# homework vi

Ayush Kumar Shah

2020-10-06

## Contents

# Introduction

In this report, the `nyc311` dataset will be analyzed to generate and communicate findings about different aspects of the data in an effective and presentable manner. The analysis of the `nyc311` dataset along with two additional related datasets will help to identify different existing problems and the causes of increasing complaints. The insights from the report will also help the agencies to reduce the frequency of complaints and also address the complaints more effectively in the future.

the two data sets introduced in the previous report are connected to the 311 data set, using `dplyr`. Before connecting them, several operations are performed on the datasets.

Also, few tables consisting of an extract of the data of each dataset as well as the final joined dataset are shown. Finally, a data dictionary for all the data in each dataset including the final joined dataset is also displayed. The connections between the columns will be shown in the next final report.

# Context

The dataset focused in this study is New York's 311 service requests. The NYC Open Data has made all NYC 311 service requests and complaints publicly available. Open Data is free public data published by New York City agencies and other partners. Each City agency has an Open Data Coordinator, who serves as the main point of contact for the Open Data Team and the public, and works to identify, document, structure, and manage the agency's public datasets. NYC Open Data curates nearly 3,000 different datasets across every facet of life in New York City.

The fundamental mission of OpenData is taking what can be complex information, generated from the operation of the largest municipal government in the country, and making it accessible to everyone: Open Data for All. Open Data is every New Yorker's digital library, one that allows them to engage on a more even footing with their government, and build online tools drawing on the information that is shared.

# About the data

## nyc311

The `nyc311` data in this report includes all 311 Service Request Calls from the residents, busineeses and visitors in New York City between the 2003 to 2015. This information is automatically updated daily. In addition, the general public can access their website to submit queries in more than 50 languages and can get help in more than 175 languages by calling 311. They provide citizens a quick, convenient way to complain about problems in their neighborhoods, and get a response.

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 fundamental goal of `nyc311` is to provide the public with swift and easy access to all New York City government services and information while offering the best customer service.

**Data extract: nyc311**

Let's view an extract of the `nyc311` data with only a few important columns.

Table 1: New York City: 311 Service Requests

| Created.Date | Agency | Complaint.Type | Descriptor | Incident.Zip | Status | Borough |
|---|---|---|---|---|---|---|
| 01/26/2010 12:00:00 AM | HPD | HEATING | HEAT | 11210 | Closed | Unspecified |
| 12/02/2010 12:00:00 AM | HPD | HEATING | HEAT | 11213 | Closed | Unspecified |
| 03/13/2014 07:57:26 AM | NYPD | Derelict Vehicle | With License Plate | 10314 | Closed | STATEN ISLAND |
| 11/13/2012 12:00:00 AM | HPD | HEATING | HEAT | 11229 | Closed | BROOKLYN |
| 07/19/2013 12:00:00 AM | HPD | PLUMBING | BASIN/SINK | 10460 | Closed | BRONX |
| 01/29/2011 05:49:00 PM | DOT | Street Light Condition | Street Light Out | 11213 | Closed | BROOKLYN |

## Additional data

The two additional datasets introduced in this report are:

1. Projected Population 2010-2040 - Total By Borough and Age Groups: Source
2. 2005 - 2011 Graduation Outcomes - Borough - Ethnicity: Source

### 1. Projected Population 2010-2040 - Total By Borough and Age Groups Groups

This dataset consists of projected total New York City population for five intervals from 2010 through 2040 by Borough, broken down by 18 age cohorts. However, the age groups may not add up to the total due to rounding.

This dataset is introduced so that the population information in each Borough is available during analysis of the nyc311 data, by joining it appropriately with the original `nyc311` data. This is achievable since the table contains the `Borough` column.

**Data extract: Population**

Table 2: Projected Population 2010-2040 by Borough and Age

| Borough | Age | 2010 | 2015 | 2020 | 2025 | 2030 | 2035 | 2040 |
|---|---|---|---|---|---|---|---|---|
| Staten Island | 5-9 | 30015 | 29595 | 29394 | 29890 | 30397 | 30253 | 29816 |
| Staten Island | 10-14 | 30797 | 30703 | 30307 | 30001 | 30541 | 31074 | 30932 |
| Queens | 10-14 | 124320 | 123672 | 129351 | 130384 | 132038 | 133473 | 133813 |
| Manhattan | 10-14 | 58229 | 55455 | 60331 | 62902 | 66322 | 66700 | 63630 |
| Brooklyn | 10-14 | 159404 | 155314 | 166761 | 174242 | 175992 | 173419 | 171265 |
| Bronx | 10-14 | 99159 | 92423 | 95703 | 100081 | 104065 | 105297 | 105010 |

**2. 2005 - 2011 Graduation Outcomes - Borough - Ethnicity**

This dataset consists of graduation results for all students by year; cohorts of 2001 through 2007 (Classes of 2005 through 2011). Graduation Outcomes are as calculated by the New York State Education Department. The New York State calculation method was first adopted for the Cohort of 2001 (Class of 2005).

Graduates are defined as those students earning either a Local or Regents diploma and exclude those earning either a special education (IEP) diploma or GED.

This dataset is introduced so that the educational status of the people in different Borough is available for further analysis in correlation with the `nyc311` data. This is achievable since the table contains the `Borough` column.

**Data extract: Graduation**

Table 3: 2005-2015 Graduation Outcomes - Borough

| Borough | Cohort Year | Cohort Category | Demographic | Total Cohort Num | Total Grads Num | Total Grads Pct of cohort |
|---|---|---|---|---|---|---|
| Bronx | 2001 | 4 Year June | Asian | 638 | 479 | 75.10% |
| Bronx | 2001 | 5 Year | Asian | 638 | 528 | 82.80% |
| Bronx | 2001 | 6 Year | Asian | 638 | 534 | 83.70% |
| Bronx | 2002 | 4 Year June | Asian | 681 | 497 | 73.00% |
| Bronx | 2002 | 5 Year | Asian | 681 | 564 | 82.80% |
| Bronx | 2002 | 6 Year | Asian | 681 | 574 | 84.30% |

Both of these datasets were obtained from the NYC OpenData. Also, they both contain the column `Borough` and `Years` which makes them connectable to the `nyc311` data.

# Tidying the data

## Tidying nyc311 data

### Checking duplicates

We check duplicates by first removing the `Unique.Key` variable since all the values are unique in the column.

Since `all_equal()` takes a very long time to compare the two data frames, we simply compare the number of rows in the main and non duplicated data frames using `nrow()`.

- Number of rows in original nyc311 dataframe = 10000
- Number of rows in non duplicated nyc311 dataframe = 10000
- Duplicate observations present FALSE

We can see that there are duplicate observations. So, we use the non duplicated data frame in the further steps which has 10000 unique observations obtained after removing 0 duplicate observations.

**Remove unspecified Borough.**

Table 4: Borough distribution

| Borough | count |
|---|---|
| BROOKLYN | 2768 |

| Borough | count |
|---|---|
| QUEENS | 2136 |
| MANHATTAN | 1859 |
| BRONX | 1549 |
| Unspecified | 1168 |
| STATEN ISLAND | 520 |

We can see that there is a significant number of observations with Unspecified Borough. Hence, those observations have been removed using `filter()`.

**Separating Created.Date to mulitple Columns**

- We separate `Created.Date` into columns `Created.Year`, `Created.Month`, `Created.Day`, and `Created.Time`.
- We again calculate `Hours` from the `Time` variable as well as `Created.Weekday` using `POSIXlt` class.
- We also calculate the `Resolve.Time.Days` variable by computing the difference of `Closed.Date` and `Created.Date`variables.

These separations are done so that we can easily analyze different trends in the data variables later based on year, month, day or hour of day. Also, the response time can by analyzed later in relation to other variables.

The table below shows the new time and date variables.

Table 5: Separation of Date and Time and Calculation of Response Time (continued below)

| Created.Date | Closed.Date | Created.Year | Created.Month | Created.Day |
|---|---|---|---|---|
| 03/13/2014 07:57:26 AM | 03/13/2014 08:08:59 AM | 2014 | 3 | 13 |
| 11/13/2012 12:00:00 AM | 11/18/2012 12:00:00 AM | 2012 | 11 | 13 |
| 07/19/2013 12:00:00 AM | 08/07/2013 12:00:00 AM | 2013 | 7 | 19 |
| 01/29/2011 05:49:00 PM | 01/31/2011 10:56:00 AM | 2011 | 1 | 29 |
| 03/02/2012 12:00:00 AM | 03/08/2012 12:00:00 AM | 2012 | 3 | 2 |
| 08/01/2011 10:49:23 PM | 08/01/2011 11:32:29 PM | 2011 | 8 | 1 |

| Created.Time | Created.Hour | Resolve.Time.Hours | Created.Weekday |
|---|---|---|---|
| 07:57:26 AM | 7 | 0.18 | Thursday |
| 12:00:00 AM | 0 | 120 | Tuesday |
| 12:00:00 AM | 0 | 432 | Friday |
| 05:49:00 PM | 17 | 41.12 | Saturday |
| 12:00:00 AM | 0 | 144 | Friday |
| 10:49:23 PM | 22 | 0.72 | Monday |

**Remove Columns**

**Redundant columns**

Let's view some columns which have redundant information.

Table 7: Redundant columns (continued below)

| Street.Name | Incident.Address | Latitude | Longitude | Location |
|---|---|---|---|---|
| WILLOW ROAD WEST | 498 WILLOW ROAD WEST | 40.62 | -74.15 | (40.61621148786262, -74.15315033162267) |
| AVENUE W | 1900 AVENUE W | 40.59 | -73.95 | (40.594748423277736, -73.95157327758106) |
| MARMION AVENUE | 1967 MARMION AVENUE | 40.84 | -73.89 | (40.843515202688536, -73.8866720963765) |
| ST JOHNS PLACE | 959 ST JOHNS PLACE | 40.67 | -73.95 | (40.671432764007186, -73.94639142012068) |
| 3 AVENUE | 3463 3 AVENUE | 40.83 | -73.91 | (40.82983679102256, -73.90640699264534) |
| ST MARKS AVENUE | 303 ST MARKS AVENUE | 40.68 | -73.97 | (40.67783529469647, -73.96512638302426) |

| Facility.Type | Location.Type | Borough | Park.Borough | Community.Board |
|---|---|---|---|---|
| Precinct | Street/Sidewalk | STATEN ISLAND | STATEN ISLAND | 01 STATEN ISLAND |
| N/A | RESIDENTIAL BUILDING | BROOKLYN | BROOKLYN | 15 BROOKLYN |
| N/A | RESIDENTIAL BUILDING | BRONX | BRONX | Unspecified BRONX |
| N/A | NA | BROOKLYN | BROOKLYN | 08 BROOKLYN |
| N/A | RESIDENTIAL BUILDING | BRONX | BRONX | 03 BRONX |
| Precinct | Street/Sidewalk | BROOKLYN | BROOKLYN | 08 BROOKLYN |

**Columns with very few data**

Let's view the counts of the non empty values in each column. We only display the columns which have counts less than 65% of the total observations.

Table 9: Columns with very less non empty values

| | Non.NA.Count |
|---|---|
| **Ferry.Direction** | 0 |
| **School.or.Citywide.Complaint** | 2 |
| **Garage.Lot.Name** | 4 |
| **Landmark** | 8 |
| **Vehicle.Type** | 9 |
| **Taxi.Company.Borough** | 11 |
| **Ferry.Terminal.Name** | 13 |
| **Road.Ramp** | 25 |
| **Bridge.Highway.Segment** | 25 |
| **Bridge.Highway.Name** | 26 |

| | Non.NA.Count |
|---|---|
| **Bridge.Highway.Direction** | 31 |
| **Taxi.Pick.Up.Location** | 76 |
| **Intersection.Street.1** | 1642 |
| **Intersection.Street.2** | 1642 |
| **School.Not.Found** | 2386 |
| **Due.Date** | 2436 |

**Removing the columns**

We remove the redundant columns seen previously, the columns with very few non empty data and also the columns which are not relevant or useful like `Created.Date`, `Due.Date` since important date and time information have already been extracted. Also, address fields like `Cross.Street.1`, `Address.Type`, and other irrelevant columns are removed. "

**Viewing columns of the tidied nyc311 dataset**

*Created.Month*, *Created.Day*, *Created.Year*, *Created.Time*, *Agency*, *Agency.Name*, *Complaint.Type*, *Descriptor*, *Location.Type*, *City*, *Status*, *Borough*, *School.Number*, *Latitude*, *Longitude*, *Created.Hour*, *Resolve.Time.Hours* and *Created.Weekday*

Hence, we tidied the `nyc311` data by removing duplicate values, gathering and spreading required columns like Created.Date using `tidyr` package. We also removed the redundant columns and columns with very little or irrelevant information.

## Tidying Population data

**Gathering the years**

Table 10: Projected Population 2010-2040 by Borough and Age

| Borough | Age | 2010 | 2015 | 2020 | 2025 | 2030 | 2035 | 2040 |
|---|---|---|---|---|---|---|---|---|
| Staten Island | 5-9 | 30015 | 29595 | 29394 | 29890 | 30397 | 30253 | 29816 |
| Staten Island | 10-14 | 30797 | 30703 | 30307 | 30001 | 30541 | 31074 | 30932 |
| Queens | 10-14 | 124320 | 123672 | 129351 | 130384 | 132038 | 133473 | 133813 |
| Manhattan | 10-14 | 58229 | 55455 | 60331 | 62902 | 66322 | 66700 | 63630 |
| Brooklyn | 10-14 | 159404 | 155314 | 166761 | 174242 | 175992 | 173419 | 171265 |
| Bronx | 10-14 | 99159 | 92423 | 95703 | 100081 | 104065 | 105297 | 105010 |

We can see that the years need to be gathered together in the `nyc_popn` data.

Table 11: Tidied Population data

| Borough | Age | Year | Population |
|---|---|---|---|
| Staten Island | 5-9 | 2040 | 29816 |
| Staten Island | 10-14 | 2040 | 30932 |
| Queens | 10-14 | 2040 | 133813 |
| Manhattan | 10-14 | 2040 | 63630 |
| Brooklyn | 10-14 | 2040 | 171265 |
| Bronx | 10-14 | 2040 | 105010 |

**Filtering by total population in each Borough**

Since the `nyc311` has no information of Age group, so only the observations with total population of each Borough is filtered.

Table 12: Total population data (Age groups removed)

| Borough | Year | Population |
|---|---|---|
| NYC Total | 2010 | 8242624 |
| Bronx | 2010 | 1385108 |
| Brooklyn | 2010 | 2552911 |
| Manhattan | 2010 | 1585873 |
| Queens | 2010 | 2250002 |
| Staten Island | 2010 | 468730 |

# Joining the datasets

## Converting graduation dataset to suitable form

Table 13: 2005-2011 Graduation Outcomes - Borough

| Borough | Cohort.Year | Cohort.Category | Demographic | Total.Cohort.Num | Total.Grads.Num | Total.Grads.Pct.of.cohort |
|---|---|---|---|---|---|---|
| Bronx | 2001 | 4 Year June | Asian | 638 | 479 | 75.10% |
| Bronx | 2001 | 5 Year | Asian | 638 | 528 | 82.80% |
| Bronx | 2001 | 6 Year | Asian | 638 | 534 | 83.70% |
| Bronx | 2002 | 4 Year June | Asian | 681 | 497 | 73.00% |
| Bronx | 2002 | 5 Year | Asian | 681 | 564 | 82.80% |
| Bronx | 2002 | 6 Year | Asian | 681 | 574 | 84.30% |

The `Cohort.Category` is parsed as number and added with the `Cohort.Year` to calculate the `Graduation.Year`. Then, only the relevant columns are selected. Finally, duplicates are removed by grouping the data according to the columns, `Graduation.Year` and `Borough`.

| Graduation.Year | Borough | Total.Cohort.Num | Total.Grads.Num | Total.Grads.Pct.of.cohort |
|---|---|---|---|---|
| 2005 | Bronx | 6150 | 2261 | 75.10% |
| 2005 | Brooklyn | 9382 | 4232 | 63.80% |
| 2005 | Manhattan | 5739 | 2924 | 81.90% |
| 2005 | Queens | 5335 | 2836 | 67.20% |
| 2005 | Staten Island | 2304 | 1700 | 80.50% |
| 2006 | Bronx | 6623 | 2922 | 82.80% |

## Check Borough's values

Since the tables will be connected by `Borough`, it is checked if there are any unspecified values in the column or mismatch in all the tables.

Table 15: Borough distribution on nyc311 data

| Borough | counts |
|---|---|
| BRONX | 1236 |
| BROOKLYN | 2338 |
| MANHATTAN | 1602 |
| QUEENS | 1895 |
| STATEN ISLAND | 430 |

Table 16: Borough distribution on graduation data

| Borough | counts |
|---|---|
| Bronx | 7 |
| Brooklyn | 7 |
| Manhattan | 7 |
| Queens | 7 |
| Staten Island | 7 |

Table 17: Borough distribution on population data

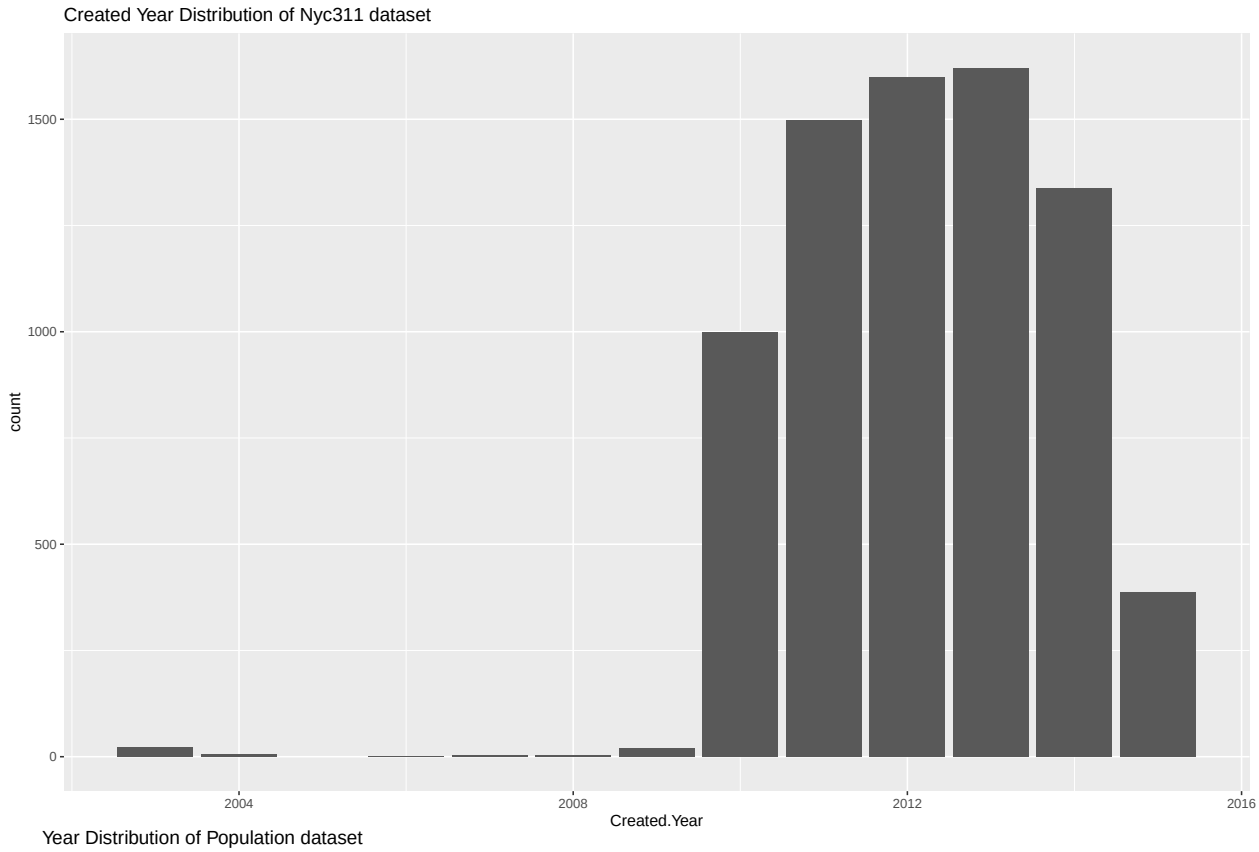
| Borough | counts |
|---|---|
| Bronx | 7 |
| Brooklyn | 7 |
| Manhattan | 7 |
| NYC Total | 7 |
| Queens | 7 |
| Staten Island | 7 |

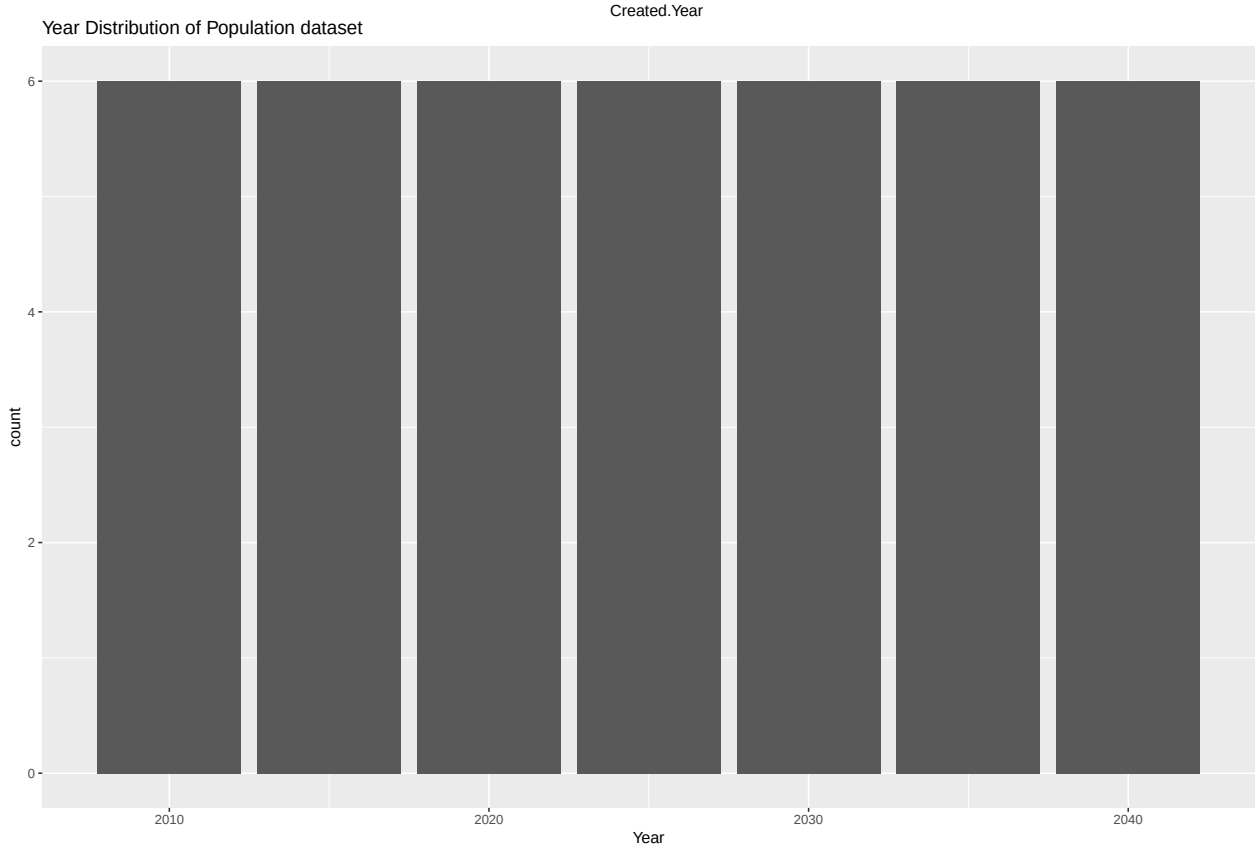
Since there are no Unspecified or null values, we are good to go. However, the values of Borough in the 2 tables table must be converted to uppercase.

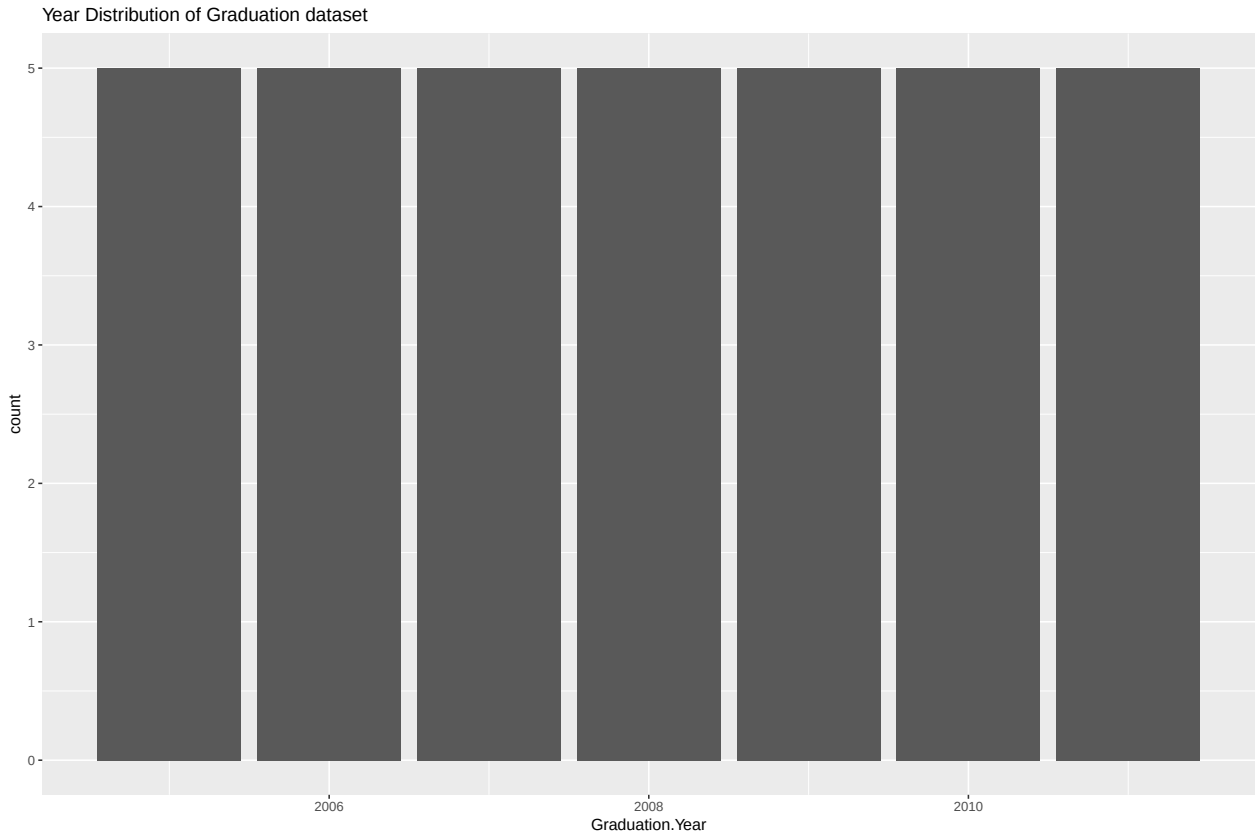
## Converting to uppercase

## Viewing the year distribution of the tables

Since these tables contain different year values, we need to sample it based on a particular year. So, the year distributions of all three datasets are checked.


Created Year Distribution of Nyc311 dataset
1500
1000
500
0
count
2004
2008
2012
2016
Created.Year


Year Distribution of Population dataset
6
4
2
0
count
2010
2020
2030
2040
Year




The results show that `2010` would be a good year to perform the analysis since it is common in all the tables.

## Joining the datasets

Now that all the datasets are processed, they are ready to be connected by the columns `Year` and `Borough` using `inner_join`. Note that the column name for `Year` differs in the three tables.

Also, 2010 is selected as the year for the analysis, hence all the tables are filtered accordingly.

The extract of final joined dataset is shown below

Table 18: Final combined dataset (continued below)

| Created.Month | Created.Day | Created.Year | Created.Time | Agency |
|---|---|---|---|---|
| 2 | 4 | 2010 | 01:00:00 PM | DEP |
| 8 | 14 | 2010 | 09:03:00 AM | DSNY |
| 6 | 29 | 2010 | 07:31:30 PM | DOB |
| 3 | 27 | 2010 | 01:39:00 PM | DOT |
| 7 | 21 | 2010 | 12:41:01 AM | DOT |
| 6 | 6 | 2010 | 09:35:00 PM | DSNY |

Table 19: Table continues below

| Agency.Name | Complaint.Type | Descriptor | Location.Type | City |
|---|---|---|---|---|
| Department of Environmental Protection | Water System | Leak (Use Comments) (WA2) | NA | NEW YORK |
| BCC - Brooklyn North | Derelict Vehicles | 14 Derelict Vehicles | Street | BROOKLYN |
| Department of Buildings | General Construction/Plumbing | Building Permit - None | NA | WOODSIDE |
| Department of Transportation | Traffic Signal Condition | Controller | NA | EAST ELMHURST |
| Department of Transportation | Street Condition | Pothole | NA | JAMAICA |
| Queens West 06 | Derelict Vehicles | 14 Derelict Vehicles | Street | REGO PARK |

Table 20: Table continues below

| Status | Borough | School.Number | Latitude | Longitude | Created.Hour |
|---|---|---|---|---|---|
| Closed | MANHATTAN | Unspecified | 40.72 | -73.99 | 13 |
| Pending | BROOKLYN | Unspecified | 40.66 | -73.92 | 9 |
| Closed | QUEENS | Unspecified | 40.73 | -73.9 | 19 |
| Closed | QUEENS | Unspecified | 40.76 | -73.9 | 13 |
| Closed | QUEENS | Unspecified | 40.72 | -73.79 | 0 |
| Pending | QUEENS | Unspecified | 40.73 | -73.85 | 21 |

Table 21: Table continues below

| Resolve.Time.Hours | Created.Weekday | Population | Total.Cohort.Num |
|---|---|---|---|
| 502.2 | Thursday | 1585873 | 7766 |
| 2.95 | Saturday | 2552911 | 11130 |
| 1612 | Tuesday | 2250002 | 6867 |
| 0.6 | Saturday | 2250002 | 6867 |
| 12.98 | Wednesday | 2250002 | 6867 |
| 0 | Sunday | 2250002 | 6867 |

| Total.Grads.Num | Total.Grads.Pct.of.cohort |
|---|---|
| 5050 | 89.80% |
| 7322 | 83.30% |
| 4289 | 83.40% |
| 4289 | 83.40% |
| 4289 | 83.40% |
| 4289 | 83.40% |

# Findings

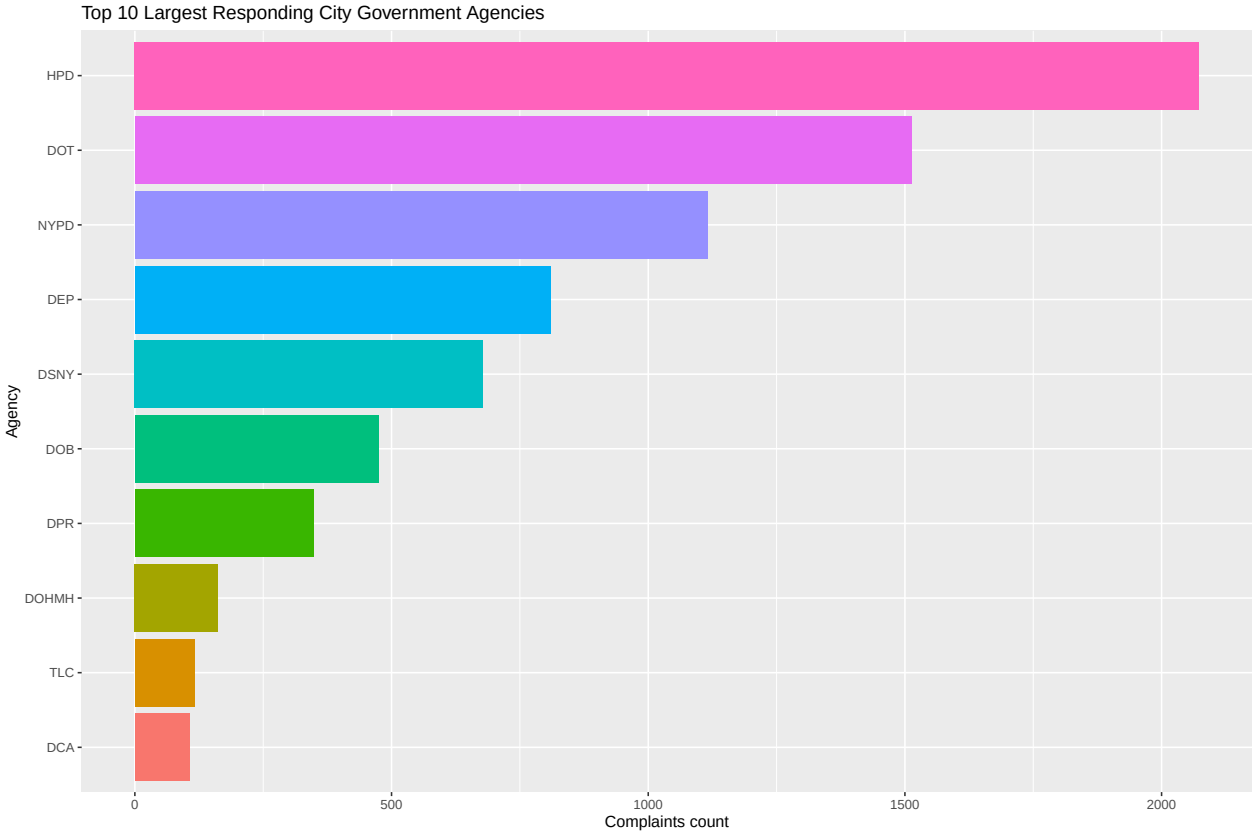
## Top 10 Largest Responding City Government Agencies

We find out the top 10 city government agencies in terms of the largest Service Requests (SR) with the count and proportion count (in percentage).

Table 23: Top 10 Largest Responding City Government Agencies

| Agency | Complaints count | Proportion in % |
|---|---|---|
| HPD | 2073 | 27.64 |
| DOT | 1513 | 20.17 |
| NYPD | 1115 | 14.86 |
| DEP | 810 | 10.8 |
| DSNY | 678 | 9.039 |
| DOB | 475 | 6.332 |
| DPR | 348 | 4.639 |
| DOHMH | 162 | 2.16 |
| TLC | 116 | 1.546 |
| DCA | 107 | 1.426 |

Then we visualize it using a bar chart.




It is clear that HPD and DOT are biggest agencies in terms of receiving the 311 calls.

## Most frequent Complaint Categories

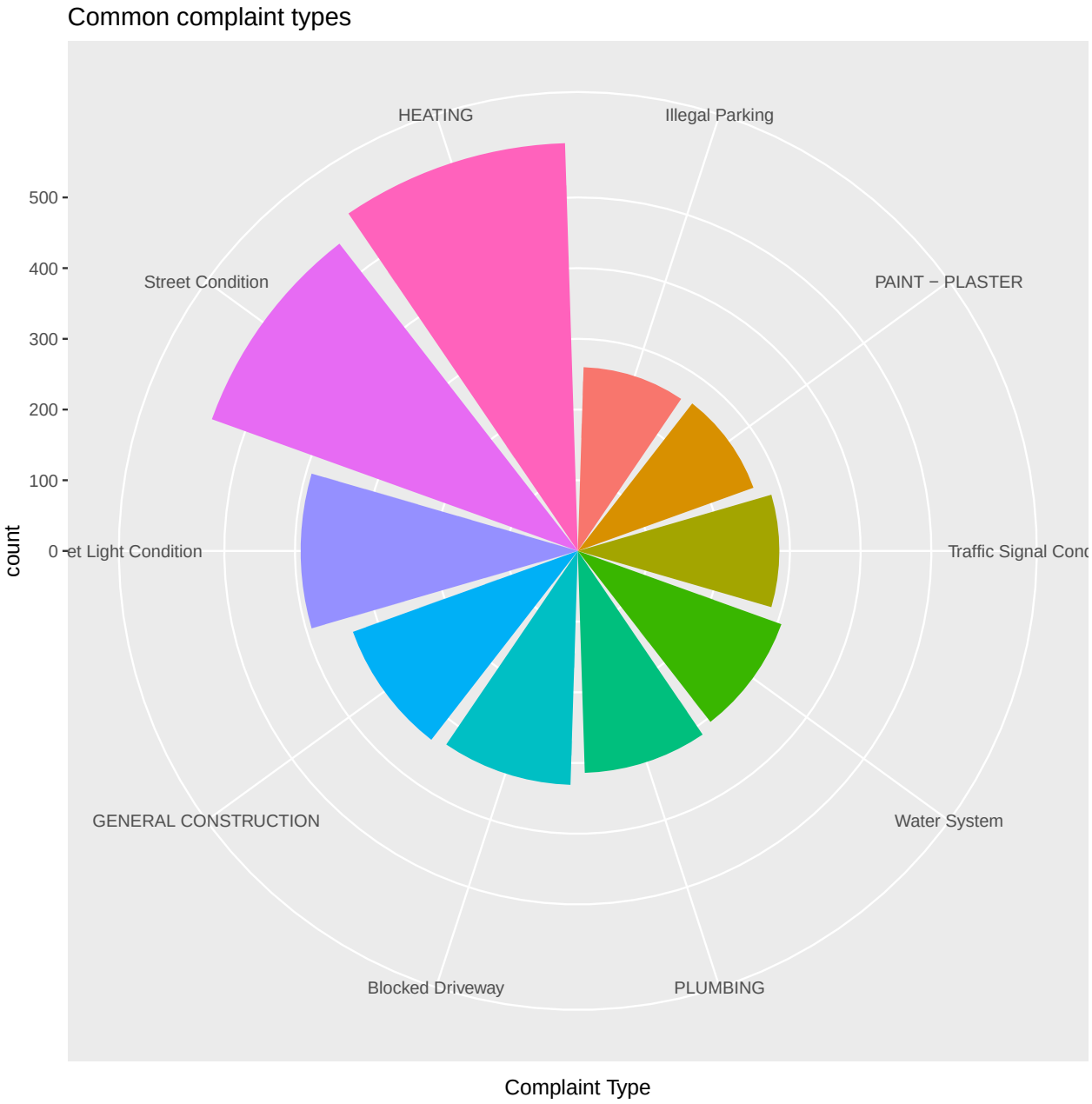
Let's view the Top 10 most frequent categories of the complaints registered along with the count and count %.

Table 24: Top 10 most frequent complaint categories

| Complaint.Type | count | proportion in % |
|---|---|---|
| HEATING | 577 | 7.692 |
| Street Condition | 550 | 7.332 |
| Street Light Condition | 392 | 5.226 |
| GENERAL CONSTRUCTION | 338 | 4.506 |
| Blocked Driveway | 331 | 4.413 |
| PLUMBING | 314 | 4.186 |
| Water System | 306 | 4.079 |
| Traffic Signal Condition | 285 | 3.799 |
| PAINT - PLASTER | 264 | 3.52 |
| Illegal Parking | 260 | 3.466 |

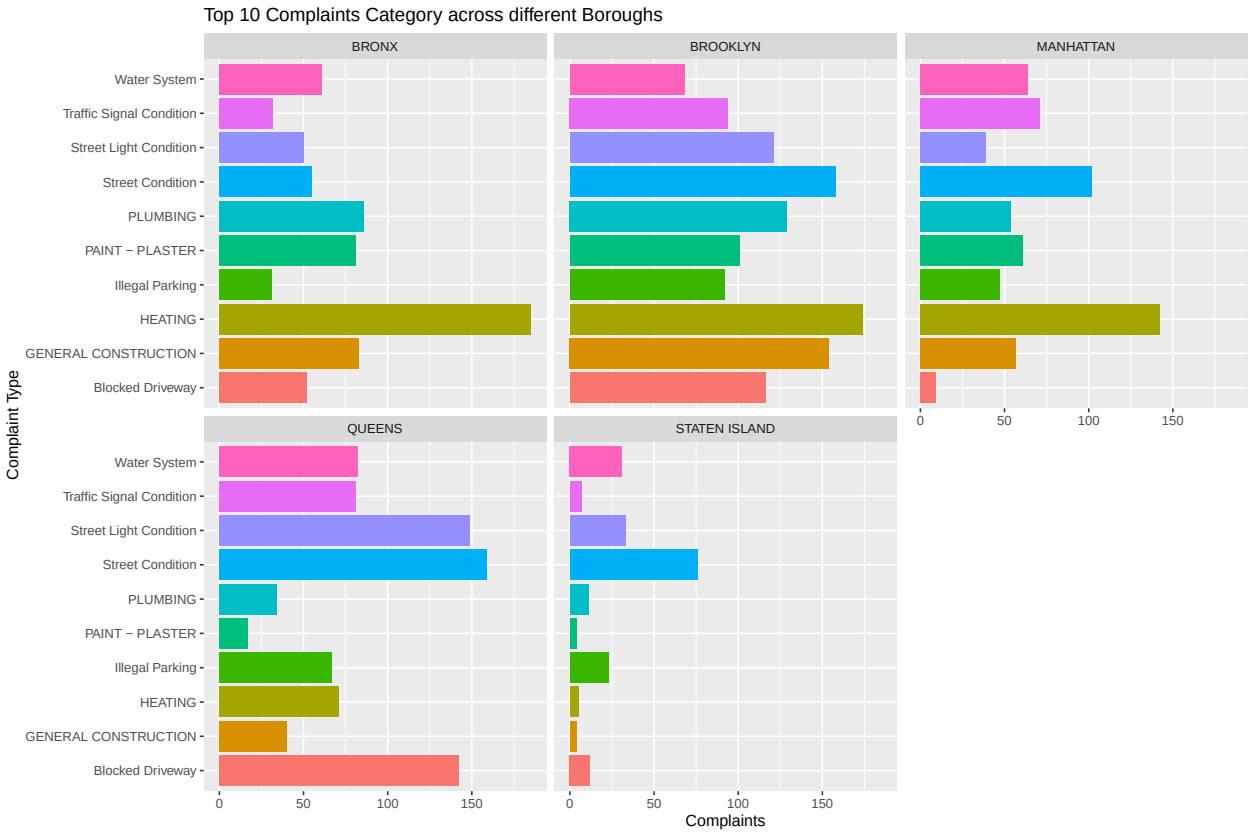
We can see that HEATING and Street Condition are very common complaints.

## Top complaint types



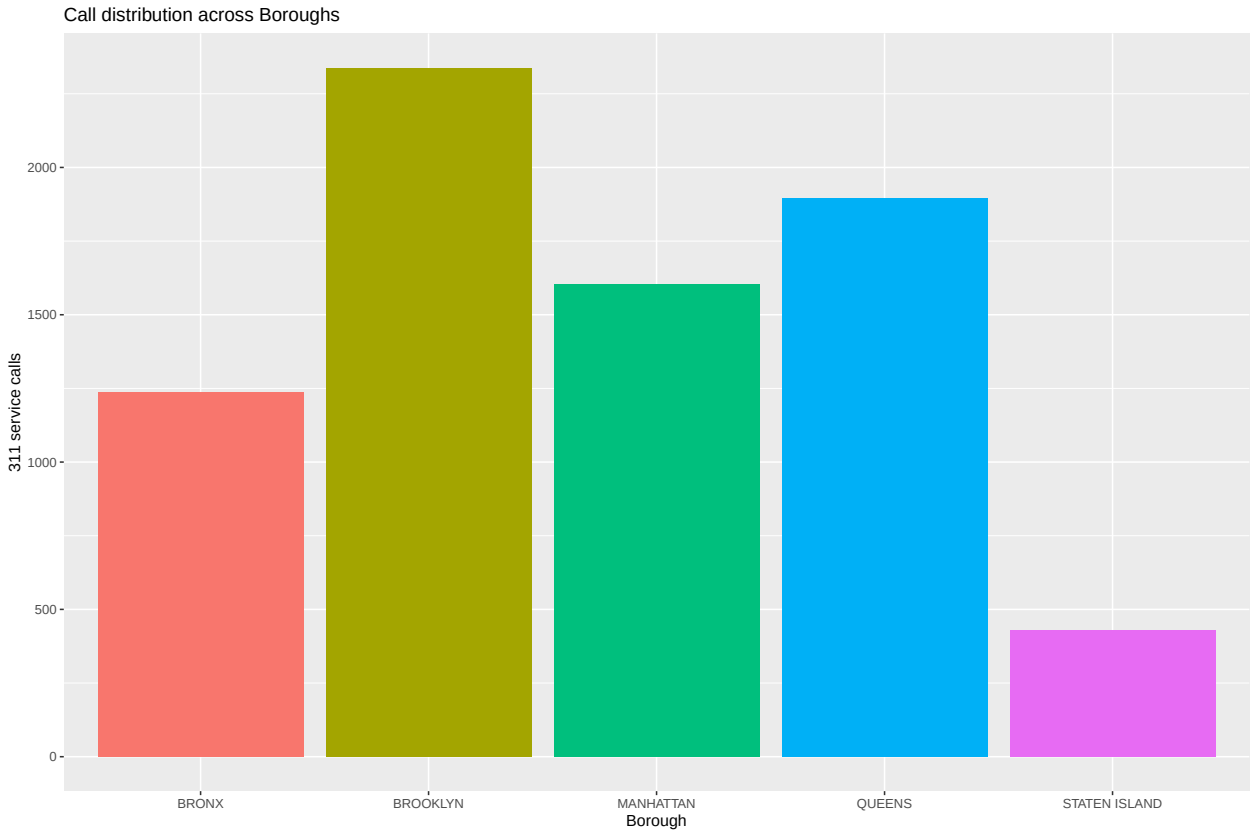

## Most freqeuent Complaint Categories across different Boroughs

Now, let's view the counts of the top 10 frequent complain categories across different Boroughs using a facet plot.




## Complaints across boroughs

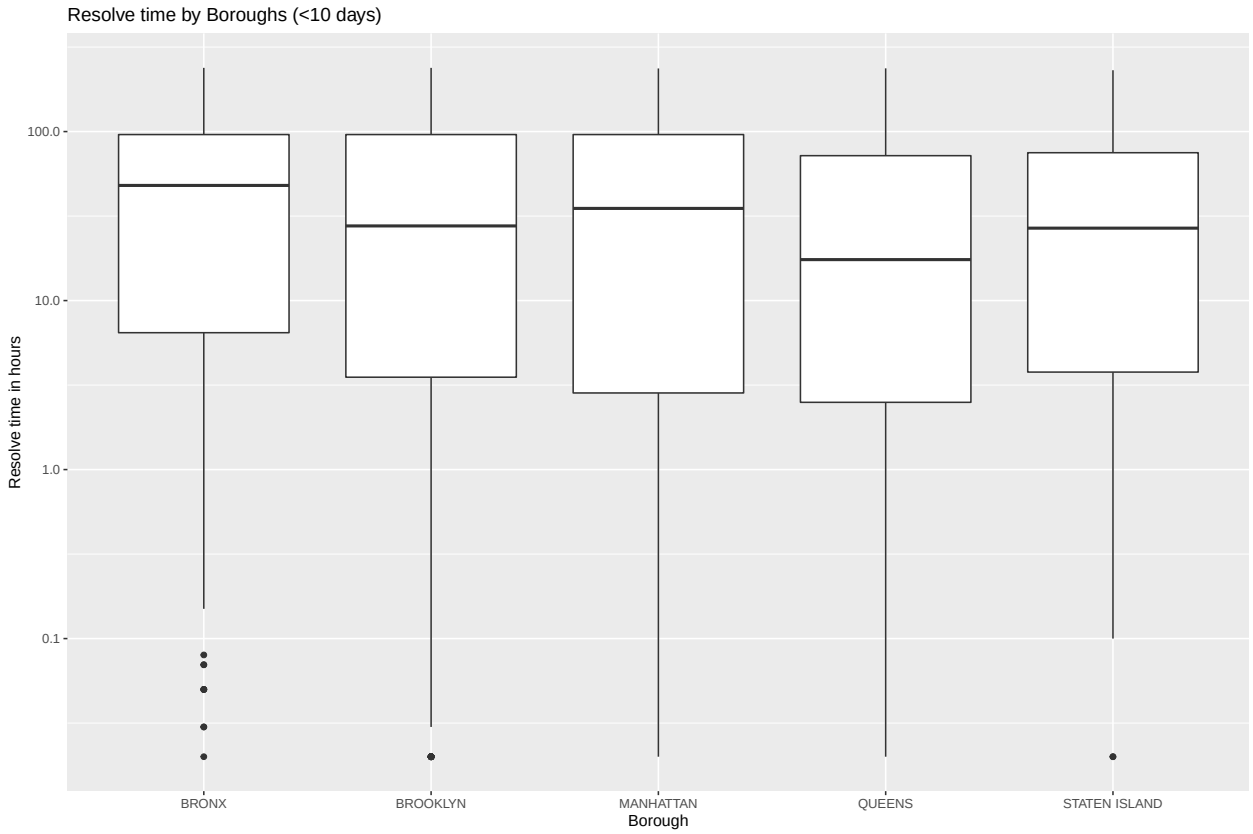
Let's find out the distribution of the calls across the Boroughs.




Hence, Brooklyn and Queens are more busy than other Boroughs.

## Resolve time across Boroughs

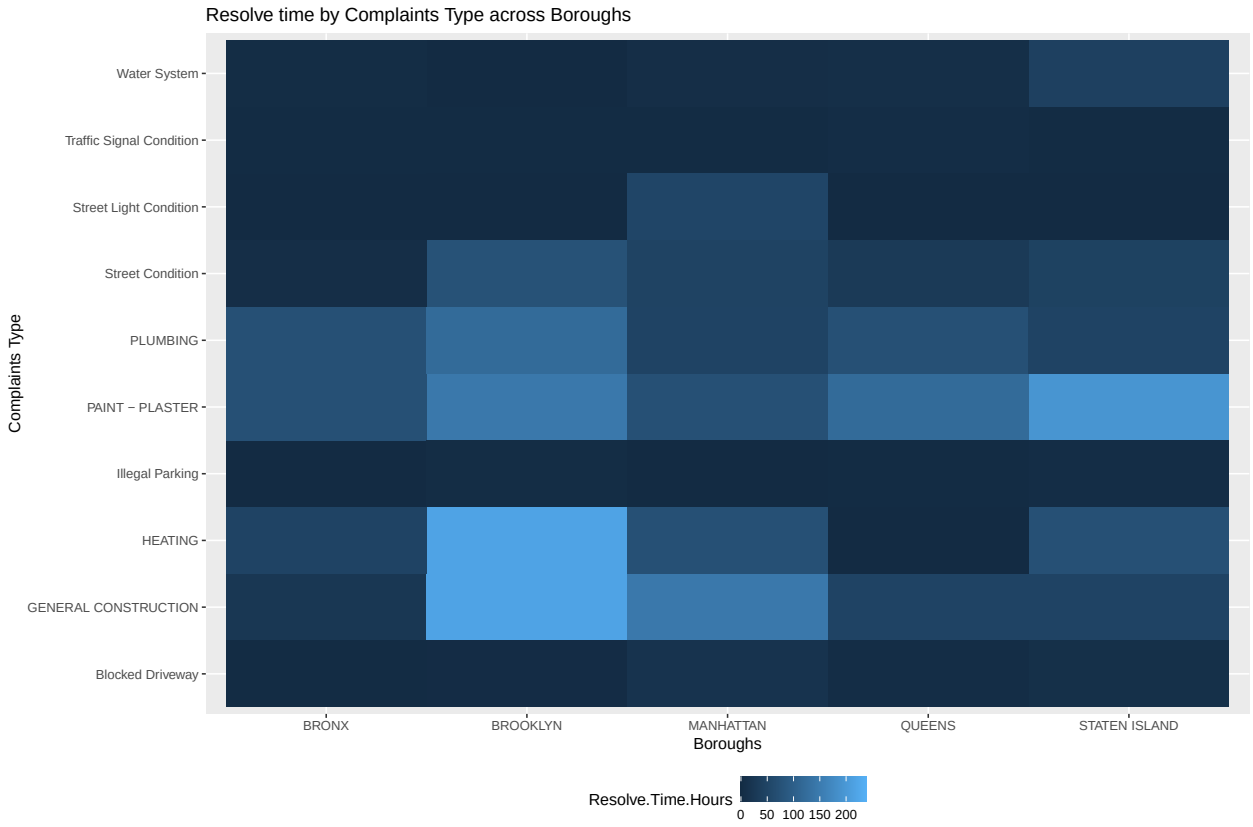
Let's analyze which Boroughs resolve the complaints quickly. We first filter the resolve time less than 10 days.




Hence, we can see that Queens has the quickest resolve time compared to others.

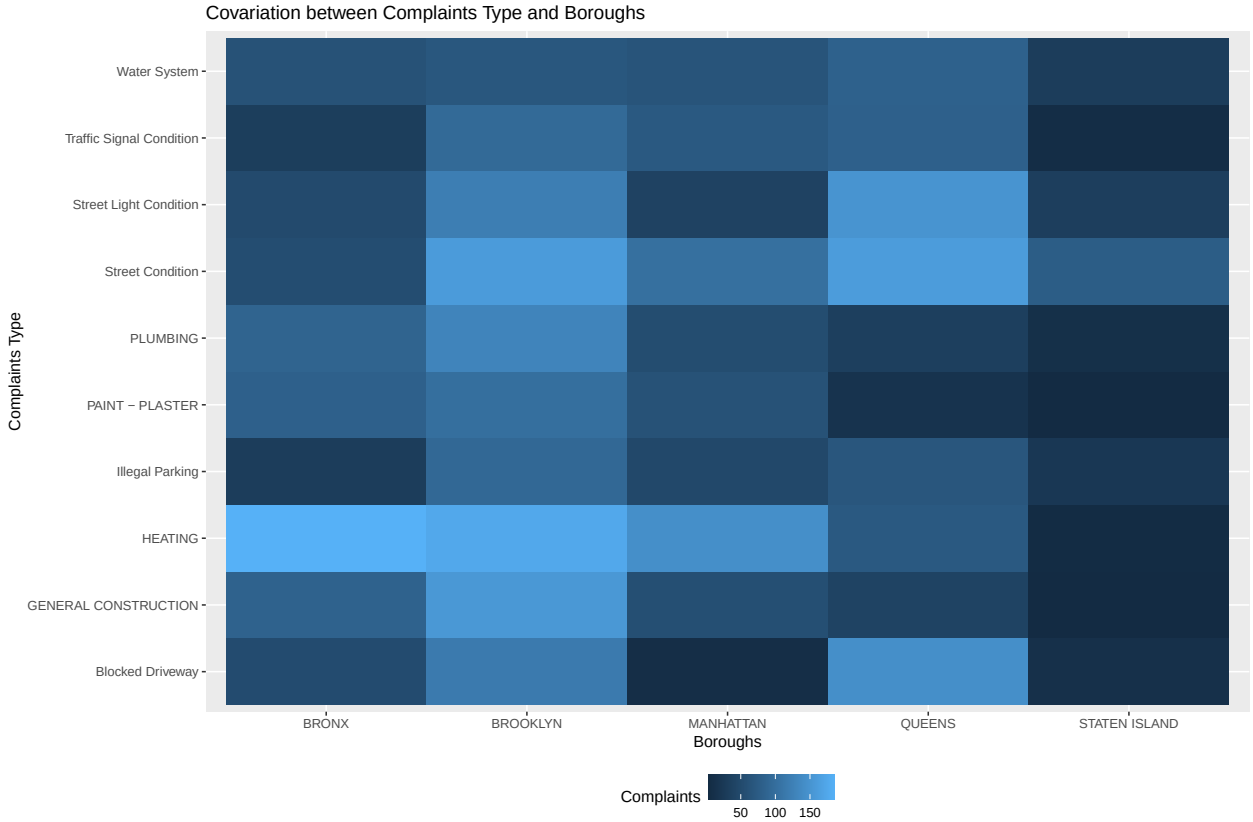
## Resolve time by Complaint Types and Boroughs

Let's see the variation of resolve time on the basis of Complaint Category across Boroughs. First, we filter resolve time less than 10 days only.




Hence, we can see that Street condition complaint takes longer to resolve by Brooklyn and Staten Island. On the other hand, complaints like Water system, Traffic Signal Condition and BLocked Driveway are quickly resolved by all Boroughs.

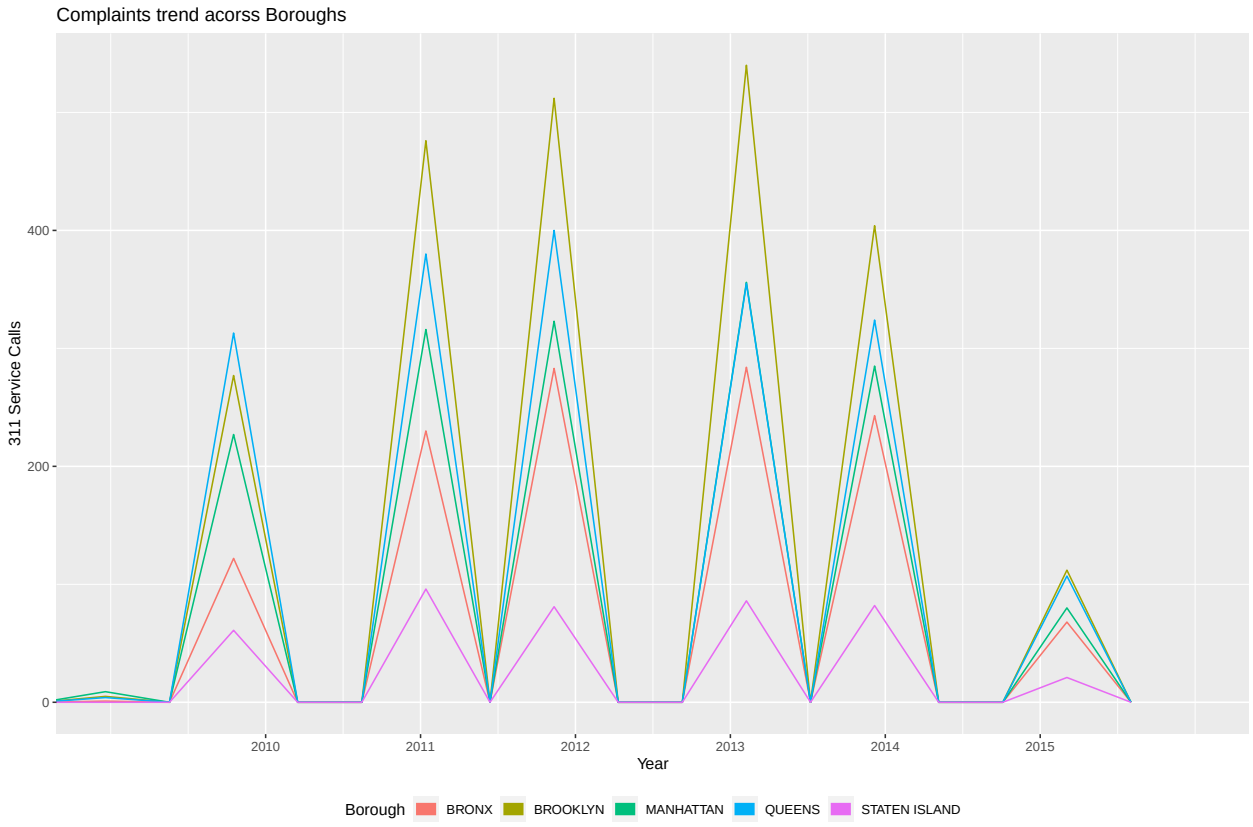
## Covariation between Complaints Type and Boroughs




We can see that Brooklyn has relatively higher number of complaints, especially heating complaints. Staten island have very few complaints.

## Complaints trend over Boroughs

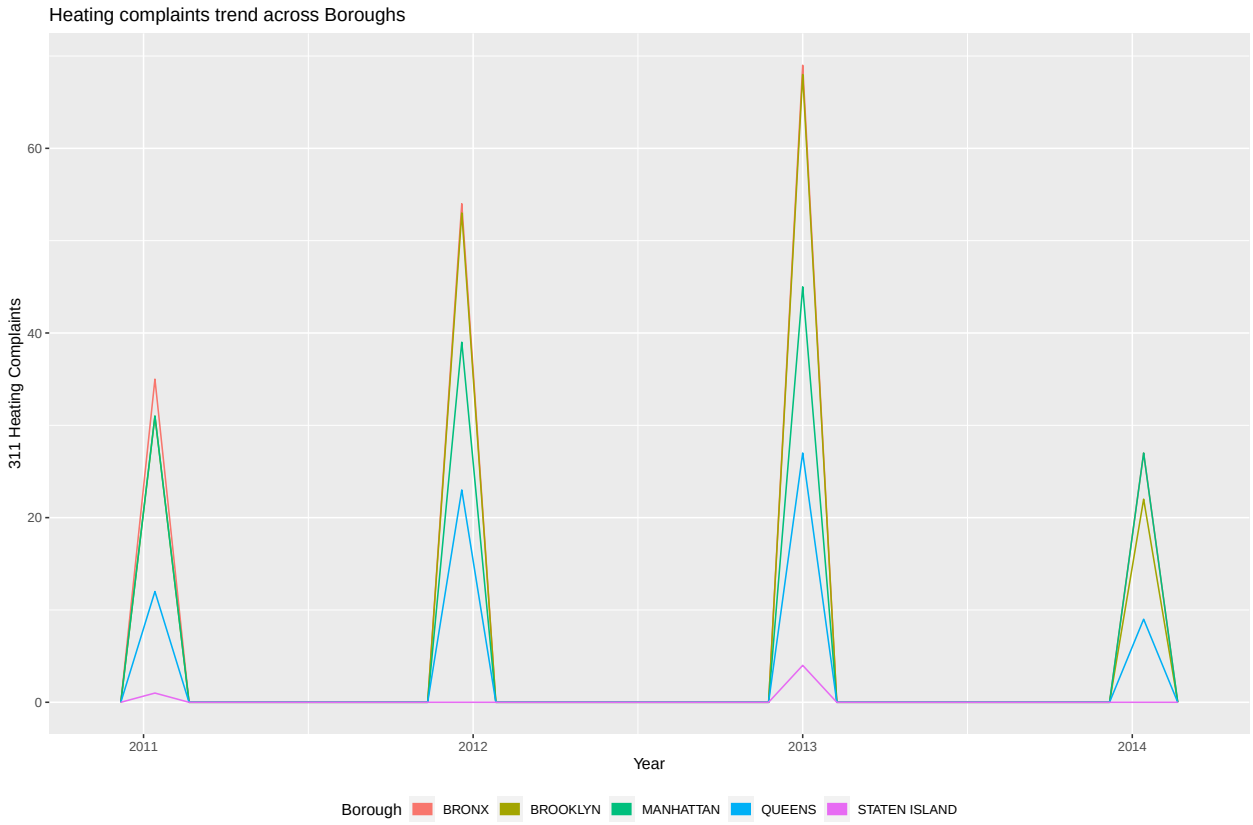
Now, let's see the frequency of complaints over the years across different Boroughs.




2012 saw a spike in complaints in Brooklyn whereas complaints in Queens was at peak in 2013. Highly effective corrective measures seem to have been employed in almost all Boroughs since the numbers dropped to less than a quarter in 2015.

## Heating Complaints trend over Boroughs

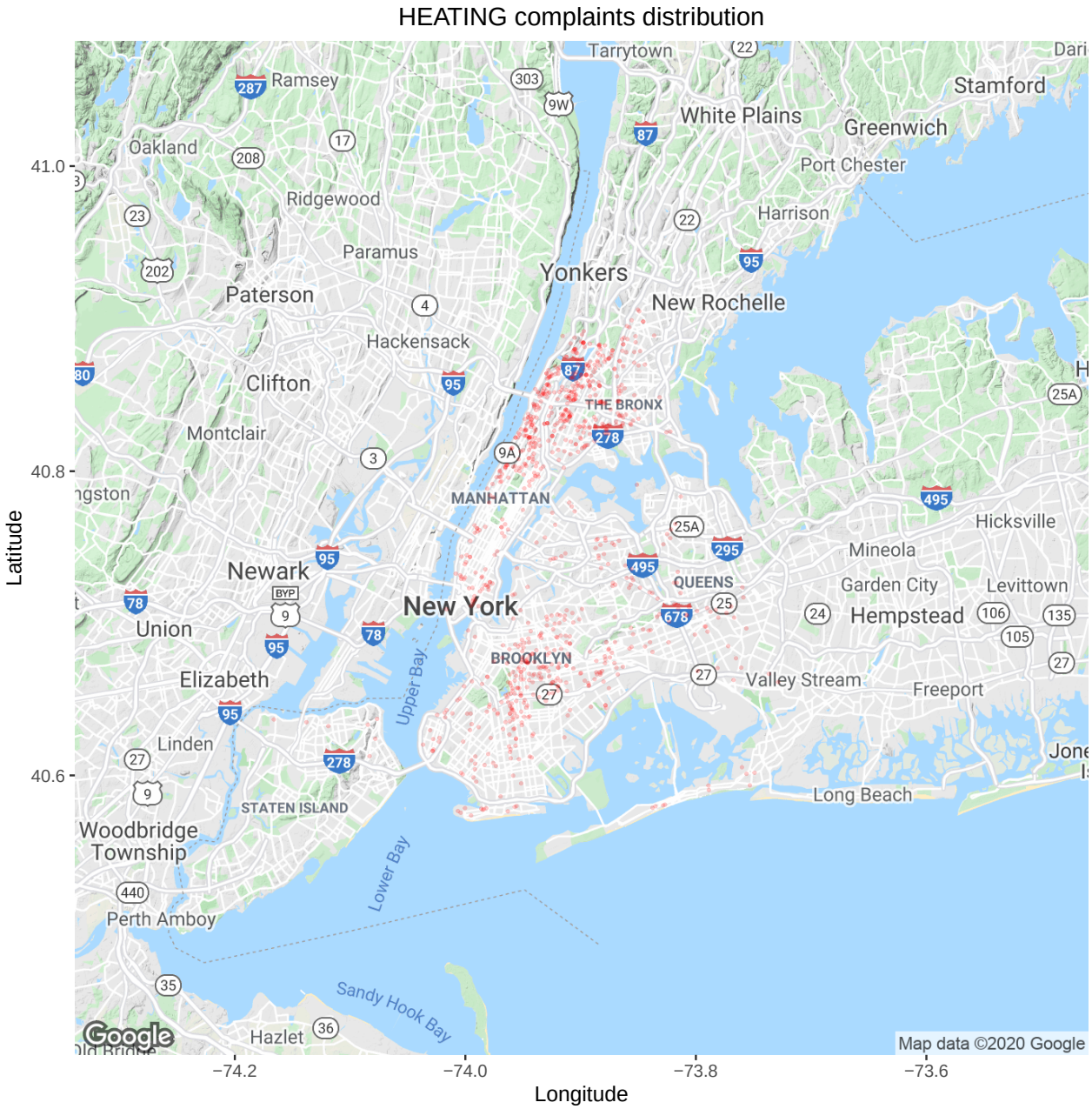
Now, let's see the evolution of heating complaints over the years across different Boroughs.




We can see that Bronx has peak heating complaints in 2013 and 2014 while Staten Island has almost flat curve u=indicating no significant rise in heating complaints.

## Distirbution of Heating complaints in maps

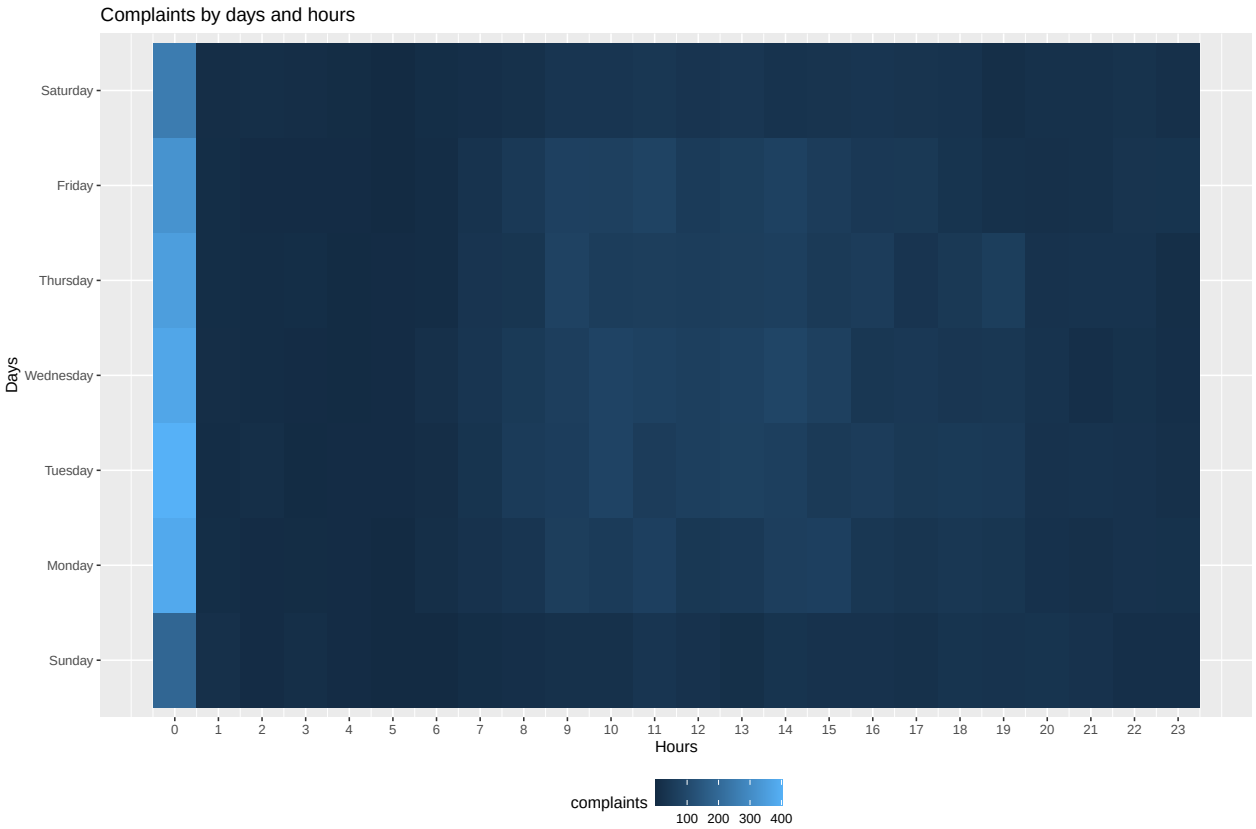
This is done on a sample (10000 observations) of nyc311 data.



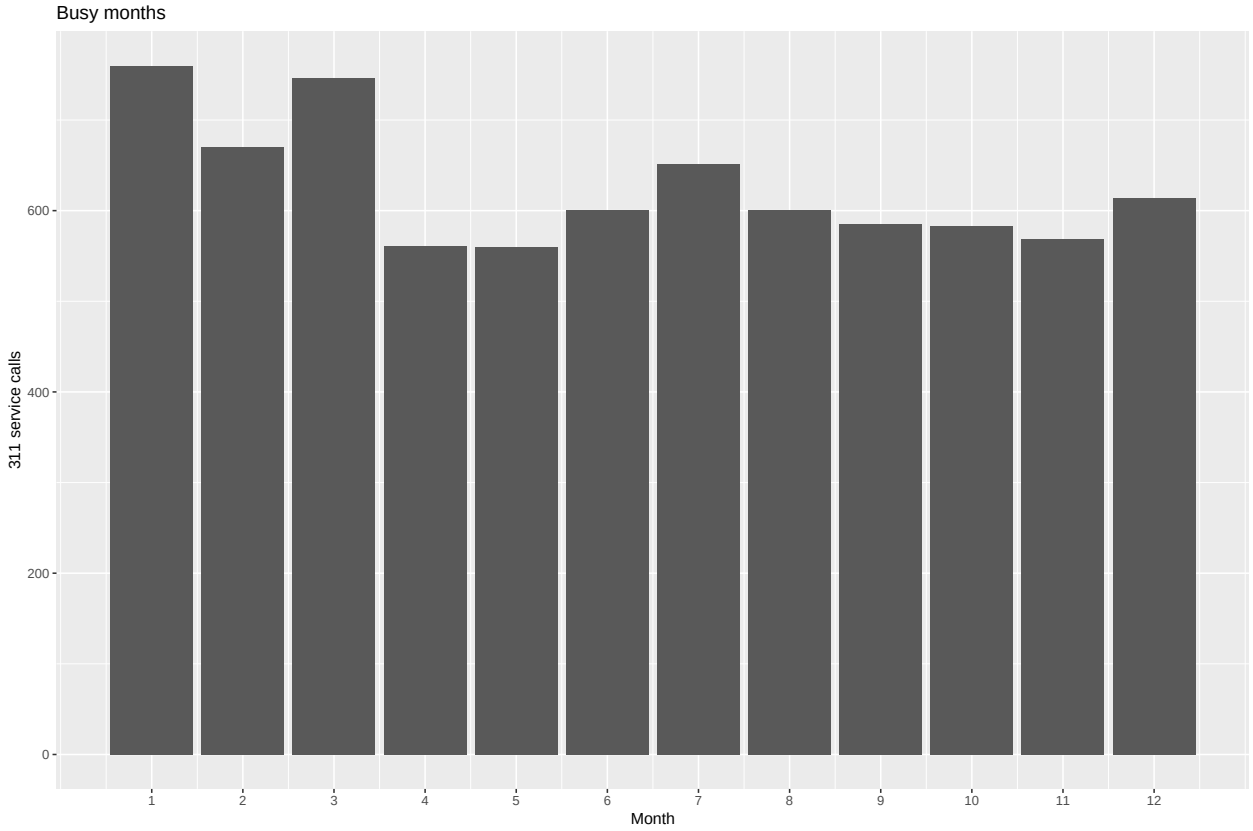

We can see the heating complaints mostly in Bronx and Brooklyn in the map.

## Busy hours and days

Let's see which days of the week and which time are more busy.

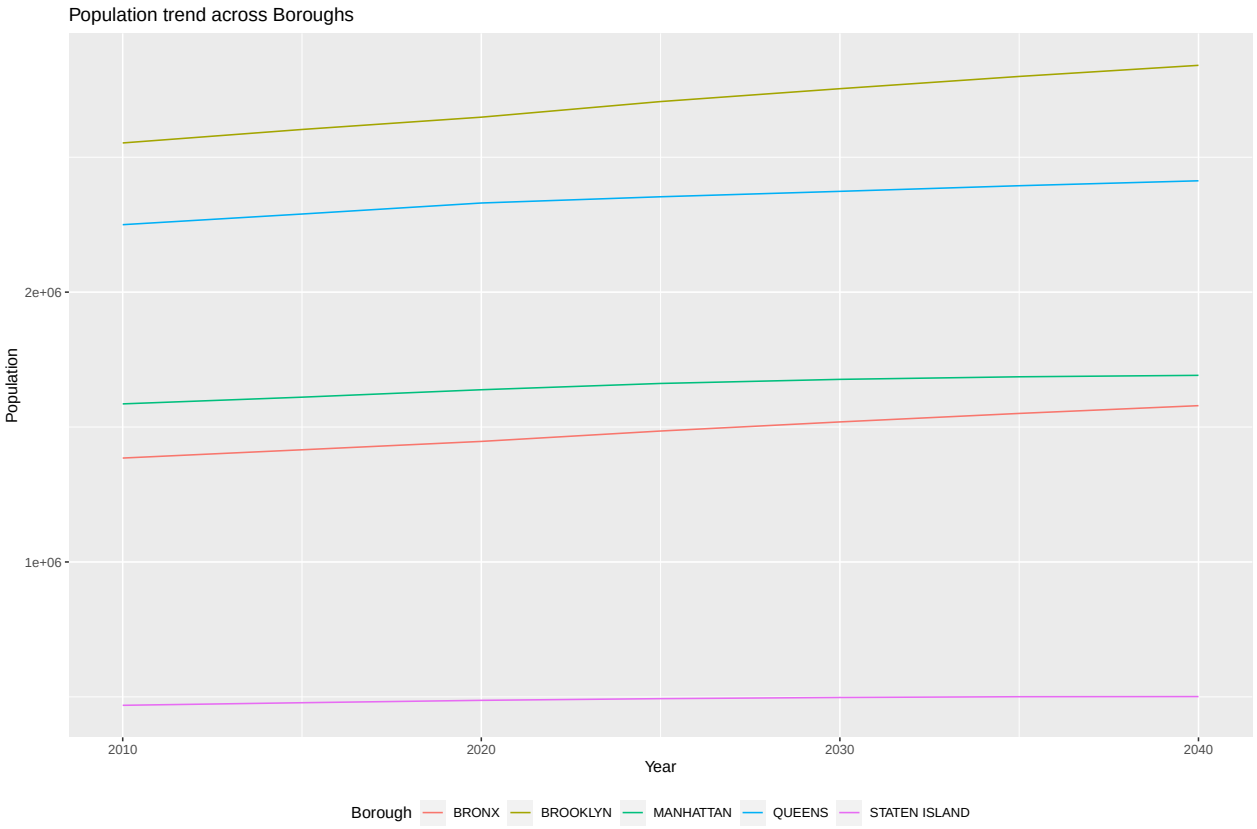

Complaints by days and hours
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
Sunday
Days
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hours
complaints
100 200 300 400

## Busy months



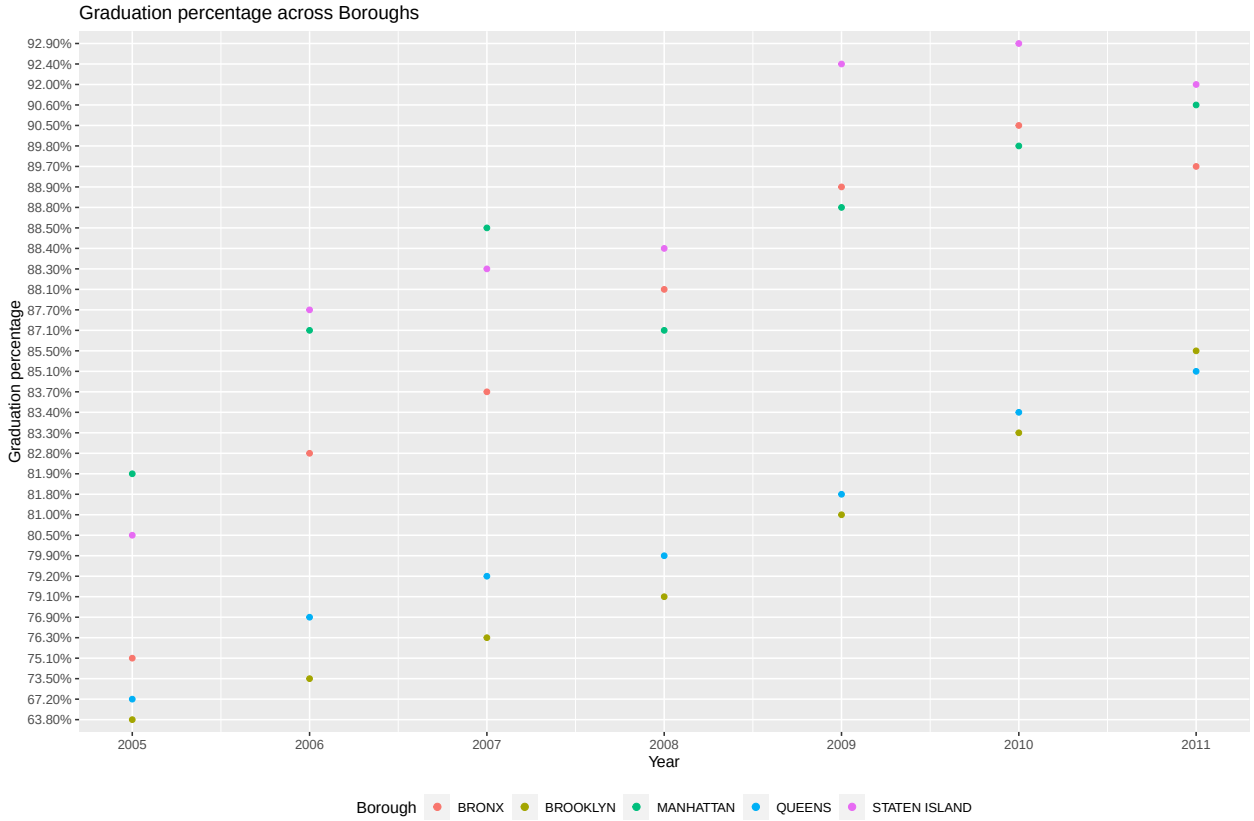

We can see that January and February are busier than other months. It may be due to the new year.

## Population trends in Boroughs




We can see that Brooklyn has the highest population over the years and is rapidly inreasing. However, Staten island's population increase is almost flat.

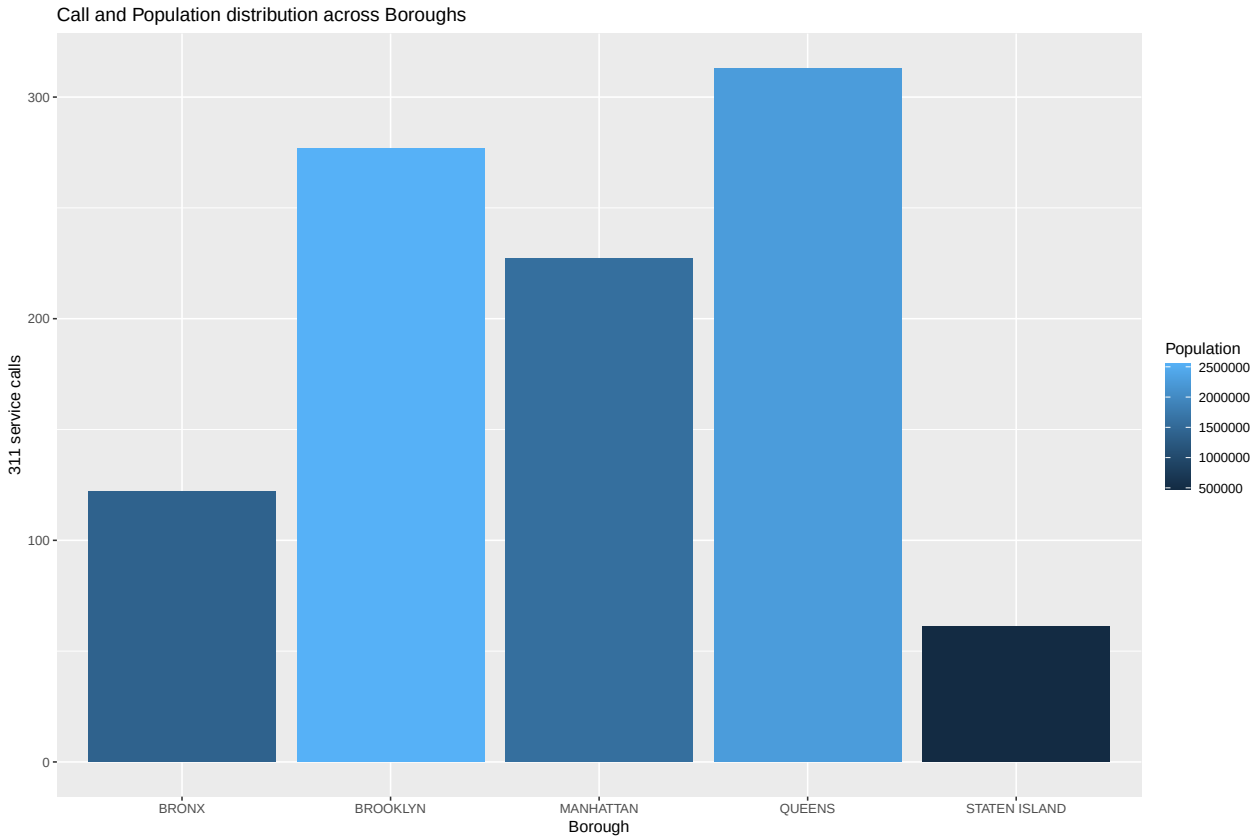
## Gradutaion percentage across Boroughs




We can see an increasing trend in percentage of graduate students every year in each Borough.

## Population and Complaints distribution across Boroughs

Let's find out the distribution of the calls and population across the Boroughs.




We can see that although Queens has a lower population than Brooklyn, it receives more 311 calls.

# Conclusions

Hence, several transformations and explorations were done on the `nyc311` data and the two additional datasets. The two datasets were also joined with the `nyc311` data by converting all the datasets into appropriate form (tidying, grouping and removal of redundant columns), performing a inner join by columns `Year` and `Borough` and finally filtering it by the `Year==2010` constraint.

We applied exploratory data analysis using various data transformation and visualization techniques. We were able to answer several questions about the data. Several analysis were done which gave the following findings.

1. It is clear that HPD and DOT are biggest agencies in terms of receiving the 311 calls.
2. We can see that HEATING and Street Condition are very common complaints.
3. Brooklyn and Queens are more busy than other Boroughs.
4. Queens has the quickest resolve time compared to others.
5. Street condition complaint takes longer to resolve by Brooklyn and Staten Island. On the other hand, complaints like Water system, Traffic Signal Condition and BLocked Driveway are quickly resolved by all Boroughs.
6. We can see that Brooklyn has relatively higher number of complaints, especially heating complaints. Staten island have very few complaints.
7. 2012 saw a spike in complaints in Brooklyn whereas complaints in Queens was at peak in
8. Highly effective corrective measures seem to have been employed in almost all Boroughs since the numbers dropped to less than a quarter in 2015
9. We can see that Bronx has peak heating complaints in 2013 and 2014 while Staten Island has almost flat curve indicating no significant rise in heating complaints.
10. We can see the heating complaints mostly in Bronx and Brooklyn in the map.

11. We can see that January and February are busier than other months. It may be due to the new year.
12. We can see that Brooklyn has the highest population over the years and is rapidly inreasing.
13. We can see an increasing trend in percentage of graduate students every year in each Borough.
14. We can see that although Queens has a lower population than Brooklyn, it receives more 311 calls. However, Staten island's population increase is almost flat.

# Appendices

## Data Dictionary

Table 25: Data dictionary for original nyc311 dataset

| Indicator | Description | DataType |
|---|---|---|
| ‘Unique Key‘ | Unique identifier of a Service Request (SR) in the open data set | Text |
| ‘Created Date‘ | Date SR was created | Floating Timestamp |
| ‘Closed Date‘ | Date SR was closed by responding agency | Floating Timestamp |
| ‘Agency‘ | Acronym of responding City Government Agency | Text |
| ‘Agency Name‘ | Full Agency name of responding City Government Agency | Text |
| ‘Complaint Type‘ | This is the first level of a hierarchy identifying the topic of the incident or condition.Complaint Type may have a corresponding Descriptor (below) or may stand alone. | Text |
| ‘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. | Text |
| ‘Location Type‘ | Describes the type of location used in the address information | Text |
| ‘Incident Zip‘ | Incident location zip code, provided by geo validation. | Text |
| ‘Incident Address‘ | House number of incident address provided by submitter. | Text |
| ‘Street Name‘ | Street name of incident address provided by the submitter | Text |
| ‘Cross Street 1‘ | First Cross street based on the geo validated incident location | Text |
| ‘Cross Street 2‘ | Second Cross Street based on the geo validated incident location | Text |
| ‘Intersection Street 1‘ | First intersecting street based on geo validated incident location | Text |
| ‘Intersection Street 2‘ | Second intersecting street based on geo validated incident location | Text |
| ‘Address Type‘ | Type of incident location information available. | Text |
| ‘City‘ | City of the incident location provided by geovalidation. | Text |
| ‘Landmark‘ | If the incident location is identified as a Landmark the name of the landmark will display here | Text |
| ‘Facility Type‘ | If available, this field describes the type of city facility associated to the SR | Text |
| ‘Status‘ | Status of SR submitted | Text |
| ‘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). | Floating Timestamp |
| ‘Resolution Description‘ | Describes the last action taken on the SR by the responding agency. May describe next or future steps. | Text |
| ‘Resolution Action Updated Date‘ | Date when responding agency last updated the SR. | Floating Timestamp |
| ‘Community Board‘ | Provided by geovalidation. | Text |
| ‘Borough‘ | Provided by the submitter and confirmed by geovalidation. | Text |
| ‘X Coordinate (State Plane)‘ | Geo validated, X coordinate of the incident location. | Number |
| ‘Y Coordinate (State Plane)‘ | Geo validated, Y coordinate of the incident location. | Number |
| ‘Park Facility Name‘ | If the incident location is a Parks Dept facility, the Name of the facility will appear here | Text |
| ‘Park Borough‘ | The borough of incident if it is a Parks Dept facility | Text |
| ‘Vehicle Type‘ | If the incident is a taxi, this field describes the type of TLC vehicle. | Text |
| ‘Taxi Company Borough‘ | If the incident is identified as a taxi, this field will display the borough of the taxi company. | Text |
| ‘Taxi Pick Up Location‘ | If the incident is identified as a taxi, this field displays the taxi pick up location | Text |
| ‘Bridge Highway Name‘ | If the incident is identified as a Bridge/Highway, the name will be displayed here. | Text |
| ‘Bridge Highway Direction‘ | If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here. | Text |
| ‘Road Ramp‘ | If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp. | Text |
| ‘Latitude‘ | Geo based Lat of the incident location | Number |
| ‘Longitude‘ | Geo based Long of the incident location | Number |
| ‘Location‘ | Combination of the geo based lat & long of the incident location | Location |

Table 26: Data dictionary for Projected Population 2010-2040 dataset

| Column.Name | Description | DataType |
|---|---|---|
| Borough | Name of the New York City Borough | Text |
| Age | One of 18 Age cohorts like '0-4', '15-19', 'Total', and so on | Text |
| Year | Year in which the population is projected | Number |
| Population | The projected population value | Number |

Table 27: Data dictionary for 2005-2011 Graduation Outcomes dataset

| Column.Name | Description | DataType |
|---|---|---|
| Borough | Name of the New York City Borough | Text |
| Graduation.Year | The cohort's year of graduation | Number |
| Total.Cohort.Num | Number of students in the cohort | Number |
| Total.Grads.Num | Number of students who graduated in the cohort | Number |
| Total.Grads.Pct.of.cohort | Percentage of students who graduated in the cohort | Number |

Table 28: Data dictionary for final nyc combined dataset

| Column.Name | Description | DataType |
|---|---|---|
| Created.Month | Month SR was created (1-12) | Number |
| Created.Day | Day of month SR was created (1-31) | Number |
| Created.Year | Year SR was created | Number |
| Created.Time | Time SR was created | Floating Timestamp |
| Created.Hour | Hour SR was created (0-24) | Number |
| Created.Weekday | Day of week SR was created corresponding to (Sunday - Saturday) | Text |
| Resolve.Time.Hours | Time taken to resolve the complaint in hours | Number |
| Population | Total population of the Borough | Number |
| Total.Cohort.Num | Number of students in the cohort | Number |
| Total.Grads.Num | Number of students who graduated in the cohort | Number |
| Total.Grads.Pct.of.cohort | Percentage of students who graduated in the cohort | Number |
| Agency | Acronym of responding City Government Agency | Text |
| Agency Name | Full Agency name of responding City Government Agency | Text |
| Complaint Type | This is the first level of a hierarchy identifying the topic of the incident or condition.Complaint Type may have a corresponding Descriptor (below) or may stand alone. | Text |
| 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. | Text |
| Location.Type | Describes the type of location used in the address information | Text |
| Incident.Zip | Incident location zip code, provided by geo validation. | Text |
| City | City of the incident location provided by geovalidation. | Text |
| Status | Status of SR submitted | Text |
| Resolution.Description | Describes the last action taken on the SR by the responding agency. May describe next or future steps. | Text |
| Borough | Provided by the submitter and confirmed by geovalidation. | Text |
| Latitude | Geo based Lat of the incident location | Number |
| Longitude | Geo based Long of the incident location | Number |