CSE 587 – Data Intensive Computing

Problem 3 – Real Direct Case Study

Author: Bhavika Jain
bhavikap@buffalo.edu

Come up with a list of research questions you think could be answered by data:

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

It is important to understand what users do when they are browsing a particular website. In our case, it would be important to observe the actions of a given user when visiting the RealDirect website in order to understand what types of homes he is looking for, how much time he spends on looking at each of the listings (linger-time). By logging such data we can gain insights into what user wants when they visit RealDirect website.

2) ***How would data be used for reporting and monitoring product usage?.***

By logging data indirectly through clicks or site-navigation, we can provide recommendation to users better suited to their needs. This data will help us in generating listings specific to each user. We can categorize homes based on zip code, neighborhood, and grossSqft, build date, price etc. Using this we can determine which kind of houses a user is more interested in based on his listing view. By recording the linger-time, we could determine how interested the user was in a particular listing and use this to recommend homes later to similar users.

3) ***How would data be built back into the product/website?.***

The logged data (linger-time, listing views, etc.) can be used in personalized listing recommendation for each user. We have evaluate what information is more useful to user and what isn't.If we put too much information on the screen a user might get confused rather than providing any information needed by him.

Answers to question 2 and 3 are summarized below using plots and observations.

2) Real Direct Data Analysis:-

PART A – BROOKLYN BOROUGH

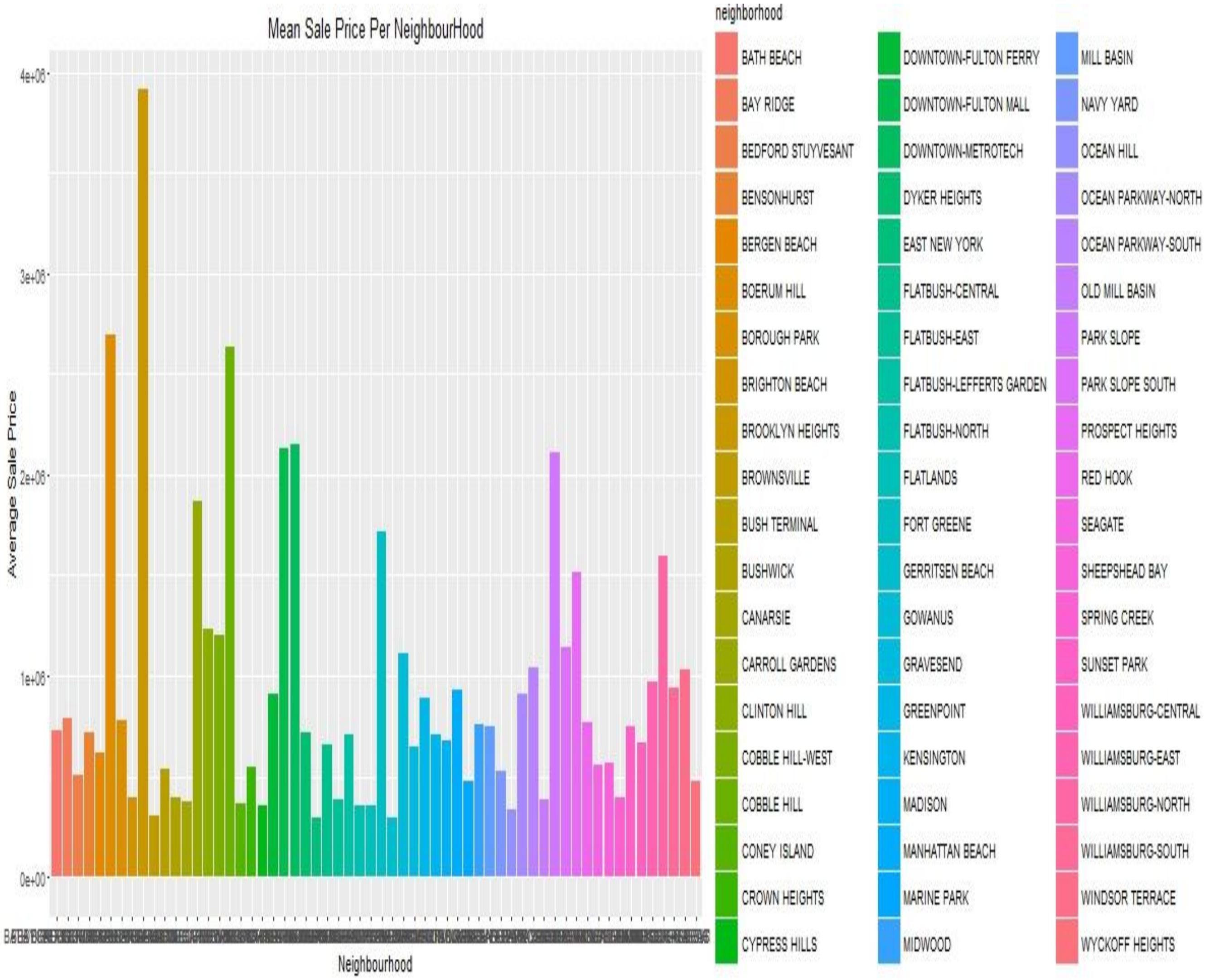
Objective: To do EDA on Brooklyn data set of Real Direct and find relevant patterns/observations.

Below observations and plots are made only for those homes that belong to One/Two/Three Family building class category.

1) Observations Based on Neighborhood

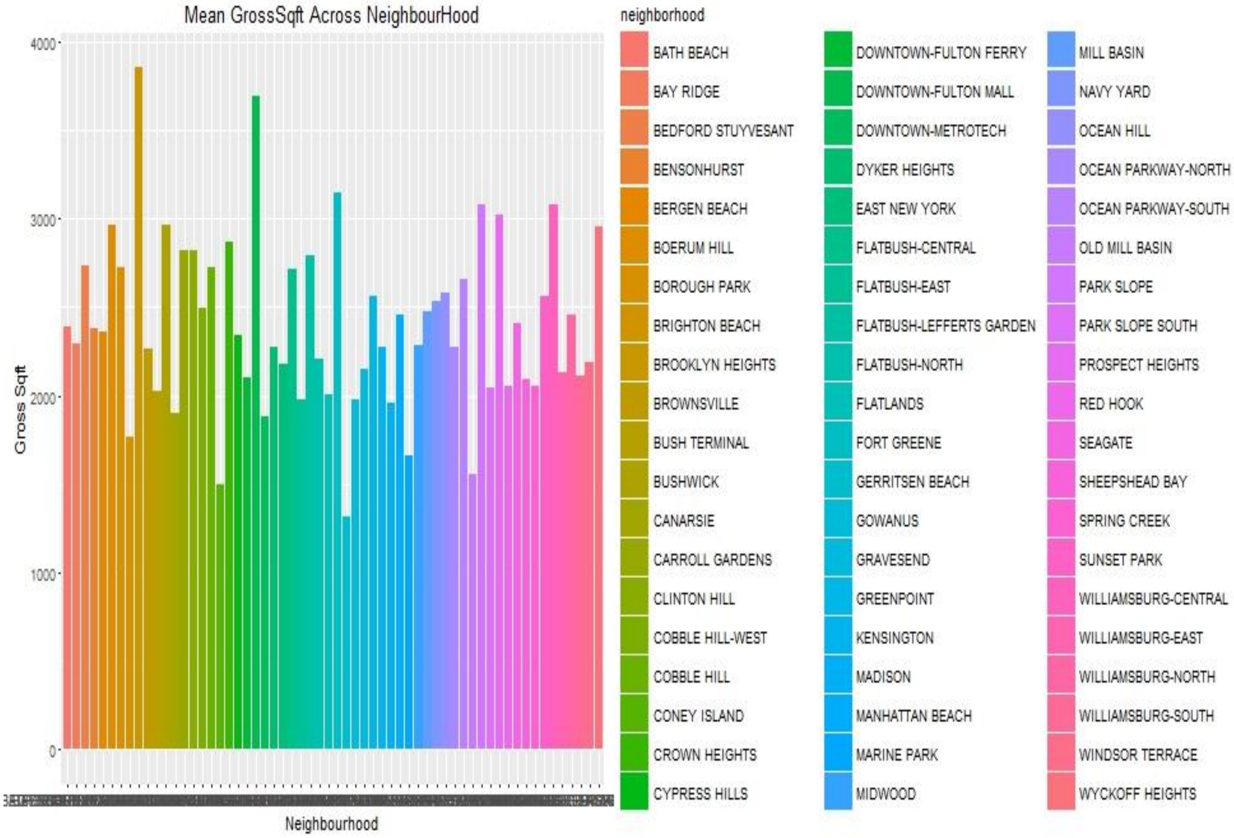
- ***In which neighborhood houses are more costly compared to others?.***

As seen in the plot below, Brooklyn Heights has an higher average sales price compared to other neighborhoods in Brooklyn



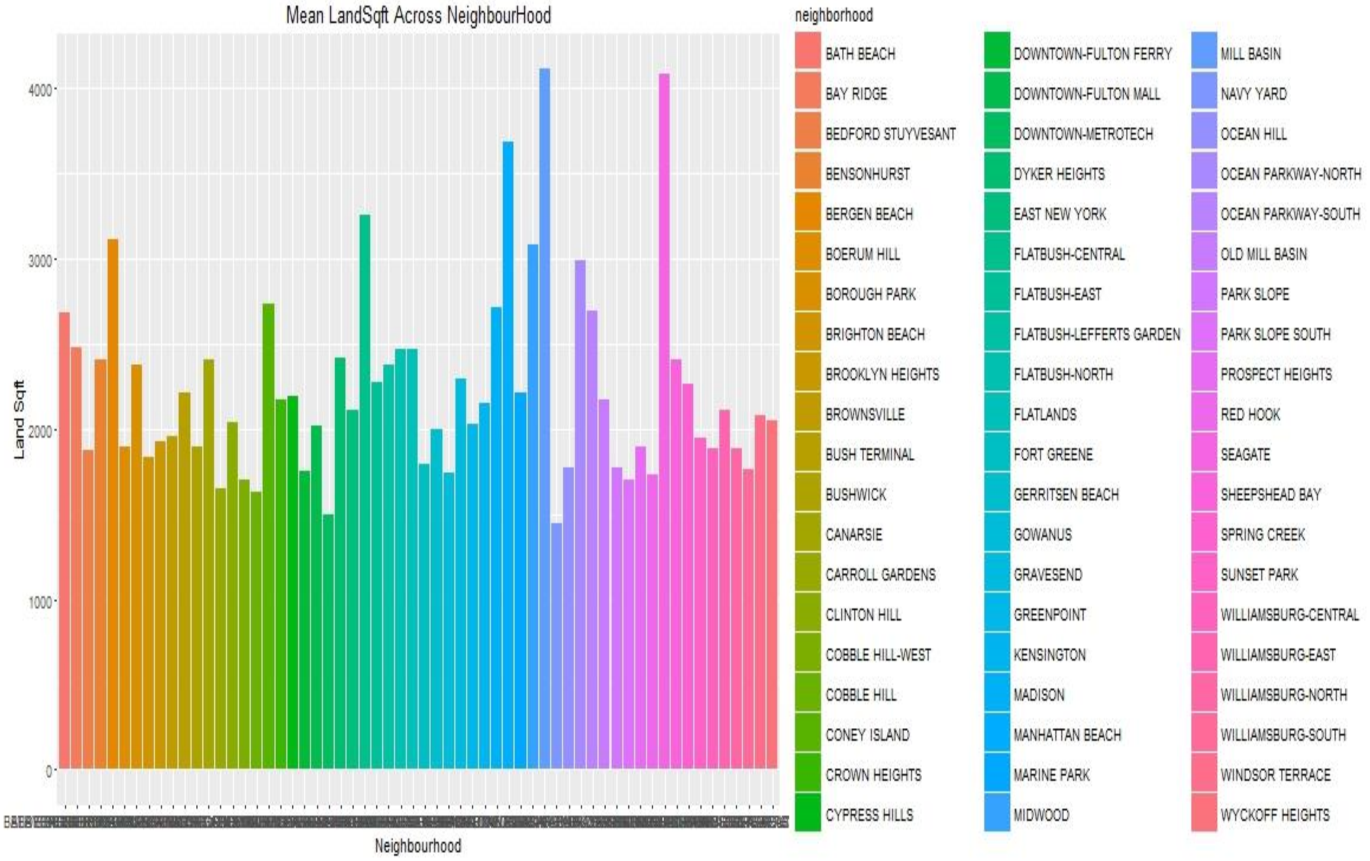
- ***In which neighborhood, houses have higher grossSqft area compared to others?.***

  Evident from the plot below, the average gross square feet of houses in Brooklyn Heights is higher compared to houses in other neighbourhoods. It can also be inferred that the reason behind the average sales price being higher in Brooklyn could be due to its houses having more grossSqft area.



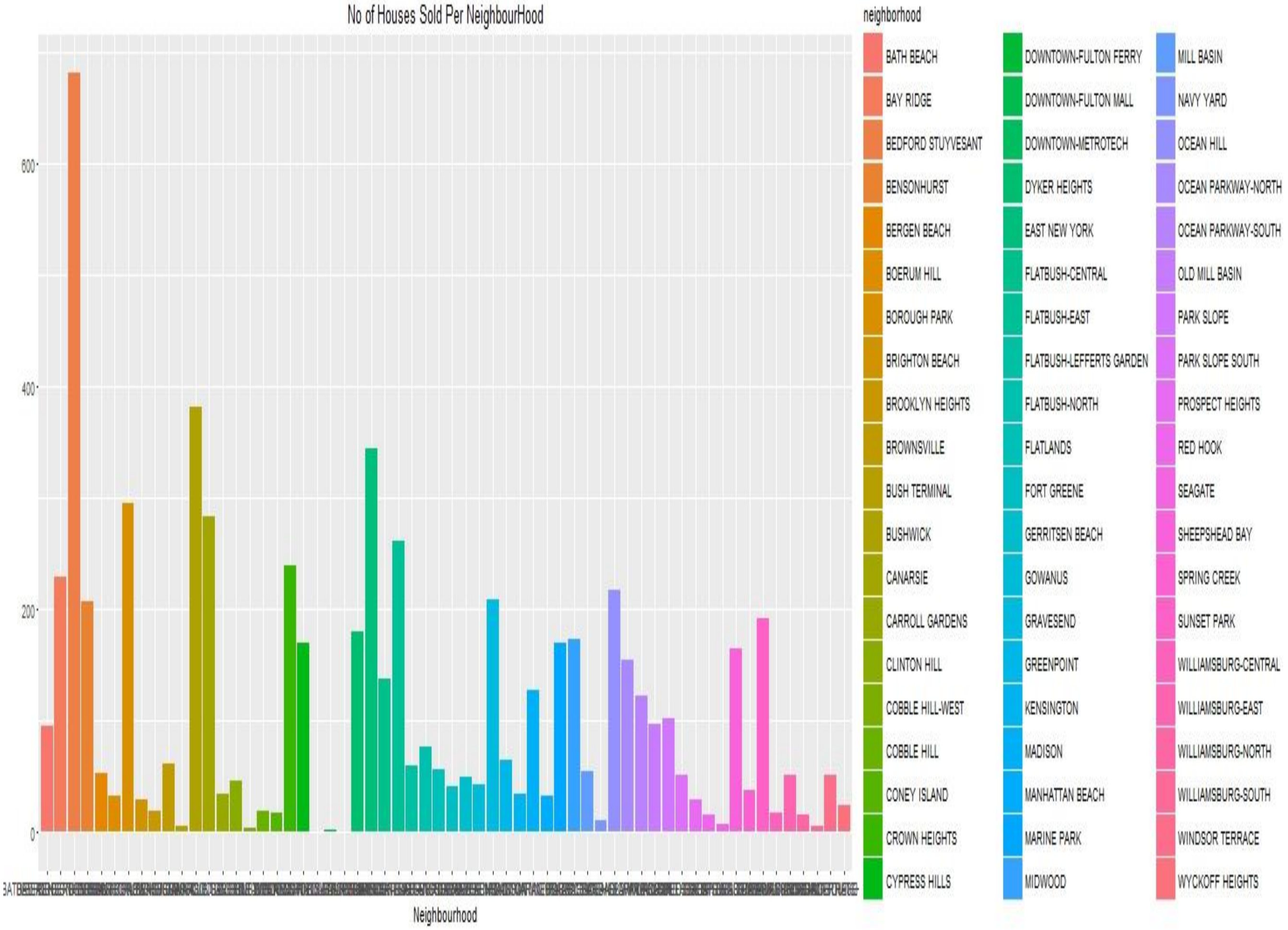
- ***Do the neighbourhoods follow similar pattern observed for GrossSqft above for LandSqft also?***

Evident from the plot below, the pattern is not same, Brooklyn Height doesn’t have higher LandSqft area compared to others. Neighbourhoods like SeaGate and Old Mill Basin on an average have higher LandSqft area compared to others.



- ***In which neighborhood most number of houses were sold?.***

As seen in the plot below, most number of houses were sold in BEDFORD STUYVESANT neighborhood. It can also be inferred from the plot below that the three Downtown neighborhood in Brooklyn are not very popular as no houses were sold in the 2012-2013 year.

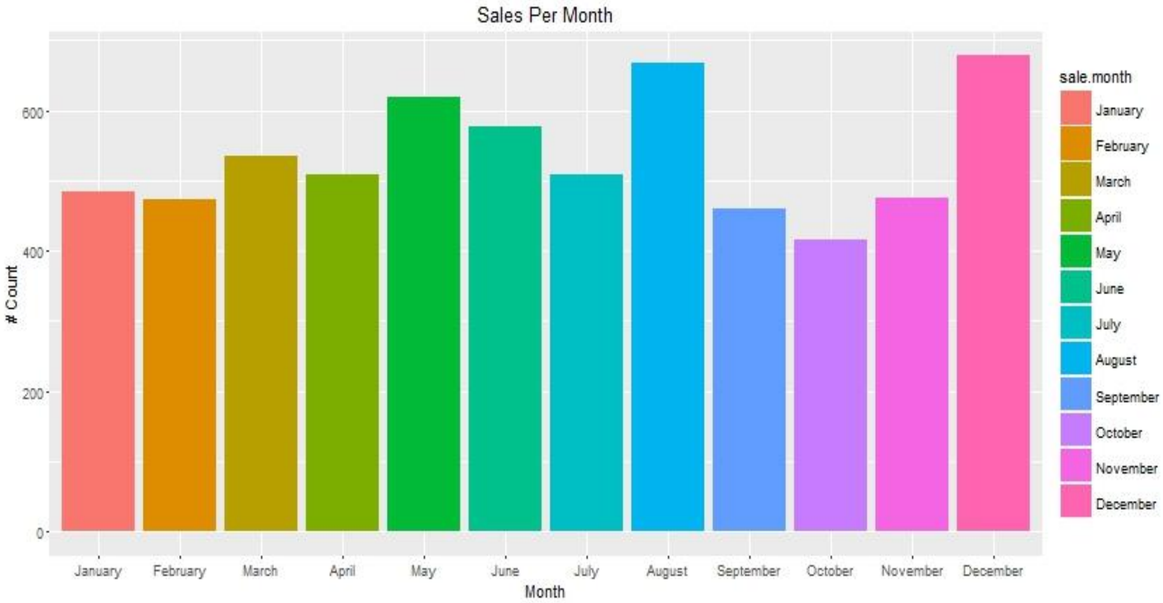


Author: Bhavika Jain
bhavikap@buffalo.edu

2) Monthly Observations :

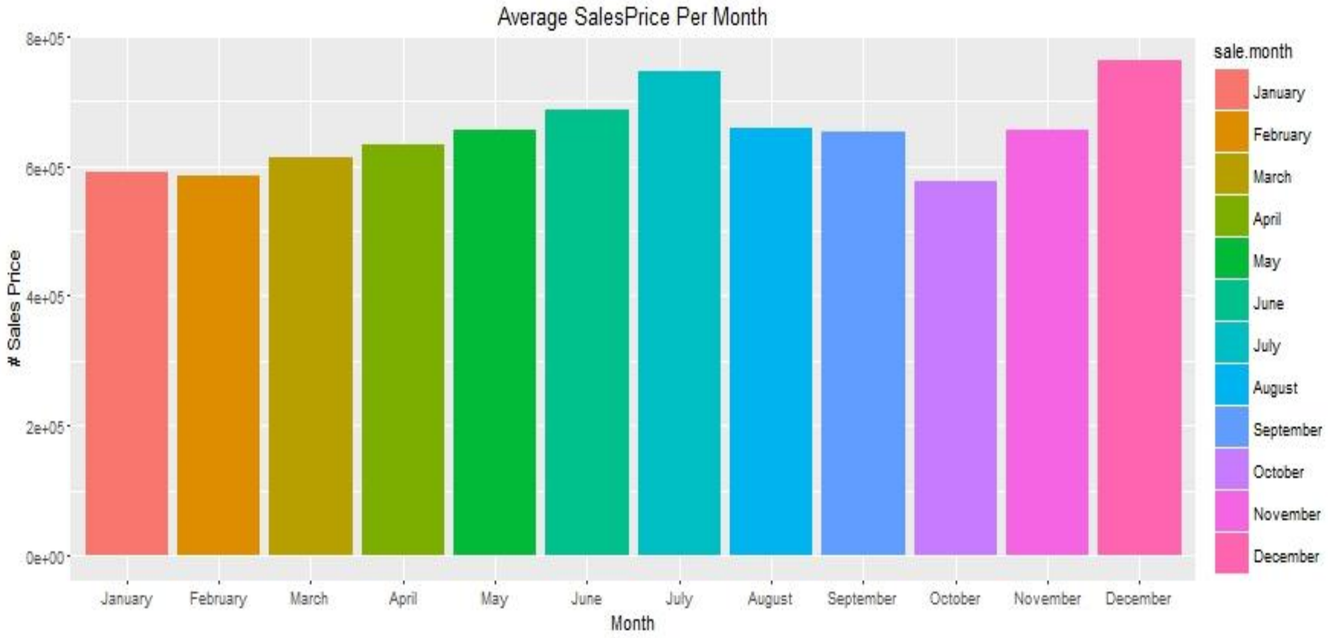
- ***Is there any peak month, when more number of houses were sold?***

As seen in the plot below, more sales happened in the month of August and December.



- ***Did more sales happened in December/August because Sales Price had been less?***

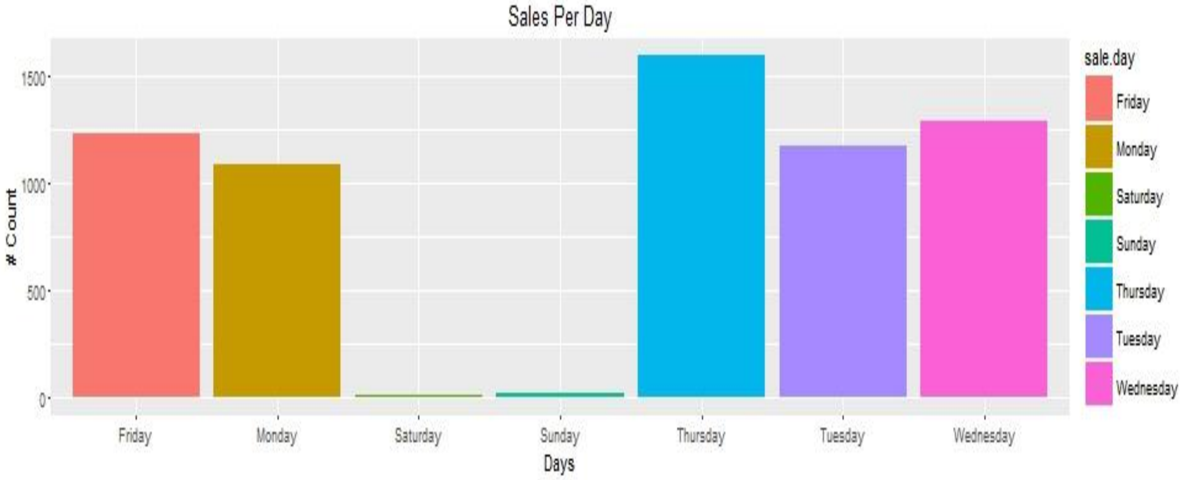
As seen in the plot below, that is not the case, the average sales price is almost equal across all months.



3) Observations based on days : -

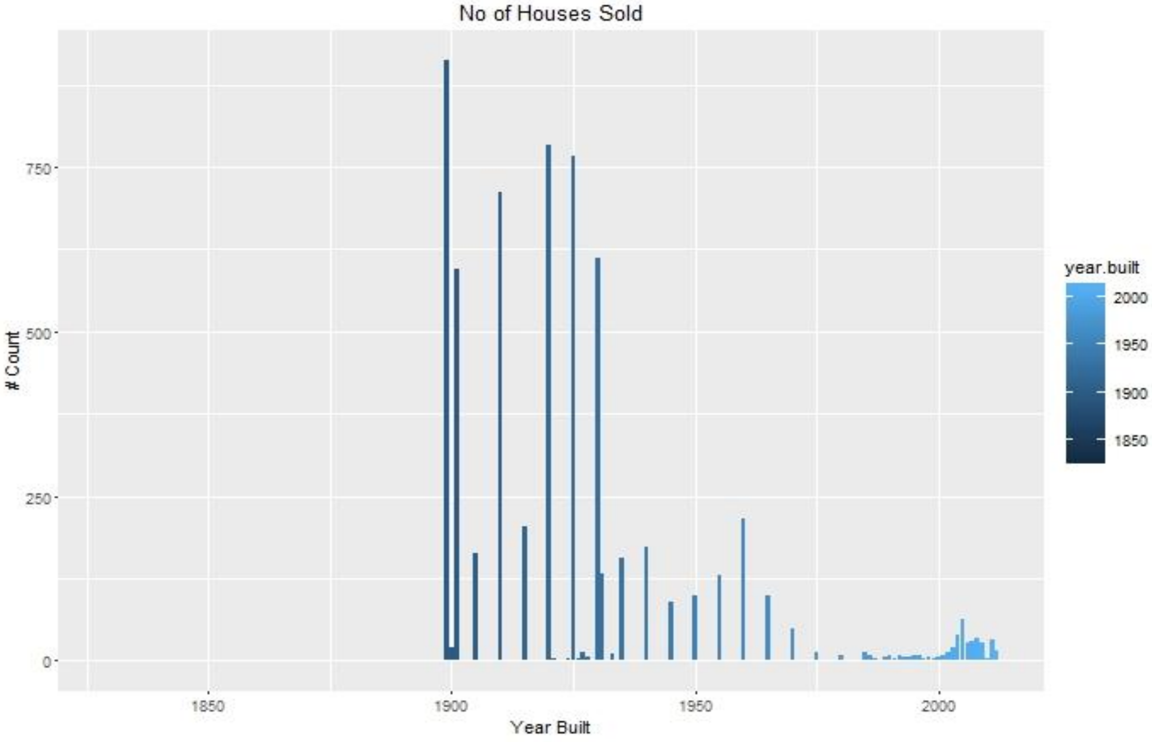
- ***Is there any particular day, when more number of houses were sold?***

Evident from the plot below, more sales happened on Thursday in the 2012-2013 year.



- ***Are houses built in particular year sold more compared to others?.***

As seen in the plot below, houses built before 1900 were sold more compared to others. We can draw an inference from this, that old houses are more popular in the Brooklyn neighborhood.



Author: Bhavika Jain
bhavikap@buffalo.edu

Part B - Extend EDA to All Boroughs

Objective: In part B we had to extend the EDA to all boroughs and find patterns and observations.

Approach:

I combined the data set of all boroughs and then did EDA on the entire data set for finding patterns/observations.

Key Variables:

- Borough
- Sales Price
- Gross Sqt
- Land Sqft

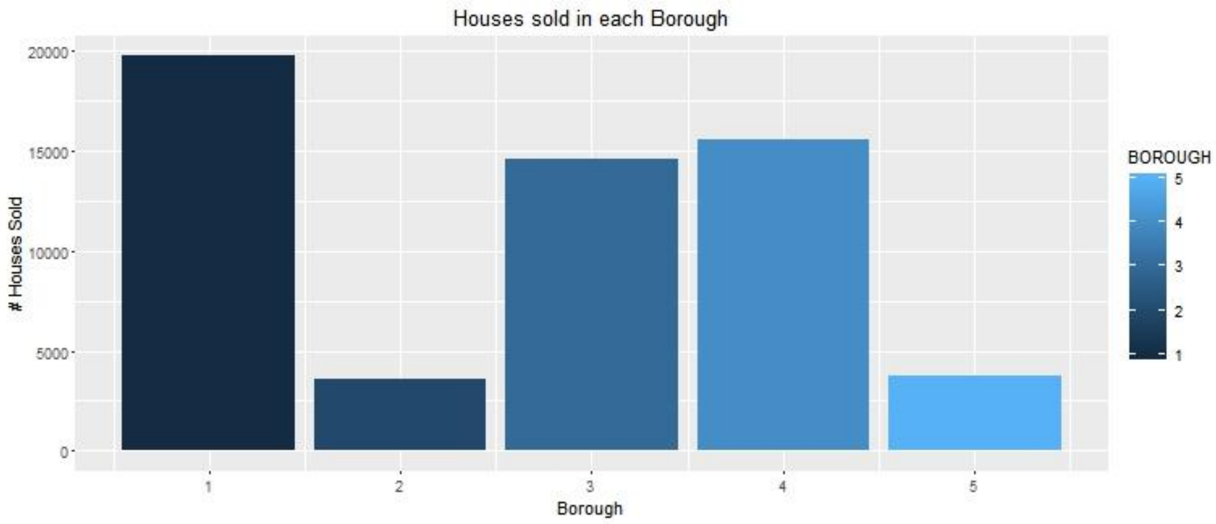
P.S:

- Borough 1 – Manhattan
- Borough 2 – Bronx
- Borough 3 – Brooklyn
- Borough 4 - Queens
- Borough 5 - Staten Island

Below are the observations done on the real direct data set for all boroughs.

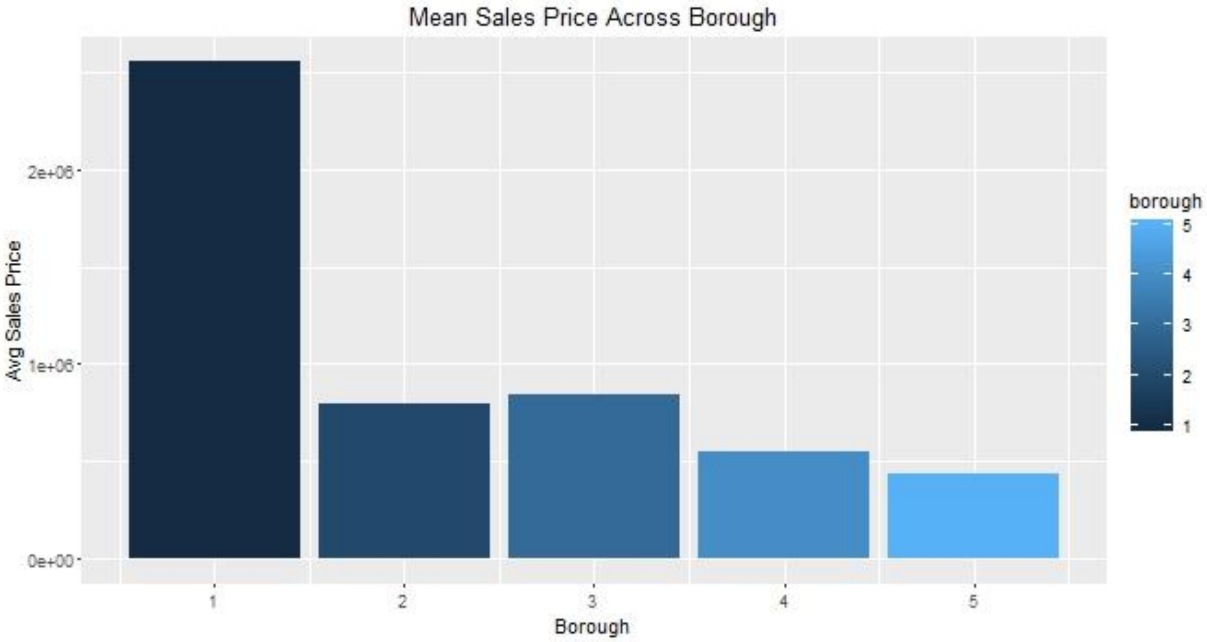
- ***Which borough has most sales?.***

Evident from the plot below, more sales happened in Manhattan Borough compared to other Boroughs. Bronx and Staten Island are less popular in NY compared to other boroughs.



➢ ***What is the Average Sales Price across each Borough?***

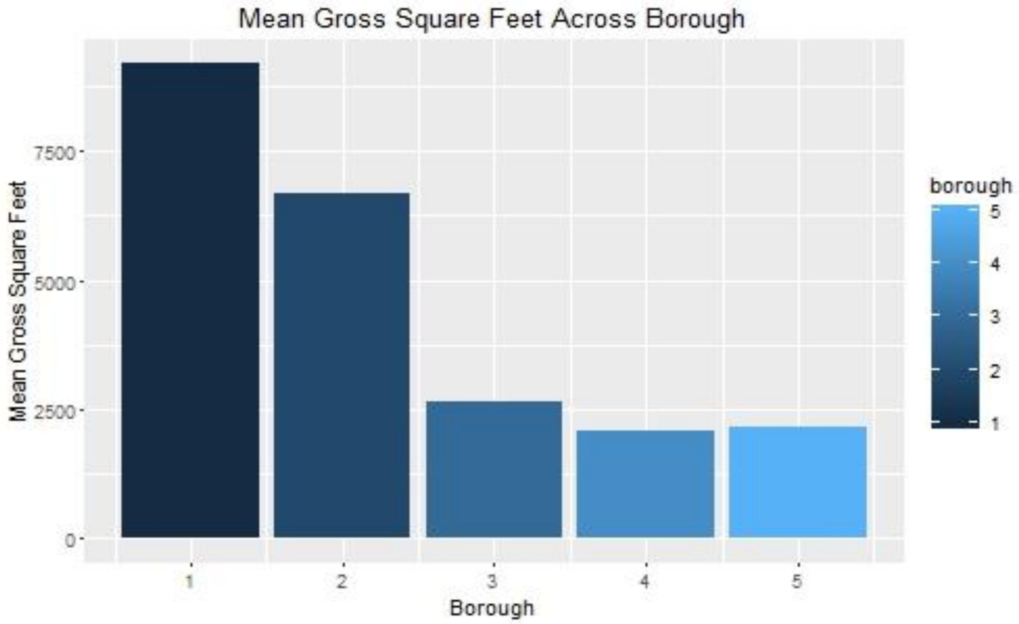
As seen in the plot, the sales price in Manhattan is higher compared to other Boroughs.



Author: Bhavika Jain
bhavikap@buffalo.edu

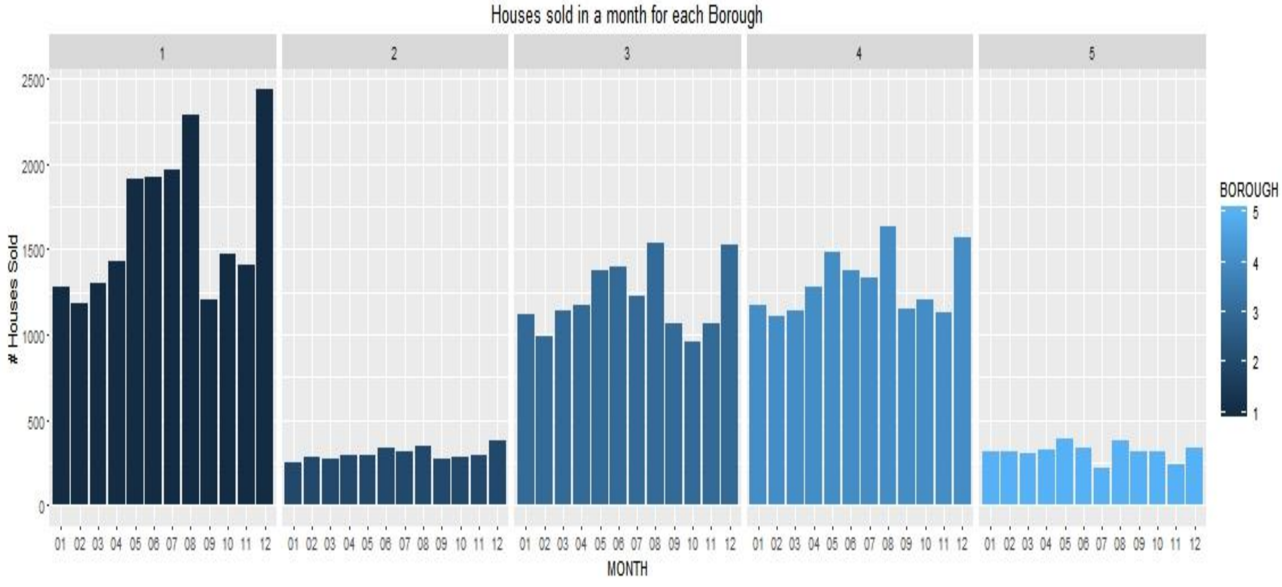
➢ ***Which borough has most gross sqt?.***

As seen from the plot below, Manhattan has the most gross sqt compared to other boroughs.



➢ ***Monthly sales of houses in each borough?***

As seen from the plot below, most sales happened in the month of December/August in each Borough.



4. 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?

We should talk to potential customers so as to understand what they look for when buying homes (school district, proximity to work place and other civic amenities, etc.) so we can ideally monitor the recommendation system that RealDirect would offer to its customers.

5. 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?

Stepping out of my comfort zone is definitely challenging, however working on a different domain allow us to look closely at the process we follow while collecting data. We take some things for granted while working in a familiar domain and in the process make mistakes and also make it difficult for people not from the same domain to understand the process.

The vocabulary wasn't really confusing. If I did not understand anything like for example, what does 'Tax Class' category meant, I simply looked it up. It is important to understand the vocabulary of the domain in which we are working to avoid making mistakes in coming to any conclusion during analysis of data.

6) Doug mentioned the company didn't necessarily have a data strategy. There is no industry standard for creating one. As you work through this assignment, think about whether there is a set of best practices you would recommend with respect to developing a data strategy for an online business, or in your own domain.

Below are the steps of best practices one must follow.

1) Collect as much data as possible.
2) Clean the data, remove any outliers.
3) Format the data if required.
4) Perform EDA on the data and find patterns.
5) Use the analysis to improve business strategy.

Author: Bhavika Jain
bhavikap@buffalo.edu