[[1]](#footnote-2)

Exploring the NYC 311 Service Request

YiHao Song, Wai Loon Chong, Tanzila Ahmed

*Abstract*— Data analysis on the New York City 311 Request Call using call request data provided by the state of New York. The dataset included the created date and closed data, thus allow us to find out how many cases were closed on a specific year. From there, we performed further analysis on on the data set to identify how many day it took to close a case. In addition, we also introduce the LA 311 call request and compare the result of the closed case with NYC 311 to determine which state perform better.

# INTRODUCTION

In the last 15 years, we have witnessed an explosion in the amount of digital data available – from the Internet, social media, scientific equipment, smart phones, surveillance cameras, and many other sources – and in the computer technologies used to process it. That is — Big Data. BIG data is suddenly everywhere. Everyone seems to be collecting it, analyzing it, making money from it and celebrating (or fearing) its powers. Whether we’re talking about analyzing zillions of Google search queries to predict flu outbreaks, or zillions of phone records to detect signs of terrorist activity, or zillions of airline stats to find the best time to buy plane tickets, big data is on the case. By combining the power of modern computing with the plentiful data of the digital era, it promises to solve virtually any problem — crime, public health, the evolution of grammar, the perils of dating — just by crunching the numbers.

In this project, we used the NYC 311 call request data from 2010 to 2017 and see which complaint type were most popular (Noise complaint) and from there we filtered the most popular complaint type every year, which was Heat/Hot Water complaint, between the 7 years, we mainly focused on the year 2015. The purpose of this is to observe the factor affecting the increasing in heat/hot water complaint.   
   
 Our next objective is to find out how efficient is the New York City in resolving the complaint. On a specific year (2015), we calculate the average closed day for all cases and observed the trend of the complaint call over a year. Finally, by comparing the result with the result from LA 311 dataset to determine which state has a better performance.

# methodologies

The data set ranging from 2010 to 2017 (approx 10gb) were used in this project. The data set contained all 5 borough of New York City. the data in year 2015 was filtered out and used the data in that particular year in our experiment.

As the data covered all five boroughs in the city, we categorize all the complaint according to their location and over a period of a year (12 months). From that we can investigate which borough has the most complaint call. To see the trend in the data, a scatter plot was drawn and we were able to see when the calls were increases and decreases. And then a function is created to find the average day to close a case for each borough and also all the borough combined.

From the data in 2015, the heat/hot water complaint was the most popular complaint. In this particular case, we analyzed to determine the factor that affect this case. First, we hypothesize that the greater the amount of restaurant the greater the complaining call would be. To get this result, we used a NYC neighborhood geojson file containing the longitude and latitude of all the location in New York City. The coordinates were extracted and spatial join was performed to on the 311 complaint call. From there, we analyzed the total heat complaint call and total restaurant in the neighborhood. With the result, spatial join was performed again to get the output of how many complaint call according to the neighborhood. The second hypothesis for this case was the greater the building age the greater the residence/business encounter the heat problem. Same approach was used, but each complaint call was mapped to exactly one building in the NYC building age dataset. A small polygon was created around each building from the buildings data set. Then, an intersection on the call location and the building location was performed using the Rtree algorithm and spatial joining. The final outcome was a list containing the building creation year and the number of complaint.

Lastly, for the LA 311 dataset, we applied the same technique like how the NYC 311call request data was process in the first part of this project. From the results, we were then able to determine and compare LA 311 with NYC 311.

# Result

After extracting the useful data from our datasets, we formalized them based on different categories and different times. After running those actions on Spark, we obtained our result and stored on HDFS.  
 There are over 15 million 311 requests from 2010 to 2016, the total number of different request type is shown as Figure 1 below. Since it is not easy to analyze all the 220 different requests one by one, we will focus on one or two major request.

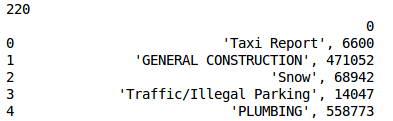


Figure 1 - Total Different Request Types

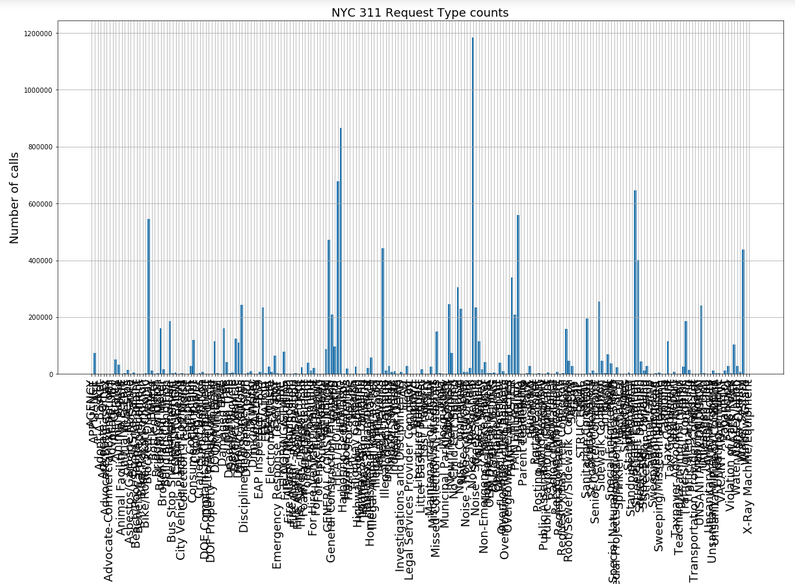


Figure 2 - Graph of all request type

Figure 2 is a messy chart to show all the request distribution from 2010 to 2016 in one plot. It is too small for a graph to show 220 request types at once. Because of this, year 2015 is picked to perform the analysis.

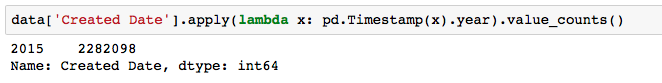


Figure 3 - Total 311 Request in 2015

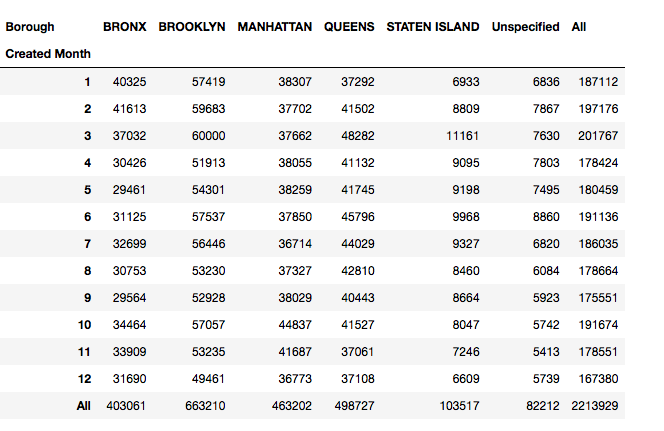


Figure 4 - 311 Requests Distribution

Figure 3 shows the total number of 311 request in 2015, which is over 2.2 million. And Figure 4 is the distribution of requests based on months and different boroughs.

From Figure 4, we can see the requests amount for different boroughs are quite stable over time. There is no month that has a sufficiently high request amount, nor a trend to show there is any relation overtime. Therefore, more conditions are needed in order to find the correlations within the 311 request data.

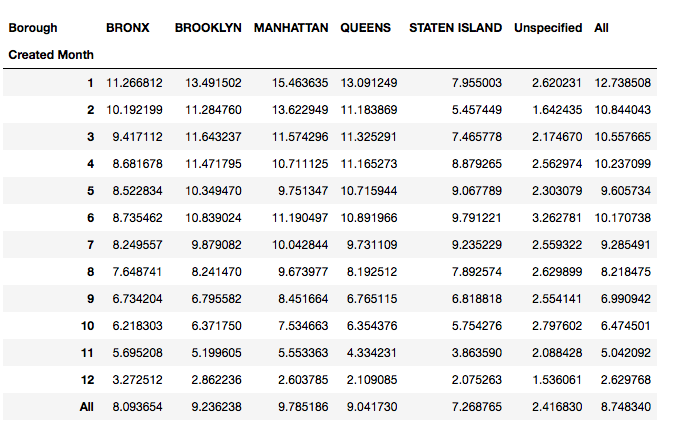


Figure 5 - Average Closed Time

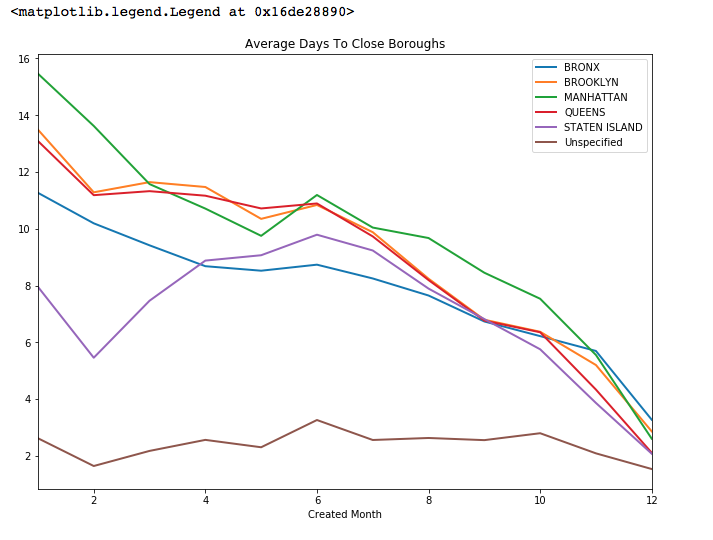


Figure 6

Figure 5 is a graph to show the average time needed to close a request. The maximum average time is more than 12 days, which is not so efficient. And the minimum average time is less than 3 days, which is acceptable considering the massive amount of request.

Figure 6 is a plot based on the average closed time. This is an intuitive graph to show the trend.

Observing the graph, ones can see there is a trend that the closed time is shorter in the later year. At the first few months of the year, the efficiency is low, but as the time goes toward the end of the year, the closed time is faster and faster. This can be due to some regulations or policies that staff needs to improve their efficiency. Or there might be another reason, the staff just closed a request, regardless the request is actually finished or not. The later reason is more possible, due to the further analysis in this report.

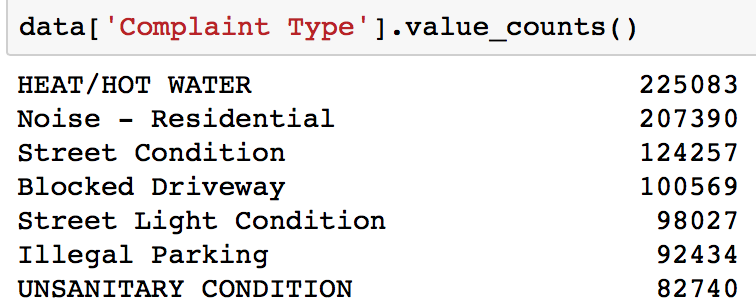


Figure 6 - Counts of Different Request Type

Figure 6 shows the most popular request in 2015 is Heat/Hot water complaint. Hence, a further analysis will base on this one category. We wonder what might be the reason to cause the large amount of the Heat/Hot water complaints. The following are the major assumptions to the cause, some large water consumption facilities such as restaurants nearby a neighborhood, the age of a building, over cold weather in winter, or pipes and the population is too high for one area.

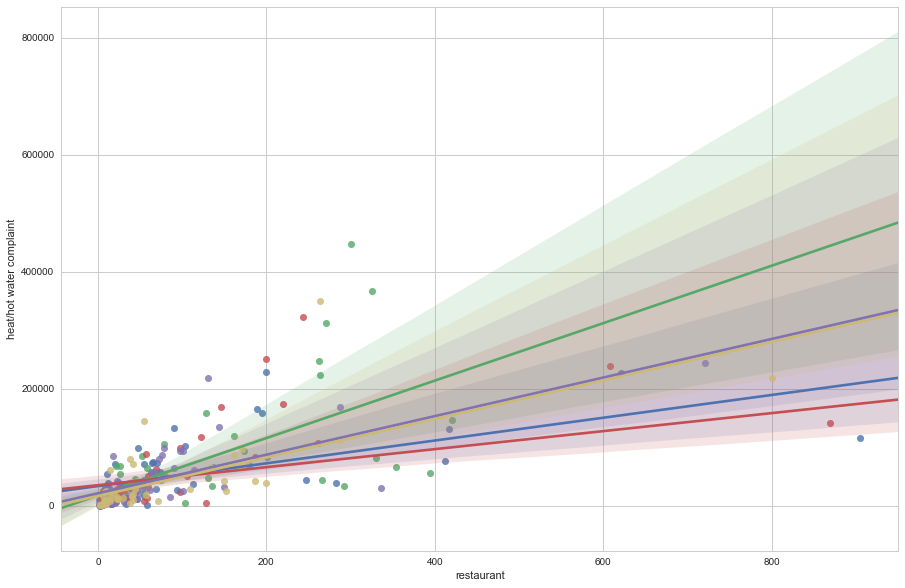


Figure 7 - Restaurants vs Heat/Hot water complaint

Figure 7 is the plot of Restaurants amount vs Heat/Hot water complaint. The different color of the dots represents the complaints over the different months. From the plot, there is no clear linear correlation. Even though the number of restaurants increased in one area, the total number of Heat/Hot water complaint doesn’t increase correspondingly.

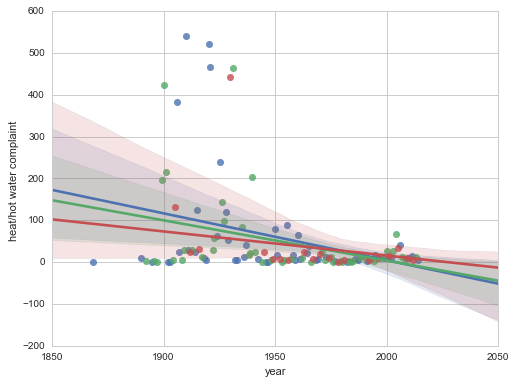
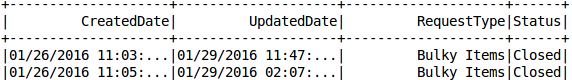


Figure 8 - Building Age vs Heat/Hot water complaint

Figure 8 is a plot of Building Age vs Heat/Hot water complaint. Ones can observe that there are some correlations. Most of the complaints lies on a linear line. Therefore, from the graph, building age could be one factor to cause the large amount of Heat/Hot water complaint.

The following part is a brief summary of the LA 311 request data, it is used as a reference to compare with the NYC 311 request data.



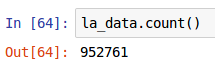
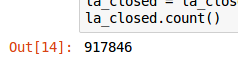
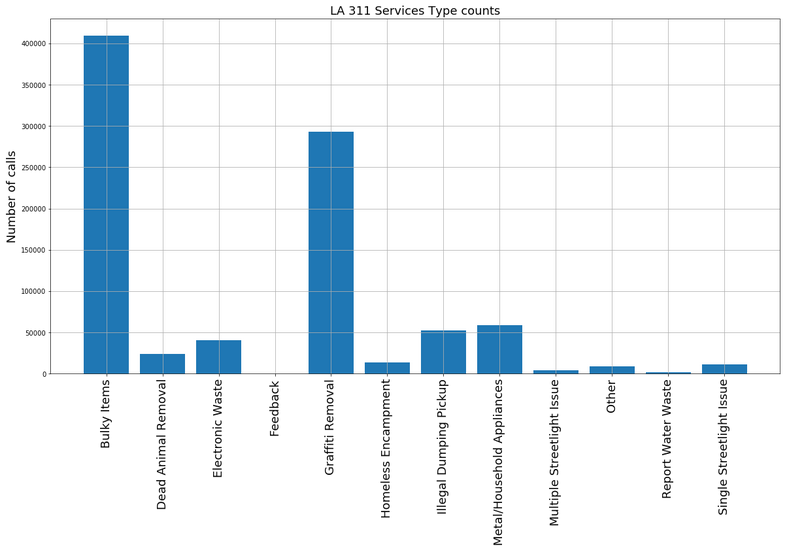
 

Figure 9 - LA 311 Data

For the LA data, the way to find the closed time for a request is using the UpdatedDate – CreatedDate. From the graph, the closed rate is 917846/962761 = 95.33%, which is unrealistically high. And for NYC, the closed rate is 97%. Because of these, we tend to believe the closed rate is fake, many requests are closed regardless of the real status.



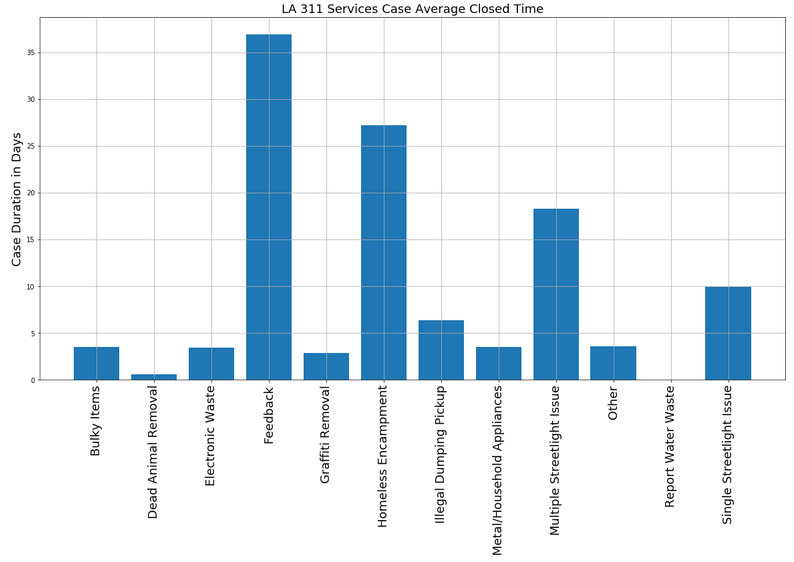


Figure 10 - LA 311 Request Type, count and closed time

Observing Figure 10, there are only 12 different categories of the 311 request, which will cause loss of detail. The most popular request type in LA is bulky items, which is about large garbage disposal. Feedback takes the maximum average time to closed.

challenges

The challenges that we faced are:

* It took long time to process the data locally because the data were huge
* It is really difficult to debug when we run our code on the cluster, it just throws a bunch of errors that we don't understand, so we have to make sure everything run correctly locally then we run it on cluster
* Most of the error come from the in-consistency of data type in the data set, we took a while to figure out what was the errors
* The LA 311 resource was very limited, we can’t really do any further analysis on the LA dataset

Although we encountered different type of errors even some of the errors we did not recognize, but overall, we learn how to run cluster on CUSP and spatial join technique, like combine two or more datasets with respect to a spatial predicate. The predicate can be a combination of directional, distance, and topological spatial relations. Last but not least, the most interesting part was we learn how to extract the coordinate out from a geojson file using Rtree. This come in handy when the data contain a lot of geometry information.

Conclusion

In conclusion, this project give us some idea on how the trend is on the NYC 311 data, based on Figure 6, the day to close a case was slowing down toward summer time and started to rise toward the end of the year. Although we did not find out what was the reason to cause the performance at the end of the year was better compare to the beginning of the year, but we are very curious to find out and would extend the project in our free time.

Besides that, based on Figure 7 the hypothesis; the greater the amount of restaurant the greater the complaint call would be rejected, because from the regression graph, we did not observe any correlation between them. While on Figure 8, the hypothesis is accepted because we can clearly see when the building age is older, the amount of complaining was also greater.

Lastly, we expect the performance of NYC 311 call request service was below average, but as we dig into the data we found out that the performance was excellent even compare to LA 311 call request (97% vs 95%). It was unfair because the dataset in LA 311 was very limited, it only contains few type of complaint type compare to NYC which had an enormous amount of complaint type (220). Overall, we would conclude that the performance on the NYC 311 call request is better.

References and Footnotes

## References

* <https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html>
* <https://www.weforum.org/agenda/2017/02/big-data-how-we-can-manage-the-risks>

1. [↑](#footnote-ref-2)