Assignment 03 – Section 013
Group 13

Proposal & Initial Study of Chicago's Traffic Crashes Data Set

Onur Onel 041074824
Pinqi Wang 041070733
Abhishek Pandav

14/07/2023

Algonquin College – School of Advanced Technology
Computer Programming
23S-CST8390

Content

Introduction

The dataset we are working with is sourced from the City of Chicago's Data Portal, specifically the "Traffic Crashes - Crashes" dataset. This dataset provides detailed information about traffic crashes occurring within the city of Chicago. The data is collected by reporting officers who respond to crash scenes and is used by the Chicago Police Department and the Department of Transportation. Each row in the dataset represents a unique traffic crash and includes a variety of information about the crash, such as the date and time of the crash, the posted speed limit, the type of traffic control device present, the weather and lighting conditions, the first type of crash (e.g., rear end, angle, pedestrian), and the type of trafficway (e.g., one-way, two-way, divided). One of the key features in the dataset is most_severe_injury, which categorizes the most severe injury that occurred in the crash. Other important features include crash_date, weather_condition, lighting_condition, first_crash_type, and trafficway_type.

Business Understanding

Objective: The main objective is to understand the factors that contribute to the severity of injuries in traffic crashes in the city of Chicago. This could involve identifying which conditions or characteristics of crashes are associated with different types of injuries. The goal is to inform interventions that reduce the severity of injuries in crashes. To achieve this, we will use classification techniques (k-Nearest Neighbors and Decision Trees) to predict the severity of injuries based on crash characteristics, and clustering (k-Means) to identify patterns in the data. We will also use linear regression to predict the total number of injuries (injuries_total) or most severe injury based on other features in the dataset. This could provide insights into which factors contribute to a higher number of injuries in a crash.

Data Mining Problem: This is a multiclass classification problem, where the target variable is most_severe_injury and the features could be various other columns in the dataset. The goal is to build models (using k-NN and Decision Trees) that can predict the most_severe_injury based on these features. Additionally, we will use k-Means clustering to explore the structure of the data and identify patterns that might not be apparent from the classification analysis.

Preliminary Plan: The preliminary plan would involve cleaning and preprocessing the data, followed by exploratory data analysis. Then, we would select appropriate features and build classification models using k-NN and Decision Trees, which we would evaluate and tune as necessary. We would also apply k-Means clustering to the data to identify patterns. Finally, we would interpret the results to understand which features are most important in predicting the severity of injuries, and to understand the structure of the data revealed by the clustering.

Data Understanding

The data utilized for the present analysis is derived from traffic collision records sourced from the City of Chicago's Data Portal. The comprehensive original dataset comprised 740,000 entries and 49 distinct variables, which posed considerable computational challenges when endeavoring to handle it directly in the Weka environment due to constraints of inaccessible memory. In order to circumnavigate this hurdle, a practical approach was embraced by importing the extensive dataset into a MySQL database. Thereafter, a concentrated subset of data was curated by selecting the 20,000 most recent entries, sorted in descending order by the CRASH_DATE attribute.

Pursuing the primary aim of reducing injury severity resulting from traffic collisions, an intentional and discerning selection of attributes was carried out. These attributes were meticulously chosen to ascertain their relevance to the analysis and their capacity to yield insights into the factors leading to injury severity. The dataset underwent supplementary preprocessing stages to refine its suitability for subsequent analysis.

Column Name	Description	Type
ID	Numeric ID assigned to each instance	Numeric
CRASH_DATE	Date and time of the crash	Date & Time
POSTED_SPEED_LIMIT	Posted speed limit at the crash location	Numeric

TRAFFIC_CONTROL_DEVICE	Traffic control device at the crash location	Nominal
DEVICE_CONDITION	Condition of the traffic control device	Nominal
WEATHER_CONDITION	Weather condition at the time of the crash	Nominal
LIGHTING_CONDITION	Lighting condition at the time of the crash	Nominal
FIRST_CRASH_TYPE	Type of the first crash in the collision	Nominal
ROADWAY_SURFACE_COND	Condition of the roadway surface	Nominal
ROAD_DEFECT	Defects in the road at the crash location	Nominal
CRASH_TYPE	Type of the crash	Nominal
INTERSECTION_RELATED_I	Whether the crash is related to an intersection	Nominal

NOT_RIGHT_OF_WAY_I	Whether the crash began outside of the public right-of-way	Nominal
HIT_AND_RUN_I	Whether the crash involved a hit and run	Nominal
DAMAGE	Estimated damage from the crash	Nominal
PRIM_CONTRIBUTORY_CAUSE	Primary contributing cause of the crash	Nominal
SEC_CONTRIBUTORY_CAUSE	Secondary contributing cause of the crash	Nominal
DOORING_I	Whether the crash involved dooring (a motor vehicle occupant opening a door into the path of a bicyclist)	Nominal
NUM_UNITS	Number of units involved in the crash	Nominal
MOST_SEVERE_INJURY	Most severe injury resulting from the crash	Nominal

INJURIES_TOTAL	Total number of injuries sustained in the crash	Nominal
INJURIES_FATAL	Total number of fatal injuries in the crash	Nominal
INJURIES_INCAPACITATING	Total number of incapacitating injuries in the crash	Nominal
INJURIES_NON_INCAPACITATING	Total number of non-incapacitating injuries in the crash	Nominal
INJURIES_REPORTED_NOT_EVIDENT	Total number of injuries reported but not evident at the scene of the crash	Nominal
INJURIES_NO_INDICATION	Total number of injuries where there is no indication of injury	Nominal

INJURIES_UNKNOWN	Total number of injuries where the injury status is unknown	Nominal

To streamline the analysis and modeling process, particular filters were instituted on the dataset. Specifically, the "filters.unsupervised.attribute.StringToNominal" filter was utilized on variables such as INJURIES_TOTAL, INJURIES_FATAL, INJURIES_INCAPACITATING, INJURIES_NON_INCAPACITATING, INJURIES_REPORTED_NOT_EVIDENT, INJURIES_NO_INDICATION, and INJURIES_UNKNOWN. This transformation streamlined the deployment of suitable analytical methodologies, enabling these attributes to be identified as nominal variables. Furthermore, it enhanced our comprehension of the inherent values encapsulated within the data.

Moreover, an identification column was added to each entry, enabling effortless pinpointing and referencing of individual instances within the dataset. This procedural enhancement optimizes data management and fosters further analysis by supplying a unique identifier for each record.

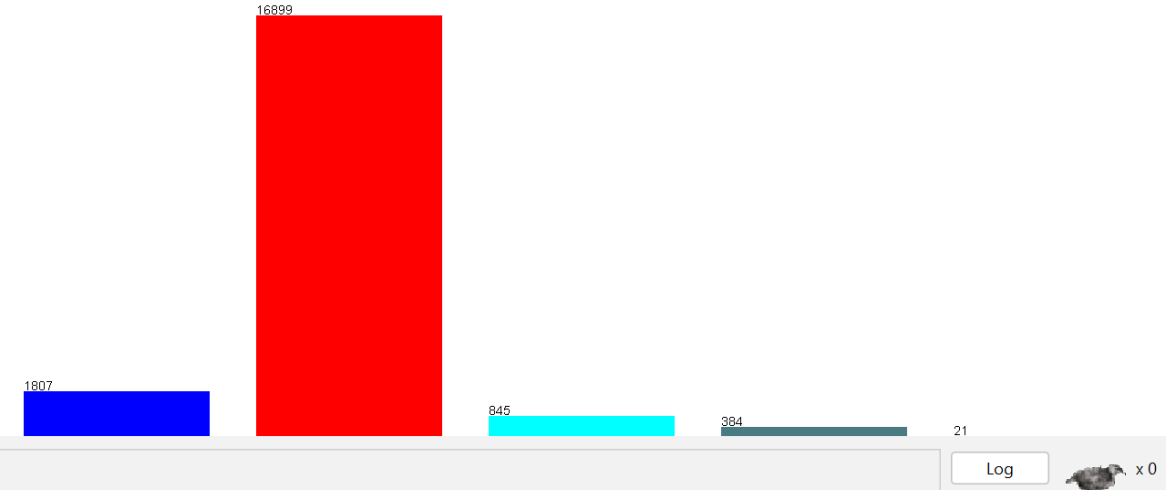
The integration of these preprocessing initiatives renders the dataset adequately prepared for in-depth analysis, classification, and predictive modeling. The refined dataset, featuring the selected attributes and the newly introduced ID column, provides a robust foundation for revealing insights and devising strategies aimed at mitigating the severity of injuries sustained in traffic collisions.

Selected attribute

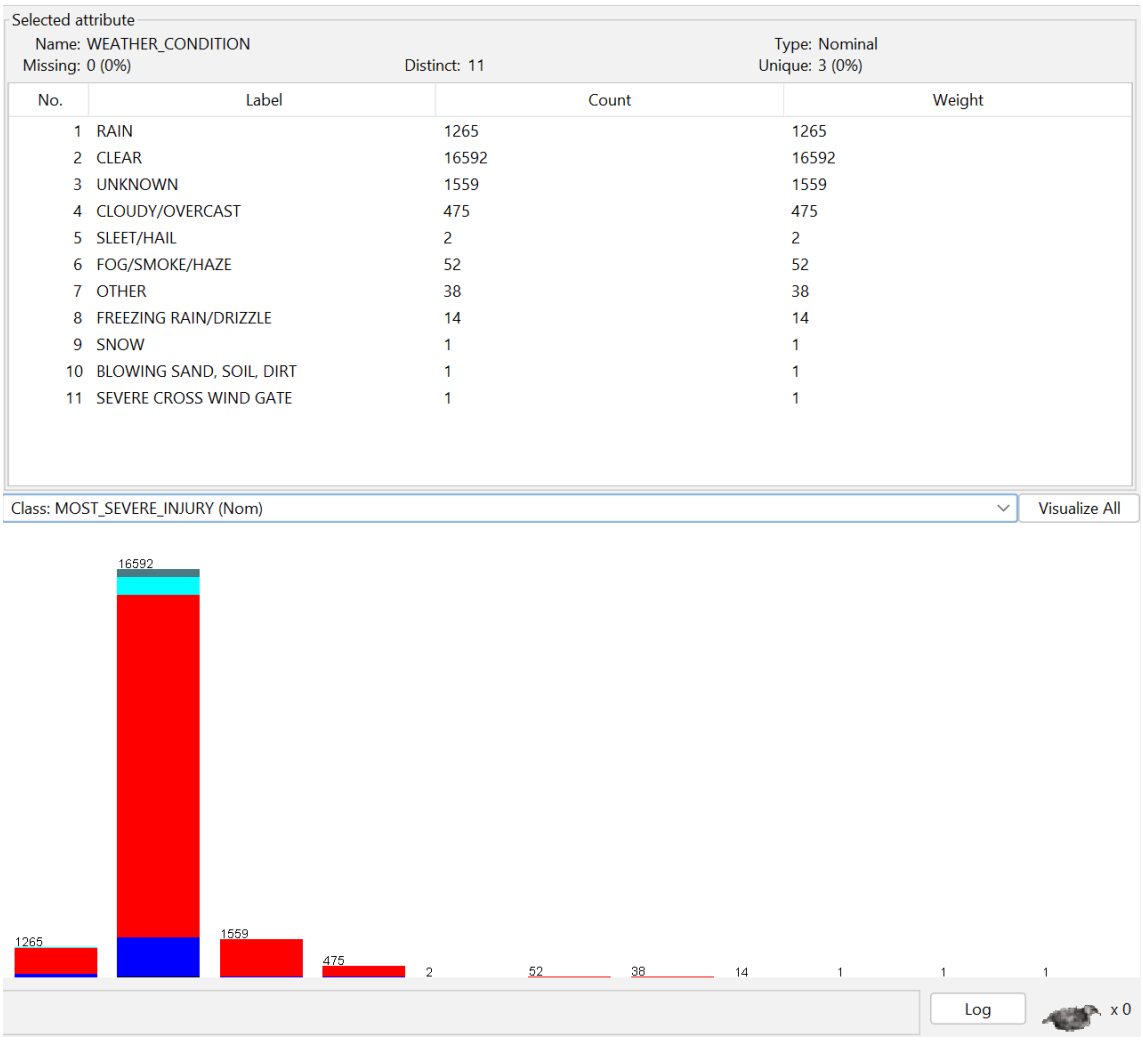
Name: MOST_SEVERE_INJURY Type: Nominal
Missing: 44 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	NONINCAPACITATING INJURY	1807	1807
2	NO INDICATION OF INJURY	16899	16899
3	REPORTED, NOT EVIDENT	845	845
4	INCAPACITATING INJURY	384	384
5	FATAL	21	21

Class: MOST_SEVERE_INJURY (Nom) Visualize All



In the dataset comprising 20,000 records from various vehicular incidents in Chicago, an important attribute is 'most severe injury'. This attribute represents the most grievous injury sustained during each individual accident. It provides a snapshot of the worst-case physical impact of these events, informing improvements in safety measures and emergency response strategies across the city.



When exploring the dataset of 20,000 records from Chicago's vehicle collision incidents, an integral factor to take into account is the weather condition at the moment of each accident, as documented by the on-scene officer. Coupling this information with the class attribute that outlines the most severe injury incurred in each incident provides substantial insights into the correlation between weather conditions and the severity of injuries from vehicular accidents in Chicago.

In the Chicago crashes dataset, the 'crash type' can be classified under one of two categories: 'Injury and/or Tow Due to Crash' or 'No Injury / Drive Away'. These categories offer a comprehensive overview of the incident outcomes, specifying if they led to physical injuries and/or necessitated vehicle towing, or on the other hand, if they were sufficiently minor for the participants to leave the scene without any discernible injuries.

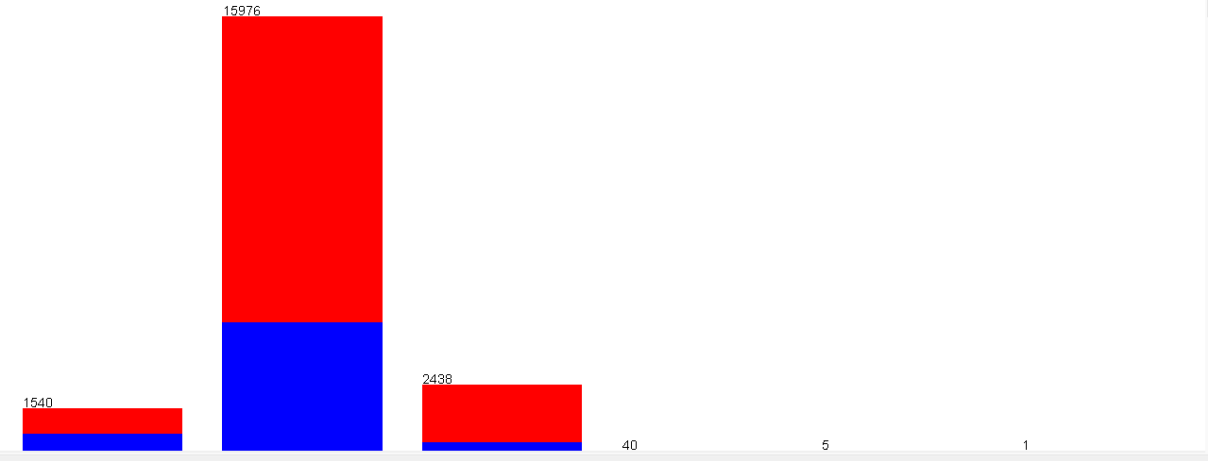
Selected attribute

Name: ROADWAY_SURFACE_COND
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)

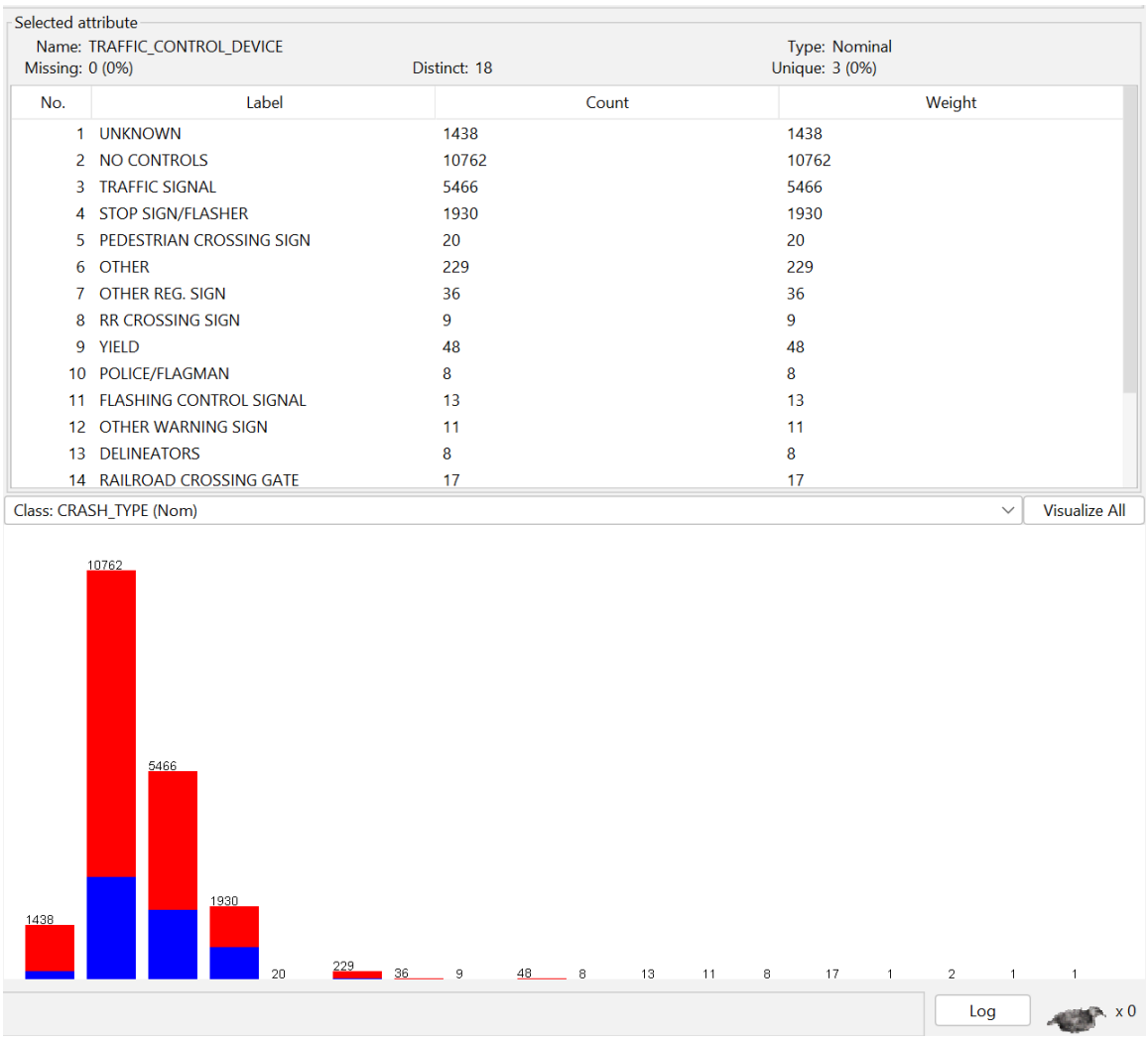
No.	Label	Count	Weight
1	WET	1540	1540
2	DRY	15976	15976
3	UNKNOWN	2438	2438
4	OTHER	40	40
5	SAND, MUD, DIRT	5	5
6	ICE	1	1

Class: CRASH_TYPE (Nom)

Visualize All



One of the key attributes to consider is the road surface condition at the time of the crash, as determined by the reporting officer. When coupled with the class attribute representing the type of crash, it provides insights into how different road conditions may influence the type and severity of vehicular accidents. Understanding these relationships can assist in efforts to prevent future accidents and improve road safety.



In an analysis of the most recent 20,000 entries from the Chicago crashes dataset, it's vital to assess the type and presence of traffic control devices at each crash location, as noted by the responding officer. Additionally, correlating this data with the 'crash type' attribute, which classifies each incident based on whether it resulted in injuries or not, can provide substantial insights. By comprehending the interplay of these factors, we can guide preventative strategies and enhance safety measures, ultimately aiming to reduce both non-injury and injury-resulting vehicular accidents in Chicago.

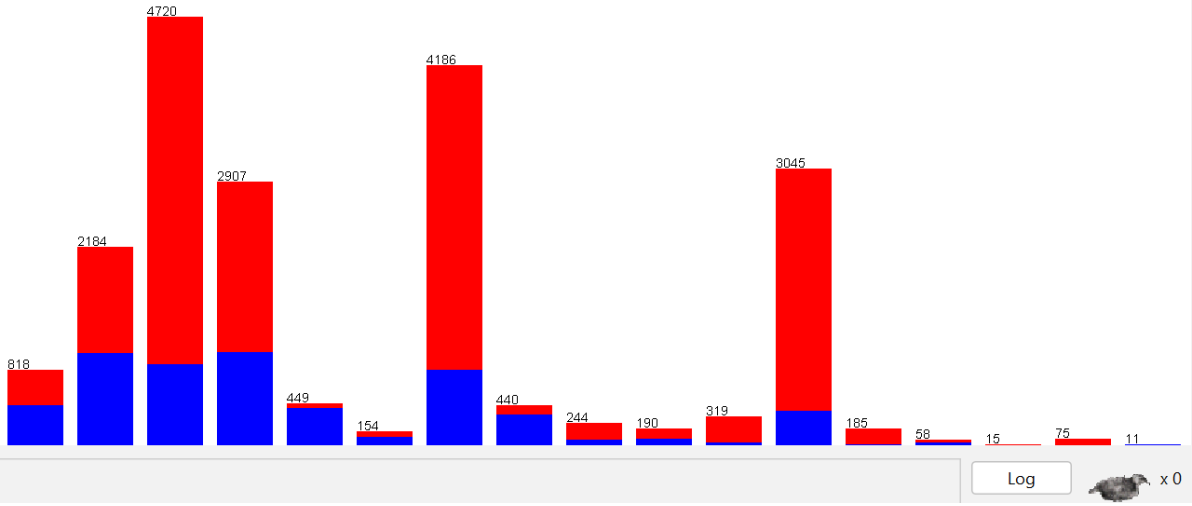
Selected attribute

Name: FIRST_CRASH_TYPE Type: Nominal

Missing: 0 (0%) Distinct: 17 Unique: 0 (0%)

No.	Label	Count	Weight
1	FIXED OBJECT	818	818
2	ANGLE	2184	2184
3	PARKED MOTOR VEHICLE	4720	4720
4	TURNING	2907	2907
5	PEDESTRIAN	449	449
6	HEAD ON	154	154
7	REAR END	4186	4186
8	PEDALCYCLIST	440	440
9	SIDESWIPE OPPOSITE DIRECTION	244	244
10	OTHER OBJECT	190	190
11	REAR TO FRONT	319	319
12	SIDESWIPE SAME DIRECTION	3045	3045
13	REAR TO SIDE	185	185
14	OTHER NONCOLLISION	58	58

Class: CRASH_TYPE (Nom) Visualize All



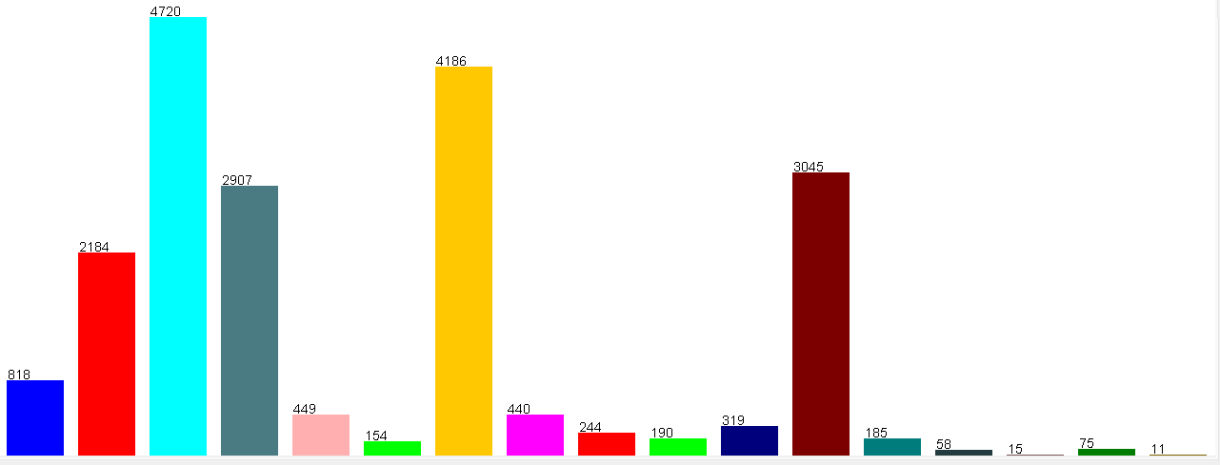
A key factor to evaluate is the 'type of first collision in crash'. This attribute details the initial nature of each accident. When cross-referenced with the 'crash type' class, which indicates whether an incident resulted in injuries, significant insights can be gleaned. By exploring the interaction between these variables, we can identify potential correlations between certain types of collisions and the likelihood of injuries, thus informing more effective safety strategies.

Selected attribute

Name: FIRST_CRASH_TYPE Type: Nominal
Missing: 0 (0%) Distinct: 17 Unique: 0 (0%)

No.	Label	Count	Weight
1	FIXED OBJECT	818	818
2	ANGLE	2184	2184
3	PARKED MOTOR VEHICLE	4720	4720
4	TURNING	2907	2907
5	PEDESTRIAN	449	449
6	HEAD ON	154	154
7	REAR END	4186	4186
8	PEDALCYCLIST	440	440
9	SIDESWIPE OPPOSITE DIRECTION	244	244
10	OTHER OBJECT	190	190
11	REAR TO FRONT	319	319
12	SIDESWIPE SAME DIRECTION	3045	3045
13	REAR TO SIDE	185	185
14	OTHER NONCOLLISION	58	58
15	ANIMAL	15	15
16	REAR TO REAR	75	75
17	OVERTURNED	11	11

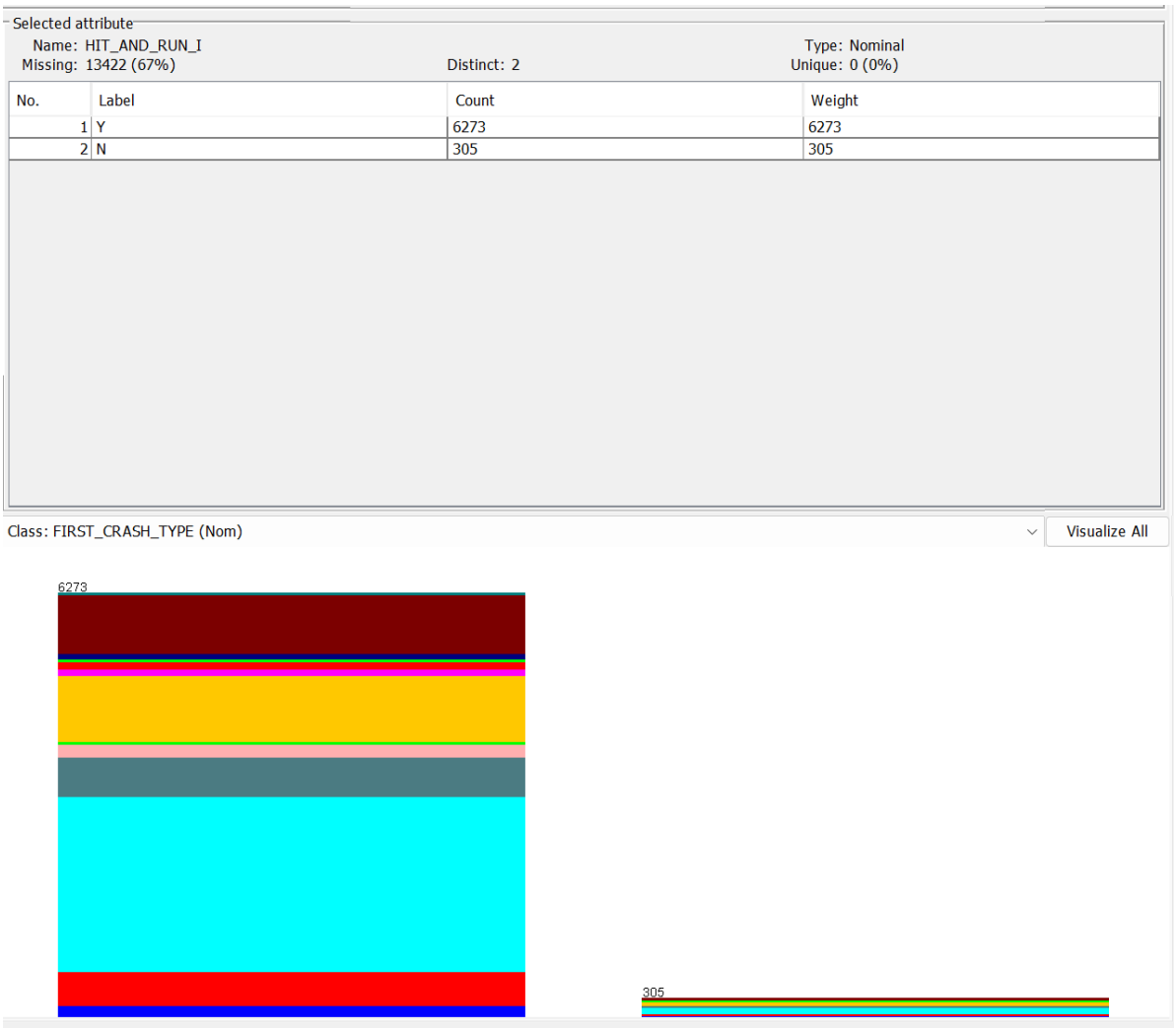
Class: FIRST_CRASH_TYPE (Nom) Visualize All



When examining the 'type of first collision in crash' attribute within the Chicago vehicular collision dataset of 20,000 records, the top three collision types are as follows:

1. Record 4720: 'Parked Motor Vehicle'
2. Record 4186: 'Rear End'
3. Record 3045: 'Sideswipe Same Direction'

These collisions involve different scenarios, with record 4720 indicating a collision with a parked motorcycle, record 4186 involving a rear-end collision, and record 3045 representing a sideswipe collision occurring in the same direction. Analyzing the prevalence of these collision types can offer valuable insights into the common scenarios and dynamics of crashes in Chicago.



In the analysis of the Chicago vehicular collision dataset, it is evident that certain collision types, specifically 'Parked Motor Vehicle', 'Rear End', and 'Sideswipe Same Direction', demonstrate a higher frequency of hit-and-run incidents. In such cases, the driver involved in the crash flees the scene without exchanging information or providing aid. Recognizing these patterns is crucial for developing focused strategies to prevent hit-and-run occurrences and improve road safety across Chicago. By understanding these correlations, effective measures can be implemented to discourage hit-and-run behavior and ensure a safer driving environment for all residents.

Intro, Business Understanding	Pinqi Wang
Data Understanding	Onur Onel, Abhishek Pandav