homework vi

Muaamar Mohammed (813007612)
Ali Mohamdi (358000432)
Abby Gozun (62809940)

June 29, 2019

Executive Summary

- The data were gathered by NYC government through NYC 311 service and started from 2003 until mid-2015.
- Overall, there are a total number of 9124937observations and 56 variables in the data on hand from March 2003 until April 2015 including the duplicates and columns that have missing values.
- The total number of duplicates are 853538 . By eliminating duplicates we found out that there are 8271377 observations only.
- Records for Boroughs was improved using official dataset of zip codes to map the respective location where the requests were coming from to reduced unspecified values.
- Some graphs and diagram were used to visualized the number of agencies that have been involved from 2003 until 2015. In this study, we proved that records from 2010 onwards shows consistencies on the record compare to the previous years.
- The NYC government increased 37.5% on agencies to support on 311 service from 2010 onwards while complaint types increased approximately 67%.
- 30% of complaints are being resolved within 24 hours while the rest takes from 1 to 20 or more days.
- The complaint with missing values takes longer time for resolution.
- Blocked Driveway and Illegal Parking were consistent to be the fastest to resolved in all boroughs which proved that NPYD is the fastest agency to resolve these complaint types.
- Although Staten Island has the lowest population, it has receded the biggest number of complaints among others (16 complaint per 100 people) , While Bronx recorded the lowest number of (12 complaint per 100 people)
- Complaint types can be clusterd into small clusters (e.g Street light Condition & Illegal Parking) or large clusters (e.g Dirty Conditions, Sewer, Street light Condition, Illegal Parking) based on the business needs and purpose of clustering.

Introduction

Data has set a trend in the world today by motivating professionals to pursue learning data analysis. One effective way to assess the business processes performance is by analyzing the data it generates and interpreting its results. Students of data science tend to practice using datasets that can enhance their analytical skills.

This data exploration is an attempt to assess the performance of NYC government on their information hub called NYC 311 service from different aspects by doing visualizations on the points that we considered important.

Due to the increasing population of New York City, it is also obvious that the demands of NYC 311 service are also increasing. Therefore, assessing its performance can point out the aspects that needs improvements to serve the residents better.

Context (Business Problem)

The dataset that we are focusing in this study is a subset of New York's 311 service where in 2010, the NYC Open Data has made all NYC 311 service requests and complaints publicly available. The information hub was launched in February 2003 with an average calls of 2126 per day and was designed to filter non-emergency calls away from the emergency phone line, 911. The data collected is widespread; we can see agency's information, incidents, complaints and geospatial coordination.

The data on hand were generated from 2003 until mid-2015 with lots of inadequacies. However, with the current population of the city and the increasing number of service requests and complaints, we aim to improve the system by using data analytics to eliminate the redundancy of some information and generate intact records in the dataset to further progress on attending the requests of the residents of New York City.

About The Data

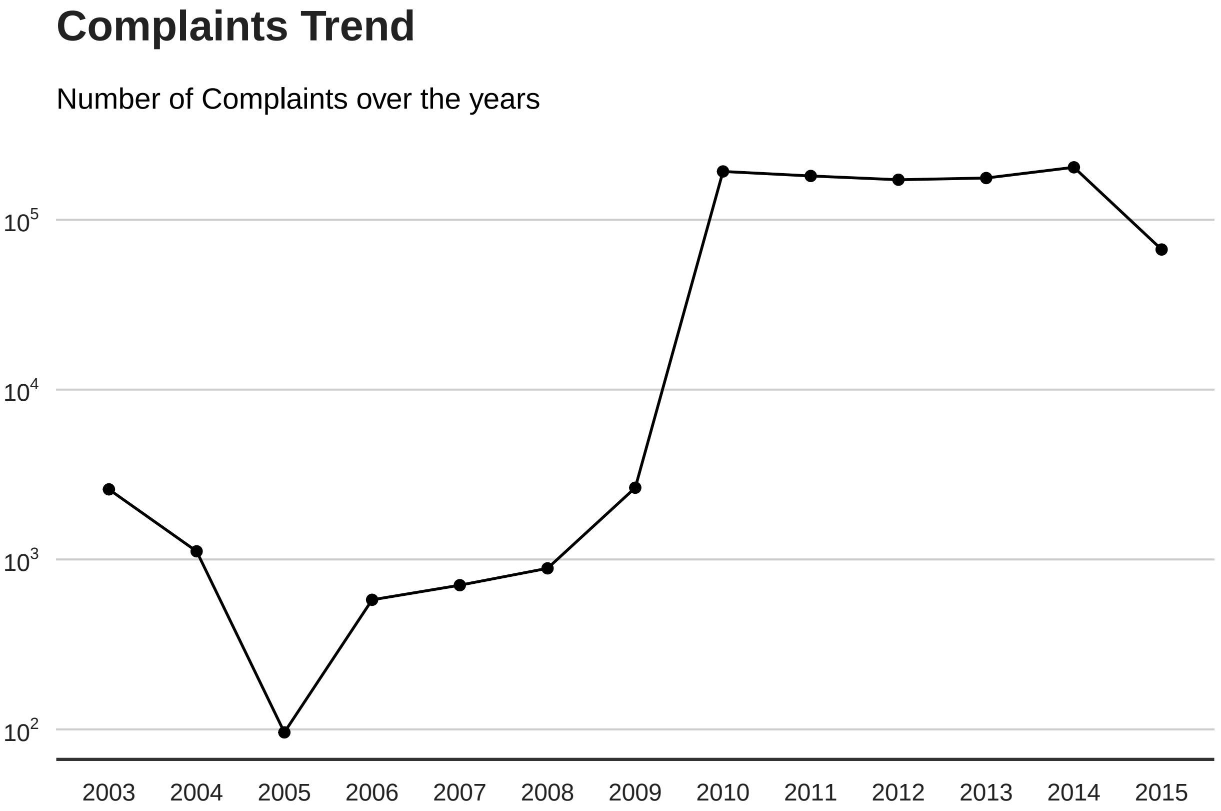
By searching online, we found that NYC 311 runs on a knowledge database that houses over 7000 pieces of information on over 3600 services from various city agencies and non-profit organizations. 311 services are available in 180 languages. Calls are answered 24 hours a day, 365 days of the year. Calls can also be made via skype and submitted over the web. Service requests can range from inquiries regarding trash collection, voting locations, businesses licensure, and even parking tickets, and complaints can include a wide range of issues including, apartment heating and cooling, street and sidewalk conditions, noisy neighbors and surroundings, and many more.

There are a total number of 9124937 observations and 56 variables in the data on hand for the period from Mar-2003 till Apr-2015 including the duplicates and columns that have missing values. In the column of unique key, there are 22 keys that were recorded more that one time and we found out that there are c contains duplicate. By eliminating duplicates, there are approximately 8271377 observations only. In addition to that, some columns like Boroughs with unspecified values was improved using official dataset of zip codes to map to the respective Borough and replace it. That helps in fixing 72% of the unspecified Boroughs.A sample of 1,000,000 incidents of the whole data set was used as a base for the analysis, with the missing values and anomalies excluded.

Findings

Data Distribution Over Years:

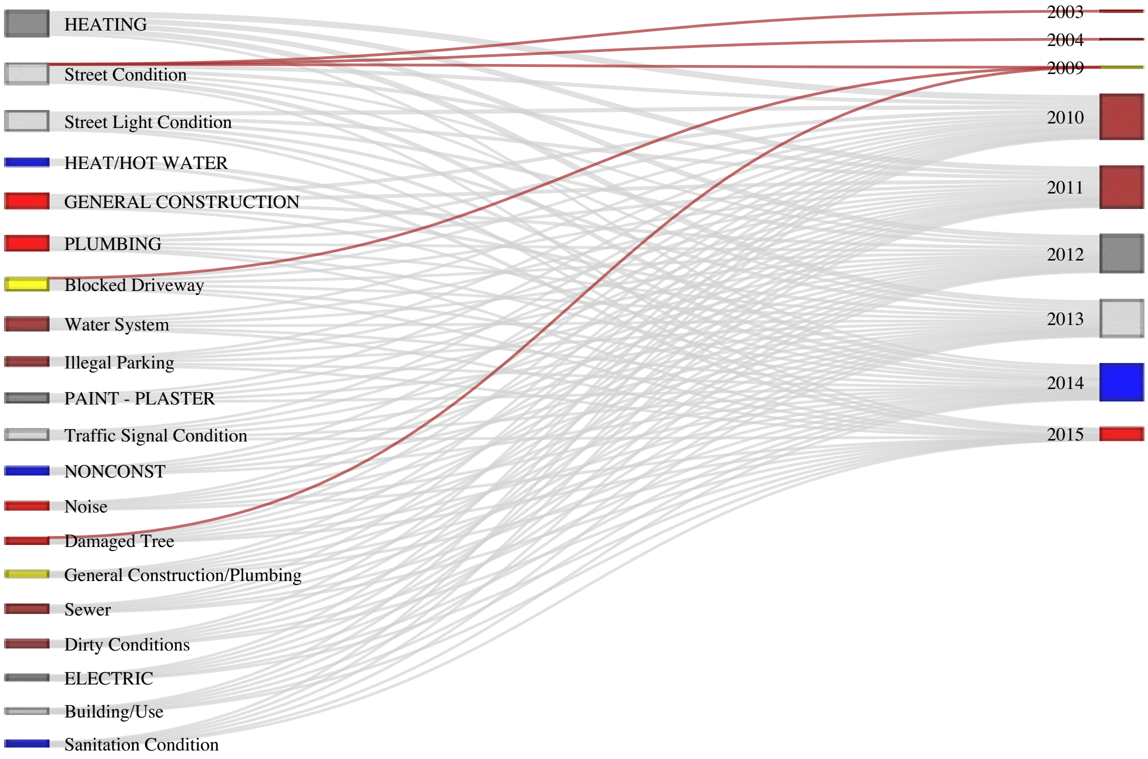
We discovered that number of complaints had increased slightly from 2003 to 2009 with a drastic escalation in 2010 continuously having same level on the years onwards.


Complaints Trend

Number of Complaints over the years

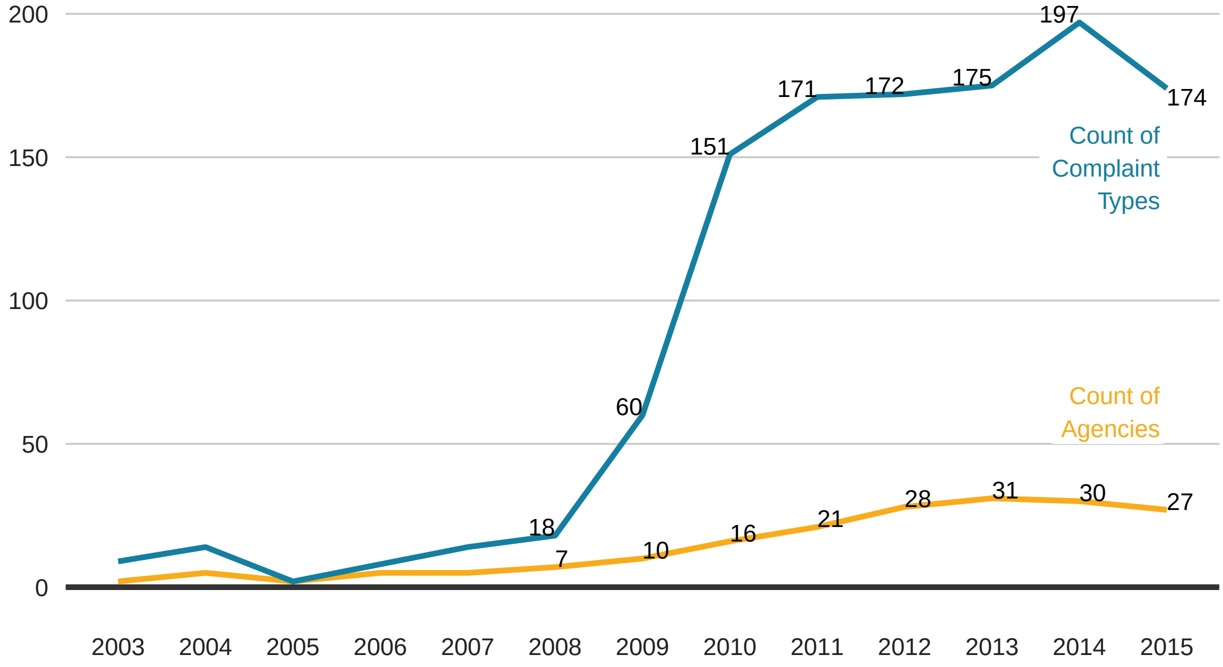
Note: In order to properly observe the change in number of complaints over the years, we used a logarithmic scale in the total complaints variable (y-axis). This is due to the difference in 2010. On a linear scale, a change between two values is perceived on the basis of the difference between the values. On a logarithmic scale, a change between two values is perceived on the basis of the ratio of the two values.

That dramatic rise in the number of complaints is confirmed to be a result of more agencies joined the NYC311 with more non-emergency complaint types handled through the shared call center. Below, Sankey diagram displays flows and complaint types and their quantities in proportion to years. The width of the lines indicates magnitudes, so the bigger the line, the larger the quantity of flow complaint type in that certain year.



Sankey diagram above confirmed clearly that less complaints types were introduced before 2010 with less agencies involved in the NYC 311. The thin brown line shows that only few number of complaint types by only 3 agencies were introduced to be handled through the shared call center. Major complaint types including top 10 complaint types like Heating, Street Light Conditions, Pluming, General Construction, Water System, Paint Plaster were introduced only from 2010 onwards.

The line chart below shows the service adoption by agencies over years:



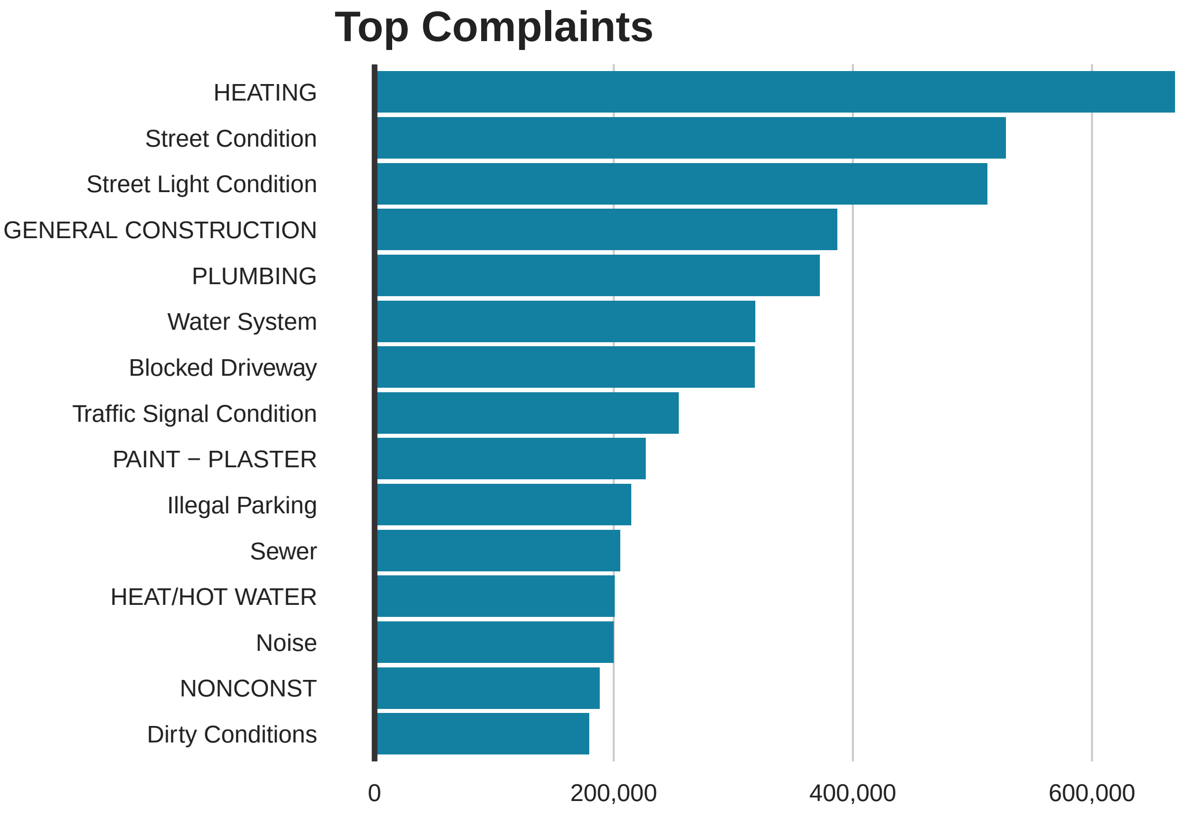
Agencies and Complaints Change

The change in number of agencies and complaint types over the years

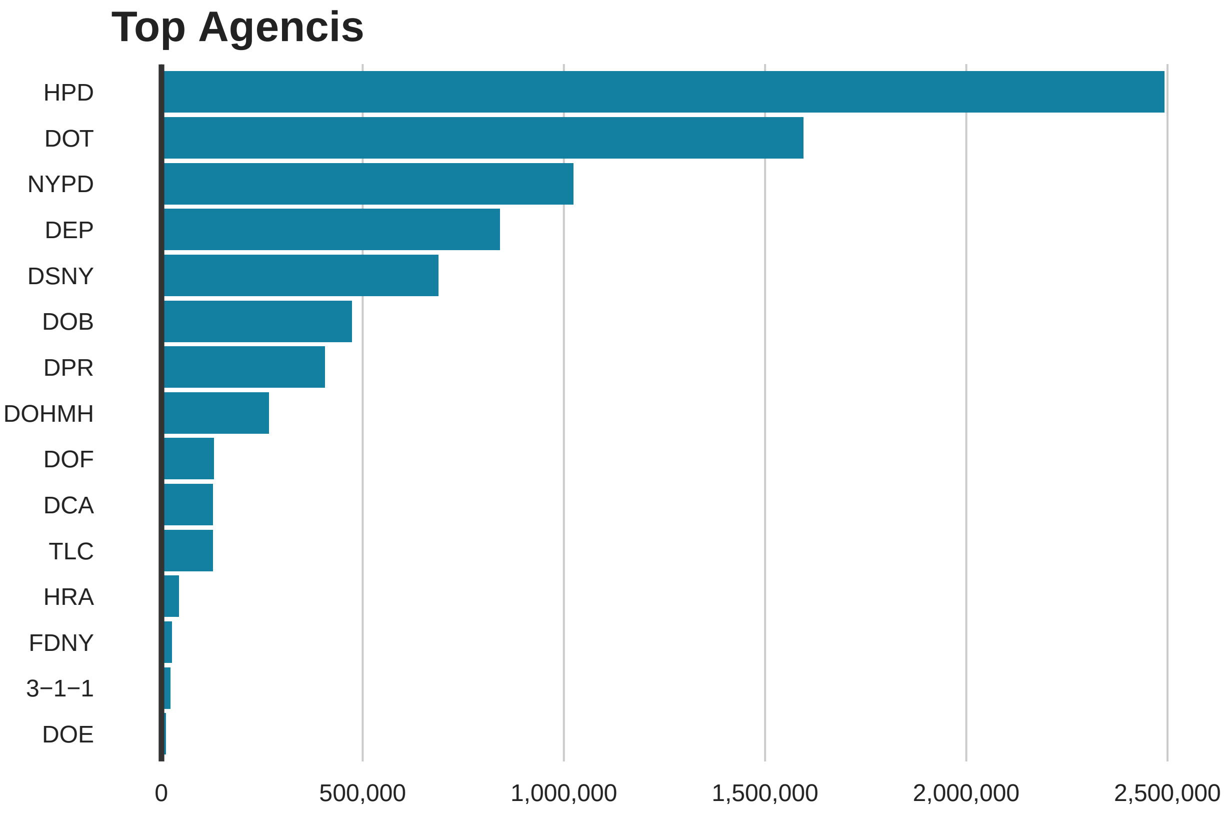
The number of agencies joined to NYC 311 in 2010 increased from 10 agencies in 2009 to 16 agencies. Therefore, the the complaint types increased from 57 in 2009 to 172 in 2010. Although the increase in number of agencies were around 37.50%, the complaint typed increased by 66.86%.

Complaint Distribution:

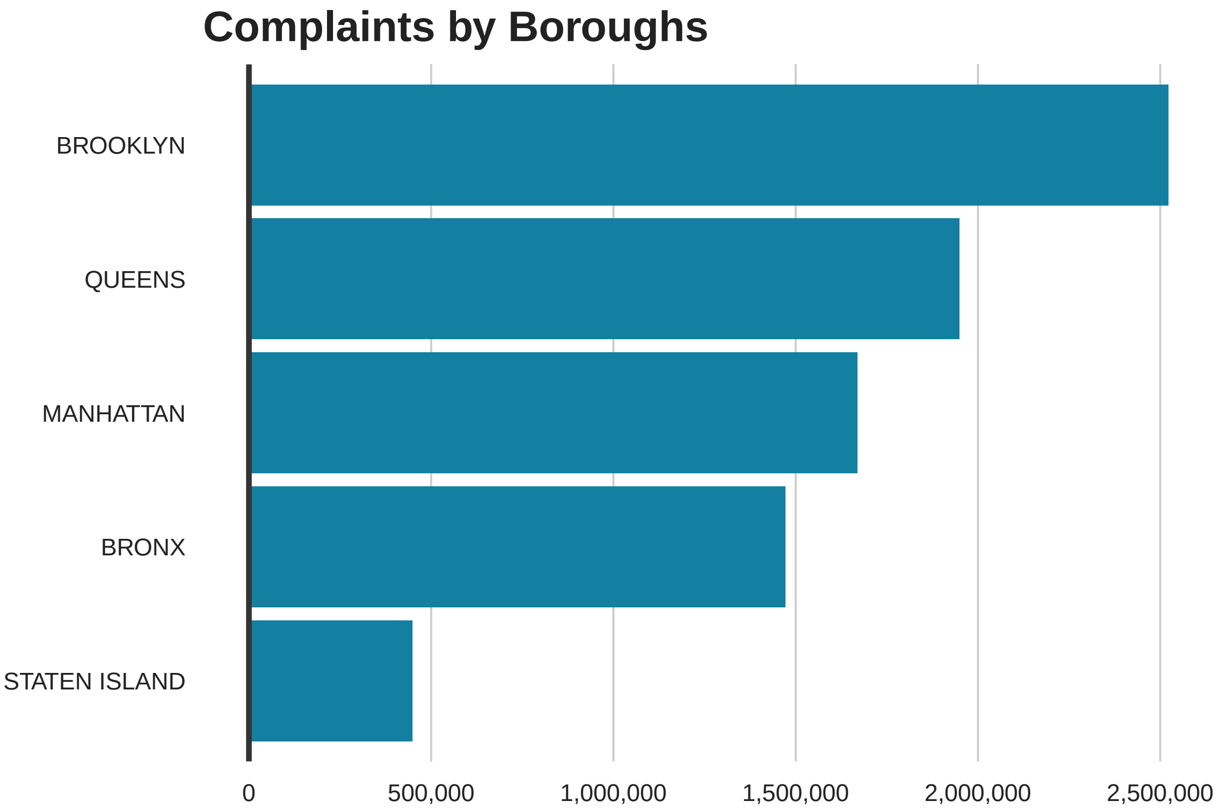
Below is a quick overview on the complaint distribution over Complaint Types, Agencies, and Boroughs:



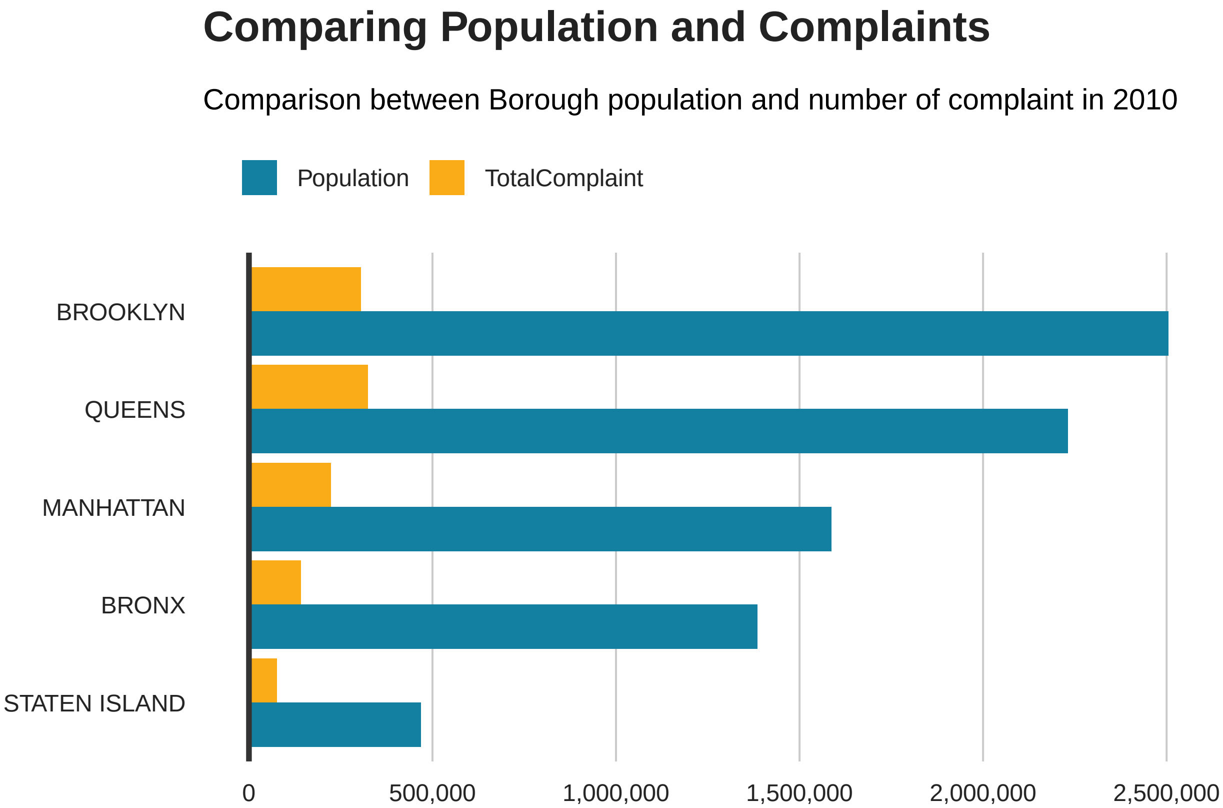
Heating is the top complaint type with more than (600000) complaint followed by street condition, Street Light Condition with more than (400000) complaints for each, then General Construction and Plumbing with more (400000) each record, then comes Blocked Driveway and Water System with around (300000) complaint for each type. The above top 15 complaints type makes ($\sim$ 57%) of the total complaints.



This graph shows the most contributing New York city agencies in terms of the number of complaints. “HPD” (with around 2.5Millions complaints) and “DOT” (with more than 1.5 million) are the two top agencies as they have registered significantly more complaints than the rest. We can also see a number of agencies not far away from the top with number of complaints between 500,000 and 1 million.



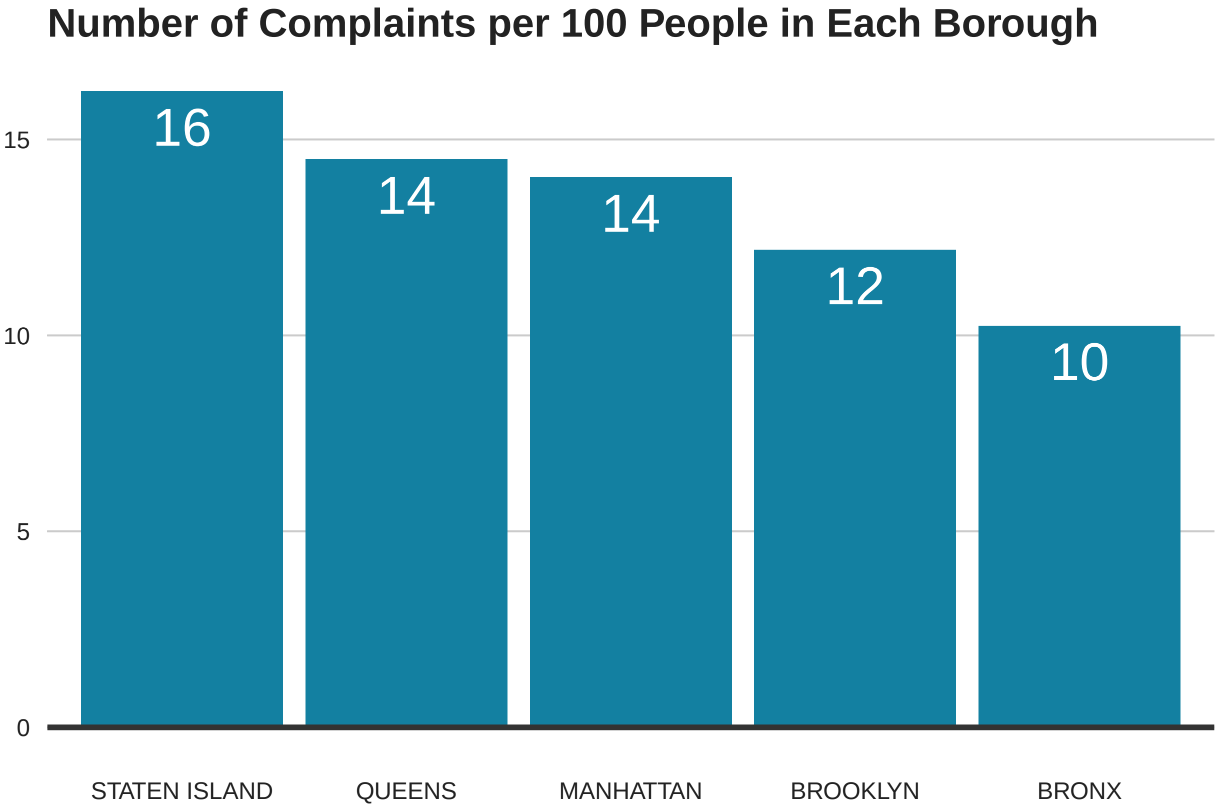
Although, Queens has the largest area, it shows that most complaints are coming from Brooklyn considering the fact that it has the highest listed population in the City. Further explanation will be in the section of correlation of population and complaints.



Comparing Population and Complaints

Comparison between Borough population and number of complaint in 2010

The above graph compares the Boroughs in two dimensions, the population and the number of complaints. Although it seems that the number of population always increases the number of complaint, however, the next graph will show how many complaints are recorded per 100 people of each Borough.



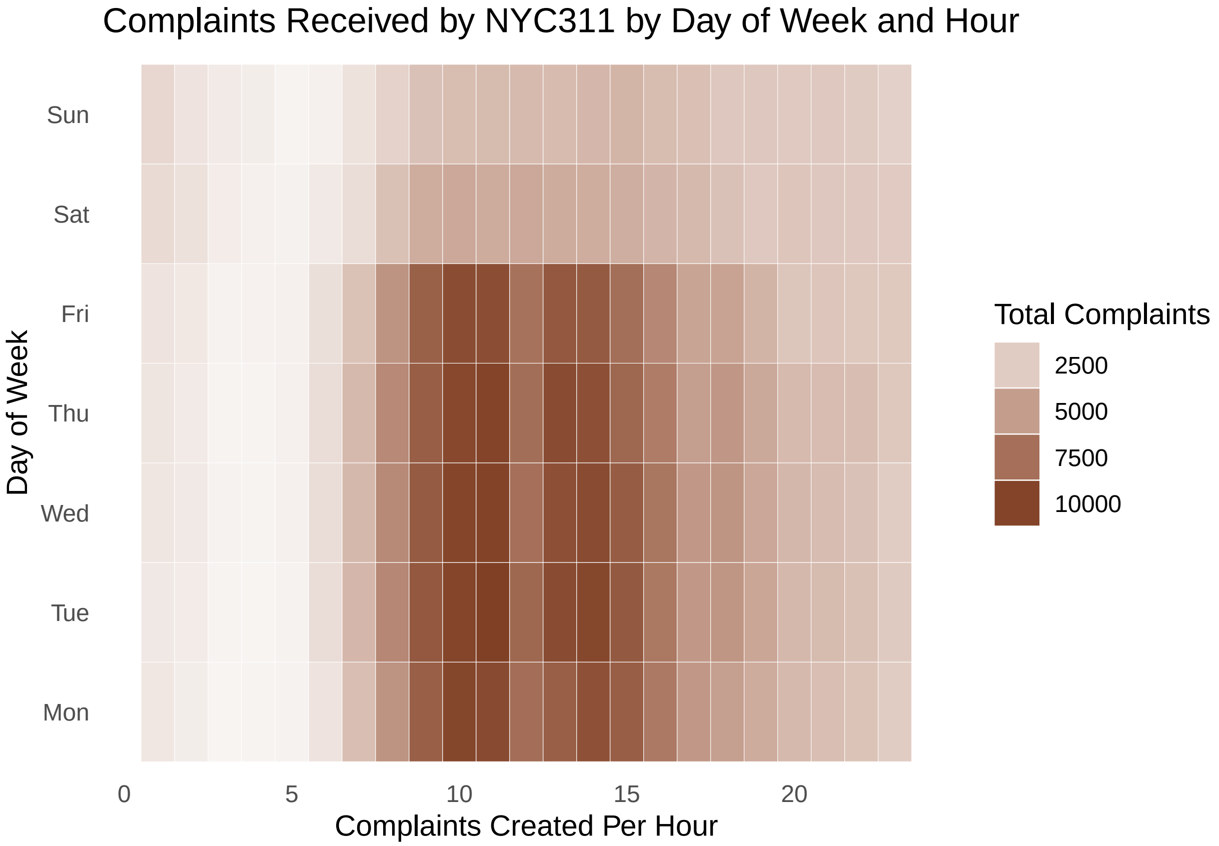
Number of Complaints per 100 People in Each Borough

An interesting finding is that, although Staten Island has the lowest population, it has receded the biggest number of complaints among others (16 complaint per 100 people). Brooklyn, which has the biggest population recorded less complaints than others (12 per 100 people) except Bronx which has recorded the lowest number of complaint per 100 people (10 complaints).

Complaints Peak Timings:

Pattern was generated with the increased of complaints during working hours as follow:

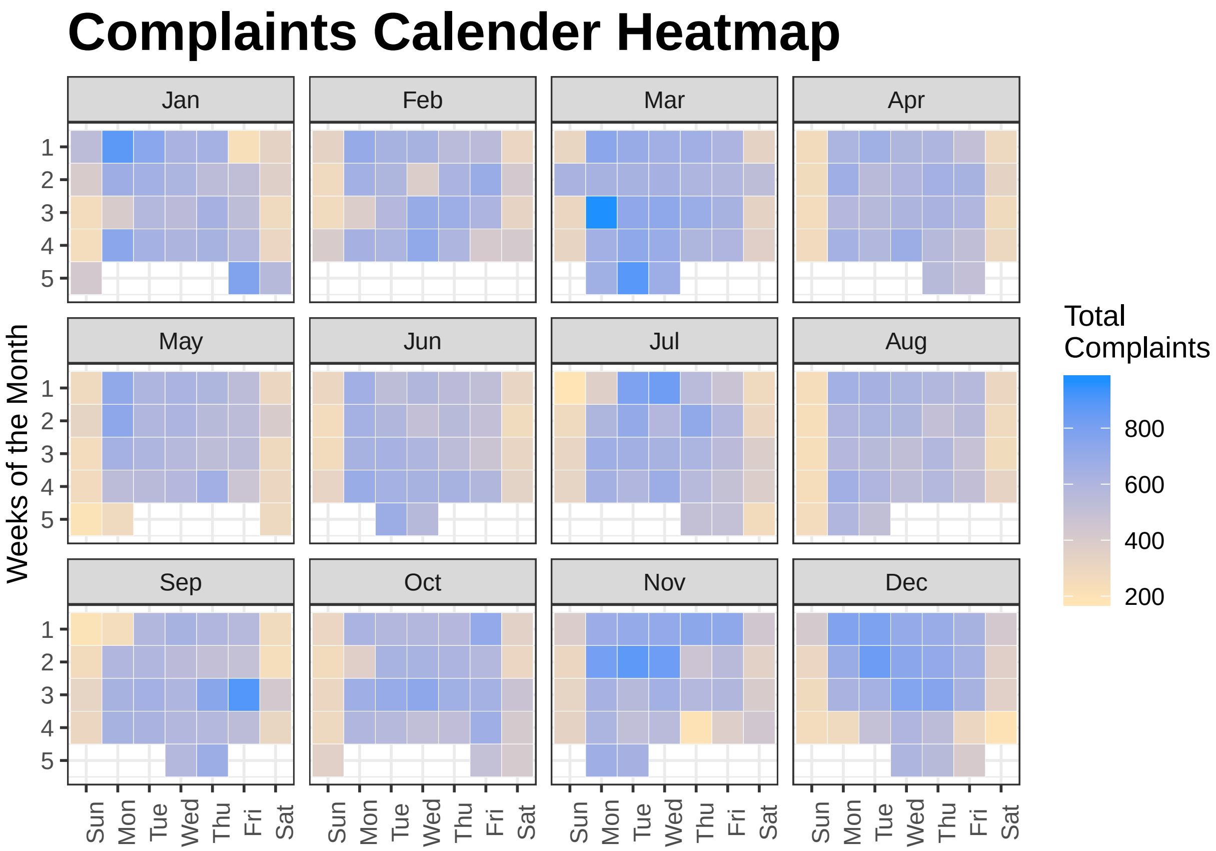
1. **Peak Hours:**



The graph above displays the timings and the days when the complaints are increasing and dropping over 24 hours in 7 days of the week. By observing the distribution over the week, complaints tend to be higher during weekdays and lesser during the weekends. In addition, it shows that the peak period starts from Monday to Friday, between 7:00AM until 7:00PM. However, we noticed the dominance of received complaints from Monday to Friday, between 9:00AM until 11:00AM and started fading in the noon time assuming the lunch time of the people, and it resumes increasing between 1:00PM until 3:00PM.

Note: Manual data entry using the default hour 12AM (midnight) created deficiencies on the data so that hour was eliminated from the analysis.

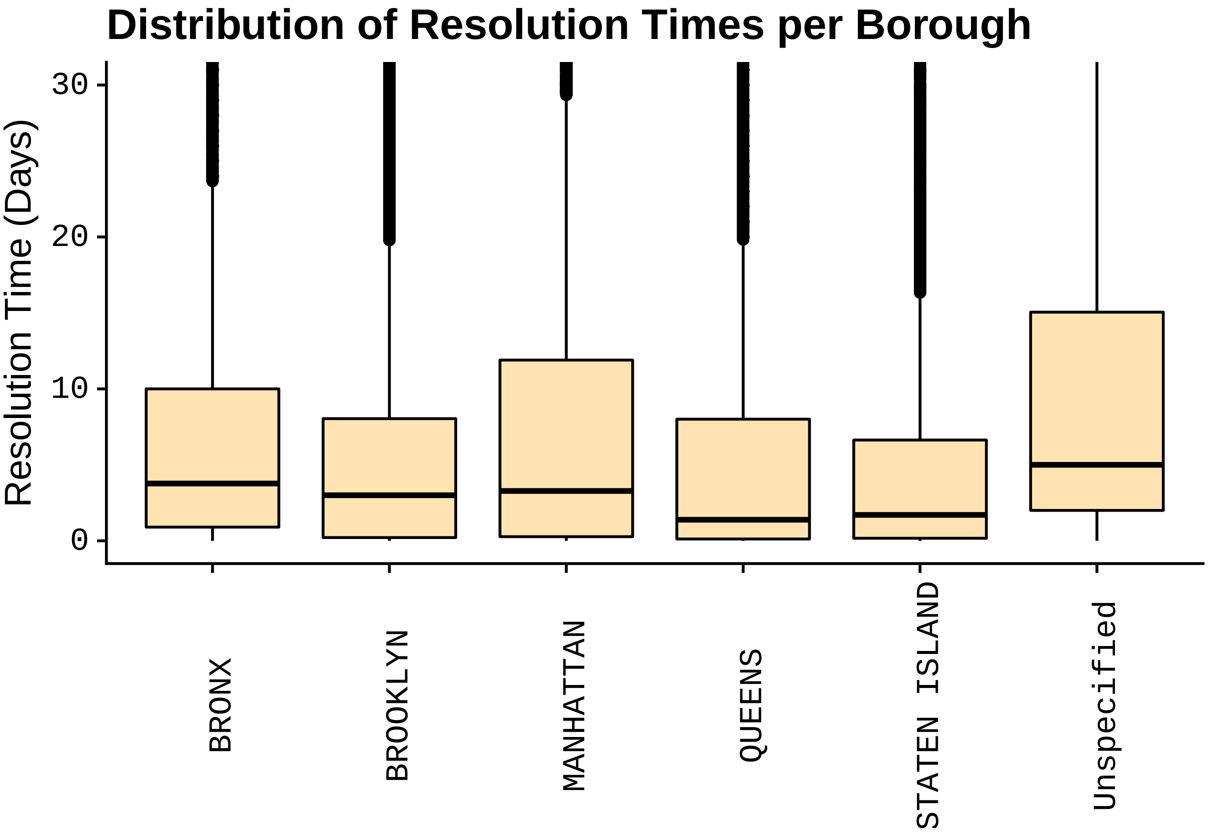
2. **Peak Days and Months:**



The graph above displays the overall complaints received in 2014 as a sample to represent the pattern of the complaints from 2010 onwards.

By observing the distribution over the week of every month, most complaints tend to be higher during weekdays and it drops during the weekends. In addition, months starting from November until March shows domination on getting more complaints assuming that it is due to the weather. Therefore, heating is the highest complaints received due to winter season during these months.

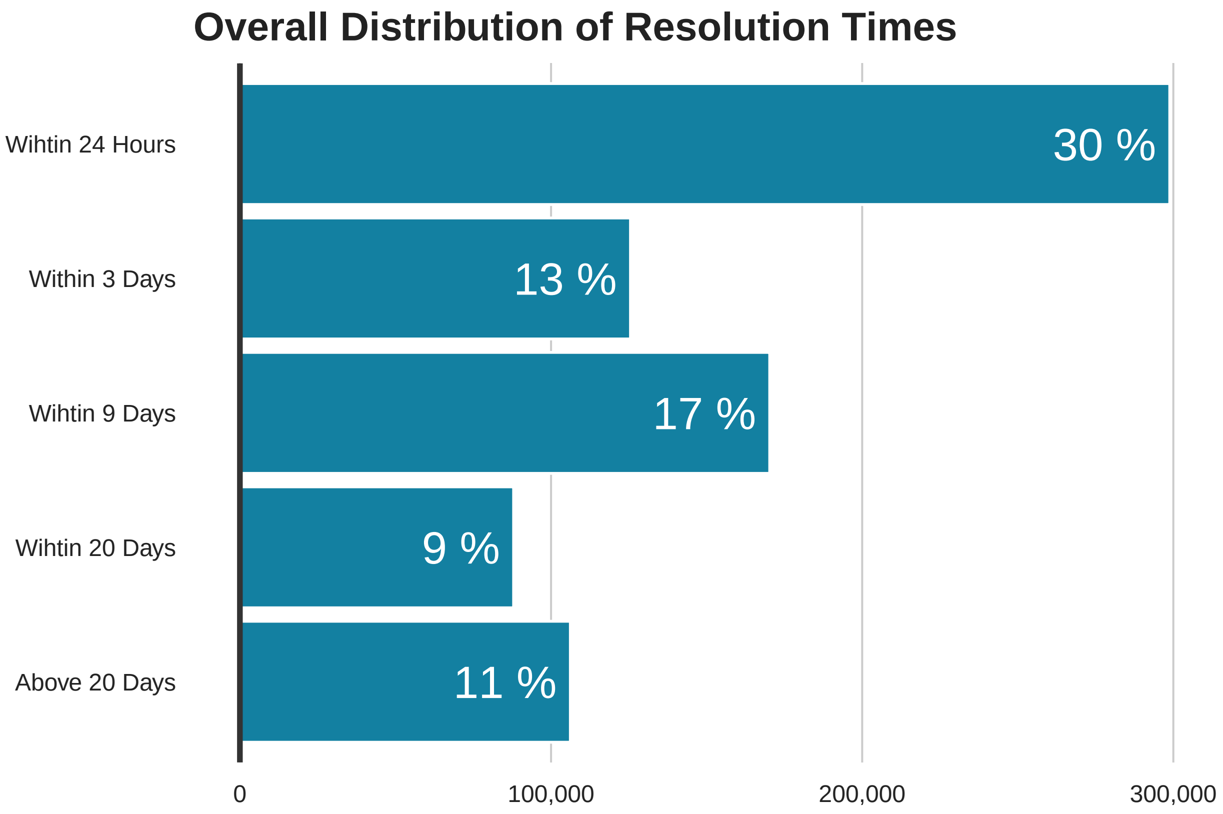
Resolution Time:



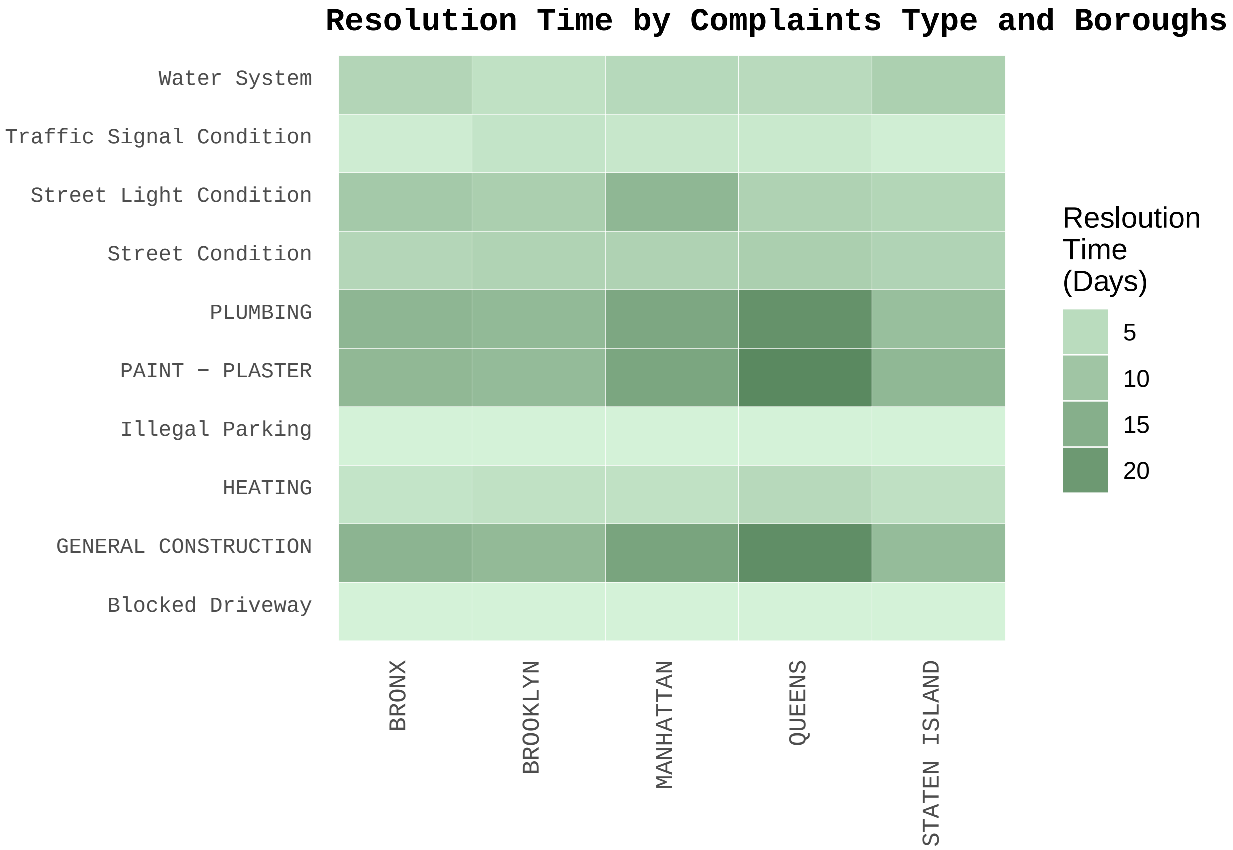
Base on the output of boxplot we can group the compaliant by resolution time as follow:

- Within 24 Hours (1st Qu.)
- Within 3 Days (Median)
- Within 9 Days (3rd Qu.)
- Within 20 Days (Around the Average (main))
- Above 20 Days (Above average and outliers)

An interesting finding revealed from the boxplot is that, the complaint with missing values takes longer time for resolution.



The above figures show the distribution of complaint resolution time as 30% of complaints are being resolved within 24 hours; 13% within 1-3 days; 17% within 3-9 days; 9% within 9-20 days and 11% more than 20 days. Overall, 3 days is the average resolution time excluding the outliers. The next heatmap shows how each Borouh is doing with each Complaint Type in terms of resolution time.



The heatmap above shows the resolution time of complaints per borough. Here we found out that Queens was the slowest borough in terms of resolving complaints mainly on general construction, paint plaster, and plumbing with an average time of 20 days for each followed by Manhattan.

On the other hand, it shows that Brooklyn was the fastest to resolve issues with an average time of 5 days on the complaints regarding water system. In addition, in all boroughs, complaints such as blocked driveway and illegal parking were consistently getting resolved within 5 days.

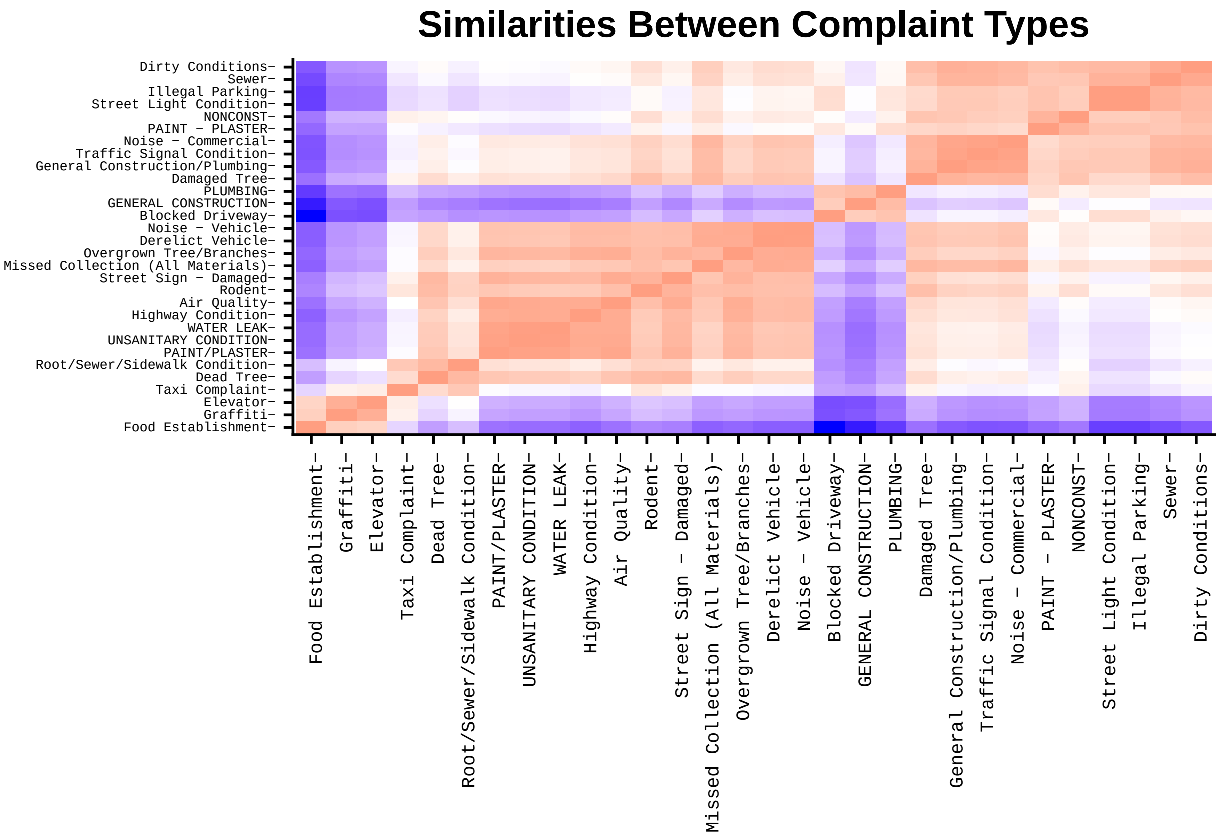
The above not only shows the differences of resolution time between complaint types or between Boroughs, but also indicates the differences between the performance of baranches of the same agency.

Complaint Types Clustering:

1. **Mesureing The Similarities:**

After the above thorough analysis of the number of complaint per each type as well as the resolution time for each type, we applied a clustering method to have a precise grouping of the complaints types using resolution time and number of complaints as data points or bases of clustering.

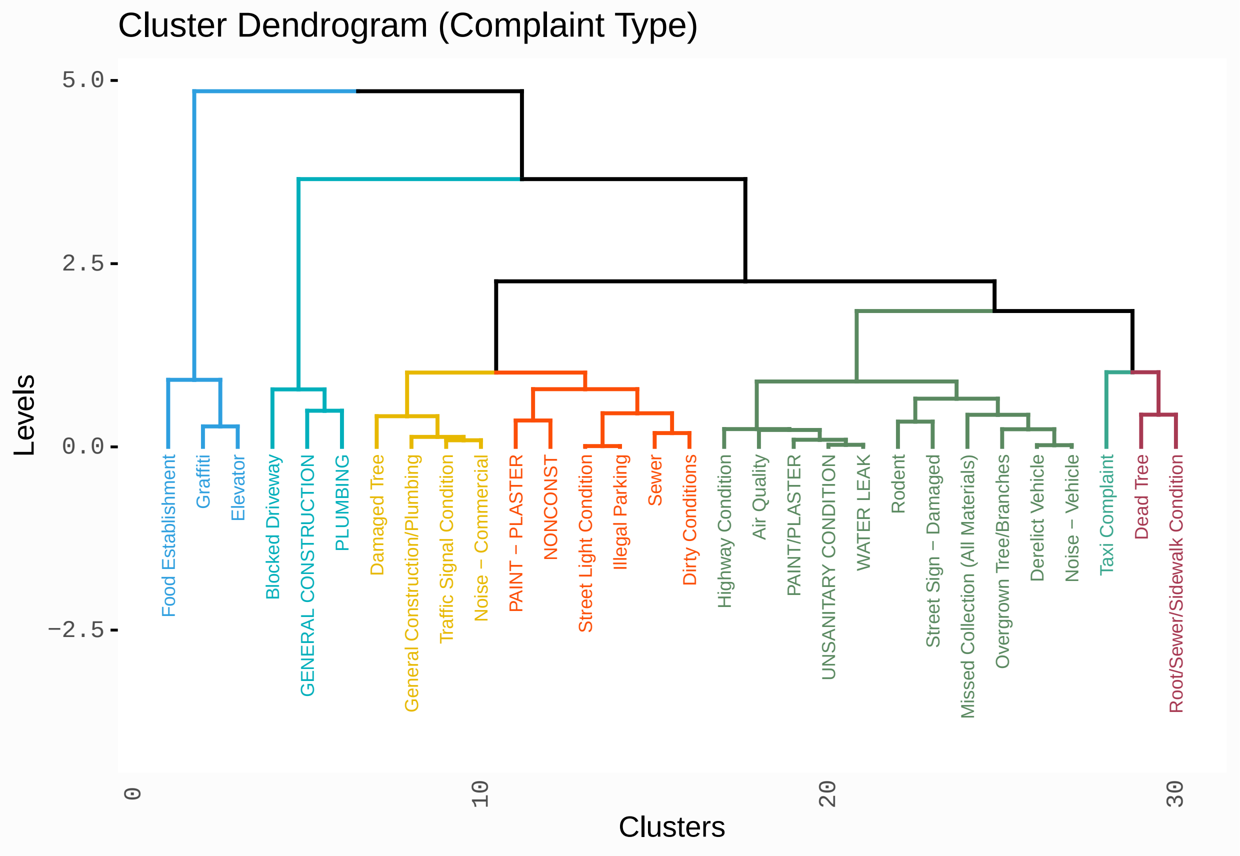
We applied *Euclidean Distance* method to measure the similarities between the difference complaint types as follow:



The above heatmap or similarity matrix shows in “Red” the similarities between complaint types with dissimilarities in “Blue”. The more the color goes towards the red color it indicates strong similarity while it indicates the opposite once it goes towards the blue color.

From the heatmap we can many different levels of dissimilarities like strong similarity: General Constructions & Plumbing, Unsanitary Condition & Paint Plaster, and Dead Tree & Street Light Damage or weak similarity: Hot Water & Air Quality. Also we can see the strong dissimilarity between General Construction and Food Establishment. Actually by spotting each area of colors on the heatmap you can create clusters of the complaint types. However, we have applied to hierarchal clustering (Dendrogram) to automatically group the complaint types into big clusters and small clusters (branches and nodes).

2. **Hierarchal Clustering (Dendrogram):**



We can see from the dendrogram many levels of clusters or branches in other words. We can cluster the complaint types into big clusters or small clusters based on the business needs and purpose of clustering. In the right of the dendrogram we observed that (Dirty Conditions & Sewer) form on cluster at the lowest level, (Street light Condition & Illegal Parking) from another cluster in the same level, and so (Nonconst, & Paint-Plaster), however all of them (Dirty Conditions, Sewer, Street light Condition, Illegal Parking) form on larger cluster share common similarities in terms of number of complaint and resolution time. In the lift of the dendrogram (Food Establishment, Graffiti and Elevator) from a separate branch of one level that doesn't share immediate common similarities with rest of the clusters. Such clustering can help in allocating special teams two certain number of the complaint type that share the same volume of complaint and same level of complexity in terms of resolution. The model can be enhanced be introducing more data points to it like peak timings for each type and any other data that can be quantified.

Conclusion

Overall, after doing analysis with the NYC 311 service, we found interesting outcomes upon exploring the data which are shown in our visualizations as well as the discrepancies and inadequacies such as multiple duplicates, missing values, and inconsistencies using 12-hour time format. Therefore, in order to get significant and useful intact results, dataset was improved by eliminating duplicates, and adding useful dataset to produce a valuable result such as adding zip codes to map the respective location where the requests were coming from to reduced unspecified values. Herewith, we achieved the results that can be shared with NYC government to further improve the NYC 311 service for the betterment of its operations.

Below are the findings (excluding discrepancies) that we got upon exploring and analyzing NYC 311 service:

1. Records from 2010 onwards shows consistencies on the record compare to the previous years as a result of having more agencies/non-government organizations involved in the NYC 311 operations.
2. 37.5% of agencies/non-government organizations were added to support on 311 service from 2010 onwards while complaint types increased approximately 67%.
3. Brooklyn is the most populated borough in New York City. Therefore, most records of complaints and service request are coming from this borough.
4. Complaints are usually increasing during working days and drastically dropping during weekends.
5. Months starting from November until March shows domination on getting more complaints assuming that it is due to the weather. Therefore, heating is the highest complaints received due to winter season during these months.
6. 30% of complaints are being resolved within 24 hours; 13% within 1-3 days; 17% within 3-9 days; 9% within 9-20 days and 11% more than 20 days.The complaint with missing values takes longer time for resolution.
7. Queens was the slowest borough to resolve the complaints with an average time of 20 days followed by Manhattan while Brooklyn was the fastest to resolve complaints with an average time of 5 days.
8. Staten Island recorded the highest number of complaint per 100 percent although it has the smallest population.
9. Complaint types can be clusterd into small clusters (e.g Street light Condition & Illegal Parking) or large clusters (e.g Dirty Conditions, Sewer, Street light Condition, Illegal Parking) based on the business needs and purpose of clustering.

From this exercise, we can provide the below recommendations:

- To have a predefined SLA for each type of complaints in order to be able to measure the performance of the agencies in terms of complaint resolution time.
- To have a predefined full mapping for each complaint type to minimize the data entry efforts and errors. Once the complaint type is selected, based on the configuration in the mapping table, it will be automatically mapped to the related agency, with the related, transferred to respective handling team, and have predefined resolution status based on the nature of the complaints.
- To implement a duplicate detection in the call center, as many people tend to report the same incident in the same time, which increases the duplicate cases and the operational activities to handle each complaint separately. The system feature can detect the duplicates based on certain criteria that can be configured in the system to notify the call center agent on that the incident already reported by other people and logged in the system.

Appendix

Additional Datasets Sources

- NYC Population Data hit this link
- NYC Zip Codes by Boroughs hit this link

Data Dictionary with Data Quality Analysis

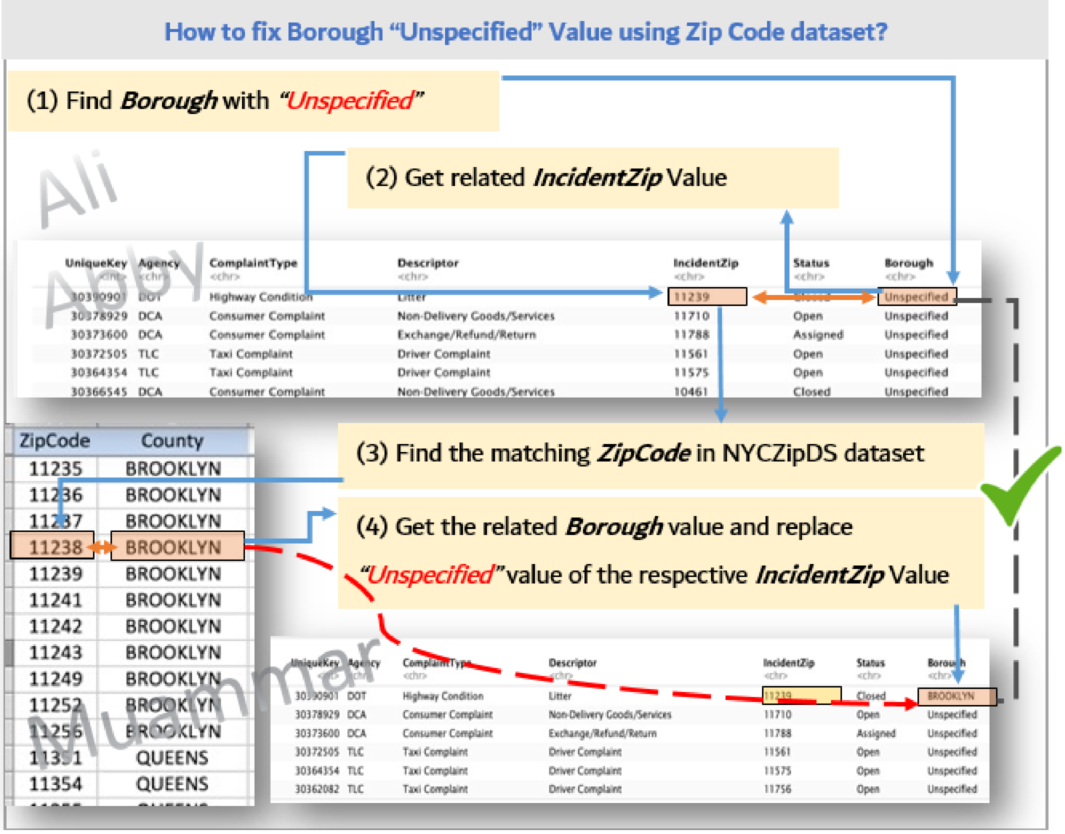
Attribute	Description	Expected Value	`describe()` function output analysis	Consider?
Unique Key	Unique identifier of a Service Request (SR) in the open data set	Sequenced integer	There are few duplicates because the number of distinct values is less that the total. We will remove the duplicate first then drop the column	No
Created Date	Date SR was created	Date in format MM/DD/YY HH:MM:SS AM/PM	No missing value, very important for time related analysis	Yes
Closed Date	Date SR was closed by responding agency	Date in format MM/DD/YY HH:MM:SS AM/PM	Missing values might mean that the request still open. Important to calculate the average time of request resolution (Closed Date - Created Date)	Yes
Agency	Acronym of responding City Government Agency	List of Gov. Agencies Acronyms (lookup)	No missing values, main attribute for analysis	Yes
Agency Name	Full Agency name of responding City Government Agency	Text	1682 agency name found while the real number is just 64. That because the agency is typed in different ways.	No
Complaint Type	This is the fist level of a hierarchy identifying the topic of the incident or condition. Complaint Type may have a corresponding Descriptor (below) or may stand alone.	List of complaint types (lookup)	No missing values, main attribute for analysis	Yes
Descriptor	This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.	List of complaint sub-type (lookup)	No missing values, main attribute for analysis (provides further detail on the incident)	Yes
Status	Status of SR submitted	Lookup (Assigned, Cancelled, Closed, Pending, +)	No missing values, main attribute for analysis, helps for status follow-up report.	Yes

Attribute	Description	Expected Value	`describe()` function output analysis	Consider?
Due Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal Service Level Agreements (SLAs).	Date in format MM/DD/YY HH:MM:SS AM/PM	Not logged, more than 60% are missing values. It would be good to use for measuring NYC311 performance (on-time resolution), but the data quality doesn't help	No
Resolution Action Updated Date	Date when responding agency last updated the SR.	Date in format MM/DD/YY HH:MM:SS AM/PM	Missing values. Not needed	No
Resolution Description	Describes the last action taken on the SR by the responding agency. May describe next or future steps.	Text	Open long text. Not needed	No
Location Type	Describes the type of location used in the address information	Text	Missing and irrelevant values	No
Incident Zip	Incident location zip code, provided by geo validation.	Integer	Poor quality data, more than 700,000 missing value, wrong entries like "??", "XX". But it is important to keep because zip code is a location identifier. To be used carefully	Yes
Incident Address	House number of incident address provided by submitter.	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Street Name	Street name of incident address provided by the submitter	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Cross Street 1	First Cross street based on the geo validated incident location	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Cross Street 2	Second Cross Street based on the geo validated incident location	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Intersection Street 1	First intersecting street based on geo validated incident location	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Intersection Street 2	Second intersecting street based on geo validated incident location	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Address Type	Type of incident location information available.	Values: Address; Block face; Intersection; LatLong; Placename	Not Need, other location identifiers like zip, park facility can be used instead	No

Attribute	Description	Expected Value	`describe()` function output analysis	Consider?
City	City of the incident location provided by geovalidation.	Tex	Misspelled. 1816 cities found while NY has only 62 city	No
Landmark	If the incident location is identified as a Landmark the name of the landmark will display here	Tex	Not Needed, other location identifiers like zip, park facility can be used instead	No
Facility Type	If available, this field describes the type of city facility associated to the SR	Tex	Not Needed, other location identifiers like zip, park facility can be used instead	No
Community Board	Provided by geovalidation.	Tex	Not Needed, other location identifiers like zip, park facility can be used instead	No
BBL	Borough Block and Lot, provided by geovalidation. Parcel number to identify the location of location of buildings and properties in NYC.	Tex	Not Needed, other location identifiers like zip, park facility can be used instead	No
Borough	Provided by the submitter and confirmed by geovalidation.	List of Borough (Lookup)	No missing values, main attribute for analysis	Yes
X Coordinate (State Plane)	Geo validated, X coordinate of the incident location.	Double	Will be used to plot the data distribution on the map	Yes
Y Coordinate (State Plane)	Geo validated, Y coordinate of the incident location.	Double	Will be used to plot the data distribution on the map	Yes
Latitude	Geo based Lat of the incident location	Double	Missing values. Other coordinates will be used	No
Longitude	Geo based Long of the incident location	Double	Missing values. Other coordinates will be used	No
Location	Combination of the geo based lat & long of the incident location	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Park Facility Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here	Tex	Not Needed, other location identifiers like zip and coordinates will be used	No
Park Borough	The borough of incident if it is a Parks Dept facility		Not Needed, other location identifiers like zip and coordinates will be used	No

Vehicle Type	If the incident is a taxi, this field describes the type of TLC vehicle.	Car Service;
		Commuter Van;
		Green Taxi
			Specific to vehicles' incidents	No

Attribute	Description	Expected Value	`describe()` function output analysis	Consider?
Taxi Company Borough	If the incident is identified as a taxi, this field will display the borough of the taxi company.		Specific to vehicles' incidents	No
Taxi Pick Up Location	If the incident is identified as a taxi, this field displays the taxi pick up location	Grand Central Station; Intersection; JKF Airport; La Guardia Airport; New York-Penn Station; Other; Port Authority Bus Terminal	Specific to vehicles' incidents	No
Bridge Highway Name	If the incident is identified as a Bridge/Highway, the name will be displayed here.	Tex	Specific to vehicles' incidents	No
Bridge Highway Direction	If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here.	Tex	Specific to vehicles' incidents	No
Road Ramp	If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.	Roadway; Ramp	Specific to vehicles' incidents	No
Bridge Highway Segment	Additional information on the section of the Bridge/Highway were the incident took place.	Tex	Specific to vehicles' incidents	No
Population	The borough population in a certain year	Double	Will be used to calculate number of complaints per CAPITA	Yes



How to fix Borough “Unspecified” Value using Zip Code dataset?

(1) Find *Borough* with *“Unspecified”*

(2) Get related *IncidentZip* Value

UniqueKey <int>	Agency <chr>	ComplaintType <chr>	Descriptor <chr>	IncidentZip <chr>	Status <chr>	Borough <chr>
30390901	DOT	Highway Condition	Litter	11239	Closed	Unspecified
30378929	DCA	Consumer Complaint	Non-Delivery Goods/Services	11710	Open	Unspecified
30373600	DCA	Consumer Complaint	Exchange/Refund/Return	11788	Assigned	Unspecified
30372505	TLC	Taxi Complaint	Driver Complaint	11561	Open	Unspecified
30364354	TLC	Taxi Complaint	Driver Complaint	11575	Open	Unspecified
30366545	DCA	Consumer Complaint	Non-Delivery Goods/Services	10461	Closed	Unspecified

(3) Find the matching *ZipCode* in NYCZipDS dataset

(4) Get the related *Borough* value and replace *“Unspecified”* value of the respective *IncidentZip* Value

ZipCode	County
11235	BROOKLYN
11236	BROOKLYN
11237	BROOKLYN
11238	BROOKLYN
11239	BROOKLYN
11241	BROOKLYN
11242	BROOKLYN
11243	BROOKLYN
11249	BROOKLYN
11252	BROOKLYN
11256	BROOKLYN
11351	QUEENS
11354	QUEENS

UniqueKey <chr>	Agency <chr>	ComplaintType <chr>	Descriptor <chr>	IncidentZip <chr>	Status <chr>	Borough <chr>
30390901	DOT	Highway Condition	Litter	11239	Closed	BROOKLYN
30378929	DCA	Consumer Complaint	Non-Delivery Goods/Services	11710	Open	Unspecified
30373600	DCA	Consumer Complaint	Exchange/Refund/Return	11788	Assigned	Unspecified
30372505	TLC	Taxi Complaint	Driver Complaint	11561	Open	Unspecified
30364354	TLC	Taxi Complaint	Driver Complaint	11575	Open	Unspecified
30362082	TLC	Taxi Complaint	Driver Complaint	11756	Open	Unspecified

Figure 1: Improve Borough Column Model

Improving Borough Column Model