# 311 Call Data Analysis Using MapReduce

***Project Objective***:

Like 911 calls, 311 calls are for non-emergency purposes and the helpline is setup in many of the cities in USA. The objective of the project is to visually analyze the New York city authority’s 311 calls. Like the 911 calls, 311 calls too have a similar response-time. However, the resources might not be as highly funded as the emergency services.

Data is extracted from NYC OpenData website and here’s the link to download the data:

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

Data Size: ~ 10GB (From 2010 - Current)

The idea here is to narrow down on the issues by consolidating them geographically and find the high-impact items. Below are the few of the points which were covered in the project proposal:

- The number of incidents reported geographically by Zip Code/ Borough

- Distribution of Complaint Types by Year, Borough and Zip

- Variation of Complaint Types by Borough and Month

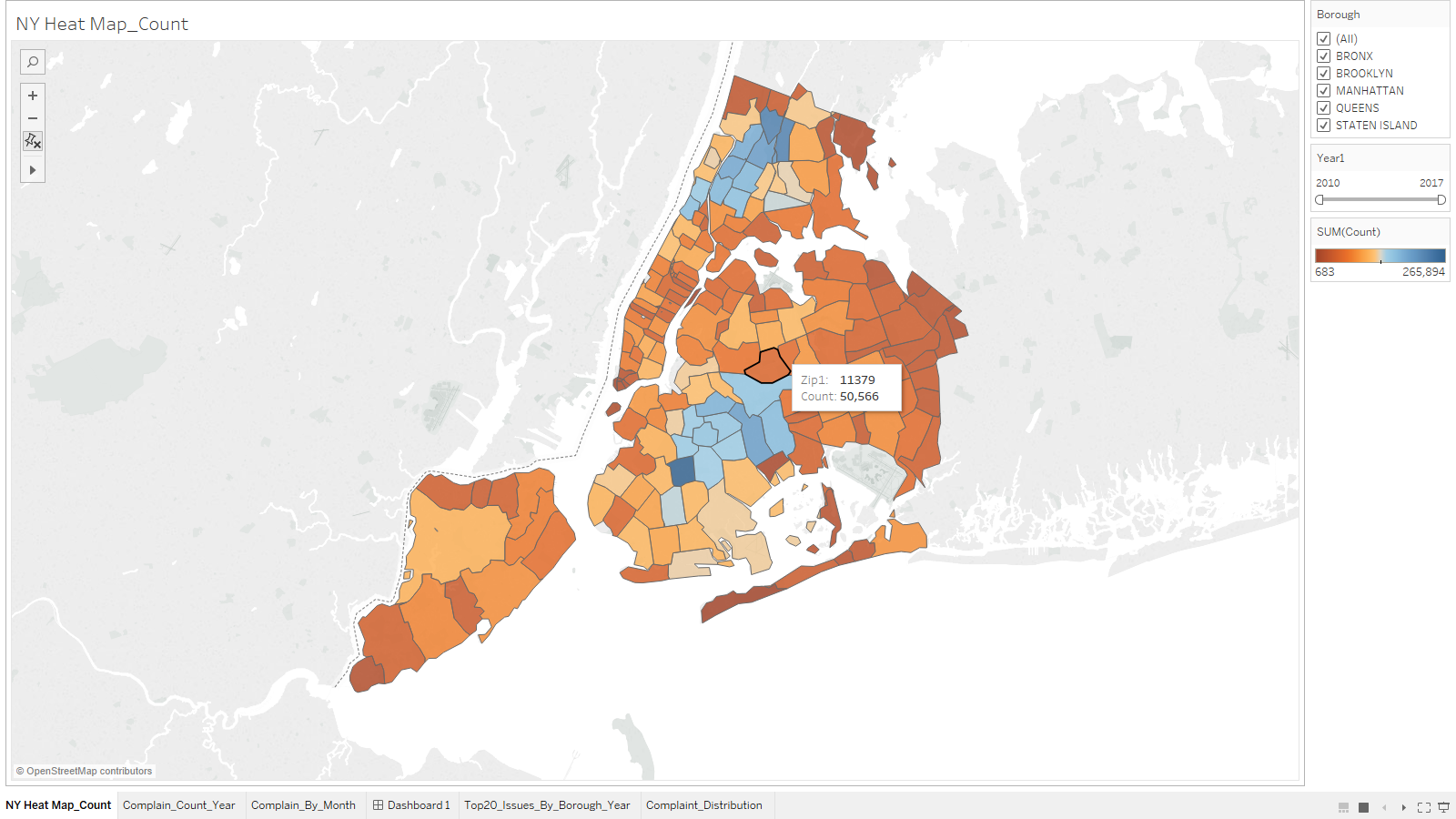
- Analyze the resolution time of Incidents by Year and Borough

- Inverted Index of Issues by Zip Code

***Analysis1****:* The number of incidents reported geographically by Zip Code/ Borough

* Data is extracted from the MR job on the 311 Dataset to extract the Zip Code and the count of Complaint Types reported by filtered by Year and zip code.
* Using this data, the NY city map can be visualized by the amount of calls reported each Zip

Below is the visual plot of number of complaints reported by zip code in Tableau

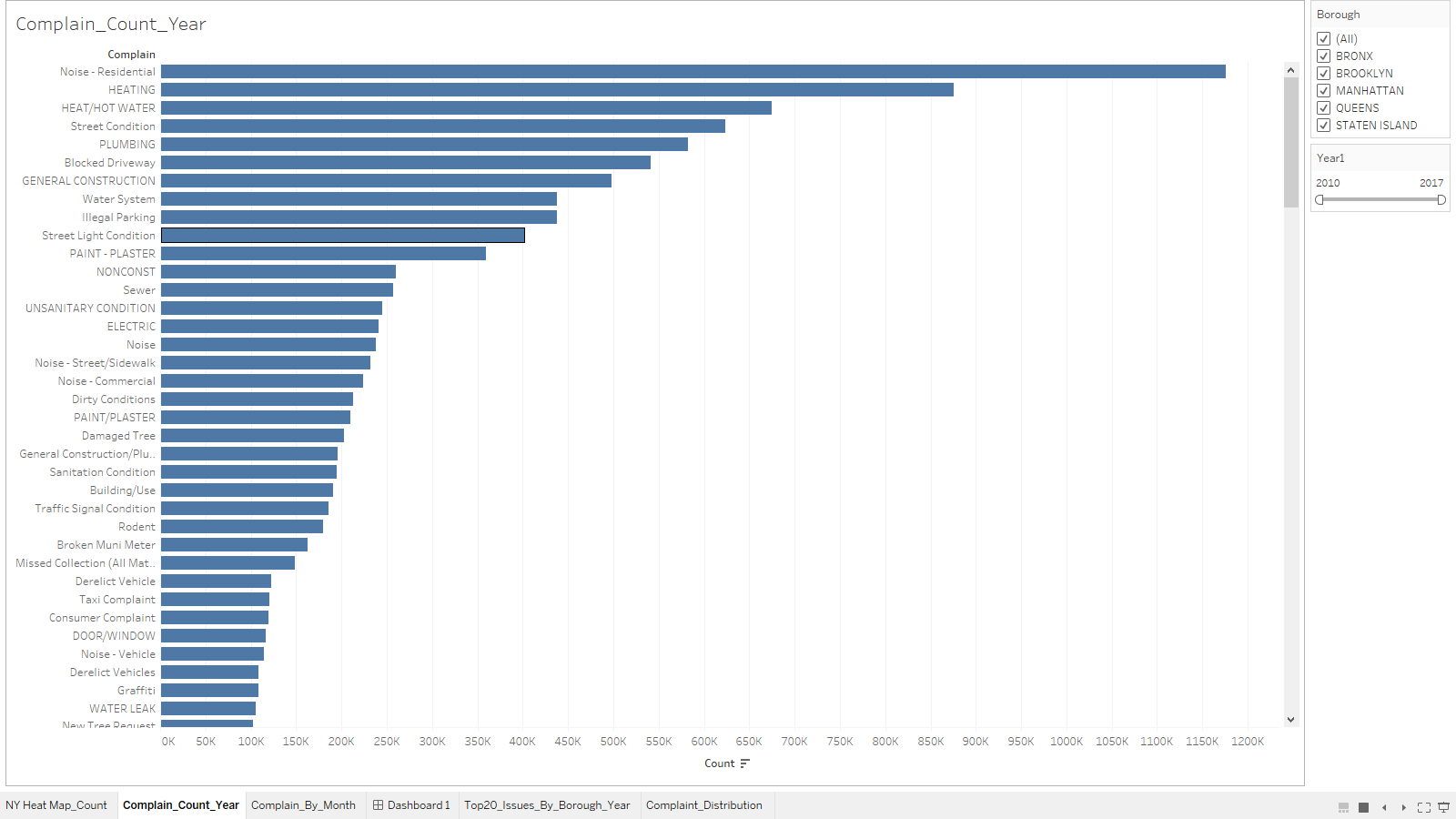


Link to MapReduce program:

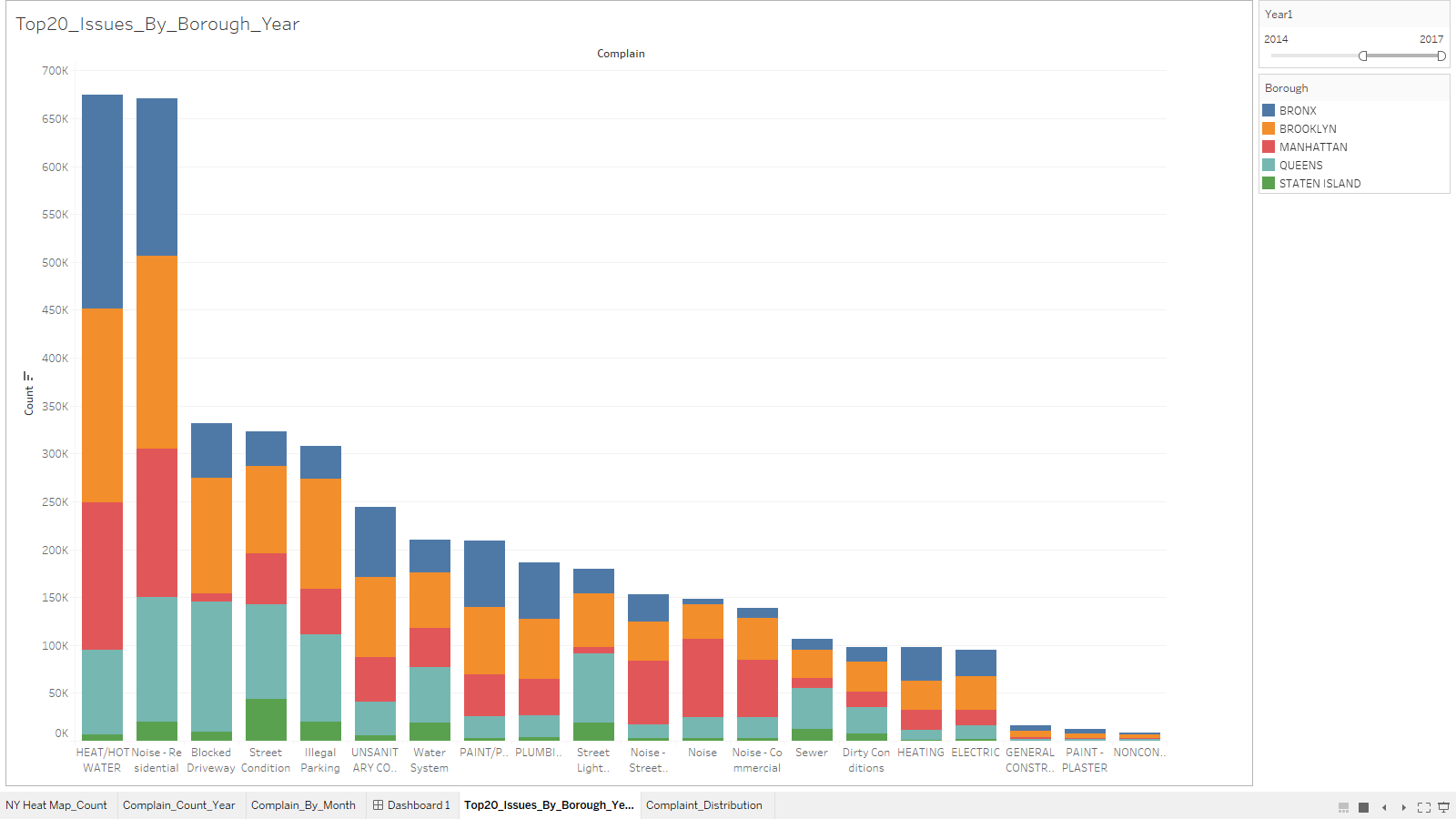
***Analysis2*:**Distribution of Complaint Types by Year, Borough and Zip

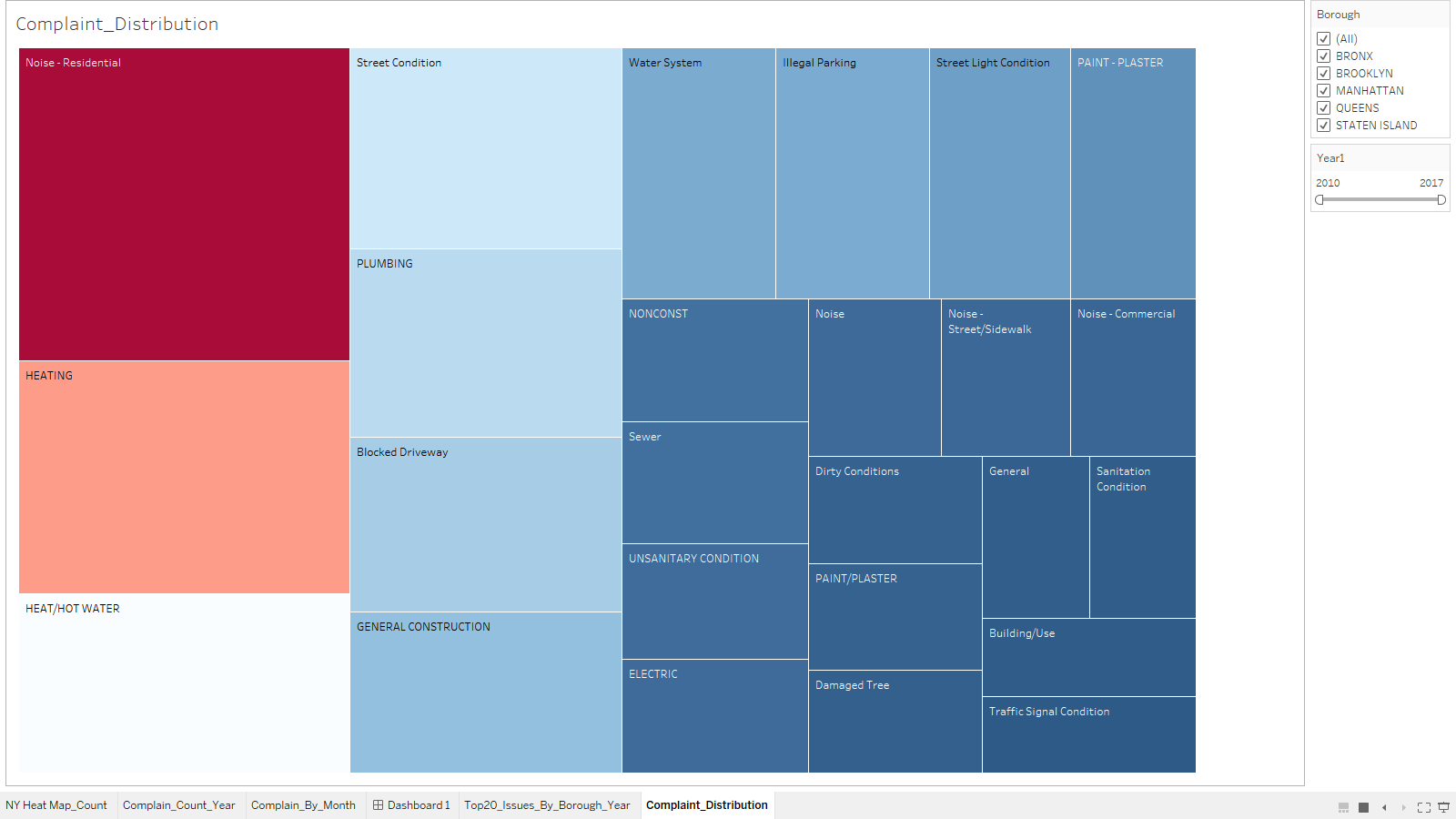
* The data is extracted from a MR job to focus on the complaint types and their distribution by zip code and year.
* The data is further drilled down into months as well.

Below is the distribution of complaint types by their count, filtered by Year and Borough



Below are few of the additional graphs to analyze the distribution of complaint type data:



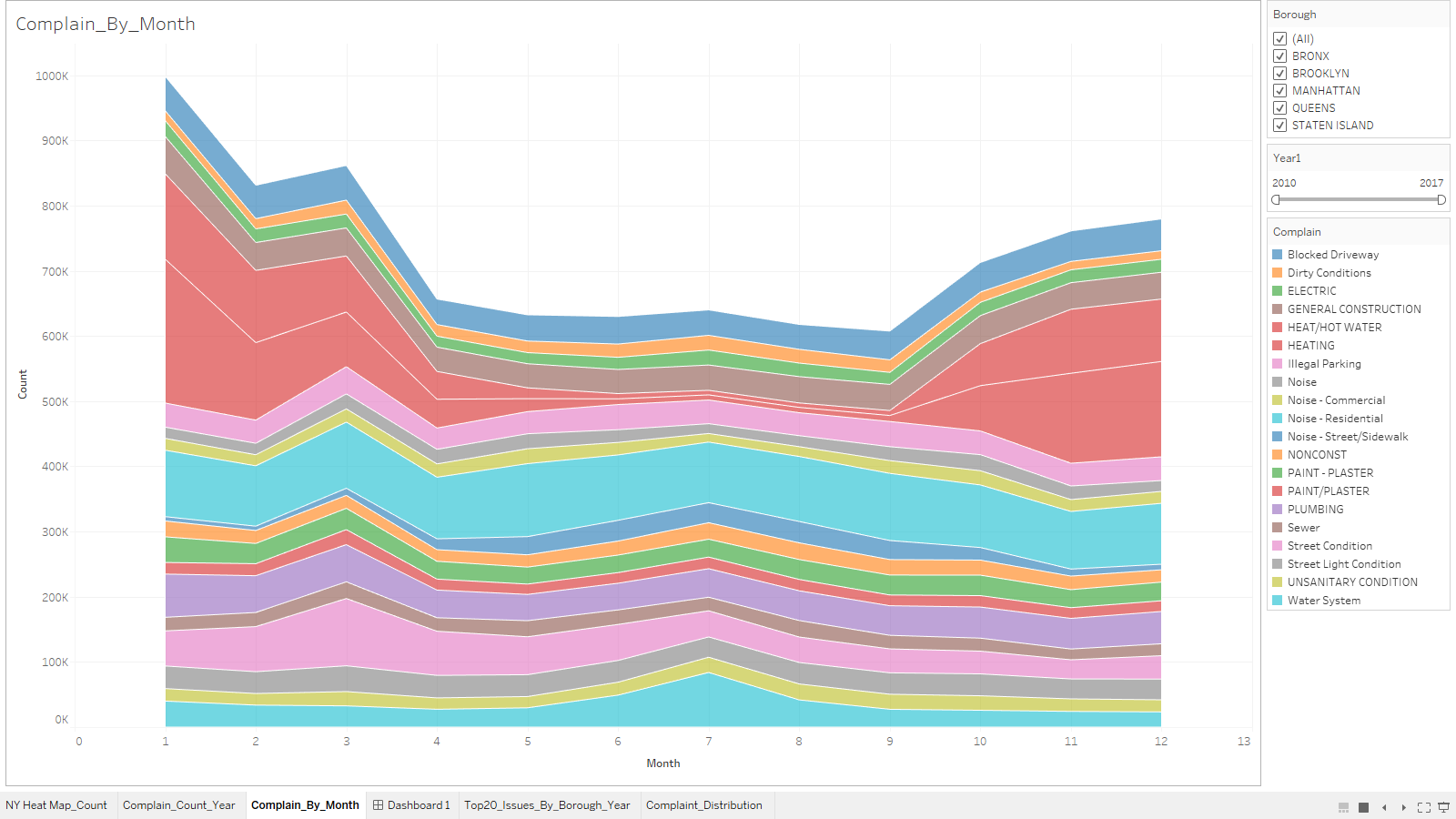


Link to MapReduce Program:

***Analysis3***: Variation of Complaint Types by Borough and Month

* The data is extracted from the MR program to get the unique count of Complaint Types by Borough and further partitioned by year and months.

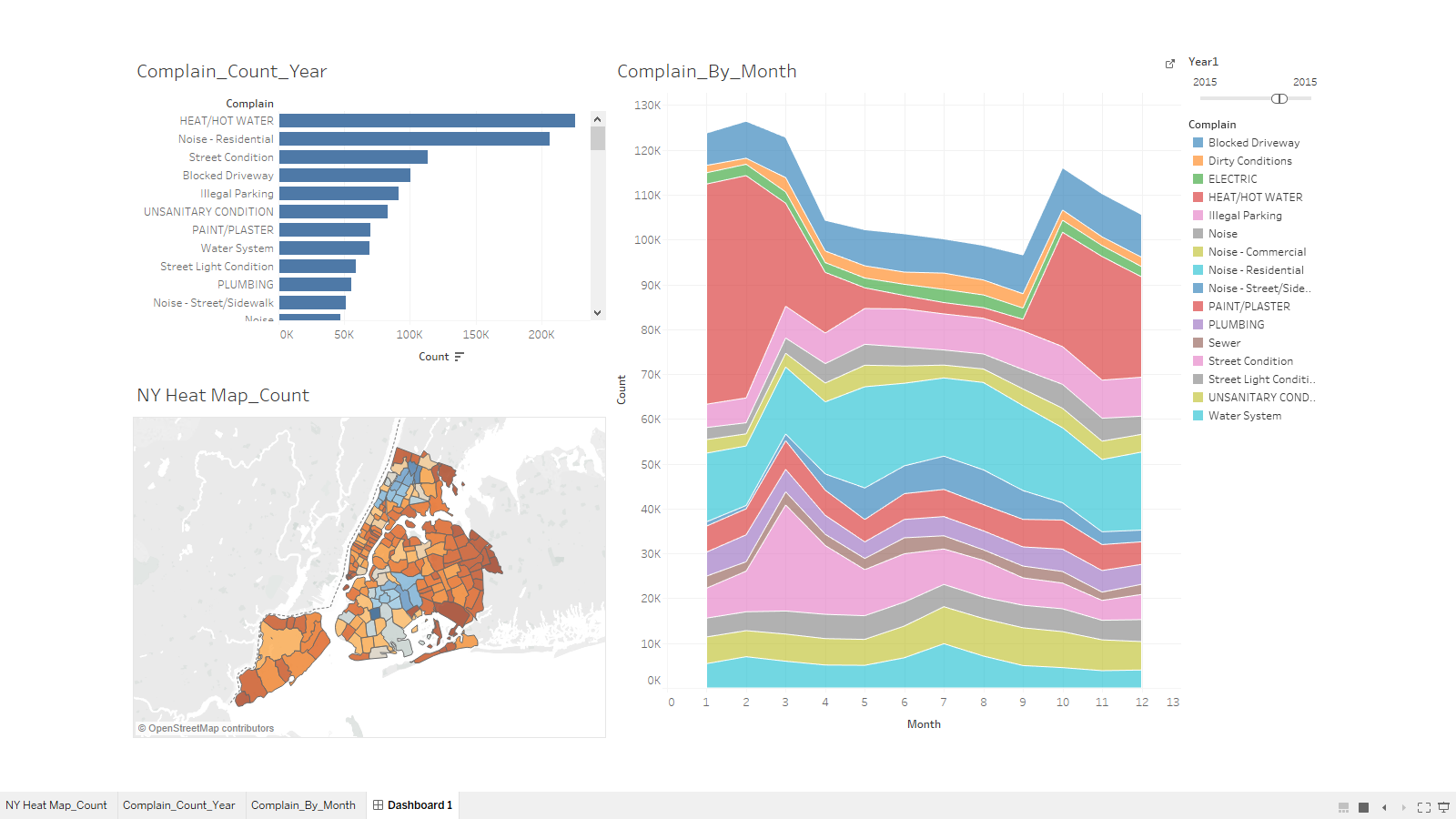
Below is the visual analysis of the same:



Link to MapReduce program:

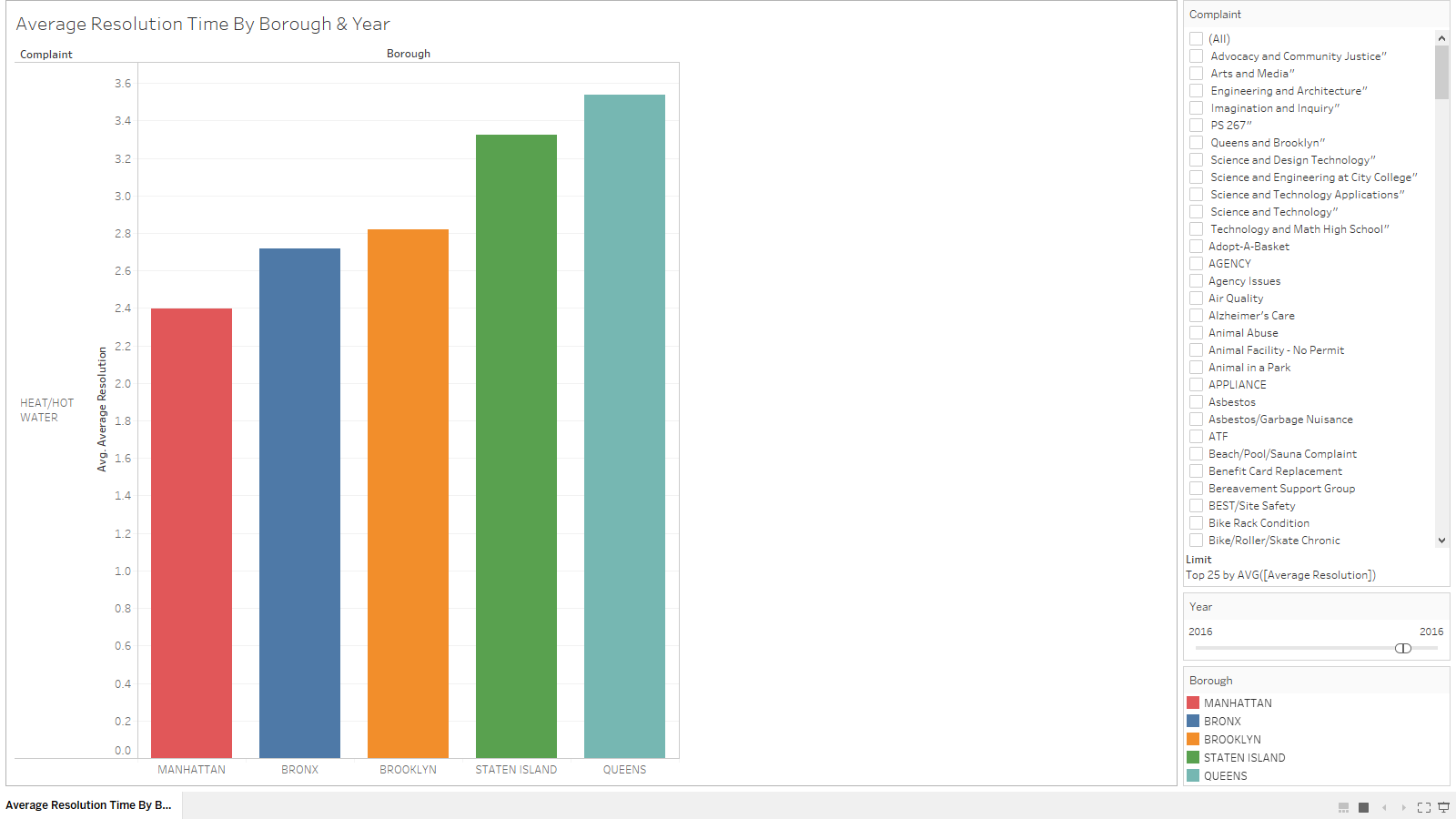
Below is the relative data visualization using the first 3 Analysis:

* Data can be drilled down by Geography, Complaint Type, Year and Month
* Pattern can be analyzed for frequently-reported incidents



***Analysis4***: From the above analysis, it was evident that ‘HEATING’ & ‘HEAT/HOT WATER’ was the most recurring reported issue. To dig deeper we find the average time taken by the NY authorities to resolve this issue.

The scope was further expanded to find the average resolution time for all the issues in NY by Borough and Year.



***Analysis5***: Inverted Index of Issues by Zip Code

* An Inverted index for each zip code in NY with the complaint types reported.



***MR Techniques Used***:

* Secondary Sorting - Partitioning Techniques
* PIG - Composite Key
* Inverted Index - Combiner