Methods of Advanced Data Engineering

An Analysis on the Effectiveness of Government Emergency Services

Syed Hammad Ali - 23401440

28 November 2024

1 Introduction

New York City, with its diverse population and dense urban environment, faces a wide range of challenges that impact its residents on a daily basis. From housing and public health issues to infrastructure concerns and public safety, the problems residents encounter are varied and often complex. Equally important is how effectively the city's government services respond to these challenges, and whether they are prepared to handle crises, particularly during times of naturally adverse conditions like hurricanes, heatwaves, or severe storms.

Understanding the distribution of these issues, the efficiency of the government's responses, and the level of preparedness for natural disasters is critical in assessing the city's resilience and capacity to support its residents.

This analysis answers the will explore the types of issues most commonly reported by NYC residents, evaluate the performance of government services in addressing these issues, and examine how well-equipped the city is to manage emergencies in the face of increasingly unpredictable climate conditions.

2 Problem Statement and Rationale

Our analysis is primarily aimed to answer the following question; *"What is the distribution of types of issues faced by the residents of NYC, what is the efficiency of the government response services in handling them, and how well prepared are these services prepared for response in times of naturally adverse conditions?"*

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 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 area do most Service Requests come from?
- Which type of issues are more common?
- 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 do we get the most number of complaints?
- What is the time required to resolve specific complaints in various areas?

With the above answers, we will better understand the dynamics of the city's issues. Our next step will be to find the following merge the 311 data set with the Storms data set and compare the average response time for complaints during a storm and otherwise.

Through these findings and analysis, a city can be better prepared for particular storm conditions. Policy-makers can use this information to allocate resources efficiently and with the help of data-driven insights know what to expect and how to overcome past shortcomings. In addition, the residents of the city can have a transparent overview of their government services' performance and a real-time sense of when to expect for their problems to be solved.

3 Data Sets

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 data set** publicly available at NYC OpenData[1].

This data set comprises of all calls made to 311 from the year 2010 - Present. The dataset contains 311 calls data regarding issues in the city, location of the report, response time, and other details. For our project, we will be taking the data of the previous 5 years only. The data contains around 13M rows spread across 41 features. Size of the data is approximately 7 GB. For the scope of this project, we will be working over the data set of years 2019-2024 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. The dataset is an open-data source and there are no restrictions on the use of Open Data. The overview and terms of use for the data are present here[2]

The second data set that we will use is the **National Storms Events data set**. We will compare the average response time for complaints during a storm and otherwise. The storm events data set is available at NOAA[3].

Storm Data is an official publication of the National Oceanic and Atmospheric Administration (NOAA) which documents the occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, significant property damage, and/or disruption to commerce. The data contains around 180 rows spread across 38 features. The storm data set contains features such as `Location`, `County`, `Date`, `Type`, `Magnitude` and few features telling about the damage caused by the event.

All data acquired in the dataset is formally dedicated to the public domain via the Creative Commons 1.0 Universal Public Domain Dedication (CC0-1.0), which removes all copyright from the data so that it may be used by anyone, for any purpose. The details can be found on their data licensing page[4].

4 Methodology

We used the following Python Packages for the pipeline.

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

For 311 calls dataset we export the data set and perform some basic data cleaning operations such as looking at the columns. We drop some columns such as road directions and taxi boroughs. The final dataset contains 37 features to be used in the analysis. We remove some null values from the data and perform the necessary data type conversion on numerical, date-time, and boolean columns. We also put in some sanity data checks such as to ensure that the created date is not later than the resolved that so that we can have accurate representations in our data.

We also create some new columns such as day of the week and month, month, year by stripping these values from the `Created Date` column. Our target variable in the analysis is the Resolution Time which we calculate in terms of days by subtracting the `created date` from `closed date`.

The Storms data was downloaded separately for each borough and then combined using the concat function in Python. Some data cleaning including checking for nulls and data types was done. We also renamed some columns for better understanding and making the merge easier.

Lastly, we performed an Left Join on the prepared 311 calls and Storm datasets to create the final data set to be used in the project.

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] NYC OpenData. *Overview and Terms of Use*. Available at`https://opendata.cityofnewyork.us/overview/`.

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

[4] NOAA. *Data Licensing in OCS*. Available at`https://nauticalcharts.noaa.gov/data/data-licensing.html`.