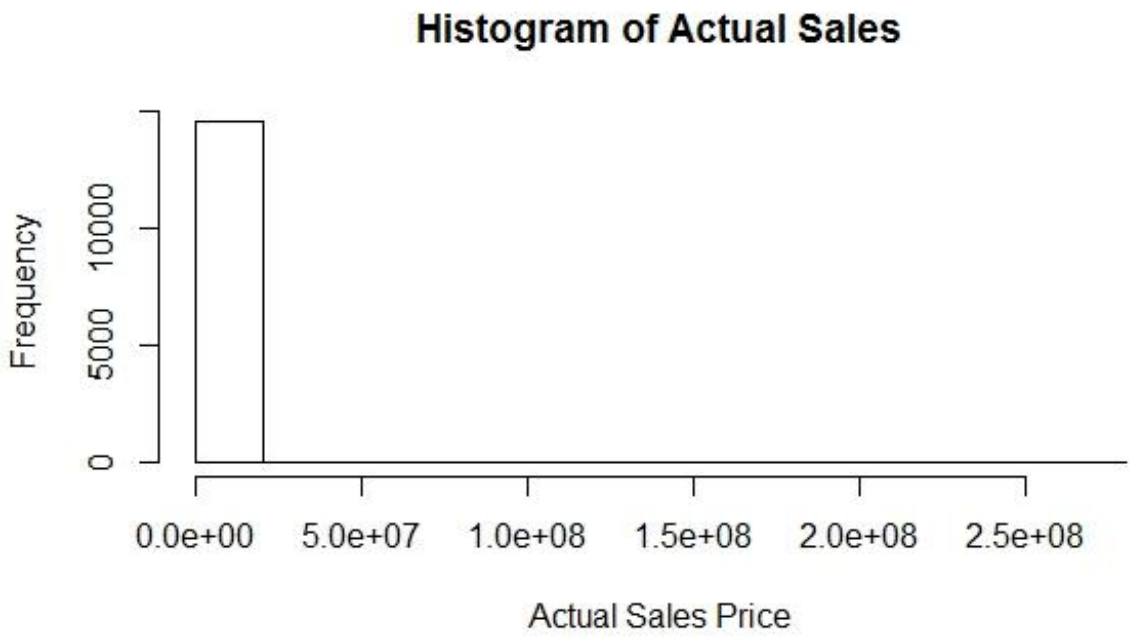
REPORT

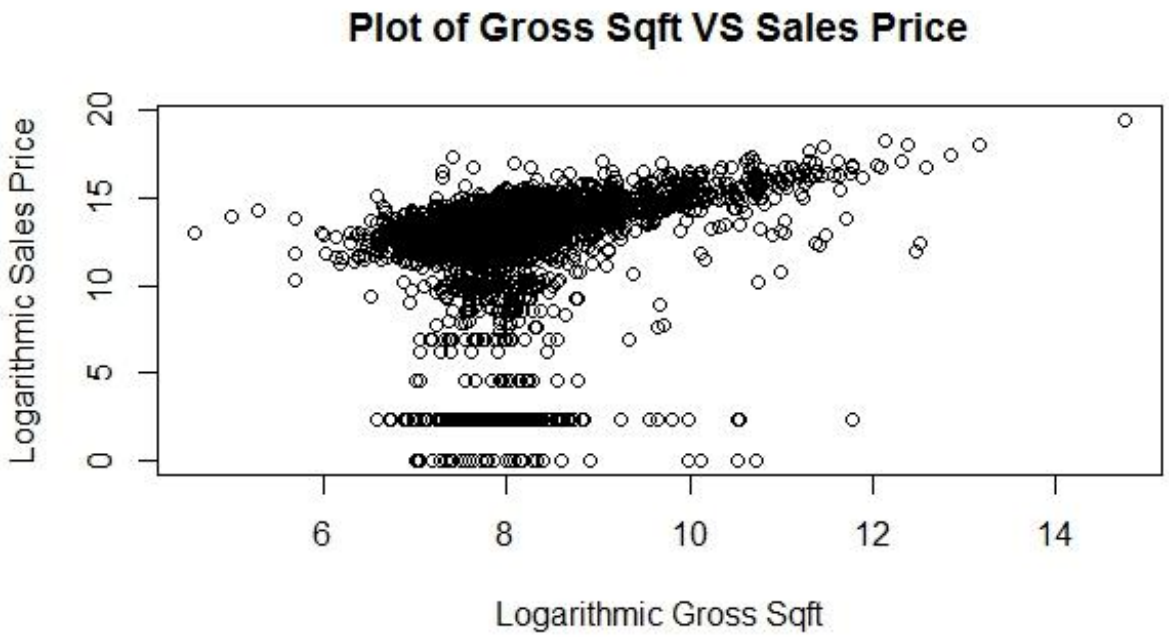
Prob 3: EDA on Brooklyn borough data and extending the analysis to the all the boroughs.

Part 1: Performing EDA on the Brooklyn borough as guided by the textbook.

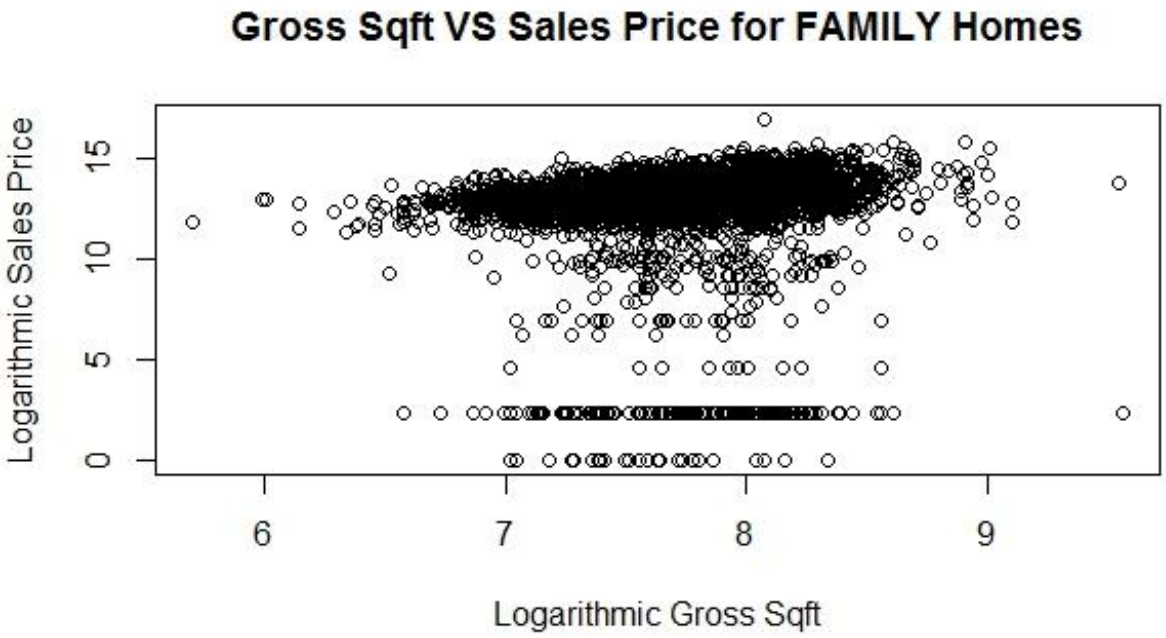
- The data is read from the csv file using **read.csv()** method which converts the object to dataframe in R
- Then I performed some cleaning and formatting such as making the price and square feet columns as numeric and selling date as Date in R.
- Then plot the histogram of sales price for all the neibourhoods in Brooklyn. The care taken here is data with sales price is removed giving actual sales. The plot is as below:

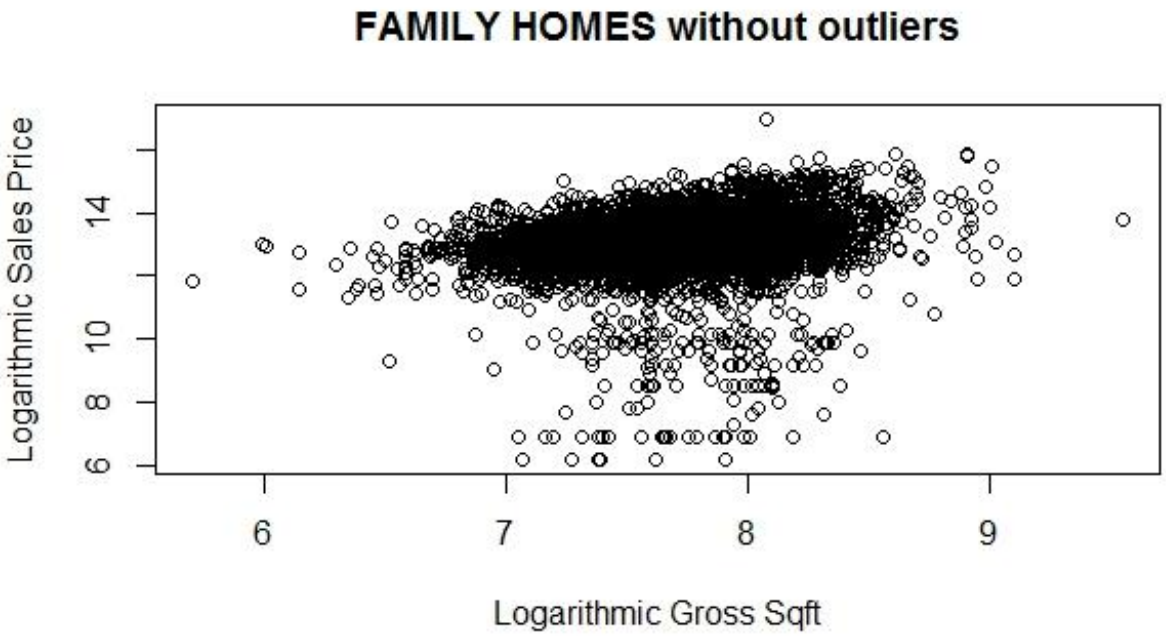


- Next I plot the graph of gross square feet area against the sales price for corresponding properties.



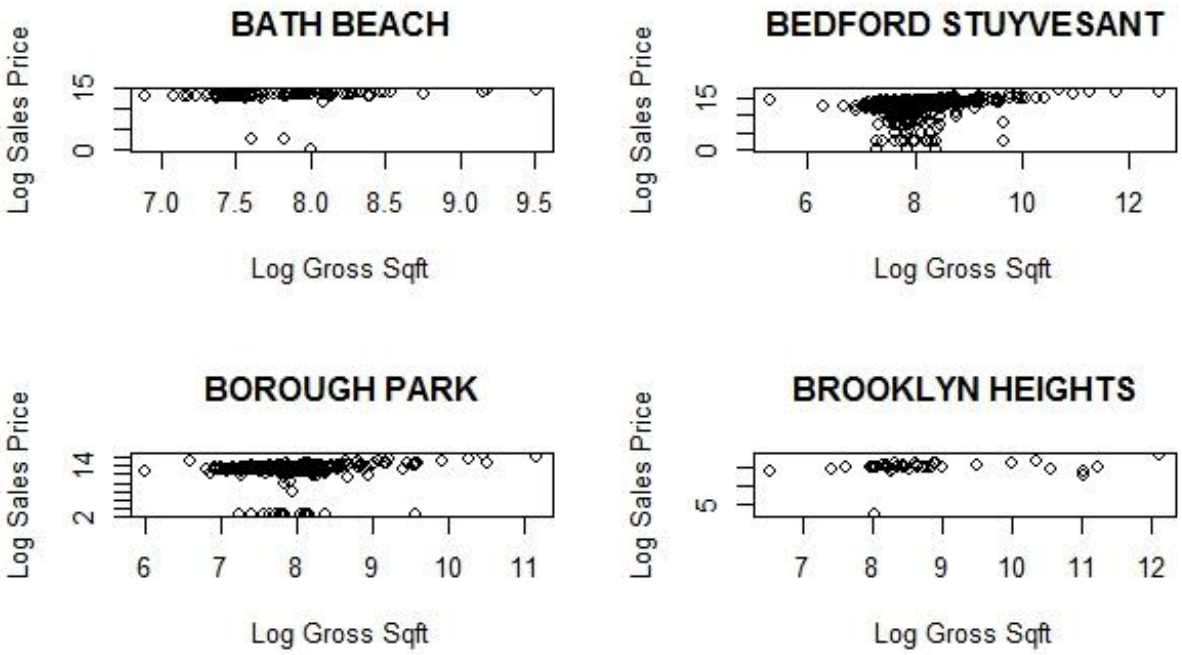
  - Inference: The logarithmic graph shows the value is crouded in area around 7 to 8 for Sqft and 10 to 15 for the price. That means these are the range the company can expect have most sales or user can expect to buy.
- Next, I plot the square feet against sales price for the Family homes and then plot by taking out the outliers. This I achieved using the grep command on the dataframe fields.





- Inference: For family homes in Brooklyn you may get bigger area houses. That means Family homes were sold cheaper as compared to the total sales in Brooklyn.

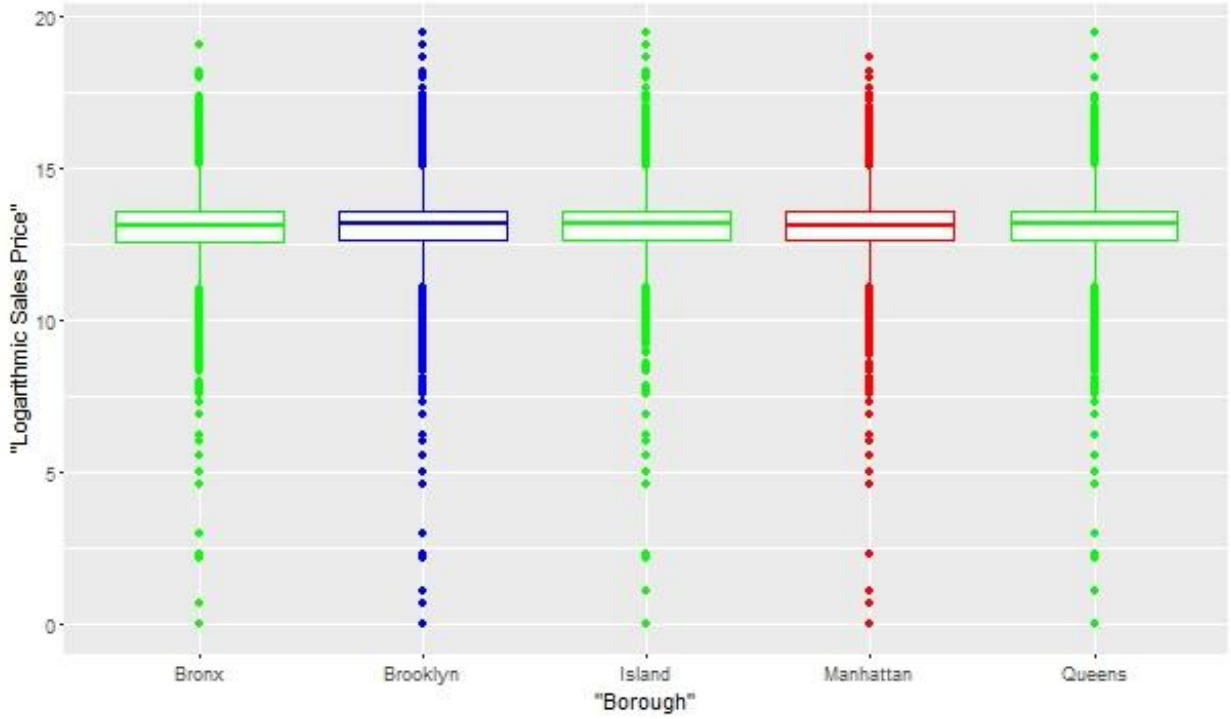
- Next, I compared the sales price against the gross square feet area for some sample neibourhoods in Brooklyn and combined the plots for them. The plot is as below:



  - Inference: The sales in Brooklyn heights were very less as compared to other neibourhoods as so the people are not interested in selling or buying properties in this area. The sales price in BATH BEACH shows that this neibourhood is very costly.

Part 2: Extending the EDA to all the boroughs.

- First, I compared the sales price in all the boroughs. The plot was taken as below:



  - Inference: This possibly shows that sales price is almost same in all the boroughs. But this can be due to less sales in costlier cities and more sales in cheaper cities. So I decided to calculate the price per sqft for each city and plot them.

- So next I calculated sales price per sqft for each sale in each city. For this I added separate column to the dataframe and then tried to plot the Sales price per sqft for each city.

  - Inference: The above plot shows that the sale in manhattan was costliest of all when taken per square feet. The Brooklyn also is cosltlier city to live in while other mentioned cities are cheaper.

REFERENCES:

- Textbook explanations
- Google search for information and examples (R-bloggers website)