CS 457 - Data Science
An Analysis of Effectiveness of Government Emergency Services

Muhammad Hozefa Haider (mh05159)
Syed Hammad Ali (sa04324)
Habab Idrees (hi04031)

3rd December 2020

1 Introduction

This project will explore the data set provided by *The Mayor's Office of Data Analytics* (MODA) and the *Department of Information Technology and Telecommunications* (DoITT), open data for NYC. The 311 calls in New York City (NYC) are publicly available at Kaggle[1].

This data set comprises of all calls made to 311 from the year 2010-Present. Our analysis would also include merging the data set with the National Storms data set to compare the average response time for complaints during a storm and otherwise. The storm events data set for New York is available at NOAA[2].

2 Problem Statement and Rationale

The analysis of 311 calls can be of great use for a wide variety of purposes, ranging from a rich understanding of the status of a city to the effectiveness of the government services in addressing such calls.
In this analysis, we want to answer following questions:

- What are different type of Service Requests? Which is most/least frequent?
- From which borough most Service Requests come from?
- How air quality issues relate across the state?
- Which agencies are more efficient in solving Service Requests?
- Which Service Requests peaks at what time of year or time of day?
- From which type of location we get most number of complaints?

With the above answers, we will have a better understanding of the city's issue. Our next step will be to predict or find the following:

- Find out the time required in terms of range of days to resolve a specific complaint in a specific borough?
- Merge the 311 data set with the Storms data set and compare the average response time for complaints during a storm and otherwise.

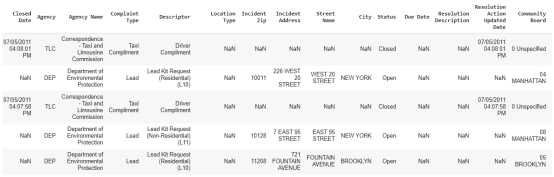
By using these answers, a city can be better prepared for a particular storm type. Policy makers can use this information to efficiently allocate resources. In addition, the residents of the city can have a real-time sense of when their problem will be solved.

3 Python Packages Used

- Numpy
- Pandas
- Matplotlib
- Scikit-learn
- Anaconda Environment

4 Data set Description

Our main focus will be on the 311 calls in New York City (NYC). 311 Service Requests encompass all non-emergency requests from the city, including but not limited to noise complaints, air quality issues and reports of unsanitary conditions etc.



Closed Date	Agency	Agency Name	Complaint Type	Descriptor	Location Type	Incident Zip	Incident Address	Street Name	City	Status	Due Date	Resolution Description	Resolution Action Updated Date	Community Board
07/05/2011 04:08:01 PM	TLC	Correspondence - Taxi and Limousine Commission	Taxi Compliment	Driver Compliment	NaN	NaN	NaN	NaN	NaN	Closed	NaN	NaN	07/05/2011 04:08:01 PM	0 Unspecified
NaN	DEP	Department of Environmental Protection	Lead	Lead Kit Request (Residential) (L10)	NaN	10011	226 WEST 20 STREET	WEST 20 STREET	NEW YORK	Open	NaN	NaN	NaN	04 MANHATTAN
07/05/2011 04:07:58 PM	TLC	Correspondence - Taxi and Limousine Commission	Taxi Compliment	Driver Compliment	NaN	NaN	NaN	NaN	NaN	Closed	NaN	NaN	07/05/2011 04:07:58 PM	0 Unspecified
NaN	DEP	Department of Environmental Protection	Lead	Lead Kit Request (Non-Residential) (L11)	NaN	10128	7 EAST 96 STREET	EAST 96 STREET	NEW YORK	Open	NaN	NaN	NaN	08 MANHATTAN
NaN	DEP	Department of Environmental Protection	Lead	Lead Kit Request (Residential) (L10)	NaN	11208	721 FOUNTAIN AVENUE	FOUNTAIN AVENUE	BROOKLYN	Open	NaN	NaN	NaN	05 BROOKLYN

Figure 1: NYC 311 Service Request 2011 Dataframe

The data contains more than 24.5M rows spread across 53 features. Size of the data is approximately 10 GB. For the scope of this project, we will be working over the data set of year 2011 only.
Fifty-three features of the data set include features related to Time such as `Created Date`, `Closed Date`, `Due Date`, and `Resolution Action Updated Date`. Location specific such as `Incident Zip`, `Incident Address`, `X-Coordinate` (State Plane), and `Y-Coordinate` (State Plane). Type such as `Complaint Type`, `Agency` and `Descriptor`. Then there are other features, which are there to support specific types of requests.

EVENT_ID	Storm_Borough	BEGIN_LOCATION	Created Date	BEGIN_TIME	EVENT_TYPE	MAGNITUDE	TOR_F_SCALE	DEATHS_DIRECT	INJURIES_DIRECT	DAMAGE_PROPERTY_NUM	DAMAGE_CROPS_NUM
277882	NORTHERN ONEIDA (ZONE)		2011-01-02	2200	Lake-Effect Snow			0	0	0	0
275983	SOUTHERN HERKIMER (ZONE)		2011-01-04	2000	Lake-Effect Snow			0	0	0	0
280866	OSWEGO (ZONE)		2011-01-04	2230	Lake-Effect Snow			0	0	30000	0
280077	NORTHERN ONEIDA (ZONE)		2011-01-05	100	Lake-Effect Snow			0	0	0	0
278443	MADISON (ZONE)		2011-01-05	1000	Lake-Effect Snow			0	0	0	0

Figure 2: NOAA Storm data set for NYC 2011

The storm data set contains features such as `Location`, `County`, `Date`, `Type`, `Magnitude` and few features telling about the damage caused by the event.

5 Methodology

5.1 Predicting time to resolve

This is a supervised learning, classification problem. The model we build for this problem contains the following:

Predicted Variables:

1. `Day of Week`
2. `Day of Month`
3. `Month`
4. `Incident Zip`
5. `Descriptor`

The first three features are stripped from the `Created Date` column of the 311 data set and `Descriptor` column contains text data so converted each unique text value into a feature for model.

Target Variable:
Resolution Time: - To calculate the resolution time in terms of days, we subtracted `created date` from `closed date`. Then divided the days into buckets such as bucket 1 for less than 2 days,bucket 2 for 2 to 6 days and bucket 3 for more than a week. Hence, our target variable represents three classes and we aim to classify input data into one of these classes. Therefore, for this problem we used the following classification models:

1. Logistic Regression
2. Decision Tree Classifier

Evaluation Metric:
We used confusion matrix for evaluation of the above model.

5.2 For merging Storms data set

We performed an inner join on Created date column between storms data set and 311 data set and performed some visual exploration to compare the average response time for complaints during a storm and otherwise.

6 Result and Analysis

By analyzing the NYC 311 Service request data set, we answered a few general but very important questions, that would later on be used for our data modeling. We found the counts of each complaint type to find the most and least Common complaints.

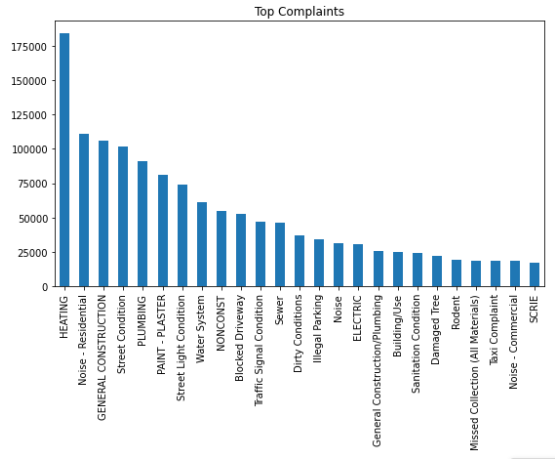


Figure 3: Most common complaints

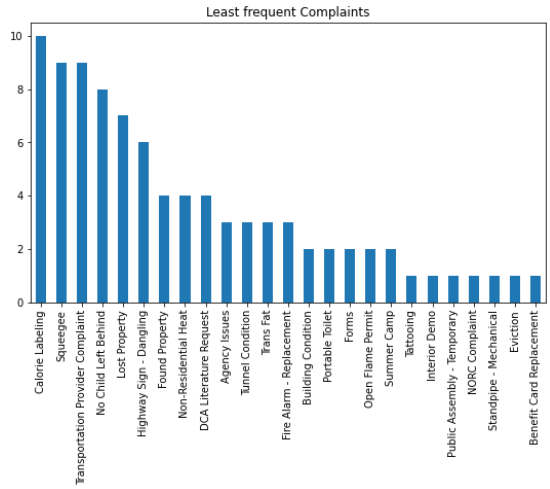


Figure 4: Least common complaints

We can see that the least frequent complaints are of calorie labeling while

the most frequent complaints are **Heating** and Residential Noise. We would now further work on only these complaint types as they are the most recurrent which means they are the most essential ones.

To have better understanding of how heating issue changes in New York, we analyze the heating issue concentration according to the months. We can see that heat issues are most common in the months of January, December and November, which are months of winter.

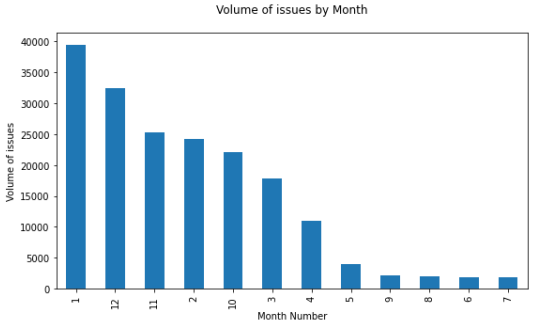


Figure 5: Volume of Heating issues in New York by months

Another major complaint is the air quality issue across NYC, so we explored the geography of spread of quality of air in New York. Since our data set provides us with Longitude and Latitude, we can scatter the complaints across them to study the issue.

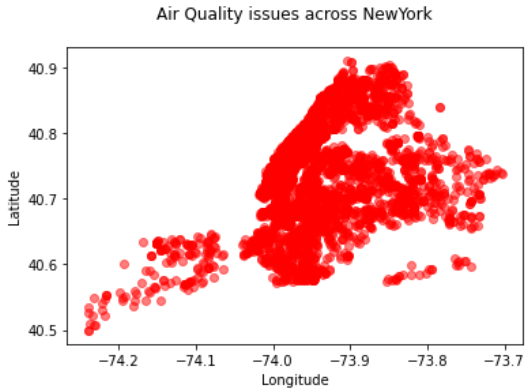


Figure 6: Air quality issues in New York

We also compared the number of complaints distributions across Boroughs so that we can focus on our desired one.

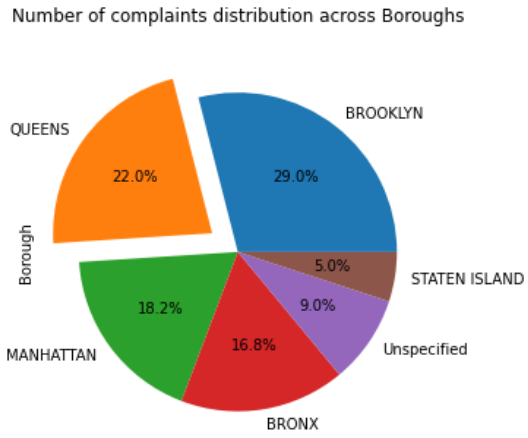


Figure 7: Number of Complaints across Boroughs

From now onwards, we will focus our study to Queens only. We will do some general analysis to find the most Frequent Complaints across Queens. Our analysis shows that Street Condition is the most common complaint type in Queens.

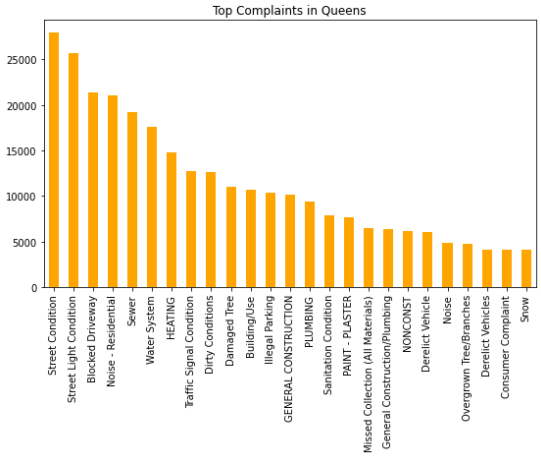


Figure 8: Most Frequent Complains in Queens

We also analyzed the average response time of complaints and checked how quickly the agencies reacted to a complaint call. Our analysis suggested that complaints like Residential noise and illegal parking are resolved within a day while issues like unsanitary condition take a rather long time to resolve. We also see that *NYPD* and *3-1-1* are the most efficient agencies among the rest. These

are also the agencies, which generally solve the most number of complaints.

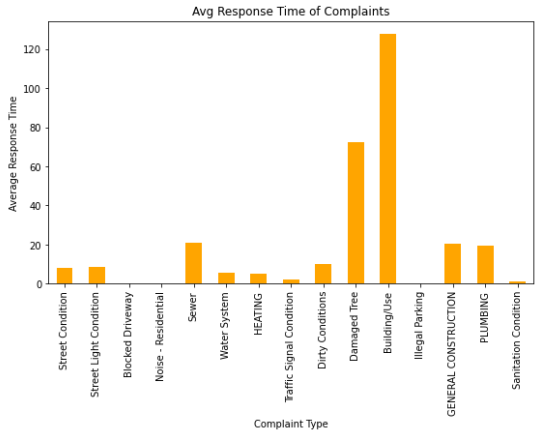


Figure 9: Average Response Time

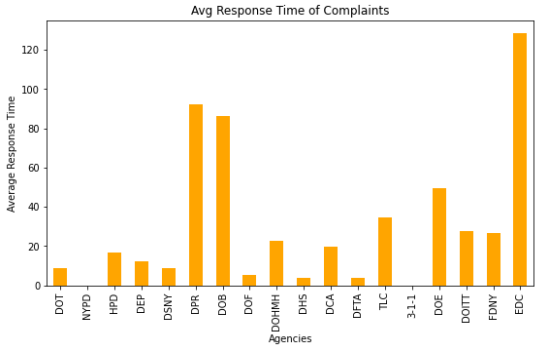


Figure 10: Average Response Time of Agencies

The above graphs tells us that most number of complaints in 2011 came from residential locations. We can relate that most common complaint type is **Street Condition** and these must be coming from residential buildings.
We will also use the Longitude and Latitude values to find the complaint concentration across Queens.

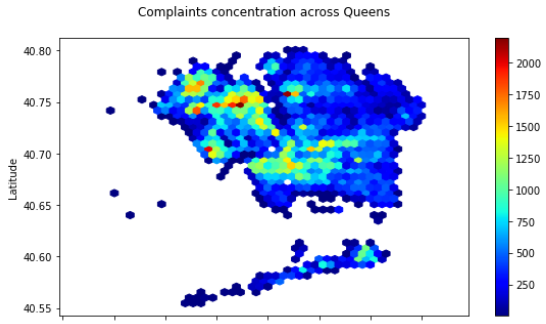


Figure 11: Complains concentration across Queens

We will now work over the most common issue i.e. **Street condition**. We will analyze the street condition issue concentration across Queens and then analyze how it changes according to the months. We can see that heat issues are most common in the months of Nov, Dec and Jan. Which are months of winter.

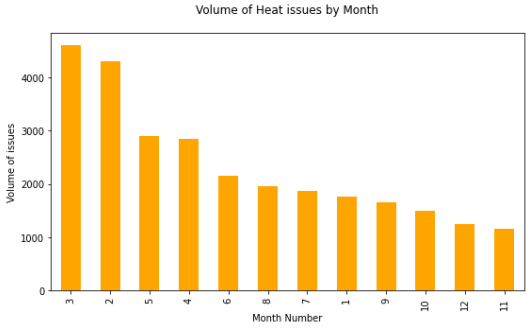


Figure 12: Volume of Street issues by months

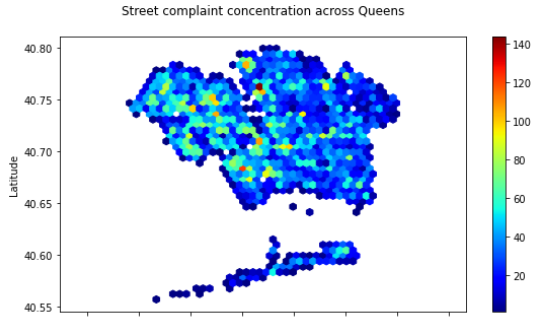


Figure 13: Street condition issues concentration in Queens

We will now predict the resolution time for Street condition complaints in Queens. In order to do this, we build a function for the data frame that extract the features for the model. The features chosen from the data set are based on intuition. The descriptor variable is broken into dummy variables so that it can be converted into categorical data. Similarly, Incident Zip column is manipulated for bringing down the scale of column.

	Incident Zip	Day of Week	Day of Month	Month	descriptor_Blocked - Construction	descriptor_Failed Street Repair	Resolution_Time
count	24915.000000	24915.000000	24915.000000	24915.000000	24915.000000	24915.000000	24915.000000
mean	376.165443	2.314469	15.157449	5.415894	0.009512	0.119486	6.524945
std	113.290810	1.698035	8.578125	3.222294	0.097068	0.324366	6.510029
min	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000
25%	361.000000	1.000000	8.000000	3.000000	0.000000	0.000000	1.000000
50%	377.000000	2.000000	15.000000	5.000000	0.000000	0.000000	5.000000
75%	419.000000	4.000000	22.000000	8.000000	0.000000	0.000000	9.000000
max	696.000000	6.000000	31.000000	12.000000	1.000000	1.000000	28.000000

Figure 14: Street Condition issues concentration in Queens

Our target variable has values ranging from 0 to 43. This can be a reason for the classifiers to perform poorly. We can try dividing our target variable Resolution Time into ranges. We can see from above that, Resolution Time's min value is 0, max value is 43 and mean is at 3.39. So we can divide it into following ranges: (0,2),(2,6),(6,31),(31,43). After binning our target variable we perform Logistic Regression over our data, for comparison purposes, we perform decision tree analysis over it too. Our Logistic Regression obtained an accuracy of 73.72% whilst the decision tree classifier obtains an accuracy of 89.94% on training dataset and an accuracy of 73.85% on training dataset .

```
Decision Tree Confusion Matrix:
  [[5146  905   10]
 [1215  921    2]
 [  13    5    5]]

Decision Tree Normalized Confusion Matrix:
  [[0.84903481 0.14931529 0.00164989]
 [0.56828812 0.43077643 0.00093545]
 [0.56521739 0.2173913  0.2173913 ]]
```

Figure 15: Decision Tree Confusion Matrices

```
Logistic Regression Confusion Matrix:
  [[6061    0    0]
 [2138    0    0]
 [  23    0    0]]

Logistic Regression Normalized Confusion Matrix:
  [[1. 0. 0.]
 [1. 0. 0.]
 [1. 0. 0.]]
```

Figure 16: Logistic Regression Confusion Matrices

We can see from the above confusion matrices that our models are doing well for the first two classes of output and the Decision Tree is more accurate. However, because of lack of data points for the third class, it is giving the wrong output for third class. We scale our data set and perform modelling again but do not see any significant improvement in the accuracy of the model.

Work on the National Centers for Environmental Information Storm data set for NYC

We merged the data set of NYC 311 service request and NOAA Storm data set for NYC of the same year.

	Unique Key	Created Date	Closed Date	Agency	Agency Name	Complaint Type	Descriptor	Location Type	Incident Zip	Incident Address	Street Name	City	Status	Due Date	Resolution Description	Resol / Up
0	19625909	2011-01-18	2011-01-25	HPD	Department of Housing Preservation and Develop...	HEATING	HEAT	RESIDENTIAL BUILDING	11233	246 SUMPTER STREET	SUMPTER STREET	BROOKLYN	Closed	NaN	The Department of Housing Preservation and Dev...	01/2 12
1	19625909	2011-01-18	2011-01-25	HPD	Department of Housing Preservation and Develop...	HEATING	HEAT	RESIDENTIAL BUILDING	11233	246 SUMPTER STREET	SUMPTER STREET	BROOKLYN	Closed	NaN	The Department of Housing Preservation and Dev...	01/2 12
2	19625909	2011-01-18	2011-01-25	HPD	Department of Housing Preservation and Develop...	HEATING	HEAT	RESIDENTIAL BUILDING	11233	246 SUMPTER STREET	SUMPTER STREET	BROOKLYN	Closed	NaN	The Department of Housing Preservation and Dev...	01/2 12

Figure 17: Merged data set of NYC 311 service request and NOAA Storm data set

We will now perform general yet very useful analysis to find the complaint distribution, most frequent complaints during a storm, Number of complaints vs Storm Types etc.

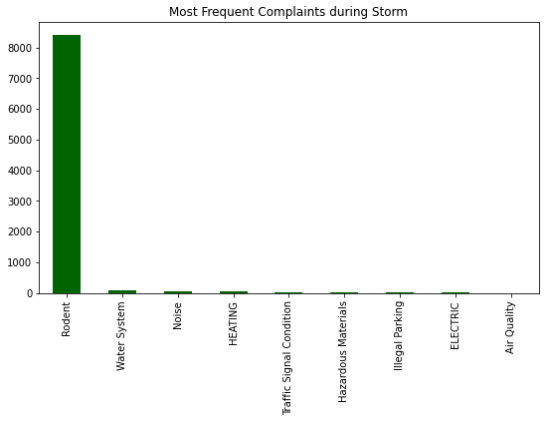


Figure 18: Most frequent complains during Storm

We can see that during a storm the most frequent complaint changes, it is no longer street condition, which makes sense, its Rodents. When a storm comes due to over flooding of sewerage pipeline, rodents come out and hence they become the biggest complaint issue. We can see by our analysis that most of the complaints are recorded during Thunderstorm and Flash Floods. This also proves why the most frequent complaint during a storm is about Rodents.

We perform a better analysis comparing the response time of complaints during storms and otherwise. We can see that the response time of standing water and Rodent increases since there are more complaints of them.

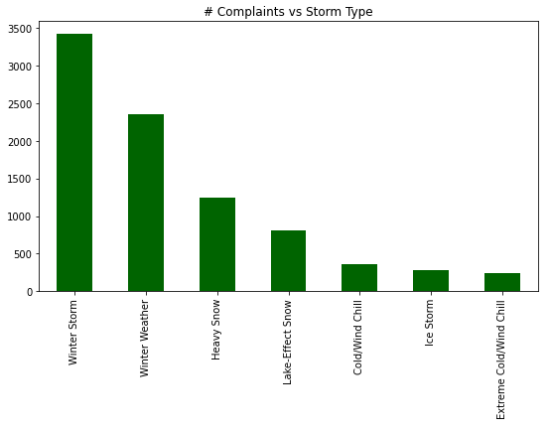


Figure 19: Comparison of Response Time during storm and otherwise

One last thing, we analyzed the response time of common complaints during storm and otherwise.

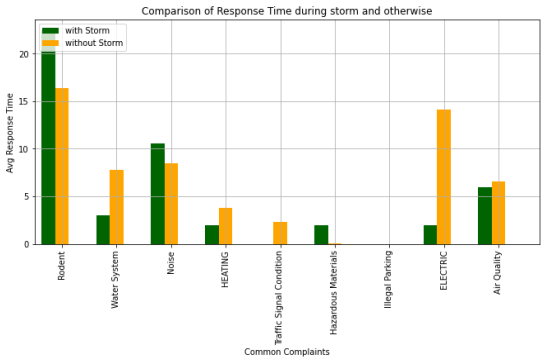


Figure 20: Comparison of Response Time during storm and otherwise

And, we can see that the average response time for rodent problems increases during storm times, however; the electric complaints are solved on top priority hence decreasing their response time.

7 Assumptions and Limitations

Our data set was very large and it was difficult to process the whole data due to limited RAM and other hardware limitations. So, we filtered the data on many levels to reduce the data size and performed all the analysis. We assumed that all the complaints were in closed state and that the impact of storm only lasted until the duration of that storm.
Nevertheless, a large-scale analysis can be performed by using the above methodology on a much larger data set and we can get more accurate results.

References

[1] Kaggle. *311 Service Requests from 2010 to Present*. Available at `https://www.kaggle.com/nidhirastogi/311-service-requests-from-2010-to-present?select=311_Service_Requests_from_2010_to_Present.csv`.

[2] National Centers for Environmental Information. *Storm Events Database(For New York Only)*. Available at`https://www.ncdc.noaa.gov/stormevents/`.