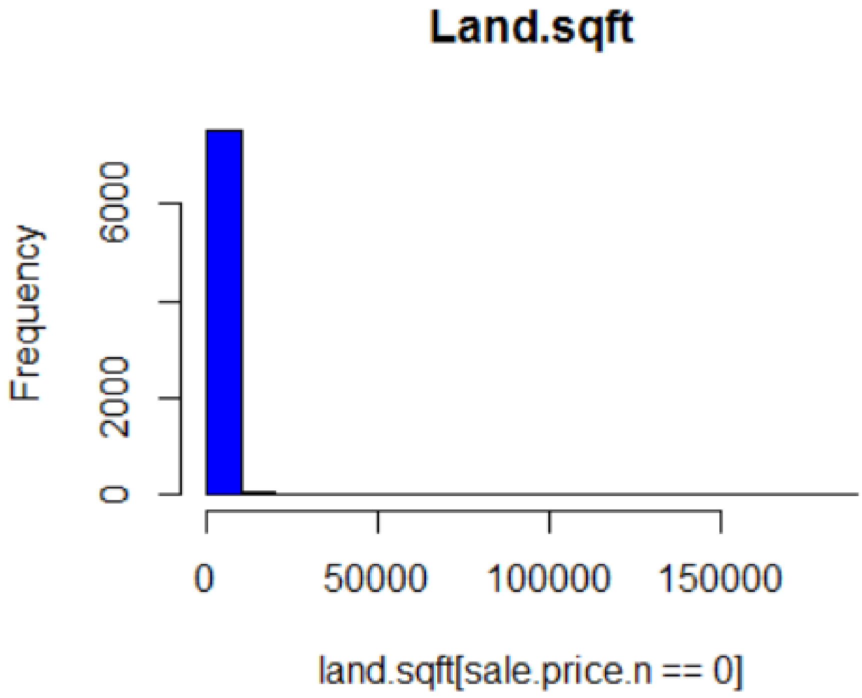
Problem 3 Case Study

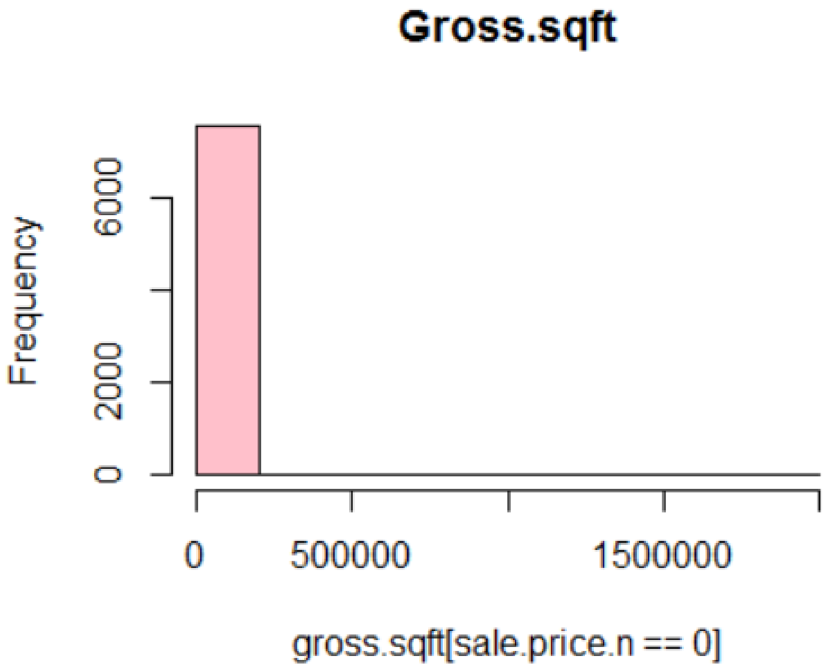
This question is quite similar to Problem 2. The difference is in the given data and also a slight variation in the approach. In this question, the main task is to perform EDA on a simple data set. The first part of this question demands an EDA on the sales of one of the Bronx of New York state. To do that, I first ensured that the data provided is free from redundant data and then followed the Linear Regression process to obtain EDA. Also to mention, in order to fetch the .xls file, Windows machine required **Perl** to be installed along with the absolute path of perl.exe mentioned as one the arguments in read.xls() command.

Following were the plots obtained:

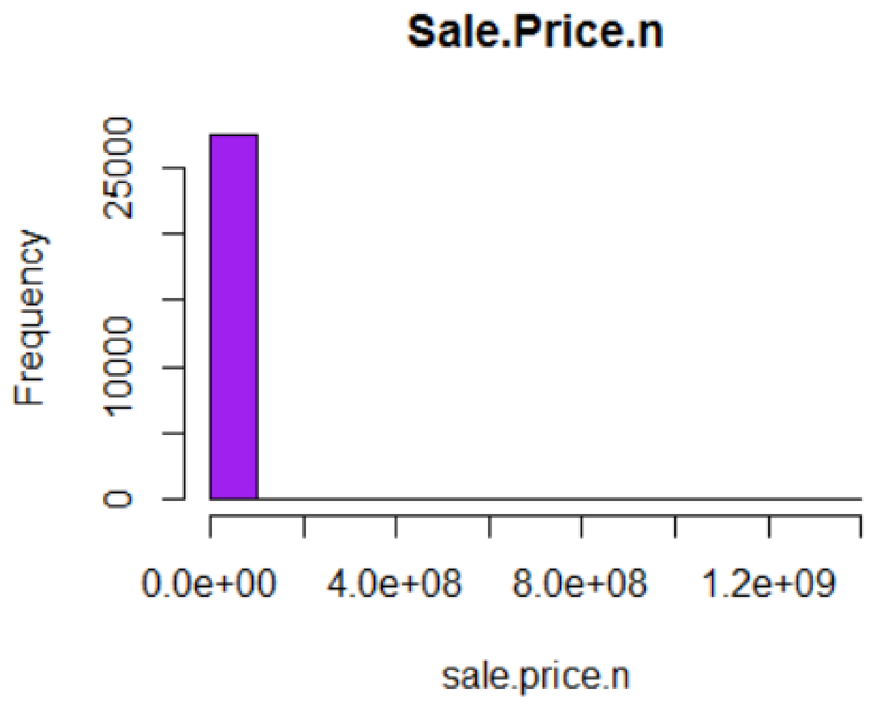
The first scenario represents Land.sqft vs Frequency when **sale.price.n==0 for one Borough of New York**



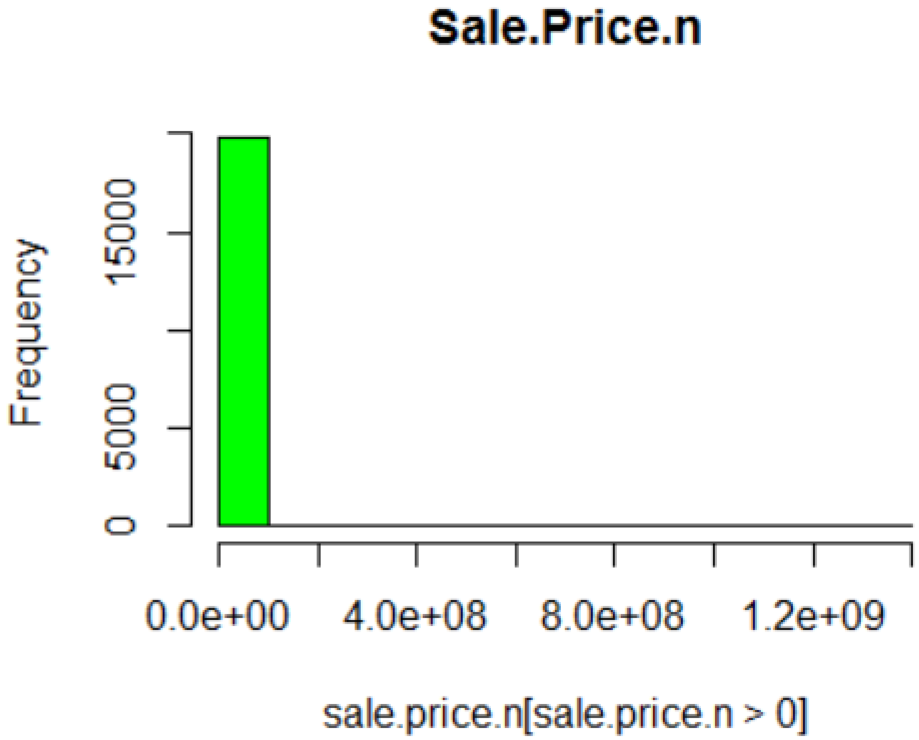
The next scenario represents Gross.sqft vs Frequency when **sale.price.n==0 for one Borough of New York**



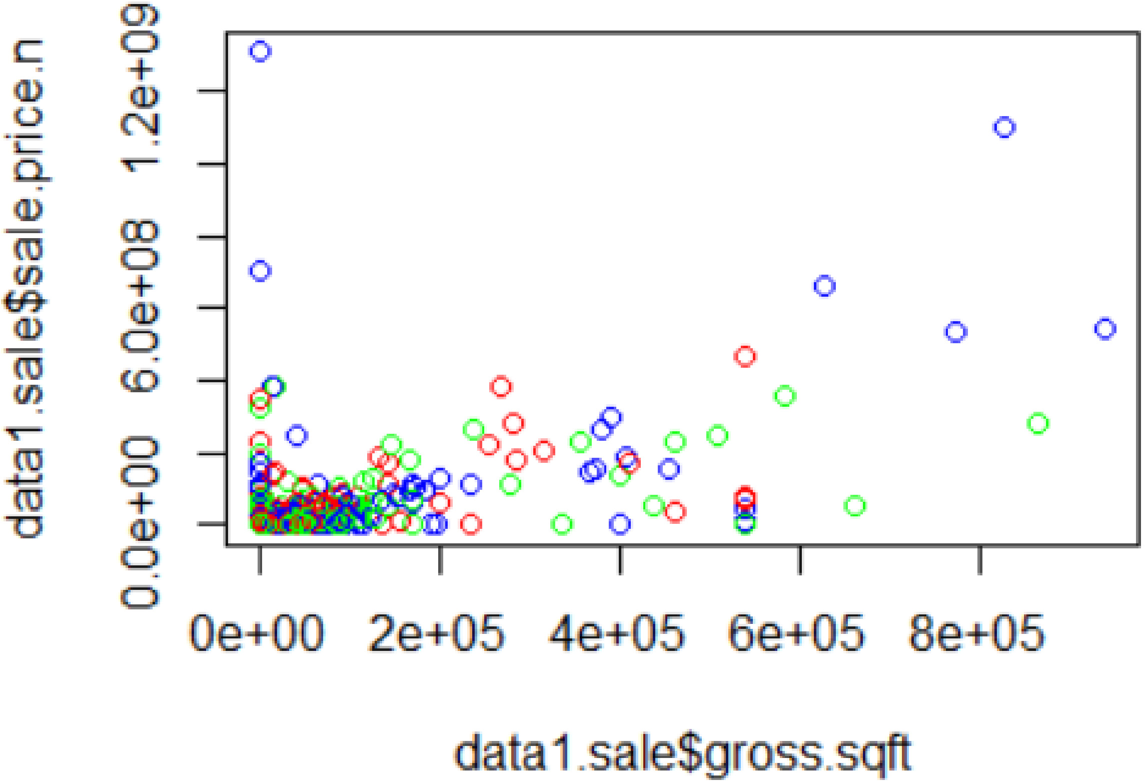
The next scenario represents Sale.Price.n vs Frequency **for one Borough of New York**



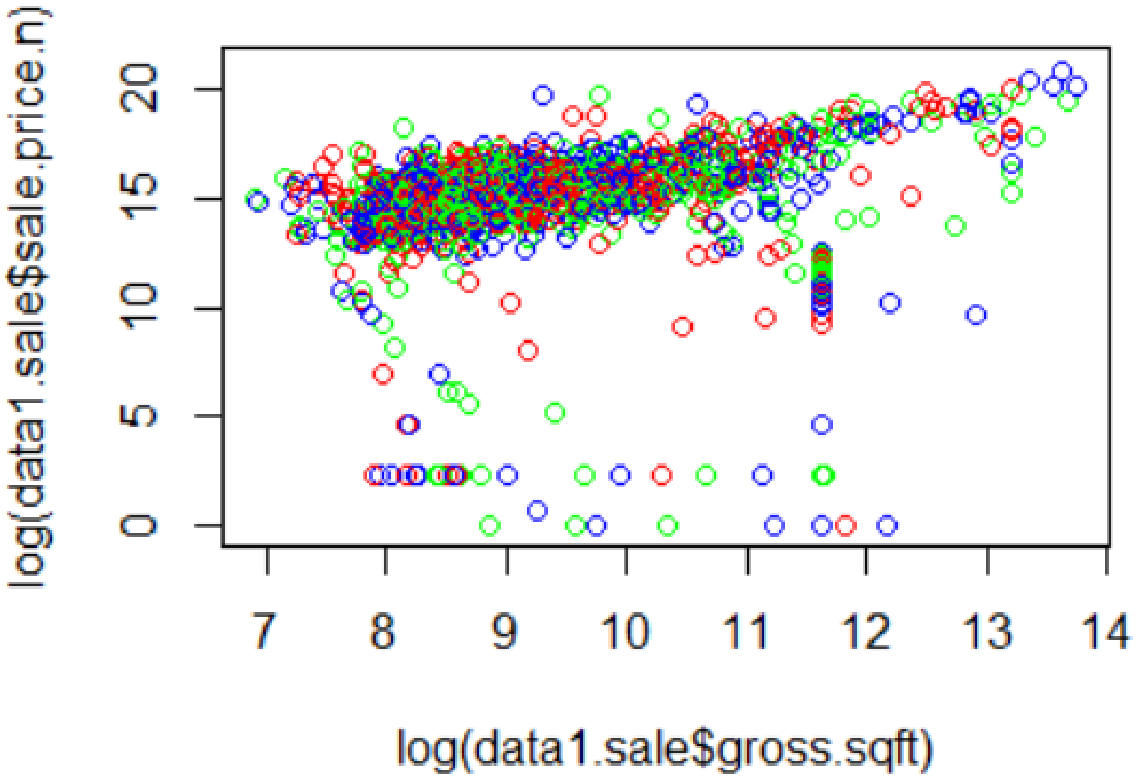
The next scenario represents Sale.Price.n vs Frequency when **Sale.Price.n>0 for one Borough of New York**

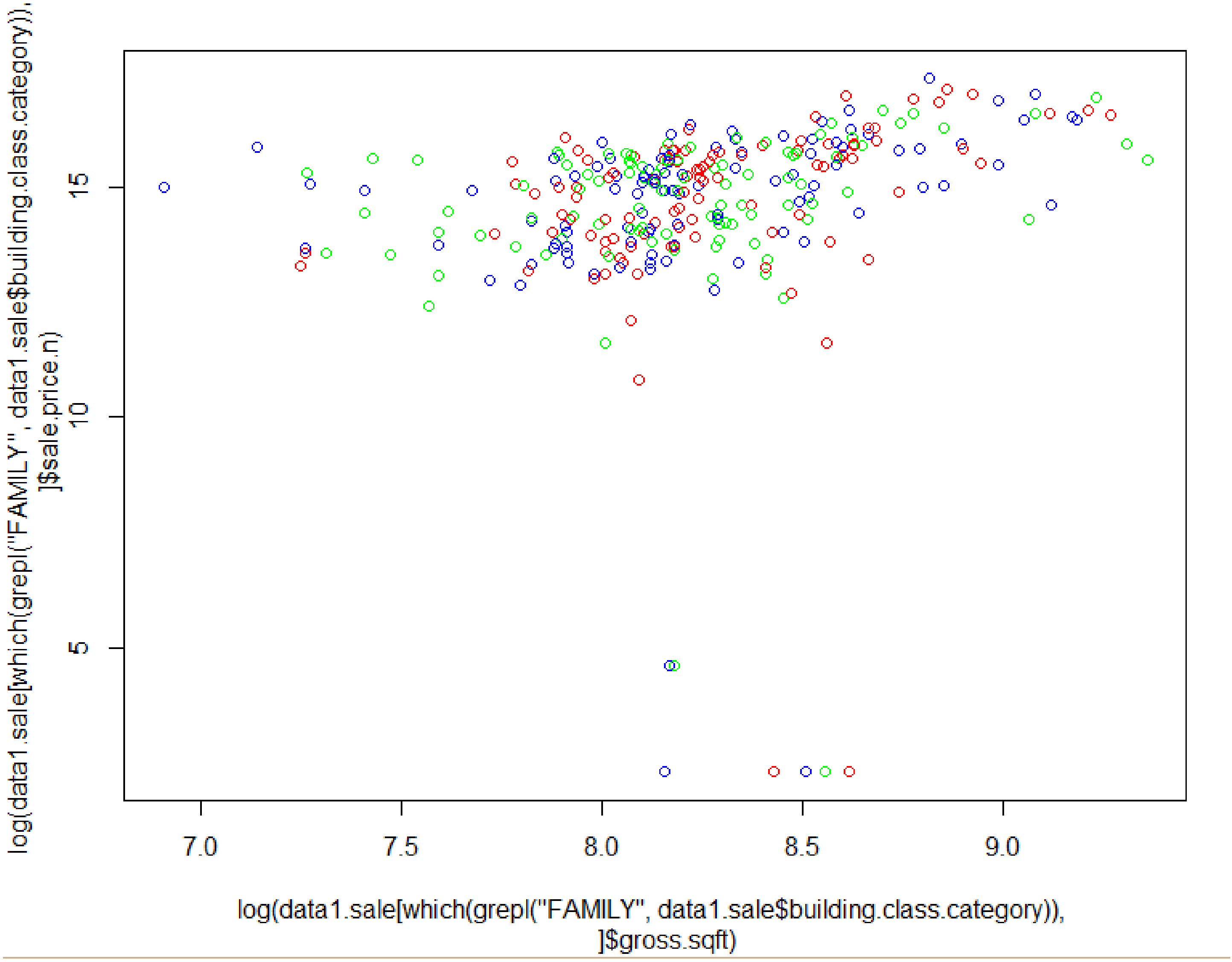

Sale.Price.n
Frequency
0
5000
15000
0.0e+00
4.0e+08
8.0e+08
1.2e+09
sale.price.n[sale.price.n > 0]

Now in the below plot, it can be clearly seen that the density of data is very high near 0 square feet which is due to the presence of noise.



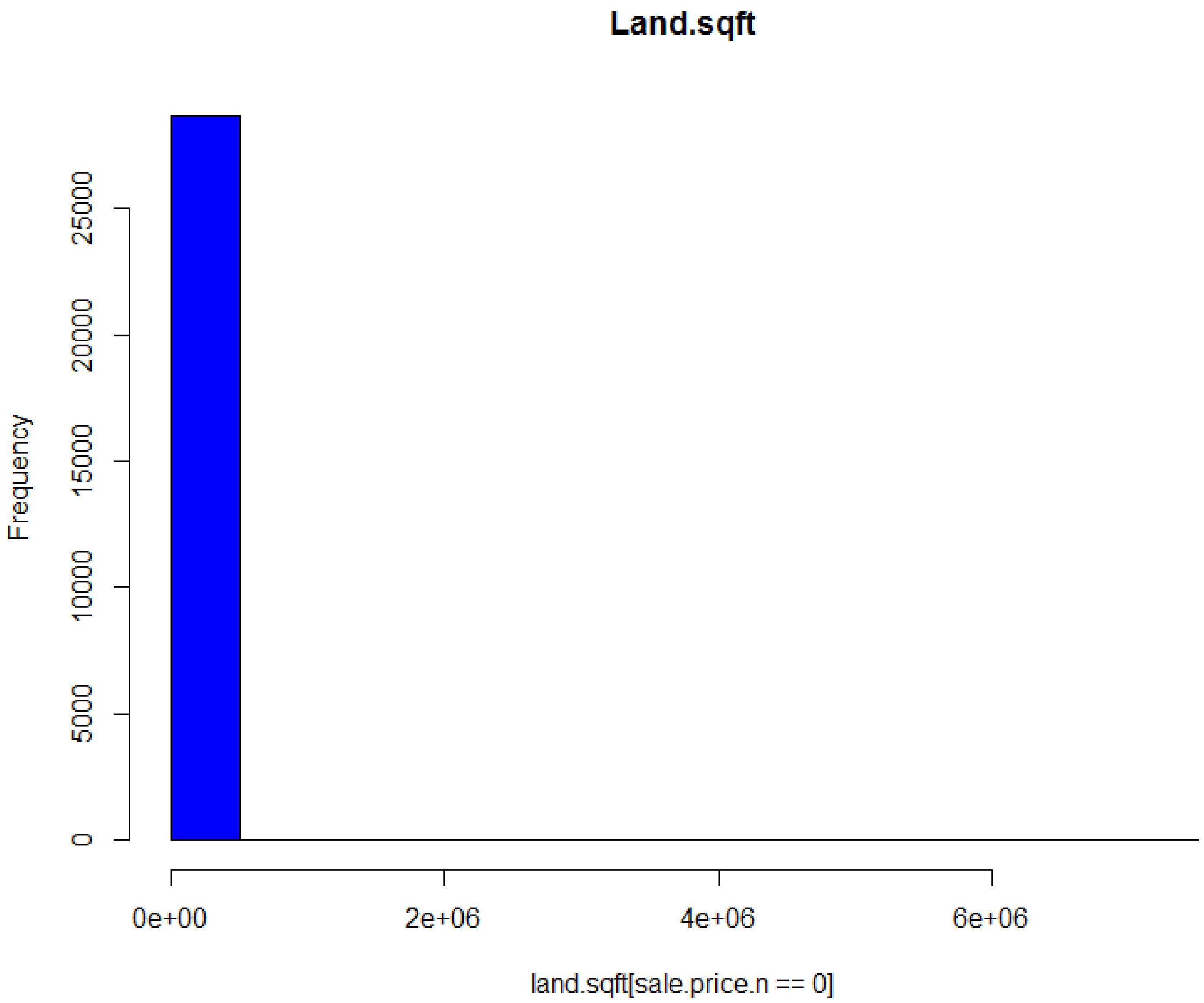
Let us try to plot the data after removing the induced noise. The below graph validates the removal of noise.



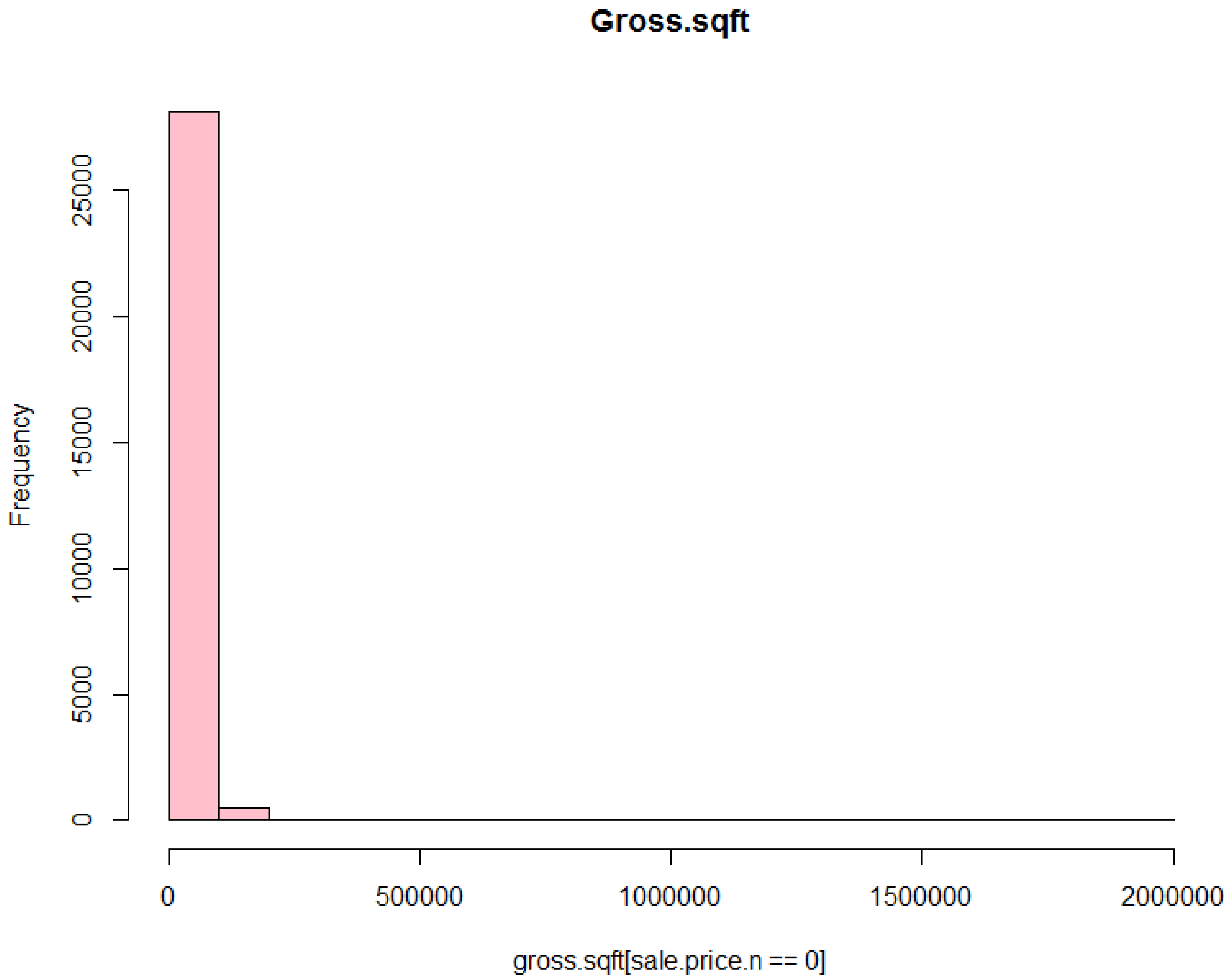


Now extending the above analysis to all the six boroughs of New York to have a combined analysis.

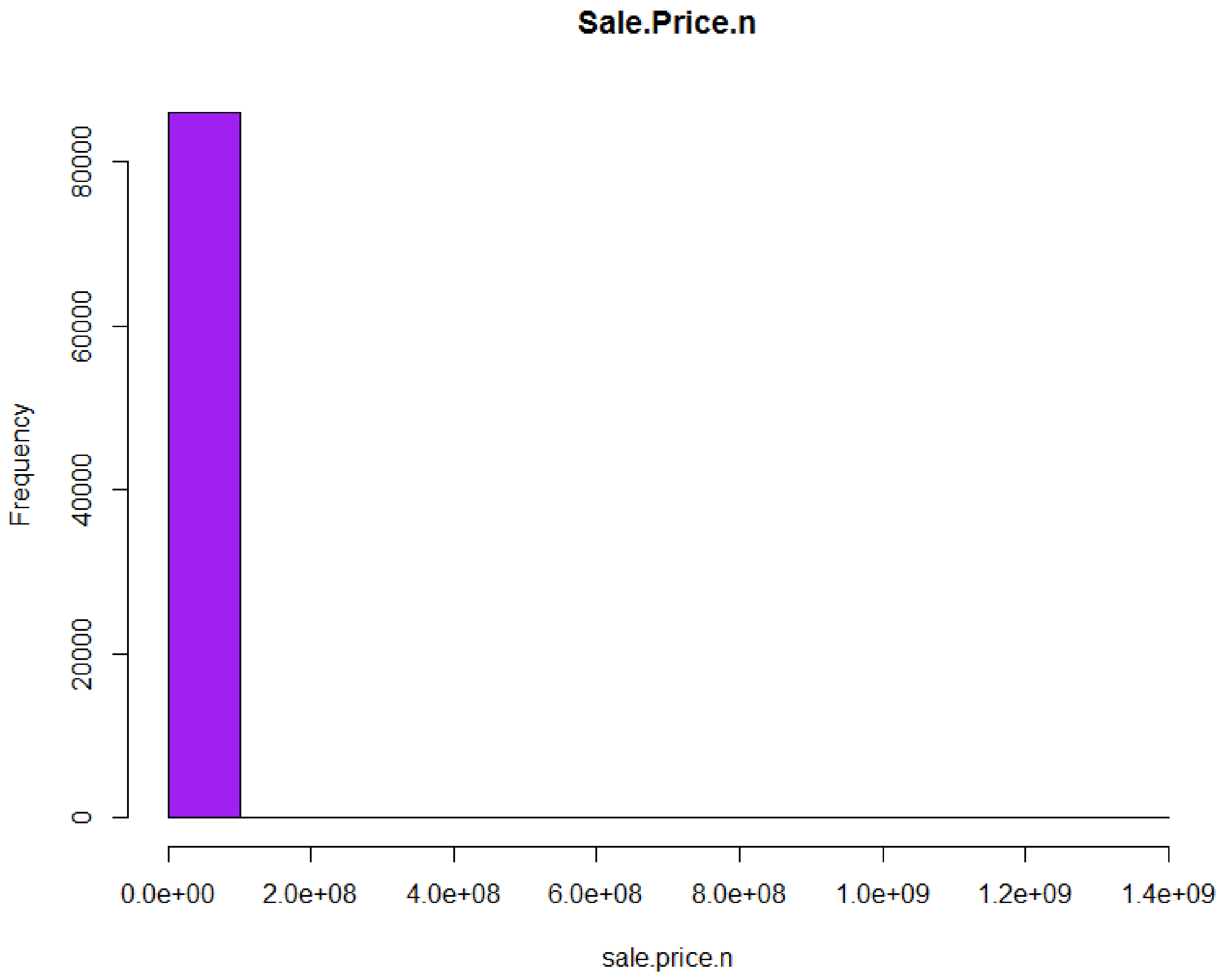
The first scenario represents Land.sqft vs Frequency when **sale.price.n==0 for all 6 Boroughs of New York**



The next scenario represents Gross.sqft vs Frequency when
sale.price.n==0 for all 6 Boroughs of New York

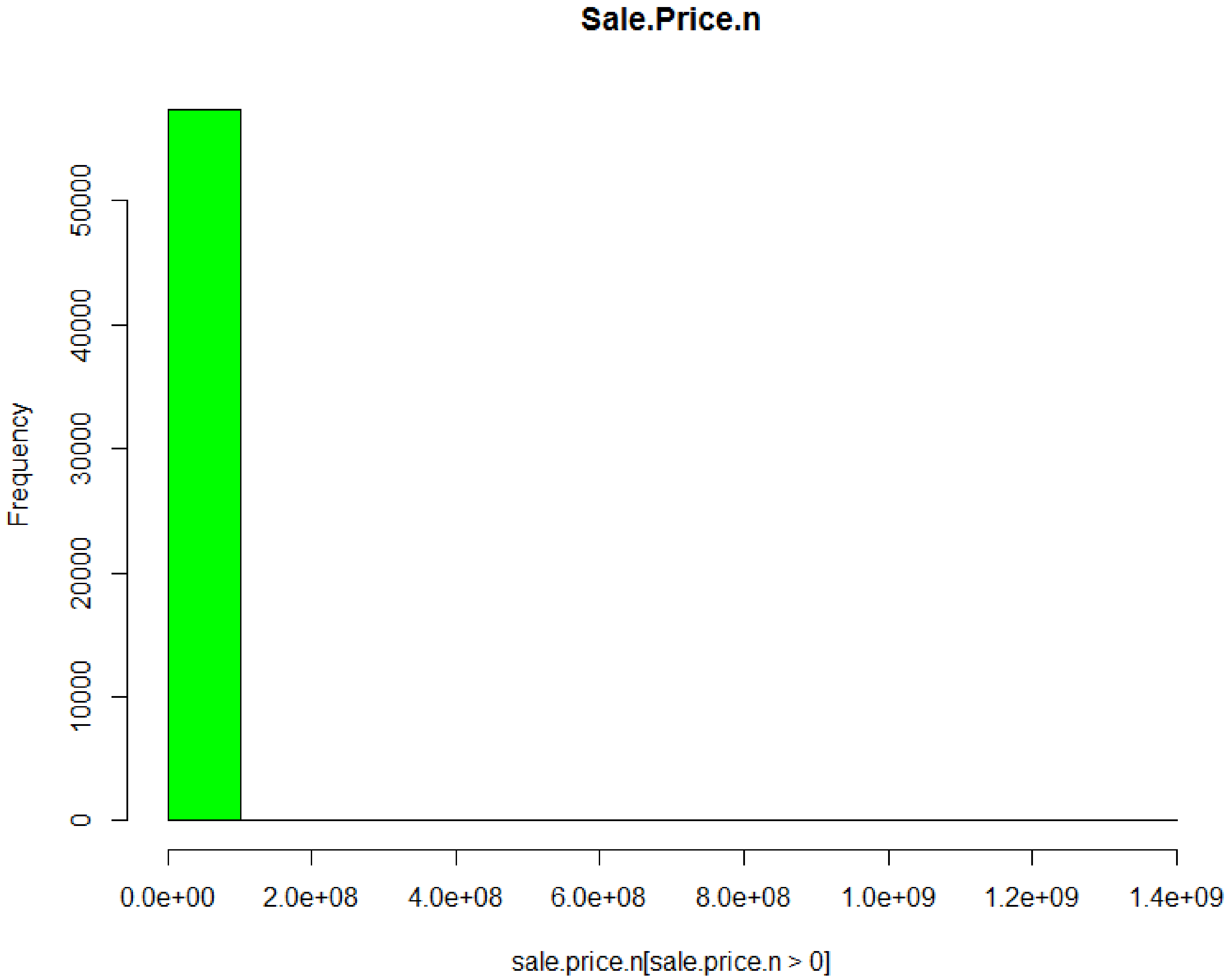


The next scenario represents Sale.Price.n vs Frequency **for all 6 Boroughs of New York**

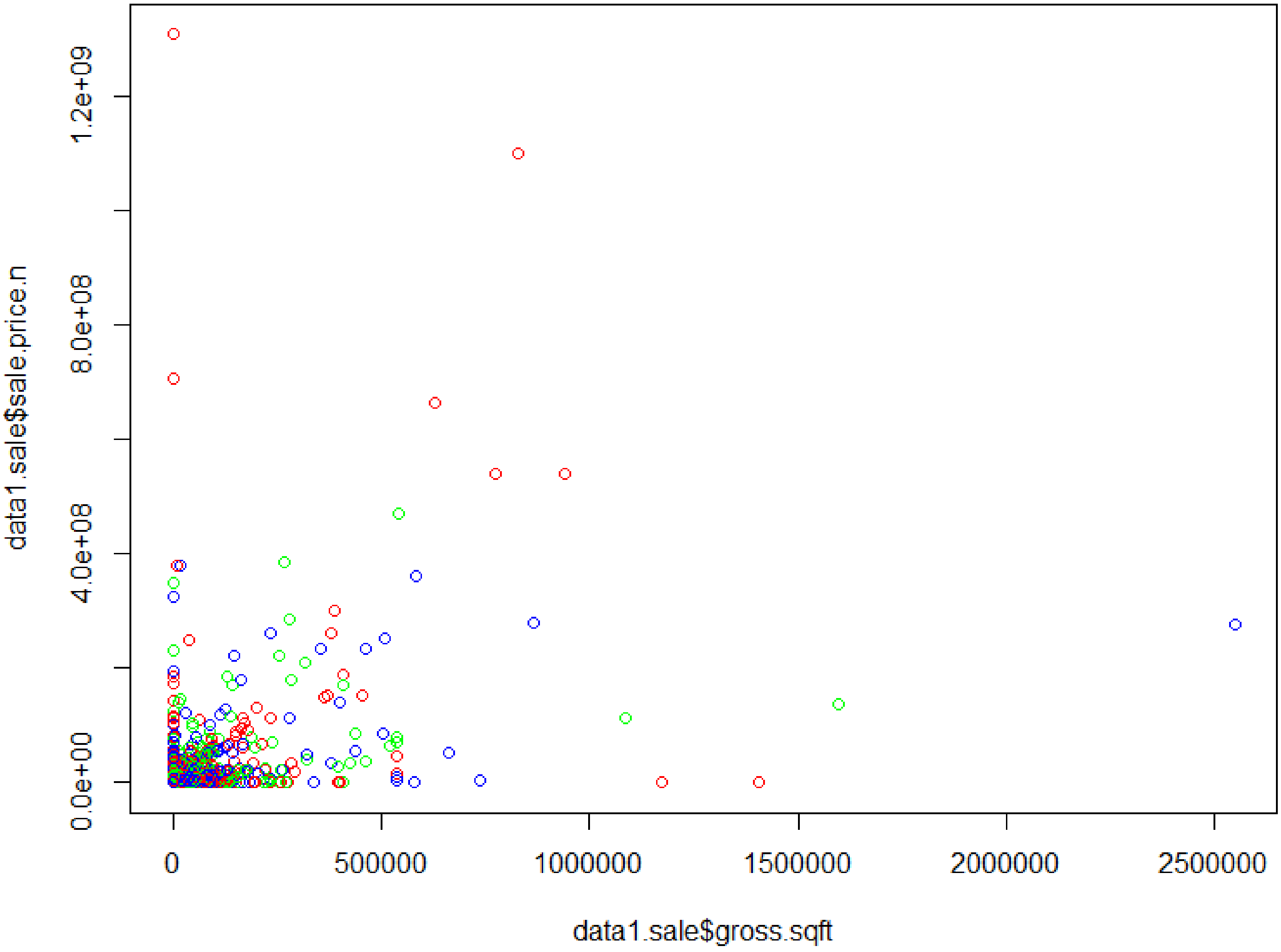


The next scenario represents Sale.Price.n vs Frequency when

Sale.Price.n>0 for all 6 Boroughs of New York

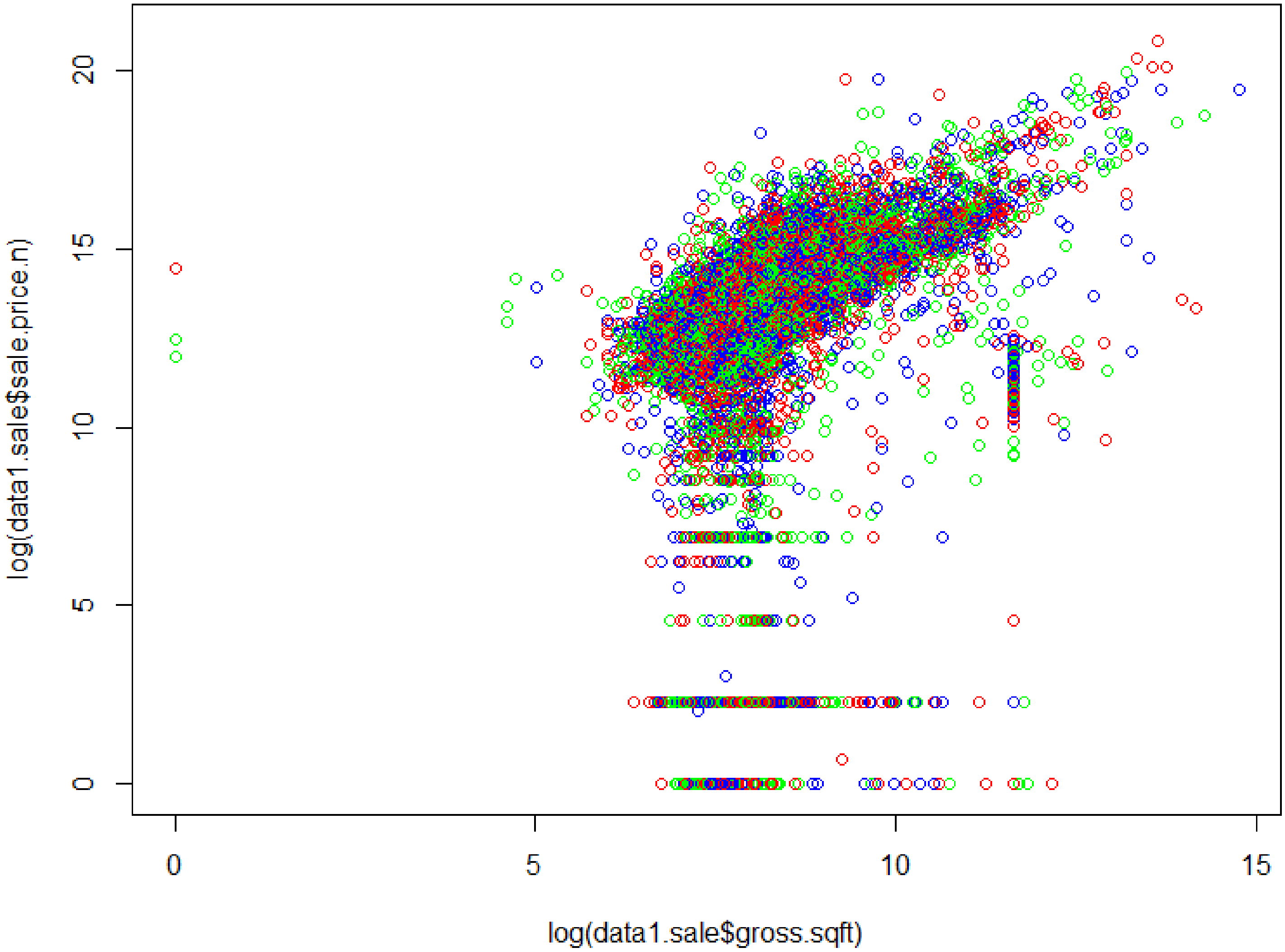


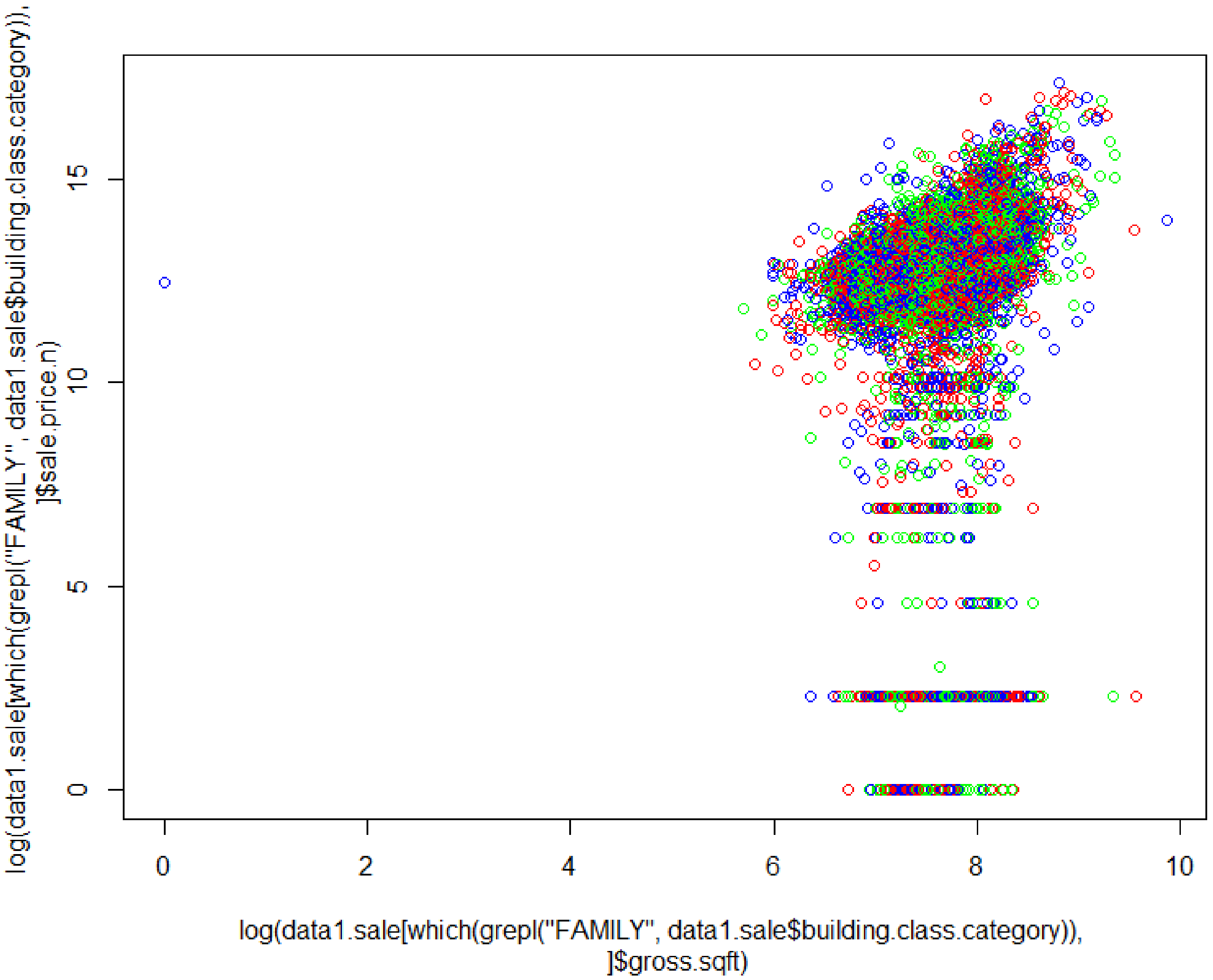
Now in the below plot for all the boroughs of New York, it can be clearly seen that the density of data is very high near 0 square feet which is due to the presence of noise.



data1.sale$sale.price.n
0.0e+00
4.0e+08
8.0e+08
1.2e+09
0
500000
1000000
1500000
2000000
2500000
data1.sale$gross.sqft

After removing the erroneous noise the plot improves as shown below




log(data1.sale[which(grepl("FAMILY", data1.sale$building.class.category)),]$sale.price.n)
0
5
10
15
0
2
4
6
8
10
log(data1.sale[which(grepl("FAMILY", data1.sale$building.class.category)),]$gross.sqft)