CSE 587: Project 1
Problem 3 - Case Study

Lalith Vikram Natarajan, UBIT ID: lalithvi, Person No: 50169243

March 5, 2016

1 Objective

Perform EDA on a real world data set

- Analysis of a single data set (Brookyln)
- Analysis on all data sets (Brookyln, Manhattan, Queens, StatenIsland)

2 Questions and Answers

Question 1. 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. Come up with a list of research questions you think could be answered by data:

- What data would you advise the engineers log and what would your ideal datasets look like?
- How would data be used for reporting and monitoring product usage?
- How would data be built back into the product/website?

Answer:

1. (a) User search queries should be logged in, along with their bounce rate, time spent on each page, keep track of users who are often visiting the site but hesitant to sign up, send a friendly alert.
   (b) Ideal dataset should include datasets similar to the ones present such as nyt1.csv

2. Logs will provide a list of properties and areas which are popular, these can be advertised and shown prominently on the website. It can also show the parts which are never clicked which would enable us to reduce the price of the listing and or rebrand it. This improves efficiency and grows revenue.

3. Using the information got from the data, we can forecast which areas are becoming popular, run targeted marketing campaigns based on the majority of the age group.

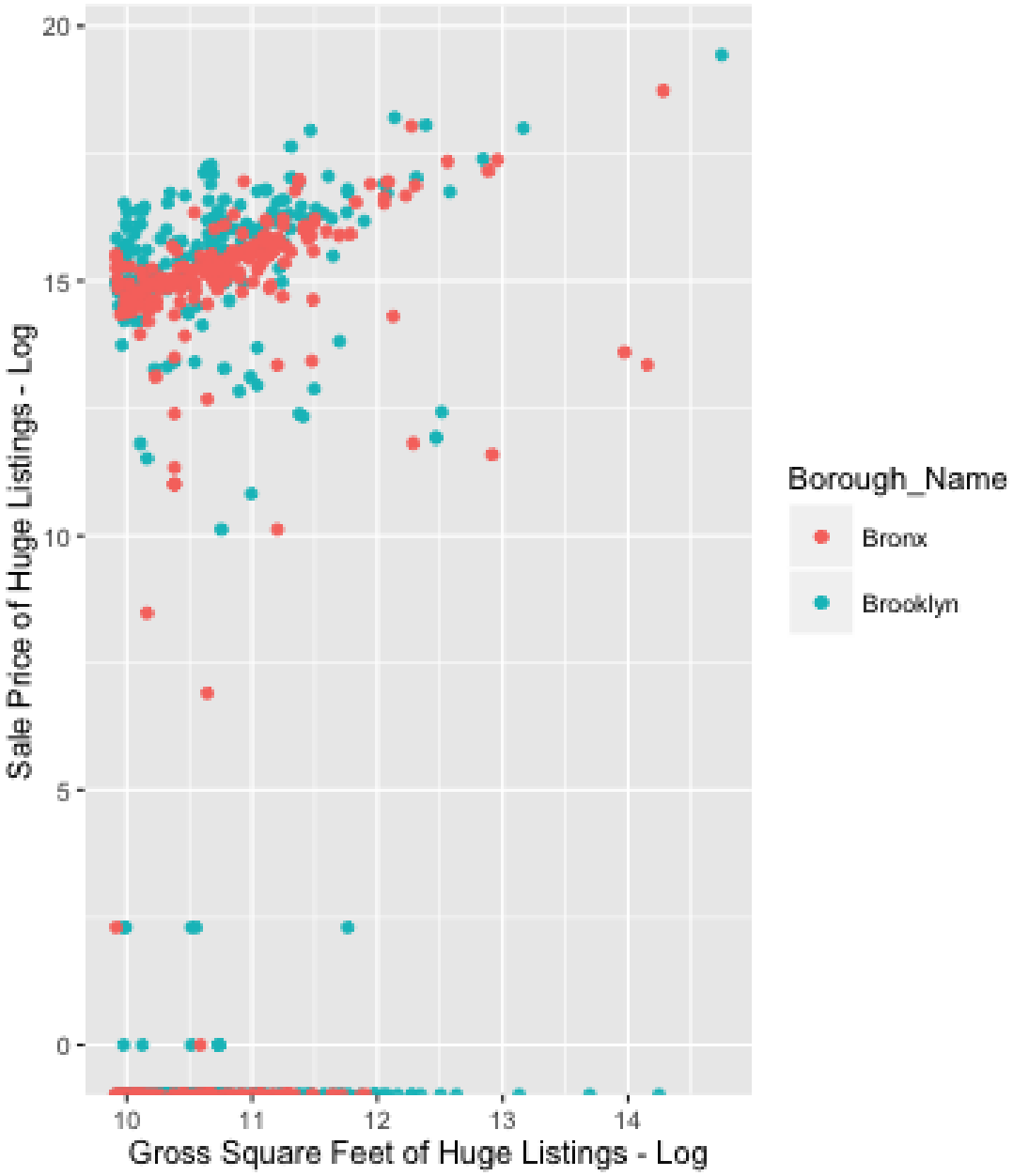
Question 2. Because there is no data yet for you to analyze (typical in a startup when its still building its product), you should get some auxiliary data to help gain intuition about this market. You can use any or all of the datasets here - start with Manhattan August, 2012-August 2013.

- 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.

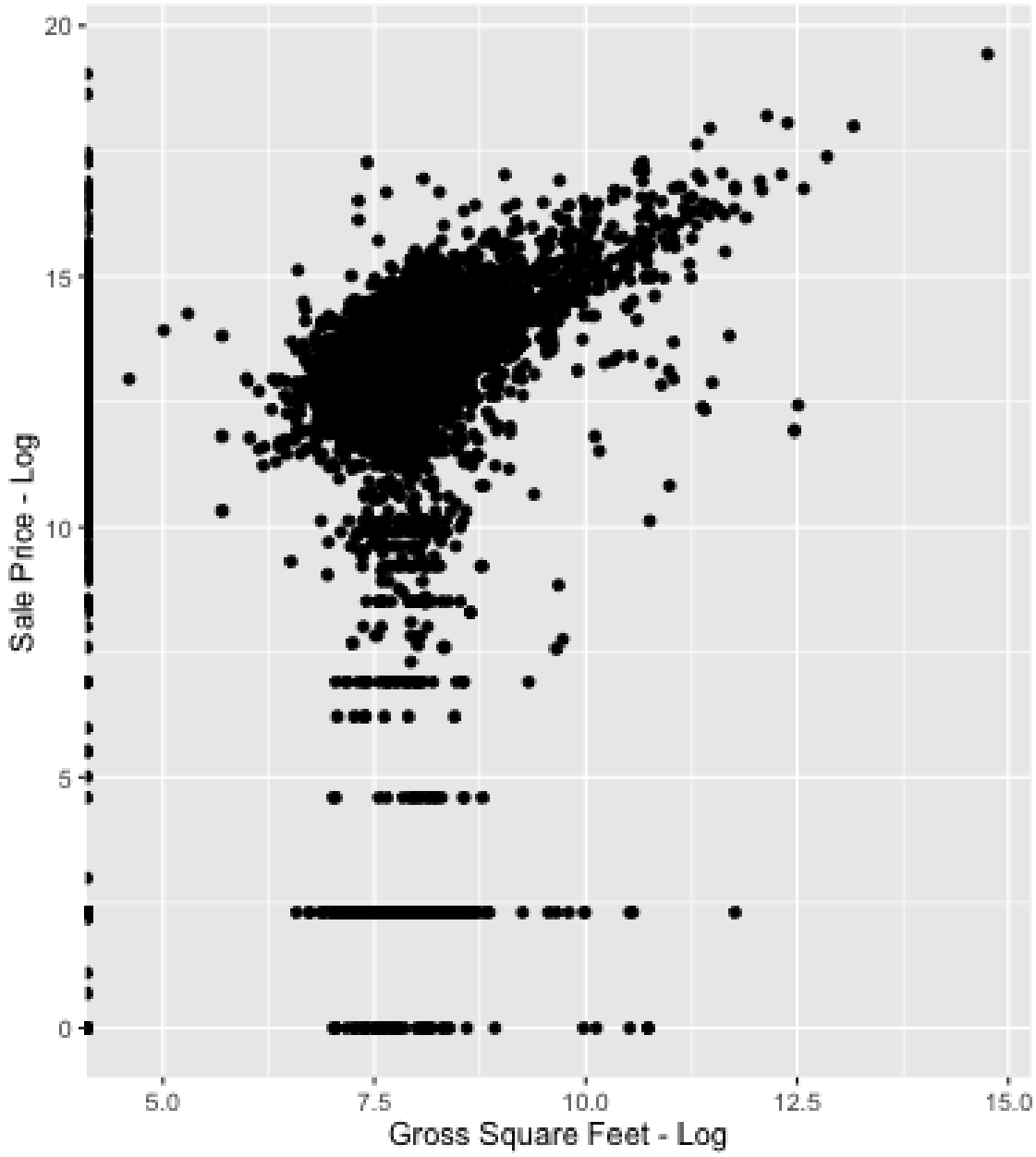
- Once the data is in good shape, conduct exploratory data analysis to visualize and make comparisons (i) across neighborhoods, and (ii) across time. If you have time, start looking for meaningful patterns in this dataset.

Answer: Please refer to the code in `lalithvi_RDP3lalithvi.R` & `lalithvi_RDP3Exlalithvi.R`.

Question 3. Summarize your findings in a brief report aimed at the CEO.
Answer: After analyzing two datasets, Brooklyn and Bronx, we found that the most popular selling neighborhood had been BEDFORD STUYVESANT, the most building class was TWO FAMILY HOMES , most of the sale price is not properly recorded, we found 10470 such records out of 28641. The average square feet of a sold house was 3594.



The above graph shows us the relationship between sale price of very huge properties (¿20000) in both Brooklyn and Bronx. It is more profitable to market to upmarket customers in Brooklyn than in Bronx



The above graph shows us the relationship between sale price of properties in Brooklyn. We should try to market our product to more people that fall within the intersection of x=7.5 & y=12 as the maximum sale has happened there.

Question 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?
Answer: RealDirect has a business model of a seller and buyer with some brokers in the middle. It would be beneficial to talk to brokers who are the domain experts and use their insights along with the data for forecasts. Search Engine Optimization executives and Marketing professionals within RealDirect would also help us in gathering more knowledge.

Question 5. Most of you are not 'domain experts' in real estate or online businesses. 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?
Answer: Yes it does, by going specific to an industry and doing EDA we are able to generalize and also figure out new ways to categorize the data.

Question 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.
Answer: Every Data Strategy should aim at attaining the following core values.

- Operational Efficiency.
- Retain and Grow Revenue.
- Figure out Barriers.
- Reduce Risk.
- Drive Insights.