BU CS591 L1 Project Summary

Property Values, Traffic Jams, and Hospitals in Boston

Joshua Mah, Joseph Muruguru, Timothy Pacius

I. Introduction

In this project, we sought to test hypotheses about relationships between hospitals, traffic jams, and property values. More specifically, we wanted to determine: a) if traffic jams were more likely in zip codes with lower property values and b) if hospitals are more prevalent in areas with higher property values.

We believe that these questions are important because they can be eventually used to determine if the relative value of each region of the city has any effect on how well served the populace is by hospitals, and more importantly if these areas are more prone to being affected by outside factors such as the Winter of 2015. To explore these hypotheses, we obtained three datasets, two from the City of Boston: Boston Hospital Locations and the 2015 Property Value Appraisals, and one from Waze which tracked local traffic from February to March of 2015. While it should be noted that the window of time for the traffic data is quite small, the window of time that it encompasses was among the worst for traffic locally last year.

In testing these hypotheses, we created several datasets via Python to help reach quantifiable conclusions, but primarily we focused on creating a dataset detailing the number of hospitals in local zip codes and the zip code's average property value, a dataset estimating the number of traffic jams occurring near hospitals in local zip codes and zip code's average property value, and a dataset estimating the number of traffic jams that occurred by traffic intersections in close proximity to hospitals.

II. Methodology

To make reductions on the initial datasets, we made extensive use of the MapReduce paradigm in Python. To reach our final datasets, we began reducing our datasets by finding pertinent data to reduce on. To reduce the Waze data and Boston Hospitals into a meaningful combination, we combined the datasets using street intersections. Similarly we reduced Hospital Locations and Property values in two different ways using their zip codes to determine the number of hospitals in each zip code and find the average property values in each zip code.

To optimize the operations on the datasets, we made sure to draw all of the data into a MongoDB collection to ensure that the data collection was not subjected to data throttling. The bulk of the operation were linear operations were only affected by the size of the data samples. A lot of our data was also limited by the dataset, some being more limiting than others, and granularity differencing among the different datasets.

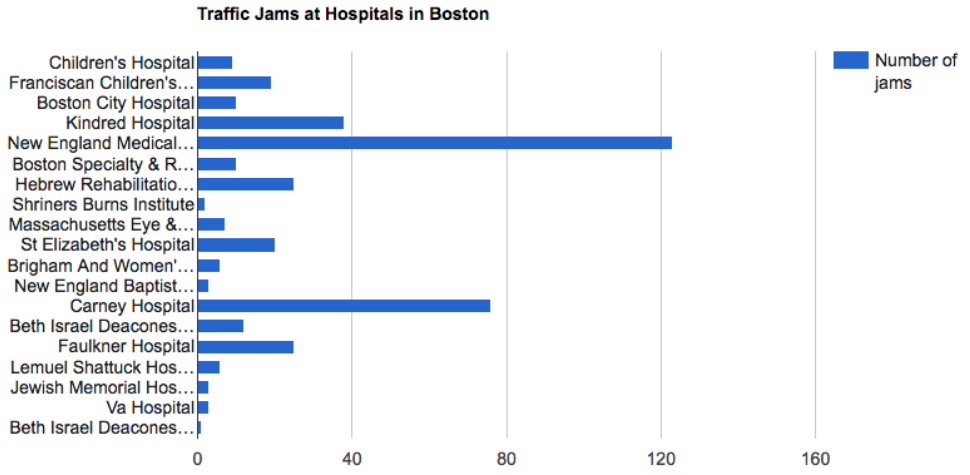
Before running any of our reductions, we had to cleanse the inputs from the Boston Hospital datasets due to their address formatting and zip code formatting differing from the other datasets. We cleaned the inputs using regular expressions, or just by pulling different keys from the data sets, including these pulls in a different dataset.

The largest limitations of the datasets were that all of the datasets simultaneously had too much and too little information. We believed that this hampered some of the conclusions that we hoped to make because zip code were a bit too coarse of a granularity to meaningfully observe trends over. Additionally, the timing of a lot of the traffic jams was inaccurate, due to a rounding by hour, which may have been a limiting factor on what data was placed.

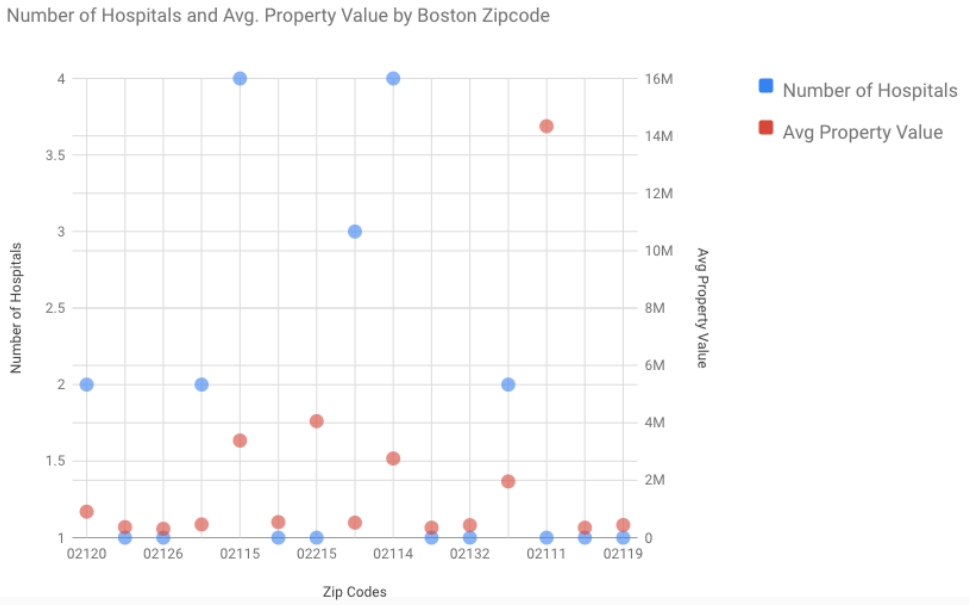
III. Results

In our hypotheses, we anticipated finding a negative correlation in the number of traffic jams and the average property value and a positive correlation in the number of hospitals and the average property. We instead found that the correlation between the number of hospitals in a zip code and zip code's average property value was negative and nearly non-existent (-0.00119) and the correlation between the number of traffic jams near hospitals and average property values was positive and somewhat significant (0.67221). The validity of these correlation coefficients is supported by the p-values of 1.0 and 0.055 respectively.

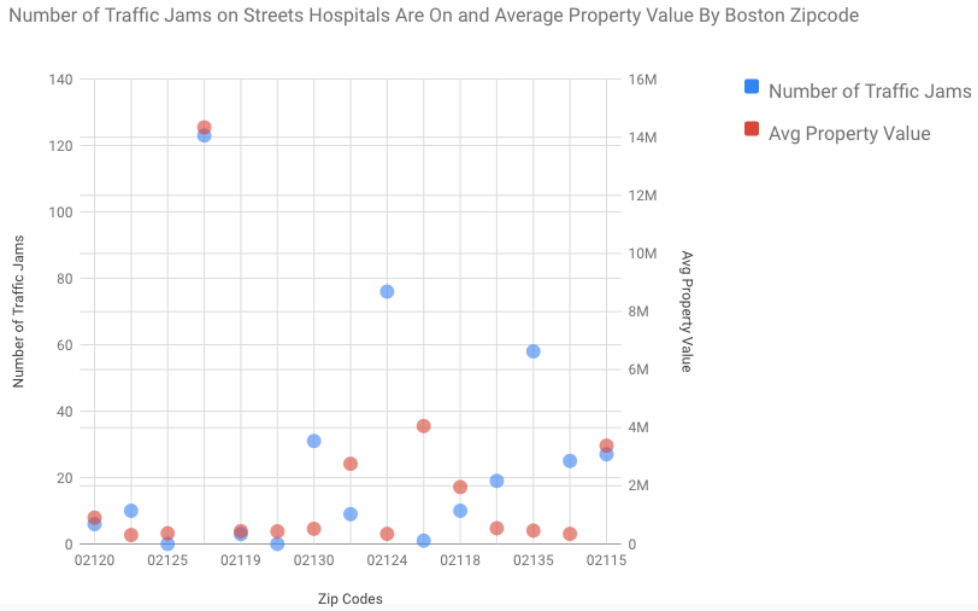
Visualization 1 (Number of Traffic Jams at Affected Hospitals)



Visualization 2 (Number of Hospitals and Average Property Value)



Visualization 3 (Number of Traffic Jams occurring near Hospitals and Average Property Value)



IV. Conclusion and Future Work

While we did take a large sample of the data, we believe that the uncertainty of randomization in sampling in property value data skewed the results. Additionally, due to the sample of traffic data being an extreme example of local traffic activity, it is hard to conclude that our findings are definitive. We believe that these findings could become better representations of our hypotheses and conclusions if the zip codes were further broken down by neighborhood, average residential property values vs. average commercial property values were taken in account, and there were more samples of traffic windows. With either a low correlation or a low p-value in our relationships, it is difficult to find a definitive response in our relationships. In the future, we can focus on different types of activities that may increase property value, some suggestions being number of police departments, number of potholes, and various other characteristics of a city. All in all, as we record more data, we will be able to see examples of characteristics that will help us to characterize neighborhoods and understand what makes certain areas greater in property value, helping to answer the question of economic conditions of neighborhoods.

V. References & Github Repository

Initial Datasets obtained from https://data.cityofboston.gov/.

Python scripts and brief project description can be found at https://github.com/Data-Mechanics/course-2016-spr-proj-two/tree/master/jmuru1_joshmah_tpacius