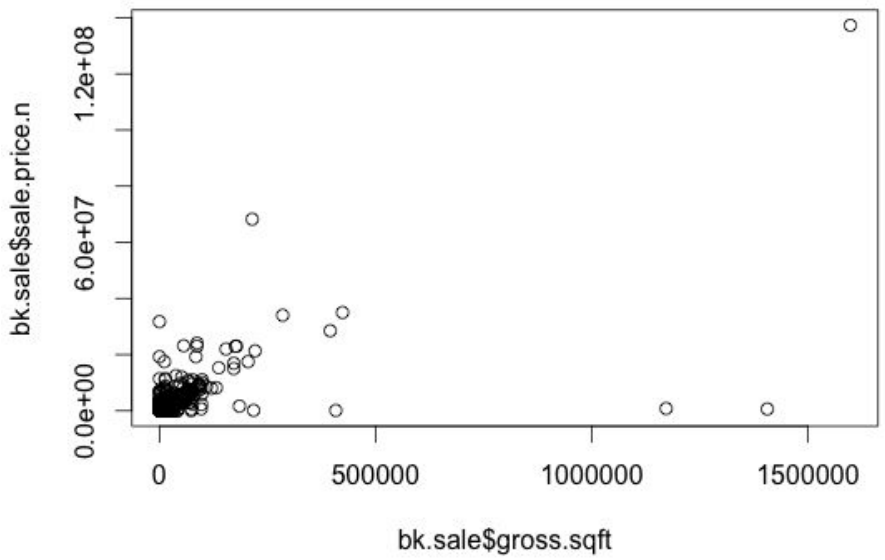
Problem 3: Data Economy: A real case study

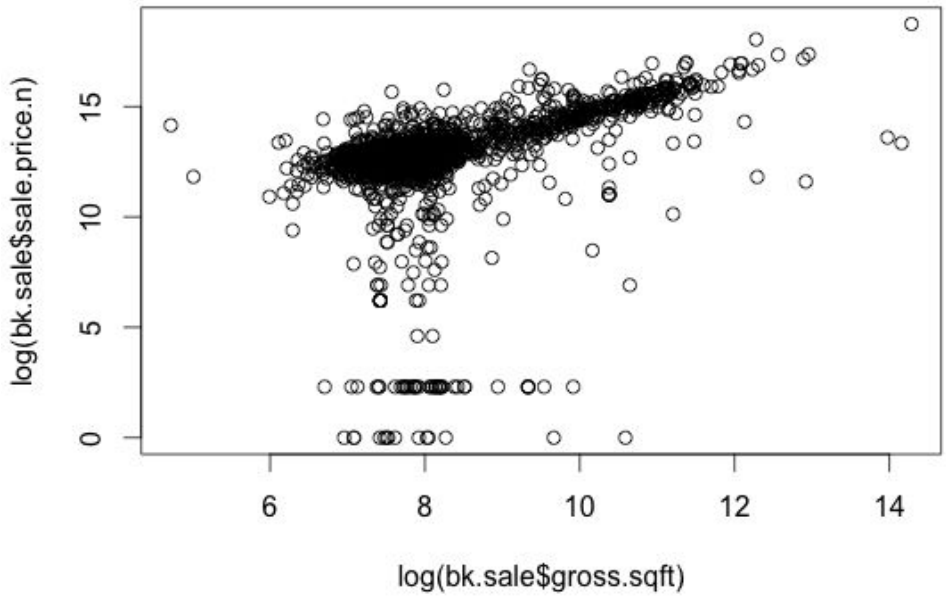
In this problem we are analyzing the data collected by a currently operational company who acts as an online middleman in the realtor business. The data is about the properties sold in a period of one year. First we analyze the data of Bronx area and then extend our analysis to more areas of New York like Manhattan, Brooklyn, Queens and Staten Island. The columns data contain are borough, neighborhood, building class category, gross square feet area, sale price, year built etc. For the analysis data needed to be cleaned first for example representing the comma separated sale price as numeric, correcting the date format, etc. by:

```
names(bk) <- tolower(names(bk))
bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", bk$gross.square.feet))
bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","", bk$land.square.feet))
bk$sale.date <- as.Date(bk$sale.date, "%m/%d/%y")
bk$year.built <- as.numeric(as.character(bk$year.built))
```

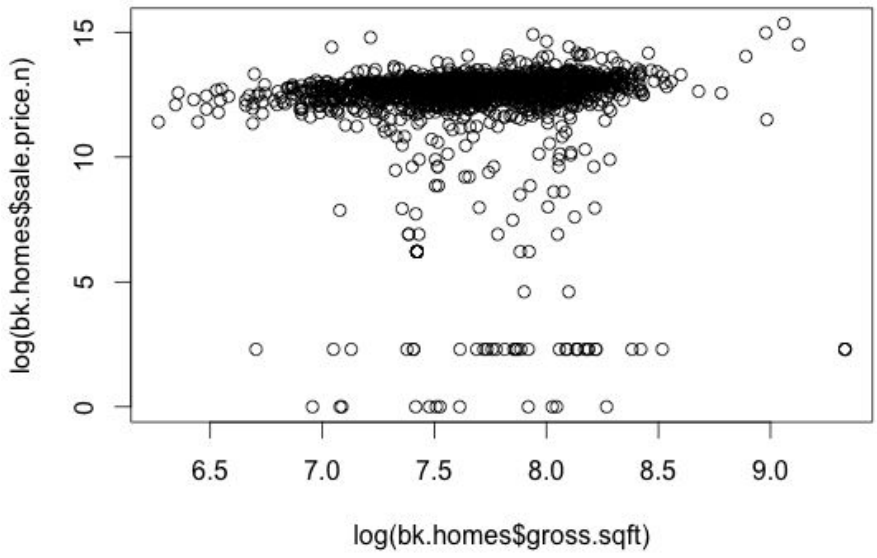
Following are the charts obtained from Bronx area:



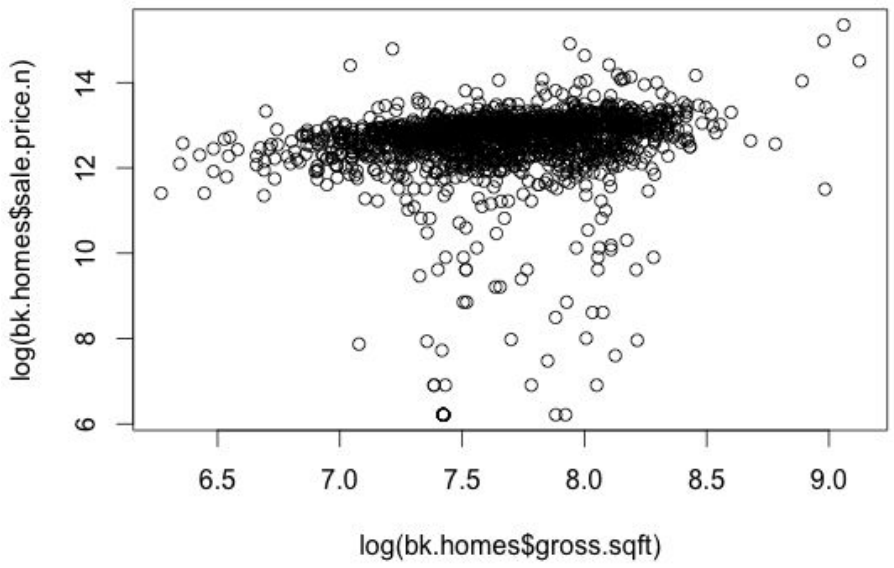

```
plot(bk.sale$gross.sqft,bk.sale$sale.price.n)
```



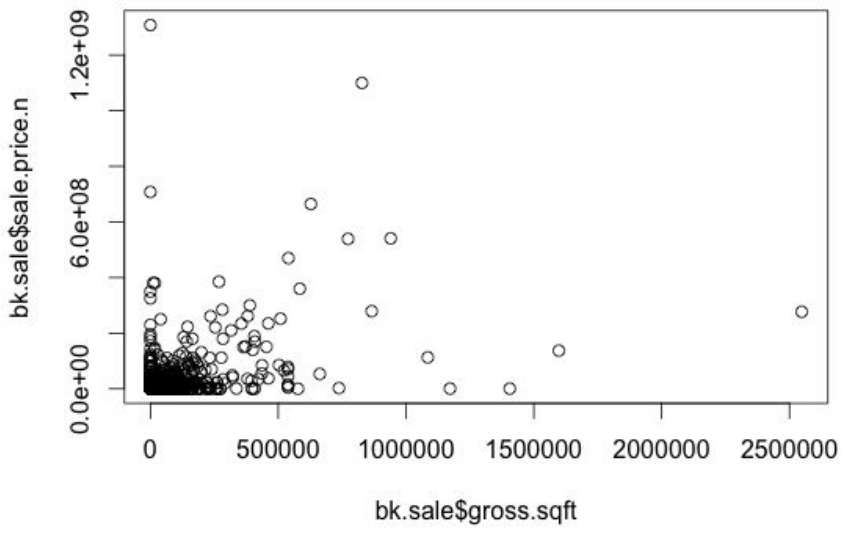
plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))



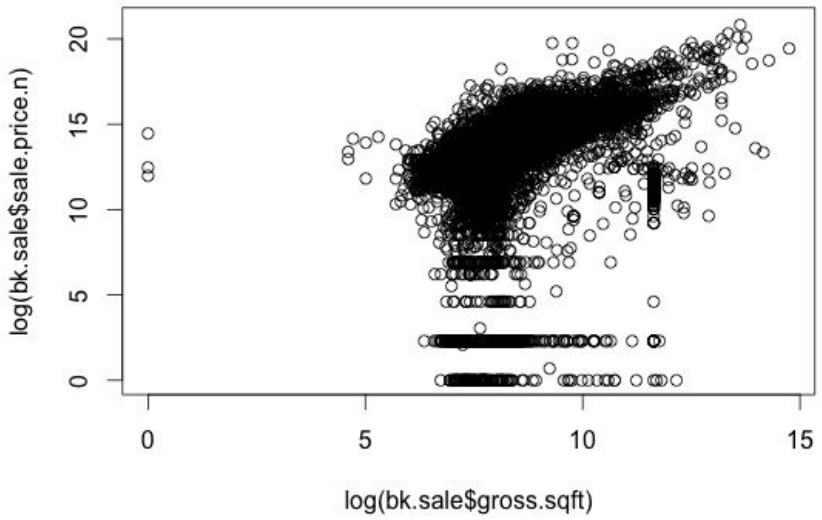
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
This plot considers only family homes



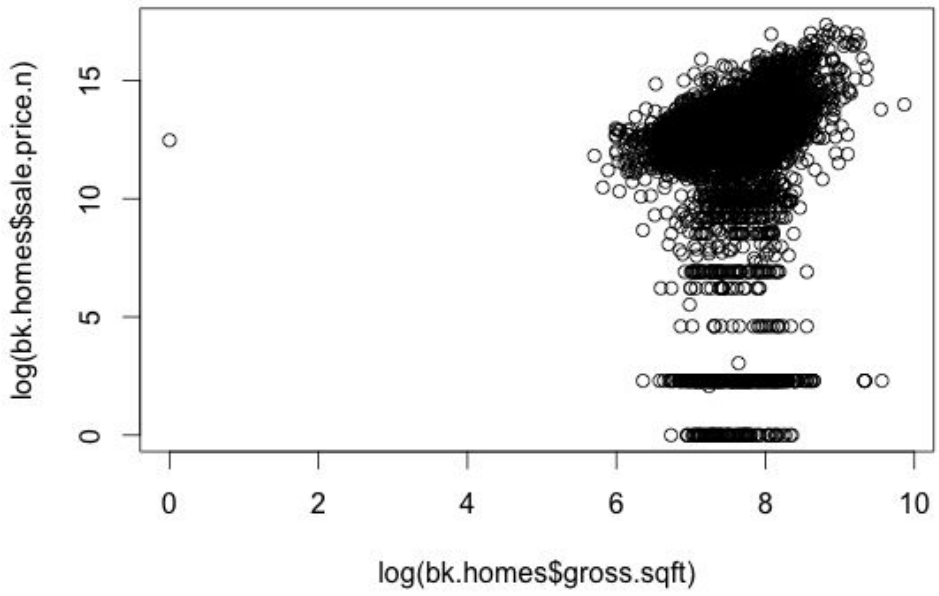
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
After removing some outliers
Now extending the analysis to the five areas, following are the charts:



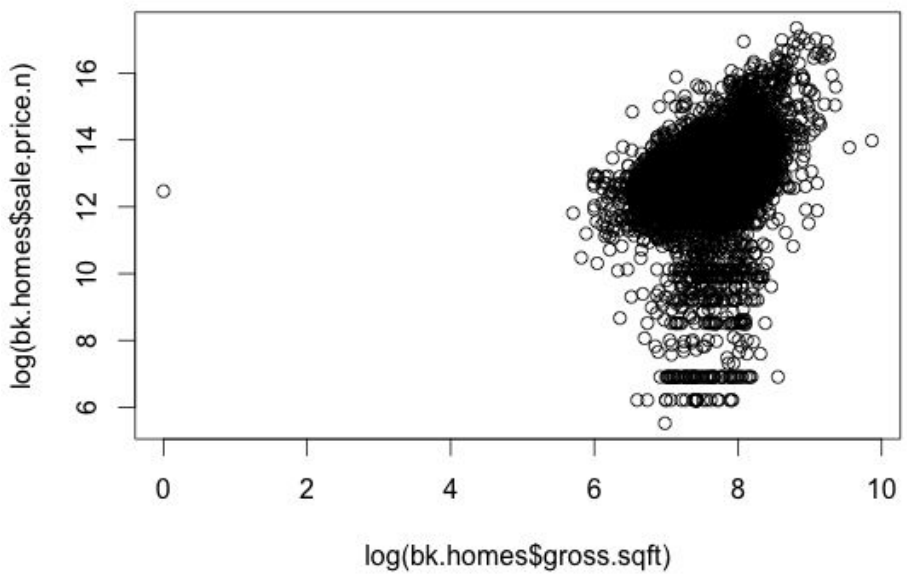
plot(bk.sale$gross.sqft,bk.sale$sale.price.n)



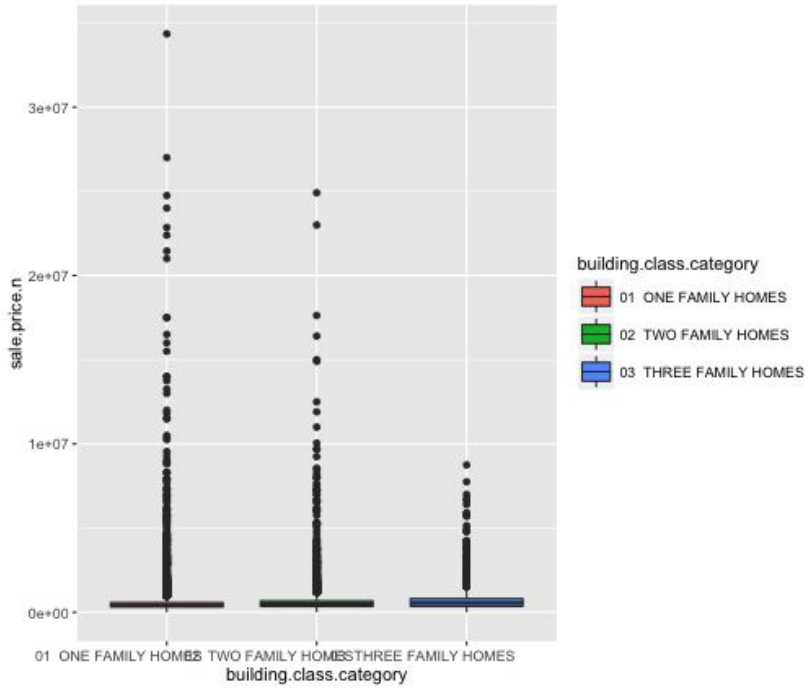
plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))



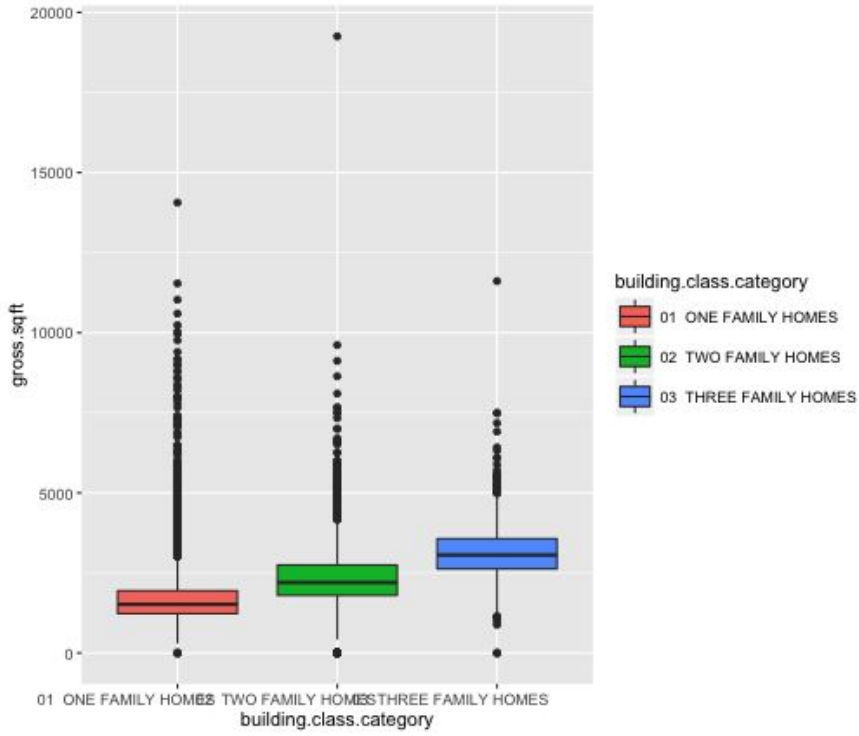
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
Family homes only



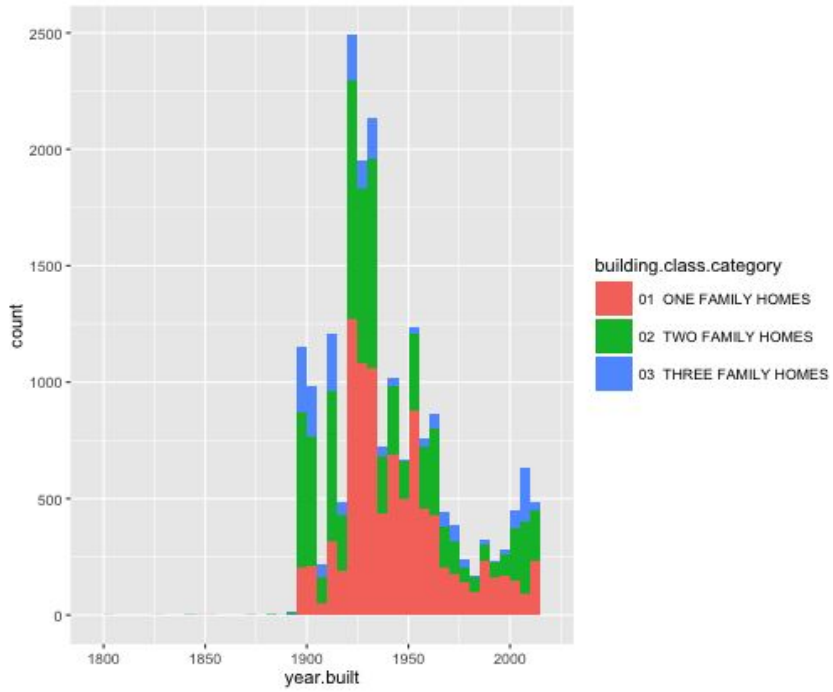
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
After removing outliers



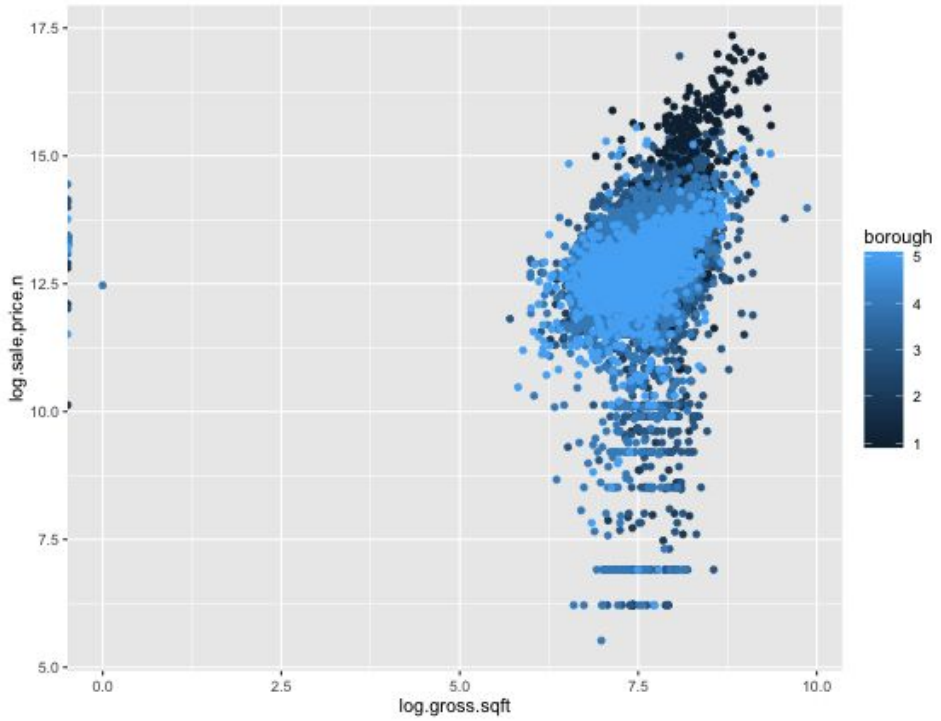
ggplot(bk.homes, aes(x=building.class.category, y= sale.price.n,
fill=building.class.category))+geom_boxplot()



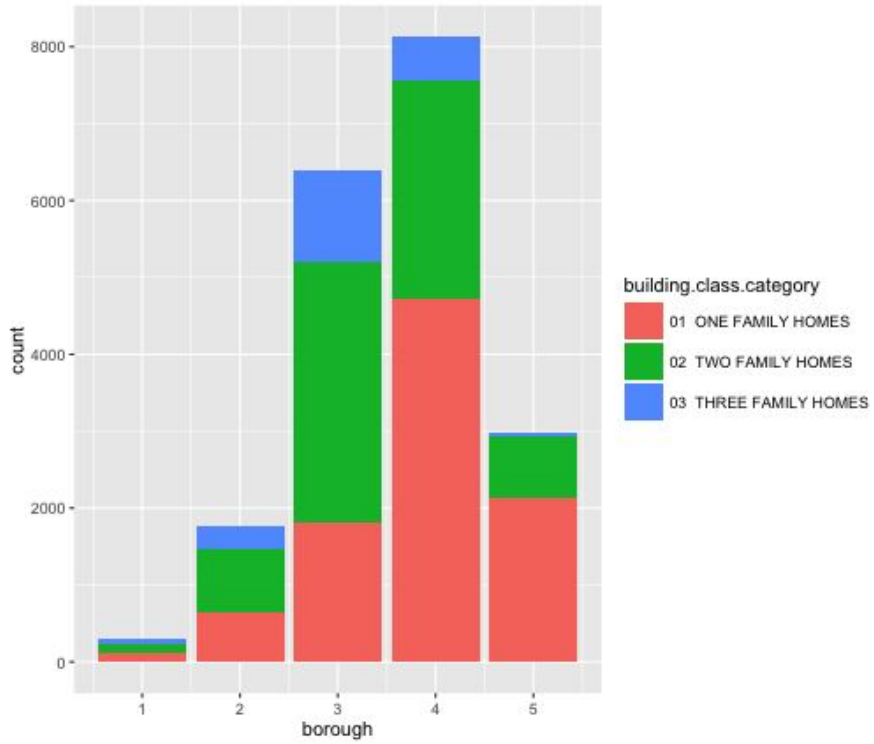
ggplot(bk.homes, aes(x=building.class.category, y= gross.sqft, fill=building.class.category))+geom_boxplot()



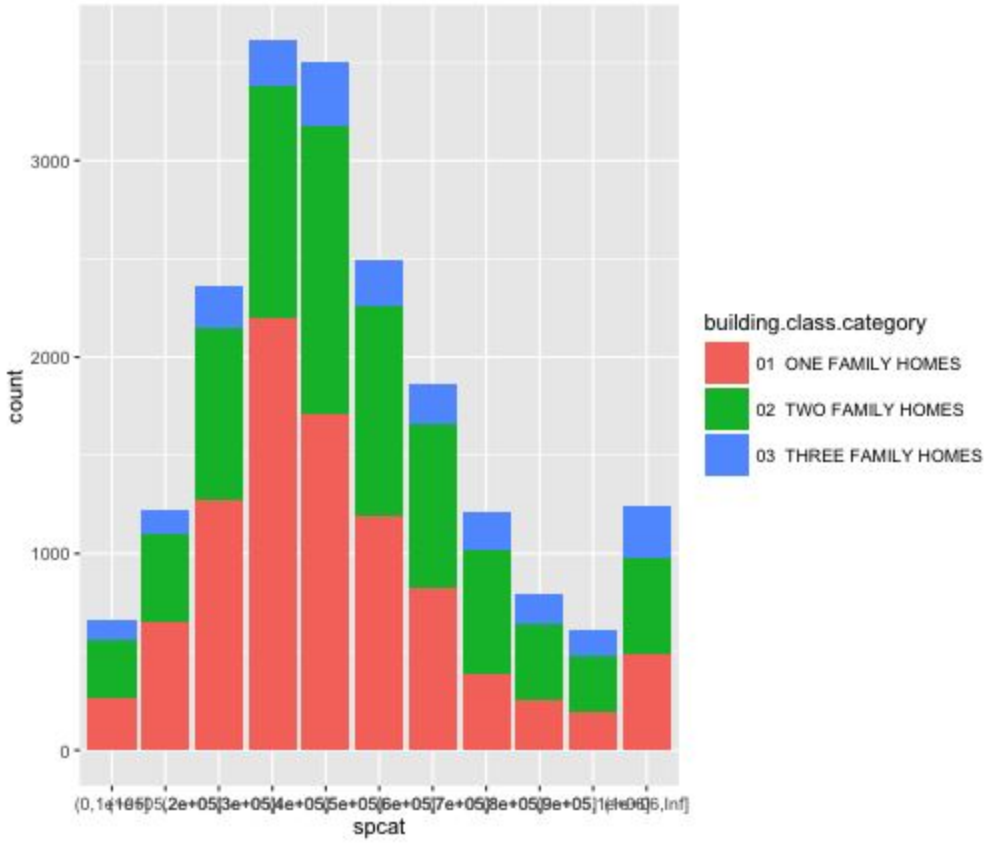
ggplot(bk.final, aes(x=year.built))+geom_histogram(binwidth = 5, aes(fill= building.class.category))



ggplot(bk.homes, aes(x=log.gross.sqft, y=log.sale.price.n))+geom_point(aes(colour= borough))



ggplot(bk.final, aes(x=borough))+geom_bar(aes(fill=building.class.category))



ggplot(bk.final,aes(x=spcat, fill= building.class.category))+geom_bar()

mean(bk.final$sale.price.n)
[1] 595121.2

To improve a product the analysis of the current product forms the basis. Enhancing the pros and degrading the cons is the only method to improvement. For improving the Real DIrect product we have analysed its current website and data.

- To analyse the current product the software engineers maintaining the website should collect the impressions and clicks data. This will lead to gaining of knowledge of user behaviour and things which attracts the user more on the website. This data can be analysed and used to further enhancement of user experience.
- By analyzing the above plots and charts it can be inferred that in most of the sale sale price and gross square feet area lie within a particular range. The major selling target is one family homes generating the most revenue. By looking at borough wise distribution of sale Manhattan area has the least number of sales and by considerably low amount.
- Not only the users currently using the website but also to attract users who are not yet familiar with the product user behaviour outside of the Real Direct environment should

be analysed. This can be done in various ways. First by contacting local realtors or circulating user based surveys. Second by collecting tweets in which people are talking about getting a house and performing a sentiment analysis of those tweets. As social media is very widespread these days and data is pouring in, this data should be put to some good use. Not only twitter but data from other social sites and blog sites should be collected and analysed.

- User behaviour is really necessary to improve the product as ultimately they are the judge.
    - Signed in and unsigned users should be handled differently, as signed users have a trust in the product and are going to stay but unsigned users need to have a reason to stay, and therefore the data should be kept separately.
    - The apartments are very expensive in New York as seen in the plots above therefore, a new product for apartment rental should be launched.
    - Also, online payment systems for rental apartments should be there to make the user experience better.