

Project Poirot

**Machine Learning based
Predictive Crime Analysis**



Team Members

- Mansi Ganatra
- Naga Ritwik Indugu
- Uma Kanumuri

Mentor:

- Prof. Ion Muslea
- Officer Wyman Thomas (USC DPS)

Data Source

- Los Angeles Open Data: Crime Data from 2010 to Present
- Size: ~ 1.8 million records
- URL: <https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq>
- Previous Work: <https://www.crimemapping.com/map/ca/losangeles>

Sample Dataset

DR Number	Date Reported	Date Occurred	Time Occurred	Area ID	Area Name	Reporting District	Crime Code	Crime Code D
181116602	09/15/2018	09/15/2018	1414	11	Northeast	1162	341	THEFT
181420300	09/15/2018	09/15/2018	0840	14	Pacific	1444	664	
181716197	09/15/2018	09/15/2018	0230	17	Devonshire	1791	510	
182017267	09/15/2018	09/15/2018	0015	20	Olympic	2049	624	
181116591	09/15/2018	09/15/2018	0435	11	Northeast	1124	442	
180123869	09/15/2018	09/15/2018	0803	01	Central	0182	745	
180123885	09/15/2018	09/15/2018	1400	01	Central	0153	480	
181222770	09/15/2018	09/15/2018	1015	12	77th Street	1233	330	
180123884	09/15/2018	09/15/2018	0615	01	Central	0119	946	
180321165	09/15/2018	09/15/2018	1125	03	Southwest	0396	410	
180219239	09/15/2018	09/15/2018	0330	02	Rampart	0275	626	

--	--

Features


Columns in this Dataset

Column Name	Description	Type	
DR Number	Division of Records Number: Official file number made up ...	Plain Text	T
Date Reported	MM/DD/YYYY	Date & Time	📅
Date Occurred	MM/DD/YYYY	Date & Time	📅
Time Occurred	In 24 hour military time.	Plain Text	T
Area ID	The LAPD has 21 Community Police Stations referred to as ...	Plain Text	T
Area Name	The 21 Geographic Areas or Patrol Divisions are also given ...	Plain Text	T
Reporting District	A four-digit code that represents a sub-area within a Geogr...	Plain Text	T
Crime Code	Indicates the crime committed. (Same as Crime Code 1)	Plain Text	T
Crime Code Description	Defines the Crime Code provided.	Plain Text	T
MO Codes	Modus Operandi: Activities associated with the suspect in c...	Plain Text	T

Features

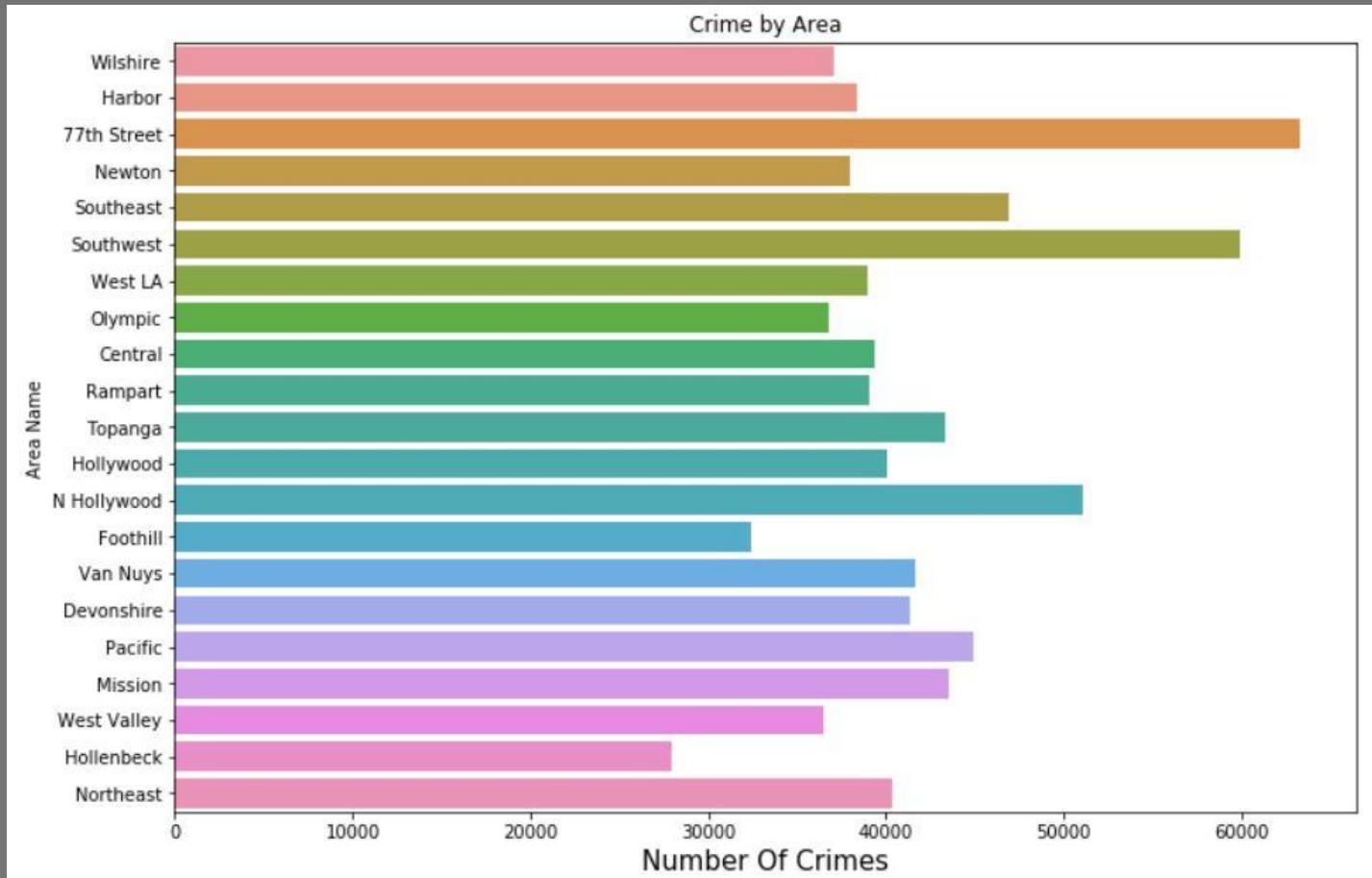
Victim Age	Two character numeric	Plain Text	T
Victim Sex	F - Female M - Male X - Unknown	Plain Text	T
Victim Descent	Descent Code: A - Other Asian B - Black C - Chinese D - Ca...	Plain Text	T
Premise Code	The type of structure, vehicle, or location where the crime t...	Plain Text	T
Premise Description	Defines the Premise Code provided.	Plain Text	T
Weapon Used Code	The type of weapon used in the crime.	Plain Text	T
Weapon Description	Defines the Weapon Used Code provided.	Plain Text	T
Status Code	Status of the case. (IC is the default)	Plain Text	T
Status Description	Defines the Status Code provided.	Plain Text	T
Crime Code 1	Indicates the crime committed. Crime Code 1 is the primar...	Plain Text	T
Crime Code 2	May contain a code for an additional crime, less serious th...	Plain Text	T
Crime Code 3	May contain a code for an additional crime, less serious th...	Plain Text	T

Features

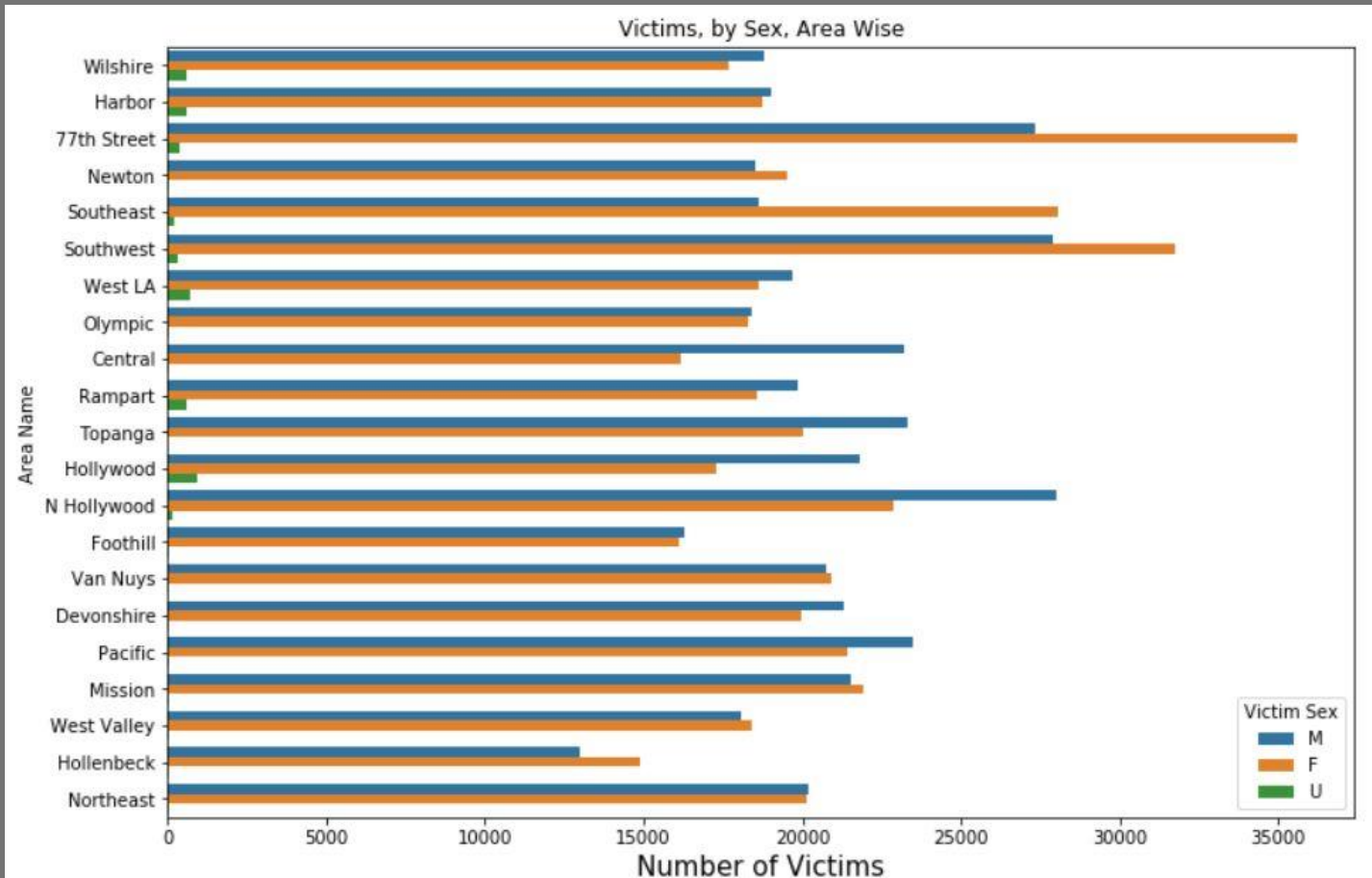
Crime Code 4	May contain a code for an additional crime, less serious th...	Plain Text	T
Address	Street address of crime incident rounded to the nearest hu...	Plain Text	T
Cross Street	Cross Street of rounded Address.	Plain Text	T
Location	The location where the crime incident occurred. Actual add...	Location	

Exploratory Data Analysis

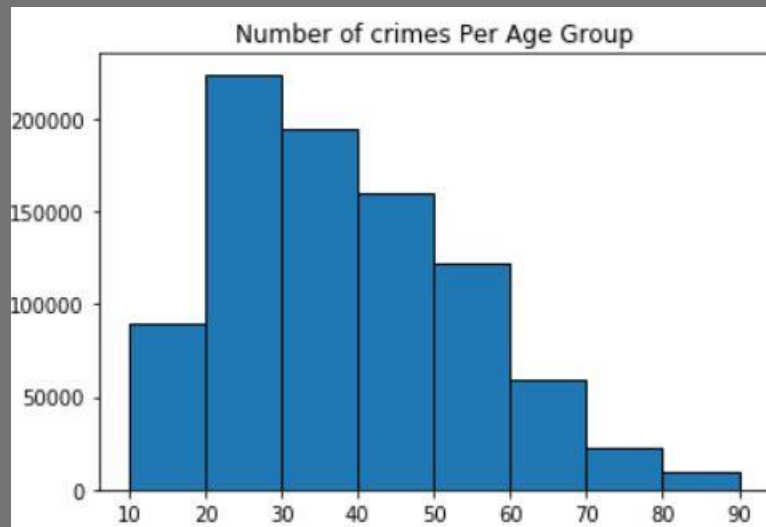
NUMBER OF CRIMES PER AREA



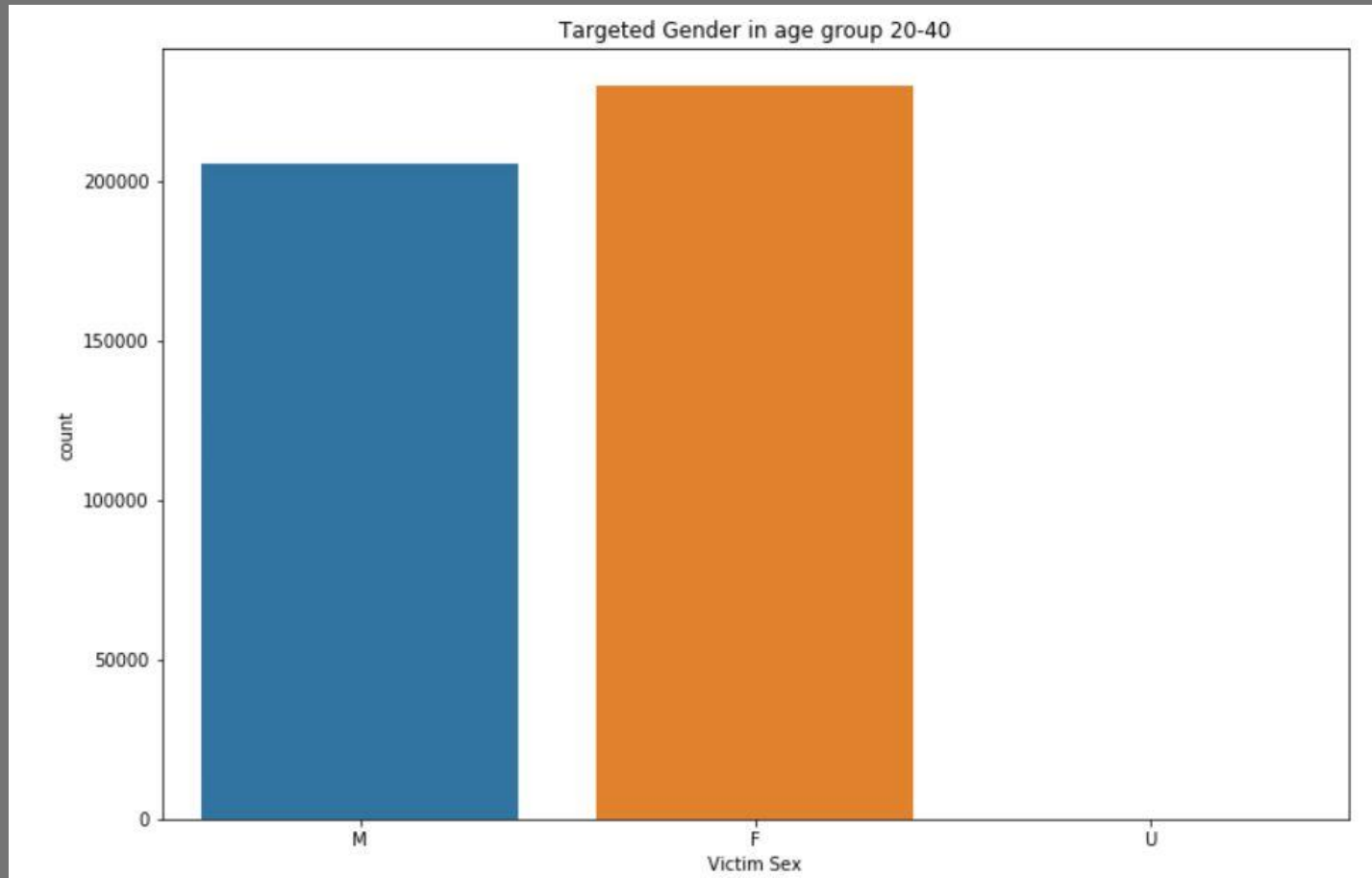
NUMBER OF VICTIMS PER AREA



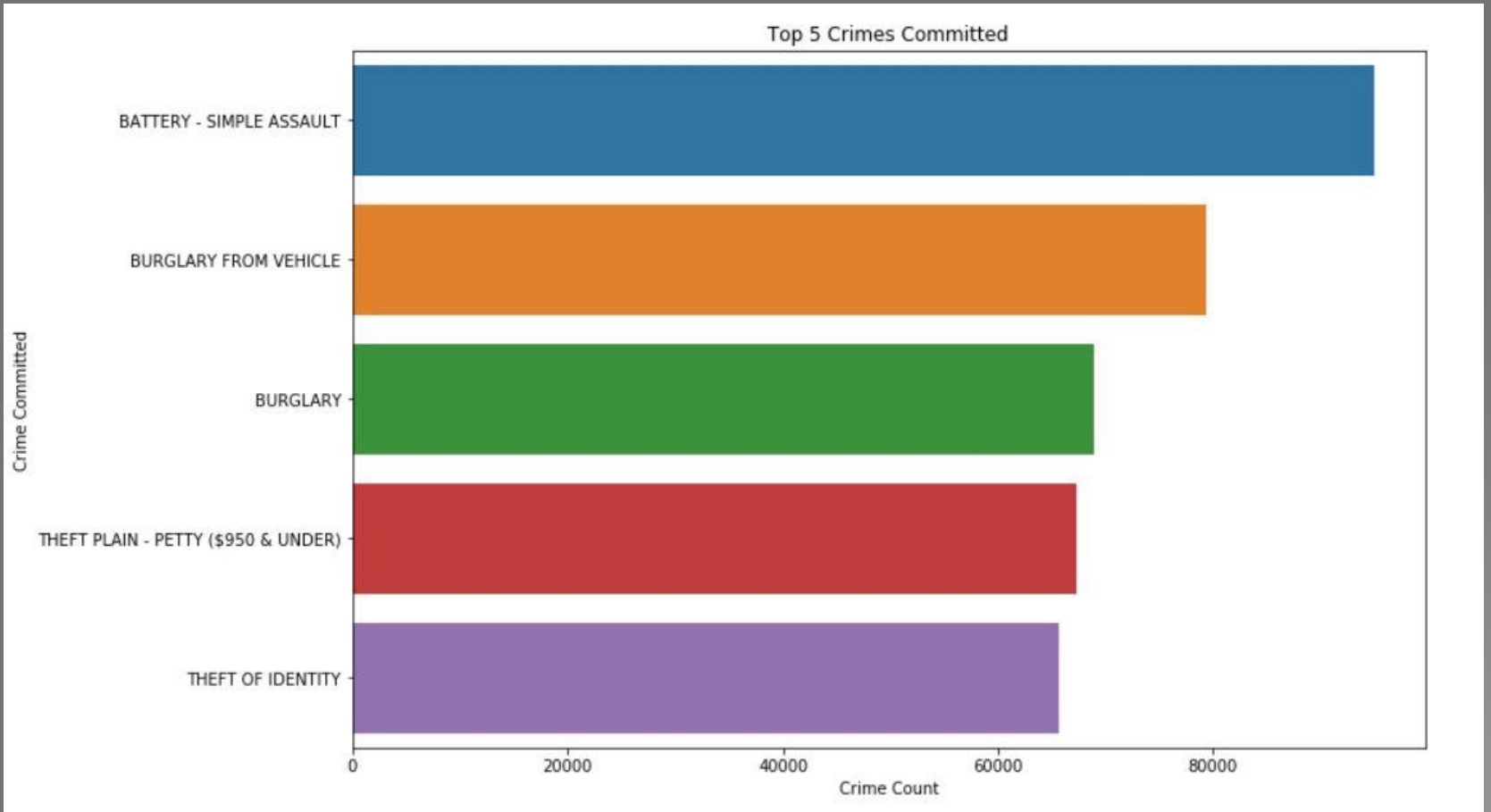
NUMBER OF CRIMES PER AGE GROUP



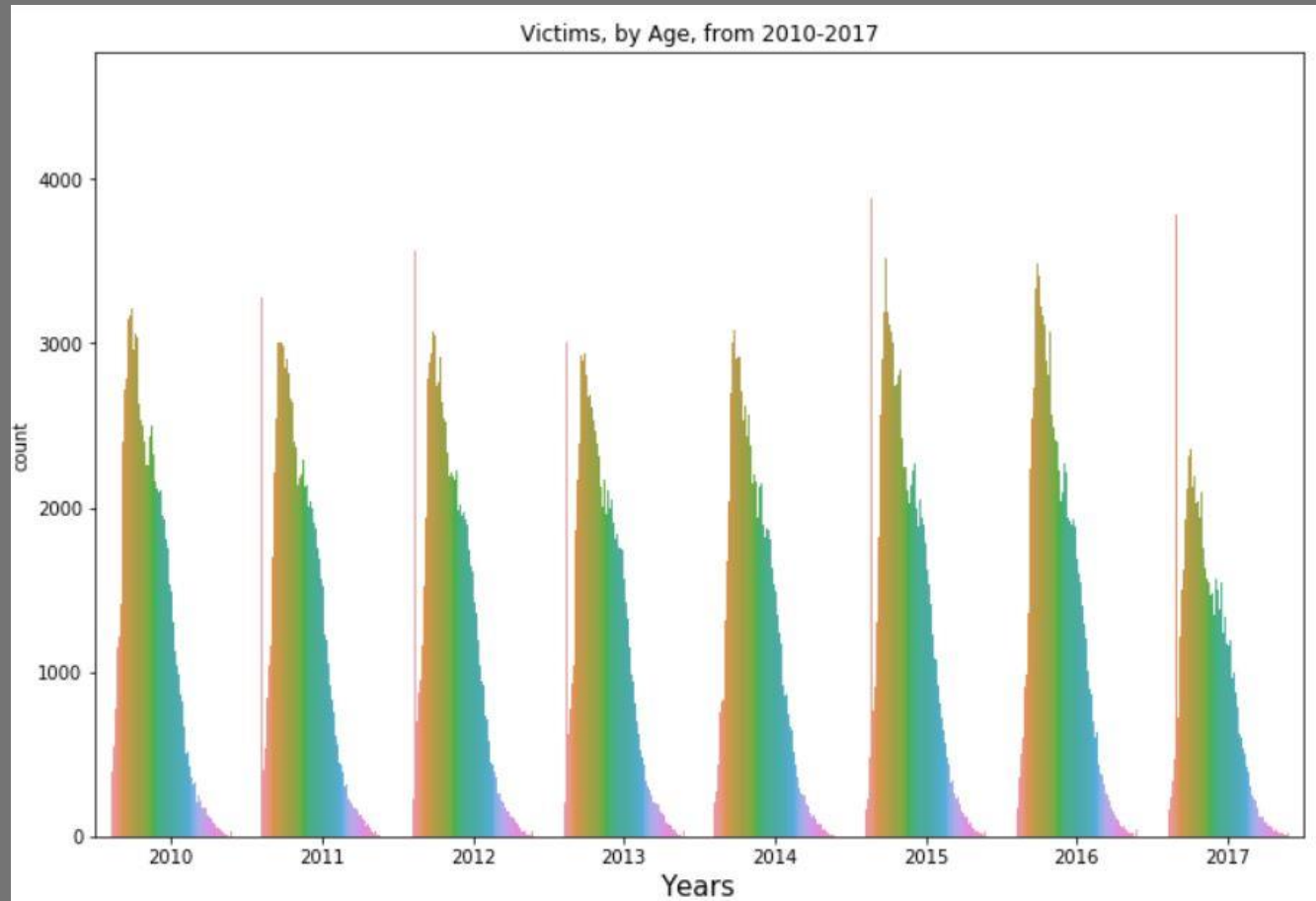
CRIME COUNT BY GENDER



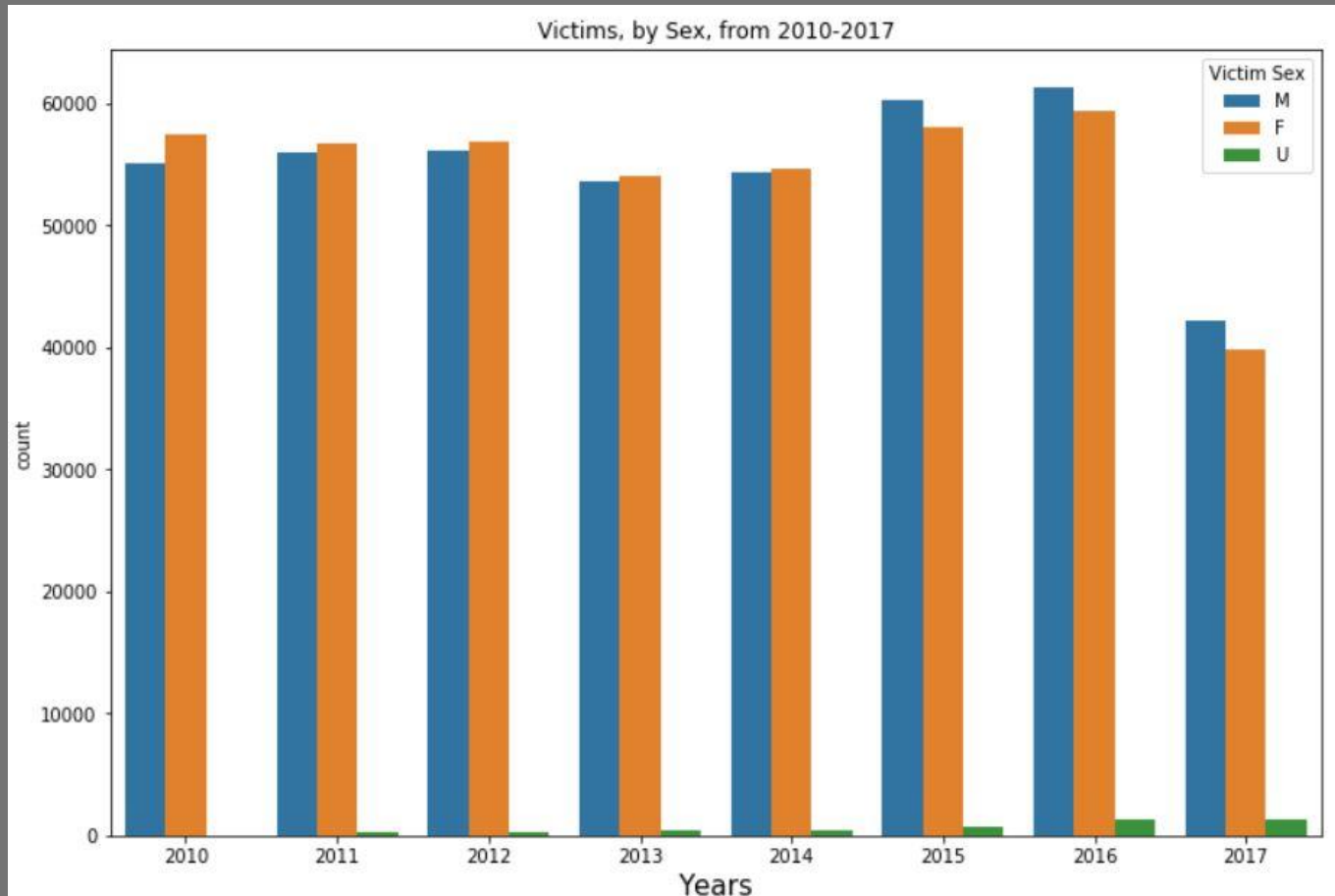
TOP FIVE CRIMES



NUMBER OF VICTIMS PER YEAR



NUMBER OF CRIMES PER GENDER FROM 2010 TO 2017



Prediction of crime code given victim age and victim sex

- From EDA we can see that there are significant amount of crimes correlated with age and gender. Hence we decided to predict the crime code given victim age and sex.
- We used Logistic Regression, SVM, KNN and Decision trees to get this prediction.

Prediction of time occurred given weekday

- Prediction of time given a day is very important to detect crimes.
- We used Linear Regression for this which did not yield very favorable results.
- We have now pre-processed the time data available to us to fit the model – we plan to apply the following algorithms which are more useful for time-series prediction:
 1. Autoregression (AR)
 2. Moving Average (MA)
 3. Autoregressive Moving Average (ARMA)
 4. Autoregressive Integrated Moving Average (ARIMA)
 5. Seasonal Autoregressive Integrated Moving-Average (SARIMA)
 6. Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX)
 7. Vector Autoregression (VAR)
 8. Vector Autoregression Moving-Average (VARMA)
 9. Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX)
 10. Simple Exponential Smoothing (SES)
 11. Holt Winter's Exponential Smoothing (HWES)
 - 12. Multivariate LSTM Forecast Model**
 - 13. Prophet**

(source: <https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/>)

Prediction of Area Code using date occurred, time occurred and crime type

- Prediction of area code given date occurred and time occurred of crime.
- We used Logistic Regression, SVM, KNN algorithms for this prediction.
- Pre-processed date occurred to its ordinal value (datatype=long)

Accuracies

Question	Logistic/Linear Regression	Decision Trees	SVM	KNN
Prediction of crime code given victim age and victim sex	With grouping: ~33.12% Without grouping: ~11.55%	With grouping: ~19.96% Without grouping: ~13.32%	H/W unable to handle	With grouping: ~21.61% Without grouping: ~16%
Prediction of time occurred given weekday	rmse: 646.66	-	-	-
Prediction of area code given date occurred, time occurred and crime type	~7%	~6.2%	H/W unable to handle	~6.7%

Milestones

TIMELINE	ACTIVITY	STATUS
SEP 1 st – SEP 7 th	Requirement Gathering	Completed
SEP 8 th – SEP 11 th	Project Proposal	Completed
SEP 12 th – SEP 23 rd	Analyze the Dataset	Completed
SEP 25 th	Project Idea Presentation	Completed
SEP 26 th – OCT 15 th	Cleaning the data and trying to answer one question each per member	Continued
OCT 16 th	Presentation and mid term report	Completed
OCT 17 th – NOV 16 th	Answer the remaining questions	To be Completed
NOV 20 th	Final report submission and presentation	To be Completed
NOV 27 th	Present the final results	To be Completed

Questions/ Suggestions



Thank You 😊