LA COUNTY CRIME RATE TREND PREDICTION

Presented by:
Thant Thiri M. Kyi
(Springboard Data Science Career Track, July 2023 Cohort)

Understanding Top 5 Frequent Crime Patterns in Los Angeles: A Data-Driven Approach

Context

Los Angeles, known for its dynamic urban landscape, faces challenges with varying crime rates.Recognizing the need for a proactive approach to community safety.

Problem Definition

- The persistent issue of crime necessitates innovative solutions.
- Utilizing data analysis to uncover crime trends and predict future patterns.

Project Objective

- To develop a predictive model that analyses historical crime data.
- Aiming to provide insights that assist in informed decision-making for community welfare.

Impact

- Highlighting the potential of data-driven strategies in enhancing public safety measures.
- Empowering communities with knowledge for better preparedness and response.

Data Acquisition

Data Source:

Los Angeles city public data website (https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/explore ,

https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8)

1. Time Range: 2012-2023
2. Record Count: ~3 million (2,993,433)

Key Variables:

1. Crime Codes, Modus Operandi, Victim's Age, Sex, Descent
2. Premise Type, Weapon Used, Case Status
3. Crime Severity (Crm Cd 1-4)
4. Location Details (Address, Latitude, Longitude)

DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA N...	Rpt Dist ...	Part 1-2	Crm Cd	Crm Cd ...	Mocodes	Vict Age	Vict Sex	Vict Des...	Premis Cd	F
010304468	2020 Jan 0	2020 Jan 0	2230	03	Southwest	0377	2	624	BATTERY -	0444 0913	36	F	B	501	S
190101086	2020 Jan 0	2020 Jan 0	0330	01	Central	0163	2	624	BATTERY -	0416 1822	25	M	H	102	S
200110444	2020 Apr 1	2020 Feb 1	1200	01	Central	0155	2	845	SEX OFFEN	1501	0	X	X	726	F
191501505	2020 Jan 0	2020 Jan 0	1730	15	N Hollywoo	1543	2	745	VANDALISN	0329 1402	76	F	W	502	M
191921269	2020 Jan 0	2020 Jan 0	0415	19	Mission	1998	2	740	VANDALISN	0329	31	X	X	409	E
200100501	2020 Jan 0	2020 Jan 0	0030	01	Central	0163	1	121	RAPE, FORC	0413 1822	25	F	H	735	N
200100502	2020 Jan 0	2020 Jan 0	1315	01	Central	0161	1	442	SHOPLIFTII	1402 2004	23	M	H	404	D
200100504	2020 Jan 0	2020 Jan 0	0040	01	Central	0155	2	946	OTHER MIS	1402 0392	0	X	X	726	F
200100507	2020 Jan 0	2020 Jan 0	0200	01	Central	0101	1	341	THEFT-GRA	1822 0344	23	M	B	502	M
201710201	2020 Jun 1	2020 May 2	1925	17	Devonshire	1708	1	341	THEFT-GRA	1300 0202	0	X	X	203	C

Data Waggling

1. **Initial Data Processing:**
    1. Merged two datasets (2010-2019, 2020-present)
    2. Standardized column names
    3. Filled missing values in key columns
2. **Data Transformation:**
    1. Renamed 'LAT' and 'LON' to 'latitude' and 'longitude'
    2. Derived date and time components from 'DATE_OCC'
    3. Mapped codes to descriptions for better readability
3. **Cleaning and Exploration:**
    1. Excluded incomplete year data (2023)
    2. Dropped duplicates, sorted, and reindexed
    3. Analysed crime frequency by area and weekdays
4. **Final Data State:**
    1. Total Records: 2,707,190
    2. Total Columns: 18
    3. Output File: 'cleaned_crime_data.csv'

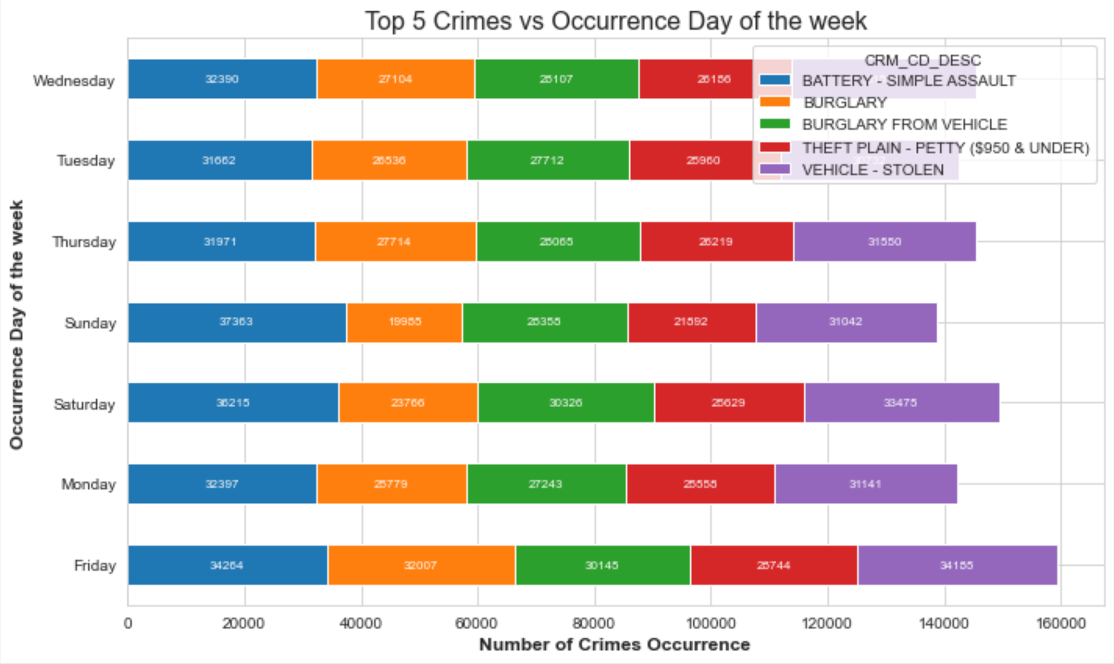
Time-Based Analysis - EDA

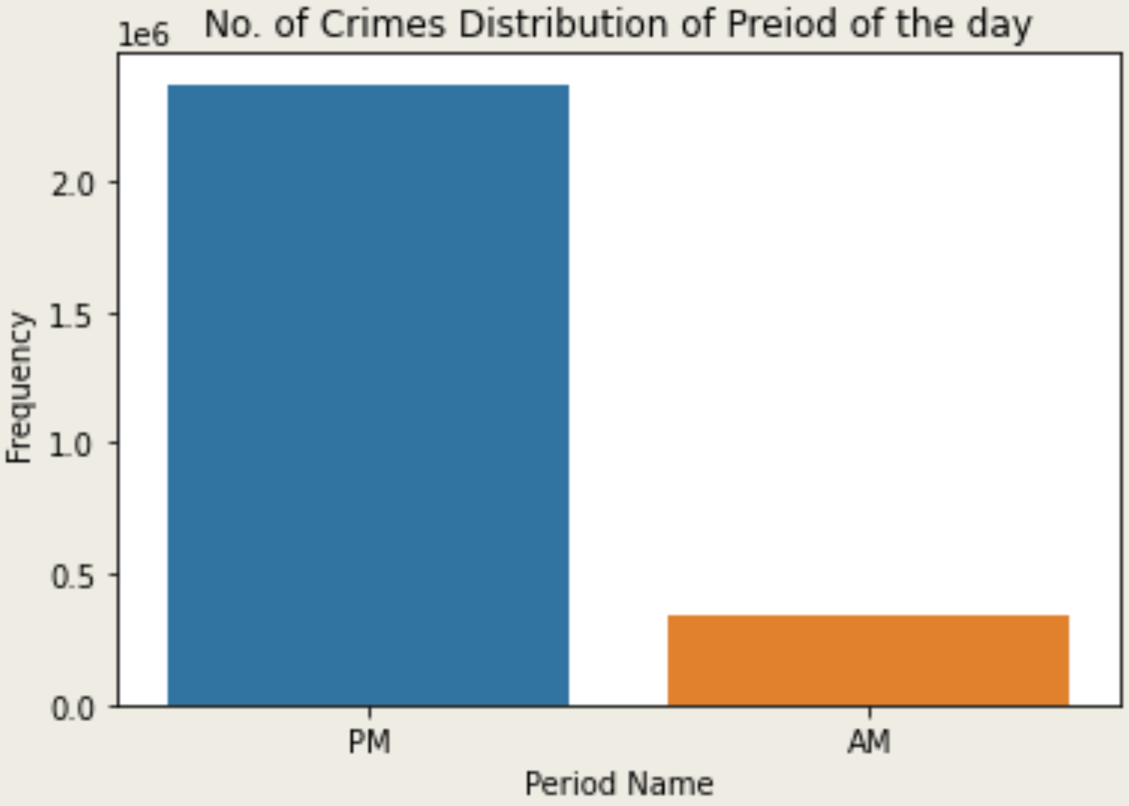
EDA Overview:

- Dataset: 'cleaned_crime_data.csv' (2010-2022)
- Focus: Timeframe, Weekdays, Crime Categories, Victim Characteristics, Area of Occurrence

Crime Occurrence by Time:

- **Weekdays:** Higher incidents on Fridays and Saturdays, lowest on Sundays (Figure 1)
- **Time of Day:** Evenings see the highest crime rates (Figure 2)

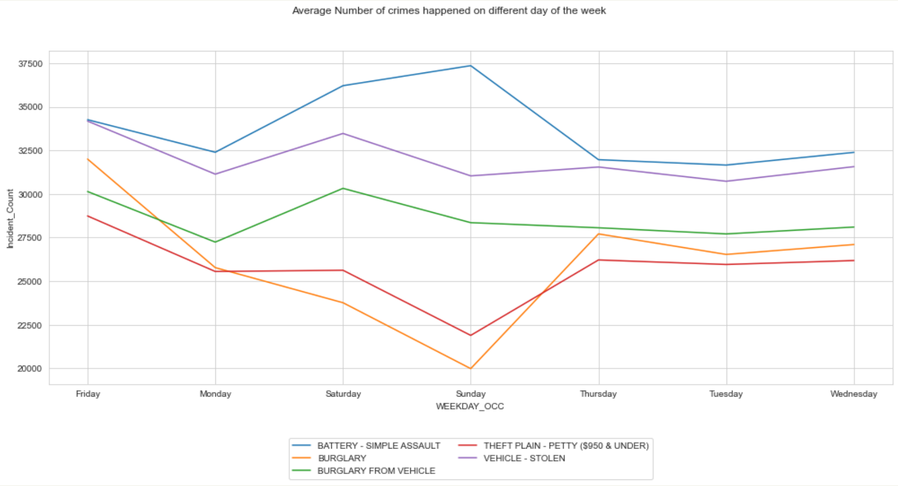


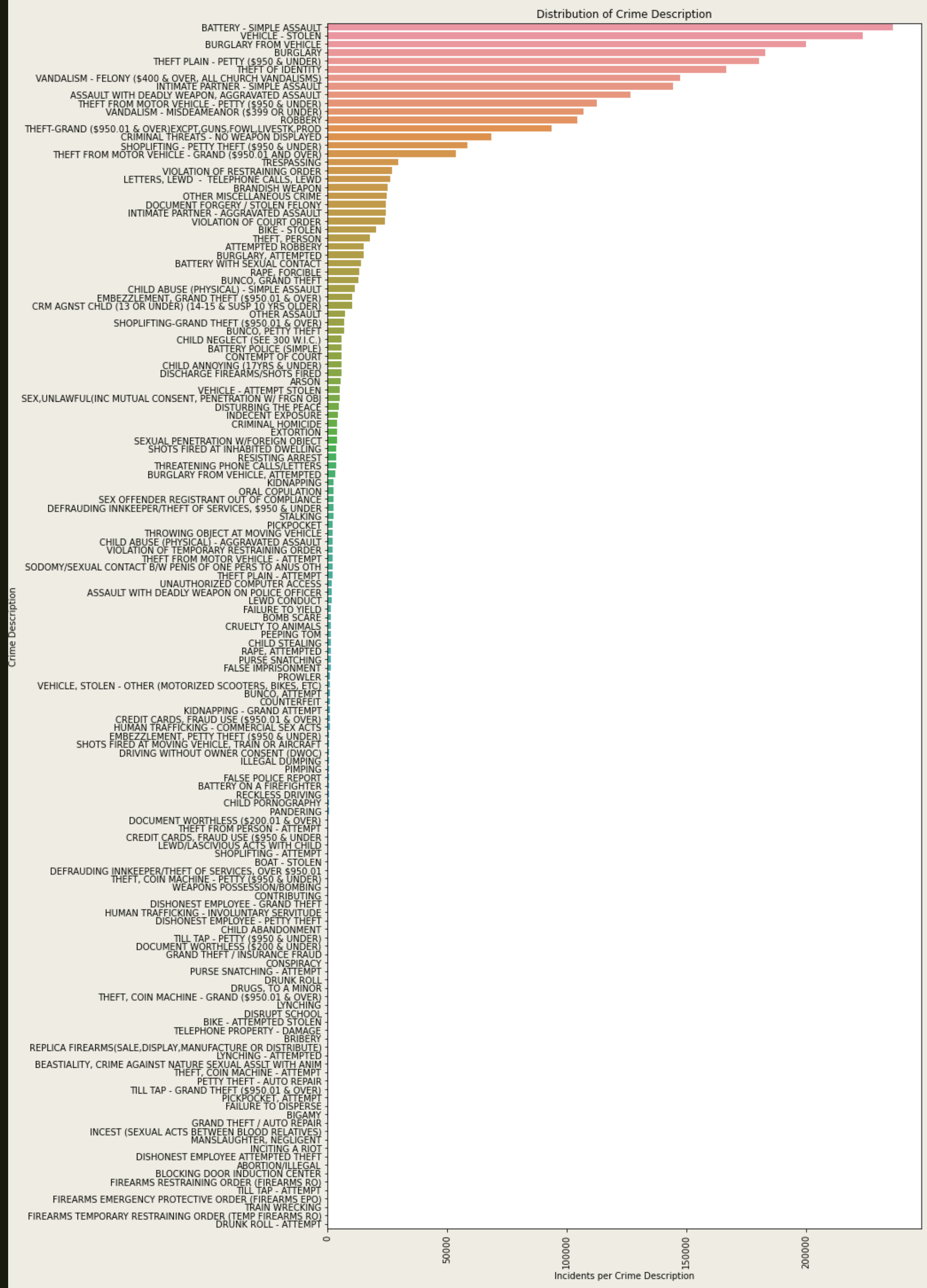


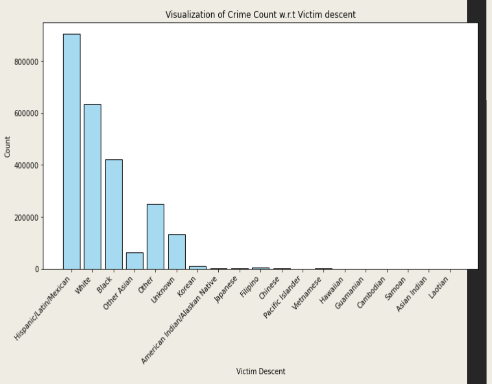
Crime Description Analysis:- EDA

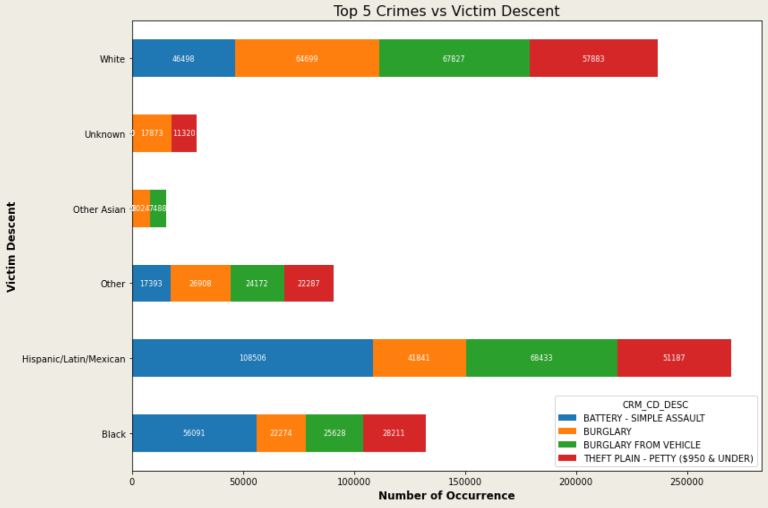
Crime Description Analysis:

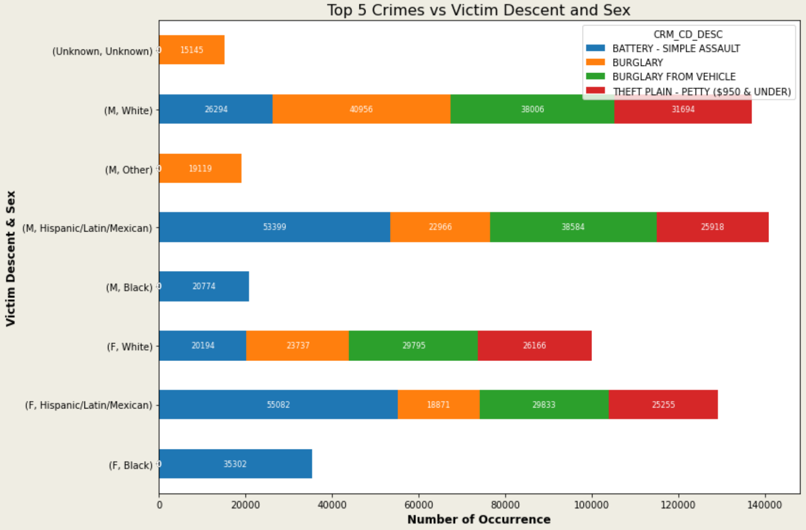
- Common Crimes: Assault, Theft, Burglary
- Trend Analysis: 'Battery - Simple Assault' highest but inconsistent; 'Assault with Deadly Weapon' on an upward trend.

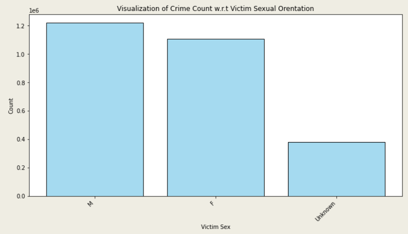








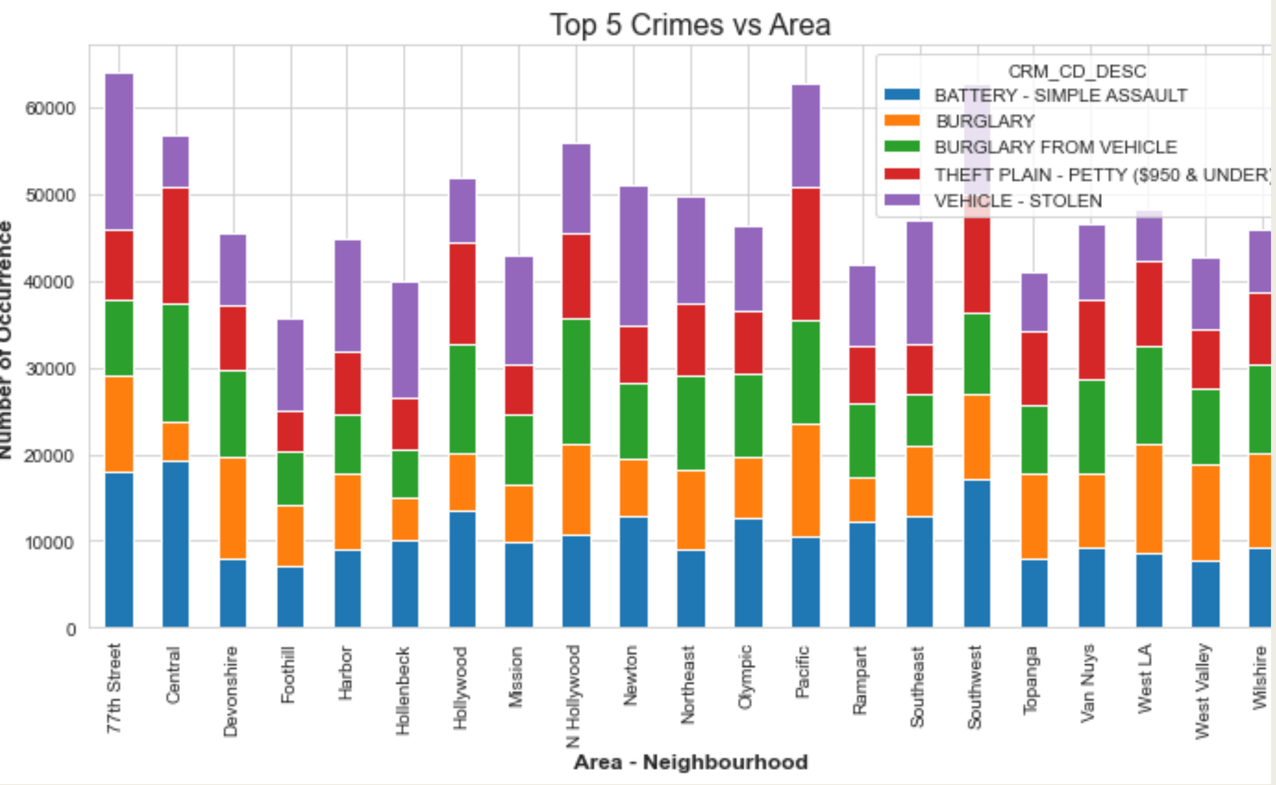


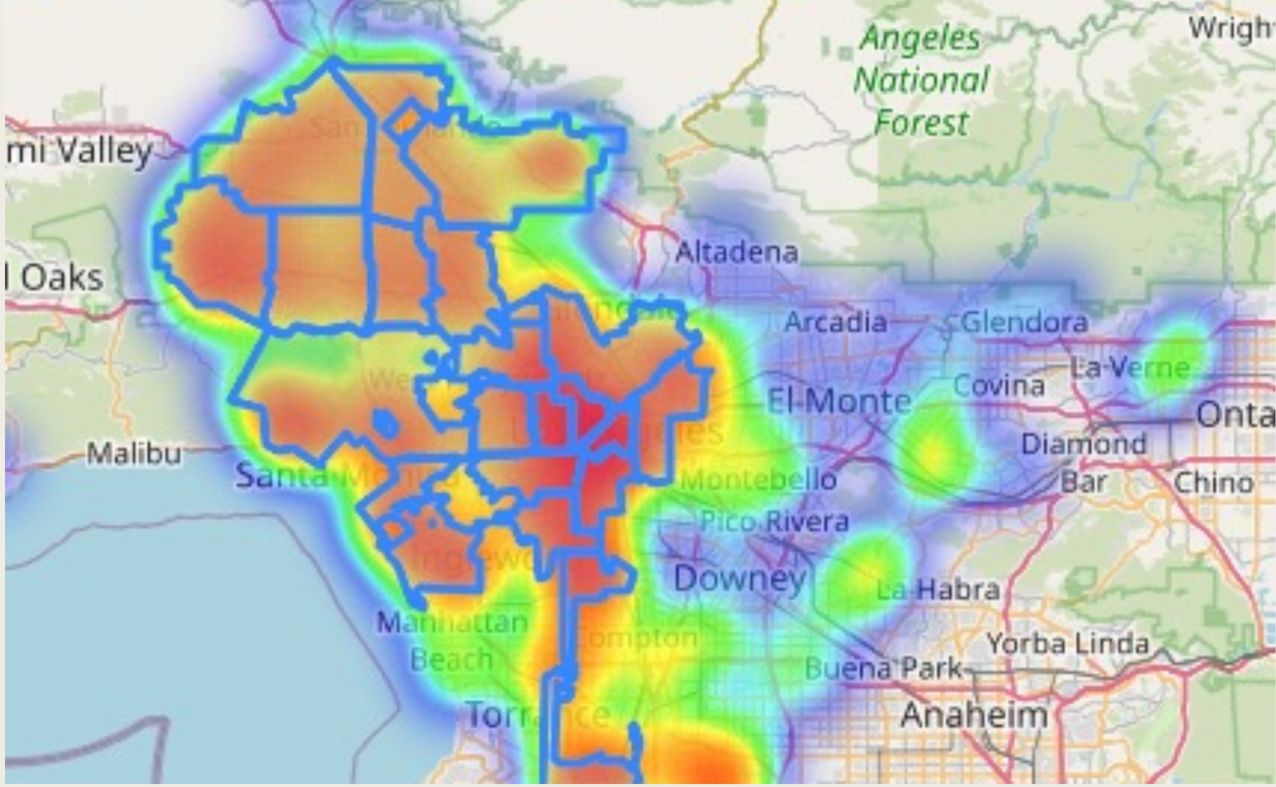


Victim Demographics - EDA

Victim Characteristics:

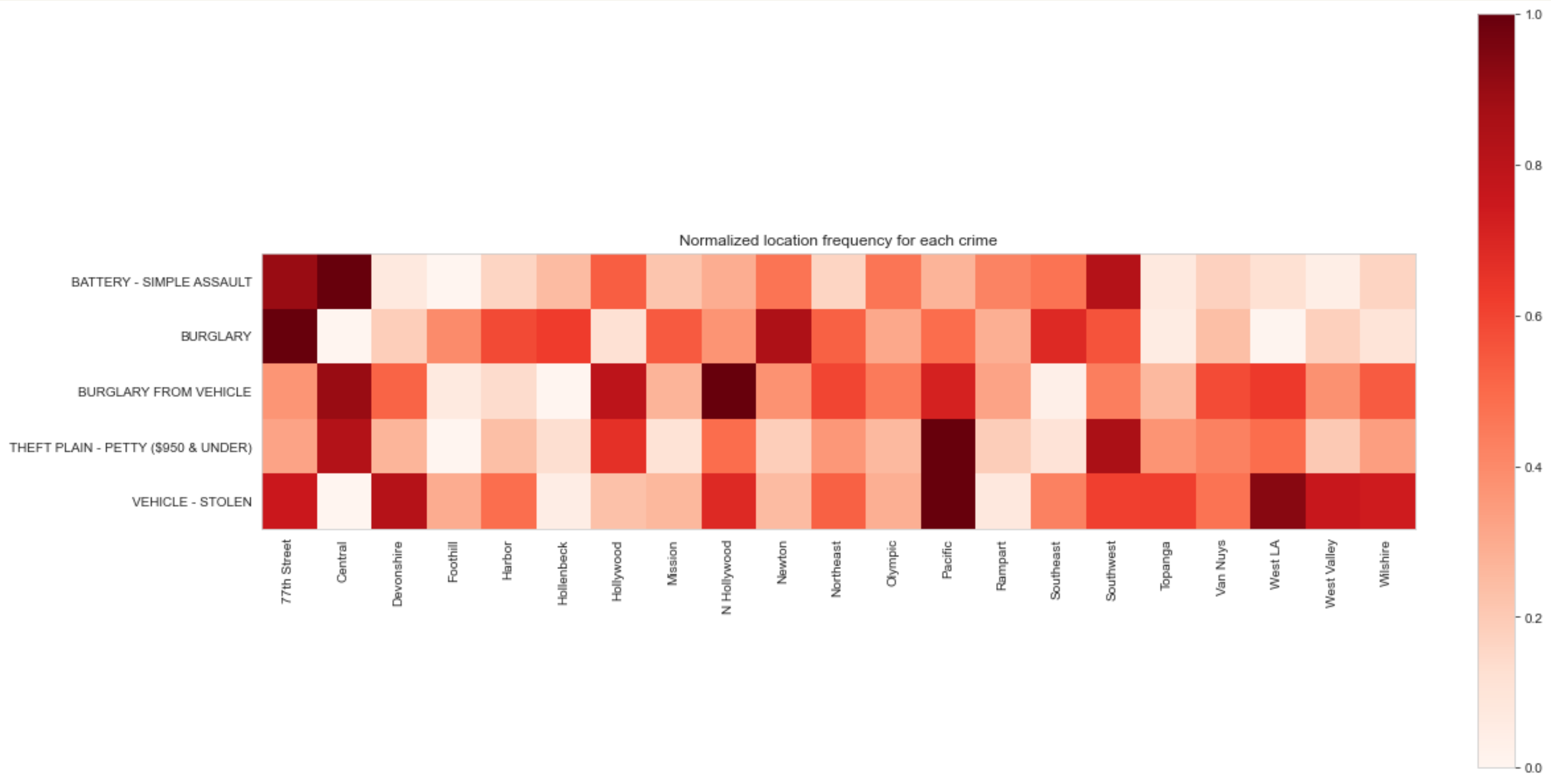
- Age, Descent, Sex Analysis
- Most victims: Hispanic/Latino/Mexican and Male





Area and Location Analysis - EDA

- Highest Crime Rate: 77th Street, LA County
- Geographic Distribution: Heatmaps and Point Maps



EDA Summary

- Crime Trends: Mix of upward and downward trends in top 5 crimes over 12 years
- Consistent High Crime Area: Downtown LA (77th Street)
- Future Forecasting: Based on area, victim characteristics, and day of the week
- Output File: 'top5_crime_data.csv'

Building the Framework for Prediction – Feature Engineering

1. Data Pre-processing Steps

1. **Data Preparation:**
   1. Loaded 'top 5 crime data' dataset
   2. Removed irrelevant columns (e.g., 'ID')
2. **Feature Engineering:**
   1. Categorical Encoding:
      1. 'VICT_SEX', 'VICT_DESCENT' via Label Encoding
      2. One-hot Encoding for binary column transformation
3. **Data Cleaning:**
   1. Duplicates removed post-encoding
4. **Feature Scaling:**
   1. Standardization using StandardScaler

2. Data Preparation for Modelling and Training

1. **Dataset Segmentation for Modelling:**
   1. Large dataset necessitated creation of smaller, focused datasets:
      1. Crime category prediction dataset with key features
      2. Crime count prediction/forecasting dataset

CRM_CD	AREA	VICT_AGE	VICT_DESCENT_Encoded	VICT_SEX_Encoded	WEEKDAY_OCC_ID	CRIME_Total
510	12	0	19	2	5	2765
510	12	0	19	2	4	2689
510	12	0	19	2	6	2562
510	13	0	19	2	5	2527
510	12	0	19	2	0	2525

2. **Model Data Setup:**
   1. Data split into features (X) and targets (y)
3. **Training and Testing Sets:**
   1. Typical split ratio applied
   2. Training set for model building
   3. Testing set for model evaluation
4. **Data Saving:**
   1. Processed data saved in **'train_test_split.pkl'**
   2. Preprocessed data for modelling saved in **'top5_crime_pre.csv'**

Model Training and Evaluation

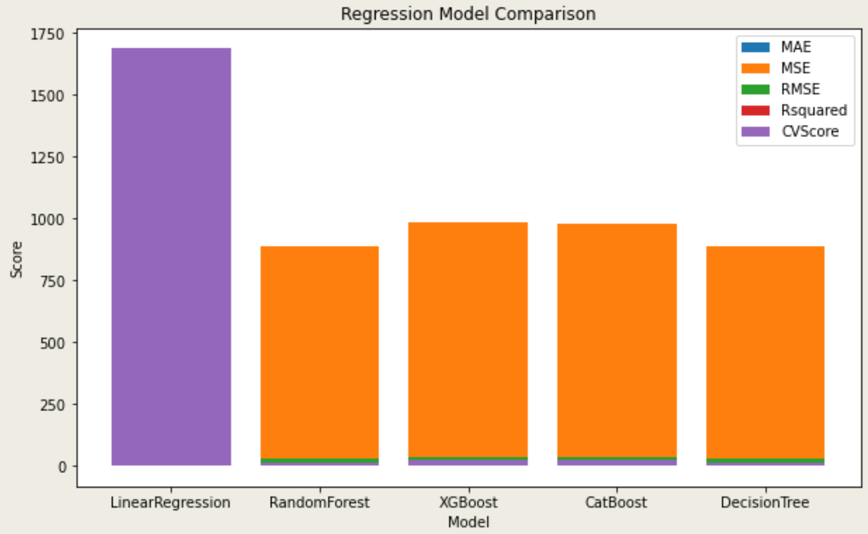
Model Building:

- Models: Linear Regression, Random Forest, XGBoost, CatBoost, Decision Tree
- Metrics: MSE, RMSE, MAE, R-squared, and Cross-Validation Scores

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

Performance Analysis:

1. Linear Regression: High errors, negative R-squared (poor performance)
2. Random Forest: Low errors, high R-squared (good performance)
3. XGBoost: Moderate errors, negative R-squared (average to below-average performance)
4. CatBoost: High errors, negative R-squared (poor performance)
5. Decision Tree: Moderate errors, high R-squared (good performance)

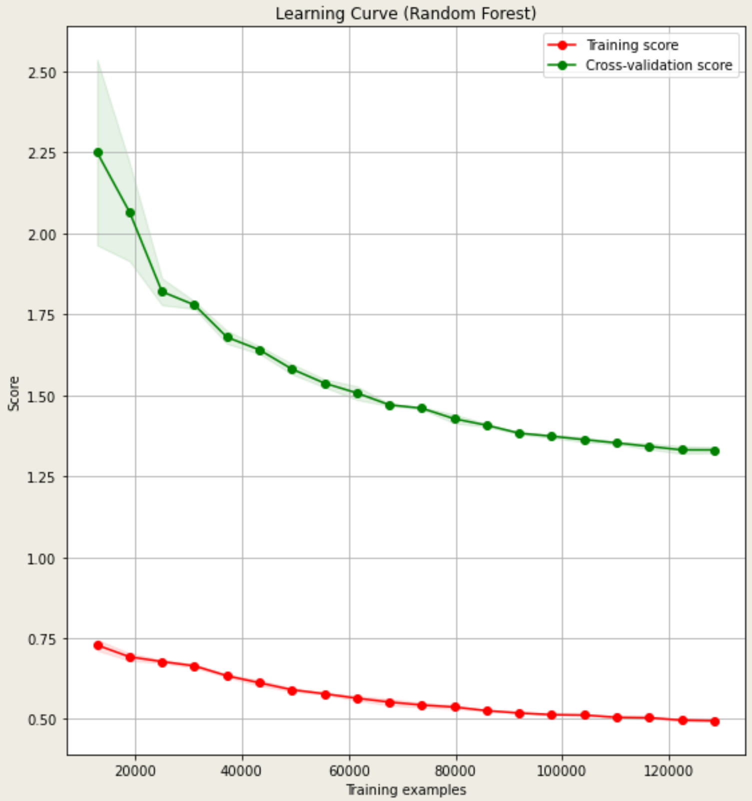


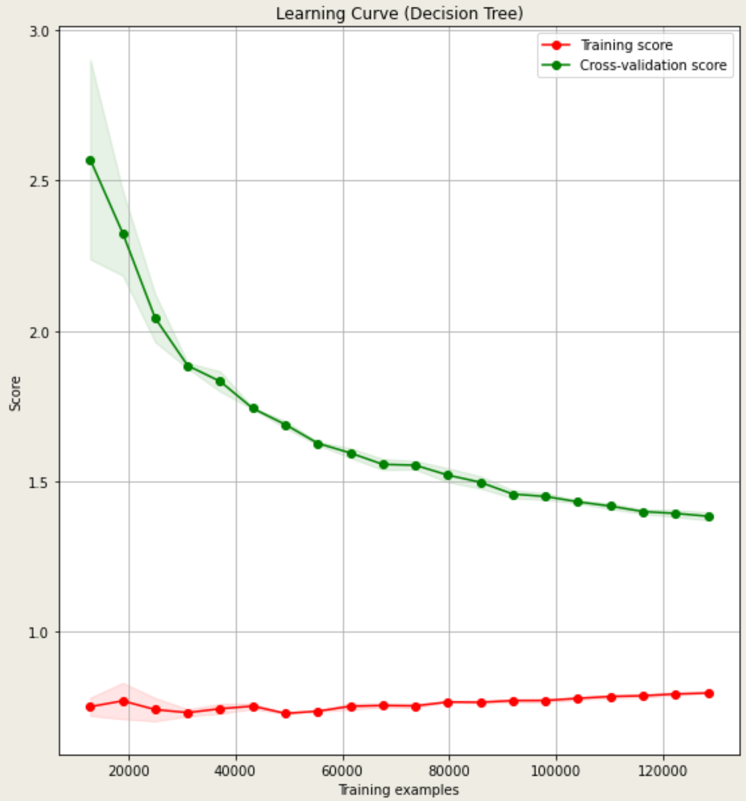
Initial Conclusion:

- Best Performers: Random Forest and Decision Tree
- Random Forest has the lowest MAE and highest R-squared
- Decision Tree has a slightly lower RMSE and the same R-squared as Random Forest

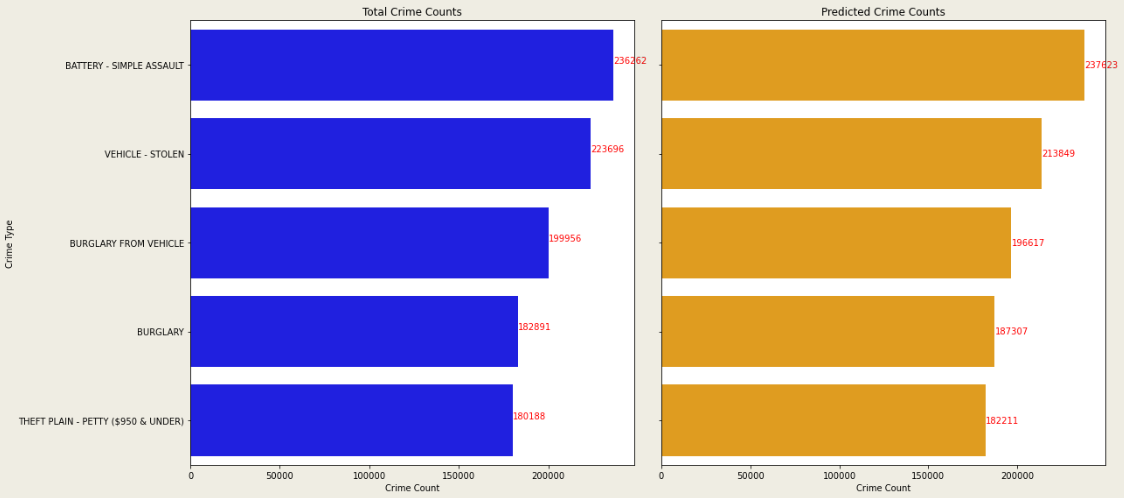
Model Validation and Final Choice

- **Overfitting and Underfitting Analysis:**
  - Learning curves are used to evaluate model generalization
  - **Random Forest Analysis:**Learning Curve: Training and CV scores are consistent, suggesting no overfitting
  - **Decision Tree Analysis:**Learning Curve: High initial training score (potential overfitting) but improves with more data, indicating good generalization
- **Final Model Selection:**
  - Decision Tree is selected based on its improving cross-validation score and ability to generalize with more data





Final Model Fitting and Feature Importance



Feature Importance:

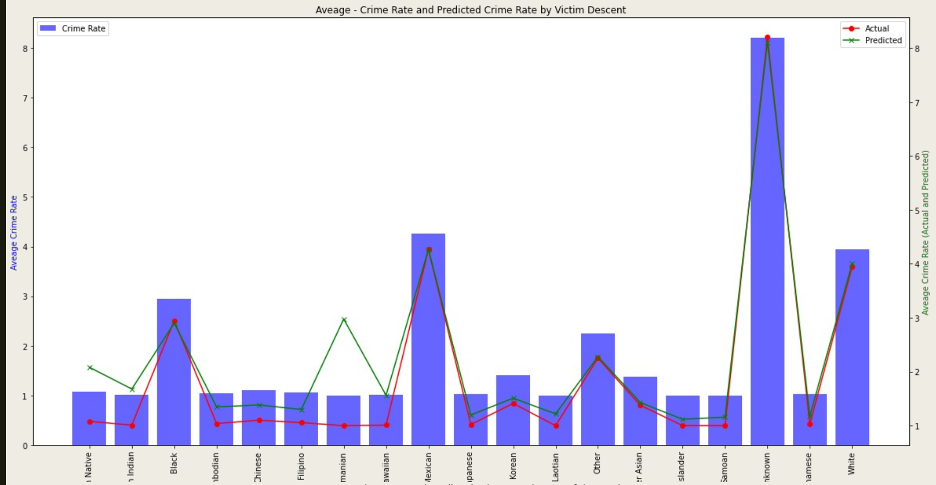
1. Key Predictive Features: Victim descent and crime type
2. High Importance: Indicating these features greatly influence crime rate predictions

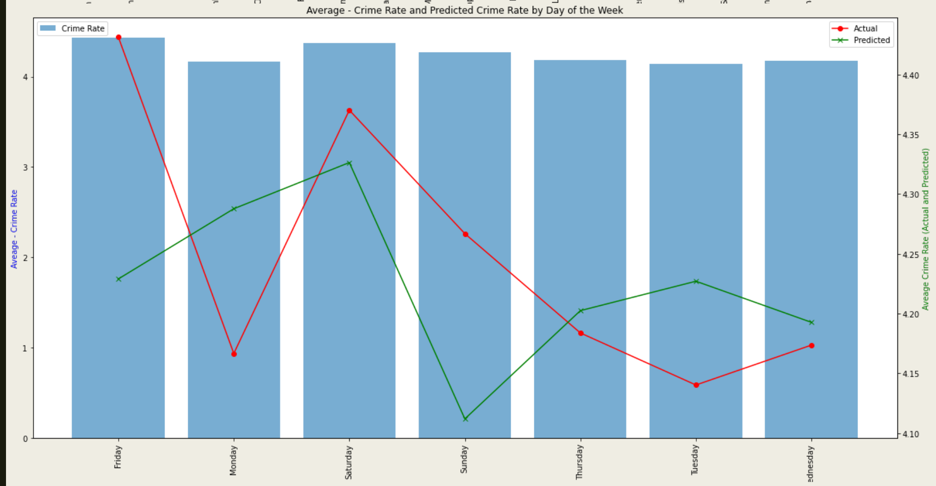
Model Prediction Accuracy by Crime Type:

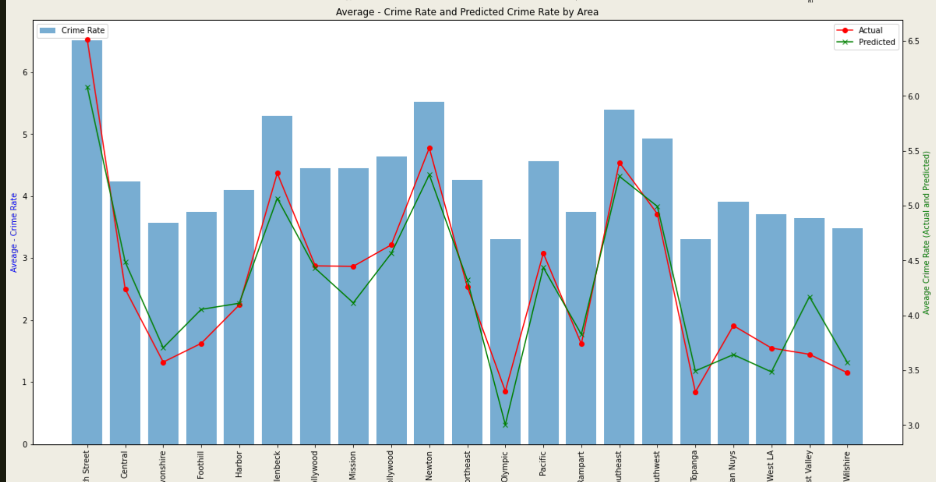
1. Battery - Simple Assault: Predictions slightly higher than actual rates
2. Burglary: Predictions slightly higher than actual
3. Burglary from Vehicle: Predictions slightly lower than actual
4. Theft Plain - Petty: Predictions slightly higher than actual
5. Vehicle - Stolen: Predictions significantly lower, indicating an area for improvement

Predictive Performance
by Descent, Area, and Day of the Week

1. **Accuracy Across Descent Groups:**
   - Overestimations: American Indian/Alaskan Native, Asian Indian, Guamanian
   - Close Predictions: Hispanic/Latin/Mexican
   - Underestimation: "Unknown" descent group
2. **Actual vs. Predicted Crime Rates by Area:**
   - Higher Actual Rates: 77th Street, Hollenbeck, Mission, Newton, Van Nuys
   - Higher Predicted Rates: Central, Foothill, Olympic, West Valley
3. **Prediction Variance by Day of the Week:**
   - Higher Actual Weekend Rates: Fridays and Sundays
   - Alternating Over and Under Predictions: During weekdays







General Trends and Future Improvements

- Model trends to overpredict for certain descent groups
- Notable underprediction for "Unknown" descent and Vehicle Theft
- Hispanic/Latin/Mexican descent group consistently targeted
- Limitations: Insufficient RAM, leading to data underutilization
- Suggestion: Investment in computational resources for better performance

THANK YOU

Q & A