311 Complaints Study

Authors: *Nikhil Reddy, William Herrera, John Zachary Martinez*

Github Repository: *https://github.com/nikhil-reddy/Team2*

Abstract

For this project, the 311 Complaints data set provided by NYC Open Source was studied and analysed in three parts. For part 1 (the pre-experiment stage), we used big data analytics and aggregation techniques to clean and analyze null/invalid values in a large dataset of complaints pertaining to 311. We employed tools such as the **Hadoop NYU Cluster, Python and pyspark** to detect, count, and replace missing or invalid data. These tools were necessary for their parallelism in dealing with data of this enormity. We found the number of invalid values for each relevant attribute of the data, many of which exceeded half the number of data items. For part 2, a more in depth analysis of the was taken where we studied general statistics of the dataset. Some analysis include the number of all complaints by type and the number of complaints by borough. For the bonus portion of the project, we downloaded weather data and demographic data from the U.S. Census to see whether or not there were interesting relationships that could explain some of the findings we made in part 2.

Introduction

Managing numerous city operations is a daunting task and with the rise of data collection and reporting services; residents of NYC as well as employees of NYC agencies have access to powerful tools to answer and identify problem spots. The 311 service is a free public service that allows individuals to register complaints on city conditions. This city government data can reveal surprising insights about life in a community as well as how that community is being served. Our goal is to make sense of this data and find interesting results that could impact citywide initiatives such as turnaround time for a variety of complaints.

The data set is raw and includes many errors such as invalid entry types or null values. To address these issues, we performed data error statistics and clean up prior to calculating statistics on complaints and making observations of the set. Example errors that were tackled were invalid zip codes where the zip code was more than 5 digits or the was non numeric. Invalid agency names was also another error detected. A statistics script was run through the data to make a count where and what type of error was made and a clean script was run after to replace invalid entries and null values with N/A which we ignored for part 2 calculations. The 311 data is in two set where one is 2009 data and the other 2010-2017. These files were merged in the code so the results from both scripts is a reflection of both datasets.

For part 2, the cleaned complaints dataset was analyzed and several statistics were derived pertaining to the complaints. A general count of all the complaint types were calculated and plotted to see which complaint type was reported most by the people of New York. Complaints were also split into the distribution of complaints by borough for all years and the general distribution of complaints between 2009-2017. More statistics were created and from the observations of the complaint types, we narrowed the scope of our project by grouping together complaint types related to similar topics into categories to get a more general categorization of the data.

The bonus part of the project was conducted using weather data and U.S. Census demographic data. From our findings in part 2, we created several hypothesis to justify our observations. We extracted data from these sets that pertained to our hypothesis and plotted them against our complaint data to observe any similar trends. Pearson correlations were also conducted to measure the strength of the correlation between our cleaned complaints data set and the data extracted from our additional sets.

Data Set Description:

- The data set used are 311 complaints from 2009 and from 2010 to 2017
- The dataset owner: NYC Open Data
- Data Set information provided by: 311, DoITT
- There are 16.6 million rows in the raw dataset where each row represents the report of a 311 complaint.
- There are 53 columns of data where each column name is a distinct feature of that complaint. See **Table1** below for the name of all column attributes with descriptions:

Table1: Column Names and Descriptions

COL_NAME	DESCRIPTION
Unique Key	Unique identifier of a Service Request (SR) in the open data set
Created Date	Date SR was created
Closed Date	Date SR was closed by responding agency
Agency	Acronym of responding City Government Agency
Agency Name	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.
Complaint Type	Full Agency name of responding City Government Agency
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.

Location Type	Describes the type of location used in the address information
Incident Zip	Incident location zip code, provided by geo validation.
Incident Address	House number of incident address provided by submitter
Street Name	Street name of incident address provided by the submitted
Cross Street 1	First Cross street based on the geo validated incident location
Cross Street 2	Second Cross Street based on the geo validated incident location
Intersection Street 1	First intersecting street based on geo validated incident location
Intersection Street 2	Second intersecting street based on geo validated incident location
Address Type	Type of incident location information available.
City	City of the incident location provided by geovalidation
Landmark	If the incident location is identified as a Landmark the name of the landmark will display here
Facility Type	If available, this field describes the type of city facility associated to the SR
Status	Status of SR submitted
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).
Resolution Action Updated Date	Date when responding agency last updated the SR.
Community Board	Provided by geovalidation.

Borough	Provided by the submitter and confirmed by geovalidation.
X Coordinate (State Plane)	Geo validated, X coordinate of the incident location.
Y Coordinate (State Plane)	Geo validated, Y coordinate of the incident location.
Park Facility Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here
Park Borough	The borough of incident if it is a Parks Dept facility
School Name	If the incident location is a Dept of Education school, the name of the school will appear in this field. If the incident is a Parks Dept facility its name will appear here.
School Number	If the incident location is a Dept of Education school, the Number of the school will appear in this field. This field is also used for Parks Dept Facilities.
School Region	If the incident location is a Dept of Education School, the school region number will be appear in this field.
School Code	If the incident location is a Dept of Education School, the school code number will be appear in this field.
School Phone Number	If the facility = Dept for the Aging or Parks Dept, the phone number will appear here. (note - Dept of Education facilities do not display phone number)
School Address	Address of facility of incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School City	City of facilities incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School State	State of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dep
School Zip	Zip of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept

School Not Found	Y' in this field indicates the facility was not found
School or City Wide Complaint	If the incident is about a Dept of Education facility, this field will indicate if the complaint is about a particualr school or a citywide issue.
Vehicle Type	If the incident is a taxi, this field describes the type of TLC vehicle.
Taxi Company Borough	If the incident is identified as a taxi, this field will display the borough of the taxi company
Taxi Pick Up Location	If the incident is identified as a taxi, this field displays the taxi pick up location
Bridge Highway Name	If the incident is identified as a Bridge/Highway, the name will be displayed here.
Bridge Highway Direction	If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here
Road Ramp	If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.
Bridge Highway Segment	Additional information on the section of the Bridge/Highway were the incident took place
Garage Lot Name	Related to DOT Parking Meter SR, this field shows what garage lot the meter is located in
Ferry Direction	Used when the incident location is within a Ferry, this field indicates the direction of ferry
Ferry Terminal Name	Used when the incident location is Ferry, this field indicates the ferry terminal where the incident took place.
Latitiude	Geo based Lat of the incident location
Longitude	Geo based Long of the incident location
Location Type	Combination of the geo based lat & long of the incident location

Part 1: Data Quality Issues

Error Statistics (PreProcessing):

- Prior to cleaning the dataset and eliminating Null values, a python script titled ***stats_surprises.py*** was written to log data quality issues by reading our dataset titled ***new_311.csv***. We decided it was necessary to categorize the data in each cell of each column as valid, invalid, unspecified, and not applicable. This approach was done so that for each column, we could easily isolate the cells and values that are not the correct type for that column feature.

- The script utilized the pyspark and numpy packages and run on Hadoop cluster with the following call: **"pyspark --packages com.databricks:spark-csv_2.10:1.4.0 stats_surprises.py".** We utilized the databricks package because it is library for parsing and querying CSV data with Apache Spark, for Spark SQL and DataFrames. Dataframes were the primary data structure used to investigate and manipulate each feature in the dataset.

- The value "NULL" took on numerous forms in the dataset such as ("N/A","","Unspecified") and a singular or combination of terms applied to every column. To generate statistics for Null Values, a for loop ran through each column and returned the count of what we deemed "NULL" for that column. See Fig1 below for a sample output of one column and see Table2 below for a list of valid type for each column as well as possible error entries for that column.

Fig1.

- Since each column has distinct parameter ranges, there were other columns that had additional discrepancies in their entries. For example, the column ***Incident Zip*** is supposed to have entries of only integers 5 digits in length or 5 digits followed by 4 digits. If there is an entry of type string (ex."X") in the column, that would be categorized as invalid.

CREATED DATE	ERROR COUNT
1/16/2009 12:00	9397
3/3/2009 12:00	8655
2/5/2009 12:00	8180

1/17/2009 12:00	8096
1/15/2009 12:00	8051
1/21/2009 12:00	7834
1/14/2009 12:00	7794
1/20/2009 12:00	7620
1/26/2009 12:00	7455
1/12/2009 12:00	7274
1/22/2009 12:00	7214
1/27/2009 12:00	7016
2/4/2009 12:00	6958
2/6/2009 12:00	6905
2/12/2009 12:00	6875
3/4/2009 12:00	6832
1/29/2009 12:00	6832
1/6/2009 12:00	6808
1/8/2009 12:00	6796

2/24/2009 12:00	6659

Fig2. Other Errors in Incident ZIP

INCIDENT ZIP	ERROR COUNT
UNKNOWN	1
NA	12
113??	1
N/A	31
?	1
X	1
402901921	1
0	1
70545020	1
198884	1
103	1

- Borough entries had additional errors. There are 5 boroughs in NYC yet there were numerous instances where the borough name was unspecified. The statistics for these entries are shown below in **Fig3.**

Fig3. Unspecified Boroughs Statistics

BOROUGH	ERROR COUNT
Unspecified	224127

- Invalid Agency Acronyms was another issue. The largest NYC city government agency only has up to 5 characters and no punctuation characters within the name. Below are the statistics in Fig4. . Note that “3-1-1” is marked as an invalid entry, it is in fact a valid agency without the dashes so that will be dealt with in Part 2 of the project.

Fig4. Invalid Agency Statistics

AGENCY	ERROR_COUNT
DESIGNCOM	3
3-1-1	25044
NYCERS	5
IA	8
DV	3
NYCPPF	9
NYCOOA	12
NYCSERVICE	4
WF1	3

- Finding invalid Complaint Descriptors. None were found.

- Finding invalid Community Board Entries. None were found
- Data Set Features with Low Frequencies
- Data Set Features with High Frequencies

Table2: Column Types and Errors

COL_NAME	ERRORS	VALID TYPE
Unique Key	NONE	int or string
Created Date	blank,N/A, unspecified, month > 12,day is greater than number in that month,	date/time
Closed Date	month > 12,day is greater than number in that month,	date/time
Agency	blank, N/A, unspecified	string
Agency Name	blank, N/A, unspecified	string
Complaint Type	blank, N/A, unspecified	string
Descriptor	blank, N/A, unspecified	string
Location Type	blank, N/A, unspecified	sting
Incident Zip	blank, N/A, unspecified	int 5 digits or 5 followed by 4 digits in length
Incident Address	blank, N/A, unspecified	string
Street Name	blank, N/A, unspecified	string

Cross Street 1	blank, N/A, unspecified	string
Cross Street 2	blank, N/A, unspecified	string
Intersection Street 1	blank, N/A, unspecified	string
Intersection Street 2	blank, N/A, unspecified	string
Address Type	blank, N/A, unspecified	string
City	blank, N/A, unspecified	string
Landmark	blank, N/A, unspecified	string
Facility Type	blank, N/A, unspecified	string
Status	blank, N/A, unspecified	string
Due Date	blank, N/A, unspecified	date/type
Resolution Action Updated Date	blank, N/A, unspecified	date/type
Community Board	blank, N/A, unspecified	string
Borough	blank, N/A, unspecified	string
X Coordinate (State Plane)	blank, N/A, unspecified	int,float
Y Coordinate (State Plane)	blank, N/A, unspecified	int, float
Park Facility Name	blank, N/A, unspecified	string
Park Borough	blank, N/A, whether or not the string is actually a borough	string

School Name	blank, N/A, unspecified	string
School Number	blank, N/A, unspecified	int
School Region	blank, N/A, unspecified	string
School Code	blank, N/A, unspecified	string
School Phone Number	blank, N/A, unspecified	10 digits integer
School Address	blank, N/A, unspecified	string
School City	blank, N/A, unspecified	string
School State	blank, N/A, unspecified	2 letter acronym string
School Zip	blank, N/A, unspecified	int 5 digits in length
School Not Found	blank, N/A, unspecified	"Y" or "N" string
School or City Wide Complaint	blank, N/A, unspecified	"School" or "Citywide Complaint" string
Vehicle Type	blank, N/A, unspecified	string
Taxi Company Borough	blank, N/A, unspecified	string
Taxi Pick Up Location	blank, N/A, unspecified	string
Bridge Highway Name	blank, N/A, unspecified	string
Bridge Highway Direction	blank, N/A, unspecified	string

Road Ramp	blank, N/A, unspecified	sting
Bridge Highway Segment	blank, N/A, unspecified	string
Garage Lot Name	blank, N/A, unspecified	string
Ferry Direction	blank, N/A, unspecified	string
Ferry Terminal Name	blank, N/A, unspecified	string
Latitiude	blank, N/A, unspecified	float
Longitude	blank, N/A, unspecified	float
Location Type	blank, N/A, unspecified	string

Part 1: Data Cleaning and Error Count

Data Cleaning :

- To clean the dataset and eliminating Null values, a python script titled ***clean.py*** was written to remove irrelevant data, replace values, and write the clean dataset to a new csv file titled ***cleaned_311.csv***.

- The script utilized the pyspark and numpy packages and it was run on Hadoop cluster with the following call: **"pyspark --packages com.databricks:spark-csv_2.10:1.4.0 clean.py".**

- Prior to replacing the values in our dataset, several columns were deleted because the data in these columns was either too widely varied, completely missing, or not necessary in the study we want to conduct for part 2 which is to analyse the turnaround time of different kinds of complaints by borough, day, season etc... and more. Table3 demonstrates which columns were eliminated from our dataset.

Table3: Deleted Columns

COL_NAME
Facility Type
School Name
School Number
School Region
School Code
School Phone Number
School Address
School City
School State
School Zip
School Not Found
School or City Wide Complaint
Bridge Highway Name
Bridge Highway Direction

Road Ramp
Bridge Highway Segment
Garage Lot Name
Ferry Direction
Ferry Terminal Name
Latitiude
Longitude

- Following the deletion of the rows, the program continued to replace all invalid zip codes with "N/A" as well as replace all "NULL" values with "N/A" in the dataset making error labeling consistent.

- The program stops here and does not delete the rows that contain "N/A" values because at this point in the project process, it has not been decided on what analysis tests we want to perform and it may be necessary to use this data so we prefer to do additional cleaning in part 2. A possible use of having these errors may be running an error analysis to see at what time of day are errors most frequently made when making 311 complaints or replying to them.

- Since the program works with data frames, it outputs a directory titled ***cleaned_311.csv*** that contains data frames which are chunks of the entire set. In order to output one csv file, the command **"hadoop hfs -getmerge cleaned_311.csv cleaned_311.csv"**.

Results:

- The raw original file ***311.csv*** ~8GB, and the ***cleaned_311.csv*** has consistent "N/A" values for non valid entries throughout the entire table and is ~6GB. Below are samples of the ***311.csv*** and the ***cleaned_311.csv***
- ***Errors by location type, address type, and number of errors in all columns were reported as well.***

New_311.csv

Location Type	Incident Zip	Incident Address
RESIDENTIAL BUILDING	11225	55 WINTHROP STREET
Restaurant/Bar/Deli/Bakery	11102	29-35 NEWTOWN AVENUE
	11220	
	11201	
	11235	
RESIDENTIAL BUILDING	11417	103-60 104 STREET
RESIDENTIAL BUILDING	11225	1211 NOSTRAND AVENUE
RESIDENTIAL BUILDING	11237	1409 HANCOCK STREET
RESIDENTIAL BUILDING	11377	31-38 68 STREET
RESIDENTIAL BUILDING	10028	227 EAST 82 STREET
RESIDENTIAL BUILDING	10467	3204 HOLLAND AVENUE

Cleaned_311.csv

Location Type	*Incident Zip*	*Incident Address*
Residential Building/House	*11225*	*421 CROWN STREET*
Street/Sidewalk	*11234*	*2057 EAST 38 STREET*
Tenant Address	*11236*	*N/A*
Street	*11358*	*41-09 161 STREET*
RESIDENTIAL BUILDING	*10455*	*665 CAULDWELL AVENUE*
Tenant Address	*11212*	*N/A*
N/A	*N/A*	*N/A*
Tenant Address	*11422*	*N/A*
N/A	*N/A*	*N/A*
N/A	*N/A*	*N/A*
Street/Sidewalk	*10039*	*235 WEST 146 STREET*
Street/Sidewalk	*11210*	*1352 FLATBUSH AVENUE*

Number of errors counted in each column:

Column_Name	Null Count
Unique Key	0
Created Date	0
Closed Date	10514
Agency	0
Agency Name	0
Complaint Type	0
Descriptor	1095
Location Type	152886
Incident Zip	30357
Incident Address	95782
Street Name	95824
Cross Street 1	96433
Cross Street 2	98455
Intersection Street 1	399431
Intersection Street 2	399425
Address Type	10915

City	29929
Status	2
Due Date	404333
Resolution Action Updated Date	1836
Community Board	257810
Borough	224127
X Coordinate (State Plane)	31513
Y Coordinate (State Plane)	31513
Park Facility Name	486803
Park Borough	224127
School or Citywide Complaint	487969
Vehicle Type	489212
Taxi Company Borough	489326
Taxi Pick Up Location	484981

Location	31513

Error count by location type:

LOCATION TYPE	ERROR COUNT
Homeless Shelter	1
Government Buildi...	1
Steam Room	1
Theater	1
Doctor's Office	1
Street Fair Vendor	1
Sports Arena	1
Store	1
School - College/...	2
Soup Kitchen	2
Other	2
School - K-12 Public	2
Nursing Home	2
Parking Lot	2
Summer Camp	2

Cafeteria - Priva...	3
Public Garden	4
Address Unknown	4
Hospital	4
Spa Pool	4

Error count by address type:

ADDRESS_TYPE	ERROR_COUNT
PLACENAME	398
	10915
BLOCKFACE	13801
INTERSECTION	89346
ADDRESS	375147

Error count by complaint type:

COMPLAINT TYPE	COUNT
Ferry Permit	1
Squeegee	1

Adopt-A-Basket	1
Harboring Bees/Wasps	1
Tunnel Condition	1
Summer Camp	2
Stalled Sites	2
Transportation Pr...	3
Poison Ivy	3
Trans Fat	3
Radioactive Material	4
X-Ray Machine/Equ...	4
Highway Sign - Da...	4
Lifeguard	4
Unsanitary Animal...	4
Illegal Fireworks	5
Special Enforcement	5
Highway Sign - Mi...	6
Legal Services Pr...	7

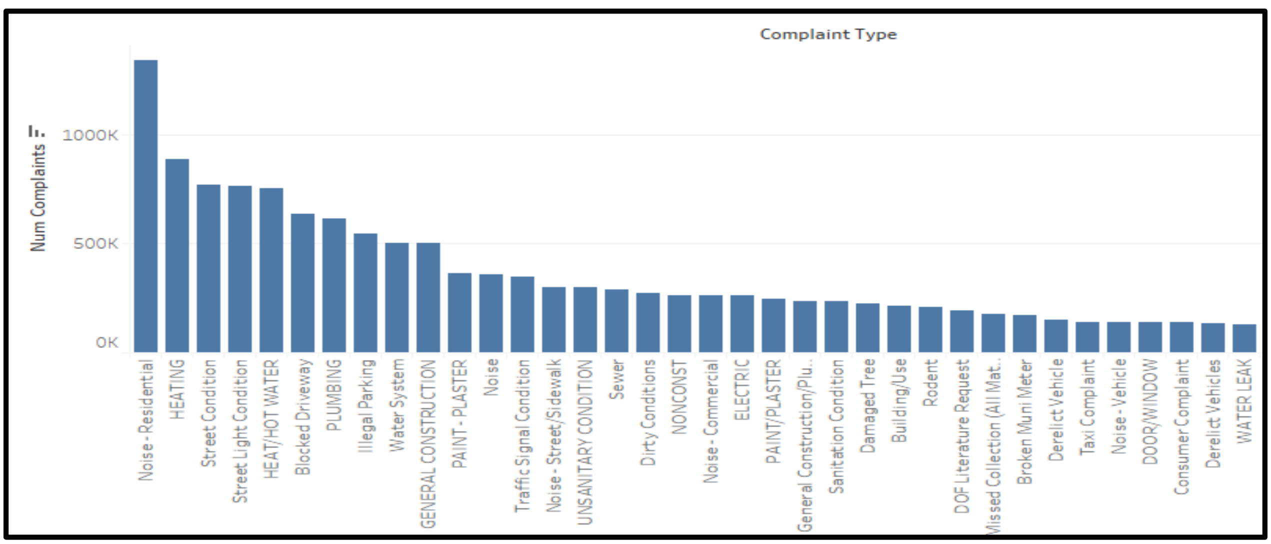
Part 2: Data Analysis

- For this part of our project, we wanted a better understanding of our dataset so we performed several analysis on our dataset which we believed would help gain insight into interesting features for future study.

Additional Data Cleaning :
Before reporting the figures of the data analysis, we noticed other errors in the data set that were not identified in part 1 of our project. We had not realized that there were errors in the closed date such as years ranging outside 2009-2017. This produced inconsistent data so we edited our clean.py script to account for this inconsistency. What we decided to do was if the read in the closed date was outside of this range, it make the value in the cell N/A. Another error detected was that the closed date was less that the created date meaning a complaint report was resolved before it was ever opened. This did not make sense so what we did to identify these dates was to calculate the turnaround time which is the time difference between the closed date and corresponding created date. If the turnaround time was negative, the closed date value would be transformed to N/A by running the clean.py file. These N/A were accounted for in the statistic table "Number of errors in each column" listed in part 1 above

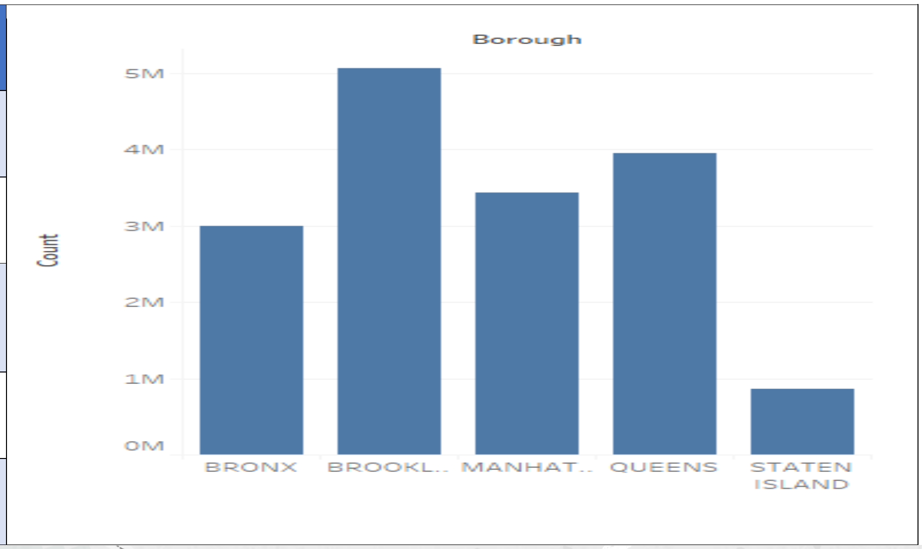
Analytics:

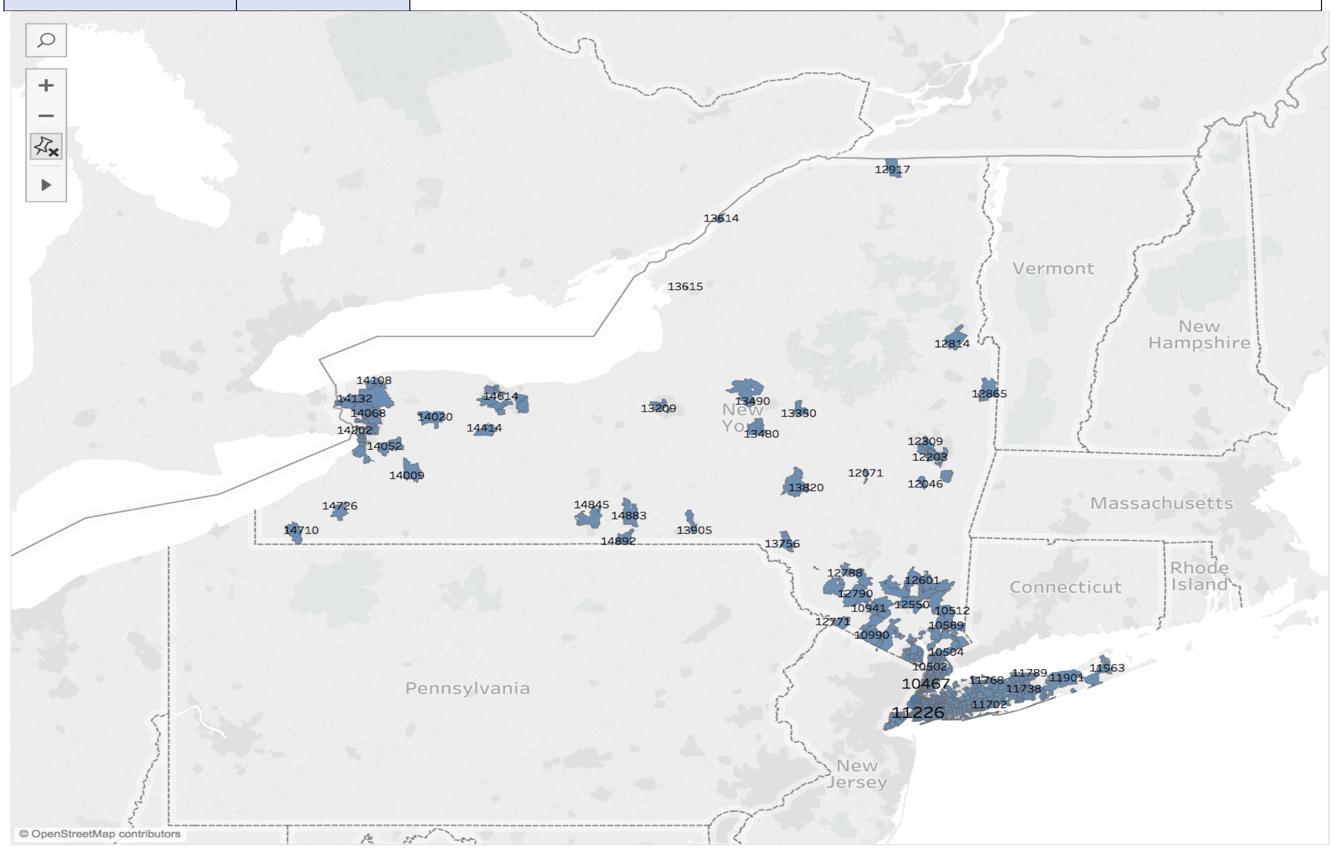
- In order to isolate a topic for a project, we decided to count the number of Complaint Types from our cleaned dataset in order of number of complaints. (below is a subset of the graph for inorder to reserve space but there were in total 306 different complaint types). From the plot below, we noticed that noise complaints were the predominant complaint type aggregating all years and heat was second. See the table in the back for more details about all of complaint types with their counts.



- In addition to discovering the number of different types of complaints, we were interested to see the distribution of complaints per borough to see whether or not there are parts of NYC that report noticeably higher rates of complaints. The table with the exact number of complaints is below along with a visual representation. Notice how Brooklyn has roughly more than 1 million more complaints than the next highest Borough between the years 2009-2017

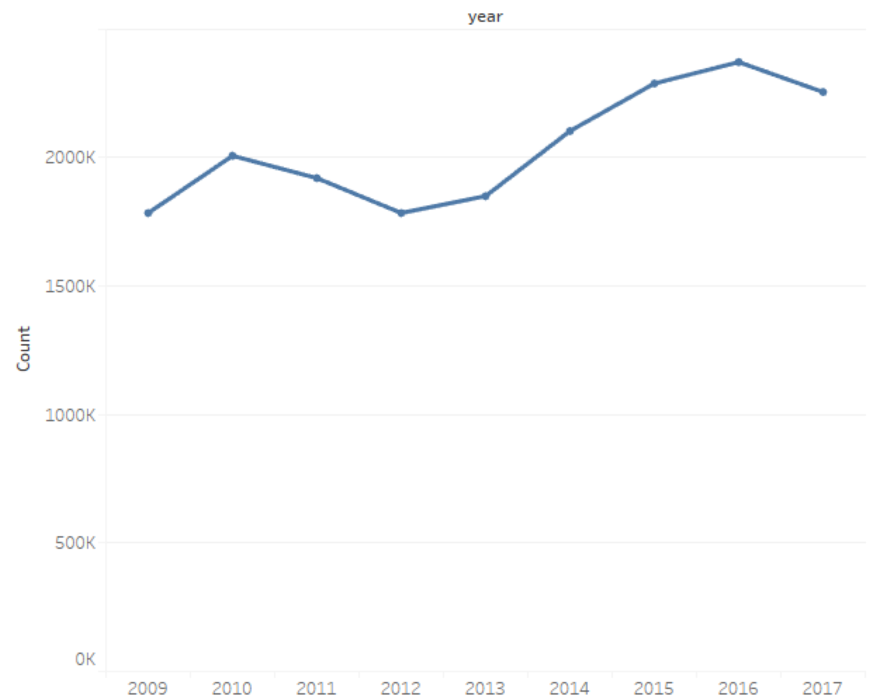
Borough	count
MANHATTAN	3434819
QUEENS	3949305
STATEN ISLAND	867557
BROOKLYN	5062681
BRONX	2998353





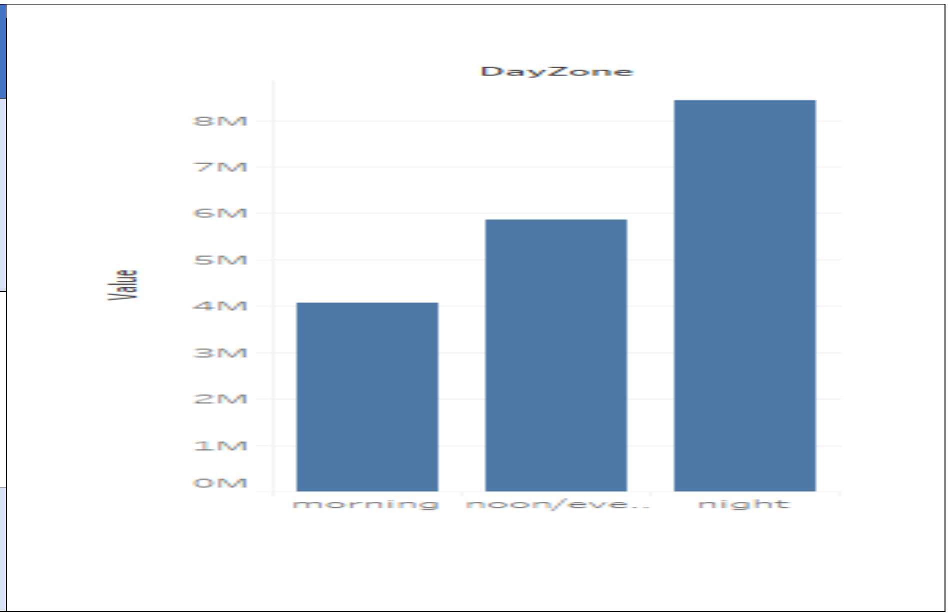
- The number of complaints created per year was also examined between the years 2009-2017. It is worth noting that the year 2017 is not over at the time of this report. The data set pertaining to 2017 only goes up to 11/30/17 so the month of December is not accounted for and is a possible explanation of the dip in 2017. There is a sharp dip in complaints between years 2010-2012 and a steady increase in complaints following except for 2017 which was explained above.Below is a table with the exact values and a line graph to visualize the trend.

year	count
2009	1783133
2010	2005760
2011	1918896
2012	1783212
2013	1849019
2014	2102226
2015	2286951
2016	2370339
2017	2253765



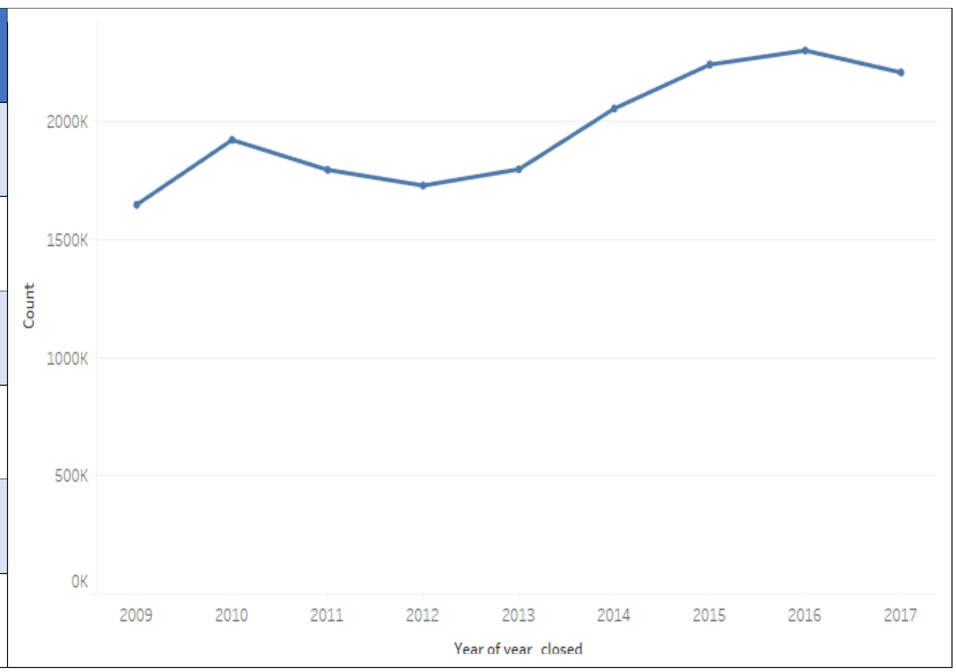
- The time of day when complaints were created was also studied. We grouped the complaints into morning, noon/evening, and night. The time ranges used were 5am-11am for the morning,12pm-7pm noon/evening, and 8pm-4am for night. By aggregating data for all years, it appears most complaints are made at night time.

DayZone	count
morning	4069336
noon/evening	5851547
night	8432418



- The number of complaints closed per year was also examined between the years 2009-2017. As previously stated, 2017 is not finished so the closed dates for recently created complaints was blank and was treated as N/A from part one so it was not included in these calculations possible. It is worth noting that for every year, it appears that there were more valid created dates than valid close dates possibly indicating not all created complaints were ever resolved. Below is a table with the exact values and above is a line graph to visualize the trend.

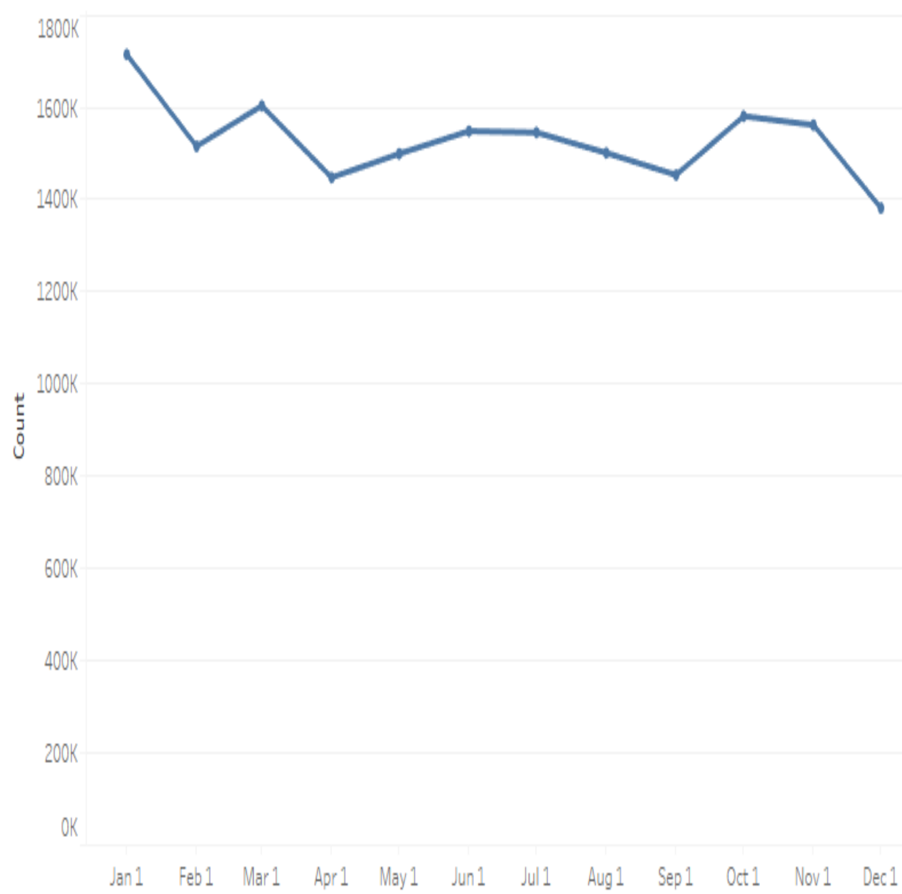
year_closed	count
2017	2210085
2016	2303115
2015	2243516
2014	2057229
2013	1799583
2012	1730635



2011	1797240
2010	1924106
2009	1649571

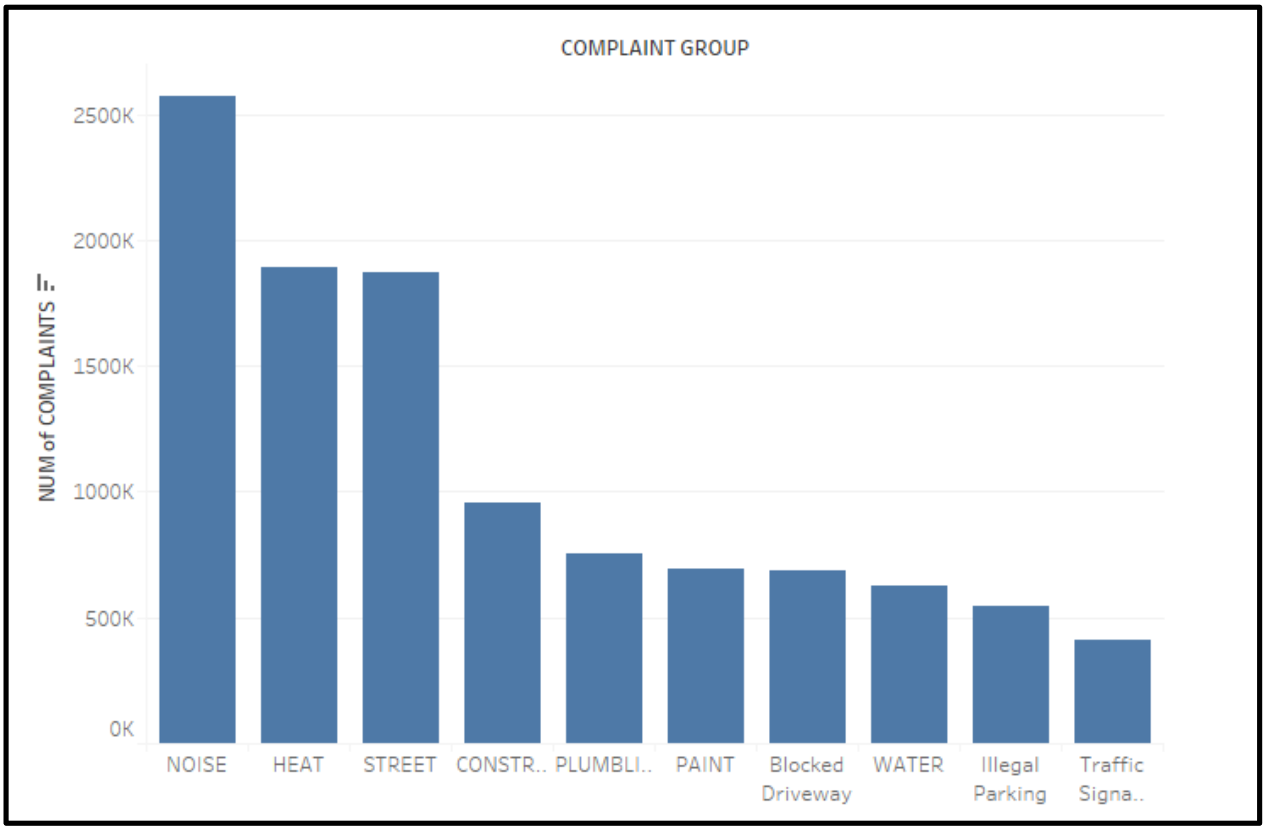
- The months in which complaints were made was also examined. We wanted to see if there was a particular time of year when more complaints are made. The highest reported month for complaints was January and it appears that the months for Fall (September-December) and Winter (December-March) have the highest number of complaints. As we can see, December has a slight decline for complaints but this can be explained by the fact that December 2017 complaints data was not available at the time this report was written

month_created	count
1	1714672
2	1515579
3	1603257
4	1447820
5	1499365
6	1548459
7	1545425
8	1501180
9	1453366
10	1580843



11	1561651	
12	1381684	

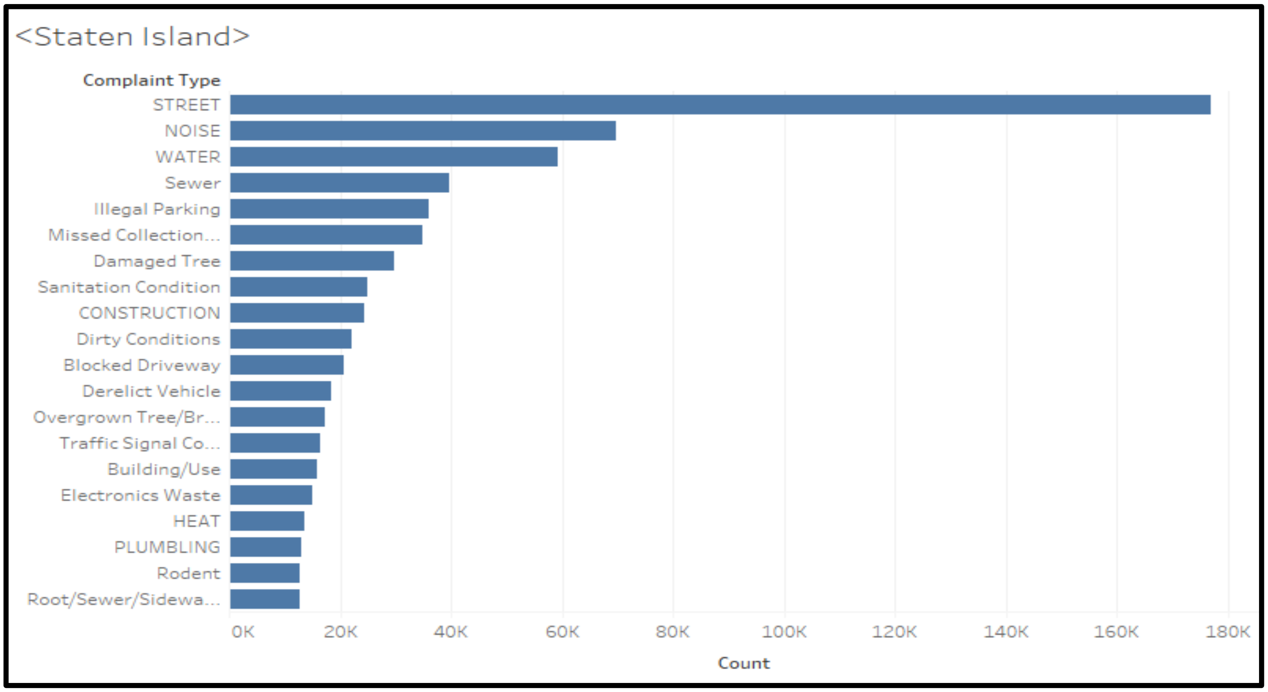
- From the above Complaints graph, we noticed numerous complaints had the same word with in the complaint such as "Noise" and "Noise Residential" etc.... To generalize the complaints dataset, we decided to group some of the complaints based on a specific set of keywords. Below are the 10 categories we defined along with the number of complaints related to that group. Noise, Heat, and Street were the three highest reported complaint types throughout the entire city for all years.

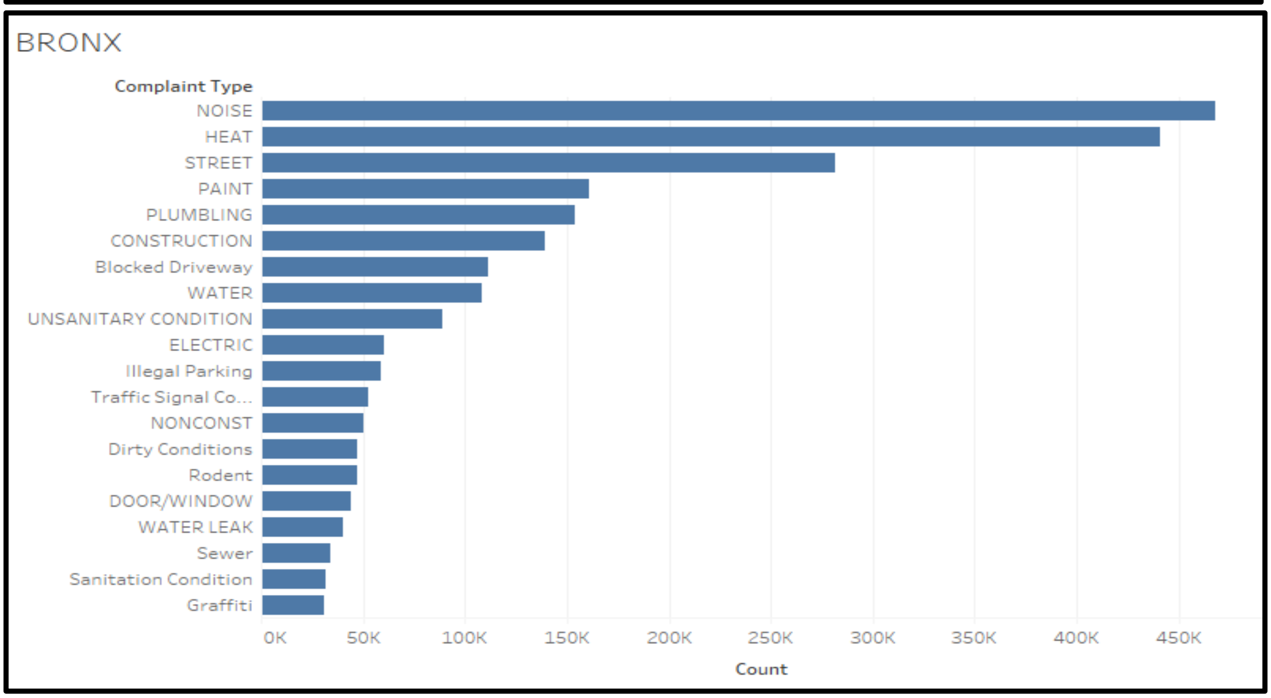


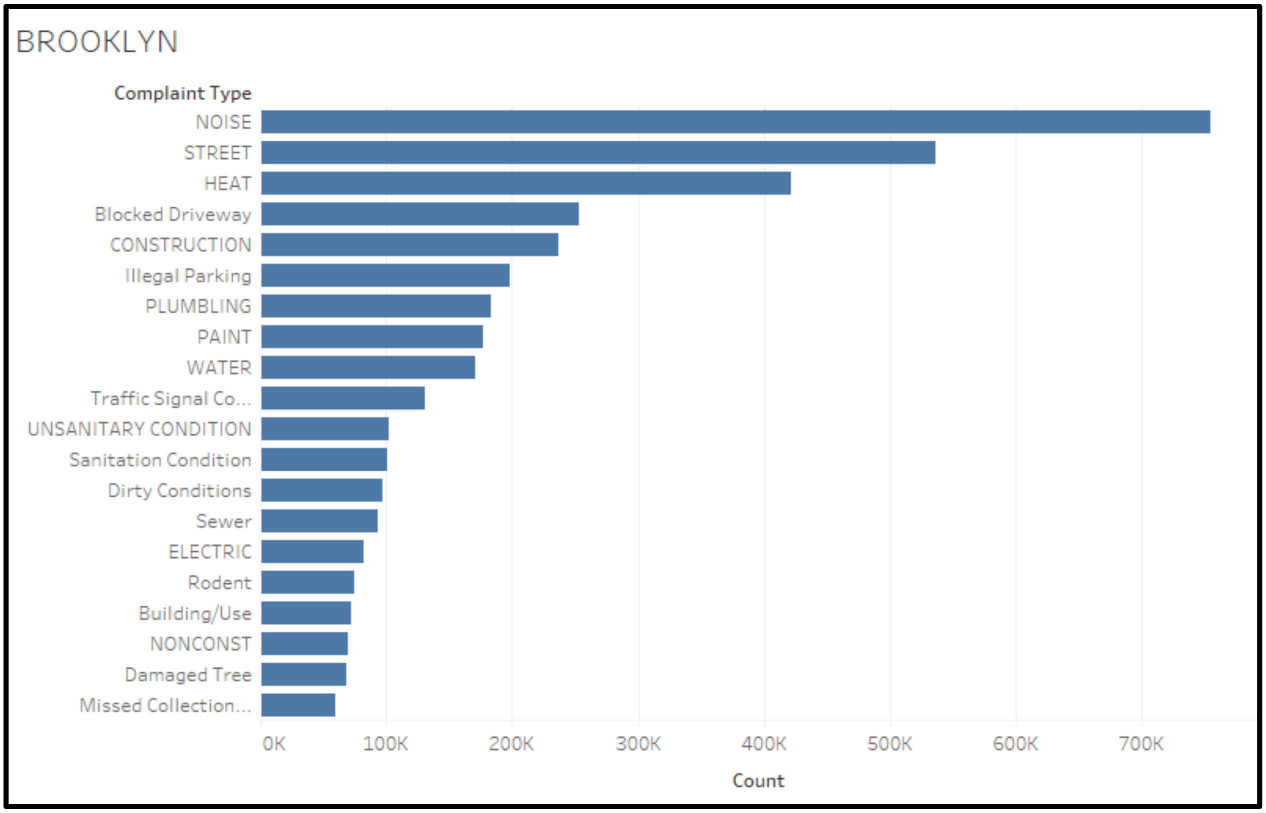
- The bar graphs below show the analytics produced by **analytics2.py.** This script displayed the count of the top 20 complaint types, grouped in the same way as above, for each of the 5 boroughs.
- From what we can observe there are slight variations in the order of the type of complaint for the borough as compared to the entire city. For Staten Island, street complaints were the significant complaint made while the rest of the city is noise. This

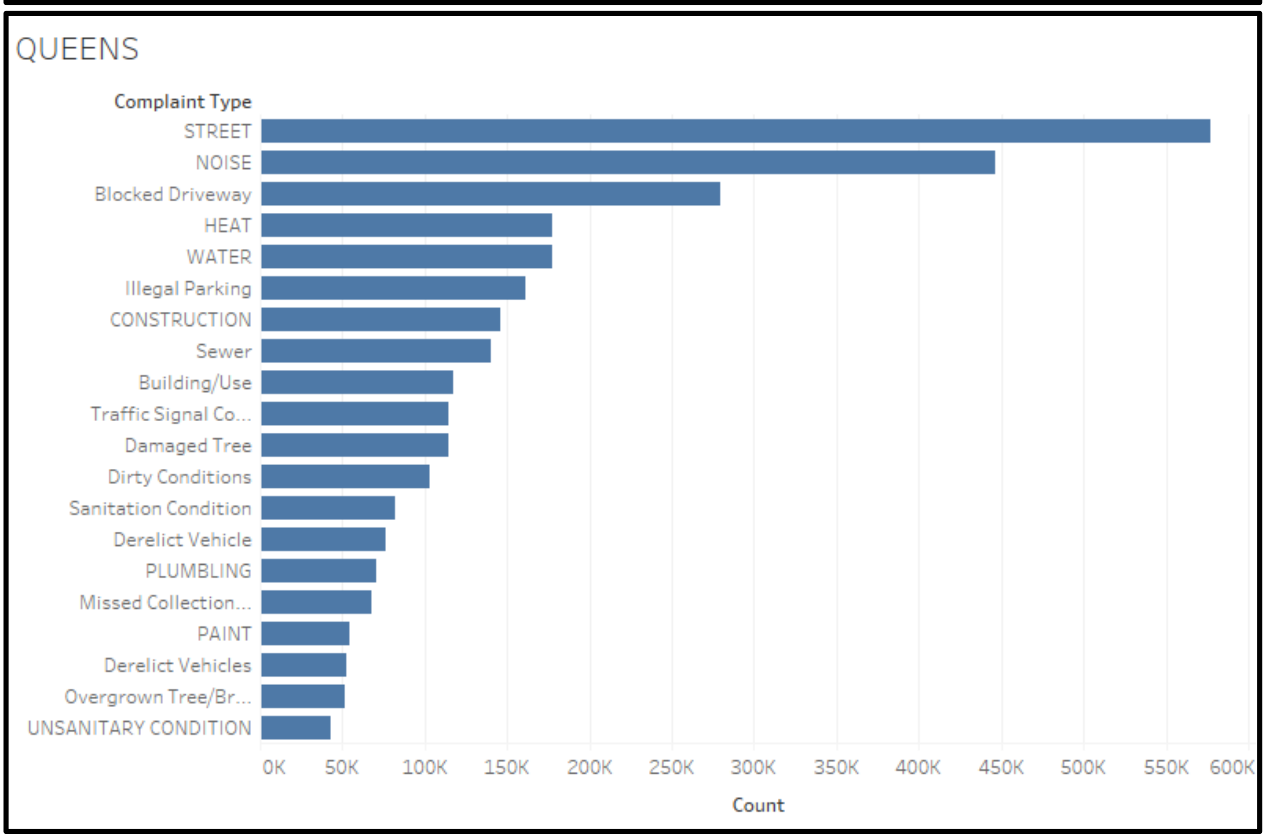
could be an indication that street conditions are worse in Staten Island possibly because cars are the predominant source of transport.

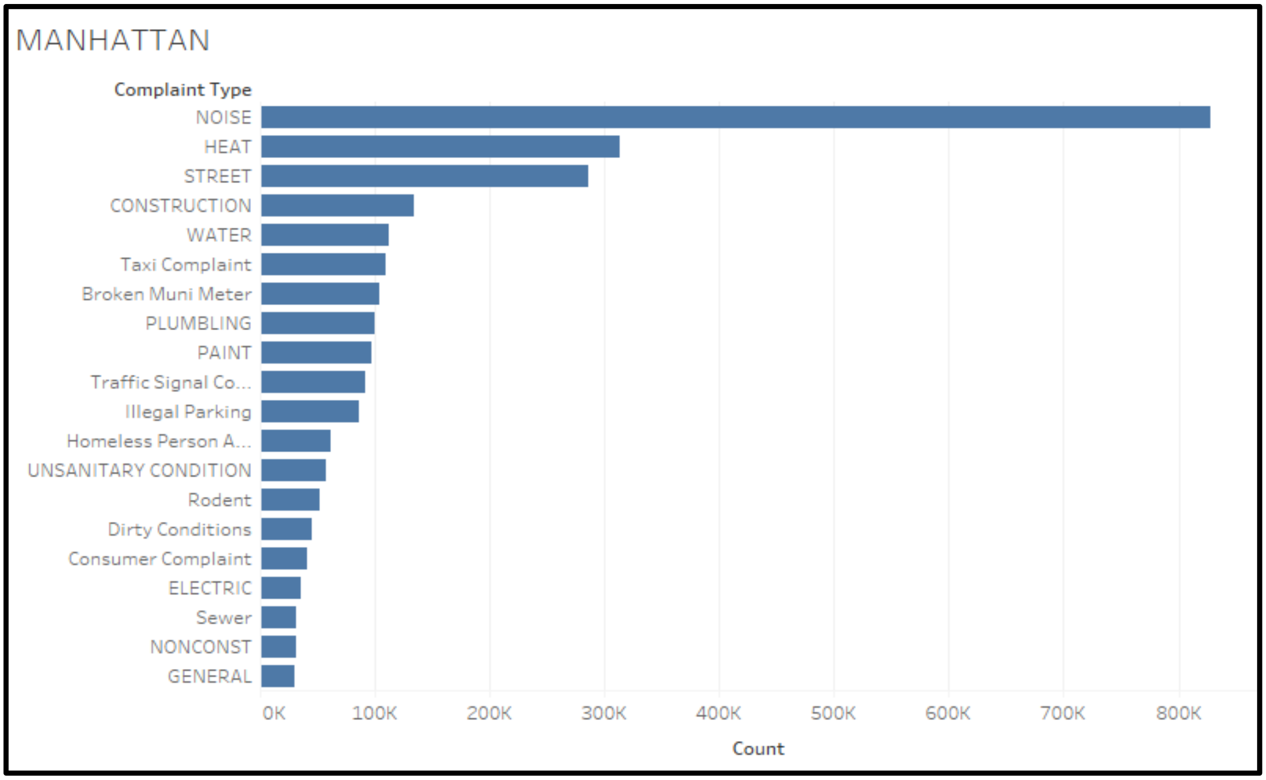
- For the Bronx, the top 3 complaint type match the order of the entire city.
- For Brooklyn, the top 3 complaint type match that of the entire city but heat is 3rd in rank while street are 2nd.
- For Queens, the top 3 complaint did not match. Street complaints was ranked 1st, noise complaints 2nd, and heat was ranked 4th where 3rd was blocked driveways.
- For Manhattan, the top 3 complaint type match the order of the entire city.






BROOKLYN
Complaint Type
NOISE
STREET
HEAT
Blocked Driveway
CONSTRUCTION
Illegal Parking
PLUMBING
PAINT
WATER
Traffic Signal Co...
UNSANITARY CONDITION
Sanitation Condition
Dirty Conditions
Sewer
ELECTRIC
Rodent
Building/Use
NONCONST
Damaged Tree
Missed Collection...
0K
100K
200K
300K
400K
500K
600K
700K
Count


QUEENS
Complaint Type
STREET
NOISE
Blocked Driveway
HEAT
WATER
Illegal Parking
CONSTRUCTION
Sewer
Building/Use
Traffic Signal Co...
Damaged Tree
Dirty Conditions
Sanitation Condition
Derelict Vehicle
PLUMBING
Missed Collection...
PAINT
Derelict Vehicles
Overgrown Tree/Br...
UNSANITARY CONDITION
0K
50K
100K
150K
200K
250K
300K
350K
400K
450K
500K
550K
600K
Count



Bonus Part: Data Exploration

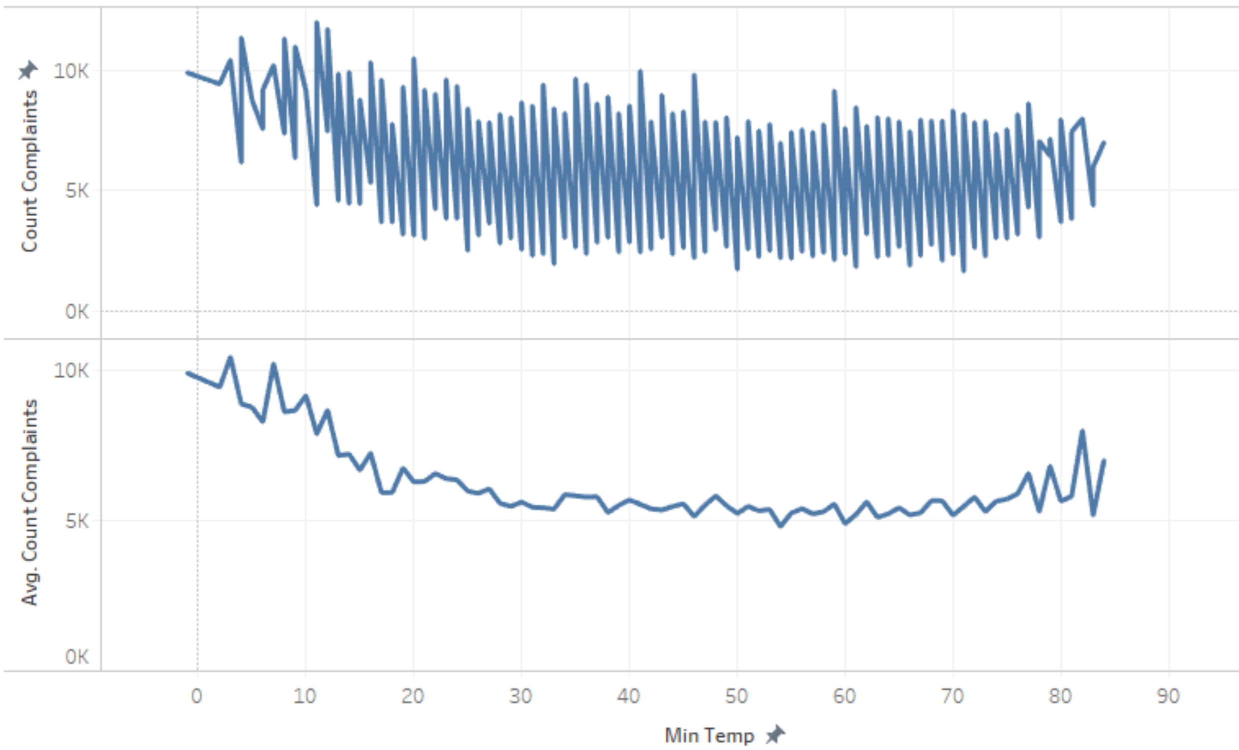
From Part 2, we noticed several interesting trends in the complaint data set. There were not sufficient attributes related to our queries so we believed it was necessary to introduce new data set that may corroborate our findings. Due to time constraints, 3 hypothesis were made and tested by analyzing certain attributes of the 311 complaint set foreign data sets and performing pearson correlations. The hypothesis made as well as the datasets and findings are listed below.

Hypothesis 1

There is a correlation between temperature and number of complaints. Specifically, that there are significantly more complaints, especially heat-related ones, in times of extreme temperatures.

- We found that a vast portion of the complaint types were related to HEATING.
- Since the 311 Complaints dataset does not provide weather related information, we obtained the Central Park weather data comes from the National Climatic Data Center(NOAA). These two datasets were merged using PySpark via a join on their date fields.
- The Pearson's correlation factor was found to be **-0.172044680055**, which indicates a low inverse relationship between temperature and number of complaints filed for heating issues. But from the below graph you may notice that this is because of the rise in complaints when temperature increases too. Hence we can conclude that the number of complaints increase

when the temperature falls below a threshold(~17 F) and increases beyond a max threshold (~82 F).

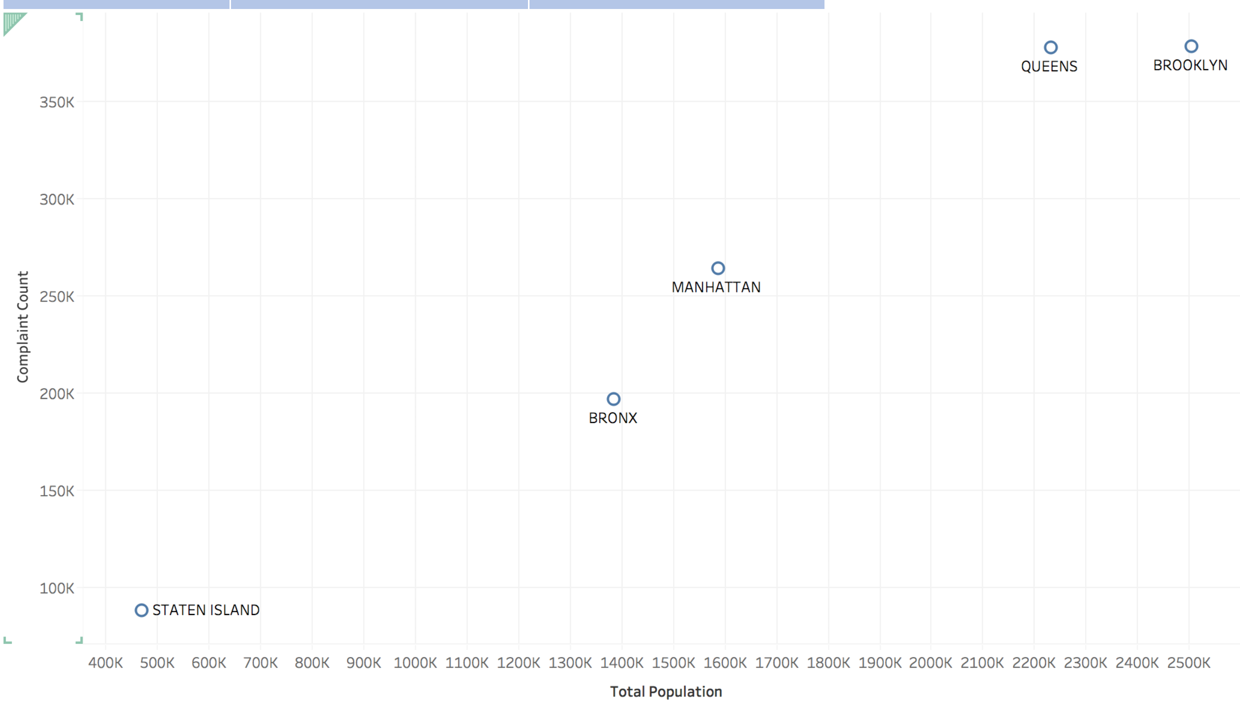


Hypothesis 2

The number of complaints in a borough increases linearly with its population. In other words, there is a semi-constant slope between population and complaint count.

- We found that a vast portion of the complaint types were registered to Brooklyn and also noticed that the population of Brooklyn was the highest amongst the five boroughs.
- Since the 311 Complaints dataset does not provide population and demographic related information, we obtained the Decennial Census from the NYC Department of City Planning. Since it contained only the population at 2010, we grouped and analysed complaints from 2010 only. These two datasets were merged using PySpark via a join on their Borough fields.
- The Pearson's correlation factor was found to be **0.985771364986**, which indicates a strong direct relationship between total population and number of complaints filed which can also be inferred from the graph. Also, note that the dataset is small in terms of number of data points since we are only analyzing the 5 boroughs as a whole

Borough	Complaint Count	Total Population
BROOKLYN	378094	2504700
QUEENS	377578	2230722
MANHATTAN	263808	1585873
BRONX	196470	1385108
STATEN ISLAND	88271	468730



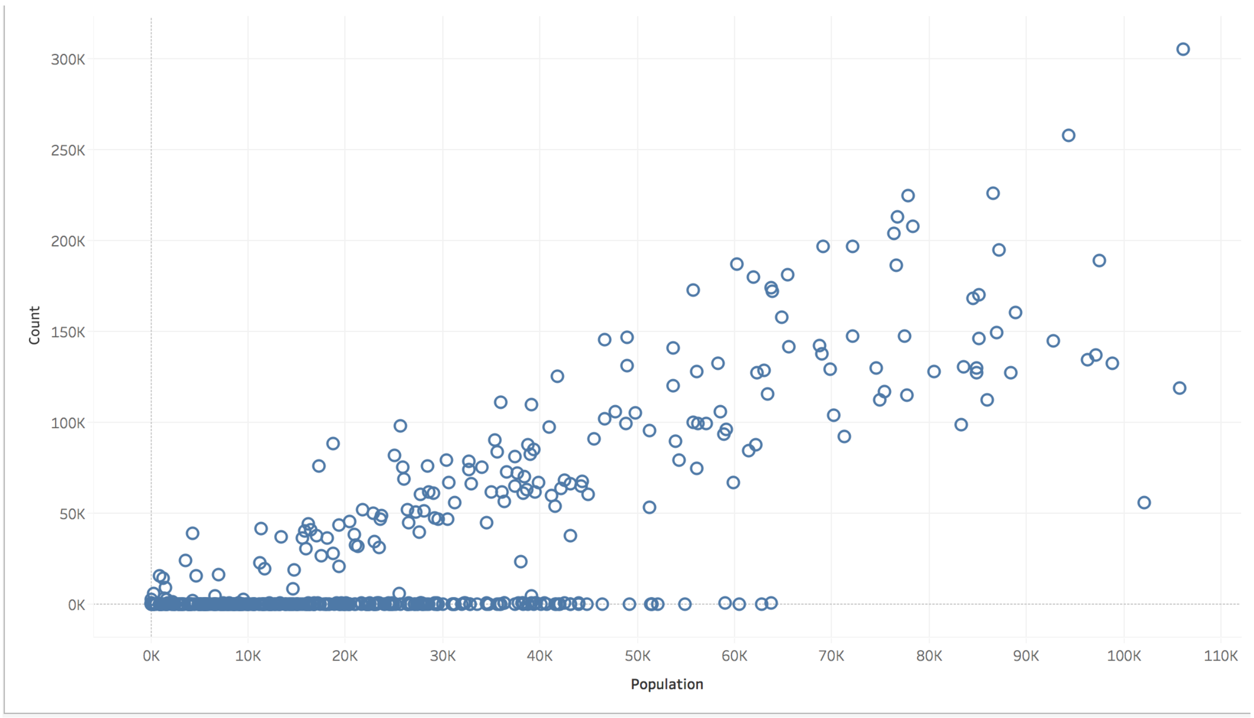
Hypothesis 3

There is a correlation between the total population of an area and the total number of complaints registered in it.

- From the previous hypothesis, we found that the correlation was very strong because of fewer data points. Hence we wanted to generalise it further by using population distribution by zip codes.
- Since the 311 Complaints dataset does not provide population and demographic related information, we obtained the Census population distribution by zip code for NYC. Since it

contained only the population at 2010, we grouped and analysed complaints from 2010 only. These two datasets were merged using PySpark via a join on their Borough fields.

- The Pearson's correlation factor was found to be **0.796369076**, which indicates a strong direct relationship between total population and number of complaints filed which can also be inferred from the graph.



Experimental Techniques and Methods

Pearson Product Moment Correlation or PPMC (Pearson's Correlation) was used to find the degree to which two columns of data were related to each other. A strong relation is one where the factor is either between -1 and -0.5, or between 0.5 and 1. A negative association signifies an inverse relationship between data.
High correlation: .5 to 1.0 or -0.5 to 1.0.
Medium correlation: .3 to .5 or -0.3 to .5.
Low correlation: .1 to .3 or -0.1 to -0.3.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. This correlation requires that the data which is being used be normally distributed.

After running the scripts and generating a csv file as output, plots were created using **Tableau**. It was used to search for interesting relationships and visualize them before codes were designed for the same.

Results/ Challenges

- The results of each part of the report were included in the table and figure description bullet points, we felt it was best to explain our findings there so the reader could look at the figure while reading our results to confirm.
- Challenges during this study occurred during the analysis stage of the project. While we would execute an idea, we would detect errors such as the closed date being less than the created date resulting in incoherent results. It took us a while to identify the errors and it impeded us from moving forward for a while.
- Finding datasets that would corroborate or disprove our hypothesis was another difficult challenge. It was challenging finding census demographic data and weather data that could go as granular as breaking down data by zip code and year. Many datasets online averaged or summarized the data so they did not provide the data points we wanted.

Summary & Conclusion

Based on our own assessment, we believe we were able to successfully gain a better understanding of the 311 complaints dataset. In part 1, we were able to detect and gain insight into the human errors made when reporting or following up a 311 complaint. After generating statistics for numerous types of errors, we resolved cells in the table with N/A if they were inconsistent with the parameters for that particular column. These N/A cells were omitted in part 2 of the analysis. When commencing part 2, we successfully were able to identify other errors that were not taken into account in the part 1 data cleaning and resolve them accordingly such as replacing closed dates that had years that were not between 2009-2017 or had a turnaround time (time difference between closed date and created date) that was negative the the closed date was before the created date with N/A.

Following the additional data clean up, we were able to analyze the complaint data itself and gain a better understanding of what the people of NYC complain about. We could calculate the variety of complaints across 2009-2017 and identify which issues were reported the most. Following this, we could narrow the scope of our study of complaints to figure out which boroughs reported the highest number of complaints, which were the predominant complaints in each borough, and the time of data most complaints were made. By observing the trends in our plots and figure, we were able to speculate about some of the observations we made and generate hypothesis that could possibly explain the calculated data.

Based on our observation of part 2, were were able to generate 3 hypothesis that could explain the trends in the complaint data. Since heat was the second major complaint, we thought that higher complaints about heat could be explained by the minimum temperature of that day. That is colder days would report more complaints pertaining to heat. By plotting the data and conducting the pearson correlation, it was determine that there was a weak correlation between temperature and heat complaints. For the second hypothesis, we believed more people living in a borough was related to more complaints from that borough. The plot and pearson correlation calculation indicated a strong correlation but we thought this data could be skewed since so few data points were used. To adjust for this, a new demographic dataset was incorporated to generate a new hypothesis that would be more granular than the 2nd hypothesis. Using zip code population data, we hypothesized that the more people living in a zip code indicated more complaints from that zip code and the plot and Pearson calculation supported the correlation with a more reasonable value.

Contributions By Each Member

- **Nikhil Reddy**- Code for merging datasets, statistical analysis of complaints dataset code, Pearson correlation code for hypothesis testing, and data cleaning code.
- **John Zachary Martinez**- Code for merging datasets, code for grouping complaint data sets into various categories, code for data cleaning.
- **William Herrera-** Report creation and figure creation using excel and Tableau from outputs of the code. Data set collection and research for statistical analysis. Assisted with code creation.

We met 2 or 3 times a week to create this project. Collaboration was constant so although what is listed above were the main roles, everyone contributed somewhat to the other ones work such as providing ideas/ solutions, searching code commands etc.

References

- Pearson's Correlation:https://docs.scipy.org/doc/numpy/reference/generated/numpy.corrcoef.html
- DataFrames & SparkSQL: http://spark.apache.org/docs/latest/sql-programmingguide.html
- PySpark: https://www.dezyre.com/apache-spark-tutorial/pyspark-tutorial
- NOAA: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-baseddatasets
- Tableau: http://onlinehelp.tableau.com/current/pro/desktop/en-us/concepts.html
- Weather Date Set: https://github.com/toddwschneider/nyc-taxi-data/blob/master/data/central_park_weather.csv.
- Demographic Data by Borough: http://www1.nyc.gov/site/planning/data-maps/nyc-population/census-2010.page
- SQL Instruction: https://datascience.stackexchange.com/questions/13123/import-csv-file-contents-into-pyspark-dataframes#13131
- Demographic Census Data by Zipcode: http://zipatlas.com/us/ny/zip-code-comparison/population-density.16.htm