Predicting Prison Sentences

Capstone 2: Supervised Learning

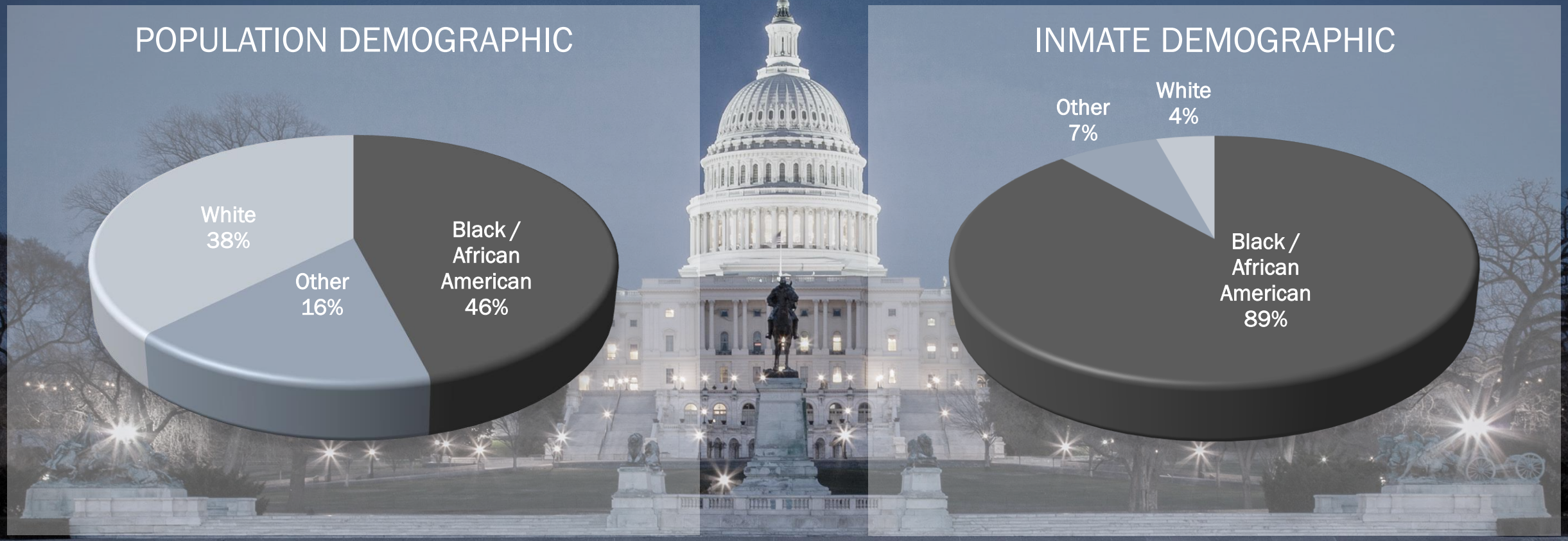
By Felix Ortega

Agenda

- Introduction
- Exploratory Data Analysis - (EDA)
- Modelling
  - Linear Regression
  - SVM
  - Random Forest
- Conclusion
- Future Work

Introduction

Washington D.C. - 2020



Introduction - Objectives

Objectives

- Primary – Predict Incarceration Sentences
- Secondary – Determine if race is significant in obtaining predictions

Introduction: Data

Data Source Information:

- Open Data DC – Site where the District of Columbia shares hundreds of datasets
- Dataset Name: Felony Sentences

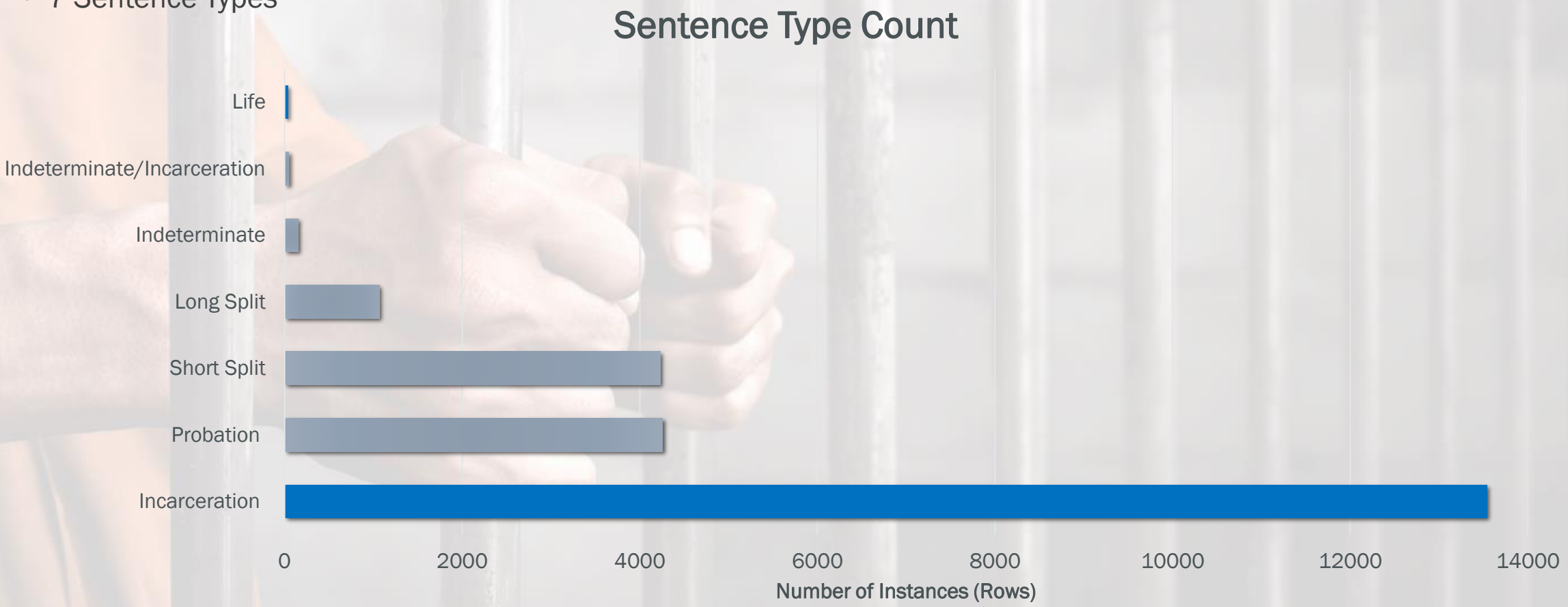
Data Description:

- Time frame: 2010 – 2018
- Instances: 23,332
- Variables: 22

Exploratory Data Analysis

Original Dataset:

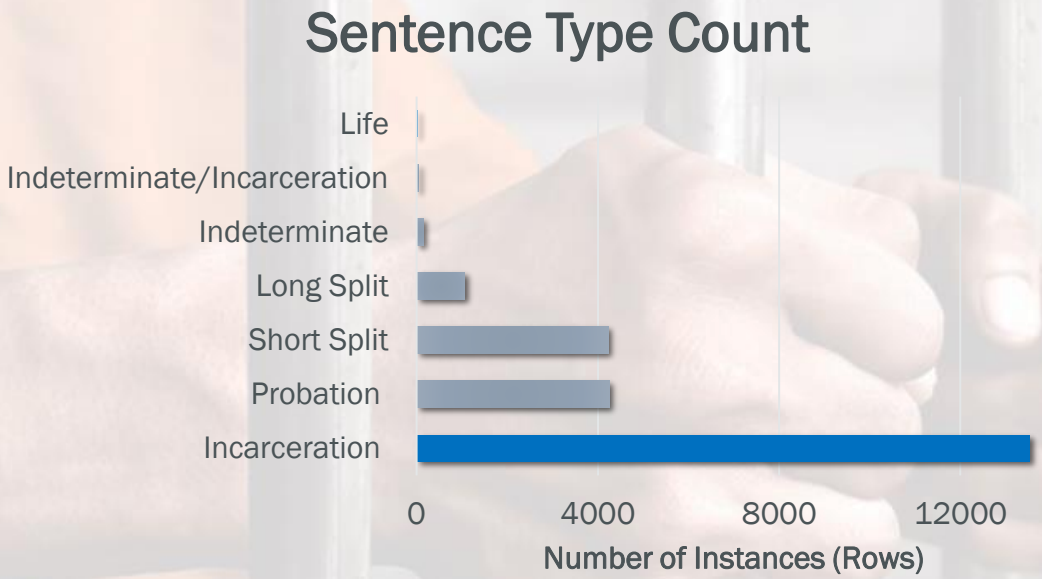
- 7 Sentence Types



Exploratory Data Analysis

Original Dataset:

- 7 Sentence Types



Modified Dataset: Sentence Types of Interest

- Incarceration and Life

Sentence Types	
Incarceration	13542
Life	37

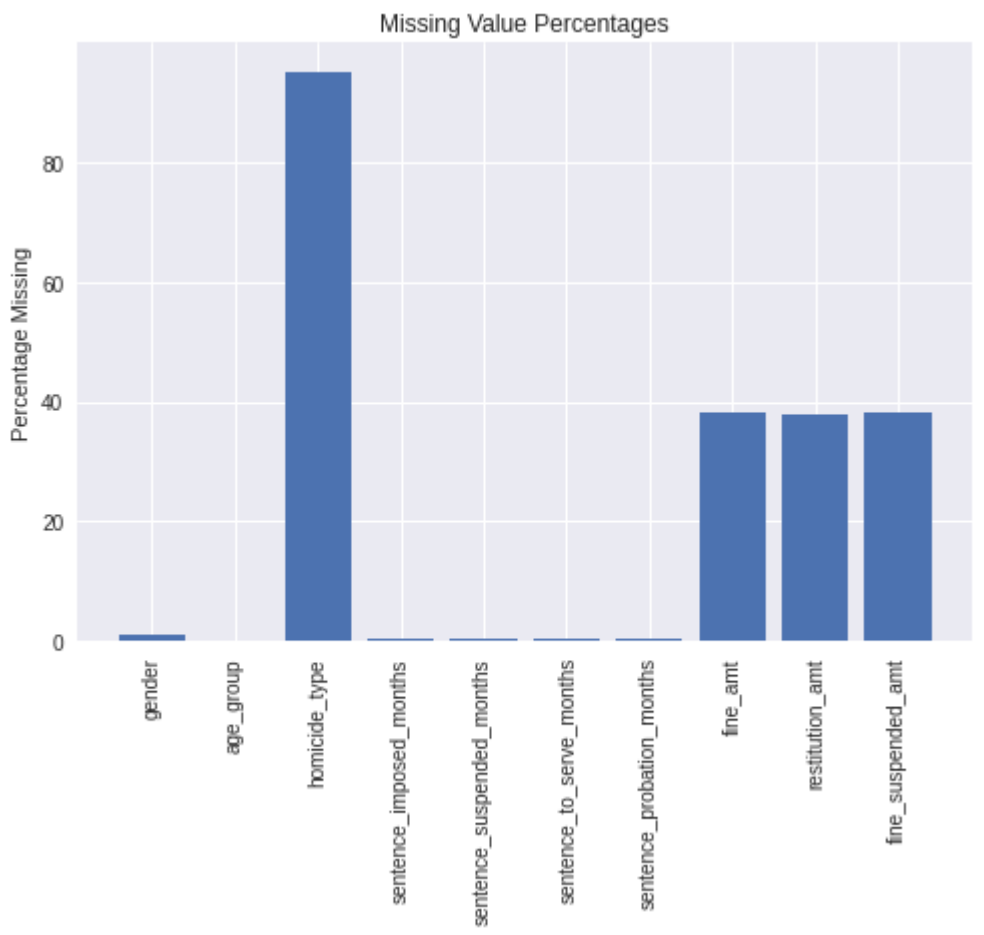
Shape before cleaning

- 13579 Rows
- 22 Columns

EDA: Cleaning

Missing Values:

- 10 columns with varying missing data



Handling Missing Values:

- Gender – replaced with ‘not_recorded’
- Age Group – dropped missing values
- Homicide Type – replaced with ‘not_homicide’
- Sentence to Serve Months – only sentence column kept
  - Dropped missing values for life sentences
- Fine Columns – replaced missing values with zeros

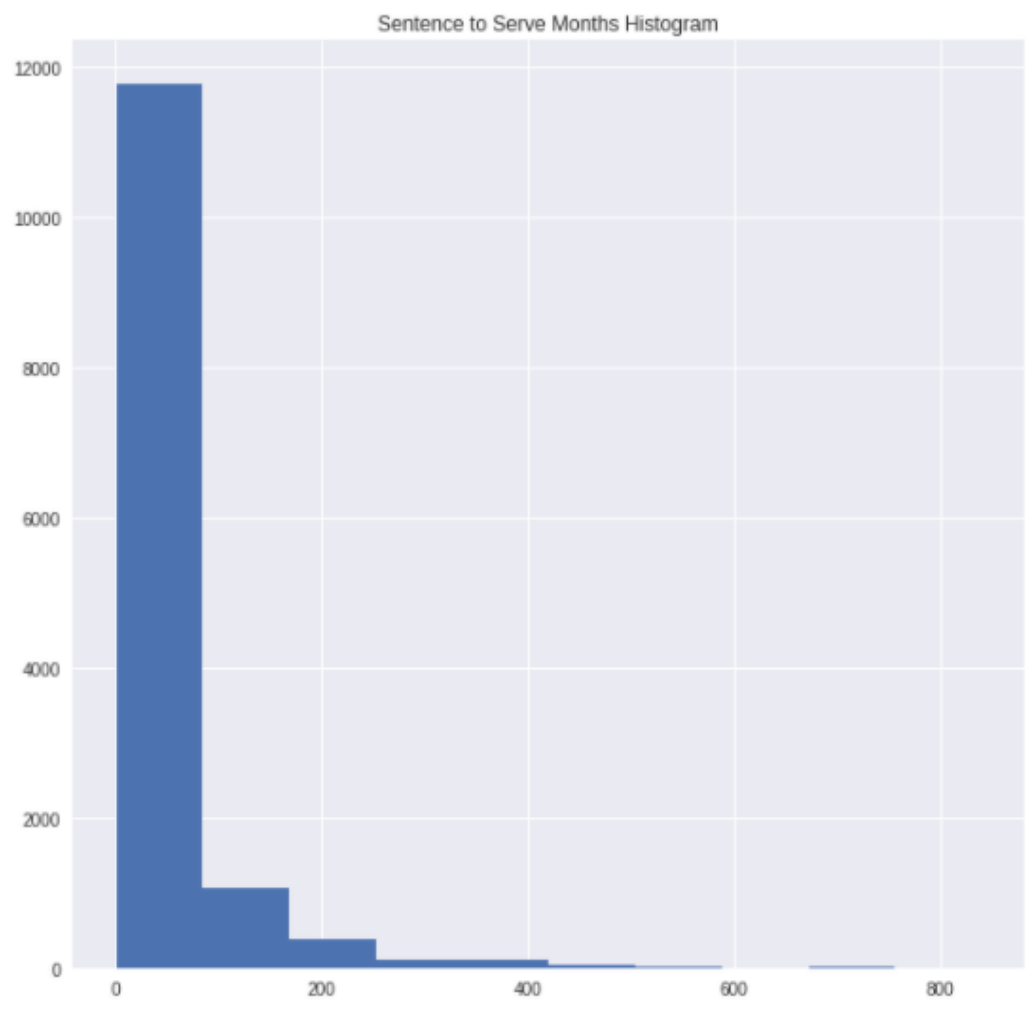
Final Dataframe Shape:

- 13509 Rows
- 19 Columns

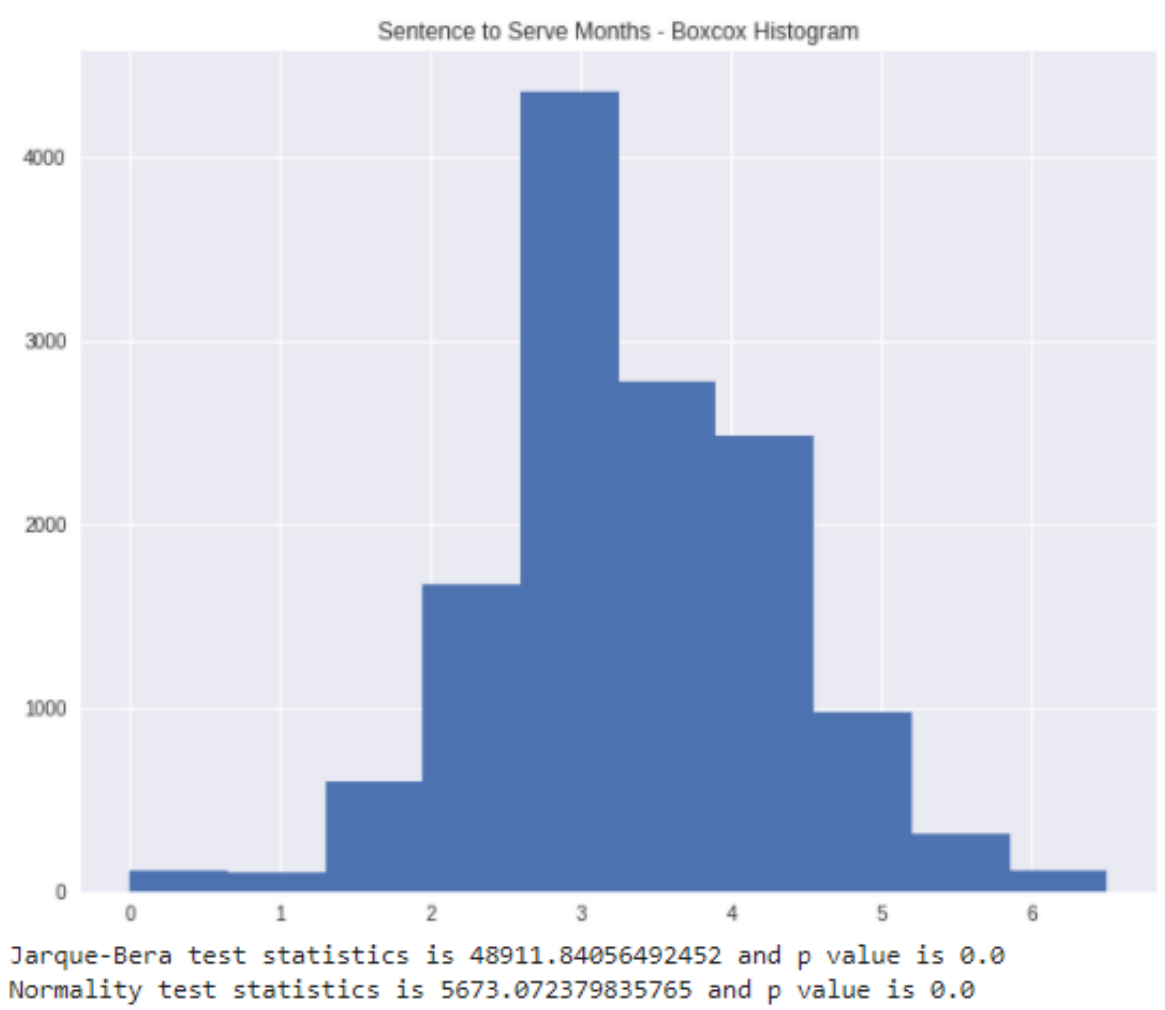
EDA: Target

Target Distribution:

- Distribution – Right skewed



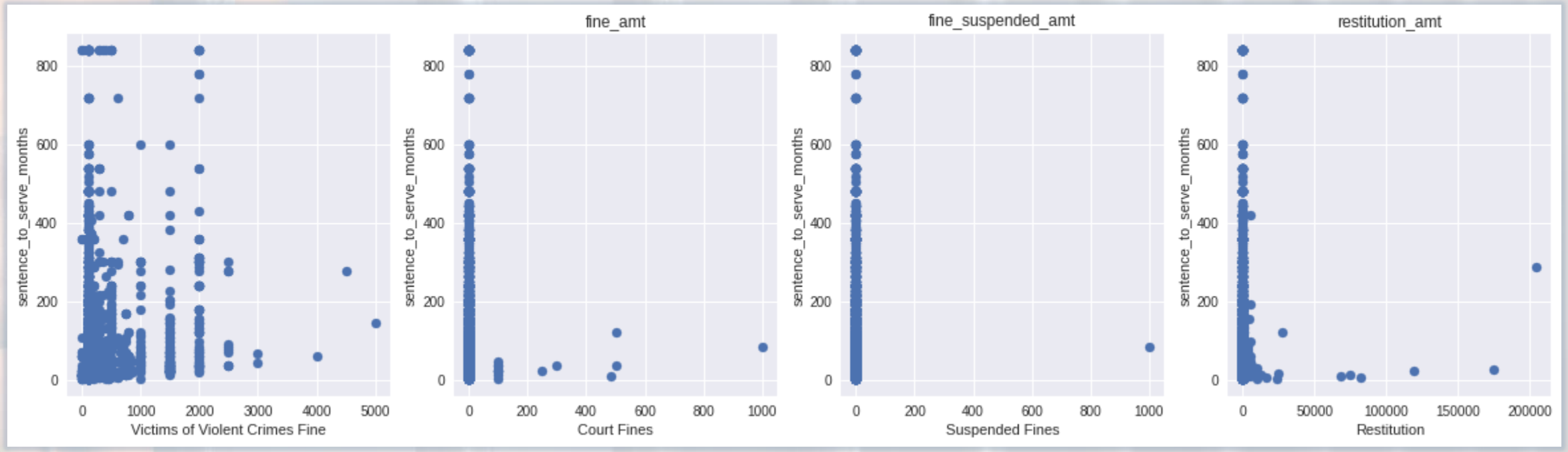
Boxcox Transformed Distribution:



EDA: Univariate Analysis

Features of Interest (continuous):

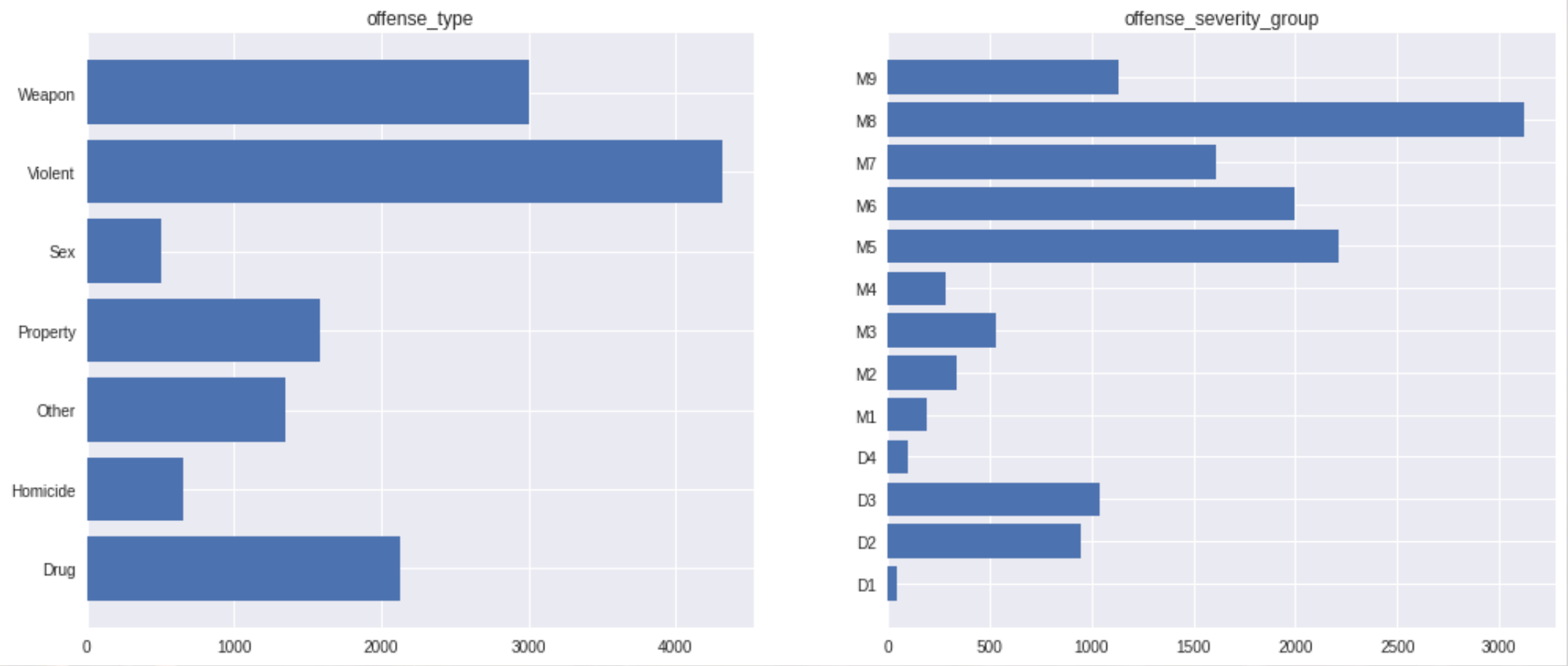
- Fine correlation with sentence



EDA: Univariate Analysis

Features of Interest (discrete):

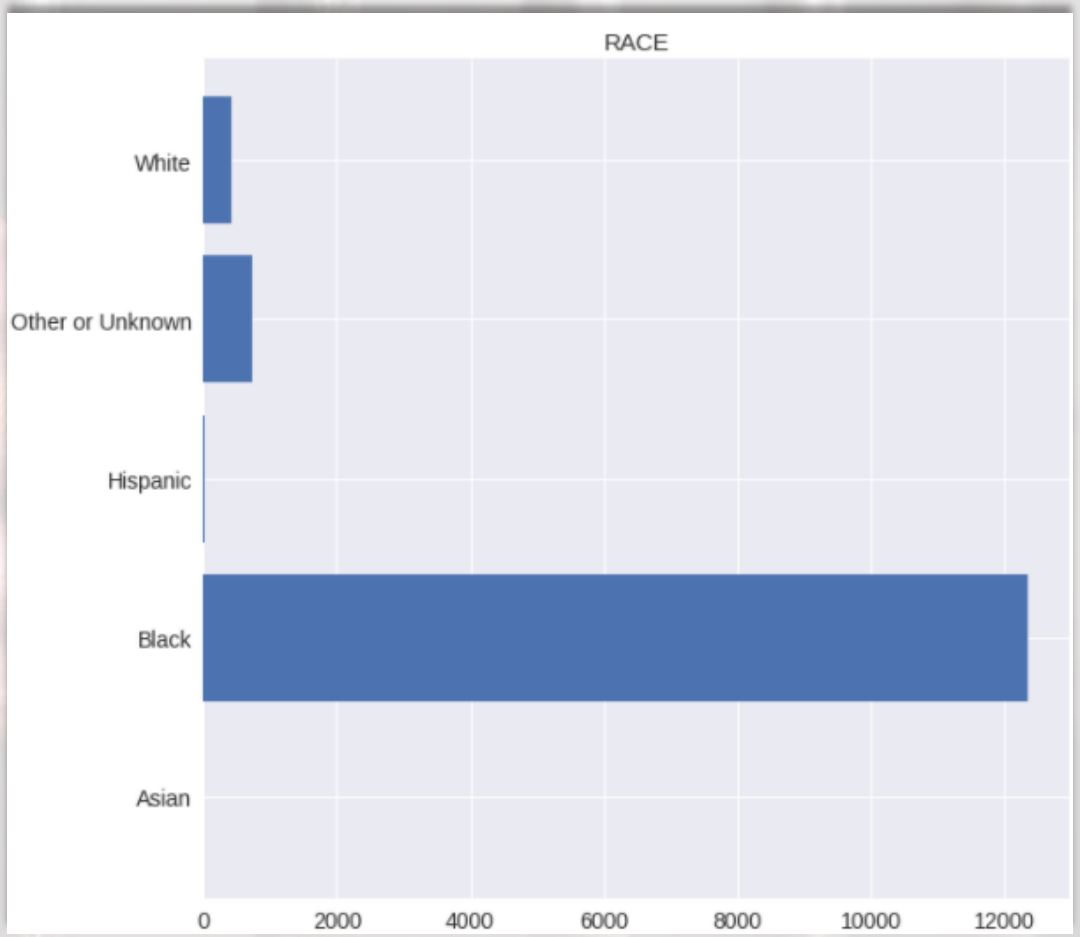
- Offense Type
- Offense Severity Group



EDA: Univariate Analysis

Features of Interest (discrete):

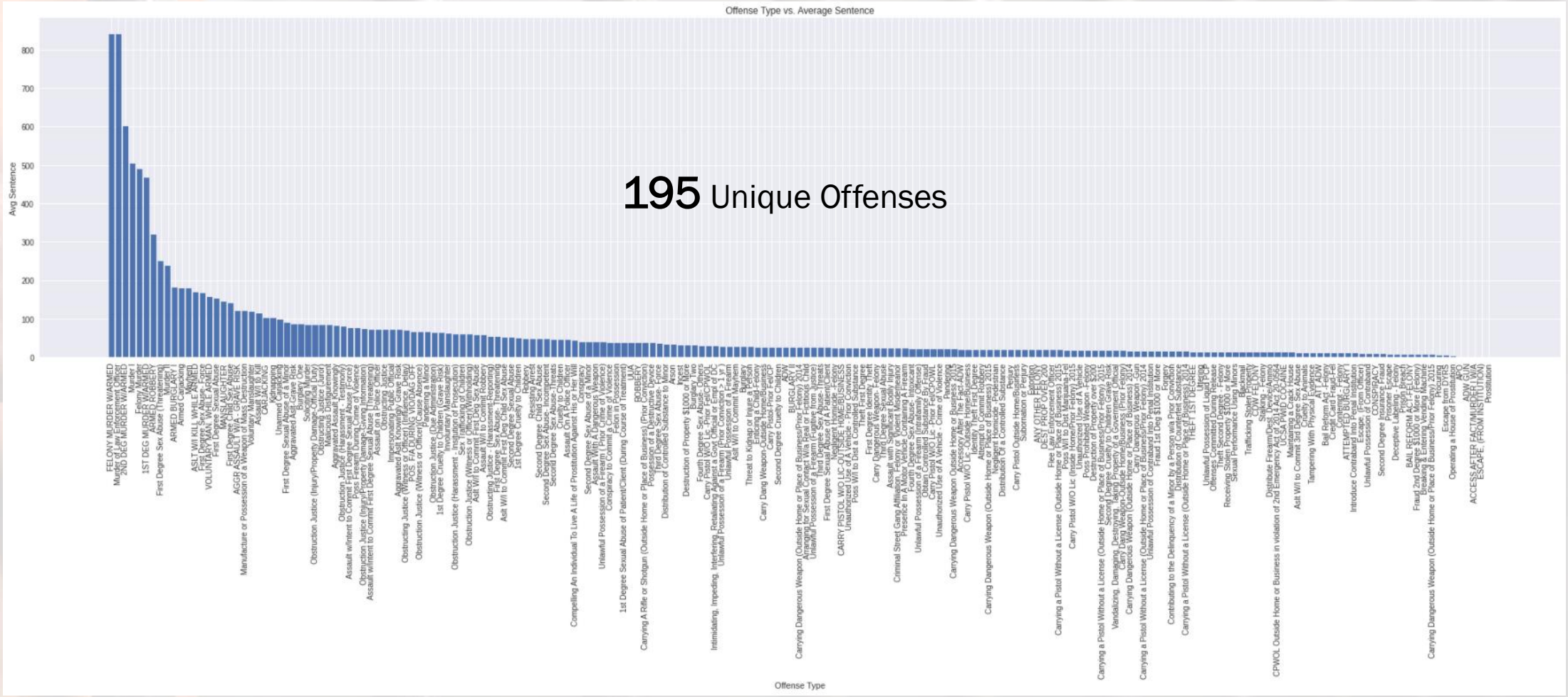
- Race



EDA: Univariate Analysis

Features of Interest (discrete):

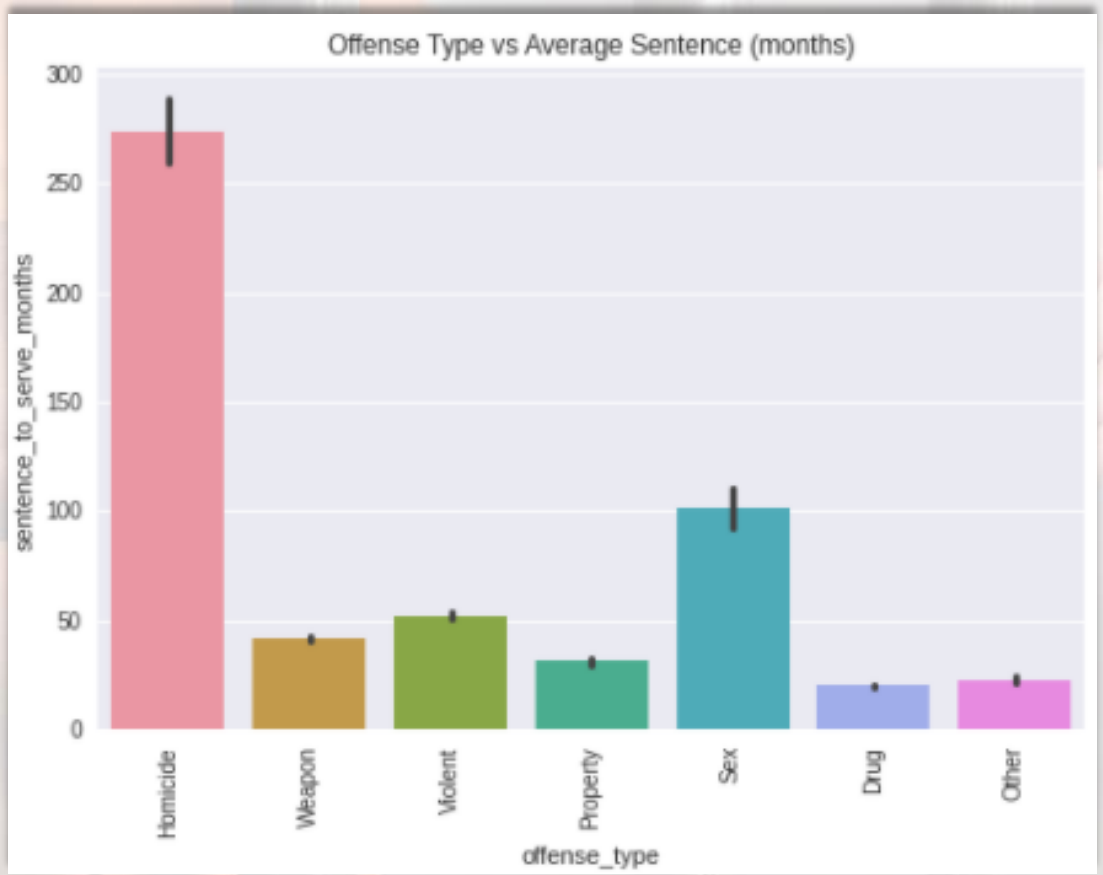
- Offense

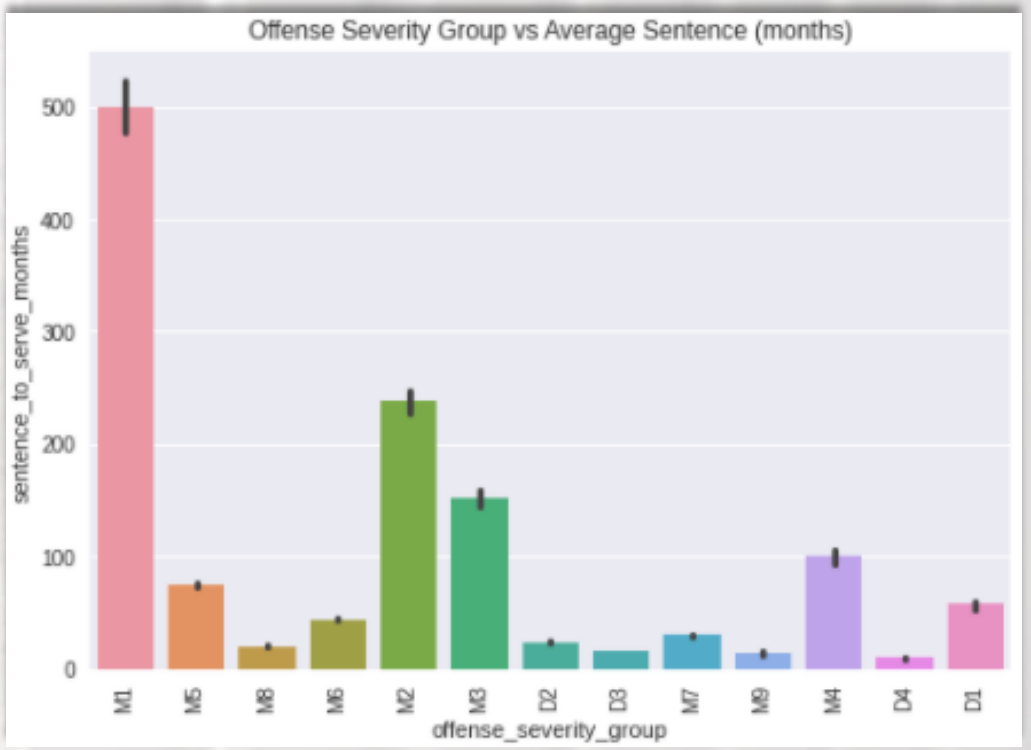


EDA: Bivariate Analysis

Average sentence length vs Offenses:

- Offense Type
- Offense Severity Group





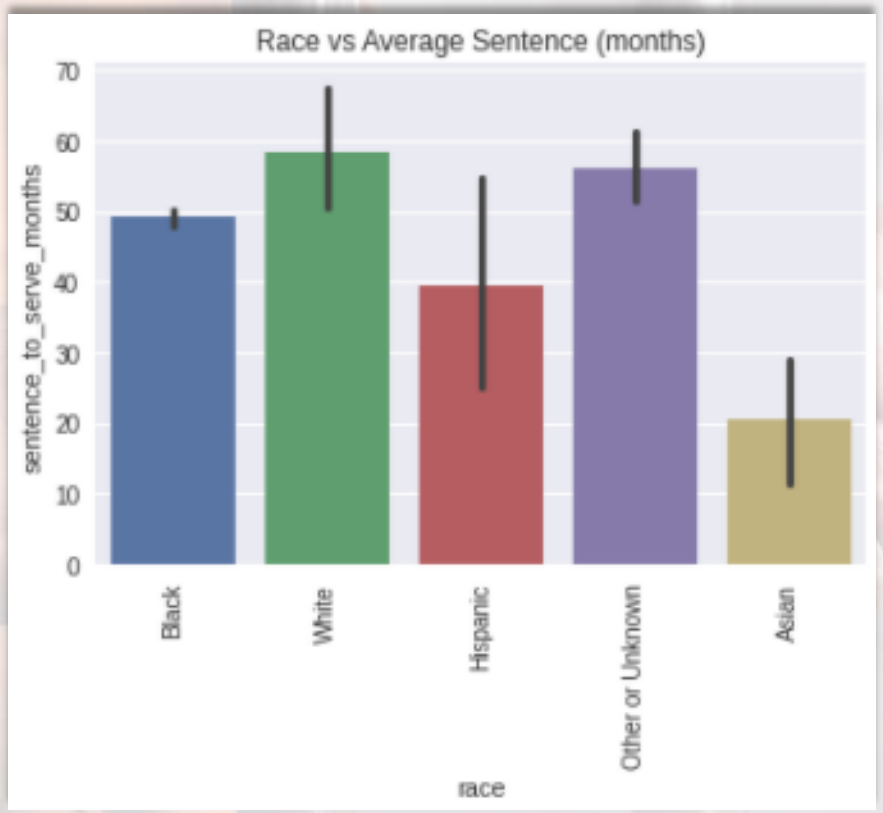
Model Performance without Offense Feature

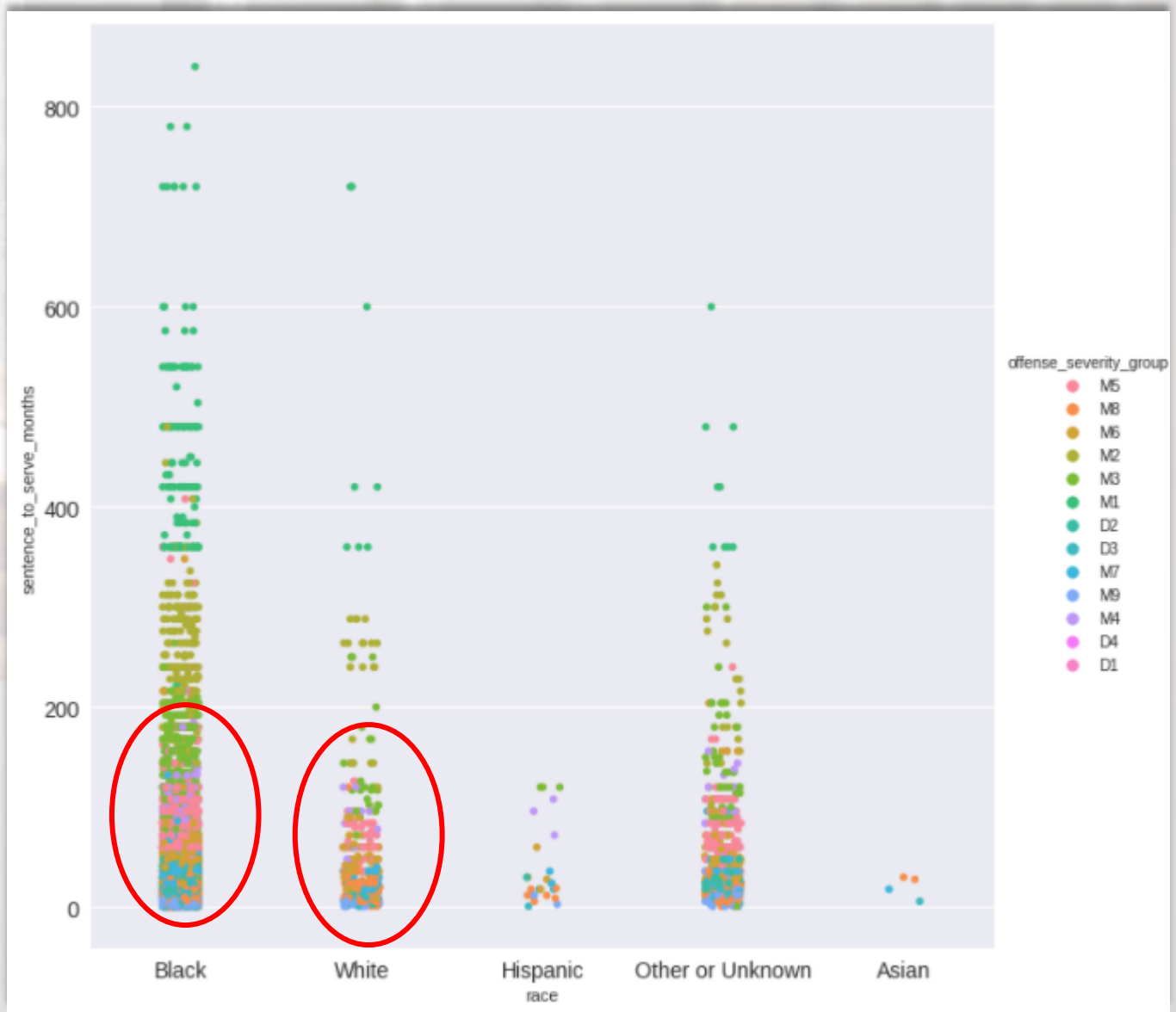
- Adjusted R^2 = 0.699
- Mean Absolute Error = 0.34

EDA: Bivariate Analysis

Average sentence length vs Race:

- Race

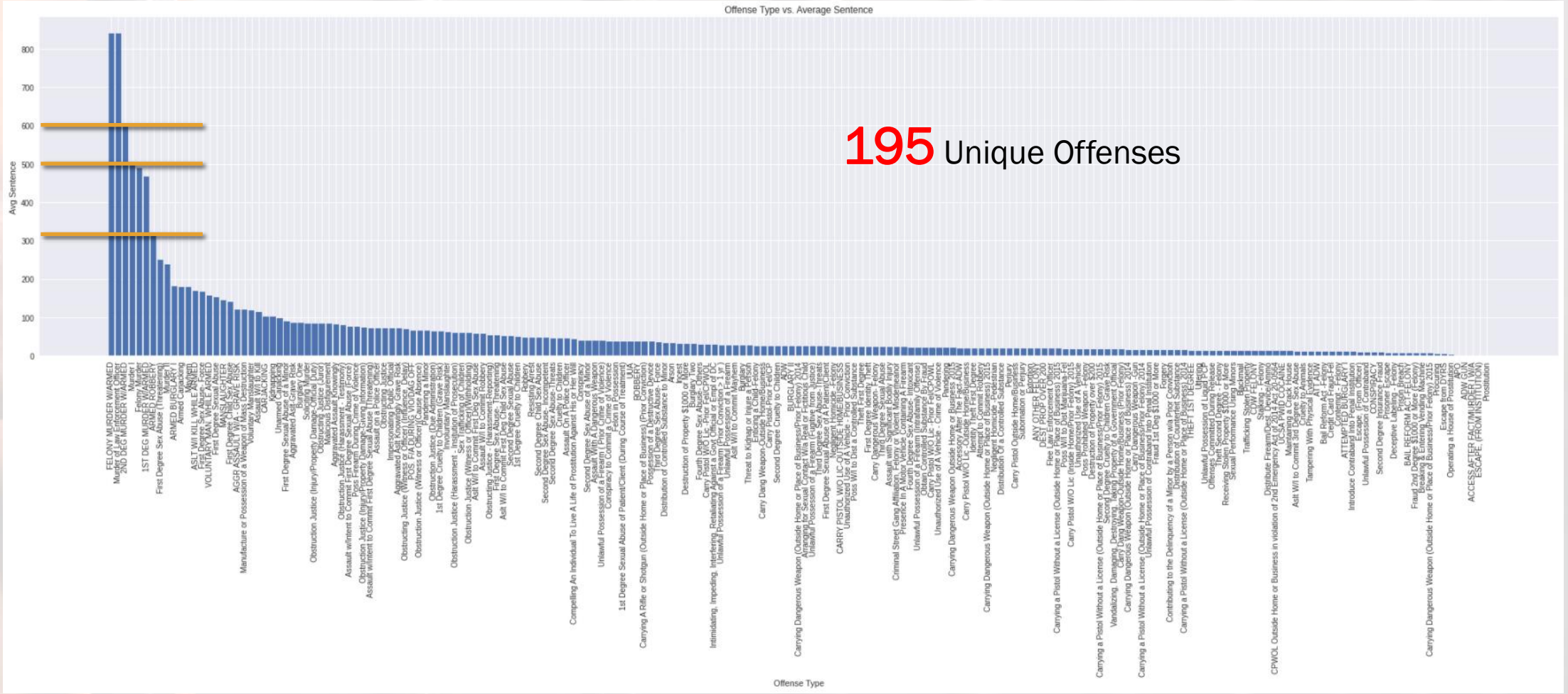




EDA: Feature Engineering

Offenses

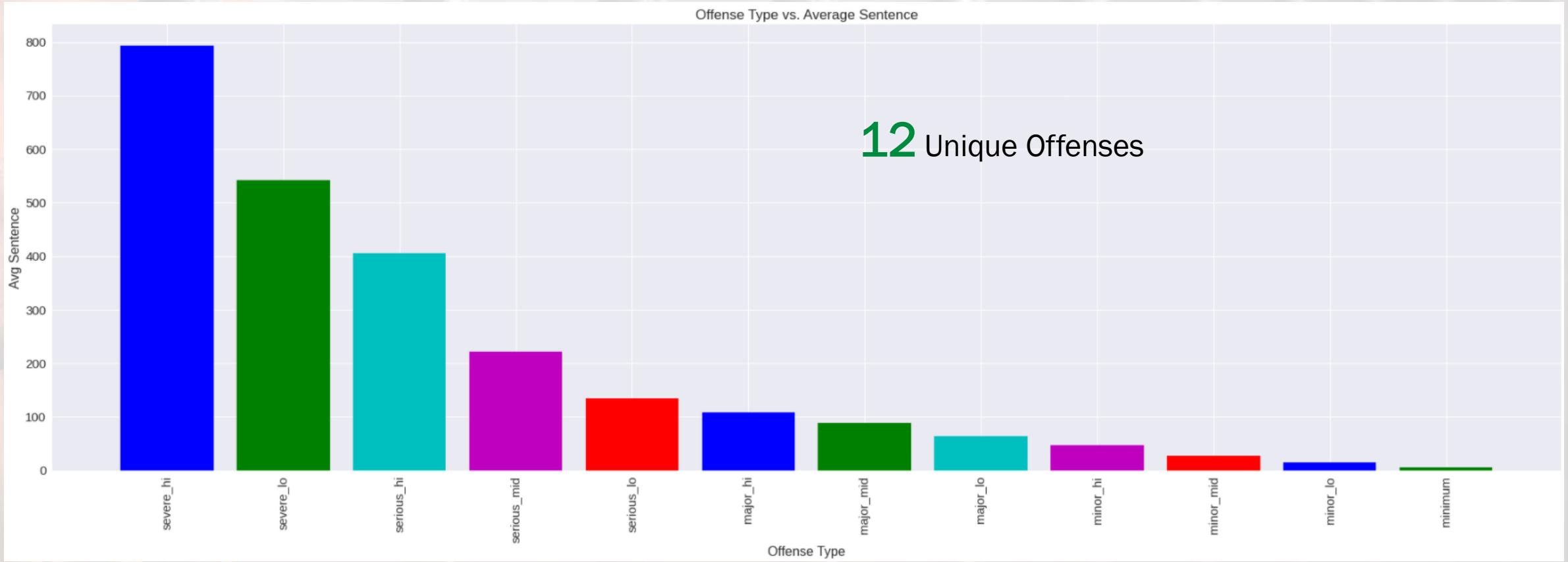
- Dimension reduction



EDA: Feature Engineering

Offenses

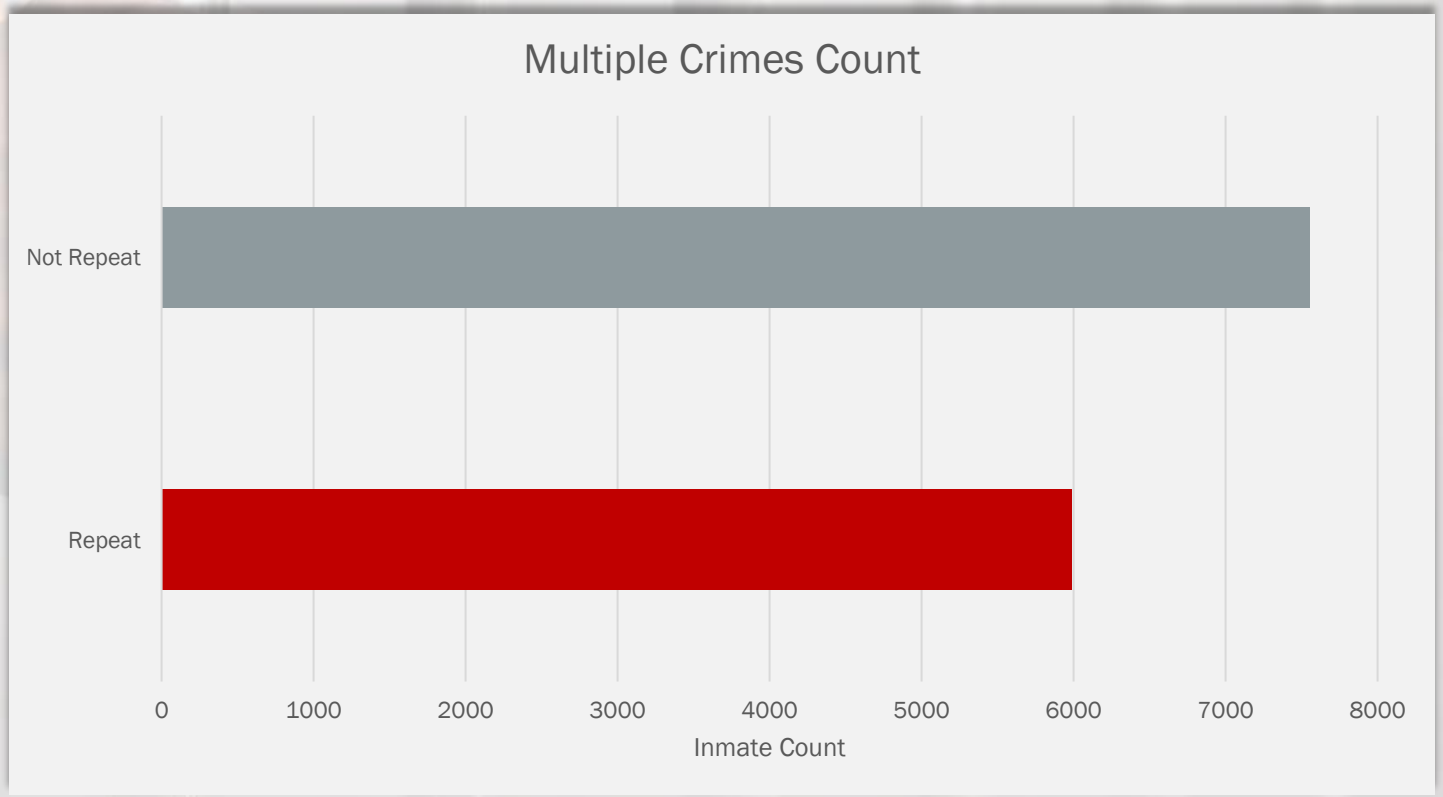
- Dimension reduction – from 195 to 12 offense labels



EDA: Feature Engineering

Created Feature:

- “Multiple_Crime” – Based on duplicate Offender ID
  - Multiple charges considered



Modelling

Model Preparation:

train_test_split – 80/20

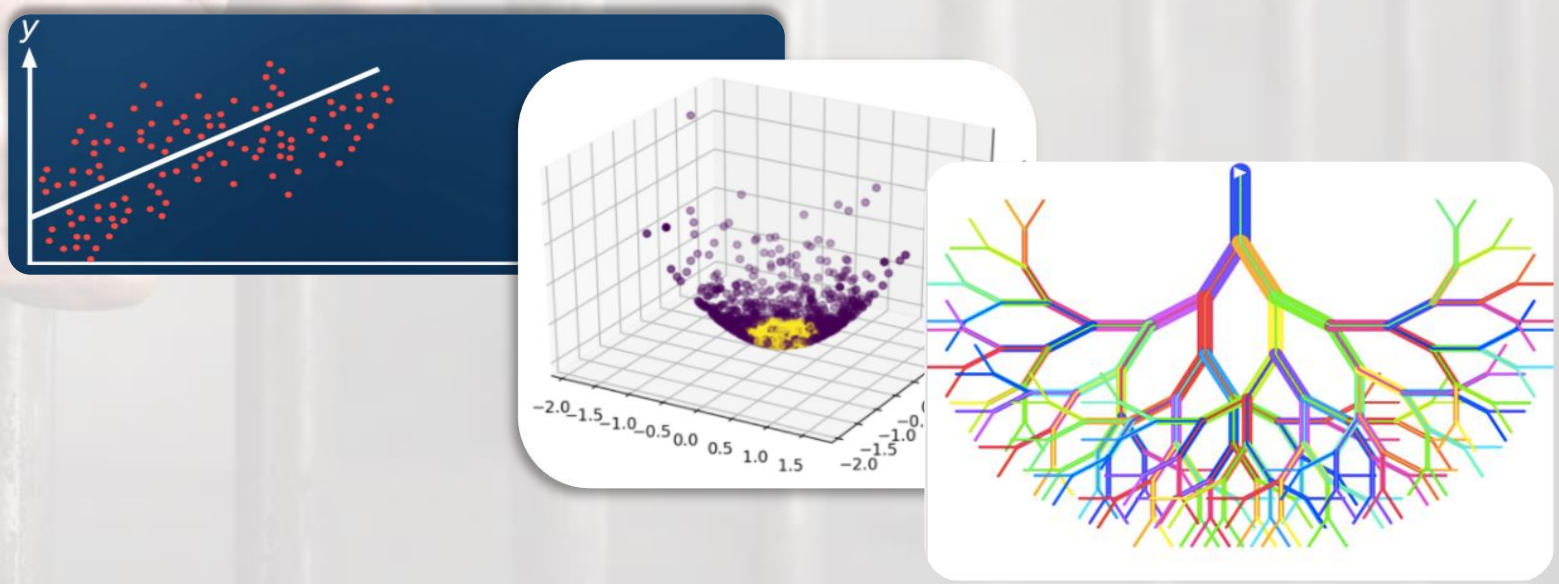
- 10807 rows in train set
- 2702 rows in test set

```
# Verify split
print(X_train.shape[0], y_train.shape[0], X_test.shape[0], y_test.shape[0])
```

```
10807 10807 2702 2702
```

Supervised Learning Models Employed:

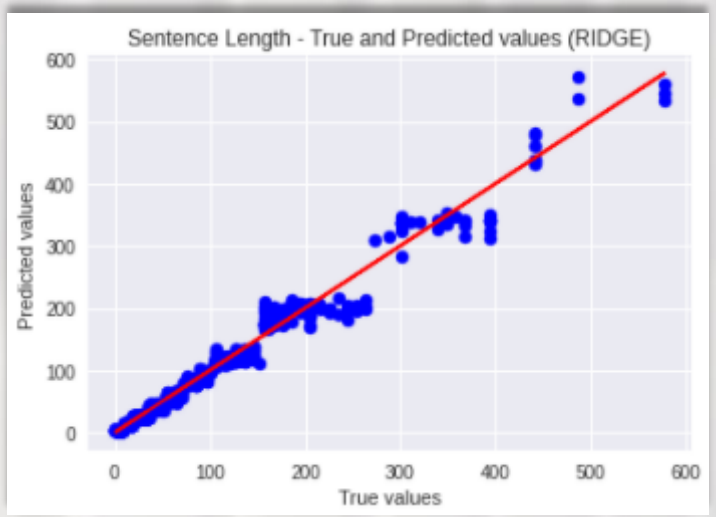
- Linear Regression
  - Lasso
  - Ridge
  - ElasticNet
  - OLS
- Support Vector Regressor
- Random Forest



Modelling – Linear Models

Linear Models - sklearn

Model	Train Score	Test Score	MSE	Mean Cross-Validation (cv = 10)
Lasso	0.941	0.946	0.049	0.93862
ElasticNet	0.941	0.946	0.049	0.93854
Ridge	0.941	0.946	0.049	0.93855

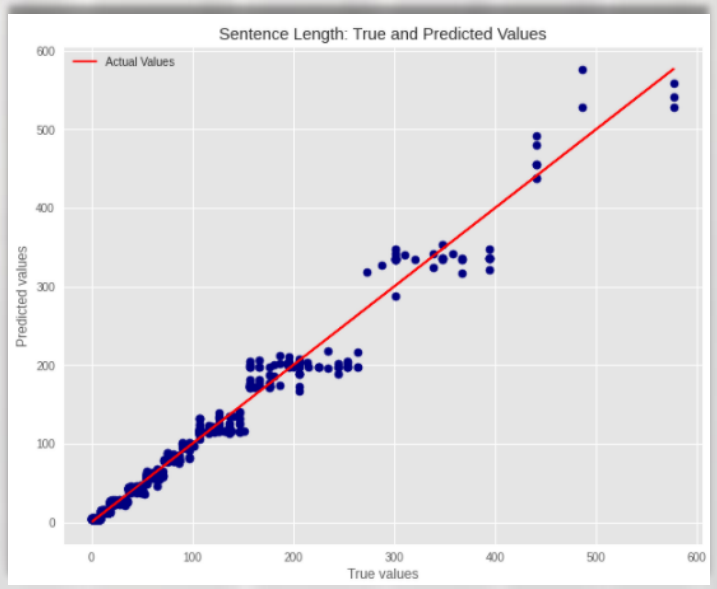


Linear Models – statsmodel

- Three OLS models varying in features
  - Race, Repeat Offender, Sentence Year
- No change in performance

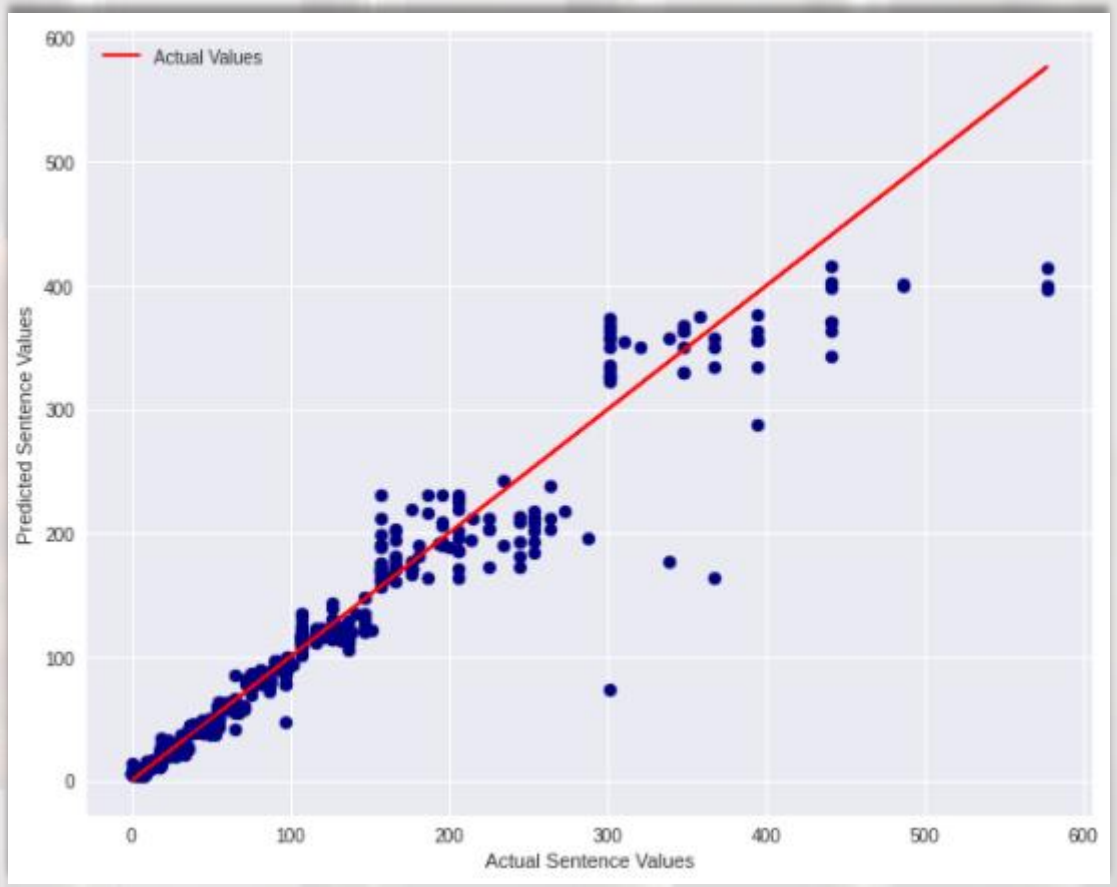
Race Features	coef	p-val
Race_Black	-0.1489	0.196
Race_Hispanic	-0.2278	0.072
Race_White	-0.1518	0.190
Race_Other	-0.1573	0.173

	R2 Adj.	Score(test)	MSE
OLS	0.940	0.947	0.049



Modelling - SVR

Actual vs Predicted Values - SVR

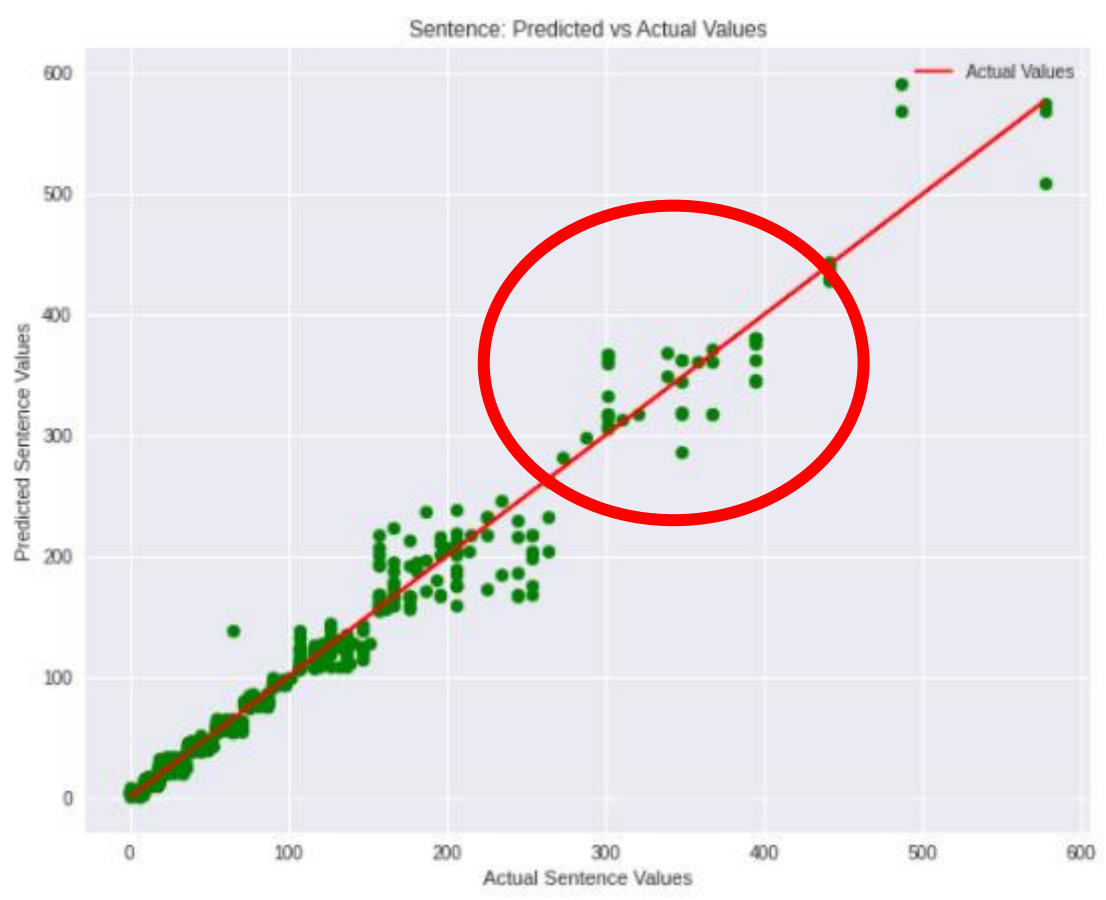


Model	Train Score	Test Score	MSE	Mean Cross-Validation (cv = 5)
SVR	0.937	0.939	0.056	0.931

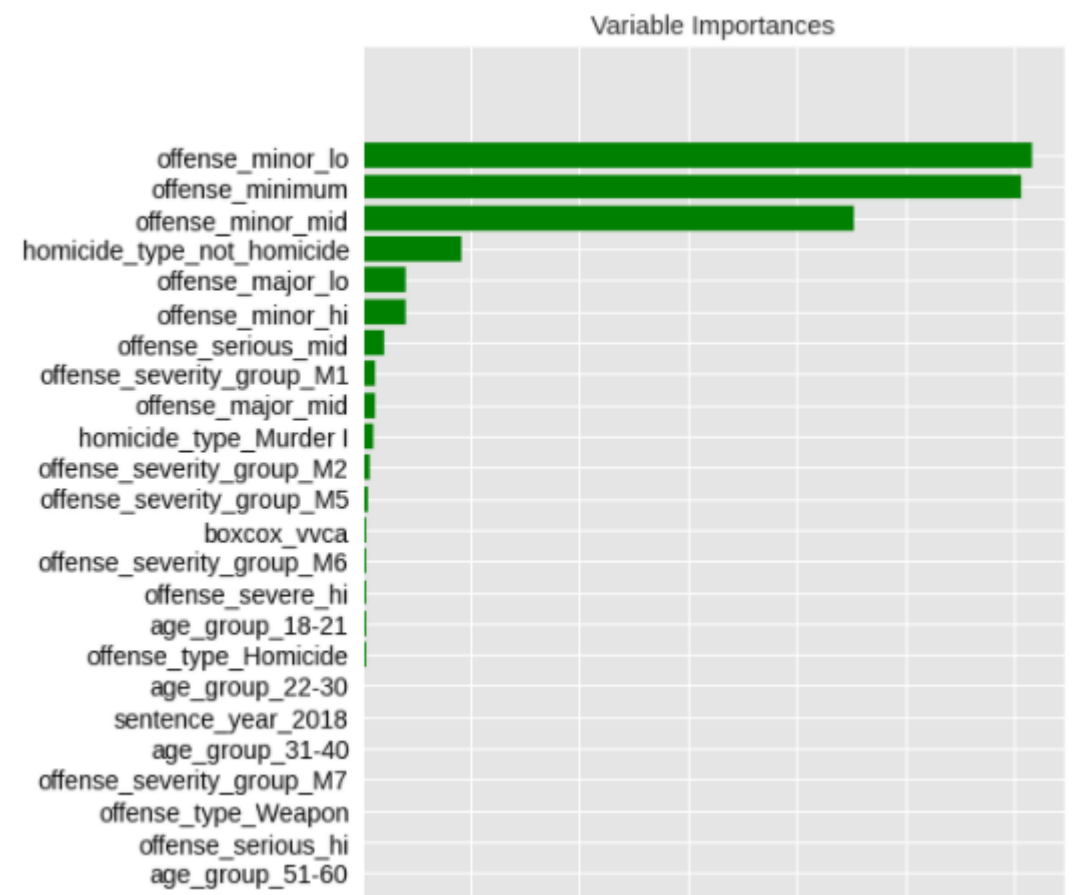
Modelling – Random Forest

Model 1: Random Forest

n - estimators	Train Score	Test Score	MSE	Mean Cross-Validation (cv = 5)
1000	0.969	0.945	0.050	0.939



Feature importance



Modelling – Random Forest

Model 2: Random Forest with Grid CV

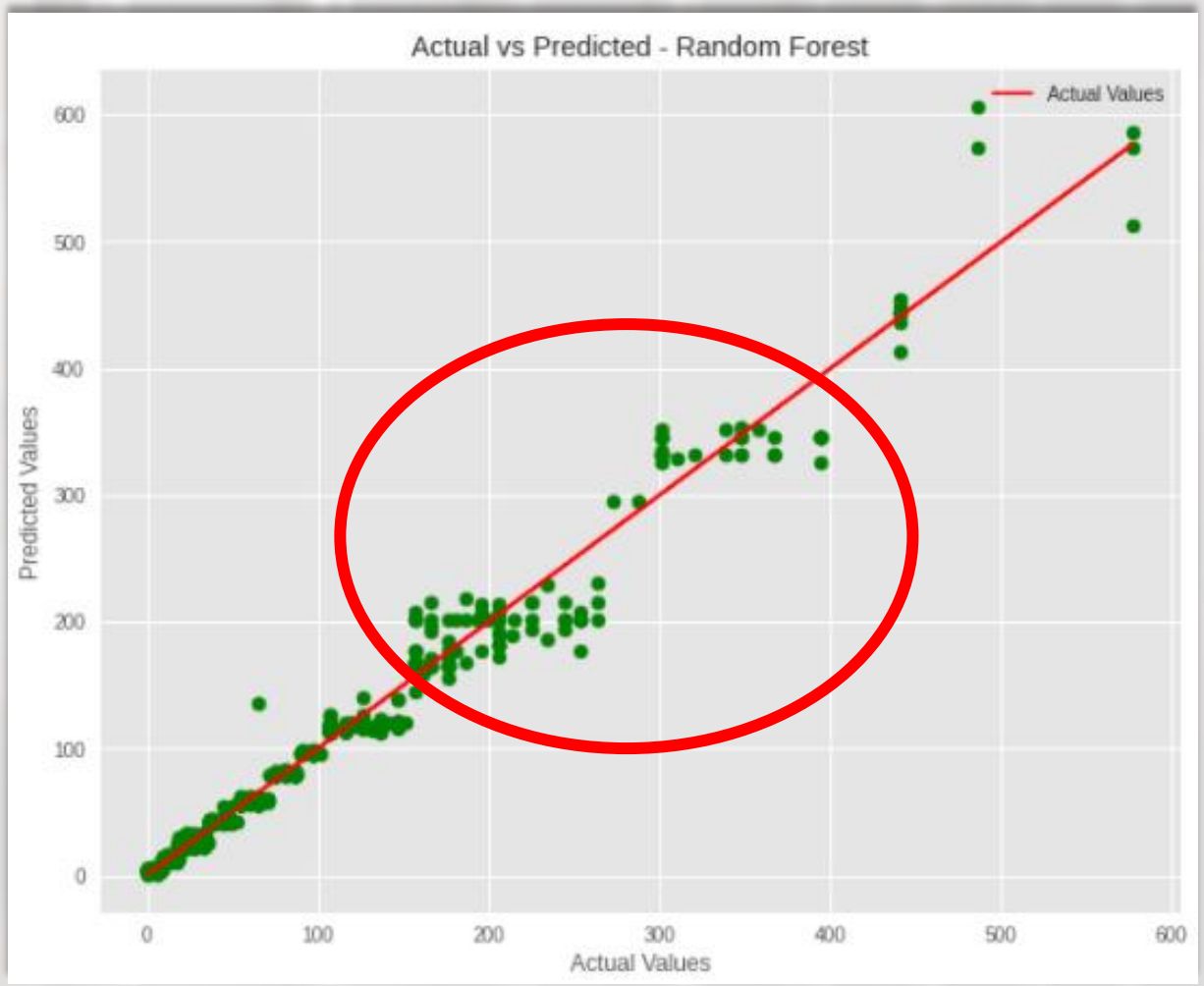
- Removed: Gender, Sentence Year, Sentence Type

Grid parameters

- n_estimators = np.arange(100, 130, 5)
- min_samples_leaf = np.arange(1, 4, 1)
- max_depth = np.arange(5, 50, 5)
- max_features = np.arange(15, 50, 5))

Best Parameters

- n_estimators = 105
- min_samples_leaf = 1
- max_depth = 10
- max_features = 40



Conclusion – Random Forest

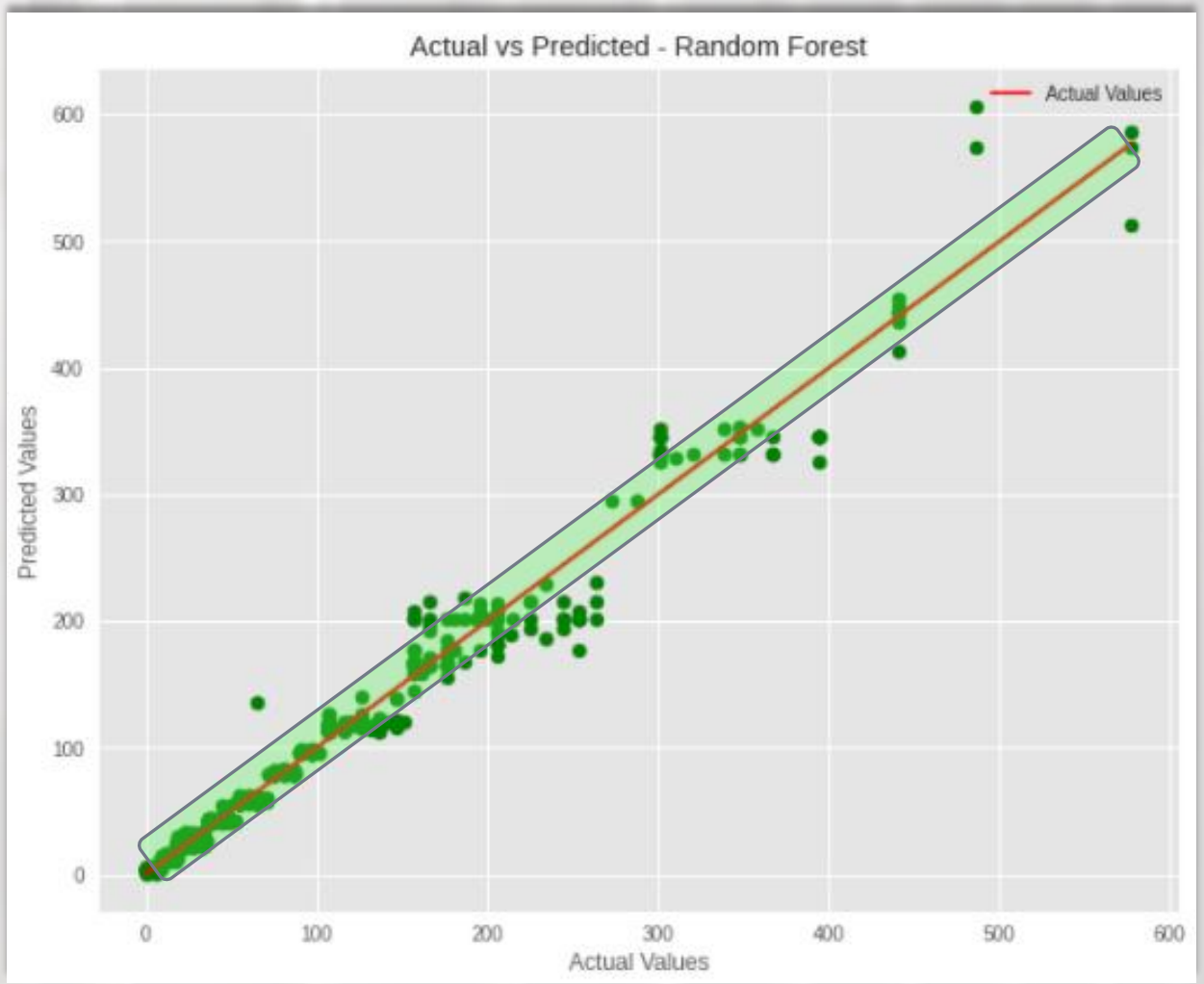
Selected Model: Random Forest with Best Parameters

Improvements observed:

Final Metrics	
Train Score	0.951
Test Score	0.951
MSE	0.046
Mean Cross-Validation (cv = 5)	0.944

Objectives

- Primary – Predict incarceration sentences
- Secondary – Determine if race is significant in obtaining predictions



Future Work

Weakness:

- Predicting mid - serious offense sentences
- Unbalanced data

Improvements:

- Improve dataset, ANOVA Testing
- Repurpose model

Audience / Application:

Trial Lawyers

U.S. Legislation

General Public

Questions

Sources

Project Data :

- https://opendata.dc.gov/datasets/9fa34e198ad240358c7c36bc063d2058?orderBy=OFFENSE_SEVERITY_GROUP&page=2

Informative Sources:

- https://doc.dc.gov/sites/default/files/dc/sites/doc/publication/attachments/DCDepartmentofCorrections_FactsandFigures_April2020_0.pdf
- https://www.census.gov/quickfacts/fact/table/DC/RHI125219
- https://www.acludc.org/en/racial-disparities-dc-policing-descriptive-evidence-2013-2017