UNIVERSITY
GWALIOR • MP • INDIA
"CELEBRATING DREAMS"

ITM University Gwalior

PBL Synopsis File

On

Crime Data Analysis

(MCA-205)

Problem Statement:

Are gender demography groups susceptible to some specific crimes?

Submitted To:

Dr. Sanjay Jain
Professor, Dept. Of CSA

Submitted By:

Prashant Singh
MCAN1CA22019

Contents

1. Introduction 1
1.1 Abstract of PBL 1
2. Data Source and Tools 2
2.1 Tools Used 2
2.2 Attributes of Dataset 2
2.1.1 CSV: Crime_Data_from_2020_to_Present.csv 3
2.3 Extraction from Data Set 4
2.2.1 CSV: Cleaned_Data.csv 5
3. Instances and Visualization from the Data 6
3.1 Top Crimes against Men 7
3.2 Top Crimes against Women 7
3.3 Top Crimes against Non-Binary (Un-Confirmed Gender Identity 7
3.4 Top Crimes against Homosexuals 7
3.5 Visualization and Conclusion 8
4. Proposed Work 9

1. Introduction

Crime data analysis is an essential tool for comprehending the patterns and trends in crime in different regions and is highly useful in the United States. With a population of over 328 million, the United States faces significant crime problems that can affect the safety and well-being of its citizens.

The US government and law enforcement agencies collect crime data from various sources and maintain a national crime database to keep track of crimes committed in the country. However, crime rates and trends can vary significantly across different US states due to factors such as demographics, socio-economic status, and culture. By analyzing crime data at the state level, it is possible to gain valuable insights into the nature and extent of crime, as well as the effectiveness of law enforcement measures in different regions. This information can then be used to develop and implement policies and strategies that can help prevent crime and improve public safety in US states. Therefore, crime data analysis is an essential tool in ensuring the security and well-being of the citizens of the United States.

1.1 Abstract of PBL

This data analysis explores whether gender demographic groups are more likely to be victims of certain types of crimes. Using data from the Gov Crime Data from Los Angeles, CA, USA i.e., National Crime Victimization Survey (NCVS), the analysis examines the prevalence of different types of crime victimization among men, women, non-binaries and others (Majorly focusing on men and women).

The results indicate that men are more likely to be victims of violent crimes committed by strangers, such as robbery and aggravated assault, as well as homicide. Women, on the other hand, are more likely to be victims of sexual assault and domestic violence. These findings have significant implications for understanding and addressing gender-based violence and victimization, and can inform policy and intervention efforts aimed at preventing and reducing these crimes.

2. Data Source and Tools

Obtaining valid and accurate data is crucial for any research or analysis. Without valid data, the results of any analysis or research are unreliable and may lead to incorrect conclusions. It is essential to ensure that the data collected is accurate, complete, and relevant to the research questions being addressed. Accurate data is also important for making informed decisions, policies, and interventions that can have significant impacts on individuals and society. Inaccurate data can result in flawed policies and ineffective interventions, which can have negative consequences for individuals and society as a whole. Therefore, ensuring the validity and accuracy of data is crucial in any research or analysis, particularly in fields such as crime data analysis, where the stakes are high and the impact can be significant.

2.1 Tools Used

I have chosen python as it is very versatile and has huge number of libraries. Specifically Pandas and Matplotlib.

2.2 Attributes of Dataset

The data source used in this project has been extracted from the existing data on Catalog.Data.Gov site. From the creator of dataset:

"*This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy.*"

Dataset Link: Gov Crime Data from Los Angeles, CA, USA

The dataset is of the shape of: (677905, 28)

677905 Observations with respect to 28 Attributes

Following is detailed description of table of dataset that is used in the PBL Project (with Dtypes):

2.1.1 CSV: Crime_Data_from_2020_to_Present.csv

Attribute	Dtype
DR_NO	int64
Date Rptd	object
DATE OCC	object
TIME OCC	int64
AREA	int64
AREA NAME	object
Rpt Dist No	int64
Part 1-2	int64
Crm Cd	int64
Crm Cd Desc	object
Mocodes	object
Vict Age	int64
Vict Sex	object
Vict Descent	object
Premis Cd	float64
Premis Desc	object
Weapon Used Cd	float64
Weapon Desc	object
Status	object
Status Desc	object
Crm Cd 1	float64
Crm Cd 2	float64
Crm Cd 3	float64
Crm Cd 4	float64
LOCATION	object
Cross Street	object
LAT	float64
LON	float64

2.3 Extraction from Data Set

Mixed data in a data set can be a challenge when performing data analysis. Mixed data refers to data sets that contain both numerical and categorical data. While numerical data can be easily analyzed using mathematical and statistical tools, categorical data requires different approaches to extract meaningful insights. One cannot perform analysis on mixed data directly, and it is essential to manipulate the data to create new data sets that can be analyzed using statistical and mathematical tools. This requires data manipulation techniques such as one-hot encoding, label encoding, and feature scaling. By creating new data sets that contain only numerical data or properly encoded categorical data, it becomes possible to perform analysis and extract meaningful insights. Therefore, data manipulation is an essential step in the data analysis process, particularly when dealing with mixed data. It ensures that the data is in the correct format for analysis and leads to more accurate results.

From the given dataset I have dropped multiple attributes that were inducing redundancy.

I have removed ‘Crm Cd Desc', 'AREA NAME', 'Premis Desc', 'Weapon Desc','Status Desc', ,'Cross Street' as these were already referenced with their respectable codes as given and table below.

I have removed 'DR_NO', 'Mocodes','LAT','LON','LOCATION','Vict Descent','Crm Cd 1','Crm Cd 2','Crm Cd 3','Crm Cd 4' as these values were not required for the analysis.

I have also removed the crimes whose count was less than the threshold value of 1000 as these were very negligible and only occur seldomly in random pattern which is why it is not usable in our analysis.

The shape of the new dataset is: (659622, 12)

659622 Attributes with respect to 12 Columns

Following are the attributes of extracted data:

2.2.1 CSV: Cleaned_Data.csv

Attributes	Dtypes
Date Rptd	object
DATE OCC	object
TIME OCC	int64
AREA	int64
Rpt Dist No	int64
Part 1-2	int64
Crime Code	int64
Vict Age	int64
Vict Sex	object
Premis Cd	float64
Weapon Used Cd	float64
Status	Object

3. Instances and Visualization from the Data

As the dataset is too large to provide its full instance, following are the head of the cleaned dataset:

	Date Rptd	DATE OCC	TIME OCC	AREA	Rpt Dist No	Part 1-2	Crm Cd	Vict Age	Vict Sex	Premis Cd	Weapon Used Cd	Status
0	01-08-2020	01-08-2020	2230	3	377	2	624	36	F	501	400	AO
1	01-02-2020	01-01-2020	330	1	163	2	624	25	M	102	500	IC
2	01-01-2020	01-01-2020	1730	15	1543	2	745	76	F	502	NaN	IC
3	01-01-2020	01-01-2020	415	19	1998	2	740	31	X	409	NaN	IC
4	01-02-2020	01-01-2020	30	1	163	1	121	25	F	735	500	IC

In the table,

Genders:

F- Female

M-Male

X- Non-Binary

H- Homo-Sexual

For description of other column check the included files.

3.1 Top Crimes against Men

Crm Cd	Crm Cd Desc	percent_m	percent_f	percent_x	percent_h
230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	71.4	26.8	1.8	0.0
220	ATTEMPTED ROBBERY	64.5	25.8	9.7	0.0
210	ROBBERY	59.9	24.1	15.9	0.1

3.2 Top Crimes against Women

Crm Cd	Crm Cd Desc	percent_m	percent_f	percent_x	percent_h
236	INTIMATE PARTNER - AGGRAVATED ASSAULT	22.449199	77.443286	0.096764	0.010752
626	INTIMATE PARTNER - SIMPLE ASSAULT	24.155965	75.65518	0.180138	0.008716
901	VIOLATION OF RESTRAINING ORDER	23.699422	75.641618	0.647399	0.011561

3.3 Top Crimes against Non-Binary (Un-Confirmed Gender Identity

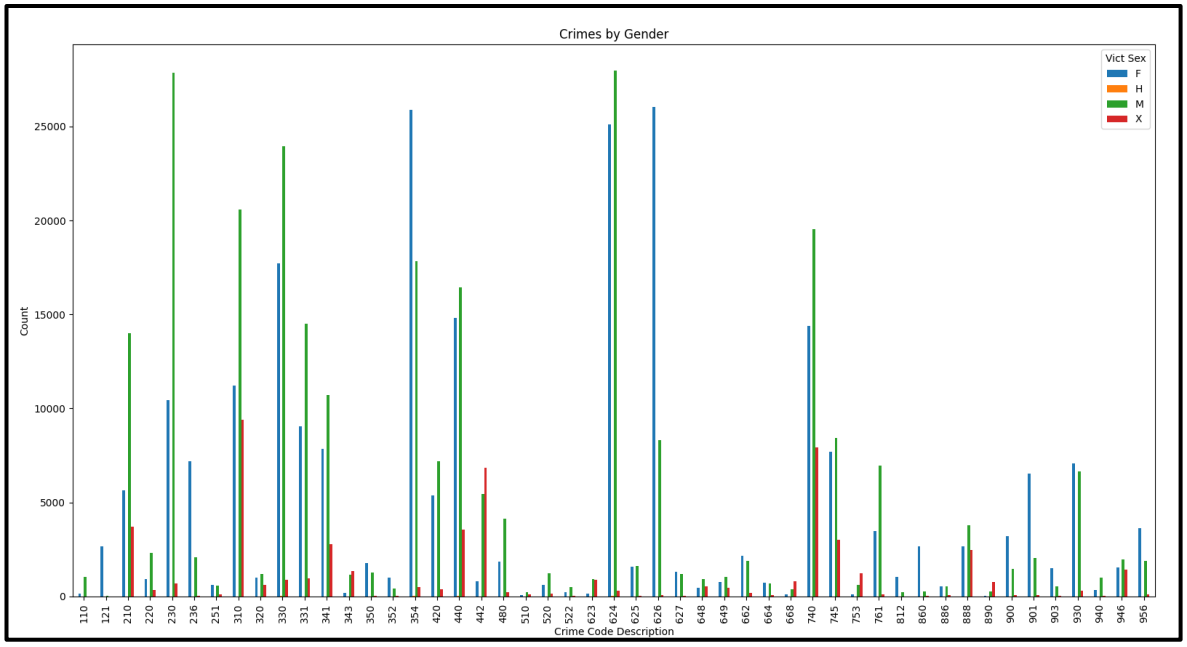
Crm Cd	Crm Cd Desc	percent_m	percent_f	percent_x	percent_h
442	SHOPLIFTING - PETTY THEFT ($950 & UNDER)	41.625842	6.085426	52.281078	0.007655
310	BURGLARY	49.951475	27.220012	22.826087	0.002426
740	VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)	46.661409	34.415926	18.913106	0.009559

3.4 Top Crimes against Homosexuals

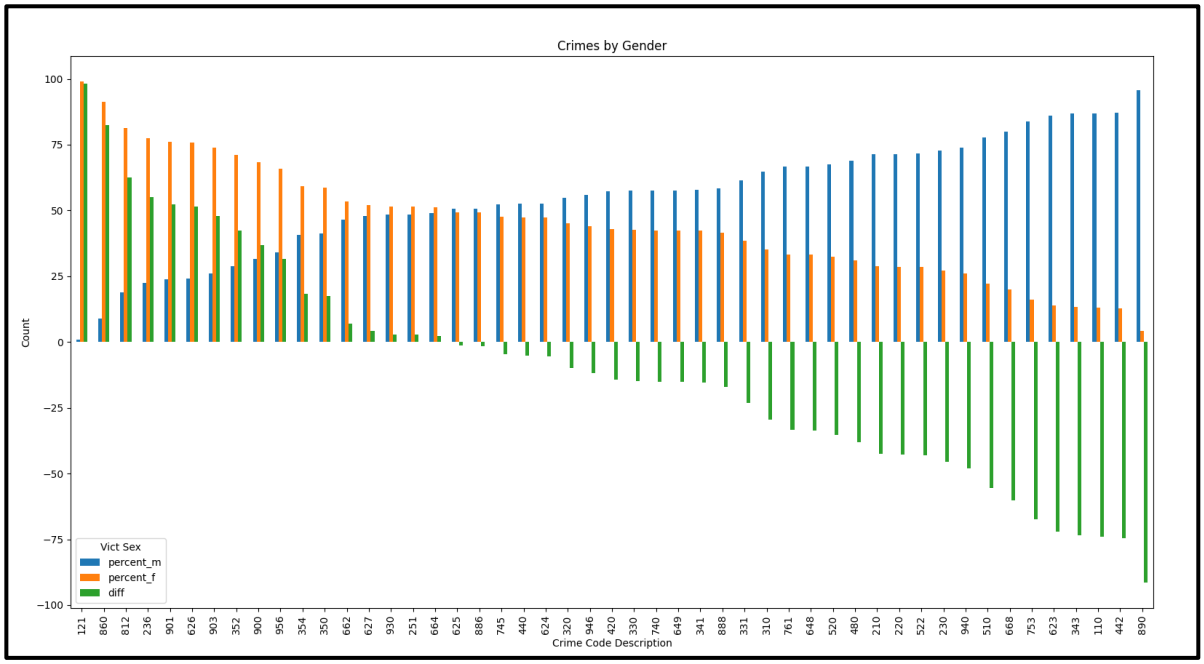
Vict Sex	Crm Cd Desc	percent_m	percent_f	percent_x	percent_h
251	SHOTS FIRED AT INHABITED DWELLING	44.40184	47.08589	8.435583	0.076687
352	PICKPOCKET	28.652482	70.638298	0.638298	0.070922
354	THEFT OF IDENTITY	40.304415	58.524064	1.130812	0.040709

3.5 Visualization and Conclusion

The following Graph shows the number of victims that are related to crime separated by genders. We can clearly see that the number of Male victims is higher than any other else followed by females and then non binaries.



The following graph shows the disparity between Criminal Cases that occur with men to that of women. On correlating the two graphs, the results indicate that men are more likely to be victims of violent crimes committed by strangers, such as robbery and aggravated assault, as well as homicide. Women, on the other hand, are more likely to be victims of sexual assault and domestic violence. These findings have significant implications for understanding and addressing gender-based violence and victimization, and can inform policy and intervention efforts aimed at preventing and reducing these crimes.



4. Proposed Work

The proposed ML project aims to develop a predictive model to determine whether gender demographic groups are more likely to be victims of certain types of crimes. The project will utilize data from the National Crime Victimization Survey (NCVS) and other relevant sources to train and validate the model.

The proposed work will involve several steps, including data pre-processing, feature engineering, model selection, and evaluation. Firstly, the collected data will be pre-processed to remove any inconsistencies and missing values. Next, feature engineering techniques such as one-hot encoding and feature scaling will be applied to transform the data into a format that can be used for training the model.

The next step will involve selecting an appropriate machine learning algorithm that can effectively predict crime victimization rates among different gender demographic groups. Several algorithms will be evaluated, including decision trees, logistic regression, and neural networks, among others. The algorithm that produces the highest predictive accuracy will be selected for use in the final model.

Finally, the model will be evaluated using various metrics such as precision, recall, and F1 score. The results of the evaluation will be used to identify any areas for improvement and fine-tune the model further. The developed model can be used to predict the likelihood of crime victimization among different gender demographic groups, and the insights gained can be used to inform policy and intervention efforts aimed at preventing and reducing crime victimization rates.

In conclusion, the proposed ML project is an important step in developing a predictive model to determine whether gender demographic groups are more likely to be victims of certain types of crimes. The insights gained from this project can be used to develop targeted interventions and policies aimed at preventing and reducing crime victimization in our communities.