Project Proposal
311 Service Requests Analysis
DS-GA-1007 Programming for Data Science

Pan Ding pd878
Liwen Tian tl1759
Maha Yaqub my1288

Introduction:

Big Data is now being used all over the world to carry out extensive analytics in order to come up with more ways to benefit the society. New York City has adopted the notion of collecting data at a very wide scale and has one of the largest databases of open data. The NYC open data portal has collects its datasets from all five boroughs which gives an elaborate source to users to get valuable data in order to develop meaningful projects that can benefit the City. Hence, taking advantage of this resource, we have decided to use the '311 Complaint Data' for the purpose of our project.

Objectives and Goals:

We inspire to build an interactive system that individuals can use to analyze the 311 complaint data in more depth via charts, graphs and maps. Python is a powerful tool that can be used to manipulate and handle big datasets. The deliverable of this project will be a system that can enable the users to make an attribute selection and then output an analysis of the complaints according to these selections. The users will be able to visualize the relationships between the number of complaints, ZIP Codes and City Agencies. For the prototype we will be using the data for the year 2014.

Our main goal is to make full use of the Python packages Pandas, Numpy and Matplotlib. Pandas will be used to organize and clean up the dataset. Pandas will also be used to create a robust dataframe which can be used to make appropriate selections. Numpy will be used to make matrices and arrays that will further be used to create informative visualizations using Matplotlib.

The end product will be a tool for city agencies, businesses and individuals to get a broad scope of the distribution of the complaints. This will help the city allocate resources according to where they are needed the most.

Datasets

The dataset being used for this project will be the NYC open data. We will be using the 311 Service Requests of 2014 for the purpose of the project. The methodology will be implemented using all the python packages that are required to be used. Pandas will be used to clean the data and make it well-structured so that it can be read with ease and make the project more interactive with an organized way to select features. Numpy arrays will be further used to help plot the data. Matplotlib will be used to plot maps representing number of complaints according to zip code. We aim to enable the user to be able to select the criteria with which they want to view the complaints. This could include bar charts that show the number of complaints per agency or a time series showing the distribution of complaints for a specific agency.

Matplotlib will also be used to make bar charts representing the number of complaints according to boroughs and agencies that is the number of complaints per borough for a selected agency. This can be interpreted in the following ways:

- Top n number of agencies by bar charts
- Time series: For each borough/each selected agencies for comparisons/selected time interval

Deliverables

This project will consist of an interactive Graphical User Interface as a tool to analyze the NYC Complaint Data. It will allow the users to select the mentioned features and visualize the data of interest. The city agencies can benefit from it by identifying which areas have a higher number of complaints which can help them consider where there is a need of more assistance and how to redistribute their resources. The citizens can also benefit from it by gaining more information about the neighborhoods which can help them make more thoughtful living arrangement .

Another aim of our project will be to make sure that all exceptions are considered and errors are given to the user if the input or selection made is not valid. This will make the project more robust and practical.