Assessment: CST8390 Assignment 03

Section Number: *021*

Professor*: Grace Pauly*

Due Date: July 29, 2024

Team Members:

Ryan Xu, 041124348

Yizhen Xu, 040787525

Yuchen Wang, 041119096

# Table of Contents

[Table of Contents 2](#_Toc172180338)

[Table of Figures 3](#_Toc172180339)

[1 Introduction 4](#_Toc172180340)

[2 Business Understanding 4](#_Toc172180341)

[2.1 Objective 4](#_Toc172180342)

[2.2 Dataset Overview 4](#_Toc172180343)

[2.3 Data Source 5](#_Toc172180344)

[2.4 Explanation of dataset 5](#_Toc172180345)

[2.5 Results from other researchers 5](#_Toc172180346)

[2.6 Team Member Responsibilities 6](#_Toc172180347)

[3 Data Understanding 6](#_Toc172180348)

[3.1 Collect initial data 6](#_Toc172180349)

[3.2 Describe data 6](#_Toc172180350)

[3.3 Explore data 7](#_Toc172180351)

[3.4 Verify data quality 9](#_Toc172180352)

[4 Data Preparation 10](#_Toc172180353)

[4.1 Select data 10](#_Toc172180354)

[4.2 Clean data 11](#_Toc172180355)

[4.3 Construct data 12](#_Toc172180356)

[4.3.1 Date, Time attributes 12](#_Toc172180357)

[4.3.2 MO Codes 13](#_Toc172180358)

[4.3.3 For tree-based method 15](#_Toc172180359)

[4.3.4 For KNN / KMEANS / Outliers 16](#_Toc172180360)

[4.4 Integrate data 18](#_Toc172180361)

[4.5 Format data 18](#_Toc172180362)

[References 19](#_Toc172180363)

# Table of Figures

[Figure 1 histogram plot for Victim Age 8](#_Toc172180452)

[Figure 2 Bar chart for Victim Sex 8](#_Toc172180453)

[Figure 3 Bar chart for Victim Descent Category 9](#_Toc172180454)

[Figure 4 Missing values 10](#_Toc172180455)

[Figure 5 operators for select data initially 10](#_Toc172180456)

[Figure 6 Operators for filter data and sampling 12](#_Toc172180457)

[Figure 7 Operators for generating date / time related attributes 12](#_Toc172180458)

[Figure 8 Operators to deal with MO Codes 13](#_Toc172180459)

[Figure 9 Counts of MO Codes 13](#_Toc172180460)

[Figure 10 Histgram plot for MO Codes 14](#_Toc172180461)

[Figure 11 Operators for save file for decision tree 16](#_Toc172180462)

[Figure 12 distribution of 'Premise Code' 16](#_Toc172180463)

[Figure 13 Operators to deal with HourOccurred and Premise Code 17](#_Toc172180464)

[Figure 14 Operators for preparation of distance algorithm 17](#_Toc172180465)

[Figure 15 Prepare the data for KMEANS and KNN 18](#_Toc172180466)

# Introduction

The goal of this project is to analyze the Los Angeles traffic accident dataset, which uses traffic accident data from the past year. The goal is to apply various data mining techniques, including clustering using k-means, outlier detection using local outlier factor (LOF) and distance-based methods, and classification using decision trees and k-nearest neighbour (kNN) algorithms. The analysis will follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to ensure a structured approach.

# Business Understanding

## Objective

The primary goal of this analysis is to gain insights into traffic collisions in Los Angeles, identify patterns, and understand the factors contributing to these incidents. The specific tasks are:

* Classification: Use decision tree and kNN algorithms to classify traffic collisions based on severity and other relevant factors.
* Clustering: Apply k-means clustering to identify patterns and common characteristics of traffic collisions, as well as potential hotspots.
* Outlier Detection: Utilize LOF and distance-based methods to detect outliers in the dataset that may represent unusual or significant traffic collisions.

## Dataset Overview

The dataset used in this project is the Los Angeles Traffic Accident Dataset, which captures detailed information about traffic accidents reported in Los Angeles over the past year. Each instance in the Los Angeles Traffic Collision dataset represents a single traffic collision event.

The key attributes of the dataset include:

* Date and Time: Timestamp of the collision
* Location: Geographic coordinates (latitude and longitude) and address
* Type and Severity: Classification of the severity of the collision
* Parties: Information about the number of vehicles and Victim involved
* Crash Factors: Conditions and factors that led to the collision, such as speeding or drunk driving

## Data Source

The dataset is sourced from the Los Angeles city open data portal and is publicly available for analysis. It can be accessed at the following link: <https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w/about_data>

## Explanation of dataset

* Name: Traffic Collision Data
* Source: Los Angeles Police Department
* Instances:10000/610484
* Features: Totally 18 attribute with 15 features, 1 Id, 2 timestamp
* Creators: Los Angeles Police Department
* Missing Value: Yes

## Results from other researchers

In similar studies, researchers have used various machine learning techniques to analyze traffic collision data and derive meaningful insights. For example, in the study "A Data Mining Approach to Study the Impact of Weather Conditions on Road Accidents in the UK" by El Faouzi et al. (2011), the authors utilized decision trees and k-means clustering to examine the influence of weather on road accidents. They found that certain weather conditions, such as heavy rain and fog, significantly increased the likelihood of severe accidents. Clustering analysis revealed distinct patterns in accident types under different weather conditions, helping to identify high-risk scenarios (El Faouzi et al., 2011). [1]

Another relevant study is "Traffic Accident Analysis Using Machine Learning Paradigms" by Akinlar et al. (2012), where the authors applied classification and clustering algorithms to traffic accident data. Their findings indicated that decision trees and kNN classifiers were effective in predicting accident severity based on attributes such as time of day, location, and involved parties. The study also highlighted that clustering techniques could identify accident hotspots, which are critical for traffic management and safety improvements (Akinlar et al., 2012). [2]

These studies demonstrate the effectiveness of machine learning techniques in analyzing traffic collision data, uncovering patterns, and providing insights that can be used to enhance road safety measures. By applying similar methods to the Los Angeles Traffic Collision dataset, this project aims to achieve comparable results, contributing to the ongoing efforts to improve traffic safety in Los Angeles.

## Team Member Responsibilities

|  |  |
| --- | --- |
| Responsibilities | member |
| Introduction | Yizhen Xu |
| Business Understanding | Yuchen Wang |
| Data Understanding | Yizhen Xu |
| Data Preparation | Ryan Xu |
| Classification | Yizhen Xu |
| Clustering | Ryan Xu |
| Outlier detection | Yuchen Wang |
| Report | Yizhen Xu, Ryan Xu, Yuchen Wang |

# Data Understanding

## Collect initial data

Downloaded the data from reference website [3]: Traffic\_Collision\_Data\_from\_2010\_to\_Present\_20240711.csv.

## Describe data

Imported the dataset into AI Studio,

* Create a subfolder “Assignment3” in “Local Repository” -> “Data”
* Specify your data format: Header Row checked; File Encoding is UTF-8
* Format your columns: Check the attribute name and type, instance numbers
  + Set “DR Number” role to “id”
  + Check “Replace errors with missing values”, or the import will fail because of “NA” value
* Where to store the data? Select “Assignment3”

There are 17 regular attributes, 610,483 instances, attributes information as this table from data website:

| **Variable Name** | **Role** | **Type** | **Description** | **Units** | **Missing Values** |
| --- | --- | --- | --- | --- | --- |
| DR Number | ID | Integer | Case ID |  | no |
| Date Reported | Attribute | nominal | Date case was reported | Date |  |
| Date Occurred | Attribute | nominal | Date incident occurred | Date |  |
| Time Occurred | Attribute | Integer | Time incident occurred | Time |  |
| Area ID | ID | Integer | Identifier for the area |  |  |
| Area Name | Attribute | nominal | Name of the case occur area |  |  |
| Reporting District | ID | Integer | Identifier for the reporting district |  |  |
| Crime Code | ID | Integer | Code of the crime |  |  |
| Crime Code Description | Attribute | nominal | Description of the crime code |  |  |
| MO Codes | Attribute | nominal | Codes representing Modus Operandi |  |  |
| Victim Age | Label | Integer | Age of the victim | Years |  |
| Victim Sex | Attribute | nominal | Sex of the victim |  |  |
| Victim Descent | Attribute | nominal | Descent of the victim |  |  |
| Premise Code | ID | Integer | Code for the type of premises |  |  |
| Premise Description | Attribute | nominal | Description of premises |  |  |
| Address | Attribute | nominal | Address where the incident occurred |  |  |
| Cross Street | Attribute | nominal | Cross street of incident |  |  |
| Location | Attribute | nominal | GPS coordinates of the incident |  |  |

## Explore data

To visualize data, identify relationships among data, query data, etc.

This is the histogram plot of ‘Victim Age’:

* + Most of victims are between 20 and 60 years old, which is consistent with our feeling.
  + There is a small peak over 90 years old, which is not in line with common sense. It may be a record error, or simply writing a higher age for those whose age is unknown.

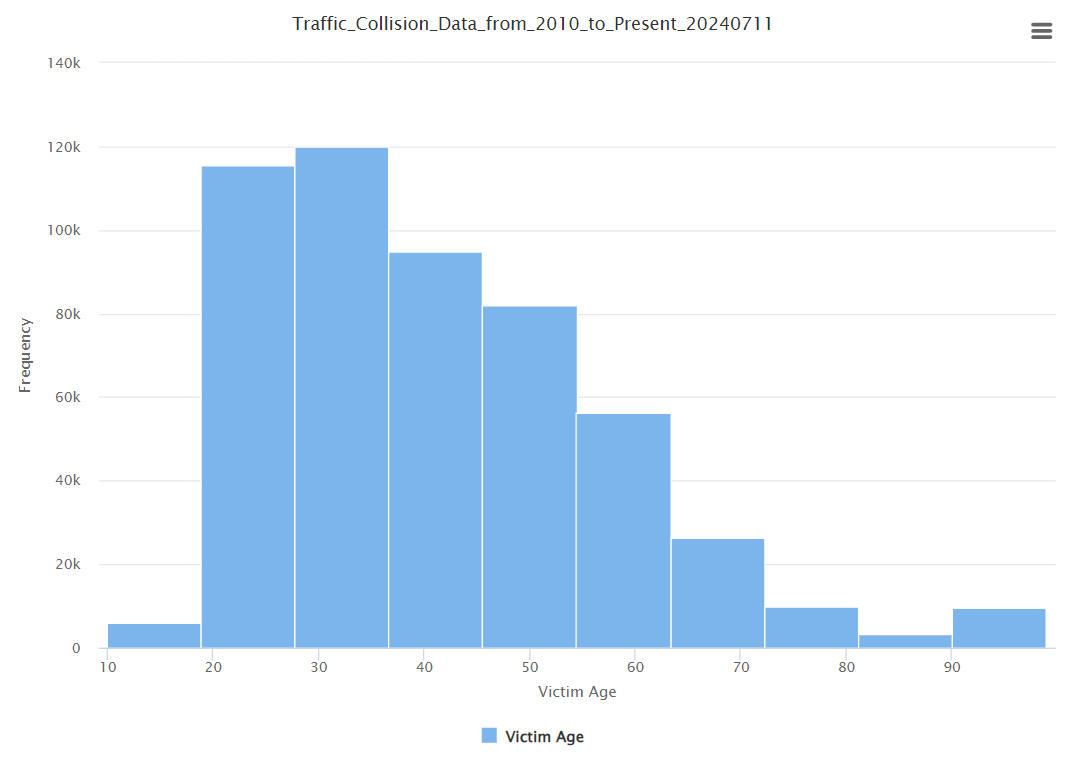


Figure 1 histogram plot for Victim Age

The Bar chart below is for Victim Sex distribution, most are Male, in addition to women, there are a small number of unknown, as the description only M, F and X can be included:

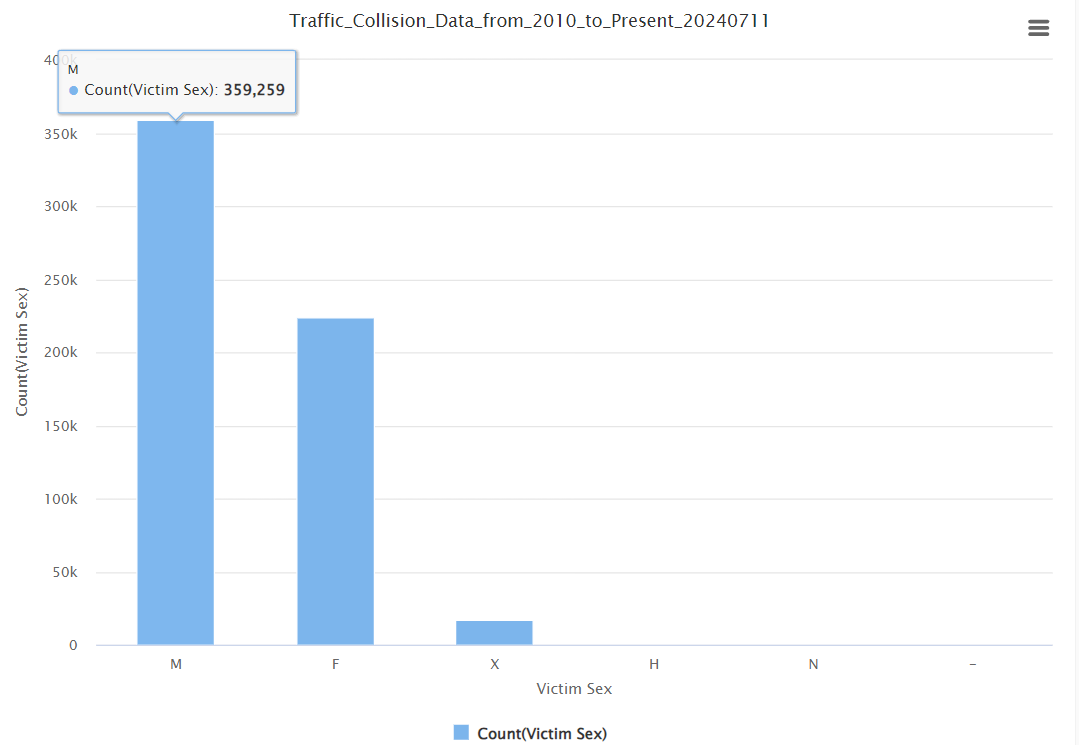


Figure 2 Bar chart for Victim Sex

The Bar chart below is for ‘Victim Descent’. H, W, O, B are the four most common categories, this may be related to the ethnic proportion of the local population.

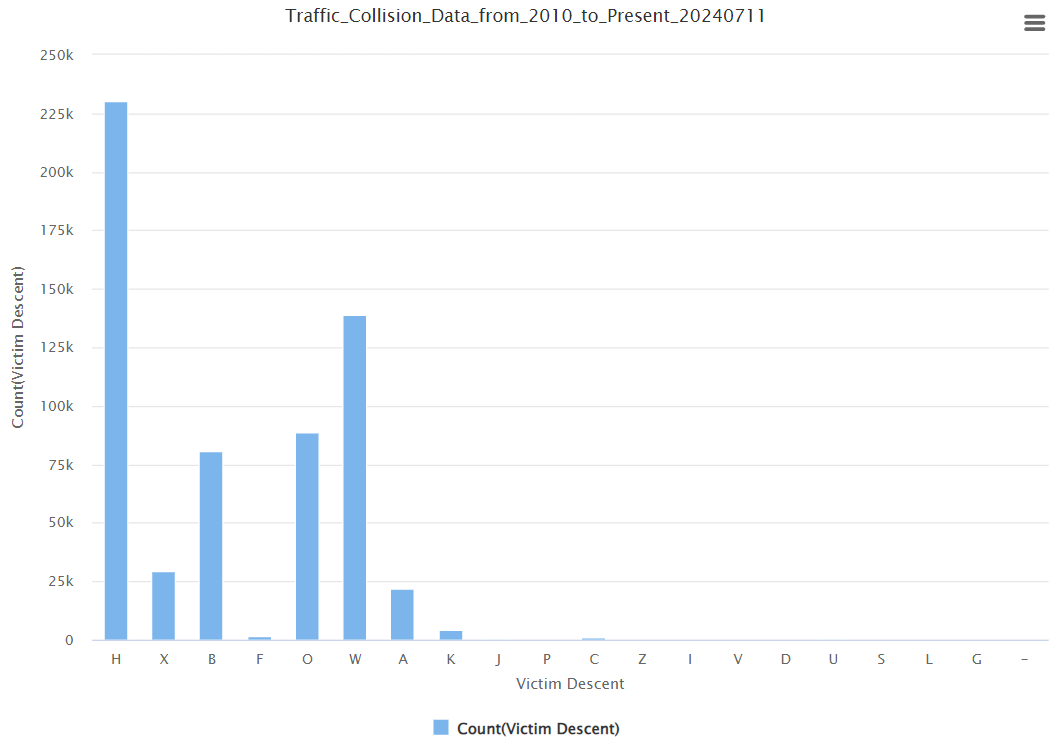


Figure 3 Bar chart for Victim Descent Category

## Verify data quality

* + **Missing Values:** After checking all attributes, we found there’re some missing values for several attributes, such as
    - ‘MO Codes’, missing 87236, 14.3% of 610k instances
    - ‘Victim Age’, missing 86912, 14.2% of 610k instances
    - ‘Victim Sex’, missing 10459, 1.7% of 610k instances
    - ‘Victim Descent’, missing 11409, 1.9% of 610k instances
    - ‘Premise Code’, missing 958, 0.2% of 610k instances

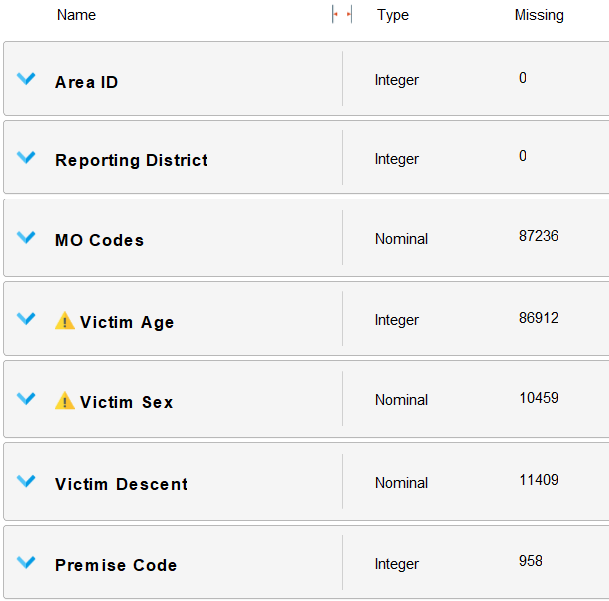


Figure 4 Missing values

* **Wrong Values:** mentioned above, there’re some wrong values in Victim Sex by H and N.
* **Duplicate:** we will sample the data without duplicate.

# Data Preparation

## Select data

These operators used to select data initially.

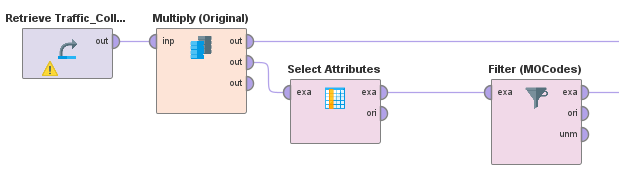


Figure 5 operators for select data initially

* **Retrieve** the data.
* **Select Attributes**, we excluded the following:
* **Address**, A specific address or street name, redundant information because this information is likely already captured by more structured fields such as Area ID, Premise Code, etc.
* **Cross Street**, same as above.
* **Area Name**, the description of Area ID.
* **Crime Code**, the entire table has only one Code, 997.
* **Crime Code Description**, description of Crime Code.
* **Premise Description**, description of Premise Code.
* **Location**, the location where the crime incident occurred. Actual address is omitted for confidentiality. XY coordinates reflect the nearest 100 block, this is also captured by more structured fields such as Area ID, Premise Code.
* **Filter examples** for select instances, as the table below which is the instances without missing data for each year, this expression is used to select instances in 2023, with at least 5 MO Codes(length>=24), and contain injury related codes:

**date\_get**([Date Occurred], DATE\_UNIT\_YEAR) **==** 2023 **&&**  
**length**([MO Codes])**>=**24 **&&**  
(**contains**([MO Codes], "3028") **||**   
**contains**([MO Codes], "3027") **||**   
**contains**([MO Codes], "3024") **||**   
**contains**([MO Codes], "3025") **||**   
**contains**([MO Codes], "3026"))

* **Sample**, we will sample after data cleaning.

## Clean data

* **Delete Redundancy instances**, no duplicates
* **Missing values:** 
  + **Filter (MissingData)**: filter examples to drop off instances with missing value, because there are enough instances, we can drop off them directly by this operator.
  + **Filter (Sex)**: filter examples to drop off instances with invalid Sex values, as the description in website, there are 3 values for ‘Victim Sex’, F, M and X, but H and N can be found in original data, so use this operator to filter H and N.
  + **Multiply**: from this operator we can know how many instances left, it’s 11,427.
  + **Sample**: sample 10,000 from 11,427 instances.

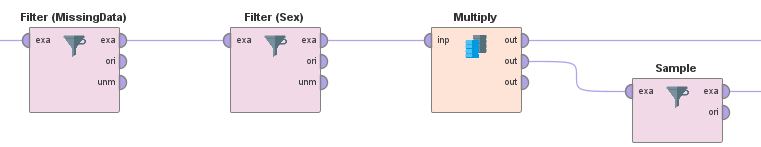


Figure 6 Operators for filter data and sampling

## Construct data

### Date, Time attributes

Several new attributes will be created for this dada analysis, such as ‘MonthOccured’, ‘HourOccurred’ and ‘ReportingDelay’,　we may find the reason for reporting delay.

* **Generate Attributes(Date/Time):** 
  + **MonthOccurred: date\_get**([Date Occurred], DATE\_UNIT\_MONTH)**+**1, because the result will be 0 to 11, so add 1 making 1 to 12
  + **HourOccurred**: floor([Time Occurred]/100), to get the hour
  + **ReportingDelay**: **date\_diff**([Date Occurred], [Date Reported], DATE\_UNIT\_DAY), this may give us some useful information.
* **Date to Numerical**: to generate ‘day of the week’, we think the traffic collisions are related to the day of week, but the default name is ‘Date Occured\_day’, the type is real
* **Real to Integer**: change the ‘Date Occured\_day’ to integer from real
* **Rename**: change ‘Date Occured\_day’ to ‘DayOfWeek’

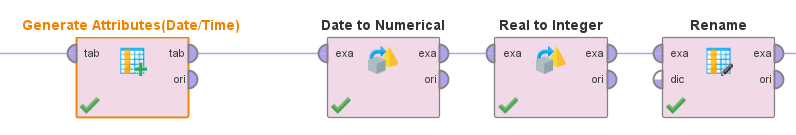


Figure 7 Operators for generating date / time related attributes

### MO Codes

In our 10000 sample instances, there are 105 different codes, the following plot is the top 50 codes, we can’t analyse all codes in this assignment, so we selected some meaningful codes as our research objects.

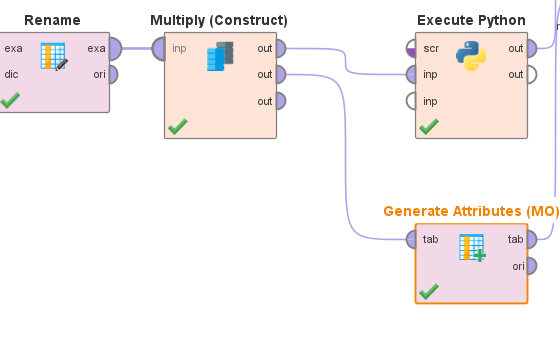


Figure 8 Operators to deal with MO Codes

* **Execute Python**: Count the code data so that we can decide which codes to choose

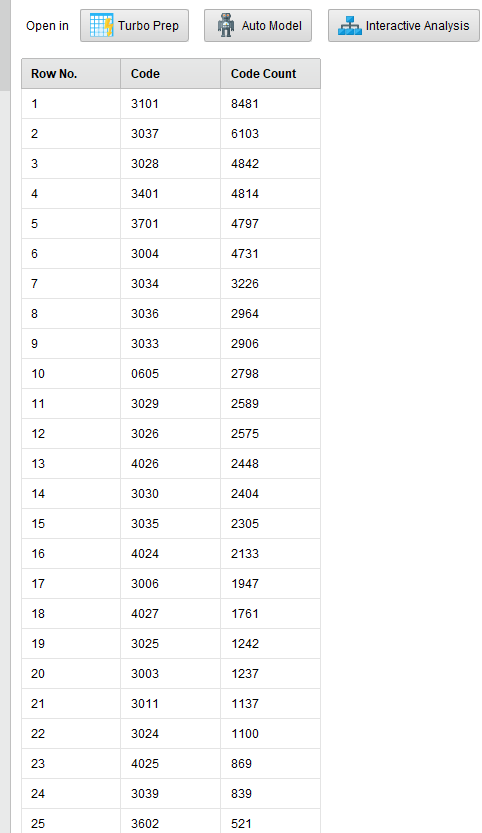


Figure 9 Counts of MO Codes

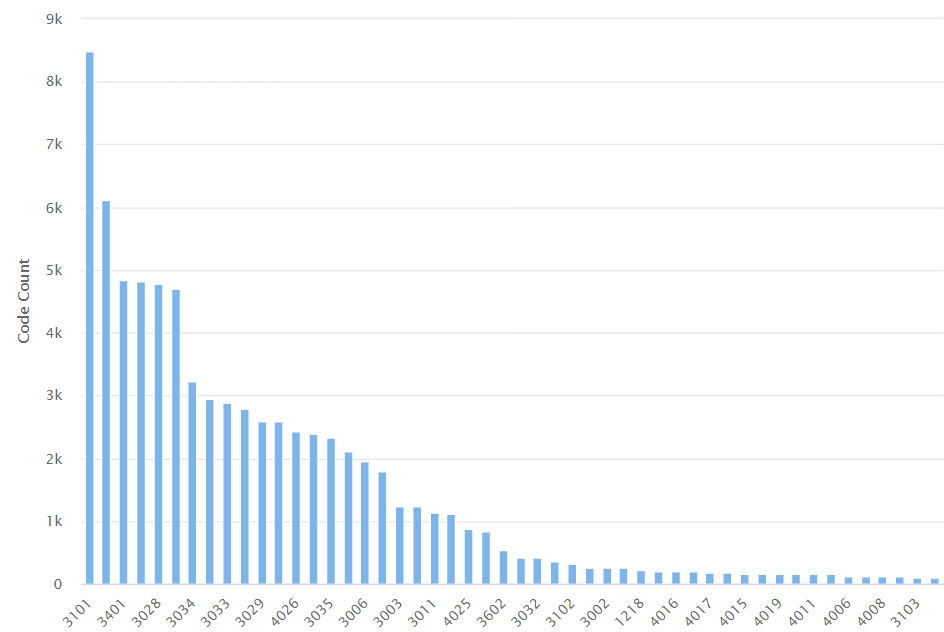


Figure 10 Histgram plot for MO Codes

* **Generate Attribute (MO)**: We decided to generate 3 attributes based the MO Codes
* **InjuryStatus,** includes these five codes, **InjuryStatus will be the label for our data analysis. New attribute was created by these codes:**

**if**(**contains**([MO Codes], "3028"), "0 - No Injury",   
 **if**(**contains**([MO Codes], "3027") **||** **contains**([MO Codes], "3024") **||** **contains**([MO Codes], "3025") **||** **contains**([MO Codes], "3026"), "1 - Injury", "9 - Unknown"))

|  |  |
| --- | --- |
| MO Code | Description |
| 3028 | T/C - (N) Non Injury |
| 3026 | T/C - (C) Complaint of Injury |
| 3025 | T/C - (B) Visible Injury |
| 3024 | T/C - (A) Severe Injury |
| 3027 | T/C - (K) Fatal Injury |

* **AtIntersection,** includes 2 codes, intersection may be a sensitive information. At the same time, we can consider two scenarios: if these two codes are in same instance, it means the quality of the instance has problem; Another is both are missing, it means one of them missed. The code is like InjuryStatus and is not shown here.

|  |  |
| --- | --- |
| MO Code | Description |
| 3036 | T/C - At Intersection - Yes |
| 3037 | T/C - At Intersection - No |

* **CollisionType,** we select 13 codes which may related to different injuries. The code is like InjuryStatus and is not shown here.

|  |  |
| --- | --- |
| MO Code | Description |
| 3003 | T/C - Veh vs Ped |
| 3004 | T/C - Veh vs Veh |
| 3006 | T/C - Veh vs Parked Veh |
| 3007 | T/C - Veh vs Train |
| 3008 | T/C - Veh vs Bike |
| 3009 | T/C - Veh vs M/C |
| 3011 | T/C - Veh vs Fixed Object |
| 3012 | T/C - Veh vs Other Object |
| 3013 | T/C - M/C vs Veh |
| 3014 | T/C - M/C vs Fixed Object |
| 3015 | T/C - M/C vs Other |
| 3016 | T/C - Bike vs Veh |
| 3018 | T/C - Bike vs Other |

### For tree-based method

We select attributes and set role; the data will be ready for decision tree.

* Select Attributes
  + **DR Number**: the ID
  + **InjuryStatus**: new attribute, the label, nominal
  + **DayOfWeek**: new attribute, the day of week, int but nominal
  + **Area ID**, **Reporting District**, **Premise Code**: regular attribute, int but nominal
  + **Victim Age**, regular attributes, int
  + **Victim Sex** and **Victim Descent**: regular attributes, nominal
  + **MonthOccurred**: new attribute, int but loop nominal
  + **HourOccurred**: new attribute, int but loop nominal
  + **ReportingDelay**: new attribute, int
  + **AtIntersection**: new attribute, nominal
  + **CollisionType**: new attribute, nominal
* **Set Role**: set **DR Number** as id, set **InjuryStatus** as label.
* **Write CSV**: save the data to csv, which will share to team members.

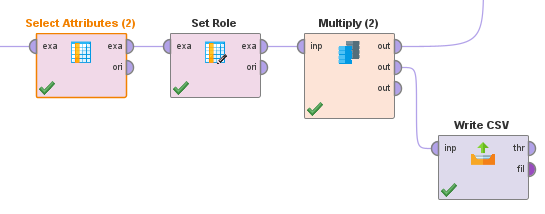


Figure 11 Operators for save file for decision tree

### For KNN / KMEANS / Outliers

Because the nominal data generated midway cannot be directly used for ‘One-hot encoding’ or ‘Nominal to Numerical’, , we should save it to CSV.

* **Generate Attributes**: generate two nominal attributes
  + **TimeOfDay** : binning hour to ‘Morning’, ‘Afternoon’, ‘Evening’ and ‘Night’
  + **NewPremiseCode**: the distribution of premise codes is extremely uneven, there’re 19 different values, to reduce the attributes generated by one-hot encoding, we simplify the 19 values to 4 categories in NewPremiseCode.

A graph with numbers and lines

Description automatically generated

Figure 12 distribution of 'Premise Code'

**if**(HourOccurred **>=** 6 **&&** HourOccurred **<** 12, "Morning",  
 **if**(HourOccurred **>=** 12 **&&** HourOccurred **<** 18, "Afternoon",  
 **if**(HourOccurred **>=** 18 **&&** HourOccurred **<** 24, "Evening", "Night")))

**if**(**contains**([Premise Code], "101"), "101",  
 **if**(**contains**([Premise Code], "108"), "108",  
 **if**(**contains**([Premise Code], "102"), "102", "Others")))

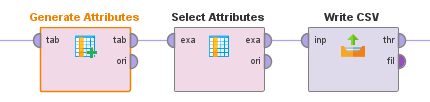


Figure 13 Operators to deal with HourOccurred and Premise Code

* **Select Attributes**: exclude ‘HourOccurred’ replaced by ‘TimeOfDay’, exclude ‘Premise Code’ replaced by ‘NewPremiseCode’.
* **Write CSV**: with this CSV file, ‘TimeOfDay’ and ‘NewPremiseCode’ can be processed by One-hot encoding or Nominal to Numerical in new process.

A close-up of a pink square

Description automatically generated

Figure 14 Operators for preparation of distance algorithm

* **Retrieve**: retrieve csv file saved in above step.
* **One-Hot Encoding**: encoding all nominal attributes except ‘InjuryStatus’ the label which should be used in KNN.
* **Select Attributes**: exclude all redundancy attributes, such as ‘Victim Sex = X’.
* **Set Role**: set ‘DR Number’ as id, ‘InjuryStatus’ as label

A diagram of a computer

Description automatically generated

Figure 15 Prepare the data for KMEANS and KNN

* **Nominal to Numerical**: for kMeans and Outlier detection, label is not needed, we can transfer the ‘InjuryStatu’ to numerical attribute which will help the models.
* **Select Attributes**: exclude the redundancy attribute ‘InjuryStatus=1 – injury’

## Integrate data

No integration data in this case.

## Format data

As part operators in Figure 14,

* **Normalize**: normalize all attributes except ‘DR Number’ which is the id
* **Write CSV(KMEANS)**: save the file which is ready for KMEANS and outliers detection
* **Normalize(2)**: normalize all attributes except special attributes for KNN.
* **Write CSV(KNN):** save the file which is ready for KNN.

Up to now, we saved 3 CSV files, one for Decision tree, one for KNN, and for KMEANS/Outliers Detection.

# References

|  |  |
| --- | --- |
| [1] | N.-E. L. J. &. K. J. El Faouzi, "High-Resolution Detector and Signal Data to Support Crash Identification and Reconstruction," SageJournals, 1 1 2011. [Online]. Available: https://doi.org/10.3141/2237-14. [Accessed 15 07 2024]. |
| [2] | C. T. C. &. E. O. K. Akinlar, "Composer Classification in Symbolic Data Using PPM," IEEE.ORG, 15 12 2012. [Online]. Available: https://doi.org/10.1109/ICMLA.2012.176. [Accessed 15 07 2024]. |
| [3] | L. A. P. Department, "Traffic Collision Data from 2010 to Present," data.lacity.org, 11 7 2024. [Online]. Available: https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w/about\_data. [Accessed 11 7 2024]. |