BOSTON UNIVERSITY

CS 777: Big Data Analytics
LAPD CRIME ANALYSIS
Term Project

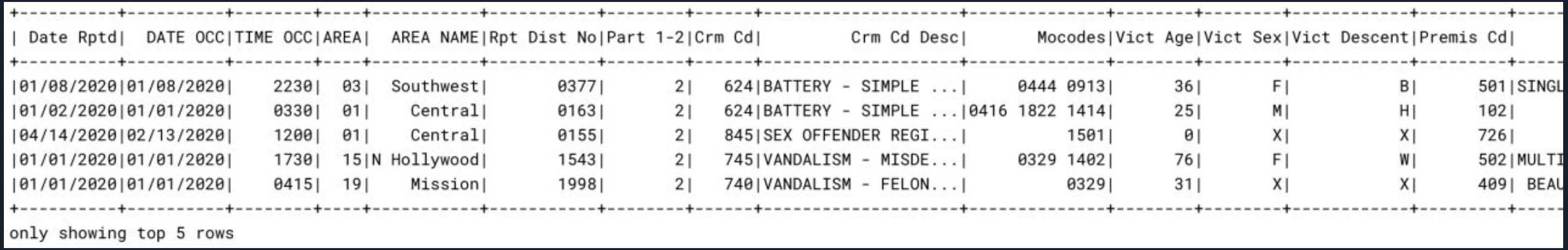
Fall 2023

Aditya Maheshwari
Sarthak Pattnaik
Vaidehi Shah

The Dataset

The LAPD police report dataset provides a comprehensive overview of crime incidents in Los Angeles from 2020 onward.

Originating from original crime reports transcribed from paper documents, potential data inaccuracies may exist due to the manual transcription process.



```
+----------+----------+--------+----+-----------+-----------+--------+------+--------------------+--------------+--------+--------+------------+---------+-----
| Date Rptd|  DATE OCC|TIME OCC|AREA|  AREA NAME|Rpt Dist No|Part 1-2|Crm Cd|         Crm Cd Desc|       Mocodes|Vict Age|Vict Sex|Vict Descent|Premis Cd|
+----------+----------+--------+----+-----------+-----------+--------+------+--------------------+--------------+--------+--------+------------+---------+-----
|01/08/2020|01/08/2020|    2230|  03|  Southwest|       0377|       2|   624|BATTERY - SIMPLE ...|     0444 0913|      36|       F|           B|      501|SINGL
|01/02/2020|01/01/2020|    0330|  01|    Central|       0163|       2|   624|BATTERY - SIMPLE ...|0416 1822 1414|      25|       M|           H|      102|
|04/14/2020|02/13/2020|    1200|  01|    Central|       0155|       2|   845|SEX OFFENDER REGI...|          1501|       0|       X|           X|      726|
|01/01/2020|01/01/2020|    1730|  15|N Hollywood|       1543|       2|   745|VANDALISM - MISDE...|     0329 1402|      76|       F|           W|      502|MULTI
|01/01/2020|01/01/2020|    0415|  19|    Mission|       1998|       2|   740|VANDALISM - FELON...|          0329|      31|       X|           X|      409| BEAU
+----------+----------+--------+----+-----------+-----------+--------+------+--------------------+--------------+--------+--------+------------+---------+-----
only showing top 5 rows
```

Data Cleaning and Processing

1\.

Column Handling:

- Dropped irrelevant columns.
- Renamed for clarity.
- Trimmed off extra white spaces.

2\.

Formatting:

- Date columns standardized.
- ‘LOCATION' and 'Cross Street' merged.
- Mapped 'VictDescent' for interpretability.

3\.

Handling Nulls:

- 'PremisCd' or 'PremisDesc' null rows dropped.

Geospatial Analysis

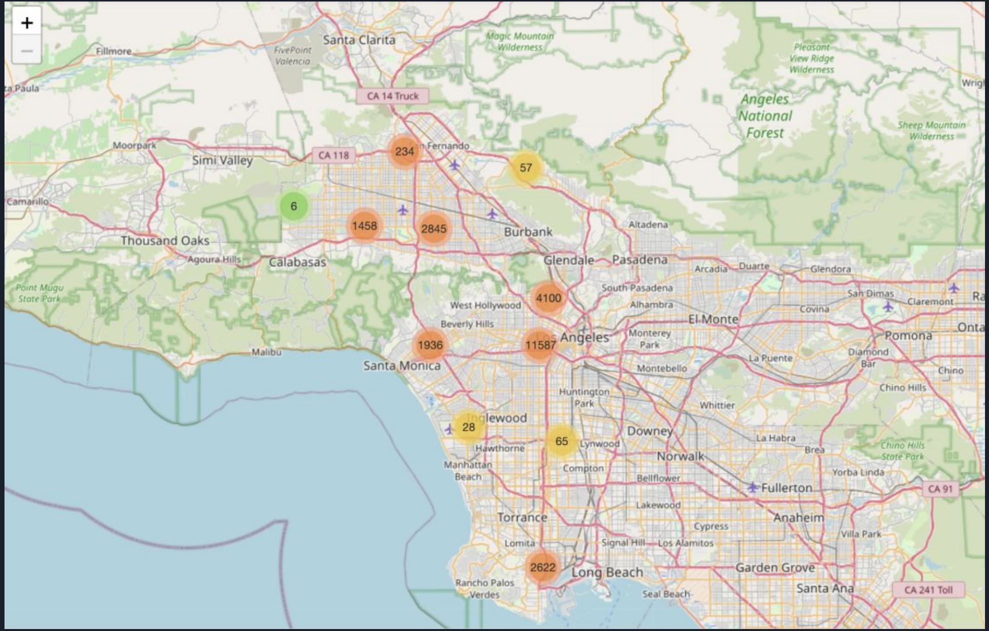
1. Extracted geographical coordinates (latitude and longitude).

2. Downloaded the geojson file for LA.

3. Plotted crime locations on the map, with each point representing an incident.

4. With the help of this geospatial map, we observed that Central LA has the highest concentration of crimes.

5. Generated heatmaps to visually represent the intensity of crime in different areas.



Geospatial Analysis

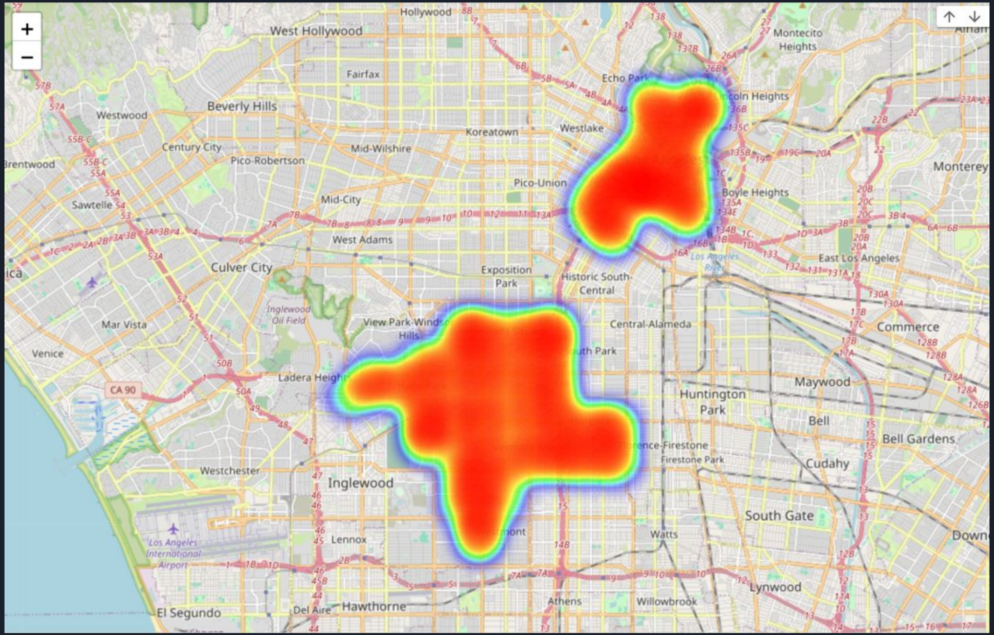
1. Extracted geographical coordinates (latitude and longitude).

2. Downloaded the geojson file for LA.

3. Plotted crime locations on the map, with each point representing an incident.

4. With the help of this geospatial map, we observed that Central LA has the highest concentration of crimes.

5. Generated heatmaps to visually represent the intensity of crime in different areas.



Integration with PostgreSQL and Tableau

- Leveraged the Python library psycopg2 to seamlessly transfer our data into PostgreSQL database.

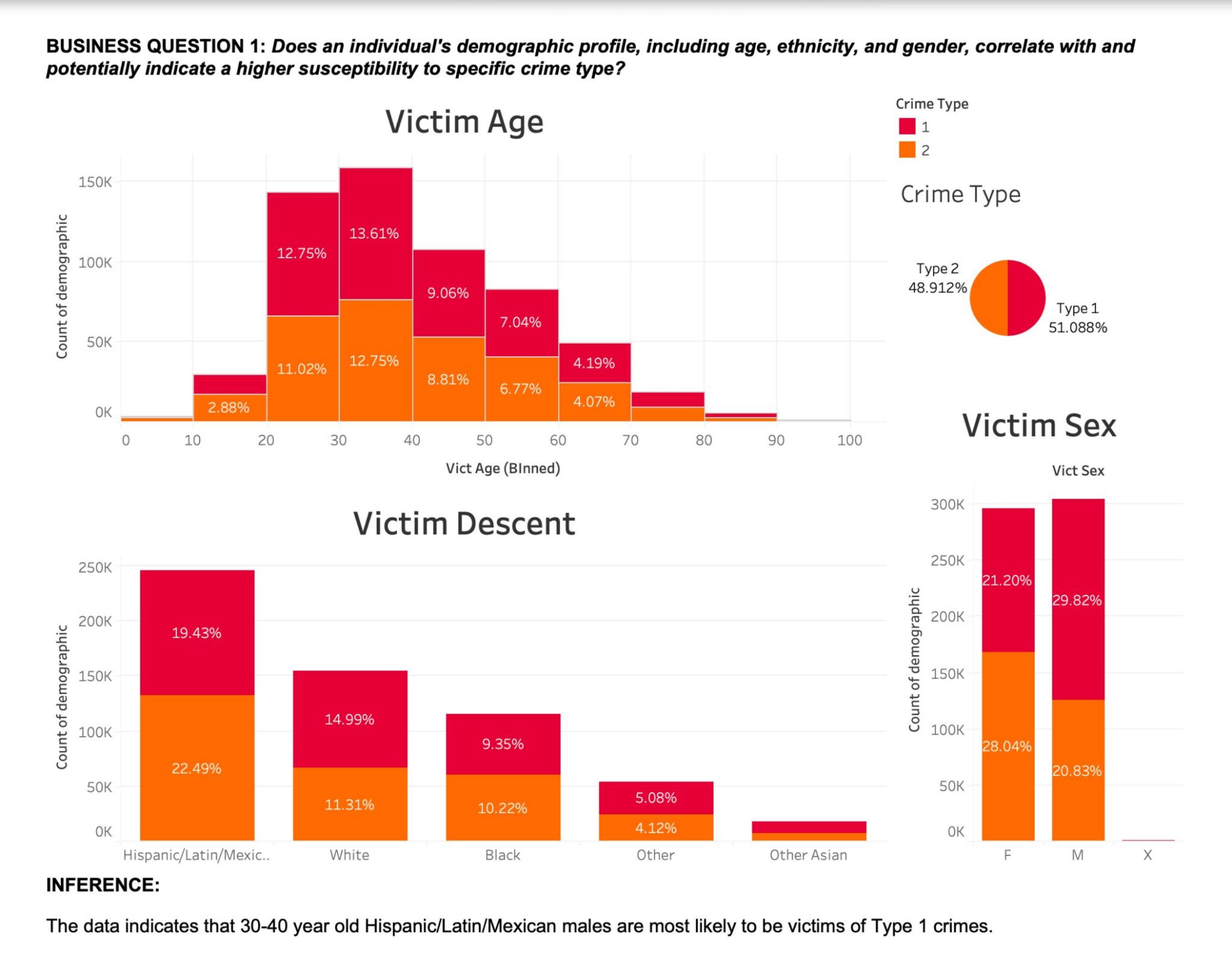
- The primary motivation was to establish a connection between Tableau and our dataset where PostgreSQL was a required intermediate. This integration facilitates diverse visualizations in alignment with our three key business questions.

Business Questions & Tableau Dashboards

Investigated whether an ***individual's demographic profile***, encompassing age, ethnicity, and gender, correlates with and potentially indicates a ***higher susceptibility to specific crime types***.

Explored how crime data can be leveraged to understand the ***most prevalent crimes and crime*** rates based on ***geographic areas***.

Examined the relationship between ***the time of occurrence and the time of reporting*** for different types of crimes, seeking to uncover any dependencies.



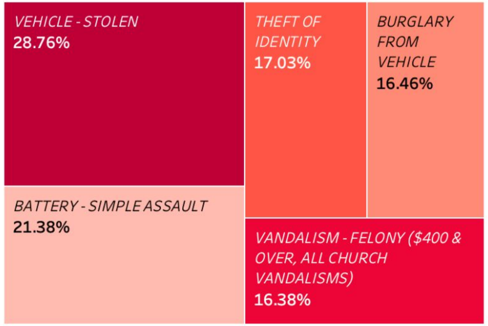
BUSINESS QUESTION 1: ***Does an individual's demographic profile, including age, ethnicity, and gender, correlate with and potentially indicate a higher susceptibility to specific crime type?***

INFERENCE:

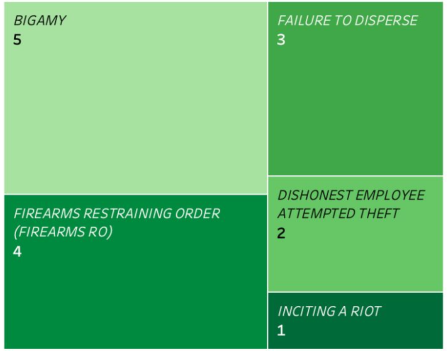
The data indicates that 30-40 year old Hispanic/Latin/Mexican males are most likely to be victims of Type 1 crimes.

BUSINESS QUESTION 2: How to levarage crime data to understand most prevelant crimes and crime rates based on the areas?

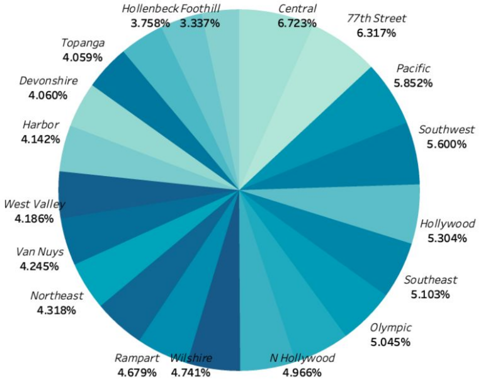
Top 5 Crimes Committed



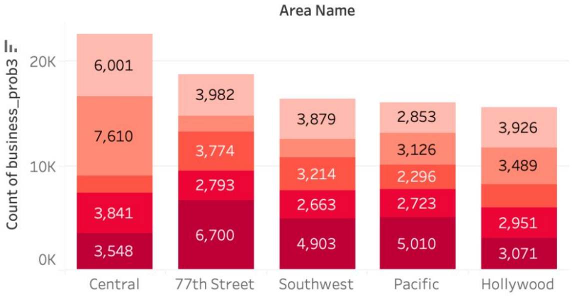
Least 5 Crimes Committed



Area-Wise Incidence Rates



Crime Hotspots: Top 5 Offenses Across Key Areas

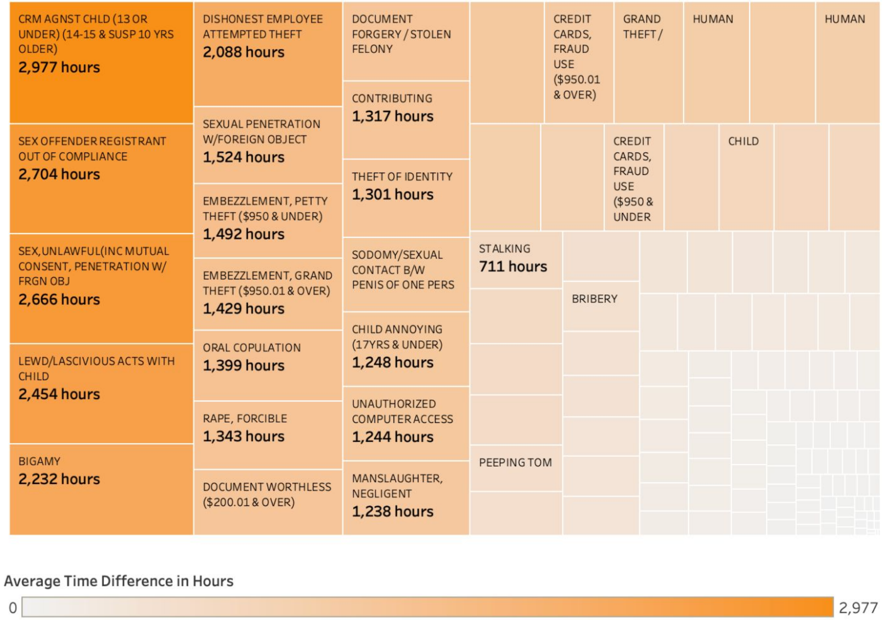


INFERENCE:

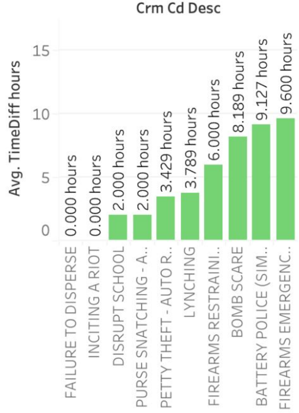
Stolen vehicles dominate crime in Los Angeles, notably in Central LA, where property crimes like burglary from vehicles are on the rise. Despite lower overall crime rates in Foothill, the challenge of stolen vehicles persists citywide. The rarity of inciting a riot suggests a generally stable societal environment.

BUSINESS PROBLEM 3: Is there a relation between the difference in the time of occurrence and report of a crime dependent on the crime?

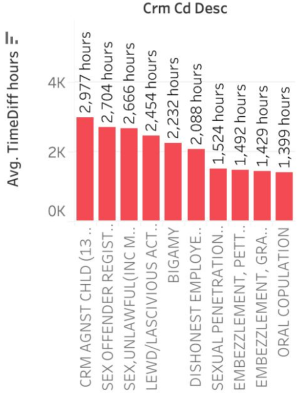
Temporal Discrepancy Heatmap: Occurrence vs. Reporting Time



Analyzing Minimum Crime Reporting Times



Analyzing Maximum Crime Reporting Times



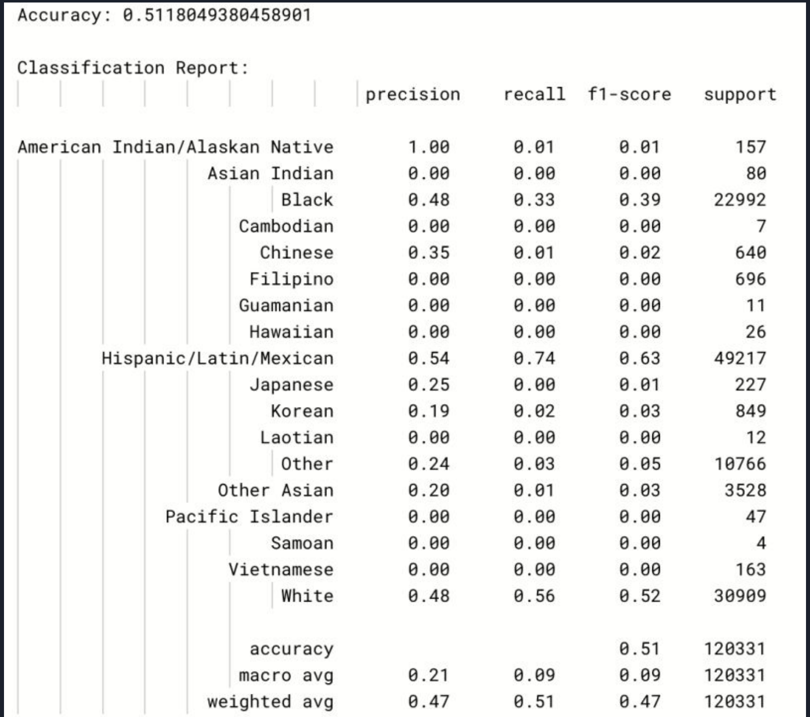
INFERENCE:

Crimes against children and sex offenses, requiring more time for reporting, underscore potential barriers in victim disclosure or societal reluctance. Conversely, swift reporting of school disruptions and purse snatching suggests heightened awareness and prompt community responsiveness to immediate threats

LA Crime Statistics Tableau Dashboard

https://public.tableau.com/LACrimeAnalytics

Machine Learning for Demographic Predictions using GCP

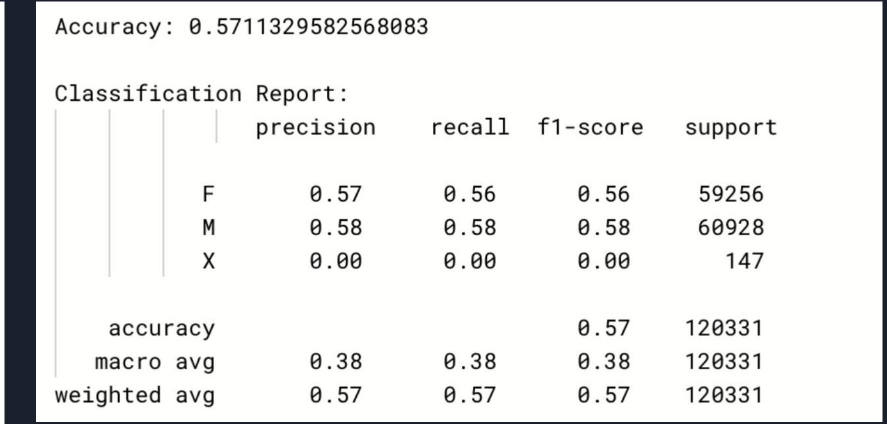


```
Accuracy: 0.5118049380458901

Classification Report:
                              precision    recall  f1-score   support

American Indian/Alaskan Native      1.00      0.01      0.01       157
                  Asian Indian      0.00      0.00      0.00        80
                         Black      0.48      0.33      0.39     22992
                     Cambodian      0.00      0.00      0.00         7
                       Chinese      0.35      0.01      0.02       640
                      Filipino      0.00      0.00      0.00       696
                     Guamanian      0.00      0.00      0.00        11
                      Hawaiian      0.00      0.00      0.00        26
        Hispanic/Latin/Mexican      0.54      0.74      0.63     49217
                      Japanese      0.25      0.00      0.01       227
                        Korean      0.19      0.02      0.03       849
                       Laotian      0.00      0.00      0.00        12
                         Other      0.24      0.03      0.05     10766
                   Other Asian      0.20      0.01      0.03      3528
              Pacific Islander      0.00      0.00      0.00        47
                        Samoan      0.00      0.00      0.00         4
                    Vietnamese      0.00      0.00      0.00       163
                         White      0.48      0.56      0.52     30909

                      accuracy                          0.51    120331
                     macro avg      0.21      0.09      0.09    120331
                  weighted avg      0.47      0.51      0.47    120331
```

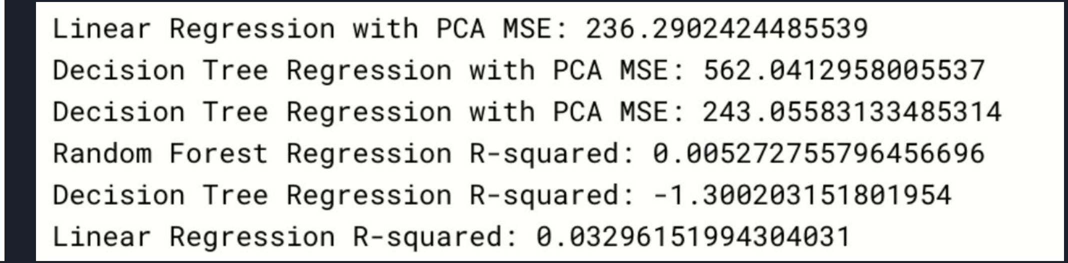


```
Accuracy: 0.5711329582568083

Classification Report:
              precision    recall  f1-score   support

           F       0.57      0.56      0.56     59256
           M       0.58      0.58      0.58     60928
           X       0.00      0.00      0.00       147

    accuracy                           0.57    120331
   macro avg       0.38      0.38      0.38    120331
weighted avg       0.57      0.57      0.57    120331
```



```
Linear Regression with PCA MSE: 236.2902424485539
Decision Tree Regression with PCA MSE: 562.0412958005537
Decision Tree Regression with PCA MSE: 243.05583133485314
Random Forest Regression R-squared: 0.005272755796456696
Decision Tree Regression R-squared: -1.300203151801954
Linear Regression R-squared: 0.03296151994304031
```

Conclusion

- Successfully conducted comprehensive analysis of LAPD crime dataset to uncover insights into crime trends, demographics, and temporal patterns

- Developed interactive Tableau dashboard to enable data-driven decision making for law enforcement

- Achieved 50-60% accuracy in predicting victim demographics using machine learning models

- Identified opportunities to optimize resource allocation and enhance public safety strategies based on analysis

Future Scope

- Refining Machine Learning Models
- Exploration of Deep-Learning Integration
- Continuous Data Collection for Trend Detection