**Clustering New York’s Boroughs based on the Citizens’ Service Requests**

Andrea Marie L. Maderazo

February 2021

1. **Introduction** 
   1. **Background**

New York is the most populous city in the United States of America with an estimated population of 8.419 million as per 2019. With a large population, the importance of the emergency hotline (911) is emphasized with an average of 4200 calls per day. To address the concerns of the citizens that are not classified as emergencies of outmost urgency, the service request hotline (311) is open for them. Using this hotline citizens of New York can contact local authorities and departments to address and solve their concerns and problems such as noise complaints, water service complaints, and a like.

* 1. **Problem**

With the various types of services and departments that the 311 service requests call addresses, it will be beneficial if the local government unit of New York can group the five boroughs of New York according to the most frequent service requests that are addressed by the different departments. This capstone project will cluster the boroughs according to the ten most frequent services.

* 1. **Interest**

This would be helpful for the local government of New York in order to anticipate the services that are needed to be rendered. The Planning and Development department of New York will find the clustered boroughs for planning out future department and to prepare for services that might be needed for each area.

1. **Data Acquisition and Cleaning**
   1. **Data Sources**

The service requests data is downloaded from [NYC Open Data](https://opendata.cityofnewyork.us/) in the format .csv. While the other geojson data were gathered from Foursquare API. The dataset for the service requests were limited to a year worth of call because of the dataset size. There are a total of 1048574 service request calls with 52 columns that logs the details of the call. Only the information regarding the service type and source of call were kept. To reduce the size of the data that was being used on python, the service request data was pre-processed by droping the columns that were not needed.

* 1. **Data Cleaning**

The data cleaning was performed using Jupyter notebook using python this involved inspecting the rows and columns of data. Using the unique caller key the dataset was tested for any duplicated values.

* 1. **Feature Selection**

From the dataset of the service requests, there are a total of 51 features that detail the recorded call. This includes the agency that addressed the call, date of call, closed date, complaint type, details, location type, zip code, and address details. For this capstone, the features that was extracted from this dataset was the complaint type and the features that details the address of the call. After creating the new dataset, the dataframe was saved using the python library pickle. This was performed to make it easier to run the code due to the size of the dataset. Removing the features that were not to be included and keeping only the important ones for the project significantly reduced the size of the dataframe.

The final dataframe for clustering included the address details and service request type for each row of recorded call. The departments and other information regarding the call that were removed are not included in the scope of this capstone.

1. **Exploratory Data Analysis**

This part of the capstone focused on analyzing and deriving the insights from the data. This exploratory data analysis is centered on the service requests calls and the boroughs of New York. The figure below and Table 1 shows the pie chart of the percentage of service requests calls from each Boroughs with 30% of the calls coming from Queens and the least count of calls from Staten Island.

**Chart, pie chart

Description automatically generated**

Figure 1. Pie chart of the sources of service request calls

Table 1. The total count of service requests calls for each boroughs of New York.

|  |  |  |
| --- | --- | --- |
| **Borough** | **Total Count of Service Request Calls** | **Percentage** |
| Queens | 195164 | 30% |
| Brooklyn | 187361 | 29% |
| Manhattan | 140978 | 21% |
| Bronx | 88976 | 14% |
| Staten Island | 44144 | 7% |

The service request call is a categorical feature with 137 types and figure 2 shows the bar plot of the ten types with the highest occurrences. The ten types of service request calls with the highest count are street related concerns, water related concerns, traffic signal condition, construction, blocked driveways, dirty conditions, building/use, noise complaint, sanitation, and noise- street/sidewalk. The most common 311 calls in New York are the street related concerns with 151803 and followed by water related concerns with 67965 calls.

**Chart, bar chart

Description automatically generated**

Figure 2. The count of the occurrences of types of service request calls

Table 2. The total count of the ten most frequent service request calls

|  |  |
| --- | --- |
| **Service Request Type** | **Count** |
| Street Related Concerns | 151803 |
| Water Related Concerns | 67965 |
| Traffic Signal Condition | 46197 |
| Construction | 34097 |
| Blocked Driveway | 28943 |
| Dirty Conditions | 22996 |
| Building/Use | 19787 |
| Noise Complaint | 19432 |
| Sanitation | 18770 |
| Noise-Street/Sidewalk | 15524 |

1. **Clustering of Boroughs**

The machine learning algorithm K mean clustering was used to group the boroughs of New York according to the service request type. This algorithm will group the boroughs based on the service request calls. The elbow method was used to determine the optimal value of k. As show in figure 3, the graph of the sum of squared distances between the data points and three will be used as k.

Figure 3, The plot of the sum of squares vs the value of k.

**Chart, line chart

Description automatically generated**

1. **Conclusions**

Table 3 details the clustered boroughs with the service request type to be anticipated by the local government unit and its agencies. From the K means algorithm, there are a total of three clusters for the 5 boroughs of New York. The first cluster with the boroughs of Bronx and Staten Island, the second cluster with Queens and Brooklyn, and the third cluster with Manhattan. For Bronx and Staten Island, the following service request types should be anticipated: street related concerns, water related concerns, traffic signal condition, blocked driveway, dirty conditions, construction, graffiti, noise-street/sidewalk, sanitation condition, and rodent. For Brooklyn and Queens, the most common service request types call that occur are street related concerns, water related concerns, traffic signal condition, blocked driveway, construction, sanitation condition, dirty conditions, derelict vehicles, building/use, and noise complaint. And for Manhattan, the most common service request types address the following concerns: street related concerns, traffic signal condition, noise complaint, water related concerns, taxi complaint, construction, broken meter, noise-commercial, noise-street/sidewalk, and noise-vehicle. And figure 4, illustrates the New York map according to the clusters of the service type pf the calls.

Table 3. The clusters of the New York boroughs according to service request calls.

**Table

Description automatically generated**

Figure 4. Map of the clustered New York of Boroughs.

**Chart, map

Description automatically generated**

1. **Future Direction**

For future capstone of the same topic or improvements to the current map, the boroughs of New York can be analyzed and clustered with the data on the departments and agencies that address the issue. Another thing that can be added is a map with the current placement of the agencies involved to have a map where future agencies can be planned for construction. Since this capstone focuses only on service requests calls of New York Boroughs due to the data restrictions in relation to privacy, specific details such as neighborhood or streets of the call would be insightful.