Problem 3: Case Study

Problem Statement

In this Problem we extend need to perform EDA on real direct data set. First we need to do analysis of realdirect data of Brooklyn only. Then we need to extend our analysis to 5 different boroughs.

Data Set

In this problem we are going to use RealDirect's data. RealDirect is a real estate company. The dataset contains following columns

(Source: http://www1.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf):

Borough:

The name of the borough in which the property is located.

Neighborhood:

Department of Finance assessors determine the neighborhood name in the course of valuing properties. The common name of the neighborhood is generally the same as the name Finance designates. However, there may be slight differences in neighborhood boundary lines and some sub-neighborhoods may not be included.

Building Class Category:

This is a field that we are including so that users of the Rolling Sales Files can easily identify similar properties by broad usage (e.g. One Family Homes) without looking up individual Building Classes. Files are sorted by Borough, Neighborhood, Building Class Category, Block and Lot.

Tax Class at Present:

Every property in the city is assigned to one of four tax classes (Classes 1, 2, 3, and 4), based on the use of the property.

- Class 1: Includes most residential property of up to three units (such as one-, two-, and three-family homes and small stores or offices with one or two attached apartments), vacant

land that is zoned for residential use, and most condominiums that are not more than three stories.

- Class 2: Includes all other property that is primarily residential, such as cooperatives and condominiums.

- Class 3: Includes property with equipment owned by a gas, telephone or electric company.

- Class 4: Includes all other properties not included in class 1,2, and 3, such as offices, factories, warehouses, garage buildings, etc.

Block:

A Tax Block is a sub-division of the borough on which real properties are located.
The Department of Finance uses a Borough-Block-Lot classification to label all real property in the City. “Whereas” addresses describe the street location of a property, the block and lot distinguishes one unit of real property from another, such as the different condominiums in a single building. Also, block and lots are not subject to name changes based on which side of the parcel the building puts its entrance on.

Lot:

A Tax Lot is a subdivision of a Tax Block and represents the property unique location.

Easement:

An easement is a right, such as a right of way, which allows an entity to make limited use of another’s real property. For example: MTA railroad tracks that run across a portion of another property.

Building Class at Present:

The Building Classification is used to describe a property’s constructive use. The first position of the Building Class is a letter that is used to describe a general class of properties (for example “A” signifies one-family homes, “O” signifies office buildings. “R” signifies condominiums). The second position, a number, adds more specific information about the property’s use or construction style (using our previous examples “A0” is a Cape Cod style one family home, “O4” is a tower type office building and “R5” is a commercial condominium unit). The term Building Class used by the Department of Finance is interchangeable with the term Building Code used by the Department of Buildings. **See NYC Building Classifications**.

Address: The street address of the property as listed on the Sales File. Coop sales include the apartment number in the address field.

Zip Code: The property's postal code **Residential Units:**

The number of residential units at the listed property.

Commercial Units:

The number of commercial units at the listed property.

Total Units:

The total number of units at the listed property.

Land Square Feet:

The land area of the property listed in square feet.

Gross Square Feet:

The total area of all the floors of a building as measured from the exterior surfaces of the outside walls of the building, including the land area and space within any building or structure on the property.

Year Built:

Year the structure on the property was built.

Building Class at Time of Sale:

The Building Classification is used to describe a property's constructive use. The first position of the Building Class is a letter that is used to describe a general class of properties (for example "A" signifies one-family homes, "O" signifies office buildings. "R" signifies condominiums). The second position, a number, adds more specific information about the property's use or construction style (using our previous examples "A0" is a Cape Cod style one family home, "O4" is a tower type office building and "R5" is a commercial condominium unit). The term Building Class as used by the Department of Finance is interchangeable with the term Building Code as used by the Department of Buildings.

Sales Price:

Price paid for the property.

Sale Date:

Date the property sold.

$0 Sales Price:

A $0 sale indicates that there was a transfer of ownership without a cash consideration. There can be a number of reasons for a $0 sale including transfers of ownership from parents to children.

Derived Columns

The given information is generally not sufficient to analyze and interpret the data set. So we derive few more columns to ease our analysis:

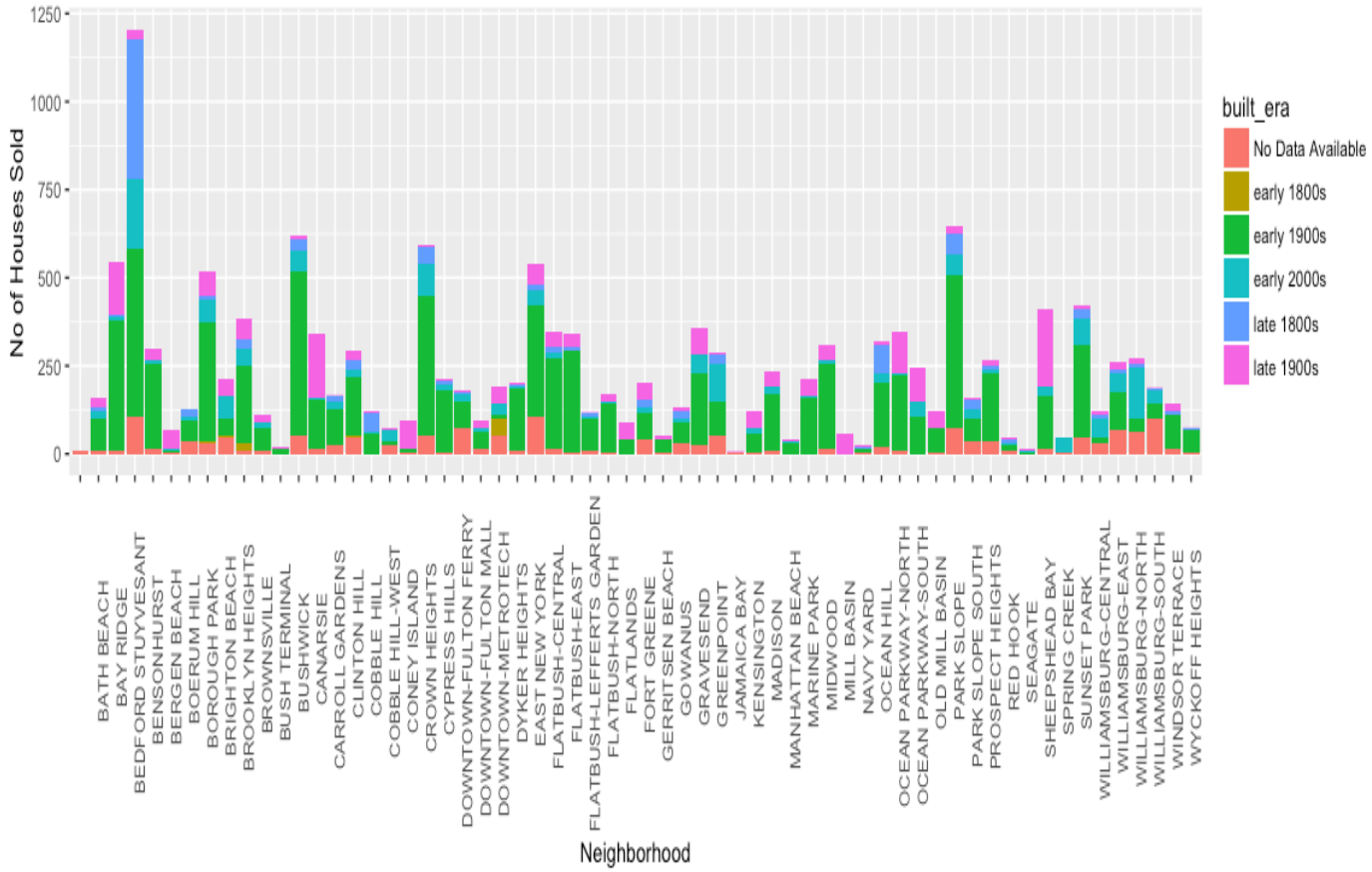
1. gross_sqft : This is numeric version of Gross Square Feet
2. sale_price_n : This is numeric version of Sale Price
3. sale_date : This is Date version of sale date
4. sale_year: extracted year of built from sale date in numeric format.
5. year_built : This is numeric version of year the property was built
6. built_era : This column represent the period in which the house was built. I defined following periods
    a. early 1800s
    b. late 1800s
    c. early 1900s
    d. late 1900s
    e. early 2000s

Cleaning Data

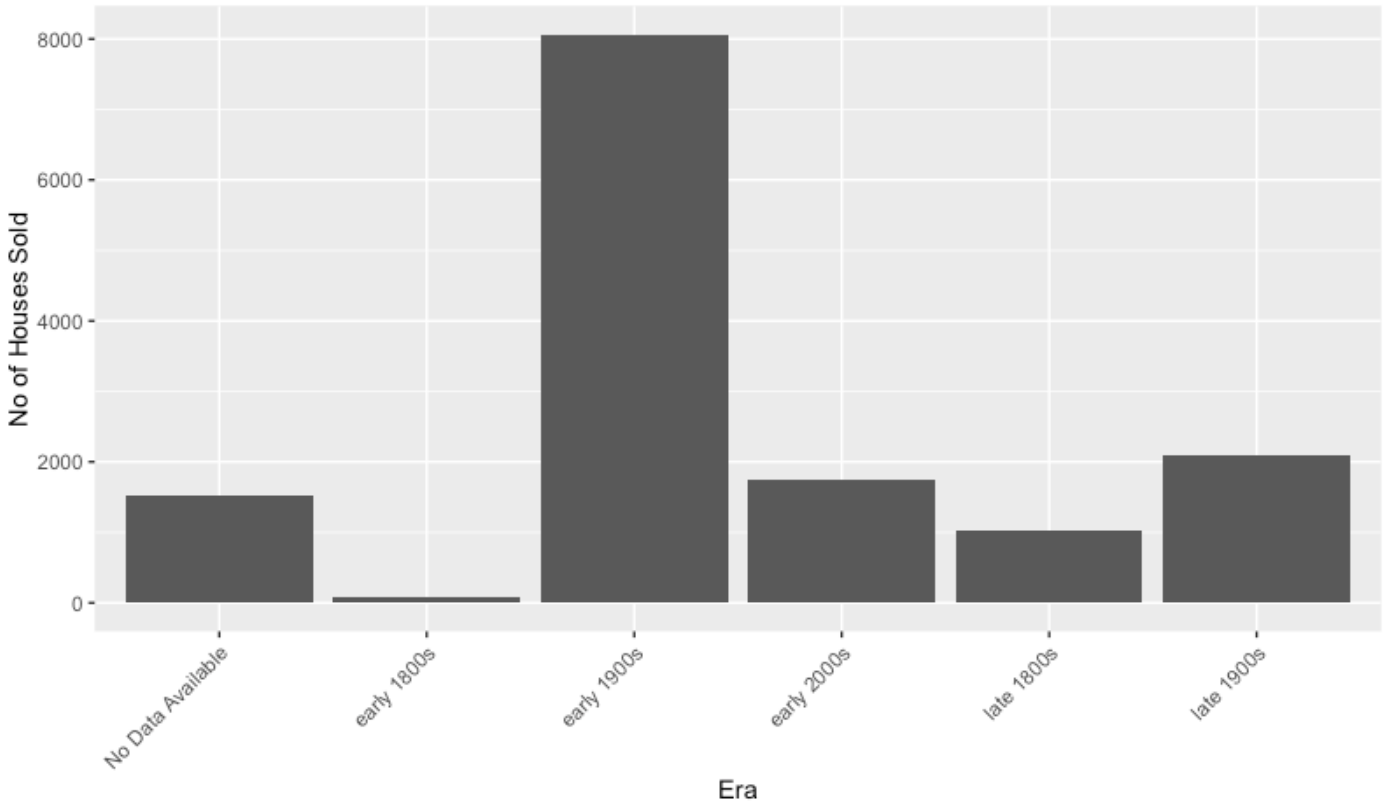
On giving a cursory look, we can identify that the data is not consistent. So we need to remove the samples that do not make sense. Following are the inconsistencies I found in data:

1. Zero Sale Price: If the sale price is zero we can assume that either the data point is wrong or no sale event happened. In both the cases, we need not consider this data point.
2. Sold Before Built: Some data points shows that the sale year is before built year. As the house need to be built to be sold, these data points do nit make sense.
3. Sale Year zero: There are data points with sale year zero. Again either they are incorrectly noted data points or no sale ever happened. In both the cases we need to remove this point.

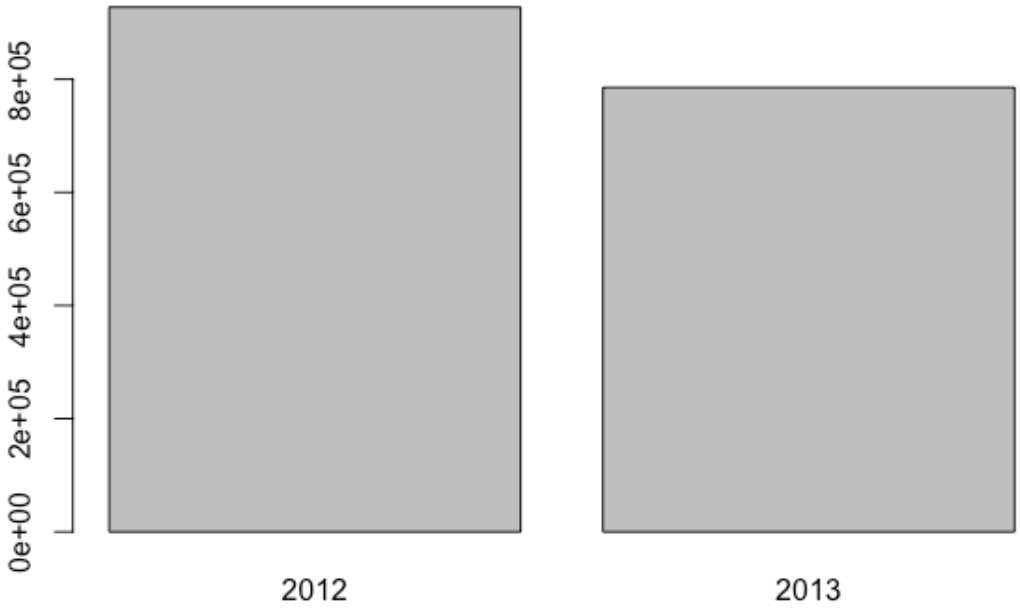
Analysis of Brooklyn Data



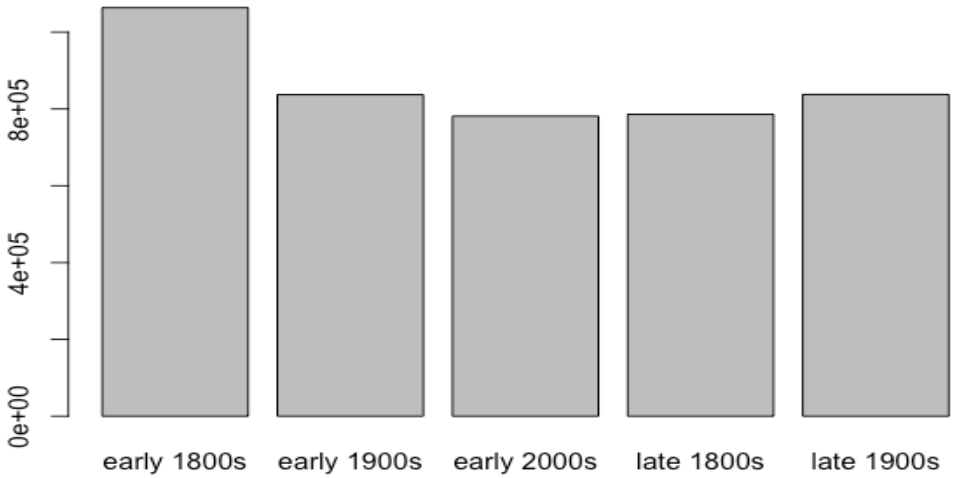
This graph gives us a good representation of how popular the neighborhoods are. Like we can see Bay Rigide is hot most neighborhood on sale. We can also see the division of timeperiod in which the graph was built. We can see that highest number of houses were built in early 1900s.



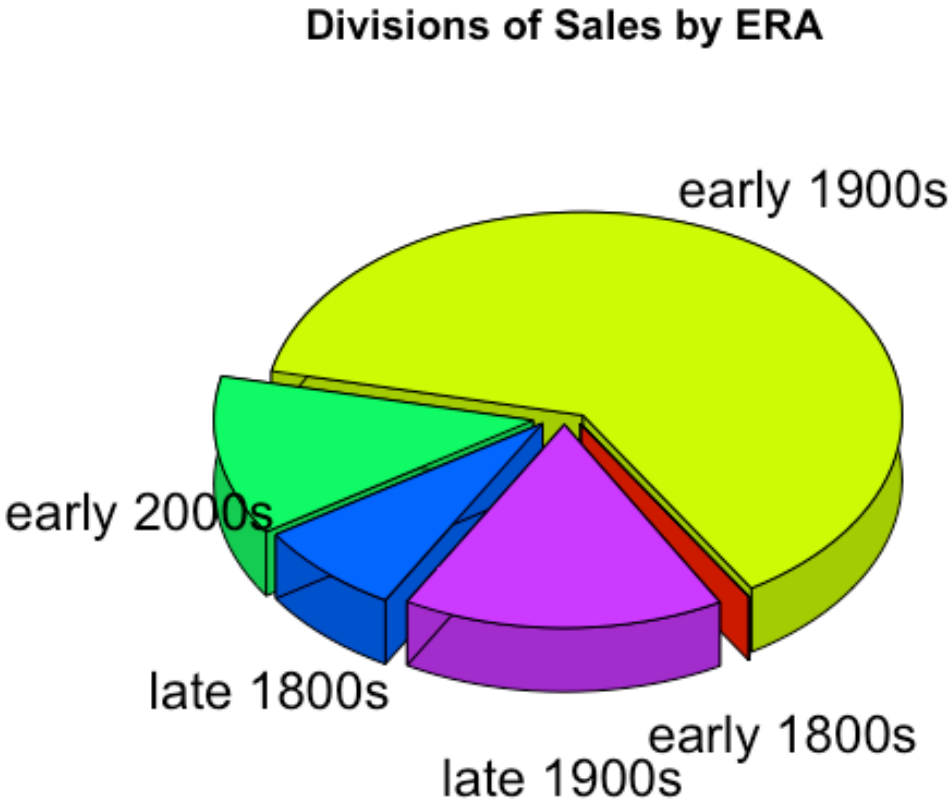
This graph is more clear representation of number of houses built per era. Again here we can see that maximum number of graphs were plot in early 1900s



The data we have corresponds to sales in only 2012 and 2013. This graphs shows the relative amount of sales made in each year.

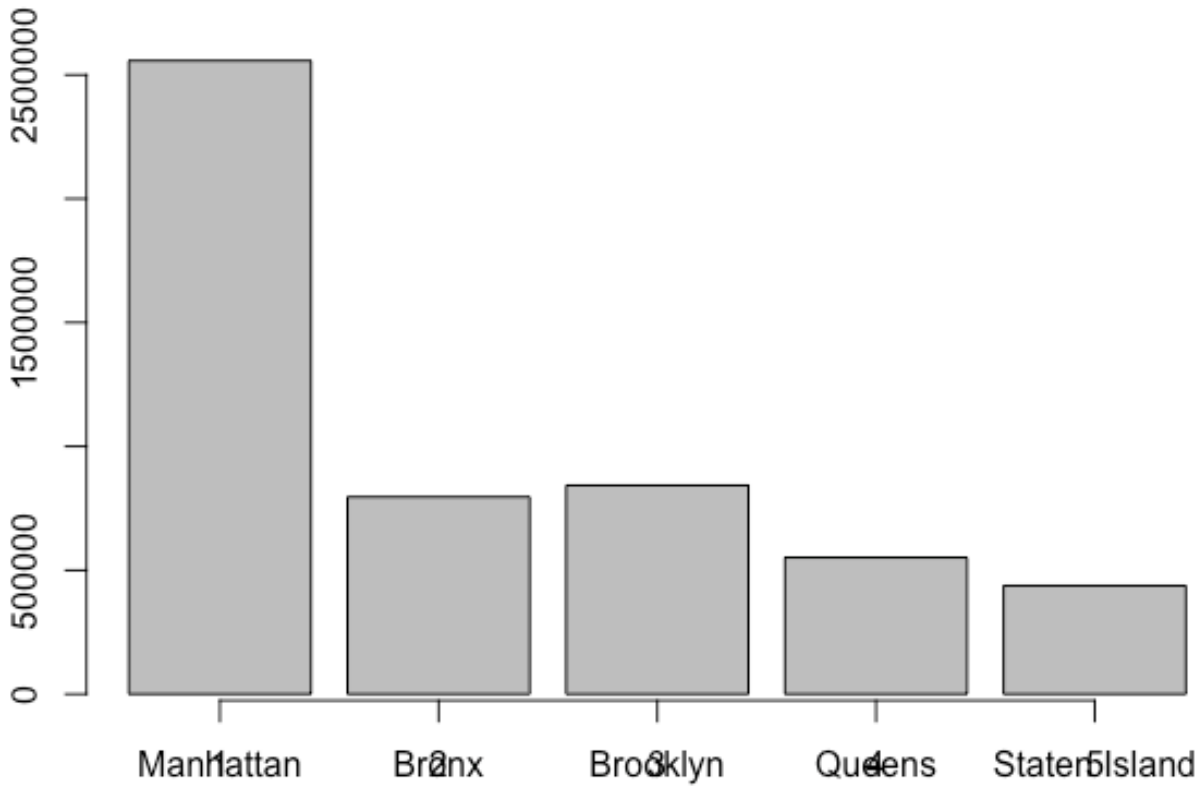


This graph compares the sales of year 2012 based on the era they were built in.

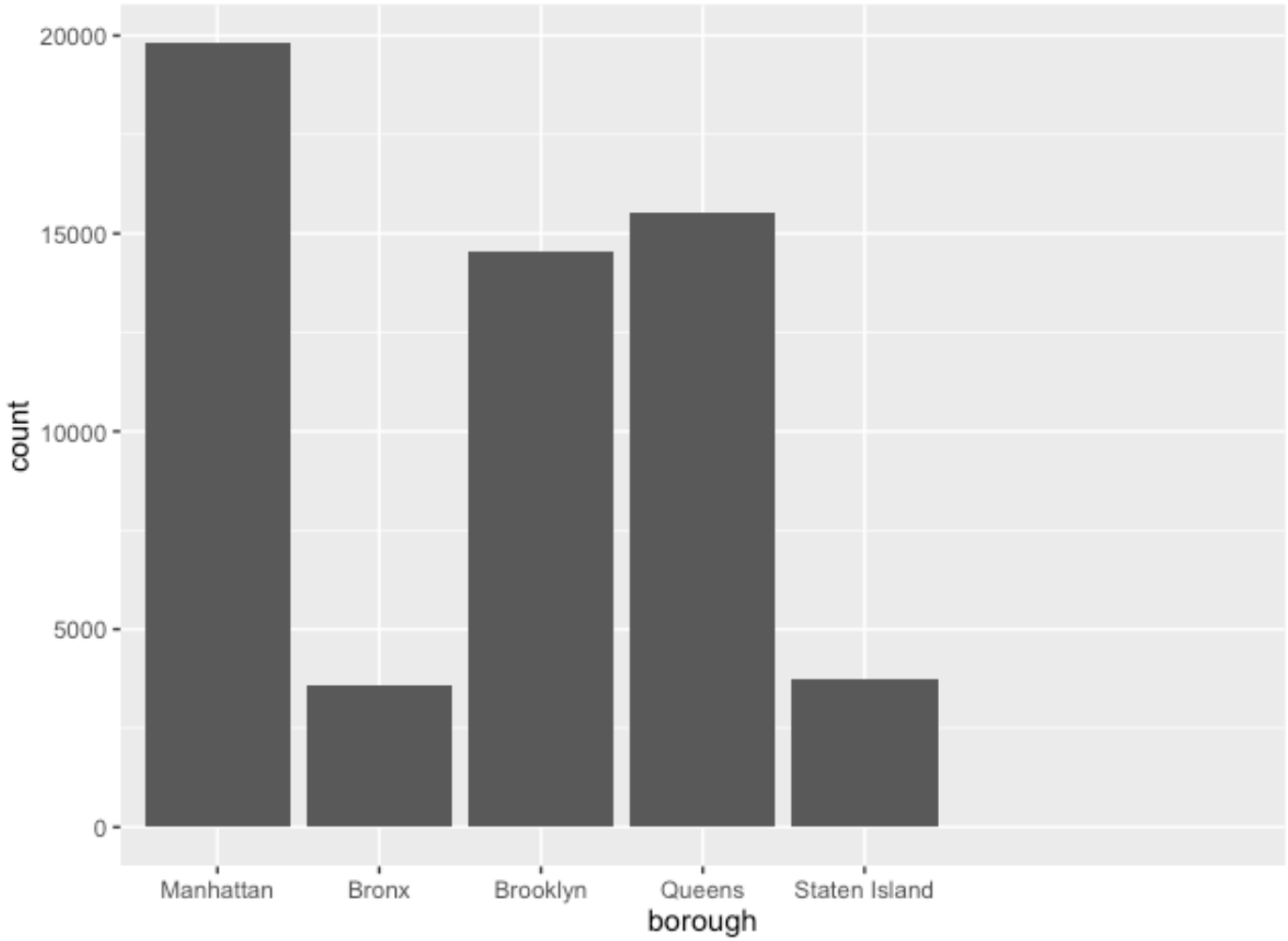


This pie chart shows the division of sales based on built year. It is evident that most of the sales recently were built in early 1900s

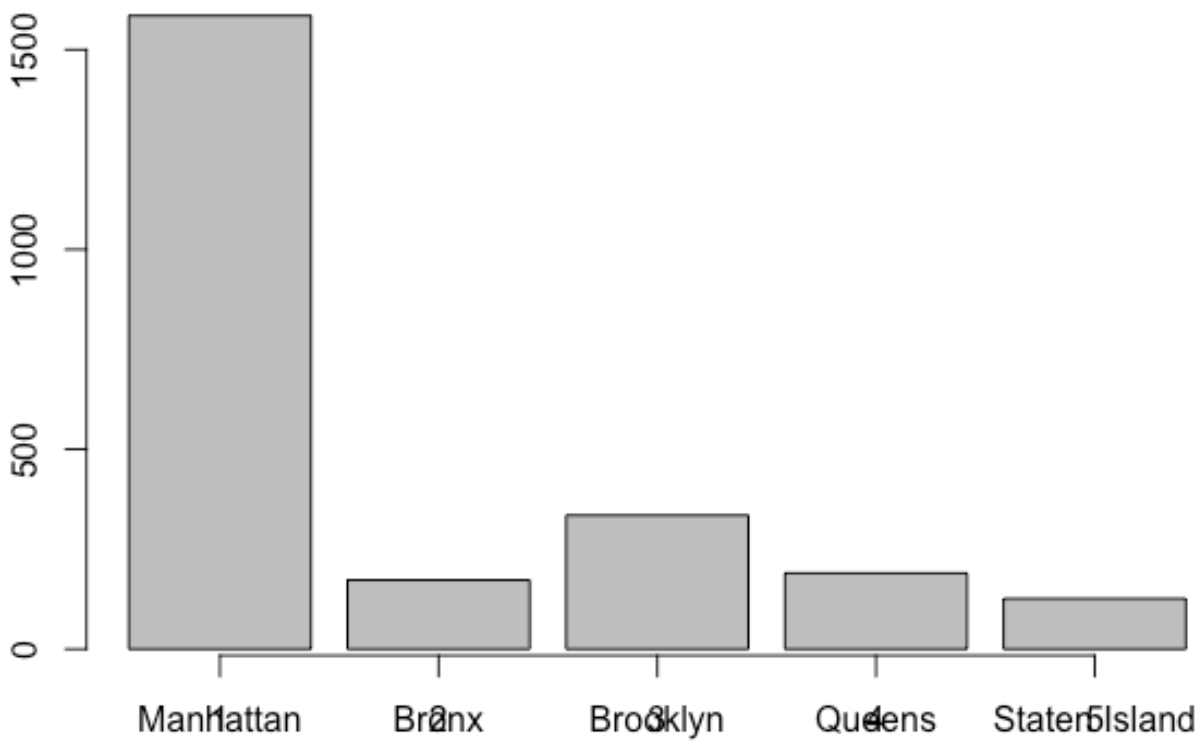
Analysis Across Borough



This bar graph shows the average sale price for each borough. We ca clearly see that Manhattan is costliest borough.



This graph shows total number of deals made per borough. We can see that again, the number of deals made in Manhattan is higher than any other borough. Though Brooklyn and Queens are not far behind.



This graph shows average rate of land, i.e. sale_price_n/land_sqft ,in different boroughs. As expected Manhattan has highest per square feet price of land.

Exercise: RealDirect Data Strategy

What data would you advise the engineers log and what would your ideal datasets look like?

I would advise the engineers to log data like facilities available in the building and and close by. Like availability of laundry, gym and Grocery Store in the property or nearby.

How would data be used for reporting and monitoring product usage?

We can record the number of clicks each user made on a given posting, or posting in a given area. We can also record number of times he filters property with given facility to identify how important these facilities are for a user or group of users.

How would data be built back into the product/website?

Using the logged data we can see what filters user is utilizing the most. Based on that data we can increase filter options related to the highly used filter option. Also we can perform surveys to know what users want the most and built it back in product.

Brief report to CEO

Based on the data from 2012 and 2013 sales we can conclude following:

- We should increase our business concentration in Manhattan area as the sales are highest in that area. Also the price is highest so out commission will also be higher.
- We should focus more on houses built in early 1900s as the older houses are more difficult to sell.

Being the "data scientist" often involves speaking to people who aren't also data scientists, so it would be ideal to have a set of communication strategies for getting to the information you need about the data. Can you think of any other people you should talk to?

To get a better information about the data we can talk to individual brokers who have first hand experience with selling property.

Does stepping out of your comfort zone and figuring out how you would go about "collecting data" in a different setting give you insight into how you do it in your own field?

Yes, collecting data in a different setting also gives us insight on how we can collect data in our field. Like in this example we understand that is is better to get data from those who are in the business from long duration, even though not on large scale, but they'll have good knowledge of small complexities which can give incorrect data.

Conclusion

From the above analysis we can conclude that most of the houses sold in 2012 and 2013 at Brooklyn were made in early 1900s. And from the second part of analysis we can show that Manhattan was costliest place to live in 2012 and 2013. Manhattan has highest average sale price and highest per square feet rate of land. Still Manhattan flaunts maximum number no housing deals in year 2012 and 2013.