

Project Report

(Application Development)

S.NO	Name	ERP
1	Farjad Ahmed	29405
2	Azfar	29398

Abstract

This report provides an in-depth analysis of Los Angeles crime data to reveal meaningful insights about the city and crime trends and patterns. Using multi-year data, the study uses advanced statistical methods to examine various dimensions of crime, including but not limited to spatial distribution, temporal variation, and prevalence of crime types.

The research examines the spatial dynamics of crime, maps the areas of occurrence, and identifies possible areas of spread. Temporal variations are examined to understand the cyclical nature of criminal activity, allowing seasonal trends and patterns to be identified.

The results of this analysis help to understand the more complex dynamics of crime in Los Angeles and provide insights that can inform policy decisions, resource allocation, and the development of targeted interventions. As cities around the world grapple with the challenge of public safety, this report serves as a model for leveraging data analytics to provide insights into crime prevention and urban safety.

Table of Contents

1.	Domain Analysis	1
1.1	Business process related to dataset	1
1.1.1	Introduction:	1
1.1.2	Business Process Overview:	1
1.2	Type of diagnosis:	2
1.2.1	Technical analysis:.....	2
1.2.2	Predictive Analytics:.....	2
1.2.3	Spatial Analysis:.....	2
1.3	The usefulness of analytics for business processes:	2
1.3.1	Proactive policy:.....	2
1.3.2	Optimization Resources:	3
1.3.3	Social Security:	3
1.3.4	Make political decisions:.....	3
2	Describing Data type	3
2.1	Classification and Regression	3
2.1.1	Classification:	3
2.1.2	Regression:.....	4
2.2	Balance or Imbalance Dataset?	4
2.2.1	Consequences of the imbalanced data set:	4
2.2.1	Data Composition	4
3	Data Cleaning	9
4	Exploratory Data Analysis	10
4.1	Univariate Analysis Of Project:	11
4.2	Bi-variate Analysis Of Project.....	24
5	Data Pre-Processing	29
6	Dash Plotly	31

1. Domain Analysis

1.1 Business process related to dataset

1.1.1 Introduction:

Our dataset of Los Angeles crime data from 2020 to the present provides a rich source of information that can be used to improve decision-making processes in law enforcement, urban planning, and public safety organizations. This analysis examines the business processes associated with this data set and examines the types of analysis that can be performed, showing how these analyses can optimize selected business processes.

1.1.2 Business Process Overview:

The main business functions related to Los Angeles crime data are maintaining public safety and maintaining order in the city. This main objective can be divided into several main areas:

1.1.2.1 *Crime detection and prevention:*

Law enforcement agencies use datasets to identify patterns and trends in criminal activity over time. Predictive analytics can be used to predict areas of potential risk and allocate resources accordingly.

1.1.2.2 *Distribution of resources:*

Police departments can use the data to optimize staffing and resource allocation based on past crime patterns. Analysis of the data can help identify areas where additional monitoring and targeted interventions are needed.

1.1.2.3 *Community participation:*

By sharing publicly available crime data, you can raise awareness and contribute to the community. Crime analysis can help you

implement community policing strategies tailored to your specific area.

1.1.2.4 *Policymaking:*

Urban planners and policymakers can use the data to make decisions related to urban planning and development. Insights from crime analysis can influence policymaking to reduce crime rates and improve public safety.

1.2 Type of diagnosis:

1.2.1 Technical analysis:

Identify and summarize key patterns and trends in crime data over a specific period. It generates summary statistics such as crime rates, frequency of different types of crime, and changes over time.

1.2.2 Predictive Analytics:

Build models to predict future crime rates and identify potential crime hotspots. Machine learning algorithms are used to predict the likelihood that certain types of crimes will occur in different locations.

1.2.3 Spatial Analysis:

Map crime data to see spatial patterns and hotspots. Implement Geographic Information System (GIS) tools to analyze the spatial distribution of crime and support resource allocation decisions.

1.3 The usefulness of analytics for business processes:

1.3.1 Proactive policy:

Predictive analytics can help law enforcement agencies speed up response times and prevent incidents by targeting high-risk areas.

1.3.2 Optimization Resources:

Efficient allocation of resources based on data-driven information can reduce costs and avoid inefficiencies.

1.3.3 Social Security:

Transparent sharing of crime data promotes trust between law enforcement and the community and fosters cooperation in crime prevention.

1.3.4 Make political decisions:

Policymakers can make decisions about urban planning and public safety initiatives, helping to create better urban environments.

In conclusion, Los Angeles crime data from 2020 to the present has great potential to improve public safety and law enforcement practices. Through description, prediction, spatialization, and documentation, stakeholders gain valuable information to inform effective policing, optimize resource allocation, and build societal trust to make informed policy. This data-driven approach is critical to addressing the dynamic challenges of maintaining urban environments and safety.

2 Describing Data type

2.1 Classification and Regression

Los Angeles Crime Data from 2020 to the present, can involve various types of analyses depending on the specific goals of our project. Generally, crime data analysis can encompass different types of tasks, and the nature of the analysis would determine whether it falls into the category of classification, regression, clustering, or other types of analyses. Here are few things considering classification and regression:

2.1.1 Classification:

- If the goal is to predict or classify types of crimes (e.g., burglary, assault, vandalism), then our analysis may involve classification. You might use machine learning algorithms to train a model that

can predict the category of a crime based on various features such as location, time, and other relevant factors.

2.1.2 Regression:

- Regression analysis might be suitable if our focus is on predicting numerical values related to crime, such as the number of incidents in a specific area over time. For example, you could predict the crime rate based on various socio-economic or demographic factors.

Hence our data can provide both classification and regression analysis

2.2 Balance or Imbalance Dataset?

It's an imbalanced data set as certain types of crimes are more occurring than others.

2.2.1 Consequences of the imbalanced data set:

Model Performance Biasedness:

Machine learning models trained on imbalanced data sets can fail in the generalization layer. The model performs well in the predictions of the majority classes but performs poorly in the minority classes. Such as in this case violent crimes like arson, murder etc.

Decision-making challenges:

Concerning crime data analysis, an unbalanced database can cause problems in resource allocation and decision-making. If a particular type of crime is underrepresented, the model may not provide sufficient information for effective crime prevention strategies.

2.2.1 Data Composition

Data	Attribute name	Type	Missing Values (percentage)	Importance
DR_NO	Division record number	int64	0	L
Date Rptd	Date report	object	0	H
DATE OCC	Date occurrence	object	0	H
TIME OCC	Time occurrence	int64	0	H
AREA	Area code	int64	0	A
AREA NAME	Area name	object	0	H
Rpt Dist No	Four-digit code that represents a sub-area within a Geographic Area	int64	0	L
Part 1-2		int64	0	D
Crm Cd	Informs the crime committed	int64	0	H
Crm Cd Desc	Crime Code	object	0	H
Mocodes	Modus Operandi: Action related to a suspect in the commission of a crime	object	13.85789	D
Vict Age	Victim age	int64	0	A
Vict Sex	Gender	object	13.17855	A
Vict Descent	Victim decent or race	object	13.17951	A
Premis Cd	The type of structure, vehicle, or location where the crime took place	float64	0.001192	H
Premis Desc	Defines the Premise Code provided.	object	0.060317	A
Weapon Used Cd	The type of weapon used in the crime.	float64	65.14595	D
Weapon Desc	Defines the Weapon Used Code provided	object	65.14595	D
Status	Status of case	object	0	L
Status Desc	Defines the Status Code provided	object	0	L
Crm Cd 1	Indicates the crime committed. Crime Code 1 is the primary and most serious one	float64	0.001311	D

Crm Cd 2	May contain a code for an additional crime, less serious than Crime Code 1	float64	92.65301	D
Crm Cd 3	May contain a code for an additional crime, less serious than Crime Code 1.	float64	99.75194	D
Crm Cd 4	May contain a code for an additional crime, less serious than Crime Code 1.	float64	99.99273	D
LOCATION	Street address of crime incident	object	0	H
Cross Street	Cross Street of rounded Address	object	84.04913	D
LAT	Latitude	float64	0	A
LON	Longitude	float64	0	A

- DR_NO: Division of Records Number: Official file number made up of a 2-digit year, area ID, and 5 digits
- Date Rptd: MM/DD/YYYY
- DATE OCC: MM/DD/YYYY
- TIME OCC: In 24-hour military time.
- AREA: The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21.
- AREA NAME: The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example, the 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.

- Rpt Dist No: A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons. Find LAPD Reporting Districts on the LA City GeoHub at http://geohub.lacity.org/datasets/c4f83909b81d4786aa8ba8a74a4b4db1_4
- Part 1-2
- Crm Cd: Indicates the crime committed. (Same as Crime Code 1)
- Crm Cd Desc: Defines the Crime Code provided.
- MO codes: Modus Operandi: Activities associated with the suspect in the commission of the crime. See the attached PDF for a list of MO Codes in numerical order. [https://data.lacity.org/api/views/y8tr-7khq/files/3a967fbd-f210-4857-bc52-60230efe256c?download=true&filename=MO%20CODES%20\(numerical%20order\).pdf](https://data.lacity.org/api/views/y8tr-7khq/files/3a967fbd-f210-4857-bc52-60230efe256c?download=true&filename=MO%20CODES%20(numerical%20order).pdf)
- Vict Age: Two-character numeric
- Vict Sex: F - Female M - Male X - Unknown
- Vict Descent: Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
- Premis Cd: The type of structure, vehicle, or location where the crime took place.
- Premis Desc: Defines the Premise Code provided.
- Weapon Used Cd: The type of weapon used in the crime.
- Weapon Desc: Defines the Weapon Used Code provided.
- Status: Status of the case. (IC is the default)

- Status Desc: Defines the Status Code provided.
- Crm Cd 1: Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
- Crm Cd 2: May contain a code for an additional crime, less serious than Crime Code 1.
- Crm Cd 3: May contain a code for an additional crime, less serious than Crime Code 1.
- Crm Cd 4: May contain a code for an additional crime, less serious than Crime Code 1.
- LOCATION: Street address of crime incident rounded to the nearest hundred blocks to maintain anonymity.
- Cross Street: Cross Street of rounded Address
- LAT: Latitude
- LON: Longitude

3 Data Cleaning

Data cleaning is a crucial step in the data analysis process to ensure the accuracy and reliability of the findings. For the given dataset with columns such as 'Crm Cd 1', 'Crm Cd 2', 'Crm Cd 3', 'Crm Cd 4', 'Cross Street', 'Weapon Used Cd', 'Weapon Desc', and 'Mocodes' dropped due to excessive missing values or irrelevance, the data cleaning process involves several key steps

Handling Missing Values:

- Identify and assess the extent of missing values in the remaining columns.
- Impute missing values using appropriate techniques, such as mean or median imputation for numerical variables, and mode imputation for categorical variables.

Date and Time Formatting:

Standardize the format of date and time columns ('Date Rptd', 'DATE OCC', 'TIME OCC') to a consistent format.

4 Exploratory Data Analysis

Exploratory Data Analysis is an essential section of the statistics evaluation technique wherein the primary intention is to summarize and visualize key characteristics, styles, and traits inside a dataset. EDA enables analysts and statistics scientists to gain insights into the underlying shape of the statistics, pick out outliers, and tell the next steps inside the evaluation. It frequently entails the usage of statistical graphics, precise statistics, and statistics visualization strategies to show styles and relationships

Univariate Analysis:

Univariate analysis focuses on examining a single variable at a time.

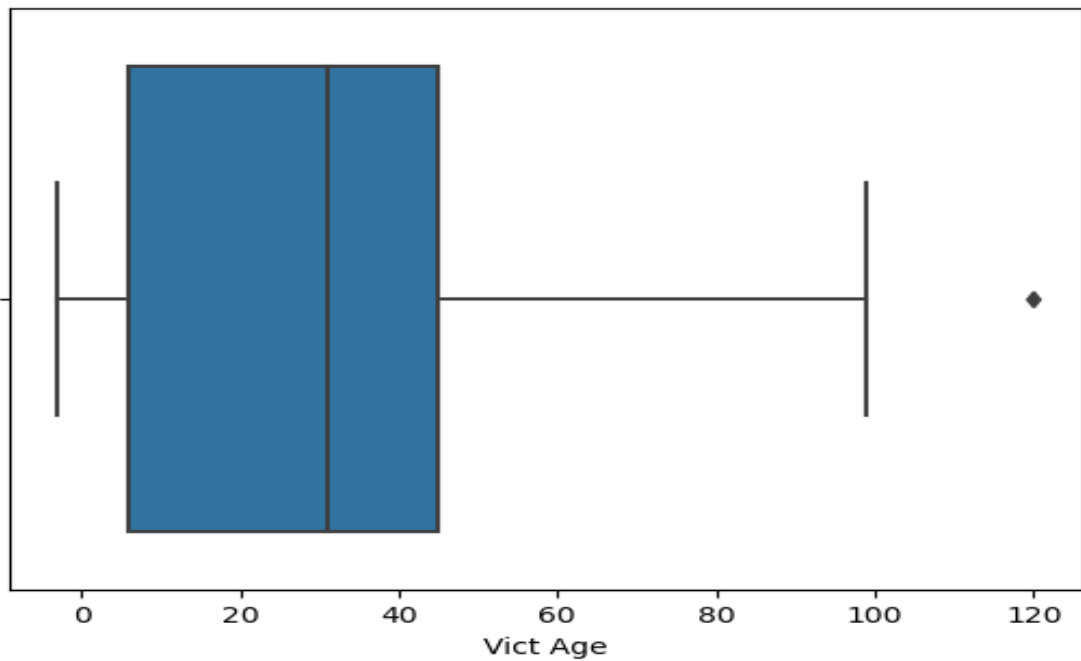
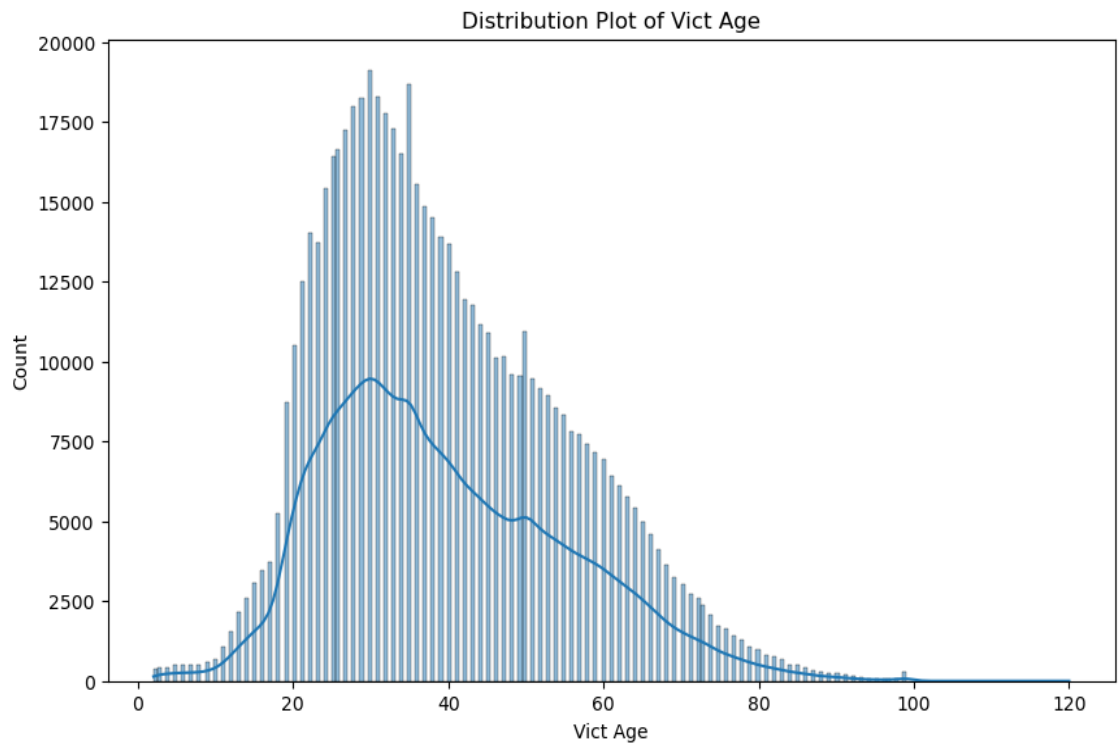
Examples of this analysis include box plots, histograms, count plots, etc.

Bivariate Analysis:

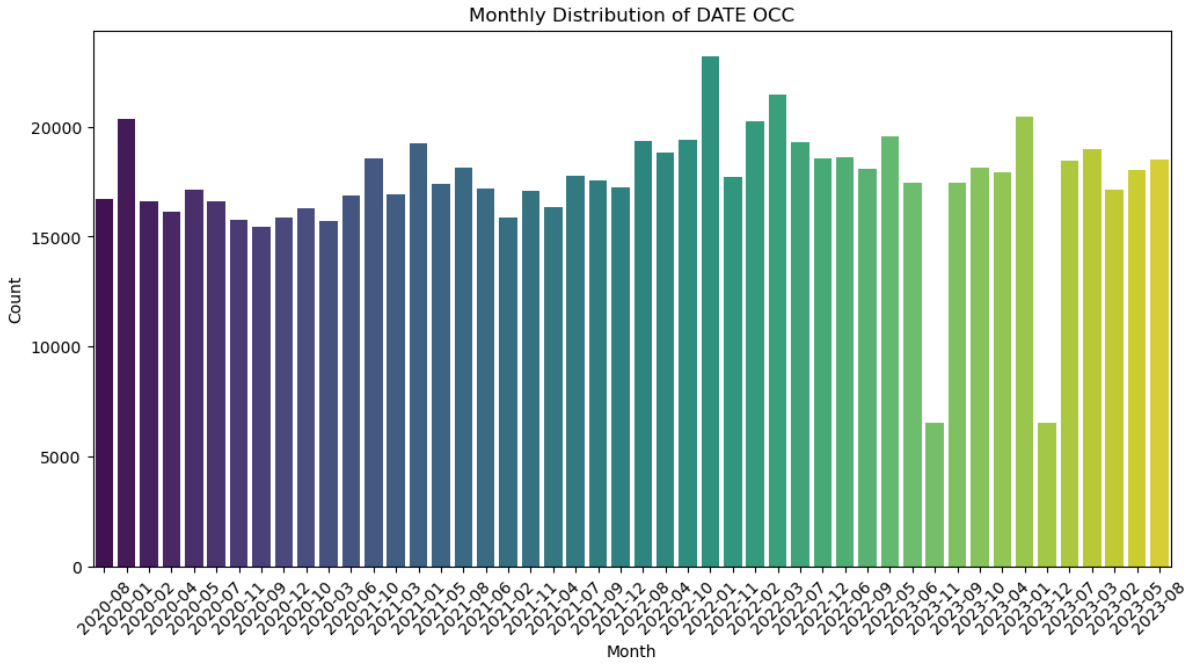
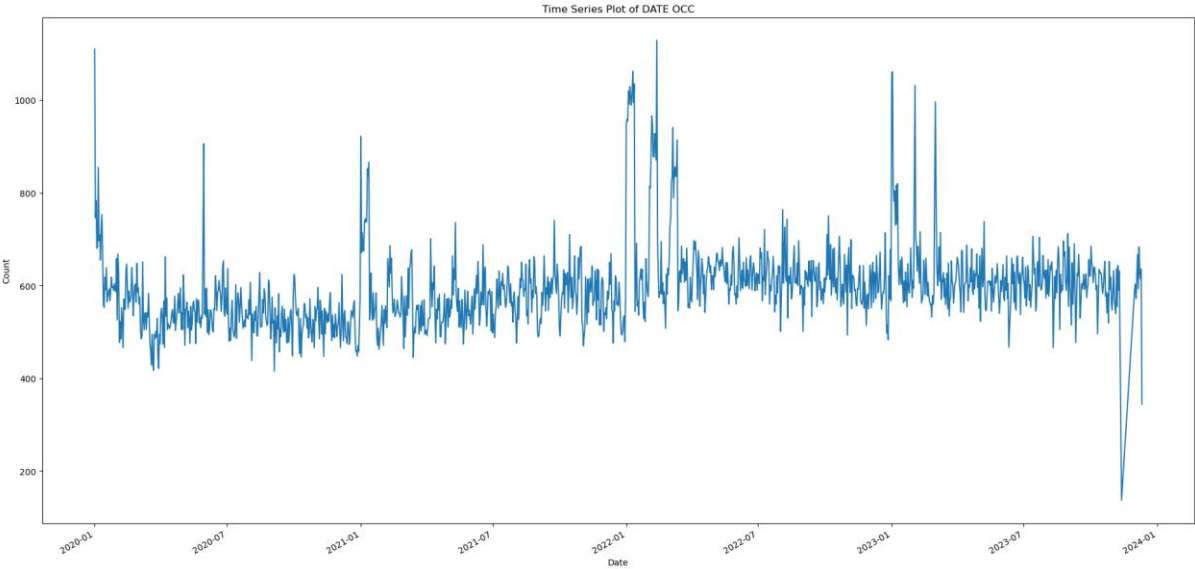
Bivariate analysis explores the relationship between two variables.

Examples of this analysis include Line Plot, Bar Plots, etc.

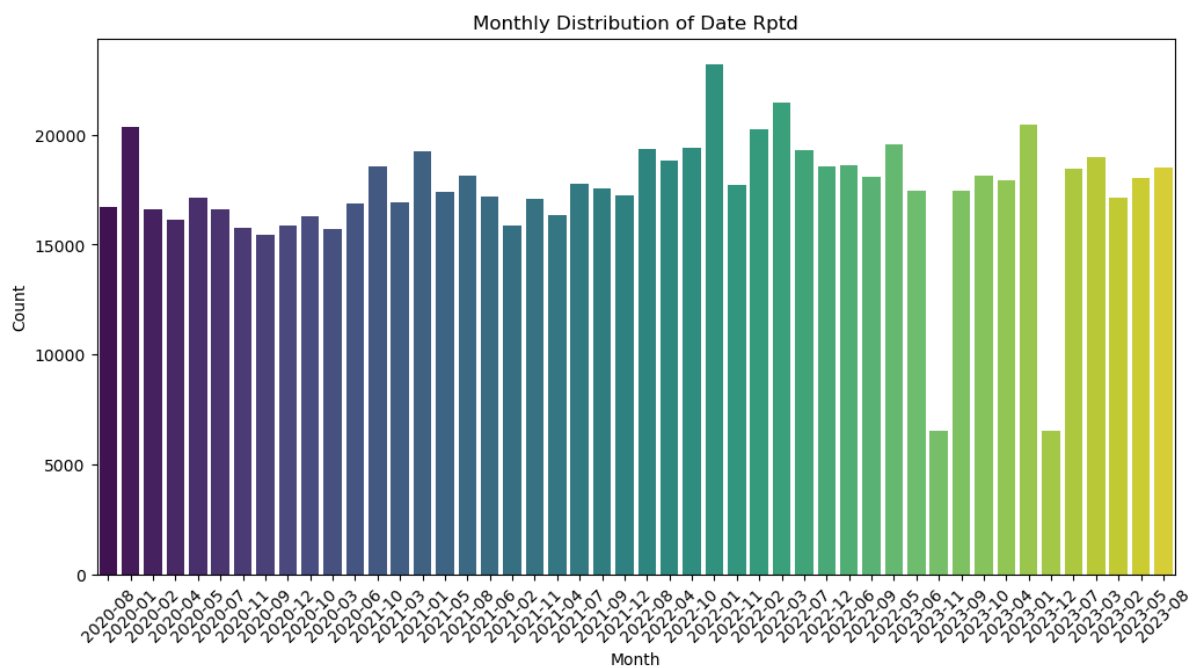
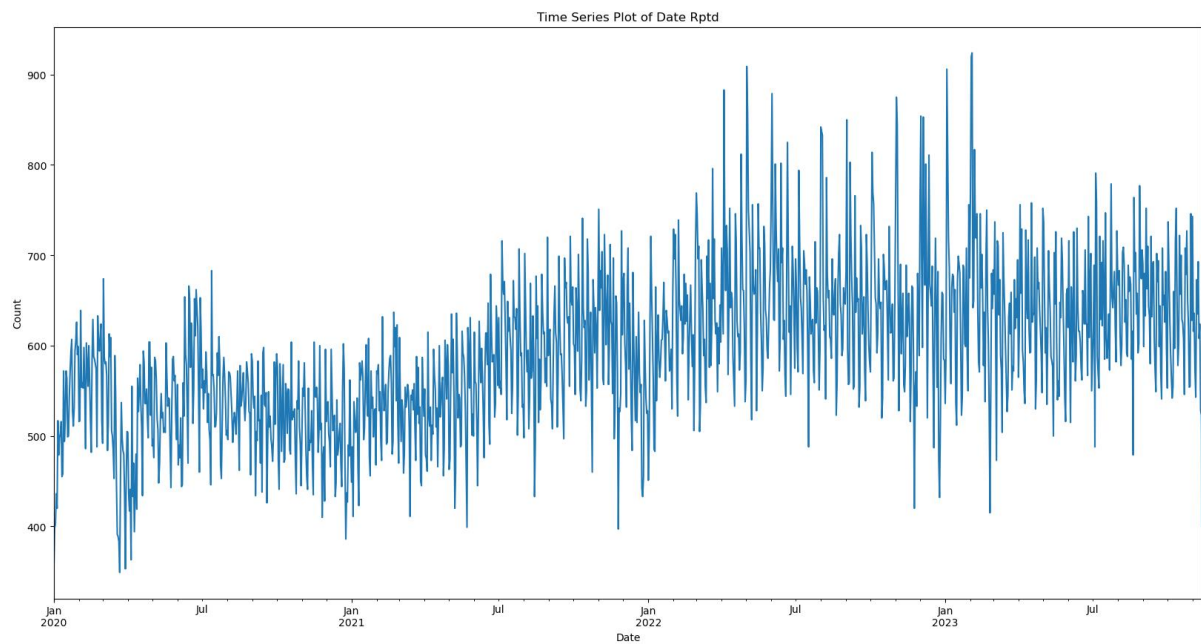
4.1 Univariate Analysis Of Project:



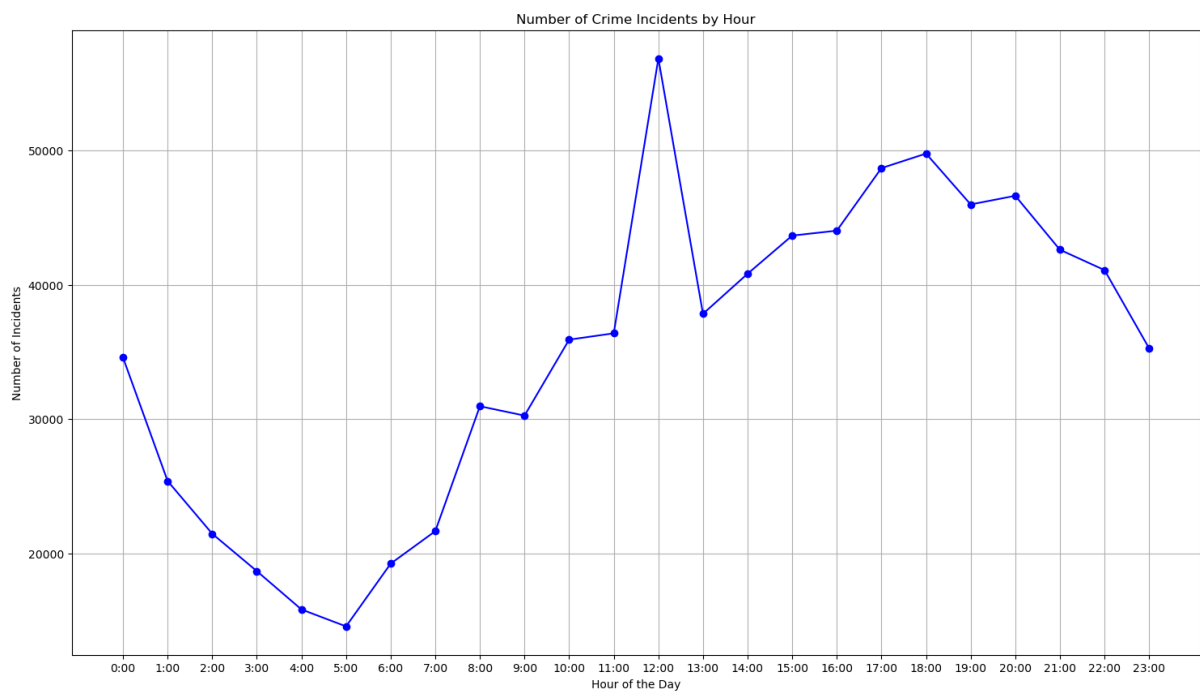
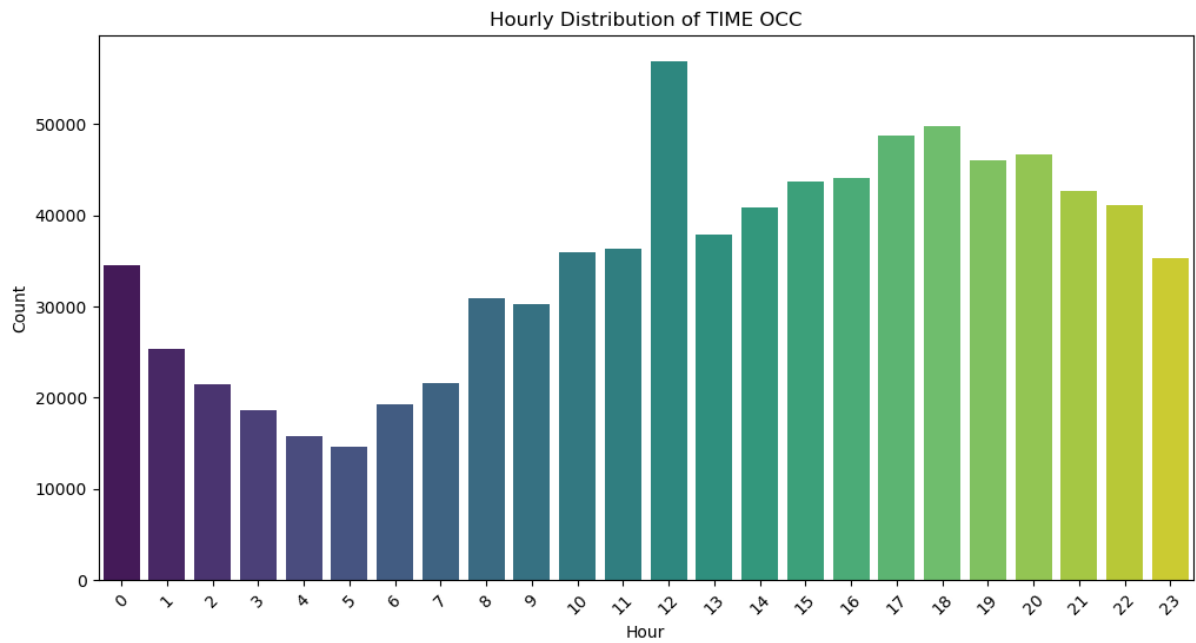
DATE OCC



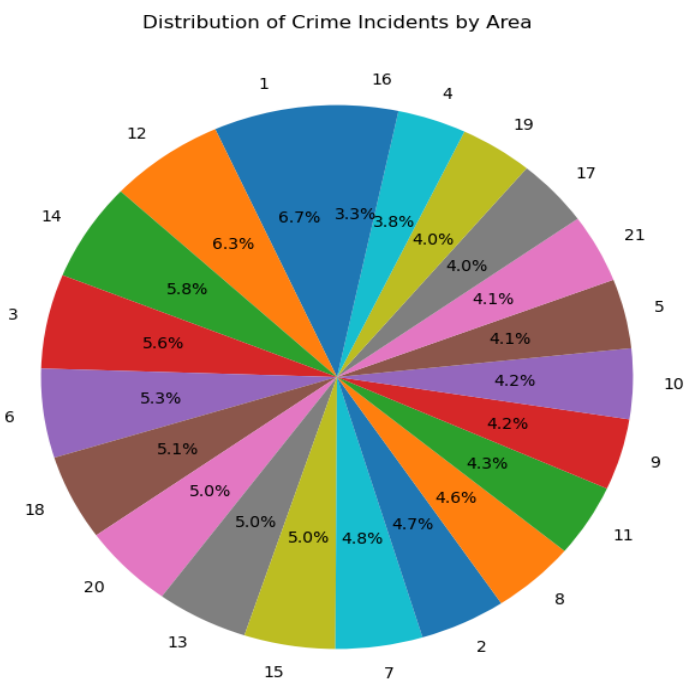
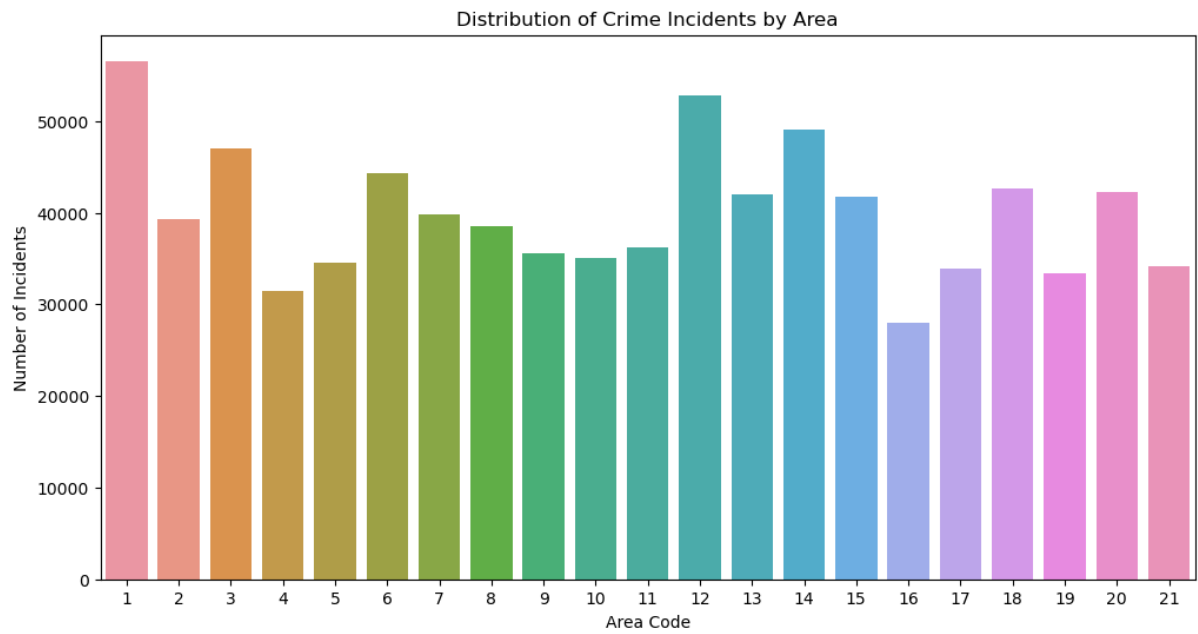
Date Rptd



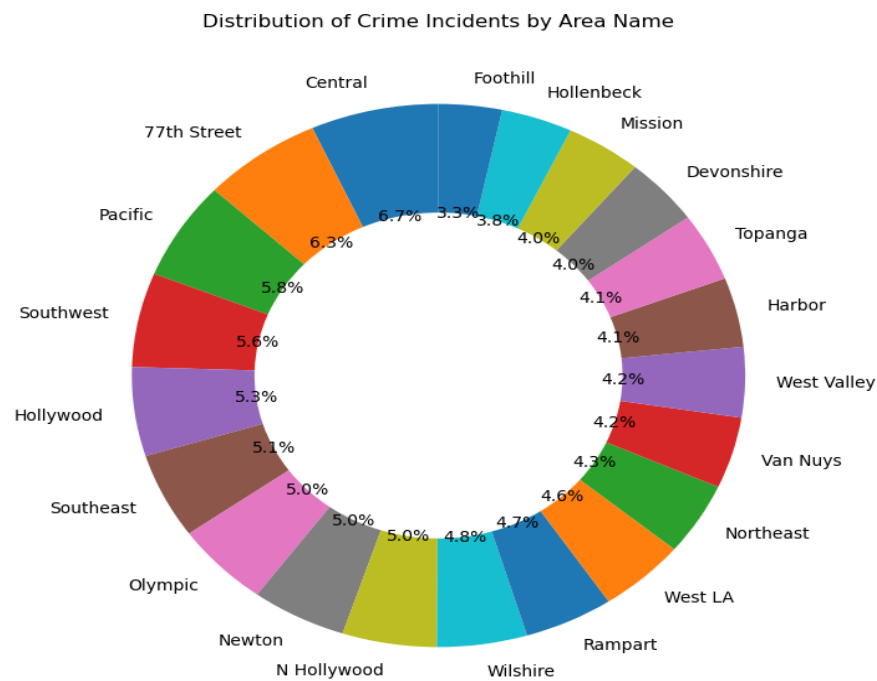
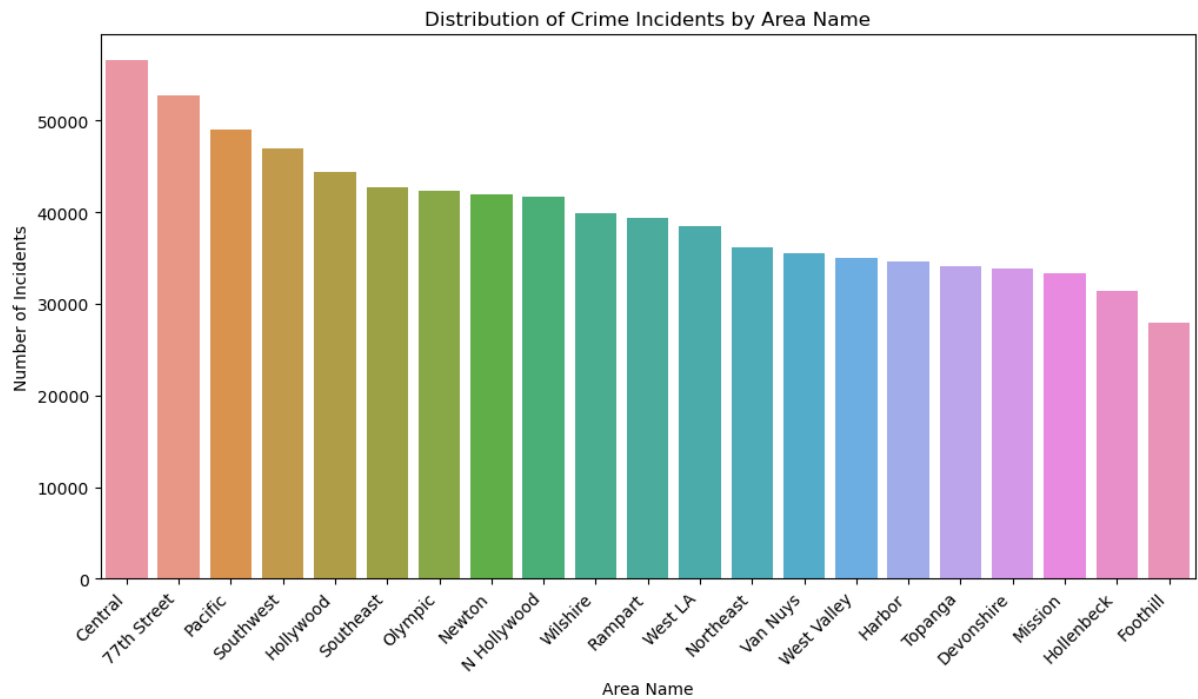
TIME OCC



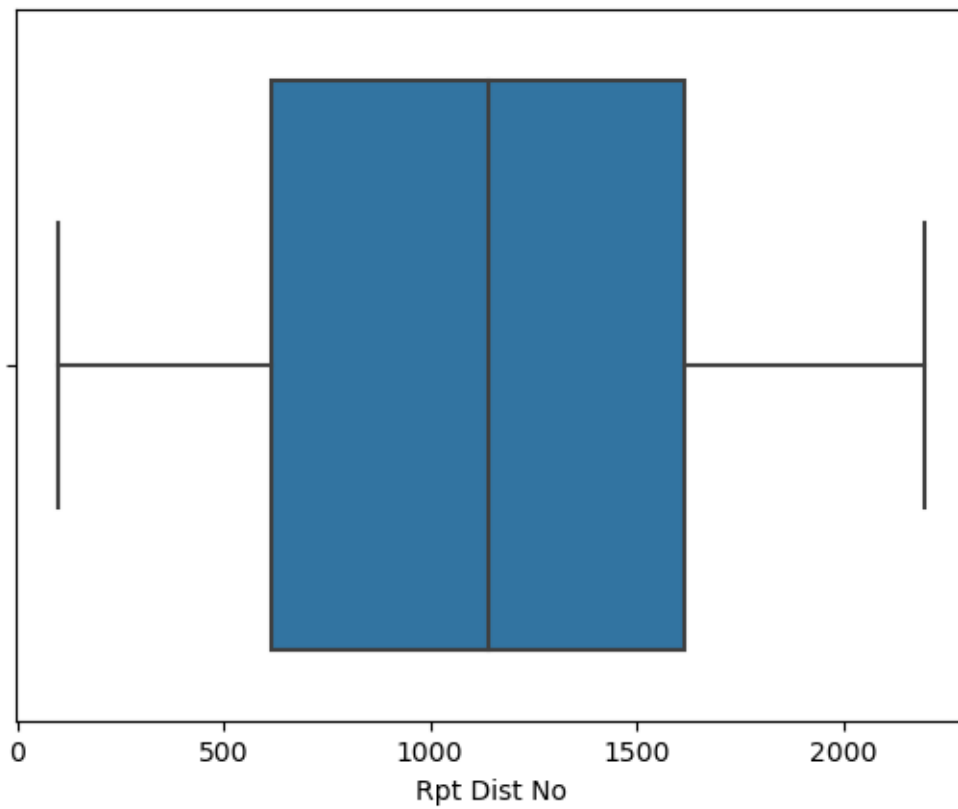
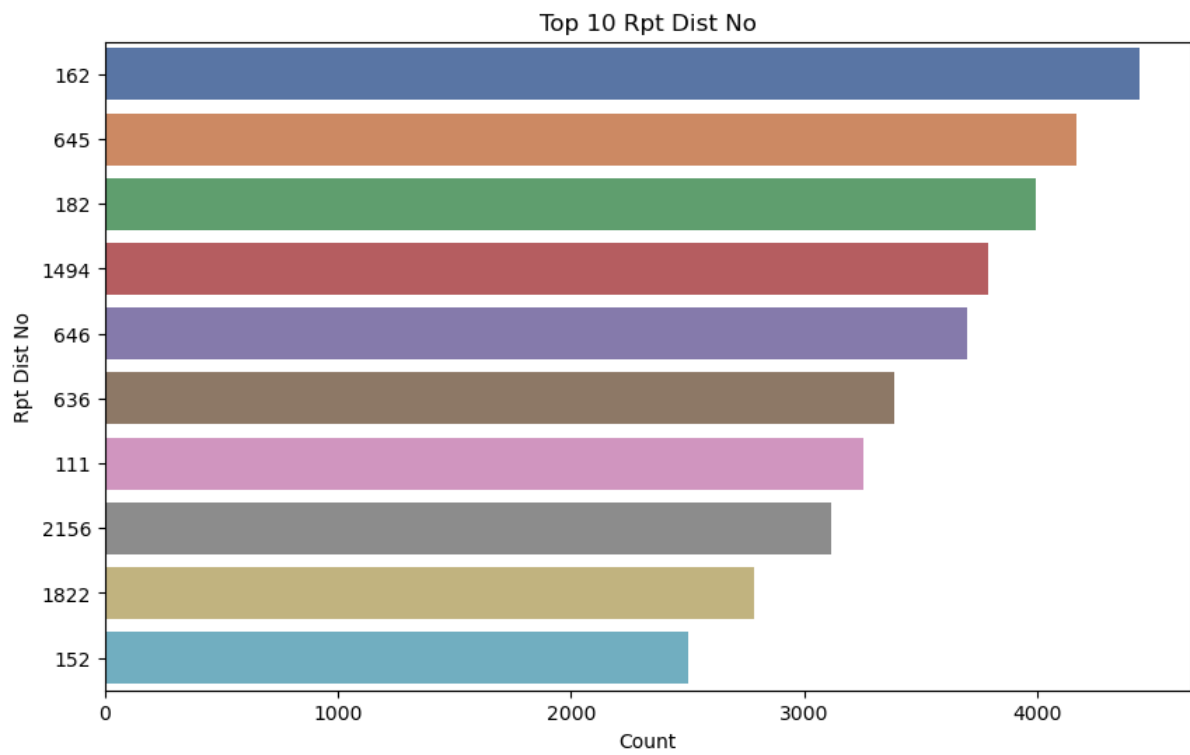
AREA



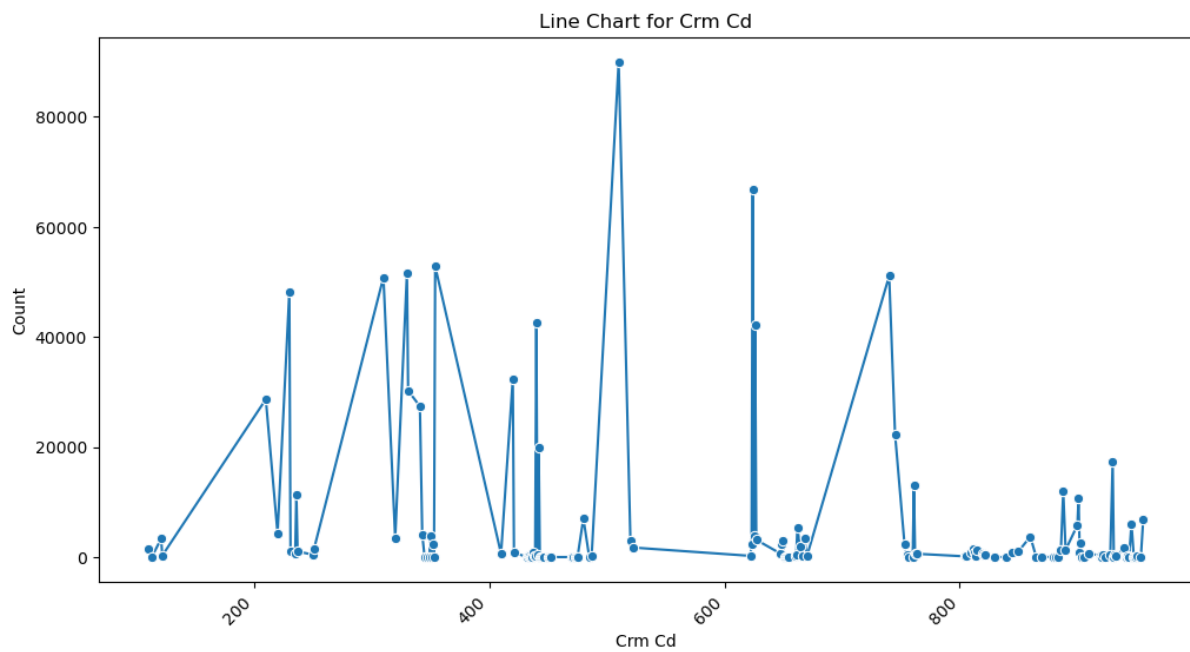
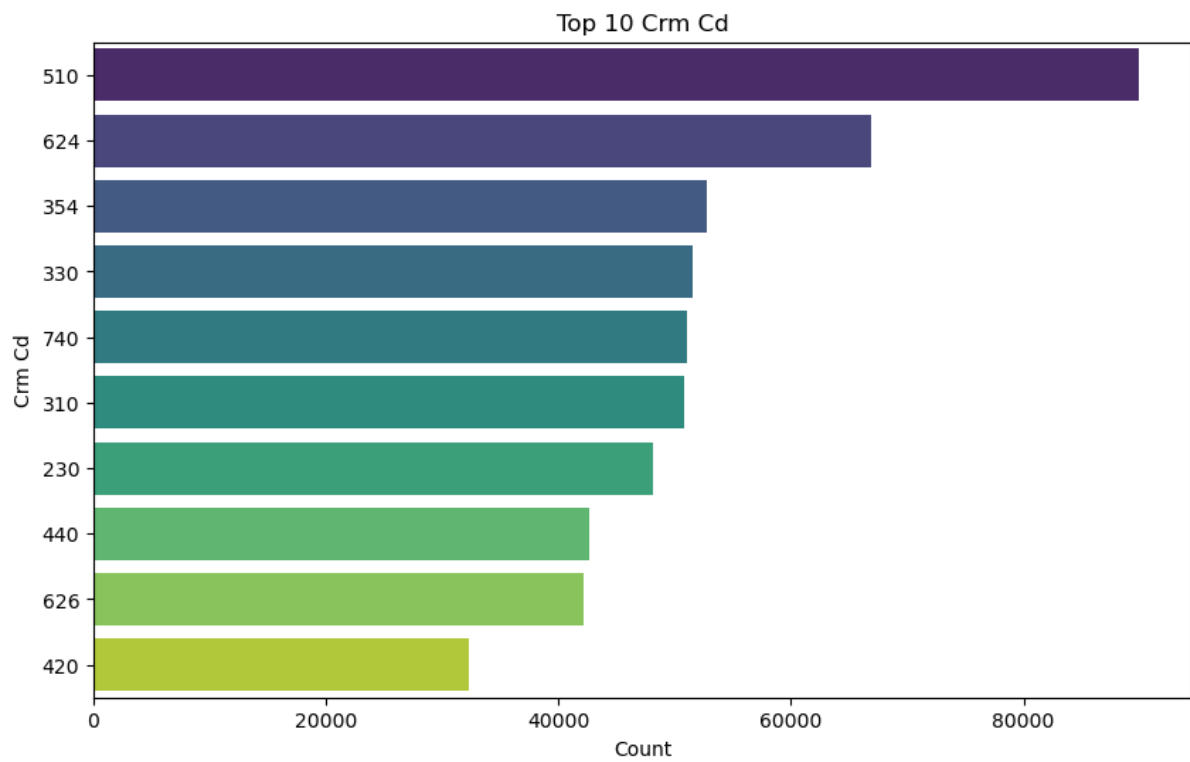
AREA NAME



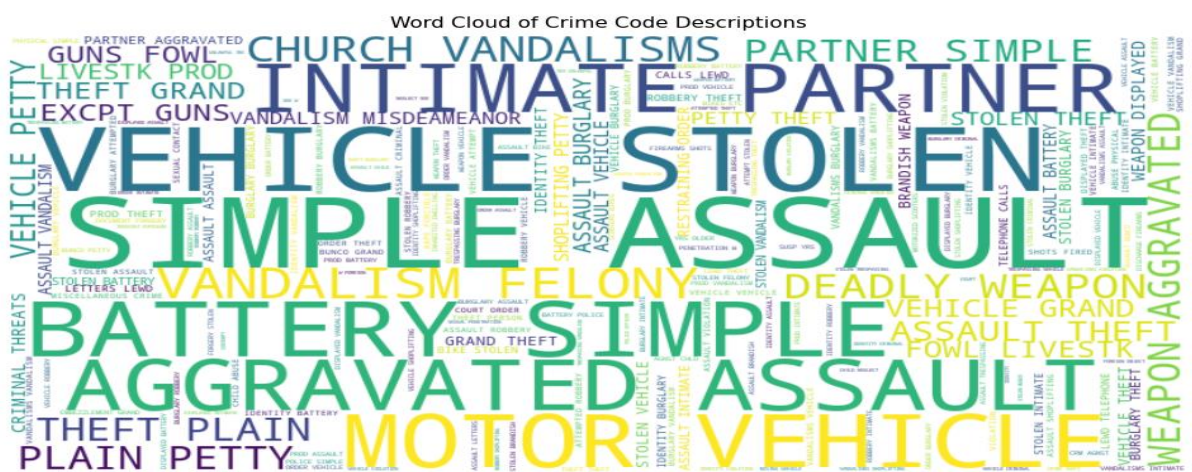
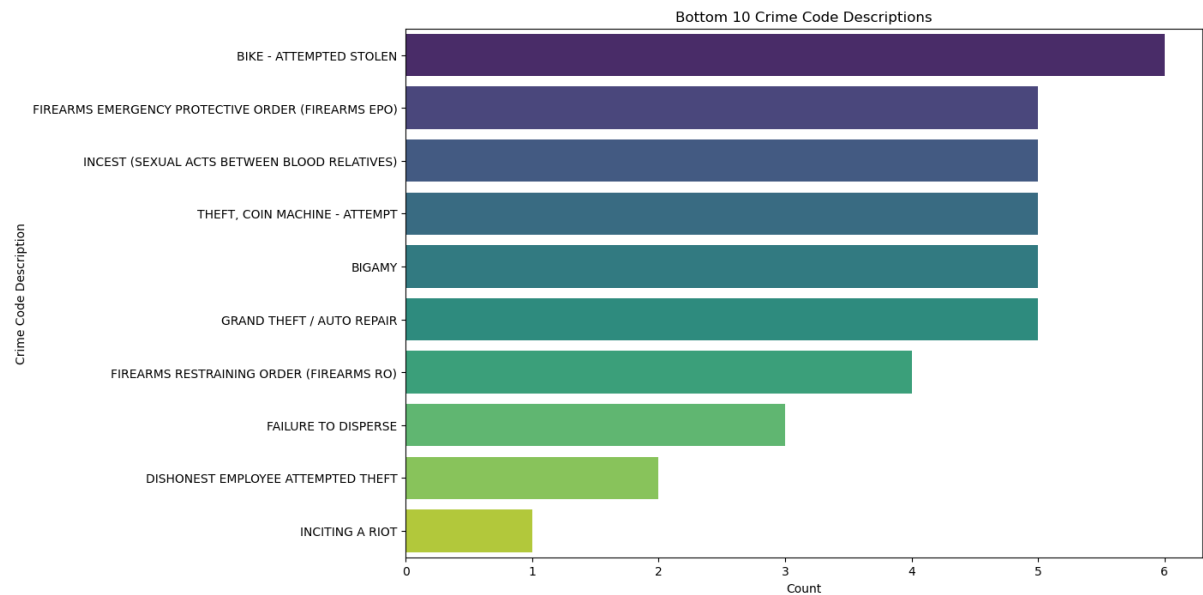
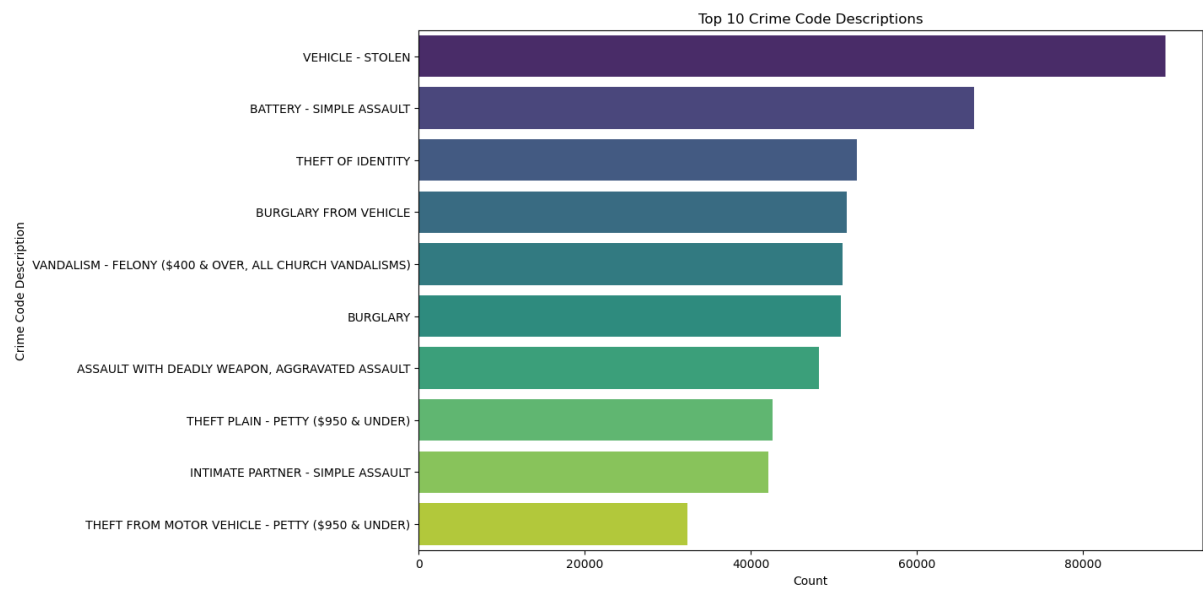
Rpt Dist No



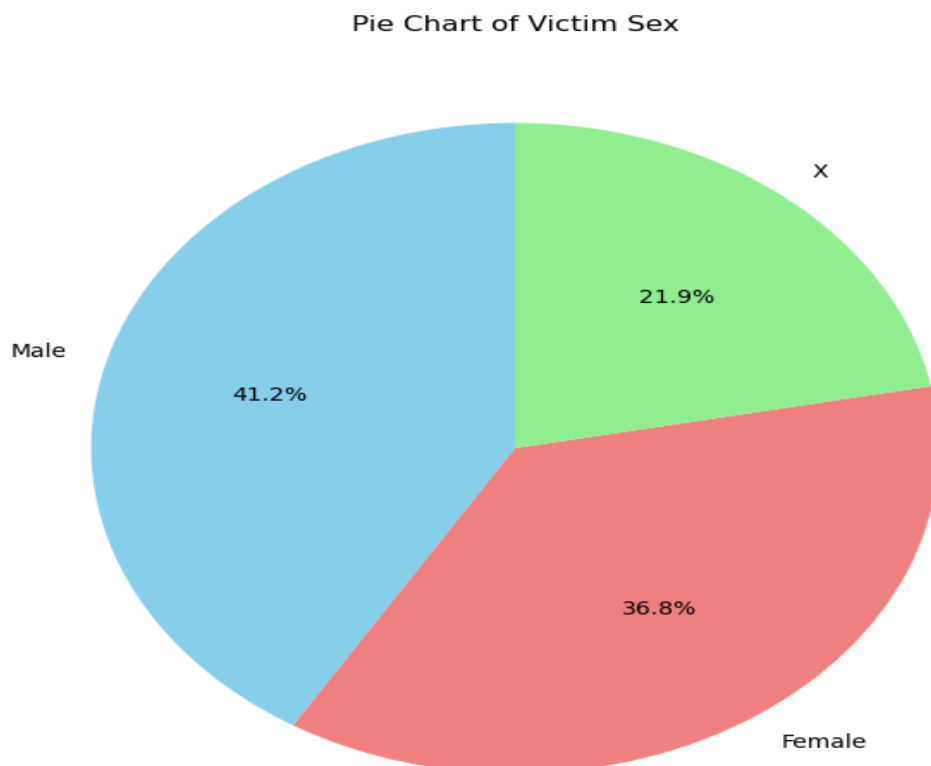
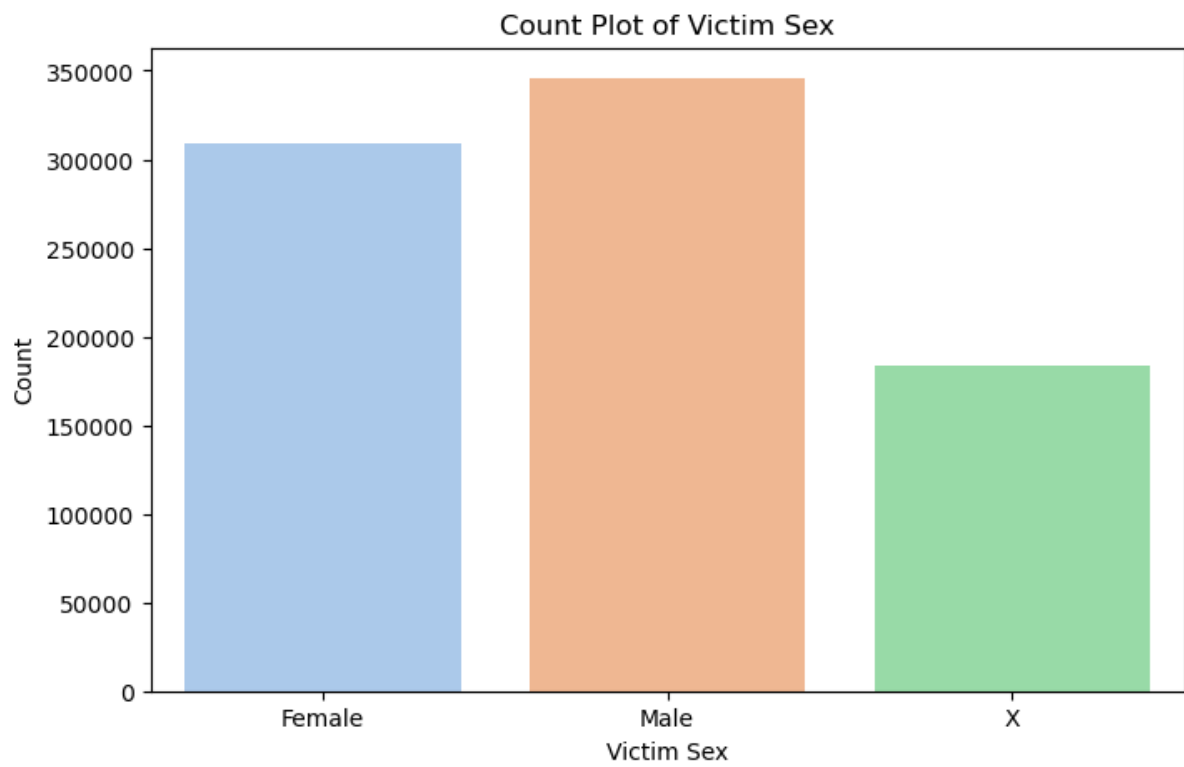
Crm Cd



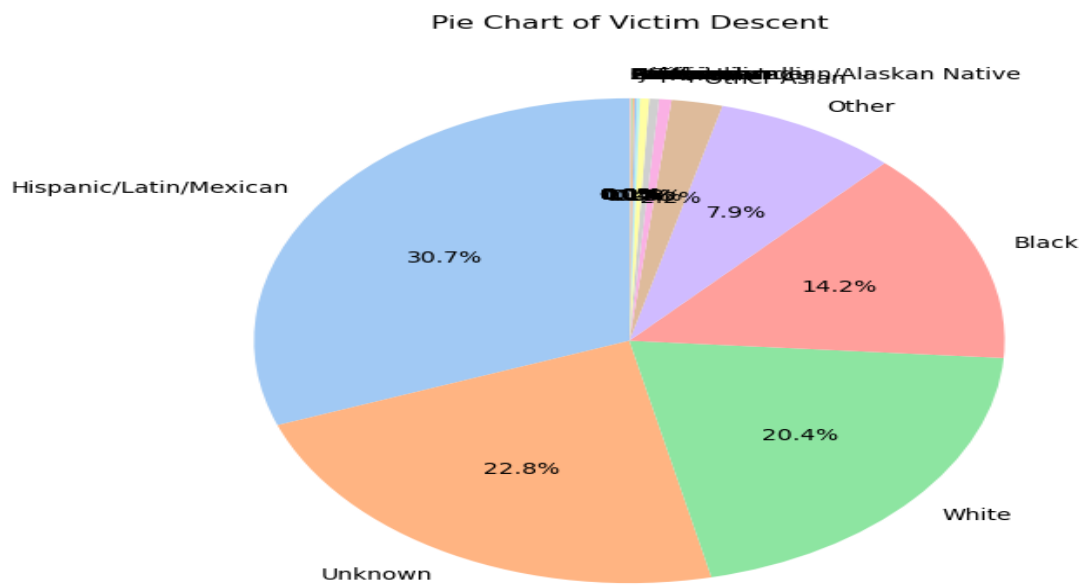
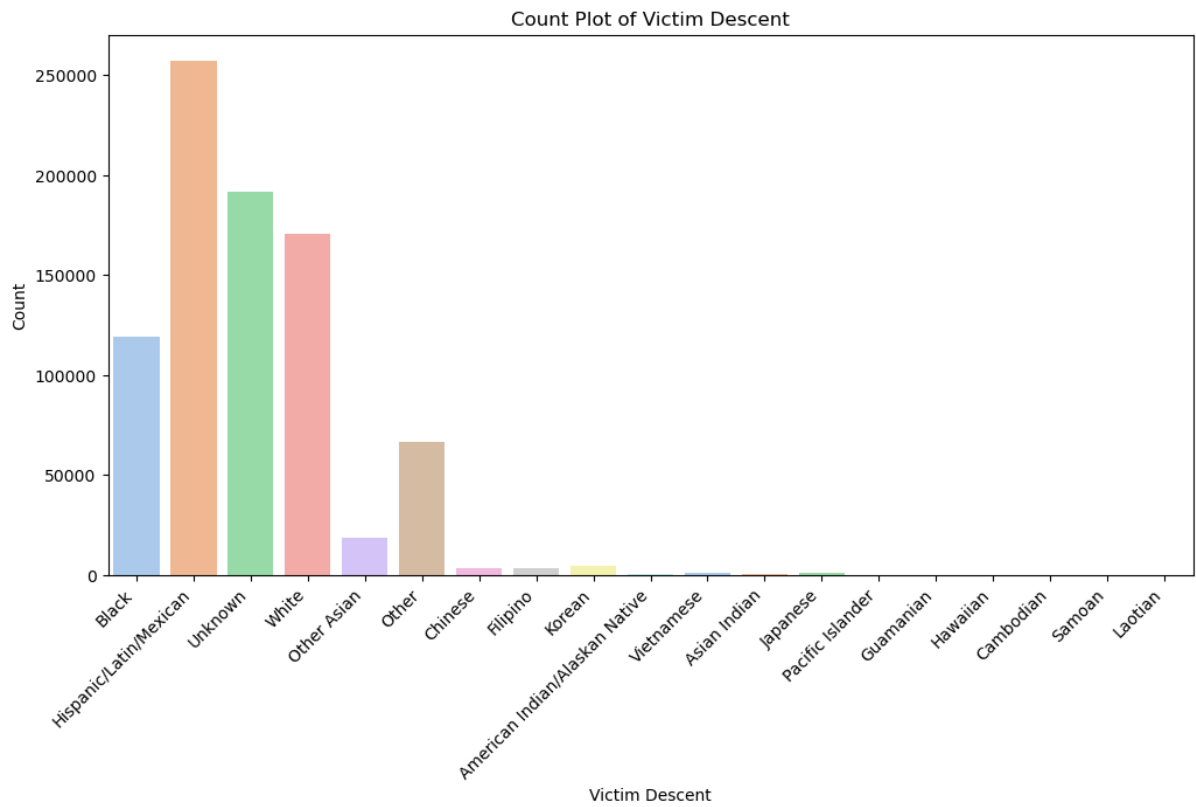
Crm	Cd	Desc
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
10	10	10
11	11	11
12	12	12
13	13	13
14	14	14
15	15	15
16	16	16
17	17	17
18	18	18
19	19	19
20	20	20
21	21	21
22	22	22
23	23	23
24	24	24
25	25	25
26	26	26
27	27	27
28	28	28
29	29	29
30	30	30
31	31	31
32	32	32
33	33	33
34	34	34
35	35	35
36	36	36
37	37	37
38	38	38
39	39	39
40	40	40
41	41	41
42	42	42
43	43	43
44	44	44
45	45	45
46	46	46
47	47	47
48	48	48
49	49	49
50	50	50
51	51	51
52	52	52
53	53	53
54	54	54
55	55	55
56	56	56
57	57	57
58	58	58
59	59	59
60	60	60
61	61	61
62	62	62
63	63	63
64	64	64
65	65	65
66	66	66
67	67	67
68	68	68
69	69	69
70	70	70
71	71	71
72	72	72
73	73	73
74	74	74
75	75	75
76	76	76
77	77	77
78	78	78
79	79	79
80	80	80
81	81	81
82	82	82
83	83	83
84	84	84
85	85	85
86	86	86
87	87	87
88	88	88
89	89	89
90	90	90
91	91	91
92	92	92
93	93	93
94	94	94
95	95	95
96	96	96
97	97	97
98	98	98
99	99	99
100	100	100



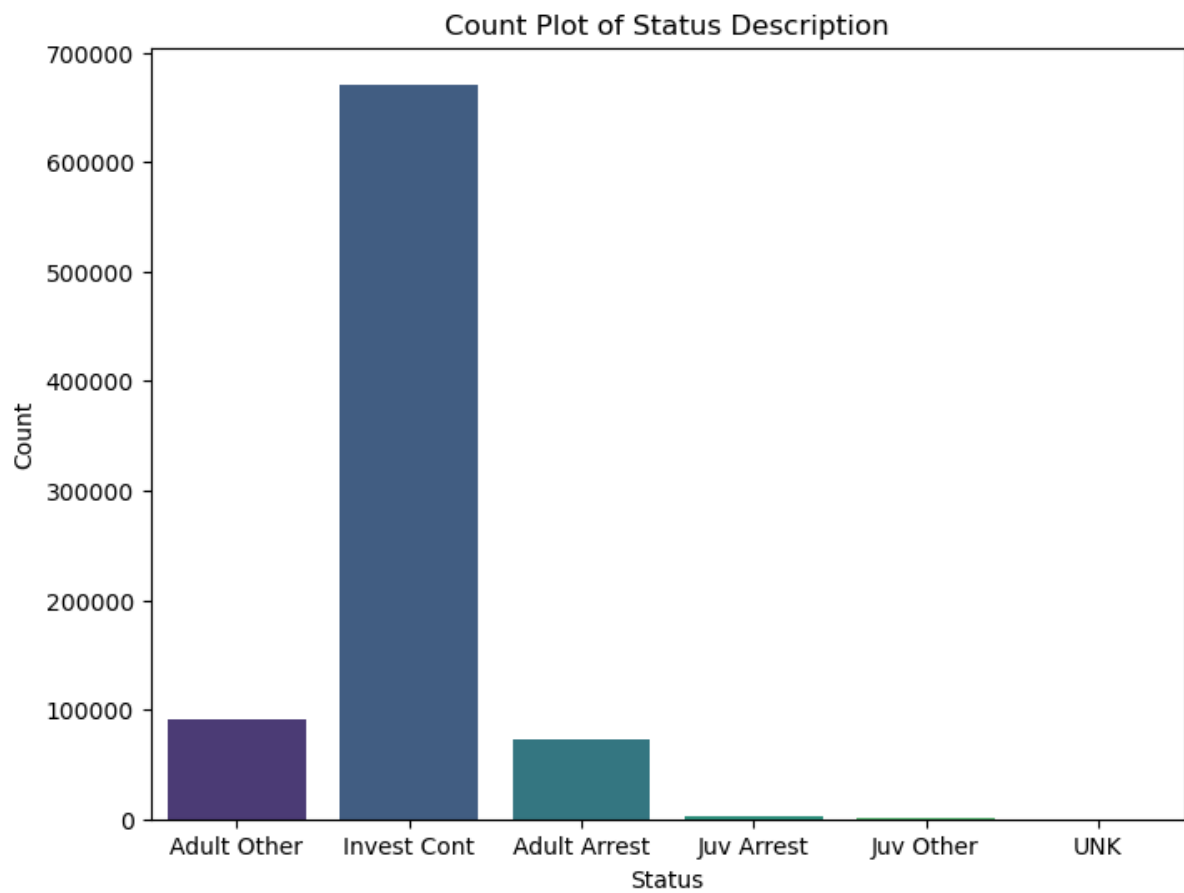
Vict Sex



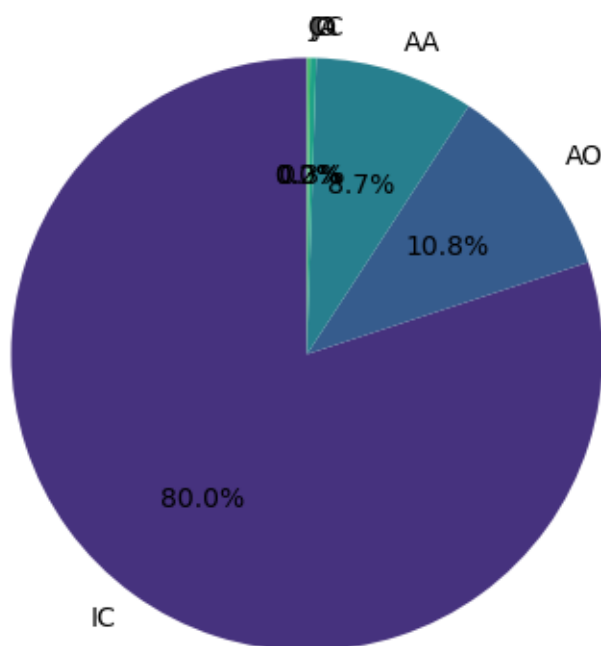
Vict Descent



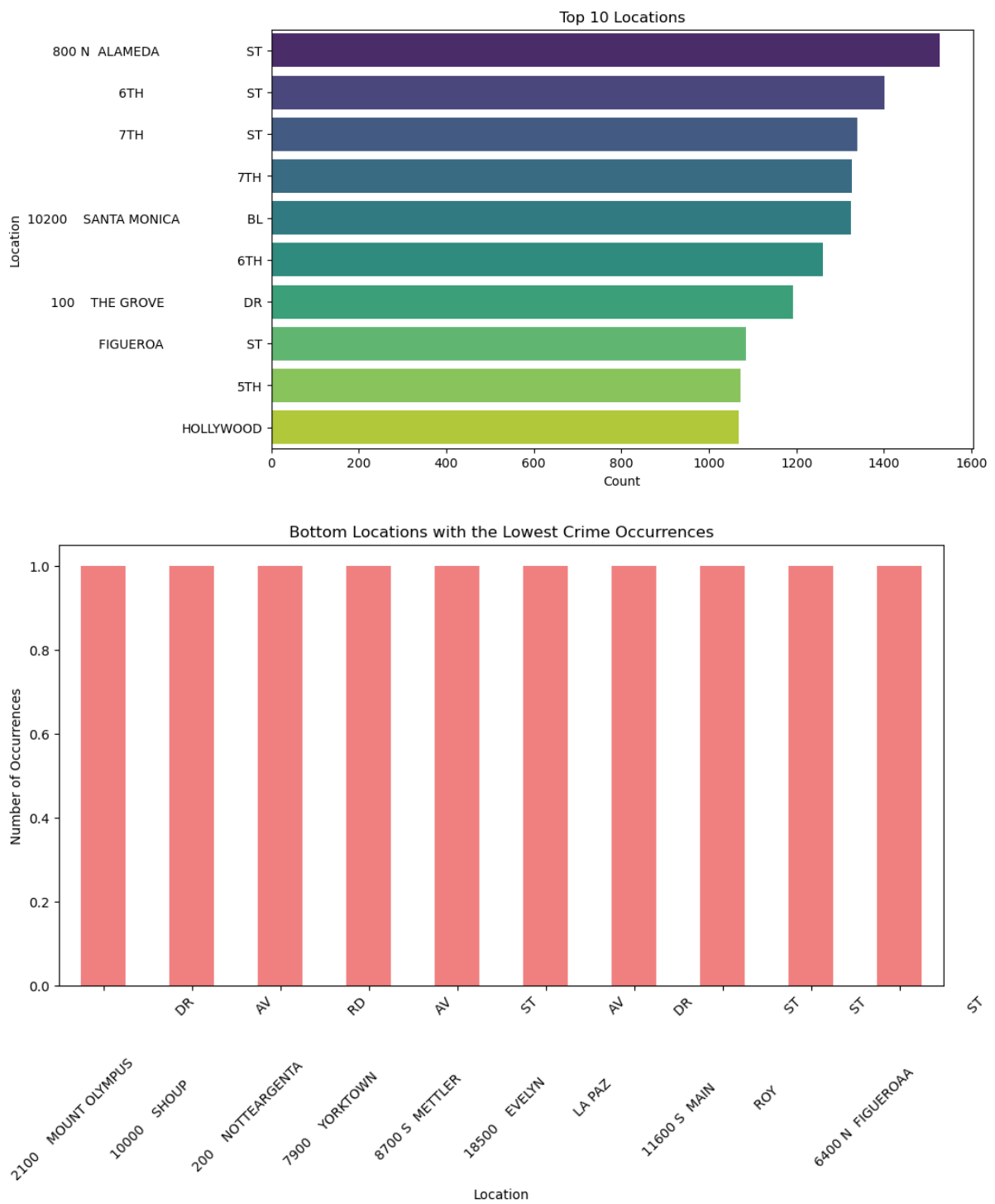
Status / Status Desc



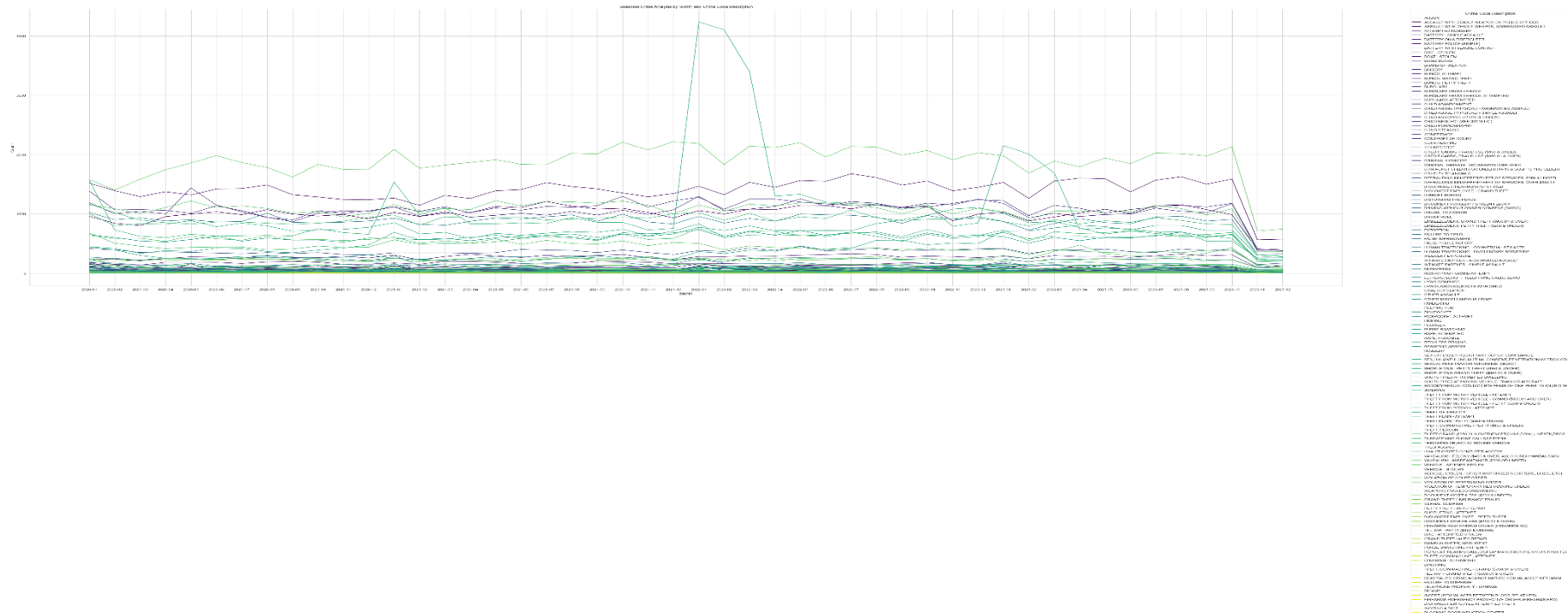
Pie Chart of Status



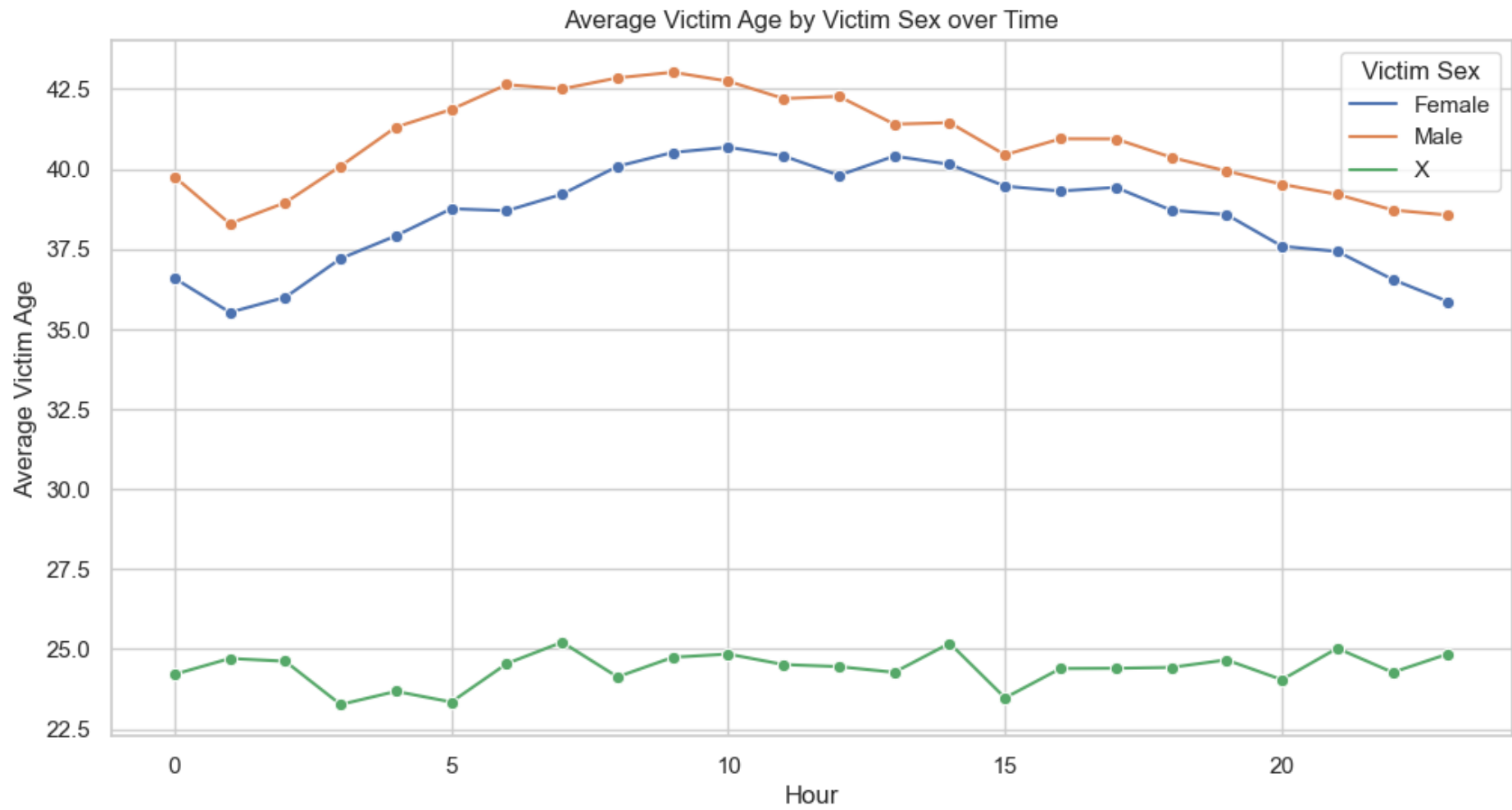
LOCATION



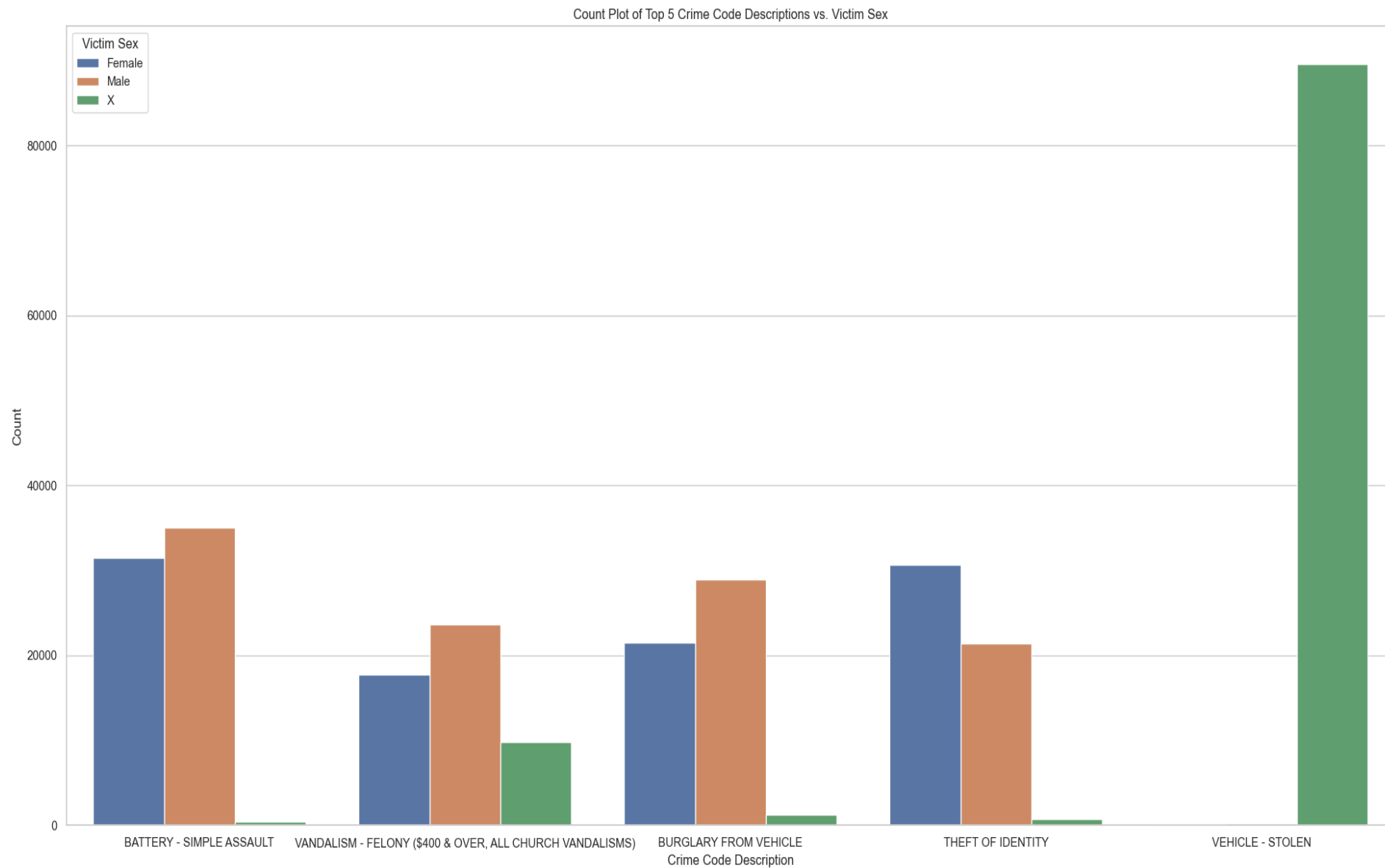
4.2 Bi-variate Analysis Of Project

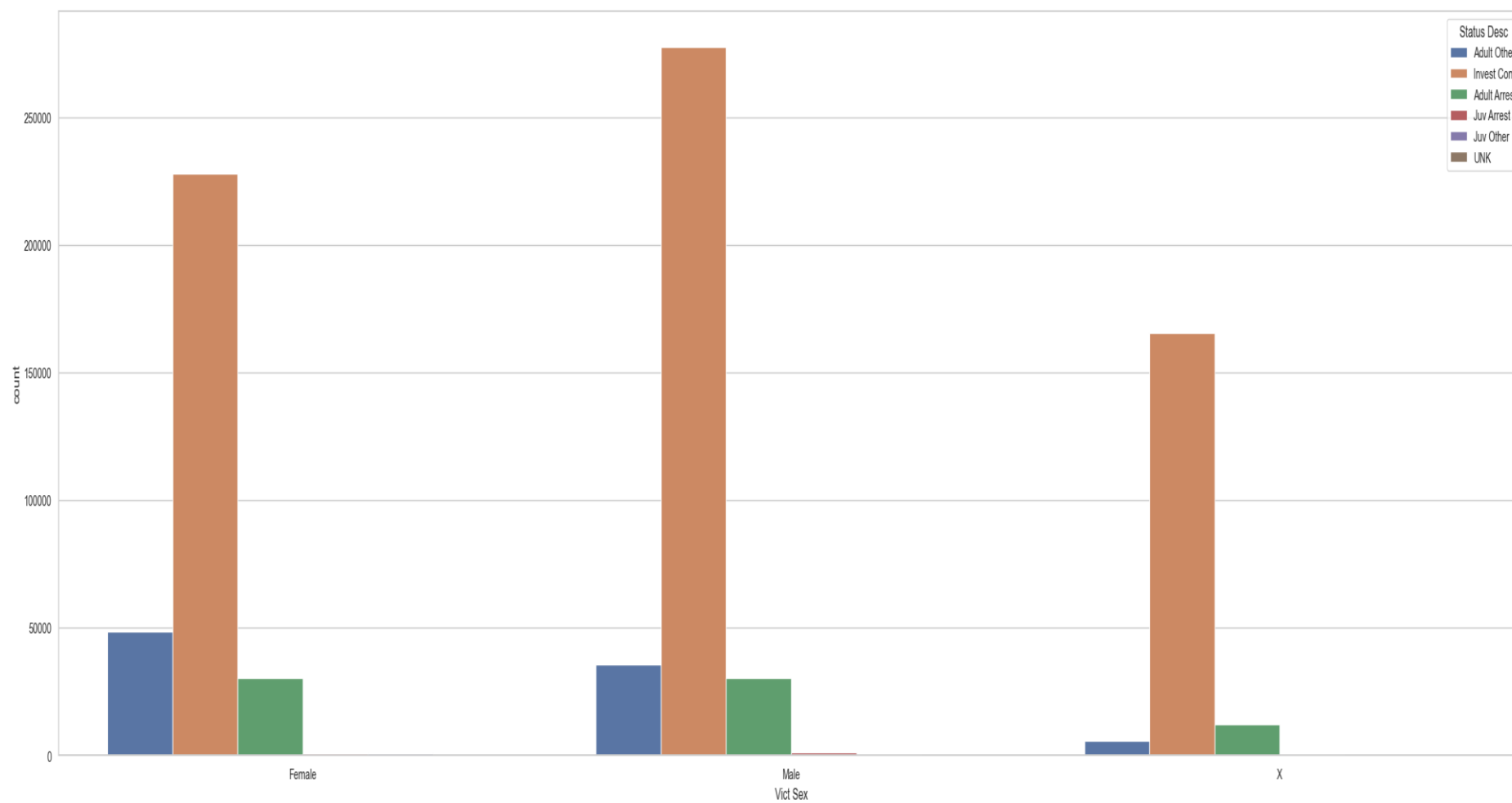


Month Vs Crime Description (showing seasonal analysis)

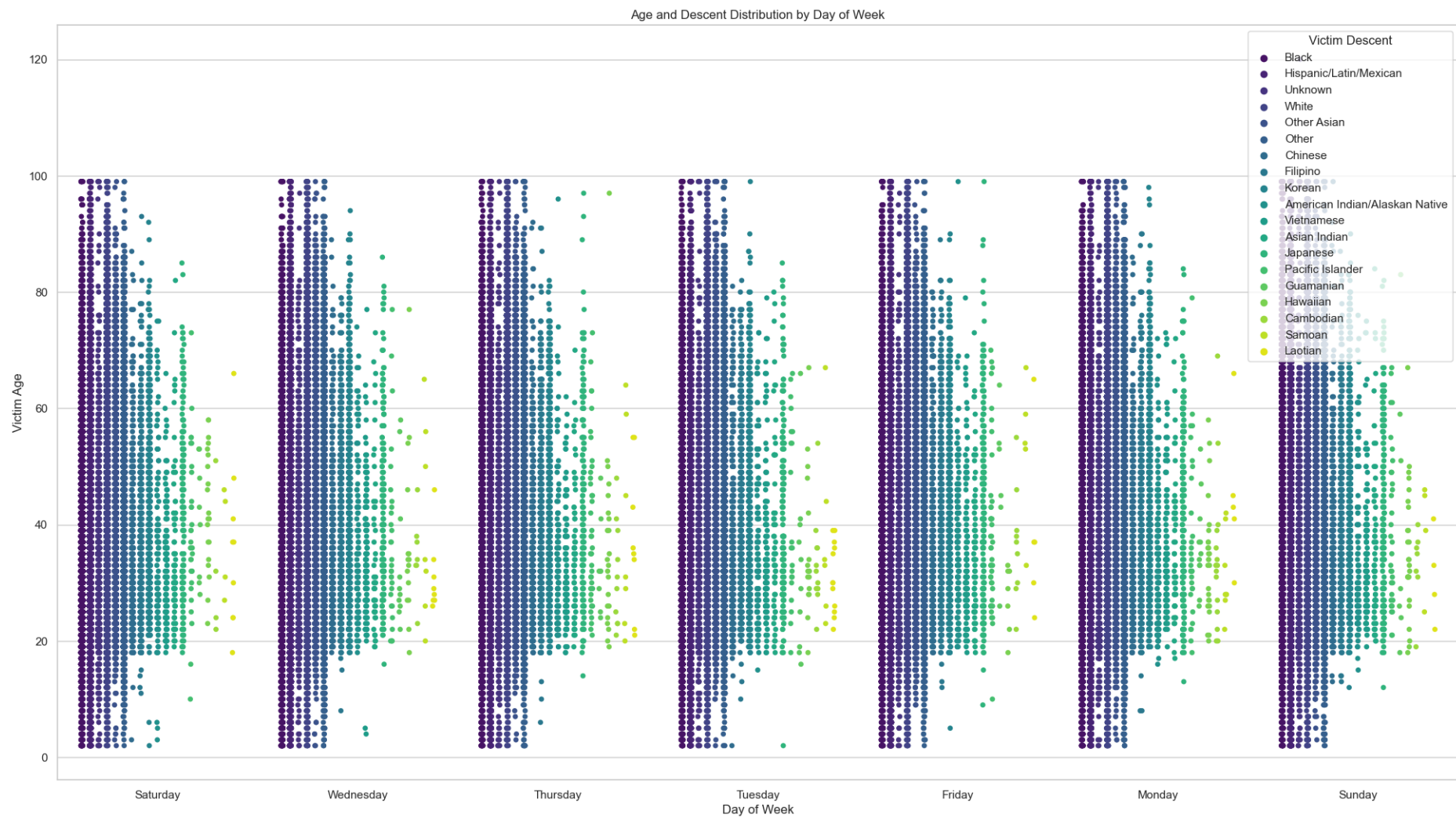


Victim sex vs Victim age





Vict Sex Vs Status Desc



5 Data Pre-Processing

Data pre-processing is a crucial step in the data analysis and machine learning pipeline. It involves cleaning and transforming raw data into a format suitable for analysis or model training. Several common techniques used in data pre-processing include discretization, normalization, one-hot encoding, and train-test splitting.

1. Discretization:

- **Definition:** Discretization is the process of converting continuous data into discrete categories or bins.
- **Purpose:** It is often applied to simplify complex datasets, reduce noise, or prepare data for algorithms that require categorical input.
- **Method:** Discretization methods include binning, where continuous values are grouped into intervals or bins, and decision tree-based methods that split data into homogenous groups.
- **Example:** Age values are discretised into categories like 'kid', 'young,' 'middle-aged,' and 'elderly.'

2. Normalization:

- **Definition:** Normalization (or scaling) is the process of rescaling numerical features to a standard range, typically between 0 and 1.
- **Purpose:** Normalization ensures that different features with different scales contribute equally to the analysis or model training. It can also help in speeding up optimization algorithms.
- **Method:** Common normalization methods include Min-Max scaling and Z-score normalization (standardization).
- **Example:** Since in my Data set there was no more numerical value I treated Lon and Lat as one Z-score both

3. One-Hot Encoding:

- **Definition:** One-hot encoding is a technique used to represent categorical variables as binary vectors.
- **Purpose:** It is necessary when working with machine learning models that require numerical input, as they cannot directly process categorical data.
- **Method:** Each category is represented by a binary vector with all zeros and a single one at the position corresponding to the category.
- **Example:** my dataset was too big and taking approx. 40+ Gigabytes so I took sample data and performed one-hot encoding (`pd.get_dummies(sampled_data)`)

4. Train-Test Split (70-30):

- **Definition:** Train-test splitting involves dividing the dataset into two subsets: one for training the model and the other for evaluating its performance.
- **Purpose:** It helps assess how well a model generalizes to new, unseen data by simulating its performance on an independent dataset.
- **Method:** The dataset is typically split into a training set (e.g., 70% of the data) and a test set (e.g., 30% of the data).
- **Example:** In a machine learning task, 70% of the data is used to train the model, and the remaining 30% is used to evaluate its performance.

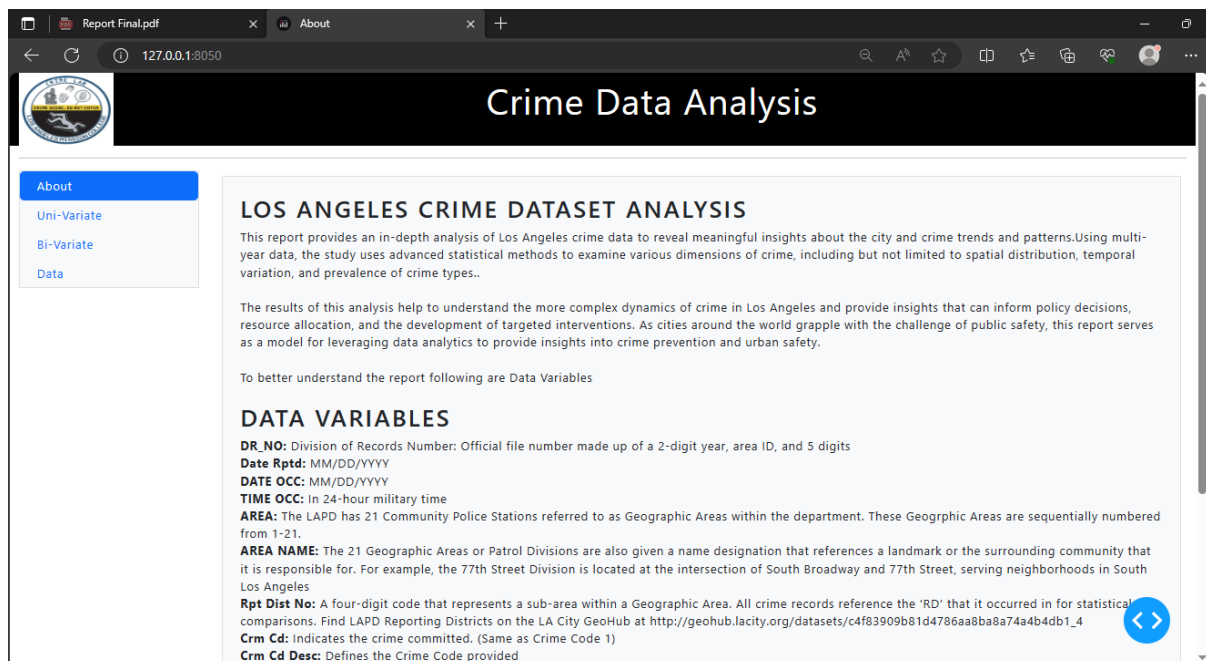
6 Dash Plotly

Dash by Plotly is an open-source Python framework for building analytical web applications. It allows developers to create interactive, web-based data visualizations with ease. The framework is particularly popular for its ability to build interactive dashboards for data analysis, reporting, and visualization.

The core of Dash comprises of three technologies:

- Flask (Provides functionality to the web server)
- React.js
- Plotly.js (used for generating charts and web Server)

Here are the end results of our dash




Main Page

Report Final.pdf

Data

127.0.0.1:8050/data

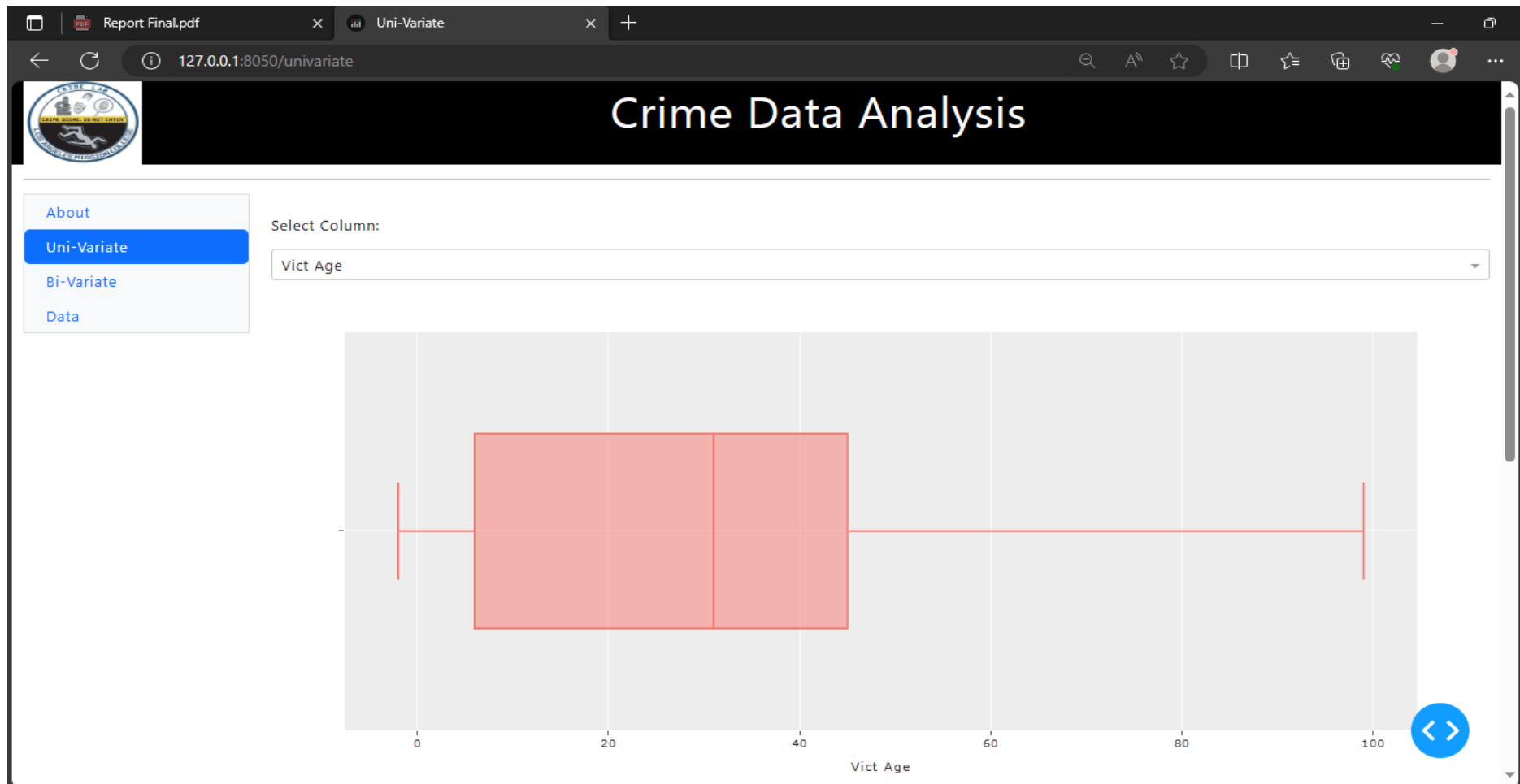


Crime Data Analysis

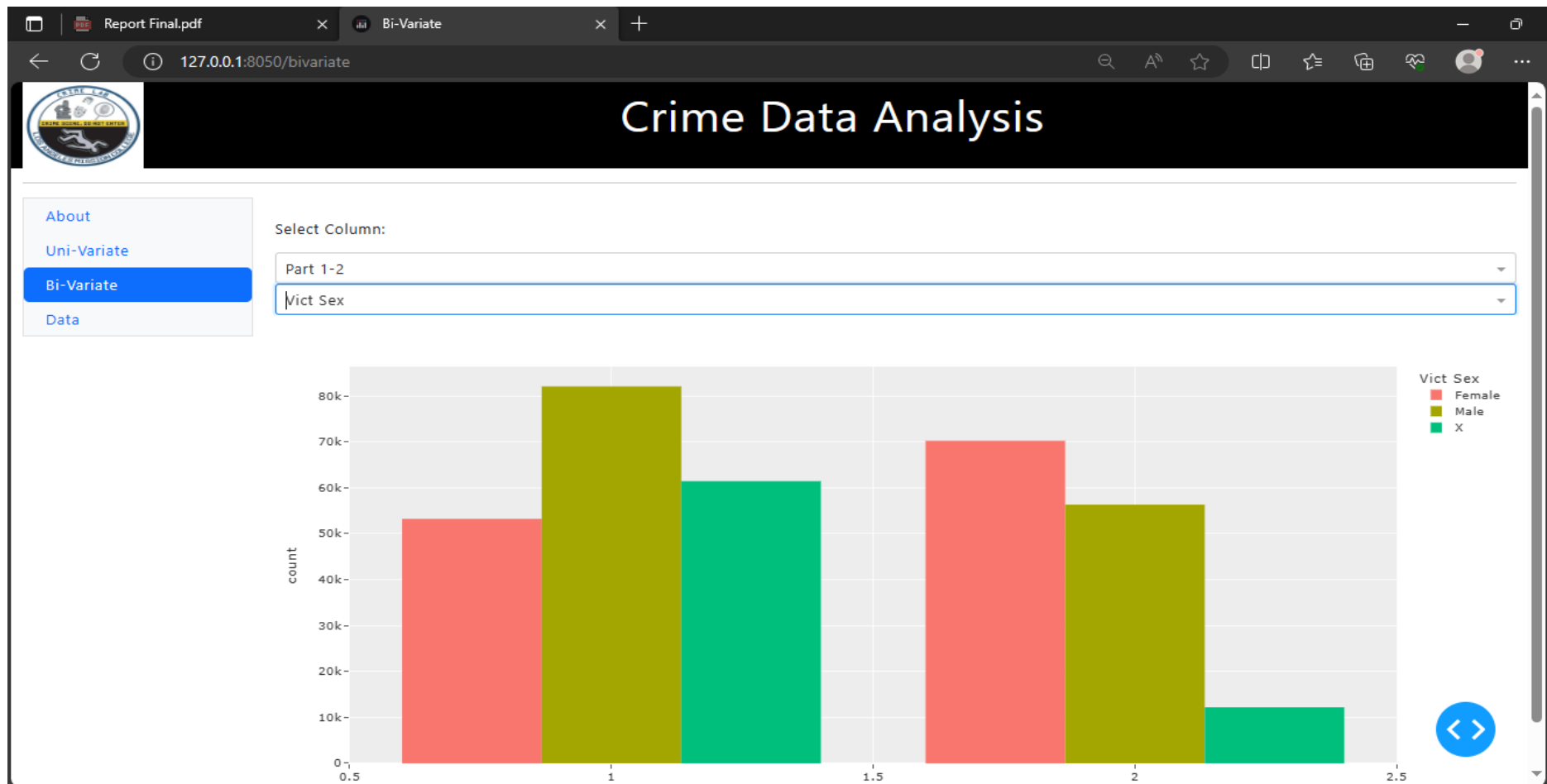
[About](#)
[Uni-Variate](#)
[Bi-Variate](#)
[Data](#)

DR_NO	DATE RPTD	DATE OCC	TIME OCC	AREA	AREA NAME	RPT DIST NO	PART 1-2	CRM CD	CRM CD D
211811565	2021-06-05	2021-05-06	1900-01-01 19:30:00	18	Southeast	1837	2	626	INTIMATE PARTN
211317689	2021-10-19	2021-10-19	1900-01-01 15:30:00	13	Newton	1313	2	850	INDECENT EXPOS
200708342	2020-04-16	2020-04-16	1900-01-01 11:08:00	7	Wilshire	743	1	440	THEFT PLAIN -
200509383	2020-05-12	2020-12-05	1900-01-01 14:50:00	5	Harbor	513	2	888	TRESPASSING
220608413	2022-03-31	2022-03-27	1900-01-01 22:00:00	6	Hollywood	637	1	330	BURGLARY FROM
201407247	2020-02-25	2020-02-24	1900-01-01 19:00:00	14	Pacific	1489	1	330	BURGLARY FROM
221017314	2022-12-05	2022-05-12	1900-01-01 19:30:00	10	West Valley	1099	1	320	BURGLARY, ATTE
202107803	2020-03-28	2020-03-27	1900-01-01 20:00:00	21	Topanga	2158	1	331	THEFT FROM MOT
210514460	2021-10-03	2021-02-10	1900-01-01 21:00:00	5	Harbor	529	1	310	BURGLARY
210118355	2021-10-09	2021-08-10	1900-01-01 10:45:00	1	Central	182	1	440	THEFT PLAIN -
200915136	2020-09-26	2020-09-24	1900-01-01 21:20:00	9	Van Nuys	985	1	331	THEFT FROM MOT
211213780	2021-06-11	2021-11-06	1900-01-01 17:00:00	12	77th Street	1266	2	930	CRIMINAL THREA
201218191	2020-08-09	2020-08-08	1900-01-01 02:30:00	12	77th Street	1267	1	510	VEHICLE - STOL
211315678	2021-09-11	2021-10-09	1900-01-01 17:35:00	13	Newton	1393	2	930	CRIMINAL THREA
220206522	2022-03-01	2022-02-28	1900-01-01 19:00:00	2	Rampart	256	1	510	VEHICLE - STOL

Data Display



Uni-Variate Analysis



Bi- Variate Analysis

THE END