Final Project Proposal – IST 652

Crimes in Boston

Project Group: Individual – Vedant D. Patil

Data Set Used:

Crime incident reports are provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond. This is a dataset containing records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Records in the new system begin from Aug 2015 to Sept 2018. I have used individual datasets for each year. Therefore, I have the 4 different datasets for each each year from 2015 until 2018 and used that data to analyze various factors regarding crimes that needs to be understood to reduce crimes in future. Also, apart from the main 4 datasets, I have used a supporting dataset that depicts which offence code belong to what type of crimes through a supporting dataset.

Dataset Source: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

Topic of Investigation

We can analyze the dataset in various ways which can definitely help us to solve the problems associated with crime. Therefore, we can analyze the datasets by answering and analyzing some of the below questions such as:

What types of crimes are most common?

Where are different types of crimes most likely to occur?

Does the frequency of crimes change over the day? Week? Year?

Which areas of Boston are safe to live in and which areas are least safe?

At what time of the day most crimes occur?

Data Exploration and Data Analysis

Data Collection/Gathering:

To read the data from the Boston Crimes website into our python environment we can pull the data directly from the website through web scrapping or we can download the CSV and combine the four datasets into one dataset and then read those CSV files using read_csv() methods.

Data Merging:

Since the dataset consists of data from 5 different tables, the data needs to be merged and arranged according in a proper tabular form using various data merging methods.

Data Converting/Data Transformation:

Once we have the data into the python environment, it s very important that we process the data into the required format, eliminate all the factors that cause biased data and eliminate introduced redundancy in the data that need to be analyzed.

Before analyze, it is very important that I clean the data and bring it to the business requirements using various below methods.

Data Cleaning Techniques That We Can Put Into Practice Right Away

Remove duplicates.

Remove irrelevant data.

Standardize capitalization.

Convert data type.

Clear formatting.

Fix errors.

Language translation.

Handle missing values.

Exploratory Data Analysis

After we have cleaned the data, we can proceed with the Exploratory Data Analysis wherein we can perform data analysis to find out meaningful story and information from the data. In this dataset, I will analyze the dataset so that I can find the appropriate data for the above topic of investigation. Along with the data in tabular format, we can also visualize the data in graphical format to further understand and analyze the data in deatil. The analysis is performed on multiple structured and semi-structured datasets.

Potential Development Task:

The Boston dataset can also be used to find out the correlation between two or more factors that can impact the rate at which crimes are increasing or decreasing. Hence, if I proceed with the prediction at a basic level, I might need further understanding and code of predictive analysis and determine the correlation factor.

Additionally, the dataset is an excellent source for analyzing data with the help of graphs, pie charts, and other visualization methods. Therefore, I might need help understanding the appropriate usage of various visualization package and their execution through code.

Last but not the least, if I try to pull multiple datasets directly from the website, I might need some additional guidance about web scrapping and merging the multiple datasets through python code.