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:

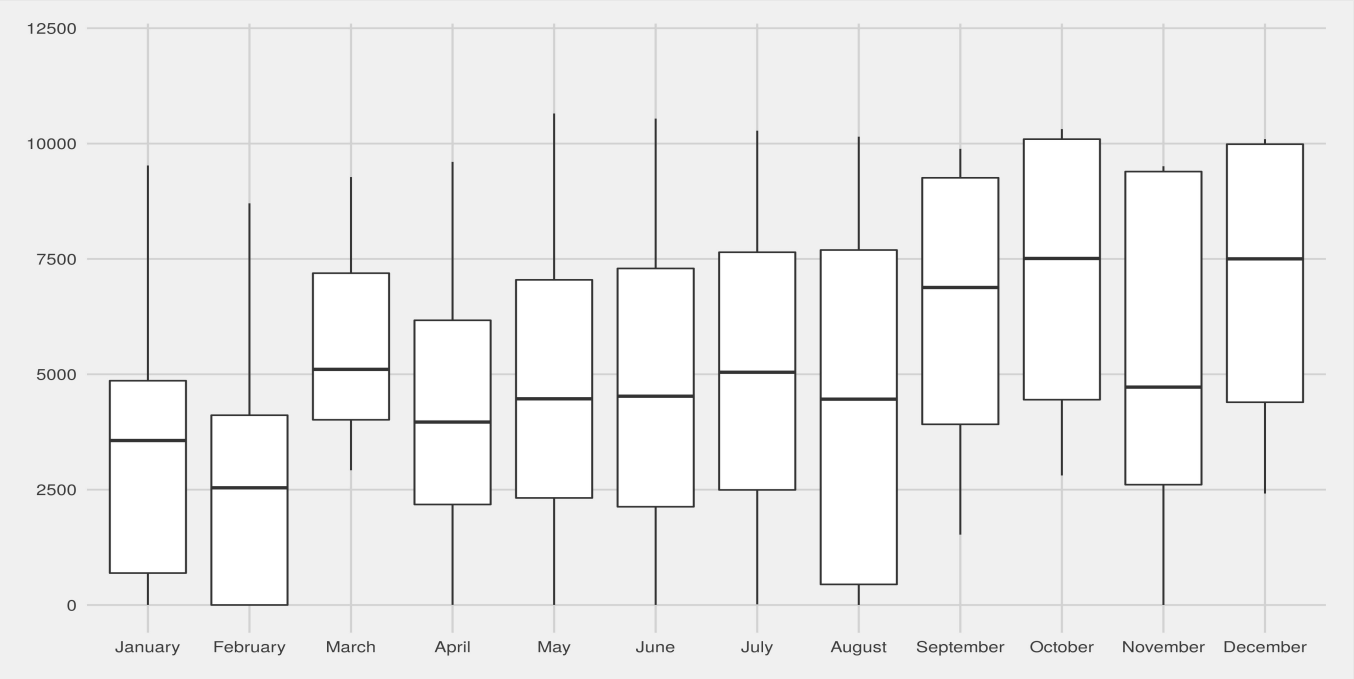
We use the Pandas describe function and the 3-sigma rule to identify outliers. Once outliers are identified, we first attempt to eliminate them by applying a log-scale transformation. If this does not suffice, we resort to replacing outliers with either the mean or median, as this approach is straightforward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

PART 03

Dataset Analysis

Dataset Analysis

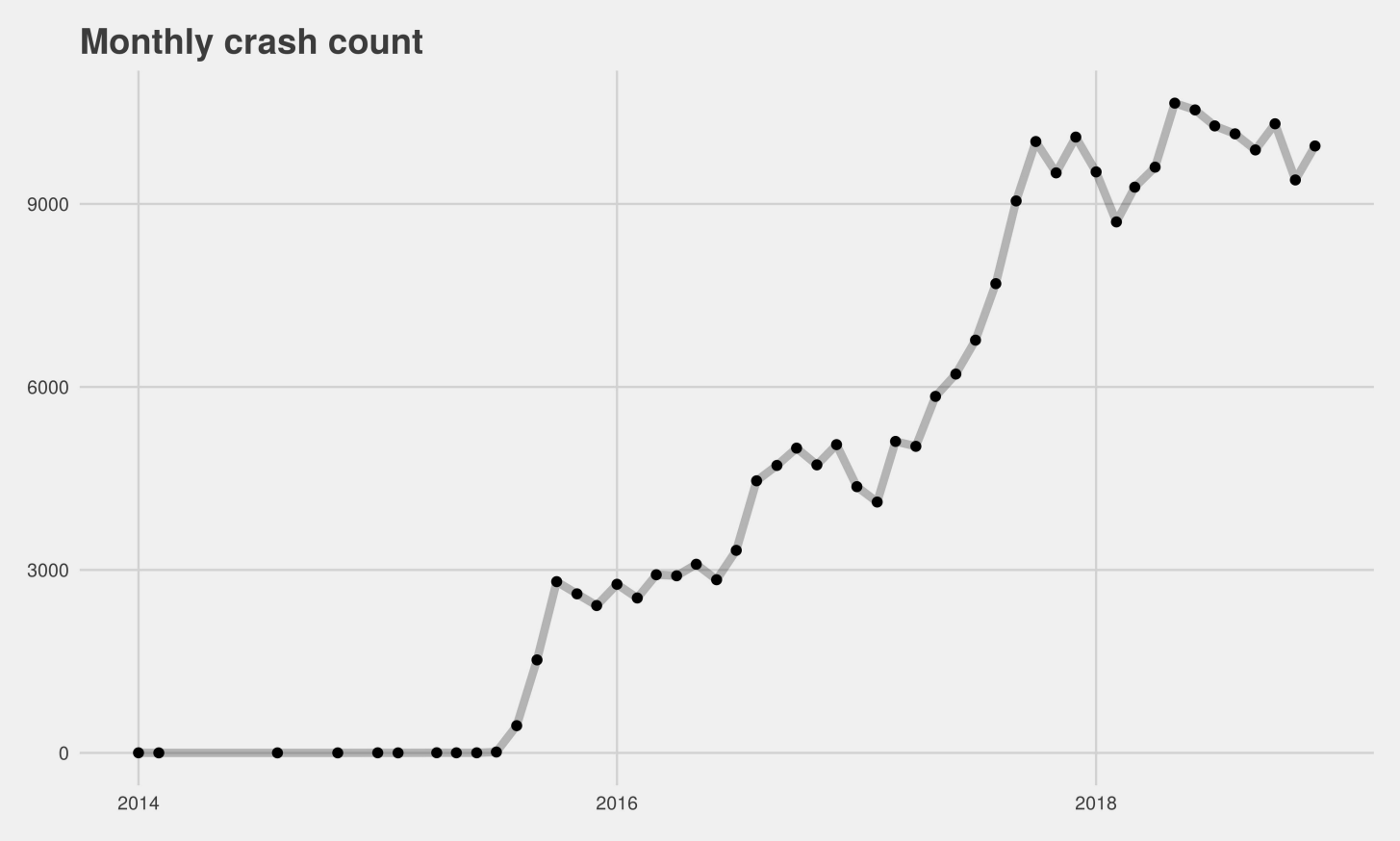
First, we analyzed the number of traffic accidents that occurred in different months of the year.



From the graph, it can be observed that the frequency of traffic accidents tends to increase roughly with the growth of the months.

Dataset Analysis

Continuing, we analyzed the variation in the number of accidents on a monthly basis, starting from the year 2014.



From the graph, it is evident that the number of accidents per month significantly increases with the progression of dates.

Dataset Analysis

Geographic Analysis

Analyzing the combination of the Chicago map and accident data from 2014 to 2019 reveals the distribution of accidents within Chicago. From the image, it can be observed that in the areas marked with red circles, accidents occur more frequently, while in the regions marked with blue circles, accidents are relatively rare.