REALDIRECT CASE STUDY

Karthik Chaganti

50169441

kchagant@buffalo.edu

CSE 587 Data Intensive Computing

University at Buffalo

Language: R

Tools: R Studio

Problem 3

Abstract

Explore its existing website, thinking about how buyers and sellers would navigate through it, and how the website is structured/organized. Try to understand the existing business model, and think about how analysis of RealDirect user-behavior data could be used to inform decision-making and product development.

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).

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

By logging data corresponding to user's requirements indirectly (i.e., via click-through and site

navigation analysis), we could recommend listings that would be better suited to his needs. This data would also help us in advertising VIP listings.

We could also classify homes into categories using BHK, Zip code/Neighborhood, Area in sq. ft., Price, Build Date and Unit Category. We could then determine what category of homes a user would prefer based on his listing views.

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.

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 as well as use collaborative filtering to determine

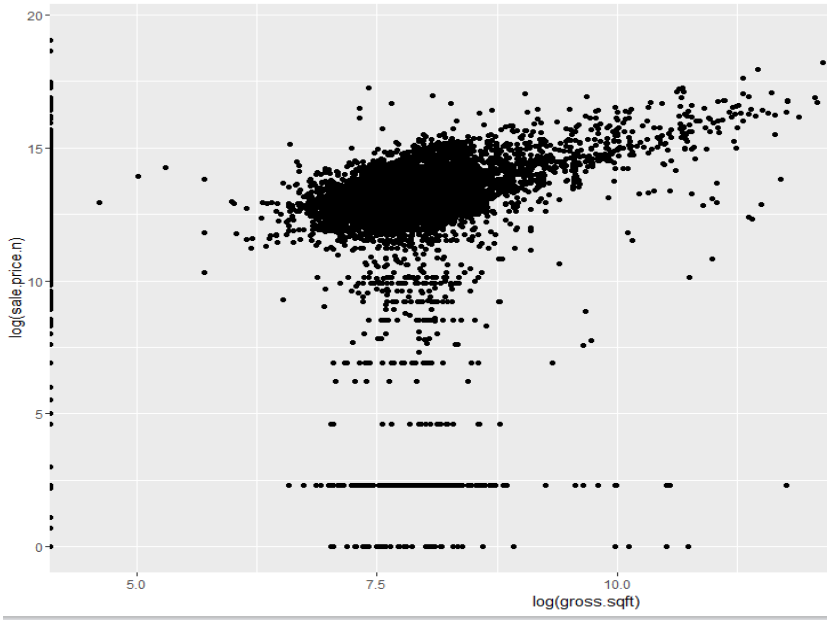
Objective

First challenge: load in and clean up the data. Next, conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.

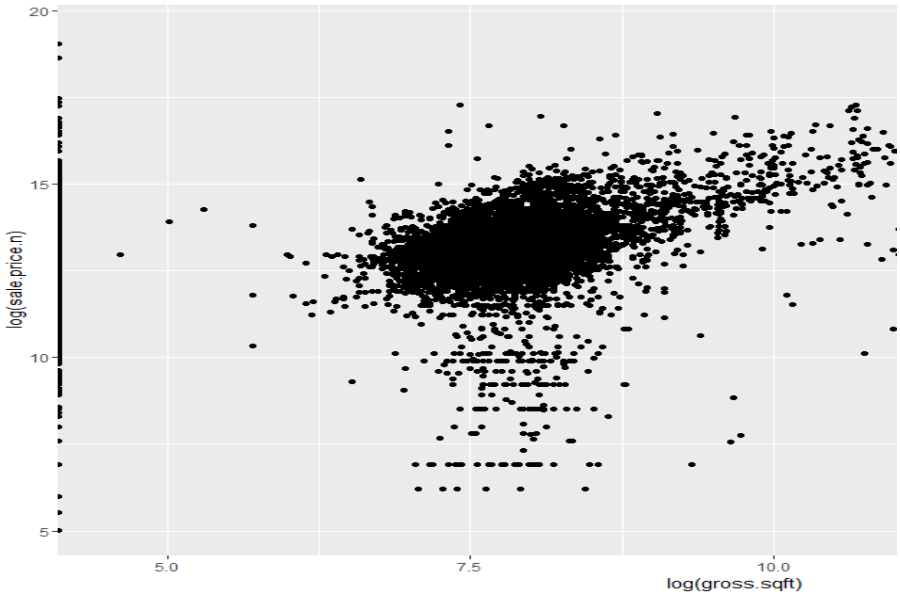
• Once the data is in good shape, conduct exploratory data analysis to visualize and make comparisons (i) across neighbor- hoods, and (ii) across time. If you have time, start looking for meaningful patterns in this dataset.

EDA ON BROOKLYN BOROUGH DATASET OVER THE YEAR

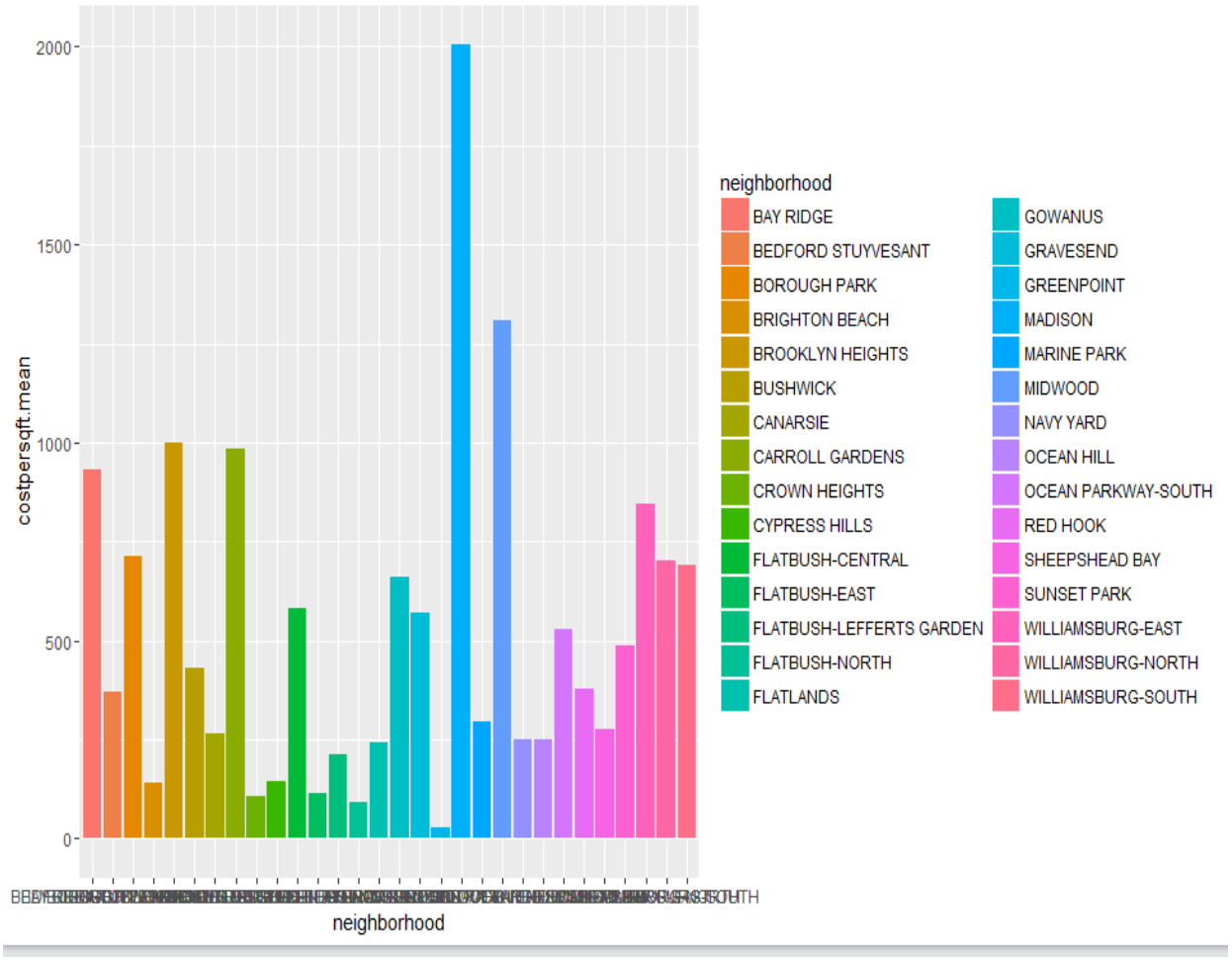
- ✓ The first step in the procedure involves removal of outliers. The only outliers that can possibly exist are in the sale prices as they are the variables that actually change. Hence the below log plot gives the potential outliers way down below in the figure:



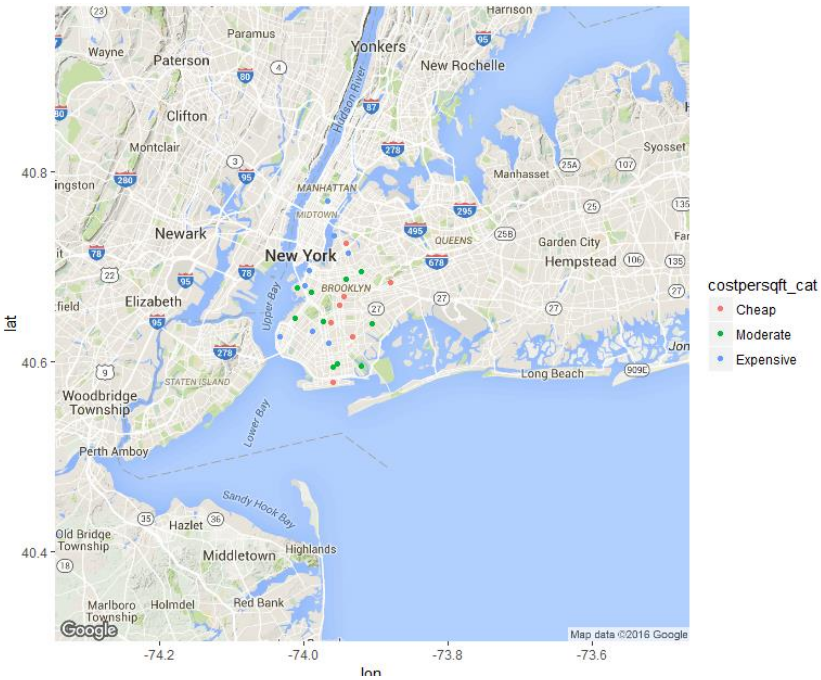
- ✓ Post removal of outliers, following plot can be a proof as the outliers have disappeared.



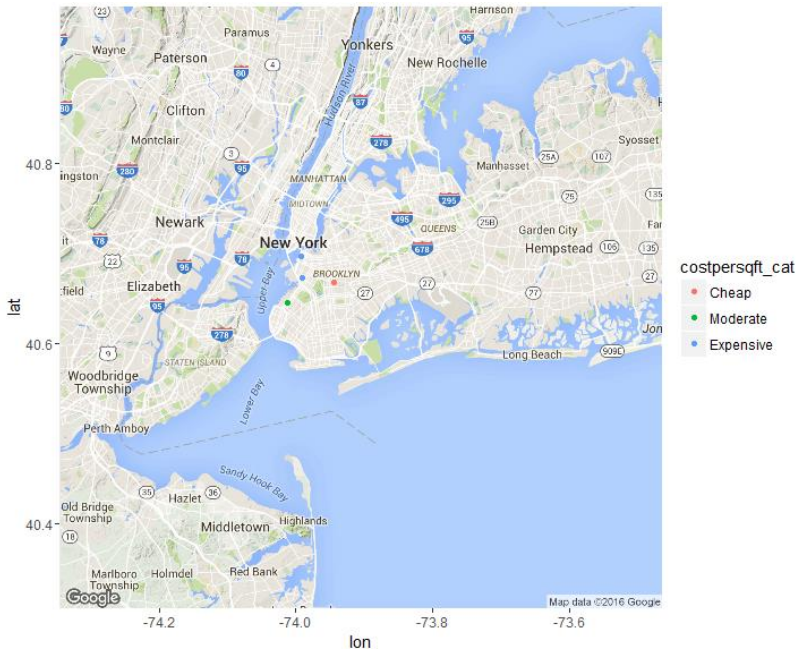
✓ **Cost per Sq. Ft. of luxury hotels distributed across neighborhoods in Brooklyn**
As we can see, the cost per sq. ft. of luxury hotels is highest in Marine Park and lowest in Madison.



✓ **Map of distribution of 'Office' spaces across neighborhoods by their cost per sq. ft.**

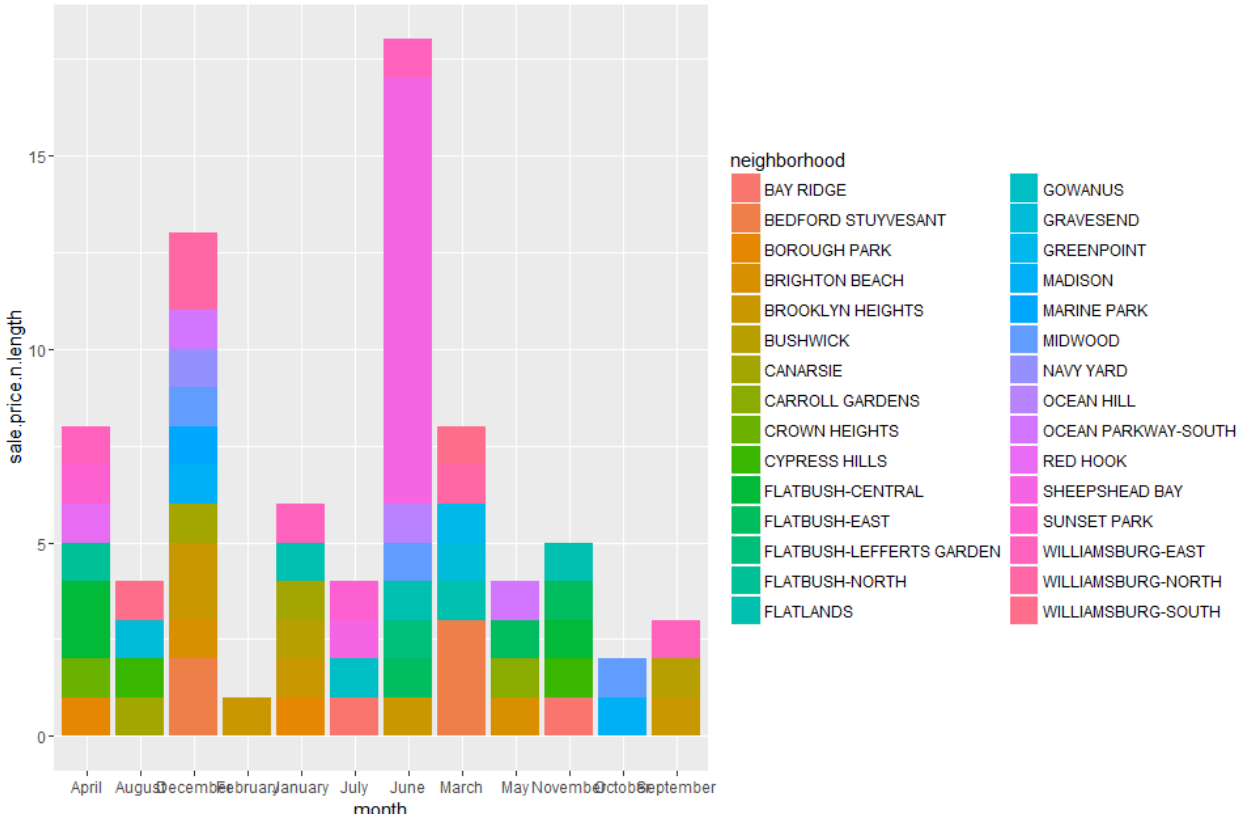


✓ **Map of distribution of 'Hotel' spaces across neighborhoods by their cost per sqft**



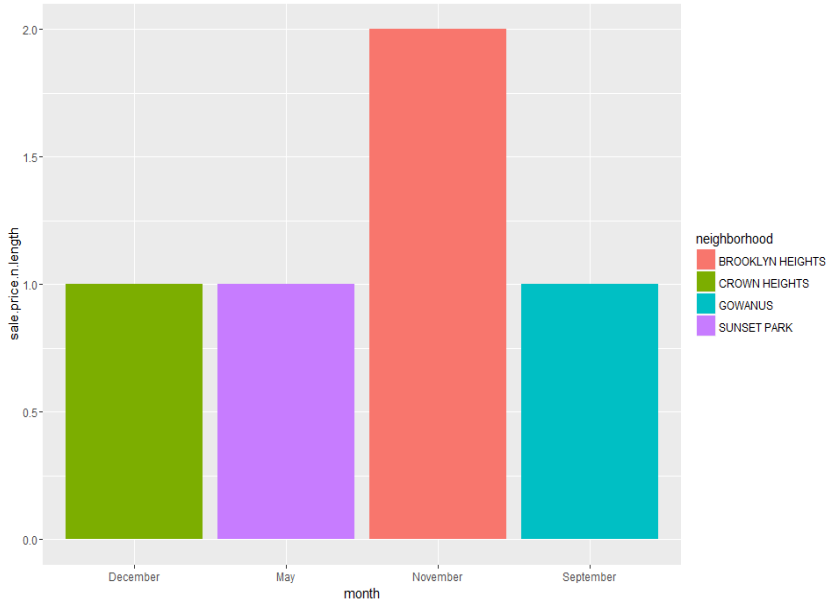
✓ We can see that there were very less luxury hotel sales in the area!

✓ **Distribution of total number of sales of offices across neighborhoods across months**



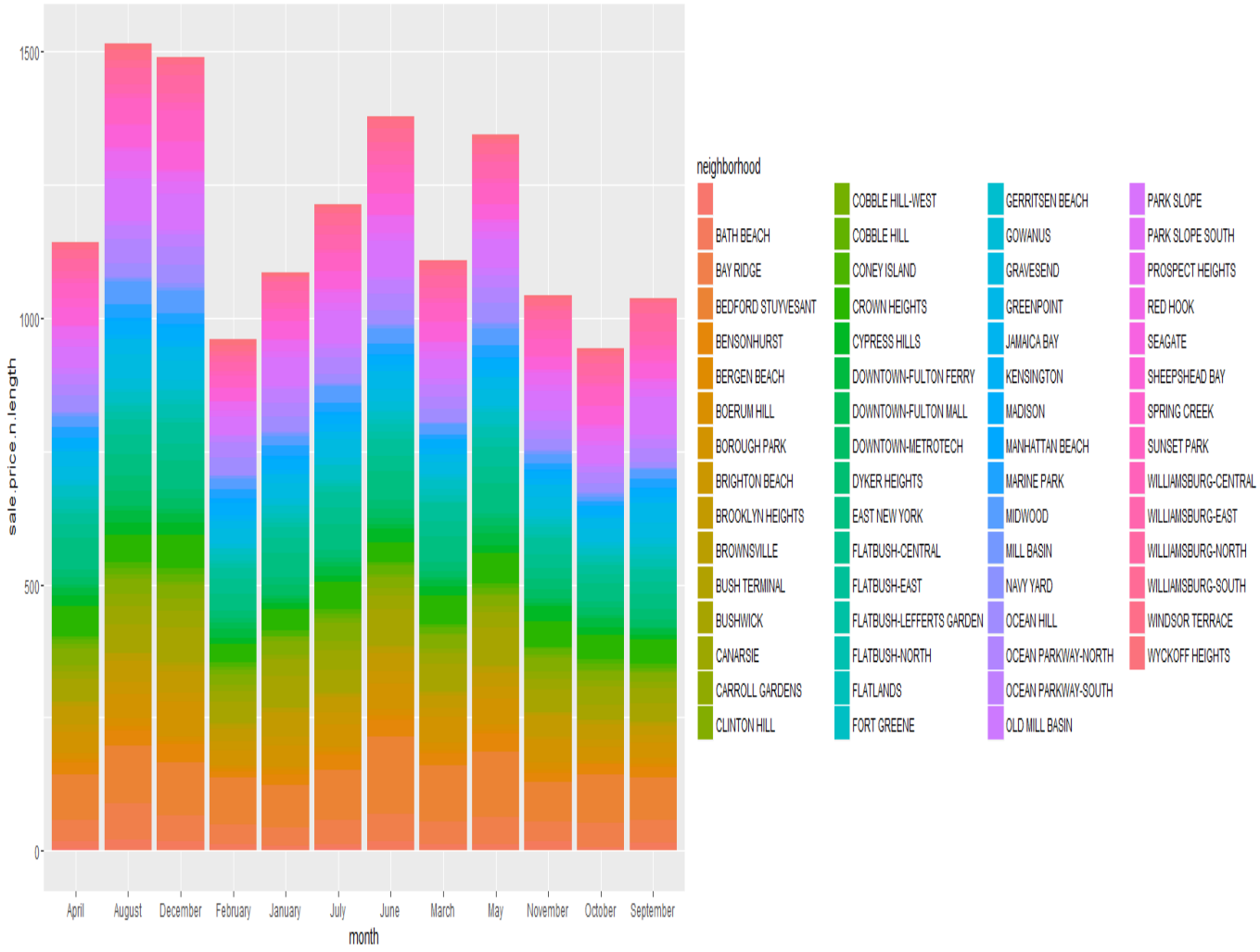
✓ **Distribution of total number of sales of hotels across neighborhoods across months**

- Brooklyn heights was the costliest sale that had taken place in November



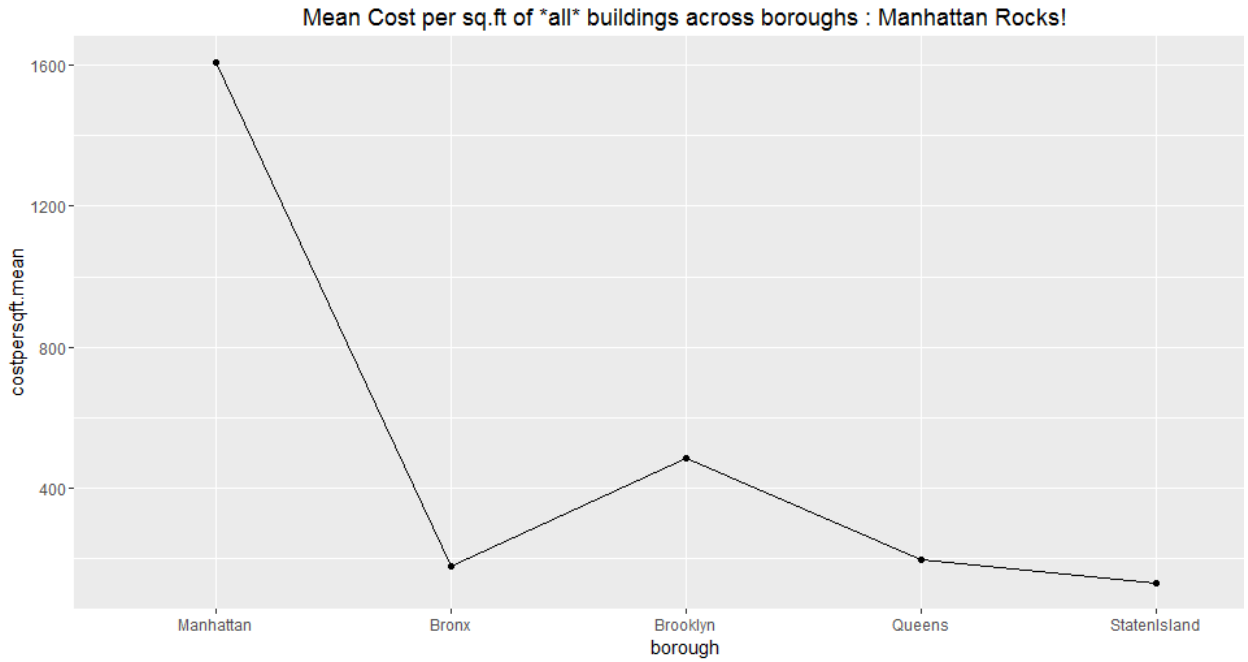
✓ **Distribution of total number of ALL sales across neighborhoods across months**

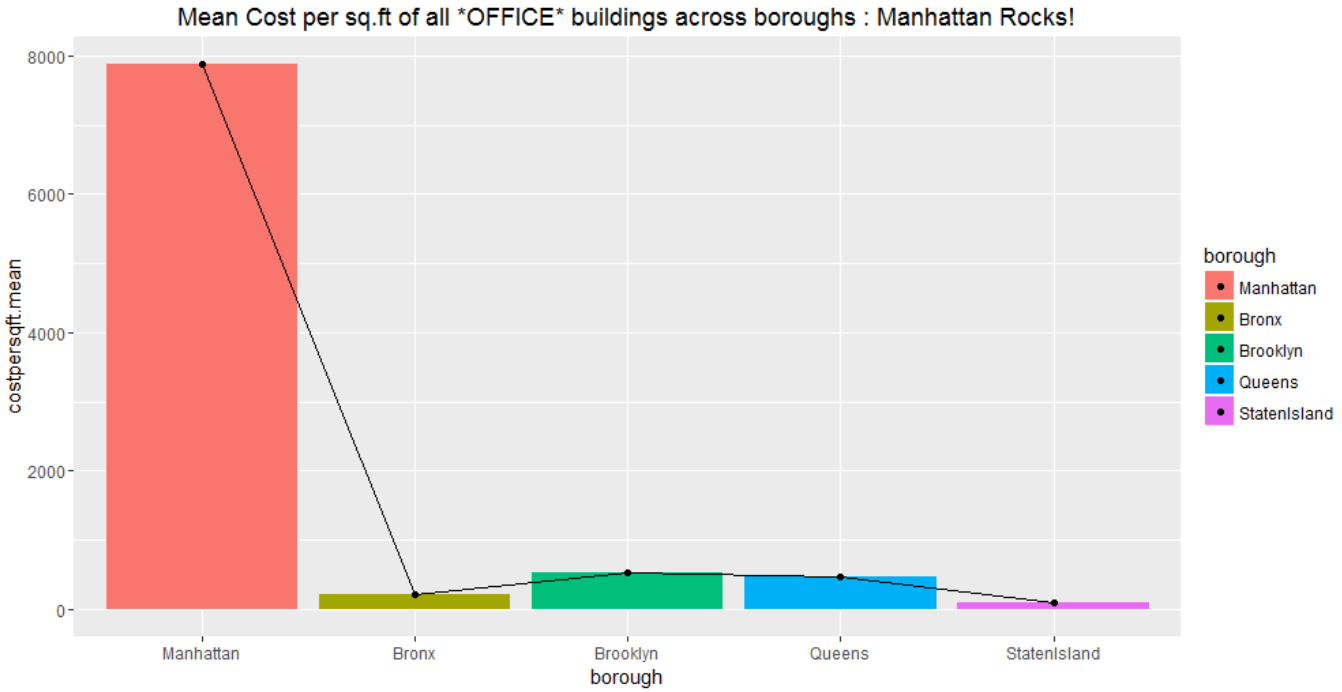
- Sales prices were highest in December and August!
- Followed by June and May
- Lowest was recorded in February suggesting an off-season

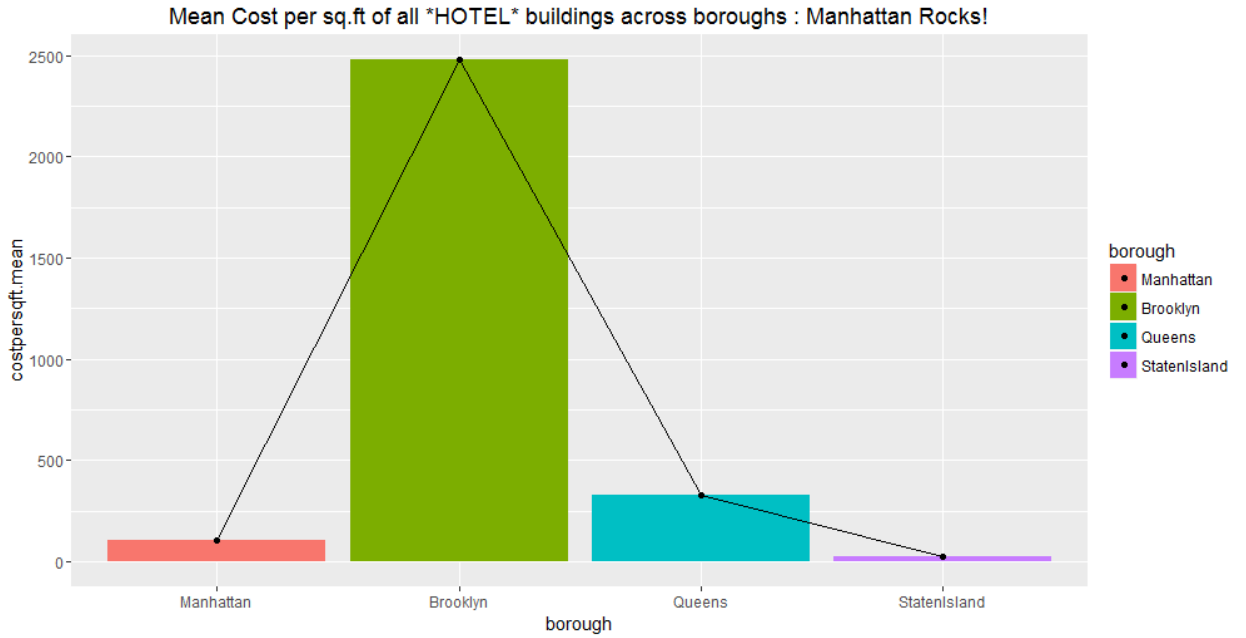


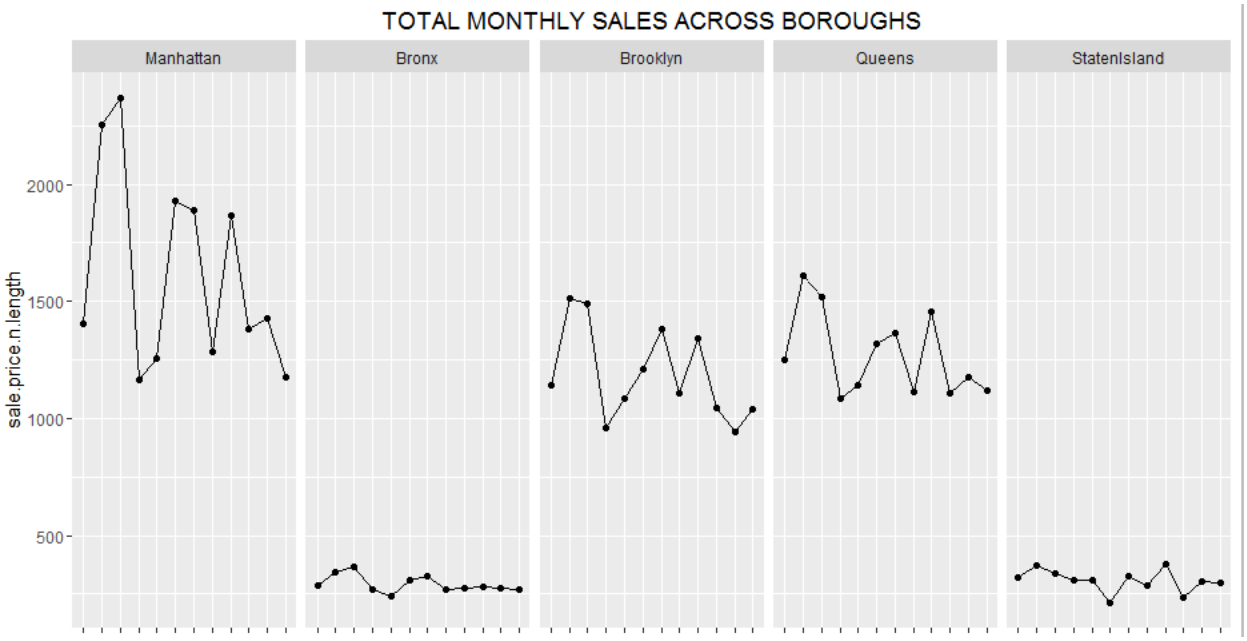
EDA ACROSS BOROUGHS OVER THE YEAR:

- From the figure we can say that the mean cost square feet of all the buildings is higest in manhattan.









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 would like to 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.

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, it does. Working on a different domain allowed us to look closely at the process we followed while collecting data. We take some things for granted while working in a familiar domain and in the process make it difficult for people not from the same domain to understand the process. Working on RealDirect helped us realize this and ensured that we document the process steps in a detailed sequential manner.

Sometimes "domain experts" have their own set of vocabulary. Did Doug use vocabulary specific to his domain that you didn't understand ("comps," "open houses," "CPC")?

Sometimes if you don't understand vocabulary that an expert is using, it can prevent you from understanding the problem. It's good to get in the habit of asking questions because eventually you will get to something you do understand. This involves persistence and is a habit to cultivate. We had to look up some of the domain vocabulary on the internet. For example, we were not familiar with the existing tax classes of real estate which initially hindered our understanding of the data. 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.

Step 1 – Understand the problem

Step 2 – Collect and clean data

Step 3 – Perform EDA to understand the data and the variables that impact the business

Step 4 – Use the analysis to improve the strategy used

Perform Steps 2 through 4 iteratively.

Summary

- - Most sales happen in Harlem and the Upper East Side.
- - Housing is most expensive in the Upper East Side.
- - Housing is least expensive in Harlem.
- - Houses have the most gross square feet in Gremich and the Upper East Side.
- - Houses seem about equal in terms of gross square feet elsewhere.
- - Houses seem about equal in terms of land square feet throughout.
- - The Upper West side has the newest housing.
- - Housing was built mostly around the early 1900s.
- - Harlem has the newest housing.
- - The amount of sales seemed equally distributed around Wednesday.
- - December has a lot more sales than any other month.
- - The least amount of sales happened in September and October.
- - Home sales according to other metrics across time seem equally distributed.