DIC

PROJECT 1 : PROBLEM 3

DATA ECONOMY REAL CASE STUDY : REAL DIRECT

NAME : ANIKET THIGALE
UB NO : 50168090

PROBLEM DESCRIPTION

Write a R script to clean data, perform EDA and analyse the RealDirect data set to find some insights.

EDA

Exploratory Data Analysis (EDA) is the first step toward building a model. It is a critical part of the data science process. The basic tools of EDA are plots, graphs and summary statistics. Generally speaking, it's a method of systematically going through the data,plotting distributions of all variables (using box plots), plotting time series of data, transforming variables, looking at all pairwise relationships between variables using scatterplot matrices, and generating summary statistics for all of them. At the very least that would mean computing their mean, minimum, maximum, the upper and lower quartiles, and identifying outliers.

DATA

The data is present in the XLS format. There are 21 variables like borough,neighborhood, block, lot etc. per record.Use perl to import the data to R.

CLEANING AND ADDING FEATURES TO DATA

1. Created a new variable, sale_price_n, to remove the $ from sale.price
2. Make land.square.feet, year.built and gross.square.feet numeric
3. Make sale.date as date in R
4. Replace the '.' in column names by '_' since sqldf package does not accept columns with '.' in them.
5. Remove outlier data for eg, data with sale.price as 0, gross.square.feet=0, where neighborhood is not mentioned, no zipcodes, year.built is 0 etc. to finally get a full dataset.
6. For the multiple boroughs data, also add a variable BoroughName according to the data that it corresponds to.

ANALYSIS FOR A SINGLE BOROUGH

Here we take the data for the brooklyn borough.

SUMMARY

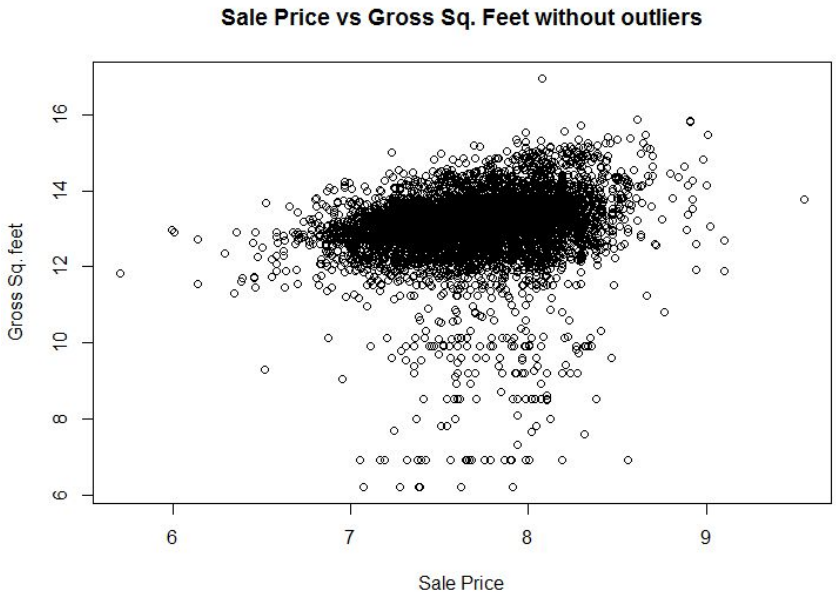
```
> #head(bk)
> summary(bk)
   BOROUGH                     NEIGHBORHOOD                                  BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT
 Min.   :3   BEDFORD STUYVESANT       : 1699   02  TWO FAMILY HOMES                    :5776       1      :10976
 1st Qu.:3   EAST NEW YORK            : 1394   01  ONE FAMILY HOMES                    :2890       2      : 6070
 Median :3   BOROUGH PARK             : 1020   13  CONDOS - ELEVATOR APARTMENTS        :2739       4      : 2445
 Mean   :3   BUSHWICK                 :  898   03  THREE FAMILY HOMES                  :2255       2A     : 1512
 3rd Qu.:3   CROWN HEIGHTS            :  886   10  COOPS - ELEVATOR APARTMENTS         :2129       2C     : 1024
 Max.   :3   PARK SLOPE               :  848   07  RENTALS - WALKUP APARTMENTS         :1755       1B     :  422
             (Other)                  :16628   (Other)                                 :5829       (Other):  924
     BLOCK          LOT         EASE.MENT      BUILDING.CLASS.AT.PRESENT                                  ADDRESS
 Min.   :  20   Min.   :   1.0   Mode:logical   R4     : 2703            163 WASHINGTON AVENUE                 :  106
 1st Qu.:1638   1st Qu.:  22.0   NA's:23373     C0     : 2258            205 WATER STREET                      :   76
 Median :3839   Median :  48.0                 D4     : 2125            380 COZINE AVENUE                     :   65
 Mean   :3984   Mean   : 305.4                 B1     : 2080            34 NORTH 7TH   STREET                 :   63
 3rd Qu.:6259   3rd Qu.: 142.0                 B3     : 1229            12399 FLATLANDS AVENUE                :   62
 Max.   :8955   Max.   :9039.0                 B2     : 1115            306 GOLD STREET                       :   62
                                               (Other):11863            (Other)                               :22939
    APART.MENT.NUMBER    ZIP.CODE     RESIDENTIAL.UNITS COMMERCIAL.UNITS   TOTAL.UNITS     LAND.SQUARE.FEET GROSS.SQUARE.FEET
             :17632   Min.   :    0   Min.   :  0.000   Min.   :  0.0000   Min.   :  0.00   0       : 8027   0       : 8934
 4           :  204   1st Qu.:11209   1st Qu.:  1.000   1st Qu.:  0.0000   1st Qu.:  1.00   2,000   : 2201   3,000   :  230
 6           :  183   Median :11218   Median :  1.000   Median :  0.0000   Median :  1.00   2,500   : 1149   3,600   :  189
 3           :  155   Mean   :11211   Mean   :  2.156   Mean   :  0.1973   Mean   :  2.37   1,800   :  597   2,400   :  185
 2           :  144   3rd Qu.:11230   3rd Qu.:  2.000   3rd Qu.:  0.0000   3rd Qu.:  2.00   4,000   :  474   2,700   :  146
 1           :  125   Max.   :11416   Max.   :509.000   Max.   :222.0000   Max.   :509.00   3,000   :  307   3,300   :  139
 (Other)     : 4930                                                                        (Other):10618   (Other):13550
   YEAR.BUILT   TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE    SALE.PRICE         SALE.DATE
 Min.   :   0   Min.   :1.000            R4     : 2739                   $0      : 8791   2012-09-27:  675
 1st Qu.:1901   1st Qu.:1.000            C0     : 2255                   $10     :  241   2012-12-27:  245
 Median :1925   Median :1.000            D4     : 2125                   $700,000:  138   2012-12-20:  222
 Mean   :1681   Mean   :1.705            B1     : 2070                   $650,000:  129   2013-03-22:  204
 3rd Qu.:1950   3rd Qu.:2.000            B3     : 1230                   $300,000:  120   2012-12-31:  179
 Max.   :2013   Max.   :4.000            B2     : 1115                   $600,000:  120   2012-12-19:  178
                                         (Other):11839                   (Other) :13834   (Other)   :21670
>
```

We can view the summary for the data for the brooklyn borough using the summary function.We can see the sales that occurred across different neighborhoods, different categories of buildings like one family homes or condos, basically across all the fields.
We can also observe some outliers in the sale.price summary,i.e., the sale price which were for 0$ or 10$ etc.

We can observe outliers in the land.square.feet and gross.square.feet
where values is 0 as well as year.built where minimum year is 0. Similarly, exclude places having no zipcodes.

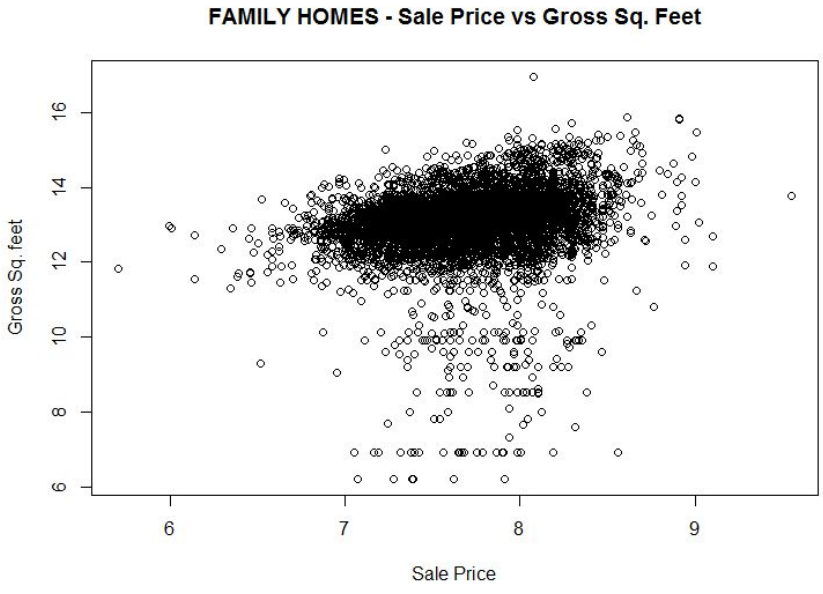
PLOTS

1)



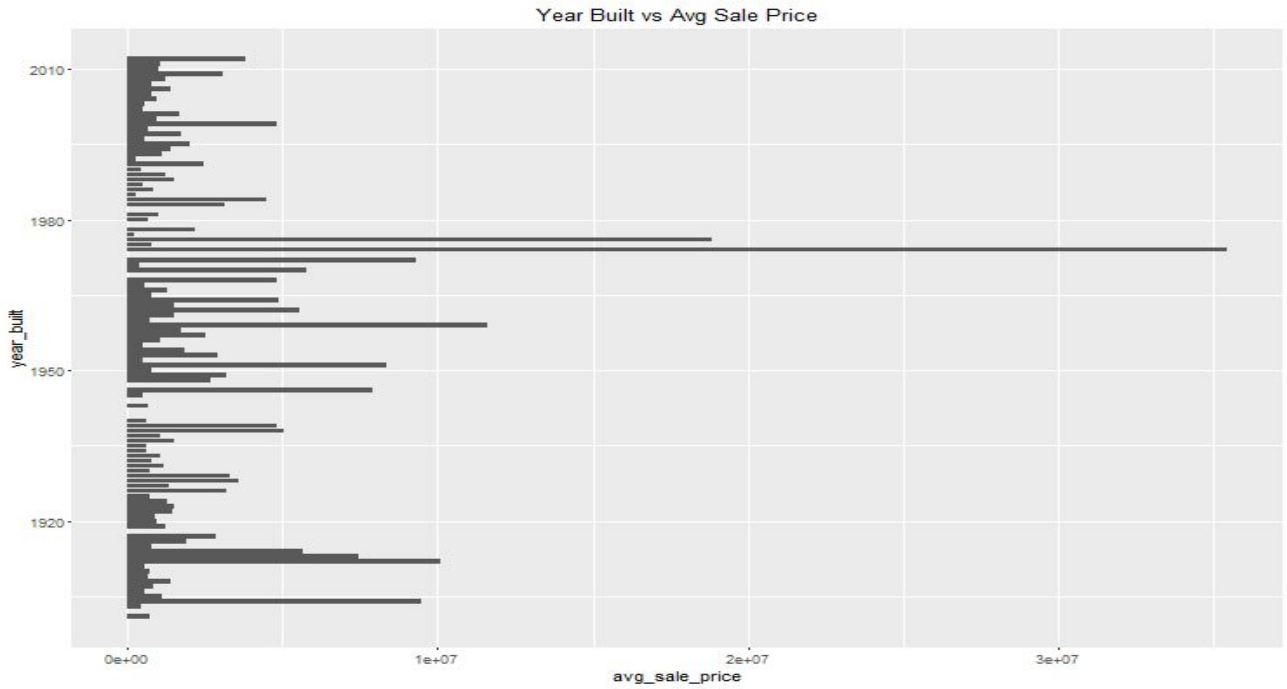
This is the graph of log(sale.price) vs log(gross.sq.feet). In general, we can say that as gross square feet increases the sale price increases.

2)



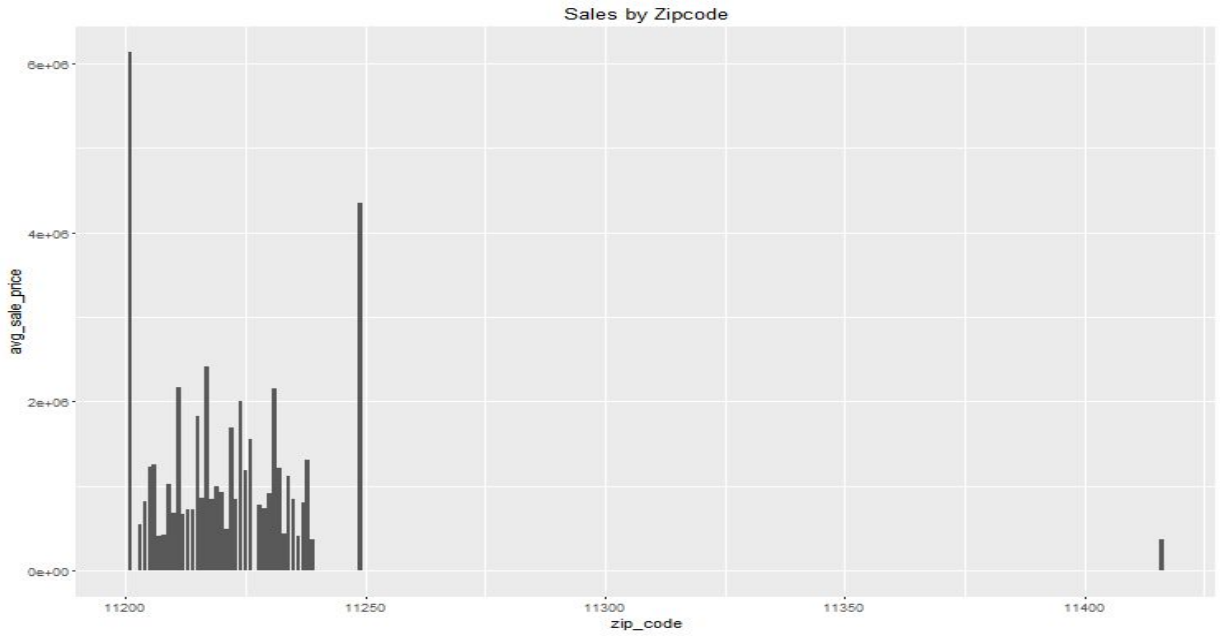
This is the graph of log(sale.price) vs log(gross.sq.feet) for family homes only. It is following the general property mentioned above.

3)



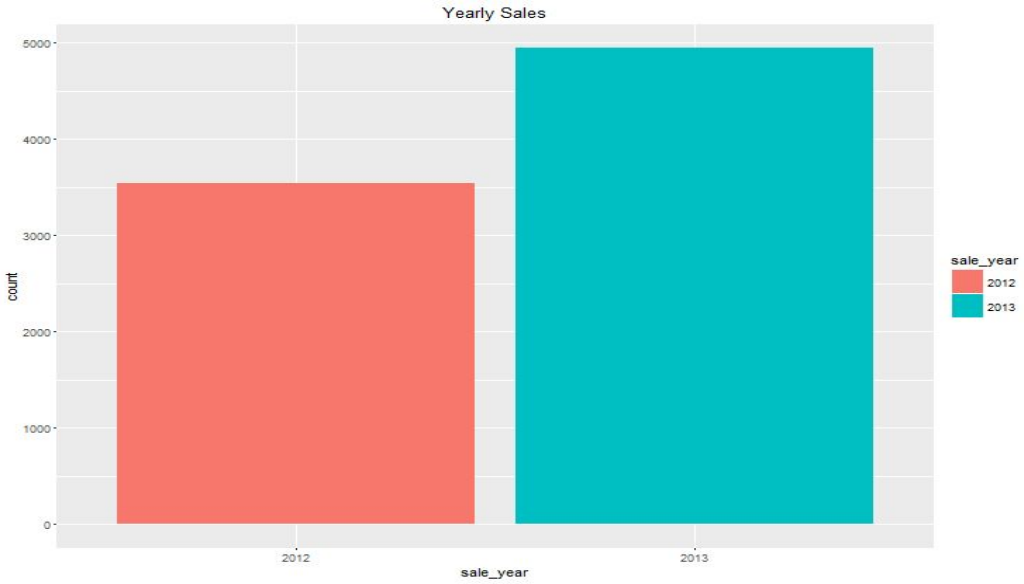
According to the graph above the houses built in 1970-1980 period had the highest sale price.

4)



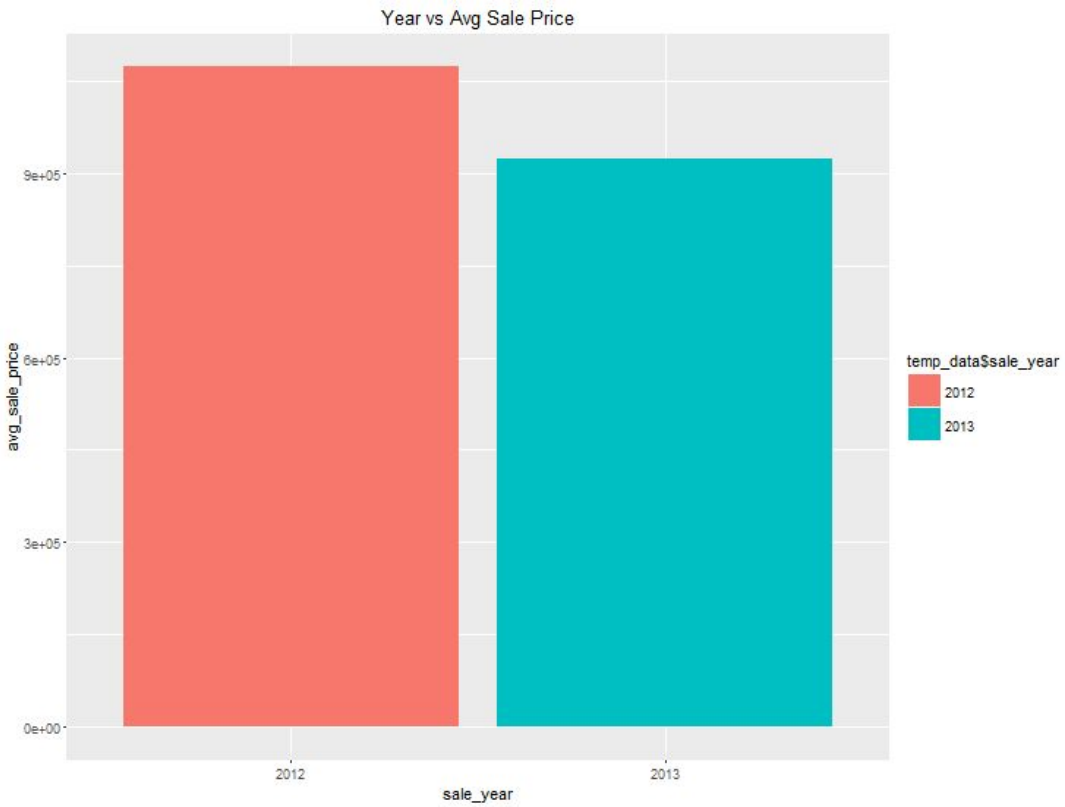
According to the graph above, places at the zipcode 11201 had the highest avg sale prices

6)



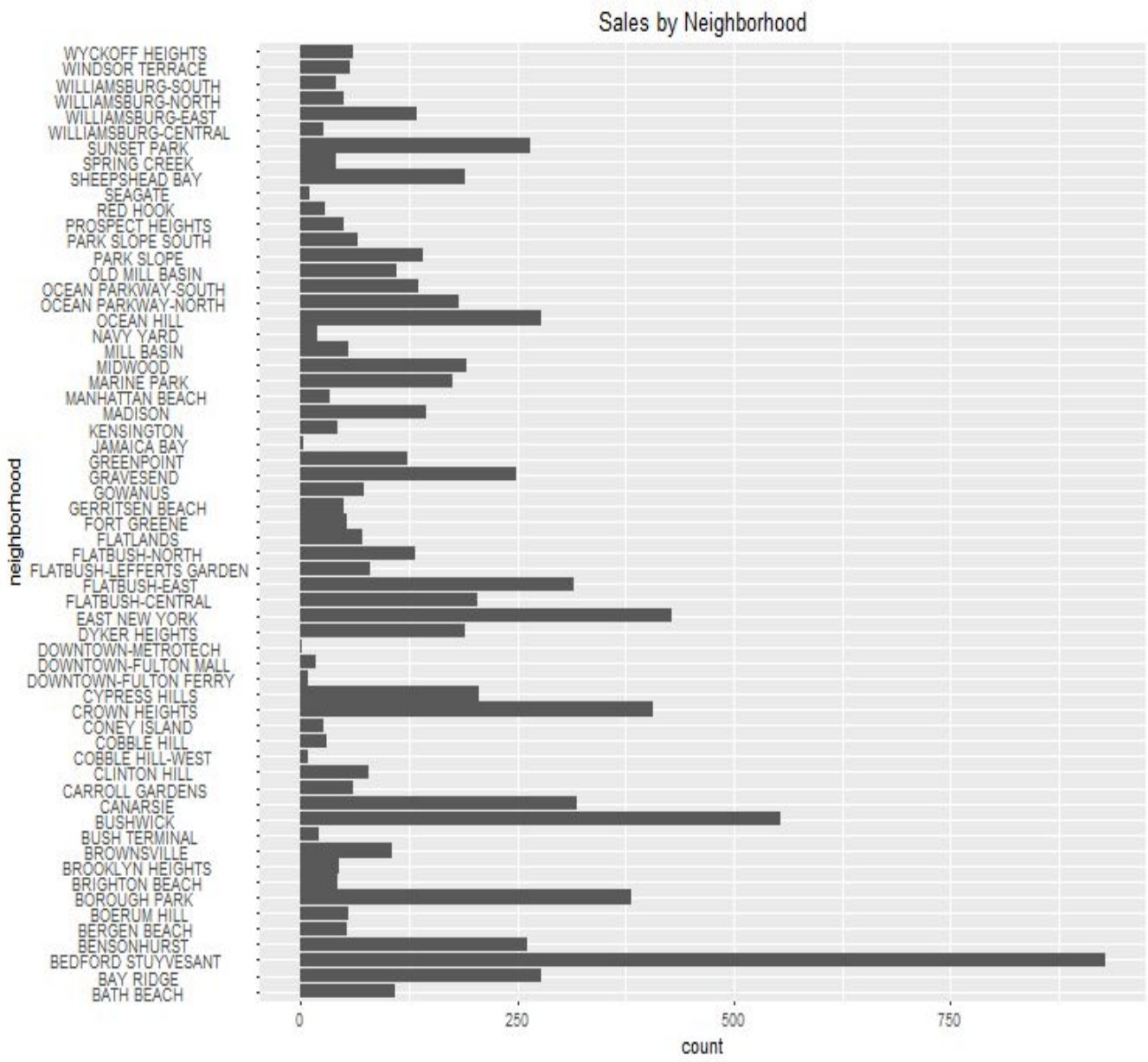
Number of sales increased in 2013.

7)



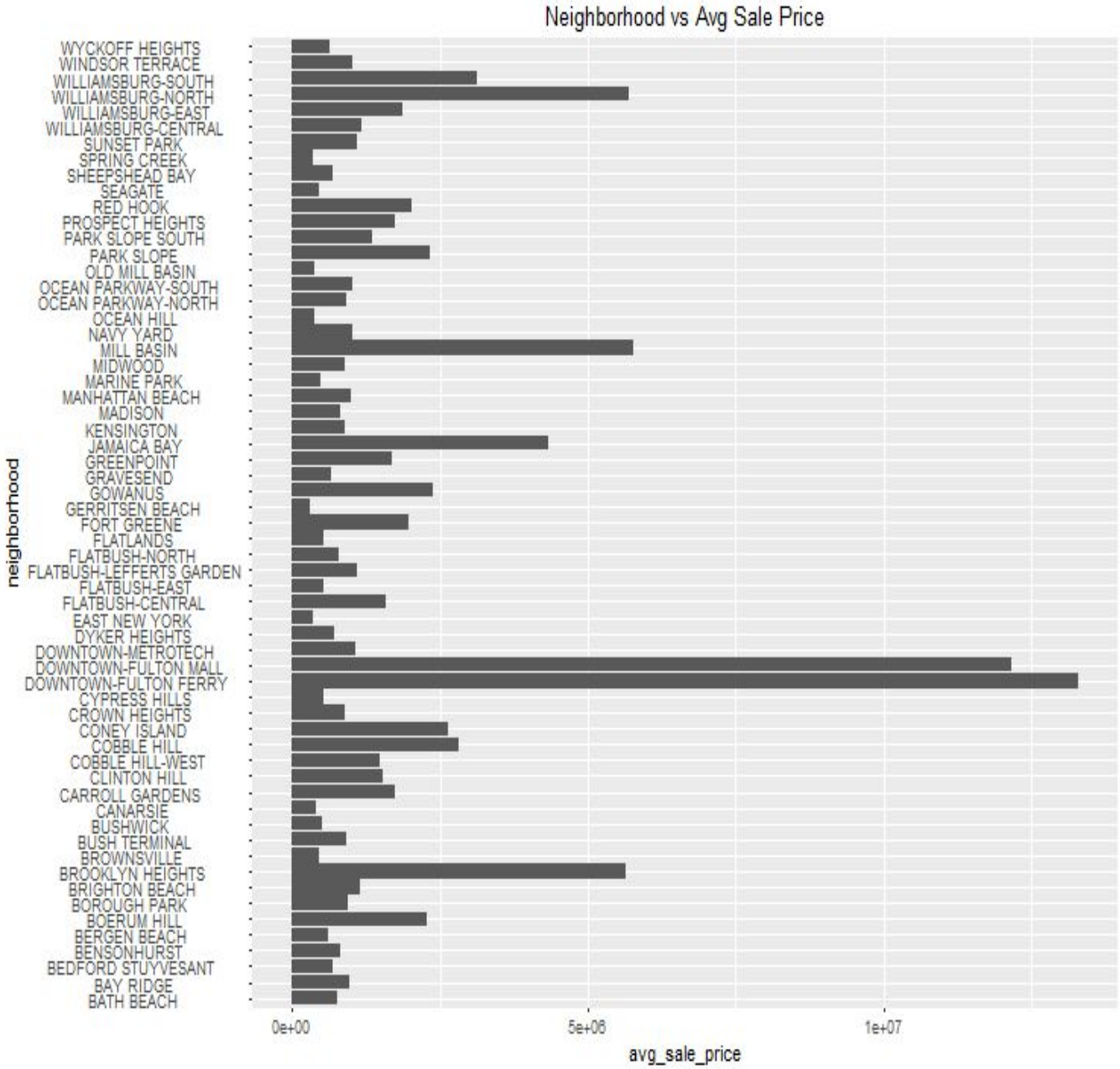
At the same time, the average sale price decreased in 2013.

5)



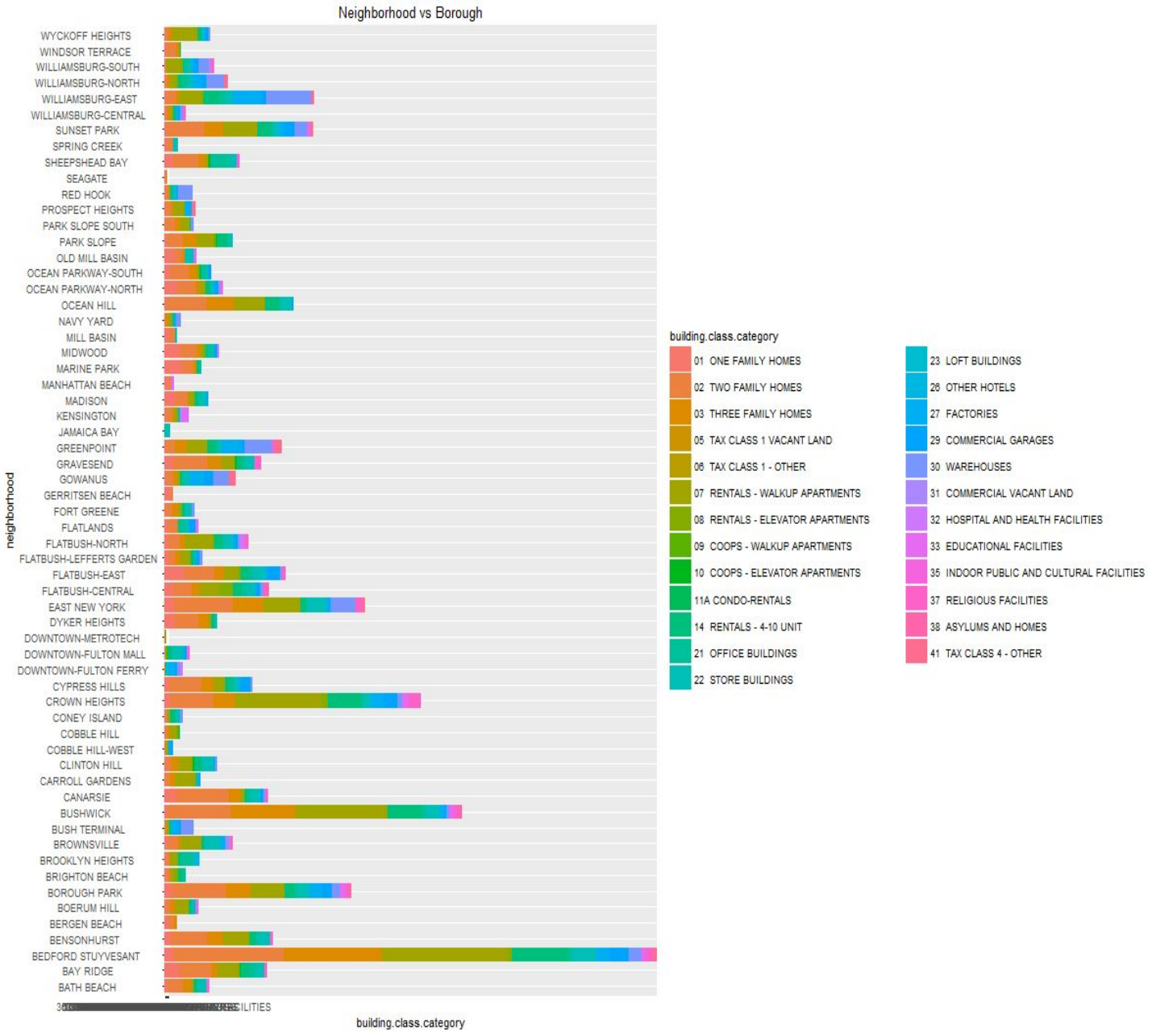
We can see the sales that happened across all the neighborhoods in Brooklyn in the graph. Highest sales took place in the neighborhood of Bedford Stuyvesant (because of low average sale price as shown in next graph) whereas min sales occured at Downtown-Metrotech.

8)



Downtown Fulton Mall had the highest avg sale price. Hence, very low sales happened in that neighborhood as can be seen in the graph above. We can see that neighborhoods having high avg sale price had lower sales.

9)



Bedford Stuyvesant has every type of building in the neighborhood. That is probably why highest sales took place in that neighborhood, whereas Downtown Metrotech only has one class, I think vacant land, because of which its price is one of the highest.

ANALYSIS FOR A MULTIPLE BOROUGHS

Here we take the data for the Brooklyn,Bronx,Staten Island,Manhattan and Queens Boroughs.

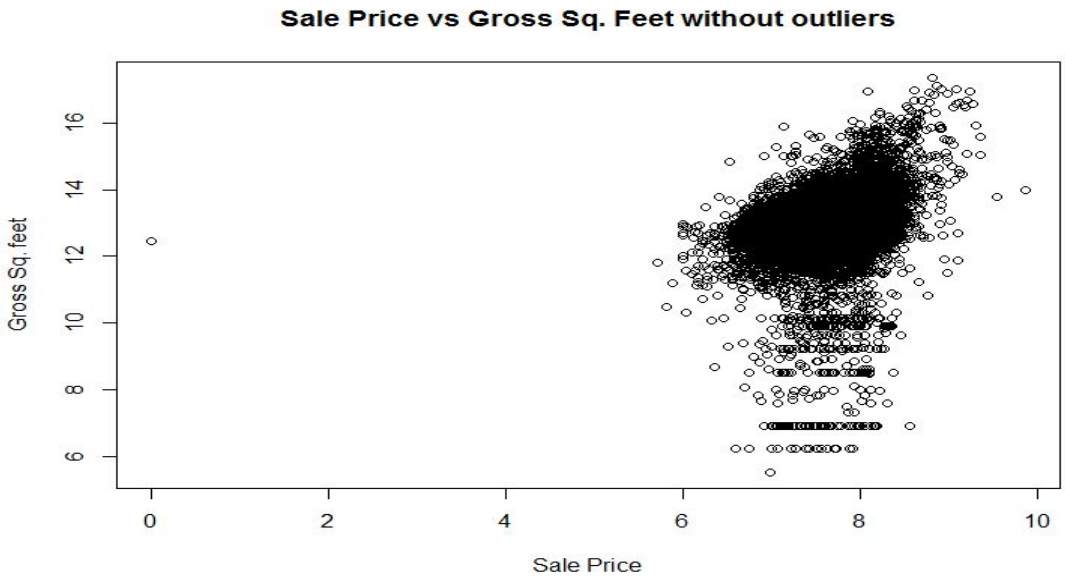
SUMMARY

```
> summary(bk)
    BOROUGH                     NEIGHBORHOOD                                BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT
 Min.   :1.000   MIDTOWN WEST            : 6264   01  ONE FAMILY HOMES                    :14846     2      :32899
 1st Qu.:1.000   FLUSHING-NORTH          : 2612   10  COOPS - ELEVATOR APARTMENTS          :13771     1      :32863
 Median :3.000   UPPER EAST SIDE (59-79)  : 2569   02  TWO FAMILY HOMES                     :13678     4      :11937
 Mean   :2.724   UPPER EAST SIDE (79-96)  : 2117   13  CONDOS - ELEVATOR APARTMENTS         :13313     2A     : 2667
 3rd Qu.:4.000   UPPER WEST SIDE (59-79)  : 2053                                            : 4204     2C     : 1905
 Max.   :5.000   BEDFORD STUYVESANT       : 1699   03  THREE FAMILY HOMES                   : 4135     1B     : 1221
                 (Other)                 :68661   (Other)                                  :22028     (Other): 2483
     BLOCK            LOT          EASE.MENT         BUILDING.CLASS.AT.PRESENT                                ADDRESS
 Min.   :    1   Min.   :   1.0   Length:85975       D4     :13461            870 7 AVENUE                          : 2087
 1st Qu.: 1052   1st Qu.:  23.0   Class :character   R4     :13171            102 WEST 57TH STREET                  : 1322
 Median : 2157   Median :  51.0   Mode  :character   A1     : 5469            200 WEST 56TH   STREET                :  608
 Mean   : 3661   Mean   : 405.4                     B1     : 4248            1335 AVENUE OF THE AMERIC             :  405
 3rd Qu.: 5599   3rd Qu.:1009.0                     A5     : 4226            102 WEST 57TH ST                      :  262
 Max.   :16323   Max.   :9117.0                     C0     : 4143            1335 AVENUE OF THE AMER               :  191
                                                    (Other):41257            (Other)                               :81100
   APART.MENT.NUMBER    ZIP.CODE     RESIDENTIAL.UNITS  COMMERCIAL.UNITS   TOTAL.UNITS       LAND.SQUARE.FEET GROSS.SQUARE.FEET
           :63830   Min.   :    0   Length:85975       Length:85975       Length:85975      0      :40340    0      :42780
 TIMES     :  599   1st Qu.:10028   Class :character   Class :character   Class :character  2,000  : 3417    112,850: 1634
 4         :  273   Median :11201   Mode  :character   Mode  :character   Mode  :character  2,500  : 3130    2,400  :  349
 3B        :  249   Mean   :10758                                                           4,000  : 2688    3,000  :  346
 3A        :  244   3rd Qu.:11238                                                           7,532  : 1635    1,800  :  293
 2         :  235   Max.   :11694                                                           1,800  : 1113    2,000  :  265
 (Other)   :20545                                                                         (Other):33652    (Other):40308
   YEAR.BUILT   TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE    SALE.PRICE        SALE.DATE
 Min.   :   0   Min.   :1.000            D4     :13461                  $0      :28638   2012-09-27:  988
 1st Qu.:1910   1st Qu.:1.000            R4     :13313                  $10     :  758   2012-12-20:  860
 Median :1931   Median :2.000            A1     : 5467                  $350,000:  394   2012-12-27:  854
 Mean   :1681   Mean   :1.868            B1     : 4238                  $400,000:  377   2012-12-17:  822
 3rd Qu.:1964   3rd Qu.:2.000            A5     : 4228                  $300,000:  370   2012-12-28:  737
 Max.   :2013   Max.   :4.000            C0     : 4135                  $450,000:  366   2012-12-21:  721
                NA's   :1                (Other):41133                  (Other) :55072   (Other)   :80993
>
```

Similar to the summary for a single borough, we can now observe all metrics and outliers like sale.price, year.built,zipcode etc. for multiple boroughs.

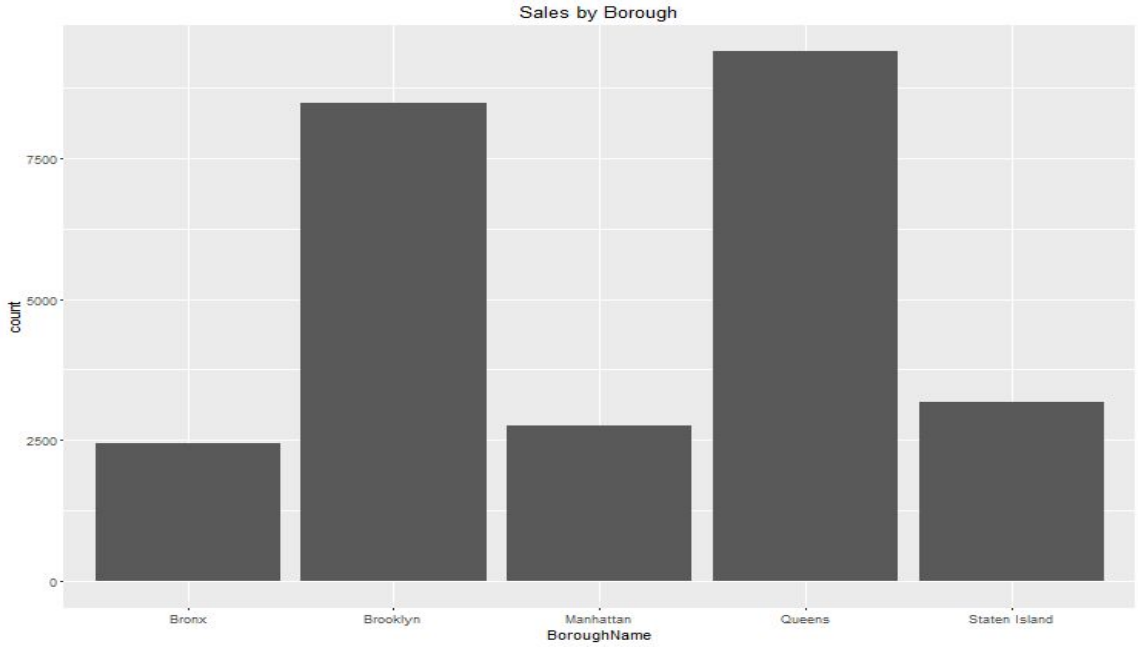
PLOTS

1)



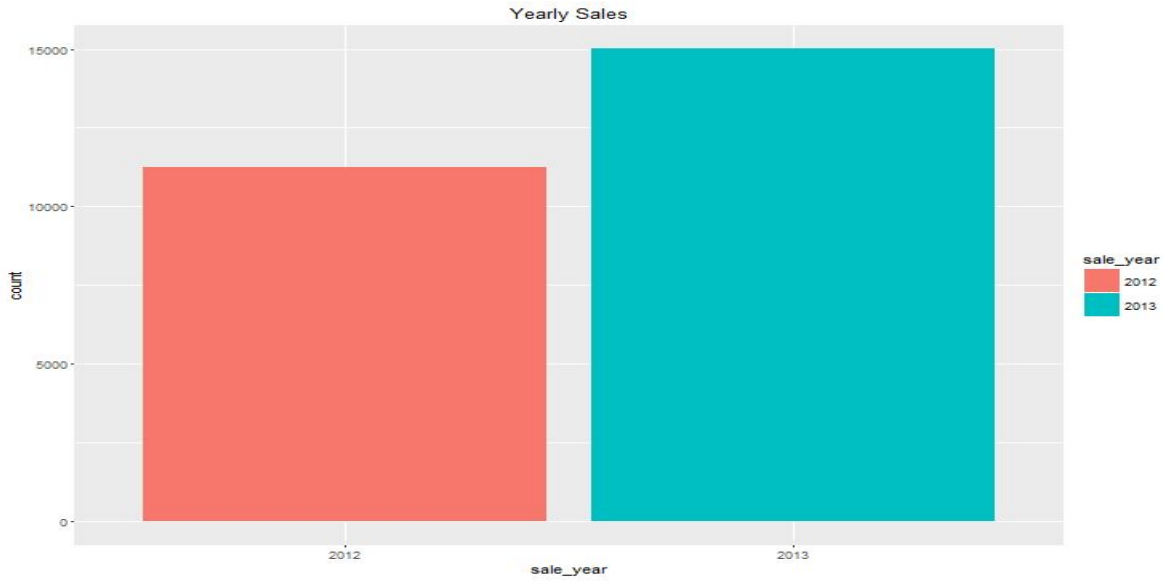
After seeing data for all boroughs, now we can see that the average sale price has increased for gross square feet.

2)



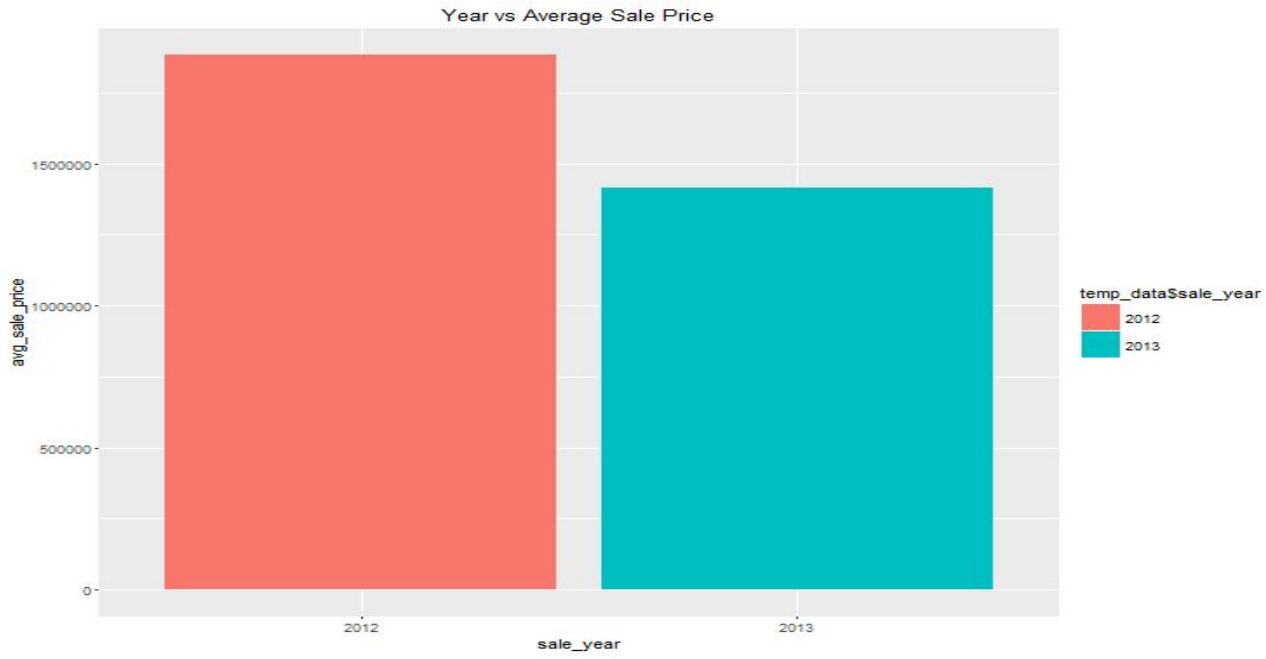
Queens had the maximum sales while Bronx had the minimum

3)



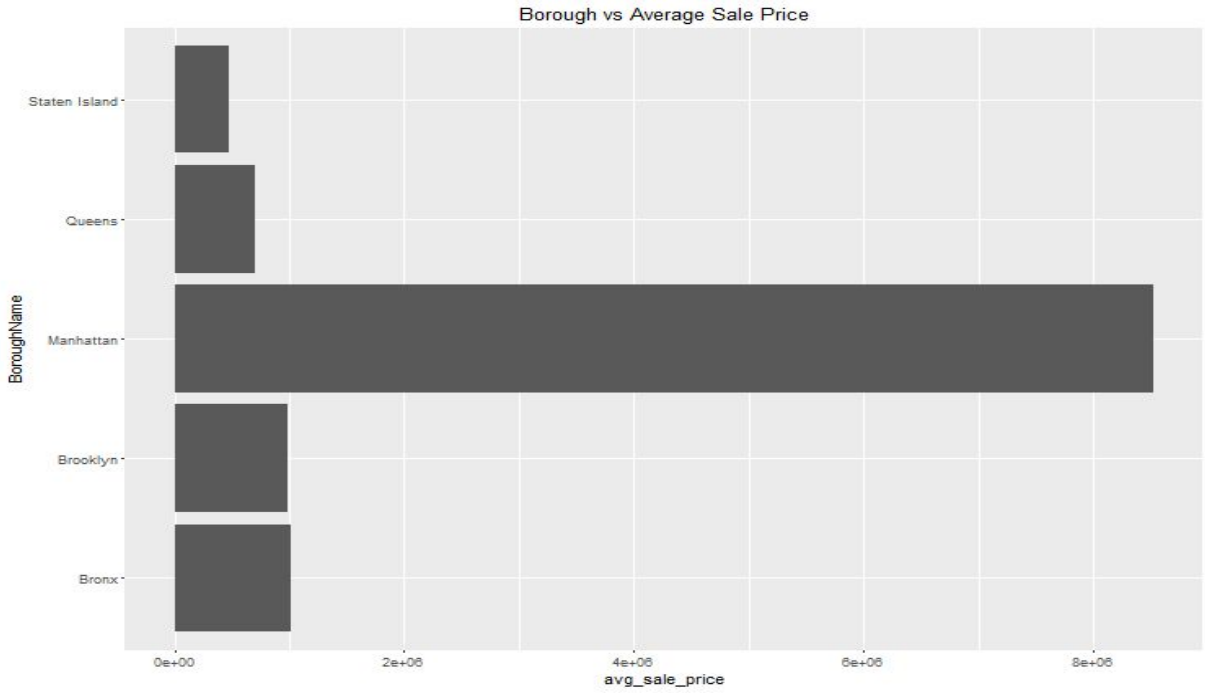
More sales happened in 2013.

4)



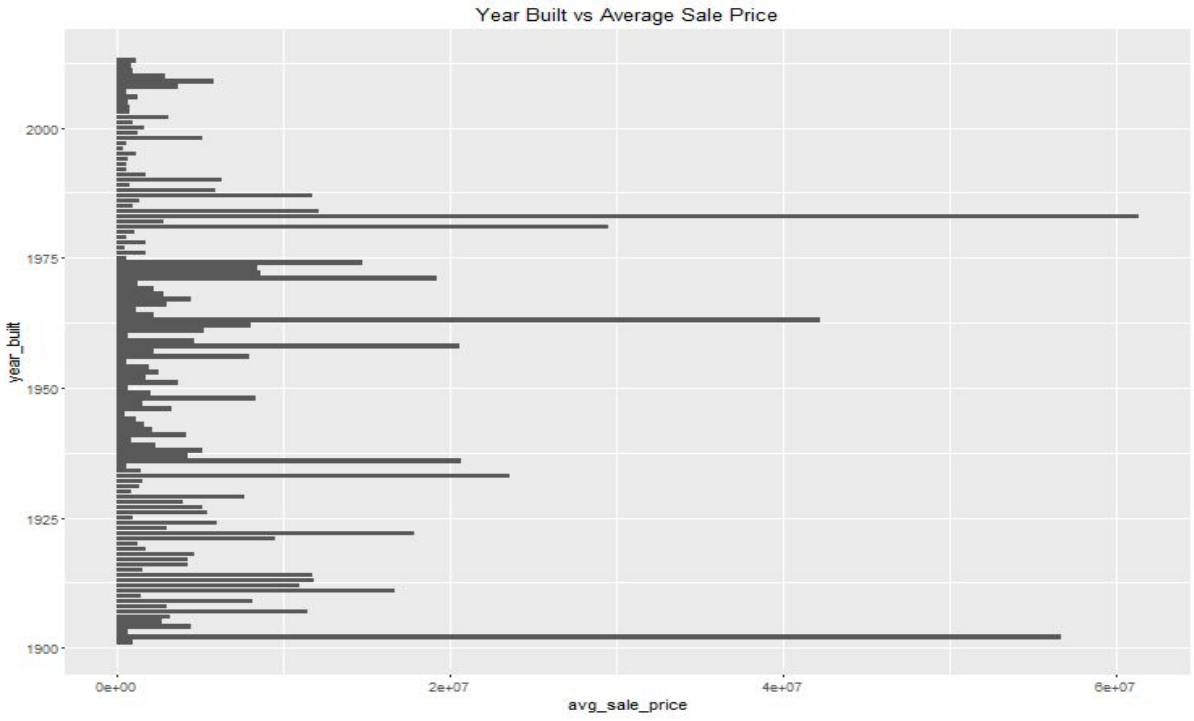
Average Sale price decreased in 2013 even after considering all boroughs data.

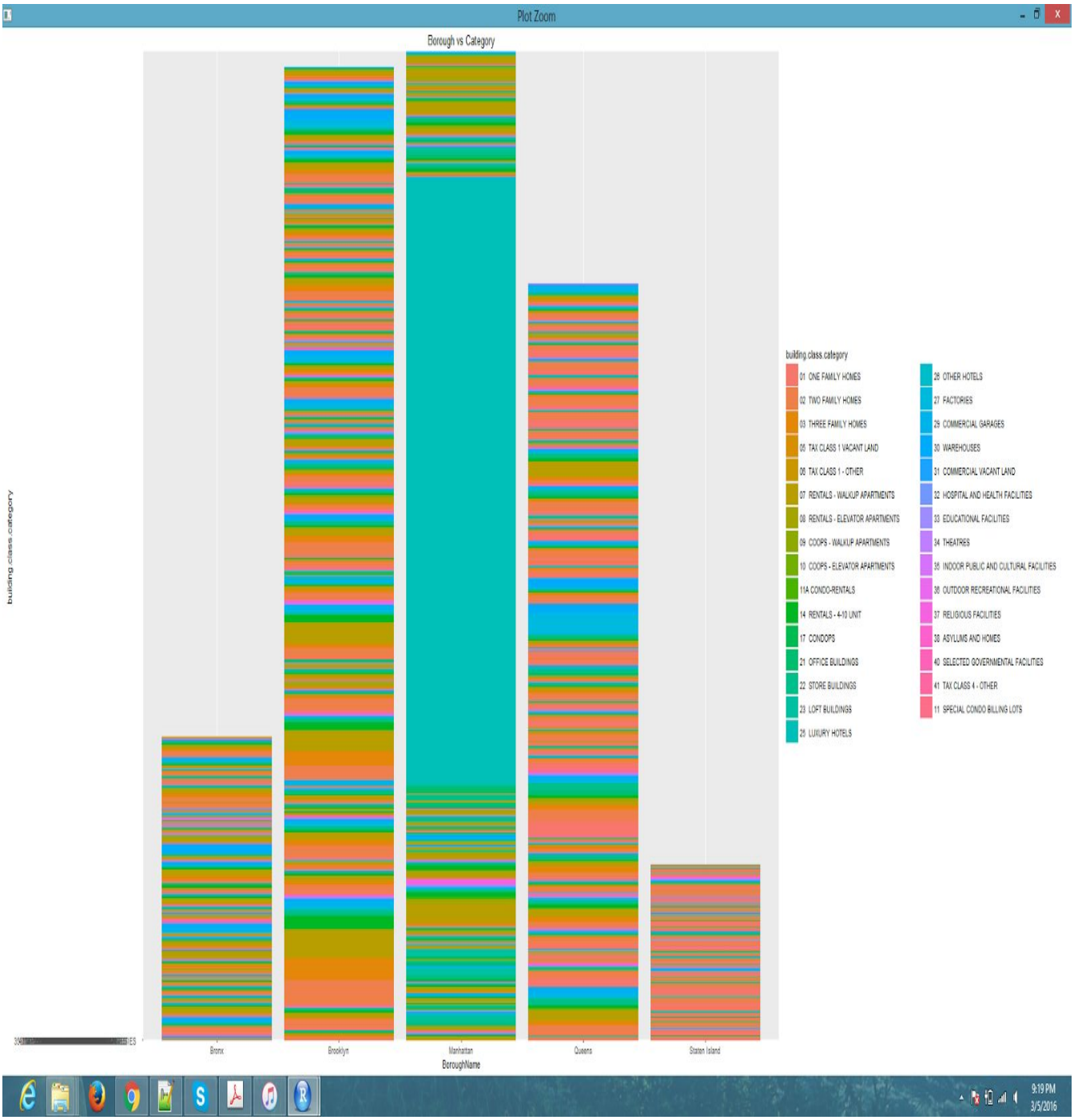
5)



Manhattan had the highest average sales price while Staten Island had the least.

6)





In this Boroughs vs Building Category graph, we can observe that Manhattan has the highest no of hotels as property sales

CONCLUSION

Thus, we have analyzed and identified interesting facts in the Real Direct Dataset