



NYC Complaints Data Report

Authors:

- Apurv Srivastav as11419
- Varun Mishra vm1420
- Surbhi Thole sst390

Guided By:

- Prof. Claudio Silva

	Index	
Abstract		Page- 3
NYC Complaint Data Summary		Page - 4
Inceptive Data Inspection		Page - 4
Data Analysis		Page - 9
Data Exploration		Page - 23
References		Page - 24

Abstract

The report delineates a thorough analysis and cleaning of the New York City Complaints data from the year 2006 to 2015. The main objective of this archive is to provide an insight into the statistics of the complaints made by the New York City people related to different issues and come up with useful inference regarding the problems people are facing due to these issues over the years. To meet the intent, The analysis is performed by data cleaning as the first step to remove inutile data using Spark script. The document recapitulate the steps of Data cleaning for all columns in Complaints data set.

NYC Complaints Data Summary

The document reiterate the New York City complaints data from the year 2009 to 2016. The data source that has been used is the open data from complaints department NYC, which can be found at- <https://data.cityofnewyork.us/Social-Services/311/wpe2-h2i5>.

The main objective of this report is to provide the analysis of the complaints made by the New York City people on different issues hampering day to day life situations and draw analysis as to which issues are faced by the people the most and what actions should be taken to avoid further grievance of the people.

Inceptive Data Inspection:

The NYC open data from the above stated source is from the year 2009-2016. The data consists of 52 columns and a large number of rows. Initially, we made a study of the percentage of null values in the columns of data and made a supposition on the efficacy of the values in the column. We concluded that if the column contains a typical maximum number of percentage of the data as null values, then it is irrelevant to keep it and hence reduced the space by removing the null values. The initial analysis scripts are provided on GitHub with the filename as “Cleaning.py”.

The following snapshots shows a sample of the data summary for each column on which the cleaning is done:

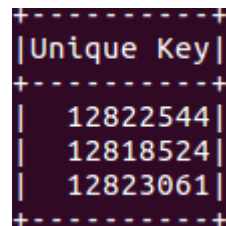
- **Column 1 :- Title → Unique ID**
Validations done on column 1 are as follows:
This column contains the unique Id's assigned to complaints.

The id length should be 8 digits and all Integers. If the entry contains any non-integer value that means the complaint is irrelevant and we have to remove it.

We found that all the entries in this column were integer values and 8 digits in length.

We removed all entries where the closed date, resolution action, action date both are null and status is either closed or null.

We fill the missing closed date values based on the resolution action update date and vice versa.



Unique Key
12822544
12818524
12823061

Fig a

- **Column 2 :- Title → Created Date**

We analysed that the format of the data entry should be -
mm/dd/yyyy hh:mm:ss AM/PM.

Created date should be any date between 2009-2016 as the data is from these years.

- **Column 3 :- Title → Closed date**

We analysed that the format of the data entry should be -
mm/dd/yyyy hh:mm:ss AM/PM

And the closed date > created date

- **Column 4 :- Title → Agency**

- **Column 5 :- Title → Agency Type**

No cleaning required in these columns as such.

- **Column 6 :- Title → Complaint Type**

In this column we analyzed that there were many similar types of complaints named differently.

For Ex: Taxi, Taxi Complaint → Taxi complaints

Which states that complains starting with similar types of names has to be grouped under same category.

Unique Key	Complaint Type
12818731	Noise - Vehicle
12818733	Noise - Commercial
12818738	Noise - Street/Si...

Fig b

Unique Key	Complaint Type
12818731	Noise Complaint
12818733	Noise Complaint
12818738	Noise Complaint

Fig c

- We assigned below mentioned types of regular expressions to find the similar categories of complaints to conclude the actual major categories of complaints made in the NYC.
- `Data = data.withColumn("Complaint Type", regexp_replace(data["Complaint Type"],"Ferry.*","Ferry Complaint"))`
- `data=data.withColumn("Complaint Type", regexp_replace(data["Complaint Type"],"Highway.*","Highway Complaint"))`
- `data=data.withColumn("Complaint Type", regexp_replace(data["Complaint Type"],"Noise.*","Noise Complaint"))`
- `data=data.withColumn("Complaint Type", regexp_replace(data["Complaint Type"],"Taxi.*","Taxi Complaint"))`
- **Column 7 :- Title → Complaint Description**

This column consist of the description of complaints made by the people of NYC. We found that there are many null values in this column and hence we replaced these NULL values by “No description” just to avoid confusion in future for the description part. As the complaint type is necessary and not description, we cannot delete the rows in description having Null Values.

Unique Key	Complaint Type	Descriptor
12825590	Asbestos	N/A
12830953	Homeless Encampment	N/A

Fig d

Unique Key	Complaint Type	Descriptor
12825590	Asbestos	No description
12830953	Homeless Encampment	No description

Fig e

- **Column 8 :- Title → Location**
- **Column 9 :- Title → Park Borough**
We deleted the above columns as it contained redundant data of no use for us to make any analysis in future.
- **Column 10 :- Title → Borough**
Replaced unspecified data with null to help in validation.
A major cleaning was done in this column. We analyzed that the data was of NYC city. NYC contains only 5 boroughs namely, Brooklyn, Manhattan, Queens, Bronx, Manhattan. But while studying the data, we found that locations other than these 5 boroughs exists in the table. As the data is NYC we have to remove data related to outside of NYC.

Unique Key	Borough
12822544	Unspecified

Fig f

Unique Key	Borough
------------	---------

Fig g

- **Column 11 :- Title → Incident Zip**

The data provided to us is of NYC complaints data. Any entry which is outside of NYC is irrelevant, and we cleaned it in a way as follows:
we calculated the minimum and maximum Zip codes of NYC which gave us a range of Zip codes of all the locations which are placed in NYC.
We then found a sub-range of zip-codes for each boroughs. Any data entry having zip-code outside this range is irrelevant to the data provided and hence we deleted it.

Unique Key	Borough	Incident Zip
12826533	Unspecified	11220

Fig h

Unique Key	Borough	Incident Zip
12826533	BROOKLYN	11220

Fig I

- Also Calculating the range of zip-codes for all boroughs made easy for future analysis to check which borough is prone to maximum or minimum complaints, how action are taken on these complaints, the time taken to solve a complaint in a particular borough etc.

Further from the remaining columns, we computed the percentage of NULL values in each of them. If the NULL values are above a specified

threshold – 50%, we deleted that column. After every cleaning cycle we saved the cleaned data in a table – temp.
Also we kept a record of the percentage of data getting cleaned at regular intervals.

Data Summary:

After the thorough process of data cleaning, a script is written which displays the following for each column:

- Base type of the column
- Semantic type of the column
- Valid/Invalid/Outlier

The following summarizes the output for each column:

Data Analysis

1) First we consider the data from each borough for the maximum number noise complaints made. These complaints are compared for each year between 2010 – 2017.

COMPLAINT TYPE	BOROUGH	YEAR	COUNTS
Noise Complaint	BRONX	2017	78258
Noise Complaint	BRONX	2016	78271
Noise Complaint	BRONX	2015	68640
Noise Complaint	BRONX	2014	59352
Noise Complaint	BRONX	2013	47631
Noise Complaint	BRONX	2012	41666
Noise Complaint	BRONX	2011	37689
Noise Complaint	BRONX	2010	40351
Noise Complaint	BROOKLYN	2017	122098
Noise Complaint	BROOKLYN	2016	125615
Noise Complaint	BROOKLYN	2015	114155
Noise Complaint	BROOKLYN	2014	99713
Noise Complaint	BROOKLYN	2013	75665
Noise Complaint	BROOKLYN	2012	65047
Noise Complaint	BROOKLYN	2011	58539
Noise Complaint	BROOKLYN	2010	56990
Noise Complaint	MANHATTAN	2017	127600
Noise Complaint	MANHATTAN	2016	132725
Noise Complaint	MANHATTAN	2015	123869
Noise Complaint	MANHATTAN	2014	108964

Fig - 1a

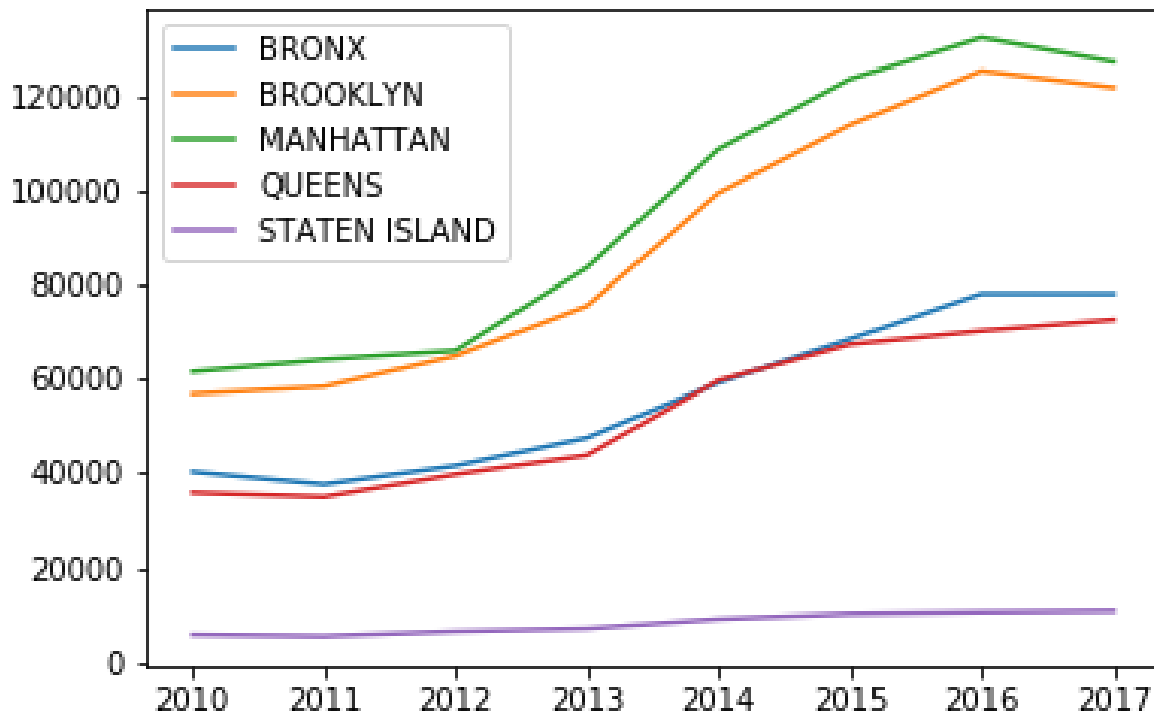


Fig - 2a

The above graph shows analysis of noise complaints made in each borough in each year. The inference that can be made from the above graph is that:

- 1) There are very less noise complaints made in Staten Island.
- 2) For Bronx and Queens the complains made are almost in the same range which are less than the complaints made in other Borough.
- 3) The complaints made in Manhattan and Brooklyn are the highest as compared to the complaints made in other borough in the New York City.

Hypothesis based on this chart:

A) We suppose that the complaints made in Manhattan and Brooklyn are due to the high population and the land area of these borough's is more due to which the there might be major complaints done in this area.

B) We analyzed that the counts of complaints are getting higher as we go from 2010 to 2017. 2017 being the highest in all the boroughs.

We suppose that this might be due to technological enhancements throughout these years due to which noise is getting created.

For ex: The number of people who can afford cars might be increased with the increase in population in these areas, creating noise due to traffic and all.

2) Next we see the total complaints count made in each of the Borough in New York City in each year.

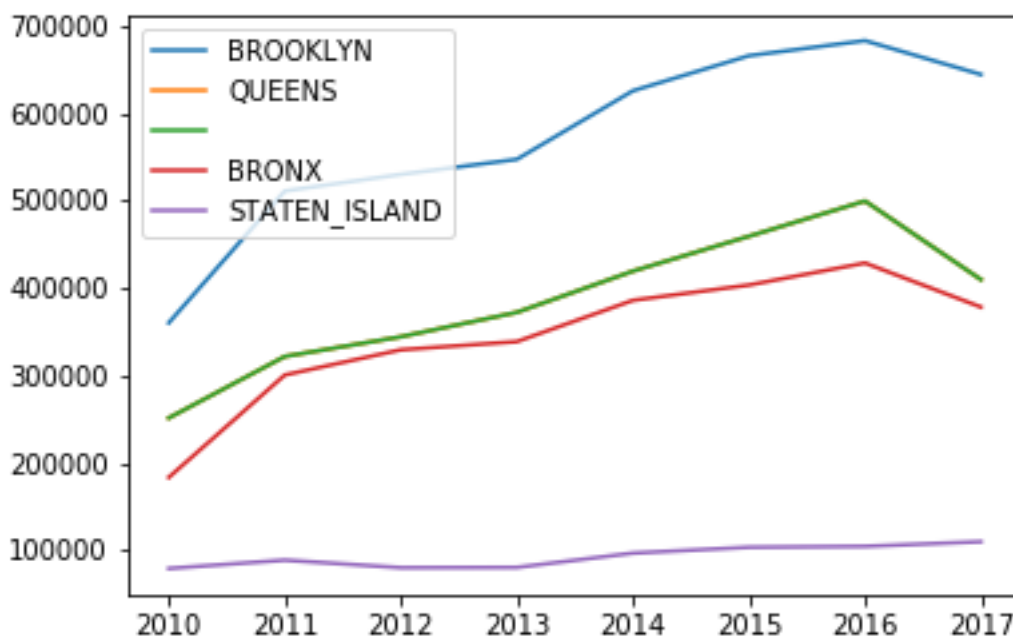


Fig 3a

From the above graph we analyzed that the number of complaints made in Brooklyn are the highest amongst all. Staten Island consist of the least amount of complaints made.

For Manhattan and Queens the number of complaints made are on an average same.

We analyzed that the number of complaints made in Staten island are on and average never increasing throughout the years.

Hypothesis:

A) The complaints which are made in Staten island are resolved much quickly as compared to the complaints made in other boroughs. This may be because once the issue is resolved no more complaints are made on the same issue keeping the complaint count same throughout the years.

And the issues for which the complaints are made in Staten Island are observed to be consistent according to the graph.

B) For all the other boroughs the number of complaints of different are analyzed to be increasing throughout the years.

3) Next we analyze the different complaints in each of the borough and the counts of the complaints made. Here we consider the Borough Brooklyn and the different complaints made there and the number of complaints count.

```
>>> spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Brooklyn','Complaint Type' FROM table WHERE Borough="BROOKLYN" GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Brooklyn' DESC LIMIT 5').show()
+-----+-----+-----+
|YEAR|Complaint counts for Brooklyn| Complaint Type|
+-----+-----+-----+
|2017|122098| Noise Complaint|
|2017|50578| Illegal Parking|
|2017|49723| HEAT/HOT WATER|
|2017|44084|Blocked Driveway|
|2017|23030|Street Condition|
+-----+-----+-----+

>>> ppp8 = spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Brooklyn','Complaint Type' FROM table WHERE Borough="BROOKLYN" GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Brooklyn' DESC LIMIT 5')
>>> ppp9 = spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Manhattan','Complaint Type' FROM table WHERE Borough="MANHATTAN" GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Manhattan' DESC LIMIT 5')
>>> spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Manhattan','Complaint Type' FROM table WHERE Borough="MANHATTAN" GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Manhattan' DESC LIMIT 5').show()
+-----+-----+-----+
|YEAR|Complaint counts for Manhattan| Complaint Type|
+-----+-----+-----+
|2017|127600| Noise Complaint|
|2017|34652| HEAT/HOT WATER|
|2017|18164| Illegal Parking|
|2017|15568|Homeless Person A...|
|2017|13571| Street Condition|
+-----+-----+-----+

>>> ppp10 = spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Queens','Complaint Type' FROM table WHERE Borough="QUEENS" GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Queens' DESC LIMIT 5')
>>> spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Queens','Complaint Type' FROM table WHERE Borough="QUEENS" GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Queens' DESC LIMIT 5').show()
+-----+-----+-----+
|YEAR|Complaint counts for Queens| Complaint Type|
+-----+-----+-----+
|2017|72734| Noise Complaint|
|2017|48761|Blocked Driveway|
|2017|42099| Illegal Parking|
|2017|28367|Street Condition|
|2017|21489| HEAT/HOT WATER|
+-----+-----+-----+

>>> █
```

Fig – 4a

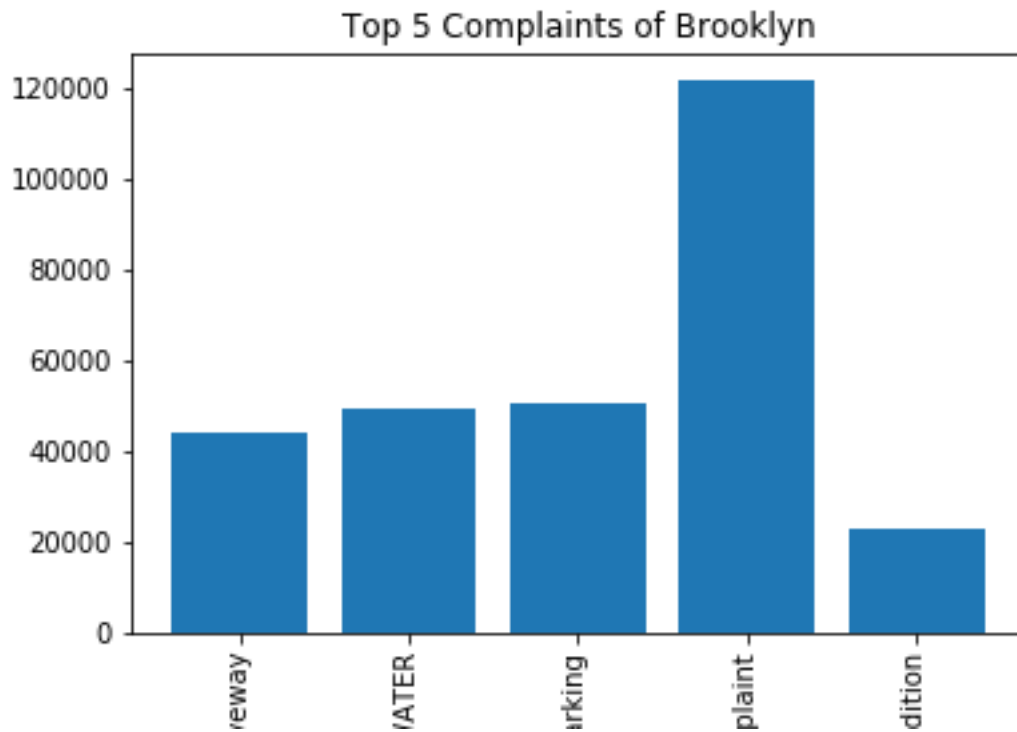


Fig 5-a

The above chart displays the number of complaints made in Borough Brooklyn. We realize that the maximum number of complaints made are noise complaints and the least number of complaints made are for street complaints.

This shows that the street complaints are either made less or are resolved in such a manner that the same issue is not persistent in the area.

The rest of the complaint type such as Water, illegal parking etc are on an average same.

In this area major concern is the noise as analyzed from the above graph.

4) Next we consider Complaint types for Manhattan and their counts.

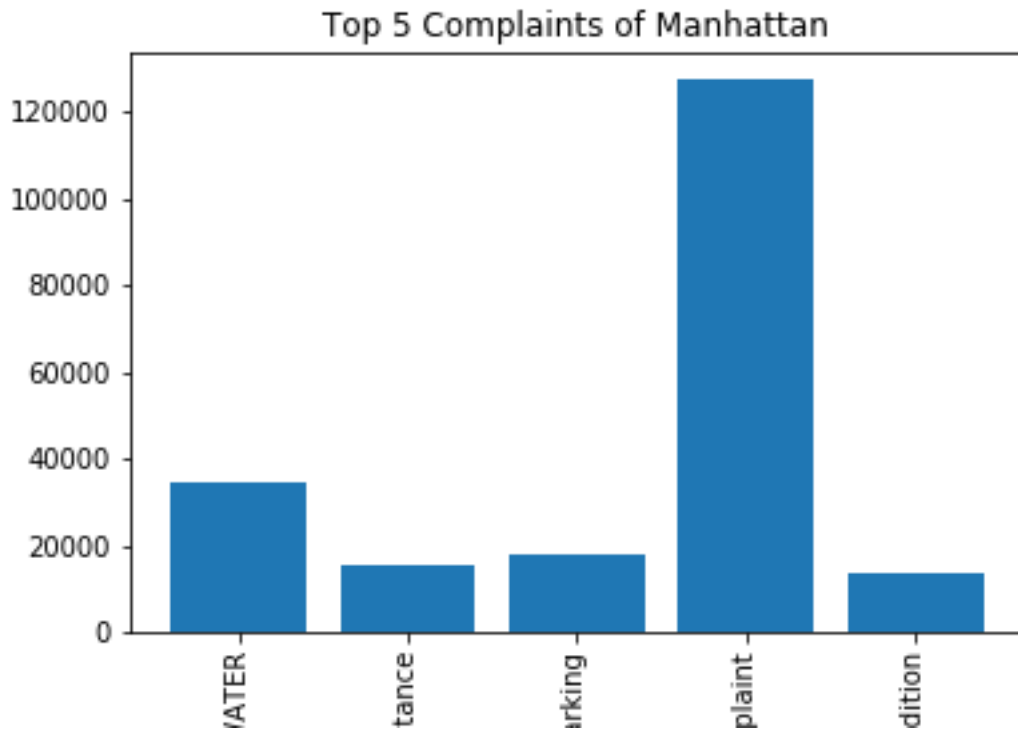


Fig 6-a

The above graph and Fig – 4a shows that the major five complaints made in Manhattan are Water, Heat, Water, Parking and Street Conditions. We analyzed that the major complaints made in Manhattan are Noise Complaints and the least complaints made are for street conditions. As Manhattan is the center of attraction for New York City, we infer that the streets will obviously be of much better conditions as compared to the other part of the city. The nominal complaints which are made can be due to traffic of streets as we know that the Manhattan area is the area with highest population and crowded for almost every part of the year.

Based on the above hypothesis we also infer that as the Manhattan city is crowded for most part of the year, The noise may be because of the the traffic and population of the people in the area.

The other complaints which are made in this city are much less than the count of the same types of complaints made in Brooklyn. We analyze that this may

be because this part is much developed than any other part of New York city, being the city's center of attraction. Hence the people will obviously not face major problems due to illegal activities as the security might be the best in this part. The city doesn't contain water complaints more as the developed part of the New York City cannot be taken risk with having water problems for obviously health concerns of the people.

Hypothesis: We hypothesize that the common complaints concerning health and good will of people are less in this part. This may be due to the richness of this area and that this part is the most expensive part of New York city. So we conclude that as you pay for services, the same amount of care is taken in return and there will be less complaints.

5) The number of different complaints made in Queens Borough.

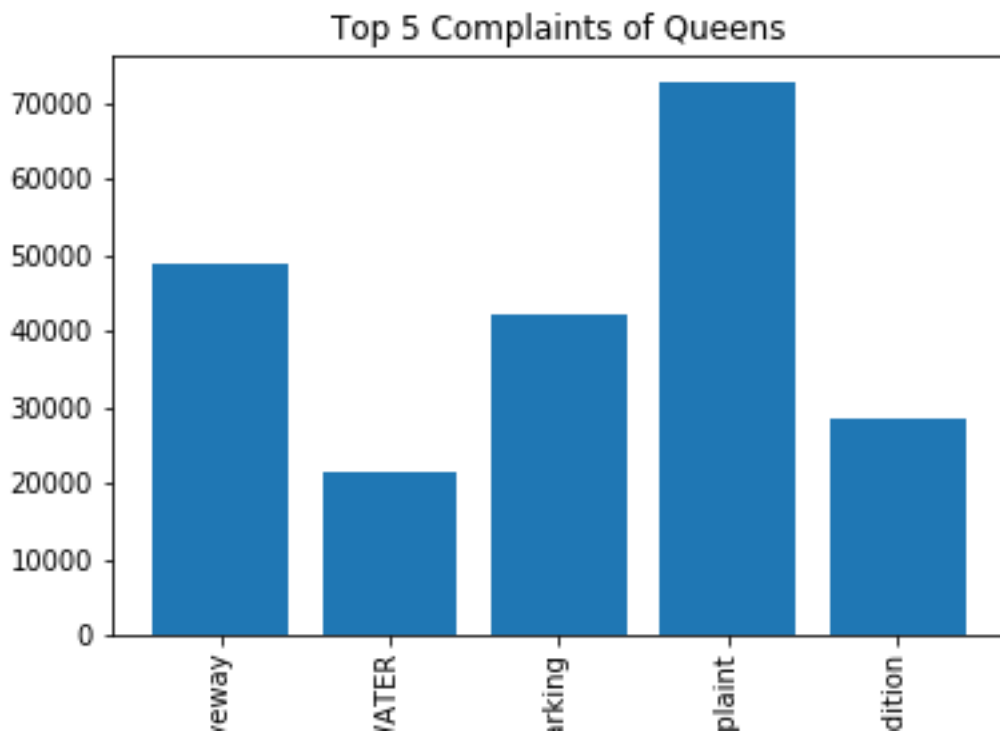


Fig 7-a

From the above graph and Fig - 4a we analyze that the maximum number of complaints made in Queens is again Noise Complaints. The least number of complaints are for Heat/ Hot Water complaints.

We analyzed that Blocked Driveway complaints and Illegal Parking complaints are very much in this part of New York city.

Hypothesis: We hypothesize that the issues related to Blocked driveway and illegal parking are inter-related to each other. This may be explained as - if people are parking in the illegal areas, there are chances that the roads get blocked and hence the complaints are made. This we inferred based on the counts of the the complaints types – Blocked driveway complaints and Illegal Parking. We see that if there is illegal parking there will be chance of block in the driveway and hence the complaints are made simultaneously giving us the on an average same amount of counts for both of them.

6) The number of different complaints made in Bronx Borough.

```
@login-2-1~
>>> ppp11 = spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Staten Island','Complaint Type' FROM table WHERE Borough='STATEN ISLAND' GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Staten Island' DESC LIMIT 5')
>>> spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Staten Island','Complaint Type' FROM table WHERE Borough='STATEN ISLAND' GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Staten Island' DESC LIMIT 5').show()
```

YEAR	Complaint counts for Staten Island	Complaint Type
2017	11043	Electronics Waste
2017	10795	Noise Complaint
2017	9316	Street Condition
2017	7377	Illegal Parking
2017	5870	Street Light Cond...

```
>>> ppp12 = spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Bronx','Complaint Type' FROM table WHERE Borough='BRONX' GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Bronx' DESC LIMIT 5')
>>> spark.sql('SELECT SUBSTRING('Created Date',7,4) AS YEAR,COUNT(*) AS 'Complaint counts for Bronx','Complaint Type' FROM table WHERE Borough='BRONX' GROUP BY Year,'Complaint Type' ORDER BY Year DESC,'Complaint counts for Bronx' DESC LIMIT 5').show()
```

YEAR	Complaint counts for Bronx	Complaint Type
2017	78258	Noise Complaint
2017	50527	HEAT/HOT WATER
2017	21953	Blocked Driveway
2017	20511	UNSANITARY CONDITION
2017	16597	PAINT/PLASTER

```
>>> spark.sql('SELECT Status,COUNT(Status) AS Counts FROM TABLE GROUP BY STATUS').show()
```

Status	Counts
Open	39620
Unspecified	1
Unassigned	3
Draft	4
Started	117
Pending	362496
Closed	14182730
Assigned	27906

Fig 8-a

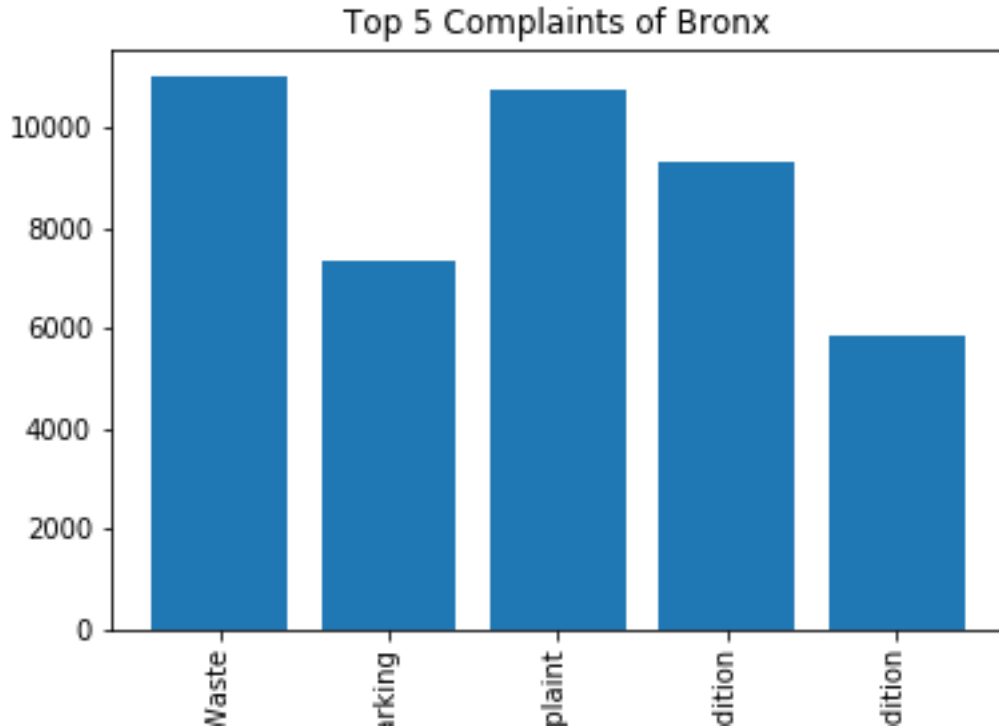


Fig - 9a

The above graph shows the complaints types and their counts in Bronx. We analyze that the top 5 complaints made in Bronx are : Noise complaints, Heat/Hot water, Blocked Driveway, Unsanitary Condition, Paint / plaster complaints.

Noise being the highest complaint and Paint/ Plaster complaints being the lowest amongst all.

We analyzed that the sanitary complaints are made in Bronx to the highest as compared to the other regions of New York City. We can also infer that maximum number of sanitary problems are made from Bronx amongst all the boroughs. This knowledge about the sanitary complaints is unusual as compared to other Boroughs, as no other borough consist of sanitary problems as their top 5 complaints.

Hypothesis:

- 1) We hypothesized that the Bronx has the most sanitary complaints throughout the New York City.
- 2) We also analyzed that that the number of complaints made in Bronx for all the complaint types are on an average the same. Which shows that count of

complaints for all types of complaints are on an average same due to which there is no as such major complaint which people are facing, all being the same avg count.

3) This sanitary complaints may be because of Bronx Borough is not as developed as other borough's of New York City, having sanitary issues as the most highest complaints type made in this borough amongst all boroughs.

7) The number of different complaints made in Staten Island Borough.

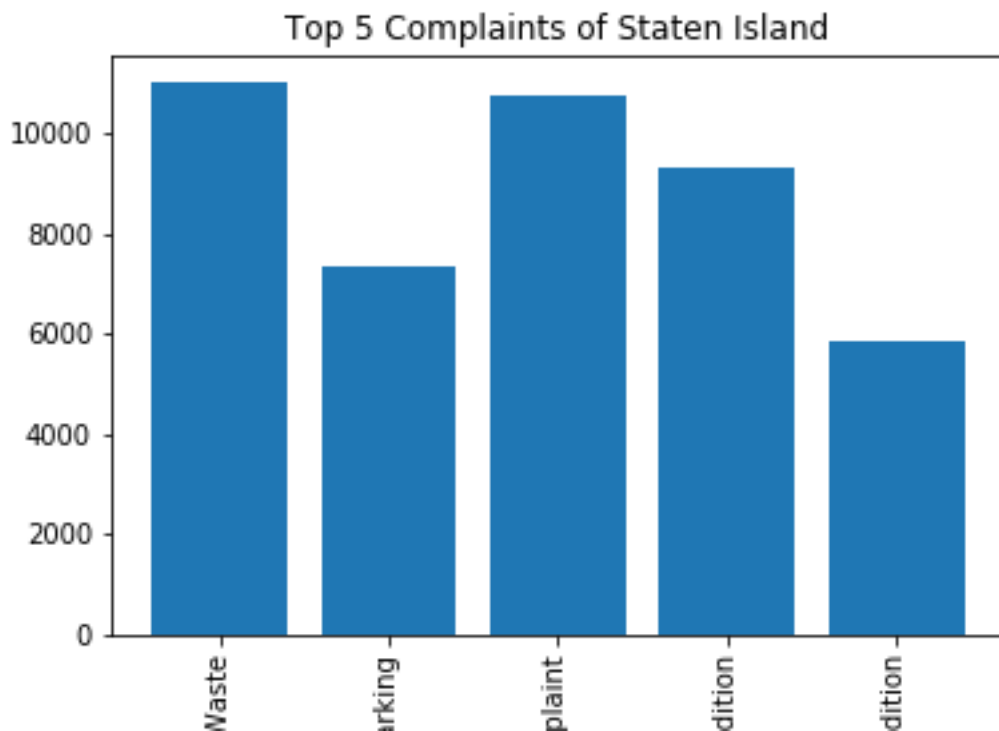


Fig 10-a

From the above figure and Fig 2-a we analyzed that the top five complaints types made in Staten Island Electronics waste, Noise complaints, Street Conditions, Illegal parking and street Light conditions. The above graph shows that the highest number of complaints are made of electronic waste. And the lowest complaints in Staten island are for street conditions in the top five complaints made.

The most unusual knowledge found from above graph is that like any other Borough of New York City, Staten Island doesn't face maximum number of Noise complaints.

On the other hand we realized that the electronics waste in Staten island is the most as compared to any other borough of New York City.

Hypothesis:

1) The lesses number of noise complaints may be due to less number of population in Staten Island as compared to the other boroughs of New York City.

2) Two reasons for e-waste are – growth of population, Increase in use of technologies. As we inferred above that the noise complaints might be less due to less population in the above area, we can infer that the growth of electronic waste and the complaints related to these might be because of the increase use of technology by the people of Staten island as compared to other people from other boroughs of New York City.

8) Complaints count for each agency per year:

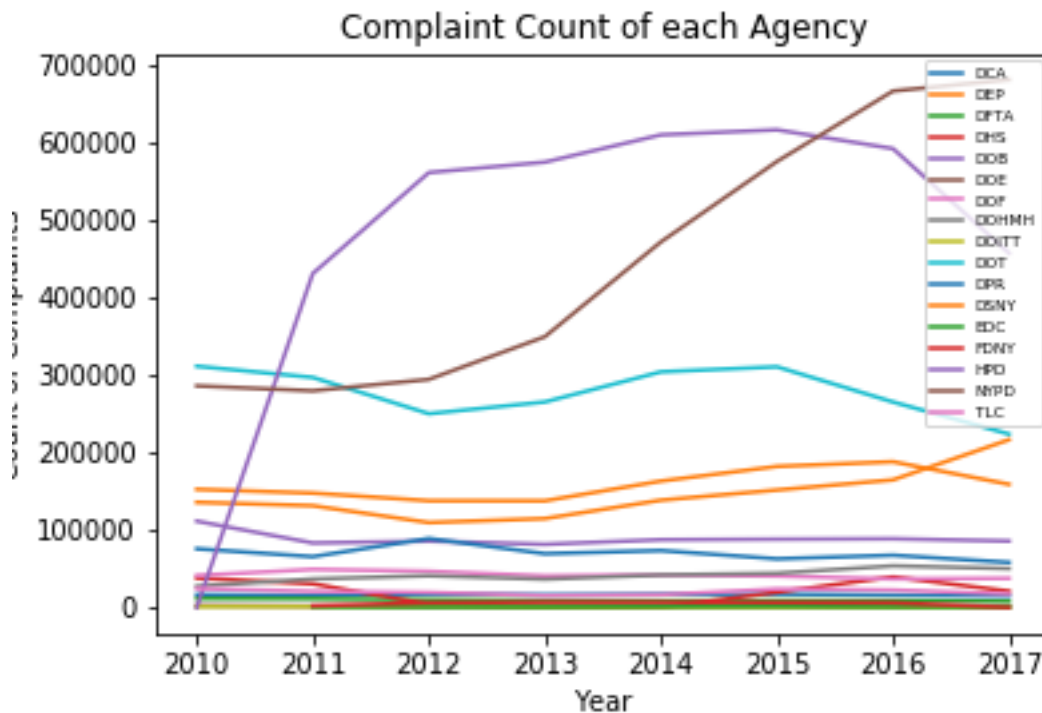


Fig – 11a

AGENCY	YEAR	COUNTS
DCA	2017	15034
DCA	2016	16063
DCA	2015	17184
DCA	2014	17008
DCA	2013	16486
DCA	2012	17118
DCA	2011	15700
DCA	2010	16526
DEP	2017	158959
DEP	2016	188083
DEP	2015	182054
DEP	2014	163156
DEP	2013	137657
DEP	2012	137772
DEP	2011	147645
DEP	2010	152484
DFTA	2017	7948
DFTA	2016	8798
DFTA	2015	9091
DFTA	2014	8701

Fig 12a

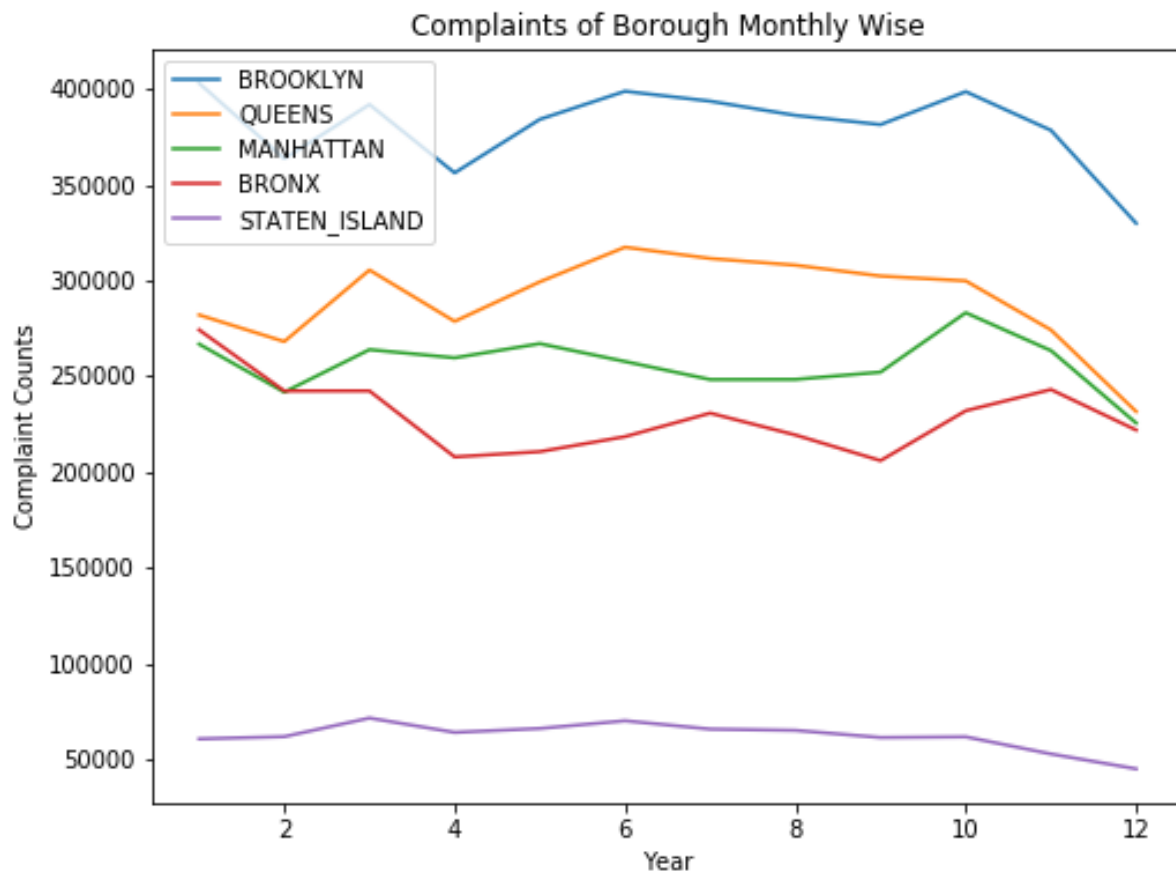
The above fig 11-a and fig 12-a shows the number of complaints made throughout the New York City for each agency per year. It has been observed that the maximum number of complaints are made against DEP agency in all years as compared to all the other agency type.

YEAR	COUNTS
HPD	3843031
NYPD	3604840
DOT	2227356
DEP	1267810
DSNY	1162831
DOB	711866
DPR	561682
DOF	334882
DOHMH	331827
TLC	160289
DHS	133368
DCA	131119
DFTA	76499
FDNY	34319
DOE	13006
EDC	6892
DOITT	5662

Fig 13-a

The above figure shows the distinct agencies in New York City against which complains are made. The figures shows that the maximum number of complains are made to HPD agency followed by NYPD throughout 2010 – 2017. These figures shows that the maximum complains are made to NYPD throughout these years and that the crime rate or other complaints involving NYPD's support are increasing per year.

9. Complaint Count of each Borough Monthly wise:

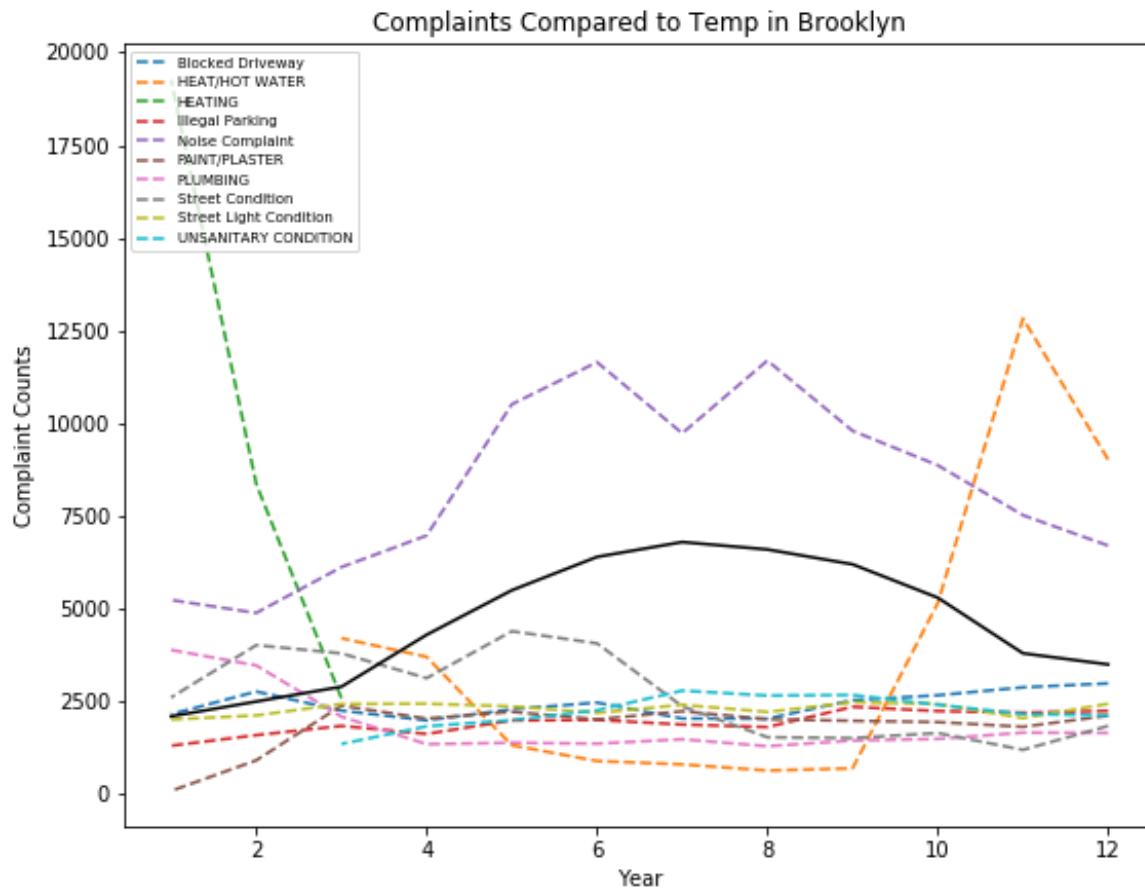


The complaints count fluctuates during summer, but it tends to decrease during winter time.

Hypothesis: The decrease in the complaint count can be attributed to the fact that there is very less movement around and everyone prefers to stay in their house. Hence we see drop in the count during the winter season.

Data Exploration

We tried exploring more insight into our data by comparing the analysis inferred from this data to that of other co-related dataset. We took historical weather dataset of New York City and tried to fetch out the co-relation with our existing dataset.



We found that there is generally good amount of snowfall during Jan to March every year. And this is the reason why there is lots of Hot water or heating complaints during that period. Moreover, the spike in Hot-Water complaints during November month can be due to the fact that most people start using hot water at this time as temperature drops to 2-4 Celsius. There is also sudden increase in Heating complaints in month of December compared to November because people start using Apartment Heating appliances.

References

1. DataUSA, 2017. BRONX COUNTY, NY. [Online]
<https://datausa.io/profile/geo/bronx-county-ny/>
2. Heard, J., 2016. Being a Data Scientist: My Experience and Toolset. [Online]
<https://jeffersonheard.github.io/2017/01/being-a-data-scientist-my-experience-and-toolset/>
3. Pereira, I., 2014. 311 beloved by New Yorkers, city survey finds. [Online]
<https://www.amny.com/news/311-beloved-by-new-yorkers-city-survey-finds-1.9513521>