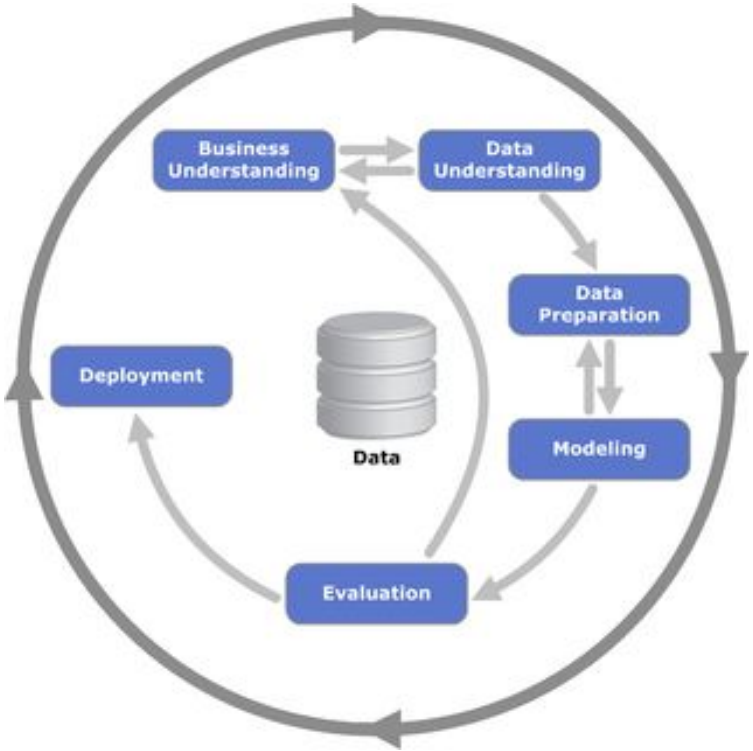
DATA-602 Final Project

Predicting Crime Types based on Historical Data: A Machine Learning Approach

By:
Kalaiarasi Kaliappan

Project Approach:

CRISP-DM(Cross-Industry Standard Process for Data Mining)



BU: Understand proj from business perspective. Understand the true goal
DU: Collect, describe and explore the data
DP: Select, clean, construct and integrate for the next step.
Modeling: Select, clean, create and Assess the model.
Evaluation: Evaluate and review the results and determine the next step.
Deployment: Plan deployment and monitoring.

Business Problem:

The dataset Crime Data contains details regarding crimes that happened in various cities around the City of Los Angeles between January 2020 and the present day. The dataset has attributes for the location, date, and time of the offense as well as the type of crime that was committed. Based on the location, date, sex, race and time of the incident, a machine-learning model can be created using the dataset to forecast the sort of crime that will be committed.

The goal of this project is to create a machine-learning model that can identify the kind of crime being committed based on the location, the time and day it occurs, and other information. This can help law enforcement officials effectively utilize their resources and prevent crimes from happening in the future.

About Dataset:

Data Source: https://catalog.data.gov/dataset/crime-data-from-2020-to-present

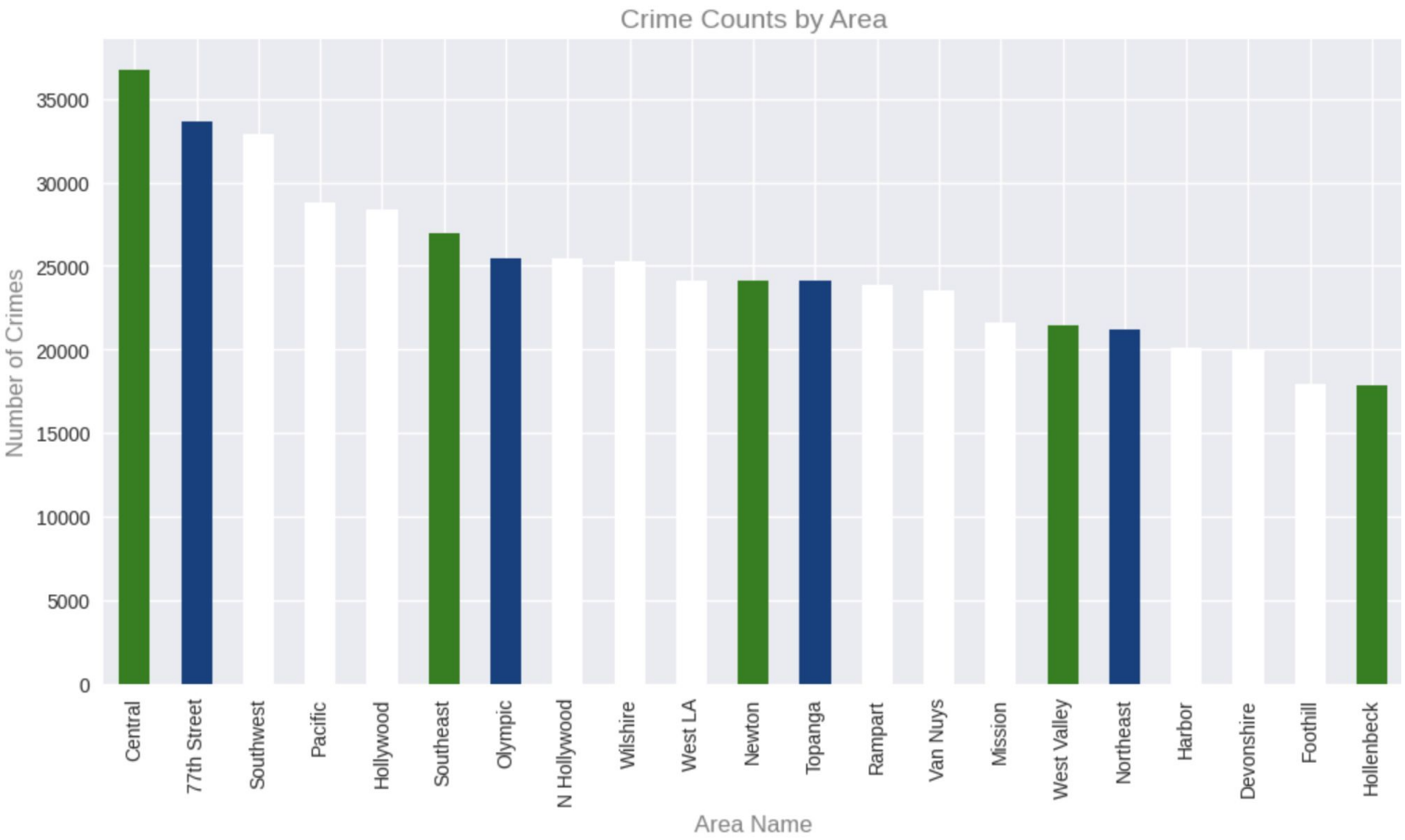
Quantity: 694,915 Records

Columns: 28

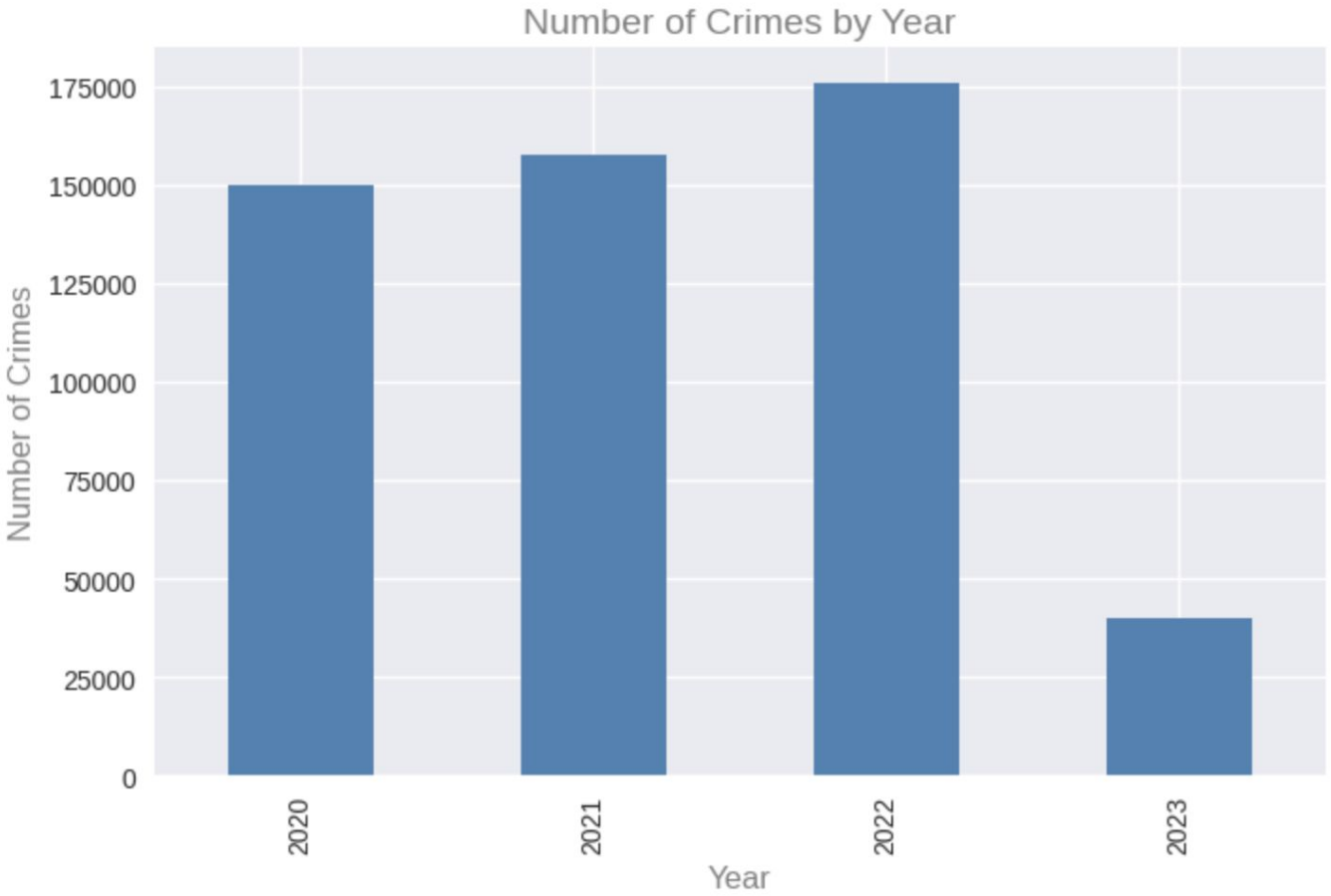
Key Attributes: Location, Vict Age, Vict Sex, Vict Race, Data_Occ, Crime description, Premis description.

Exploratory data analysis

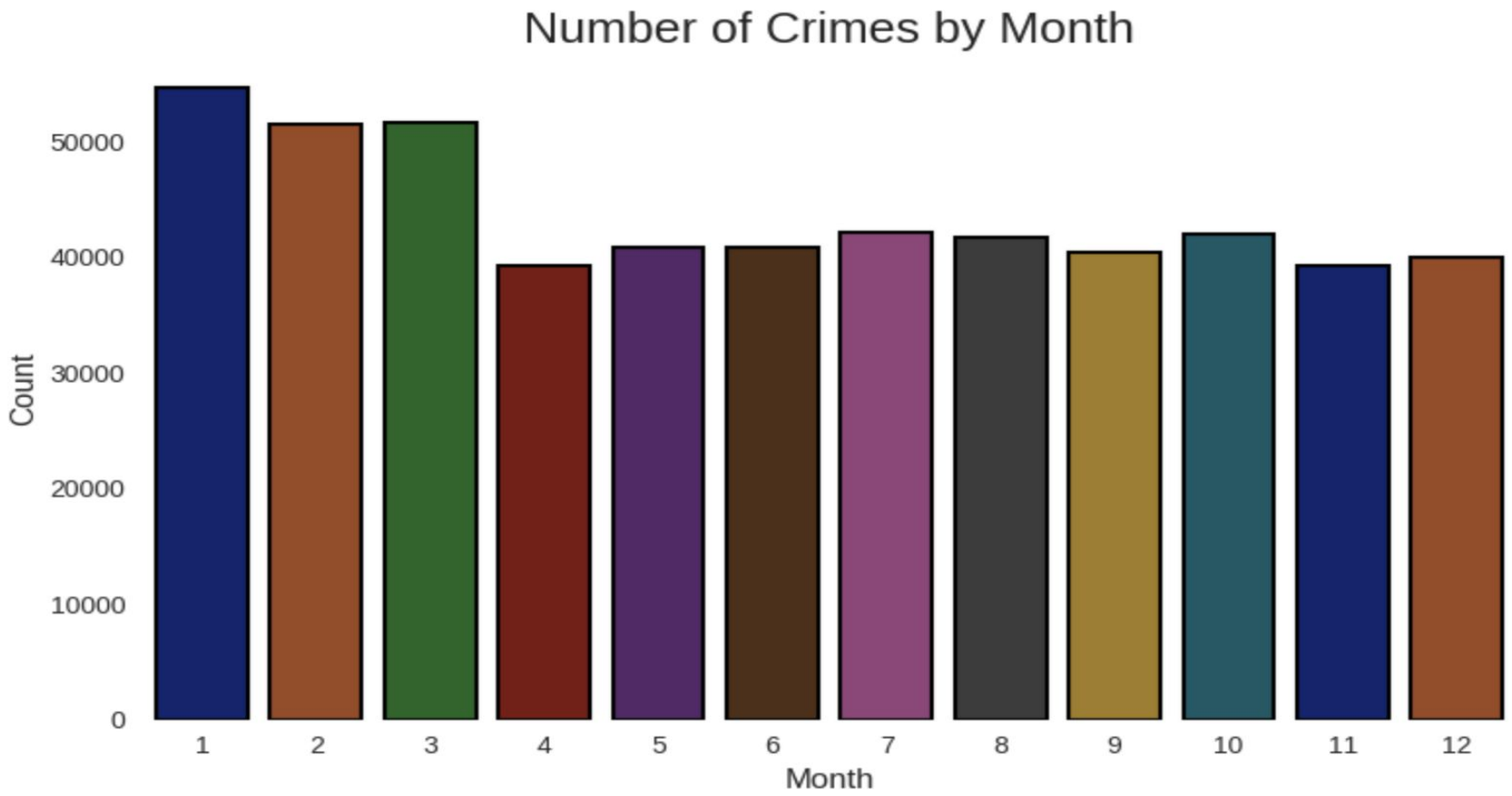
Crime Counts By Area

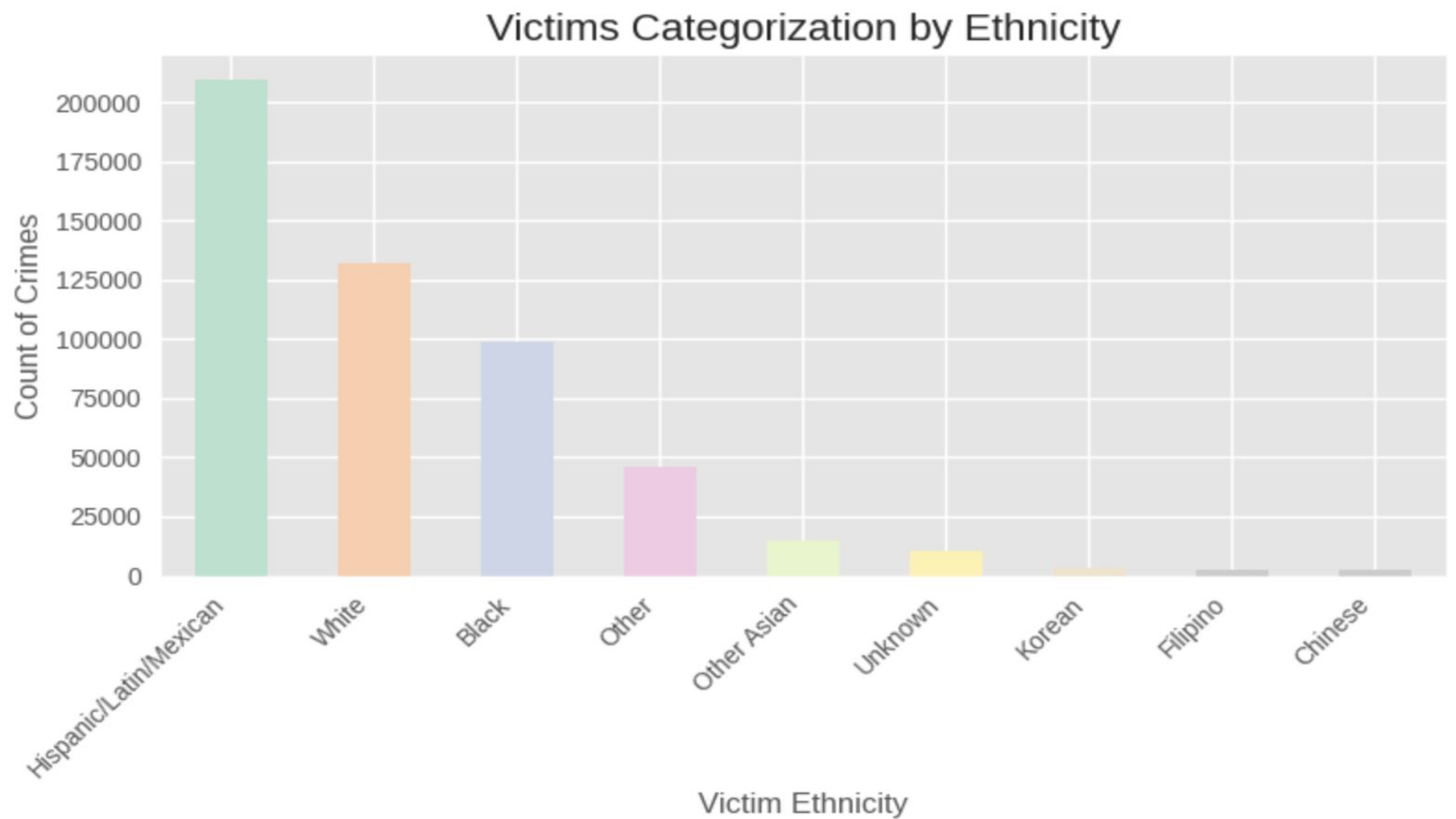


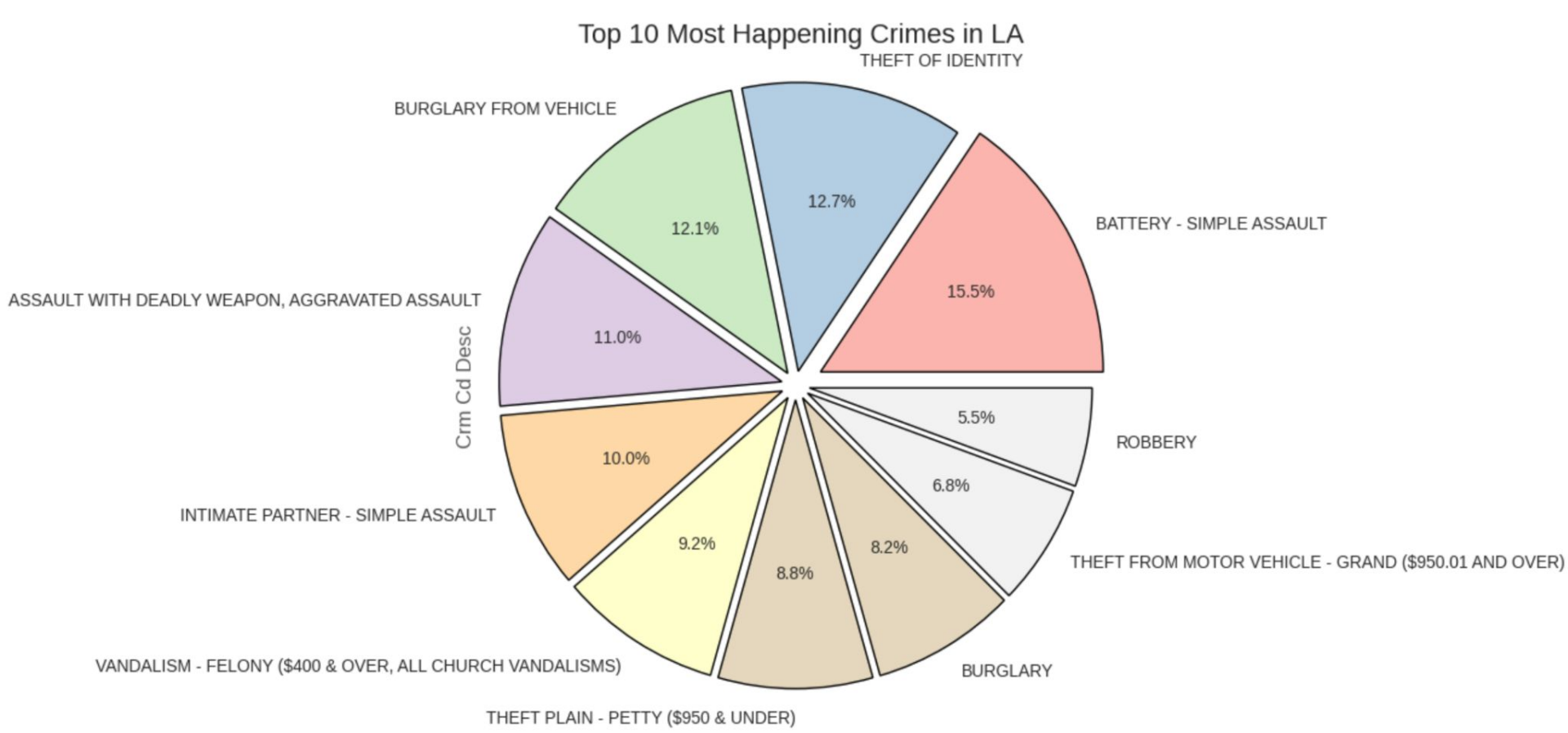
Number of Crimes by Year



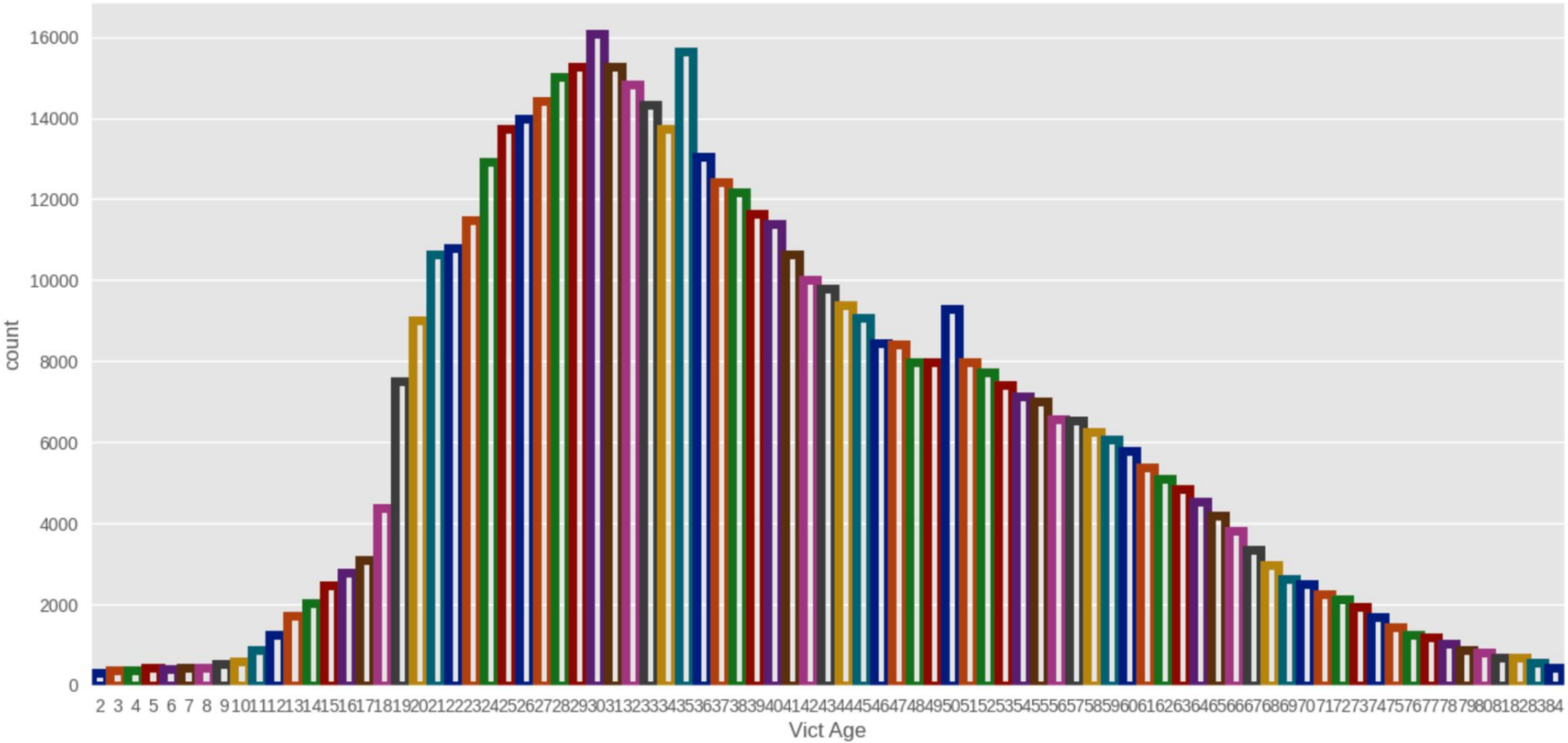
Number of Crimes by Month:

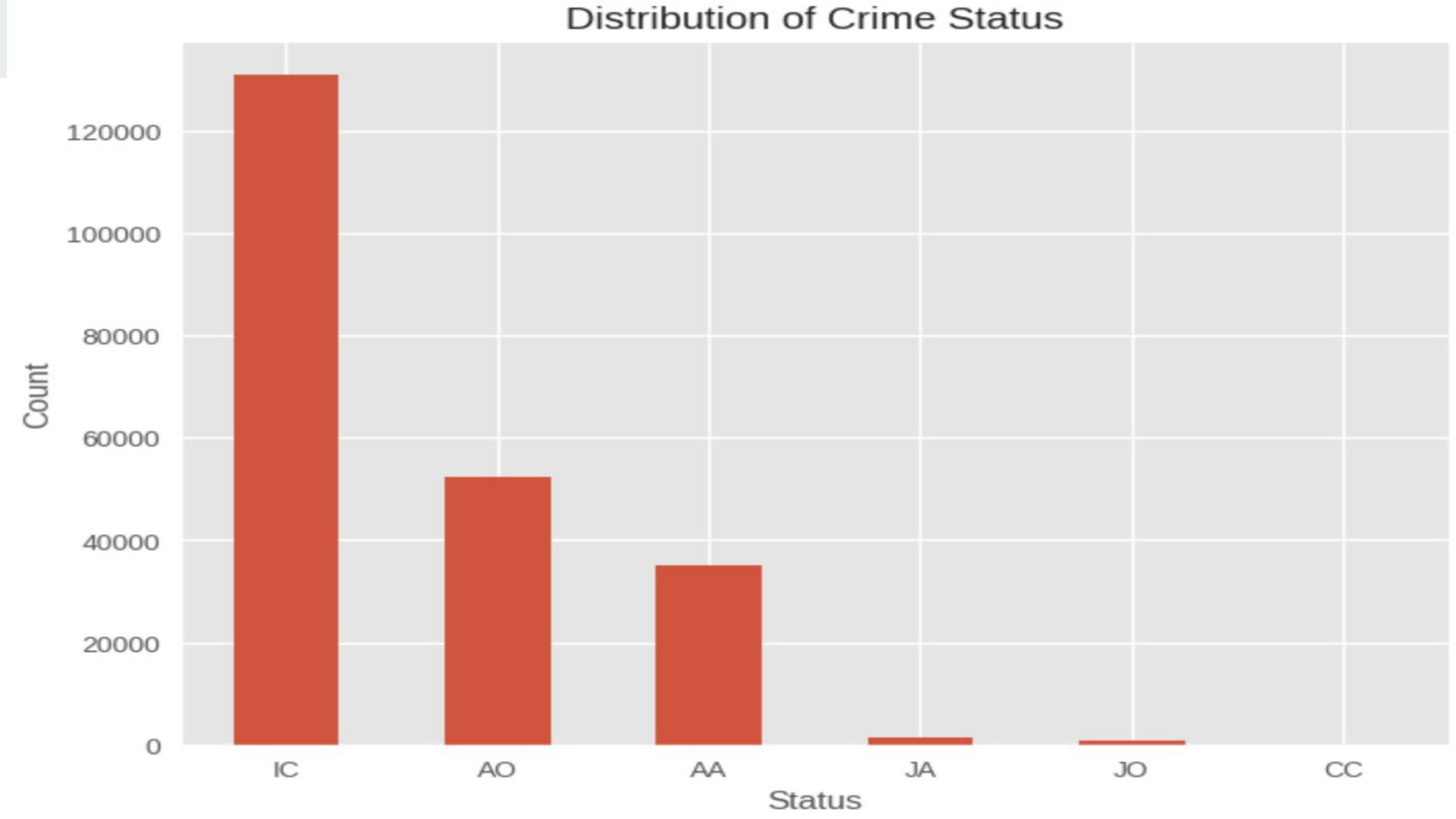



Victims Categorization by Ethnicity
Count of Crimes
200000
175000
150000
125000
100000
75000
50000
25000
0
Hispanic/Latin/Mexican
White
Black
Other
Other Asian
Unknown
Korean
Filipino
Chinese
Victim Ethnicity

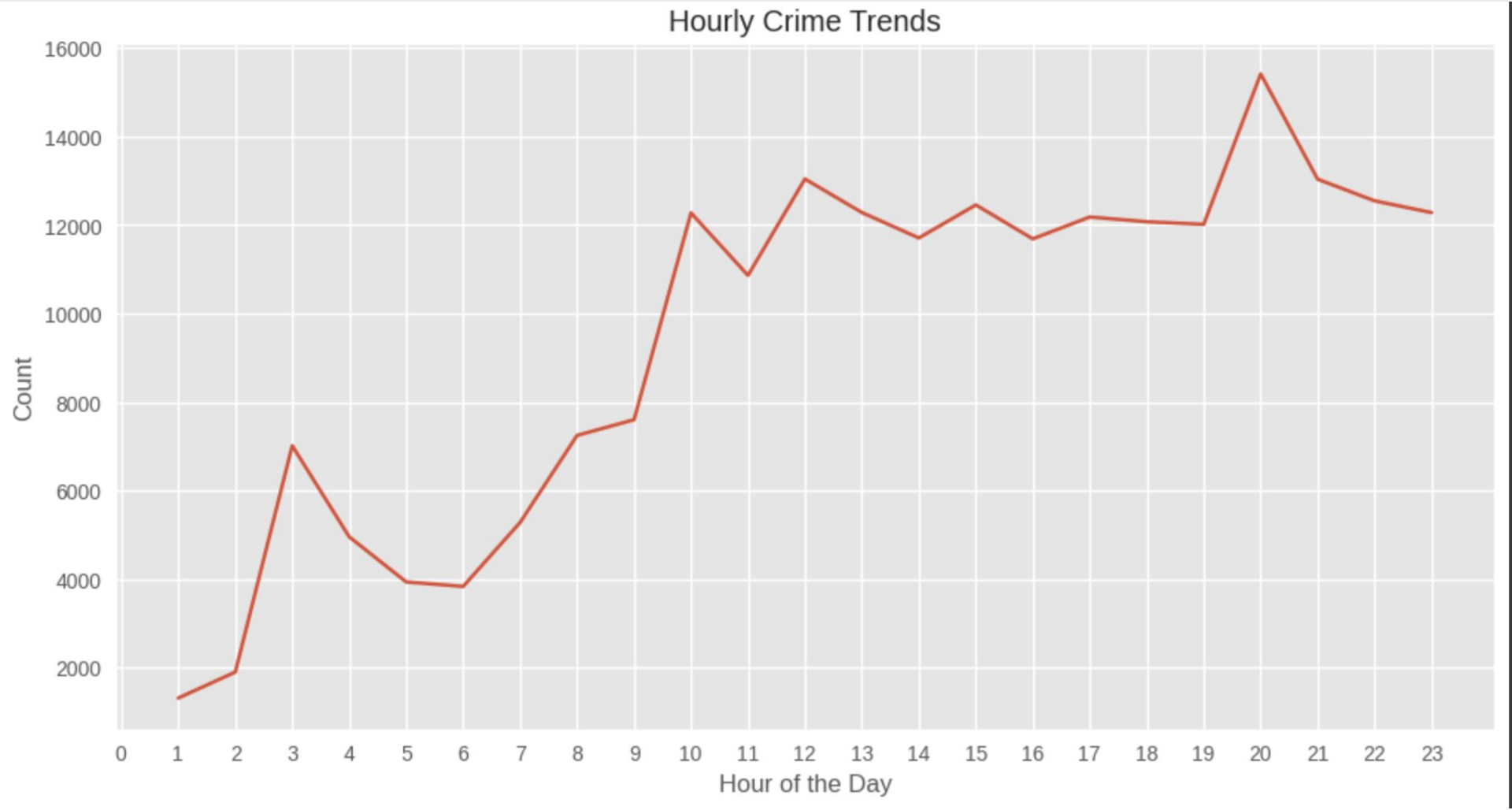

Top 10 Most Happening Crimes in LA
THEFT OF IDENTITY
BURGLARY FROM VEHICLE
BATTERY - SIMPLE ASSAULT
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
Crm Cd Desc
ROBBERY
INTIMATE PARTNER - SIMPLE ASSAULT
THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
BURGLARY
THEFT PLAIN - PETTY ($950 & UNDER)
12.7%
12.1%
15.5%
11.0%
5.5%
10.0%
6.8%
9.2%
8.2%
8.8%

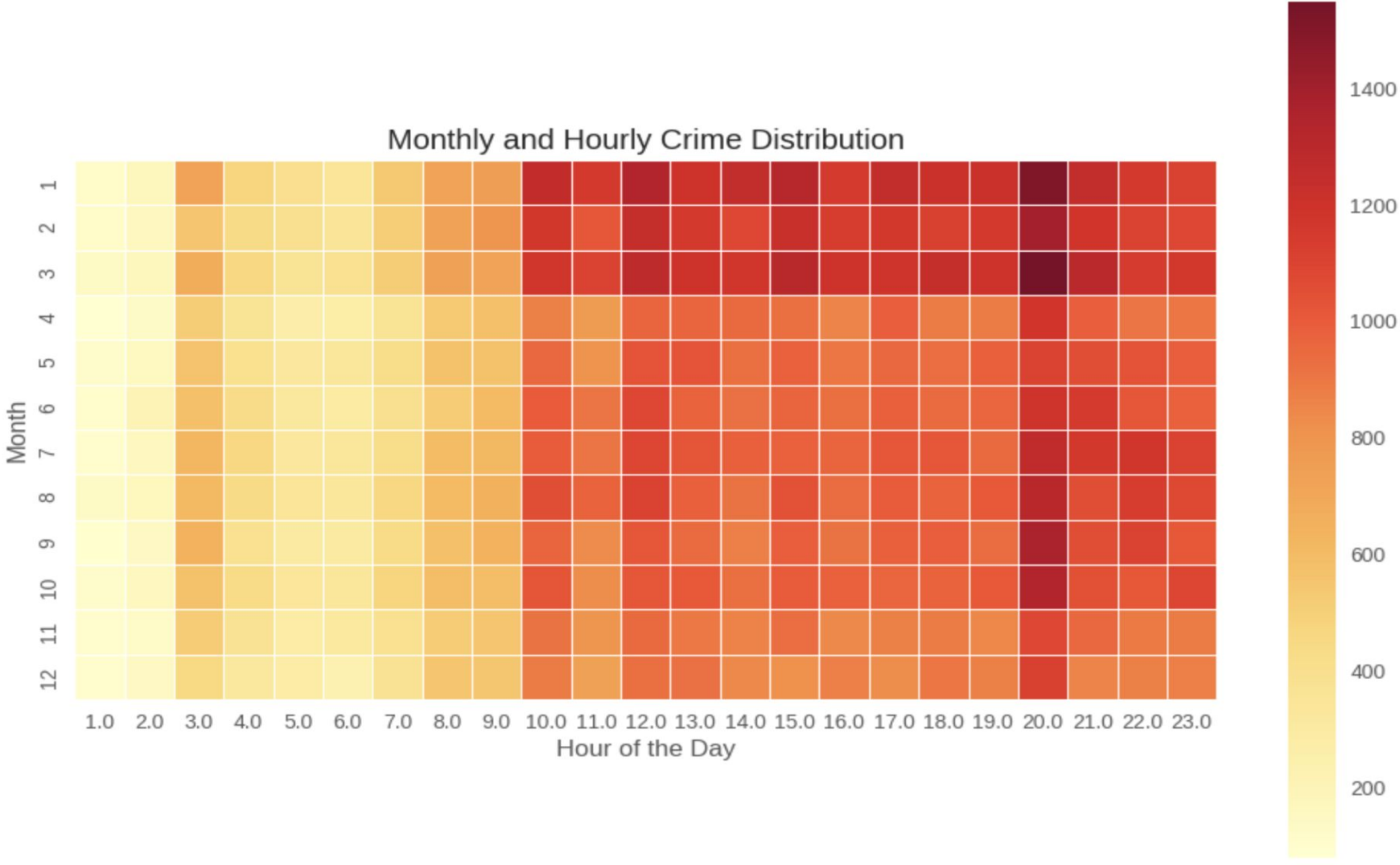
Affected victim age distribution




Distribution of Crime Status
120000
100000
80000
60000
40000
20000
0
Count
IC
AO
AA
JA
JO
CC
Status


Distribution of Gender of Victims
100000
80000
60000
40000
20000
0
Count
M
F
X
H
Gender


Hourly Crime Trends
Count
16000
14000
12000
10000
8000
6000
4000
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of the Day


Monthly and Hourly Crime Distribution
Month
1
2
3
4
5
6
7
8
9
10
11
12
1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0
Hour of the Day
1400
1200
1000
800
600
400
200

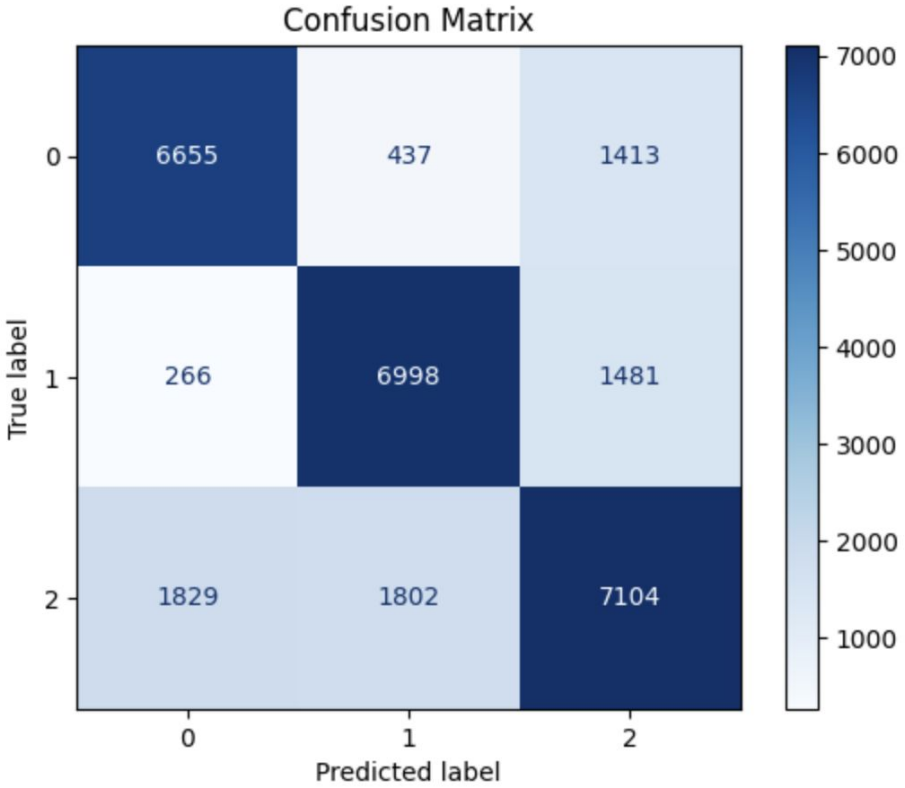
Modeling:

Type of Modeling: Classification

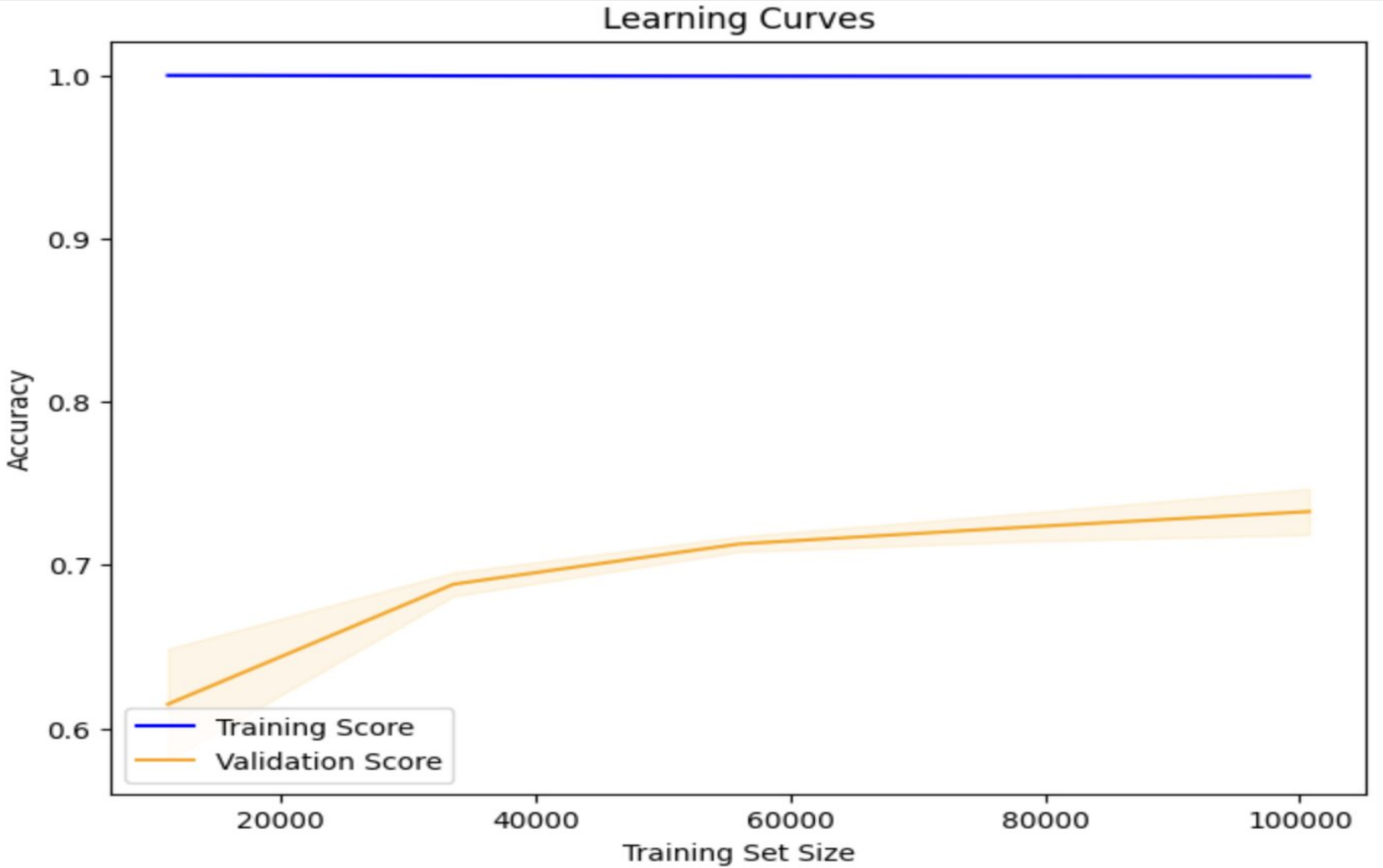
Features: Area, Vict_Age, Vict_Sex, Vict_descent, Premis cd, Time_occ, Day_of_week

Target Variable: crm cd

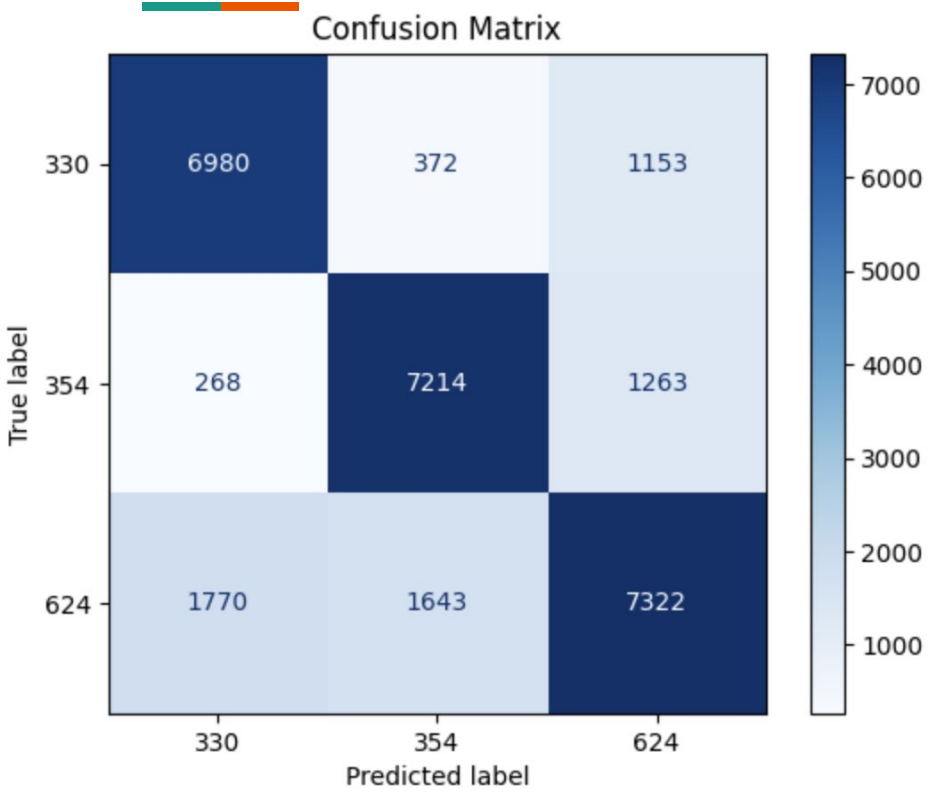
Random Forest Classification:



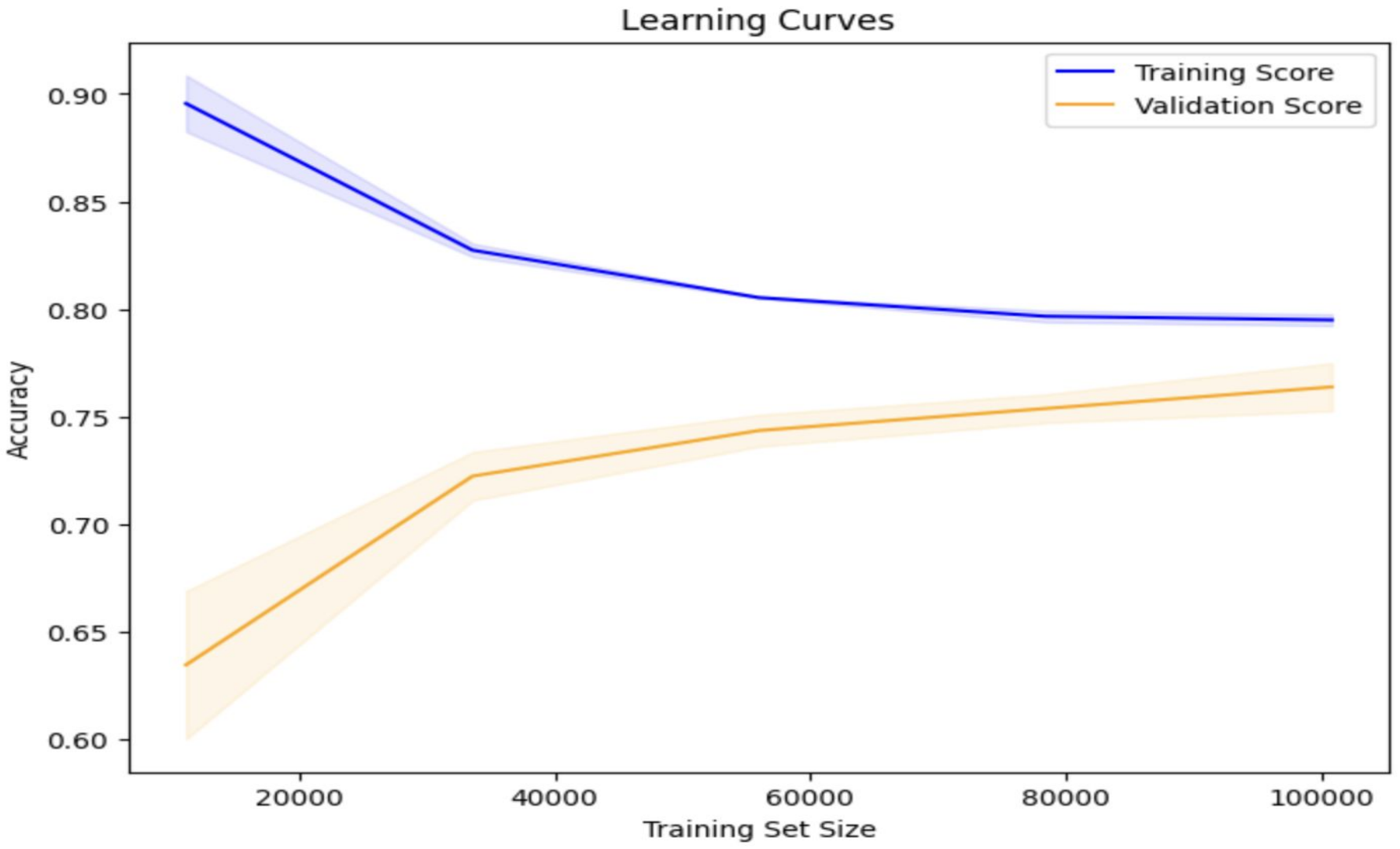

```
Accuracy: 0.742
```


Learning Curves
1.0
0.9
0.8
0.7
0.6
Accuracy
Training Score
Validation Score
20000
40000
60000
80000
100000
Training Set Size

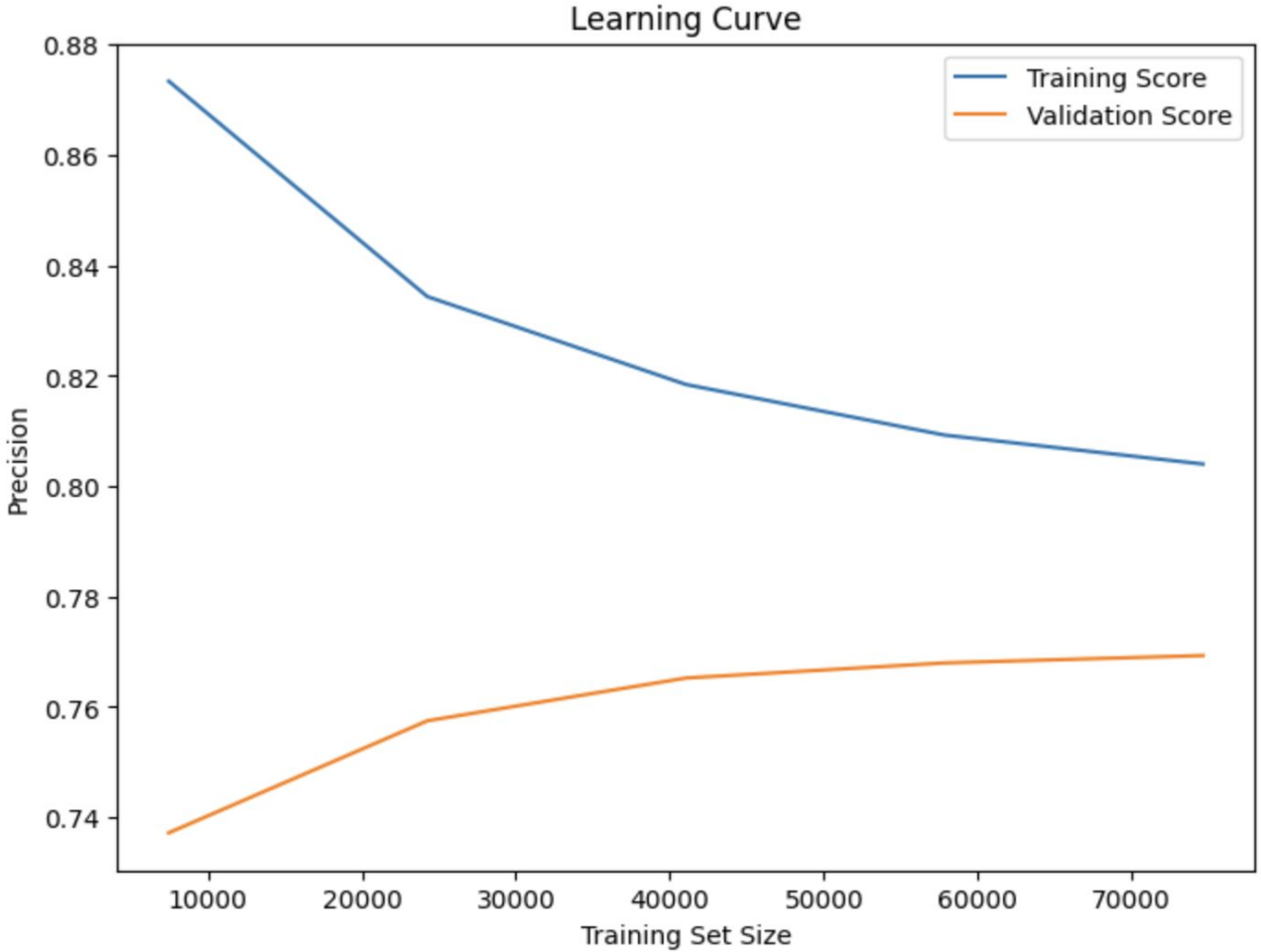
XGBoost (Extreme Gradient Boosting)




```
Accuracy: 0.76
```


Learning Curves
Training Score
Validation Score
0.90
0.85
0.80
0.75
0.70
0.65
0.60
Accuracy
20000
40000
60000
80000
100000
Training Set Size

Stacking:



Precision:
0.7701786972292215

Advantages of Random Forest and XGboost model:

- Ensemble Learning (capture more complex relationships and reduce the impact of individual tree errors)
- Handling Non-linearity (handle non-linear relationships between features and the target variable)
- Robust to Overfitting
- Feature Importance
- Handling Imbalanced Data
- Flexibility and Tunability

Thank You!!!