Problem Solving

**(NAME)**

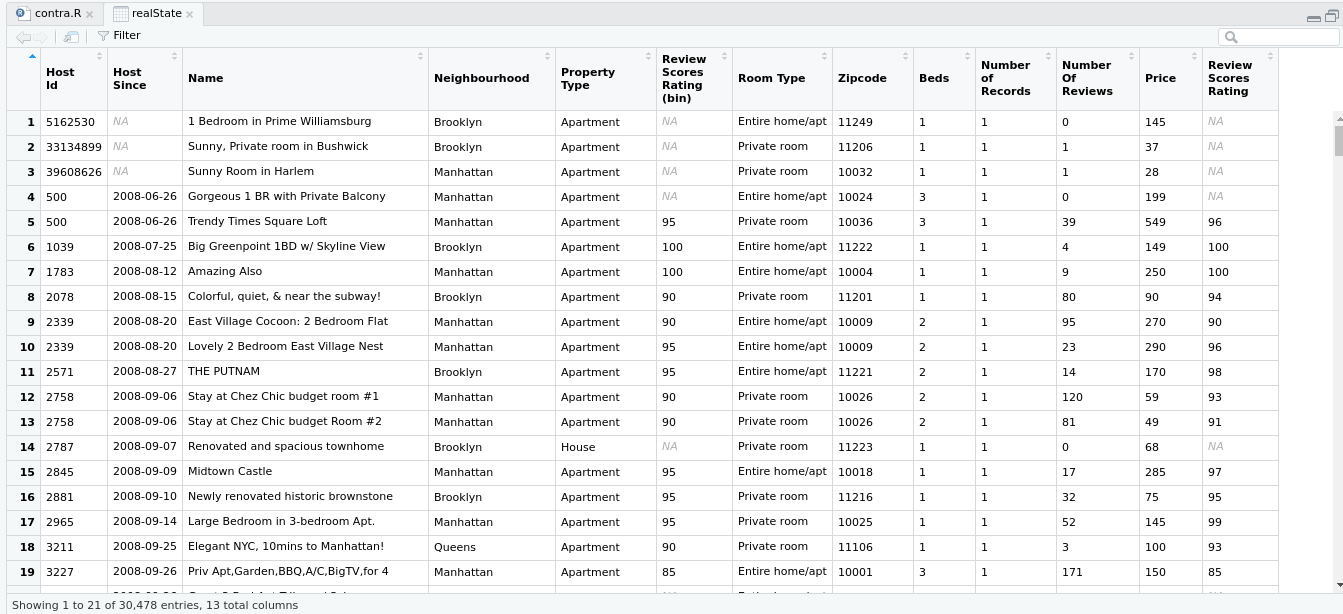
**(REGISTRATION NUMBER)**

# 1.0 Introduction

For the data analysis purpose there are number of popular tools and application that can be used. Considering the performance ratings and user experiences R-Studio is effective and poplar tool among the community of the data science. R Studio contains more attractive and effective features when we compare with the competitors. Considering the programming languages like python and other analyzing languages R programming stand at the top of the performance list because of the incredible performance. Specialty with the R programming is users can simply run the commands line by line and do no have to compile the entire document for the outputs. That saves a lot of time and effort in the programming manner.

Apart from all that R Studio allow users to power up the R programming platform and in that way users can freely manage the way of using the visualization tools and other additional features with a minimum time period. Analyzing features in the R studio tool is very reliable and put the flags on the precise locations in the data sets that are questionable. Along with that R programming and R studio consist of vector operations which can be helpful in a considerably larger operations. Which means users do not have to so many loops and repetitive operations to go through the document. Well supported community and so many packages are to make sure that the users are not left alone in the data science.

I have selected a data set from a real state business in USA. This data set represents all the apartments that have been owned by the business and the customer feedback as well as the description. The data set has imported to the R Studio and rest of the analyzing portions are based on this data set.

  
Image 1: Data Set for the Analyzing

# 2.0 Visualizations of the data-set

## 2.1 Mean

In this data-set we have to look in to the pricing and rating of the Apartments. There are more than 30,000 inputs and the mean of the categories are calculated as follows.

  
Image 2: Mean of the data set

Here I have exercised the mean syntax in R programming in R Studio. Simply it gives the mean of the column that I have selected in the syntax in this case the prices of the apartments. There are limited number of objects that the visualization tools can be applied because some of the columns contain unfilled rows and lines which can leads to a false analytical summary.

## 2.2 Standard Deviation

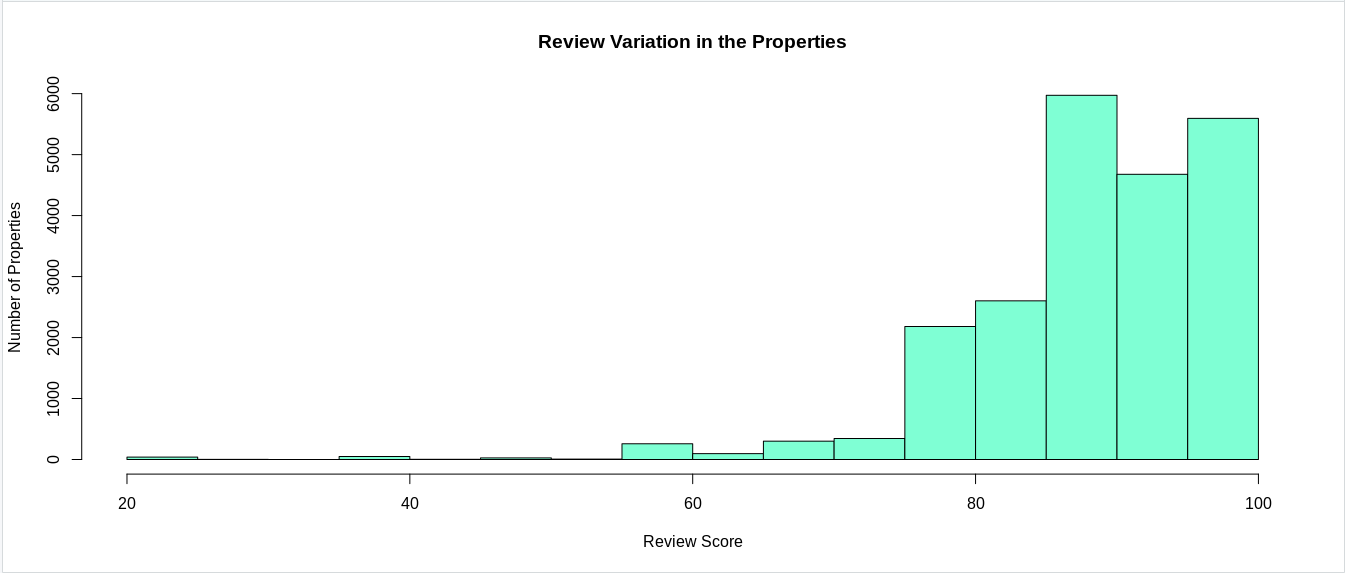
Standard deviation is a comparison measurement that can be used to analysis the behavior of the data set in the future and mostly used in forecasting.

  
Image 3: Standard Deviation

This data set has a standard deviation of 197.7855 in the price section and I used sd syntax for the calculation.

## 2.3 Histogram

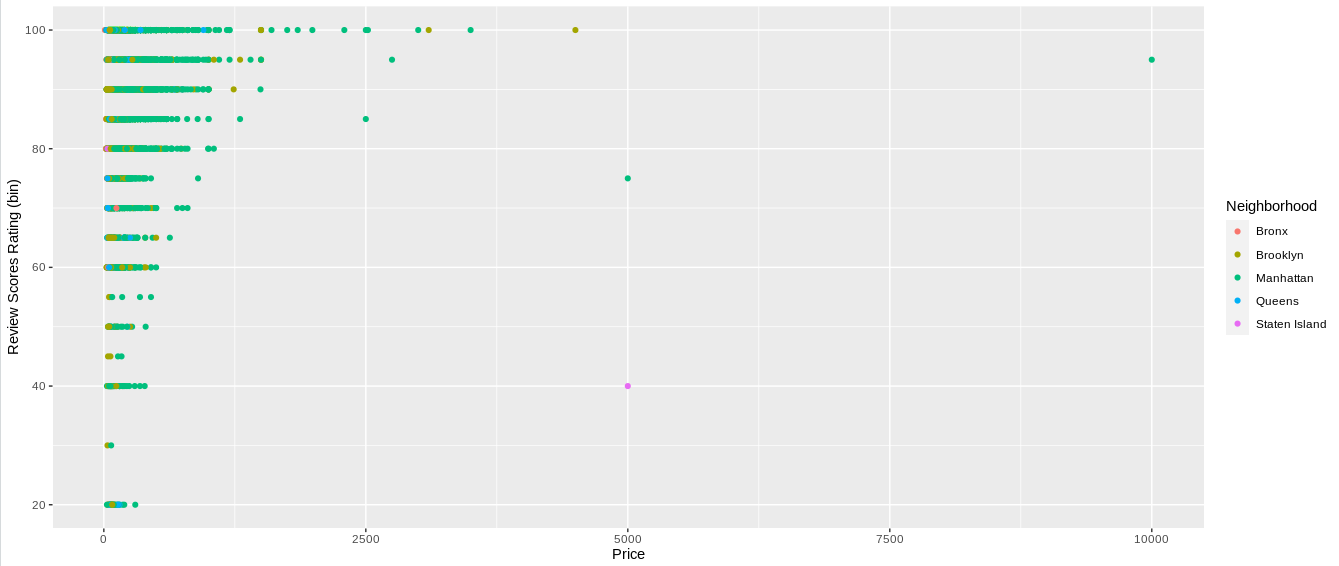
Histogram is a really good method of data visualization and in this scenario we have used several data columns for the representation purpose. Considering the amount of information given in the data set we can see that the variation between the reviews on the properties and the price of the properties are proportional to each other. So that if the business organization could more focus on the customer satisfaction and their feedback on the properties then they can achieve a better outcome in their business.

  
Image 4: Histogram - Review Score Vs Price

As you can see in this representation all the low reviewed places of the company owns are at a low level of the pricing scale. When we explore the data set there are considerably higher amount of customers are in that potion of the customers are in the low reviewed are of properties and because of the low demand on that segment organization cannot charge the higher amount of pay rates on those properties.

## 2.3 Scatter Diagram

Scatter diagram extracted from the R Studio by using the ggplot2 package and the visualizing environment. Same as in the histogram we can see the variation of the data along with the review and rating mechanism.

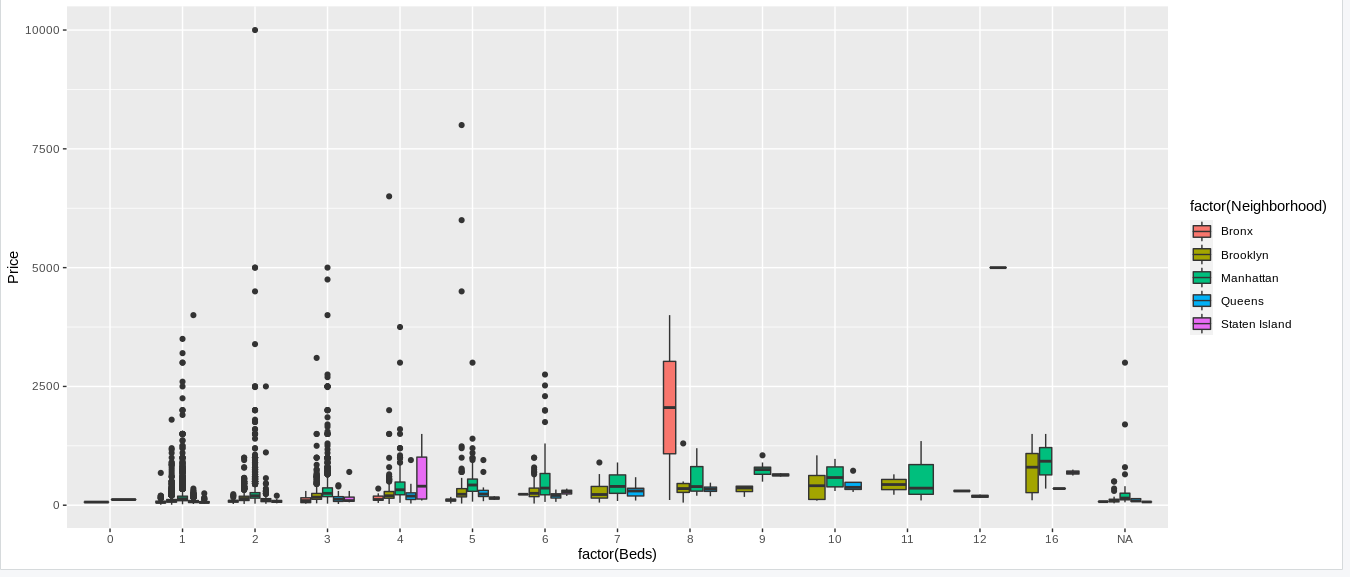
  
Image 5: Scatter Diagram

## 2.4 Box-plot

The box plot is very effective way of visualizing data according to the categorical factors. Mostly in the common and basic data visualizations we use non categorical data which has a controlled flow in the data paths and easy represent in the outputs. But categorical data should br converted as the non categorical data using the factor key word in the R programming.

In this scenario the location of the properties are counted as categorical data and a box plot is used represent them in a understandable manner to the receivers by not using the factoring feature.

In this figure we have obtained the box plot between price variation according to the size of the apartment and also we have considered the location of the apartment. The positive side of this representation is we can have a clear picture on our data set because we have used more than two variables and that gives a wider angle on the data set. At the same time there are some significant amount of outliers in the data set and using the box plot we can understand them in a more effective way.

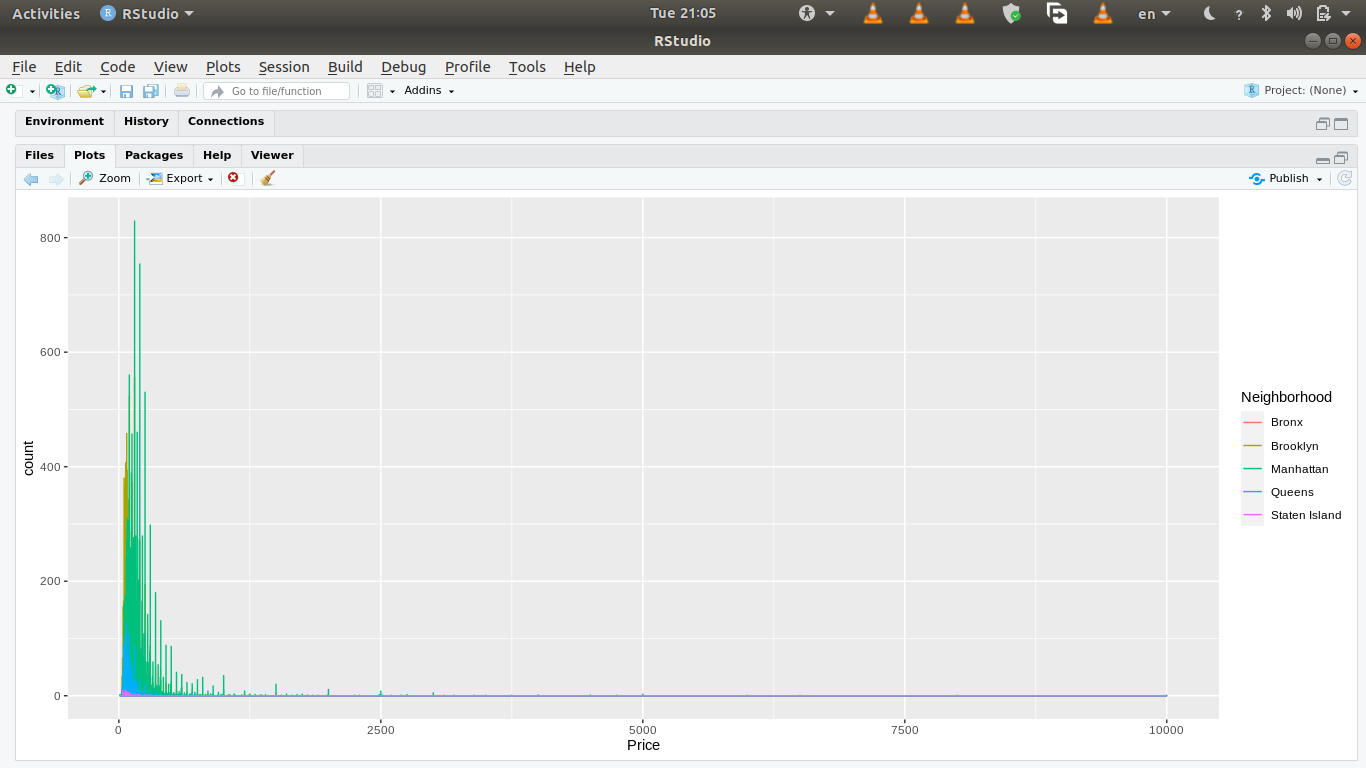
  
Image 6: Box Plot - Price Vs No of Beds Vs Location

Considering the outliers we can determine the customer choice according to the respective city. So that we can see when the size of the apartment is smaller than the average then the price variation is going higher in urban arias.

## 2.5 Polygon

Polygon is created using the histogram itself in a more delicate manner. In this case the it is created considering the amount of money people have to pay for a apartment. To make this information more useful the locations of the apartments has been added to the polygon.

As we can see in the figure there are very high amount of demand for the apartments which are in the price range of the first 500 units. In that segment most of the places are located in the Manhattan Island or Queens.

  
Image 7: Polygon - Price Vs Count and the Location

# 3.0 Discussion

## 3.1 Problem Identification