Project Report

LA COUNTY CRIME RATE PREDICTION

Thant Thiri M. Kyi

Data Science Career Track (July-2023 Cohort), Springboard.

SCHOOL OF DATA
by Springboard

Table of Contents

1 Introduction

1.1 Background

Los Angeles is the second-most populous city in the United States and has a high crime rate. By providing a data-driven crime prediction model , policy makers and certain communities can be empowered to implement targeted security measures and make informed decisions.

1.2 Problem Statement

How can develop a predictive crime model for specific areas within Los Angeles Country using the past 12 years of historical crime data to help policy makers to enhance security, ensure the safety of locals, and assist the community in making informed decisions about the crime trend in certain areas.

1.3 Goal

The project aims to working on two major directions:

- Predicting average surges of crime rate trend based on the area, nature of crimes and victim characteristic.
- Understanding patterns that could help communities in making informed decisions.

2 Data Acquisition and Cleaning

2.1 Dataset

The dataset used was found on the Los Angeles city public data website (data.lacity.org). The data contains nearly 3 millions records (2993433) ranging from 2012 to 2023. The following variables are found in the dataset:

1. Crm Cd Desc -Defines the Crime Code provided.
2. Mocodes-Modus Operandi: Activities associated with the suspect in commission of the crime.
3. Vict Age
4. Vict Sex
    a. F - Female
    b. M - Male

c. X - Unknown

5. Vict Descent
   a. A - Other Asian
   b. B - Black
   c. C - Chinese
   d. D - Cambodian
   e. F - Filipino
   f. G - Guamanian
   g. H - Hispanic/Latin/Mexican
   h. I - American Indian/Alaskan Native
   i. J - Japanese
   j. K - Korean
   k. L - Laotian- Other
   l. P - Pacific Islander
   m. S - Samoan
   n. U - Hawaiian
   o. V - Vietnamese
   p. W - White
   q. X - Unknown
   r. Z - Asian Indian
6. Premis Cd -The type of structure, vehicle, or location where the crime took place.
7. Premis Desc - Defines the Premise Code provided.
8. Weapon Used Cd - The type of weapon used in the crime.
9. Weapon Desc - Defines the Weapon Used Code provided.
10. Status - Status of the case. (IC is the default)
11. Status Desc -Defines the Status Code provided.
12. Crm Cd 1 - Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
13. Crm Cd 2 -May contain a code for an additional crime, less serious than Crime Code 1.
14. Crm Cd 3 -May contain a code for an additional crime, less serious than Crime Code 1.
15. Crm Cd 4 -May contain a code for an additional crime, less serious than Crime Code 1.
16. LOCATION - Street address of crime incident rounded to the nearest hundred block to maintain anonymity.
17. Cross Street - Cross Street of rounded Address

18. LAT - Latitude
19. LON -Longitude

2.2 Data wrangling

The data wrangling process involved thorough cleaning, transformation, and premilitary data exploration, setting a solid foundation for deeper statistical exploration and data-driven insights into crime patterns in Los Angeles City.

The following steps performed at this stage:

1. Two Crime datasets for 2010-2019 and 2020-present were loaded from CSV files. Each dataset contained 28 columns with data types ranging from int64, object, to float64.
2. Converted column names to uppercase and removed spaces for standardisation.
3. Concatenated the two datasets vertically.
4. Renamed 'LAT' and 'LON' to 'latitude' and 'longitude'.
5. Filled missing values in 'VICT_SEX' and 'VICT_DESCENT' with 'Unknown'. Derived month, day, year, and weekday from the 'DATE_OCC' column. Categorized 'TIME_OCC' into 'AM' and 'PM'.
6. Removed columns not essential for initial crime analysis like 'MOCODES', 'DR_NO', etc.
7. Excluded data for the year 2023 for data consistency since 2023 is not ended yet.
8. Mapped codes to their corresponding descent descriptions for 'VICT_DESCENT' and 'VICT_SEX' .
9. Dropped duplicate records.
10. Sorted and reindexed the whole dataset.
11. Explored on some categorial features columns ('TIME_OCC', 'AREA_NAME', 'CRM_CD_DESC', 'VICT_SEX', 'VICT_DESCENT') and analysed crime frequency by area and weekdays.
12. The cleaned data contains 2,707,190 records and 18 columns.
    1. DATE_OCC
    2. TIME_OCC
    3. AREA
    4. AREA_NAME

5. CRM_CD
6. CRM_CD_DESC
7. VICT_AGE
8. VICT_SEX
9. VICT_DESCENT
10. latitude
11. longitude
12. MTH_OCC
13. DAY_OCC
14. YEAR_OCC
15. WEEKDAY_OCC_ID
16. WEEKDAY_OCC
17. TIME_OCC_TYPE_ID
18. TIME_OCC_TYPE

13. Finally, the cleaned data is saved as **'cleaned_crime_data.csv'** for further statistical analysis in the next step.

3 Exploratory Data Analysis (EDA)

As the 2nd step of the project , EDA is performed on the cleaned dataset file ,'cleaned_crime_data.csv', containing crime data from 2010 to 2022. Initial exploration included checking data types, descriptive statistics, null values, and duplicates.
Data Exploration focus was on various aspects like the occurrence timeframe(date/time), day of the week, crime category description, area occurred and victim characteristics.

3.1 Crime occurrence influenced by aspect of time

- Friday and Saturday recorded the highest number of incidents, while Sunday had the least (See Figure 1).

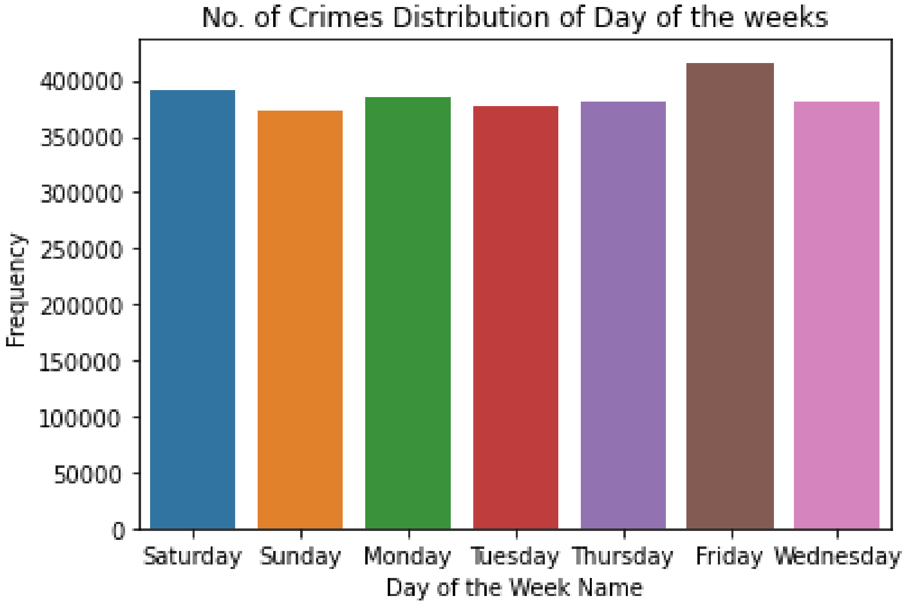
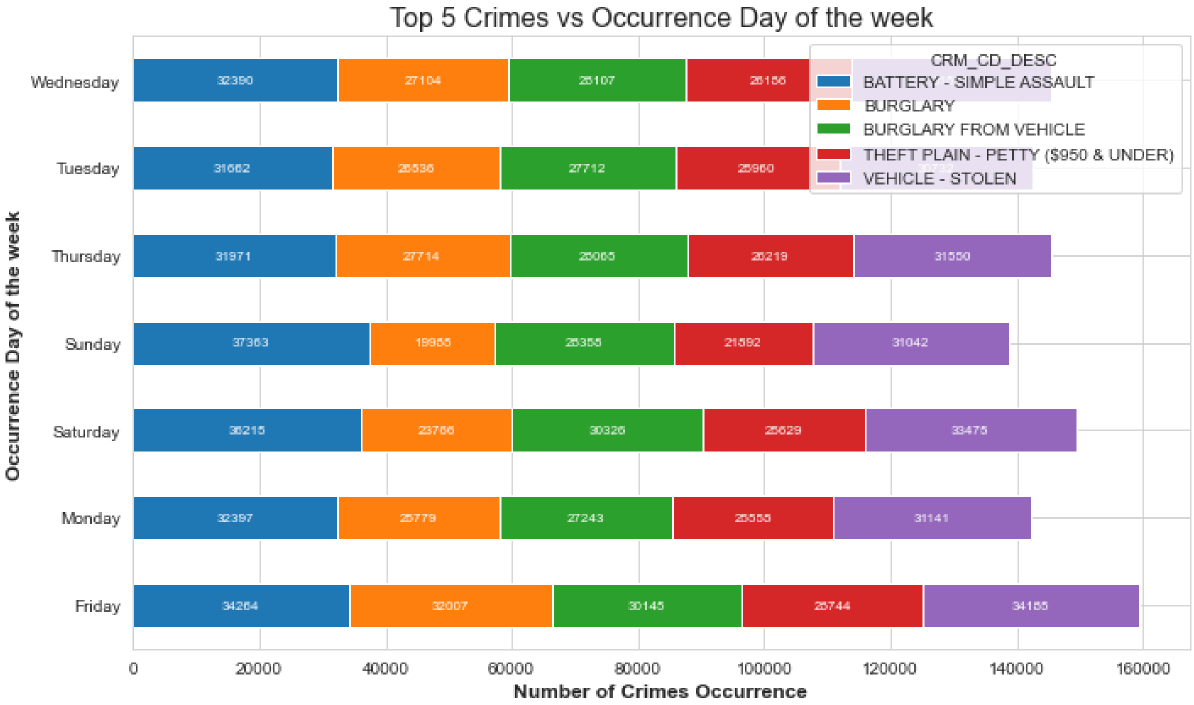


Figure 1



- Evenings saw the highest crime rates (See Figure 2).

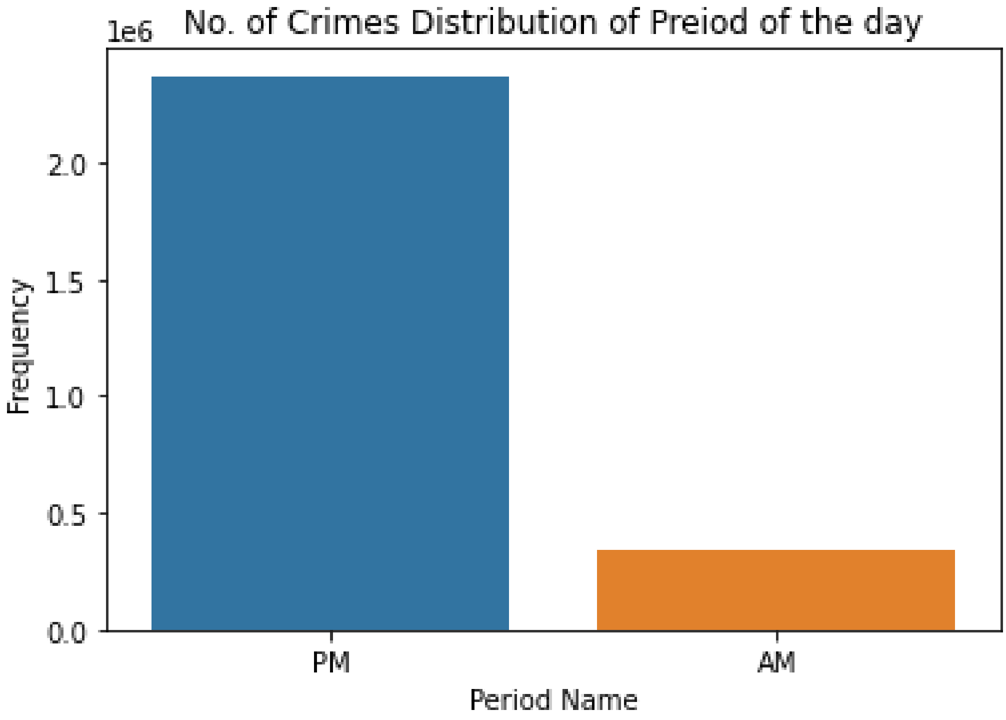


Figure 2

- January had the highest average crime rate over 12 years (See Figure 3).

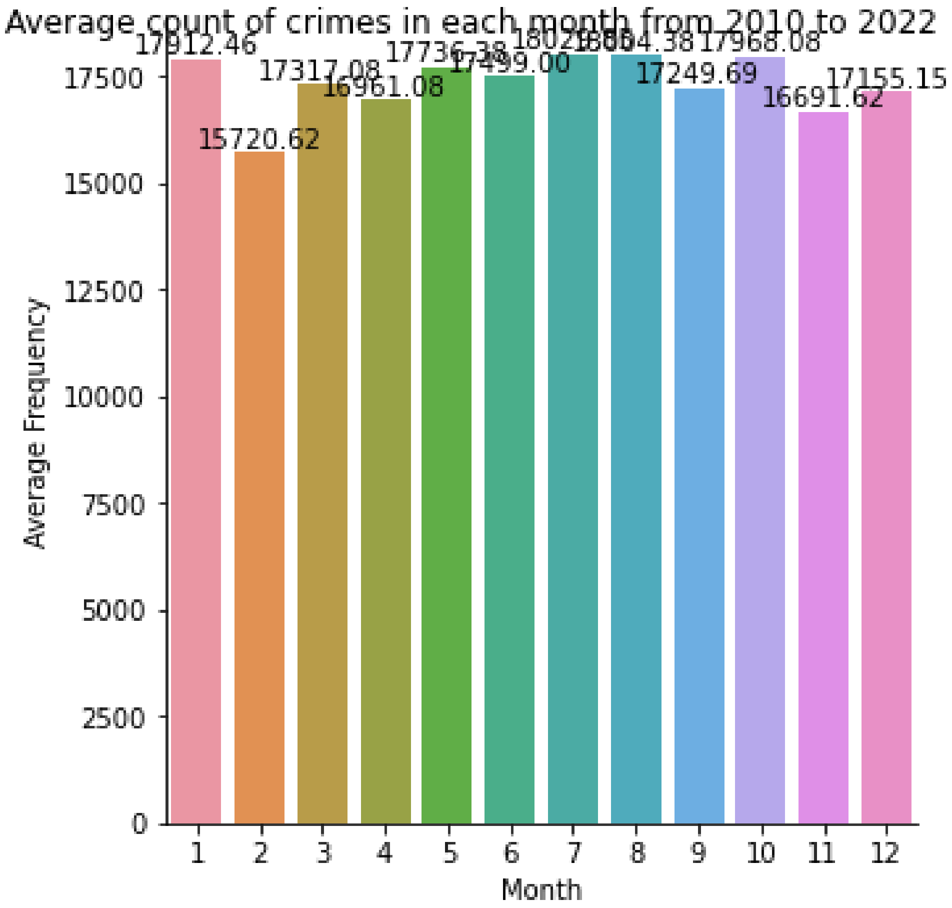


Figure 3

3.2 Crime Description Analysis

- Assault, theft, and burglary were the most common types of crime (See Figure 4).

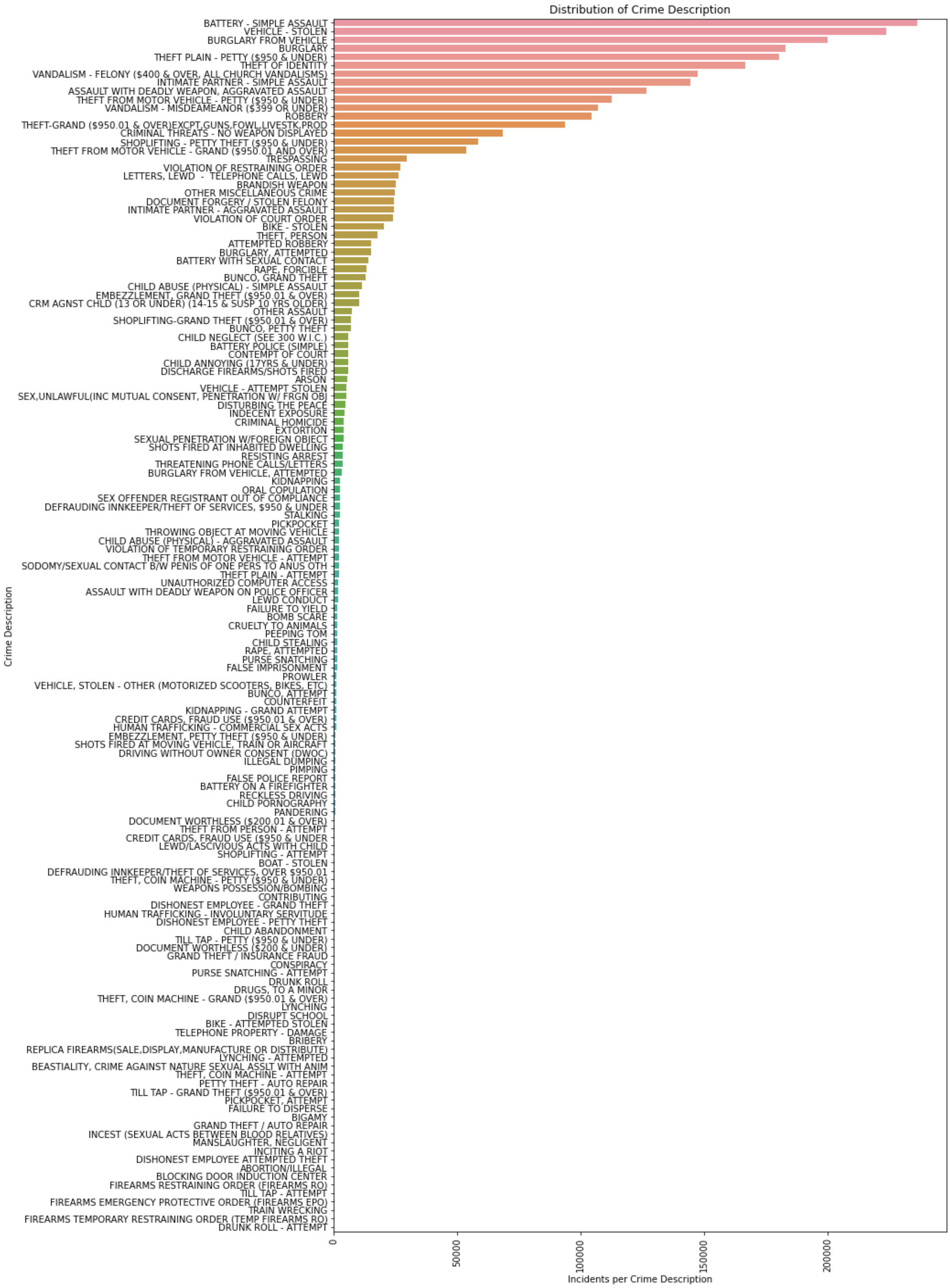


Figure 4

- Detailed analysis of the top 5 crimes was conducted using radar charts and count plots. Regardless "Battery - simple assult" has highest incidents from the past 12 years,the trend is not consistent in recent years. On the other hands, "Assult with deadly weapon" incidents are heading to upward trends among the other crimes (See Figure 5).

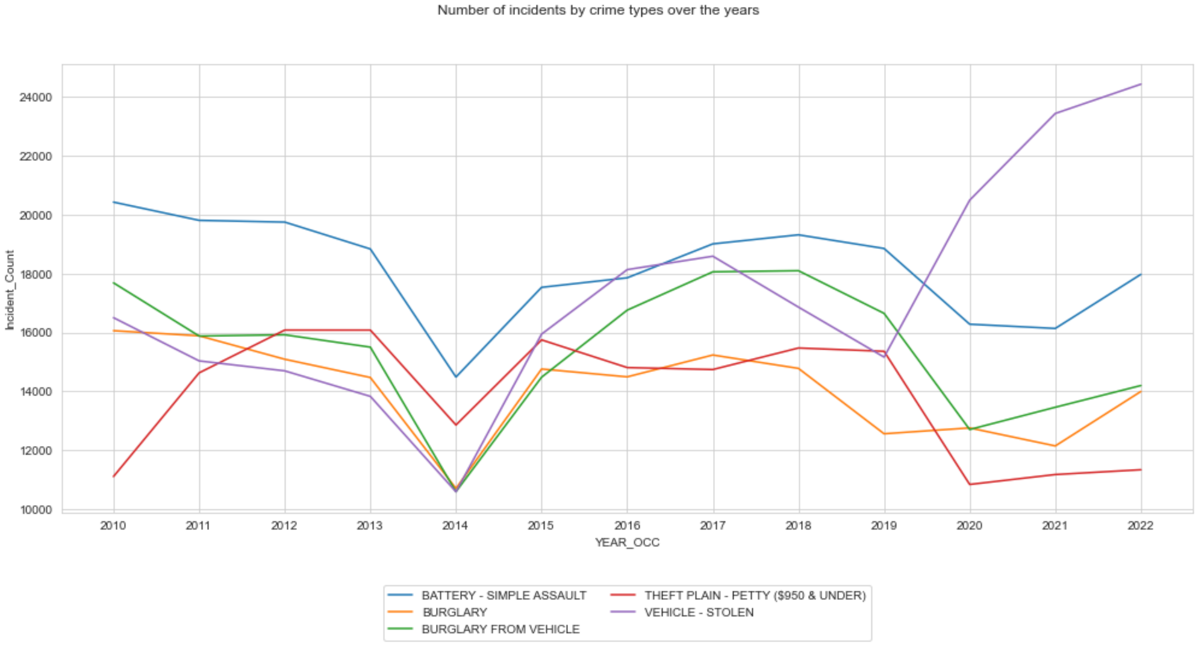


Figure 5

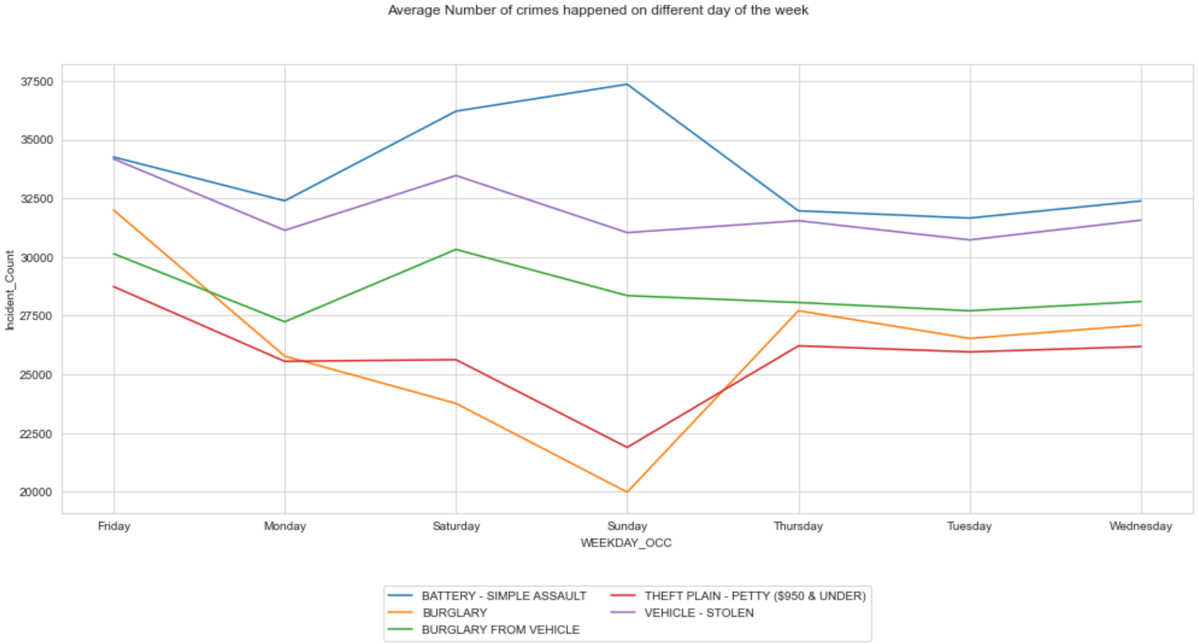


Figure 6

3.3 Victim Characteristics

- Age distribution, descent, and sex of victims were visualized. In general, the most of the victims are Hispanic/Latino/Mexican and Male (See Figure 7,8,9 &10).

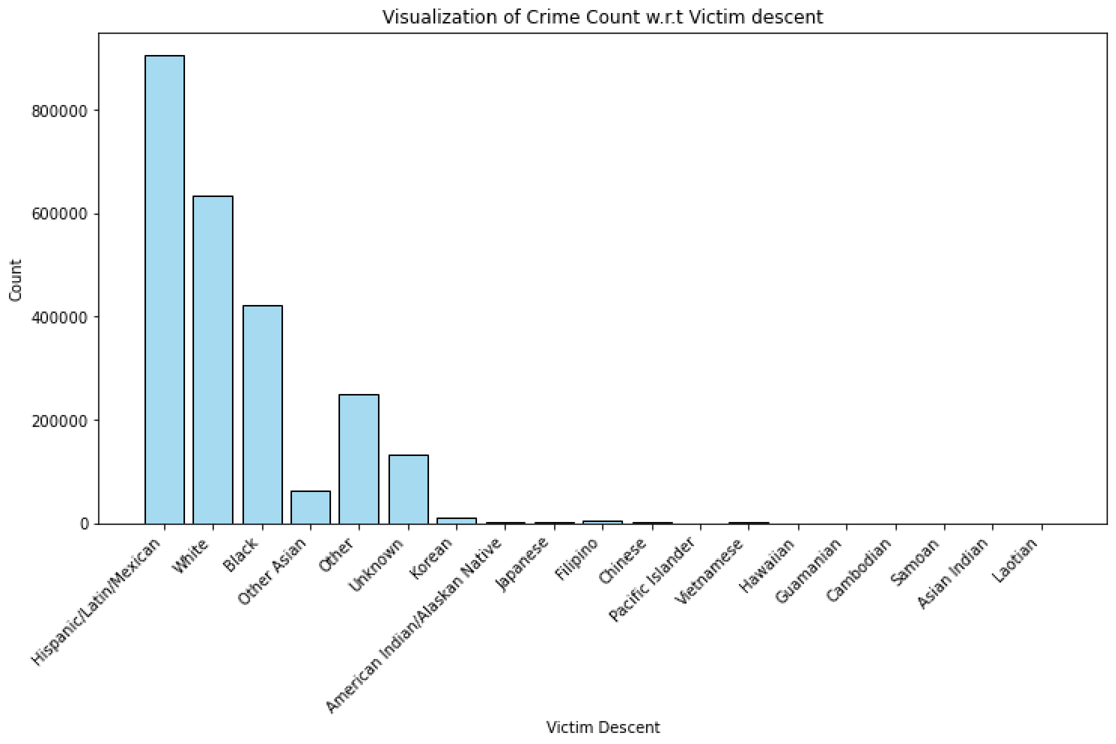


Figure 7

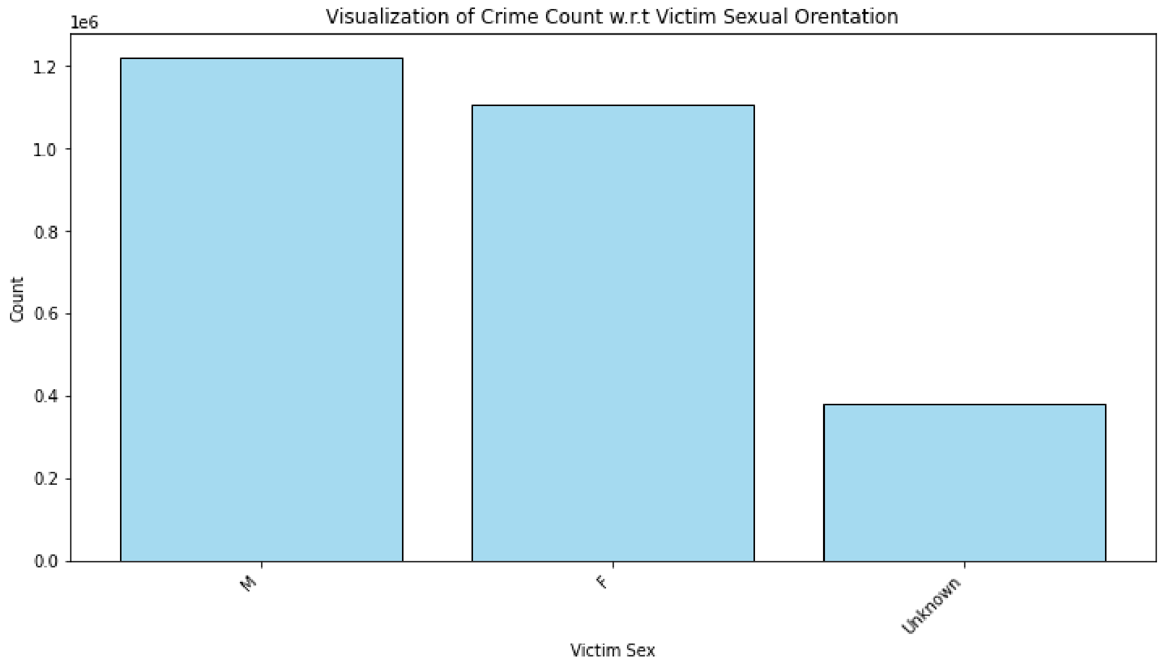


Figure 8

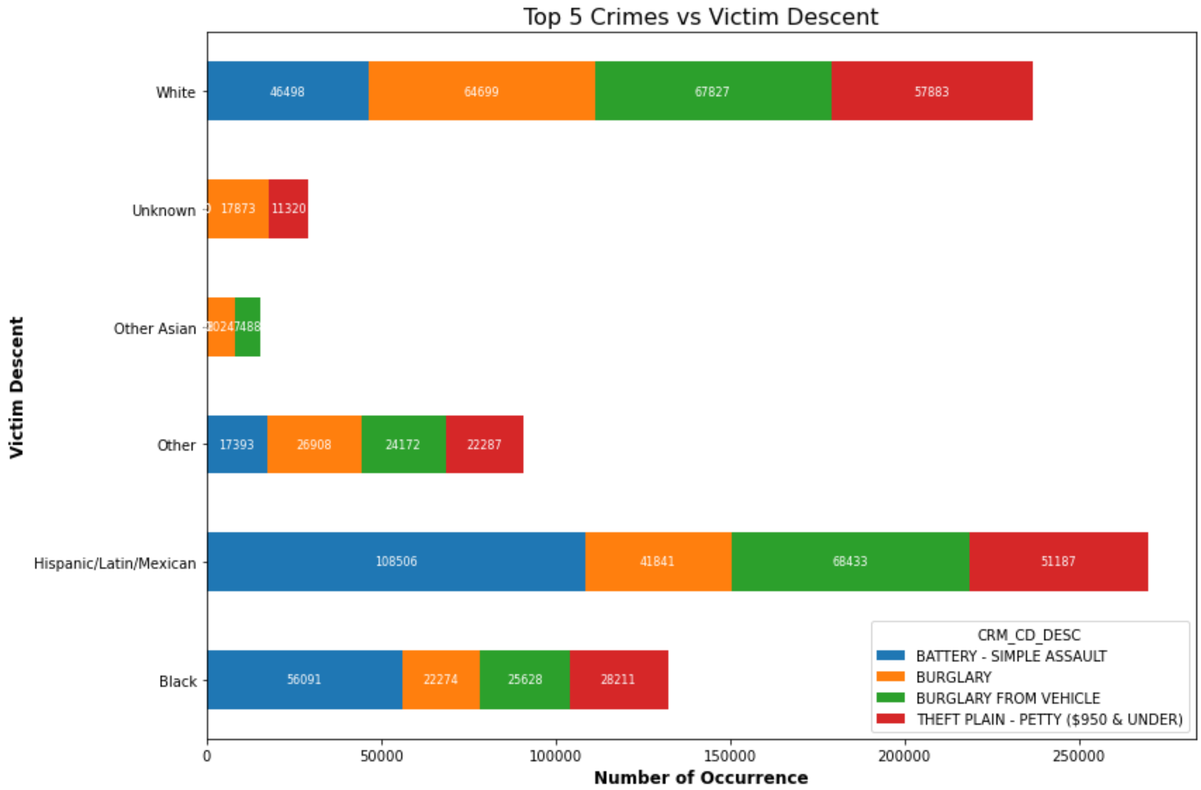


Figure 9

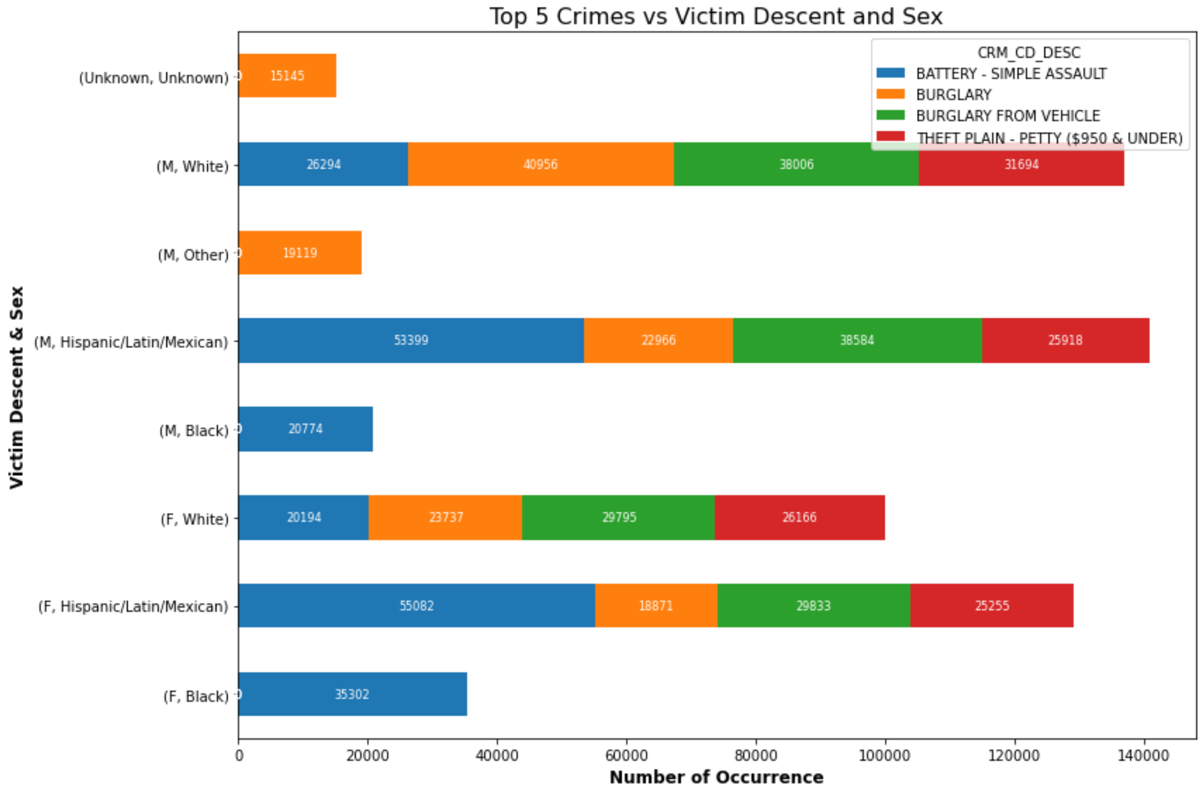


Figure 10

3.4 Area and Location

- Analysis of crime distribution by area within LA County revealed that 77th Street had the highest crime rate (Figure 11 & 12)
- Heatmaps and point maps were utilized to visualize crime occurrences geographically(Figure 13).

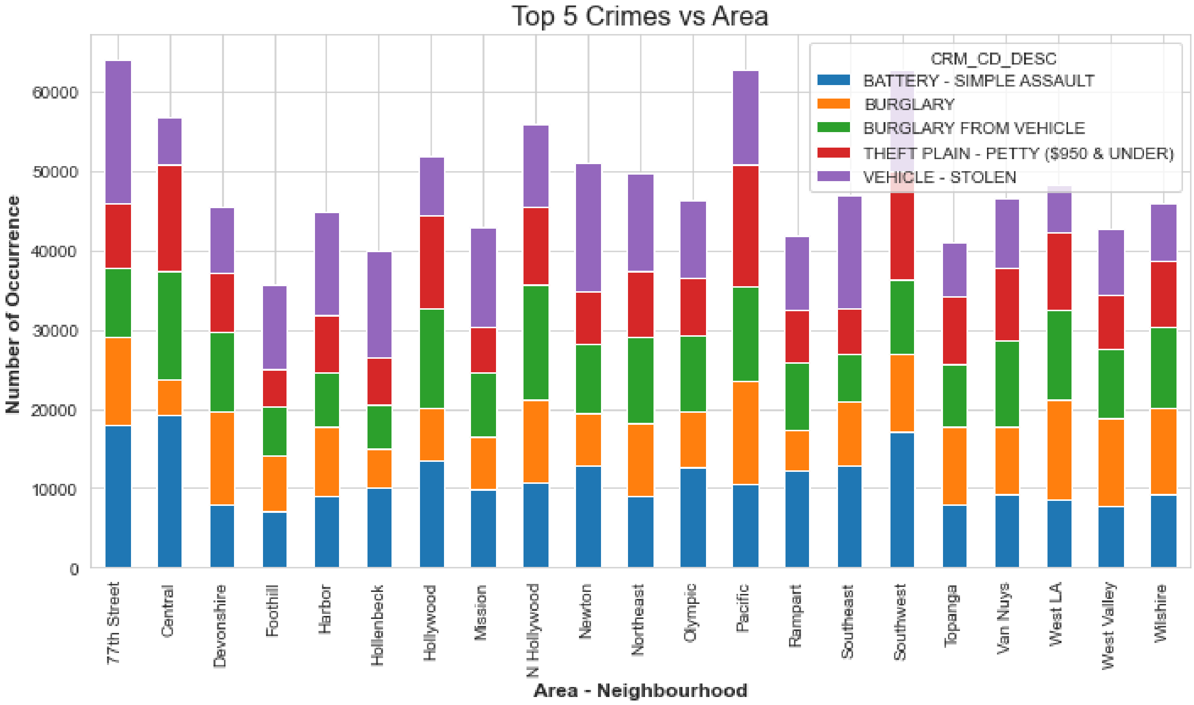


Figure 11

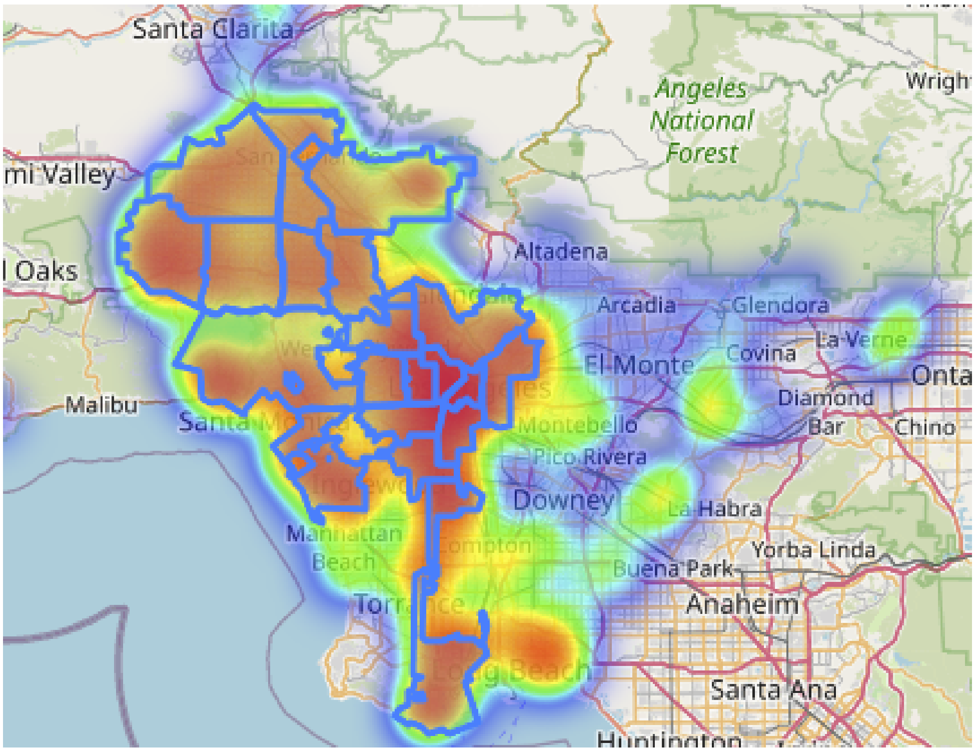


Figure 12

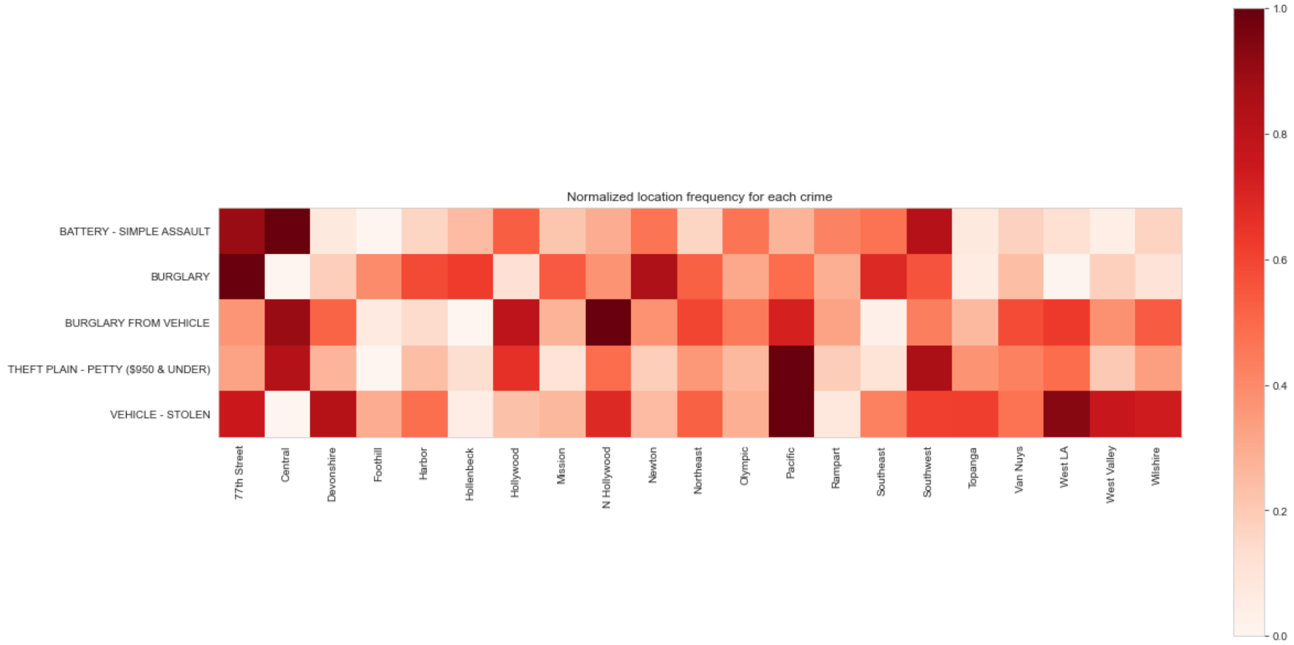


Figure 13

3.5 EDA Summary

EDA visualisations concluded with observations on crime trends with top 5 frequent crimes over 12 years and the charts, numbers, and data indicate

some of the crimes have been reported an upward trend, some are downward. But Downtown LA (77th Street) area is all-time-high crime rate consistently. Data suggested forecasting future crime probabilities based on area ,victim characteristic and day of the week. The analysed data of the top 5 crime types is saved to a CSV file named top5_crime_data.csv for future reference.
In next step, *the crime rate trend of top 5 frequent crime type will be predicted and investigated the relationships between feature categories such as the area ,crime type and victim descent and the crime rate.*

4 Training and Modelling

4.1 Prepressing and Training

The pre-processing and training steps undertaken on a dataset of top 5 crime data. It includes data loading, feature engineering, scaling, and preparation for machine learning modelling. For data preparation for models, actions below were performed.

1. Initial pre-processing involves removing unnecessary columns like 'ID'. Categorical variables such as 'VICT_SEX' and 'VICT_DESCENT' are identified for encoding. Label Encoding is applied to these variables to convert them into numerical form.
2. Categorical columns are converted to string type. One-hot encoding is then applied, transforming these variables into a set of binary columns.
3. Duplicate records are removed post-encoding to maintain data integrity.
4. Numeric features are standardized using StandardScaler. This step is crucial for many machine learning algorithms to interpret the features uniformly.
5. Actual Dataset is extremely huge and not able to processed. Hence, two more datasets are prepared for different modelling purposes, focusing on crime category prediction and count prediction/forecasting. These include:
    a. Total count of crimes with feature columns (VICTIM DESCENT,SEX AGE and AREA and CRIME TYPE).
    b. Total count of crimes pivoted by crime type.

6. The datasets are split into features (X) and target variables (y) for various modelling approaches, including count prediction and forecasting.
7. The datasets are further split into training and testing sets with a typical split ratio, ensuring a portion of the data is reserved for model evaluation.
8. The processed data and split datasets are saved in Pickle format (train_test_split.pkl) and as CSV (top5_crime_pre.csv)files for further use in the modelling phase.

4.2 Models trainings

The modelling phase begins as the final step of project work. 5 regressors models were built and evaluated.

1) Linear Regression: A simple linear regression model is fitted to the training data. Evaluation metrics like MSE, RMSE, MAE, and R-squared are calculated, along with cross-validation scores. Coefficients and intercepts are also examined.
2) Random Forest Regressor: This model is known for its robustness and is fitted next. Similar to linear regression, evaluation metrics and feature importance are determined.
3) Gradient Boosting Machines (XGBoost): This involves using GridSearchCV for hyperparameter tuning. The best model is then evaluated using the standard metrics.
4) CatBoost Regressor: Another gradient boosting model, CatBoost, is used. It undergoes a similar process of grid search and evaluation.
5) Decision Tree Regressor: Finally, a decision tree model is built. The best parameters are determined through grid search, and the model is evaluated.

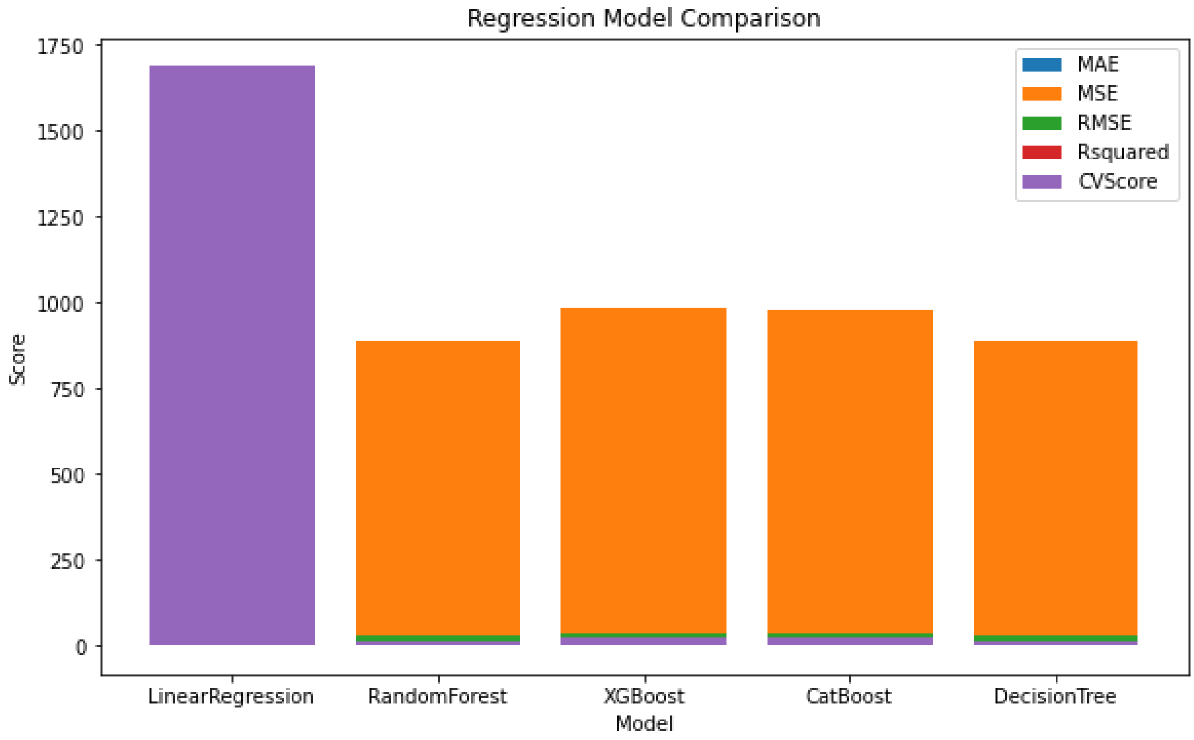
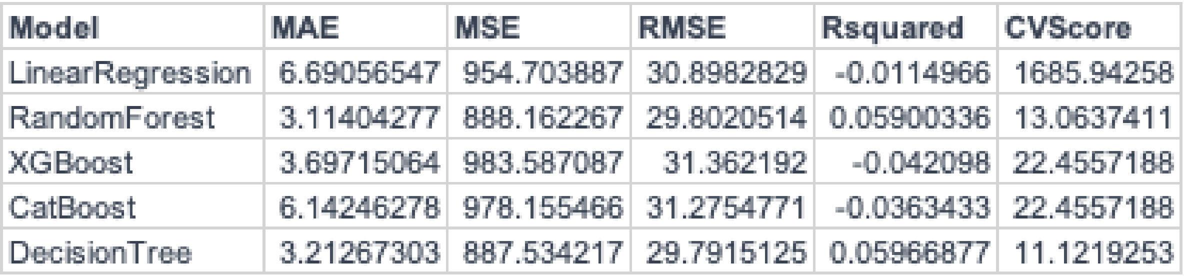


Figure 14



Model	MAE	MSE	RMSE	Rsquared	CVScore
LinearRegression	6.69056547	954.703887	30.8982829	-0.0114966	1685.94258
RandomForest	3.11404277	888.162267	29.8020514	0.05900336	13.0637411
XGBoost	3.69715064	983.587087	31.362192	-0.042098	22.4557188
CatBoost	6.14246278	978.155466	31.2754771	-0.0363433	22.4557188
DecisionTree	3.21267303	887.534217	29.7915125	0.05966877	11.1219253

Figure 15

To choose the most optimal model, Models are compared based on various metrics like MAE, MSE, RMSE, R-squared, and CV scores.

1. Linear Regression has high errors (MAE, MSE, RMSE) and a negative R-squared, suggesting poor performance.
2. Random Forest shows relatively low errors and the highest R-squared, indicating good performance.
3. XGBoost has moderate errors and a negative R-squared, indicating average to below-average performance.
4. CatBoost shows high errors and a negative R-squared, similar to Linear Regression.
5. Decision Tree has moderate errors and the highest R-squared (similar to Random Forest), suggesting good performance.

Overall, the Random Forest and Decision Tree models seem to perform the best. Random Forest has the lowest MAE and the highest R-squared, indicating it might be the most accurate and consistent. However, Decision Tree also shows strong performance with a slightly lower RMSE and the same R-squared as Random Forest.

Further validation on model is carried out to determine overfitting or underfitting.

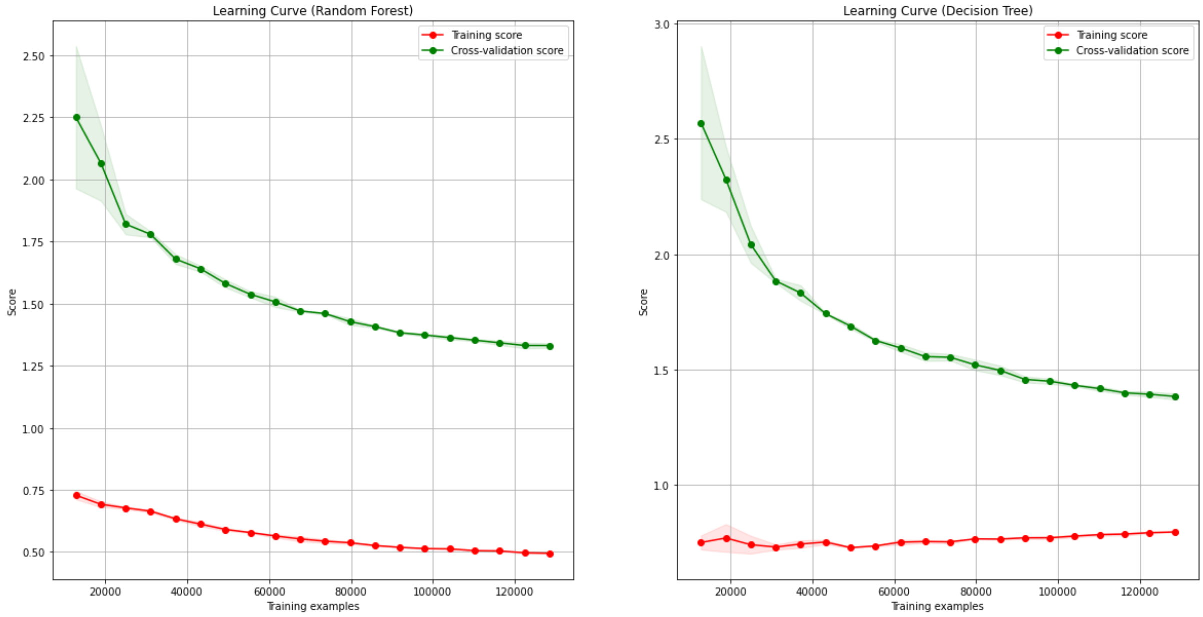


Figure 16

On the Random Forest Learning curve plot (Figure 16 - left) ,The gap between the training and cross-validation scores is narrowing but remains relatively constant in the latter part of the curve.

The Decision Tree Learning curve plot (Figure 16 - right) has a very high training score initially, suggesting potential overfitting. However, as more data is used for training, the training score decreases, and the cross-validation score increases, closing the gap between them.

Based on the information visible, **Decision tree model** is chosen as the best model because the cross-validation score is improving significantly with more data, and the gap between the training and validation scores is narrowing, suggesting that the model is benefiting from additional data and is likely generalizing well.

4.3 Final model fitting

The decision tree model has shown high feature importance for victim descent and crime type in predicting crime rates (Figure 17).

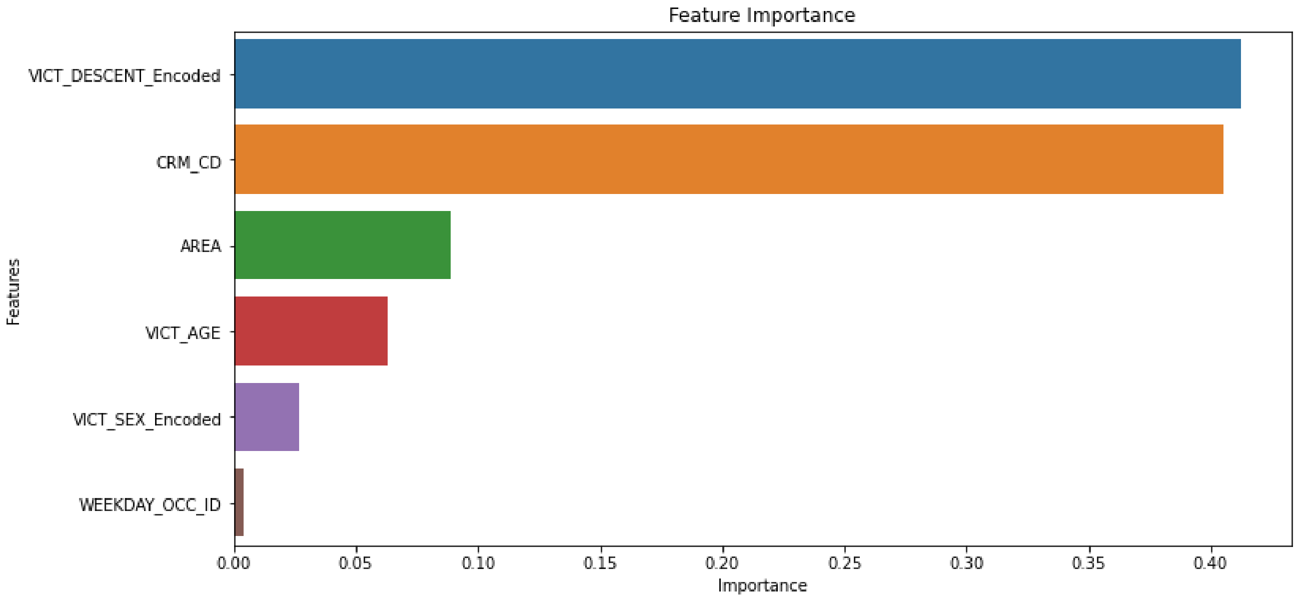


Figure 17

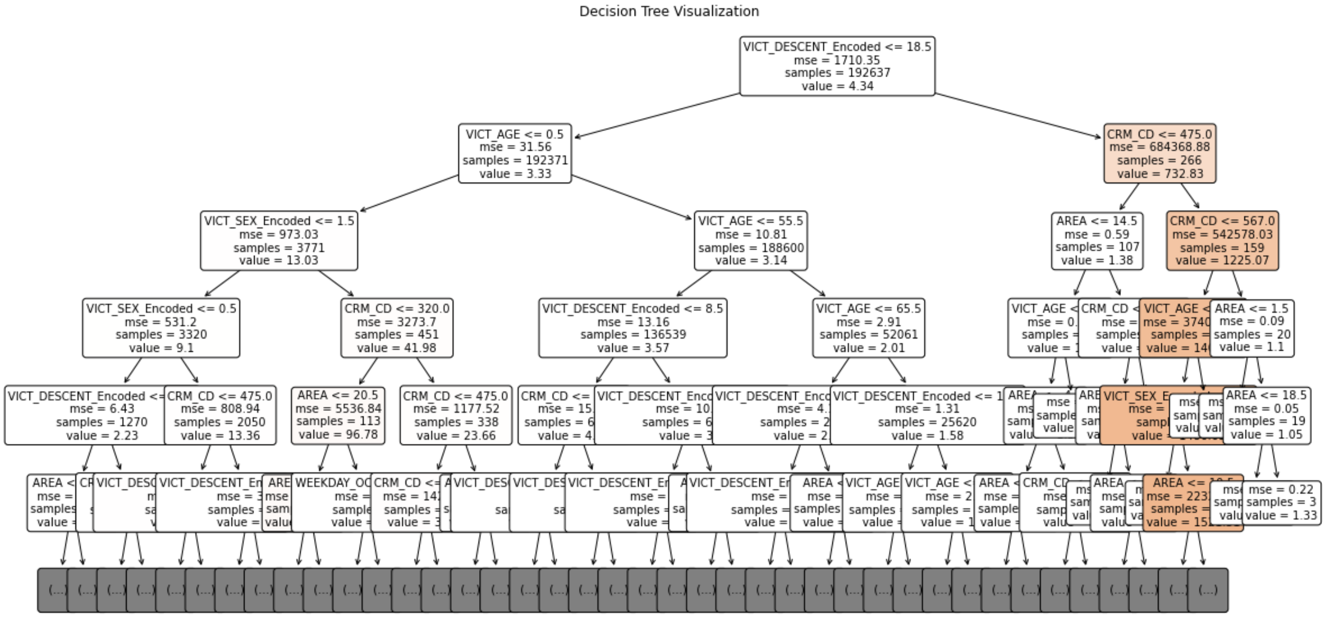


Figure 18

4.4 Final model Prediction

1. How Accurate Are the Predictions for Different Crime Types?

- Battery - Simple Assault: Predicted rates are slightly higher than actual rates.
- Burglary: Predicted rates are slightly higher than actual rates.

- Burglary from Vehicle: Predicted rates are slightly lower than actual rates.
- Theft Plain - Petty: Predicted rates are slightly higher than actual rates.

- Vehicle - Stolen: Predicted rates are significantly lower than actual rates, indicating a possible area for model improvement.

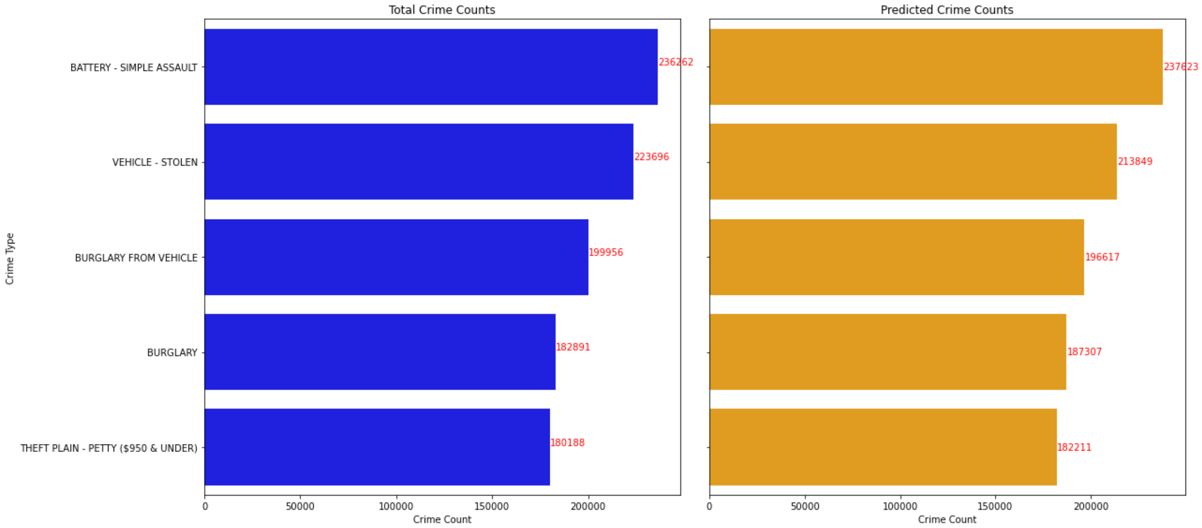


Figure 19

2. How Does the Model's Accuracy Vary Across Different Descent Groups?

- Significant Overestimations: American Indian/Alaskan Native, Asian Indian, and Guamanian groups have predicted crime rates much higher than actual rates.
- Close Predictions: Hispanic/Latin/Mexican descent group predictions are nearly identical to actual rates.
- Notable Underestimation: The "Unknown" descent group has the highest actual average crime rate, with predictions being slightly lower.

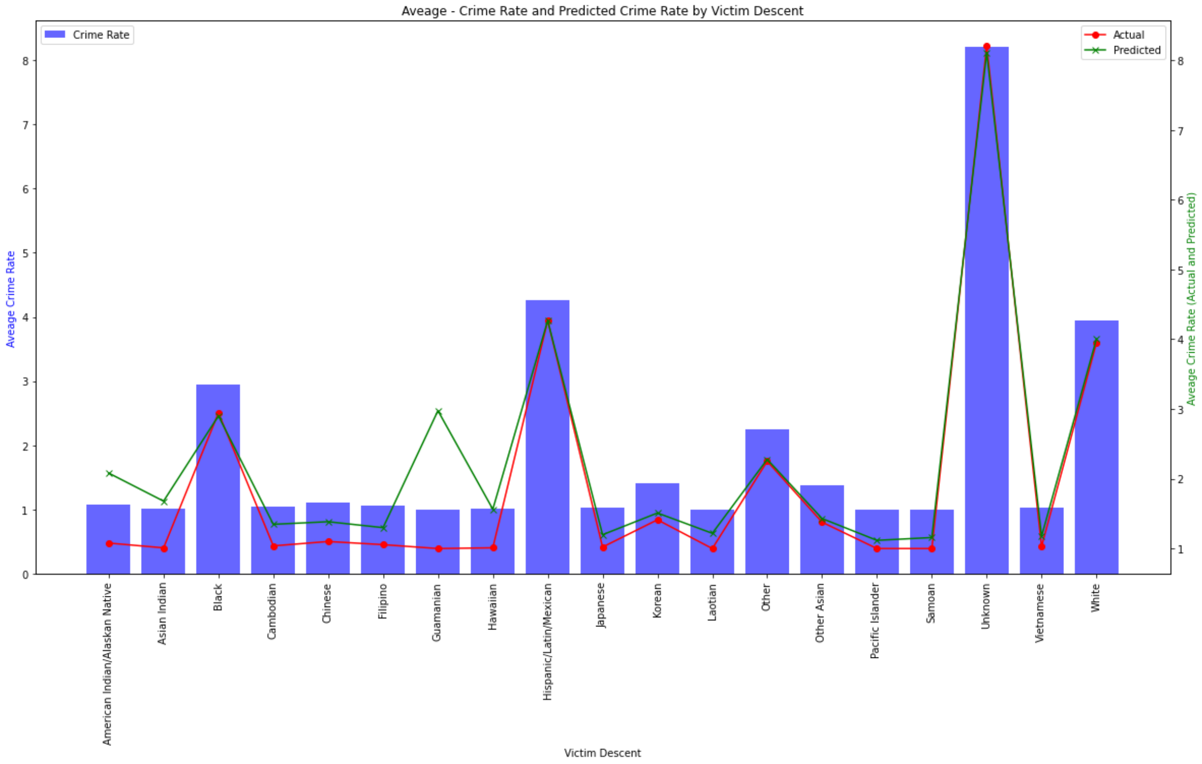


Figure 20

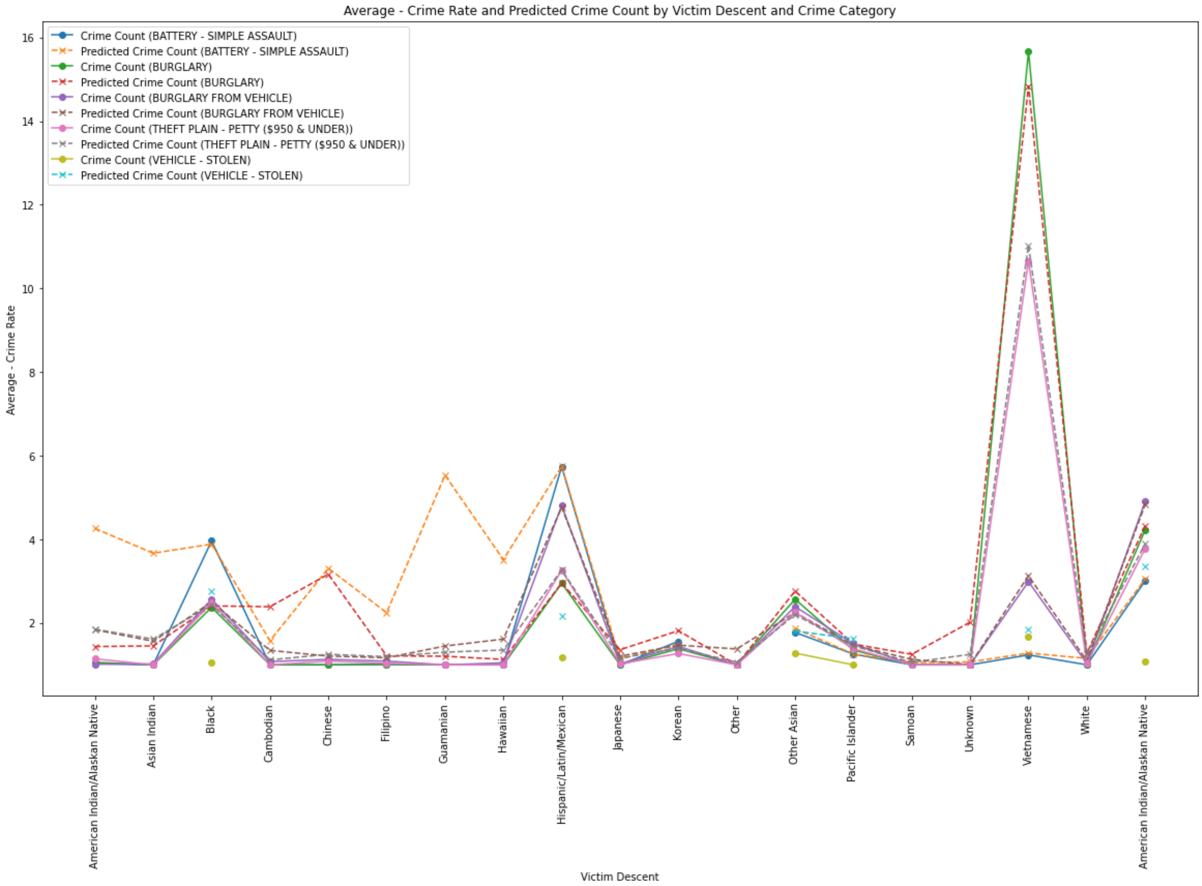


Figure 21

3. What Are the Differences Between Actual and Predicted Crime Rates by Area?

- Higher Actual Rates: 77th Street, Hollenbeck, Mission, Newton, and Van Nuys areas have higher actual crime rates than predicted.
- Higher Predicted Rates: Central, Foothill, Olympic, and West Valley have higher predicted rates than actual.

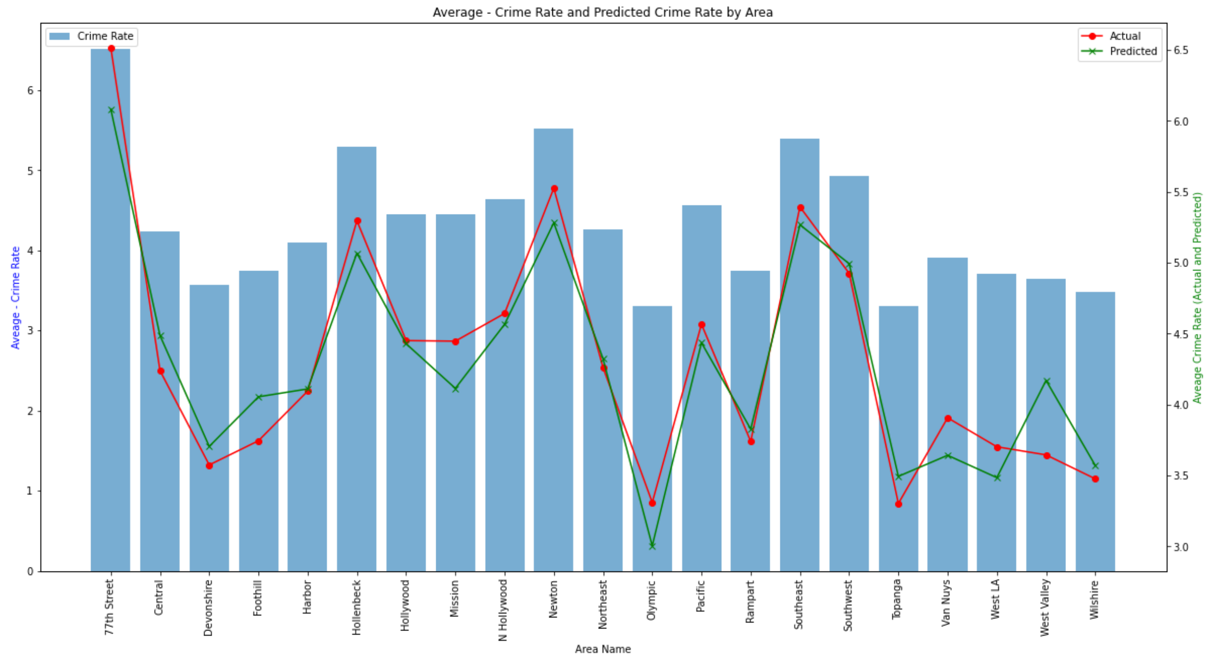


Figure 22

4. How Well Does the Model Predict Crime Rates Based on the Day of the Week?

- Higher Actual Weekend Rates: Fridays and Sundays show higher actual crime rates than predicted, potentially indicating a trend that the model may not be capturing effectively.
- Variable Predictions During the Week: There is an alternating pattern of over and under-predictions on weekdays.

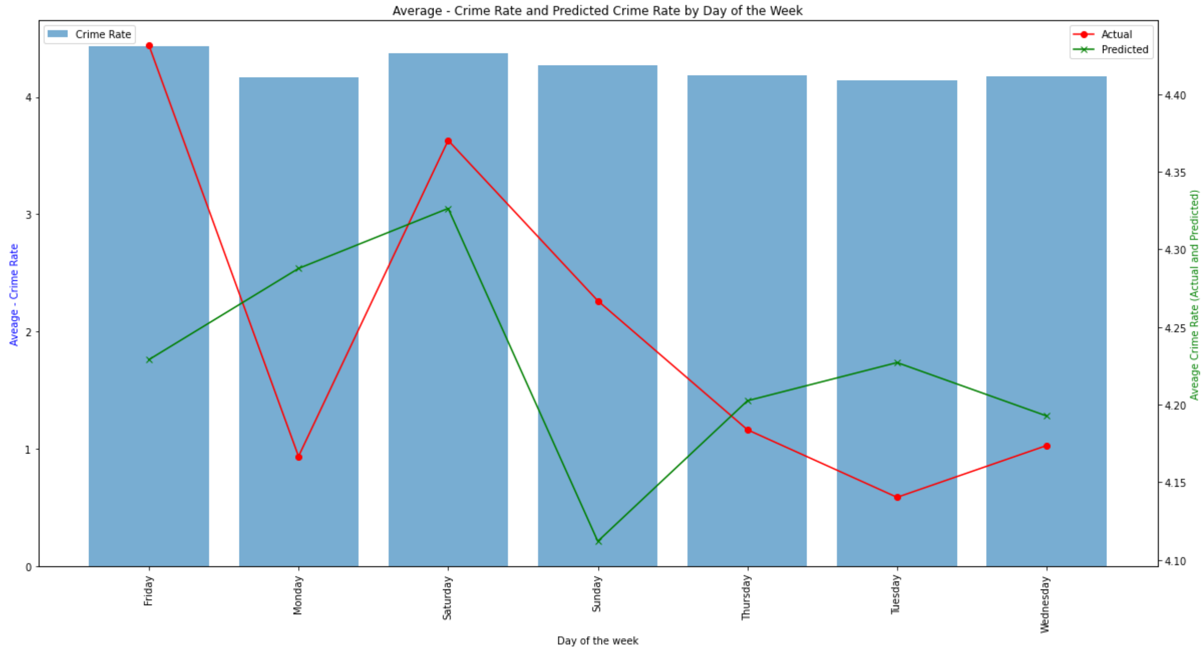


Figure 23

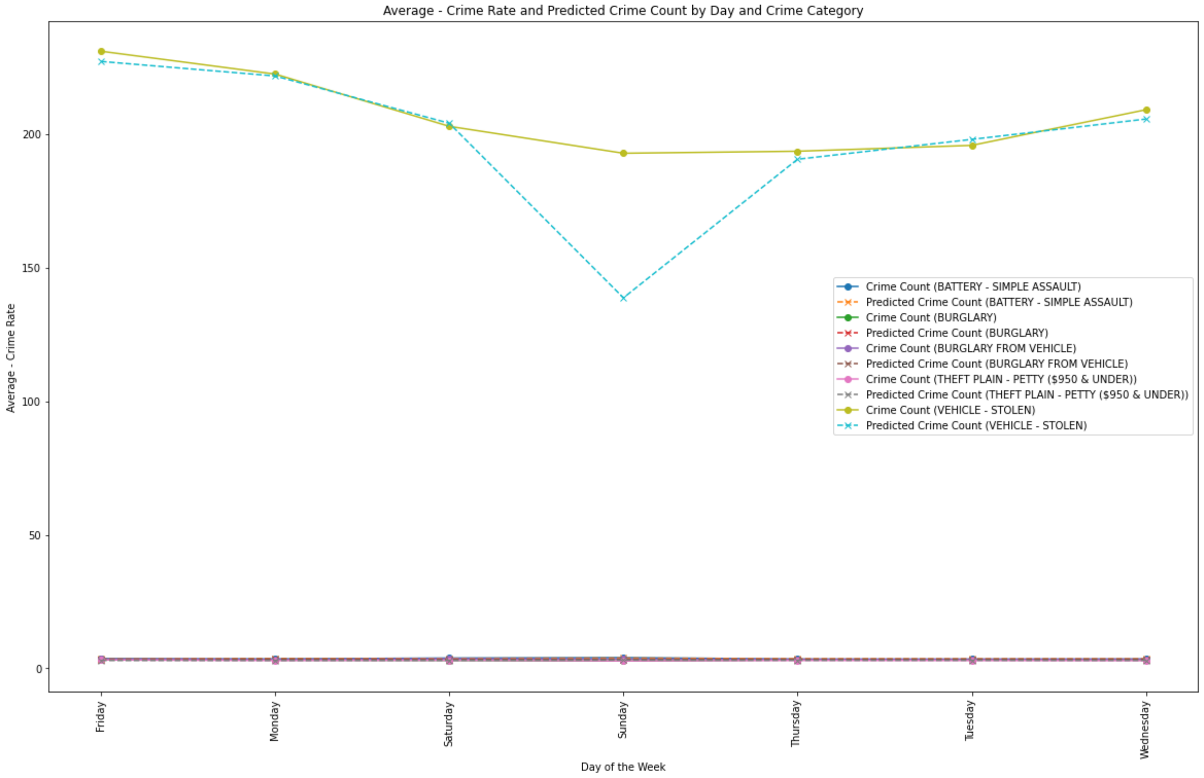


Figure 24

To summarise, the prediction suggested that a general trend where the model predict upward crime rates for certain descent groups, particularly American Indian/Alaskan Native and Guamanian. Predictions for the "Unknown" descent group are lower than the actual rates, which may require model recalibration. The tends to suggest that vehicle theft, especially on Sundays are likely to lessen. The victim descent Hispanic/Latin/Mexican descent group are remarkably the highest targeted victim constantly.

5 Future improvement

The primary constraint we encountered was insufficient Random Access Memory (RAM). This limitation compelled us to use only a portion of the available data, restricting the model's ability to fully leverage the dataset for analysis. Specifically, it hindered the incorporation of comprehensive time series data and precise geographical coordinates, such as latitude and longitude.

Additionally, investing in more powerful computational capabilities would likely yield significant returns in model accuracy and performance through advanced hyperparameter tuning.