NYPD Complaint Data Analysis (2013–2023)

Our team analyzed the NYPD Complaint Data, which contained over 8 million criminal complaint records from New York City from 2008 to 2023. We were particularly interested in this dataset because we were curious about crime trends in New York and the broader significance of public safety from 2013 to 2023.

Our analysis began with a comprehensive data cleaning process. We systematically removed 13 irrelevant columns and excluded rows containing more than three missing values to uphold the integrity of the dataset. This process resulted in a refined dataset of approximately 3.5 million records. Additionally, missing demographic data was imputed using modal values, and categorical variables were standardized to ensure consistency throughout the dataset.

We applied statistical methods and a variety of visualizations to extract meaningful insights. Key findings revealed the following:

- Crime peaked between 3 and 5 PM daily, with Fridays showing the highest complaints.
- A notable decline in crime was observed in 2020, which could be attributed to the COVID-19 pandemic.
- Geographically, Brooklyn and Manhattan accounted for most incidents, with residential areas being the most common crime locations.
- Individuals aged 25–44 made up the majority of both victims and suspects. Female victims were disproportionately affected across all crime types.
- Over 95% of reported crimes were completed, with attempted crimes decreasing from 1.95% in 2013 to 1.44% in 2023.

Throughout the project, we faced several technical challenges, particularly in handling the scale and complexity of large dataset. To address these issues, we optimized data processing using efficient Panda's operations, applied targeted data imputation strategies based on borough- and crime-specific mode values, and enhanced the clarity of our visualizations by fine-tuning parameters in our Folium heatmaps.

To explore predictive potential, we developed an XGBoost classifier to predict crime types based on available features. The model achieved an overall accuracy of 67% and a weighted F1-score of 0.63. It performed particularly well on frequent crime categories such as "Harassment 2" and "Petit Larceny," but struggled with less common categories like "Possession of Stolen Property" and "Unknown."

The findings have practical applications for law enforcement in optimizing patrol routes and staffing, urban planners in improving infrastructure in high-risk areas, public safety researchers

studying crime trends, and community organizations designing targeted crime prevention programs.

Task Distribution:

- **Ali Adnan**- Cleaning data, Visualization, Predictive Model
- **Shreya Karki** – Kaggle API, Cleaning data, Statistics, Visualization
- **Ayushman Shrestha** –Statistics, Visualization
- **Hongyan Zhang** – Statistics, Visualization

References

- Pandas Documentation: https://pandas.pydata.org/docs/
- Matplotlib Histograms: Official Documentation
- Seaborn Distribution Plots: Seaborn Tutorial
- Folium Heatmaps: https://python-visualization.github.io/folium/
- NYC Open Data Portal: https://opendata.cityofnewyork.us/
- XGBoost Official Documentation
- XGBoost in Python (towardsdatascience)
- Scikit-learn Metrics Guide
- https://thesai.org/Downloads/Volume15No1/Paper_23-Crime_Prediction_Model_using_Three_Classification_Techniques.pdf