Los Angeles Crime Data : Statistical Analysis Plan

Group 13: Titus Robin Arun(tra29), Revanth Ganga(rg361), Suim Park(sp699), Meixiang Du(md480)

1. Data Overview

- The Dataset used originates from the crime records of the Los Angeles Police Department(LAPD) from 2020 up to present day. The data is created by transcribing original crime reports that are typed on paper and is updated on a weekly basis.
- The Dataset contains **820,599** observations, encompassing **28** variables as of 18-Oct-2023.
- Each row of the Dataset represents a crime reported in Los Angeles, and contains the following categories of information(with selected examples):
  1. **Location**: Latitude, Longitude, Area, Street, District
  2. **Victim Demographic**: Age, Gender, Ethnicity
  3. **Crime Description**: Type of Crime, Investigation Outcomes, Weapon Usage
  4. **Date and Time**: Date Reported, Date Occurred, Time Occurred
  5. **Identifier/Classifier**: Crime Record Identifier, Mocodes

2. Modeling

- **Research Question 1: What are the strongest indicating factors that influence the seriousness of crime committed.**
  - **A. Model Type**: Logistic Regression
  - **B. Research Question Type**: Inference
  - Given the inferential nature, we will use Priori variable selection and use confusion matrix, F1 score, Kappa value, and the area under the ROC curve to assess and fine-tune our model.
  - **C. List of Predictor Variables**
    - Independent Variables
      - Time and Date: Time Occurred, Date Occurred
      - Victim Demographic: Age, Sex, Race
      - Geographic Factor: Area, Distance to Precinct
      - Weapons Used
    - Outcome Variable
      - Crime Type: Serious Crime / Non-serious Crime
  - **D. Interaction Term**
    - Investigate any potential interaction between the day of the week and time of occurrence.
    - Examine the interaction between location and victim's race to understand if crime severity varies for different demographics based on locations.

Los Angeles Crime Data : Statistical Analysis Plan

Group 13: Titus Robin Arun(tra29), Revanth Ganga(rg361), Suim Park(sp699), Meixiang Du(md480)

- **Research Question 2: Likelihood of the number of the crime for a specific area in a specific time period is 10,000 (tentative number).**
    A. **Model Type**: Poisson Regression
    B. **Research Question Type**: Prediction
    - Given the prediction nature, we use deviance or pearson chi-squared tests, check for dispersion with residual analysis, and interpret coefficients as expected count changes.
    C. **List of Predictor Variables**
        - Independent Variables
            - Time and Date: Time Occurred, Date Occurred
            - Geographic Factor: Area
        - Outcome Variable
            - Count of Crimes (Generated by grouping the dataset by the selected predictor variables)
    D. **Interaction Term**
        - Examine any possible interaction between the day of week and date(e.g holidays) and the time of crime committed.

3. Potential Challenges

A. **Missingness**: The absence of data, particularly in victim demographic variables such as "age," "sex," and "decent," can introduce bias and significantly impact the model's performance. Depending on the extent of missingness and the type of variable., we will try to impute or drop the variable from the modeling to ensure the integrity and accuracy of the analysis.

B. **Data**: While we have sufficient size of Data (~800k Records) to use for the analysis, the dynamic nature of the Data (updated weekly) might possess a challenge as we may see new discrepancies which were not present earlier during EDA. To overcome this, we will be using the same Dataset we used for the EDA phase and not update it to the latest one at the time of final modeling.

C. **Categorical Variable Handling**: Certain categorical variables, such as crime description, area, comprise a substantial number of levels, potentially complicating the modeling process. To overcome this we will be combining them under suitable umbrella categories to reduce the number of levels and overfitting.

D. **Outliers**: Some variables such as the distance to precinct have outliers which may act as leverage points and affect the model. We will drop the outliers after investigation to reduce their impact.