Final Project Machine Learning

24F CST8502

Predicting Crime Categories and Patterns in Dallas:

A Machine Learning Approach to Enhance Law Enforcement Resource Allocation and Prevention Strategies

Table of Content

1 Introduction

This assignment is based on the Police Incidents dataset (Police Incidents | Dallas OpenData) published by the City of Dallas, USA. The assignment will go through CRISP-DM methodology and all the steps will be listed in this report.

2 Business Understanding

2.1 Determine Business Objectives

The Dallas police department is dedicated to serving the people of Dallas and strives to reduce crime and provide a safe city and their missions emphasize the department's commitment to public safety and crime reduction as its primary objectives, while highlighting the importance of serving the community in achieving these goals.

Dallas Police Department has 7 Patrol Divisions: CENTRAL, NORTHCENTRAL, NORTHEAST, NORTHWEST, SOUTHCENTRAL, SOUTHEAST, SOUTHWEST, JACKEVANSHEADQUARTERS.

2.2 Assess Situation

In this part, we determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.

1. Resource Availability Analysis
    - Data Resources: Dallas Police Incidents dataset from Dallas OpenData with free, publicly available data, regular updates available or historical data accessible. And it contains many fields (i.e., incident details, locations, dates, offense types, etc.) which can be used for many data analysis questions.
    - Human Resources (6 team members) can divide into 2 Data Preparation Specialists, 2 Classification Modeling Analysts and 2 Clustering/Outlier Detection Analysts
    - Technical Resources: Python programming environment required libraries: pandas, numpy, matplotlib, scikit-learn, seaborn, scipy and pyproj.
2. Project Requirements
    - Technical Requirements: Knowledge of machine learning algorithms (Classification, Clustering and Outlier Detection) and Python programming skills.
    - Dataset Requirements: minimum 10 relevant attributes are required, having clean, preprocessed data with adequate sample size (10,000 instances if needed)
    - Timeline Requirements:
        - Part 1 (Business Understanding, Data Understanding & Preparation) due November 15
        - Part 2 (Modeling & Evaluation) due November 22
        - Presentations: November 25 – December 6

3. Risk Assessment
    - At the highest priority level, data quality presents a significant concern, as the accuracy and completeness of our analysis heavily depend on the quality of the input data. To address this, we will implement thorough data cleaning and validation procedures, and if necessary, create features derived from existing data points to compensate for any deficiencies in the primary data. Technical challenges also pose a high-priority risk, which we plan to address through regular team meetings focused on technical problem-solving, while ensuring team members are cross-trained to provide support across different areas when needed.
    - Moving to medium-priority risks, timeline constraints represent a significant challenge given the project's fixed deadline. Our mitigation strategy includes establishing clear project milestones and implementing regular progress tracking mechanisms. If time becomes particularly tight, we will prioritize essential features and analyse to ensure core project objectives are met. Model performance is another medium-priority risk, as the effectiveness of our analysis depends on the quality of our predictive models. We plan to iterate through different modeling approaches to optimize performance, while maintaining a focus on interpretability even if we need to trade off some degree of accuracy.
    - At the lower priority level, team coordination has been identified as a potential risk, though one that can be effectively managed through established communication channels and clear task assignments. Our contingency plan includes maintaining flexibility in task allocation to ensure work continues smoothly even if adjustments need to be made to the original assignments. This comprehensive risk management approach will help ensure project success while maintaining the ability to adapt to challenges as they arise.
4. Costs:
    - Time Investment: 6 team members × approximately 20 hours each. Regular meetings and coordination time.
    - Technical Resources: Computing resources (using personal computers) and Software (using free, open-source tools/libraries)
    - Training/Learning Curve: Time spent learning new techniques and documentation reading and research.
5. Benefits
    - The project offers a comprehensive range of benefits across different timeframes and impact levels. In terms of direct benefits, our analysis will provide crucial insights into violent crime patterns, allowing us to identify key contributing factors that influence criminal activity. Through the development of predictive models for crime classification and the detection of unusual crime patterns, we will create valuable tools for understanding and responding to criminal behavior in Dallas.

- The project also yields significant indirect benefits for our team members. Through hands-on experience with a real-world dataset, team members will develop and enhance their data science skills while building their professional portfolios. The project's collaborative nature provides valuable team experience, preparing participants for future professional endeavors in data science and analytics.
- Looking at long-term benefits, this project can have a lasting impact on law enforcement operations. By providing insights that could improve police resource allocation and deepening our understanding of crime patterns, we lay the groundwork for more effective crime prevention strategies. These findings could serve as a foundation for future initiatives aimed at enhancing public safety and law enforcement effectiveness in Dallas.

2.3 Determine Data Mining Goals

The goal of this project is to analyze the Dallas Police Incidents dataset to identify key factors that contribute to violent crime incidents. By understanding the characteristics and patterns of violent crimes, we can help the Dallas Police Department develop more targeted strategies and interventions to reduce such incidents.

2.4 Produce Project Plan

Select technologies and tools and define detailed plans for each project phase.

- Week 1-2 (Business Understanding, Data Understanding & Data Preparation): Dataset exploration, and cleaning. Understand the features and set up the initial analysis.
- Week 3-4 (Modeling): Implementation of classification models, Clustering analysis, Outlier detection and evaluate the models.
- Week 5-6 (Evaluation & Presentation): Results analysis, Report writing, Presentation preparation and Final delivery

Item	Sub-task
Business Understanding	
Data Understanding	Using data from the year 2022. Explored 14 attributes using visualizations, checked for data distribution, verified data quality. At the end of the analysis, only 2 attributes ‘Type Location’ and ‘Division’ deemed useful.

	Explored and filtered attributes from 15 to 28 and 85 – 86. Date and Time attributes were selected
	From attribute 43 to attribute 56 of the dataset
	Explored and filtered 14 attributes from columns 57 to 70 of the dataset for their data quality and purpose in answering the question. Only 1 attribute is used after filtering: 'UCR Disposition'
	Explored 14 attributes from columns 71 – 84, observing quality of data and its distribution. Only 4 attributes relevant in answering the business question were kept – NIBRS Group, Zip Code, X Coordinate and Y Coordinate
Data Preparation - General (Write the name of the attribute and who prepared it - include those attributes that need preparation). If you can consider an attribute as-is, there is no need to include that in this table	Dropping rows in 'Division' and 'Type Location' since only 10% of the data is missing. 'Type Location' has 71 unique values, which is mapped to 9 general values and a new attribute 'Crime Scene' is created. Remove Nan

	values.
	From the 4 columns selected, Zip Codes were binned based on regional zones, missing values in all 4 attributes were also handled accordingly
	Preparation rules for: Victim Race Victim Gender
	Data Preparation Code integration along with respective attributes
Task 1 - Classification	kNN
	Random Forest
	Decision Tree
Task 2 - Clustering	kMeans & Elbow charts and Clusters
Data Prep - Outlier Detection	Approach 1: **Keep the categorical features** Removed irrelevant attributes (e.g., X and Y coordinates). Binning Dates in quarter and time of the day Encoded categorical

	variables using one-hot encoding.
	Approach 2: **Anomaly detection for Time Series** Construct new column: `Total crime case` that calculate the total number of incidents in a day The `Datetime of Occurrence` is format to pandas date format in order to transformed into Date as Number later on. Removed all other features. Using `StandardScaler` for the dataset before applying on LOF as LOF is sensitive to the scale of features
Task 3 - Outlier Detection	Approach 1: **Keep the categorical features**
	LOF, ISF and common Outliers
	Approach 2: **Anomaly detection for Time Series**
	LOF, ISF and common

	Outliers
Data Prep - Classification	Month, DayOfWeek, Hour, TimeOfDay was created
Data Prep - Clustering	List prepared columns
Merge processes (Give details)	Google Colab for code Word document for report Powerpoint for presentation
Interpretation of results (Give details)	Interpreation results in Outliers, compare results, plot results, write reports.
Interpretation of results – Classification	Accuracy, Classification report and confusion matrix compared results with the other Classification Models

	Feature Importance analysis Decision Tree Rule Generation
Data Science Question	Finding insights, patterns and conclusion

3 Data Understanding

3.1 Collect Initial Data

Dallas Police Incidents dataset from Dallas OpenData with free, publicly available data.

3.2 Describe Data

	Attribute Name	Description
1	Incident Number w/year	An RMS generated incident number (report number) with the year
2	Year of Incident	Year associated with the incident number
3	Service Number ID	Incident number plus year code plus offense number (Ex: -02 means there is two offenses with this one incident) Internal use
4	Watch	Police watch 1st 2nd or 3rd (1st watch = Late Night, 2nd watch = Days and 3rd watch = Evenings)
5	Call (911) Problem	Police call signal generated by Communications (Type of 911 call dispatched)
6	Type of Incident	Type of Incident
7	Type Location	Location type where incident took place for example, Apartment Parking, Residence
8	Type of Property	The target item… Parkinglot, Motor Vehicle
9	Incident Address	Address where incident occurred
10	Apartment Number	Apartment number

11	Reporting Area	Geographic area comprised of reporting areas where incident occurred
12	Beat	Geographic area comprised of beats where incident occurred
13	Division	Geographic area comprised of census blocks where incident occurred (smallest police geography)
14	Sector	Geographic area comprised of Sectors where incident occurred
15	Council District	Geographic area comprised of city council districts where incident occurred
16	Target Area Action Grids	Geographic areas targeted for higher-than-average crime
17	Community	Community Prosecution Areas as designated by the City Community Prosecutors
18	Date1 of Occurrence	The first date of the date occurrence of the incident (Ex: incident occurred between 1/1/2016 and 1/2/2016)
19	Year1 of Occurrence	Year of the indent based on the Date of Occurrence (Date1). Internal use
20	Month1 of Occurence	Month (starting) of the indent based on the Date of Occurrence (Date1). Internal use
21	Day1 of the Week	Day of the indent based on the Date of Occurrence (Date1). Internal use
22	Time1 of Occurrence	The first (starting) time of the time occurrence of the incident (Ex: incident occurred between 8:00am and 5:00pm)
23	Day1 of the Year	The calendar number of the year 1-365 is based on Date1. Internal use
24	Date2 of Occurrence	The second date of the date occurrence of the incident (Ex: incident occurred between 1/1/2016 and 1/2/2016)
25	Year2 of Occurrence	Year of the indent based on the Date of Occurrence (Date2)

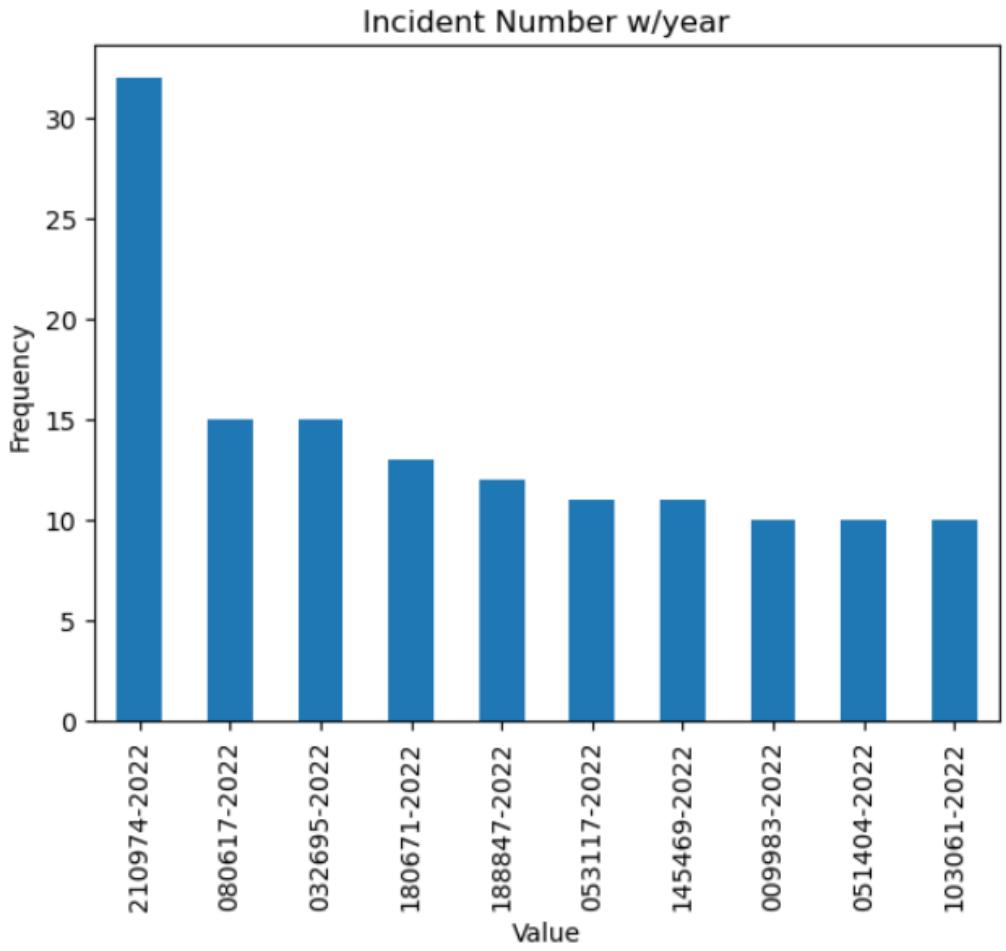
26	Month2 of Occurrence	Month (end) of the indent based on the Date of Occurrence (Date2)
27	Day2 of the Week	Day of the indent based on the Date of Occurrence (Date1)
28	Time2 of Occurrence	The second(end) time of the time occurrence of the incident (Ex: incident occurred between 8:00am and 5:00pm)
29	Day2 of the Year	The calendar number of the year 1-365 based on Date2. Internal use
30	Date of Report	The date of the incident as reported to the police
31	Date incident created	The date the incident record was created. Internal use
32	Offense Entered Year	The year the offense entered the system. Internal use
33	Offense Entered Month	The month the offense entered the system. Internal use
34	Offense Entered Day of the Week	The day the offense entered the system. Internal use
35	Offense Entered Time	The time the offense was entered into the system. Internal use
36	Offense Entered Date/Time	The calendar number of the year the offense was entered. Internal use
37	CFS Number	CFS Number
38	Call Received Date Time	The date the related call was received
39	Call Date Time	Date and time of the related call
40	Call Cleared Date Time	Date and time related call was cleared
41	Call Dispatch Date Time	Date and time related call was dispatched
42	Special Report (Pre-RMS)	No longer applies. PreRMS
43	Person Involvement Type	Person can be; victim, reporting person, witness
44	Victim Type	Victim Type
45	Victim Race	Victim Race

46	Victim Ethnicity	Victim Ethnicity
47	Victim Gender	Victim Gender
48	Responding Officer #1 Badge No	Responding Officer #1 Badge No
49	Responding Officer #1 Name	Responding Officer #1 Name
50	Responding Officer #2 Badge No	Responding officer #2 Badge number
51	Responding Officer #2 Name	Responding officer #2 Name
52	Reporting Officer Badge No	Reporting Officer Badge No
53	Assisting Officer Badge No	Assisting Officer Badge No
54	Reviewing Officer Badge No	Reviewing Officer Badge No
55	Element Number Assigned	Reporting officers assigned element number
56	Investigating Unit 1	1st Assigned investigative unit
57	Investigating Unit 2	Investigating Unit 2
58	Offense Status	Status of the offense
59	UCR Disposition	UCR Disposition of the incident
60	Modus Operandi (MO)	Short description of the offense
61	Family Offense	Yes or no if the offense is family violence
62	Hate Crime	Yes or no if offense is a hate crime
63	Hate Crime Description	Hate Crime Description
64	Weapon Used	Weapon Used
65	Gang Related Offense	Yes or no if offense is gang related
66	Drug Related Intervenient	Yes or no if incident is drug related
67	RMS Code	UCR Offense code
68	Criminal Justice Information Service Code	Criminal Justice Information Services Code (CJIS)
69	Penal Code	State Penal Violation Code number
70	UCR Offense Name	UCR Offense Name
71	UCR Offense Description	UCR Offense description. Internal use

72	UCR Code	UCR Code
73	Offense Type	Offense category Part1 or Part2 or Not coded. Internal use
74	NIBRS Crime	NIBRS Crime
75	NIBRS Crime Category	NIBRS Crime Category
76	NIBRS Crime Against	NIBRS Crime Against
77	NIBRS Code	NIBRS Code
78	NIBRS Group	NIBRS Group
79	NIBRS Type	NIBRS Type
80	Update Date	Date incident was last updated. Internal use
81	X Coordinate	X Coordinate
82	Y Coordinate	Y Coordinate
83	Zip Code	Zip code in which incident occurred
84	City	City in which incident occurred
85	State	State in which incident occurred
86	Location1	Location1

3.3 Explore Data

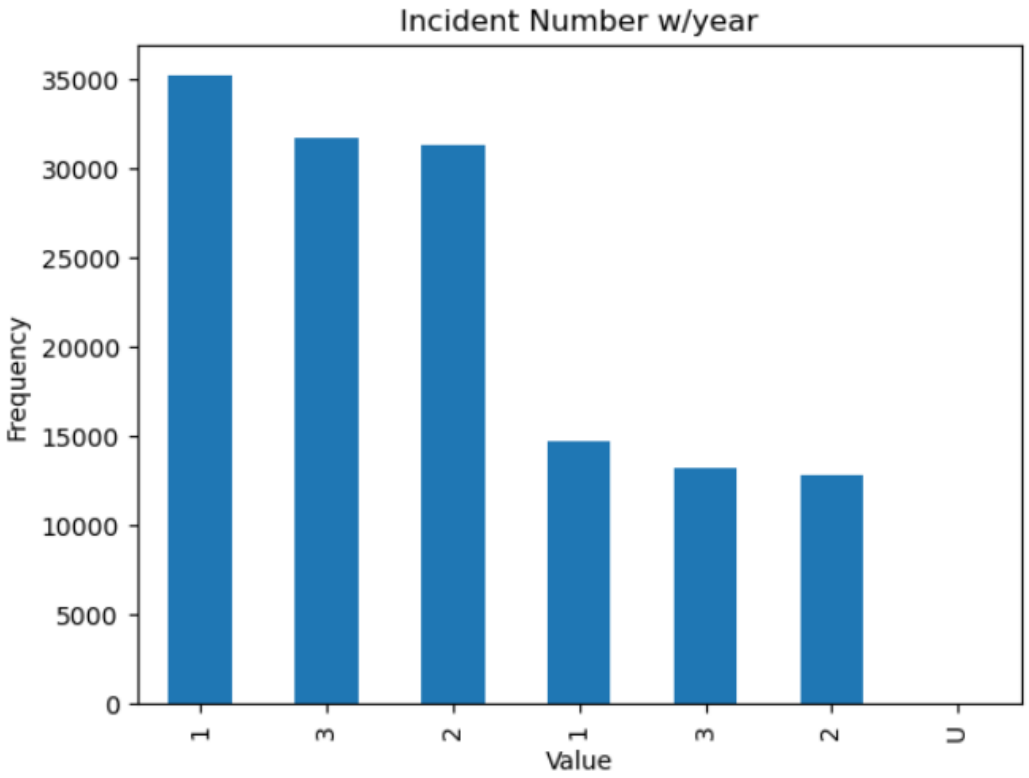
1 Attribute: ‘Incident Number w/year’ - It is a unique ID appended with year



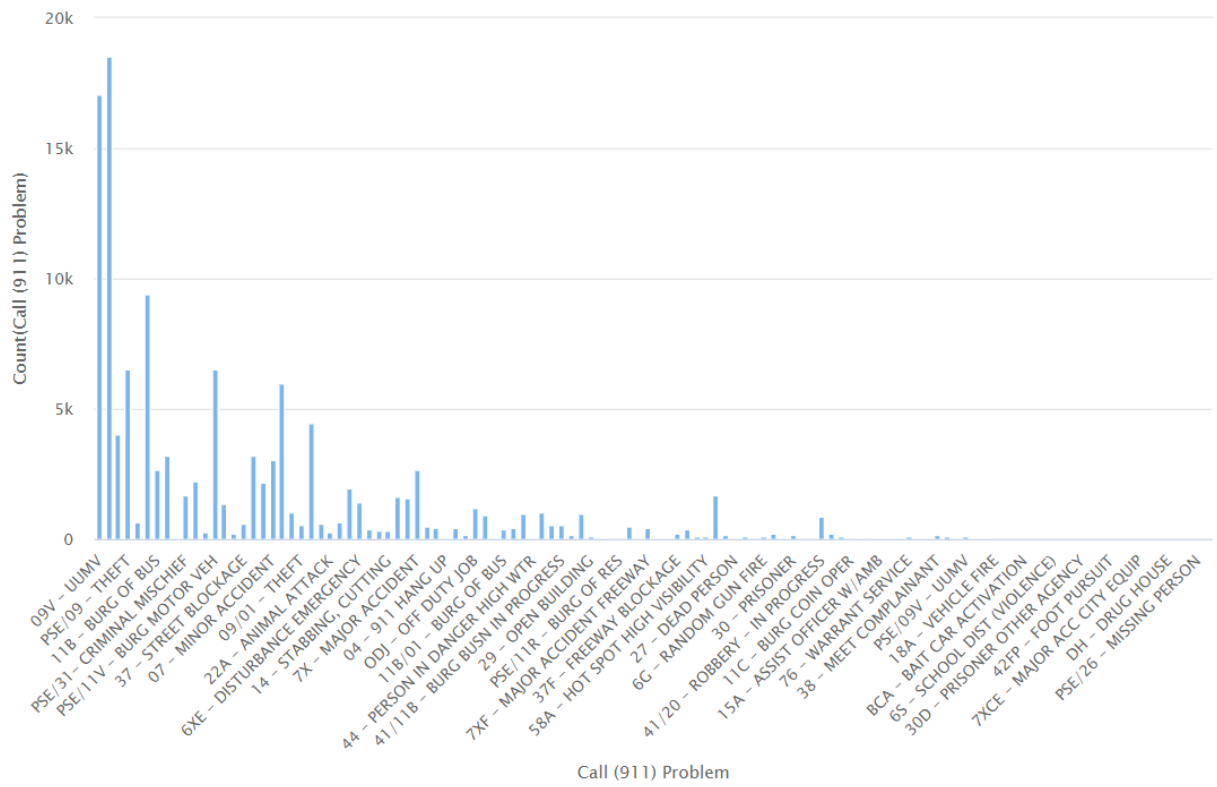
2 Attribute: ‘Year of Incident’, unique value corresponding to the year of occurrence of crime.

3 Attribute: ‘Service Number ID’, It is a unique ID appended with year and offense number.

4 Attribute: ‘Watch’ - Police Patrol Late Night, Day and Evenings. Night patrol is high.

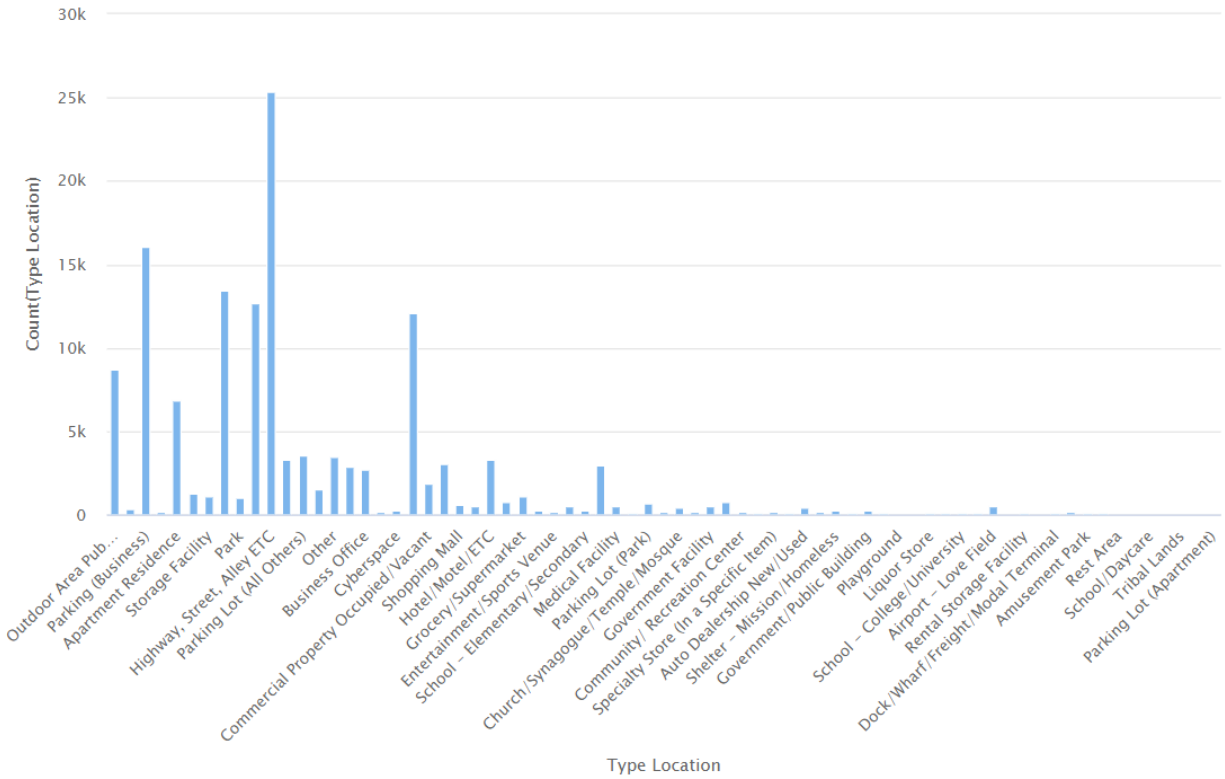


5 Attribute: ‘Call (911) Problem’ - Motor Vehicle Theft seems to be the one of the highest 911 calls.



6 Attribute: ‘Type of Incident’, type of crime which is similar to NIBR crime.

7 Attribute: ‘Type Location’, location of the crime occurred, not in terms of coordinates, actual place as in parking lot etc.

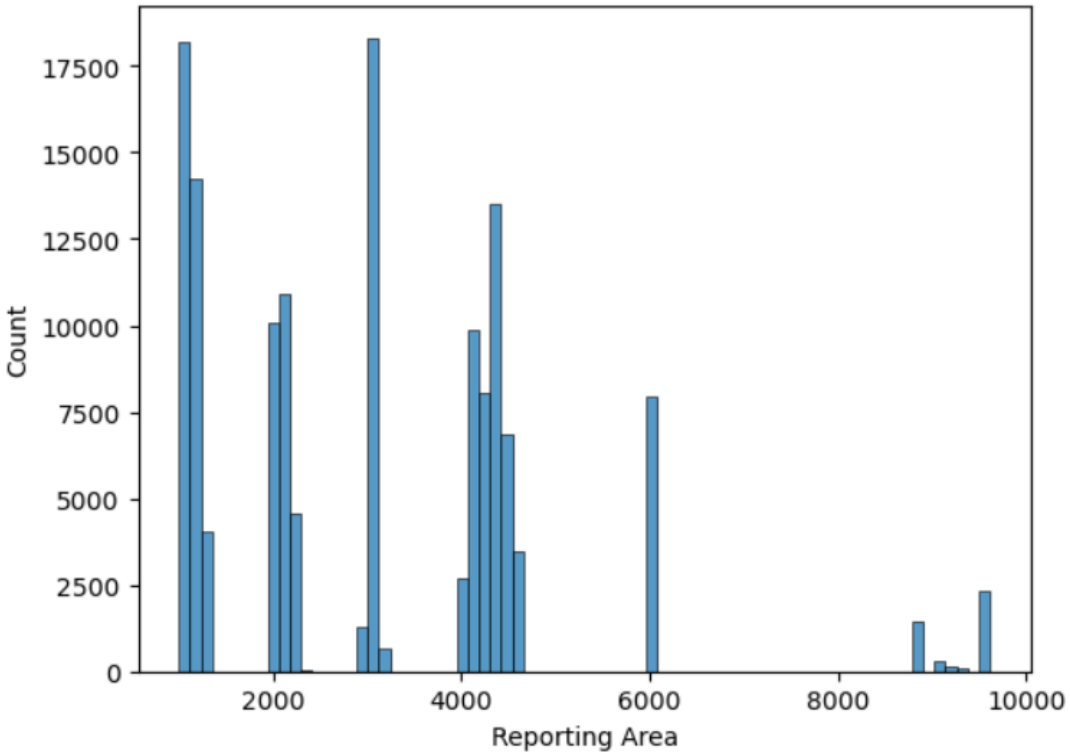


8 Attribute: ‘Type of Property’, very similar to the above attribute, ‘Type Location’.

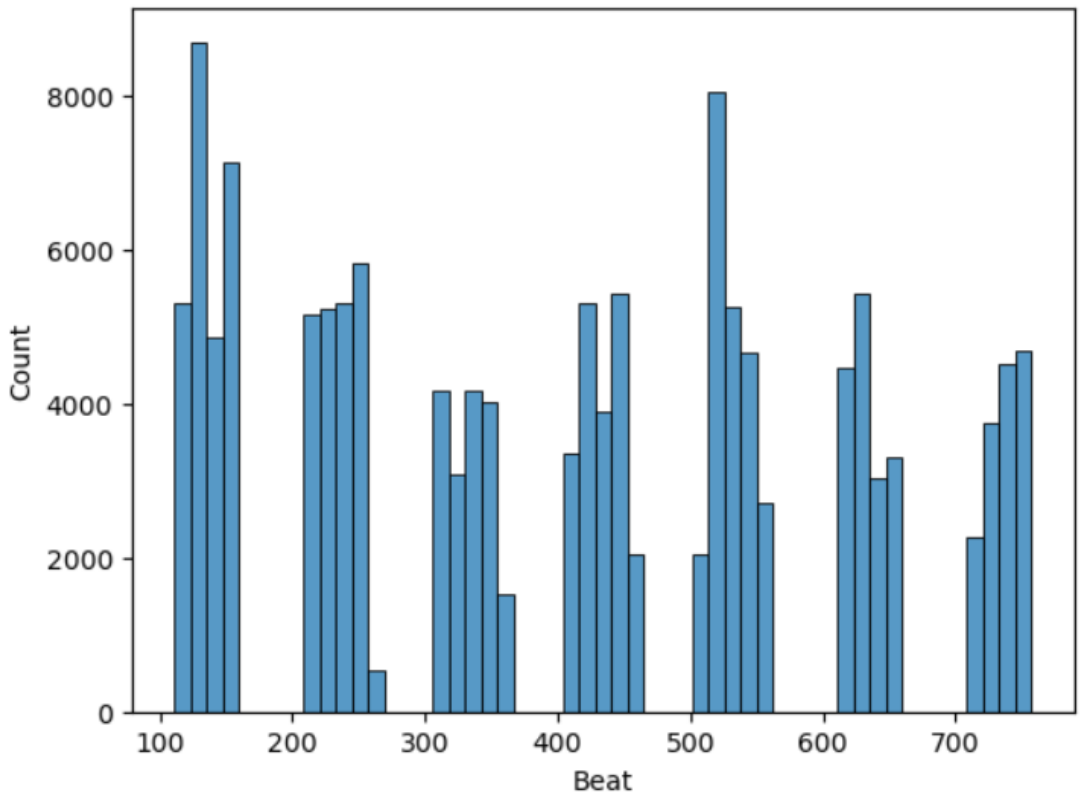
9 Attribute: ‘Incident Address’, very similar to the other location attributes present in the dataset.

10 Attribute: ‘Apartment Number’, part of the address attribute.

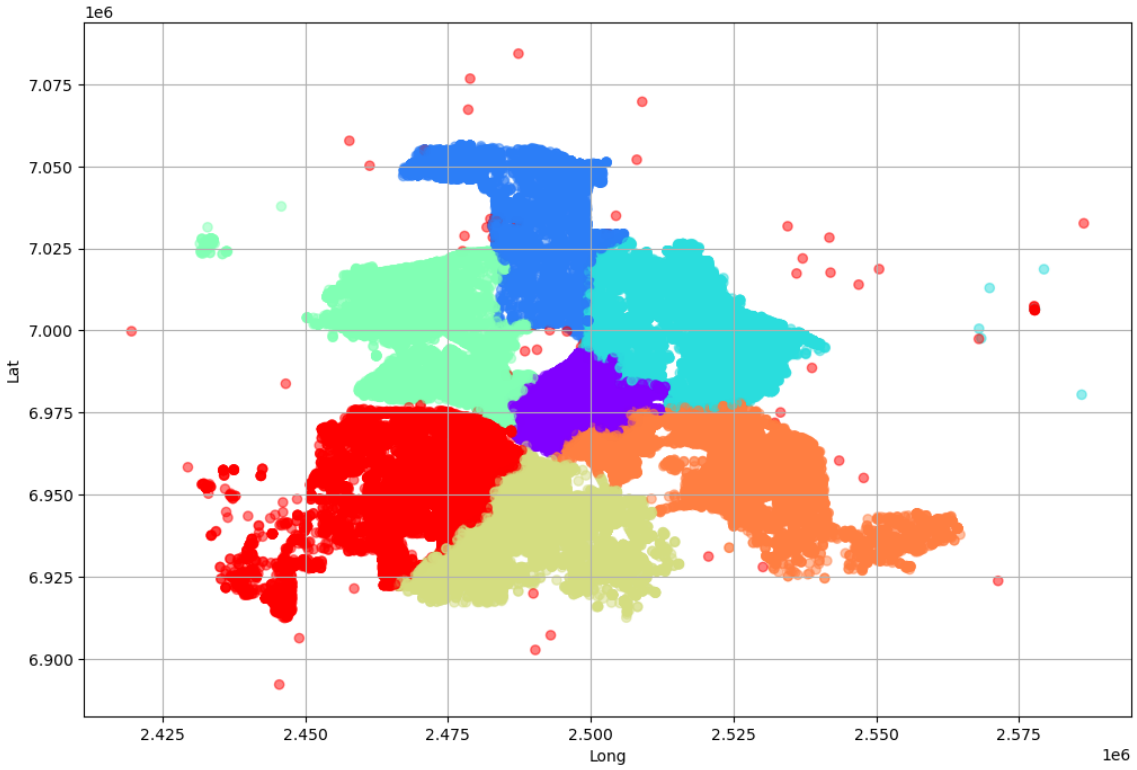
11 Attribute: ‘Reporting Area’, In geographic hierarchy, Reporting Area is on the 3rd level, used in finer data analysis.



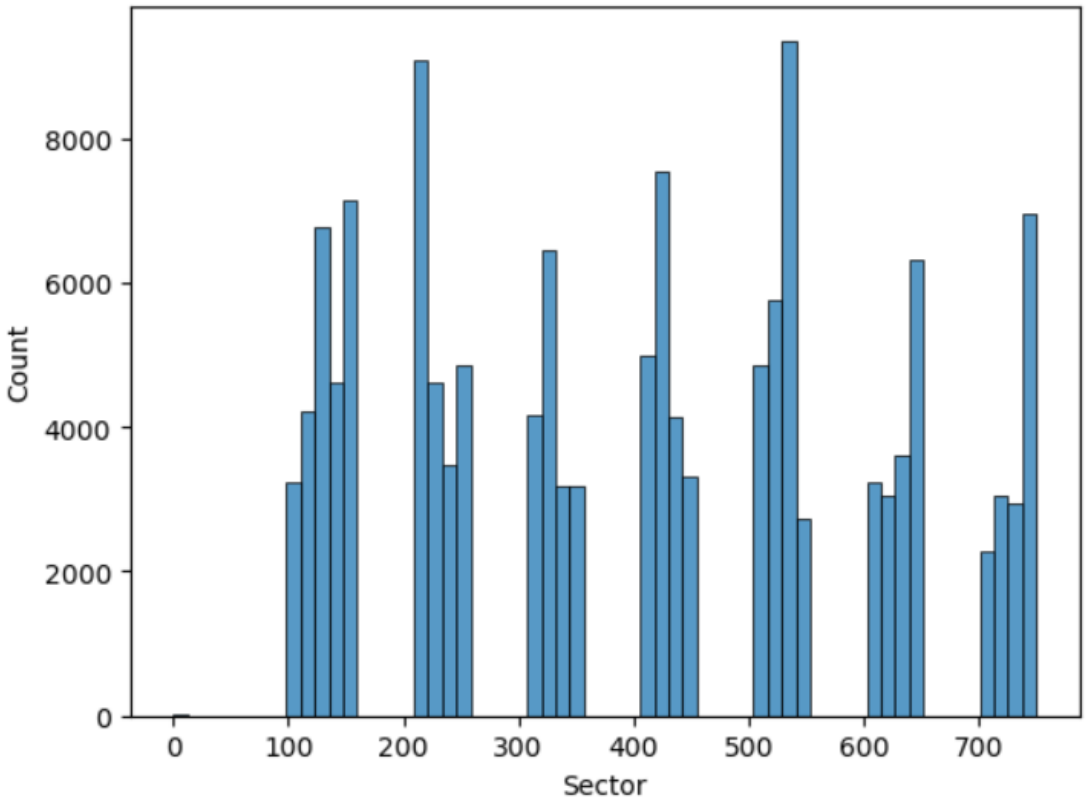
12 Attribute: ‘Beat’, In geographic hierarchy Beat is at the bottom level, used in finer data analysis.



13 Attribute: ‘Division’, In geographic hierarchy, Division is the topmost level, used in broader data analysis.



14 Attribute: ‘Sector’, In geographic hierarchy, Sector is the 2nd topmost level, used in broader data analysis.



15 Council District

Index	Nominal value	Absolute count	Fraction
1	D2	17269	0.124
2	D6	17026	0.122
3	D14	15451	0.111
4	D7	11293	0.081
5	D4	8908	0.064
6	D8	8903	0.064
7	D11	8895	0.064
8	D13	8692	0.063
9	D1	8329	0.060
10	D10	7841	0.056
11	D9	7703	0.055
12	D3	7616	0.055
13	D5	6076	0.044
14	D12	5010	0.036
15	8	2	0.000
16	10	1	0.000
17	9	1	0.000

16 Target Area Action Grids

17 Community

18 Date1 of Occurrence

19 Year1 of Occurrence

20 Month1 of Occurrence

21 Day1 of the Week

22 Time1 of Occurrence

23 Day1 of the Year

24 Date2 of Occurrence

25 Year2 of Occurrence

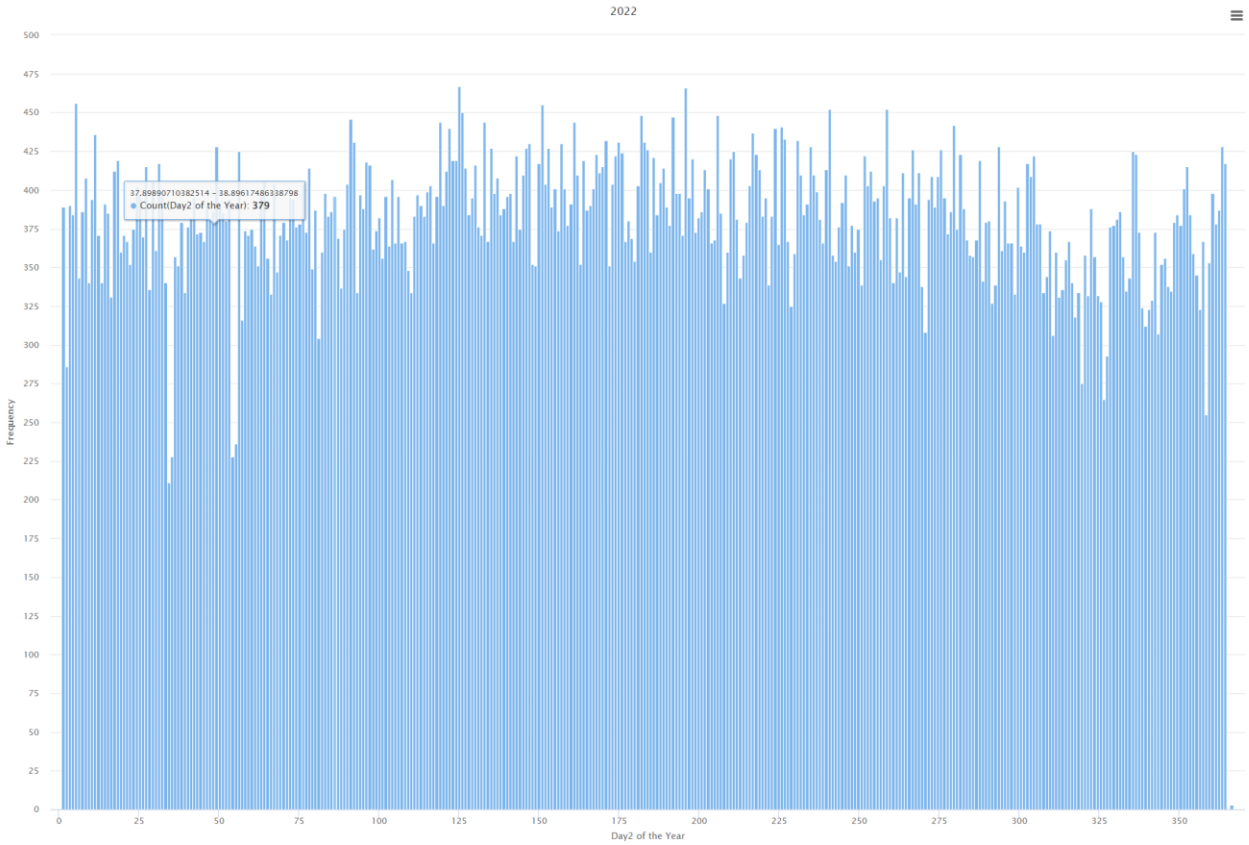
26 Month2 of Occurrence

27 Day2 of the Week

28 Time 2 of Occurrence

29

Day2 of the year: this attribute has 366 unique values represents the calendar day of the year (1–365 or 366 for leap years) for the second recorded date of an incident.

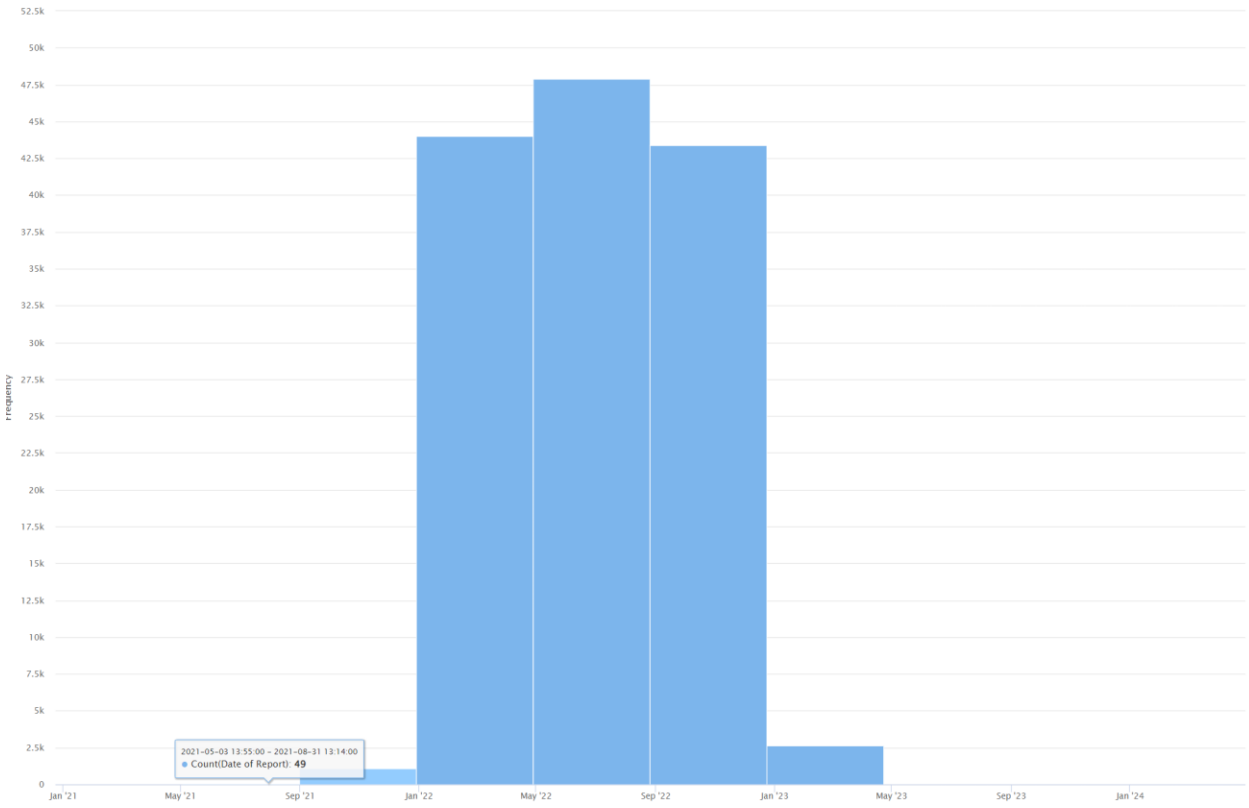


30.

Date of Report: this attribute represents the date on which the incident was officially reported to the authorities (it a string composing with the date and time of the report).

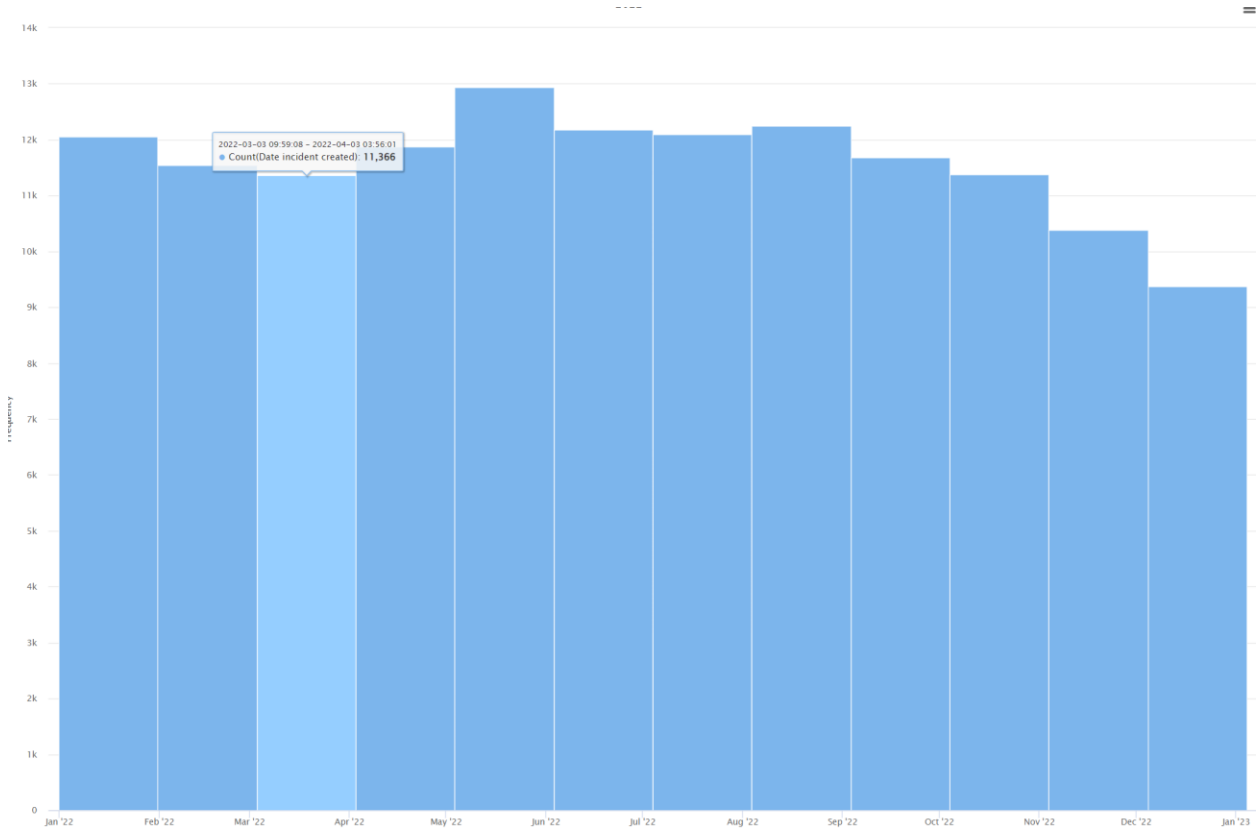
It has 101462 unique values

Data seems to be inconsistent as some incidents are reported before the actual incident is created .



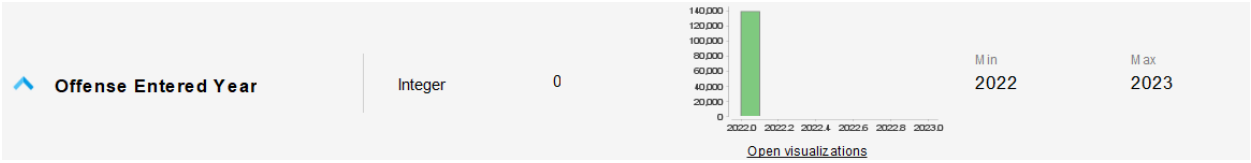
31.

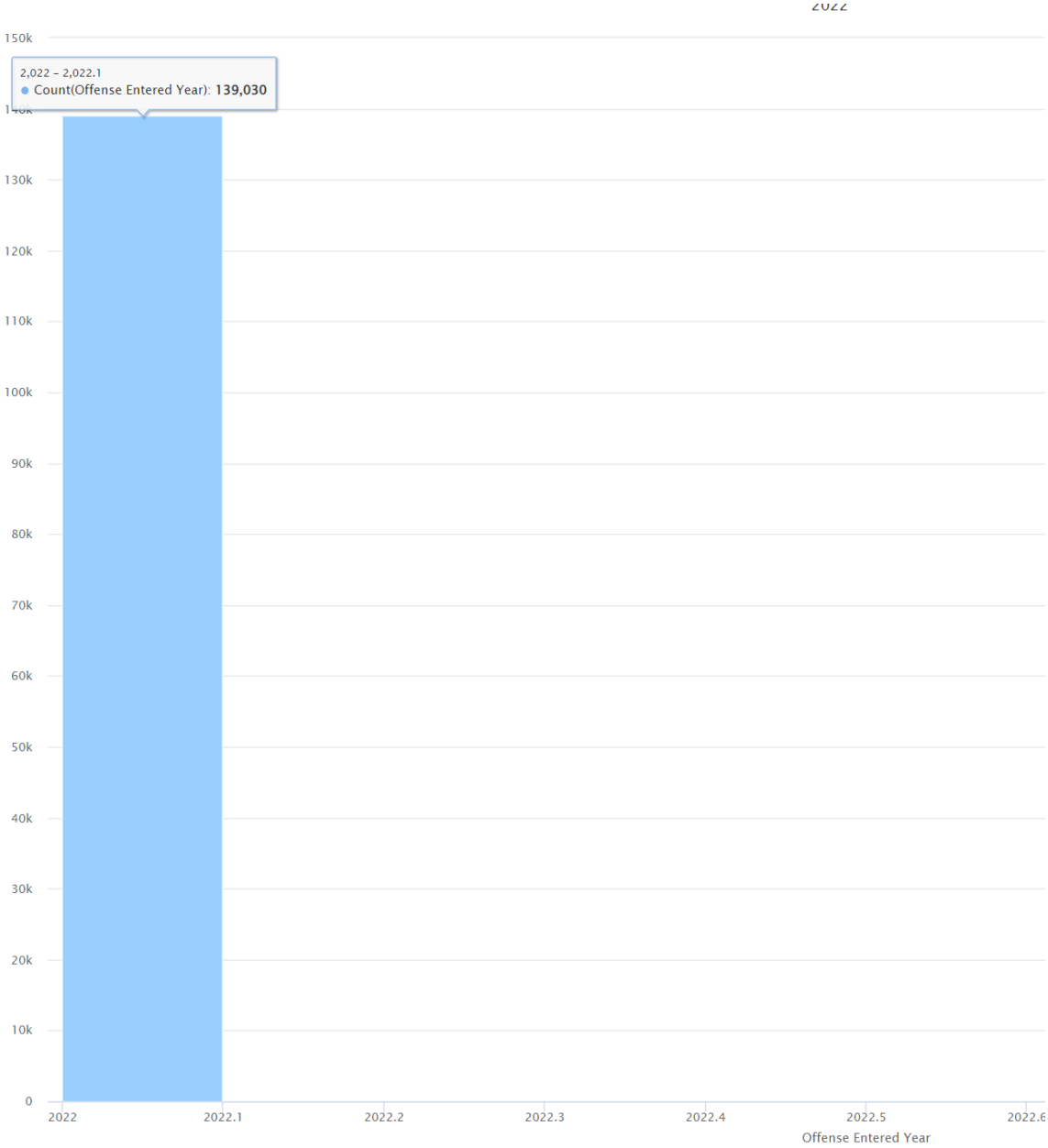
Date incident created : this attribute is of type string , it refers to the date on which the incident was logged in the system. It has 109845 unique values



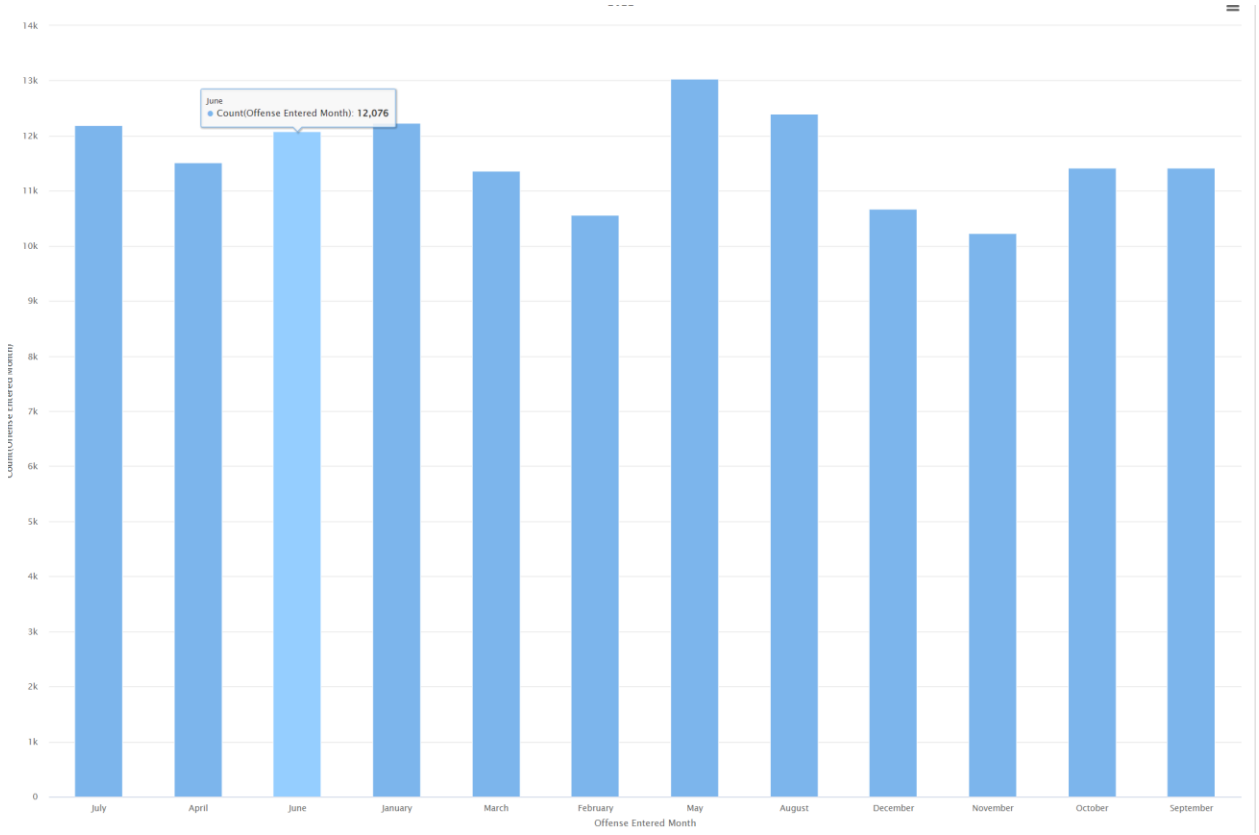
32

Offense Enter Year : this attribute means the year in which the offense was officially entered into the system. It has 2 unique values (2022, 2023).





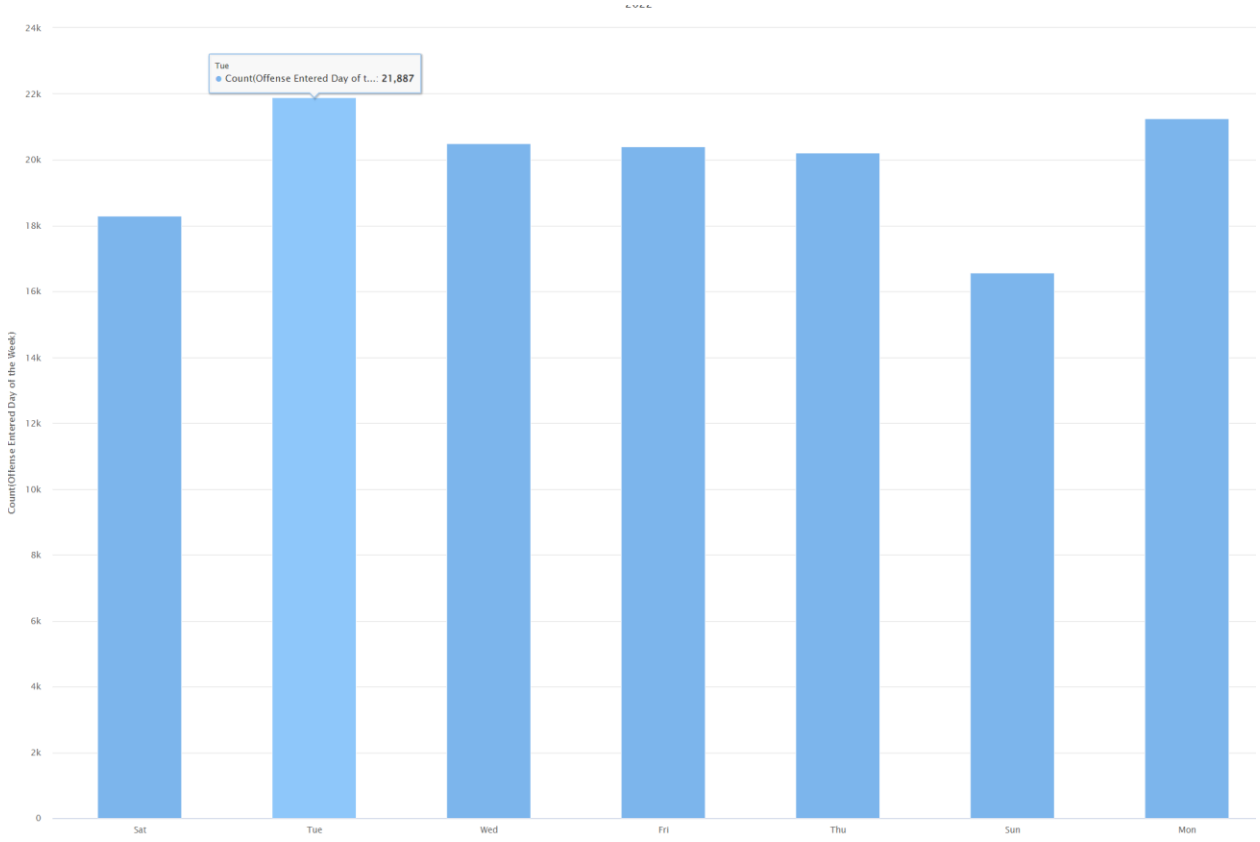
33 Offence Entered Month: this attribute means the month in which the offense was entered into the system. It has 12 unique values which is equivalent to 12 months of the year and we can see from bar chart the may has the most entries



34. Offense Entered Day of the week :

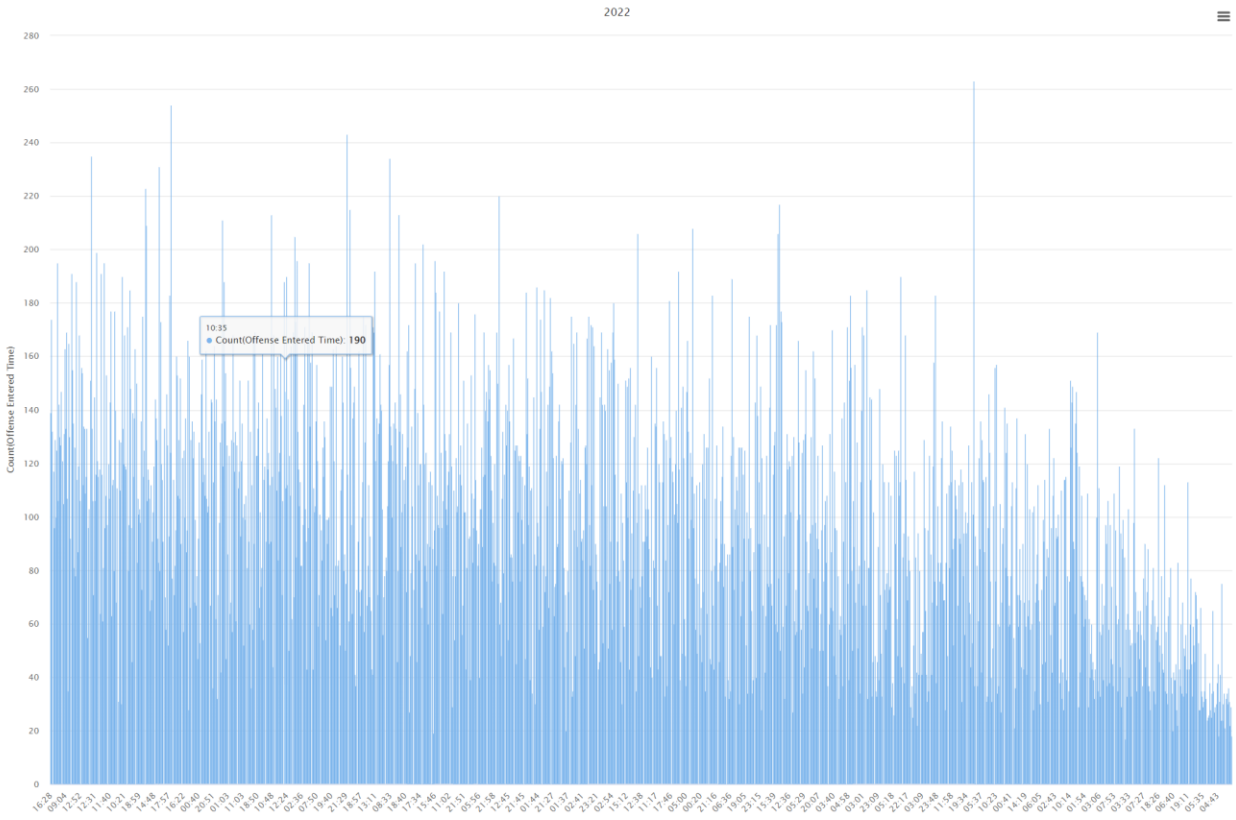
The day of the week (e.g., Monday, Tuesday) when the offense was entered into the system. This attribute has 7 unique values.

From the bar chart we can see that Tuesday has the most entries.

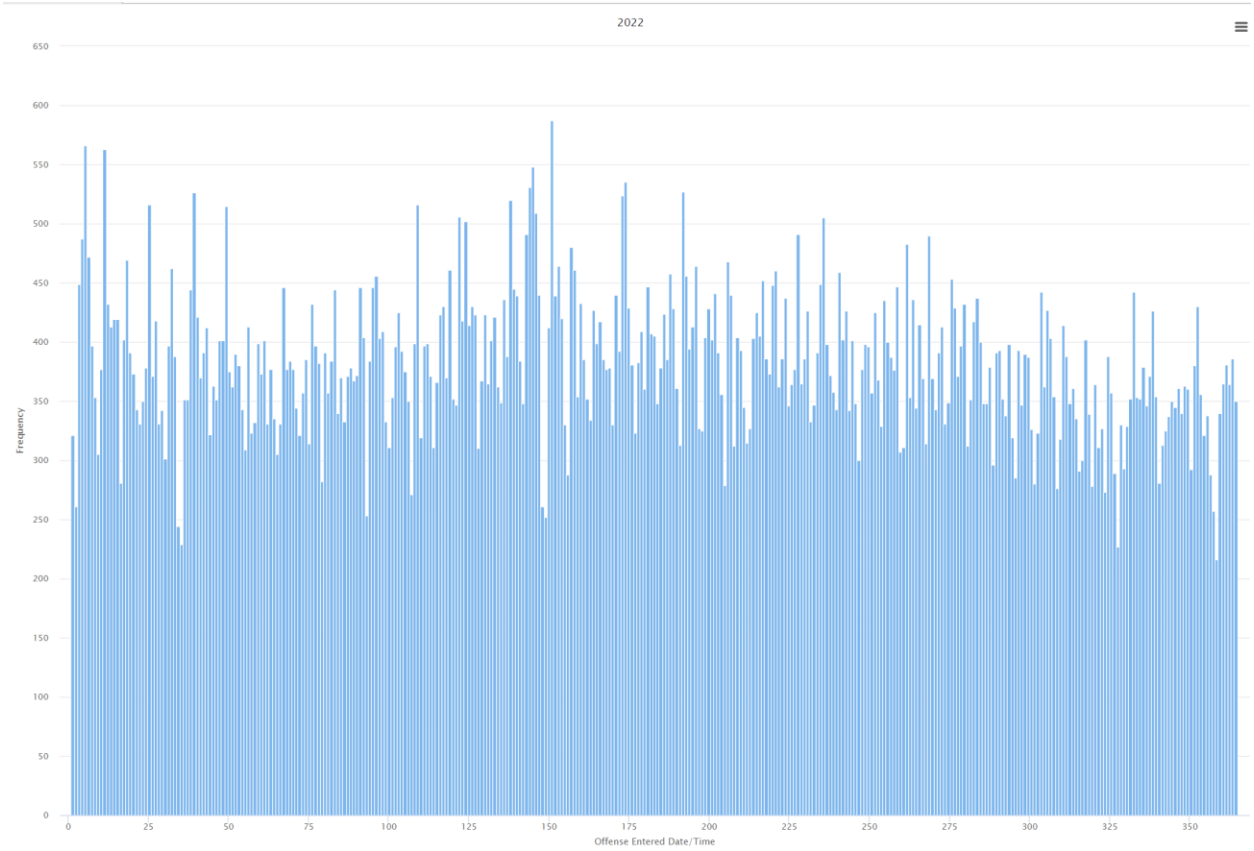


35 Offence Entered Time this attribute means the time of day when the offense was entered into the system.

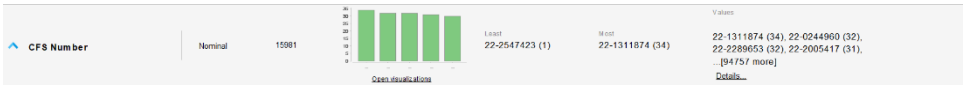
It has 1440 unique values . 15:15 has the most entries.



36 Offense Entered Time: The exact date and time when the offense was entered into the system. This attribute has 365 unique values , we can see that on day 150 was the most entries with 587 being entered in the system .

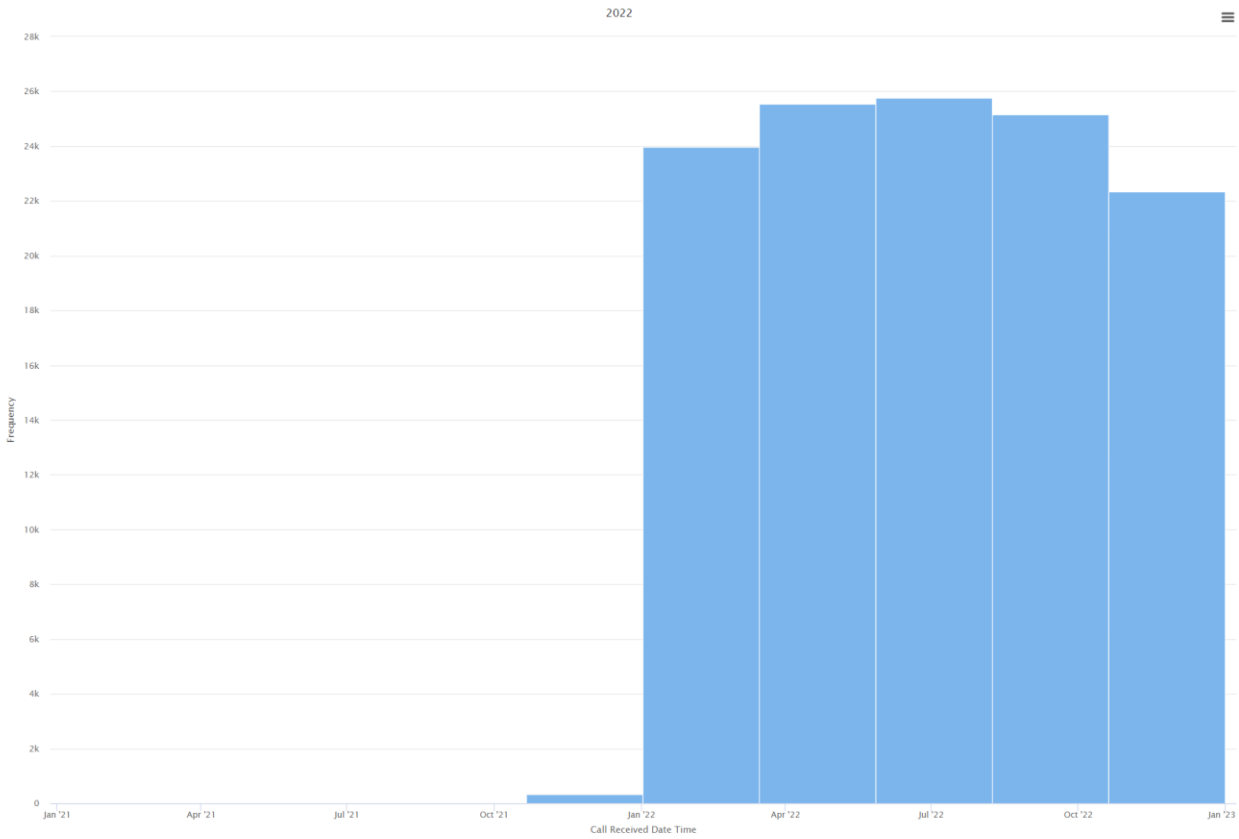


37 .CFS Number: this attribute Call For Service (CFS) number, a unique identifier for a dispatched incident. It has 94763 unique values . the most is 22-1311874 with 34 repetitions .

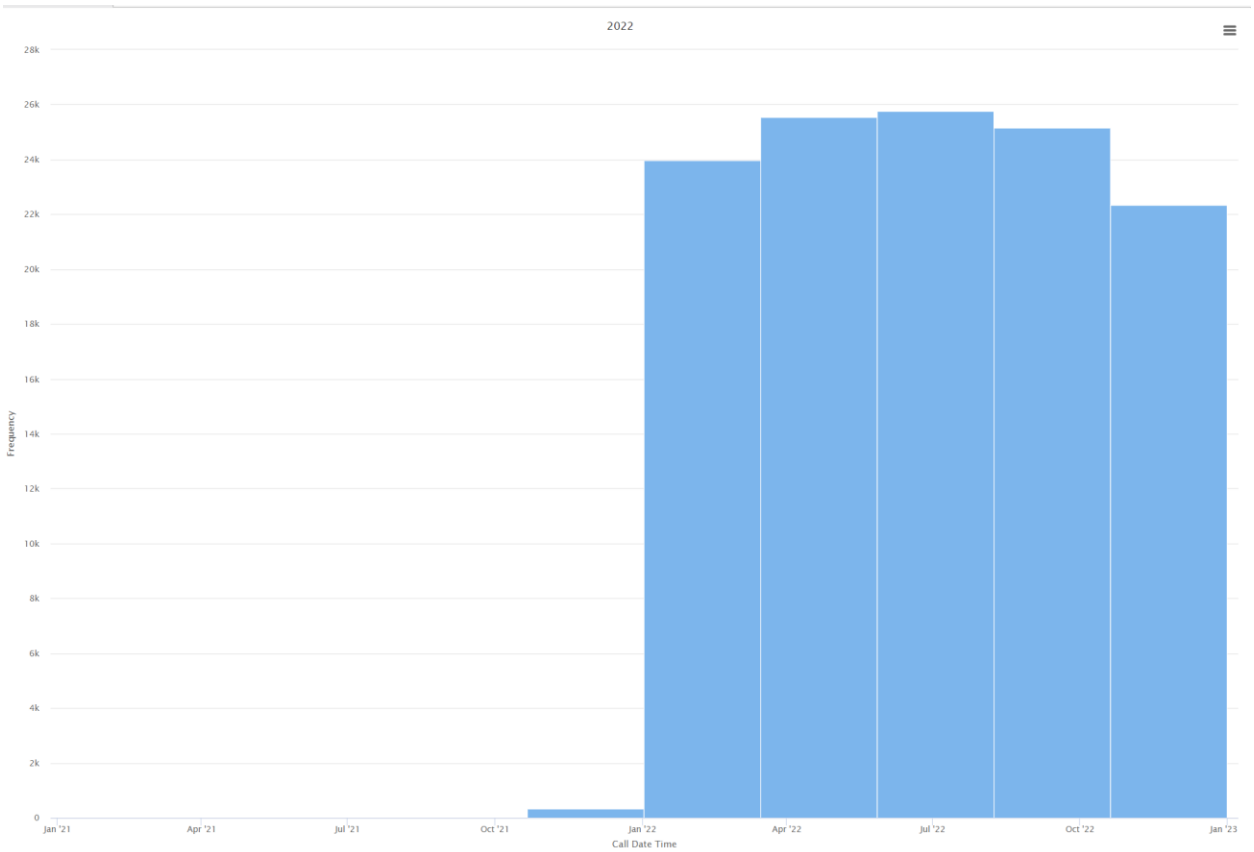


38.

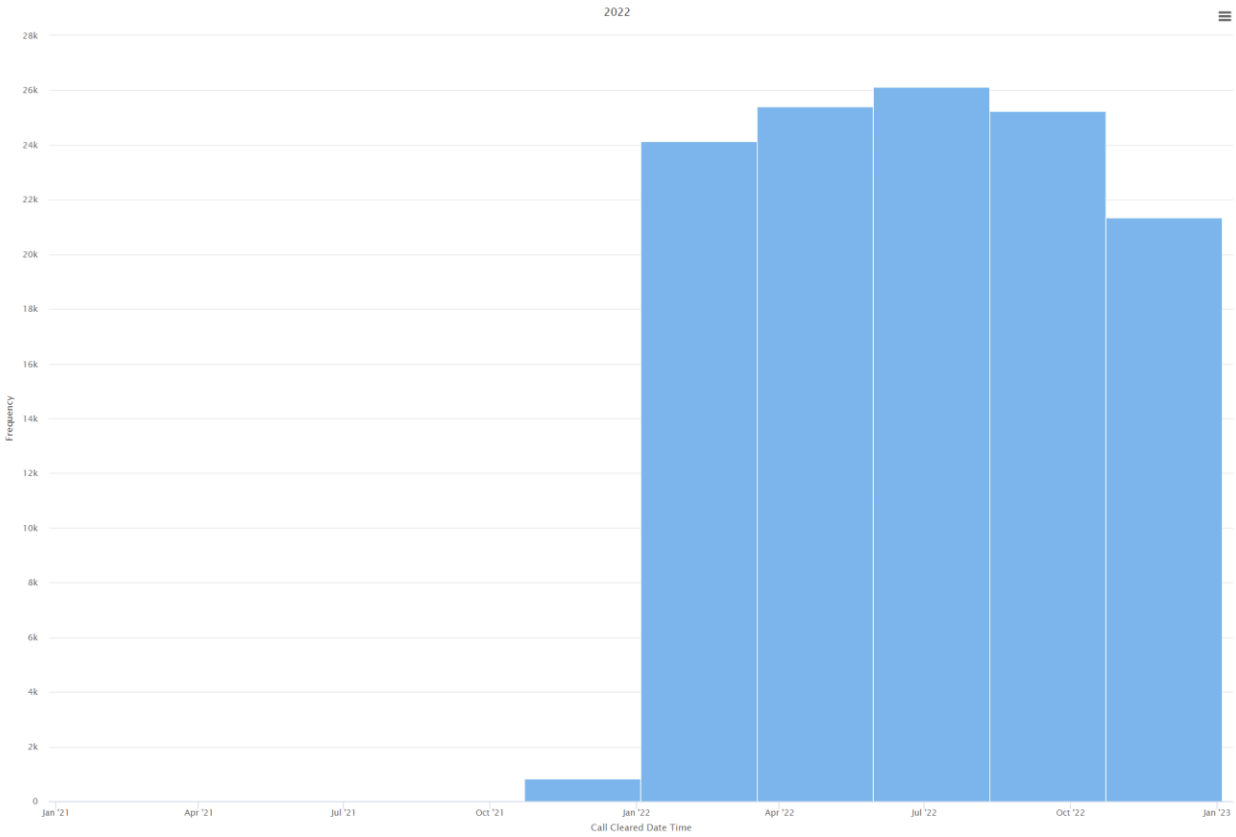
Call Received Date Time : The exact date and time the call was received by emergency services. This attribute is in a string format and has 94666 unique values, it is in string format with date and time



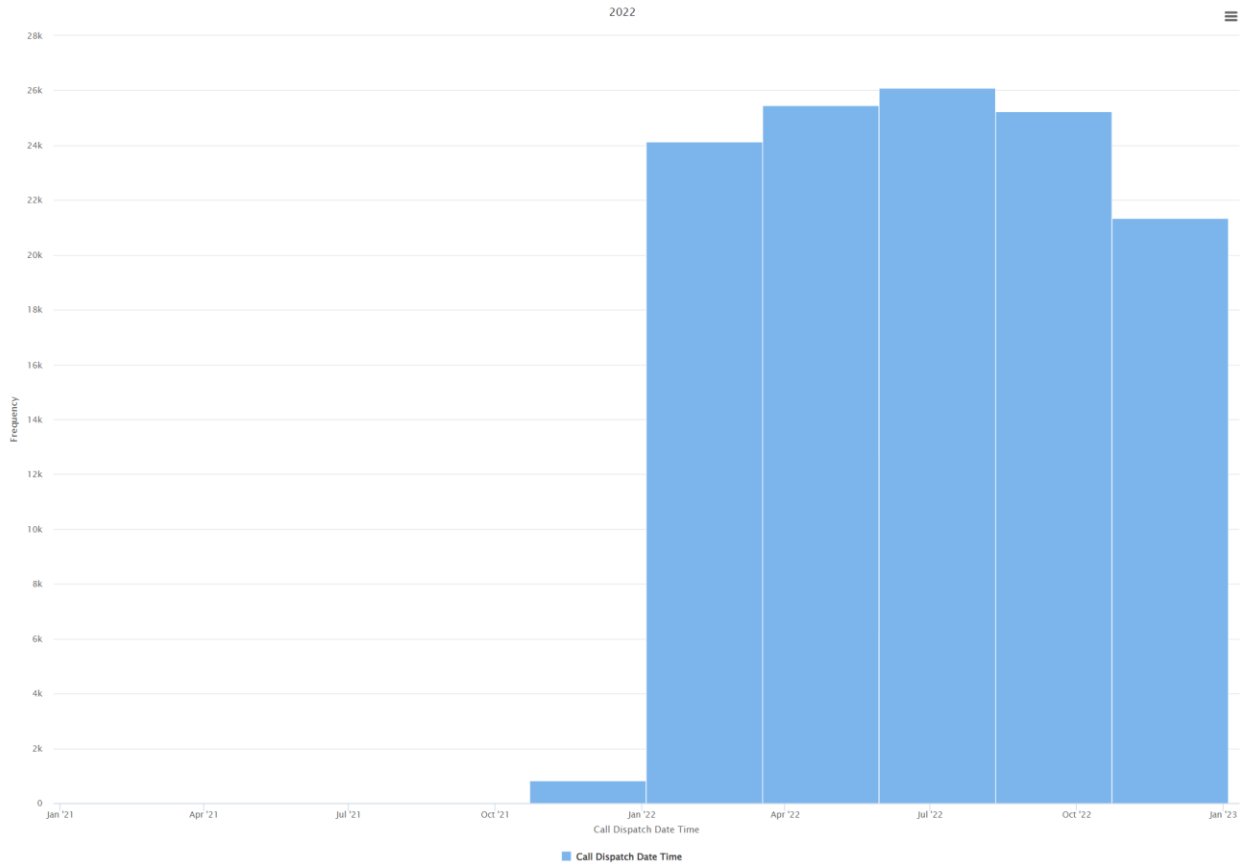
39. Call date time : The date and time associated with the call. It has 94642 unique values, it is in string format with date and time



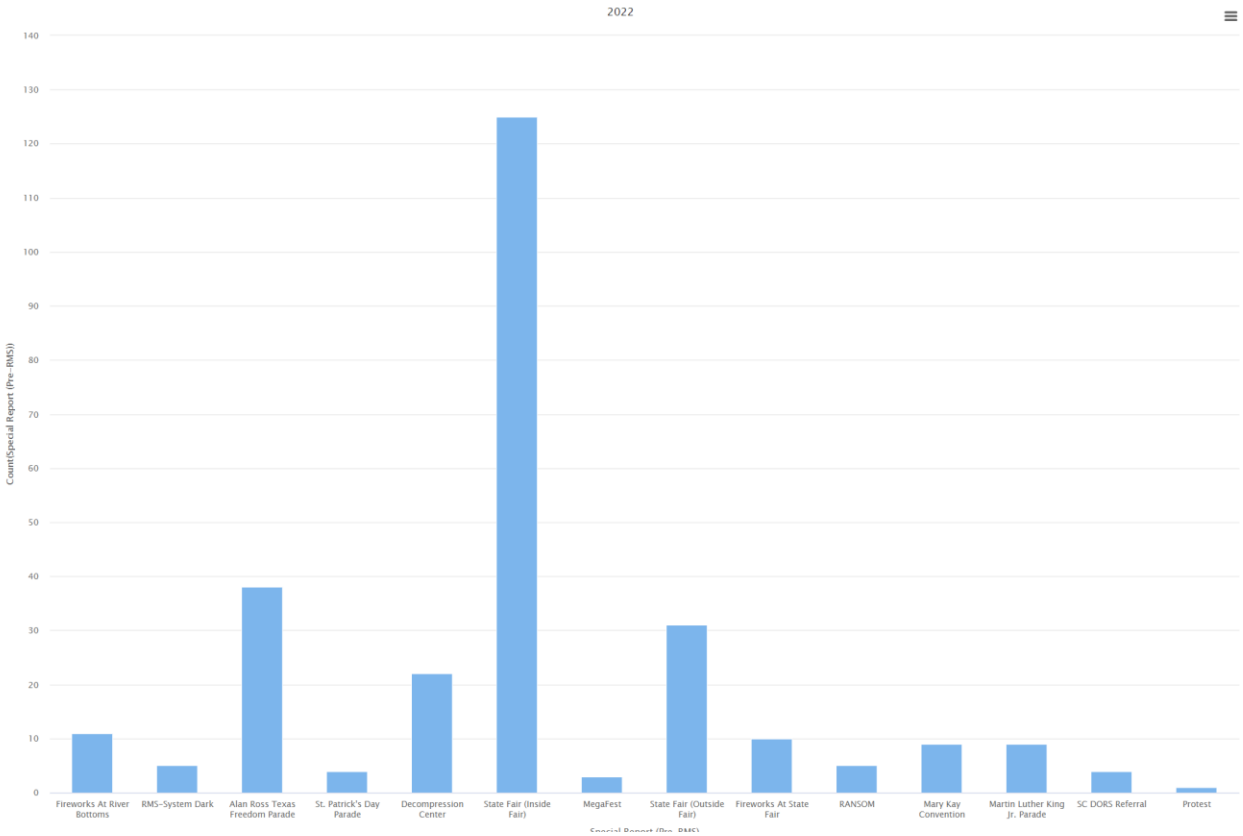
40. Call Cleared Date Time : The date and time when the call was officially cleared or resolved. It has 94642 unique values it is in string format with date and time



41. Call dispatched Date time : The date and time when officers were dispatched to respond to the call. It has 94549 unique values it is in string format with date and time

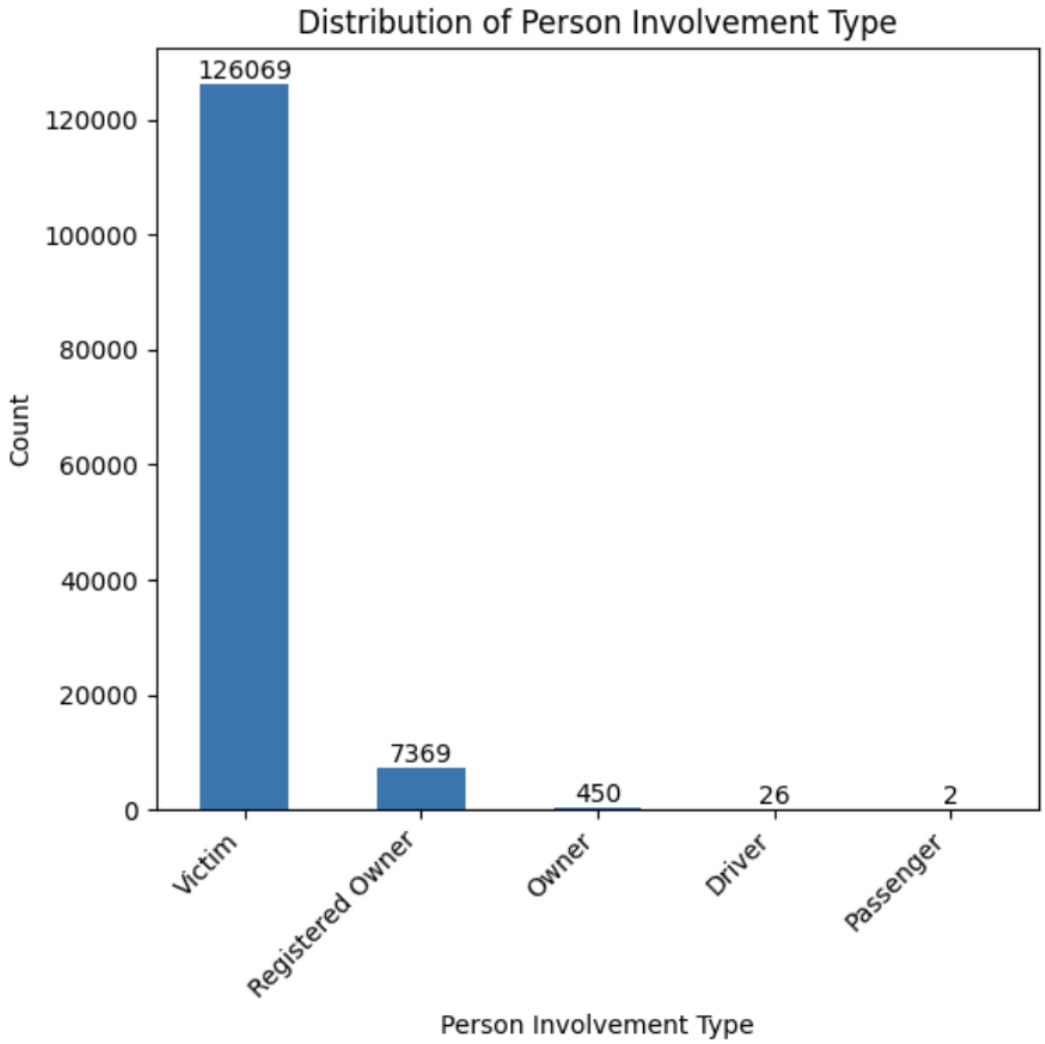


42 special report (Pre-Rms) : Pre-Record Management System (Pre-RMS) special report, possibly containing legacy data or additional notes. This attribute has 14 unque ivalues and 99 missing values .


2022
Count(Special Report (Pre- RMS))
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
Fireworks At River Bottoms
RMS-System Dark
Alan Ross Texas Freedom Parade
St. Patrick's Day Parade
Decompression Center
State Fair (Inside Fair)
MegaFest
State Fair (Outside Fair)
Fireworks At State Fair
RANSOM
Mary Kay Convention
Martin Luther King Jr. Parade
SC DORS Referral
Protest

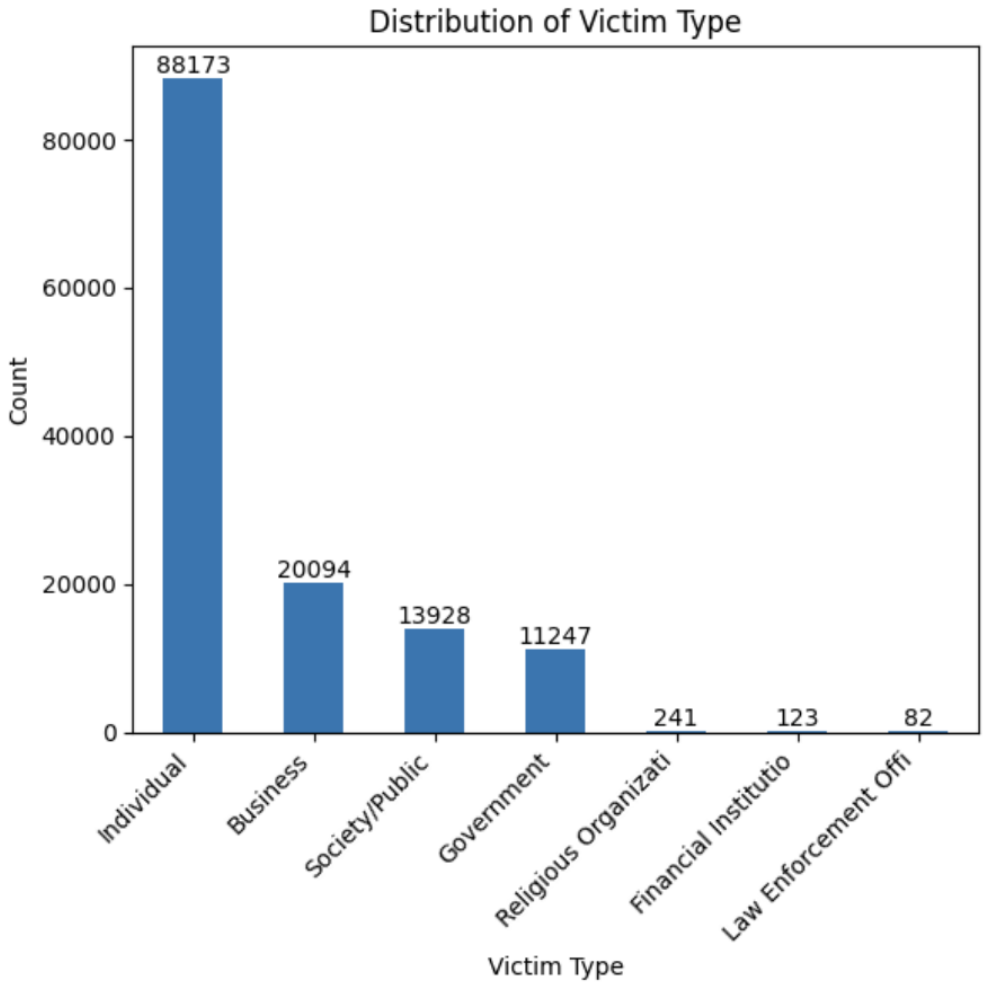
43 Attribute: “Person Involvement Type” has 5 distinct values,

```
Attribute: Person Involvement Type: has 5 distinct values
  'Victim': 126069 instances (90.62%)
  'Registered Owner': 7369 instances (5.30%)
  'Owner': 450 instances (0.32%)
  'Driver': 26 instances (0.02%)
  'Passenger': 2 instances (0.00%)
```



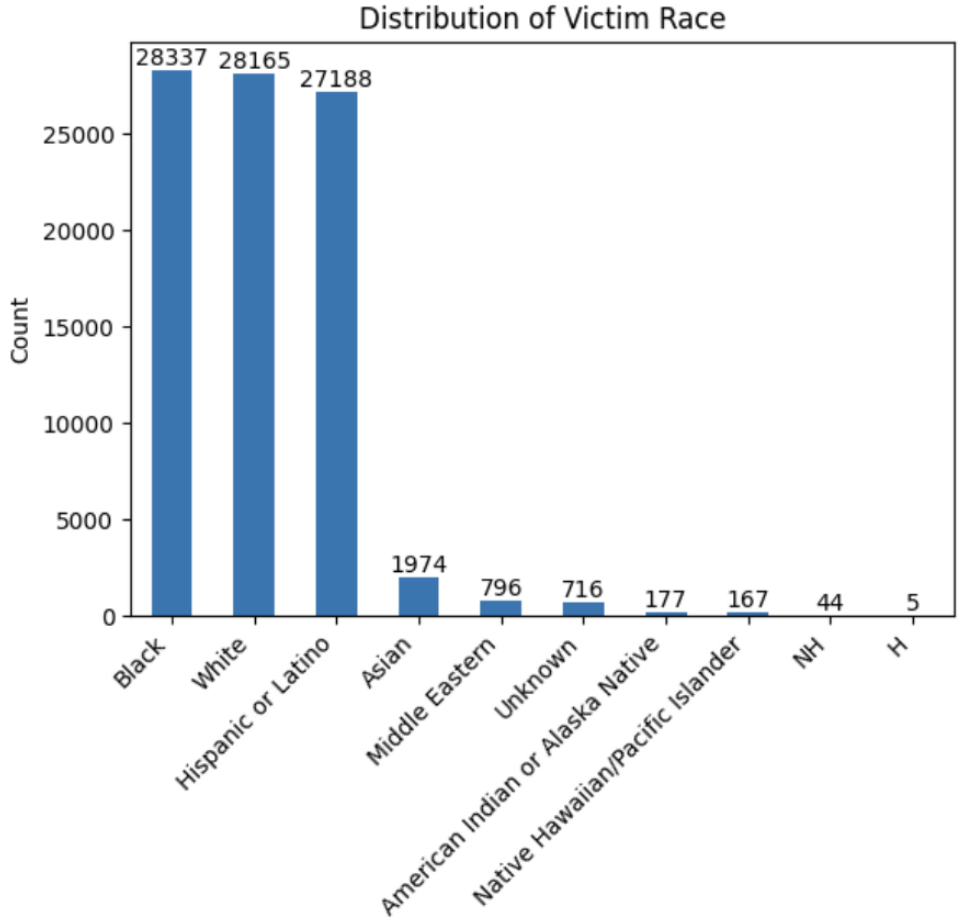
44 Attribute: Victim Type: has 7 distinct values

```
Attribute: Victim Type: has 7 distinct values
  'Individual': 88173 instances (63.38%)
  'Business': 20094 instances (14.44%)
  'Society/Public': 13928 instances (10.01%)
  'Government': 11247 instances (8.08%)
  'Religious Organizati': 241 instances (0.17%)
  'Financial Institutio': 123 instances (0.09%)
  'Law Enforcement Offi': 82 instances (0.06%)
```



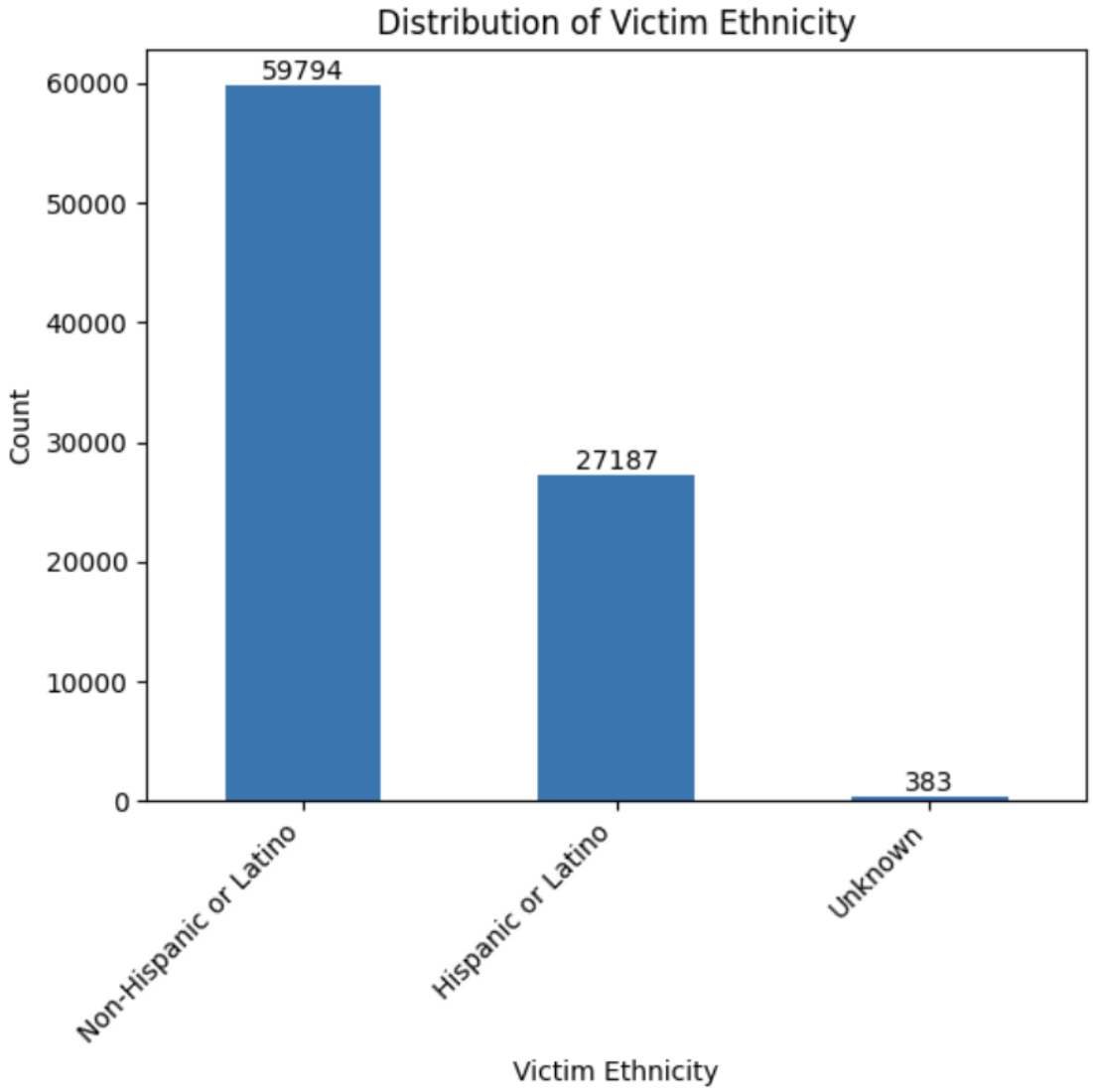
45 Attribute: Victim Race: has 10 distinct values

```
Attribute: Victim Race: has 10 distinct values
  'Black': 28337 instances (20.37%)
  'White': 28165 instances (20.25%)
  'Hispanic or Latino': 27188 instances (19.54%)
  'Asian': 1974 instances (1.42%)
  'Middle Eastern': 796 instances (0.57%)
  'Unknown': 716 instances (0.51%)
  'American Indian or Alaska Native': 177 instances (0.13%)
  'Native Hawaiian/Pacific Islander': 167 instances (0.12%)
  'NH': 44 instances (0.03%)
  'H': 5 instances (0.00%)
```



46 Attribute: Victim Ethnicity: has 3 distinct values

```
Attribute: Victim Ethnicity: has 3 distinct values
  'Non-Hispanic or Latino': 59794 instances (42.98%)
  'Hispanic or Latino': 27187 instances (19.54%)
  'Unknown': 383 instances (0.28%)
```

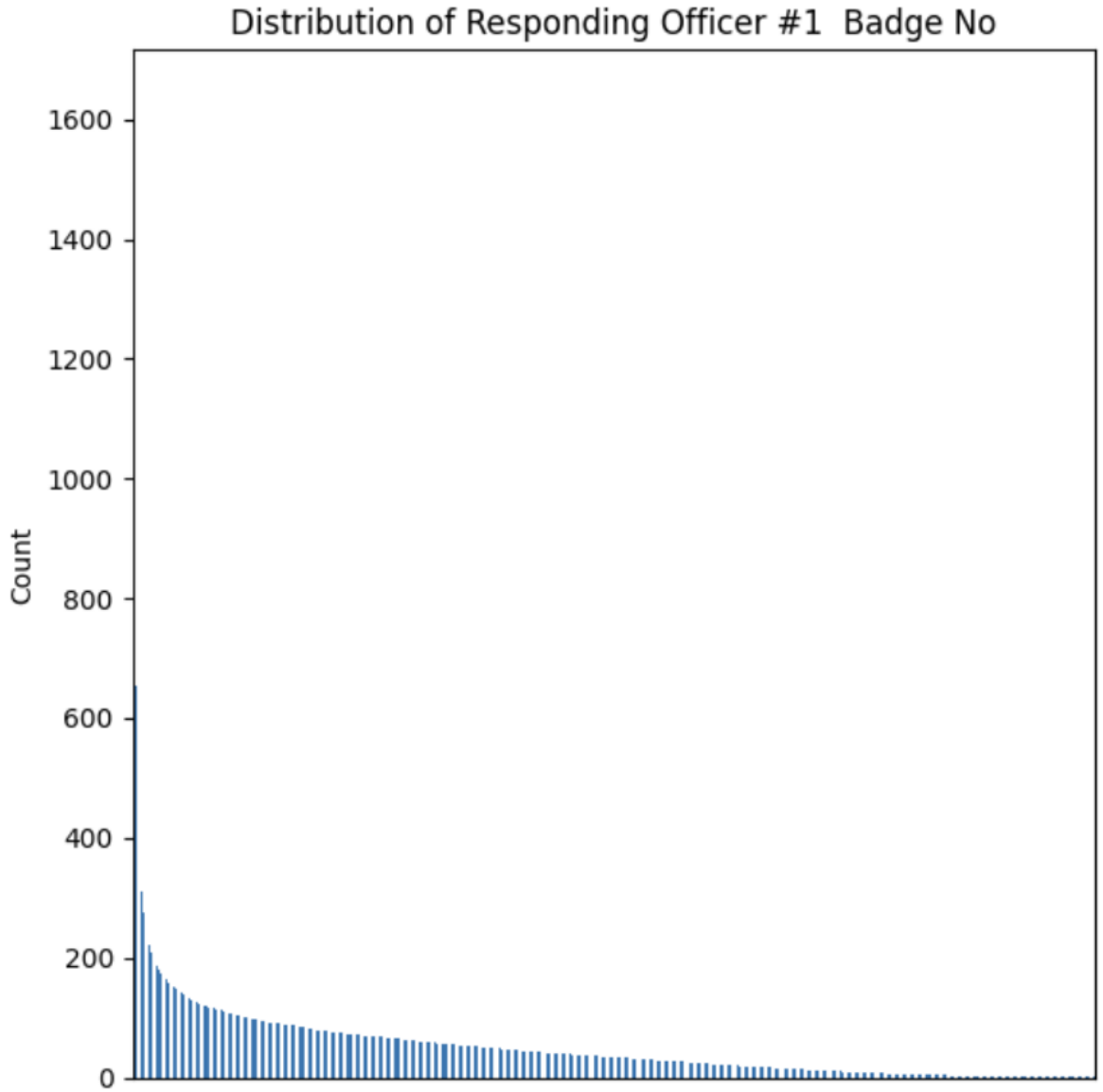


47 Attribute: Victim Gender: has 3 distinct values

```
Attribute: Victim Gender: has 3 distinct values
  'Male': 51026 instances (36.68%)
  'Female': 36303 instances (26.10%)
  'Unknown': 275 instances (0.20%)
```

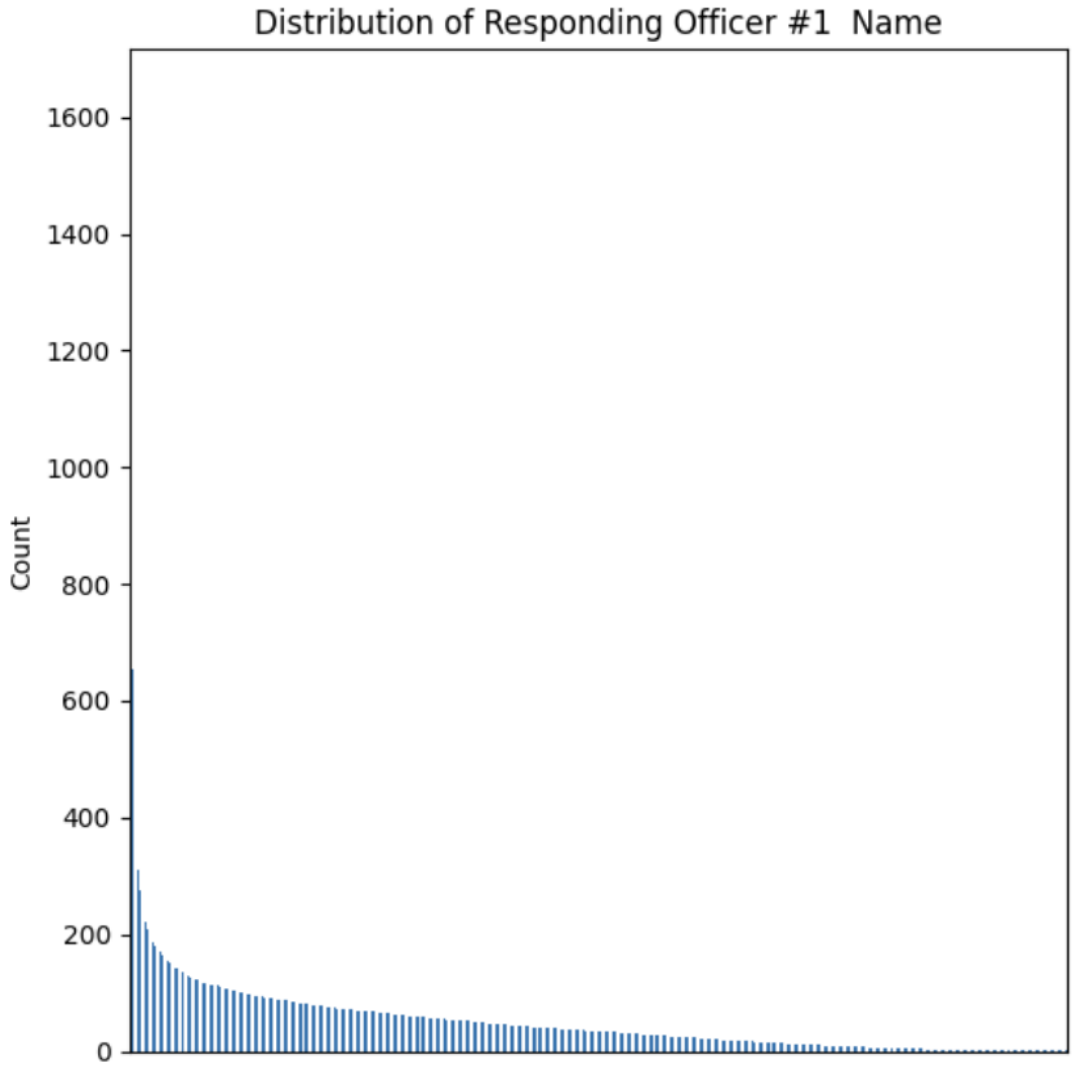
48 Responding Officer #1 Badge No

```
Attribute: Responding Officer #1  Badge No: has 2406 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: '122756': 1634 instances (1.17%)
  Least common: '7309': 1 instances (0.00%)
```



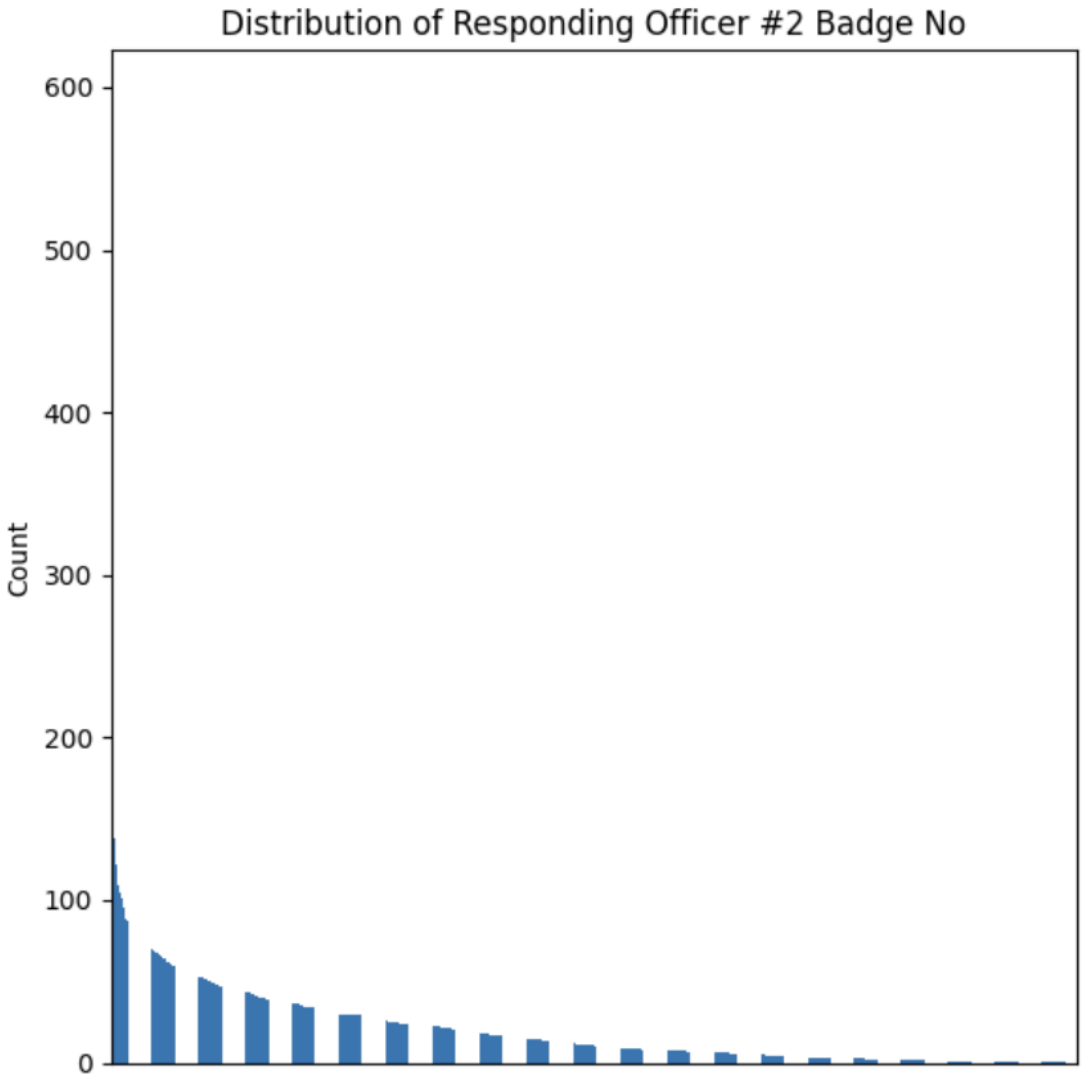
49 Responding Officer #1 Name

```
Attribute: Responding Officer #1  Name: has 2399 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: 'HORTON,JONQUIL': 1634 instances (1.17%)
  Least common: 'MILTON,JUSTIN,MIKAL': 1 instances (0.00%)
```



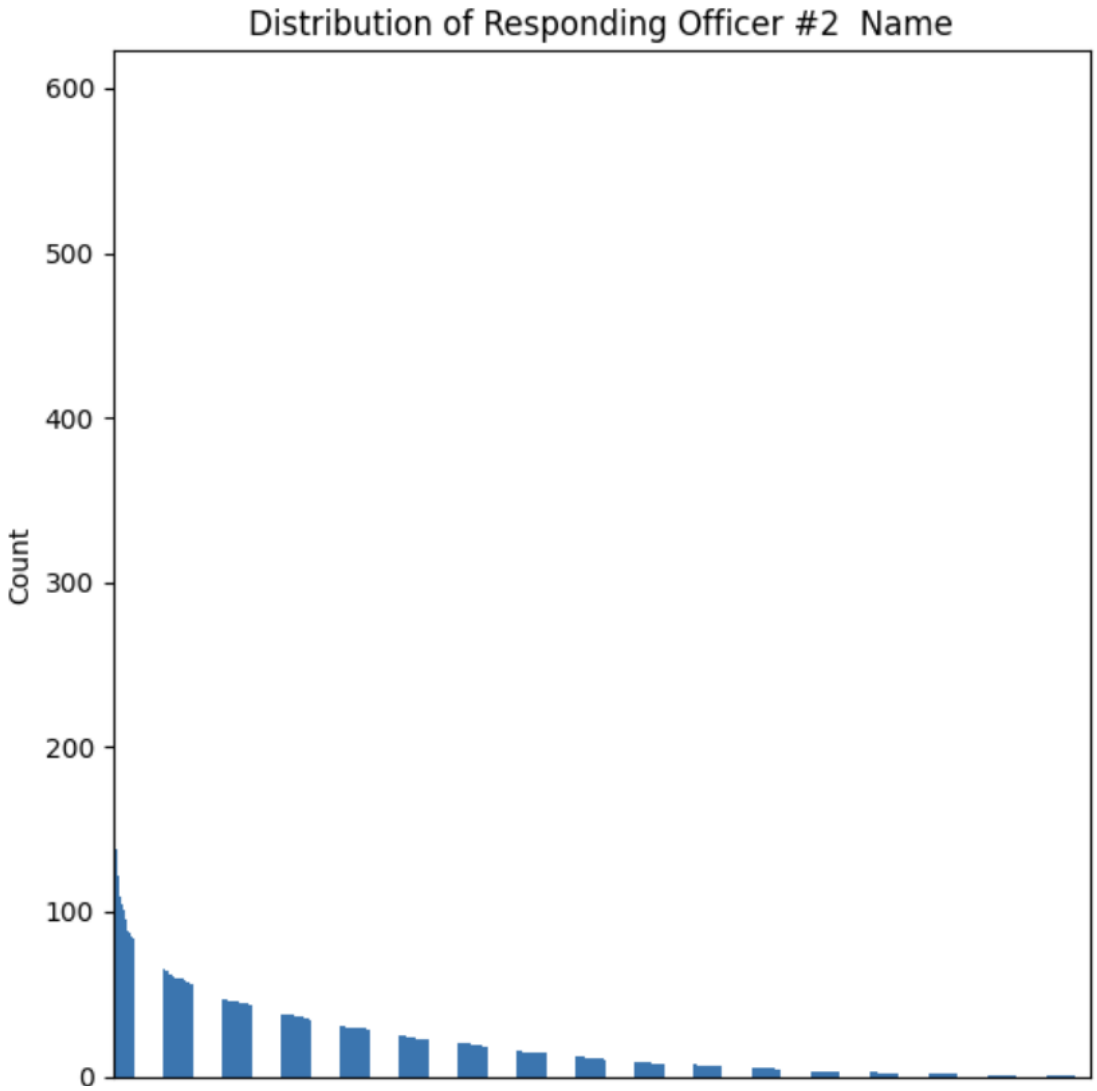
50 Responding Officer #2 Badge No

```
Attribute: Responding Officer #2 Badge No: has 2078 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: '9241': 593 instances (0.43%)
  Least common: '9139': 1 instances (0.00%)
```



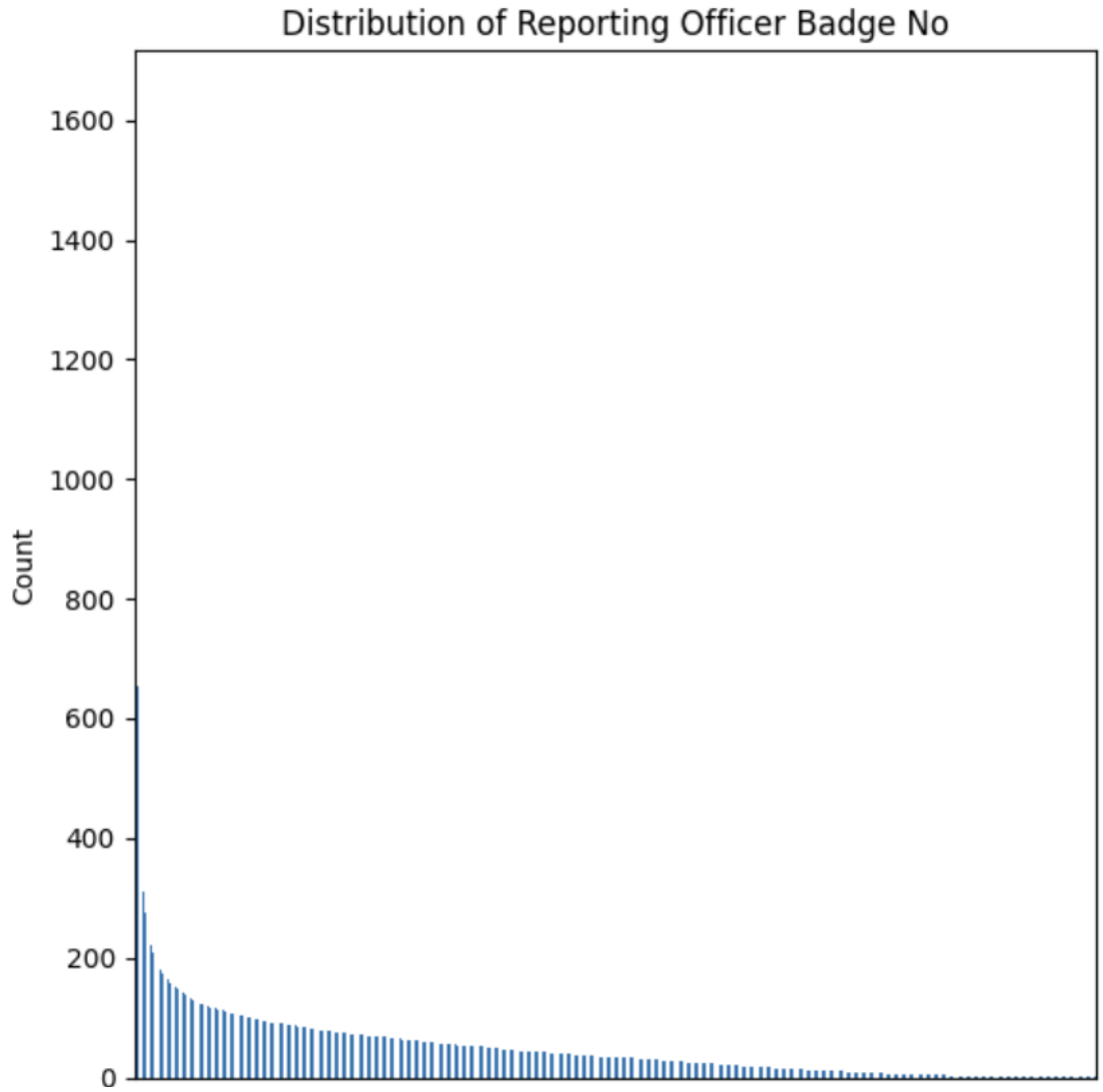
51 Responding Officer #2 Name

```
Attribute: Responding Officer #2  Name: has 2074 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: 'CARNEY,MARY,KATHRYN': 593 instances (0.43%)
  Least common: 'SHEERIN,MILES,HENRY': 1 instances (0.00%)
```



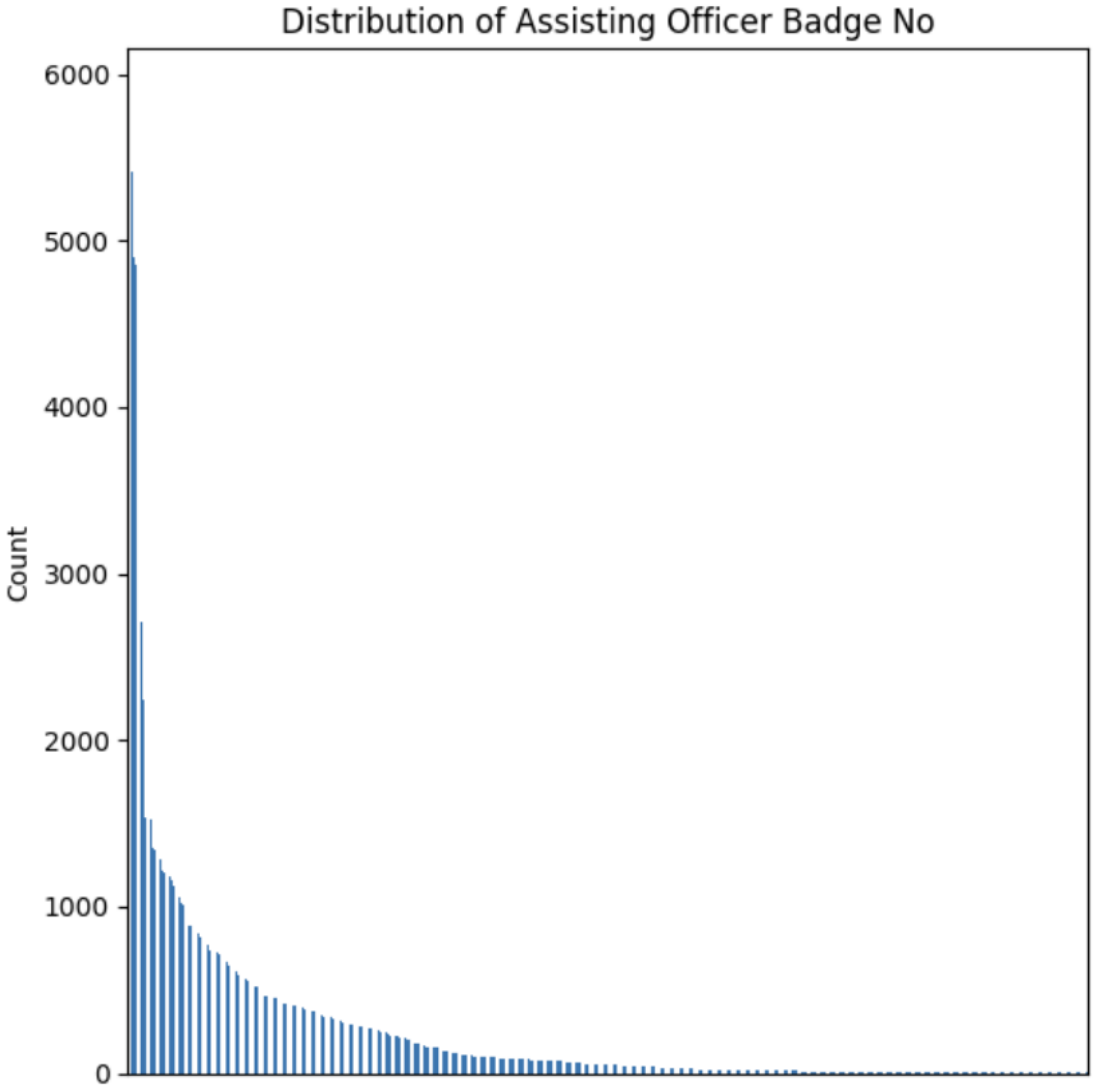
52 Reporting Officer Badge No

```
Attribute: Reporting Officer Badge No: has 2407 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: '122756': 1634 instances (1.17%)
  Least common: '7309': 1 instances (0.00%)
```



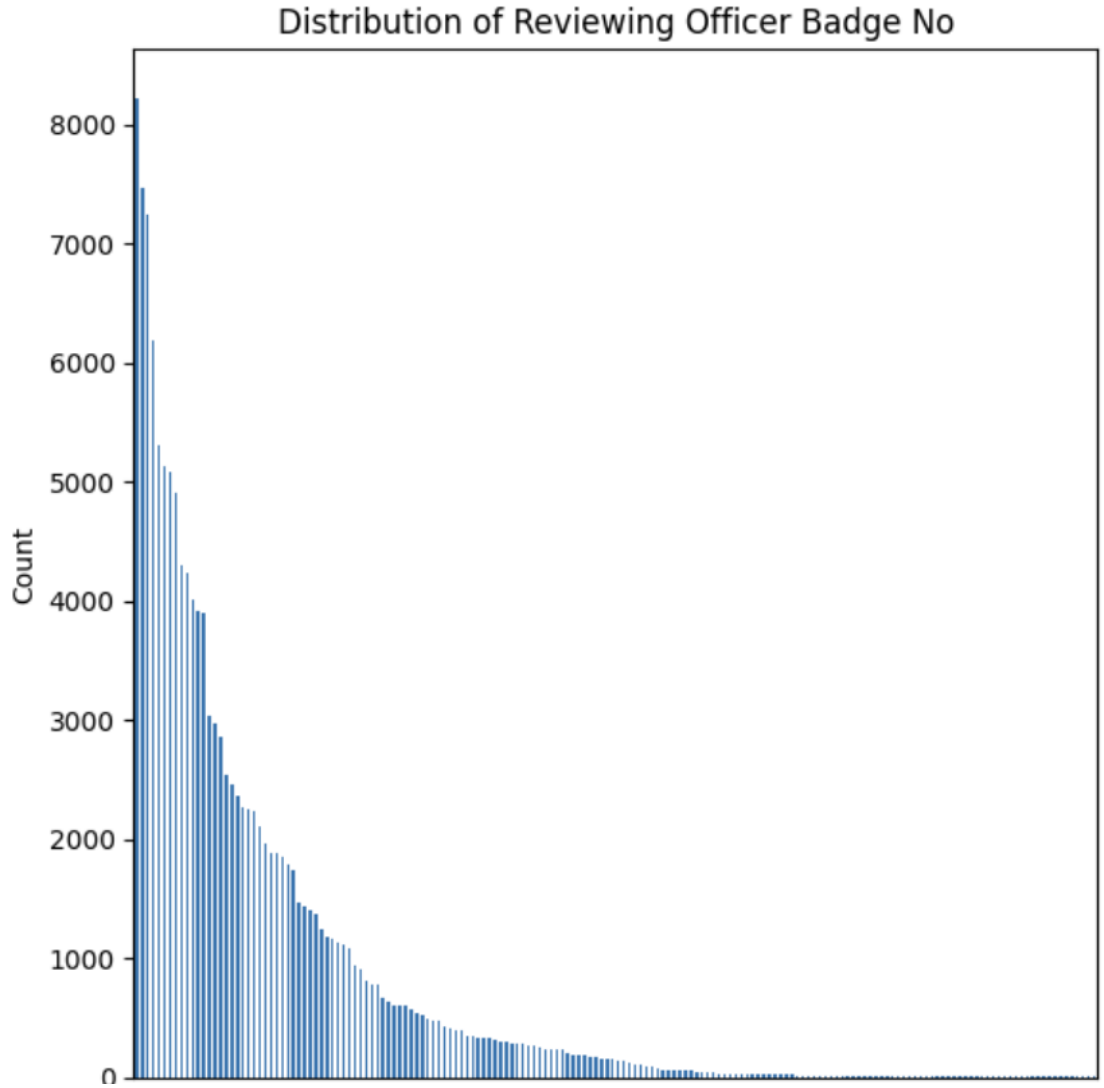
53 Assisting Officer Badge No

```
Attribute: Assisting Officer Badge No: has 404 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: 'T168': 5860 instances (4.21%)
  Least common: '8374': 1 instances (0.00%)
```



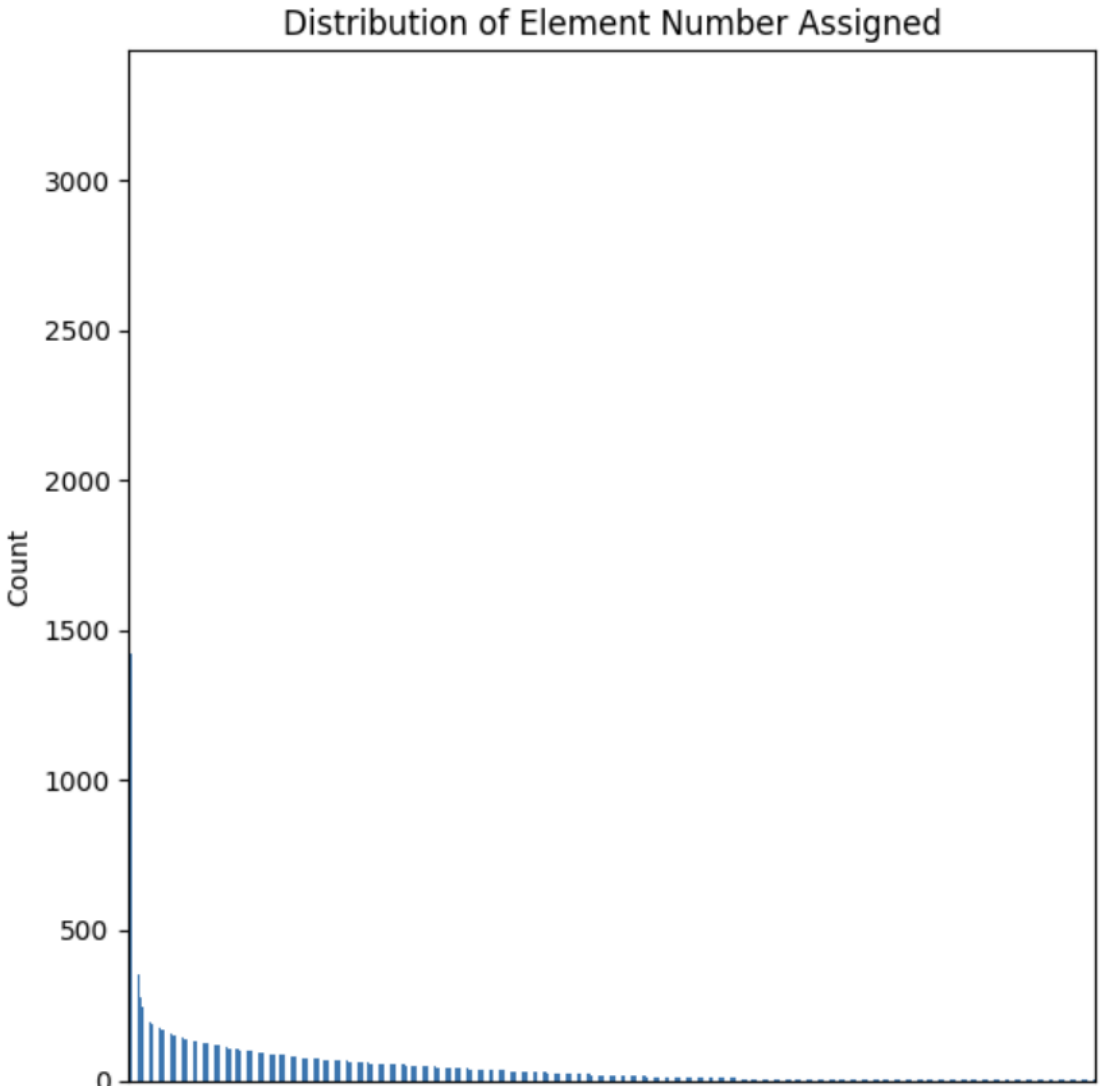
54 Reviewing Officer Badge No

```
Attribute: Reviewing Officer Badge No: has 172 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: '81075.0': 8221 instances (5.91%)
  Least common: 'M262': 1 instances (0.00%)
```



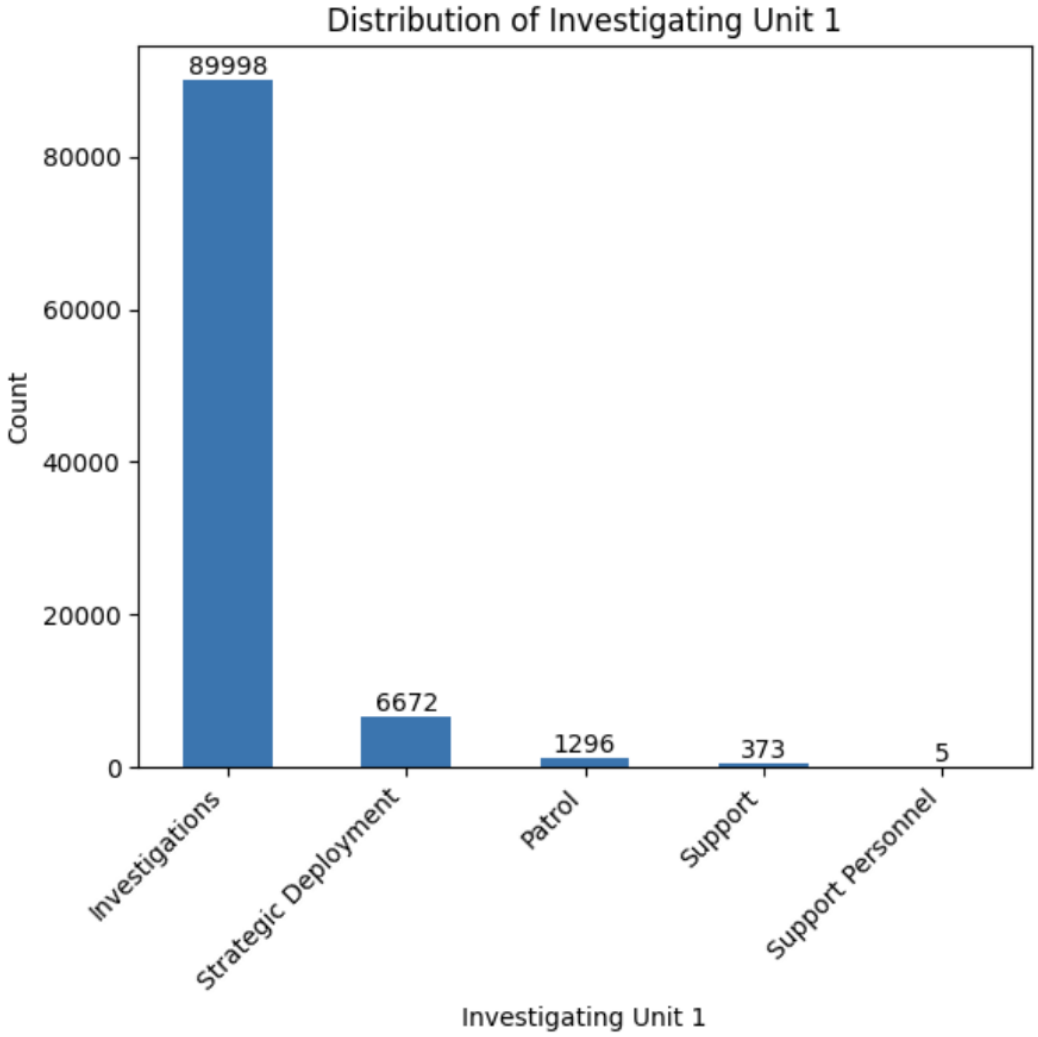
55 Element Number Assigned

```
Attribute: Element Number Assigned: has 2616 distinct values
  (Too many unique values to show detailed breakdown)
  Most common: 'EX10': 3269 instances (2.35%)
  Least common: 'A237': 1 instances (0.00%)
```



56 Attribute: Investigating Unit 1: has 5 distinct values

```
Attribute: Investigating Unit 1: has 5 distinct values
  'Investigations': 89998 instances (64.69%)
  'Strategic Deployment': 6672 instances (4.80%)
  'Patrol': 1296 instances (0.93%)
  'Support': 373 instances (0.27%)
  'Support Personnel': 5 instances (0.00%)
```



57 – Investigating Unit 2

46 distinct values (including missing) with the most common units are investigating vehicle or property crimes.

Row No.	Investigating Unit 2	count(Investigating Unit 2) ↓
43	Special Investigations / Auto Theft	32012
35	Property Crime Division / NW Property Crimes	8356
4	Capers / Assaults	7856
33	Property Crime Division / NC Property Crimes	7323
38	Property Crime Division / SW Property Crimes	6474
20	Field Services / Vehicle Crimes Unit	6129
34	Property Crime Division / NE Property Crimes	5283
32	Property Crime Division / CE Property Crimes	5133
36	Property Crime Division / SC Property Crimes	4745

58 – Offense Status

7 distinct values (including missing) with most cases being in suspended status.

Row No.	Offense Status	count(Of... ↓
7	Suspended	113174
2	Clear by Arrest	15020
5	Open	4318
4	Closed/Cleared	3066
3	Clear by Exceptional Arrest	2813
6	Returned for Correction	3

59 – UCR Disposition

10 distinct values (including missing) with most cases being in suspended status.

Row No. ↓	UCR Disposition	count(UCR D...
10	Suspended	113221
9	Open	4334
8	Closed	3071
7	CBEA (Under Age 17)	18
6	CBEA (Over Age 17)	2824
5	CBEA (Age 17)	32
4	CBA (Under 17)	13
3	CBA (Over Age 17)	14690
2	CBA (Age 17)	266

60 – Modus Operandi

67598 distinct values (including missing) with the most common modus operandi is ‘CRIMINAL TRESPASS WARNING’. Some entries looks similar to each other as shown below.

Row No.	Modus Operandi (MO)	count(M... ↓
12294	CRIMINAL TRESPASS WARNING	1525
13632	FOUND PROPERTY	1383
12493	CT WARNING	710
38536	UNEXPLAINED DEATH	639
52643	UNK SUSP TOOK COMPS VEHICLE WITHOUT CONSENT	610
66832	UUMV	594
51859	UNK SUSP TOOK COMP'S VEHICLE WITHOUT CONSENT	502
15914	LOST PROPERTY	446
16332	NATURAL DEATH	441

61 – Family Offense

2 distinct values (including missing) with every single value are false. This is supposed to be binary but there is no case considered to be family offense in the dataset.

Row No.	Family Offense	count(Fa... ↓
2	false	123163

62 – Hate Crime

3 distinct values (including missing) with the most common value are ‘yes’. However, the vast majority actually have missing data therefore it can be assumed they are actually NOT hate crimes.

Row No.	Hate Crime	count(Ha... ↓
3	Yes	81
2	No	19

63 – Hate Crime Description

12 distinct values (including missing) with the most common hate crime are against ‘Anti Asian/Pacific Islander’.

Row No.	Hate Crime Description	count(Ha... ↓
2	Anti Asian/Pacific Islander	18
4	Anti Homosexual (Gays and Lesbians)	5
3	Anti Black Or African American	4
6	Anti Male Homosexual (Gay)	4
7	Anti Multi-Racial Group	2
5	Anti Jewish	1
8	Anti Multi-Religious Group	1
9	Anti Other Ethnicity/Natl Origin	1
10	Anti Other Religion	1
11	Anti Transgender	1
12	Anti White	1

64 – Weapon Used

12 distinct values (including missing) with the most common weapons are firearm related or using their own body parts (considered to be a weapon).

Row No.	Weapon Used	count(Weapon Used) ↓
14	Handgun	5331
23	Personal Weapons (Hands-Feet ETC)	3017
13	Firearm (Type Not Stated)	1338

65 - Gang Related Offense

5 distinct values (including missing) with vast majority being UNK/unknown.

Row No.	Gang Related Offense	count(Gang Related Offense) ↓
5	UNK	15166
2	G	101
3	J	12
4	No	1

66 – Drug Related Istevencident

5 distinct values (including missing) with most cases are not related to drugs.

Row No.	Drug Related Istevencident	count(Dr... ↓
3	No	110909
4	UNK	7127
5	Yes	5098
2	NA-99999999-W1	1

67 - RMS Code

559 distinct values (including missing) with most common value is FS-24110003-G13 which is related to Texas Department of Public Safety criminal classification of UNAUTH USE OF VEHICLE.

Row No.	RMS Code	count(R... ↓
239	FS-24110003-G13	13613
240	FS-24110003-G14	9113
316	MA-22990004-F1	8059
507	NA-99999999-MSC11	5949
320	MA-22990011-F287	5198

68 - Criminal Justice Information Service Code

361 distinct values (including missing) with most common value being 99999999 which is related to ‘unknown’ value in the CJIS code.

Row No.	Criminal Justice Information Service Code	count(Cr... ↓
360	99999999	40747
79	24110003	23638
44	22990004	8192
47	22990011	5258
143	29990042	4974

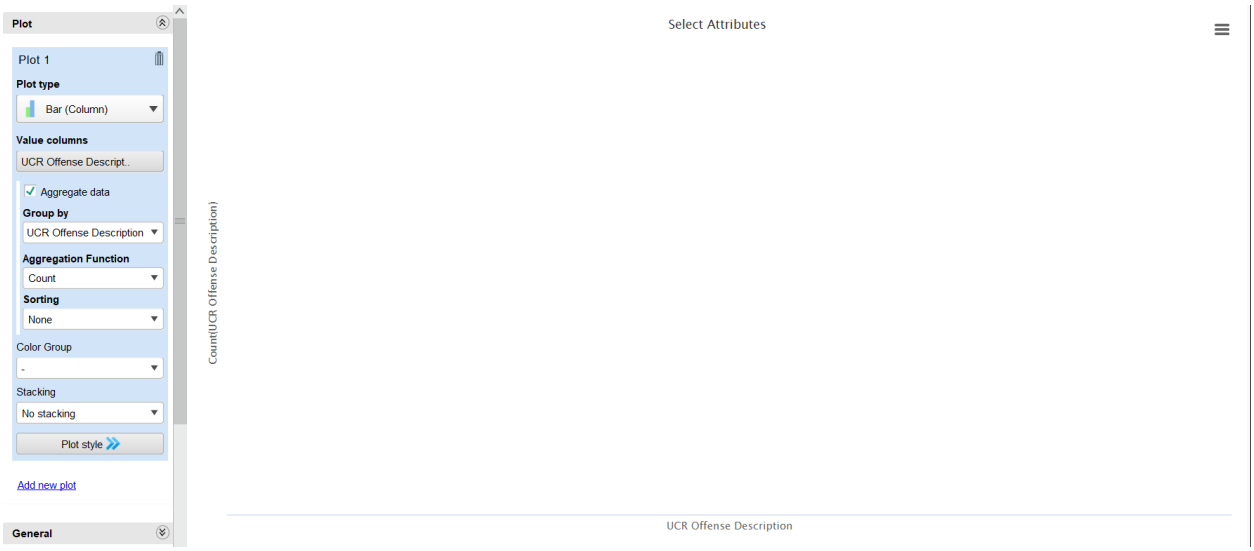
69 - Penal Code

382 distinct values (including missing) with the most common value is PC 31.07 which is related to the Texas penal code crime of UNAUTH USE OF VEHICLE.

70 - UCR Offense Name

It only has missing values.

71 - UCR Offense Description



This only has missing values as shown above, hence the empty graph

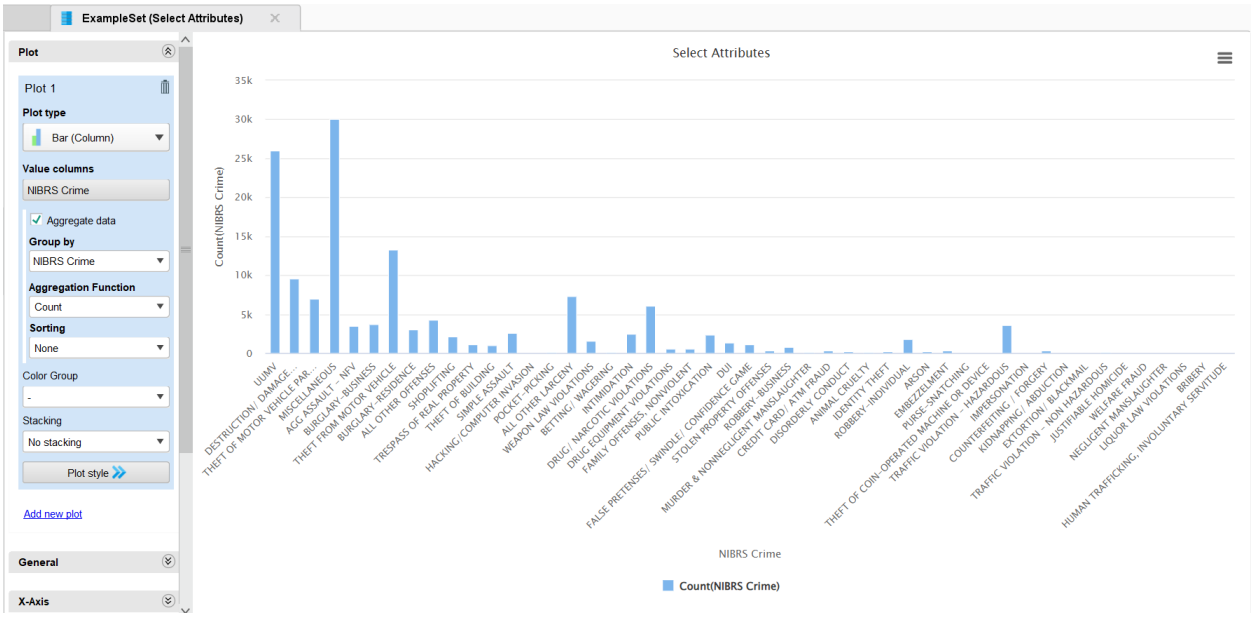
72 - UCR Code

Like UCR Offense Name and Description, this also only has missing values

73 - Offense Type

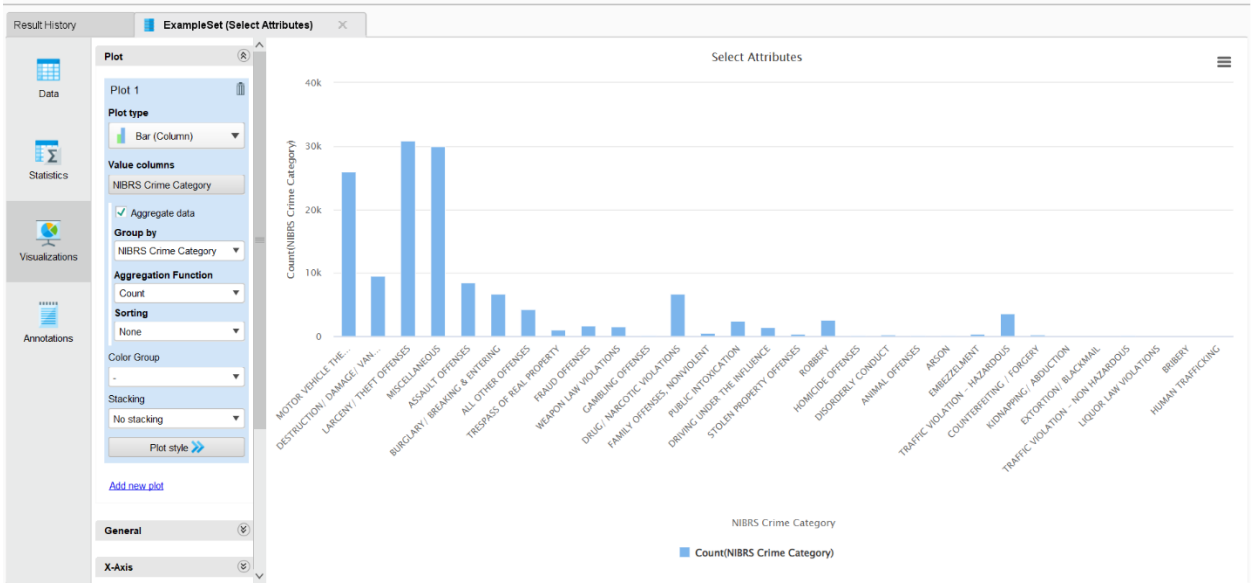
This has 99% missing values, with only one value as shown above

74 - NIBRS Crime



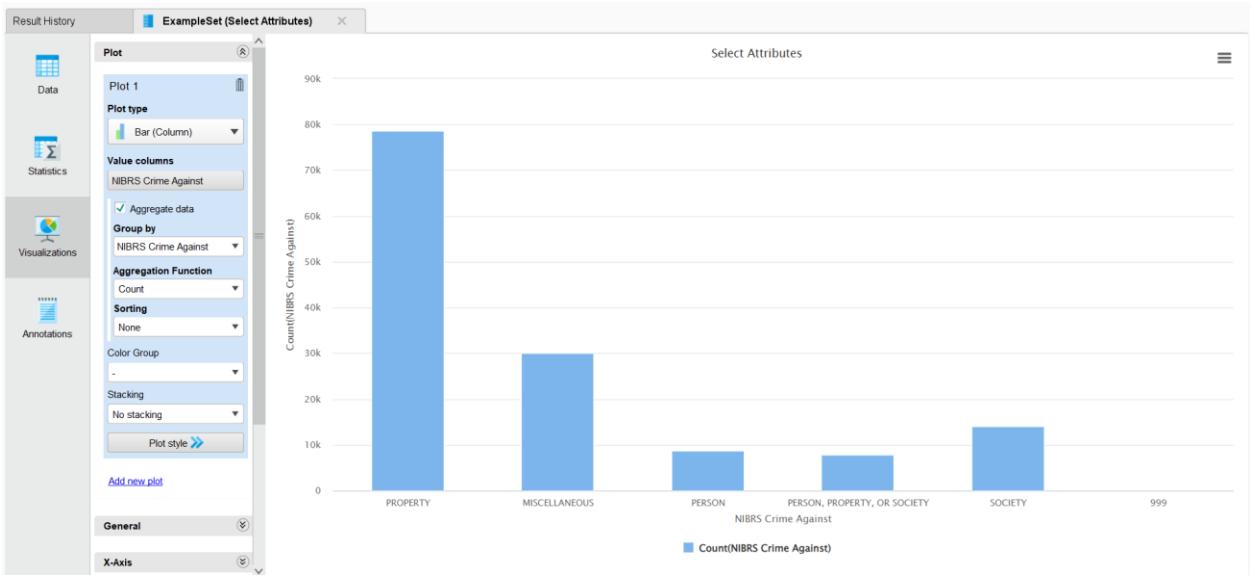
The NIBRS Crime represents crimes under the National Incident-Based Reporting System (NIBRS). UUMV (Unauthorized Use of a Motor Vehicle) and MISCELLANEOUS have the highest count, with ANIMAL CRUELTY, ARSON etc having much lower counts

75 - NIBRS Crime Category



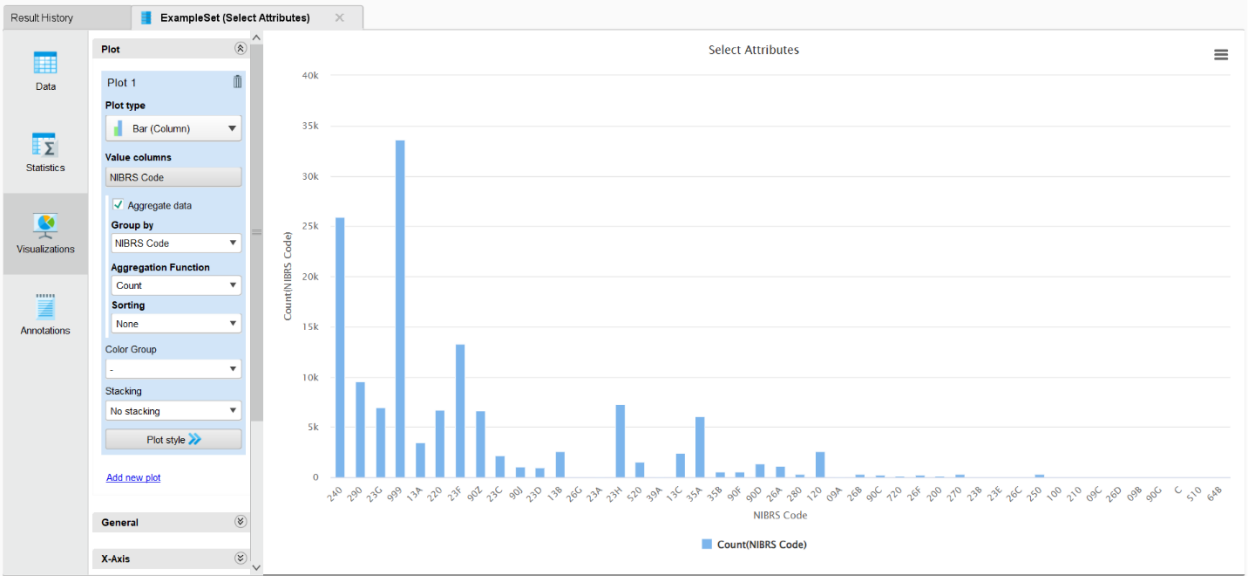
This has similar values to the NIBRS Crime attribute with ANIMAL OFFCENSES and ARSON appearing having low values here as well.

76 - NIBRS Crime Against



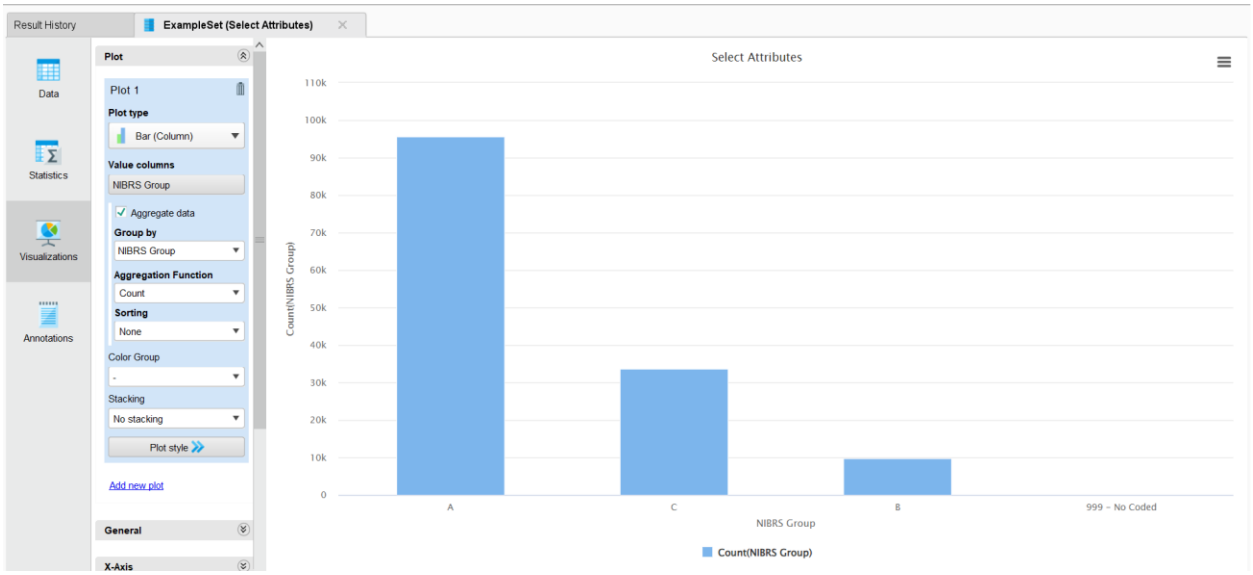
This shows the nature of the victims of crime according to NIBRS. As shown above, crimes against PROPERTY have the highest frequency, with person-based crimes appearing less often

77 - NIBRS Code



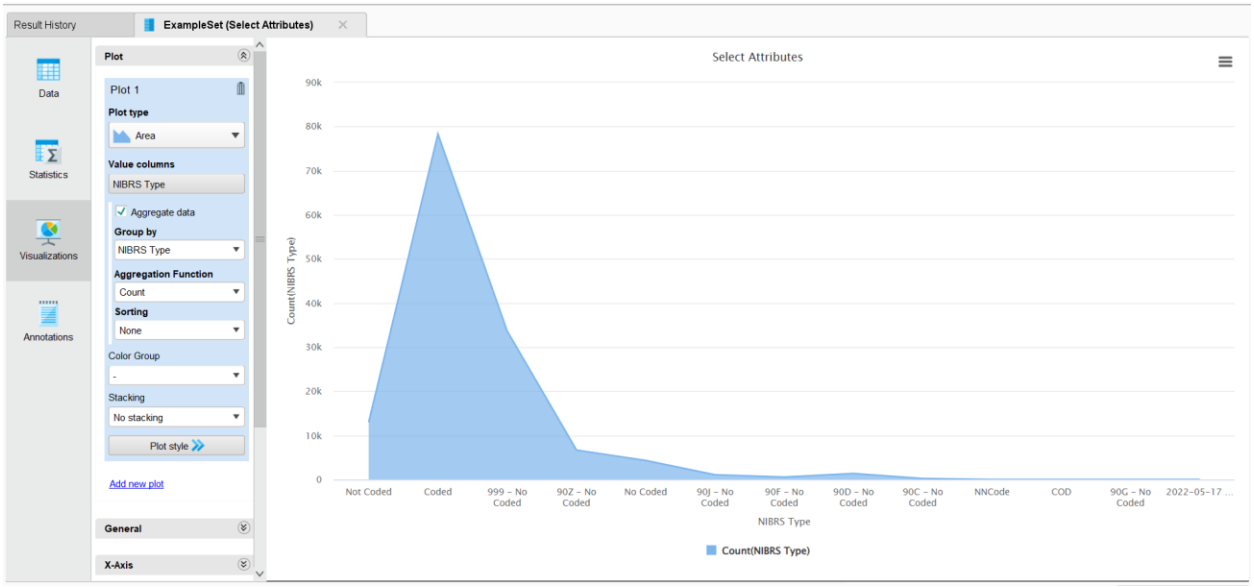
The NIBRS codes correspond to specific crime classifications under the National Incident-Based Reporting System (NIBRS) with 999 (Unknown or Miscellaneous Crime) having the highest frequency – this is like both NIBRS Crime Category and NIBRS Crime (having MISCALLANEOUS with the highest occurrence)

78 - NIBRS Group



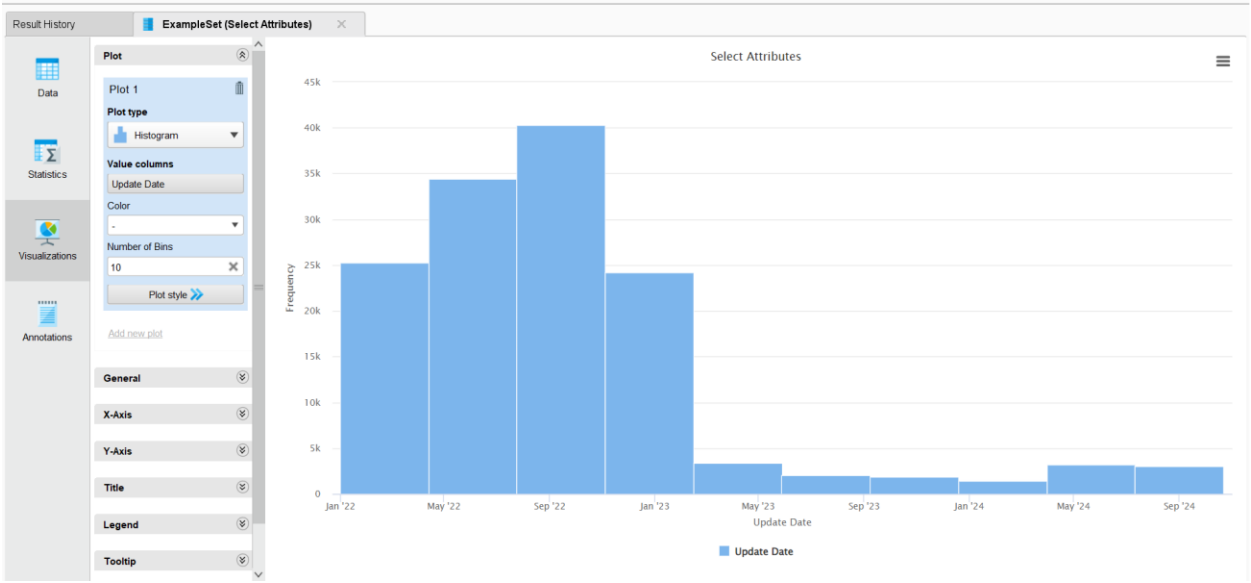
NIBRS offenses are organized into Groups A (more severe) and B (less severe), which categorize crimes based on their severity and the type of reporting required. C groups could sometimes be a placeholder for administrative or miscellaneous non-criminal incidents. In the above, A has the highest frequency, which implies that higher-severity crimes are more recurring

79 - NIBRS Type



This area graph shows that offenses that fall into a code (hence having the ”Coded” value) have the highest frequency

80 - Update Date

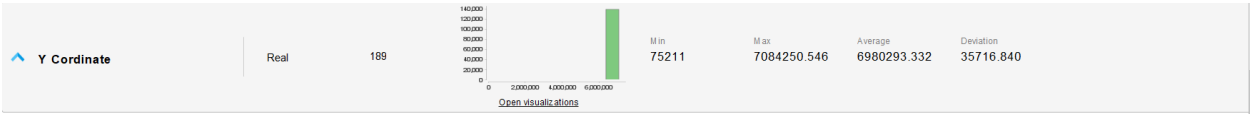


This shows that most crimes were updated in the September 2022, and least updates happened in January 2024

81 - X Coordinate

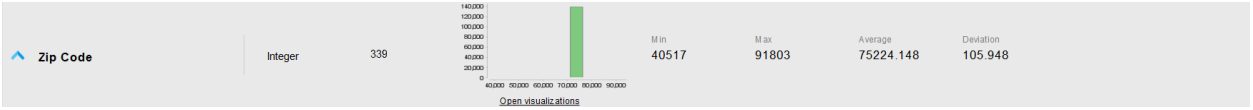
This is a numeric attribute that ranges between 2,400,000 and 6,700,00

82 - Y Cordinate



This is also a numeric attribute that ranges between 75,000 and 7,000,000. Which is a wider range than the X Coordinate

83 - Zip Code



This ranges between 40,000 and 92,000 with the average being 75224

84 – City

Most crimes happened within DALLAS (which is expected, being the case study area), and the least being from WAXAHACHIE

85 State

86 Location1

3.4 Verify Data Quality

1 Attribute: ‘Incident Number w/year’ - Less than 10% of missing data.

2 Attribute: ‘Year of Incident’ - Less than 10% of missing data.

3 Attribute: ‘Service Number ID’ - Less than 10% of missing data.

4 Attribute: ‘Watch’ - Less than 10% of missing data, discrepancy in the values, both string and integer values are present.

5 Attribute: ‘Call (911) Problem’ - Less than 50% of missing data.

6 Attribute: ‘Type of Incident’ - Less than 10% of missing data.

7 Attribute: ‘Type Location’ - Less than 10% of missing data.

8 Attribute: ‘Type of Property’ - More than 50% of missing data.

9 Attribute: ‘Incident Address’ - Less than 10% of missing data.

10 Attribute: ‘Apartment Number’ - More than 50% of missing data.

11 Attribute: ‘Reporting Area’ - Less than 10% of missing data.

12 Attribute: ‘Beat’ - Less than 10% of missing data.

13 Attribute: ‘Division’ - Less than 10% of missing data, duplicate values present.

14 Attribute: ‘Sector’ - Less than 10% of missing data.

15 Council District – More than 10 % missing data

16 Target Area Action Grids - More than half missing values

17 Community - More than half missing values

18 Date1 of Occurrence - No missing values

19 Year1 of Occurrence - Redundant

20 Month1 of Occurrence - Redundant

21 Day1 of the Week - Redundant

22 Time1 of Occurrence – No Missing values

23 Day1 of the Year – Redundant

24 Date2 of Occurrence - 73.9 % are the same values as Date1 of Occurrence

25 Year2 of Occurrence - 73.9 % are the same values as Date1 of Occurrence

26 Month2 of Occurrence - 73.9 % are the same values as Date1 of Occurrence

27 Day2 of the Week - 73.9 % are the same values as Date1 of Occurrence

28 Time 2 of Occurrence - 73.9 % are the same values as Date1 of Occurrence

29 Attribute ‘Daye 2 of the year’: missing 14 instances (0.1%)

30 Attribute ‘Date of Report’: missing 0 instances (0%).

31 Attribute ‘Date Incident created’: missing 0 instances (0%).

32 Attribute ‘offense entered year’: missing 0 instances (0%).

33 Attribute ‘offence entered month’ : missing 0 instances (0%).

34 Attribute ‘offence entered day of the week’ : missing 0 instances (0%).

35 Attribute ‘offence entered time’: missing 0 instances (0%).

36 Attribute ‘CSF Number’: missing 15981 instances (11.5%).

37 Attribute ‘Date Received date time’: missing 15981 instances (11.5%).

38 Attribute ‘call date time’: missing 15981 instances (11.5%).

39 Attribute ‘call date time’: missing 15981 instances (11.5%).

40 Attribute ‘call cleared date time’: missing 16018 instances (11.5%)

41Atrribute ‘call dispatch date time ‘: missing 16010 instances (11.5%).

42Attribute ‘special report(pre-Rms): missing 138835 instances (99.8).

43 Attribute 'Person Involvement Type': missing 5196 instances (3.74%)

44 Attribute 'Victim Type': missing 5224 instances (3.76%)

45 Attribute 'Victim Race': missing 51543 instances (37.05%)

46 Attribute 'Victim Ethnicity': missing 51748 instances (37.20%)

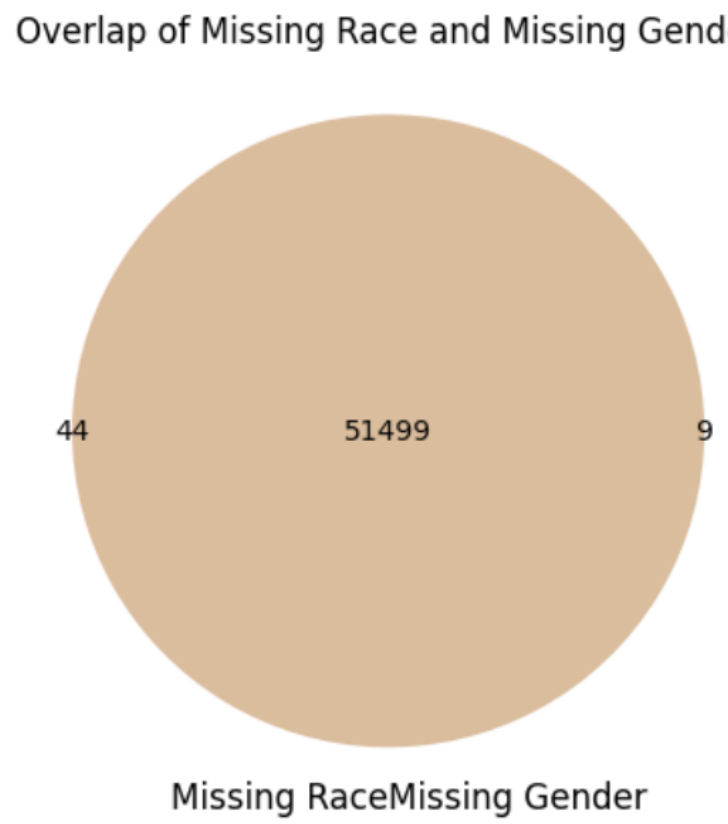
47 Attribute 'Victim Gender': missing 51508 instances (37.03%)

Both “Victim Race” and “Victim Gender” has the similar form of missing data:

1. Instances where both race and gender are missing:
Count: 51499
Percentage: 37.02%

2. Instances where only race is missing:
Count: 44
Percentage: 0.03%

3. Instances where only gender is missing:
Count: 9
Percentage: 0.01%



```
Group by Person Involvement Type:
Victim: 45103 instances (87.58%)
Registered Owner: 1191 instances (2.31%)
Owner: 10 instances (0.02%)


Victim Type:
Business: 20088 instances (39.01%)
Society/Public: 13926 instances (27.04%)
Government: 11247 instances (21.84%)
Individual: 671 instances (1.30%)
Religious Organizati: 241 instances (0.47%)
Financial Institutio: 123 instances (0.24%)
```

48 Attribute 'Responding Officer #1 Badge No': missing 16121 instances (11.59%)

49 Attribute 'Responding Officer #1 Name': missing 16210 instances (11.65%)

50 Attribute 'Responding Officer #2 Badge No': missing 97585 instances (70.15%)

51 Attribute 'Responding Officer #2 Name': missing 97610 instances (70.17%)

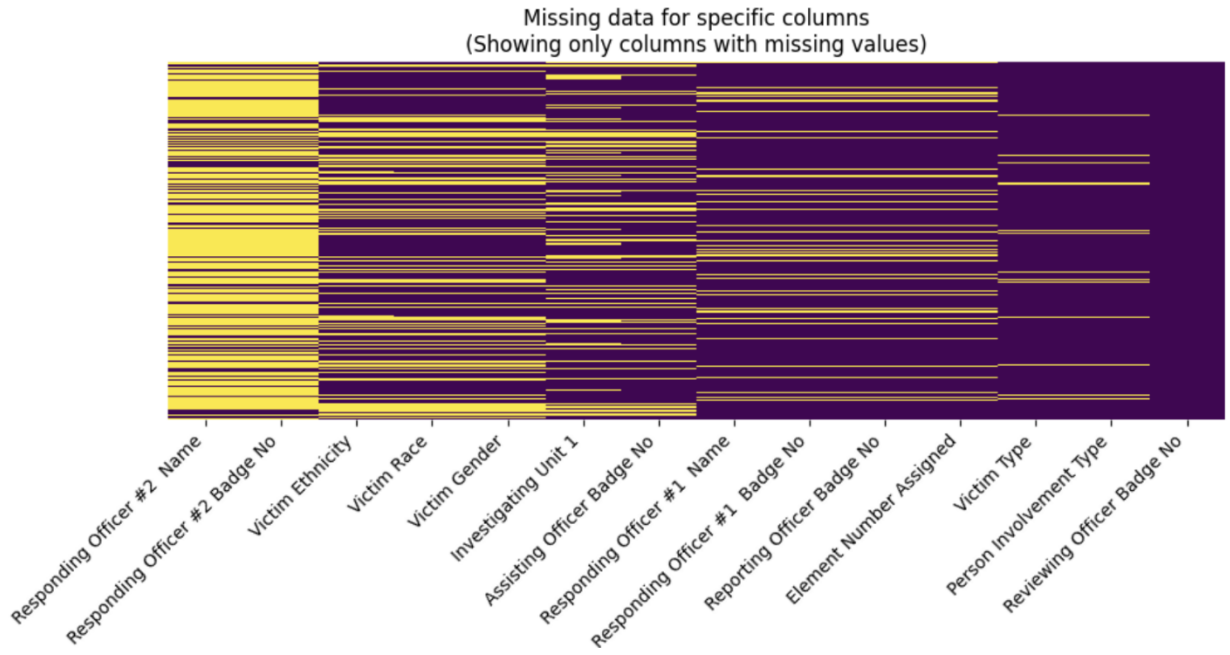
52 Attribute 'Reporting Officer Badge No': missing 16117 instances (11.59%)

53 Attribute 'Assisting Officer Badge No': missing 35715 instances (25.67%)

54 Attribute 'Reviewing Officer Badge No': missing 28 instances (0.02%)

55 Attribute 'Element Number Assigned': missing 15861 instances (11.40%)

56 Attribute 'Investigating Unit 1': missing 40768 instances (29.31%)



57 – Attribute ‘Investigating Unit 2’: missing 40764 instances (29.30%)

58 – Attribute ‘Offense Status’: missing 715 instances (0.51%)

59 – Attribute ‘UCR disposition’: missing 640 instances (0.46%)

60 – Attribute ‘Modus Operandi (MO)’: missing 16089 instances (11.56%)

61 – Attribute ‘Family Offense’: missing 15946 instances (11.46%)

62 – Attribute ‘Hate Crime’: missing 139009 instances (99.93%)

63 – Attribute ‘Hate Crime Description’: missing 139070 instances (99.97%)

64 – Attribute ‘Weapon Used’: missing 127009 instances (91.30%)

65 – Attribute ‘Gang Related Offense’: missing 123829 instances (89.02%)

66 – Attribute ‘Drug Related Istevencident’: missing 15974 instances (11.48%)

67- Attribute ‘RMS Code’: no missing instance (0%)

68 – Attribute ‘Criminal Justice Information Service Code’: missing 1 instance (< 0.001%)

69 – Attribute ‘Penal Code’: missing 1 instance (< 0.001%)

70 – Attribute ‘UCR Offense Name’: all missing (100%)

71 – Attribute ‘UCR Offense Description’: all missing (100%)

72 – Attribute ‘UCR Code’: all missing (100%)

73 – Attribute ‘Offense Type’: 139108 missing values (99.999%)

74 – Attribute ‘NIBRS Crime’: 14 missing values (0.01%)

75 – Attribute ‘NIBRS Crime Category’: 14 missing values (0.01%)

76 – Attribute ‘NIBRS Crime Against’: 14 missing values (0.01%)

77 – Attribute ‘NIBRS Code’: 14 missing values (0.01%)

78 – Attribute ‘NIBRS Group’: 14 missing values (0.01%)

79 – Attribute ‘NIBRS Type’: 0 missing values (0%)

80 – Attribute ‘Update Date’: 1 missing value (0.0007%)

81 – Attribute ‘X Coordinate’: 189 missing values (0.136%)

82 – Attribute ‘Y Cordinate’: 189 missing values (0.136%)

83 – Attribute ‘Zip Code’: 339 missing values (0.244%)

84 – Attribute ‘City’: 324 missing values (0.233%), with duplicate values like “dallas” and “DALLAS”, which logically mean the same thing

85 State

86 Location1

4 Data Preparation

4.1 Select Data

Suvitha: Attributes Type Location and Division are selected.

Alex: Attributes Date and Time of Occurrance

Max: Attribute CFS Number is selected.

Thao: Attributes Victim Race and Victim Gender are selected

Alvin: Attribute UCR Disposition is selected. The other attributes have data quality problems or better attributes can be chosen.

David: Attributes NIBRS Group, X Coordinate, Y Coordinate and Zip Code are selected

4.2 Clean Data

Suvitha: Dropping rows in ‘Division’ and ‘Type Location’ since only 10% of the data is missing.

Alex: No need for cleaning

Max: The Incident Created attribute is in a string format (2022-01-27T16:04:47-05:00) and may contain missing or invalid entries.

Thao:

Filling missing data for 'Victim Race' and 'Victim Gender'

(Attribute 'Victim Race': missing 51543 instances (37.05%) Attribute 'Victim Gender': missing 51508 instances (37.03%))

Suggesting to fill missing data for Race as “Unknow” for Gender: 50% Male, 50% Female.

Alvin: There is 0.46% missing data which may be replaced with “Unknown” for the UCR Disposition attribute.

David:

Zip Code - 339 missing values – were replaced with “00000”, which was further binned as “Unknown”. Further recommendations for how this attribute could have been prepared is outlined in 6.4 (Determine next steps)

NIBRS Group – Replaced missing values with most recurrent - "A". A single instance also had a value “999 - No Coded”, which was replaced with “C”

X Coordinate – Replaced missing values with the code “0" which was further processed in the clustering step

Y Coordinate – Similar to X Coordinate, replaced with the same code

4.3 Construct Data

Suvitha: ‘Type Location’ has 71 unique values, which is mapped to 9 general values and a new attribute ‘Crime Scene’ is created. Remove Nan values. For Classification, from ‘Date of Occurrence’ feature, new features ‘Date’, ‘Month’, ‘DayofWeek’, ‘Hour’ and ‘TimeOfDay’ are created.

Alex: Construction of Datetime attribute where Date and Time are joint

Max: new column (multiple crimes at the same time) it will be generated from CFS number column

Thao:

Attribute: “Victim Race” has 10 distinct values, should be mapped to more general and NIBRS-standard values

Mapping from

Victim Race: 10 distinct values
'Black': 28337 instances (20.37%)
'White': 28165 instances (20.25%)
'Hispanic or Latino': 27188 instances (19.54%)
'Asian': 1974 instances (1.42%)
'Middle Eastern': 796 instances (0.57%)
'Unknown': 716 instances (0.51%)
'American Indian or Alaska Native': 177 instances (0.13%)
'Native Hawaiian/Pacific Islander': 167 instances (0.12%)
'NH': 44 instances (0.03%)
'H': 5 instances (0.00%)

To Valid Data Values for Race:

W = White

B = Black or African American

I = American Indian or Alaska Native

A = Asian

P = Native Hawaiian or Other Pacific Islander

U = Unknown

Alvin: No new data is constructed.

David: New columns (timeOfDay, dayOfWeek and Month) were generated from Date Time of Occurence attribute in classification data preparation

4.4 Integrate data

Suvitha:

Alex: not needed

Max: not necessarily

Thao: Not necessarily

Alvin: No outside data source is necessary for integration.

David: Unncessary

4.5 Format Data

Suvitha: ‘Division’ attribute has discrepancy, the values will be changed to all lowercase. ‘Crime Scene’ and ‘Division’ will be one-hot encoded. Out of 14 attributes, only ‘Crime Scene’ and ‘Division’ are selected, all the other columns will be dropped. Remove outliers if any.

Alex: Proper Date time Format and post format of timestamp for clustering

Max: CFS number column was you used to create a new column multiple crimes at the same time where it will be false if the CFS number is unique and true if its not.

Thao: Not necessarily.

Alvin: Binning is done for all the different CBA (cleared by arrest) values into just one CBA value.

David: Zip Codes were binned based on regional zones. These were the various categories: Downtown/Central Dallas, North Dallas, South Dallas, East Dallas, West Dallas, Suburbs and Outside Dallas Metro Area

5 Modeling

5.1 Select modeling techniques

Classification

- Decision Tree
  A decision tree was selected as a classification technique for predicting crime type according to its NIBRS group. The decision tree is particularly useful because it does not make assumptions about the specific form or distribution for the underlying data, which makes it more flexible in capturing complex relationships.

- KNN- KNN model does not make any assumptions about the underlying data distribution making it suitable for this dataset with complex relationships. It can also handle multi-class classification tasks.
- Random Forest- It provides high accuracy and robust predictions for classification tasks. It also reduces the risk of overfitting. It works very well with a large number of features.

Clustering

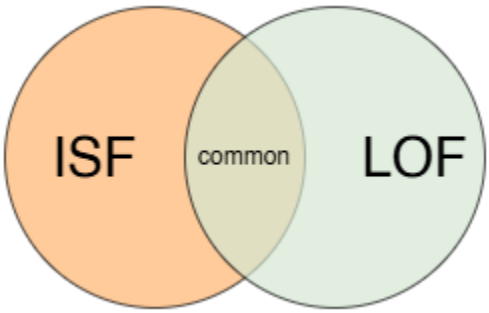
- k-Means:
- The K-means algorithm is a clustering technique used to divide a given dataset into K distinct clusters based on the similarity of data points. K-means aims to group data so that the points within each cluster are more similar to each other than to points in other clusters. K-means is used for several purposes such as: organizing data into groups (clustering), identifying underlying patterns in the data (pattern recognition), reducing the number of groups compared to the original data help to simplify further analysis (data simplification).

Outliers Detection

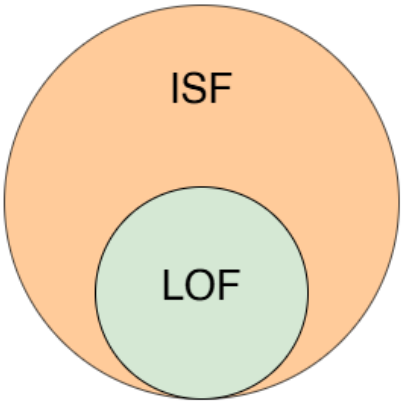
- Local Outlier Factor (LOF):
  - Identifies anomalies based on the density of data points in local neighbourhoods.
  - Sensitive to scaling; requires normalisation and one-hot encoding for categorical variables.

- Isolation Forest (ISF):
  - Detects anomalies by isolating data points using random partitions.
  - Less sensitive to scaling and can handle missing values, but preprocessing improves performance.

- Combination Approach:
  - Approach 1: First use LOF and ISF for determining list of outliers and find the common instances of those sets.



- Approach 2: Use ISF as an initial filter to reduce the candidate set, followed by LOF for more precise outlier detection. This method leverages the strengths of both algorithms to ensure robust detection.



5.2 Generate test design

Classification

-The Datetime of Occurrence column was processed to derive three new attributes:

- **Time of Day**: Categorized into morning, afternoon, evening, and night, to capture temporal patterns in crime occurrence.
- **Day of Week**: Encoded as categorical features to reflect weekly patterns in criminal activity.
- **Month**: Used to identify seasonal trends that may influence crime rates. This feature engineering step provided the decision tree with meaningful attributes that directly align with its ability to split data based on categorical thresholds. These features improve interpretability and enhance the model's capacity to detect patterns.

- NIBRS Group was set as target class
- Irrelevant columns were dropped – X Coordinate, Y Coordinate (Zip code suffices for our purposes) and temporary columns used during binning. For example, the Date and Hour columns. The initial Datetime of Occurrence attribute was also dropped
- One-hot encoding was applied to all categorical attributes – which was all attributes
- train_test_split was used to split the dataset into Train and Test which is 70% and 30% respectively.

Clustering

Several clustering methods with the K-Means algorithm were done in order to see how some attributes relate to each other.

We performed 4 clustering methods with different pairs of attributes with victim race/victim gender, x-coordinates/y-coordinates, crime scene/division, and datetime/area zone. An overall clustering is also done to see the ideal number of k with all the attributes together. The class attribute is removed before doing the overall clustering.

For data preparation, normalization is done on the numerical values to reduce outliers, and categorical values are converted into numerical values so that k-means can be performed.

Outliers

- Approach 1: Keep the categorical features
  - Data Preprocessing:
    - Removed irrelevant attributes (e.g., X and Y coordinates).
    - Handled missing data with imputation for LOF.
    - Datetime Features:
      - Converted Datetime of Occurrence into numeric features such as:
      - Year, Month, Day, Hour, Minute.
      - Quarter (e.g., Q1 for January-March).
      - Time of Day (e.g., afternoon, evening).
    - Binning Dates:
      - Grouped monthly data into quarters for trend analysis.
      - Categorized Time of Day into logical bins for better feature interpretation.
    - Encoding Features:
      - Applied one-hot encoding for categorical variables like Division, Victim Race, and Crime Scene.
    - Encoded categorical variables using one-hot encoding.

- Approach 2: Anomaly detection for Time Series
  - Data Preprocessing:
    - Construct new column: Total crime case that calculate the total number of incidents in a day
    - The Datetime of Occurrence is format to pandas date format in order to transformed into Date as Number later on.

- Removed all other features.
- Using StandardScaler for the dataset before applying on LOF as LOF is sensitive to the scale of features.

5.3 Build model

Classification

- Decision Tree

  A DecisionTreeClassifier was trained using Gini impurity as the criterion to optimize splits.
  The maximum depth was set to ensure model complexity did not lead to overfitting.

- KNN- KNeighborsClassifier was trained with K as 14 because of low error rate.

- Random Forest- RandomForestClassifier was first trained with 10 decision trees. To improve the accuracy, it was trained again with 200 decision trees with max_depth = 70, max_features= "sqrt". This improves generalization by preventing overfitting and making the forest more robust.

Clustering

k-means parameters:

Number of iterations: 15 tries were done to check the ideal k with values of k ranging between 1 to 15
Ideal k: different clustering methods turns to have different k values but most of them is 3
Plotting parameters:
Dot size: varies between 1 to 10
Jitter: added to make the clustering easier to see and the values varies

Outliers

- General setup
  - LOF:
    - Parameters:
      - Contamination: 0.05 (assumed 5% of the data are outliers).
      - Neighbors: 10 (default value for neighborhood size).

        - Outcome: Identified initial outliers based on local density differences.
    - ISF:
        - Parameters:
            - Contamination: 0.05.
            - Random State: 42 (ensures reproducibility).
    - Outcome: Detected outliers using partition-based techniques.

- Result:
    - Approach 1: Keep the categorical features
        - Found common outliers by combining results from both methods: 92 instances identified as common outliers across both algorithms.
        - Using ISF as input for LOF found 25 instances for outliers
    - Approach 2: Anomaly detection for Time Series
        - Found common outliers by combining results from both methods found 10 instances identified as common outliers across both algorithms.

5.4 Assess model

Classification

- Decision Tree

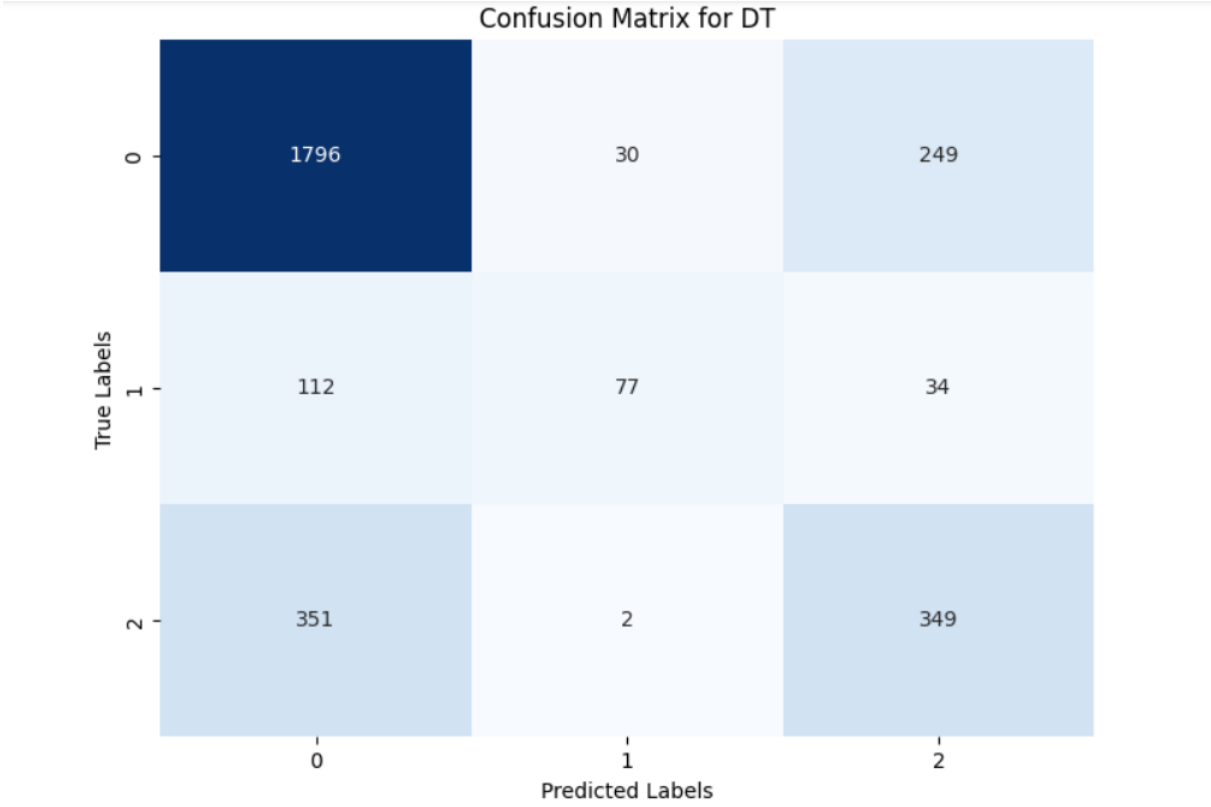
Classification Report

DT Classification report

	precision	recall	f1-score	support
A	0.80	0.87	0.83	2075
B	0.71	0.35	0.46	223
C	0.55	0.50	0.52	702
accuracy			0.74	3000
macro avg	0.68	0.57	0.61	3000
weighted avg	0.73	0.74	0.73	3000

As shown above, the model achieved a weighted average accuracy of 73%, with precision, recall, and F1-scores provided for each class (A, B, C). Class A showed the highest recall (0.87), while class B had the lowest (0.35), indicating difficulty in distinguishing certain crime types.

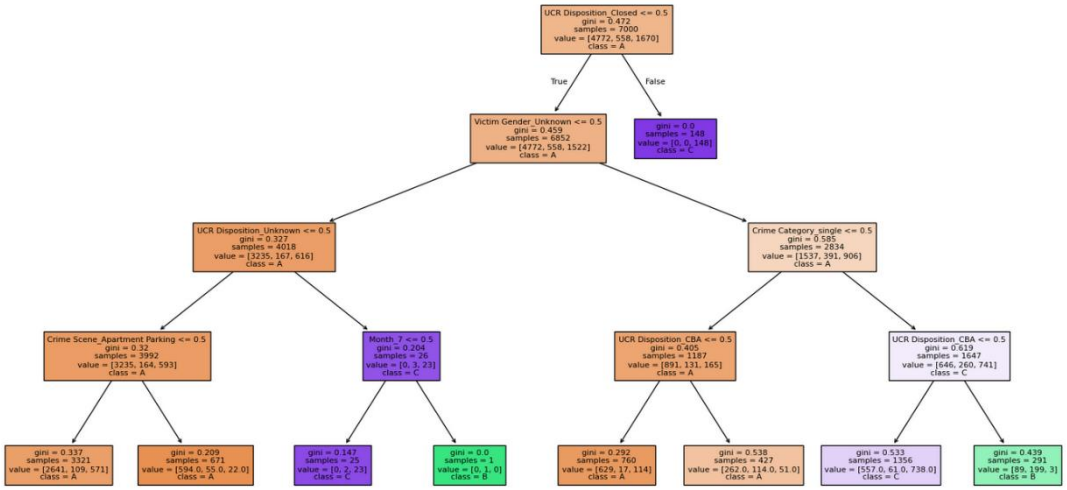
Confusion Matrix



The highlighted lighter areas show where the model struggled, such as misclassifications between class B (1) and C (2).

True positives for NIBRS Group A = 1796
True positives for NIBRS Group B = 77
True positives for NIBRS Group C = 349

Decision Tree Plot



Picture can be zoomed to better see features

Some rules from decision tree (from the left split on Victim Gender):
If UCR Disposition is not closed (false), the NIBRS group is C

If UCR Disposition is closed (true), Victim Gender is unknown (true), UCR Disposition is unknown (true), Crime Scene has apartment parking (true), the NIBRS group is A

If UCR Disposition is closed (true), Victim Gender is unknown (true), UCR Disposition is unknown (true), Crime Scene does not have apartment parking (false), the NIBRS group is still A

If UCR Disposition is closed (true), Victim Gender is unknown (true), UCR Disposition is not unknown (false), the month of crime occurrence is not July (7), the NIBRS group is B

If UCR Disposition is closed (true), Victim Gender is unknown (true), UCR Disposition is not unknown (false), the month of crime occurrence is July (7), the NIBRS group is C

Sample of varying the max depth and how it affects accuracy:

```
[ ]  # Print accuracy
     print("Accuracy with a max depth of 3:", accuracy_score(y_test, y_pred_3))
```

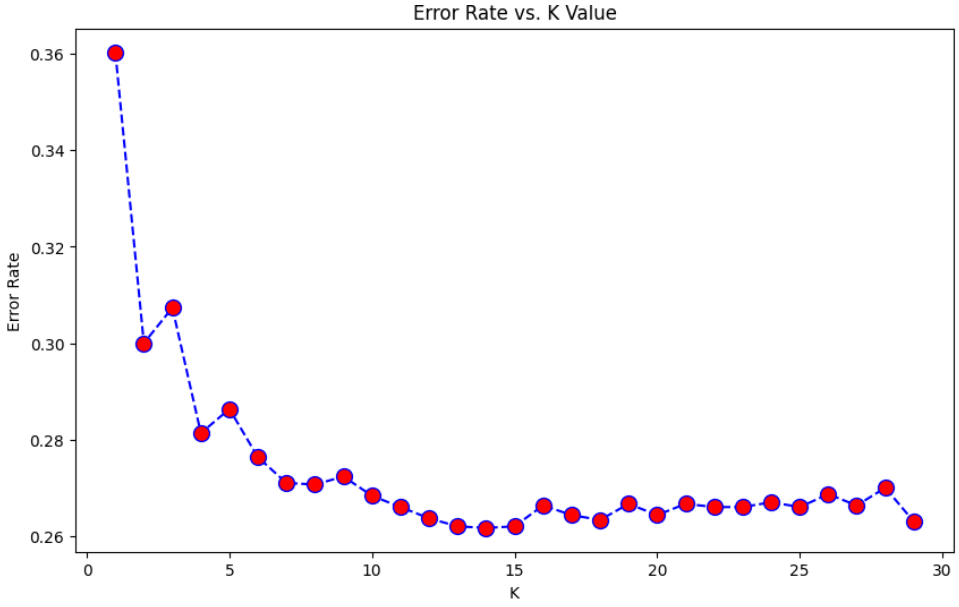
```
Accuracy with a max depth of 3: 0.7156666666666667
```

```
     # Print accuracy
     print("Accuracy with a max depth of 4:", accuracy_score(y_test, y_pred_dt))
```

```
Accuracy with a max depth of 4: 0.7406666666666667
```

- KNN

  K = 14, the error rate is at the lowest.



KNN Model Accuracy

```
KNN Model Accuracy for Training Data: 0.76
KNN Model Accuracy for Test Data: 0.74
```

KNN Classification Report

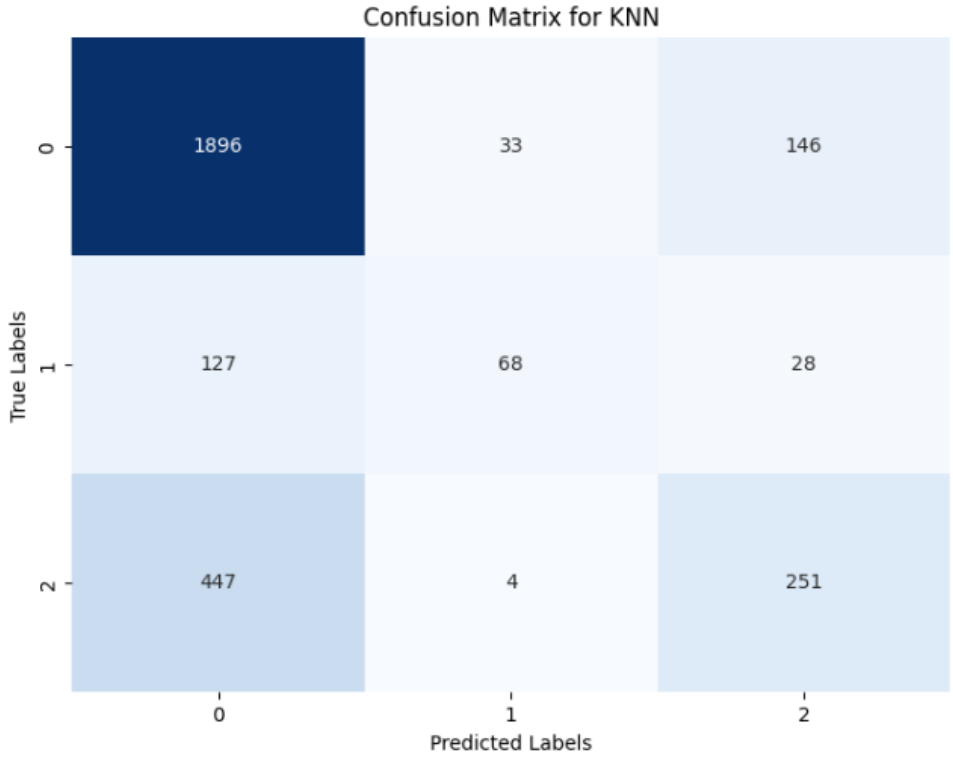
```
KNN Classification report
              precision    recall  f1-score   support

           A       0.77      0.91      0.83      2075
           B       0.65      0.30      0.41       223
           C       0.59      0.36      0.45       702

    accuracy                           0.74      3000
   macro avg       0.67      0.53      0.56      3000
weighted avg       0.72      0.74      0.71      3000
```

The overall performance of the KNN model shows a 74% accuracy, reflecting decent predictions overall but clearly struggles with minority classes (B and C). Class A (the majority class) achieves significantly better precision (0.77), recall (0.91), and F1-score (0.83).

Classes B and C, with fewer examples, show lower recall (0.30 and 0.36, respectively), indicating many actual instances of these classes are misclassified.

In summary, the model performs well for the majority class but struggles with minority classes.

KNN Confusion Matrix



Class A: High recall (91%) with 1896 true positives, indicating the model captures the majority of Class A instances effectively.
Class B: Low recall (30%) with only 68 true positives, suggesting significant misclassification or underrepresentation.
Class C: Moderate recall (36%) with 251 true positives, showing room for improvement in identifying this class.

- Random Forest

Random Forest Model Accuracy with 10 Decision Trees = 73%

```
Model accuracy score with 10 decision-trees : 0.73
```

Random Forest Model Accuracy with 200 Decision Trees = 77%

```
Model accuracy score with 200 decision-trees : 0.771
```

Random Forest Classification Report

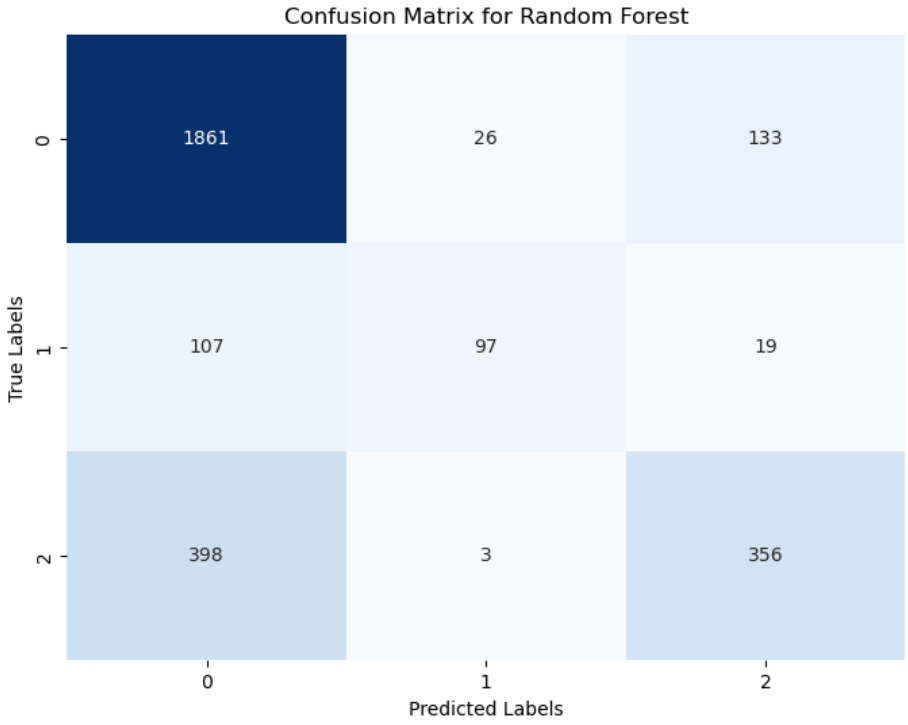
Random Forest Classification report

	precision	recall	f1-score	support
A	0.79	0.92	0.85	2020
B	0.77	0.43	0.56	223
C	0.70	0.47	0.56	757
accuracy			0.77	3000
macro avg	0.75	0.61	0.66	3000
weighted avg	0.76	0.77	0.75	3000

The Random Forest model achieves 77% accuracy, performing well on Class A but struggling with Classes B and C.

Class A: High performance with a precision of 0.79, recall of 0.92, and F1-score of 0.85.
Class B: Moderate precision (0.77) but low recall (0.43), leading to an F1-score of 0.56.
Class C: Precision of 0.70 and recall of 0.47, with an F1-score of 0.56.
In summary, the model performs well for the majority class but struggles with minority classes.

Random Forest Confusion Matrix



Class A has high recall (0.92), with 1861 true positives, showing that the model correctly identified most instances of this class.
Class B has a recall of 0.43, with 97 true positives, meaning many actual Class B instances were missed.
Class C has a recall of 0.47, with 356 true positives, indicating a moderate performance but still missing a significant portion of Class C instances.

Random Forest Feature Importance

	Feature	Importance
18	UCR Disposition_Closed	0.078515
17	UCR Disposition_CBA	0.077608
20	UCR Disposition_Suspended	0.069273
32	Crime Category_single	0.066465
16	Victim Gender_Unknown	0.058877
12	Victim Race_Unknown	0.047577
31	Crime Category_multiple	0.040460
33	Crime Category_unknown	0.023721
27	Crime Scene_Public Roads	0.022574
22	Crime Scene_Apartment Parking	0.021152
15	Victim Gender_Male	0.018467
63	TimeOfDay_Night	0.017927
62	TimeOfDay_Morning	0.017646
19	UCR Disposition_Open	0.016486
0	Division_central	0.016477
60	TimeOfDay_Afternoon	0.014438
30	Crime Scene_Stores	0.013635
36	Area Zone_North Dallas	0.013146
14	Victim Gender_Female	0.012982

Feature Relevance:

The UCR Disposition-related features are clearly the most influential, reflecting the importance of how crimes are categorized or handled in the system.

Demographic details like Victim Gender and Victim Race have varying levels of importance.

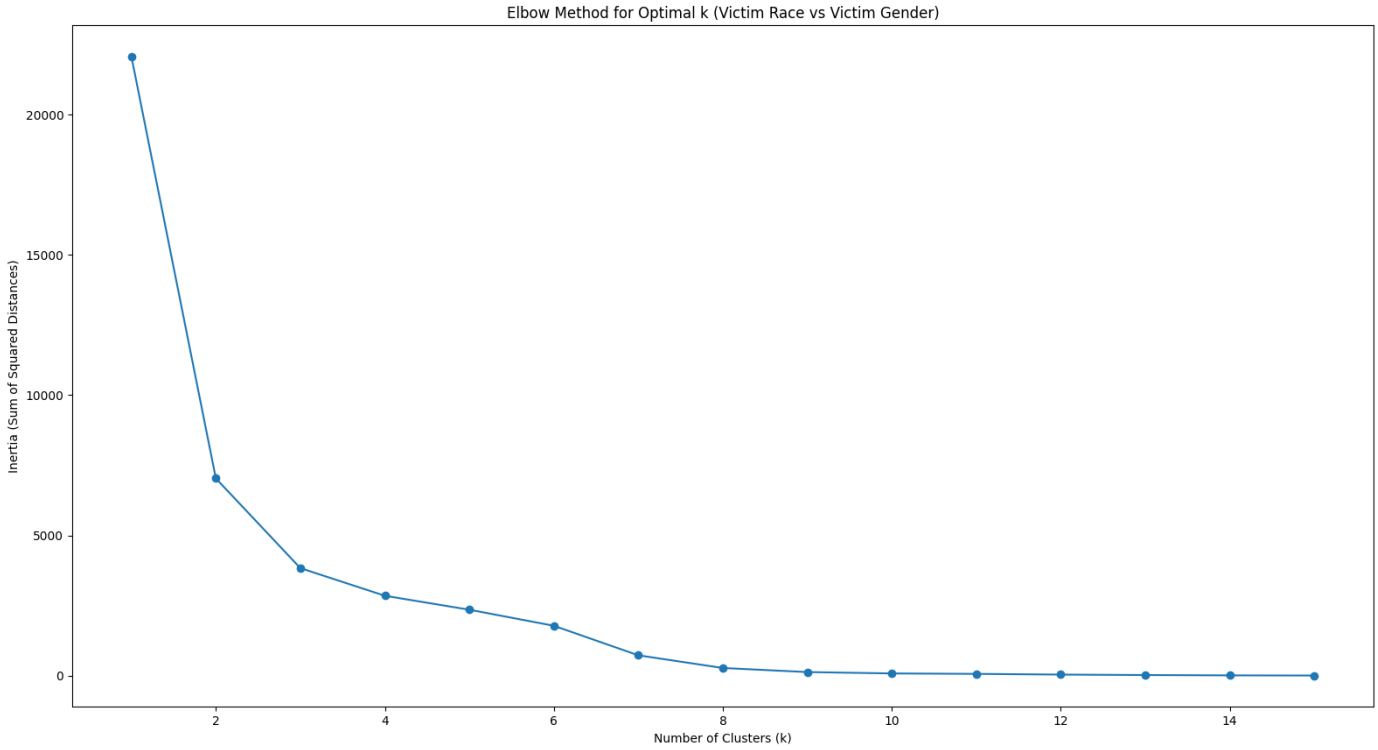
Features with "Unknown" values (e.g., Victim Gender_Unknown and Victim Race_Unknown) have relatively higher importance, which could indicate that missing or unknown data carry inherent predictive value possibly due to data imbalance.

Crime scene categories (e.g., Apartment Parking, Public Roads) and time-related features (e.g., TimeOfDay_Night) provide moderate importance, reflecting situational influences on predictions.

Day of the Week has low importance, implying weak patterns related to the day crimes occurrence.

Clustering

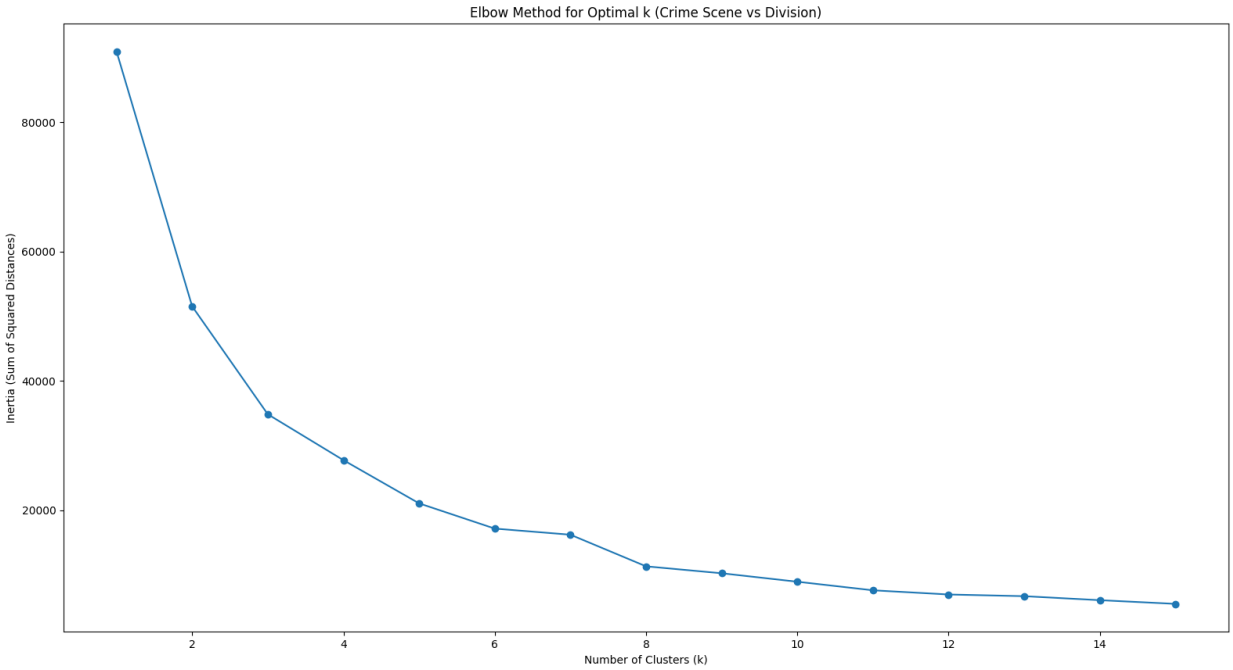
Optimal K for Race and Gender



By looking at the graph, the reduction of inertia can be seen to plateau when increasing the number of K to values larger than 3. The ideal K is considered to be 3.

CRIME SCENE AND DIVISION

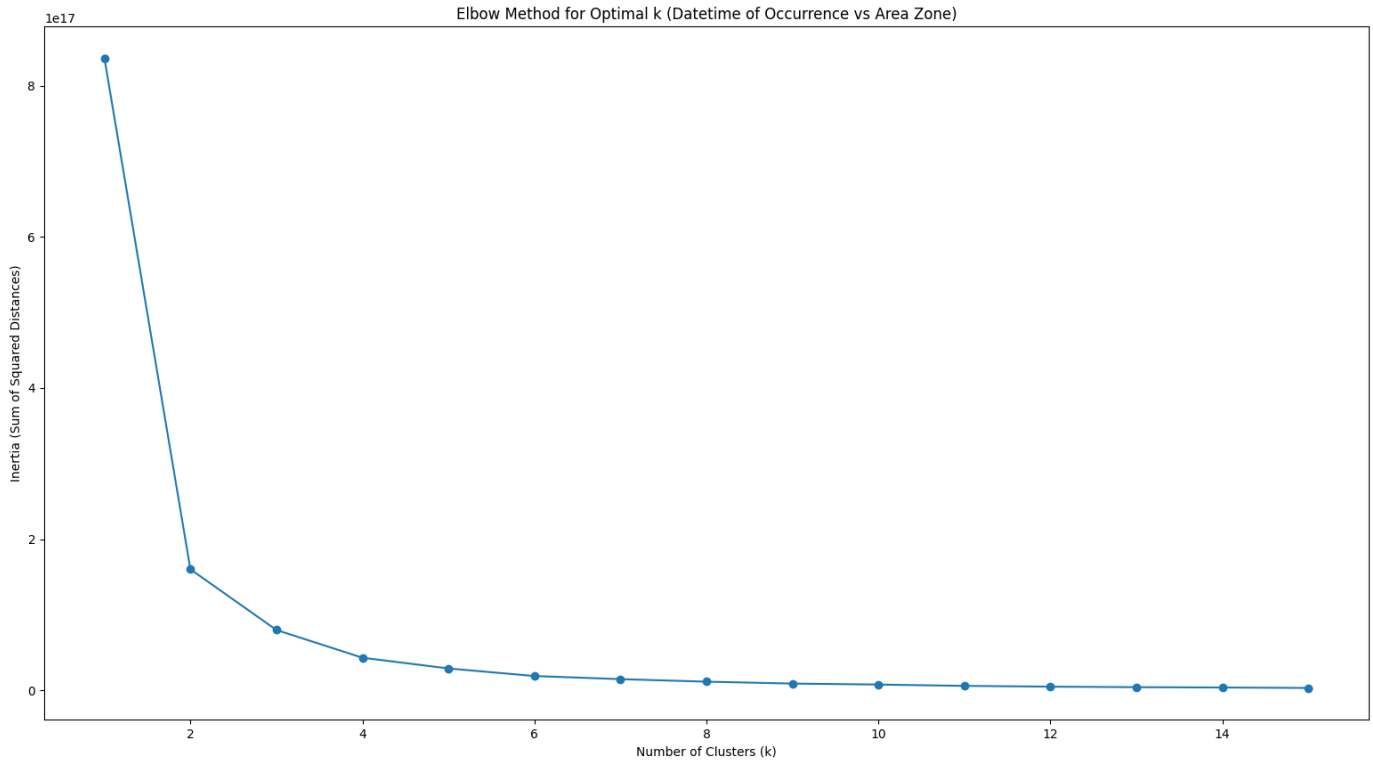
Optimal K for Crime Scene and Division



By looking at the graph, the reduction of inertia can be seen to plateau when increasing the number of K to values larger than 3. The ideal K is considered to be 3.

DATETIME OF OCCURRENCE AND AREA ZONE

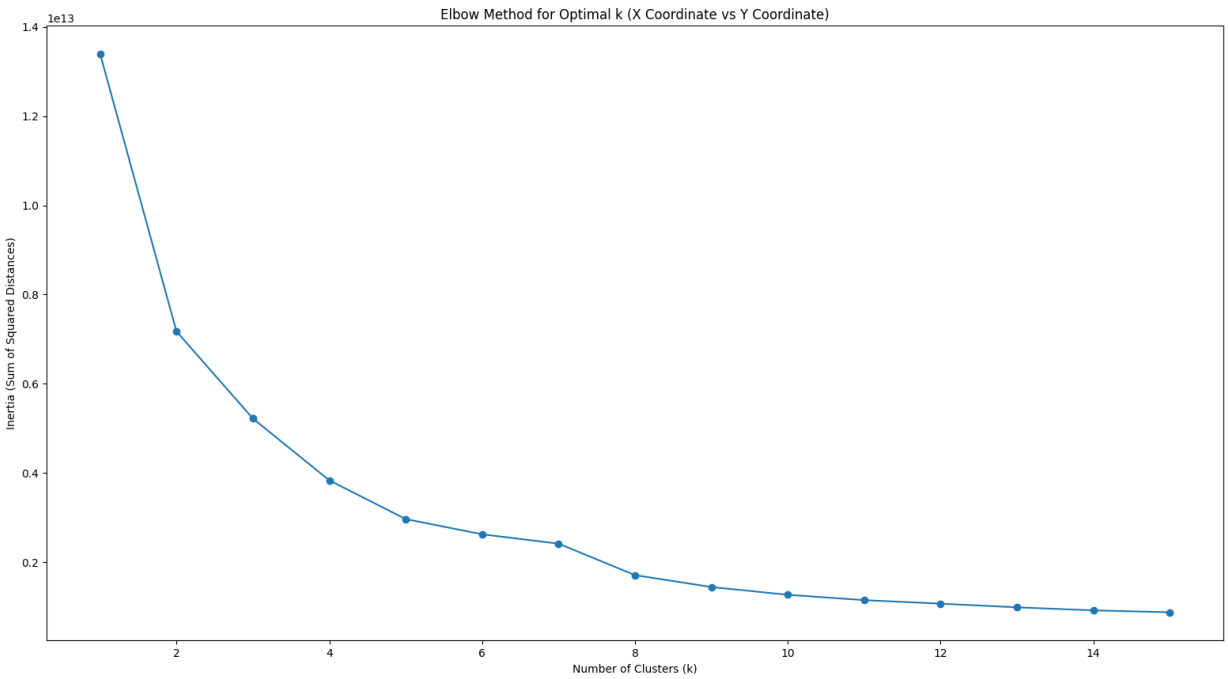
Optimal K for Datetime of Occurrence and Area Zone



By looking at the graph, the reduction of inertia can be seen to plateau when increasing the number of K to values larger than 3. We consider the ideal K to be 3.

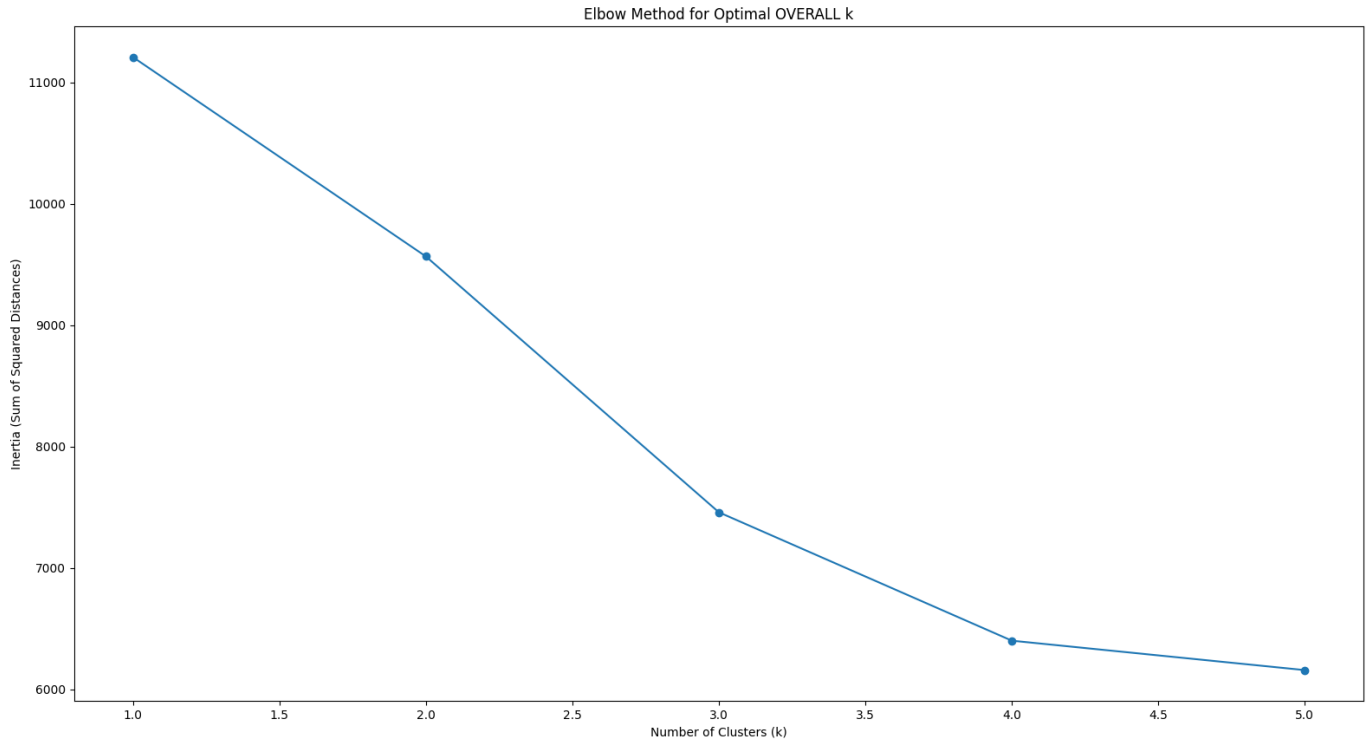
X AND Y COORDINATES

Optimal K for X and Y Coordinates



By looking at the graph, the reduction of inertia can be seen to plateau when increasing the number of K to values larger than 3. We consider the ideal K to be 3.

Elbow Method for Overall k



By looking at the graph, the reduction of inertia can be seen to plateau when increasing the number of K to values larger than 3. We consider the ideal K to be 3.

Outliers

Analyze Outliers

- Using approach 1: Keep the categorical features

  Key attributes of these outliers:

  - Division: Some divisions (e.g., North Dallas) had unexpected crime spikes.
  - Time_Period: Nighttime incidents or early morning crimes were often flagged.
  - Area Zone: Zones like Downtown/Central Dallas showed unusual patterns.
  - Crime Scene: Specific crime scenes, such as Apartment Residences and Stores, were overrepresented.
- Using approach 2: Anomaly detection for Time Series

  Key attributes of these outliers: These patterns suggest seasonal or event-related anomalies.

  - January 2, 2022: High post-New Year activity.
  - July 1–5, 2022: Independence Day celebrations.
  - October 13, 2022: Fall holiday-related activity.
  - November 21, 2022: Pre-Thanksgiving incidents.

Evaluate Impact

- Using approach 1: Keep the categorical features does not show much of the outliers features and characteristic. This result can be inherited from data that has almost categorical features.
- Using approach 2: Anomaly detection for Time Series show patterns that indicate high/low incident time.

6 Evaluation

6.1. Evaluate results

Classification

Model	Accuracy

Decision Tree	74% (Maxdepth of 4)
KNN	74%
Random Forest	77%

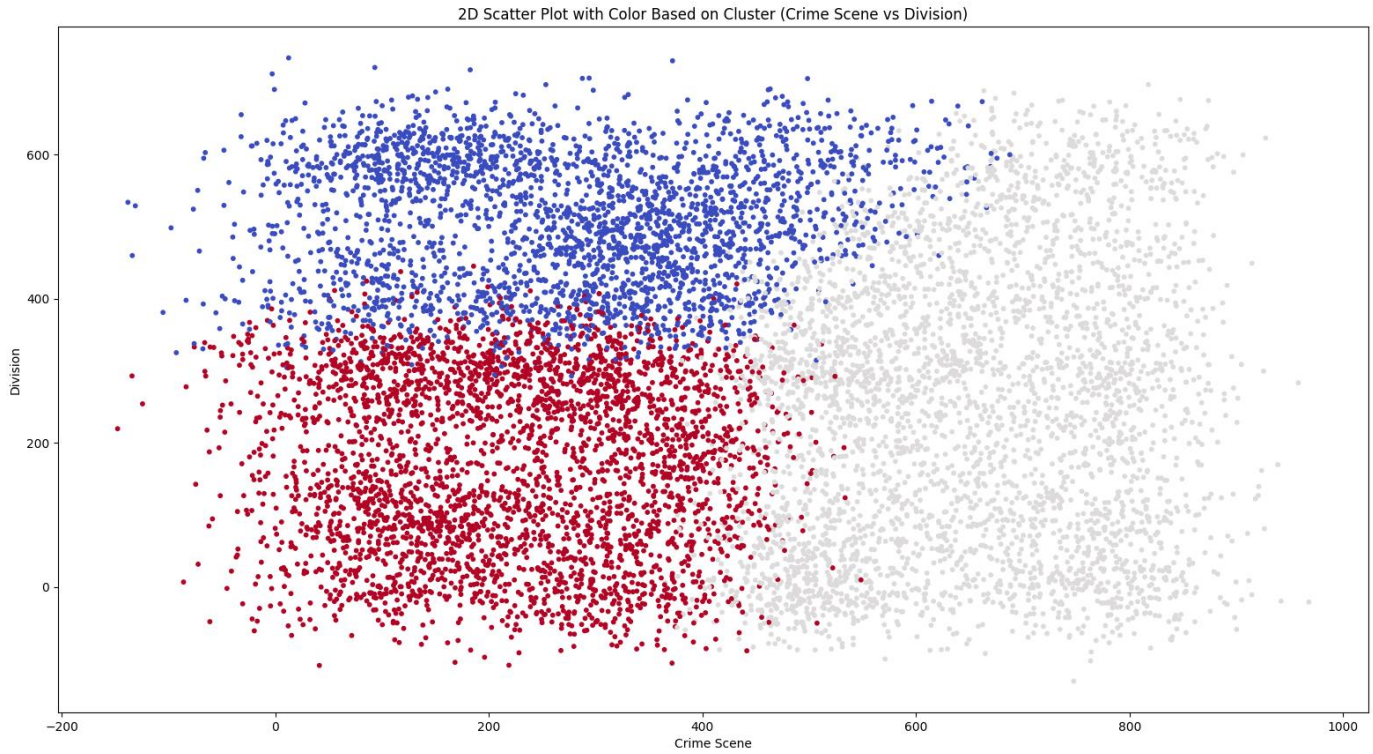
Clustering

Race and Gender Cluster

Race and Gender Map

```
Victim Race Map:
    Code            Category
0      0               White
1      1  Hispanic or Latino
2      2               Black
3      3             Unknown
4      4               Asian
5      5     American Native
6      6      Middle Eastern
Victim Gender Map:
    Code Category
0      0     Male
1      1   Female
2      2  Unknown
```

Crime Scene and Division Cluster



Crime Scene and Division Map

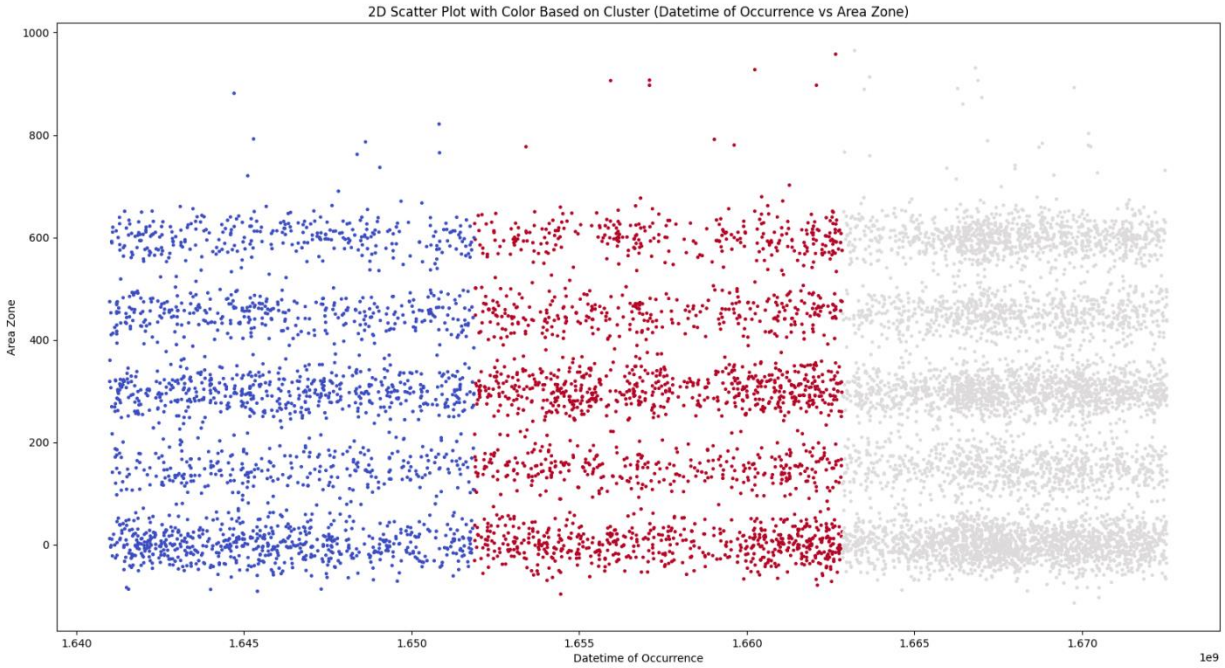
Crime Scene Map:

	Code	Category
0	0	Misc
1	1	Apartment Residence
2	2	Apartment Parking
3	3	Public Roads
4	4	Single Family Residence
5	5	Commercial Establishments
4	4	Single Family Residence
5	5	Commercial Establishments
6	6	Public Spaces
7	7	Stores
8	8	Business Parking

Division Map:

	Code	Category
0	0	northwest
1	1	northeast
2	2	south central
3	3	central
4	4	southwest
5	5	southeast
6	6	north central

Datetime of Occurrence and Area Zone Cluster

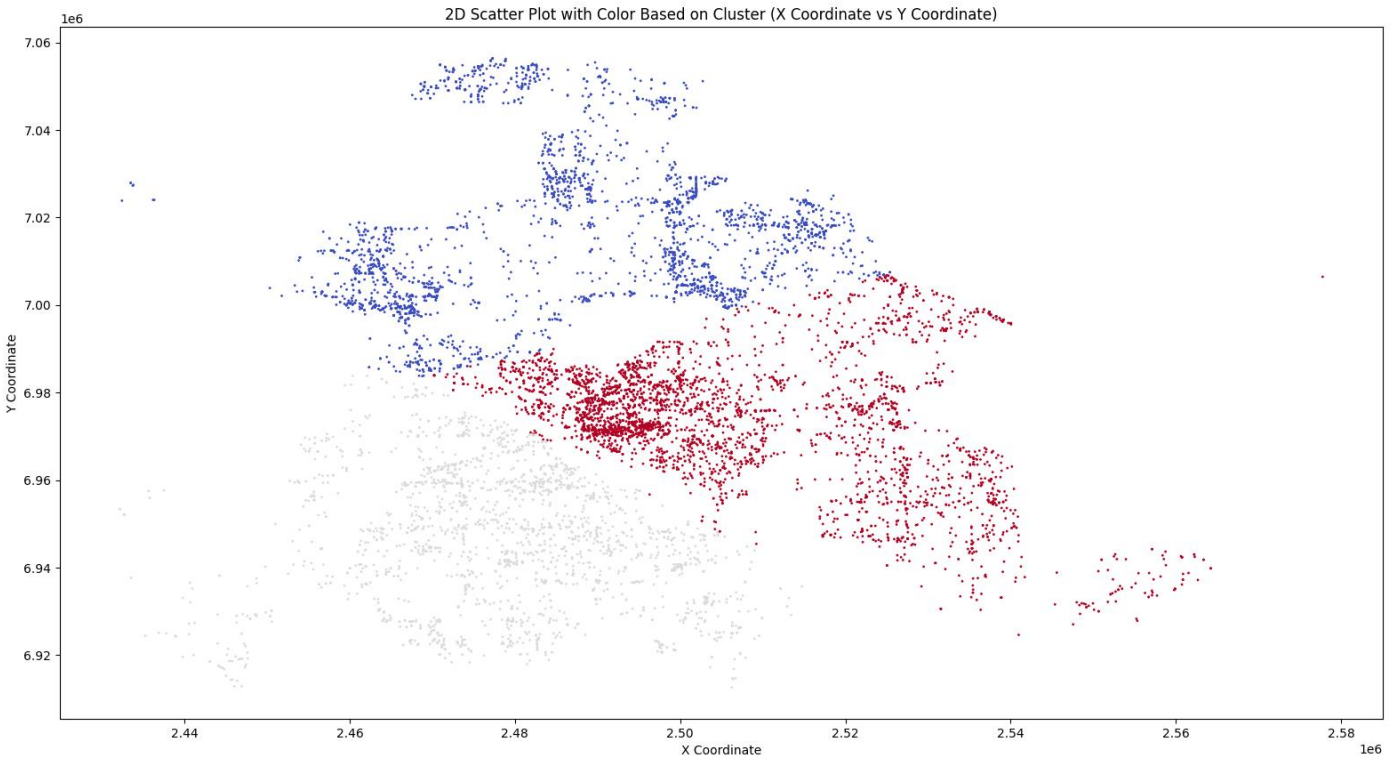


Datetime of Occurrence and Area Zone Map

Area Zone Map:

	Code	Category
0	0	North Dallas
1	1	East Dallas
2	2	South Dallas
3	3	West Dallas
4	4	Downtown/Central Dallas
5	5	Suburb
6	6	Unknown

X and Y Coordinates Cluster



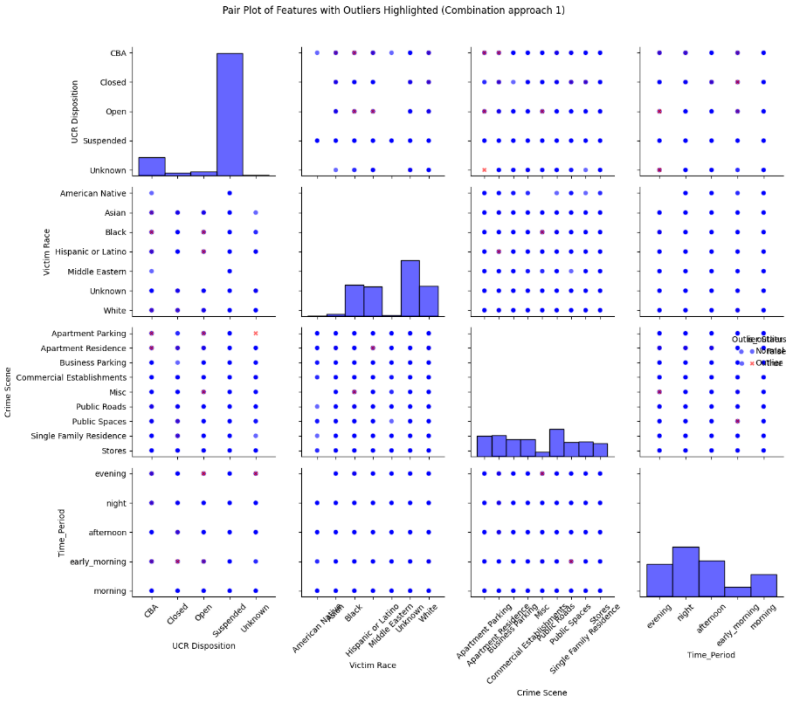
Outliers

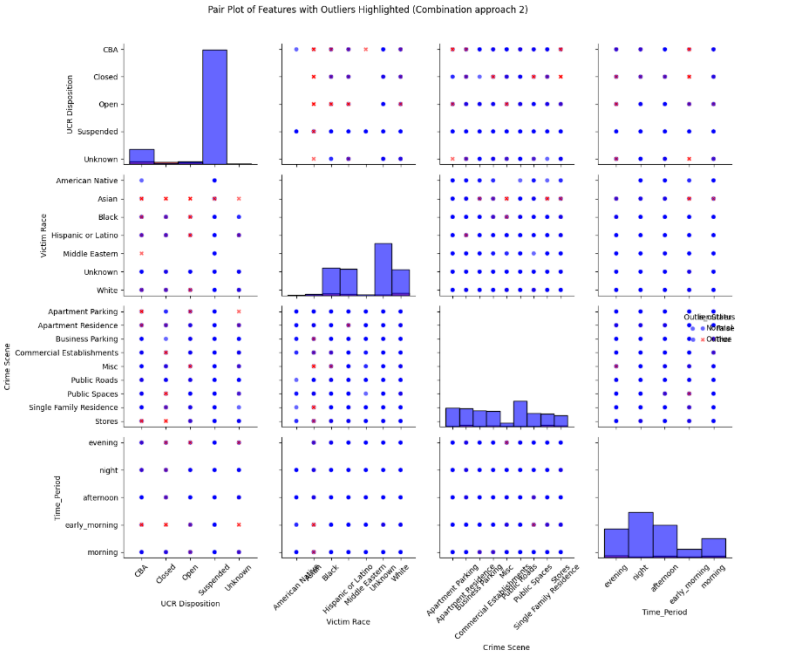
Observations

- **Compare Against Data Distribution**:
  - Example: A Single Family Residence in a low-crime zone during a high-crime time is an anomaly.
- **Contextual Patterns**:
  - Incidents during holidays (e.g., New Year or Independence Day) align with expected outliers.
- **Initial Analysis**:
  - Visual inspection showed some outliers that models did not detect.
  - This indicates potential gaps in model sensitivity for complex, contextual anomalies.
- **Model Performance**:
  - LOF detected anomalies tied to local density changes but missed broader patterns.
  - ISF flagged global anomalies but struggled with nuanced cases (e.g., clustered outliers).

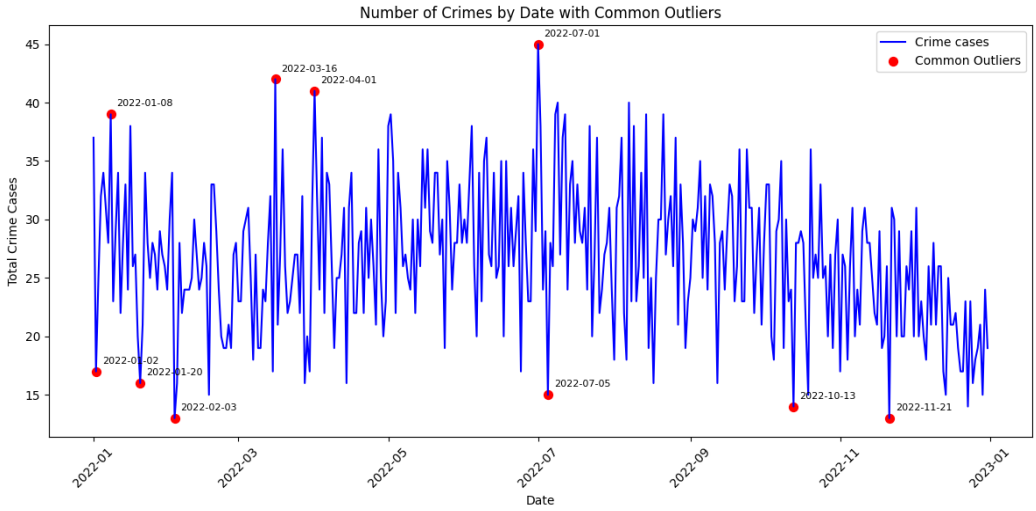
Visualizations

- Using approach 1: **Keep the categorical features**





- Using approach 2: **Anomaly detection for Time Series**



6.2. Interpret results

Classification

- Strengths:
  The Random Forest model performs well for Class A, which is the majority class, achieving high recall (92%) and high F1-score (85%).
  Precision for all classes is above 60%, this indicates that when the model predicts a class, it is reasonably confident.
  For the decision tree, it had easy interpretability of decision-making rules, as shown in the visualized tree. It also efficiently handles encoded data and engineered time-based attributes
- Weaknesses:
  Class A Dominance: Class A has the most samples, so the overall performance may be biased.
  Recall for Minority Classes: Indicating that the model struggles to correctly identify Class B and C.

Clustering

RACE AND GENDER

By performing k-means clustering on victim race and victim gender, we are gaining information of how close the relations among each victim's race and gender are. The clusters also shows that the algorithm considers the Hispanic and White victims of both sexes are related to each other. The Black victims of both sexes are related to each other while there is a large cluster of victims with unknown races and gender.

CRIME SCENE AND DIVISION

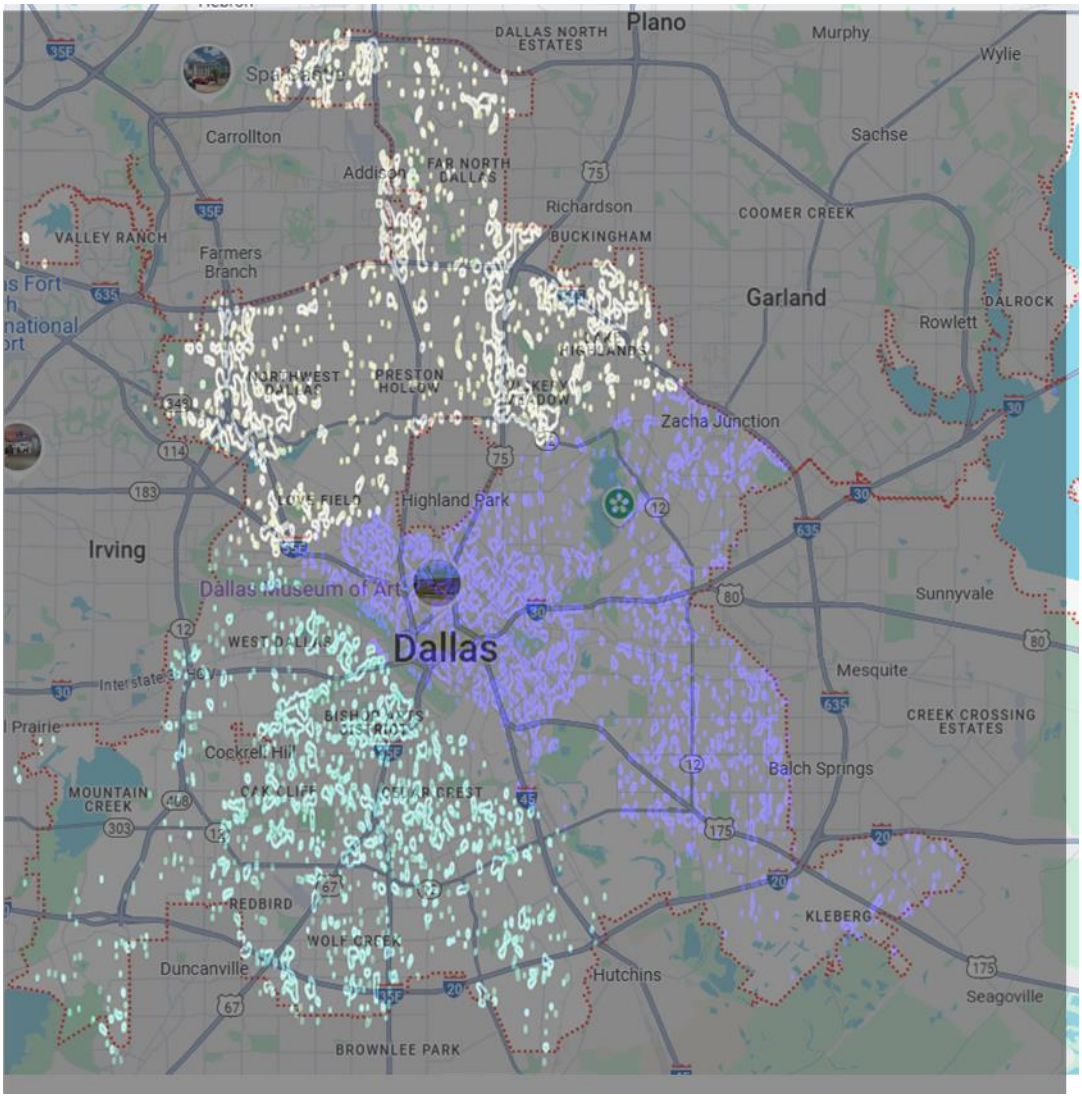
Before performing the clustering algorithm, we hoped to gain some insights as to how the type of locations where crime is committed to be related to the police division area. After performing k-means clustering on crime scene and division, we did not think there is meaningful information to be gained.

DATETIME OF OCCURRENCE AND AREA ZONE

By performing k-means clustering on the datetime and area zone attributes, it can be seen that the intensity of crimes happening is mostly on the last one-third of the year. Future investigations can be done to find out the reason for this happening.

X AND Y COORDINATES

By performing k-means clustering on the X- and Y-coordinates, we can see that the intensity of crimes is high in the center and south of Dallas. The north part of Dallas is sparsely affected by crimes relative to the other parts of Dallas.



Outliers

- Using approach 1: **Keep the categorical features**
    - **LOF** and **ISF** show similar numbers of outliers detected: 500 instances. However, the common instances between those models are 92 or 25 instances (using different combination approaches). It shows the models are resulting not high similar detection for outliers.
- Using approach 2: **Anomaly detection for Time Series**

- **LOF** and **ISF** show similar numbers of outliers detected: 19 instances. And the common instances between those models are 10 instances (more than 50%). It shows the models are resulting in high similar detection for outliers.

6.3. Review of process

Classification

For classification, the modeling process was well-structured, focusing on clear objectives:

- Appropriate preprocessing steps, including feature extraction and encoding, provided a strong foundation.
- Evaluation metrics highlighted the model's performance, helping identify improvement areas. However, some steps could benefit from further refinement, such as better handling of class imbalance.

Clustering

For clustering, just finding the ideal number of k does not really give much important information. We performed additional clustering methods with only pairs of attributes to see if we can get meaningful information. There were a lot of trials and errors when trying to ensure that the pairs cluster map is easy to grasp.

Outliers

To make outlier detection using Local Outlier Factor (LOF) and Isolation Forest (ISF) more effective and useful, certain data requirements and considerations should be taken into account. Here are the key factors:

- **Data quality:** The preparation process should ensure that the dataset is free of noise, missing data should be better managed than just filling in 'unknown' values.
- **Feature selection:** Both LOF and IsF can perform better when the data is approximately normally distributed. The dataset and the feature we are working on are skewed, so the models do not perform well.

6.4. Determine next steps

Classification

Feature Selection with high feature importance can be used to build a model and train it. Since it is an unbalanced dataset, we can try different sampling methods.

In addition, the Zip Code attribute can be further enhanced by replacing missing values with the mode of the distribution (distribution is non-uniform). In this case, 3 values dominate (75220, 75243 and 75228), so the missing values can be replaced with all 3 according to their percentage weights. This ensures a single value doesn’t over dominate the dataset

35.6% of missing values will be replaced with 75220

32.5% of missing values will be replaced with 75243

31.9% of missing values will be replaced with 75228

Clustering

The pairs clusters map can be further refined in the future to see if actions can be taken by the Dallas police department based on the data. Other pairs cluster maps can also be produced for the same purpose.

Outliers

> Improve the effectiveness of outlier detection using LOF and ISF methods by constructing more numbered columns that indicate specific areas of observation. For example: Can detect outliers on date time series for investigating the number of crime incidents by filtering by victim race or area of incident.

7. Conclusion

Random Forest is marginally better than Decision Tree.

Insights

Resource Allocation and Crime Prevention

Time-Location Analysis: Crime patterns based on time (morning vs. night) and location (public roads vs. apartment parking) can help law enforcement plan shifts, patrols, and surveillance in high-risk areas. Our analysis also identifies interesting trends during holiday periods. Outliers suggest that during holidays such as New Year's, Independence Day, and winter events, the number of incidents tends to decrease. However, there is a noticeable surge in incidents immediately following these holidays or events. This insight can assist in preemptive resource allocation and post-event monitoring to address the rise in incidents effectively.

Geospatial Analysis: Geographic insights can lead to better planning of crime prevention measures like security infrastructure, lighting, or community engagement strategies such as neighbourhood watch in specific regions like North Dallas and South Dallas.

Resource Optimization: By analyzing crime frequency patterns across time and location, authorities can allocate police personnel more effectively, ensuring that areas with higher crime rates receive more focused attention.

8. References

https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7/about_data