Mini Project 2: Open NYC Data

Analysis in this report is for 311 calls database, which stores the complaints registered with various agencies during the course of time.
Some points regarding dataset and report:

A) Report is organized in the form of data analysis and Findings with that analysis using the visuals.
B) Three year data (2015, 2014, and 2013) has been used for the analysis.
C) Data of 2015 is from January to October while 2014 and 2013 data is for all 12 months.

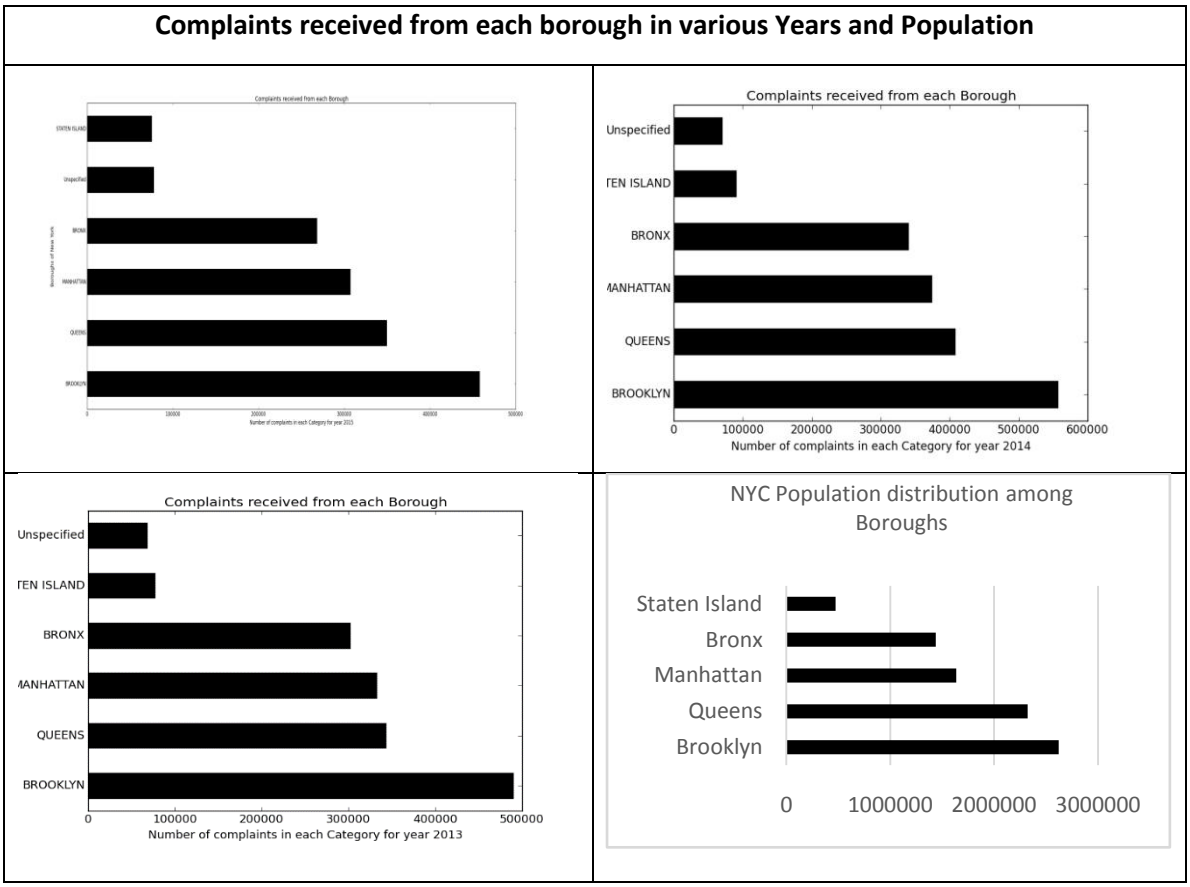
Let's begin:

1) To Understand the pattern of complaints viz a viz each Borough chart has been drawn with Boroughs at each scale and total number of complaints at other.From this visual it is clear that Brooklyn and Queens always have more complaints while Staten Island fall at least side.

   Findings:

   a) When we compare complaint graph with population graph, there is a congruence between the two. So number of **complaints are more or less proportional to Population in different years**.
   b) One important observation here is Queens still have less number of queries compare to it's population. So **study of Queens can help reduce problems in other boroughs**.
   c) 2015 total complaints till October: 1535266
      2014 total complaints: 1836978
      2013 total complaints: 1610644

Complaints received from each borough in various Years and Population

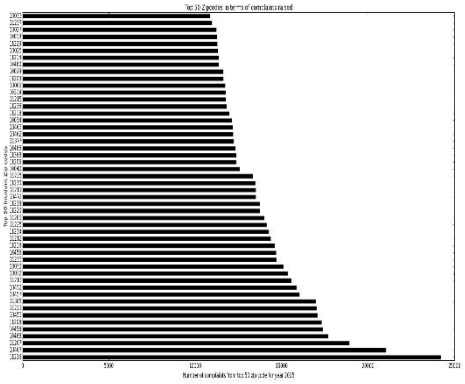
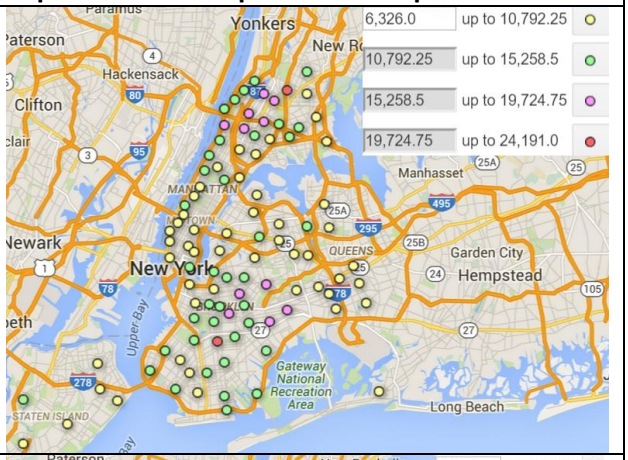
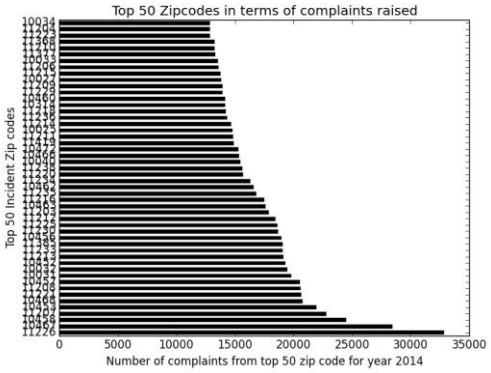
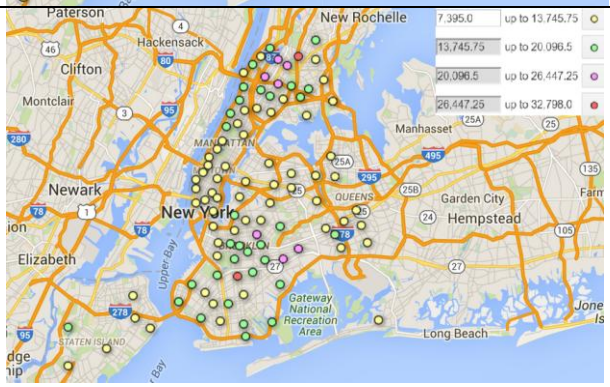
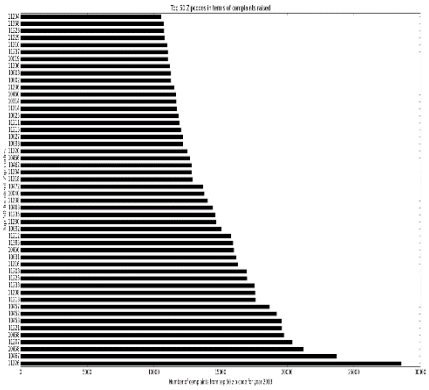
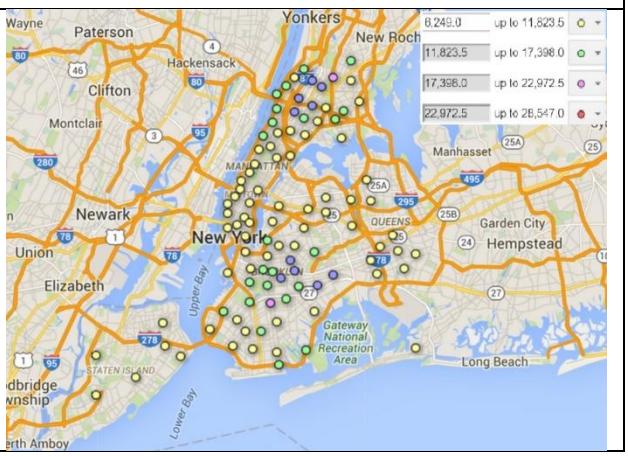


Rishabh Agrawal (110467487)

2) Next let's find out the problem areas in NYC in recent years using map and top 50 list:
Finding:

a) Onequick observation here is that zip 11226 is topping the charts in all years consecutively.

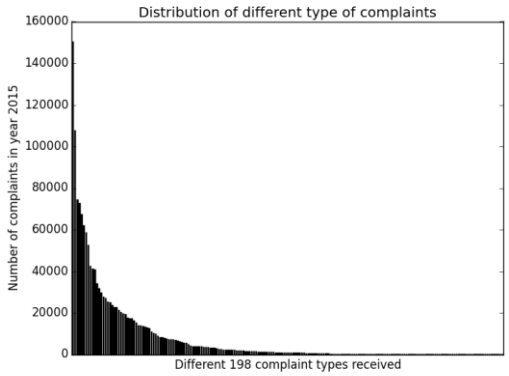
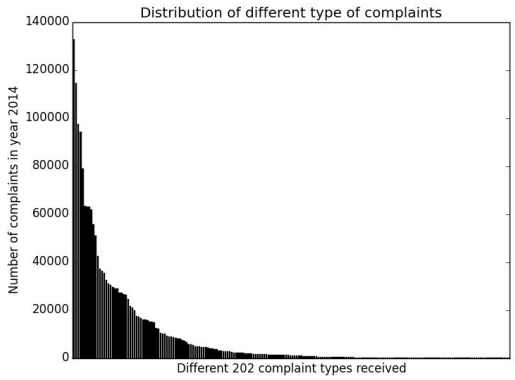
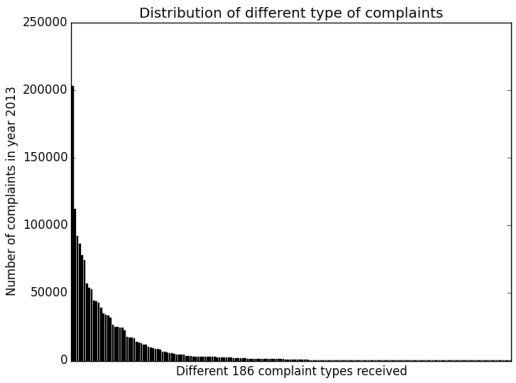
b) Also there are some other repeating Zip codes, which should be the concern areas for various agencies.

	Top 50 Incident Zip codes vs Complaints	**Top 100 Incident Zip codes on map**
2015		
2014		
2013		

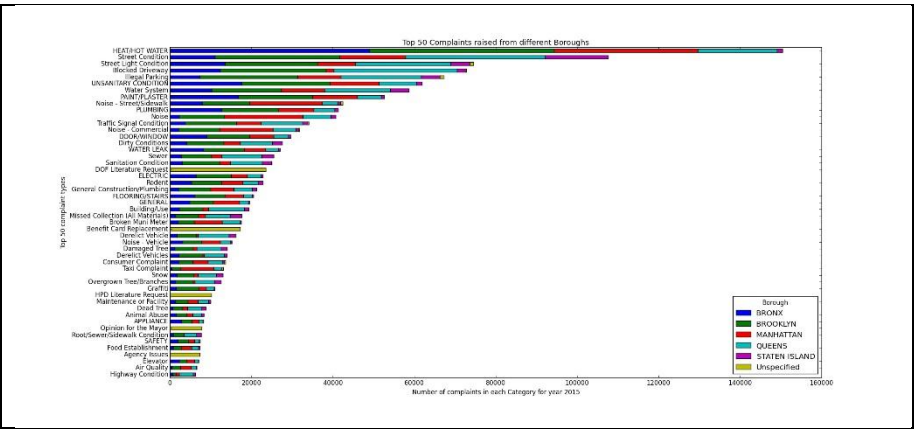
3) Next target is to find what bothers the people! So let's draw charts of distribution of complaints vs type of complaints.
Findings:

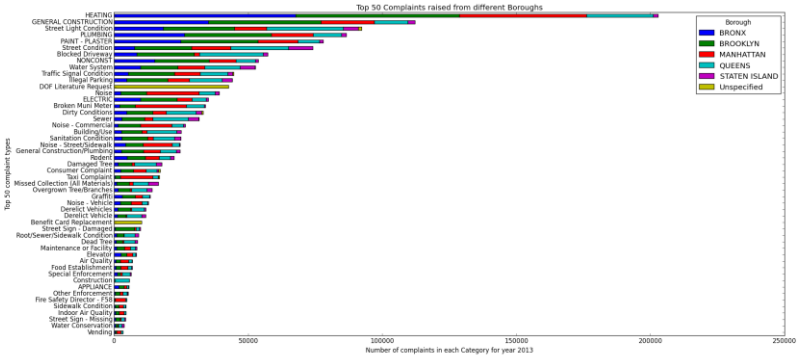
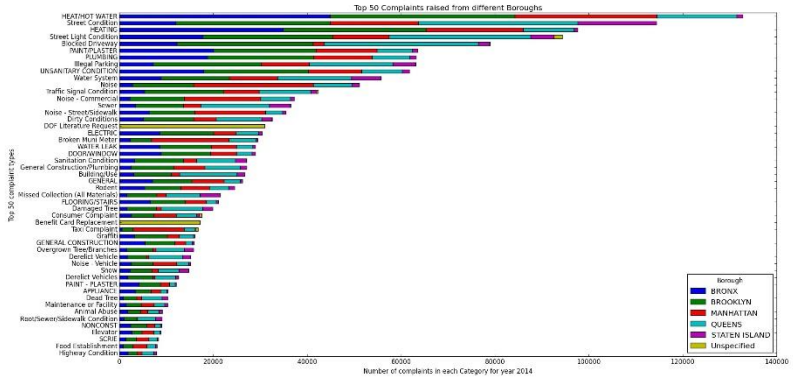
a) Distribution shows there are some problems which are getting high amount of complaints while some other are getting very few complaints.

b) Also some new complaints are getting added in each successive year.
Year: Different type of complaints.
[2013: 186] [2014: 202] [2015: 198 (Up to October)]

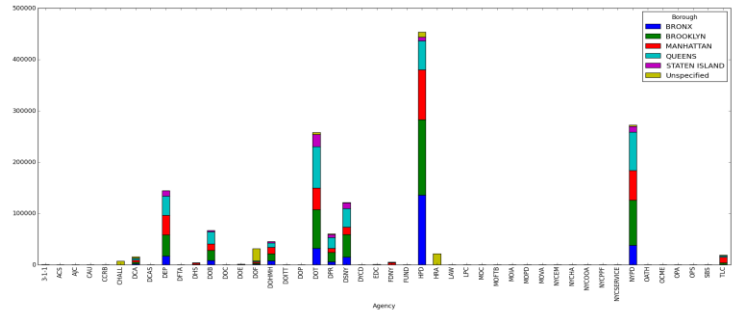
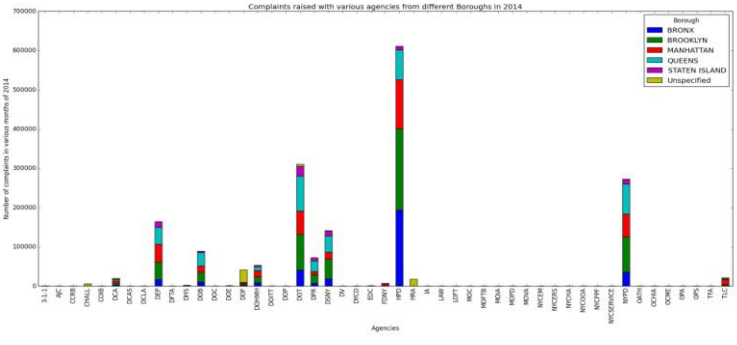
2015	2014
	
	**2013**

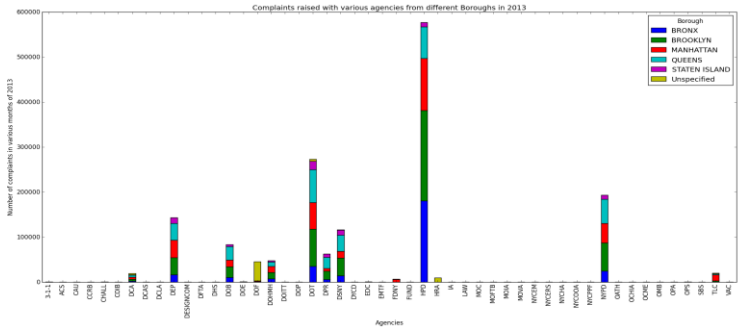
4) Now as distribution is not uniform, let's plot Top 50 complaints stacked for different boroughs.
Findings:
   a) Recurrent problems which people faced are Heat, Hot water, street conditions, Blocked Driveway.
   b) There are changes in the order but these are some problems which keep bothering people in different years.
   c) Also it can be seen in the graph that some problems are dominant in specific borough like Illegal parking and blocked driveways are big problems in Brooklyn while Sewer is a big problem in Queens.
   d) Similar observations can be drawn from the graphs and agencies can choose the focus area of their respective work.

2015	

2014	
2013	

5) Ok, Let's move to find **distribution of complaints among various agencies correspond to boroughs Findings:HPD, NYPD and DOT** are the one getting most complaints. Either time to allocate more resources or to fix problems in advance for getting less complaints!!

2015	
2014	

2013	

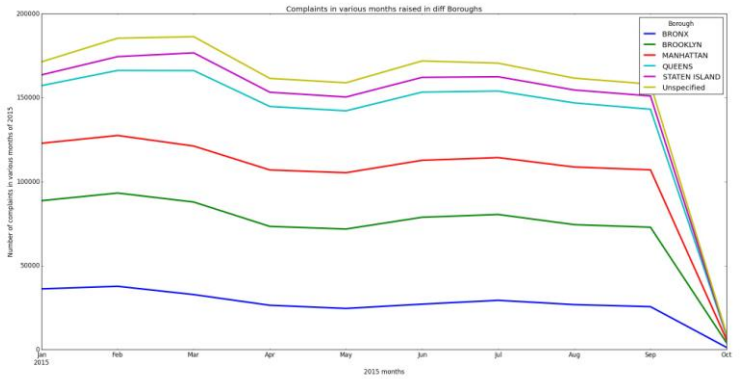
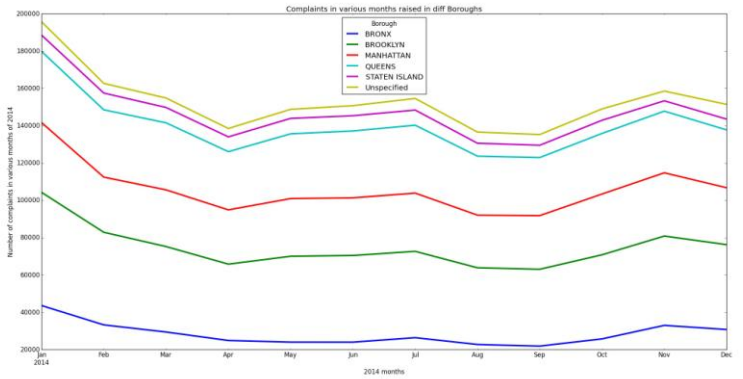
6) Now try to do some Time Series analysis. So here comes the graph of complaints in various months stacked for Boroughs.

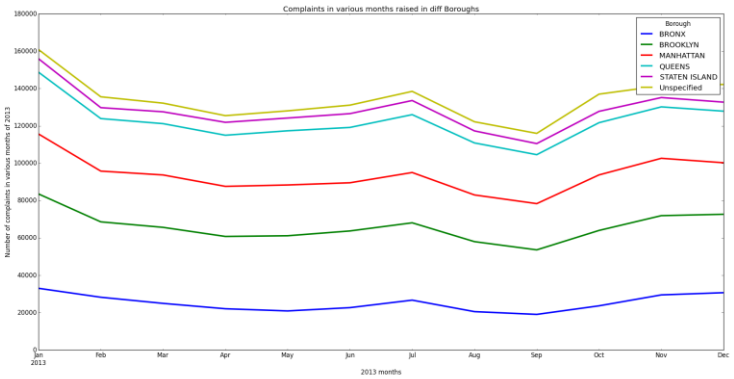
Findings:

1) April, August, September are the months in which Agencies personals can take a break as there are least number of complaints in these months. Also these are the months of spring and autumn when weather is not too harsh.

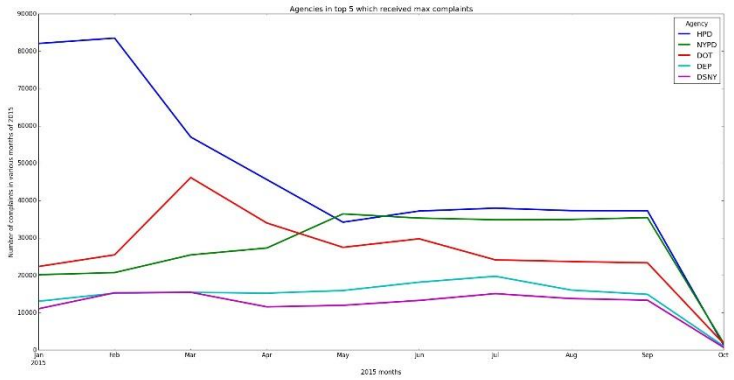
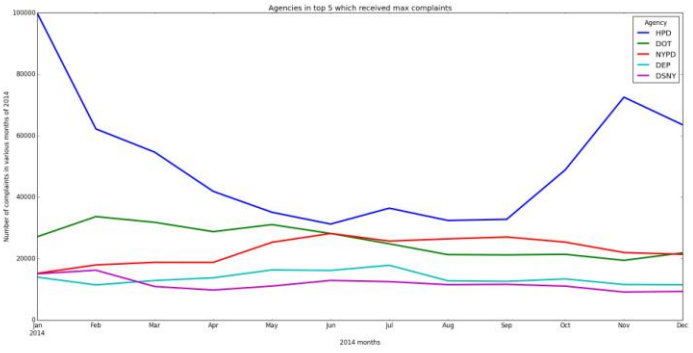
2) While it seems all leaves need to be canceled during peak of winters and summers i.e. Dec, Jan and Jun, July. As these are the months getting surge in the complaints.

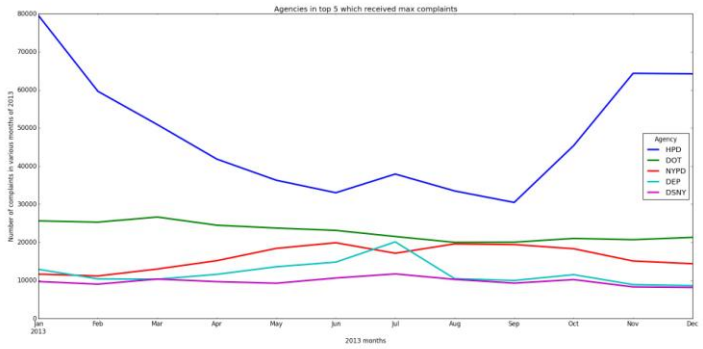
3) So there is a proportionality relation between weather conditions and number of complaints as per the graphs.

2015	
2014	

2013	

7) Next target is to see which agencies are getting flooded with the complaints in various months:
Findings:
   a) Complaints for **HPD** (which gathers highest complaints among all agencies) need to gather strength for winter as this is the time where we can see peaks in graphs for all years.
   b) While curve for rest four other agencies are more or less flat means they are getting same amount of calls throughout the year. So they can plan more easily as they know that calls are not going to flood.

2015	
2014	

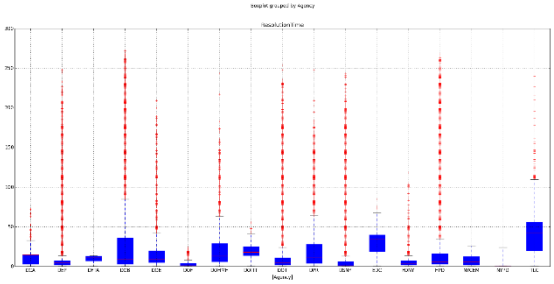
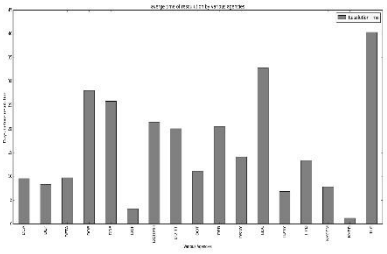
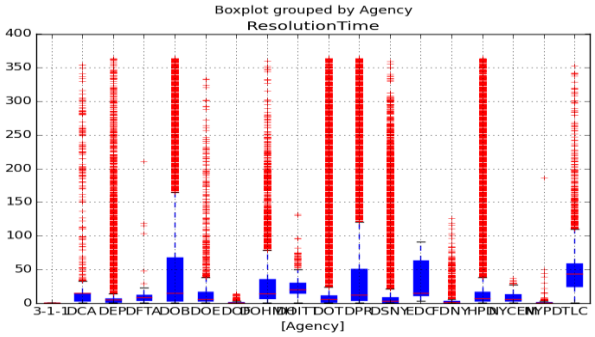
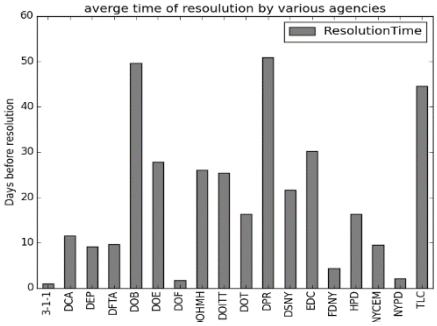
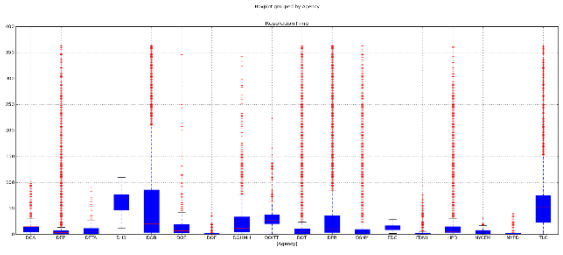
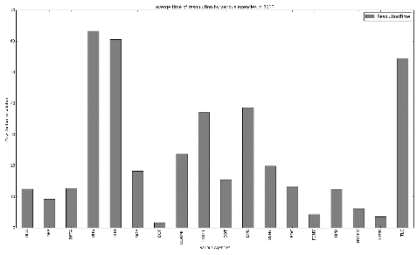
2013	

8) Now let's try to find out the **resolution time of a problem for different agencies**:
Findings:

a) From Average resolution time we can see that there are some agencies which are consistently doing poor on this front and average timing are varying from 1 day to 60 days.

2013: DHS, DOB, TLC
2014: DOB, TPR, TLS
2015: EDC, TLC, DOB

b) There are outliers which we can see in box plots drawn but average is in approx. 50 days' time frame.

Year	Distribution of resolution time	Average resolution time
2015		
2014		
2013		

Some database anomalies found:

1) Agency DSNY- Department of Sanitation is not reporting correct closed date for various service requests of Graffiti. All these closed date has been reported in future.

Unique key	Created Date	Closed Date	Status
16774392	06/01/2010	06/17/2201	Closed
24762224	04/15/2013	08/26/2016	Closed
30130047	05/12/2015	08/04/2016	Closed

Also some of the End dates are mentioned as 1900/01/01, which is not meaningful.

2) Also there are many Zip codes information reported as Unknown or N/A.

References:

Time Series:

http://earthpy.org/pandas-basics.html

Stacking and unstackinggroupby:

http://pbpython.com/simple-graphing-pandas.html

NYC data:

https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

Visualization:

http://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-box-return

Rishabh Agrawal (110467487)