Crime rate analysis for the city of Los Angeles

Authors: J V N Ajay Raj: ajayrajv@iisc.ac.in, Nithin Poovanna: pnithin@iisc.ac.in, Sandhya K M: sandhyakm@iisc.ac.in, Gayathri Ramasubramanian: rgayathri@iisc.ac.in

Problem:

Definition

This study aims to analyze crime rate in a particular state to figure out types of crimes, value of crimes while understanding trends. We seek to provide actionable insights on the data to enhance crime prevention efforts and public safety.

Problem motivation

In recent years, urban areas have experienced an uptick in the crime rates. This trend not only poses a threat to the safety and well-being of residents but also strained law enforcement agencies and resources. Hence, there is a need to analyze the crime data to understand the contributing factors.

Our research aims to address this issue by a data centric approach. By identifying trends, patterns and potential factors of crime, we seek to provide evidence-based insights than can help in interventions and policy decisions.

Goal would be to reduce crimes and contribute to a safer and more secure future for the residents while also helping to guide policymakers, law enforcement agencies and NGO's in allocating resources effectively.

Design Goals

Methodology

While downloading the data from gov.in and understanding the data. Data cleaning followed by data munging using data processing tools.

Once data munging is completed, prepare patterns by forming relationships across various features.
With this analysis we hope to identify hotspots and understand the sentimental analysis of culprit.

Scope

Focus on the dataset for the city of Los Angeles from a time frame of 2020 to 2023 with a total crime on a daily basis. Most of the dataset contains criminal cases committed in the city. The dataset is publicly accessible from the data.gov

Features supported

Data collection

- First step is to extract data manually from the below link.
    - Crime Data from 2020 to Present - Catalog
- Upon extraction of the data, we move the data to Google drive.
- Using data processing tools, we plan to clean and prepare the data for further processing.

Visualization

Using python libraries, matplot, seaborn we plan to generate charts for the below points while generating further insights.

- Classify the crime based on people, property or public domain.
- The average age of victim in the central region
- Rank the Area with the highest recorded crimes.
- 10AM to 5 AM identify the burglaries happened where the age of victim >60

https://github.com/poovannanithin/DA-231-Data-Engineering-Project

Crime rate analysis for the city of Los Angeles

Authors: J V N Ajay Raj: ajayrajv@iisc.ac.in, Nithin Poovanna: pnithin@iisc.ac.in, Sandhya K M: sandhyakm@iisc.ac.in, Gayathri Ramasubramanian: rgayathri@iisc.ac.in

Scalability

Create standard formats and standard procedure which can help in the scalability for data of other urban cities.

Approach

High level design

By following the high-level design of data flow, we plan to implement the model.

- Google drive – Raw data storage
- Python(pandas & NumPy) – Data cleaning
- Pyspark – Read + processing
- Python(with additional libraries matplot & seaborn) – visualization
- ML model for sentimental analysis

Big data platforms used

- Google collab & Spark/python

Data sources and data models

- Crime Data from 2020 to Present - Catalog
    - Data to be downloaded in a csv format and then used across the analysis.
- Data frames, schemas, partitions and joins using Pyspark.

Evaluation approach

Experiment plan

- Validate data types, rows and columns of the input file, error detection

Performance metric for success

- Time series model: Since this data contains the daily crimes, this model should handle the temporal data efficiently allowing us to track changes and patterns in crime rates.
- Geospatial analysis: Identifying hotspots and patterns along with the analysis of data.

Features metrics for success

- Accuracy: We plan to monitor data such as missing data, data consistency and data accuracy
- Data availability and uptime, Data recovery time & Error handling

Summary

Success

- Classified the crime based on people, property or public domain
    - Crime rate was higher for Males with total number of crimes at 3,05,325 and against Females 2,92,753

- The average age of victim in the central region
  - Victims around 40-43 age are mostly targeted for crime

- Rank the Area with the highest recorded crimes.
  - Central region has the highest number of crimes with crime rate at 53,909 reported cases stands number for crimes in LA
  - Hollenbeck has the lowest number of crimes with 30134 cases standing at 20th rank

- 10AM to 5 AM identify the burglaries happened where the age of victim >60
  - Crime rate was higher for Senior citizens above 60 during day time with 9357 crimes reported and the number of cases reported at night was 1984 as compared to night time throughout the LA areas

Achieved design goals as per project proposals

Future Extensions

- Fine tune the Spark ML code to scale the model
- Resizing the current model to accommodate the complete historical data of US crimes, while ensuring seamless data processing without encountering any performance constraints
- Identify patterns of similar crimes across the country which would help in improving the Patrolling during the crime interval observed