Traffic Crashes

CONTENTS

PART 01

Dataset Description

Dataset Description

Data Source

The dataset we choosed is from the website Chicago Data Portal .

Data Description

This dataset records traffic accidents in Chicago from 2014 to the present. It has 780k rows, and every instance has 49 features. Each row is a traffic crash.

Dataset Description

Features Description

The row in this dataset has 49 features. These features primarily depict information about the time and location of the accidents, the extent of injuries to individuals, the number of vehicles involved, the severity of vehicle damage, the geographic environment of the accident location, and the traffic configuration at the accident site. After preprocessing the data, we can analyze it from various perspectives.

CRASH_DATE	Date and time of crash as entered by the reporting officer	Date & Time	
POSTED_SPEED_LIMIT	Posted speed limit, as determined by reporting officer	Number	#
TRAFFIC_CONTROL_DEVICE	Traffic control device present at crash location, as determined ...	Plain Text	T
DEVICE_CONDITION	Condition of traffic control device, as determined by reporting ...	Plain Text	T
WEATHER_CONDITION	Weather condition at time of crash, as determined by reporting...	Plain Text	T
LIGHTING_CONDITION	Light condition at time of crash, as determined by reporting o...	Plain Text	T
FIRST_CRASH_TYPE	Type of first collision in crash	Plain Text	T
TRAFFICWAY_TYPE	Trafficway type, as determined by reporting officer	Plain Text	T
LANE_CNT	Total number of through lanes in either direction, excluding tur...	Number	#
ALIGNMENT	Street alignment at crash location, as determined by reporting ...	Plain Text	T
ROADWAY_SURFACE_COND	Road surface condition, as determined by reporting officer	Plain Text	T
ROAD_DEFECT	Road defects, as determined by reporting officer	Plain Text	T
REPORT_TYPE	Administrative report type (at scene, at desk, amended)	Plain Text	T
CRASH_TYPE	A general severity classification for the crash. Can be either Inj...	Plain Text	T
INTERSECTION_RELATED_I	A field observation by the police officer whether an intersectio...	Plain Text	T
NOT_RIGHT_OF_WAY_I	Whether the crash begun or first contact was made outside of ...	Plain Text	T
HIT_AND_RUN_I	Crash did/did not involve a driver who caused the crash and fl...	Plain Text	T
DAMAGE	A field observation of estimated damage.	Plain Text	T
DATE_POLICE_NOTIFIED	Calendar date on which police were notified of the crash	Date & Time	

PART 02

Data Preprocessing

Data Propessing

The purpose of data preprocessing

In this phase, the data is prepared for the analysis purpose which contains relevant information. Pre-processing and cleaning of data are one of the most important tasks that must be one before dataset can be used for machine learning. The real-world data is noisy, incomplete and inconsistent. So, it is required to be cleaned.

Data Propessing

These are the following steps taken to clean the data:

1. Data normalization
2. Remove duplicate data
3. Handling missing values:

For variables with a high missing rate (greater than 80%), low coverage, and low importance, they can be directly deleted. For variables with a low missing rate (less than 95%) and low importance, missing values can be imputed based on the data distribution. If the data follows a uniform distribution, missing values can be filled with the mean of that variable; for skewed distributions, the median can be used for imputation.

4. Outlier handling:

We use the Pandas describe function and the 3-sigma rule to identify outliers. Once outliers are identified, we first attempt to eliminate them by applying a log-scale transformation. If this does not suffice, we resort to replacing outliers with either the mean or median, as this approach is straightforward and minimizes information loss.

Data Propessing

Delete unimportant features

For our analysis, the following features are unimportant and can be removed:

CRASH_RECORD_ID, RD_NO, CRASH_DATE_EST_I, REPORT_TYPE, STREET_NO, PHOTOS_TAKEN_I, STATEMENTS_TAKEN_I, WORKERS_PRESENT_I, INJURIES_UNKNOWN, INJURIES_INCAPACITATING, INJURIES_NON_INCAPACITATING. INJURIES_REPORTED_NOT_EVIDENT, INJURIES_NO_INDICATION, DAMAGE, DATE_POLICE_NOTIFIED, NUM_UNITS, STREET_DIRECTION, STREET_NAME, LANE_CNT, SEC_CONTRIBUTORY_CAUSE, DOORING_I

After removing all unnecessary features, the shape of the dataframe has become:

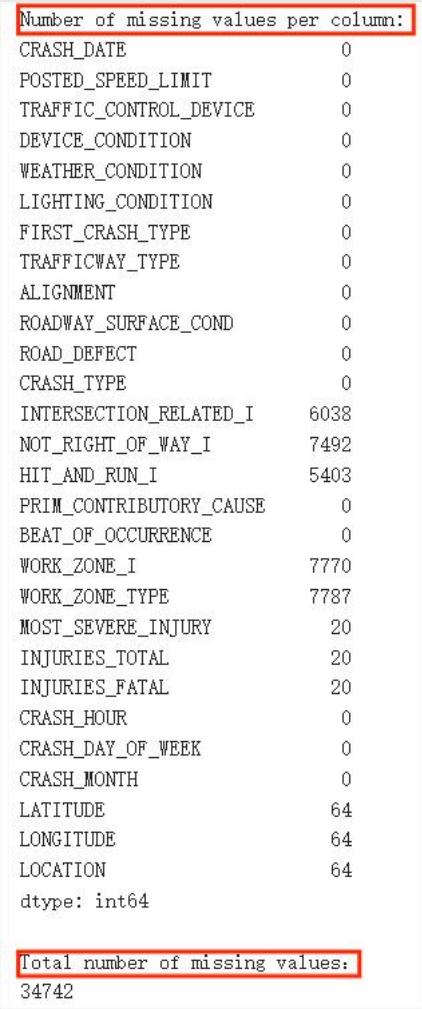
```
The initial shape of the dataframe is:
(7822, 49)
The shape of the dataframe after drop the unneccessary columns is:
(7822, 28)
```

This has saved us a lot of time for our subsequent analysis.

Data Propessing

Handling missing values

Firstly, by using the code, we obtained the number of missing values for each column in the dataset, as shown in the figure on the right: Among these, the features INTERSECTION_RELATED_I, NOT_RIGHT_OF_WAY_I, HIT_AND_RUN_I, WORK_ZONE_I, WORK_ZONE_TYPE have excessive missing values and could be considered for deletion. For features with fewer missing values, such as INJURIES_TOTAL and INJURIES_FATAL, mean imputation could be applied. For the remaining features, filling with 'unknown' can be considered.



```
Number of missing values per column:
CRASH_DATE                     0
POSTED_SPEED_LIMIT             0
TRAFFIC_CONTROL_DEVICE         0
DEVICE_CONDITION               0
WEATHER_CONDITION              0
LIGHTING_CONDITION             0
FIRST_CRASH_TYPE               0
TRAFFICWAY_TYPE                0
ALIGNMENT                      0
ROADWAY_SURFACE_COND           0
ROAD_DEFECT                    0
CRASH_TYPE                     0
INTERSECTION_RELATED_I      6038
NOT_RIGHT_OF_WAY_I          7492
HIT_AND_RUN_I               5403
PRIM_CONTRIBUTORY_CAUSE        0
BEAT_OF_OCCURRENCE             0
WORK_ZONE_I                 7770
WORK_ZONE_TYPE              7787
MOST_SEVERE_INJURY            20
INJURIES_TOTAL                20
INJURIES_FATAL                20
CRASH_HOUR                     0
CRASH_DAY_OF_WEEK              0
CRASH_MONTH                    0
LATITUDE                      64
LONGITUDE                     64
LOCATION                      64
dtype: int64

Total number of missing values:
34742
```

Data Propessing

Processing of numerical features POSTED_SPEED_LIMITh and INJURITES_TOTAL

During the analysis, we observed inaccuracies in certain records for the feature POSTED_SPEED_LIMIT, leading us to remove these samples. Subsequently, we performed Min-Max scaling on two numerical features.

Min-Max scaling rescales data to a specified range (usually [0, 1] or [-1, 1]). The formula is as follow:

$$x_{\text{norm}} = \frac{(x - \min(x))}{(\max(x) - \min(x))}$$

Where x is the original data, min(x) and max(x) are the minimum and maximum values of the data, respectively.

Data Propessing

Category classification

To facilitate subsequent analysis, categorize the features MOST_SEVERE_INJURY and CRASH_TYPE.

```
# Classify the MOST_SEVERE_INJURY feature into categories
severity_mapping = {
    'NO INDICATION OF INJURY': 'NO_INJURY',
    'NONINCAPACITATING INJURY': 'MINOR_INJURY',
    'REPORTED, NOT EVIDENT': 'NOT_EVIDENT',
    'INCAPACITATING INJURY': 'INCAPACITATING',
    'UNKNOWN': 'UNKNOWN',
    'FATAL': 'FATAL'
}
df['MOST_SEVERE_INJURY'] = df['MOST_SEVERE_INJURY'].map(severity_mapping)

# Classify CRASH_TYPE features into categories
crash_type_mapping = {
    'NO INJURY / DRIVE AWAY': 'NO_INJURY',
    'INJURY AND / OR TOW DUE TO CRASH': 'INJURY_OR_TOW'
}
df['CRASH_TYPE'] = df['CRASH_TYPE'].map(crash_type_mapping)
```

Data Propessing

One-hot encoding and Label Encoding

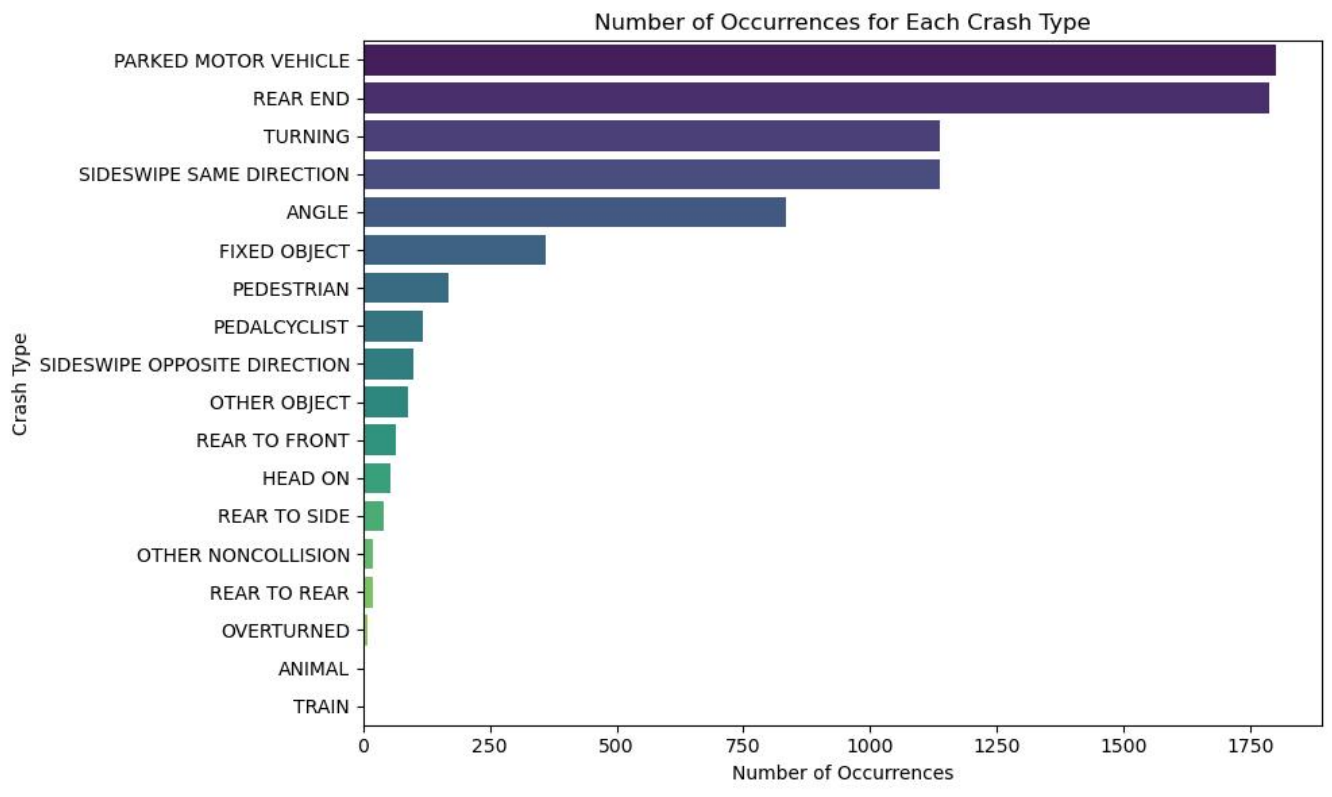
Perform one-hot encoding on the feature 'TRAFFIC_CONTROL_DEVICE', and apply Label Encoding to the features 'DEVICE_CONDITION', 'WEATHER_CONDITION', 'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'ALIGNMENT', 'ROADWAY_SURFACE_COND', and 'ROAD_DEFECT'.

PART 03

Dataset Analysis

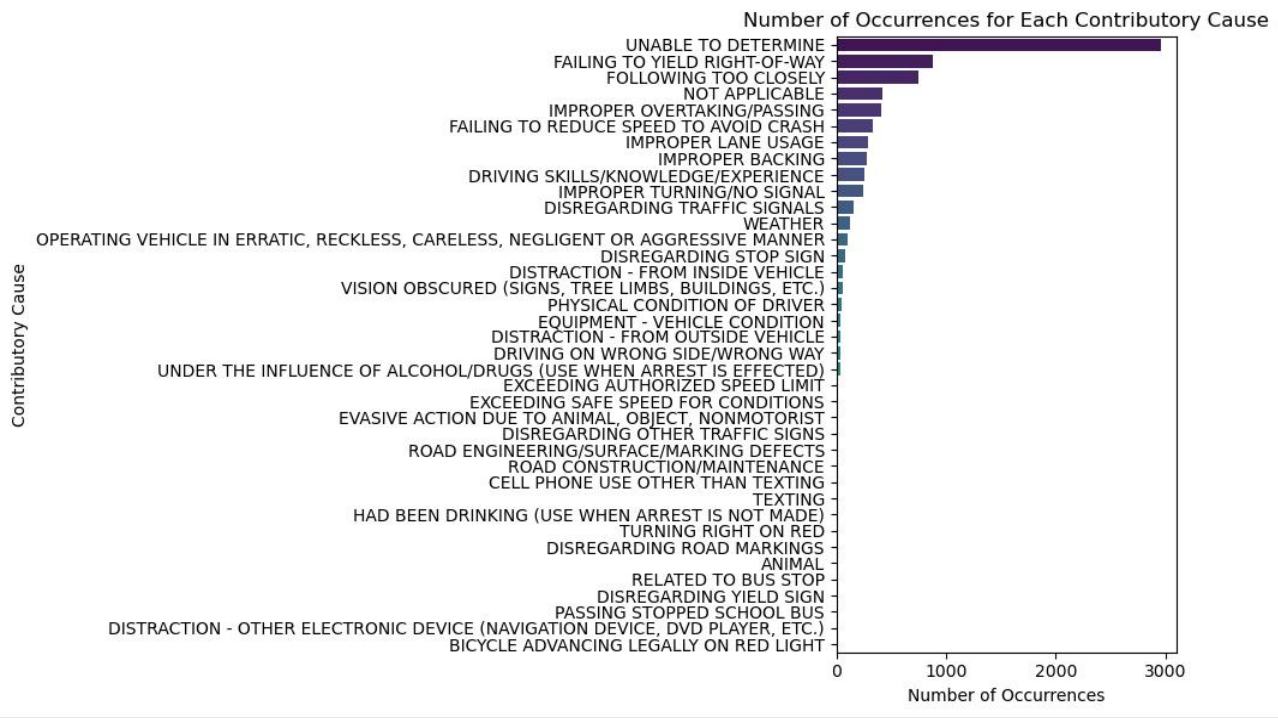
Dataset Analysis

Visualize the feature 'FIRST_CRASH_TYPE' for insights; this reveals that the most frequent accident types are 'PARKED MOTOR VEHICLE' and 'REAR END'.



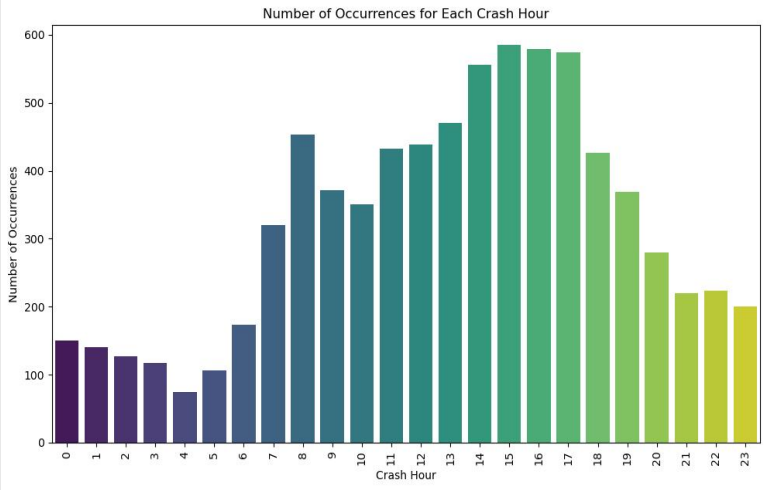
Dataset Analysis

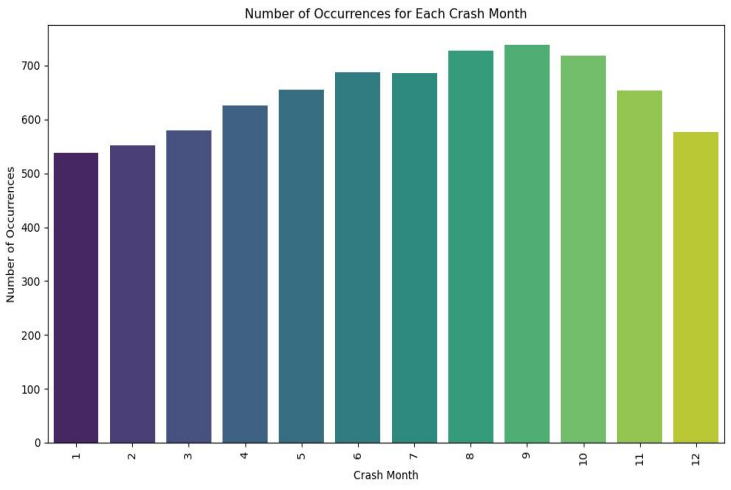
Analyze the feature 'PRIM_CONTRIBUTORY_CAUSE' visually; it unveils that apart from 'UNABLE TO DETERMINE', the primary cause of accidents is 'FAILING TO YIELD RIGHT-OF-WAY'.



Dataset Analysis

The relationship between the occurrence of accidents and the 'HOUR' and 'MONTH'.





From the bar chart above, it's evident that accidents predominantly occur between 7 AM and 7 PM daily. There isn't a significant variance observed across different months.

Dataset Analysis

Geographic Analysis

Based on the map analysis, it's apparent that accidents are more frequent in downtown Chicago. This aligns with our intuition, considering the higher volume of vehicles in urban centers.