**Analysis of Crime Occurrences**

**in Los Angeles dating back to 2020**

**1. Abstract (10 pts)**

This project examined the patterns in crime occurrences and predicted crime types based on demographic factors, location and time. Using crime data from 2020 to present, I conducted exploratory data analysis to find out patterns and relationships. The research used statistical summaries, visualization techniques and different machine learning models to identify crime types and evaluate predictive accuracy. Findings show that significant demographic and geographic trends can aid in developing targeted crime prevention strategies.

**2. Introduction (20 pts)**

Understanding crime patterns is essential for law enforcement agencies and making policies. This project discovers key patterns in crime data to answer the question: ***What are the key patterns in crime occurrences, and How can we predict crime types and identify whether crime reports are timely or delayed based on demographic factors, location, and times?*** The dataset contains specific records of crimes, including demographic information of victims, location and time of crime occurrences.

This project aims to analyze two sides: (1) to discover actionable information about crime trends, (2) develop predictive models classifying crime types effectively and identifying whether crime reports are timely or delayed. The report is structured as follows:

Firstly, the crime data set is introduced, and exploration data analysis is followed.

Next, the EDA process will be summarized, and specific information is included in the Rmd file

Finally, models will be listed, and specific information is included in the Rmd file

**3. Dataset (30 pts)**

This dataset contains information about crime incidents in the City of Los Angeles from 2020 to present, with columns such as crime descriptions, victim demographics, location coordinates and time of crime occurrences. The dataset was downloaded as a CSV file from a public repository provided by the City of Los Angeles, ensuring accessibility, accuracy, and reliability.

This dataset provides a comprehensive insight into crimes, with more than 20 attributes capturing important details. Features such as Victim Age, Descent, Sex is necessary in analyzing demographics, while attributes about Area, Longitude, Latitude, and Time occurs facilitates to explore spatial and temporal trends.

Link I got the data: [Dataset](https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data)

The original dataset has 28 columns and up to 987,000 rows, which is quite large. However, in the preprocessing, I reduced the dataset size by focusing on the top 3 crime types for Crm.Cd.Desc, removed some unnecessary columns and invalid rows for my project and I also added new columns based on other columns to improve the performance in the process of building models such as:

* Time.to.reports: to count how many days a crime occurrence took to report
* Delayed\_Report (1-0): to check whether a crime occurrence was delayed (>30 days) or not in reporting
* Time\_Slots\_Happening: to convert Date.Occ to time slots (midnight, morning, afternoon, evening)
* Weekdays\_of\_DateRptd: to transform Date.Occ to weekdays (Monday, Tuesday, etc).

Finally, I do this project based on the new dataset with 24 columns and 9,999 rows

**4. Exploratory Data Analysis (EDA) (60 pts)**

In the EDA process, I used techniques:

* To understand the target variable (Crm.Cd.Desc): ggplot2 library (geom\_bar).
* To understand categorical variables: chisq.test function, ggplot2 library (geom\_bar), functions in dplyr library.
* To understand continuous variables: ggplot2 library(geom\_density, geom\_smooth)
* T-test function.
* ANOVA test.

**5. Models (75 pts)**

I built different models:

* **SVM, Decision Tree** and **Random Forest** to Classify crime types based on Victim demographics and crime locations, and I did a comparison about the performance between these models.

*It is crucial to predict crime types based on Victim demographics and crime locations due to various reasons:*

* Enacting preventative measures: gaining insights related to time slots, weekdays can highlight high risk times
* Investigating effectively: police can investigate quickly patterns and focus on situations which have similar profiles
* Improving resource allocation: law enforcement agencies can allocate resources effectively by understanding insights into crime patterns
* **Logistic Regression** to Classify Time.to.reports as Delayed (>30 days) or Timely (<= 30days)

*It is necessary to identify whether a crime report is timely or delayed because of various reasons:*

* Allocating resources: Knowing the time of reports may help stakeholders allocate their resources effectively.
* Improving the function of law enforcement agencies: Reports timely allow faster investigation while delayed reports can reduce the probability of solving crimes because of the loss of necessary data.
* Preventing Crimes: Understanding factors causing delayed reports provides information for policies which encourage timely reporting.
* To evaluate its performance, I created confusion matrix and Summary metrics for each model
* *I got the highest accuracy of Random Forest model compared to Decision Tree and SVM due to the concept of Random Forest is Combining predictions from multiple decision trees, which decreases overfitting problems and increases good generalization of data. While Decision tree is just a single tree and, in this case, may cause overfitting because it might create too many branches, leading to the model learning the specific of the data train without generalizing the common rules of the dataset. With SVM model, even using radial kernel, SVM can still face challenges with complex data where the boundary is not easy to separate, which make it may still not capture the patterns as effectively as Random Forest does.*

**6. Summary of learning (5 pts)**

I feel like my professional skills have significantly improved after doing this project. I have learned much useful knowledge about the R language and the way to build machine learning models to predict the future which is necessary for the success of a business. Besides that, I also faced many challenges such as I did not know where I should start with this project, or finding a too large dataset made me frustrated because my laptop was unable to handle it. Moreover, choosing suitable models for the dataset and encountering errors while building models also took me a lot of time to deal with. But thanks to instructions from my professor, TA and friends, I gradually found out solutions to solve these issues.

**Project Development Directions**: For the further improvement of this project in the future, I am going to focus on incorporating additional variables such as details about neighborhood characteristics, or temporal trends such as seasonality, etc by combining with other datasets. Moreover, utilizing advanced machine learning models such as neural networks or gradient boosting can improve performance.