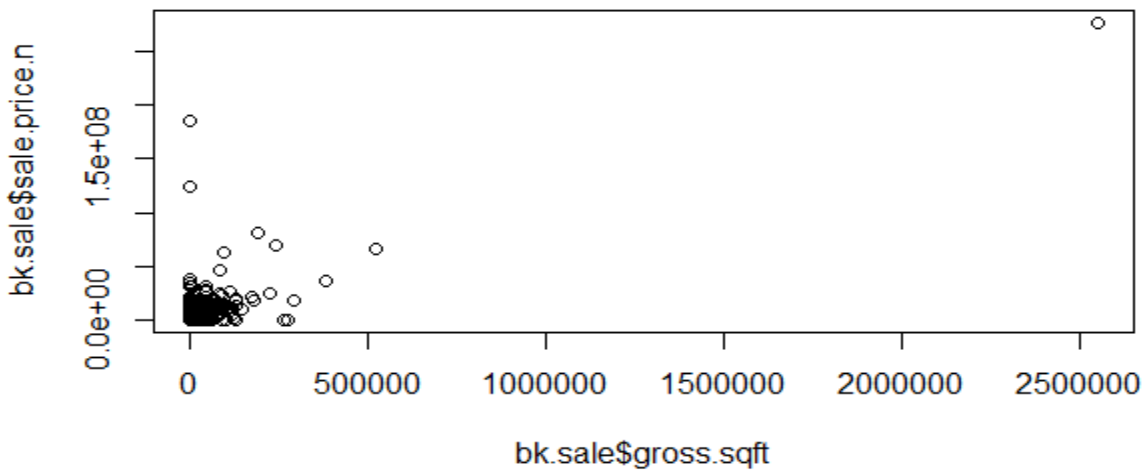
Problem 3: Data economy: A *real* case study

Problem Statement: For this problem you will work the RealDirect problem discussed in Ch.2. Work on the sample code discussed in p.49,50. Then extend it to the entire RealDirect data set to perform EDA to find some more insights. **You will realize the data given as well as the code require some cleaning and editing.** This particular problem is a "data problem" from a company is that is currently operational, see http://www.realdirect.com/ . Read the details and understand the **business model**. The sample code is given in pages 49-50 for Brooklyn borough. Work on it and make sure understand the EDA process. Then answer the questions in pages 48-49 in your report. Repeat the analysis for different boroughs. Save the scripts, plots, **matrix of plots to compare** various boroughs, and data environments (for a possible TA demo). (a) Name the R script RD3Username.R. (b) Now extend the EDA to monthly data and work on the questions given in page 48-49. Save the script as RDP3ExUsername.R. (c) For both problems save the charts and interpret the outcomes (charts): write a paragraph to interpret the results. Save this in a single pdf document RDP3Username.pdf. zip and tar and submit RDP3Username.tar

By answering the questions in the textbook the R scripts for all the rollingsales areas was written. The data was cleaned and exploratory data analysis was carried out. The followings plots were used to summarize the findings for a brief report aimed at the CEO.

The following graphs or plots are for the all the sub-urban areas.



This graph shows the Number of houses for a particular price and a particular square feet for Brooklyn area. It can be observed that there are more houses with area less than 500000 Sqft and less price.

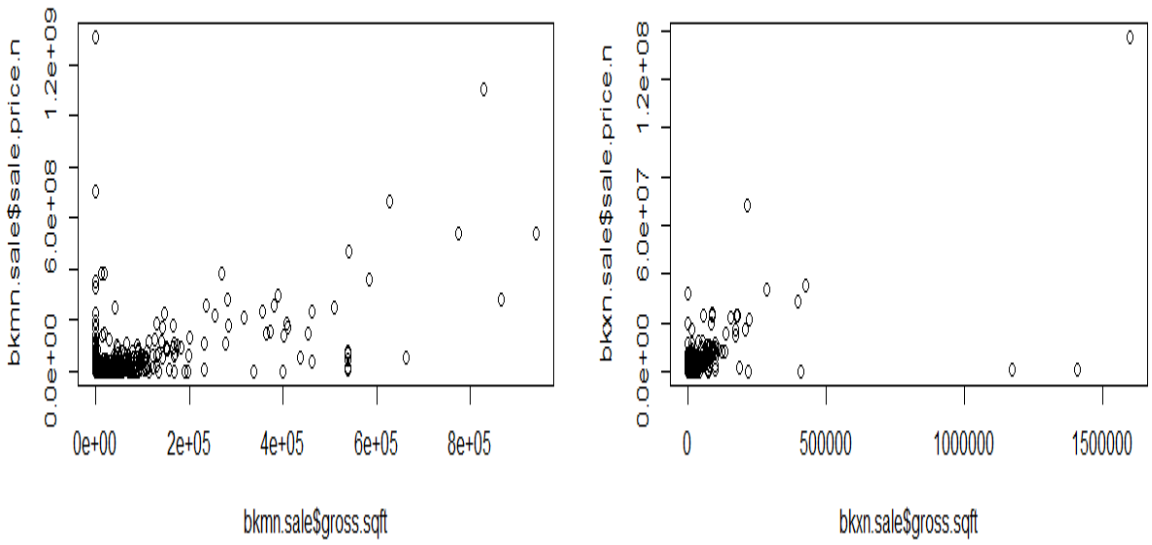
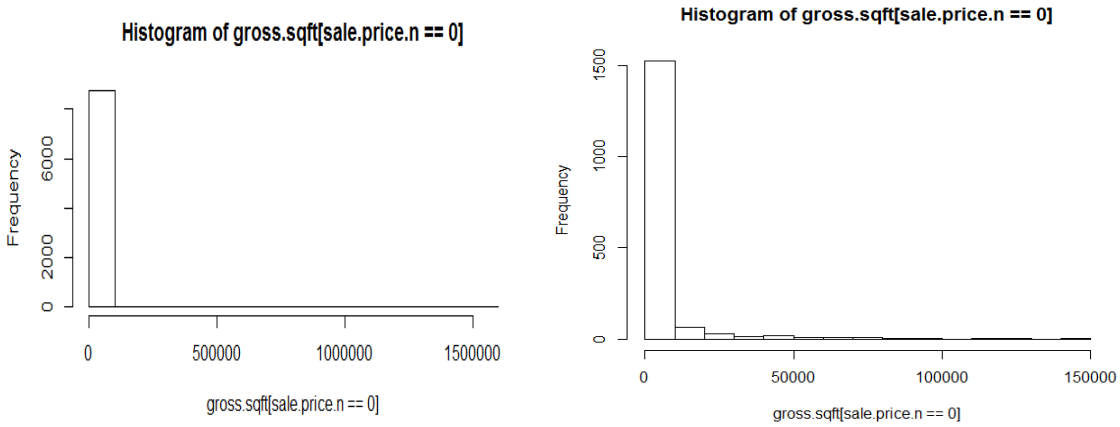


Fig: Manhattan and Bronx

The previous same plot was applied for other areas too. The above 2 plots are for Manhatten and Bronx areas. These 2 graphs also show that more number of houses with area less than 500000 Sqft are sols.



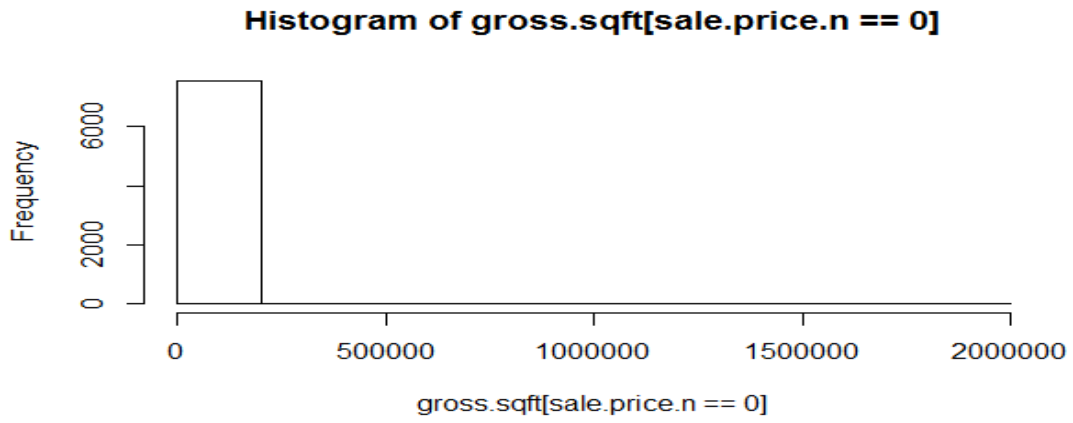
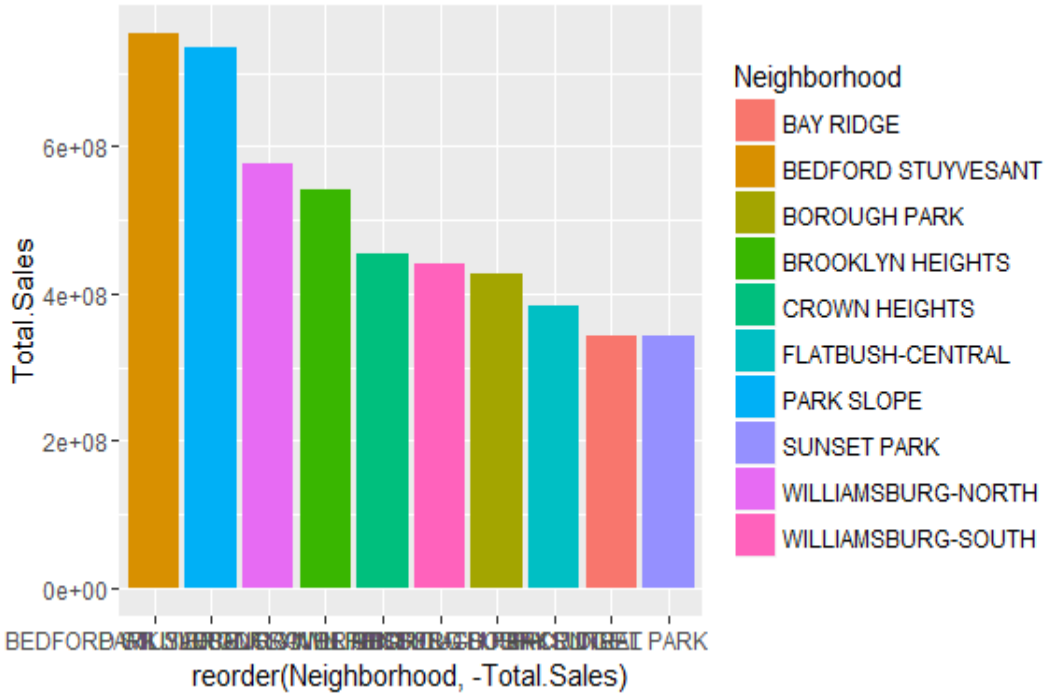


Fig: Histogram for Brooklyn Bronx, Manhattan

The above histograms gives the frequency of houses for sale for different square feet. It again can be observed that there are more houses under 500000 Sqft are sold.



The above graph shows the number of sales in different Neighborhoods of Brooklyn area. The same was applied for all the other 4 Sub-urban areas of New York and the following graphs were obtained.

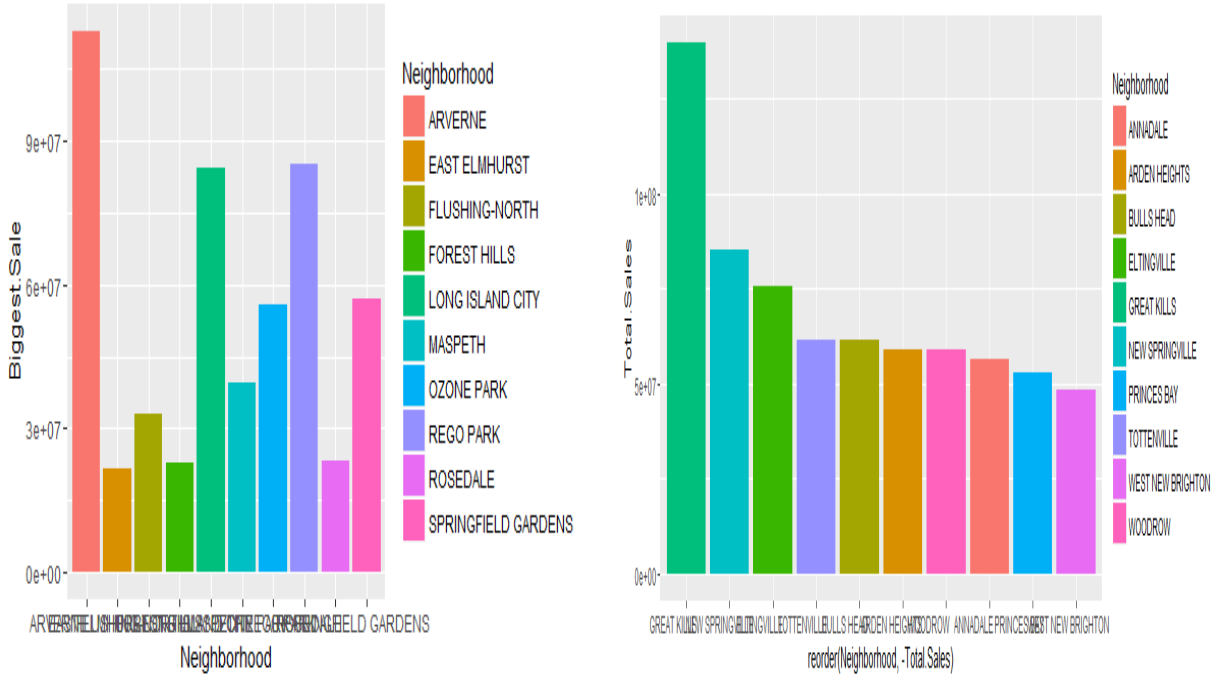


Fig: Queens and Statenisland

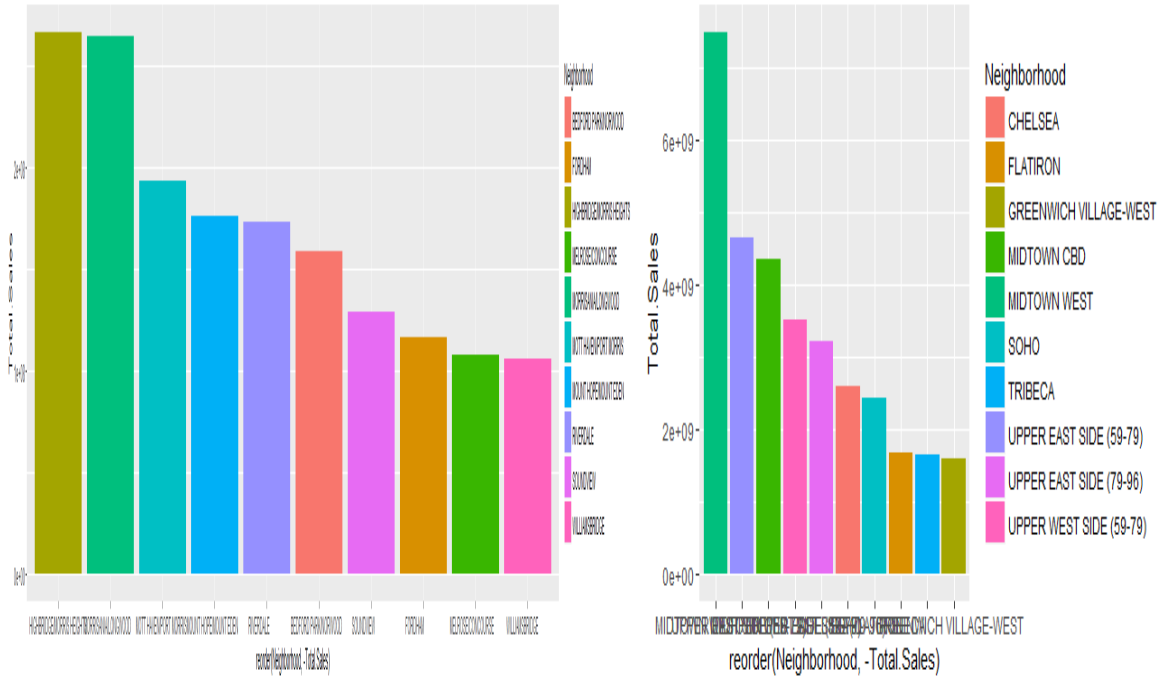


Fig: Bronx and Manhattan

The next set of graphs for different areas are the logarithmic representation of sales of houses based on their price and Squrefeet.

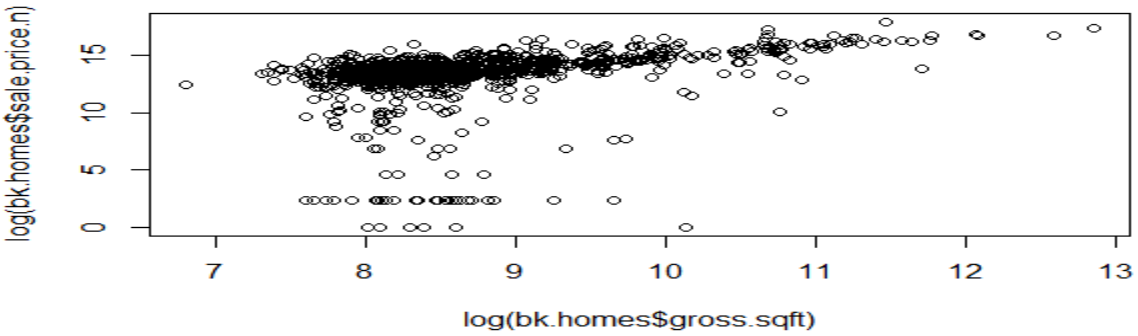


Fig: Broklyn

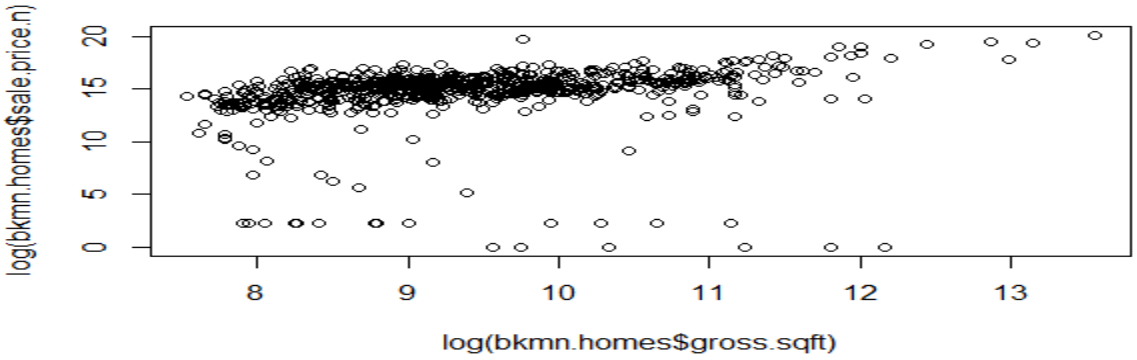


Fig: Manhattan

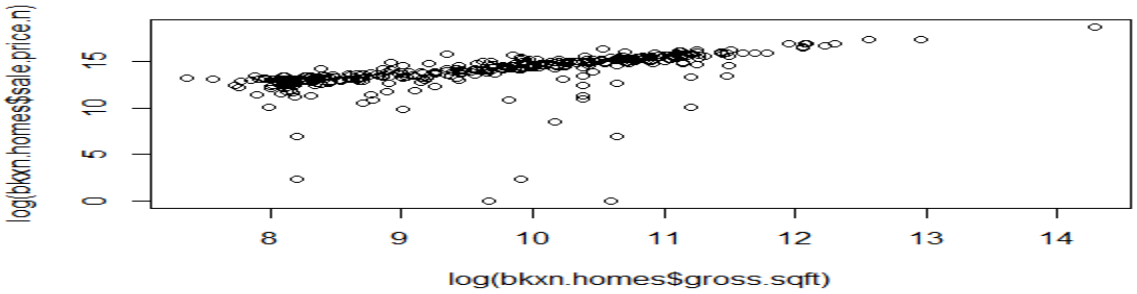


Fig: Bronx

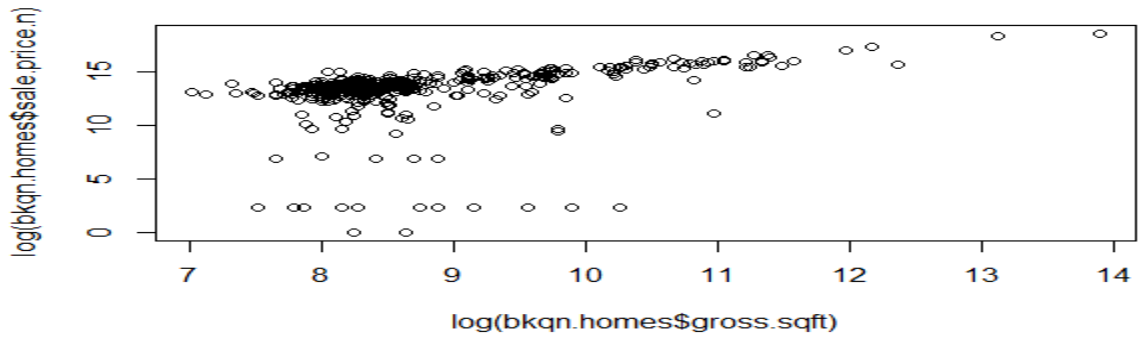


Fig: Queens

From all the above graphs it can be observed that the more number of houses are sold between the price log 10 to 15 are and area log 8 to 10 Sqft.

Next we have plots based on the biggest sales in various Neighbourhoods in different areas. These plots can help in observing which neighbourhood in more desired for living, houses in which neighbourhood cost more as compaired to others etc..

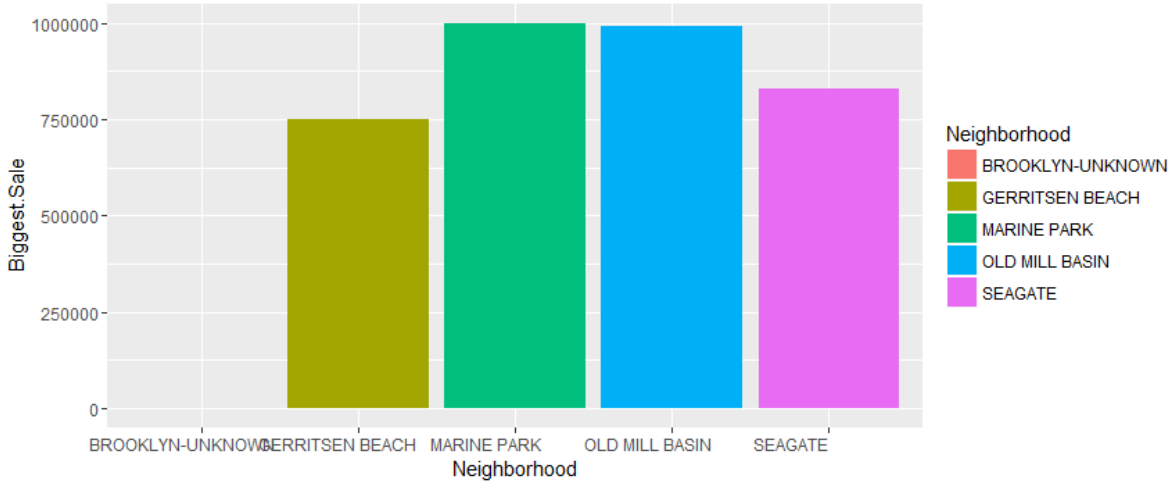


Fig: Sales in Neighbourhoods of Broklyn area based on Sqft.

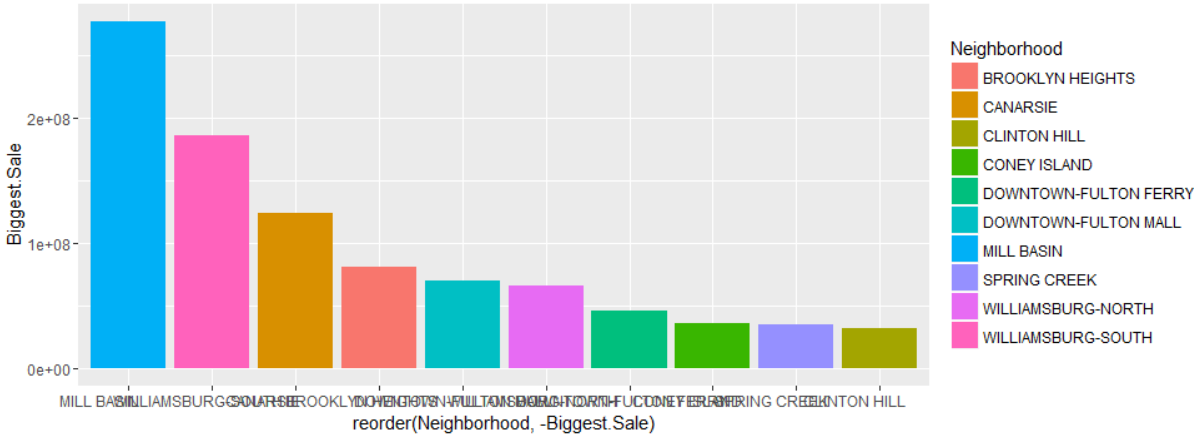


Fig: Sales in Neighbourhoods of Broklyn area based on Price.

The same was applied for all the different Sub-urban areas neighnourhood.

Next type of plot is for the number of sales for the different months of the year 2012 and 2013 in Broykln. The same was applied for all the other Sub-urban areas.

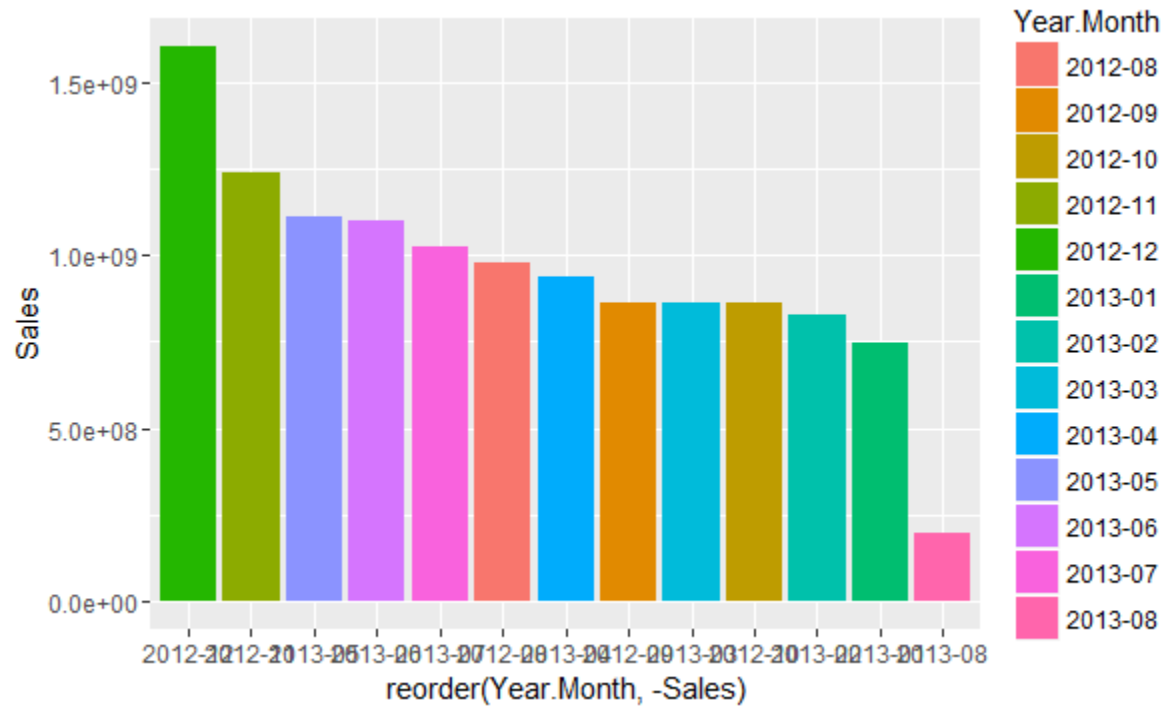


Fig: Broklyn

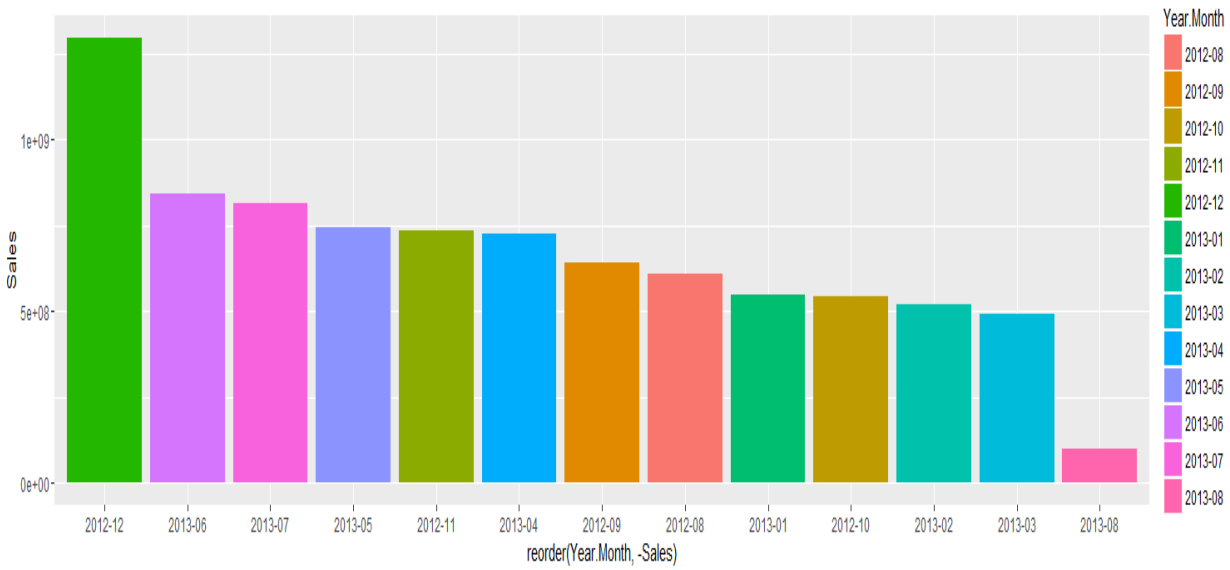


Fig: Queens

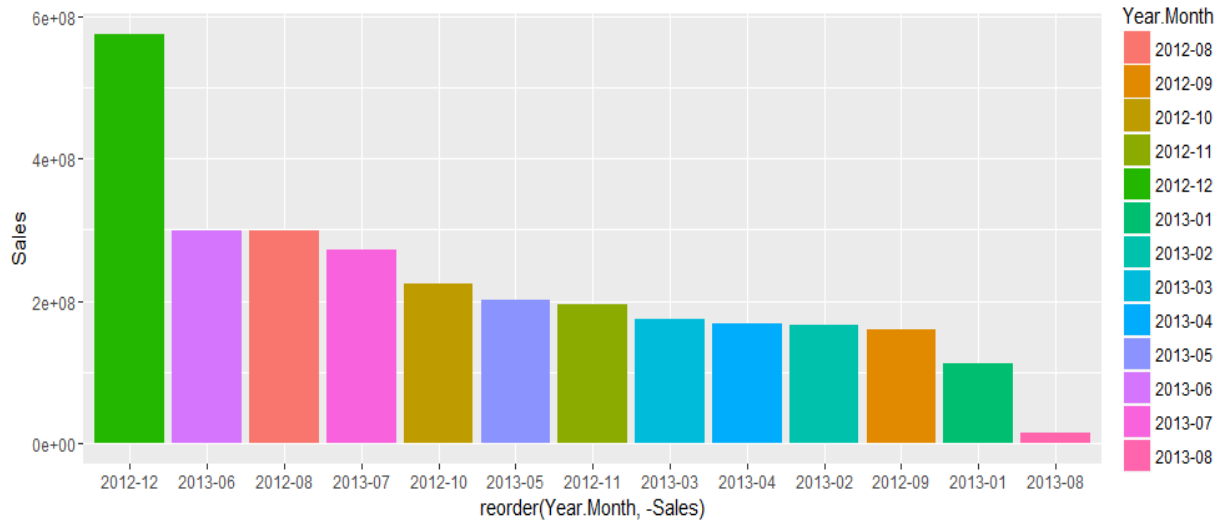


Fig: Bronx

From the above graphs its clear that more number of houses are bought during the holiday season December.

From all the above types of graphs and plots a lot of inference can be obtained like the price range of houses that sell more and the Area and neighbourhood in which more houses are sold and also the month when the sales are more. This can help the CEO to make decisions on the rate, area and many more factors of the houses.