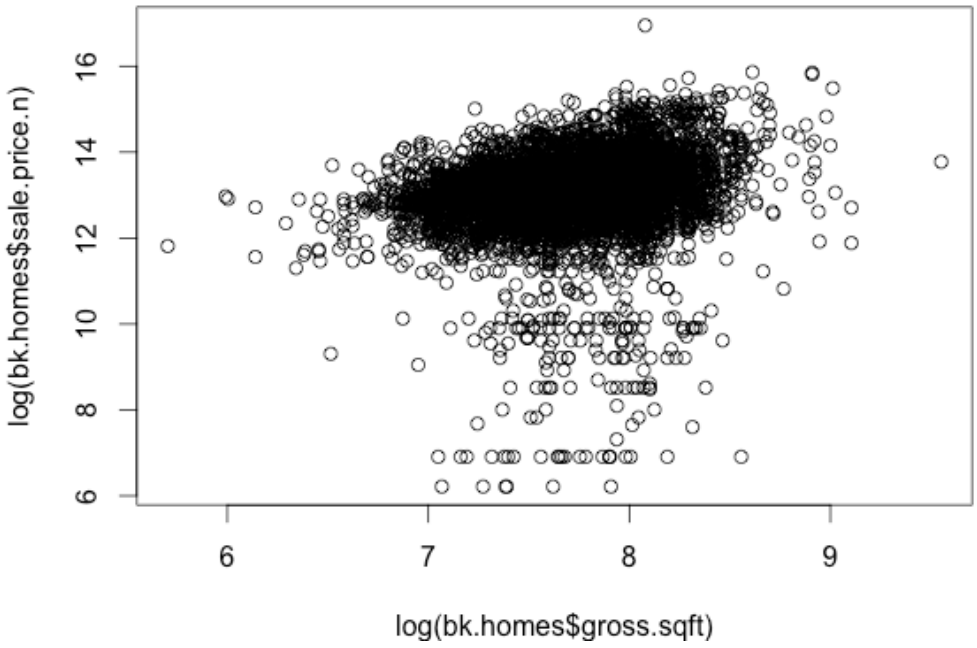
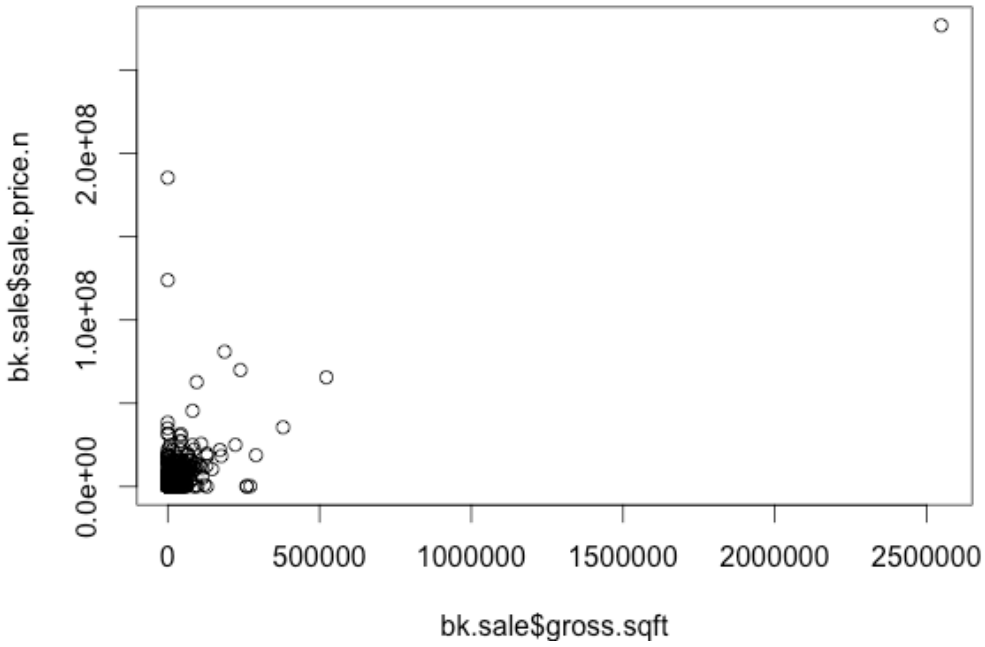
CSE587:DATA INTENSIVE COMPUTING

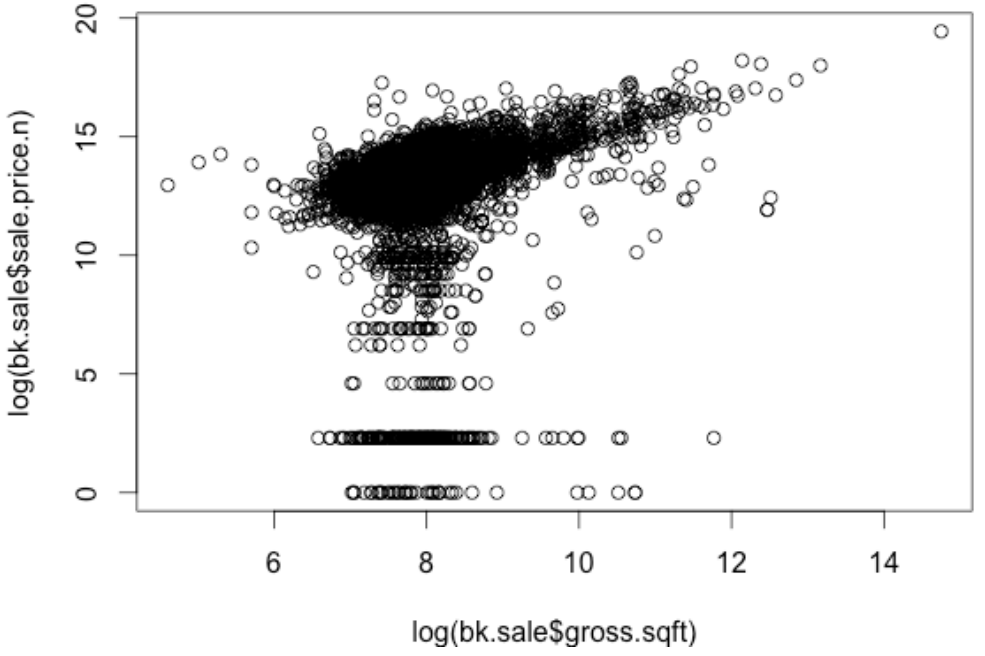
PROBLEM3:



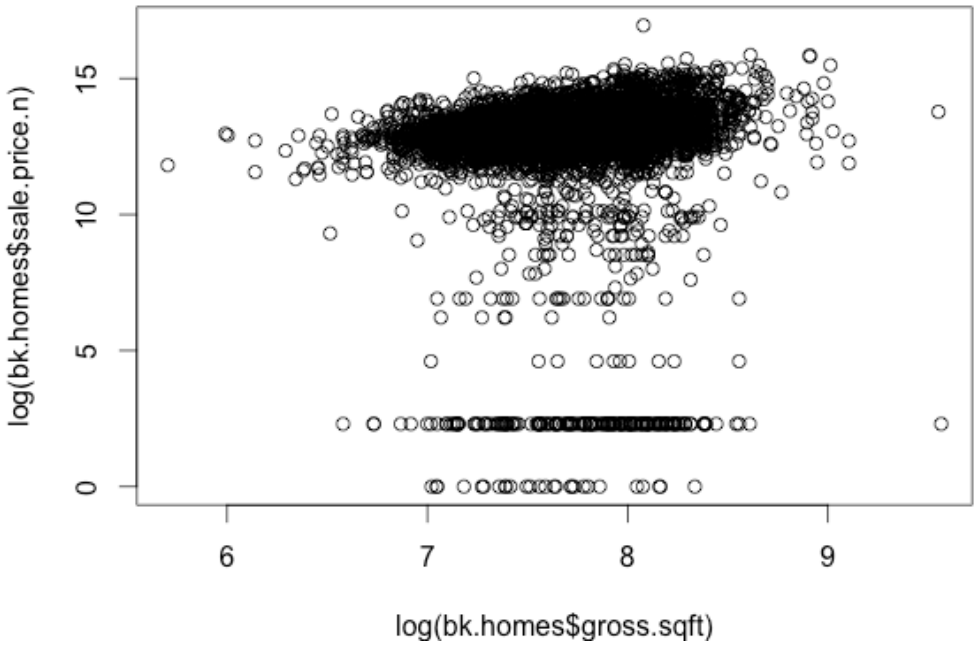
The above chart helps in identifying the outliers in the Brooklyn dataset which helps in cleaning the outliers. A cleaner data is required for accurate analysis of the data.

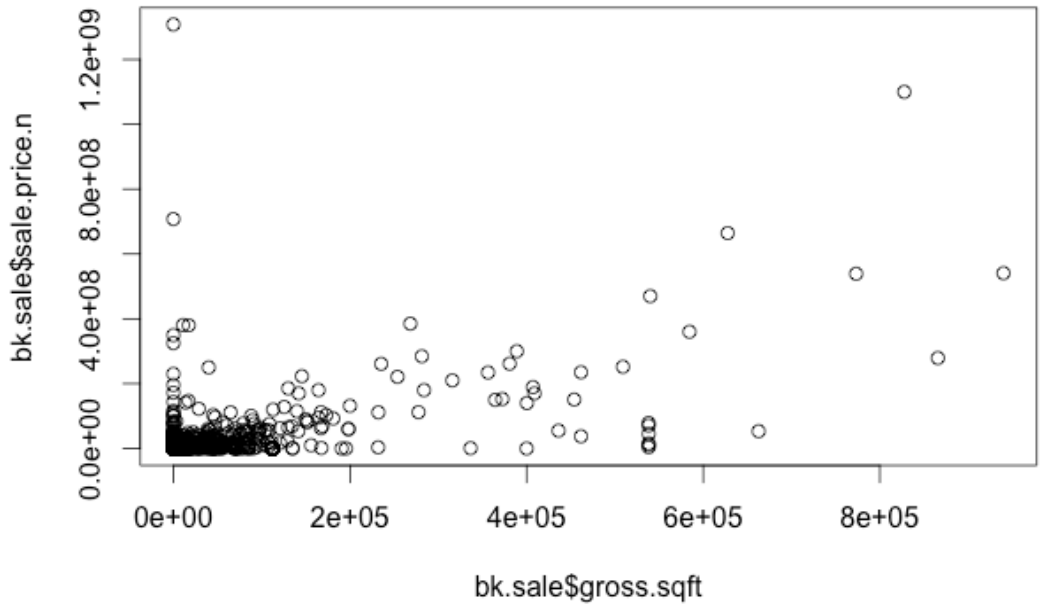


This chart shows the relation between the sale price and the gross square feet. We can see that the dots are more concentrated in the 0 to 500000 square feet range.

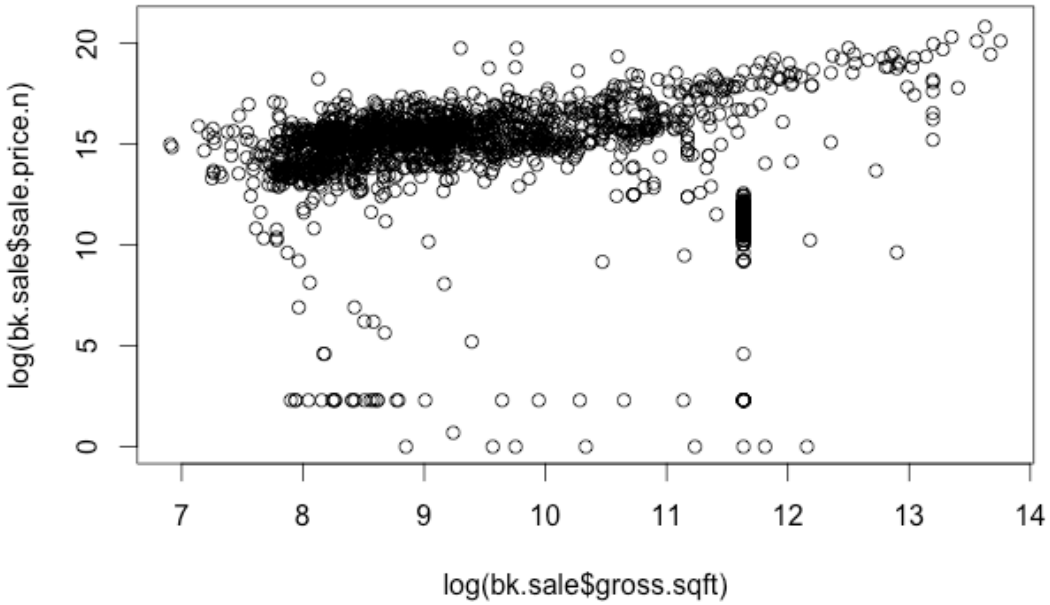


Here we can see that the sale price ranges from 0 to 20 and the corresponding gross square feet is in range of 6 to 8. This gives a much more exact correlation than the previous graph and shows that the price corresponds to the square feet.

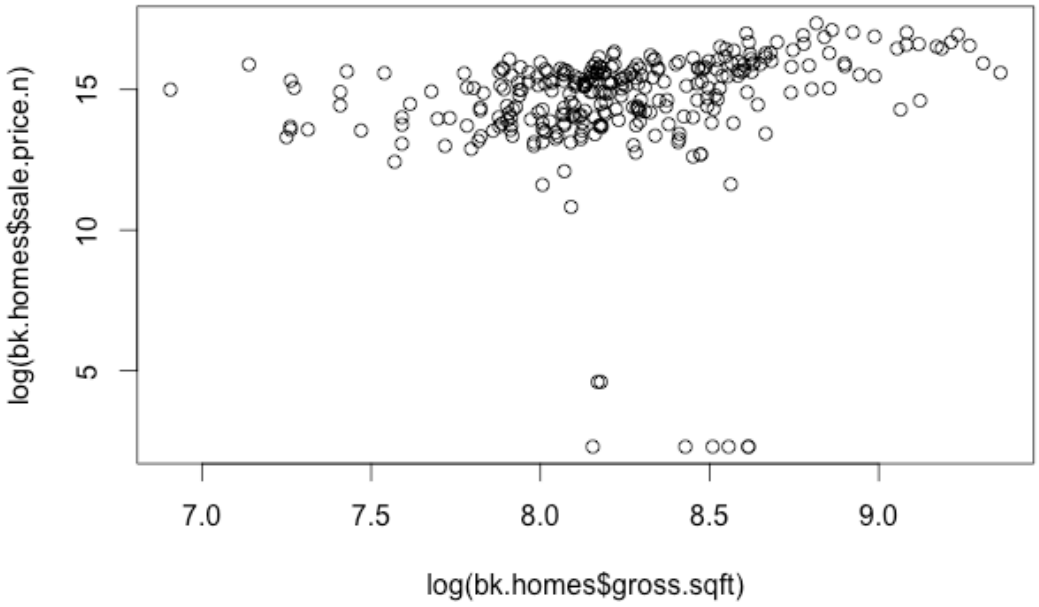




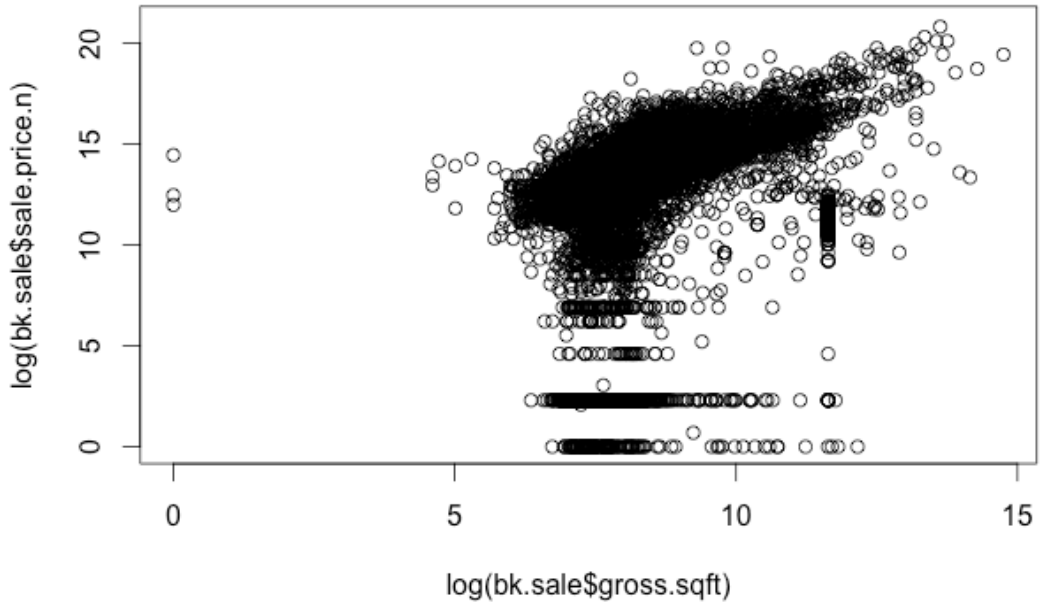
We can see that the Manhattan graph is much more scattered than the Brooklyn graph which shows the nuances in the sale price with respect to gross square feet variation.



When we take log, we can observe the correlation better between the two values.



Extending the dataset, we get the following plots,



The plot becomes much more denser than was for the previous one day data and is mainly concentrated in the middle