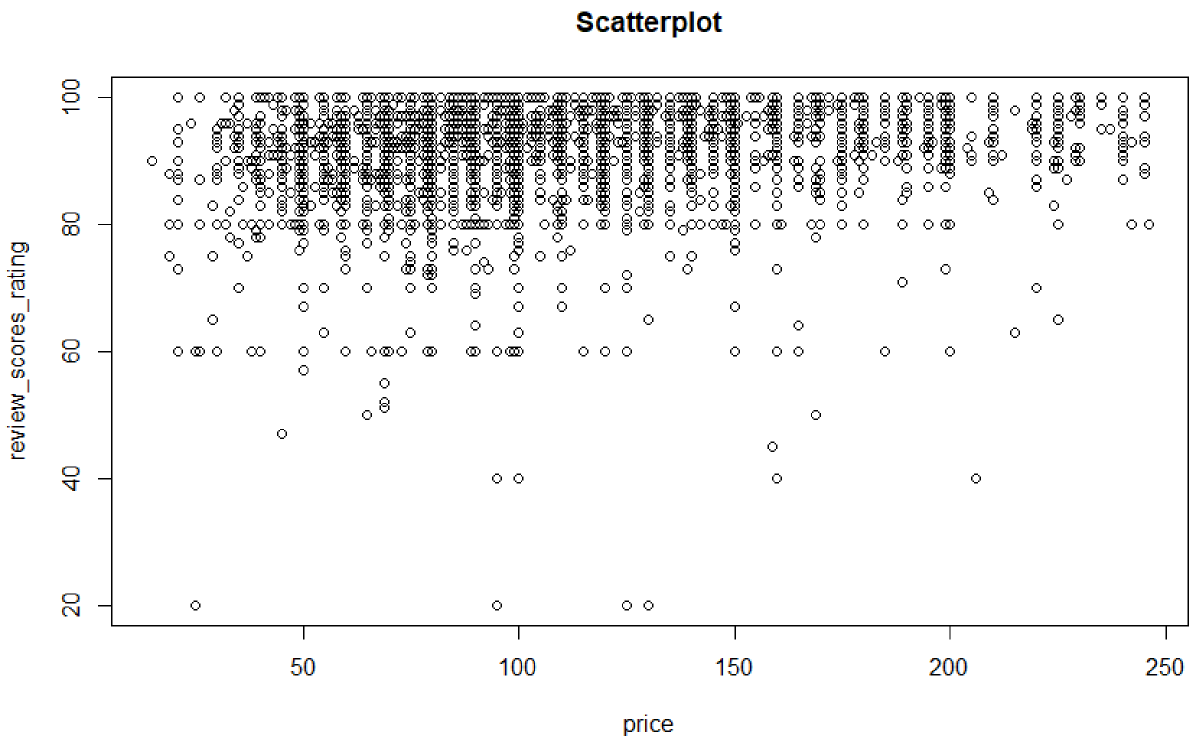


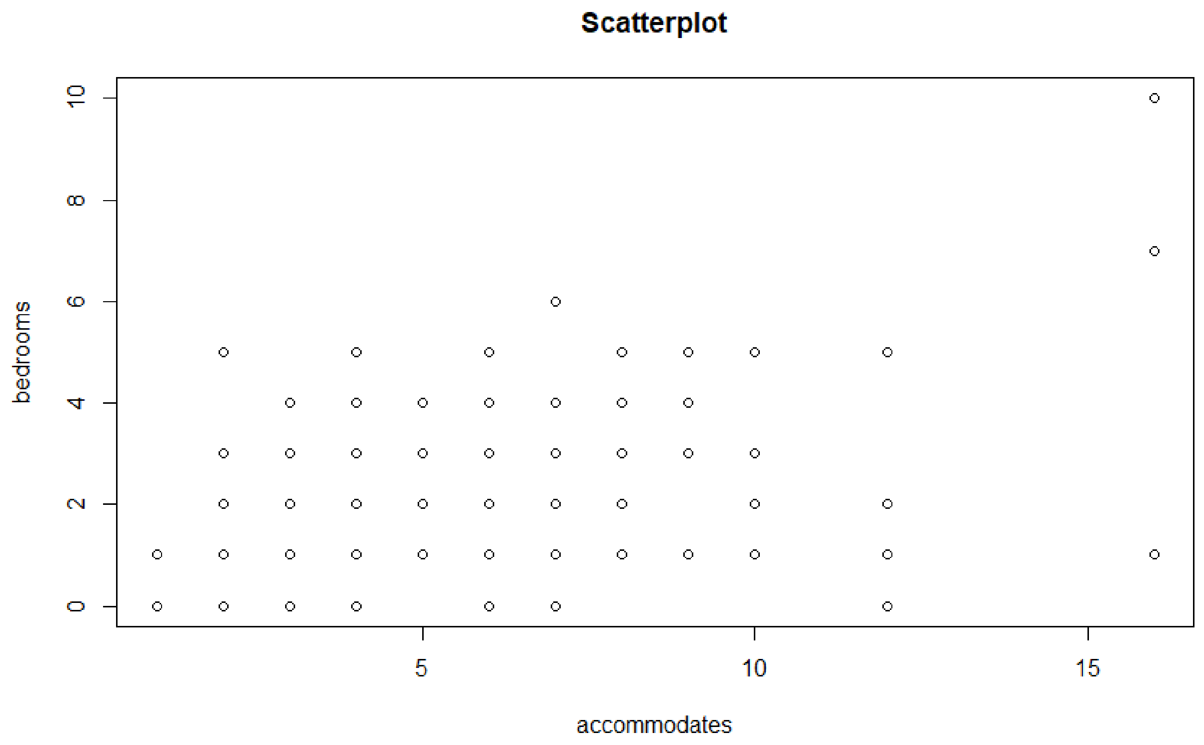
Scatterplots

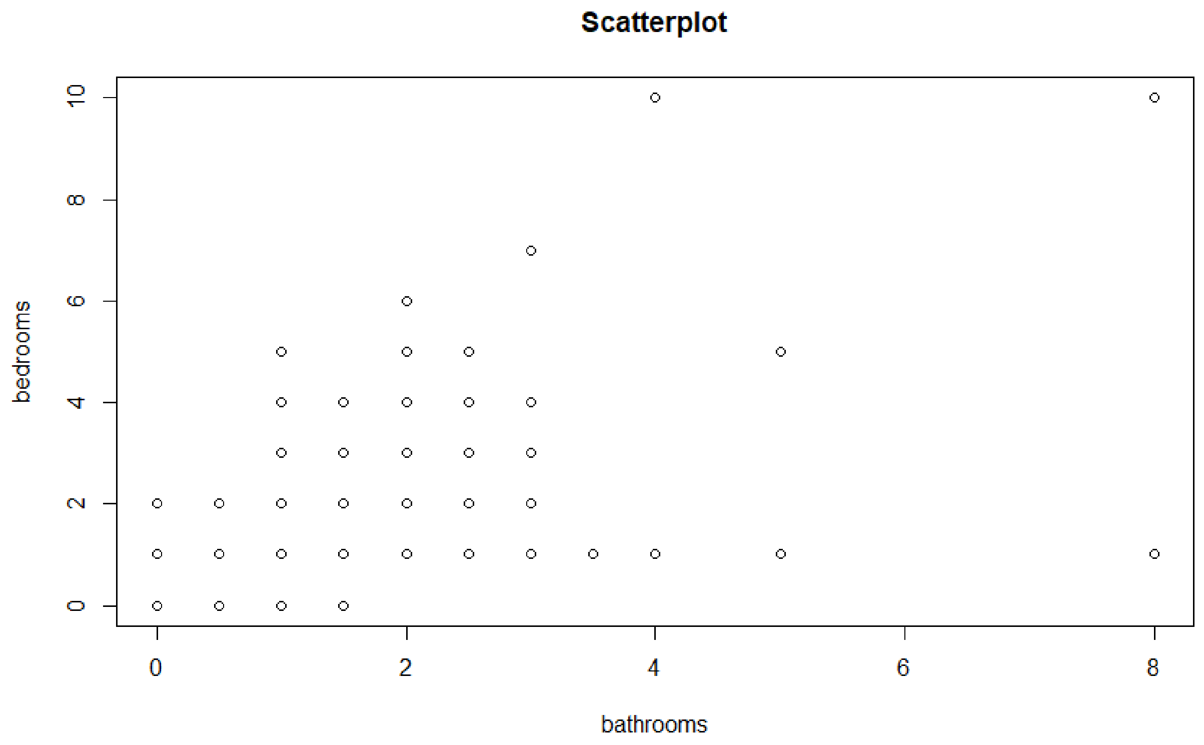


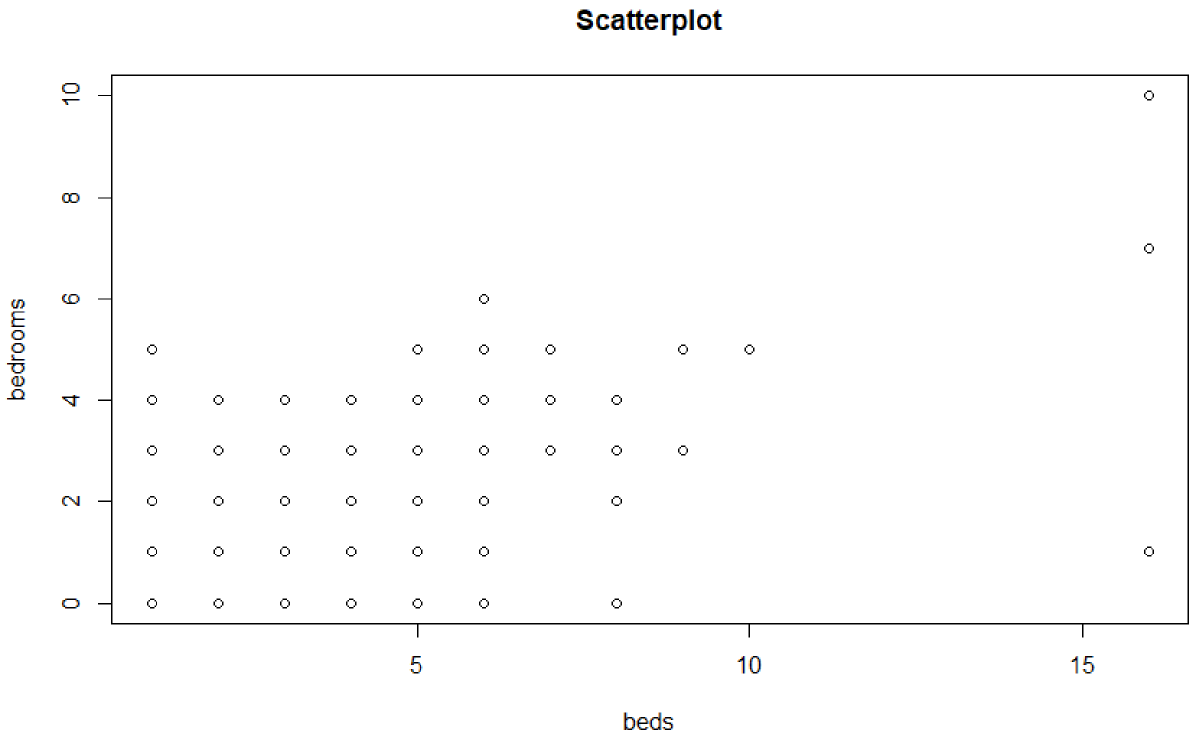



Scatterplot
review_scores_location
2
4
6
8
10
50
100
150
200
250
price

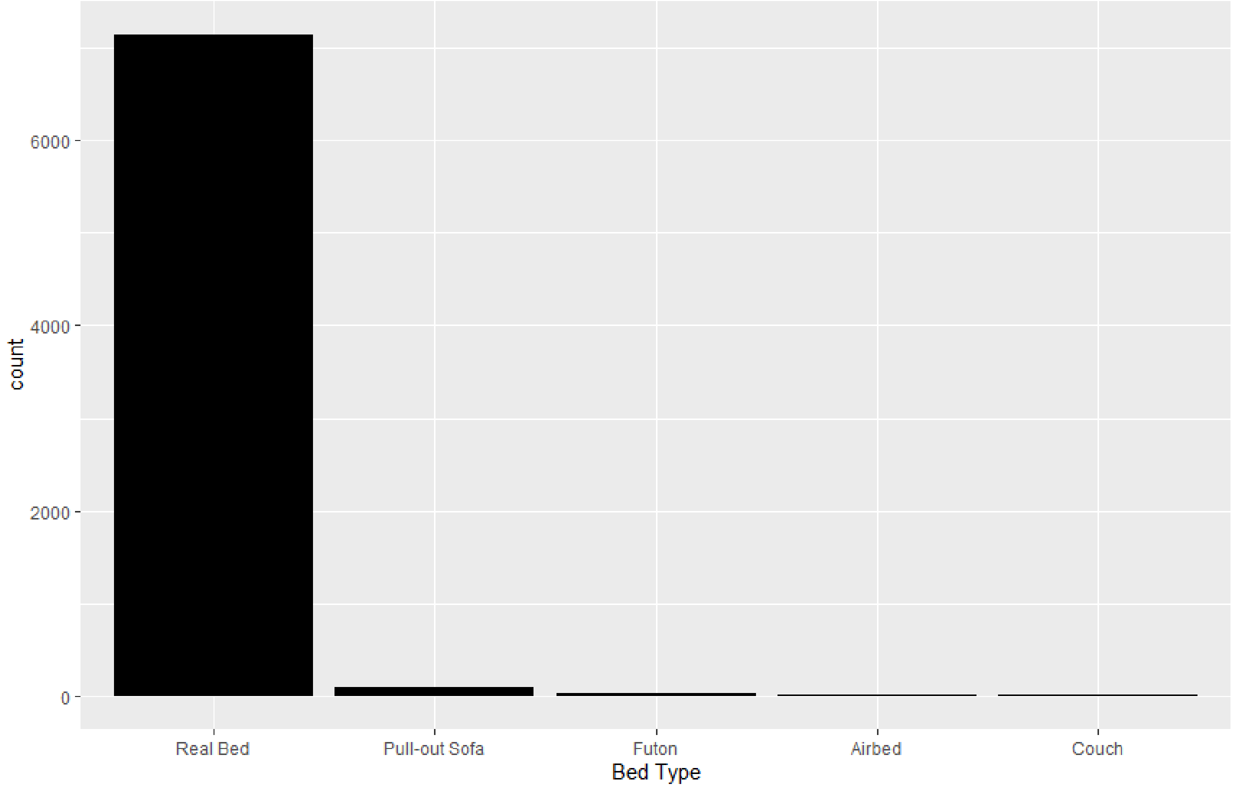

Scatterplot
review_scores_rating
20
40
60
80
100
50
100
150
200
250
price

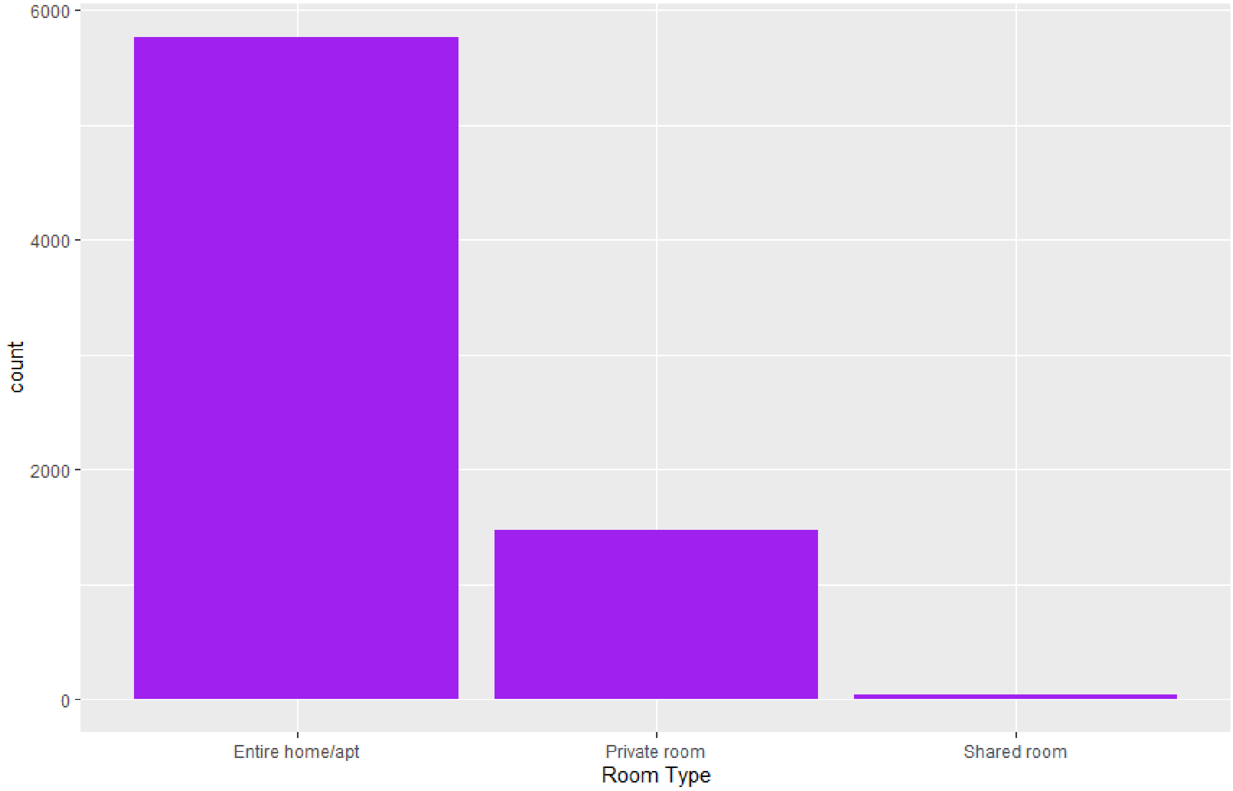

Scatterplot
bedrooms
0
2
4
6
8
10
5
10
15
accommodates


Scatterplot
bedrooms
0
2
4
6
8
10
0
2
4
6
8
bathrooms


Scatterplot
bedrooms
0
2
4
6
8
10
5
10
15
beds

Classification Visuals





Some of the visuals are attached above, if needed more, will be provided.

Looking forward to your feedback.

Thanks