Milestone 3: Model Implementation

For Milestones 1 and 2, we successfully completed several foundational steps that have set the stage for the progress in Milestone 3. We selected our project topic, developed a structured website with dedicated tabs to organize our findings, and clearly defined the aim and end goals of the project. These milestones also included performing comprehensive Exploratory Data Analysis (EDA) on the dataset, where we gained deeper insights into the variables, relationships, and patterns by creating and analyzing 10 detailed visualizations. Furthermore, we formulated 10 key questions that drive our analysis, providing a roadmap for answering critical aspects of the project. These efforts have equipped us with an in-depth understanding of our dataset and the project's scope, enabling us to approach Milestone 3 with a clear direction and actionable insights.

Website Link: https://nikhil02jk.github.io/Crime-data-analysis/

Screenshot of the data after the cleaning step for Milestone 2

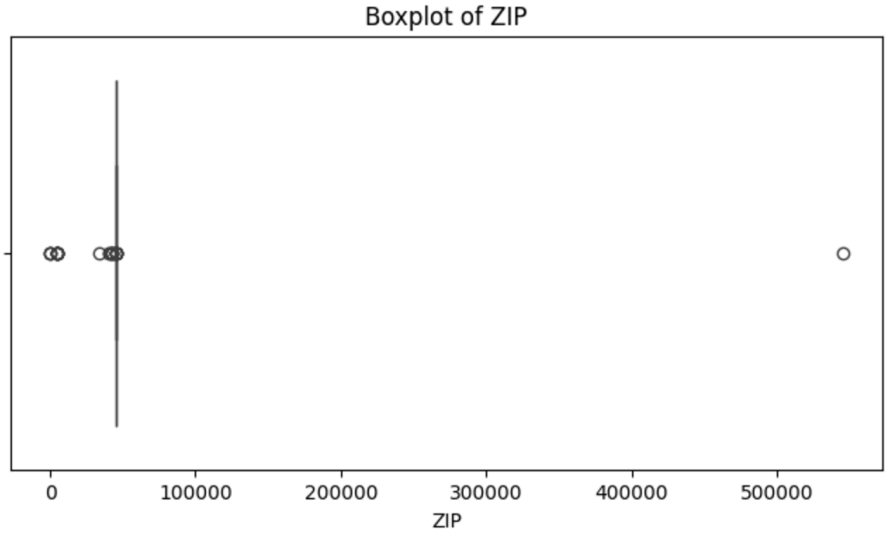
	INCIDENT_NO	CLSD	UCR	DST	BEAT	OFFENSE	LOCATION	THEFT_CODE	HATE_BIAS	DAYOFWEEK	CPD_NEIGHBORHOOD	WEAPONS	DATE_OF_CLEARANCE	HOUR_FROM	HOUR_TO	ADDRESS	LONGITUDE	LATITUD
0	229033241	F--CLEARED BY ARREST - ADULT	1493.0	3	3	CRIMINAL DAMAGING/ENDANGERING	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	TUESDAY	WESTWOOD	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	09/30/2023 12:00:00 AM	220.0	220.0	28XX MONTANA AV	0.3381376199617140	0.55986
1	229033250	Z--EARLY CLOSED	810.0	4	3	ASSAULT	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	TUESDAY	AVONDALE	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	12/13/2022 12:00:00 AM	30.0	330.0	9XX GLENWOOD AV	0.6575302196218620	0.54524
2	229033231	D--VICTIM REFUSED TO COOPERATE	303.0	4	6	AGGRAVATED ROBBERY	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	CARTHAGE	12 - HANDGUN	01/25/2023 12:00:00 AM	1920.0	1935.0	X W SEYMOUR AV	0.6710199154857720	0.83676
3	229033231	D--VICTIM REFUSED TO COOPERATE	303.0	4	6	AGGRAVATED ROBBERY	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	CARTHAGE	12 - HANDGUN	01/25/2023 12:00:00 AM	1920.0	1935.0	X W SEYMOUR AV	0.6710194315734350	0.8360
4	229033225	H--WARRANT ISSUED	802.0	3	2	AGGRAVATED MENACING	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	TUESDAY	EAST PRICE HILL	99 - NONE	12/13/2022 12:00:00 AM	1845.0	1850.0	0XX WOODLAWN AV	0.4097542466338170	0.27413
5	229033225	H--WARRANT ISSUED	600.0	3	2	THEFT	02-MULTI FAMILY APARTMENT	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	TUESDAY	EAST PRICE HILL	99 - NONE	12/13/2022 12:00:00 AM	1845.0	1850.0	0XX WOODLAWN AV	0.4099019295230770	0.2647
6	229033232	J--CLOSED	600.0	2	1	THEFT	02-MULTI FAMILY APARTMENT	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	TUESDAY	EVANSTON	99 - NONE	01/12/2023 12:00:00 AM	1810.0	1811.0	3XXX TRIMBLE AV	0.7168420902777660	0.5417
7	229033221	J--CLOSED	802.0	4	3	AGGRAVATED MENACING	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	TUESDAY	AVONDALE	13 - HANDGUN	01/12/2023 12:00:00 AM	1730.0	1740.0	0XX ROCKDALE AV	0.6276748600726910	0.54091
8	229034052	H--WARRANT ISSUED	701.0	CENTRAL BUSINESS	1	UNAUTHORIZED USE OF MOTOR VEHICLE	62-GAMBLING FACILITY / CASINO / RACE TRACK		N--NO BIAS/NOT APPLICABLE	FRIDAY	C. B. D. / RIVERFRONT	99 - NONE	12/23/2022 12:00:00 AM	119.0	130.0			
9	229034052	H--WARRANT ISSUED	2706.0	CENTRAL BUSINESS	1	VIOLATE PROTECTION ORDER/CONSENT AGREEMENT	62-GAMBLING FACILITY / CASINO / RACE TRACK		N--NO BIAS/NOT APPLICABLE	FRIDAY	C. B. D. / RIVERFRONT	99 - NONE	12/23/2022 12:00:00 AM	119.0	130.0			
10	229034050	F--CLEARED BY ARREST - ADULT	862.0	5	1	DOMESTIC VIOLENCE	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	THURSDAY	CLIFTON/UNIVERSITY HEIGHTS	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/10/2023 12:00:00 AM	2326.0	2315.0	6XX RIDDLE RD	0.5319808806603820	0.49601
11	229034046	J--CLOSED	810.0	4	6	ASSAULT	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	THURSDAY	HARTWELL	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/12/2023 12:00:00 AM	2111.0	230.0	1XX W GALBRAITH RD	0.6717547606625940	0.95153
12	229034043	D--VICTIM REFUSED TO COOPERATE	802.0	1	3	AGGRAVATED MENACING	01-SINGLE FAMILY HOME		N--NO BIAS/NOT APPLICABLE	THURSDAY	WEST END	U - UNKNOWN	02/16/2023 12:00:00 AM	190.0	2018.0	4XX ELIZABETH ST	0.5534013566498240	0.30636
13	229034041	H--WARRANT ISSUED	810.0	1	3	ASSAULT	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	THURSDAY	WEST END	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/24/2023 12:00:00 AM	2030.0	2038.0	140X DUDLEY WALK	0.5282364971249891	0.32473
14	229034035	F--CLEARED BY ARREST - ADULT	802.0	2	1	AGGRAVATED MENACING	01-SINGLE FAMILY HOME		N--NO BIAS/NOT APPLICABLE	THURSDAY	EVANSTON	20 - KNIFE/CUTTING INSTRUMENT (ICEPICK, AX, ETC.)	01/05/2023 12:00:00 AM	1845.0	1849.0	3400X FERNSIDE PL	0.6866201942070800	0.50954
15	229033782	J--CLOSED	600.0	2	4	THEFT	48-PARKING LOT	24O-MOTOR VEHICLE THEFT	N--NO BIAS/NOT APPLICABLE	TUESDAY	LINWOOD	99 - NONE	01/03/2023 12:00:00 AM	342.0	342.0	540X WOOSTER RD	0.9048058874567180	0.468611
16	229033775	J--CLOSED	301.0	1	2	ROBBERY	47-STREET		N--NO BIAS/NOT APPLICABLE	TUESDAY	PENDLETON	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/02/2023 12:00:00 AM	317.0	320.0	0XX E 12TH ST	0.5900418672723330	0.32245
17	229033779	B--PROSECUTION DECLINED	551.0	3	2	BURGLARY	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	MONDAY	EAST PRICE HILL	99 - NONE	12/21/2022 12:00:00 AM	110.0	10.0	340X W 8TH ST	0.4158240166306660	0.29689
18	229033773	D--VICTIM REFUSED TO COOPERATE	810.0	5	4	ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	MOUNT ARY	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	02/06/2023 12:00:00 AM	123.0	158.0			
19	229033773	D--VICTIM REFUSED TO COOPERATE	810.0	5	4	ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	MOUNT ARY	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	02/06/2023 12:00:00 AM	123.0	158.0			
20	229033770	Z--EARLY CLOSED	810.0	3	3	ASSAULT	17-BARBER/BEAUTY SHOP		N--NO BIAS/NOT APPLICABLE	SUNDAY	WESTWOOD	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	12/20/2022 12:00:00 AM	170.0	210.0	340X GLENMORE AV	0.2800727706981480	0.55821
21	229033765	D--VICTIM REFUSED TO COOPERATE	600.0	5	4	THEFT	02-MULTI FAMILY APARTMENT	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	MONDAY	MOUNT AIRY	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/01/2023 12:00:00 AM	2245.0	235.0	27XX HILLVISTA LN	0.3971945965808250	0.82237
22	229033510	K--UNFOUNDED	810.0	1	2	ASSAULT	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	FRIDAY	PENDLETON	80 - OTHER WEAPON	01/14/2023 12:00:00 AM	1915.0	1920.0	0XX E 13TH ST	0.5955455474526449	0.331398
23	229033510	K--UNFOUNDED	810.0	1	2	ASSAULT	02-MULTI FAMILY APARTMENT		N--NO BIAS/NOT APPLICABLE	FRIDAY	PENDLETON	80 - OTHER WEAPON	01/14/2023 12:00:00 AM	1915.0	1920.0	0XX E 13TH ST	0.5989173040116070	0.334316
24	229033513	Z--EARLY CLOSED	303.0	4	4	AGGRAVATED ROBBERY	47-STREET		N--NO BIAS/NOT APPLICABLE	FRIDAY	NORTH AVONDALE	20 - KNIFE/CUTTING INSTRUMENT (ICEPICK, AX, ETC.)	12/16/2022 12:00:00 AM	1920.0	1925.0	400X BEECHWOOD AV	0.6631374308340290	0.61951
25	229033513	Z--EARLY CLOSED	303.0	4	4	AGGRAVATED ROBBERY	47-STREET		N--NO BIAS/NOT APPLICABLE	FRIDAY	NORTH AVONDALE	20 - KNIFE/CUTTING INSTRUMENT (ICEPICK, AX, ETC.)	12/16/2022 12:00:00 AM	1920.0	1925.0	400X BEECHWOOD AV	0.6640530338473240	0.61842
26	229033514	F--CLEARED BY ARREST - ADULT	862.0	5	4	DOMESTIC VIOLENCE	02-MULTI FAMILY 4 PLEX		6B--DOMESTIC VIOLENCE	FRIDAY	MOUNT AIRY	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/27/2023 12:00:00 AM	1915.0	1925.0	250X FLANIGAN CT	0.3930678218236610	0.89781
27	229033506	B--PROSECUTION DECLINED	401.0	3	3	FELONIOUS ASSAULT	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WESTWOOD	12 - HANDGUN	01/11/2023 12:00:00 AM	1815.0	1830.0			
28	229033503	Z--EARLY CLOSED	1493.0	5	3	CRIMINAL DAMAGING/ENDANGERING	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WINTON HILLS	11 - FIREARM (TYPE NOT STATED)	12/22/2022 12:00:00 AM	1811.0	1813.0	5XX DUTCH COLONY DR	0.5709855280148450	0.80258
29	229033503	Z--EARLY CLOSED	1521.0	5	3	IMPROPERLY DISCHARGING FIREARM AT/INTO HABITATION/SCHOOL	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WINTON HILLS	11 - FIREARM (TYPE NOT STATED)	12/22/2022 12:00:00 AM	1811.0	1813.0	5XX DUTCH COLONY DR	0.5797918850129640	0.80359
30	229033503	Z--EARLY CLOSED	1521.0	5	3	IMPROPERLY DISCHARGING FIREARM AT/INTO HABITATION/SCHOOL	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WINTON HILLS	11 - FIREARM (TYPE NOT STATED)	12/22/2022 12:00:00 AM	1811.0	1813.0	5XX DUTCH COLONY DR	0.5823791569553980	0.80092
31	229033503	Z--EARLY CLOSED	1493.0	5	3	CRIMINAL DAMAGING/ENDANGERING	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WINTON HILLS	11 - FIREARM (TYPE NOT STATED)	12/22/2022 12:00:00 AM	1811.0	1813.0	5XX DUTCH COLONY DR	0.5827040468117120	0.80128
32	229033219	F--CLEARED BY ARREST - ADULT	2706.0	3	2	VIOLATE PROTECTION ORDER/CONSENT AGREEMENT	77-OTHER		N--NO BIAS/NOT APPLICABLE	TUESDAY	WEST PRICE HILL	99 - NONE	01/26/2023 12:00:00 AM	113.0	113.0	10XX ROSEMONT AV	0.3687328752719740	0.32290
33	229033216	Z--EARLY CLOSED	600.0	3	3	THEFT	02-MULTI FAMILY APARTMENT	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	SATURDAY	WESTWOOD	99 - NONE	12/14/2022 12:00:00 AM	70.0	2359.0	25XX HARRISON AV	0.3700835257291800	0.61514
34	229033249	F--CLEARED BY ARREST - ADULT	401.0	4	5	FELONIOUS ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7318856063918830	0.79462
35	229033249	F--CLEARED BY ARREST - ADULT	105.0	4	5	MURDER	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7338749060978040	0.79049
36	229033249	F--CLEARED BY ARREST - ADULT	303.0	4	5	AGGRAVATED ROBBERY	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7326339956281670	0.78976
37	229033249	F--CLEARED BY ARREST - ADULT	303.0	4	5	AGGRAVATED ROBBERY	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7324863179389640	0.79408
38	229033249	F--CLEARED BY ARREST - ADULT	303.0	4	5	AGGRAVATED ROBBERY	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7345247278104610	0.78958
39	229033249	F--CLEARED BY ARREST - ADULT	401.0	4	5	FELONIOUS ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7352711486093270	0.79854
40	229033249	F--CLEARED BY ARREST - ADULT	401.0	4	5	FELONIOUS ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7311867525139940	0.78733
41	229033249	F--CLEARED BY ARREST - ADULT	105.0	4	5	MURDER	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7315721695357990	0.79620
42	229033249	F--CLEARED BY ARREST - ADULT	105.0	4	5	MURDER	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	ROSELAWN	12 - HANDGUN	12/20/2022 12:00:00 AM	174.0	175.0	62XX JOYCE LN	0.7299167232669610	0.79474
43	229033217	Z--EARLY CLOSED	600.0	3	2	THEFT	48-PARKING LOT	24O-MOTOR VEHICLE THEFT	N--NO BIAS/NOT APPLICABLE	SUNDAY	EAST PRICE HILL	99 - NONE	12/14/2022 12:00:00 AM	1730.0	1730.0	36XX WARSAW AV	0.4035965382695470	0.31862
44	229033211	D--VICTIM REFUSED TO COOPERATE	251.0	4	1	RAPE	11-HOSPITAL		N--NO BIAS/NOT APPLICABLE	SUNDAY	MOUNT AUBURN	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/08/2023 12:00:00 AM	0.0	0.0			
45	229033212	J--CLOSED	600.0	5	4	THEFT	43-YARD	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	FRIDAY	COLLEGE HILL	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	12/24/2022 12:00:00 AM	1230.0	1232.0	560X COLERAIN AV	0.3902965529810780	0.85928
46	229034096	F--CLEARED BY ARREST - ADULT	810.0	3	3	ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	THURSDAY	WESTWOOD	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	03/11/2023 12:00:00 AM	1545.0	160.0	500X CROOKSHANK RD	0.3010320735109820	0.43400
47	229034096	F--CLEARED BY ARREST - ADULT	810.0	3	3	ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	THURSDAY	WESTWOOD	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	03/11/2023 12:00:00 AM	1545.0	160.0	500X CROOKSHANK RD	0.2990304032602430	0.42999
48	229034096	F--CLEARED BY ARREST - ADULT	810.0	3	3	ASSAULT	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	THURSDAY	WESTWOOD	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	03/11/2023 12:00:00 AM	1545.0	160.0	500X CROOKSHANK RD	0.2988779578201600	0.42859
49	229034022	H--WARRANT ISSUED	810.0	5	2	ASSAULT	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	THURSDAY	CAMP WASHINGTON	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	02/01/2023 12:00:00 AM	150.0	1511.0	11XX DRAPER ST	0.5137269935699940	0.42626
50	229034022	H--WARRANT ISSUED	1493.0	5	2	CRIMINAL DAMAGING/ENDANGERING	02-MULTI FAMILY		N--NO BIAS/NOT APPLICABLE	THURSDAY	CAMP WASHINGTON	99 - NONE	02/01/2023 12:00:00 AM	150.0	1511.0	11XX DRAPER ST	0.5139155462999521	0.42301
51	229034060	J--CLOSED	2609.0	2	3	EXTORTION	77-OTHER		N--NO BIAS/NOT APPLICABLE	MONDAY	PLEASANT RIDGE	U - UNKNOWN	01/18/2023 12:00:00 AM	170.0	190.0	630X MONTGOMERY RD	0.8517589968249190	0.77941
52	229034015	B--PROSECUTION DECLINED	862.0	5	3	DOMESTIC VIOLENCE	37-GROCERY/SUPERMARKET		6B--DOMESTIC VIOLENCE	THURSDAY	SPRING GROVE VILLAGE	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	02/15/2023 12:00:00 AM	130.0	1330.0	47XX KENARD AV	0.5715762610039020	0.65442
53	229034017	Z--EARLY CLOSED	600.0	4	3	THEFT	01-SINGLE FAMILY HOME	23D-THEFT FROM BUILDING	N--NO BIAS/NOT APPLICABLE	TUESDAY	AVONDALE	99 - NONE	12/22/2022 12:00:00 AM	0.0	0.0	6XX MAPLE AV	0.629228461002640	0.51741
54	229033753	Z--EARLY CLOSED	1493.0	3	3	CRIMINAL DAMAGING/ENDANGERING	47-STREET		N--NO BIAS/NOT APPLICABLE	MONDAY	WESTWOOD	U - UNKNOWN	12/20/2022 12:00:00 AM	1831.0	1831.0	23XX BOUDINOT AV	0.3044980060532330	0.44004
55	229033758	H--WARRANT ISSUED	810.0	1	2	ASSAULT	47-STREET		N--NO BIAS/NOT APPLICABLE	MONDAY	OVER-THE-RHINE	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	12/21/2022 12:00:00 AM	1845.0	1850.0	17XX VINE ST	0.5630882458524240	0.35338
56	229033757	F--CLEARED BY ARREST - ADULT	2706.0	3	2	VIOLATE PROTECTION ORDER/CONSENT AGREEMENT	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	MONDAY	WEST PRICE HILL	99 - NONE	01/24/2023 12:00:00 AM	180.0	1849.0	12XX 1ST AV	0.3835246175497390	0.35420
57	229033752	F--CLEARED BY ARREST - ADULT	600.0	3	2	THEFT	01-SINGLE FAMILY HOME	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	SUNDAY	EAST PRICE HILL	99 - NONE	12/28/2023 12:00:00 AM	90.0	90.0	21XX QUEBEC RD	0.4207909261506590	0.38702
58	229033745	Z--EARLY CLOSED	810.0	2	2	ASSAULT	47-STREET		N--NO BIAS/NOT APPLICABLE	MONDAY	HYDE PARK	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/21/2023 12:00:00 AM	1642.0	1644.0	27XX ERIE AV	0.7811814774593980	0.50799
59	229033745	Z--EARLY CLOSED	810.0	2	2	ASSAULT	47-STREET		N--NO BIAS/NOT APPLICABLE	MONDAY	HYDE PARK	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/21/2023 12:00:00 AM	1642.0	1644.0	27XX ERIE AV	0.7836924656998674	0.49863
60	229033745	Z--EARLY CLOSED	810.0	2	2	ASSAULT	47-STREET		N--NO BIAS/NOT APPLICABLE	MONDAY	HYDE PARK	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/21/2023 12:00:00 AM	1642.0	1644.0	27XX ERIE AV	0.7800156572598170	0.50580
61	229033745	Z--EARLY CLOSED	810.0	2	2	ASSAULT	47-STREET		N--NO BIAS/NOT APPLICABLE	MONDAY	HYDE PARK	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/21/2023 12:00:00 AM	1642.0	1644.0	27XX ERIE AV	0.7808851316809360	0.50197
62	229033747	H--WARRANT ISSUED	862.0	1	3	DOMESTIC VIOLENCE	02-MULTI FAMILY APARTMENT		6B--DOMESTIC VIOLENCE	SUNDAY	OVER-THE-RHINE	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	12/18/2022 12:00:00 AM	220.0	10.0	16XX RACE ST	0.5459235557911450	0.34450
63	229033519	Z--EARLY CLOSED	802.0	3	4	AGGRAVATED MENACING	29-GAS STATION		N--NO BIAS/NOT APPLICABLE	FRIDAY	SOUTH FAIRMOUNT	U - UNKNOWN	12/02/2022 12:00:00 AM	189.0	189.0	150X QUEEN CITY AV	0.4737435506112320	0.41848
64	229033317	F--CLEARED BY ARREST - ADULT	1493.0	4	6	CRIMINAL DAMAGING/ENDANGERING	02-MULTI FAMILY DUPLEX		N--NO BIAS/NOT APPLICABLE	FRIDAY	HARTWELL	99 - NONE	12/28/2022 12:00:00 AM	1735.0	1742.0			
65	229033479	K--UNFOUNDED	551.0	3	2	BURGLARY	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	TUESDAY	EAST PRICE HILL	99 - NONE	12/19/2022 12:00:00 AM	0.0	1430.0	36XX LEHMAN RD	0.4459627570540580	0.35792
66	229033496	F--CLEARED BY ARREST - ADULT	810.0	2	2	ASSAULT	22-OTHER BUSINESS OFFICE		N--NO BIAS/NOT APPLICABLE	THURSDAY	OAKLEY	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/03/2023 12:00:00 AM	2025.0	2025.0	44XX APPLETON ST	0.8182788720906730	0.60539
67	229033494	F--CLEARED BY ARREST - ADULT	810.0	5	4	ASSAULT	48-PARKING LOT		6B--DOMESTIC VIOLENCE	FRIDAY	COLLEGE HILL	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/27/2023 12:00:00 AM	160.0	170.0	19XX SAWANNAH WY	0.4483074215685900	0.93280
68	229033494	F--CLEARED BY ARREST - ADULT	862.0	5	4	DOMESTIC VIOLENCE	48-PARKING LOT		6B--DOMESTIC VIOLENCE	FRIDAY	COLLEGE HILL	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/27/2023 12:00:00 AM	160.0	170.0	19XX SAWANNAH WY	0.4458346447129880	0.93189
69	229033490	D--VICTIM REFUSED TO COOPERATE	551.0	3	3	BURGLARY	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WESTWOOD	99 - NONE	02/07/2023 12:00:00 AM	40.0	40.0	29XX HARRISON AV	0.3329812036605200	0.53741
70	229033490	D--VICTIM REFUSED TO COOPERATE	551.0	3	3	BURGLARY	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	FRIDAY	WESTWOOD	99 - NONE	02/07/2023 12:00:00 AM	40.0	40.0	29XX HARRISON AV	0.3329521023660870	0.53595
71	229033210	J--CLOSED	600.0	5	1	THEFT	47-STREET	23H-ALL OTHER LARCENY	N--NO BIAS/NOT APPLICABLE	TUESDAY	CLIFTON/UNIVERSITY HEIGHTS	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	01/17/2023 12:00:00 AM	1550.0	165.0	3XX CALHOUN ST	0.5589938022302880	0.43042
72	229033207	H--WARRANT ISSUED	1493.0	3	3	CRIMINAL DAMAGING/ENDANGERING	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	MONDAY	WESTWOOD	30 - BLUNT OBJECT (CLUB, HAMMER, ETC.)	12/14/2022 12:00:00 AM	430.0	435.0	35XX FIELDCREST DR	0.2665994062309224	0.55701
73	229033207	H--WARRANT ISSUED	1493.0	3	3	CRIMINAL DAMAGING/ENDANGERING	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	MONDAY	WESTWOOD	30 - BLUNT OBJECT (CLUB, HAMMER, ETC.)	12/14/2022 12:00:00 AM	430.0	435.0	35XX FIELDCREST DR	0.2643788939706160	0.55981
74	229033207	H--WARRANT ISSUED	1493.0	3	3	CRIMINAL DAMAGING/ENDANGERING	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	MONDAY	WESTWOOD	30 - BLUNT OBJECT (CLUB, HAMMER, ETC.)	12/14/2022 12:00:00 AM	430.0	435.0	35XX FIELDCREST DR	0.2662691706528820	0.55428
75	229033207	H--WARRANT ISSUED	1493.0	3	3	CRIMINAL DAMAGING/ENDANGERING	01-SINGLE FAMILY 2 STORY		N--NO BIAS/NOT APPLICABLE	MONDAY	WESTWOOD	30 - BLUNT OBJECT (CLUB, HAMMER, ETC.)	12/14/2022 12:00:00 AM	430.0	435.0	35XX FIELDCREST DR	0.2650877478389480	0.55952
76	229033205	F--CLEARED BY ARREST - ADULT	600.0	5	1	THEFT	01-SINGLE FAMILY HOME	23D-THEFT FROM BUILDING	N--NO BIAS/NOT APPLICABLE	MONDAY	FAIRVIEW	40 - PERSONAL WEAPONS (HANDS, FEET, TEETH, ETC.)	12/17/2022 12:00:00 AM	1330.0	1518.0	23XX OHIO AV	0.5666151525269530	0.4139
77	229033214	J--CLOSED	1493.0	2	4	CRIMINAL DAMAGING/ENDANGERING	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	TUESDAY	MT. WASHINGTON	U - UNKNOWN	01/30/2023 12:00:00 AM	1330.0	1430.0	24XX SPINDLEHILL DR	0.9145559656669210	0.27634
78	229033200	Z--EARLY CLOSED	1493.0	3	5	CRIMINAL DAMAGING/ENDANGERING	48-PARKING LOT		N--NO BIAS/NOT APPLICABLE	MONDAY	FAY APARTMENTS	11 - FIREARM (TYPE NOT STATED)	12/14/2022 12:00:00 AM	2330.0	110.0	38XX PRESIDENT DR	0.4422380382727500	0.50098

Improvements on Data Cleaning from the Previous Review:

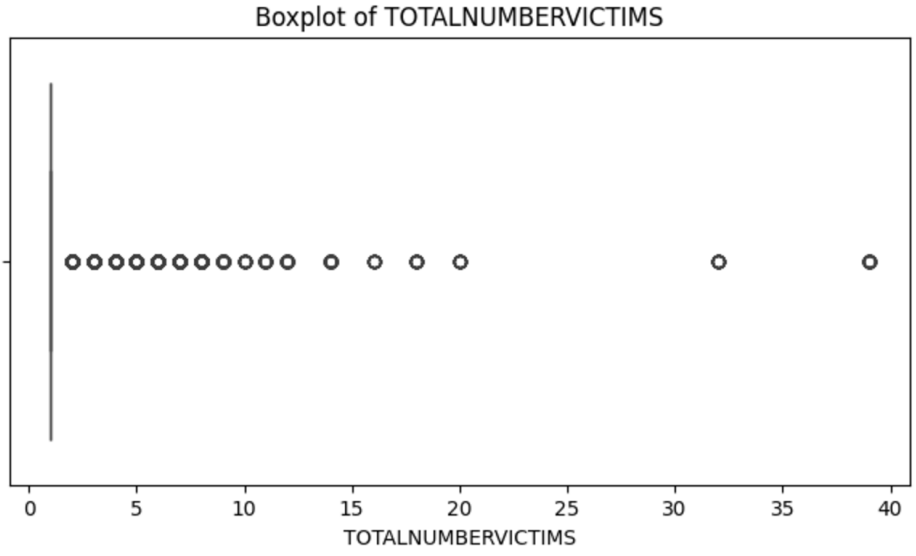
Outlier Detection

We have performed outlier detection on the numerical columns of our data, and these are the results

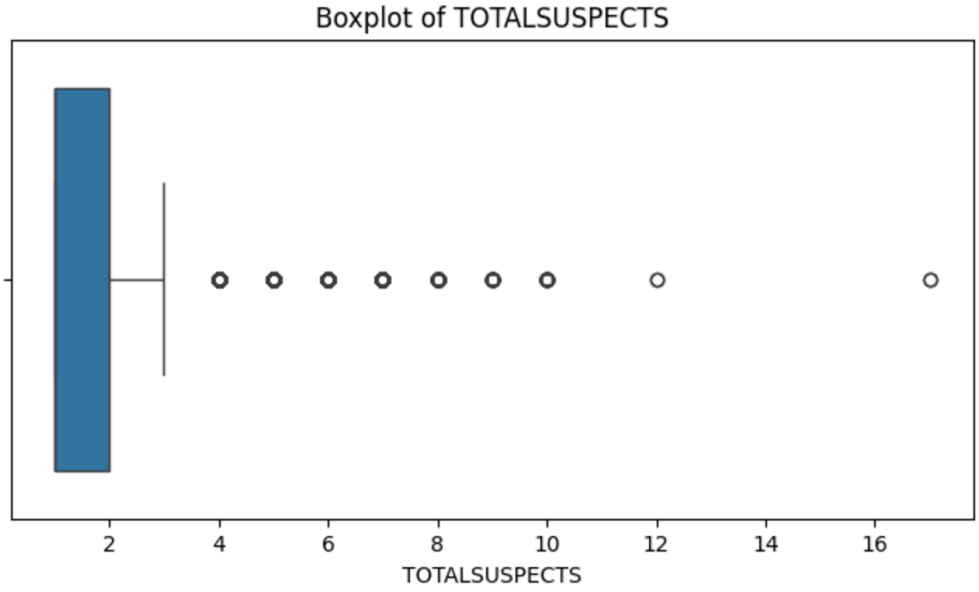
```
Column: ZIP
  Number of outliers: 24
  Lower bound: 45171.5
  Upper bound: 45263.5
```



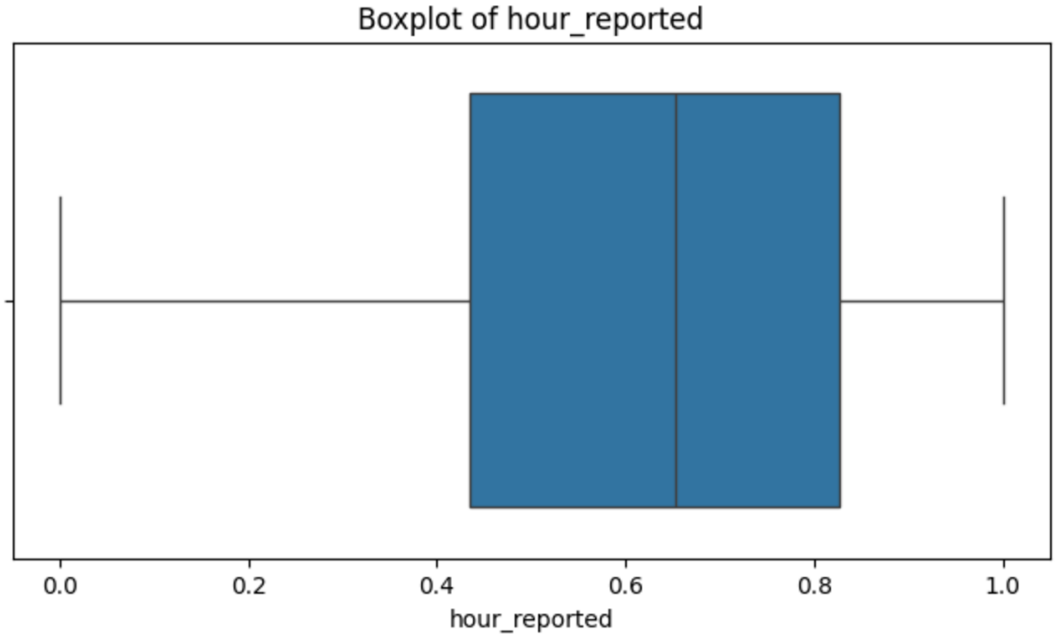

```
Skipping column HATE_BIAS because it contains only NaN values.
Column: TOTALNUMBERVICTIMS
  Number of outliers: 28131
  Lower bound: 1.0
  Upper bound: 1.0
```



```
Column: TOTALSUSPECTS
  Number of outliers: 6054
  Lower bound: -0.5
  Upper bound: 3.5
```




```
Skipping column OFFENSE because it contains only NaN values.
Column: hour_reported
  Number of outliers: 0
  Lower bound: -0.15217391304347833
  Upper bound: 1.4130434782608696
```



Outlier Removal

Outlier detection revealed anomalies in the TOTALNUMBERVICTIMS and TOTALSUSPECTS columns. However, these values were retained, as such occurrences could represent valid cases involving a high number of suspects or victims, which are plausible in the context of criminal incidents. For the other columns used in modeling, outlier detection and handling were meticulously performed to ensure the accuracy and reliability of the models.

```
# Function to drop outliers using IQR method
def drop_outliers(df):
    numeric_cols = df.select_dtypes(include=['number']).columns  # Select numeric columns
    for col in numeric_cols:
      if col != 'TOTALNUMBERVICTIMS' and col != 'TOTALSUSPECTS' and col != 'HATE_BIAS':
        Q1 = df[col].quantile(0.25)  # First quartile
        Q3 = df[col].quantile(0.75)  # Third quartile
        IQR = Q3 - Q1                # Interquartile range
        lower_bound = Q1 - 1.5 * IQR # Lower bound
        upper_bound = Q3 + 1.5 * IQR # Upper bound

        # Remove rows outside the bounds
        df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]

    return df
```

Perform Null Value checks for the columns that we are using for modelling

[ ]:
```
# Calculate and print the number of null values in each column
null_counts = data_subset.isnull().sum()

print("Number of null values in each column:")
print(null_counts)
```

```
Number of null values in each column:
LOCATION              0
SNA_NEIGHBORHOOD      0
ZIP                   0
HATE_BIAS             0
TOTALNUMBERVICTIMS    0
TOTALSUSPECTS         0
UCR_GROUP             0
OFFENSE               0
hour_reported         0
dtype: int64
```

Screenshot of the data after cleaning

data_before_modelling

	LOCATION	SNA_NEIGHBORHOOD	ZIP	HATE_BIAS	TOTALNUMBERVICTIMS	TOTALSUSPECTS	UCR_GROUP	OFFENSE	time_reported
0	02-MULTI FAMILY APARTMENT	WESTWOOD	45211.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	CRIMINAL DAMAGING/ENDANGERING	11:11:00 PM
1	02-MULTI FAMILY	AVONDALE	45229.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	ASSAULT	08:55:20 PM
2	48-PARKING LOT	CARTHAGE	45216.0	N--NO BIAS/NOT APPLICABLE	1.0	2.0	ROBBERY	AGGRAVATED ROBBERY	07:36:00 PM
3	48-PARKING LOT	CARTHAGE	45216.0	N--NO BIAS/NOT APPLICABLE	1.0	2.0	ROBBERY	AGGRAVATED ROBBERY	07:36:00 PM
4	02-MULTI FAMILY APARTMENT	EAST PRICE HILL	45205.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	AGGRAVATED MENACING	06:59:00 PM
5	02-MULTI FAMILY APARTMENT	EAST PRICE HILL	45205.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	THEFT	THEFT	06:59:00 PM
6	02-MULTI FAMILY APARTMENT	EVANSTON	45207.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	THEFT	THEFT	06:12:00 PM
7	02-MULTI FAMILY APARTMENT	AVONDALE	45229.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	AGGRAVATED MENACING	06:00:00 PM
8	62-GAMBLING FACILITY / CASINO / RACE TRACK		45202.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	UNAUTHORIZED USE	UNAUTHORIZED USE OF MOTOR VEHICLE	01:22:00 AM
9	62-GAMBLING FACILITY / CASINO / RACE TRACK		45202.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	VIOLATE PROTECTION ORDER/CONSENT AGREEMENT	01:22:00 AM
10	02-MULTI FAMILY APARTMENT	CUF	45220.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	DOMESTIC VIOLENCE	11:26:00 PM
11	02-MULTI FAMILY APARTMENT	HARTWELL	45216.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	ASSAULT	11:05:00 PM
12	01-SINGLE FAMILY HOME	WEST END	45203.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	AGGRAVATED MENACING	10:10:00 PM
13	02-MULTI FAMILY APARTMENT	WEST END	45214.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	ASSAULT	08:38:00 PM
14	01-SINGLE FAMILY HOME	EVANSTON	45207.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	AGGRAVATED MENACING	07:28:34 PM
15	48-PARKING LOT	LINWOOD	45226.0	N--NO BIAS/NOT APPLICABLE	2.0	2.0	THEFT	THEFT	03:58:00 AM
16	47-STREET	PENDLETON	45202.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	ROBBERY	ROBBERY	03:21:02 AM
17	02-MULTI FAMILY APARTMENT	EAST PRICE HILL	45205.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	BURGLARY/BREAKING ENTERING	BURGLARY	02:07:00 AM
18	48-PARKING LOT		45223.0	N--NO BIAS/NOT APPLICABLE	1.0	2.0	PART 2 MINOR	ASSAULT	01:23:00 AM
19	48-PARKING LOT		45223.0	N--NO BIAS/NOT APPLICABLE	1.0	2.0	PART 2 MINOR	ASSAULT	01:23:00 AM
20	17-BARBER/BEAUTY SHOP	WESTWOOD	45211.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	ASSAULT	11:34:00 PM
21	02-MULTI FAMILY APARTMENT	MT. AIRY	45239.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	THEFT	THEFT	11:17:00 PM
22	02-MULTI FAMILY APARTMENT	PENDLETON	45202.0	N--NO BIAS/NOT APPLICABLE	2.0	1.0	PART 2 MINOR	ASSAULT	07:36:46 PM
23	02-MULTI FAMILY APARTMENT	PENDLETON	45202.0	N--NO BIAS/NOT APPLICABLE	2.0	1.0	PART 2 MINOR	ASSAULT	07:36:46 PM
24	47-STREET	NORTH AVONDALE - PADDOCK HILLS	45229.0	N--NO BIAS/NOT APPLICABLE	2.0	1.0	ROBBERY	AGGRAVATED ROBBERY	07:29:32 PM
25	47-STREET	NORTH AVONDALE - PADDOCK HILLS	45229.0	N--NO BIAS/NOT APPLICABLE	2.0	1.0	ROBBERY	AGGRAVATED ROBBERY	07:29:32 PM
26	02-MULTI FAMILY 4 PLEX	MT. AIRY	45239.0	88--DOMESTIC VIOLENCE	1.0	1.0	PART 2 MINOR	DOMESTIC VIOLENCE	07:28:00 PM
27	02-MULTI FAMILY		45211.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	AGGRAVATED ASSAULTS	FELONIOUS ASSAULT	06:30:00 PM
28	02-MULTI FAMILY	WINTON HILLS	45232.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	PART 2 MINOR	CRIMINAL DAMAGING/ENDANGERING	06:14:00 PM
29	02-MULTI FAMILY	WINTON HILLS	45232.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	PART 2 MINOR	IMPROPERLY DISCHARGING FIREARM AT/INTO HABITATION/SCHOOL	06:14:00 PM
30	02-MULTI FAMILY	WINTON HILLS	45232.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	PART 2 MINOR	IMPROPERLY DISCHARGING FIREARM AT/INTO HABITATION/SCHOOL	06:14:00 PM
31	02-MULTI FAMILY	WINTON HILLS	45232.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	PART 2 MINOR	CRIMINAL DAMAGING/ENDANGERING	06:14:00 PM
32	77-OTHER	WEST PRICE HILL	45205.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	PART 2 MINOR	VIOLATE PROTECTION ORDER/CONSENT AGREEMENT	05:36:00 PM
33	02-MULTI FAMILY APARTMENT	WESTWOOD	45211.0	N--NO BIAS/NOT APPLICABLE	1.0	1.0	THEFT	THEFT	05:31:38 PM
34	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	AGGRAVATED ASSAULTS	FELONIOUS ASSAULT	05:11:00 PM
35	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	HOMICIDE	MURDER	05:11:00 PM
36	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	ROBBERY	AGGRAVATED ROBBERY	05:11:00 PM
37	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	ROBBERY	AGGRAVATED ROBBERY	05:11:00 PM
38	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	ROBBERY	AGGRAVATED ROBBERY	05:11:00 PM
39	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	AGGRAVATED ASSAULTS	FELONIOUS ASSAULT	05:11:00 PM
40	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	AGGRAVATED ASSAULTS	FELONIOUS ASSAULT	05:11:00 PM
41	48-PARKING LOT	ROSELAWN	45237.0	N--NO BIAS/NOT APPLICABLE	3.0	1.0	HOMICIDE	MURDER	05:11:00 PM

Screenshot of the data after cleaning and preprocessing and before building the model

```
[11]: data_subset.head()
```

[11]:

	LOCATION	SNA_NEIGHBORHOOD	ZIP	HATE_BIAS	TOTALNUMBERVICTIMS	TOTALSUSPECTS	UCR_GROUP	OFFENSE	hour_reported
0	8	49	0.082921	16	0.0	0.0000	3	15	1.000000
1	6	0	0.082954	16	0.0	0.0000	3	10	0.869565
2	54	4	0.082930	16	0.0	0.0625	5	6	0.826087
3	54	4	0.082930	16	0.0	0.0625	5	6	0.826087
4	8	12	0.082910	16	0.0	0.0000	3	4	0.782609

These are the only few columns that we are using for model building in our data.we have converted categorical values to numerical values for model building.

Label Encoding

```
categorical_columns = ['LOCATION', 'SNA_NEIGHBORHOOD', 'HATE_BIAS', 'UCR_GROUP', 'OFFENSE']
label_encoders = {col: LabelEncoder() for col in categorical_columns}
for col in categorical_columns:
    data_subset[col] = label_encoders[col].fit_transform(data_subset[col])
```

```
scaler = MinMaxScaler()
numeric_columns = ['ZIP', 'TOTALNUMBERVICTIMS', 'TOTALSUSPECTS', 'hour_reported']
data_subset[numeric_columns] = scaler.fit_transform(data_subset[numeric_columns])
```

```
data_subset.head()
```

	LOCATION	SNA_NEIGHBORHOOD	ZIP	HATE_BIAS	TOTALNUMBERVICTIMS	TOTALSUSPECTS	UCR_GROUP	OFFENSE	hour_reported
0	8	49	0.082921	16	0.0	0.0000	3	15	1.000000
1	6	0	0.082954	16	0.0	0.0000	3	10	0.869565
2	54	4	0.082930	16	0.0	0.0625	5	6	0.826087
3	54	4	0.082930	16	0.0	0.0625	5	6	0.826087
4	8	12	0.082910	16	0.0	0.0000	3	4	0.782609

The provided code preprocesses the dataset by transforming categorical and numerical features to prepare it for machine learning. Categorical columns (LOCATION, SNA_NEIGHBORHOOD, etc.) are label-encoded, converting unique categories into integer values. Numerical columns (ZIP, TOTALNUMBERVICTIMS, etc.) are scaled using MinMaxScaler to normalize their range between 0 and 1. These steps ensure that categorical data is converted into a format suitable for modeling, and numerical features are on a consistent scale, improving model performance and interpretability. Finally, the processed dataset is displayed to confirm the transformations.

Splitting the data to testing and training

```
[ ]: X = data_subset.drop(columns=['OFFENSE'])
     y = data_subset['OFFENSE']
```

```
[ ]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[ ]: X_train = X_train.values
     X_test = X_test.values
```

```
[ ]: X_train = X_train.reshape(X_train.shape[0], -1)
     X_test = X_test.reshape(X_test.shape[0], -1)
```

```
[ ]: scaler = StandardScaler()
     X_train = scaler.fit_transform(X_train)
     X_test = scaler.transform(X_test)
```

```
[ ]: time_steps = 1
     X_train = X_train.reshape(X_train.shape[0], X_train.shape[1], time_steps)
     X_test = X_test.reshape(X_test.shape[0], X_test.shape[1], time_steps)
```

```
[ ]: print("Reshaped X_train shape:", X_train.shape)
     print("Reshaped X_test shape:", X_test.shape)
```

```
Reshaped X_train shape: (94940, 8, 1)
Reshaped X_test shape: (23736, 8, 1)
```

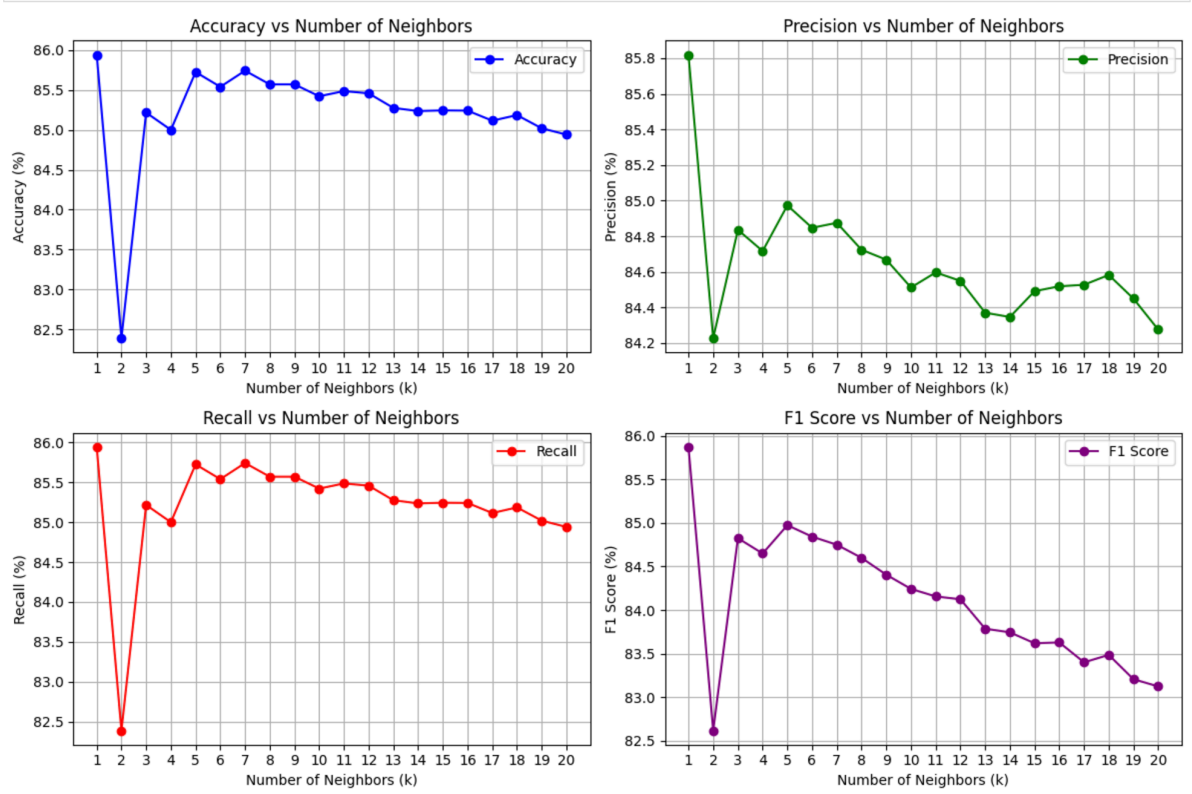
The code prepares the dataset for model training by separating features and the target variable. The feature set X is created by dropping the target column OFFENSE from the dataset, while the target variable y contains the OFFENSE values. The data is then split into training and testing subsets using an 80-20 split via train_test_split, ensuring reproducibility with a fixed random_state. The features in X_train and X_test are converted into NumPy arrays for further processing. A StandardScaler is applied to normalize the training and testing features, transforming them to have a mean of 0 and a standard deviation of 1. Next, the data is reshaped to ensure compatibility with models that require specific input dimensions, such as recurrent neural networks. The final shapes of X_train and X_test are printed to confirm the transformations, with the reshaped data now having dimensions suitable for time-series analysis or similar tasks.

DATA MODELLING

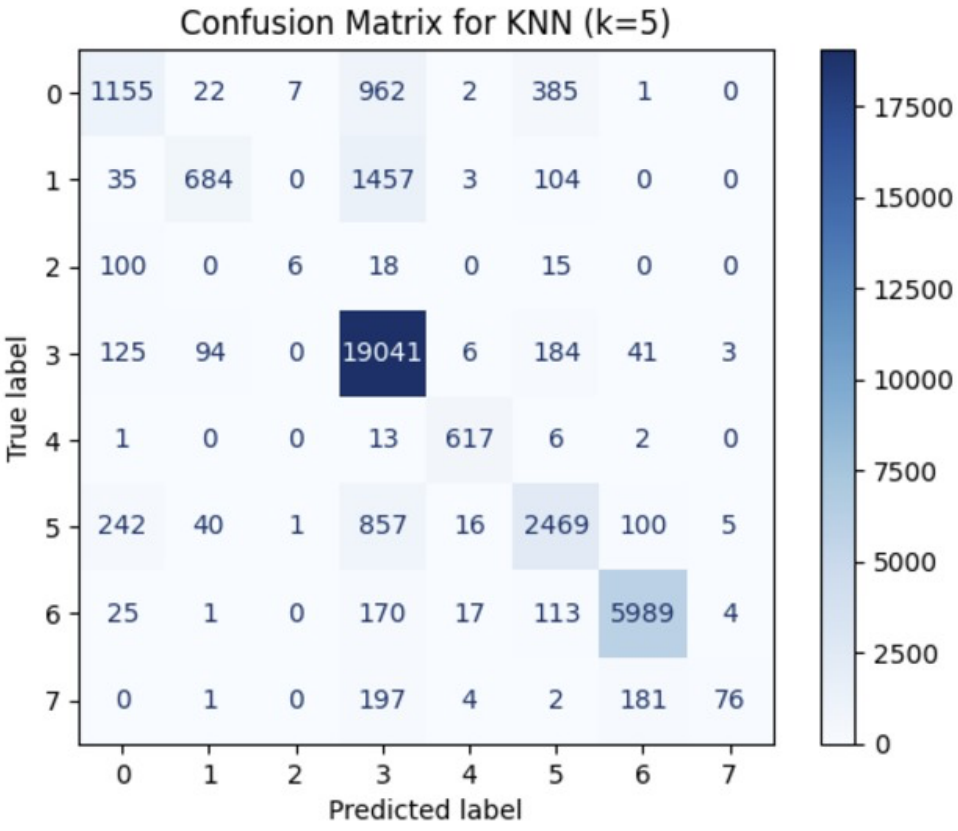
1.K NEARESRT NEIGHBOUR

TARGET VARIABLE – UCR_GROUP

In our KNN modeling, we have used UCR (Uniform Crime Reporting) as our target variable because it provides a standardized way of categorizing crimes into broad classifications, such as violent crimes or property crimes. Using UCR as the target allows us to focus on understanding crime patterns and trends, which can help in resource allocation, law enforcement planning, and crime prevention efforts. Since UCR is a universally recognized crime reporting system, it makes our analysis more relevant and actionable, aligning it with real-world applications like policymaking and identifying crime hotspots. This ensures that our findings can be effectively used to address and understand criminal activity.



```
KNN Accuracy (using ADDRESS): 82.14%
KNN F1-Score (using ADDRESS): 79.96%

Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.46      0.55      2534
           1       0.82      0.30      0.44      2283
           2       0.46      0.04      0.08       139
           3       0.85      0.95      0.90     19494
           4       0.92      0.95      0.94       639
           5       0.74      0.63      0.68      3730
           6       0.80      0.92      0.86      6319
           7       0.50      0.00      0.00       461

    accuracy                           0.82     35599
   macro avg       0.72      0.53      0.56     35599
weighted avg       0.81      0.82      0.80     35599
```



The K-Nearest Neighbors (KNN) model achieved an accuracy of **82.14%** and an F1-score of **79.96%** on the test dataset. The classification report provides detailed metrics for each class, including precision, recall, and F1-score. The model performed exceptionally well for

class 3 and class 4, with F1-scores of **0.90** and **0.94**, respectively, indicating that these classes are predicted with high precision and recall. However, performance for class 2 was poor, with an F1-score of **0.08**, and class 7 had a score of **0.00**, indicating challenges in identifying these classes. The macro-average F1-score is **0.56**, reflecting the imbalance in performance across classes, while the weighted average F1-score is **0.80**, indicating the overall performance weighted by class support. These results suggest that while the model works well for majority classes, it struggles with underrepresented or minority classes, which might require further optimization or rebalancing of the dataset.

High Accuracy for Class 3: Class 3 has the highest correctly classified instances with 19,041, indicating excellent performance for this class.

Misclassifications: Class 0 shows significant misclassification, with 962 instances predicted as Class 3. Class 5 has 857 instances misclassified as Class 3 and 242 instances misclassified as Class 0. Other classes like 1, 4, 6, and 7 also show some degree of misclassification, though less severe.

Confusion for Neighboring Classes: The confusion is noticeable for classes that are "neighboring" or similar in data space, e.g., misclassification between Classes 5 and 3.

Sparse Misclassifications: Certain classes have very few non-zero values for off-diagonal elements, such as Class 7, where most predictions fall under Class 7 with minor misclassifications.

Dominant Class Sizes: Class 3 dominates the dataset in terms of size, as reflected by the large count of true and predicted values compared to other classes.

2. RANDOM FOREST

```
[ ]: from sklearn. ensemble import RandomForestClassifier
label_encoder = LabelEncoder()
data['VICTIM_ETHNICITY'] = label_encoder.fit_transform(data['VICTIM_ETHNICITY'].astype(str))
```

```
[ ]: data['UCR_GROUP'] = label_encoder.fit_transform(data['UCR_GROUP'])

X = data[['UCR_GROUP']]
y = data['VICTIM_ETHNICITY']
```

```
[ ]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)
```

```
129]: RandomForestClassifier(random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

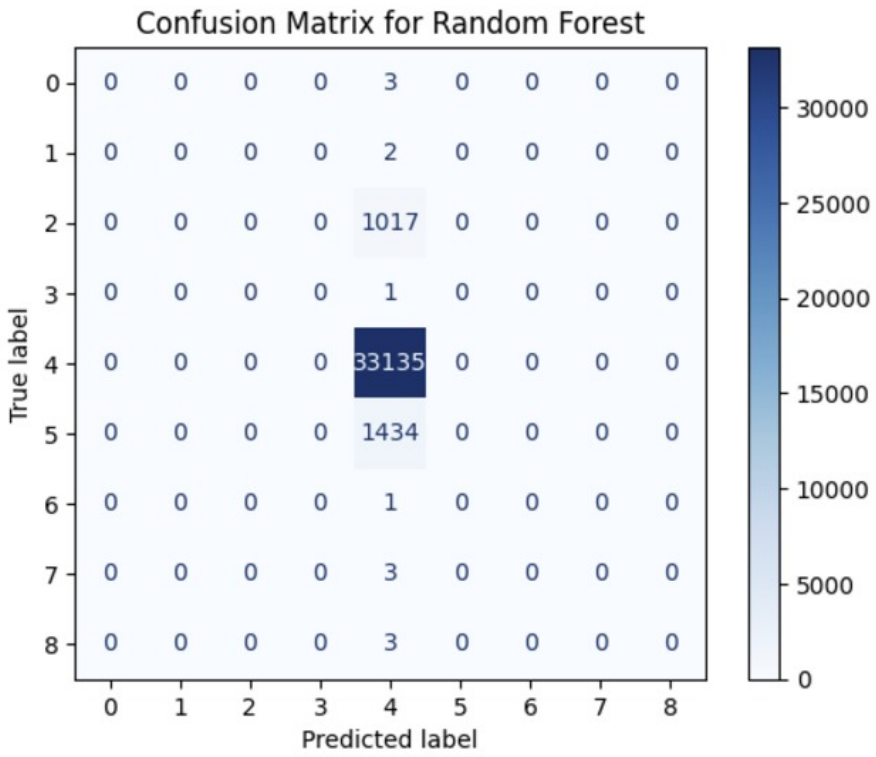
```
[ ]: from sklearn.metrics import accuracy_score, classification_report, f1_score
y_pred = rf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred) * 100
f1 = f1_score(y_test, y_pred, average='weighted') * 100
```

```
[ ]: print(f"Random Forest Accuracy (using ADDRESS): {accuracy:.2f}%")
print(f"Random Forest F1-Score (using ADDRESS): {f1:.2f}%")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

```
Random Forest Accuracy (using ADDRESS): 93.08%
Random Forest F1-Score (using ADDRESS): 89.74%
```

In this model, we are predicting the VICTIM_ETHNICITY using UCR_GROUP as the input feature. The significance of calculating this variable lies in its potential to provide insights into how different crime categories are associated with specific victim demographics. This can help law enforcement and policymakers identify trends, address potential biases, and allocate resources effectively to support vulnerable groups. By understanding these relationships, interventions can be tailored to improve community safety and inclusivity.

```
Random Forest Accuracy (using ADDRESS): 93.08%
Random Forest F1-Score (using ADDRESS): 89.74%

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         3
           1       0.00      0.00      0.00         2
           2       0.00      0.00      0.00      1017
           3       0.00      0.00      0.00         1
           4       0.93      1.00      0.96     33135
           5       0.00      0.00      0.00      1434
           6       0.00      0.00      0.00         1
           7       0.00      0.00      0.00         3
           8       0.00      0.00      0.00         3

    accuracy                           0.93     35599
   macro avg       0.10      0.11      0.11     35599
weighted avg       0.87      0.93      0.90     35599
```



The model's output shows an accuracy of **93.00%**, meaning that the predictions are highly accurate for the test dataset. The F1-score is **0.90**, which indicates that the model has a strong balance between precision and recall, ensuring reliable results across various

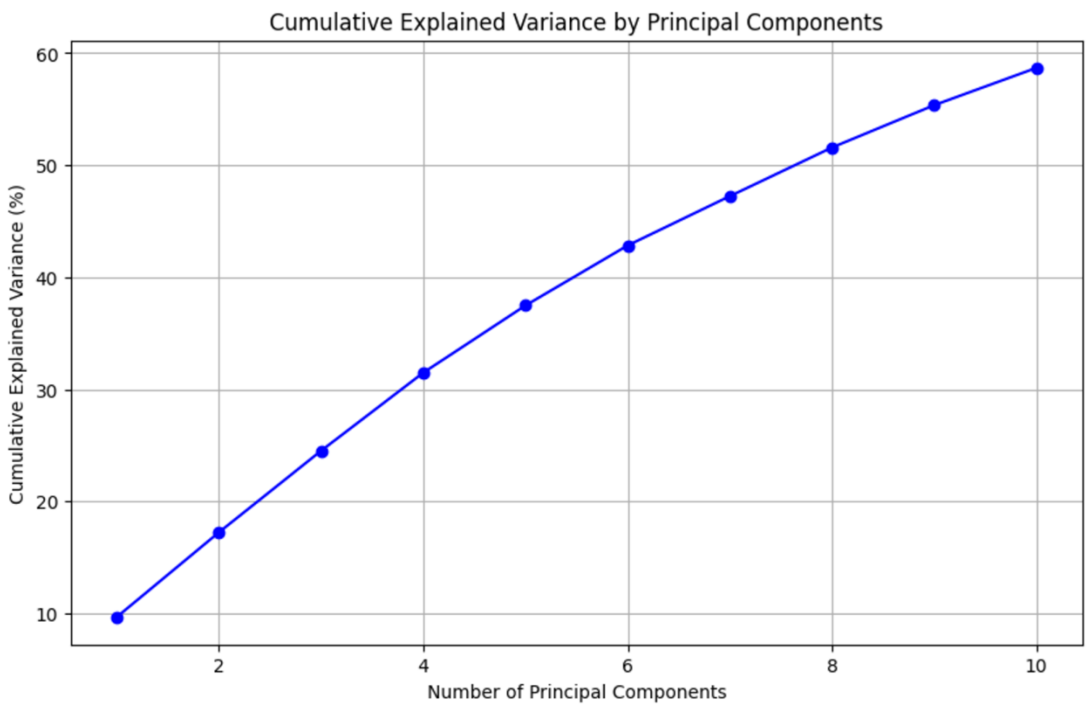
categories. Overall, the Random Forest classifier demonstrates its effectiveness in classifying victim ethnicity based on crime group data. However, further fine-tuning, feature engineering, or incorporating additional variables might enhance the model's performance and provide even deeper insights.

True labels are shown on the vertical axis, and predicted labels are on the horizontal axis. Most correct predictions are concentrated on the diagonal of the matrix, particularly for the class corresponding to True Label 4 (Predicted Label 4) with 33,135 correct predictions. Misclassifications are minimal but present:1,017 instances of True Label 2 were predicted as Label 4.1,434 instances of True Label 5 were predicted as Label 4. Classes with very low sample counts (e.g., Labels 0, 1, 3, 6, 7, and 8) show either no predictions or very minimal counts in both correct and incorrect predictions.

Overall, the model seems highly biased toward predicting Label 4, as most predictions fall in this category.

PRINCIPAL COMPONENT ANALYSIS

In the next step, we performed Principal Component Analysis (PCA) on the dataset to reduce dimensionality while retaining the maximum variance in the data. The graph above shows the cumulative explained variance as a function of the number of principal components. Each additional component contributes to the total variance captured, with the cumulative explained variance reaching around 60% after including 10 principal components. This indicates that these components collectively retain a significant portion of the data's variability, making PCA an effective preprocessing step. By using PCA, we aim to reduce computational complexity, eliminate redundant features, and enhance the performance of our models, especially when dealing with high-dimensional data. The curve also helps identify the optimal number of components to balance variance retention and dimensionality reduction



3. PRINCIPAL COMPONENT ANALYSIS – RANDOM FOREST

```
[ ]: from sklearn.decomposition import PCA
     import matplotlib.pyplot as plt
     label_encoders = {}
     for column in data.select_dtypes(include=['object']).columns:
         le = LabelEncoder()
         data[column] = le.fit_transform(data[column].astype(str))
         label_encoders[column] = le
```

```
[ ]: data.fillna(data.mean(), inplace=True)
```

```
[ ]: y = data['THEFT_CODE']
     X = data.drop(columns=['THEFT_CODE'])
```

```
[ ]: X_scaled = scaler.fit_transform(X)
```

```
[ ]: n_components = min(X.shape[1], 10)
     pca = PCA(n_components=n_components)
     X_pca = pca.fit_transform(X_scaled)
```

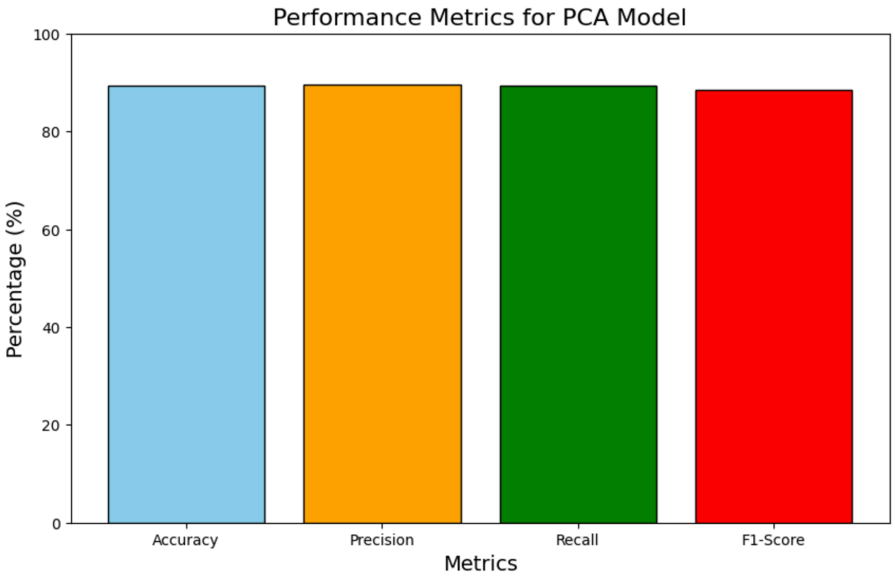
```
[ ]: X_train, X_test, y_train, y_test = train_test_split(X_pca, y, test_size=0.3, random_state=42)
```

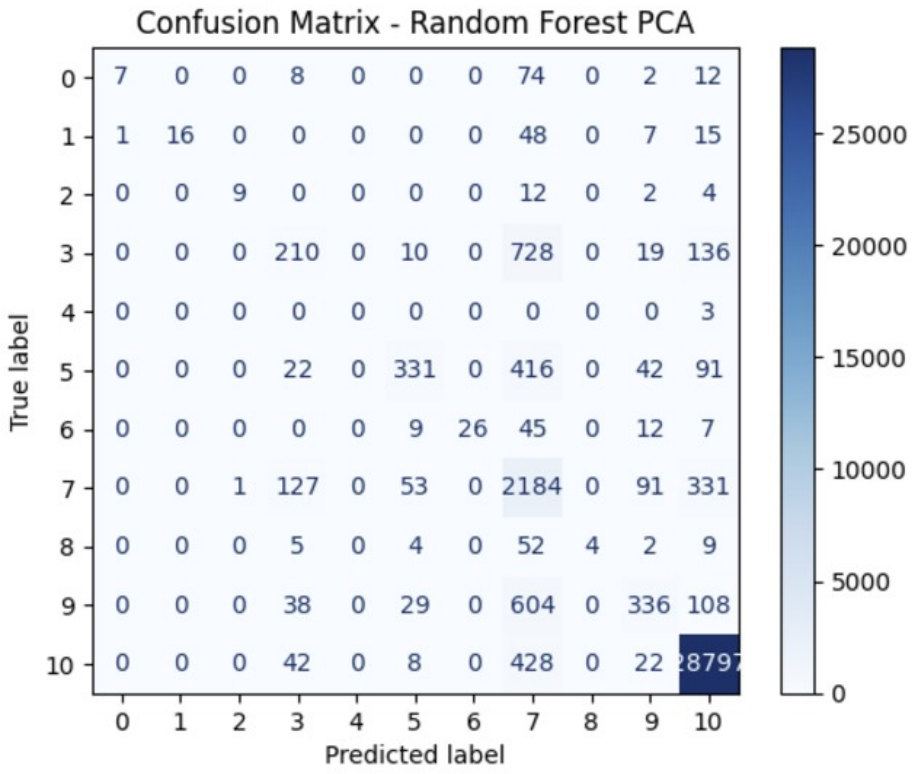
```
[ ]: from sklearn. ensemble import RandomForestClassifier
     rf = RandomForestClassifier(random_state=42)
     rf.fit(X_train, y_train)
```

The output shows the performance metrics for the Random Forest model trained on the PCA-transformed dataset. The model achieved an **accuracy of 89.37%**, indicating that nearly 9 out of 10 predictions were correct. The **precision of 89.57%** reflects the model's ability to minimize false positives and correctly identify true positives. The **recall of 89.37%** shows the model's capability to detect most of the actual positive cases, while the **F1-score of 88.43%** provides a balanced metric that considers both precision and recall.

it likely improved the computational efficiency and performance by reducing noise and redundancy in the features while retaining essential variance. This combination of PCA with Random Forest provides a robust approach for handling high-dimensional datasets effectively.

```
Accuracy: 89.37%
Precision: 89.57%
Recall: 89.37%
F1-Score: 88.43%

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.08      0.14       103
           1       1.00      0.20      0.33        87
           2       0.90      0.33      0.49        27
           3       0.45      0.18      0.26      1103
           4       0.00      0.00      0.00         3
           5       0.71      0.36      0.48       902
           6       0.96      0.23      0.37        99
           7       0.48      0.73      0.58      2787
           8       1.00      0.05      0.10        76
           9       0.65      0.32      0.43      1115
          10       0.97      0.98      0.98     29297

    accuracy                           0.89     35599
   macro avg       0.73      0.32      0.38     35599
weighted avg       0.90      0.89      0.88     35599
```





Diagonal dominance:

Correct predictions (values on the diagonal) are most significant for True Label 10 (Predicted Label 10), with 28,797 correct predictions, indicating strong performance for this class.

Other classes have fewer correct predictions, such as: True Label 7: 2,184 correct predictions. True Label 6: 728 correct predictions.

Misclassifications: Many misclassifications occur across different labels: True Label 6 is often confused with Labels 7, 8, and 10. True Label 7 has significant misclassifications, with 331 instances predicted as Label 10.

True Label 5 shows notable misclassifications across Labels 6, 7, and 10. Some rows (e.g., True Labels 0, 1, 2) show very low correct prediction counts and significant misclassification spread.

Class bias: Predictions are skewed heavily towards Label 10, which has the largest count overall, indicating a potential class imbalance.

Low performance on rare classes: Classes with True Labels 0, 1, 2, and 4 have very few correct predictions, indicating poor performance on these less frequent labels.

4. SUPPORT VECTOR MACHINE

```
[ ]: from sklearn.svm import SVC
     from sklearn.linear_model import SGDClassifier
     from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

[ ]: label_encoders = {}
     for column in data.select_dtypes(include=['object']).columns:
         le = LabelEncoder()
         data[column] = le.fit_transform(data[column].astype(str))
         label_encoders[column] = le

[ ]: X.fillna(X.median(), inplace=True)  # Fill numerical NaNs with median
     categorical_cols = X.select_dtypes(include=['object']).columns
     for col in categorical_cols:
         X[col].fillna("Unknown", inplace=True)

[ ]: y = data['TOTALSUSPECTS']
     X = data.drop(columns=['TOTALSUSPECTS'])

[ ]: data.fillna(data.mean(), inplace=True)

[ ]: scaler = StandardScaler()
     X_scaled = scaler.fit_transform(X)

[ ]: sgd_svc = SGDClassifier(loss='hinge', max_iter=1000, tol=1e-3, random_state=42)

[ ]: X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)

[ ]: sgd_svc.fit(X_train, y_train)
```

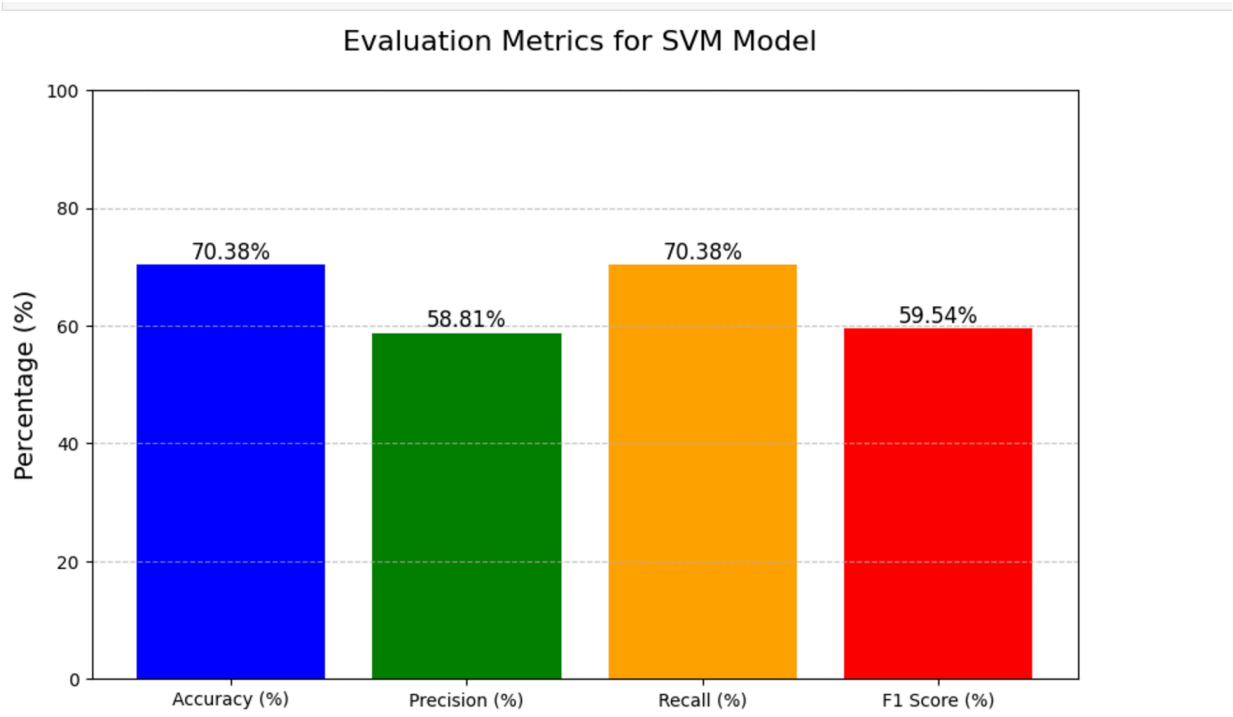
The code implements a Support Vector Machine (SVM) classifier using the SGDClassifier with a hinge loss function, which is commonly used for linear SVMs. The variable being predicted in this model is **TOTALSUSPECTS**, which represents the total number of suspects associated with each data point.

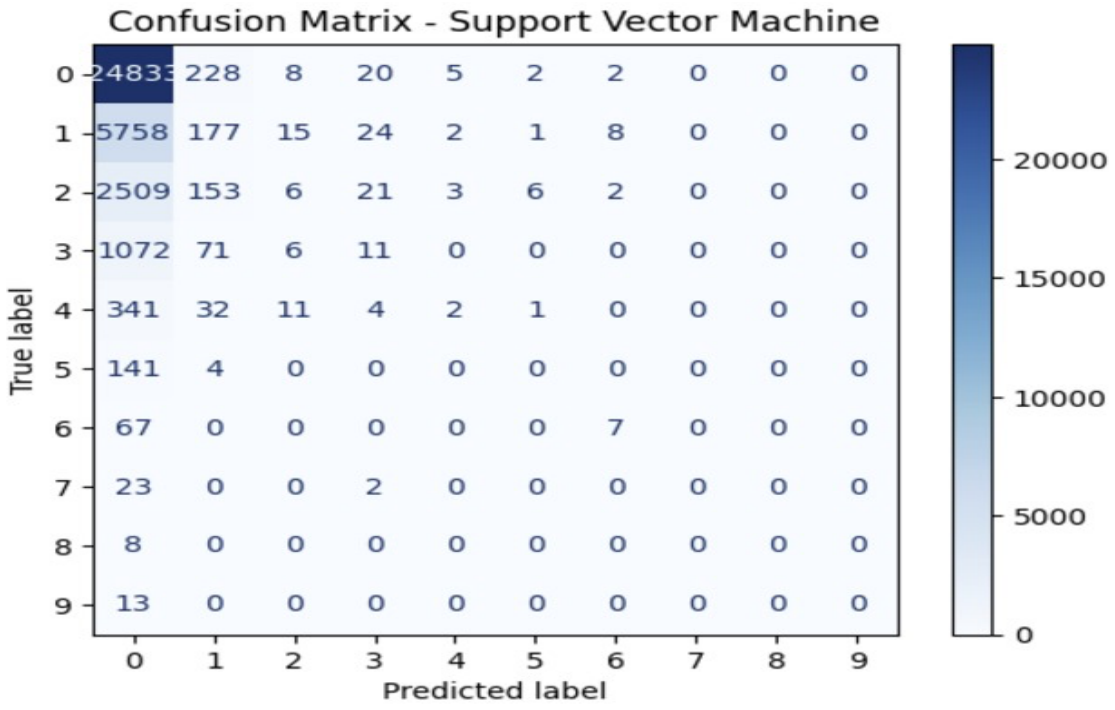
Predicting **TOTALSUSPECTS** is significant because it provides valuable insights into the patterns and characteristics associated with criminal incidents. By accurately forecasting the number of suspects involved in a case, law enforcement agencies can better allocate resources, prioritize investigations, and identify trends that could indicate larger issues, such as organized crime or recurring offenders. This information can also support data-driven decision-making in crime prevention strategies, helping policymakers address underlying factors contributing to higher numbers of suspects in certain areas or scenarios. Additionally, such predictions can enhance public safety by enabling more focused and effective interventions.

The preprocessing steps involve encoding categorical variables with LabelEncoder, filling missing numerical values with their median, and replacing missing categorical values with "Unknown." The target variable, TOTALSUSPECTS, is separated from the features, and all missing values in the dataset are replaced with the mean for numerical columns. The features are then standardized using StandardScaler to ensure that they are on the same

scale. Finally, the dataset is split into training and testing sets with a 70-30 split, and the SVM classifier is trained on the scaled training data.

The performance metrics show that the model achieved an **accuracy of 70.38%**, meaning it correctly predicted the total number of suspects in approximately 7 out of 10 cases. The **precision of 58.81%** reflects moderate success in minimizing false positives, while the **recall of 70.38%** indicates a good ability to identify most true cases. The **F1-score of 59.54%** provides a balanced measure of precision and recall. These results suggest reasonable performance but highlight opportunities for improvement through techniques like feature engineering or dataset balancing.





Correct Predictions (Diagonal Values): The largest number of correct predictions is for True Label 0, with 24,833 correctly classified instances. True Labels 1 and 2 also have relatively high correct predictions, with 5,758 and 2,509, respectively. Correct predictions for other labels drop significantly, with minimal correct classifications for labels like 3 through 9.

Misclassifications: Most misclassifications occur in the neighboring classes: For True Label 0, 228 instances are classified as Label 1, and 20 are classified as Label 3.For True Label 1, 177 instances are classified as Label 0, and 24 as Label 3. Misclassifications decrease significantly for labels with fewer samples (e.g., Labels 5, 6, 7, 8, and 9).

Class Imbalance or Bias: The model performs significantly better for dominant labels like 0, 1, and 2, suggesting potential class imbalance or insufficient separability for other labels. Rare or less-represented labels (5 through 9) have negligible correct predictions and are mostly misclassified as Label 0 or 1.

Model Performance: The overall trend shows that the model struggles to generalize for minor classes. The diagonal dominance is visible for major classes, but the confusion for less-represented classes highlights performance issues.

5.NAIVE BIAS

```
]: from sklearn.naive_bayes import GaussianNB
   from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report
   import matplotlib.pyplot as plt
```

```
]: label_encoders = {}
   for column in data.select_dtypes(include=['object']).columns:
       le = LabelEncoder()
       data[column] = le.fit_transform(data[column].astype(str))
       label_encoders[column] = le
```

```
]: data.fillna(data.mean(), inplace=True)

   y = data['THEFT_CODE']
   X = data.drop(columns=['THEFT_CODE'])
```

```
]: scaler = StandardScaler()
   X_scaled = scaler.fit_transform(X)
```

```
]: X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
```

```
]: nb = GaussianNB()
   nb.fit(X_train, y_train)
```

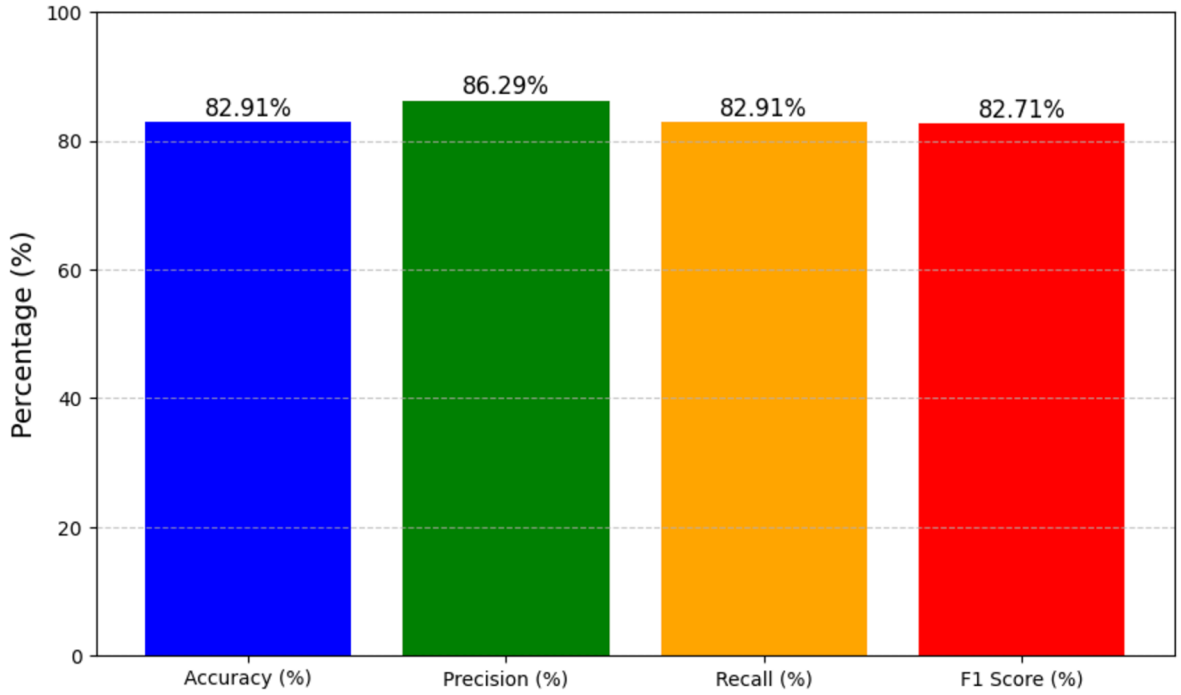
```
GaussianNB()
```

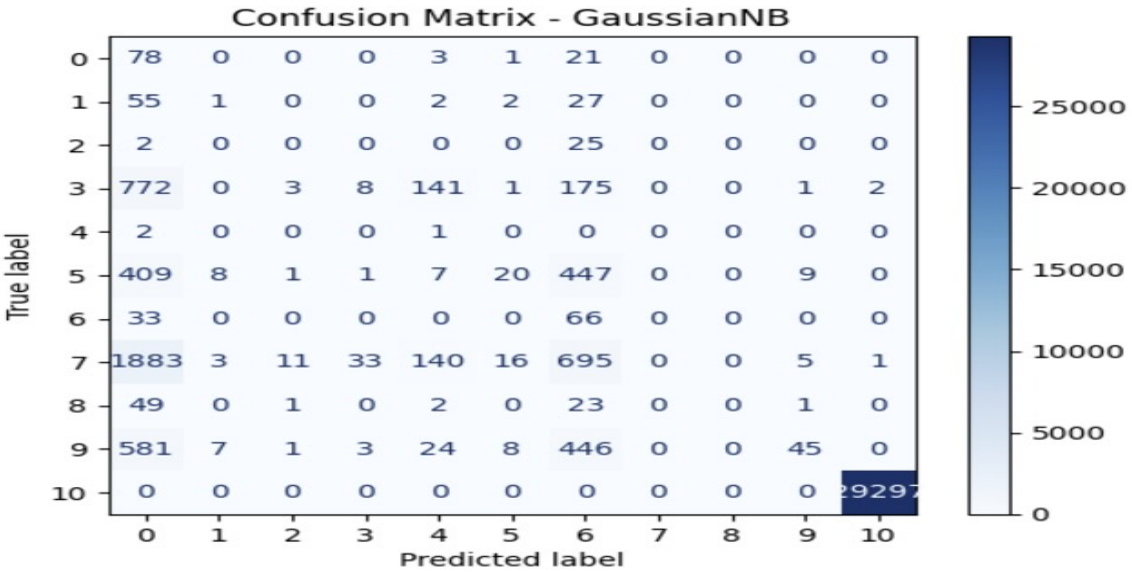
The variable being predicted by the Naive Bayes model is TOTALSUSPECTS, which represents the total number of suspects involved in each case. This prediction is significant because it provides valuable insights into crime patterns and allows law enforcement agencies to better understand the scale of incidents, allocate resources efficiently, and identify potential trends in criminal activity. By forecasting the number of suspects, this model can aid in strategic decision-making and help prioritize cases that may require more immediate attention.

Explanation

The Naive Bayes model achieved an accuracy of 82.91%, meaning it correctly predicted the number of suspects in approximately 83% of the test cases. The precision of 86.29% highlights the model's ability to avoid false positives, ensuring the predictions are specific and reliable. The recall of 82.91% indicates that the model successfully identified most true cases, capturing the majority of actual incidents involving suspects. The F1-score of 82.71% shows a balanced performance in terms of precision and recall, making the model robust for practical applications.

These results demonstrate that the Naive Bayes model effectively predicts TOTALSUSPECTS, providing a probabilistic approach to understanding and analyzing crime data. While the performance is strong, further optimization or additional features might improve the results even further, enhancing its utility for law enforcement and crime prevention efforts.





Strong Class 10 Performance: The model shows perfect classification for class 10 (29297 correct predictions), which appears to be the majority class based on the color intensity.

Major Misclassification Patterns: Class 7 has significant misclassifications: 1883 instances were incorrectly predicted as class 0, and 695 as class 6Class 3 also shows notable confusion: 772 instances predicted as class 0, and 175 as class 6. Class 5 has 409 instances misclassified as class 0, and 447 as class 6.

Common Error Pattern: Classes 0 and 6 appear to be frequent incorrect predictions across multiple true classes, suggesting the model might be biased towards these classes.

Weak Performance Areas: Classes 7 and 8 show very few correct predictions The model seems to rarely predict classes 7, 8, and 9, as indicated by the mostly zeros in these columns.

Diagonal Performance: The diagonal elements (representing correct predictions) are notably weaker for most classes compared to the misclassifications, indicating suboptimal overall performance except for class 10.

6. XGBOOST

[ ]:
```
from xgboost import XGBRegressor
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report
import matplotlib.pyplot as plt
```

[ ]:
```
label_encoders = {}
for column in data.select_dtypes(include=['object']).columns:
    if column != 'ZIP':  # Don't encode the target
        le = LabelEncoder()
        data[column] = le.fit_transform(data[column].astype(str))
        label_encoders[column] = le

# Handle missing values
data.fillna(data.mean(), inplace=True)

# Step 2: Select ZIP as the target variable
y = data['SNA_NEIGHBORHOOD']  # Use the 'ZIP' column as the target
X = data.drop(columns=['SNA_NEIGHBORHOOD'])  # Drop the target column from features

# Step 3: Check if ZIP needs label encoding
if y.dtype == 'object' or isinstance(y[0], str):
    le = LabelEncoder()
    y = le.fit_transform(y)  # Encode target as numeric if necessary

# Step 4: Standardize features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Step 5: Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
```

```
from sklearn.metrics import mean_absolute_error, mean_squared_error
xgb_regressor = XGBRegressor(random_state=42)
xgb_regressor.fit(X_train, y_train)

# Step 7: Evaluate the Model
y_pred = xgb_regressor.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)

print(f"XGBoost Model Mean Absolute Error (MAE): {mae:.2f}")
print(f"XGBoost Model Mean Squared Error (MSE): {mse:.2f}")
print(f"XGBoost Model Root Mean Squared Error (RMSE): {rmse:.2f}")
```

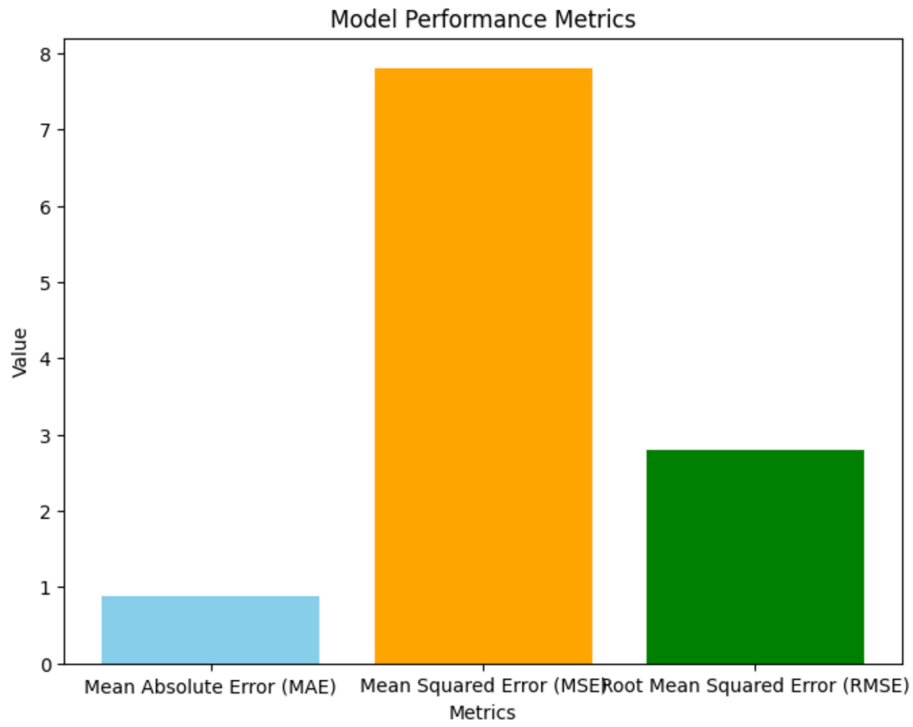
```
XGBoost Model Mean Absolute Error (MAE): 0.88
XGBoost Model Mean Squared Error (MSE): 7.81
XGBoost Model Root Mean Squared Error (RMSE): 2.79
```

The code implements the XGBoost Regressor model to predict the target variable SNA_NEIGHBORHOOD, which likely represents a specific category or neighborhood-related feature. Predicting this variable is significant as it can help uncover localized patterns, trends, or relationships in the data. For example, understanding neighborhood-specific metrics could aid in better resource allocation, identifying hotspots for intervention, or gaining insights into regional variations that inform decision-making processes.

The data preprocessing involves several key steps to ensure the dataset is clean and ready for modeling. Categorical variables are label-encoded to transform them into numerical formats suitable for machine learning models. Missing numerical values are filled with their mean to maintain data completeness and prevent training issues. The SNA_NEIGHBORHOOD column is designated as the target variable, while the remaining columns serve as input features. Standardization is then applied to scale all numerical features, ensuring that the model processes features on a uniform scale, which is critical for gradient-boosting algorithms like XGBoost. The dataset is then split into training and testing sets, with 70% of the data used for training and 30% reserved for evaluation.

After training the XGBoost Regressor on the scaled training data, the model's performance is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The MAE of 0.88 indicates that the predictions, on average, deviate by less than one unit from the actual values. The MSE of 7.81 measures the squared average error, penalizing larger deviations more heavily, while the RMSE of 2.79 provides an interpretable measure of error in the same units as the target variable. These metrics demonstrate that the XGBoost model effectively captures patterns in the data with relatively low error.

Overall, the ability to predict SNA_NEIGHBORHOOD with this level of accuracy highlights the value of the model in providing actionable insights for neighborhood-specific analysis. The robustness of XGBoost, combined with effective preprocessing, ensures that the model can handle complex relationships and feature interactions.



7. LSTM

The code implements an LSTM (Long Short-Term Memory) model to predict the target variable HATE_BIAS, which categorizes hate-related biases. This prediction is highly significant as it provides insights into patterns of hate-related incidents, enabling law enforcement agencies and policymakers to identify and address targeted discrimination and violence effectively. By understanding these biases, proactive measures can be implemented to foster safer, more inclusive communities and improve support systems for those affected.

The dataset is preprocessed by separating the HATE_BIAS column as the target variable, while the remaining columns serve as features. The data is split into training and testing sets, with 80% used for training and 20% for testing. Since LSTM models require 3D input, the datasets are reshaped into the required format with dimensions of samples, time steps, and features. The target variable is also converted into categorical format, corresponding to the number of unique hate bias categories, to enable multi-class classification.

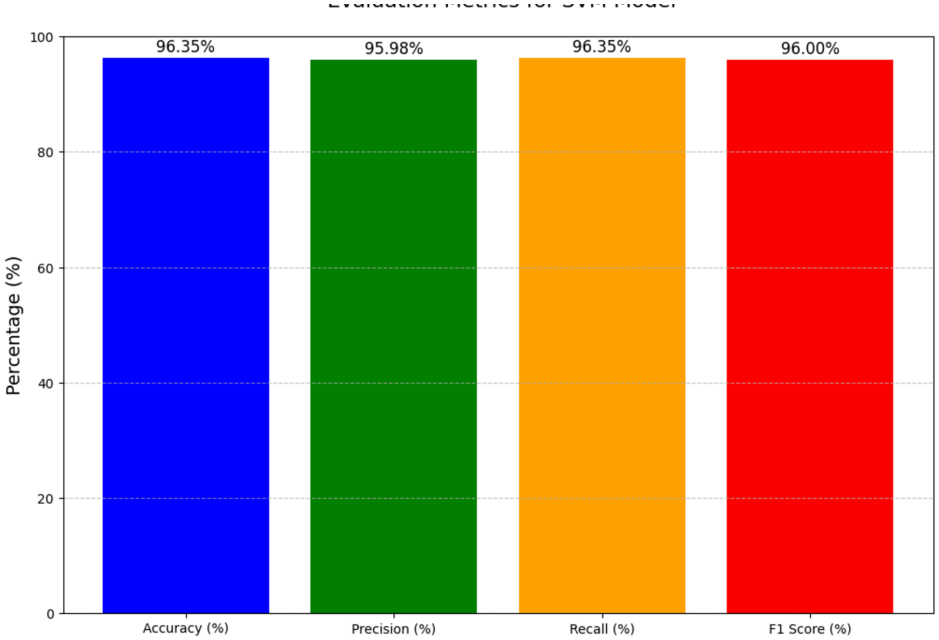
The LSTM model architecture includes an LSTM layer with 64 units and ReLU activation to capture temporal dependencies, followed by dropout layers to reduce overfitting. Dense layers handle the classification task, with the final output layer using softmax activation to predict probabilities for each hate bias category. The model is compiled with the adam optimizer and the categorical_crossentropy loss function and is trained for 10 epochs with a batch size of 64, using the test data for validation.

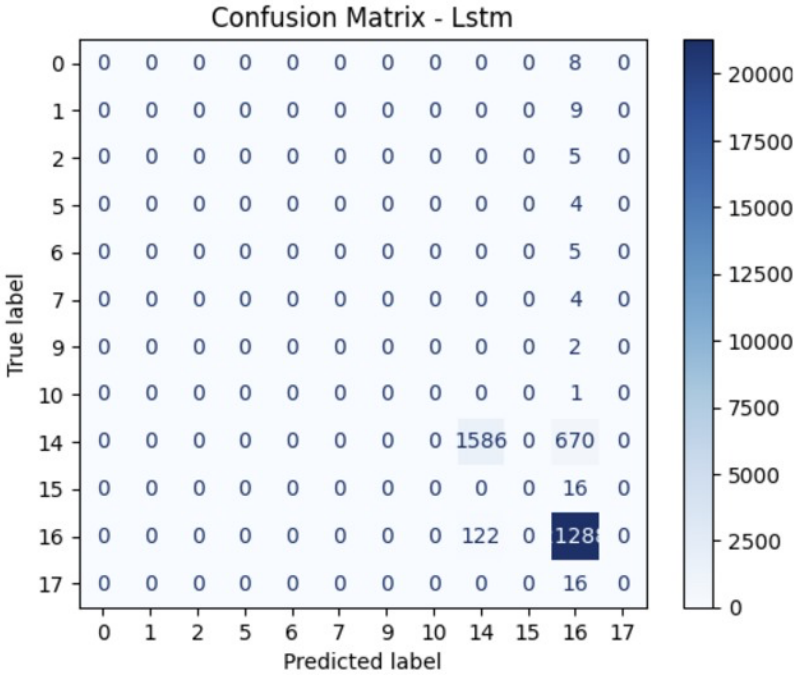
The model performed exceptionally well, achieving an accuracy of 96.35%, a precision of 95.98%, a recall of 96.35%, and an F1-score of 95.99%, indicating a strong balance between precision and recall. These metrics demonstrate that the model can effectively predict hate bias categories with high reliability. The use of LSTM ensures that temporal patterns and relationships in the data are captured, making it particularly suitable for sequential datasets. The results highlight the potential of this approach to provide actionable insights, supporting efforts to reduce hate-related incidents and improve community safety.

```
# Prepare features (X) and target (y)
X = data_subset.drop(columns=['HATE_BIAS'])
y = data_subset['HATE_BIAS']

# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Reshape the data for LSTM input: (samples, time_steps, features)
time_steps = 1  # LSTM input requires 3D data
X_train_lstm = np.reshape(X_train.values, (X_train.shape[0], time_steps, X_train.shape[1]))
X_test_lstm = np.reshape(X_test.values, (X_test.shape[0], time_steps, X_test.shape[1]))

# Convert the target to categorical format for classification
num_classes = len(np.unique(y))  # Total number of hate bias categories
y_train_cat = to_categorical(y_train, num_classes)
y_test_cat = to_categorical(y_test, num_classes)

# Define the LSTM model architecture
model = Sequential([
    LSTM(64, input_shape=(time_steps, X_train.shape[1]), activation='relu', return_sequences=False),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dropout(0.2),
    Dense(num_classes, activation='softmax')  # Output layer for classification
])

# Compile the model
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])

# Train the model
history = model.fit(
    X_train_lstm, y_train_cat,
    epochs=10,  # Adjust the epochs as needed
    batch_size=64,
    validation_data=(X_test_lstm, y_test_cat)
)
```

```
]: results = {
    "Accuracy (%)": accuracy,
    "Precision (%)": precision,
    "Recall (%)": recall,
    "F1 Score (%)": f1
}
print(results)
```

```
{'Accuracy (%)': 96.35153353555779, 'Precision (%)': 95.97671223840118, 'Recall (%)': 96.35153353555779, 'F1 Score (%)': 95.9982270634602
8}
```





Label Distribution: The matrix shows predictions across labels 0-17 (with some numbers missing in the sequence). The model appears to heavily favor predicting class 16, followed by class 14

Major Patterns: Class 14 shows mixed performance: 1586 correct predictions (class 14), but 670 instances misclassified as class 16. Class 16 has 1283 correct predictions, with 122 instances misclassified as class 14

Most other classes (0-10) are consistently misclassified as class 16, though with relatively small numbers (1-9 instances each)

Model Behavior: The model makes predictions primarily in classes 14 and 16. There's a clear binary confusion between classes 14 and 16. Many classes receive zero predictions, indicating the model isn't utilizing its full output space

8. GRADIENT BOOSTING

```python
def build_and_train_model_with_metrics(data, target_column):
    data, target_encoder = preprocess_data(data, target_column)

    X = data.drop(columns=[target_column])
    y = data[target_column]

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)

    model = GradientBoostingClassifier(random_state=42, n_estimators=50)
    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    cm = confusion_matrix(y_test, y_pred)

    print(f"Performance on {target_column}:")
    print(classification_report(y_test, y_pred))
    print(f"Accuracy: {accuracy}")

    plt.figure(figsize=(25, 17))
    sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=target_encoder.classes_, yticklabels=target_encoder.classes_)
    plt.title(f"Confusion Matrix for {target_column}")
    plt.xlabel("Predicted")
    plt.ylabel("Actual")
    plt.show()

    return model, target_encoder
```

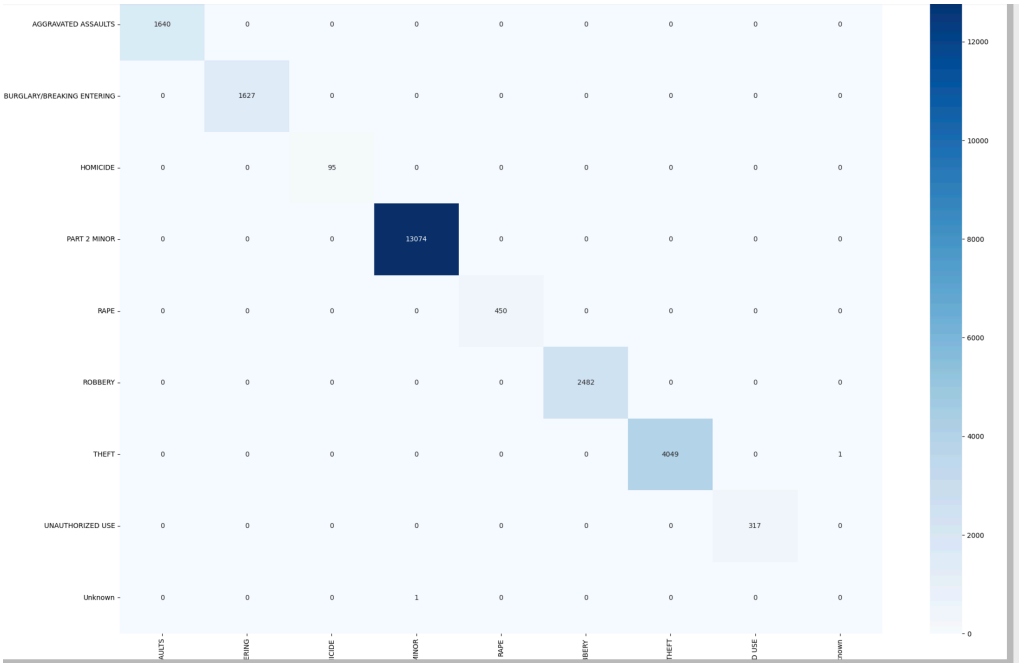
The code implements a Gradient Boosting Classifier to predict the target variable UCR_GROUP, which classifies crimes into predefined categories based on the Uniform Crime Reporting (UCR) system. This classification is vital for organizing and analyzing crime data, allowing law enforcement and policymakers to better understand and respond to various types of criminal activities.

The function build_and_train_model_with_metrics first preprocesses the dataset by separating features (X) and the target variable (y), in this case, UCR_GROUP. The data is split into training and testing sets (80-20 split) to train the model and evaluate its

performance. Standard scaling is applied to the features to normalize them, ensuring consistent input ranges for the model.

The Gradient Boosting Classifier, configured with 50 estimators and a fixed random_state=42, is trained on the scaled training data. Gradient Boosting operates by building an ensemble of weak learners (decision trees) iteratively, improving classification performance with each step by focusing on misclassified samples.

```
Predicting UCR_GROUP...
Performance on UCR_GROUP:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1640
           1       1.00      1.00      1.00      1627
           2       1.00      1.00      1.00        95
           3       1.00      1.00      1.00     13074
           4       1.00      1.00      1.00       450
           5       1.00      1.00      1.00      2482
           6       1.00      1.00      1.00      4050
           7       1.00      1.00      1.00       317
           8       0.00      0.00      0.00         1

    accuracy                           1.00     23736
   macro avg       0.89      0.89      0.89     23736
weighted avg       1.00      1.00      1.00     23736
```



After training, the model makes predictions on the test set. The predicted results are compared with the actual labels using metrics such as accuracy, macro-average precision, recall, and F1-score. The confusion matrix, visualized as a heatmap, provides a detailed view of true positives, false positives, and false negatives for each crime category. The results demonstrate an accuracy of 99.99%, with a weighted average of 1.00 for precision, recall, and F1-score, indicating the model is almost perfect in its predictions.

Predicting UCR_GROUP using the Gradient Boosting Classifier has substantial implications. It automates the classification of crimes, saving time and reducing human error. The high accuracy of the model ensures reliable categorization, which is crucial for identifying trends, allocating resources, and devising crime prevention strategies. However, the near-perfect accuracy might indicate potential overfitting, suggesting that further evaluation on unseen data is necessary to confirm the model's generalizability. Nonetheless, this model provides a robust framework for enhancing data-driven decision-making in crime analysis and reporting.