Big Data Project

Han Zhao (hz1411)
Hezhi Wang (hw1567)
Jin U Bak (jub205)

Abstract

This report is an exploration of the NYC 311 Dataset. We first look at its basic structures, summary statistics and detect any data quality issues. Then we test and verify several hypotheses. Our findings are, firstly, number of complaints is negatively correlated with temperature. Also, cold weather will bring about more days with extremely large number of complaints. Besides, out of all the type of complaints, heating problems have strong seasonality, which contributes to the rise of complaints in winter times. Further, Complaints are distributed unevenly in different zip code, and is correlated with population, income, age and sex ratio of the corresponding district. We also look at the correlation between the number of noise-related complaints and the number of motor vehicle accidents in NYC. It is found to be no significant relationship between the two.

Introduction

311 is a service line provided by the City of New York. It acts as a hotline for the New York city and provides various information about the government agencies and city services. It also provides a platform to file a complain to various government agencies in the city such as Department of Public Health(DPH), NYPD, Department of Transportation(DOT), etc. 311 dataset has records of complaints filed from 2009 to present and is updated daily. In this report, we aim to have a deeper understanding of the NYC311 Dataset. The first part is to explore its basic structure with some summary statistics and to address data quality issues. This study allows us to learn more about the distribution of complaints across the city by borough, complaint types, date. Based on findings from the first part, we further focus on some hypothesis of interest of the dataset. Since the data is huge and we need to analyze and aggregate each column, big data infrastructure can improve efficiency greatly. For the analysis of 311 data, we have used Pyspark as the main framework for our data processing and analysis as it enabled us to process millions of rows of data efficiently in timely manner.

Part I Data Summary and Data Quality Issues

NYC 311 provides 24/7 access to New York City government services. The 311 dataset we are using contains all 311 Service Requests which is updated daily from 2009 to present with more than 15m records and 52 columns. Dataset is downloaded from the following links:

Dataset for 2009:
https://data.cityofnewyork.us/Social-Services/new-311/9s88-aed8
Dataset for 2010-present:
https://data.cityofnewyork.us/Social-Services/311/wpe2-h2i5

Two datasets are combined as one and analyzed.

Dataset is also available on NYU HPC HDFS, /user/jub205/311all.csv

1. Data Quality Issues

1.1 Summary for Base Type

Each column is assigned a base type, and values that does not match the base type will be marked as invalid. Below is the count for base types across columns.

Base Type	**Number of Columns**
BOOLEAN	1
FLOAT	2
DATETIME	4
INT	5
NOMINAL	6
TEXT	34

1.2 Summary for Semantic Type

Each column is assigned a base semantic type, and values that does not match the format of the given base semantic type will be marked as invalid. Below is the count for semantic types across columns.

Semantic Type	Number of Columns	Column Name
KEY	1	Unique_Key
DATE	4	Created_Date Closed_Date Due_Date Resolution_Action_Updated_Date
AGENCY	2	Agency Agency_Name
STATE	1	School_State
CITY	2	City School_City
BOROUGH	3	Borough Taxi_Company_Borough Park_Borough
SCHOOL	6	School_Name School_Region School_Number School_Code School_Not_Found School_or_Citywide_Complaint
TYPE	5	Complaint_Type Location_Type Address_Type Facility_Type Vehicle_Type
ZIP	2	Incident_Zip School_Zip
ADDRESS	8	Street_Name Incident_Address Cross_Street_1 Cross_Street_2

		Intersection_Street_1 Intersection_Street_2 School_Address Taxi_Pick_Up_Location
GEOLOCATION	5	X_Coordinate_(State_Plane) Y_Coordinate_(State_Plane) Latitude Longitude Location
LOCATION	10	Landmark Community_Board Park_Facility_Name Garage_Lot_Name Bridge_Highway_Direction Road_Ramp Bridge_Highway_Segment Ferry_Direction Ferry_Terminal_Name Bridge_Highway_Name
PHONE	1	School_Phone_Number
TEXT	2	Descriptor Status

1.3 Summary for Labels

For each value, we first check whether it is empty, or is non-empty missing values, like ’N/A’, ‘NULL’ & ‘UNSPECIFIED’. We then check if it matches the given base type, or formatted in accordance with the semantic type of the column. For some columns, we also checked if it is beyond time or geographical range.

To be specific, for columns with base type DATETIME, we checked if it is written in the correct date format, and whether it lies within year 2009 to 2017. For a valid zip code, it should have the correct format, which we checked by regular expression, and also lies within New York, which is from 00501 to 12975. For columns with semantic type ADDRESS, they are validated by checking whether there are NA or special characters like ‘@‘ included.

Column Name	**Number of Valid Value**	**Number of NULL Value**	**Number of Invalid Value**	**Number of Non-Empty Missing Value**
Unique_Key	15413188	0	0	0
Created_Date	15413188	0	0	0
Closed_Date	14873571	533541	6076	0
Agency	15413188	0	0	0
Agency_Name	15413188	0	0	0
Complaint_Type	15413187	0	0	1
Descriptor	15270034	818	0	142336
Location_Type	11345040	3983362	0	84786
Incident_Zip	14317979	1092972	1270	967
Incident_Address	11964601	3427426	20219	942
Street_Name	11924042	3427542	60823	781
Cross_Street_1	10229980	4525081	653011	5116
Cross_Street_2	10201481	4588542	613503	9662

Intersection_Street_1	2333261	12967433	112433	61
Intersection_Street_2	0	12970374	0	2442814
Address_Type	14694494	718694	0	0
City	14325630	1086448	216	894
Landmark	8491	15404691	0	6
Facility_Type	3608878	18752	0	11785558
Status	15413121	34	0	33
Due_Date	6069842	9333649	9697	0
Resolution_Action_Updated_Date	15097944	314952	292	0
Community_Board	13893757	0	0	1519431
Borough	13893757	0	0	1519431
X_Coordinate_State_Plane	13773444	1639744	0	0
Y_Coordinate_State_Plane	13773444	1639744	0	0
Park_Facility_Name	93538	0	0	15319650
Park_Borough	13893757	0	0	1519431
School_Name	93538	0	0	15319650
School_Number	90350	2850	0	15319988
School_Region	15227	78311	0	15319650
School_Code	15235	78303	0	15319650
School_Phone_Number	76242	0	0	15336946
School_Address	93538	0	0	15319650

School_City	93538	0	0	15319650
School_State	93538	0	0	15319650
School_Zip	0	0	0	15319651
School_Not_Found	5924640	9488548	0	0
School_or_Citywide_Complaint	3992	15409196	0	0
Vehicle_Type	7974	15405214	0	0
Taxi_Company_Borough	13368	15399820	0	0
Taxi_Pick_Up_Location	123117	15290071	0	0
Bridge_Highway_Name	42691	15370497	0	0
Bridge_Highway_Direction	42629	15370513	0	46
Road_Ramp	42285	15370536	0	367
Bridge_Highway_Segment	42188	15364089	6517	394
Garage_Lot_Name	5086	15408102	0	0
Ferry_Direction	3564	15407314	0	2310
Ferry_Terminal_Name	10290	15341563	0	61335
Latitude	13773441	1639744	3	0
Longitude	13773441	1639744	3	0
Location	13773441	1639744	3	0

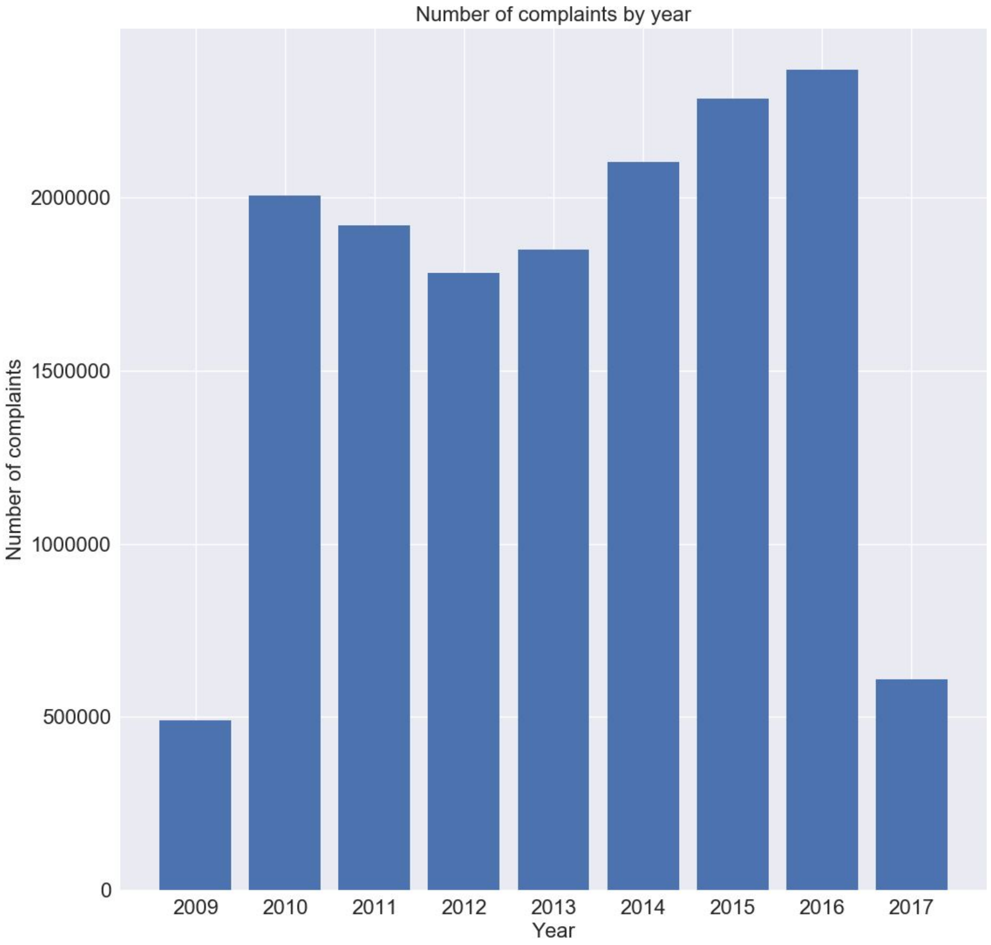
As we can see from the above table, the dataset we are working with consists of 52 columns. Some columns have high number of null values due to the nature of the columns. For example, “Ferry_Direction” has very few non-null values as the column is used when the incident location is within a Ferry. We noticed that there is a significant increase in the number of incidents reported from 2009 to 2010. This was due to missing data in 2009 as we can see from the figure3, number of complaints filed by day. There is a discontinuity of data from early 2009 to 2010. This could be a reason why dataset was separated in two parts, 2009 and 2010-present. For our future analysis, we can focus on analyzing data from 2010-present. For obvious reason, 2017 has lower number of complaints.

Some columns are quite noisy, which makes it hard to check their validity. For example, “City” contains so many different city/district names, together with many apparent mis-spellings, like ‘BROOKLN’, together with some suspicious outliers, like ‘Boston’ and even ’LA’. As 311 service is within New York, these ought to be considered invalid. However, we are unable to find a complete list of cities/twons in New York, so we cannot detect them right now. For our future analysis, we intend to explore the relation between ‘Incident_Zip’ and ‘City’, and probably remove all rows without a non-empty and valid zipcode.

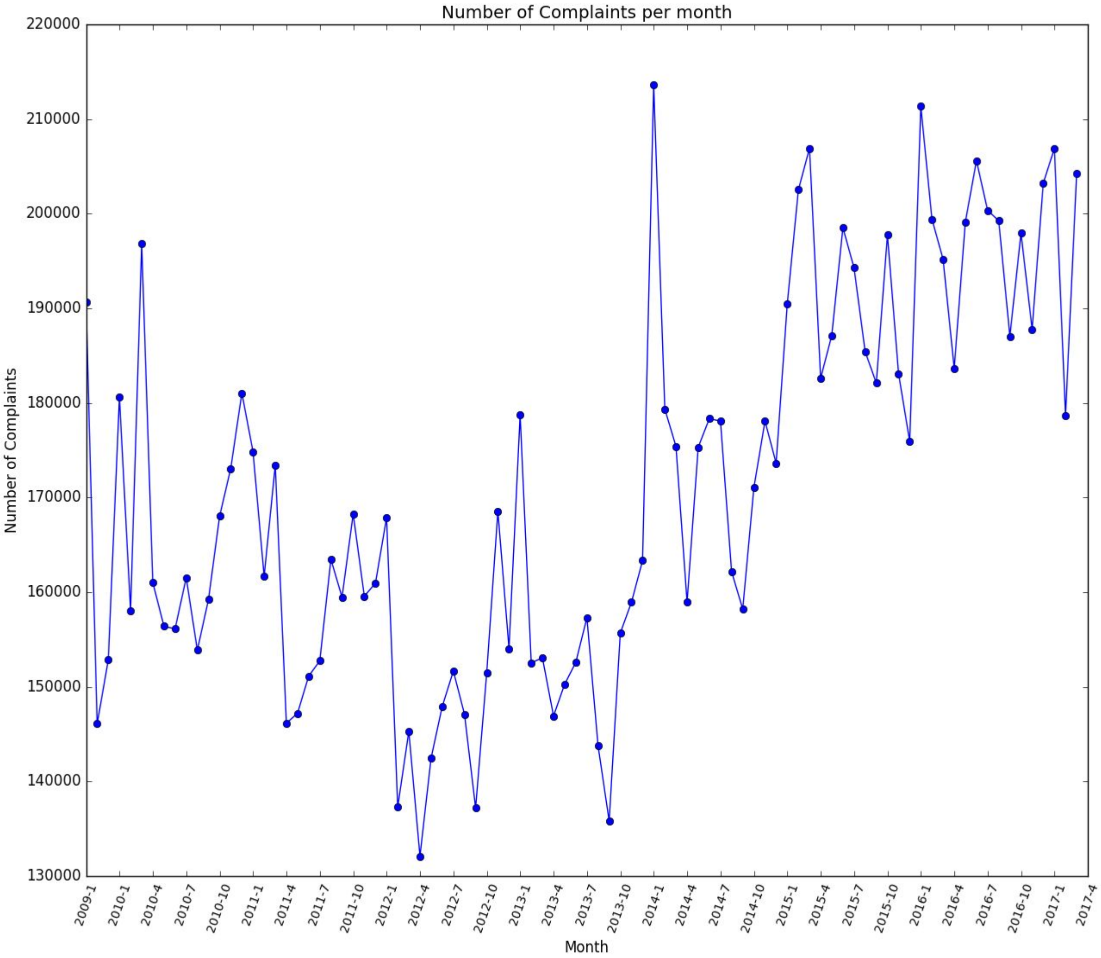
2. Data Analysis

In order to better understand the data collection we have chosen, we have done some basic analysis and summarized it.

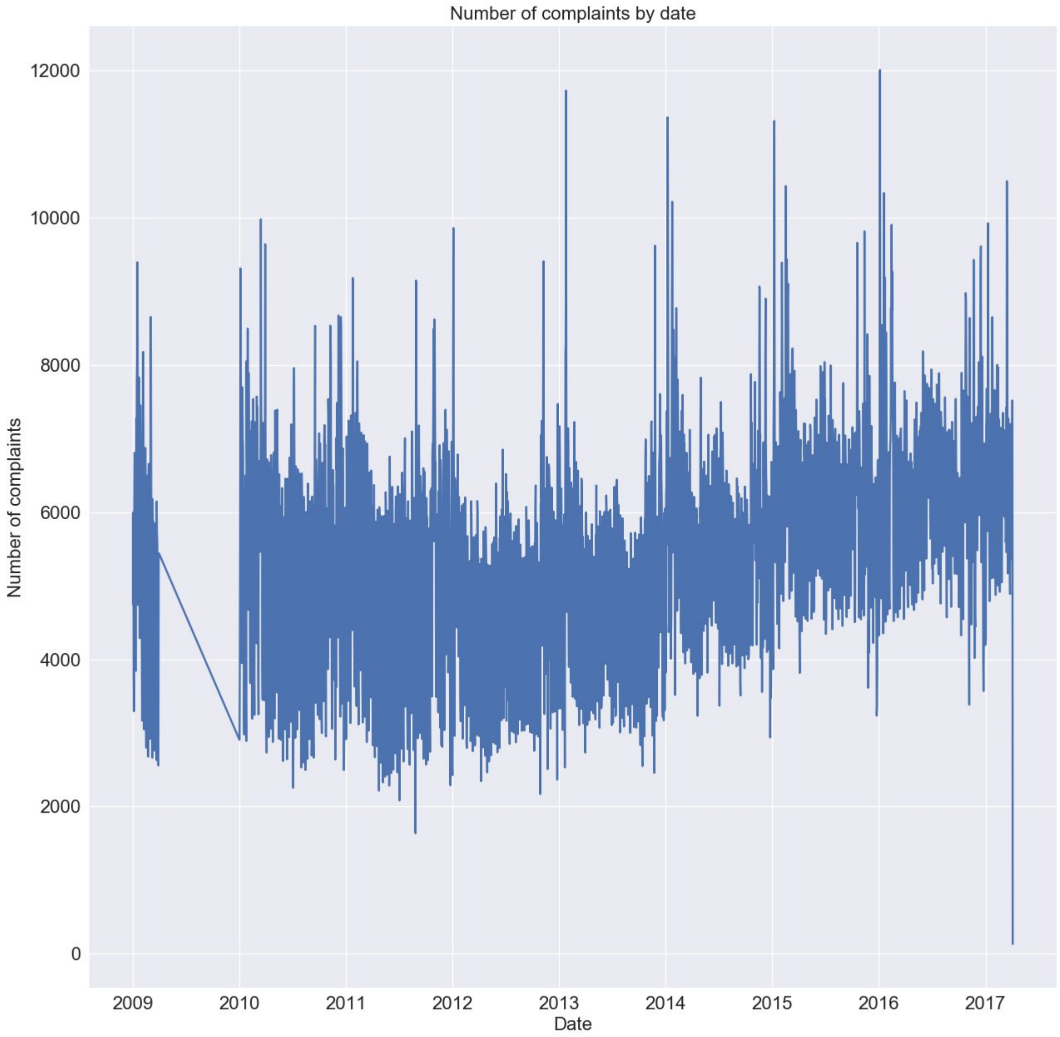
2.1 Number of complaints filed in each year



2.2 Number of complaints filed by Month

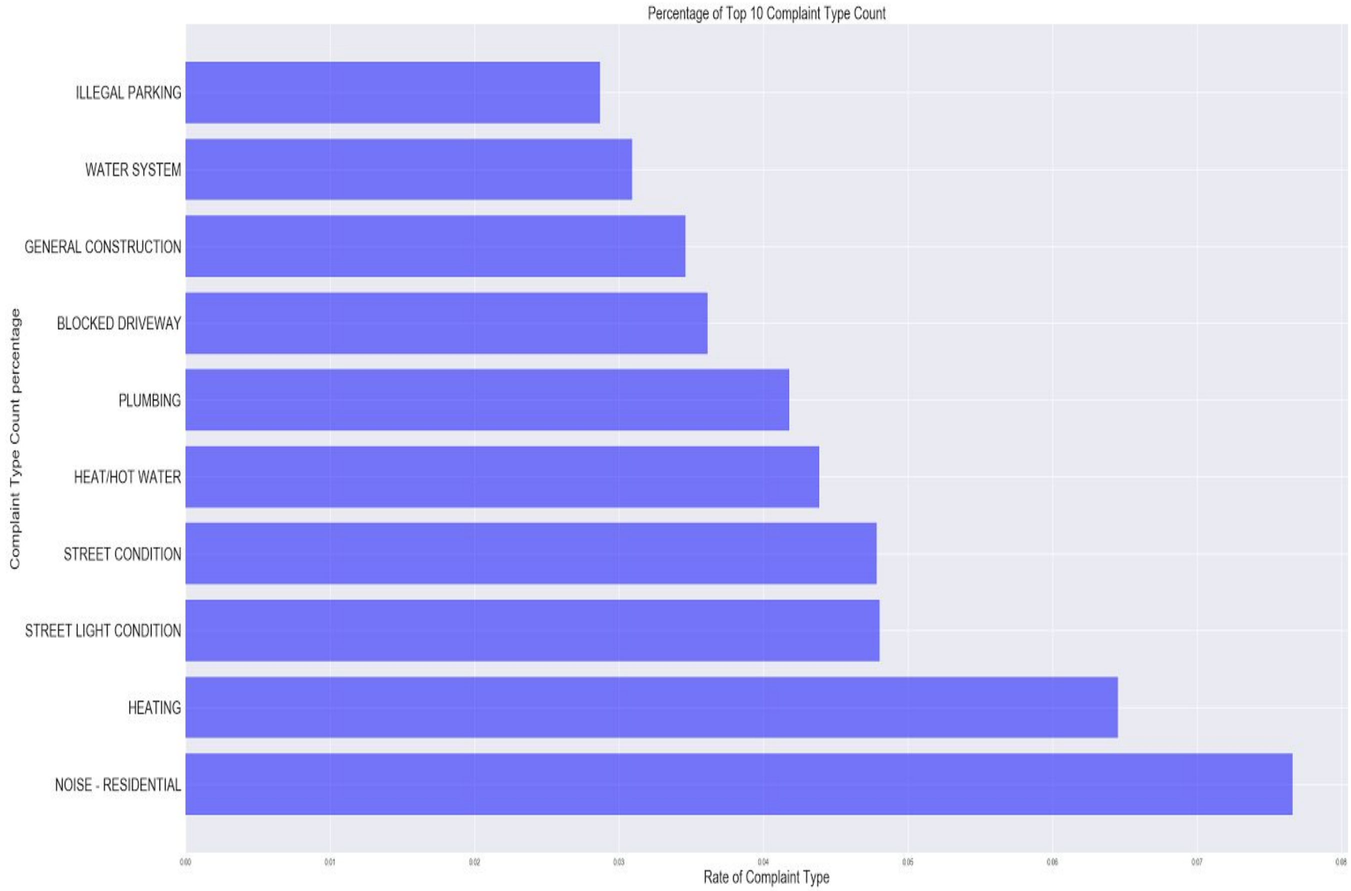


2.3 Number of complaints filed by day

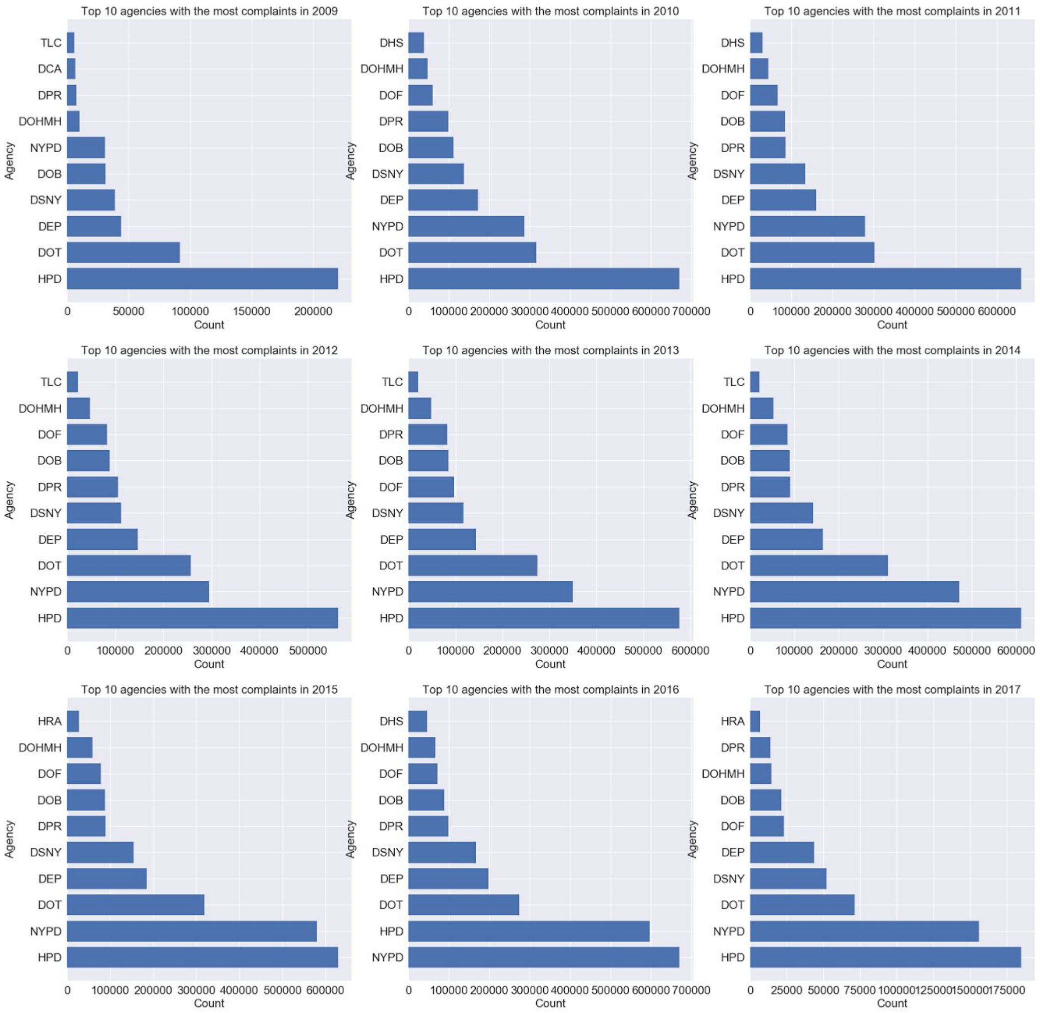


As mentioned earlier in the report, we see there is a missing chunk of data in 2009. It seems like data is available for only first few months of 2009.

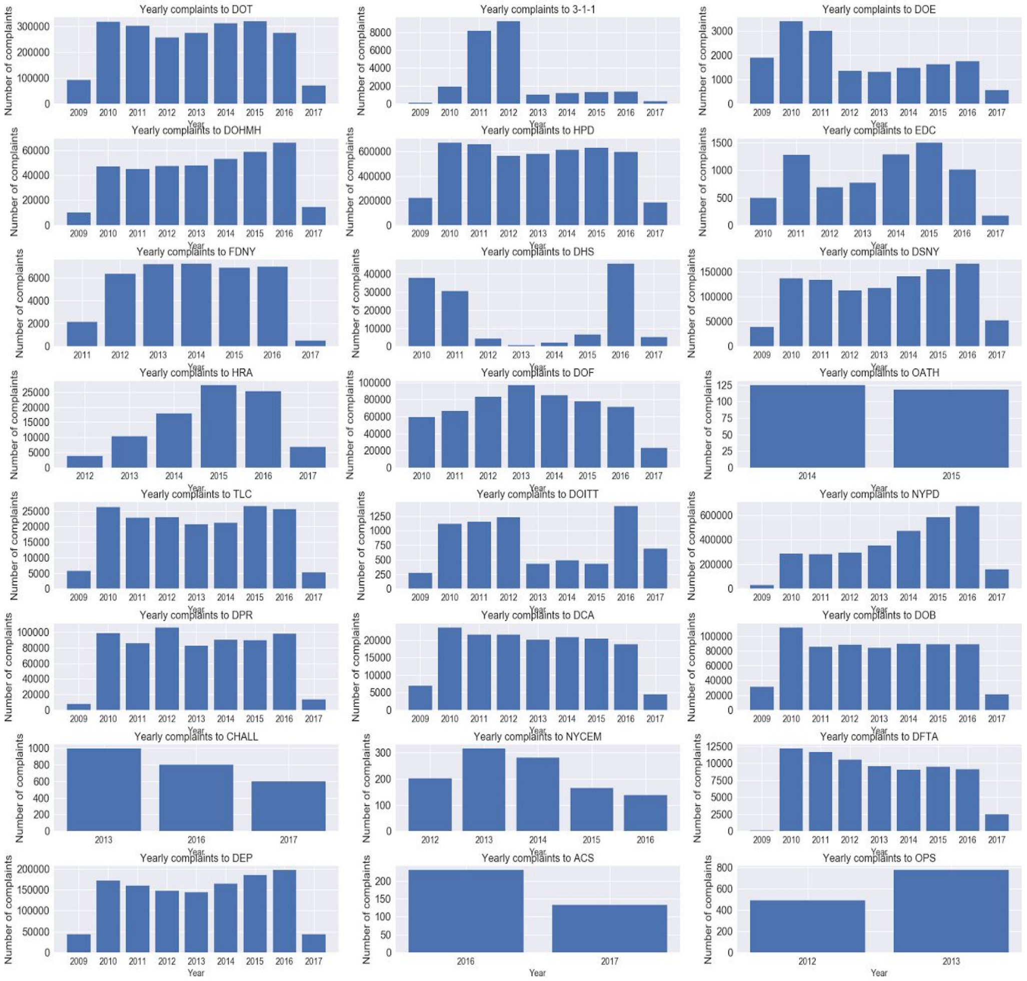
2.4 Top 10 Complaint Type from 2009 to present in NYC



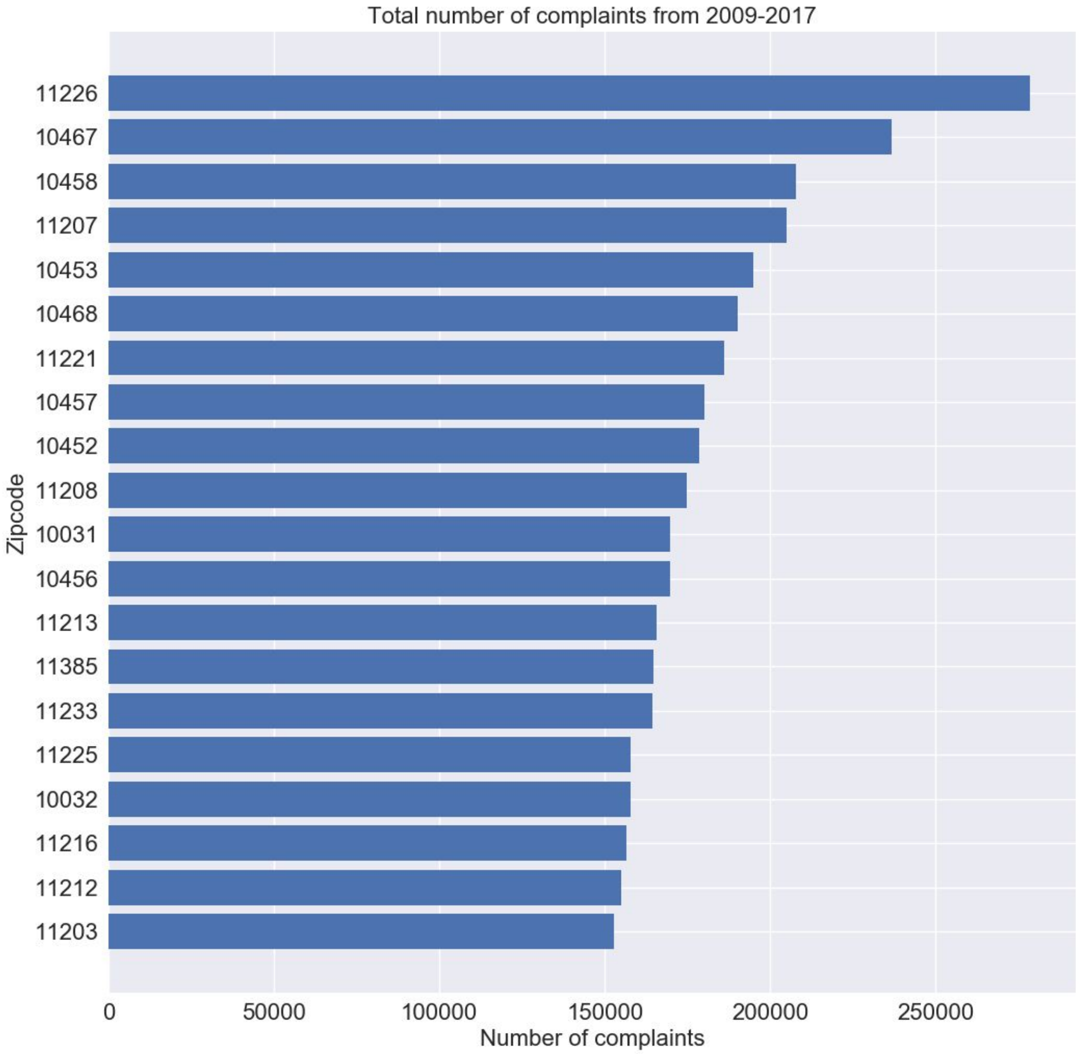
2.5 Top 10 Agencies that received the most complaints in each year



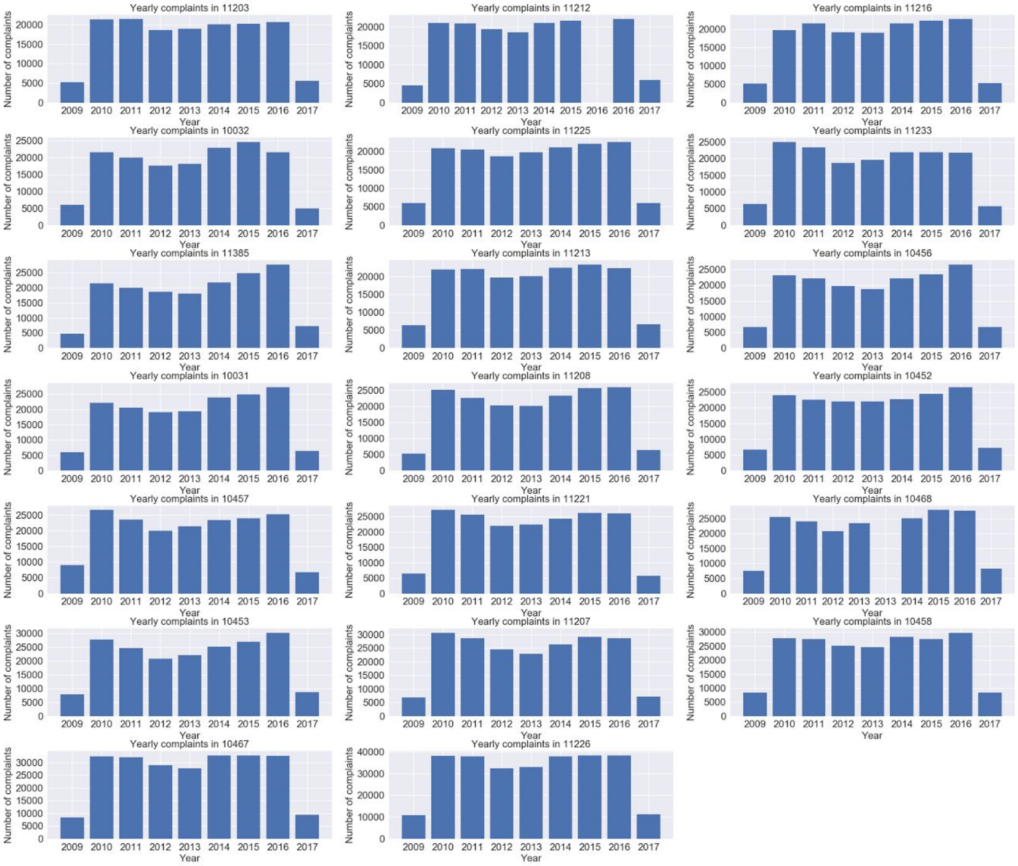
2.6 Yearly complaints filed with each agency



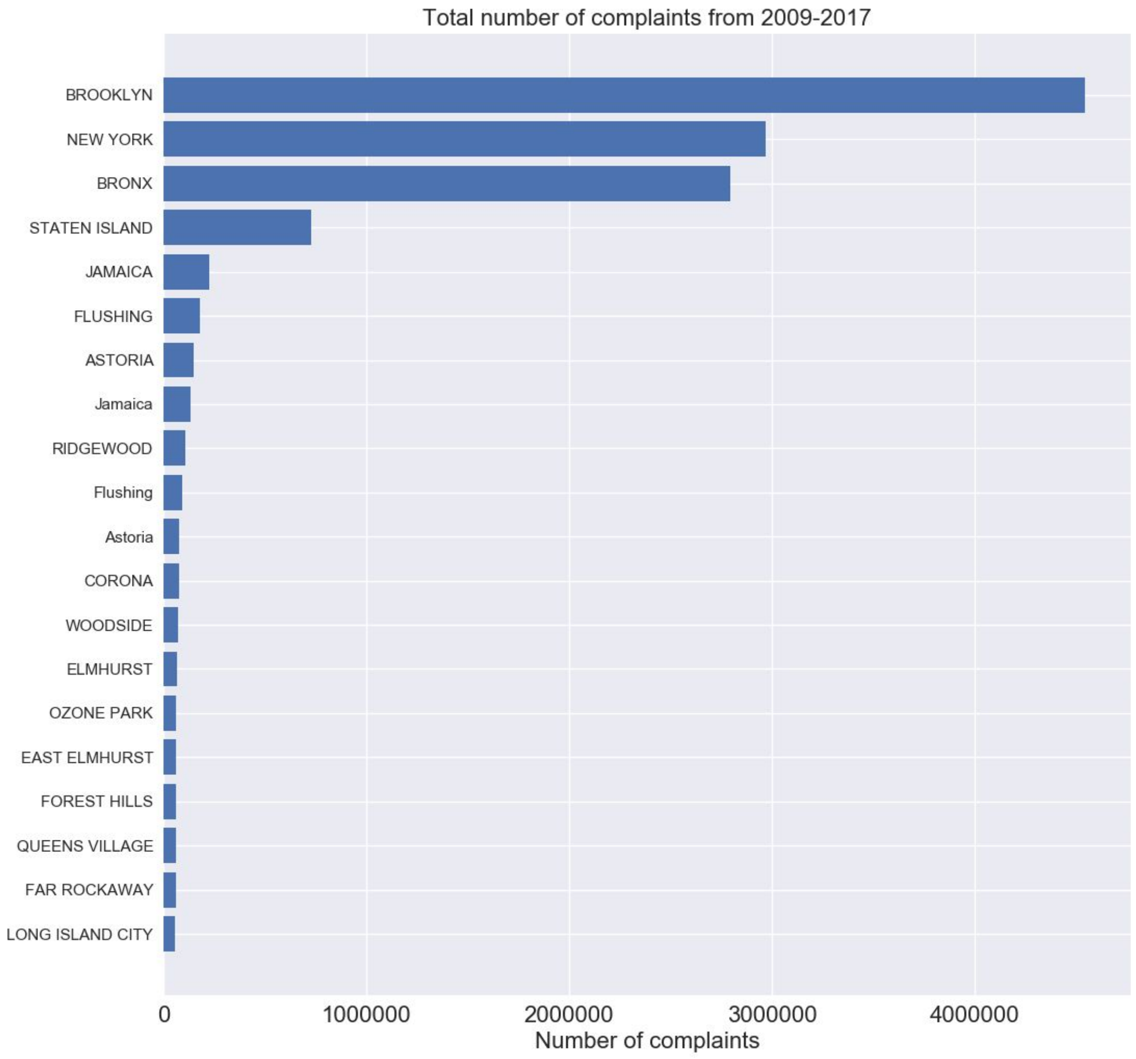
2.7 Top 20 zipcode with the most complaints from 2009-2017



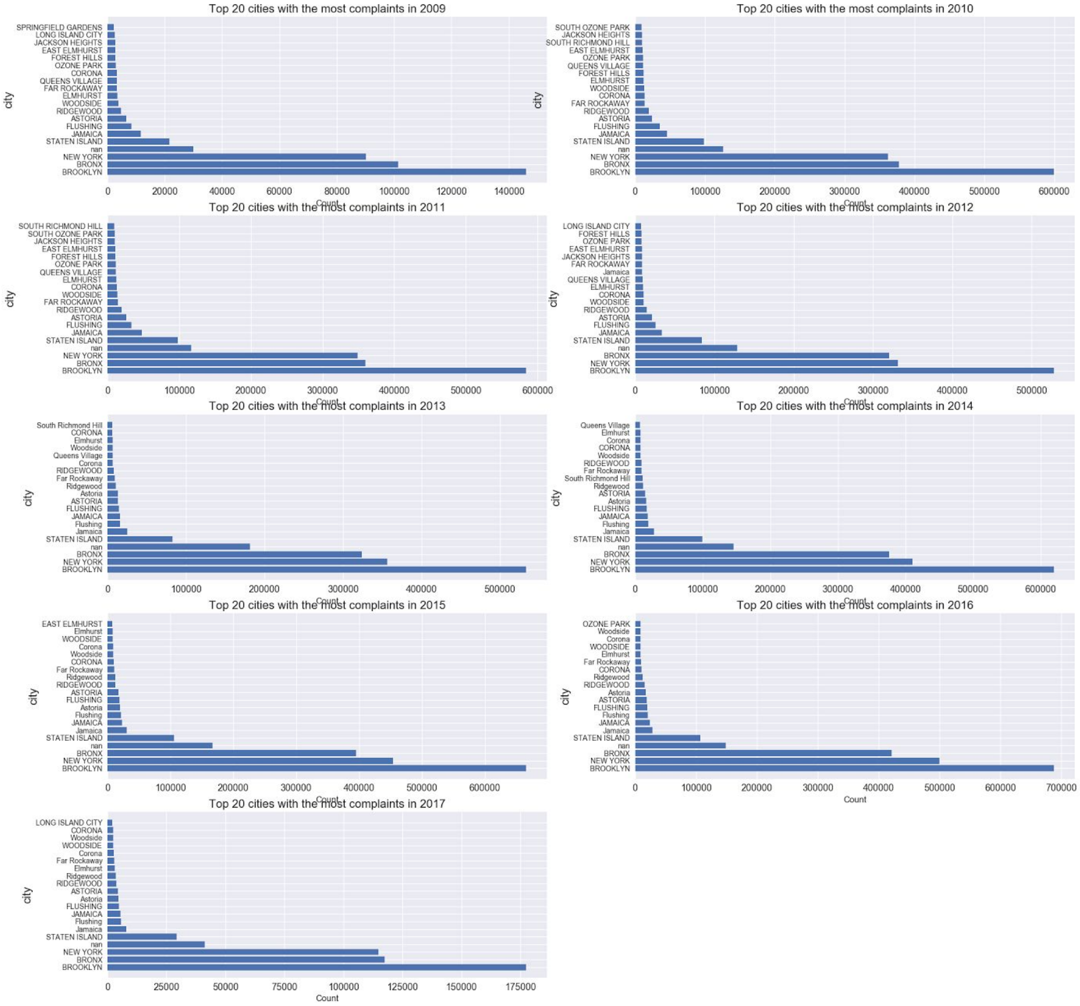
2.8 Yearly change in number of complaints in top 20 zipcode



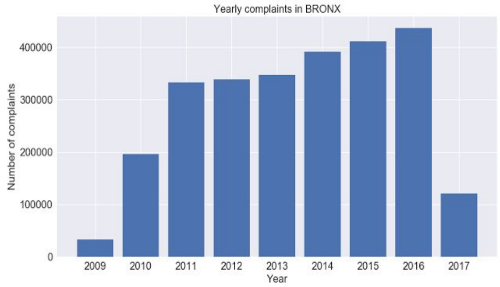
2.9 Top 20 cities with the most complaints from 2009 to 2017

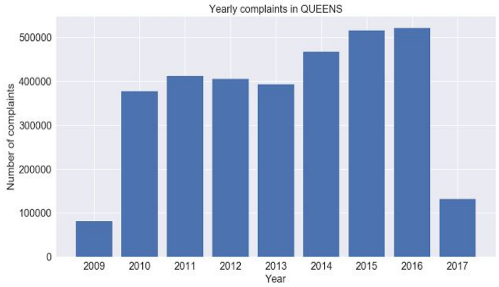


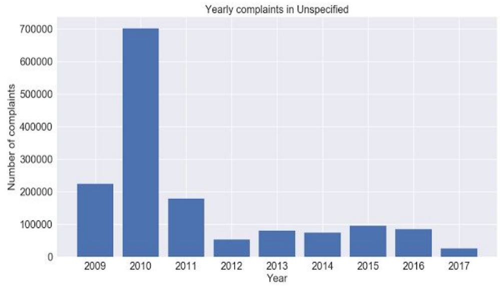
2.10 Yearly top 20 cities with the most complaints

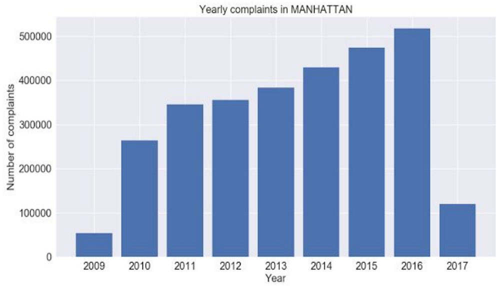


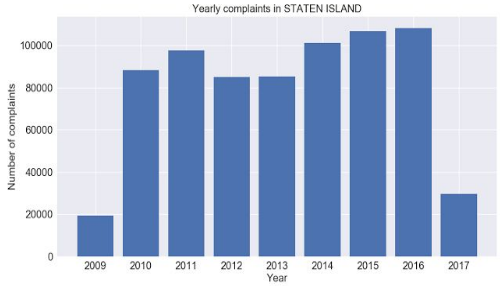
2.11 Yearly complaints filed in each borough

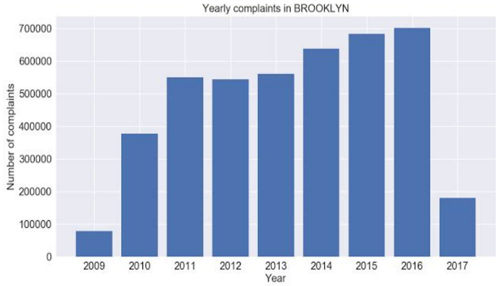




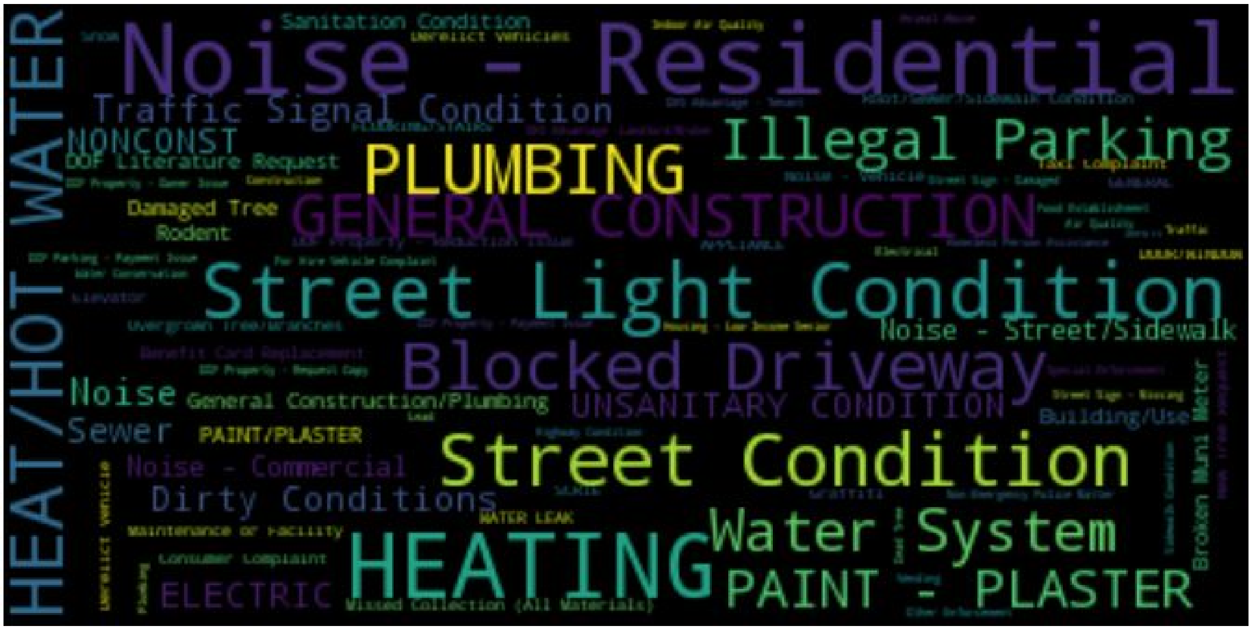








2.12 **Word cloud built from complaint types and its count**



For better visualization of complaint types, word cloud is built by getting the frequency of complain types. We see some common, i.e. more frequent complaints such as plumbing, noise, heating, heat/hot water issues. As we see heating and heat/hot water appearing more than other complaint types, it gives us some hints on possibility of correlation between temperature and number of complaints.

Part II Data Exploration

1 Complaints and Temperature

By looking at figure 2.2 in Part 1, we notice that numbers of complaints are not distributed evenly across different months. It seems that the months at the end and the start of the year usually have a large number of complaints compared to other months. And from Figure 2.3 in Part 1, we see that there is large variance among the number of complaints in each day. Days with extremely high number of complaints also seems to cluster around the same time period. Those extreme values can have a large effect on the number of complaints of the whole month, which might explain why those months have high complaints count. In New York, this happens to be the time when temperature is at the lowest point of the year. So we make a hypothesis that number of complaints is negatively correlated with temperature.

For this part, we count the number of complaints each day sorted in descending order on Hadoop. Then we used the output data count_day.out to do our analysis locally.

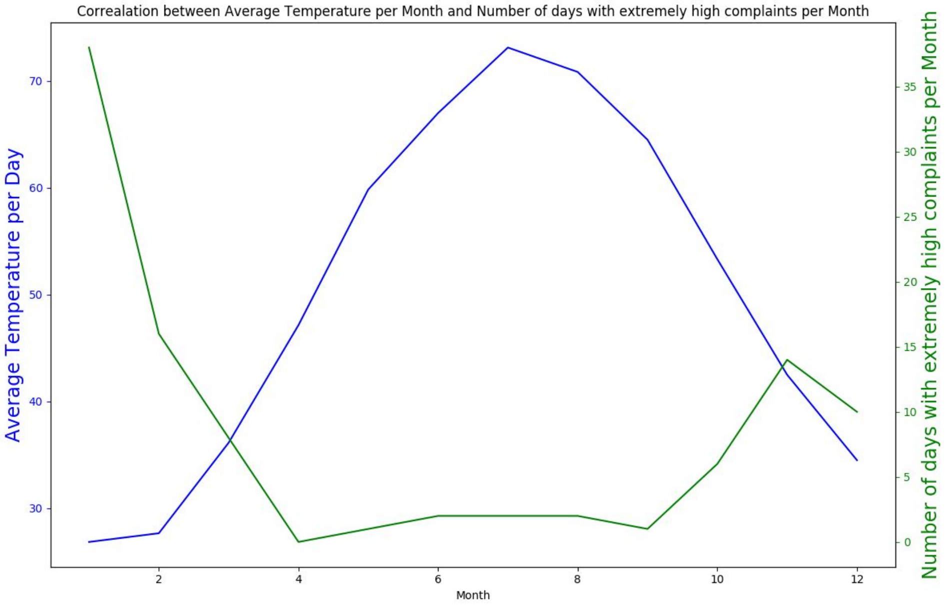
The daily weather data of the same time period is from http://www7.ncdc.noaa.gov/CDO/dataproduct.

For part 1.1, we first obtained the daily average temperature of all the weather station in NY state, and then group these temperature by date and year to get average temperature of NY state per month from 2010 Jan to 2017 Apr.

For part 1.2, we first obtained the daily average temperature for all the weather station in NY state, and then averaged those to represent the average temperature for the whole state for a specific day.

1.1 Count of Extreme Days at Month Level

We first look at this hypothesis at month level. We pick the top 100 counts as days with extreme values. Then we assign them to 12 months according to their date. For the weather data, we compute the average temperature of each month. A plot of count of extreme days per month and average temperature per month is shown below.

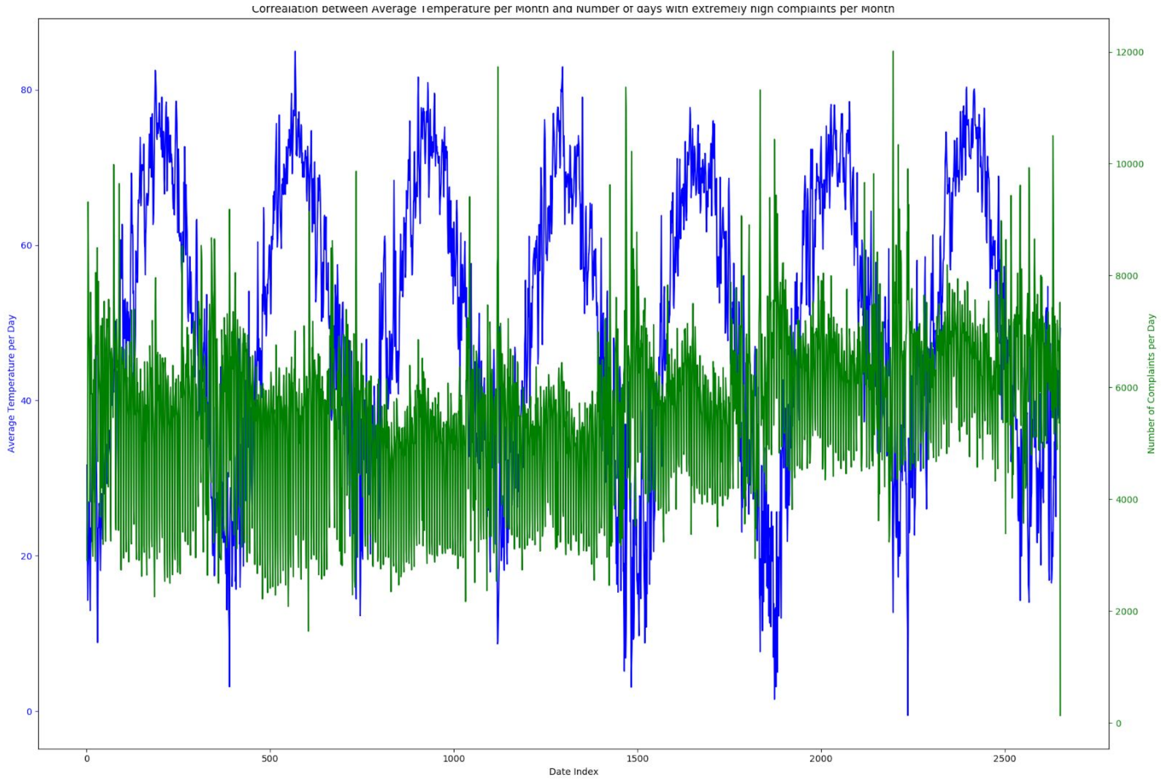


We can see pretty clearly that when temperature is low, the number of extreme days per month is pretty high, and vice versa.

We can further test their correlation using Pearson's Correlation. The result here is (count of days with extreme values per month, monthly average temperature, -0.7382). The p-value here is 0.0061. So they do have a significantly negative correlation and our hypothesis is true.

1.2 Daily Count and Temperature

Now we look directly at the count of complaints each day and its relationship with daily temperature. Their relationship is shown explicitly in the figure below.



Here the blue line is the temperature per day, and green line the count of complaints. Notice that the the peaks of counts seems to coincide with the troughs of weather, indicating a negative relationship between them.

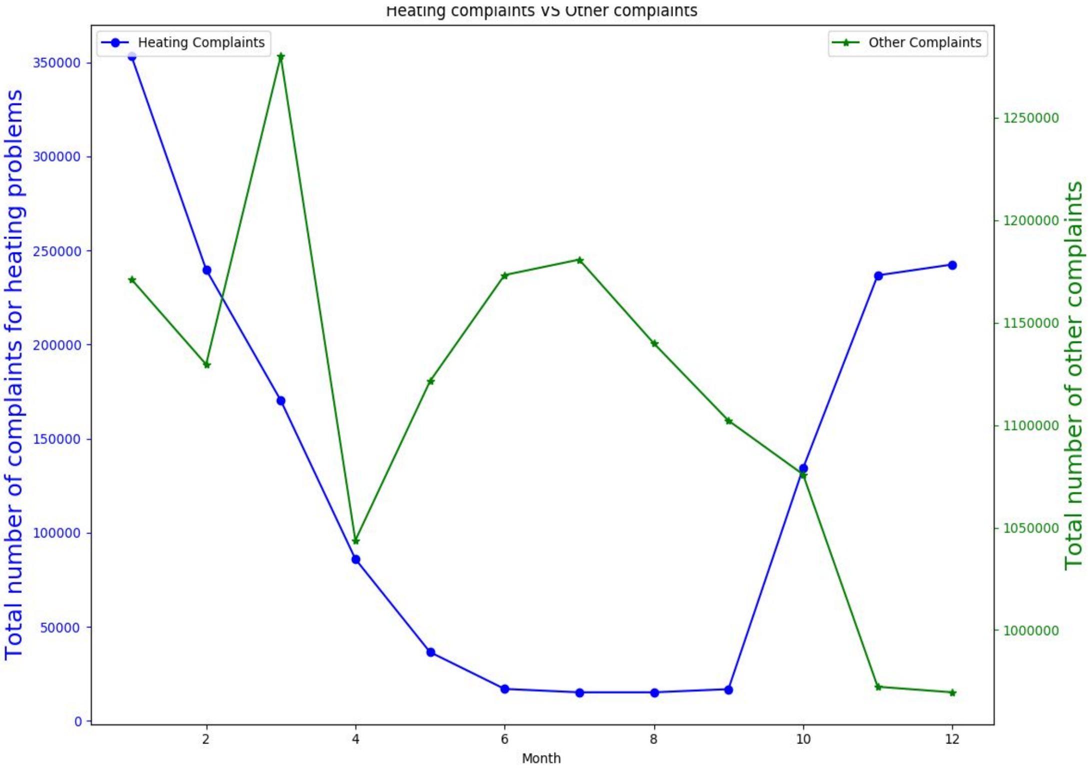
Again we test their correlation using Pearson's Correlation. The correlation here is -0.1982, with p-value 6.6e-25. So they have a very significantly negative correlation: (count of complaints each day, daily temperature, -0.1982). Since the p-value is extremely close to 0, our hypothesis is true.

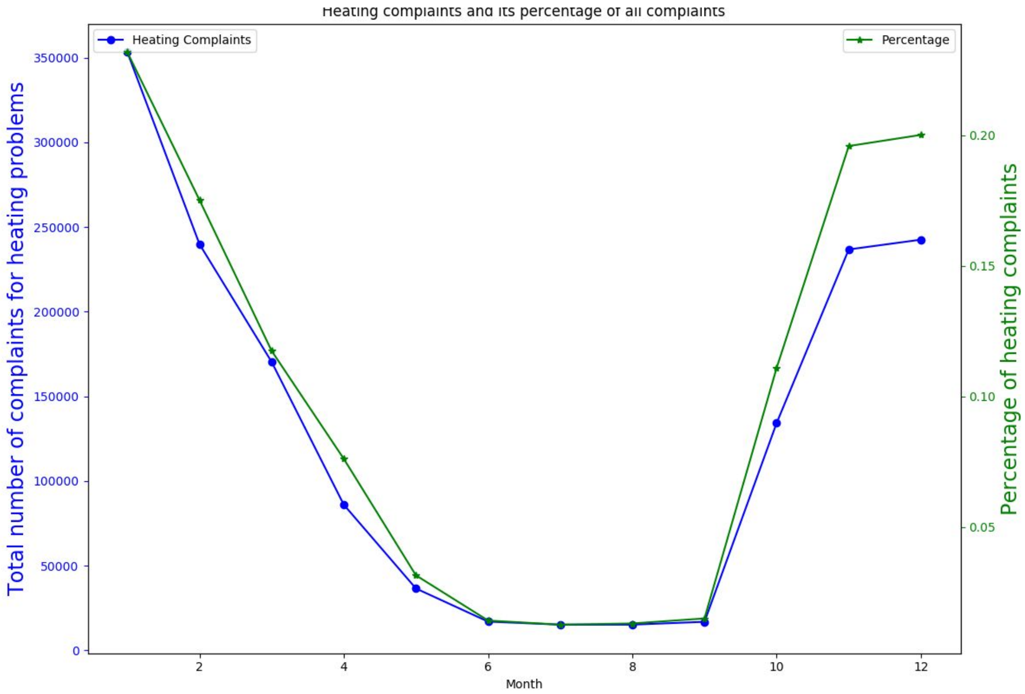
2 Seasonality of Complaints Type

The previous hypothesis leads us to think about the reason behind the correlation of number of complaints and temperature. If cold weather makes people complain more, it may be that some typical problems are only prominent when temperature is low. This leads to a new hypothesis that larger number of complaints in cold days is from the rise in heating problems, while other complaints remain stable.

To test this, we divide the complaints into 2 types: heating problems and all others. We then use mapreduce to calculate the count of the 2 types by month, averaged by the number of that month during the whole time span.

Below are 2 plots showing the result.





From the first figure we can see that heating complaints do have strong seasonality pattern. There is a leap of heating complaints in winter times, while in summer times the count is fairly low. From the second figure, we also notice that, the percentage of heating complaints rise and decrease simultaneously with its absolute count. This shows that the other complaints do not have strong seasonality, and that our hypothesis is true.

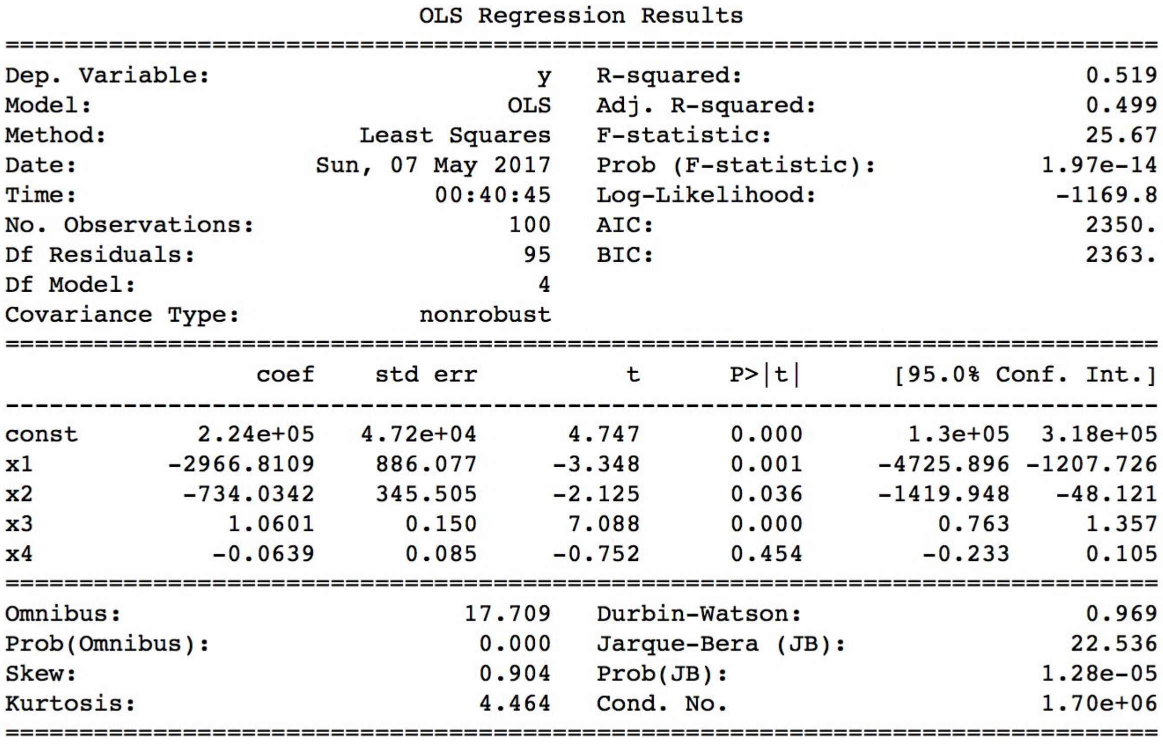
3 Complaints and Zipcode

From Figure 2.7 in part 1, the distribution of complaints varies across different zip code. The data for the plot is used for further analysis in this part. For simplicity, we pick the top 100 zipcode. Our basic assumption is that number of complaints is affected by demographic features within each region. To be specific, it is correlated with Median age, Sex ratio (males per 100 females), Mean income, and Population of each region.

The census data used was extracted from the following link:
https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

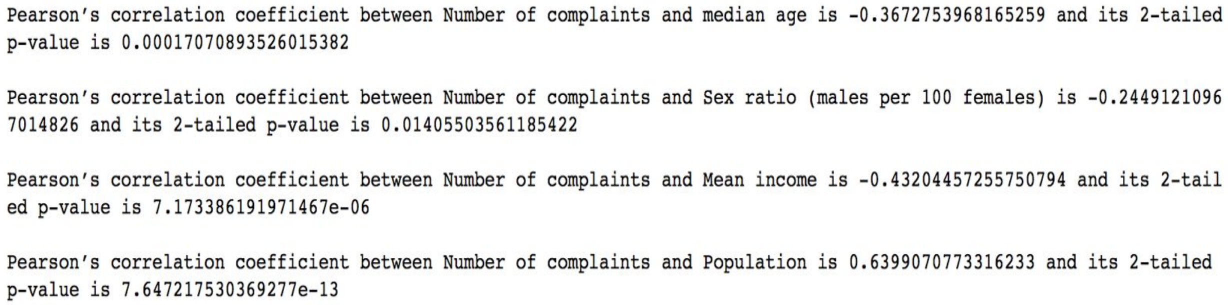
From the above website, we obtained the Median age, Sex ratio (males per 100 females), Mean income, and Population information of all zipcode for NY state. Then, we merged these four datasets by zipcode. After that, we kept rows that its zipcode is in our top 100 zipcode list for analysis. And then this dataset is merged with our counts of complaints by zipcode.

To explore their relationship, here we run an OLS regression. The dependent variable is the count of complaints, and the independent variables are Median age, Sex ratio (males per 100 females), Mean income, and Population. Below is the result of the regression.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.519
Model:                            OLS   Adj. R-squared:                  0.499
Method:                 Least Squares   F-statistic:                     25.67
Date:                Sun, 07 May 2017   Prob (F-statistic):           1.97e-14
Time:                        00:40:45   Log-Likelihood:                -1169.8
No. Observations:                 100   AIC:                             2350.
Df Residuals:                      95   BIC:                             2363.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        2.24e+05   4.72e+04      4.747      0.000        1.3e+05  3.18e+05
x1         -2966.8109    886.077     -3.348      0.001      -4725.896 -1207.726
x2          -734.0342    345.505     -2.125      0.036      -1419.948   -48.121
x3             1.0601      0.150      7.088      0.000          0.763     1.357
x4            -0.0639      0.085     -0.752      0.454         -0.233     0.105
==============================================================================
Omnibus:                       17.709   Durbin-Watson:                   0.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.536
Skew:                           0.904   Prob(JB):                     1.28e-05
Kurtosis:                       4.464   Cond. No.                     1.70e+06
==============================================================================
```

From the regression result, we see that these 4 variables are jointly significant (F - statistic close to 0). And in this regression, X1 (median age) is negatively correlated with Y(number of complaints), which is pretty significant, indicating that older people seem to complain less. X2 (sex ratio) is negatively correlated with Y(number of complaints), which shows that men might have a lower tendency of filing a complaint. This is significant at 5% level. X3 (Mean Income) is positively correlated with Y(number of complaints) and very significant. This might show that people with higher income tends to complain more. X4(population) is not significant here. However, none of the above results can be interpreted as causal effect. Also, there could be multicollinearity among the variables.



```
Pearson’s correlation coefficient between Number of complaints and median age is -0.3672753968165259 and its 2-tailed
p-value is 0.00017070893526015382

Pearson’s correlation coefficient between Number of complaints and Sex ratio (males per 100 females) is -0.2449121096
7014826 and its 2-tailed p-value is 0.01405503561185422

Pearson’s correlation coefficient between Number of complaints and Mean income is -0.43204457255750794 and its 2-tail
ed p-value is 7.173386191971467e-06

Pearson’s correlation coefficient between Number of complaints and Population is 0.6399070773316233 and its 2-tailed
p-value is 7.647217530369277e-13
```

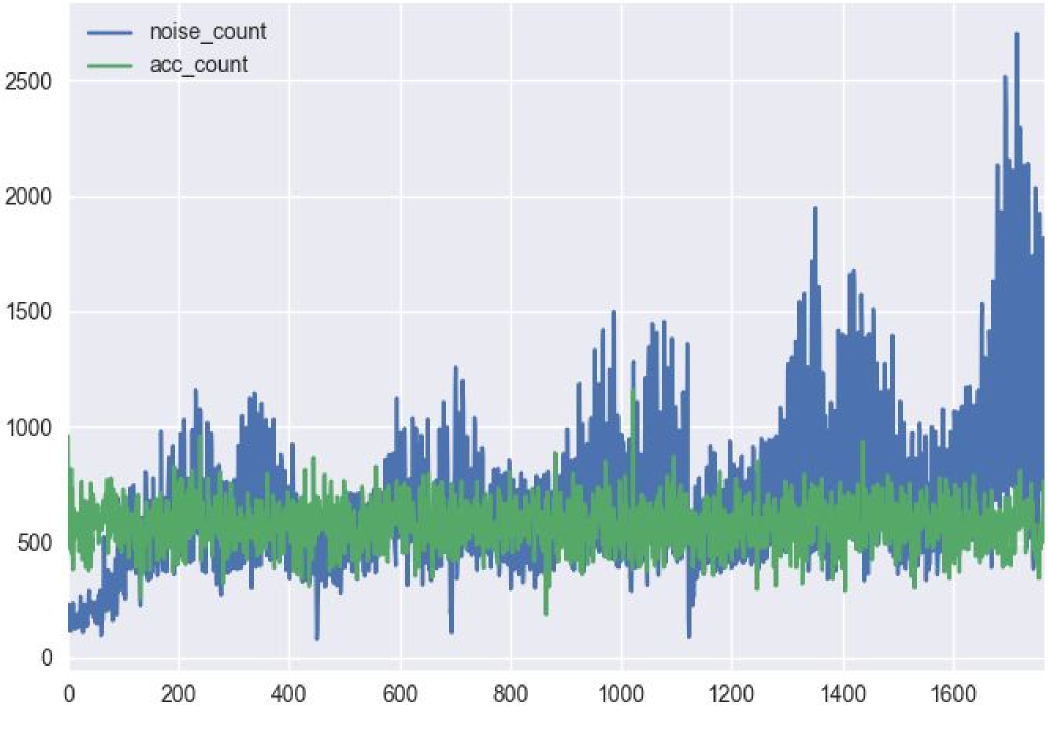
We can also calculate Pearson's Correlation between each pair. Our list of correlated attributes and their correlation and p-value are:

(Number of complaints, median age, -0.367275, 0.00017)
(Number of complaints, sex ratio (males per 100 females), -0.2449121, 0.014055)
(Number of complaints, Mean income, -0.4320, 7.17339e-06)
(Number of complaints, Population, 0.639907, 7.647218e-13)

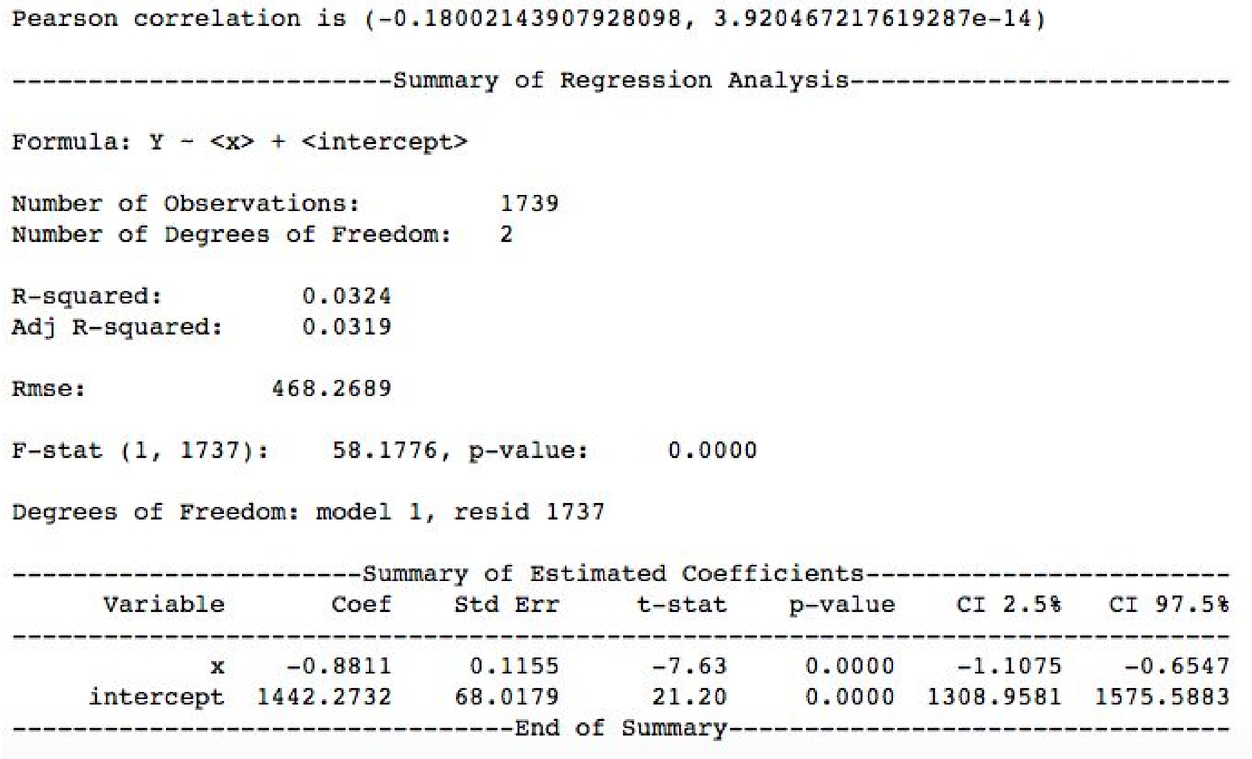
Considering the results, our hypothesis that number of complaints is correlated with median age, sex ratio, and mean income is true. However, we cannot be sure that number of complaints is indeed correlated with population, given the conflicting result given by 2 analysis tools.

4 NYPD Vehicle Accident

In this part, we are trying to see if there's any relationship between number of noise complaints that's filtered from 311 data and motor vehicle accident data provided from NYPD. Noise being one of the top 10 complaints as shown in our data analysis part, we are proposing that there is some correlation between noise and motor vehicle accidents. When accidents happen outside, it will create noise from the collision leading to ambulance and police cars coming with siren on and crowd building around that location. In order to test our hypothesis, we have filtered 311 data, extracting only the rows that's related to noise complain. We also downloaded data from https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95 for motor vehicle accidents. Accident dataset had starting date from July 2012, so when 311 data was merged to the accident data, earlier dates in 311 data were excluded. To test their relationship based on their daily count of complaints filed, we have plotted simple line plot of data points along the date range. Below graph shows the plots of daily count of noise(noise_count) and accident count(acc_count).



As we can see from the above plot, number of motor vehicle accidents tend to stay in the range with less variations, i.e. values are within certain range with less spikes over time. But noise count tend to vary with more frequent spikes over time. Now, to quantify their relationship, I have computed Pearson correlation coefficient and also ran OLS regression analysis as shown below.



```
Pearson correlation is (-0.18002143907928098, 3.920467217619287e-14)

-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <x> + <intercept>

Number of Observations:         1739
Number of Degrees of Freedom:   2

R-squared:         0.0324
Adj R-squared:     0.0319

Rmse:            468.2689

F-stat (1, 1737):    58.1776, p-value:     0.0000

Degrees of Freedom: model 1, resid 1737

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
             x    -0.8811     0.1155      -7.63     0.0000    -1.1075    -0.6547
     intercept  1442.2732    68.0179      21.20     0.0000  1308.9581  1575.5883
---------------------------------End of Summary---------------------------------
```

For the computation of the above, we have used Python Pandas built-in functions. Looking at Pearson correlation coefficient, it does not support our hypothesis proposed earlier as its value is found to be about -0.18. Our hypothesis was expecting some positive relationship between the two but Pearson correlation coefficient found to be negative and relationship is not strong enough to show correlation between the two.
(Number of noise complaints, Number of motor vehicle accidents, -0.18002, 3.9205e-14)

Analyzing the summary of regression analysis, the intercept value is overwhelming compared to the coefficient of x. X in this case is number of motor vehicle accidents and Y is the number of noise-related complaints. Here again, the line has negative slope and little variations in x won’t affect the y value as its intercept value is large. This is acceptable as the daily counts of noise complaints and accidents are in the units of hundred as shown in the very first graph in this part. So, looking at the result, the number of car accidents have not much effect on the number of complaints, noise complaints could be due to other factors.

Individual Contributions

Hezhi Wang: Part 1 - 1 Data Quality issues, Part 1 - 2.4, Part 2 - 1&3, writing report

Han Zhao: Part 1 - 1 Data Quality issues, Part 1 - 2.2, Part 2 - 1&2, writing report

Jin U Bak: Part 1 -Data Analysis 2.1, 2.3, 2.5-2.12, Part 2 - 4, writing report

Conclusions

From part 1, we can see that the most common data quality issues are missing data. Also, for columns with datetime or street address format, there are many rows that do not conform to the given format, or are out of valid range. These should be due to errors while inputting each 311 record.

Besides, from summary figures in part 1, some interesting patterns and trends can be observed. For example, day and month counts of complaints are not stable. Complaints are not distributed evenly across different zipcode or city. Noise and heating problems consists a large part of total complaints.

From our explorations in Part 2, we can conclude that, number of complaints is negatively correlated with temperature. Lower temperature can lead to more days with extremely large number of complaints. This negative correlation can be partly explained by the fact that, there is a huge leap in number of heating problems when it is cold, while other complaints remain stable without significant seasonality.

Further, complaints are not distributed evenly in space. The number of complaints in each district (identified by zipcode) is correlated with population, income, median age and sex ratio. The specific correlations between zip code and census data are mentioned above in part II 3.

Finally, we have looked at the relationship between noise-related complaints and NYPD motor vehicle accidents. With the assumption that car accidents will cause noise around the area with ambulance and police car siren and crowd gathering around the accident site, we conducted the experiment and have not found any significant relationship between the two. It is more likely that noise-related complaints are filed due the other factors such as dog barking and loud music in the neighborhood.

Throughout this experiment, we have used Python and Pyspark as the main tool for big data processing and data analysis & visualization. Even though we were dealing with millions rows of data, NYU HPC was able to process the data within few seconds with the default setting on the number of mappers and reducers. For reproducing the work we have done, this can be found on our Github repository(link provided in Reference section) and run on NYU HPC Dumbo cluster with default settings.

References

1. 311 Data set for 2009:
   https://data.cityofnewyork.us/Social-Services/new-311/9s88-aed8

2. 311 Dataset for 2010-present:
   https://data.cityofnewyork.us/Social-Services/311/wpe2-h2i5

3. Weather dataset from 2010-present:
   http://www7.ncdc.noaa.gov/CDO/dataproduct

4. Census dataset for NY state:
   https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

5. NYPD motor vehicle accident data:
   https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

6. Github Repository
   https://github.com/HezhiWang/Big_Data_project

7. Report on Google Docs
   https://drive.google.com/open?id=1EBsUfZedlkybdVxFeVuVaW3t6FK3X6R09_gyXXRlhAU