IS507 - DATA, STATISTICAL MODELS AND INFORMATION

PREDICTIVE MODELING IN CRIME DATA ANALYSIS

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

CONTENTS

DATASET OVERVIEW

The dataset appears to be a ***record of crime incidents, with each row representing a unique case***.
Here are the columns and some of their respective data:

1. **CASE**: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident
2. **DATE OF OCCURRENCE**: Date and time when the crime occurred.
3. **BLOCK**: The partially redacted address where the incident occurred.
4. **IUCR:**The Illinois Uniform Crime Reporting code. There are around 400 IUCR codes used by state of Illinois
5. **PRIMARY DESCRIPTION**: The primary description of the IUCR code
6. **SECONDARY DESCRIPTION**: he secondary description of the IUCR code, a subcategory of the primary description
7. **LOCATION DESCRIPTION**: Description of the location where the crime occurred.
8. **ARREST**: Indicates if an arrest was made (Y/N)
9. **DOMESTIC**: Indicates if the crime was domestic-related (Y/N).
10. **BEAT :** Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car.

DATASET OVERVIEW

11. **DISTRICT** : Indicates the police district where the incident occurred.
12. **FBI CODE**: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
13. **X COORDINATE, Y COORDINATE:** Geographic coordinates.
14. **LATITUDE, LONGITUDE**: Latitude and longitude of the incident.
15. **LOCATION**: A tuple combining latitude and longitude.
16. **WARD**: The ward (City Council district) where the incident occurred.

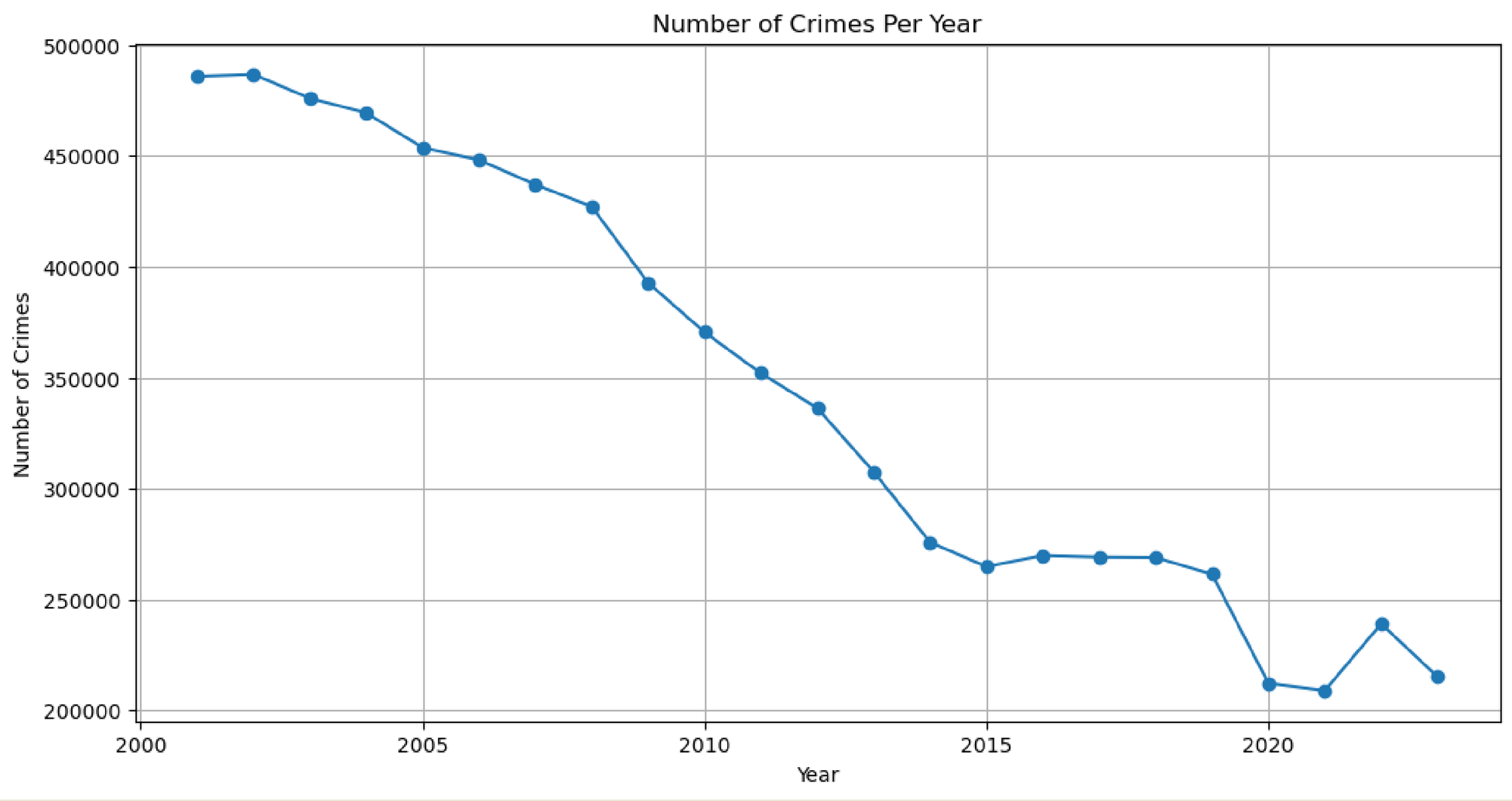
DATASET OVERVIEW

	ID	Case.Number	Date	Block	IUCR	Primary.Type	Description	Location.Descrip
12	12045583	JD226426	05/07/2020 10:24:00 AM	035XX S INDIANA AVE	0820	THEFT	$500 AND UNDER	APARTMENT
13	12031001	JD209965	04/16/2020 05:00:00 AM	005XX W 32ND ST	0460	BATTERY	SIMPLE	APARTMENT
14	12093529	JD282112	07/01/2020 10:16:00 AM	081XX S COLES AVE	051A	ASSAULT	AGGRAVATED - HANDGUN	STREET
15	12178140	JD381597	09/27/2020 11:29:00 PM	065XX S WOLCOTT AVE	0460	BATTERY	SIMPLE	RESIDENCE - POR
16	4144897	HL474854	07/10/2005 03:00:00 PM	062XX S ABERDEEN ST	0430	BATTERY	AGGRAVATED: OTHER DANG WEAPON	STREET
20	12126129	JD321064	08/04/2020 08:28:00 PM	081XX S LOOMIS BLVD	143A	WEAPONS VIOLATION	UNLAWFUL POSSESSION - HANDGUN	STREET
27	12010314	JD186932	03/15/2020 09:00:00 PM	051XX W HURON ST	2820	OTHER OFFENSE	TELEPHONE THREAT	APARTMENT
28	12067286	JD251718	06/02/2020 10:00:00 PM	042XX S EMERALD AVE	0820	THEFT	$500 AND UNDER	STREET
29	4229528	HL545852	08/12/2005 11:00:00 PM	063XX S COTTAGE GROVE AVE	3730	INTERFERENCE WITH PUBLIC OFFICER	OBSTRUCTING JUSTICE	SIDEWALK
35	12163191	JD364357	09/11/2020 10:44:00 PM	008XX N TRUMBULL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE
36	12142075	JD339709	08/21/2020 12:00:00 AM	015XX N MILWAUKEE AVE	0420	BATTERY	AGGRAVATED - KNIFE / CUTTING INSTRUMENT	SIDEWALK
42	12005594	JD181480	03/11/2020 12:00:00 AM	040XX W 26TH ST	0860	THEFT	RETAIL THEFT	SMALL RETAIL ST
43	12070567	JD249245	06/01/2020 01:20:00 AM	024XX N MILWAUKEE AVE	0620	BURGLARY	UNLAWFUL ENTRY	SMALL RETAIL ST
44	12057551	JD240361	05/23/2020 12:20:00 PM	068XX S LOOMIS BLVD	0486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET
45	12057185	JD239952	05/22/2020 10:29:00 PM	0000X N LOREL AVE	2024	NARCOTICS	POSSESS - HEROIN (WHITE)	STREET
46	12107263	JD298884	07/15/2020 10:15:00 PM	070XX S CHAPPEL AVE	0560	ASSAULT	SIMPLE	APARTMENT
63	12178347	JD382022	09/28/2020 10:23:00 AM	038XX W JACKSON BLVD	2820	OTHER OFFENSE	TELEPHONE THREAT	OTHER (SPECIFY)
64	12164381	JD361415	09/09/2020 09:23:00 AM	081XX S VERNON AVE	0560	ASSAULT	SIMPLE	RESIDENCE
65	12072697	JD257284	06/06/2020 07:45:00 AM	065XX S UNIVERSITY AVE	2820	OTHER OFFENSE	TELEPHONE THREAT	OTHER (SPECIFY)
66	12126191	JD321188	08/05/2020 01:00:00 AM	025XX S MICHIGAN AVE	0454	BATTERY	AGGRAVATED P.O. - HANDS, FISTS, FEET, NO / MINO...	HOSPITAL BUILDI
67	12018869	JD196508	03/27/2020 11:20:00 PM	066XX S SPAULDING AVE	141A	WEAPONS VIOLATION	UNLAWFUL USE - HANDGUN	STREET
68	12068919	JD253635	06/04/2020 02:15:00 PM	010XX W DIVISION ST	2820	OTHER OFFENSE	TELEPHONE THREAT	OTHER (SPECIFY)
69	12044629	JD225380	05/05/2020 10:00:00 PM	079XX S THROOP ST	0810	THEFT	OVER $500	STREET
70	12157580	JD357987	09/05/2020 11:50:00 PM	022XX W MAYPOLE AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	CHA APARTMENT
71	12045468	JD225478	05/05/2020 10:00:00 PM	104XX S AVENUE M	0910	MOTOR VEHICLE THEFT	AUTOMOBILE	DRIVEWAY - RESI
72	12132877	JD329246	08/12/2020 02:20:00 AM	102XX S DR MARTIN LUTHER KING JR DR	1310	CRIMINAL DAMAGE	TO PROPERTY	RESIDENCE
73	12097474	JD287186	07/05/2020 05:18:00 PM	007XX N CENTRAL AVE	0454	BATTERY	AGGRAVATED P.O. - HANDS, FISTS, FEET, NO / MINO...	STREET
74	12005082	JD180933	03/10/2020 03:40:00 PM	016XX W 63RD ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET
75	12007650	JD183972	03/12/2020 08:00:00 PM	092XX S LUELLA AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET
76	12153480	JD353103	08/20/2020 10:00:00 AM	087XX S STATE ST	1120	DECEPTIVE PRACTICE	FORGERY	BANK
78	12038537	JD218490	04/27/2020 10:00:00 AM	048XX W HIRSCH ST	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	RESIDENCE
79	12028031	JD204153	04/07/2020 04:50:00 PM	064XX S DR MARTIN LUTHER KING JR DR	0420	BATTERY	AGGRAVATED - KNIFE / CUTTING INSTRUMENT	PARKING LOT / G

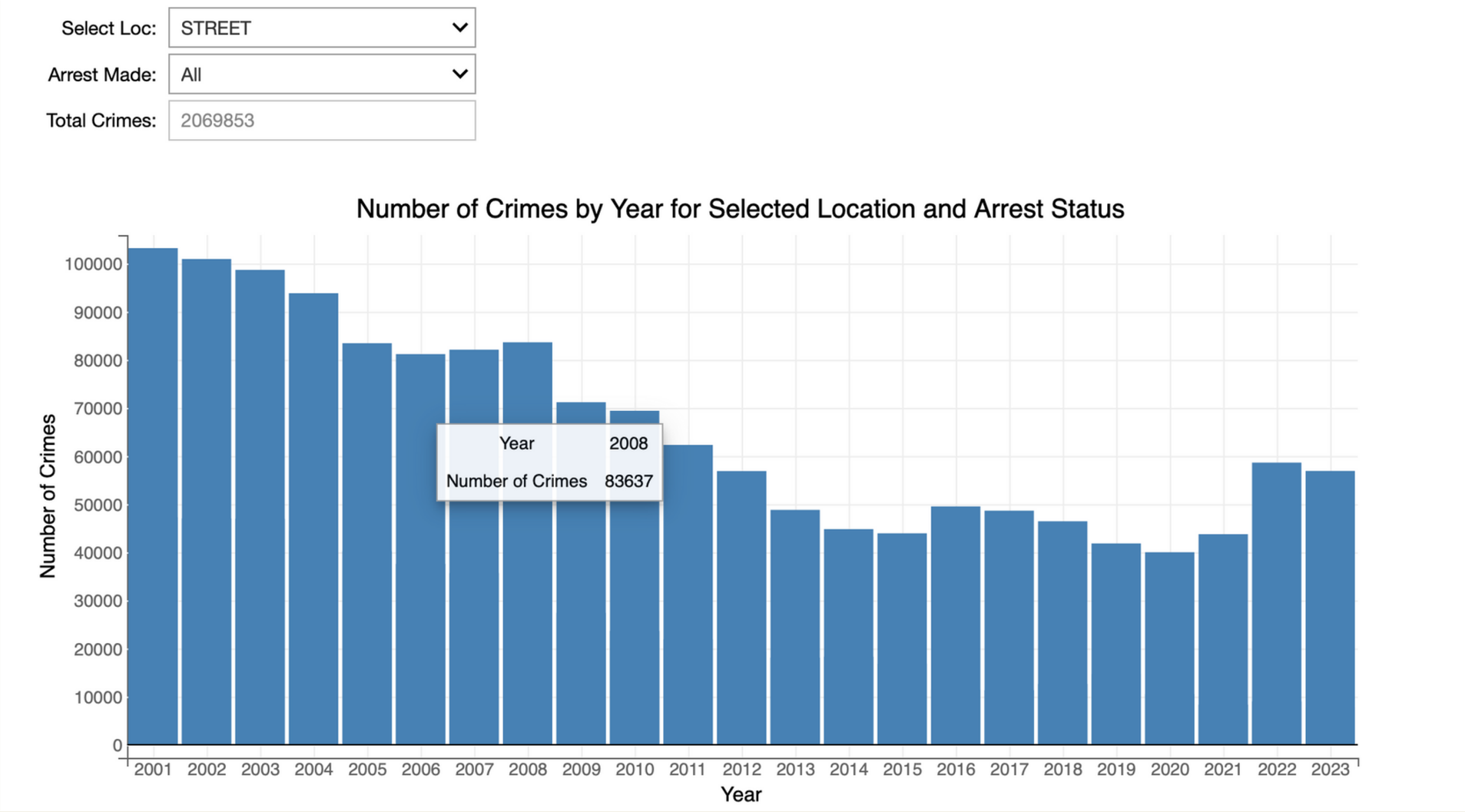
Showing 1 to 32 of 7,232,620 entries, 44 total columns

DATA EXPLORATION

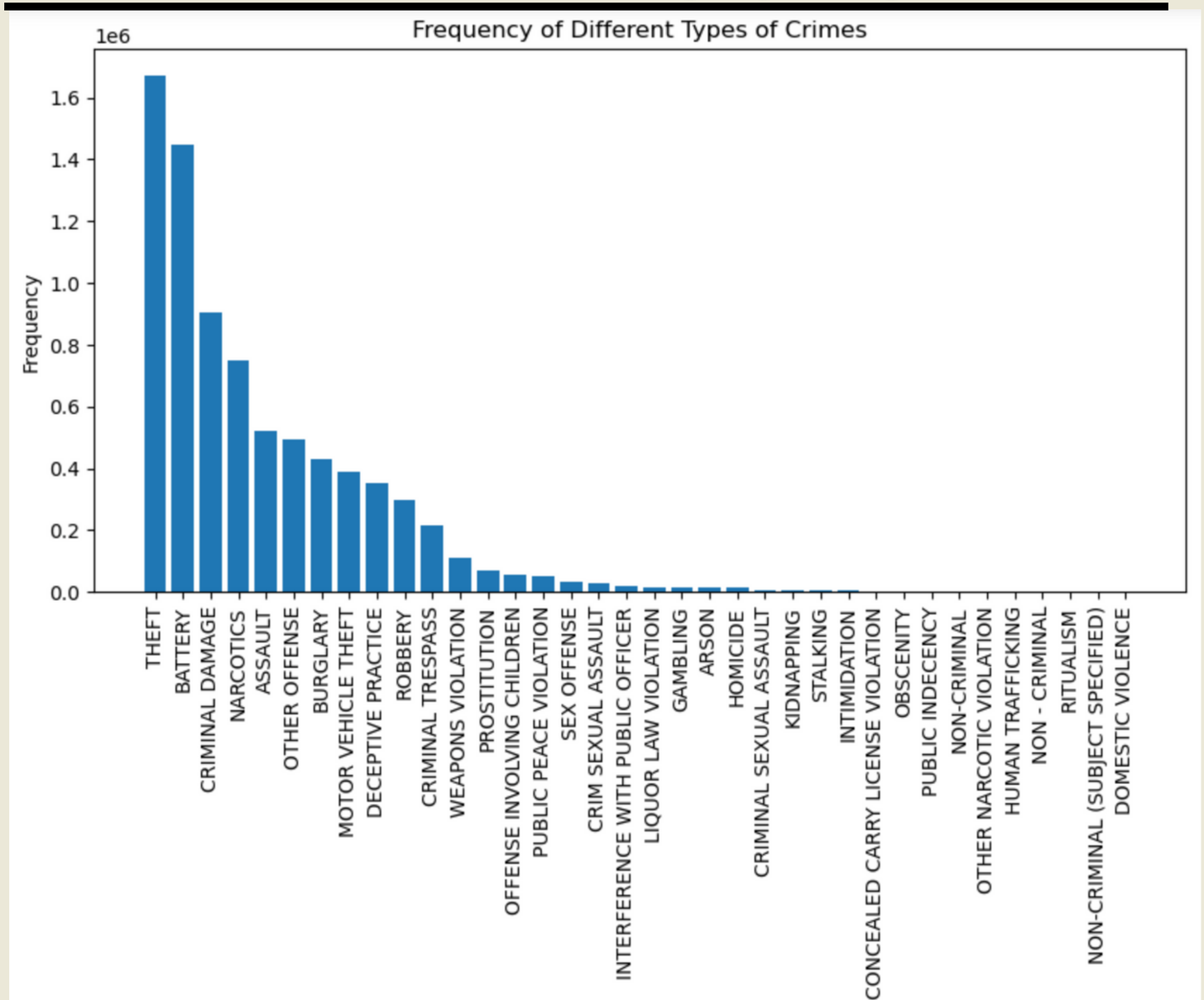
NUMBER OF CRIMES PER YEAR



LOCATION WITH MAXIMUM CRIMES- STREETS

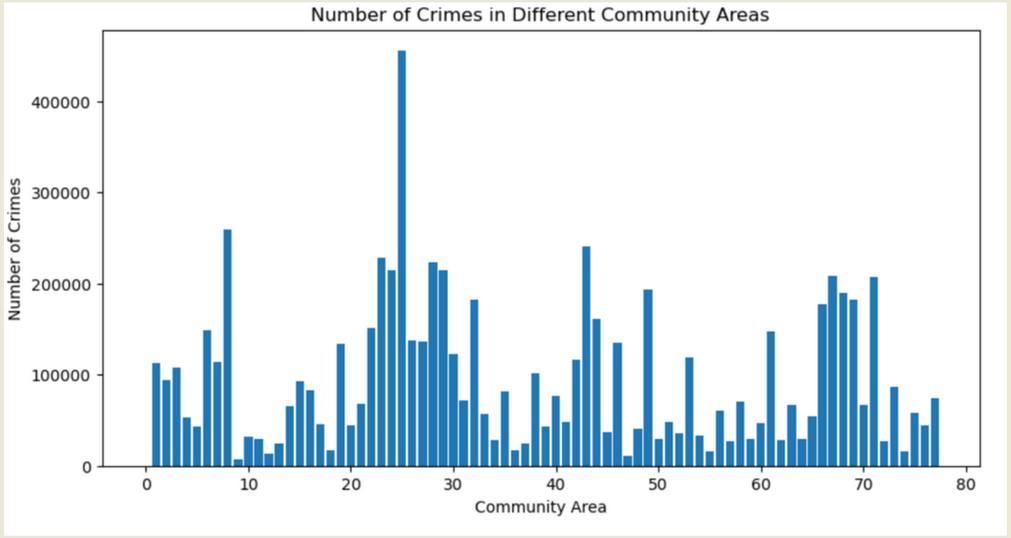


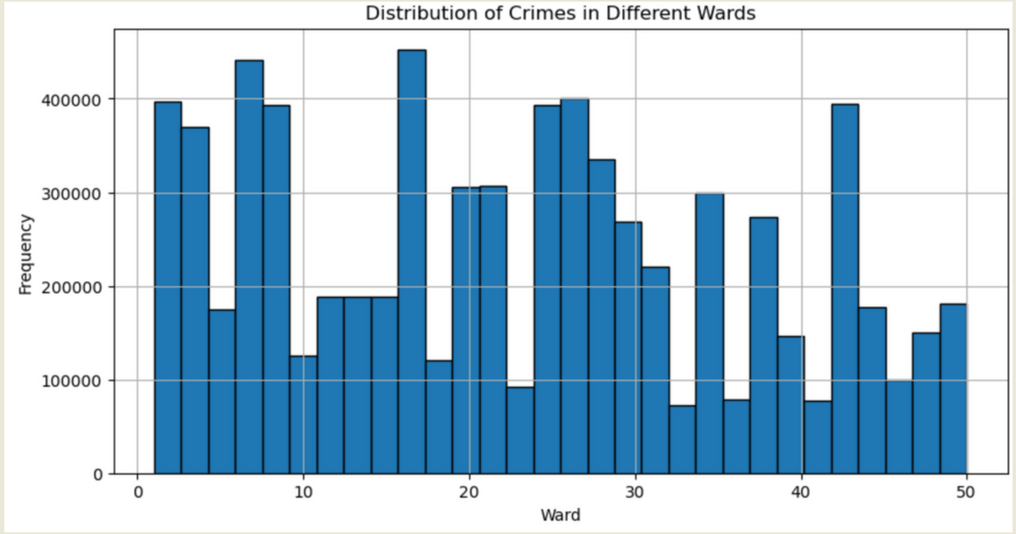
DIFFERENT TYPES OF CRIMES



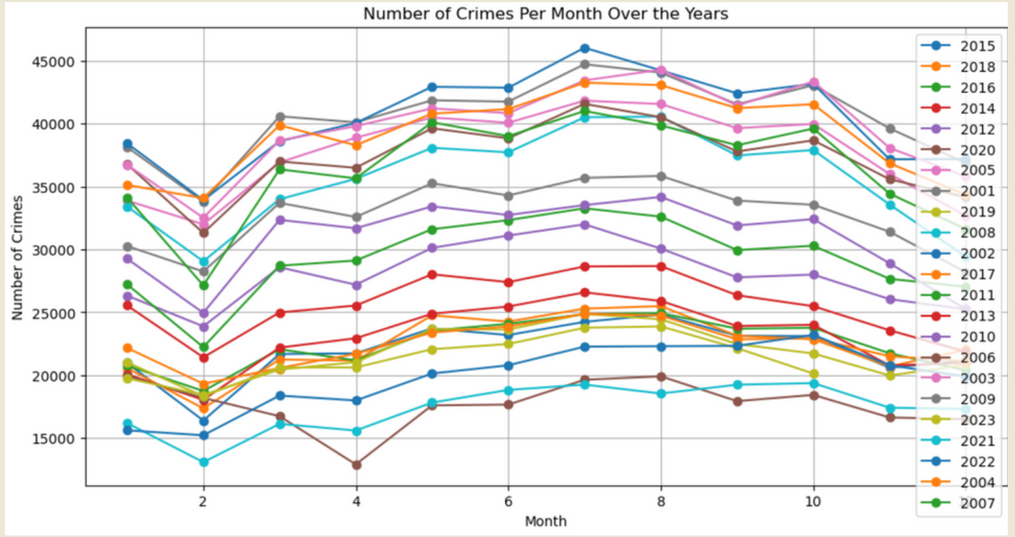
OTHER VISUALISATIONS

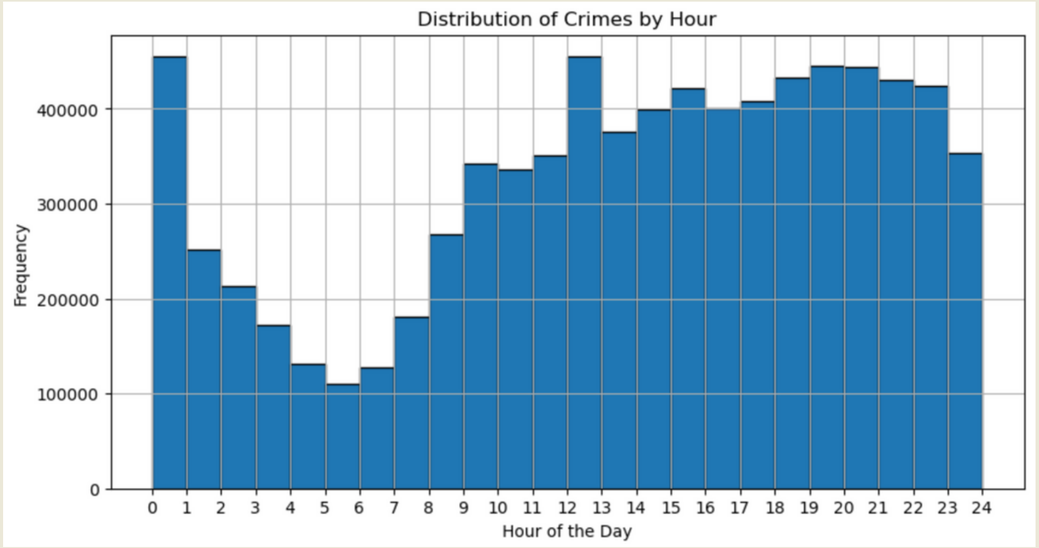
COMMUNITY AREA AND WARD





MONTH AND YEAR





DATA PRE-PROCESSING

- ## IMPUTING AND CLEANING MISSING VALUES

  Imputing missing NULL values with NA across the dataset, excluding the LOCATION DESCRIPTION column., followed by removal of rows containing these NA values

- ## COLUMN SEGREGATION

  Extracting and segregating information from the BLOCK column into distinct columns for ZIPCODE, DIRECTION, STREET AVENUE in the database.

DATA PRE-PROCESSING

- ***REMOVING DATA POINTS OUTSIDE THE BOUNDING BOX OF CHICAGO***

  Eliminated data points outside the Chicago city bounds, restricting coordinates to the geographical box: 41.6439,-87.9401 (southwest) to 41.9437,-87.58782 (northeast)."

- ***REMOVE DUPLICATES BASED ON THE 'CASE ID' :***

  By eliminating redundant entires, the dataset is streamlined to ensure accurate and non-repetitive representation of cases

ENCODING

WHY WE USED LABEL ENCODING:

- **Space Efficiency**
- **Simplicity and Interpretability**

ENCODING

Converting rows to factors:

filtered_df$IUCR <- as.factor(filtered_df$IUCR)

Converting factors to numericals

df_dataset$IUCR_Encoded <- as.numeric(filtered_df$IUCR)

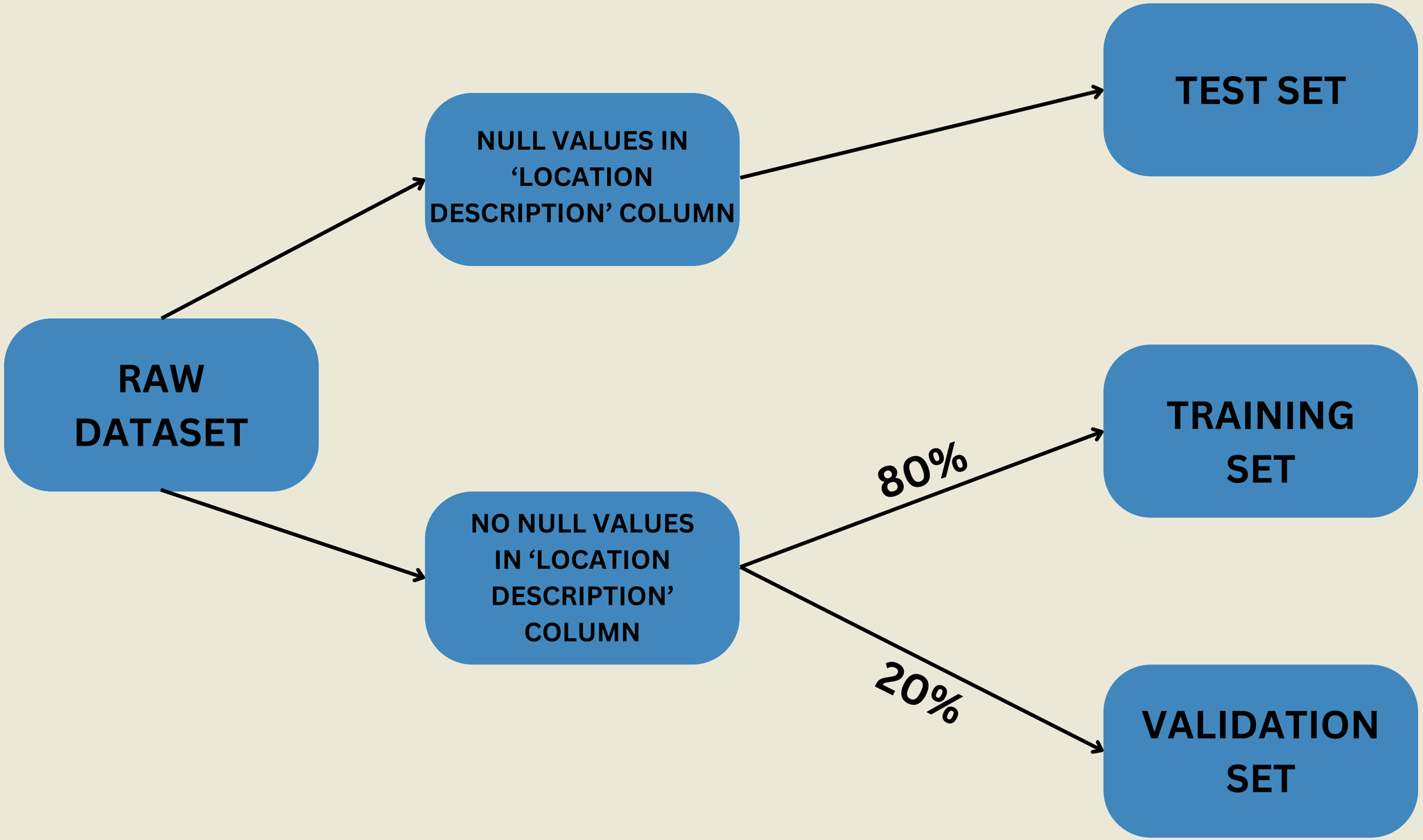
FEATURE SCALING

- Feature scaling is a method used to standardize the range of independent variables or features of data.

- scale(): This function is used to perform z-score normalization on the selected columns.

IUCR_Encoded	Primary.Type_Encoded	Description_Encoded	Arrest_Encoded	Domestic_Encoded	Beat_Encoded	District_Encoded	Community.Area_Encoded	FBI.Code_Encoded
-0.57705247	-1.160888289	0.87567943	-0.6344746	-0.4874732	0.401729028	0.38238716	-0.87027769	-0.4872719
0.38411460	-0.710594680	1.13971691	-0.6344746	-0.4874732	-0.700815204	-0.63153496	0.52505810	0.5767167
0.38411460	-0.710594680	1.13971691	-0.6344746	2.0513683	-0.396665071	-0.42875054	0.15786447	0.5767167
0.74111951	0.820403589	0.91968567	1.5760869	-0.4874732	-1.252087321	-1.23988824	-0.13589043	0.8807134
-1.03017409	-1.070829567	0.87567943	1.5760869	2.0513683	1.561301411	1.59909370	-1.45778750	-0.3352735
-0.82420972	-1.070829567	-1.63267670	-0.6344746	2.0513683	0.306682112	0.38238716	-0.87027769	-1.0952653
1.97690575	0.730344867	-0.32129051	-0.6344746	-0.4874732	-0.738833971	-0.63153496	0.52505810	1.9447019
-0.89286451	-1.070829567	-0.63813550	-0.6344746	2.0513683	-0.244590005	-0.22596612	1.25944536	-0.3352735
-0.49466672	-0.980770845	1.21892816	1.5760869	-0.4874732	-1.347134237	-1.23988824	-0.28276788	-0.9432669
-1.03017409	-1.070829567	0.87567943	-0.6344746	-0.4874732	1.504273261	1.59909370	-1.45778750	-0.3352735
1.31781976	0.280051258	0.39161070	1.5760869	-0.4874732	1.770404627	2.00466255	1.62663899	1.1847101
-1.03017409	-1.070829567	0.87567943	-0.6344746	-0.4874732	1.846442161	2.00466255	1.70007772	-0.3352735
-0.27497139	0.189992537	-0.99018548	-0.6344746	-0.4874732	1.295170044	1.19352485	-1.16403259	-0.6392702
-1.34598613	1.180638475	-1.28942797	-0.6344746	-0.4874732	-0.073505555	-0.22596612	1.18600664	-1.3992620
0.39784556	-0.710594680	1.18372316	-0.6344746	-0.4874732	-1.480199921	-1.44267266	-0.42964533	0.5767167
-1.03017409	-1.070829567	0.87567943	1.5760869	-0.4874732	-0.985955954	-1.03710381	0.08442575	-0.3352735
-0.57705247	-1.160888289	0.87567943	-0.6344746	-0.4874732	-1.138031021	-1.03710381	0.01098702	-0.4872719
-0.45347385	1.450814640	0.25959196	-0.6344746	-0.4874732	-0.301618155	-0.22596612	1.18600664	-0.7912686
-0.26124044	0.189992537	1.19252441	-0.6344746	-0.4874732	-1.023974721	-1.03710381	0.08442575	-0.6392702
-0.30243331	1.450814640	-0.41810426	-0.6344746	-0.4874732	-0.909918421	-0.83431939	0.23130320	-0.7912686
-0.89286451	-1.070829567	-0.63813550	-0.6344746	2.0513683	0.230644578	0.17960273	0.96569046	-0.3352735
-0.89286451	-1.070829567	-0.63813550	-0.6344746	2.0513683	-1.081002871	-1.03710381	0.01098702	-0.3352735
-1.03017409	-1.070829567	0.87567943	-0.6344746	-0.4874732	0.306682112	0.38238716	-0.87027769	-0.3352735
-0.89286451	-1.070829567	-0.63813550	-0.6344746	2.0513683	-0.073505555	-0.22596612	1.18600664	-0.3352735
1.31781976	0.280051258	0.39161070	1.5760869	-0.4874732	0.268663345	0.17960273	0.96569046	1.1847101
-0.89286451	-1.070829567	-0.63813550	-0.6344746	2.0513683	0.420738412	0.38238716	-0.87027769	-0.3352735
0.08203352	-0.440418515	-0.24207927	-0.6344746	-0.4874732	1.561301411	1.59909370	-1.45778750	0.1207216
1.94944383	0.730344867	0.97249317	-0.6344746	-0.4874732	-0.909918421	-0.83431939	0.23130320	-0.4872719
-0.30243331	1.450814640	-0.41810426	-0.6344746	-0.4874732	1.257151278	1.19352485	-1.16403259	-0.7912686
1.30408880	0.009875093	-0.89337174	1.5760869	-0.4874732	1.067057445	0.78795600	-1.23747132	1.4887068
-0.43974289	1.450814640	-1.94952168	-0.6344746	-0.4874732	-0.700815204	-0.63153496	0.52505810	-0.7912686
0.09576448	-0.440418515	-0.49731551	-0.6344746	-0.4874732	-1.023974721	-1.03710381	0.08442575	0.1207216

SPLITTING THE DATASET



CHOOSING A MODEL:
GAUSSIAN NAIVE BAYES CLASSIFIER

ASSUMPTION OF INDEPENDENCE

FAST TRAINING AND PREDICTION

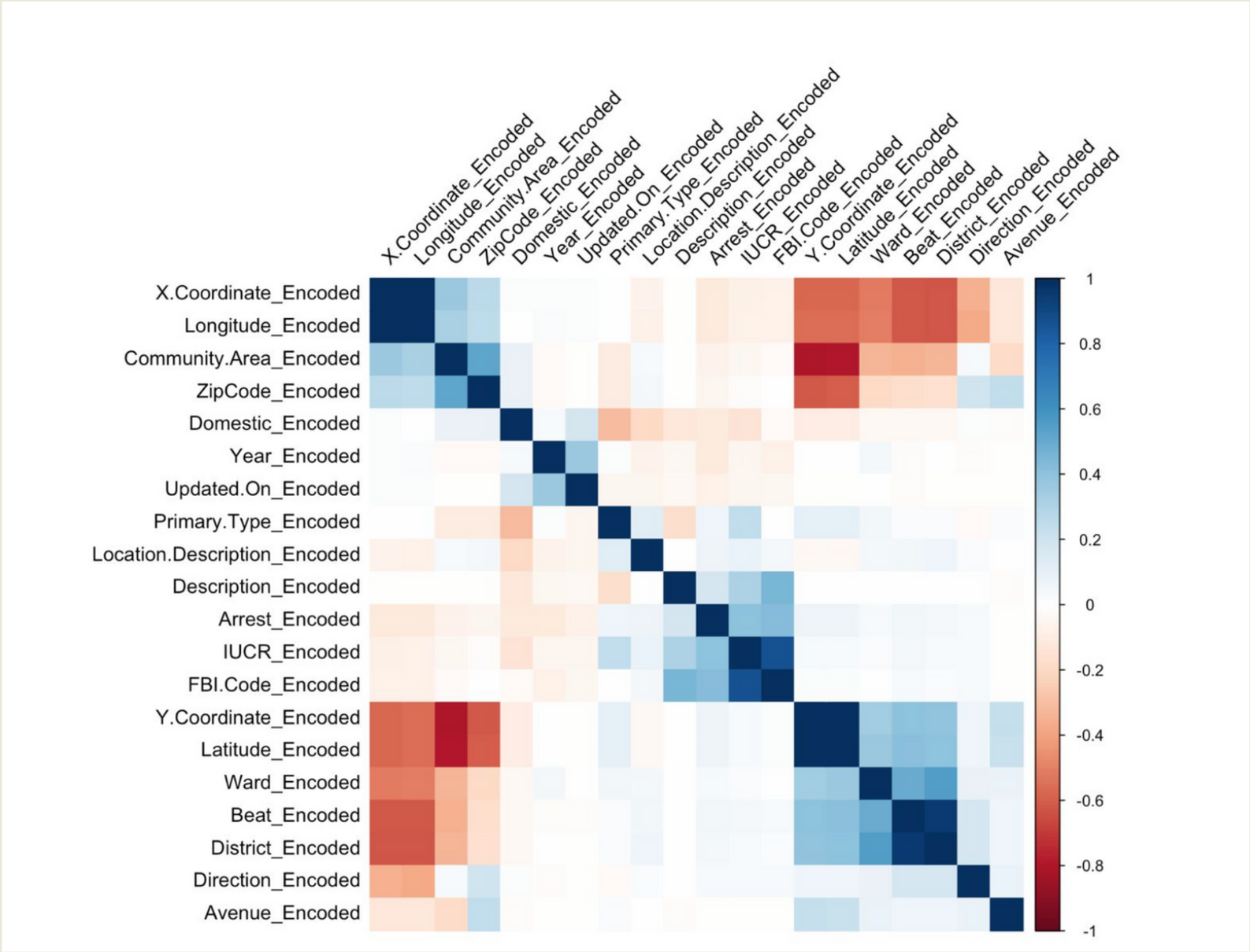
USEFUL FOR IMBALANCED DATASETS

ROBUSTNESS TO IRRELEVANT FEATURES

LESS PRONE TO OVER-FITTING

SUITED FOR TEXT CLASSIFICATION

CORRELATION ANALYSIS



As observed from the graph, these columns show a strong correlation

Longitude_Encoded
X.Coordinate_Encoded
Community.Area_Encoded
District_Encoded
Latitude_Encoded
Beat_Encoded

Predictive Feature Selection:

Features that have a strong correlation with the target variable might be good predictors.

CONFUSION MATRIX

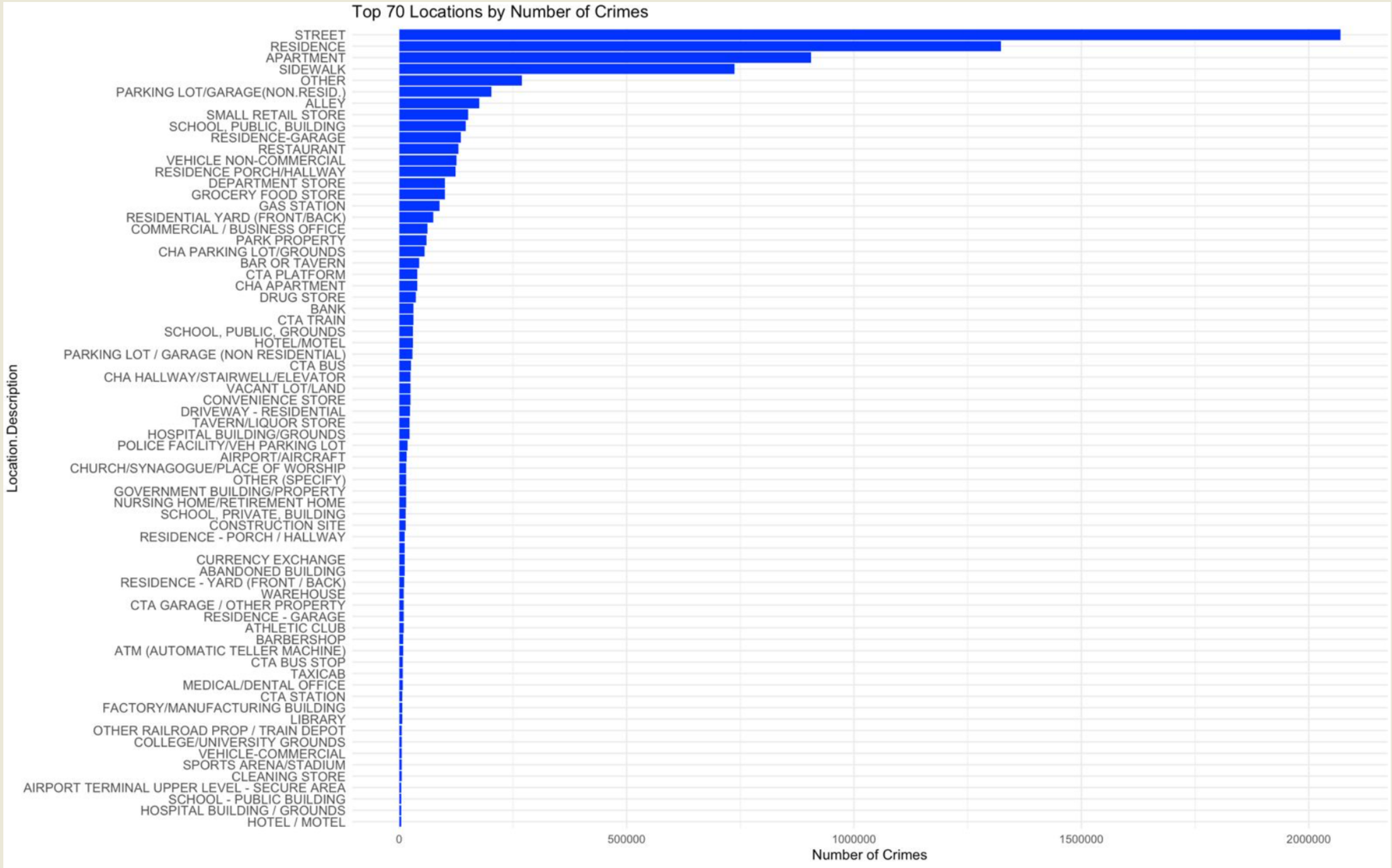
- A Confusion matrix is an N x N matrix used for ***evaluating the performance of a classification model***, where N is the total number of target classes.
- The matrix compares the ***actual target values with those predicted*** by the machine learning model.

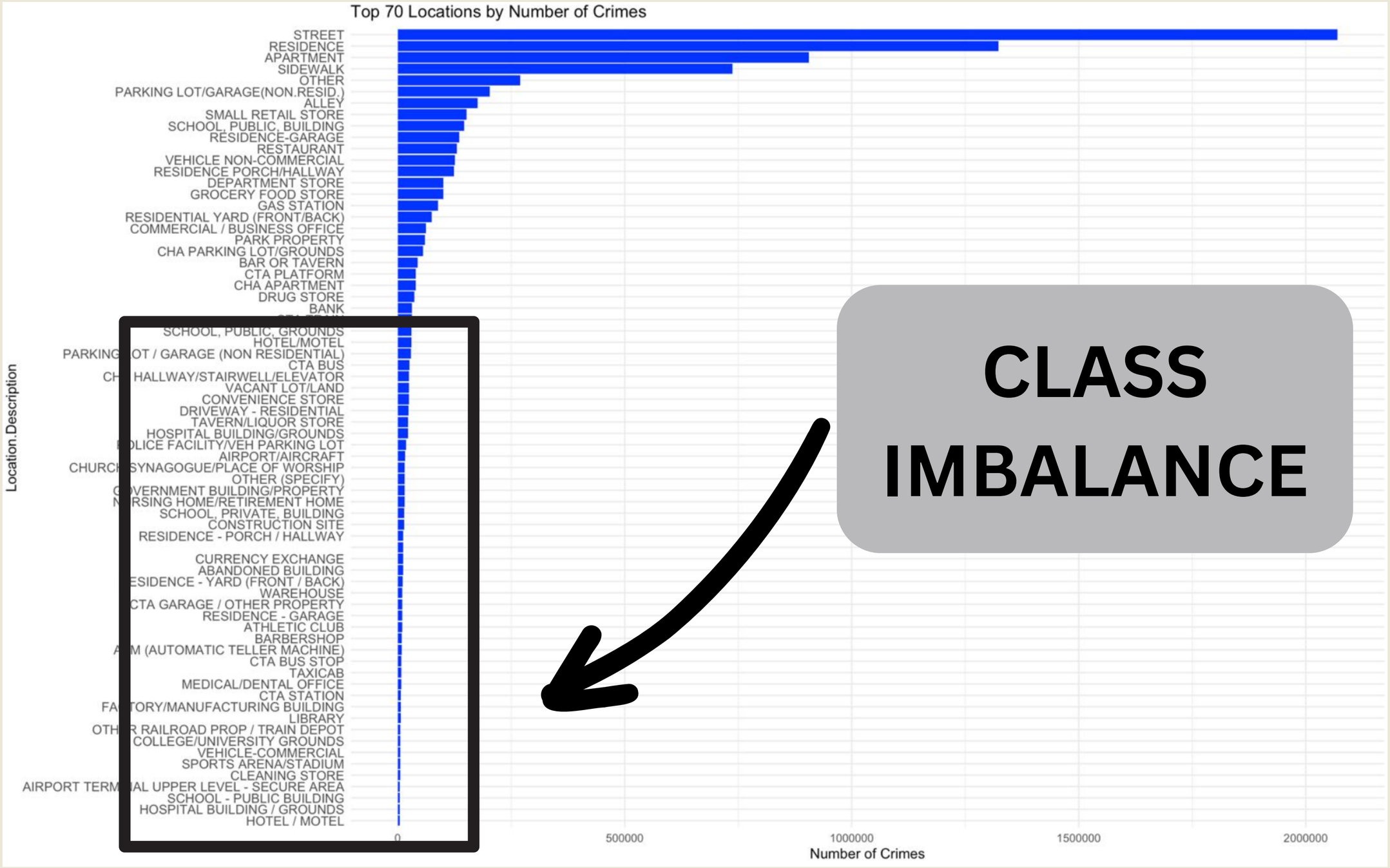
Confusion Matrix and Statistics

	Reference																	
Prediction	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	248	13	0	0	0	0	0	0	0	0	0	0	0	20	11	5	0	34
2	59	370	10	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
3	0	0	297	14	49	24	0	0	0	0	49	0	0	0	0	0	0	0
4	0	0	0	294	0	8	2	0	0	30	17	21	0	0	0	0	0	0
5	0	0	0	0	377	68	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	17	47	7	318	4	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	18	0	8	363	0	0	0	0	20	0	0	0	158	0	0
8	0	0	0	0	0	0	0	314	83	44	0	0	0	0	0	0	0	0
9	0	49	0	1	0	0	0	87	370	0	92	0	0	0	0	0	0	0
10	0	0	0	128	0	0	0	9	5	385	3	63	0	0	0	0	0	0
11	0	8	36	4	0	0	0	0	7	0	341	0	0	0	0	0	0	0
12	0	0	0	3	0	0	2	0	0	2	0	359	0	0	0	4	0	0
13	21	0	0	0	0	0	0	2	0	0	0	0	498	0	86	0	0	0
14	25	0	0	0	0	0	0	0	0	0	0	0	0	436	16	0	25	25
15	151	0	0	0	0	0	0	0	0	0	0	0	80	44	520	0	16	0
16	0	0	0	0	0	0	13	0	0	0	0	13	0	0	0	231	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	32	32	0	366	0
18	20	0	0	0	0	0	0	0	0	0	0	0	0	9	0	4	0	597

RESULTS-
ATTEMPT 1

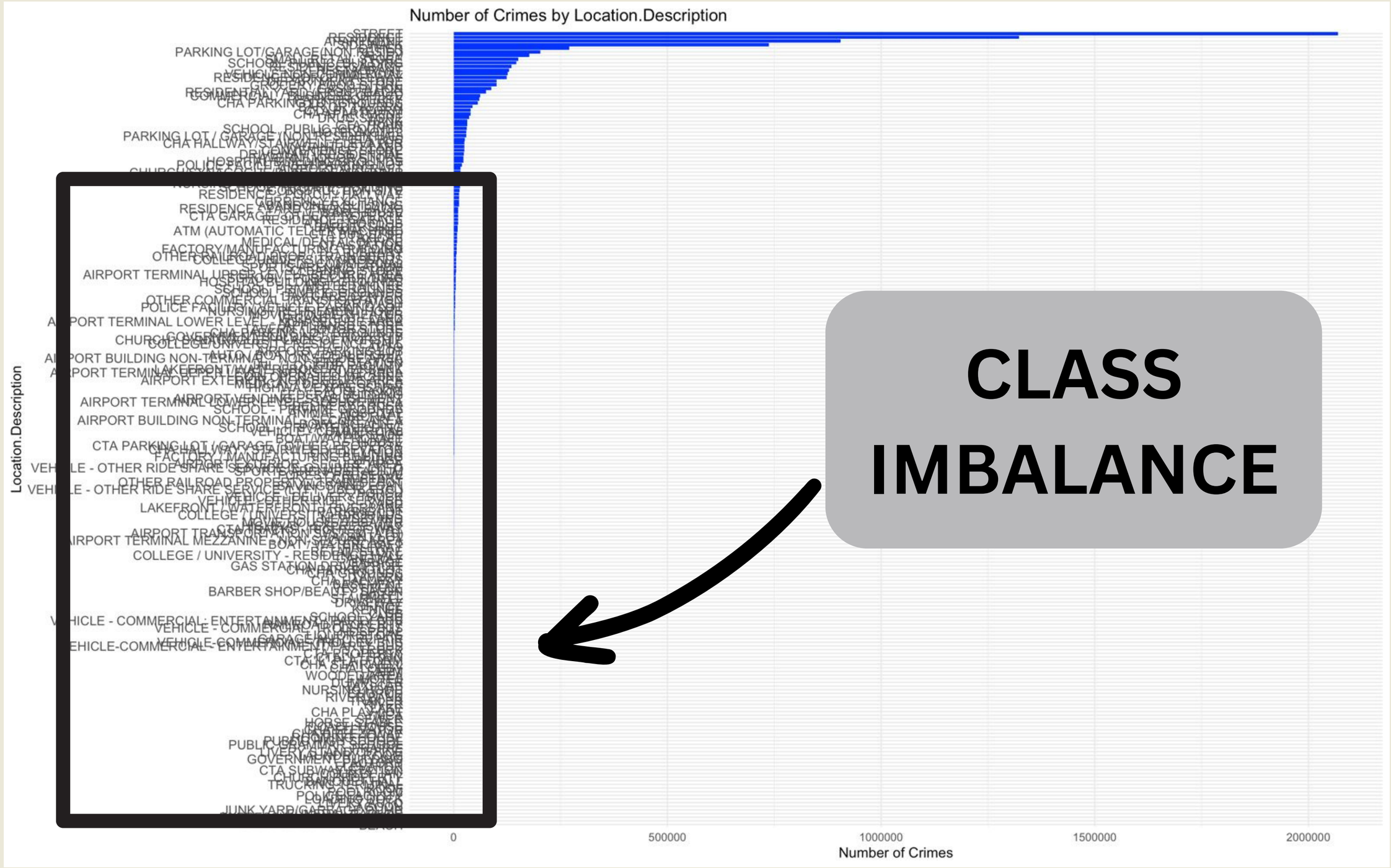
```
Overall Statistics

               Accuracy : 0.4479
                 95% CI : (0.4377, 0.4582)
    No Information Rate : 0.3779
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2162

 Mcnemar's Test P-Value : < 2.2e-16
```


Top 70 Locations by Number of Crimes
Location.Description
STREET
RESIDENCE
APARTMENT
SIDEWALK
OTHER
PARKING LOT/GARAGE(NON.RESID.)
ALLEY
SMALL RETAIL STORE
SCHOOL, PUBLIC, BUILDING
RESIDENCE-GARAGE
RESTAURANT
VEHICLE NON-COMMERCIAL
RESIDENCE PORCH/HALLWAY
DEPARTMENT STORE
GROCERY FOOD STORE
GAS STATION
RESIDENTIAL YARD (FRONT/BACK)
COMMERCIAL / BUSINESS OFFICE
PARK PROPERTY
CHA PARKING LOT/GROUNDS
BAR OR TAVERN
CTA PLATFORM
CHA APARTMENT
DRUG STORE
BANK
CTA TRAIN
SCHOOL, PUBLIC, GROUNDS
HOTEL/MOTEL
PARKING LOT / GARAGE (NON RESIDENTIAL)
CTA BUS
CHA HALLWAY/STAIRWELL/ELEVATOR
VACANT LOT/LAND
CONVENIENCE STORE
DRIVEWAY - RESIDENTIAL
TAVERN/LIQUOR STORE
HOSPITAL BUILDING/GROUNDS
POLICE FACILITY/VEH PARKING LOT
AIRPORT/AIRCRAFT
CHURCH/SYNAGOGUE/PLACE OF WORSHIP
OTHER (SPECIFY)
GOVERNMENT BUILDING/PROPERTY
NURSING HOME/RETIREMENT HOME
SCHOOL, PRIVATE, BUILDING
CONSTRUCTION SITE
RESIDENCE - PORCH / HALLWAY
CURRENCY EXCHANGE
ABANDONED BUILDING
RESIDENCE - YARD (FRONT / BACK)
WAREHOUSE
CTA GARAGE / OTHER PROPERTY
RESIDENCE - GARAGE
ATHLETIC CLUB
BARBERSHOP
ATM (AUTOMATIC TELLER MACHINE)
CTA BUS STOP
TAXICAB
MEDICAL/DENTAL OFFICE
CTA STATION
FACTORY/MANUFACTURING BUILDING
LIBRARY
OTHER RAILROAD PROP / TRAIN DEPOT
COLLEGE/UNIVERSITY GROUNDS
VEHICLE-COMMERCIAL
SPORTS ARENA/STADIUM
CLEANING STORE
AIRPORT TERMINAL UPPER LEVEL - SECURE AREA
SCHOOL - PUBLIC BUILDING
HOSPITAL BUILDING / GROUNDS
HOTEL / MOTEL
0
500000
1000000
1500000
2000000
Number of Crimes

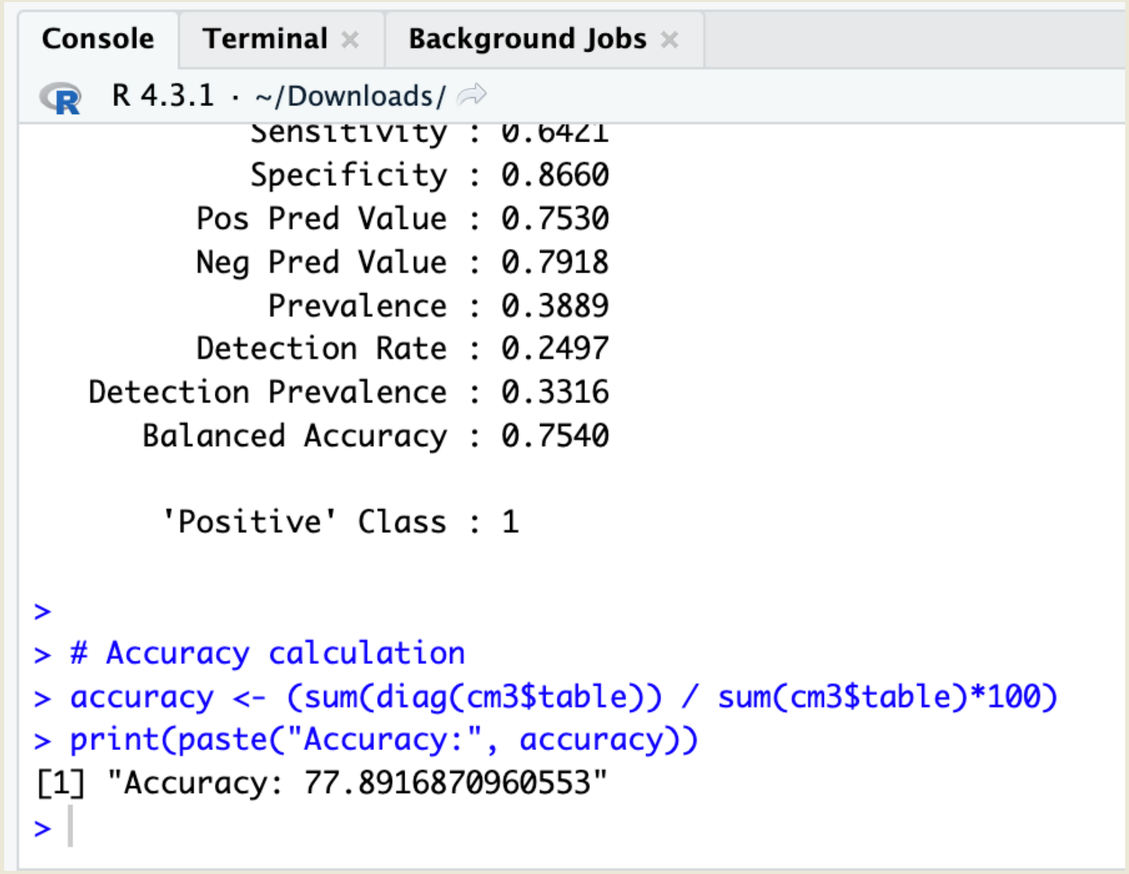

Top 70 Locations by Number of Crimes
STREET
RESIDENCE
APARTMENT
SIDEWALK
OTHER
PARKING LOT/GARAGE(NON.RESID.)
ALLEY
SMALL RETAIL STORE
SCHOOL, PUBLIC, BUILDING
RESIDENCE-GARAGE
RESTAURANT
VEHICLE NON-COMMERCIAL
RESIDENCE PORCH/HALLWAY
DEPARTMENT STORE
GROCERY FOOD STORE
GAS STATION
RESIDENTIAL YARD (FRONT/BACK)
COMMERCIAL / BUSINESS OFFICE
PARK PROPERTY
CHA PARKING LOT/GROUNDS
BAR OR TAVERN
CTA PLATFORM
CHA APARTMENT
DRUG STORE
BANK
SCHOOL, PUBLIC, GROUNDS
HOTEL/MOTEL
PARKING LOT / GARAGE (NON RESIDENTIAL)
CTA BUS
HALLWAY/STAIRWELL/ELEVATOR
VACANT LOT/LAND
CONVENIENCE STORE
DRIVEWAY - RESIDENTIAL
TAVERN/LIQUOR STORE
HOSPITAL BUILDING/GROUNDS
POLICE FACILITY/VEH PARKING LOT
AIRPORT/AIRCRAFT
SYNAGOGUE/PLACE OF WORSHIP
OTHER (SPECIFY)
GOVERNMENT BUILDING/PROPERTY
NURSING HOME/RETIREMENT HOME
SCHOOL, PRIVATE, BUILDING
CONSTRUCTION SITE
RESIDENCE - PORCH / HALLWAY
CURRENCY EXCHANGE
ABANDONED BUILDING
RESIDENCE - YARD (FRONT / BACK)
WAREHOUSE
CTA GARAGE / OTHER PROPERTY
RESIDENCE - GARAGE
ATHLETIC CLUB
BARBERSHOP
ATM (AUTOMATIC TELLER MACHINE)
CTA BUS STOP
TAXICAB
MEDICAL/DENTAL OFFICE
CTA STATION
FACTORY/MANUFACTURING BUILDING
LIBRARY
OTHER RAILROAD PROP / TRAIN DEPOT
COLLEGE/UNIVERSITY GROUNDS
VEHICLE-COMMERCIAL
SPORTS ARENA/STADIUM
CLEANING STORE
AIRPORT TERMINAL UPPER LEVEL - SECURE AREA
SCHOOL - PUBLIC BUILDING
HOSPITAL BUILDING / GROUNDS
HOTEL / MOTEL
Location.Description
0
500000
1000000
1500000
2000000
Number of Crimes
CLASS IMBALANCE

CLASS IMBALANCE

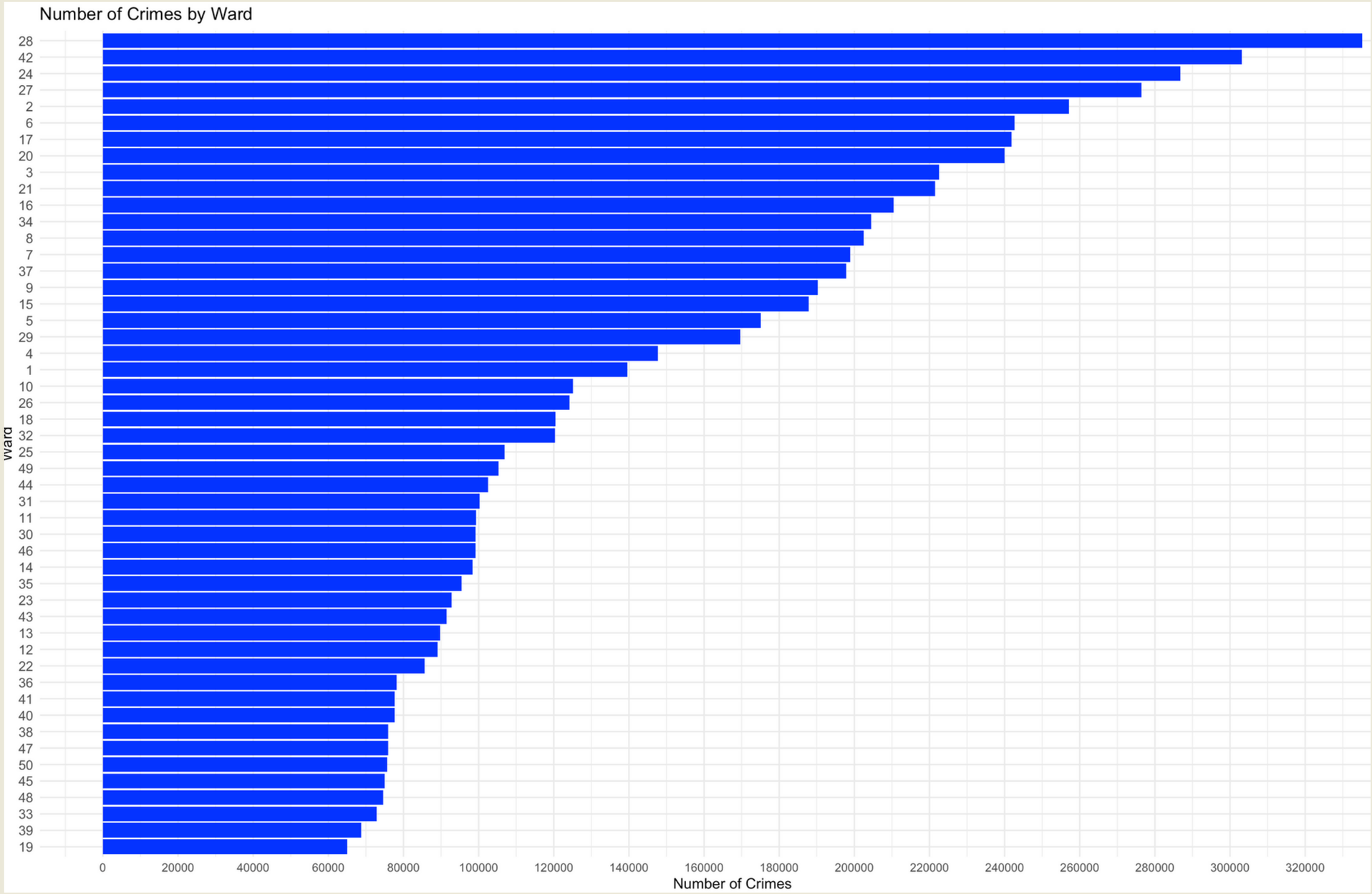
- A distribution of categories (or classes) where ***some classes have a much higher frequency than others***
- The location labeled "**STREET**" has a disproportionately high number of crimes, as indicated by the length of its bar.
- This is followed by "**RESIDENCE**" and "**APARTMENT**", which also have long bars but significantly fewer crimes than "STREET".
- The majority of other locations have much shorter bars, indicating fewer crimes, with many locations at the bottom of the chart having relatively very few crimes reported.


Number of Crimes by Location.Description
Location.Description
Number of Crimes
0
500000
1000000
1500000
2000000
CLASS IMBALANCE

UPDATED RESULTS



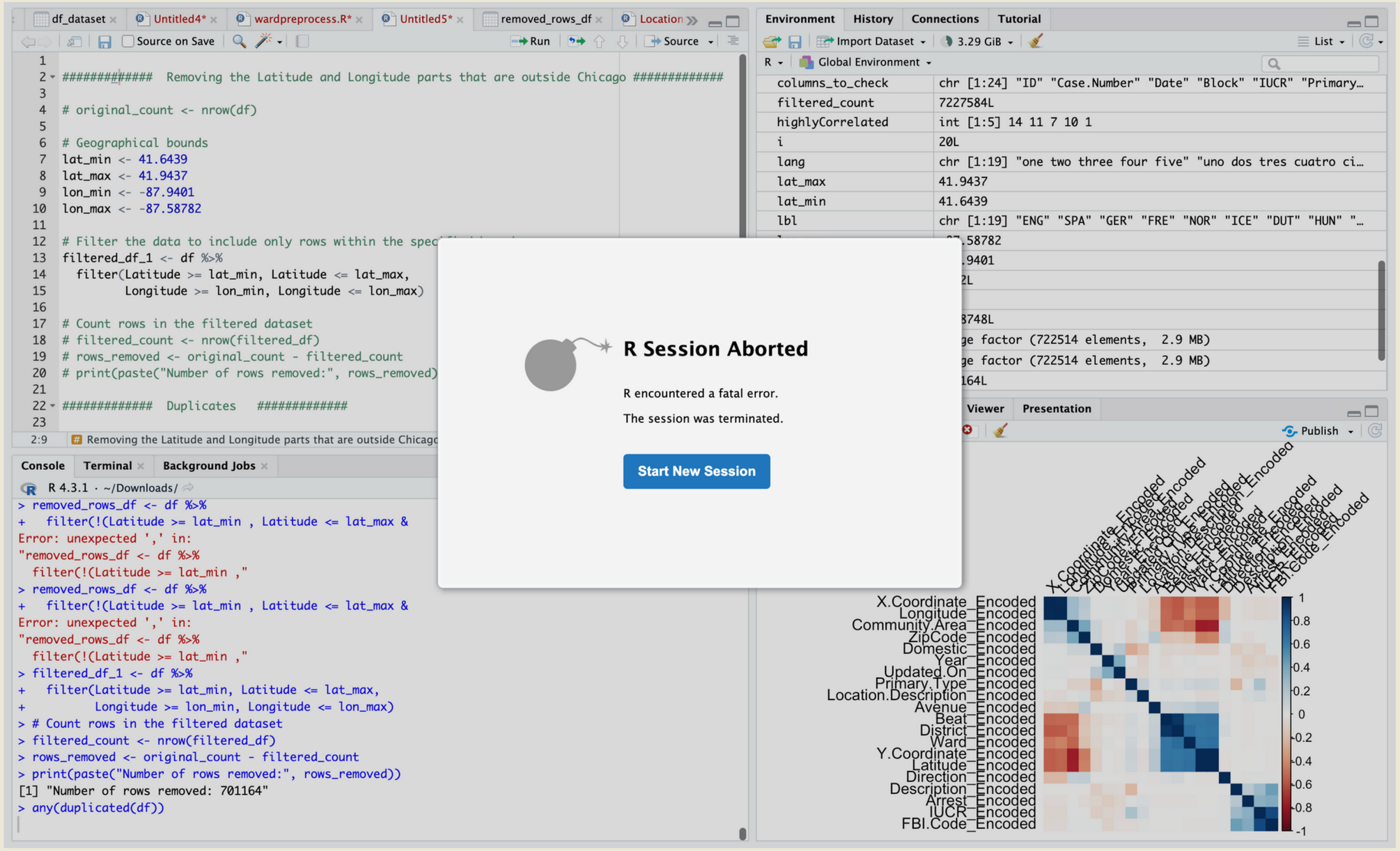
Predicting the null values in WARD column



RESULTS

```
Overall Statistics

               Accuracy : 0.8146
                 95% CI : (0.8061, 0.8229)
    No Information Rate : 0.1044
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7987

 Mcnemar's Test P-Value : NA
```

Thank You


df_dataset
Untitled4*
wardpreprocess.R*
Untitled5*
removed_rows_df
Location
Source on Save
Run
Source
Environment
History
Connections
Tutorial
Import Dataset
3.29 GiB
List
R
Global Environment
columns_to_check chr [1:24] "ID" "Case.Number" "Date" "Block" "IUCR" "Primary…
filtered_count 7227584L
highlyCorrelated int [1:5] 14 11 7 10 1
i 20L
lang chr [1:19] "one two three four five" "uno dos tres cuatro ci…
lat_max 41.9437
lat_min 41.6439
lbl chr [1:19] "ENG" "SPA" "GER" "FRE" "NOR" "ICE" "DUT" "HUN" "…
############ Removing the Latitude and Longitude parts that are outside Chicago ############
# original_count <- nrow(df)
# Geographical bounds
lat_min <- 41.6439
lat_max <- 41.9437
lon_min <- -87.9401
lon_max <- -87.58782
# Filter the data to include only rows within the spec
filtered_df_1 <- df %>%
filter(Latitude >= lat_min, Latitude <= lat_max,
Longitude >= lon_min, Longitude <= lon_max)
Count rows in the filtered dataset
filtered_count <- nrow(filtered_df)
rows_removed <- original_count - filtered_count
print(paste("Number of rows removed:", rows_removed))
############ Duplicates ############
2:9 Removing the Latitude and Longitude parts that are outside Chicago
R Session Aborted
R encountered a fatal error.
The session was terminated.
Start New Session
Console
Terminal
Background Jobs
R 4.3.1 · ~/Downloads/
> removed_rows_df <- df %>%
+ filter(!(Latitude >= lat_min , Latitude <= lat_max &
Error: unexpected ',' in:
"removed_rows_df <- df %>%
filter(!(Latitude >= lat_min ,"
> removed_rows_df <- df %>%
+ filter(!(Latitude >= lat_min , Latitude <= lat_max &
Error: unexpected ',' in:
"removed_rows_df <- df %>%
filter(!(Latitude >= lat_min ,"
> filtered_df_1 <- df %>%
+ filter(Latitude >= lat_min, Latitude <= lat_max,
+ Longitude >= lon_min, Longitude <= lon_max)
> # Count rows in the filtered dataset
> filtered_count <- nrow(filtered_df)
> rows_removed <- original_count - filtered_count
> print(paste("Number of rows removed:", rows_removed))
[1] "Number of rows removed: 701164"
> any(duplicated(df))
Viewer
Presentation
Publish
X.Coordinate_Encoded
Longitude_Encoded
Community.Area_Encoded
ZipCode_Encoded
Domestic_Encoded
Year_Encoded
Updated.On_Encoded
Primary.Type_Encoded
Location.Description_Encoded
Avenue_Encoded
Beat_Encoded
District_Encoded
Ward_Encoded
Y.Coordinate_Encoded
Latitude_Encoded
Direction_Encoded
Description_Encoded
Arrest_Encoded
IUCR_Encoded
FBI.Code_Encoded
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

QUESTION DURING PRESENTATION

Why did you use Label Encoding instead of One Hot Encoding

ANSWER

In One Hot Encoding, each and every unique variable is converted into a new column and has the values of true or false making the dataset with more than 1500 columns. It takes so much computaional power to train the model and also R was crashing with one hot method. That is why we used Label Encoding.