Data Mining Project - Spring 2022
(Boston Crime Analysis)

Introduction

The given dataset contains about 567,000 incident reports of crimes in Boston, USA starting from 2015 to 2022. 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 in June of 2015.

The dataset can be downloaded from the following link:
https://drive.google.com/file/d/16D8GBCj5iOn3GGo_oWNmw1K3_BFS0t_7/view?usp=sharing

Details of the dataset fields are as follows:

Column Name	Description
INCIDENT_NUMBER	Unique ID for each incident
OFFENSE_CODE	Code assigned to a offense
OFFENSE_CODE_GROUP	Group assigned to a particular offense
OFFENSE_DESCRIPTION	Description of the offense
DISTRICT	District where the offense took place
REPORTING_AREA	Area code where offense took place
SHOOTING	Whether shooting took place in the incident. 1 and Y represent that shooting took place while 0 and empty values represent shooting did not take place
OCCURRED_ON_DATE	Date on which the incident occurred
YEAR	Year during which the incident occurred
MONTH	Month during which the incident took place
DAY_OF_WEEK	Day of week on which the incident took place
HOUR	Hour at which the incident took place
STREET	Street where the indecent took place
LAT	Latitude of the location where the indecent took place
LONG	Longitude of the location where the indecent took place
LOCATION	GPS coordinates of the location where the incident took place

Deliverables

There are going to be a total of three deliverables for this project. The details of each deliverable is mentioned below.

Deliverable 1: Pre-processing & Exploratory Data Analysis (EDA) [40%]:

This deliverable is primarily focused on extracting as getting your hands dirty with the dataset. This will consists of data cleaning / preprocessing, initial data exploration, visualizations etc. This could include but not limited to the following points:

- Change types of columns as per requirements
- Find and handle missing values or incomplete rows
- Correlation between attributes
- Most frequently occurring incident
- District wise analysis based on number of incidents
- Using maps to visualize incidents

These are just some suggestions of tasks you can do. You need to figure out what other useful information can you figure out from this data.

You need to elaborate your technique along with reasoning for each task you perform in your report.

Deliverable 2: Clustering and Frequent Pattern Mining [40%]:

- Perform cluster analyses on the data, primarily on location and offense type. Other clustering could be done based on time of day etc.
- Find frequent patterns in the dataset and explain your findings

Your report for this deliverable must justify your choice of clustering algorithm.

NOTE: You are free to implement your own algorithm or use any library to perform selected tasks.

Deliverable 3: Final Report [20%]:

In this deliverable, you will be incorporating feedback from the previous deliverables and presenting your findings. You need to explain what information you have extracted from the data and suggest ways to help the police department overcome the rising crime. In your report, also mention other information, if any, that could be extracted from the data and how you plan to do that.

Administrivia

- Use jupyter notebook for this project (Google Colab or local). For each deliverable you will be submitting the notebook (ipynb) and HTML file of the notebook along with a report in PDF format.

- For each deliverable, submit a zip folder containing all the files. Only one member of the group should be submitting all deliverables.

- Zip file should be named as GroupNumber_ProjectName.zip

- For each deliverable, one group member per deliverable has to appear for a viva. Every member has to appear for the viva at least one.