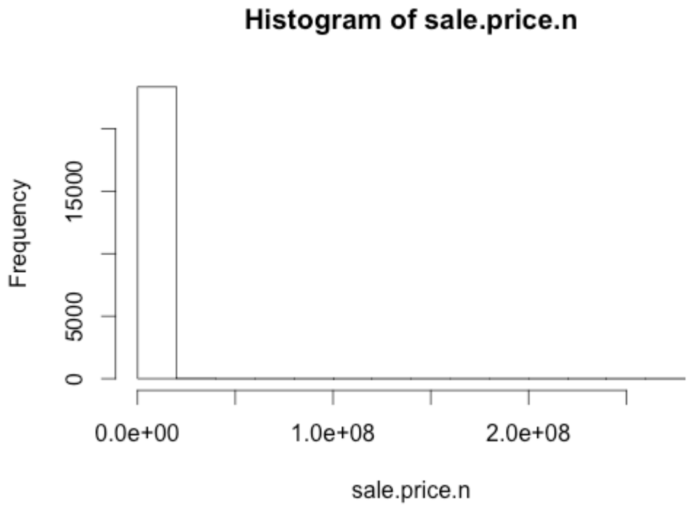
Problem 3
Part A

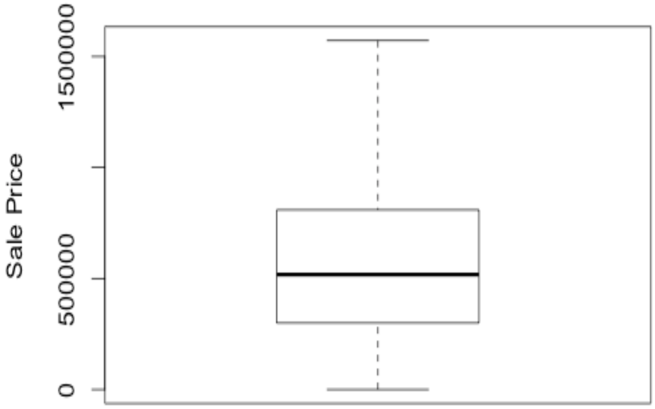
Problem Description:
In this part we analyze the data for the Brooklyn area and try to visualize the patterns in the data. Data is first loaded into R and then cleaned to remove extra formatting which can cause inconsistencies. We then create a new data frame containing only houses whose sale price is available and valid. We also create another data frame based on the criteria that they are “Family” homes.

Analysis of Plots:

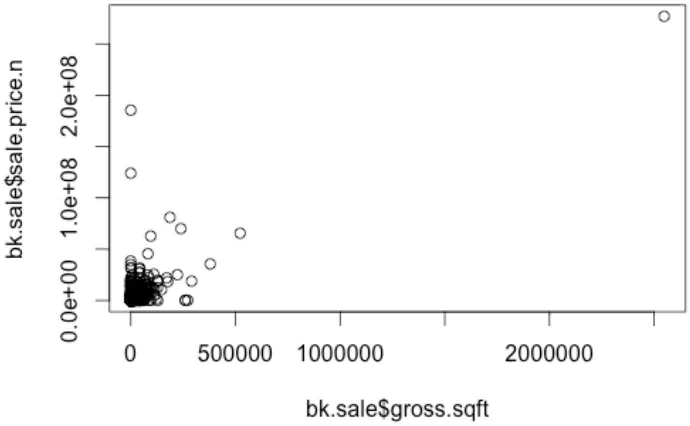
1. The first plot below is used to plot the sale prices of the properties and their frequencies but since most houses don’t have their sales prices listed we do not get a good plot.



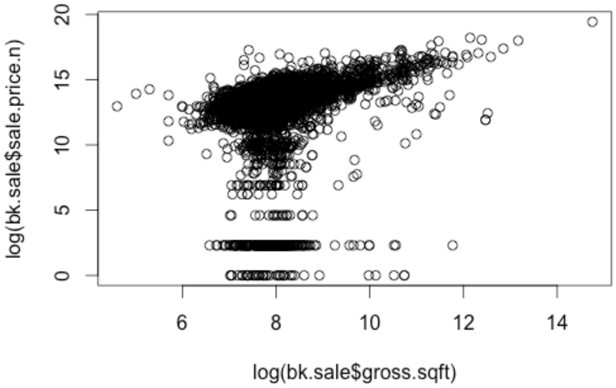
2. The next plot below is a boxplot of the sale prices with much more clean data set (outliers have been removed). We can see that the average sale price is close to $500,000.



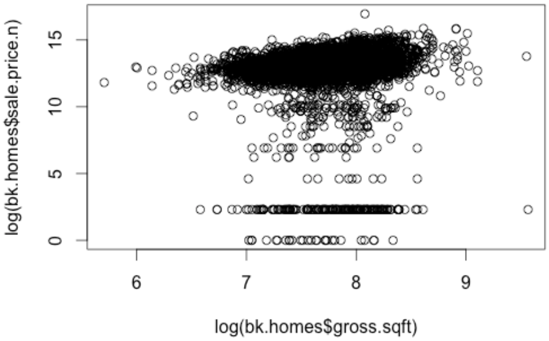
3. The next plot below visualizes the gross square feet versus the sale price. Due to the outliers present in the data we don’t get a good plot.



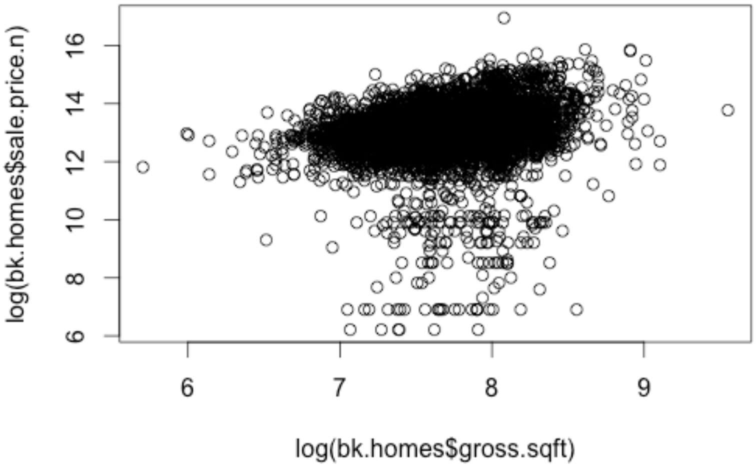
4. To fix the problems of having outliers in the data we use the log function to get a better plot. We can see that there is almost a linear relationship between sale price and gross square feet.



5. The plot below visualizes the gross square feet versus sale price of only “Family” homes. There is only a slight increase in sale price with increase in area.



6. The final plot visualizes the same data as above without outliers.



Part B

Problem Description:

In the second part of the problem we read in all the datasets for different boroughs in New York City and try to visualize the data patterns. Similar to the above problem we clean the data and create different data sets for properties with sale prices. We also divide the data into one, two and three bedroom houses and try to analyze them individually.

Answers to Questions on Page 48:

1. Answers to the following questions would help us understand the market better and make much more informed decisions:
   a. *How do the prices of properties compare in the different boroughs?*
   b. *What is the average cost of a property in the various neighborhoods of New York?*
   c. *Where are the most number of properties on sale?*
   d. *How many bedrooms does the property have? What is the difference between the costs of one, two and three bedroom houses?*
   e. *Where can you find the cheapest one bedroom house in New York?*
   f. *During what period of the year are most houses sold?*
   g. *Where in New York are most apartments sold?*
   h. *Which are is the biggest market in New York?*

2. Once the data is loaded into R, it is cleaned by removing formatting in numbers such as prices, areas, dates, etc. All the properties with valid price tags are placed into a separate data frame for further analysis. Price per square feet is calculated for each property with a correct value in the gross square feet column.

   The plots below visualize the data according to neighborhoods and over time.

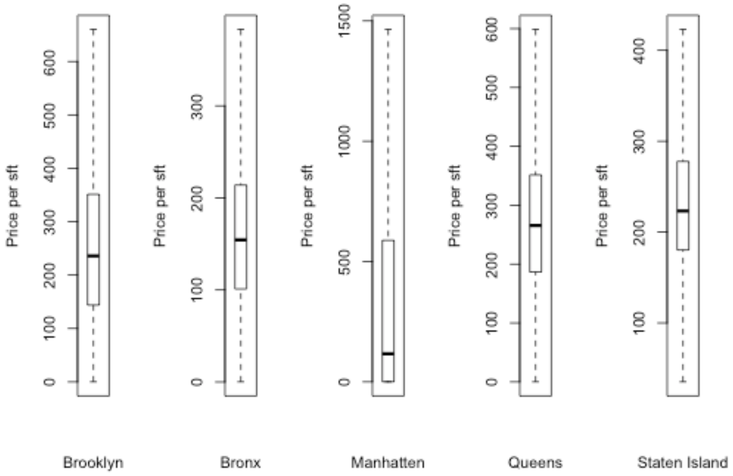


Figure 1: Cost Comparison

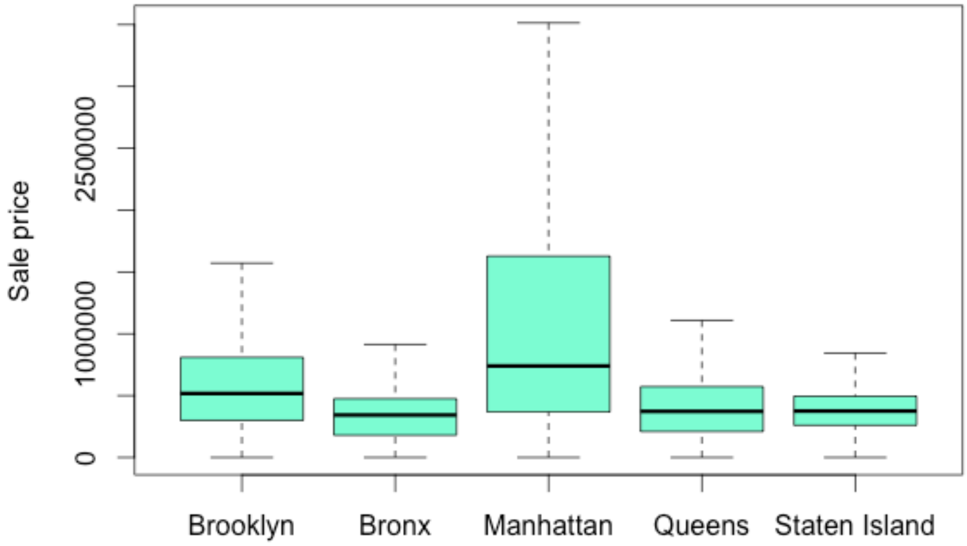


Figure 2: Actual Sale Price Value

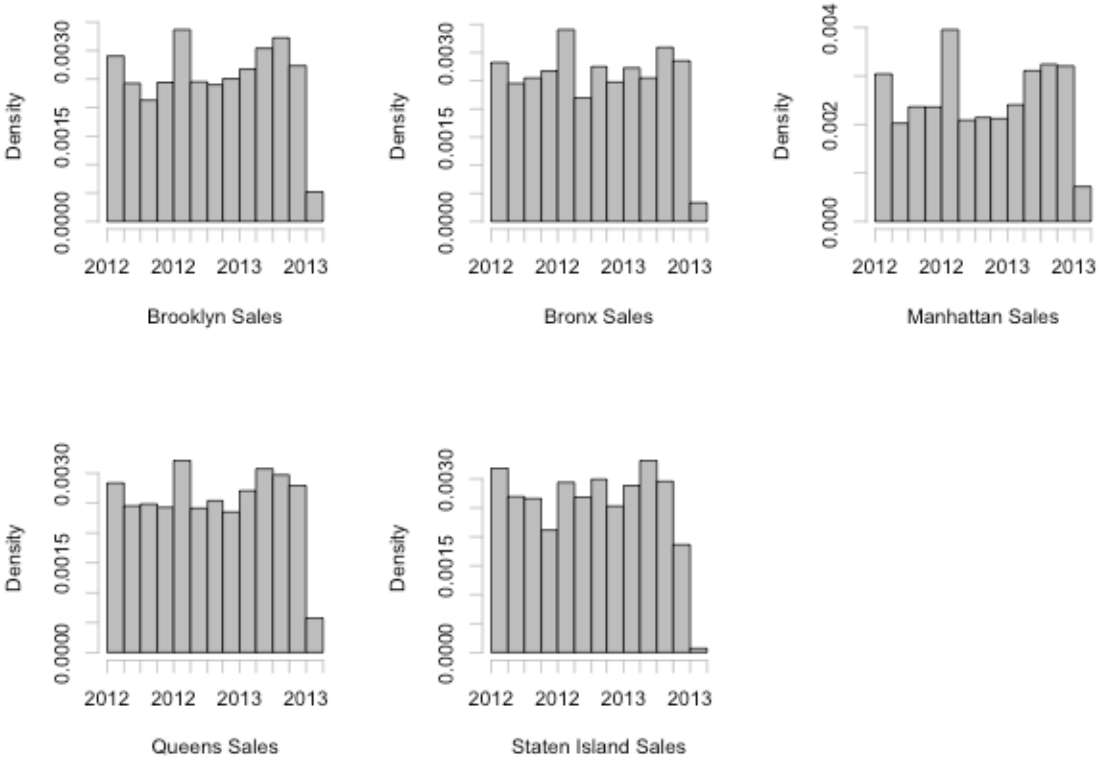


Figure 3: Monthly Sales Figures (Histogram)

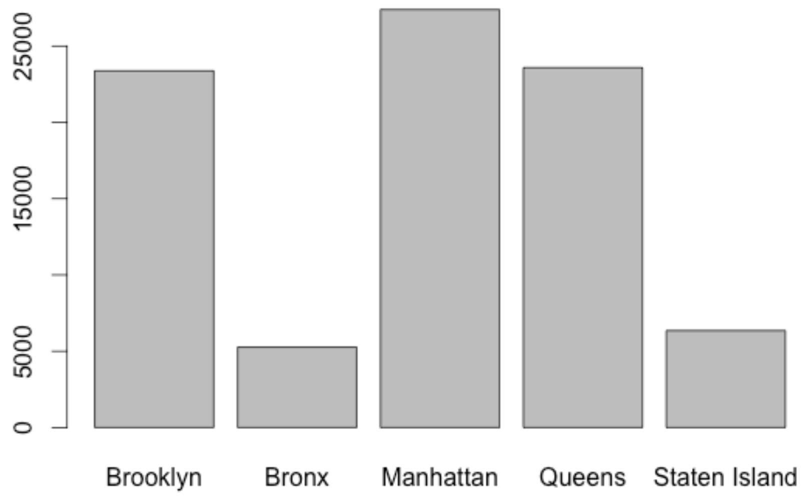


Figure 4: Number of Listings

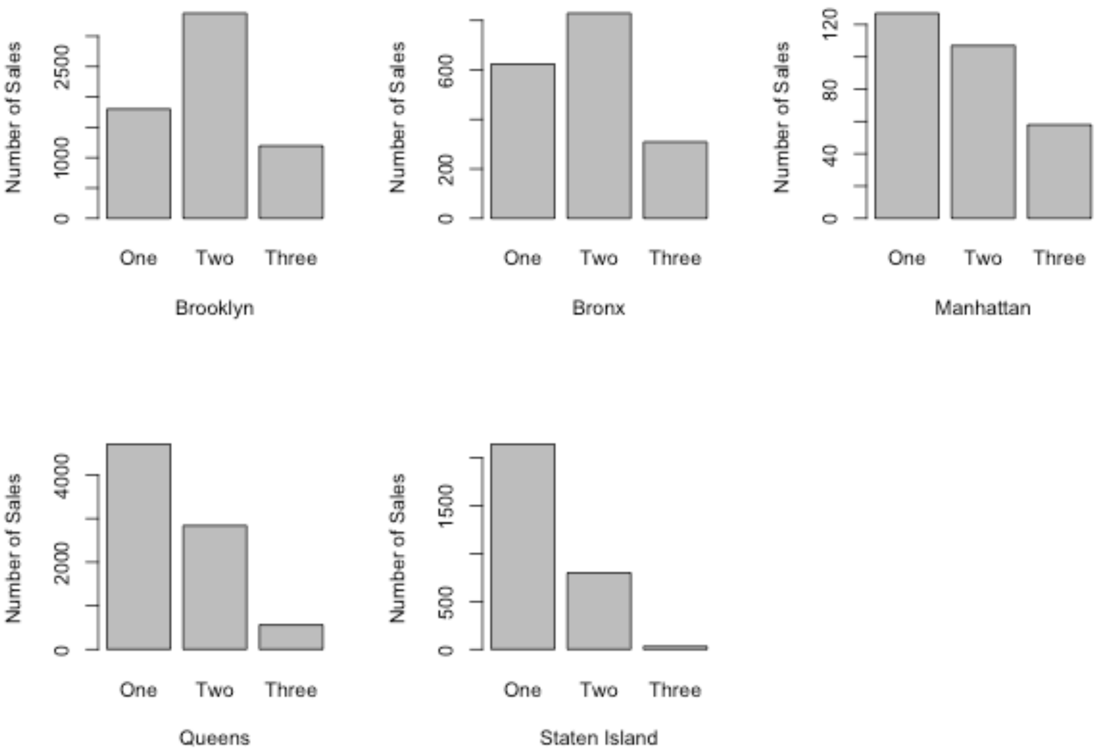


Figure 5: Number of One, Two and Three Bedroom House Listings

3. Summary:

From the above data the following points can be noted:

- From Figure 4, we can see that most number of listings is for Manhattan, Brooklyn and Queens.
- Customers who require three bedroom houses should check the Brooklyn and Bronx areas. There are more single bedroom houses in Manhattan than two and three. (Figure 5)
- Monthly Sales are highest during December for all the boroughs. (Figure 3)
- The cheapest houses can be found in Bronx and the most expensive in Manhattan. (Figure 2)

4. To get more information regarding property sales it would be helpful if you could get information from someone already working in the real estate domain. By talking to some one who specializes in this domain it, it would be easier to find out what kind of data the consumer expects. Also, it is important to maintain constant communication with the CEO to figure out what kind of data is required by the company to improve the sales and how to market the product.

5, 6. To develop a data strategy for an online business it is important to note the following points:

a. Accurate Data is required to make correct analysis.

b. Complete data is required to ensure there are no gaps in the data set.

c. Data samples should be random to get a clear picture.