Name – Anshul Roonwal
SBU ID – 110554783
Course – CSE 590 Data Science Fundamentals
Stony Brook University
Submission Date – 8th October 2015

Introduction to 311 Data Set

Unlike 911 which is primarily meant for handling emergency calls, 311 is meant for non-emergency calls in US. In this report, we are only taking into consideration the complaints/request made by the citizens or the visitors of New York City only. There are various departments maintained by the government of US to handle different kind of requests. Whenever a request is made, the system captures a variety of data about the request. For instance, it stores the 'Created Date', 'Closed Date', 'Complaint Type', 'Agency Name' and other useful information. This information have been analyzed to observe interesting things about these complaints and aim at making the New York City better.

Source of data

The data of 311 requests is available for the public to use at
https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9?

Data Cleaning

Data received has a column that stores the zip code of the place from where the request was made. In some of the rows, I observed that the zip code did not belong to New York City and in some cases the zip codes did not even exist.

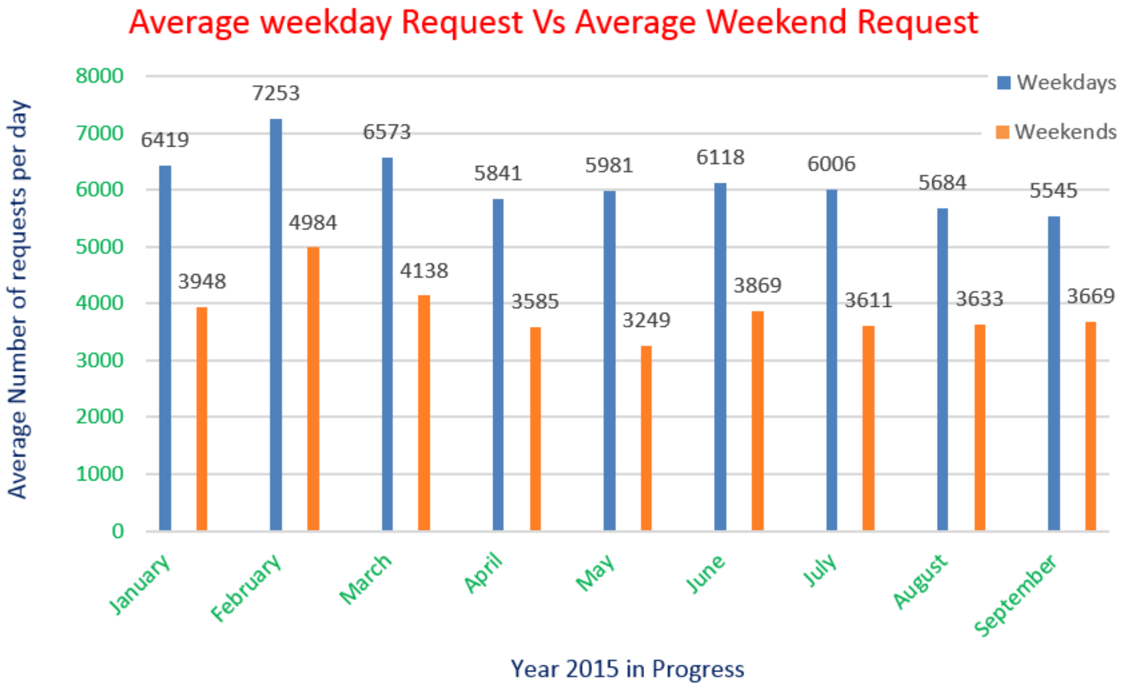
For instance, → Some rows mentions zip code as '99999'
However, the highest available zip code is "99950" KETCHIKAN.

Therefore while parsing the data, I filtered out the rows which did not belong to NYC. To know the zip codes of NYC, I referred the web @
http://nyc.pediacities.com/New_York_City_ZIP_Codes

Striking Observations

1) "NYC is Happy on Holidays"
2) 8th January 2015 saw a surge in 311 requests with most of them being the Heat/Hot Water Complaints. Very low temperature (8° F) was recorded that day.
3) Last day of each month receives less requests.
4) People of NYC are making most complaints about the Hot Water.
5) Economic Development Corporation needs more resources.
And a few more…

NYC complains less on Holidays and Weekends…..



Performing Hypothesis Testing

I have the data for the requests made on weekend (including holidays) and requests made on weekdays. I want to perform **hypothesis testing** to find out how significant is the difference in the average number of request made on weekends and the average number of requests made on Weekdays.

$A_{working} =$ The average number of requests made on Weekdays

$A_{Non\text{-}Working} =$ The average number of requests made on Weekends including holidays.

When to say, the difference is significant?

$$\text{If } A_{Non\text{-}Working} \geq (A_{working} + 50\% \text{ of } A_{working}) = 1.5 \text{ times } A_{working}$$

If the above equation holds true, we say that the difference is significant since a difference is of 50% which is big enough for its significance.

H_0 = Null Hypothesis→ “There is **NO SIGNIFICANT DIFFERNCE** in the average of the number of requests made on working days and non-working days (consisting of public holidays and weekends).”

H_1 = Alternative Hypothesis→ “The average of the number of requests made on working days are around 50% higher than the average of the number of requests made on Non-working days ”

Significance Level (α) = 0.05

Steps followed

1) Starting from the 01-January-2015, calculated the average of request of both working days and non-working days for every month in 2015 till date.
2) Perform ***T-Test*** **-** Two sample assuming equal variance for testing the difference between two means→
   a) Mean of the average number of requests raised on Working days.

      ⇨ 6157.77

   b) Mean of the average number of requests raised on Non-Working days.

      ⇨ 3854
3) 40% of Mean of $A_{working}$ = (0.5)x3854 = 1927 [Our hypothesized mean of difference]

Perform the T-Test: Two-Sample test

	Working	*non*
Mean	6157.777778	3854
Variance	274542.1944	243389.75
Observations	9	9
Pooled Variance	258965.9722	
Hypothesized Mean Difference	1930	
df	16	
t Stat	2.516877223	
P(T<=t) one-tail	0.011441604	
t Critical one-tail	1.745883676	
P(T<=t) two-tail	0.022883208	
t Critical two-tail	2.119905299	

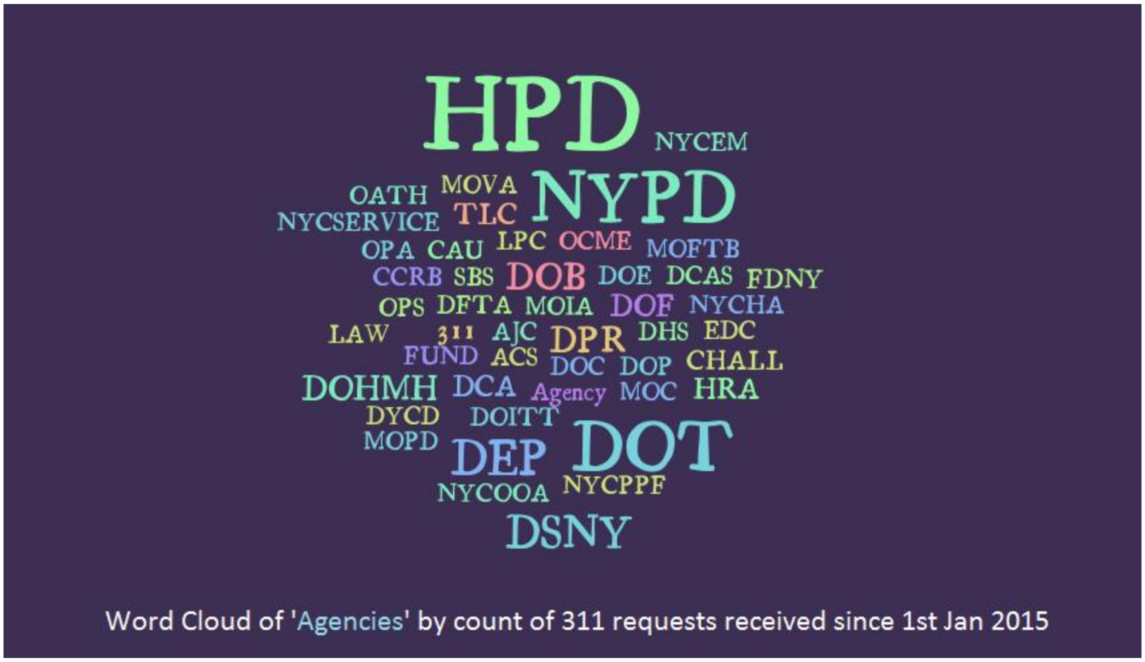
Tool used to perform T-Test is MS-Excel

Conclusion →

Since ‘**p-value**’ of 0.022 is significantly lower than the significance level of α = 0.05, therefore we can reject the null hypothesis and we can conclude that our alternate hypothesis is true.

So, it is safe to say that – “The average of the number of requests made on working days are around 50% higher than the average of the number of requests made on Non-working days.”

Ranking the Agencies on amount of requests received by them.



Word Cloud of 'Agencies' by count of 311 requests received since 1st Jan 2015

A few of the highlighting agency names are →

HPD - Housing Preservation & Development; DOT – Department of transportation

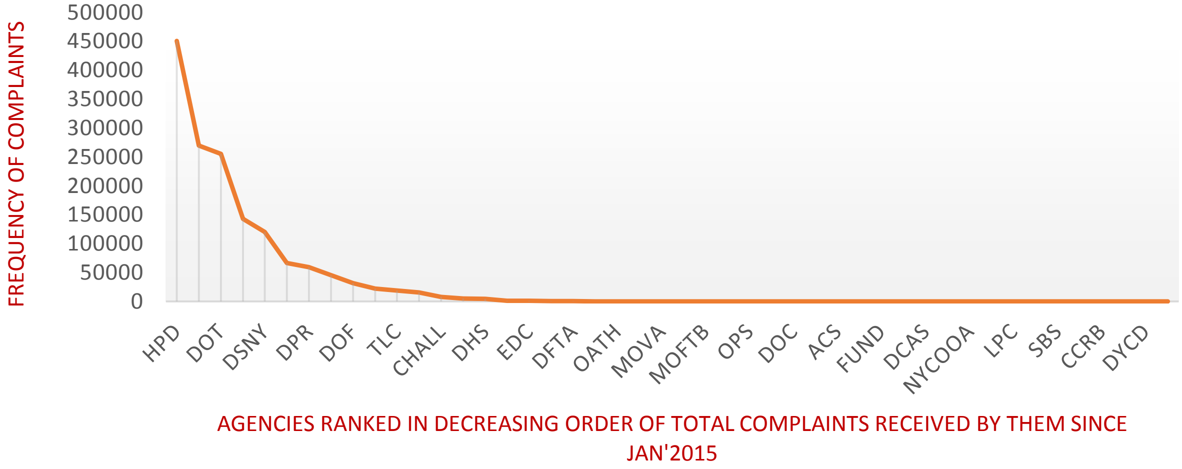
NYPD – New York City Police Department; DSNY – Department of Sanitation, New York

Reference for Word Cloud - http://worditout.com/word-cloud/

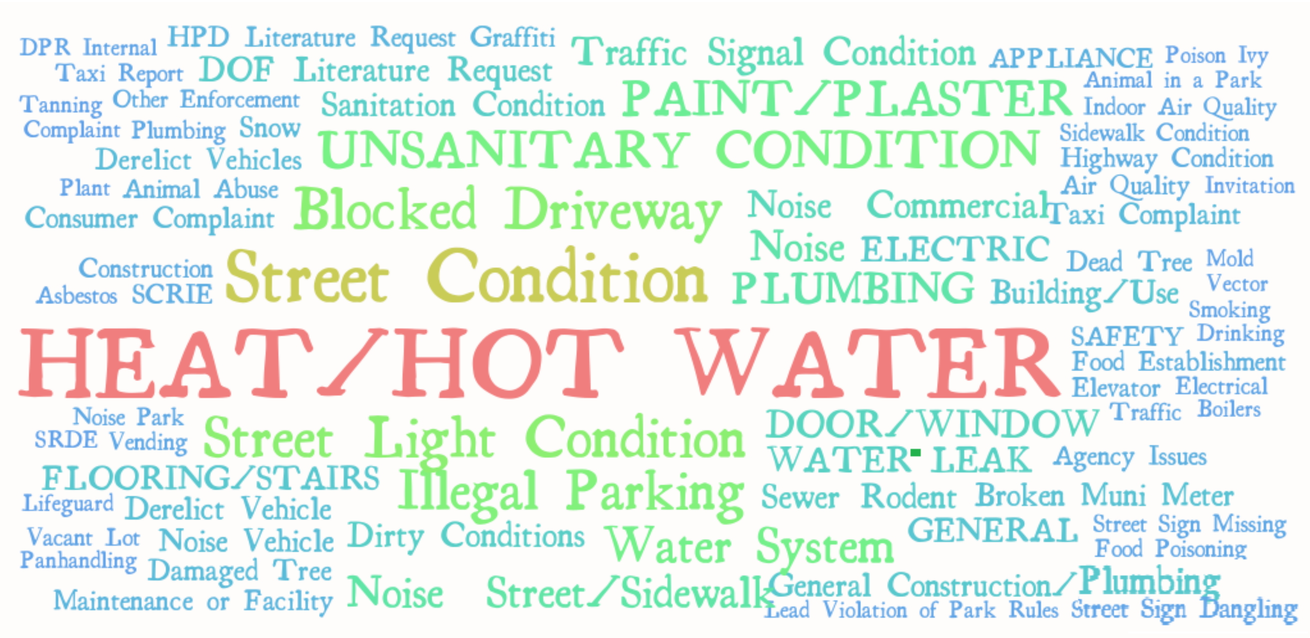
Power Law is a good fit

The frequency of complaint to different Agencies from the beginning of 2015 follows a power law distribution.

Frequncy of complaints to different Agencies follows Power Law Distribution

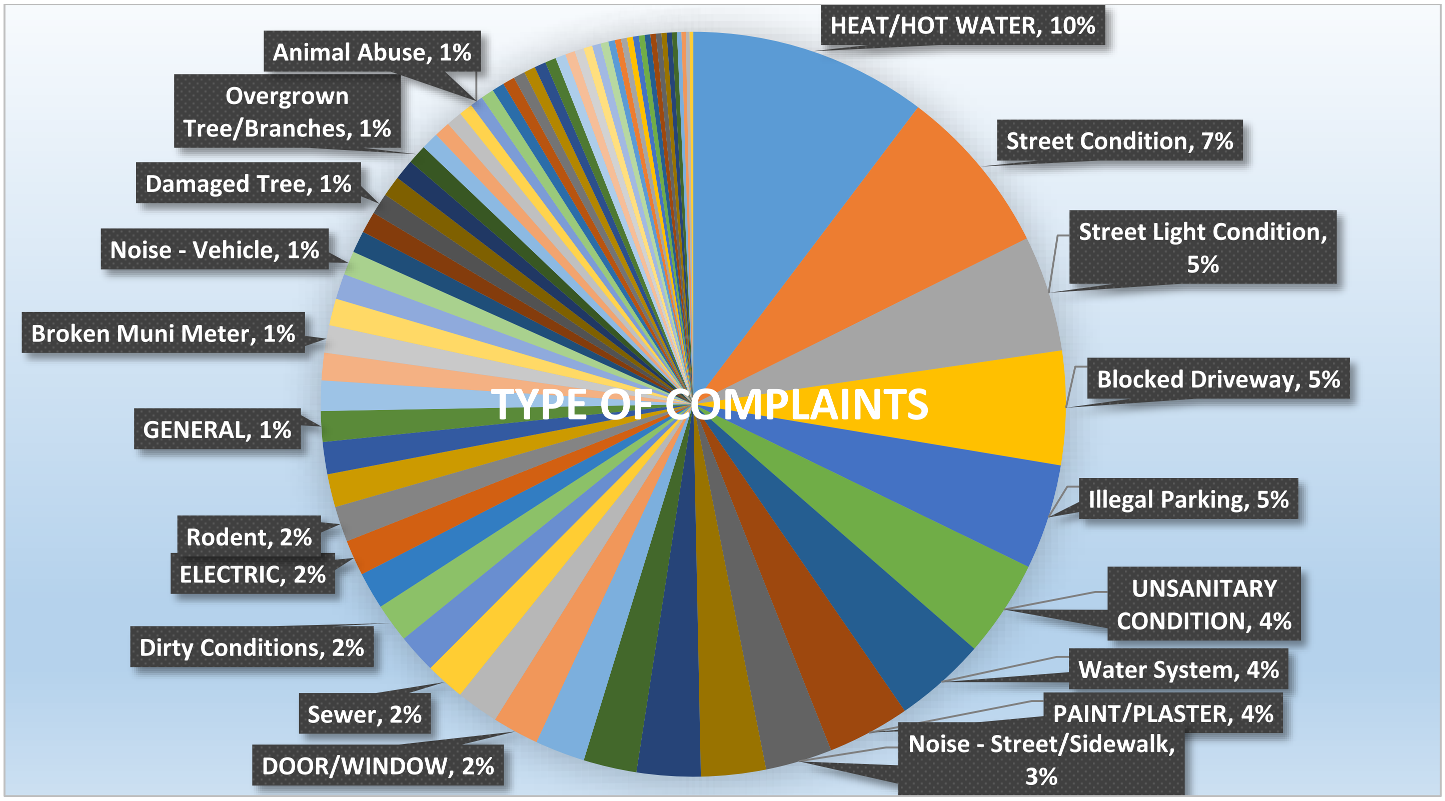


What is it that people of NYC are complaining about the most?



Word Cloud on "311 Complaint Types" since 1st Jan 2015

The highest number of Requests received were related to Heat and Hot Water.
Since January 2015, total such requests received = 150162 (10% of the overall requests made).



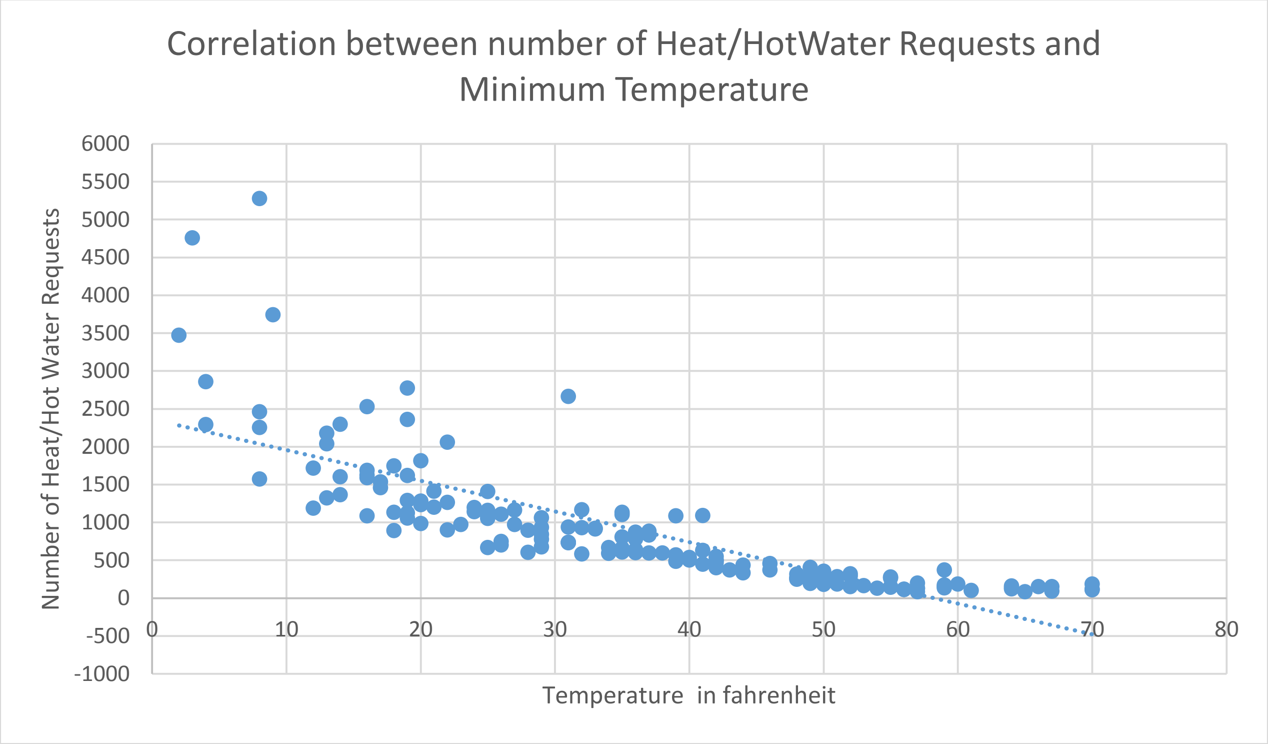
The chart shows the major type of complaints that were made using 311 and their distribution.

Using Linear Regression

After finding out that the maximum requests raised were of Heat/Hot Water, I tried to find if there exists any correlation between the numbers of Heat/Hot Water requests made and the minimum temperature for that day.

For the temperature of each day since the beginning of this year, I had to refer to source other than NYC open data @ http://www.accuweather.com/en/us/new-york-ny/

Collecting readings of minimum temperature over all the days along with the number of Heat/Hot Water requests made since January 2015, I found a linear correlation between them which has been plotted below.



Why the slope of the line is negative?

The correlation here is given by this expression below →

$$\boxed{\text{Number of Heat/Hot Water Requests} \quad \alpha \quad \frac{1}{Temperature\ for\ that\ day}}$$

Since there exists an inverse proportionality between these two parameters, the slope is negative. But it still shows all the properties of linear regression.

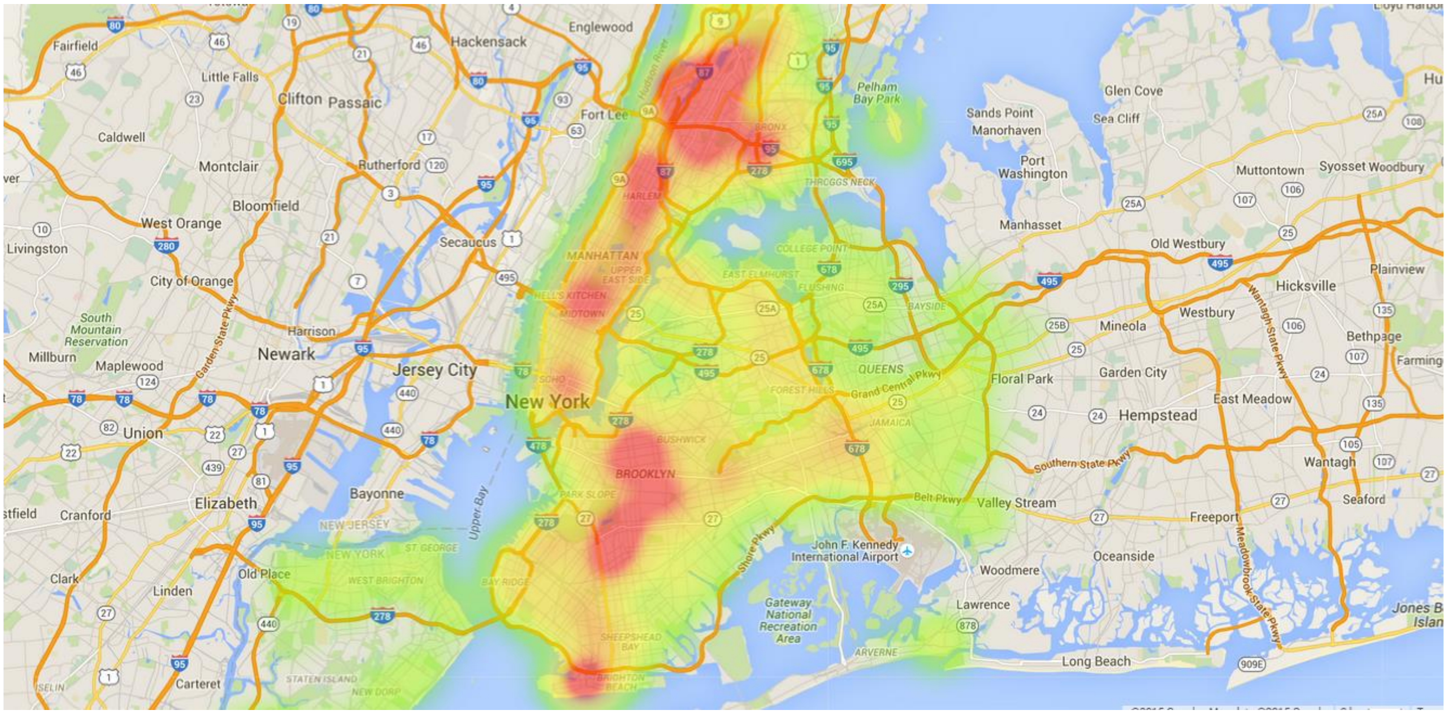
Which Department needs more resources?

The department that takes the maximum time to close a request is the one which needs more resources.
When a request is made, its creation date is logged in a 'Created Date' and the day it was resolved/closed, is logged as 'Closed Date'. The one that has the highest average difference in the closing date and created date is the one that is taking most time.

Name of the department	Average time to close a request (in hrs)
Economic Development Corporation	388.15
Fire Department New York	285.71
Department of Education	264.91
Department for the Aging	210.02

Economic Development Corporation need more resources. What these resources could be?
One of the reasons could be that these departments are understaffed. If so, the government of NYC can appoint more human resource in those departments.

Heat Map shows the regions making highest complaints using 311.



The data Set provides latitude and longitude values of the location from where a request was made. Using that information and using Google Fusion Tables, the above heat map was prepared.

Conclusion→

If NYC wants to take any action on improving the quality of services of its department, this heat map list outs the regions where it should start from. The regions shown in red color are the ones from where the number of requests made were more in comparison to the regions in yellow and green.