Statistical Analysis Of US Accidents (2016 - 2023) Data

Mohammad Rouintan

Computer Science, Shahid Beheshti University

Abstract

The US Accidents dataset provides a rich and extensive collection of car accident data spanning 49 states in the United States. Collected from February 2016 to March 2023 through multiple Traffic APIs, this dataset offers a unique opportunity to delve into the realm of traffic incidents and their underlying causes. With approximately 7.7 million accident records, it presents a valuable resource for a wide range of applications, including real-time accident prediction, hotspot identification, casualty analysis, and the study of environmental influences on accident occurrence. This article investigates the potential of the US Accidents dataset in uncovering deep insights and patterns within car accidents. The dataset's comprehensive coverage, incorporating data from government transportation departments, law enforcement agencies, traffic cameras, and road sensors, ensures a holistic representation of accidents nationwide. By examining variables such as location, weather conditions, and time of occurrence, researchers can develop predictive models to enhance road safety and optimize traffic management strategies.

Introduction

Our task for this project is to perform statistical tests on the dataset as well as exploratory data analysis to find important data behaviors and characteristics. In carrying out this project, we have used data science libraries such as NumPy, Pandas, SciPy for statistical data analysis, as well as Matplotlib for data visualization.

This dataset contains approximately 7.7 million samples collected from accidents in the United States, each of which has the same features that can help us find different reasons for the occurrence of accidents, the conditions of the accident location, etc.

first, for data analysis we should read description of dataset and know features of dataset in detail. This dataset contains:

- ***ID***: (This is a unique identifier of the accident record.)
- ***Source***: (Source of raw accident data)
- ***Severity***: (Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).)
- ***Start_Time***: (Shows start time of the accident in local time zone.)
- ***End_time***: (Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.)
- ***Start_Lat***: (Shows latitude in GPS coordinate of the start point.)
- ***Start_Lng***: (Shows longitude in GPS coordinate of the start point.)
- ***End_Lat***: (Shows latitude in GPS coordinate of the end point.)
- ***End_Lng***: (Shows longitude in GPS coordinate of the end point.)
- ***Distance(mi)***: (The length of the road extent affected by the accident in miles.)
- ***Description***: (Shows a human provided description of the accident.)
- ***Street***: (Shows the street name in address field.)
- ***City***: (Shows the city in address field.)
- ***County***: (Shows the county in address field.)
- ***State***: (Shows the state in address field.)
- ***Zipcode***: (Shows the zipcode in address field.)
- ***Country***: (Shows the country in address field.)
- ***Timezone***: (Shows timezone based on the location of the accident (eastern, central, etc.).)
- ***Airport_Code***: (Denotes an airport-based weather station which is the closest one to location of the accident.)
- ***Weather_Timestamp***: (Shows the time-stamp of weather observation record (in local time).)
- ***Temperature(F)***: (Shows the temperature (in Fahrenheit).)
- ***Wind_Chill(F)***: (Shows the wind chill (in Fahrenheit).)
- ***Humidity(%)***: (Shows the humidity (in percentage).)
- ***Pressure(in)***: (Shows the air pressure (in inches).)
- ***Visibility(mi)***: (Shows visibility (in miles).)
- ***Wind_Direction***: (Shows wind direction.)
- ***Wind_Speed(mph)***: (Shows wind speed (in miles per hour).)
- ***Precipitation(in)***: (Shows precipitation amount in inches, if there is any.)
- ***Weather_Condition***: (Shows the weather condi-

tion (rain, snow, thunderstorm, fog, etc.))

- ***Amenity***: (A POI annotation which indicates presence of amenity in a nearby location.)
- ***Bump***: (A POI annotation which indicates presence of speed bump or hump in a nearby location.)
- ***Crossing***: (A POI annotation which indicates presence of crossing in a nearby location.)
- ***Give_Way***: (A POI annotation which indicates presence of give_way in a nearby location.)
- ***Junction***: (A POI annotation which indicates presence of junction in a nearby location.)
- ***No_Exit***: (A POI annotation which indicates presence of no_exit in a nearby location.)
- ***Railway***: (A POI annotation which indicates presence of railway in a nearby location.)
- ***Roundabout***: (A POI annotation which indicates presence of roundabout in a nearby location.)
- ***Station***: (A POI annotation which indicates presence of station in a nearby location.)
- ***Stop***: (A POI annotation which indicates presence of stop in a nearby location.)
- ***Traffic_Calming***: (A POI annotation which indicates presence of traffic_calming in a nearby location.)
- ***Traffic_Signal***: (A POI annotation which indicates presence of traffic_signal in a nearby location.)
- ***Turning_Loop***: (A POI annotation which indicates presence of turning_loop in a nearby location.)
- ***Sunrise_Sunset***: (Shows the period of day (i.e. day or night) based on sunrise/sunset.)
- ***Civil_Twilight***: (Shows the period of day (i.e. day or night) based on civil twilight.)
- ***Nautical_Twilight***: (Shows the period of day (i.e. day or night) based on nautical twilight.)
- ***Astronomical_Twilight***: (Shows the period of day (i.e. day or night) based on astronomical twilight.)

Methodologies

Handling Missing Values and Inefficient Features

The two features of ID and Source are not useful for us, so we delete them. This dataset has many missing values. We delete those columns that have missing value more than 1 million. There are still a number of columns that have significant missing values, but if we want to delete them, we cannot analyze the data properly. Therefore, we either have to fill them in some way or use samples that do not have missing values. In the start time column, the time of the accident along with its date is mentioned. We have expanded this, that is, we have added four columns to the dataset with the titles: Year, Month, Day, Hour. This allows us to more easily deal with the time of the accident and analyze it.

Hypothesis Testing

Question1: Average of Temperature(F) is 60 (one sample T-test)

$$H_0 : \mu = 60 \tag{1}$$

$$H_1 : \mu \neq 60 \tag{2}$$

reject null hypothesis

Question2: Average of Humidity(%) is 65 (one sample T-test)

$$H_0 : \mu = 65 \tag{3}$$

$$H_1 : \mu \neq 65 \tag{4}$$

reject null hypothesis

Question3: Average of Pressure(in) is 30 (one sample T-test)

$$H_0 : \mu = 30 \tag{5}$$

$$H_1 : \mu \neq 30 \tag{6}$$

reject null hypothesis

Question4: Average of Wind_Speed(mph) is 7 (one sample T-test)

$$H_0 : \mu = 7 \tag{7}$$

$$H_1 : \mu \neq 7 \tag{8}$$

reject null hypothesis

Question5: Severity as categorical feature and Timezone as categorical feature (chi2 Test) Reject H0,There is a relationship between 2 categorical variables

Question6: Severity as categorical feature and City as categorical feature (chi2 Test) Reject H0,There is a relationship between 2 categorical variables

Question7: Severity as categorical feature and State as categorical feature (chi2 Test) Reject H0,There is a relationship between 2 categorical variables

Question8: State Effect on Temperature(F) (ANOVA)

$$H_0 : \mu_{California} = \mu_{Florida} = \mu_{Texas} \tag{9}$$

$$H_1 : \mu_{California} \neq \mu_{Florida} \neq \mu_{Texas} \tag{10}$$

reject null hypothesis

Question9: Timezone Effect on Temperature(F) (ANOVA)

$$H_0 : \mu_{US/Eastern} = \mu_{US/Pacific} = \mu_{US/Central} = \mu_{US/Mountai} \tag{11}$$

$$H_1 : \mu_{US/Eastern} \neq \mu_{US/Pacific} \neq \mu_{US/Central} \neq \mu_{US/Mounta} \tag{12}$$

reject null hypothesis

Exploratory Data Analysis (EDA)

In the EDA section of this data, we review and analyze a series of categorical data. We will examine this using different types of plots.

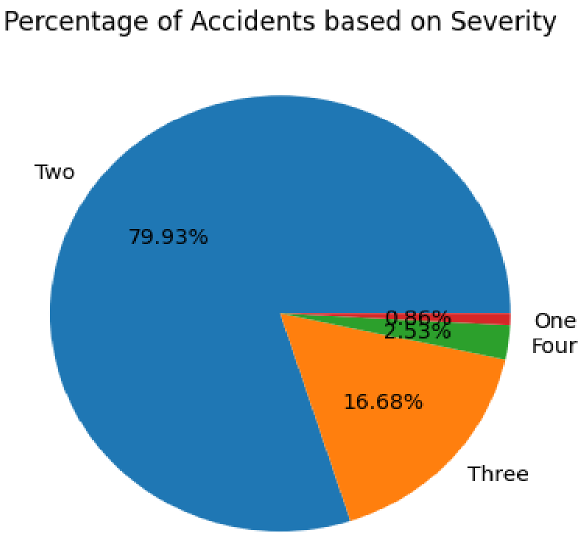

Figure 1: *PiePlot Of Severity feature. Result: Approximately 80% of accidents belong to class 2 and 17% belong to class 1*

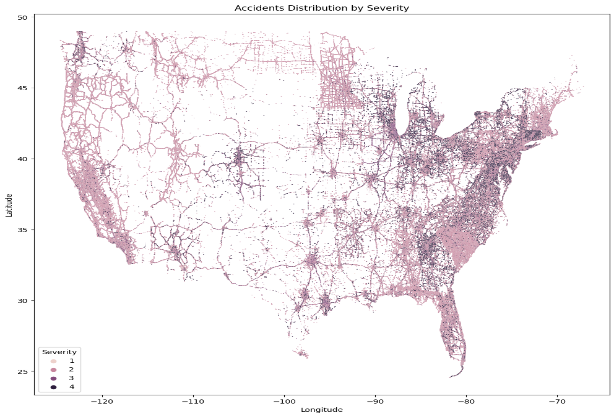

Figure 2: *Density of the number of accidents by different severities on a map of the United States*

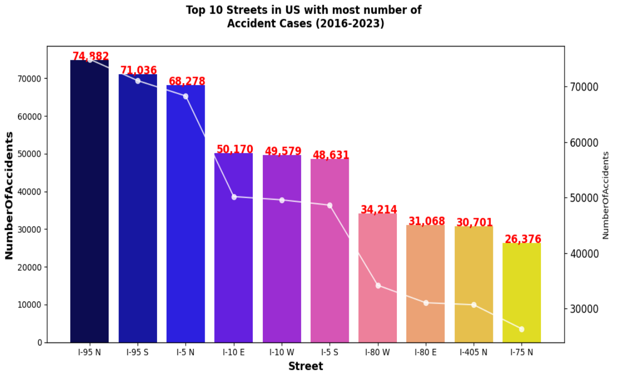

Figure 3: *Top10 Streets in US with most number of Accident Cases(2016 - 2023)*

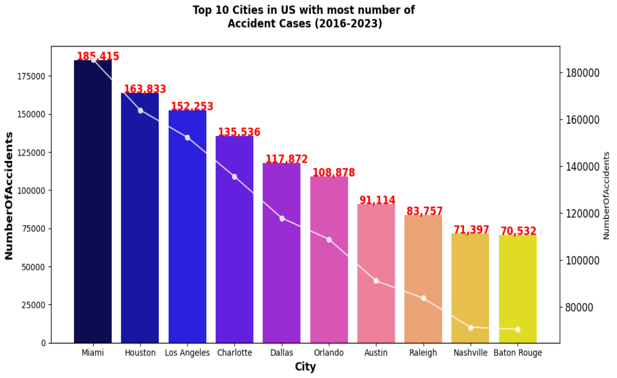

Figure 4: *Top10 Cities in US with most number of Accident Cases(2016 - 2023)*

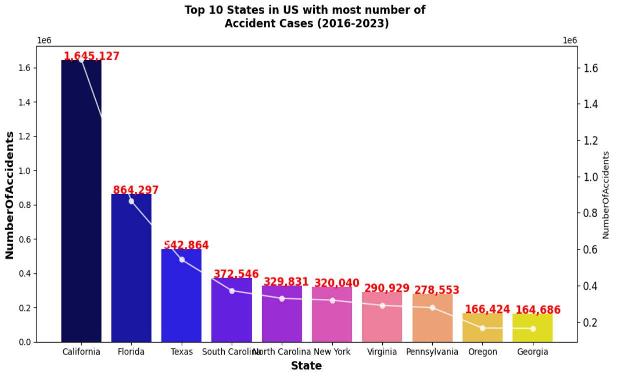

Figure 5: *Top10 Cities in US with most number of Accident Cases(2016 - 2023)*

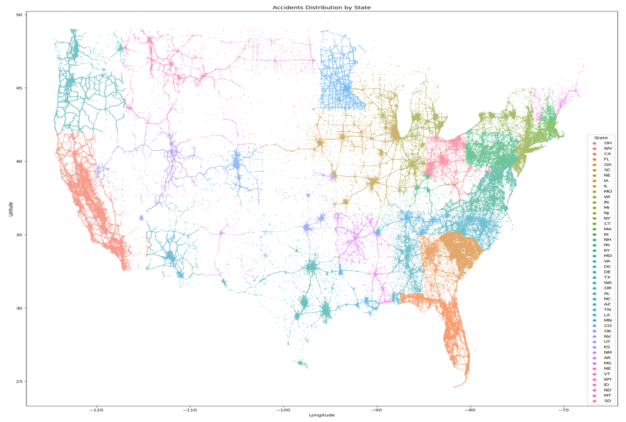

Figure 6: *Density of the number of accidents by different state on a map of the United States*

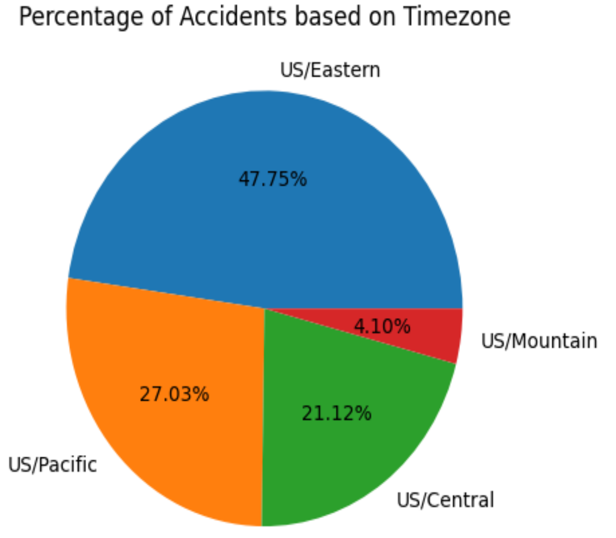

Figure 7: *PiePlot Of Timezone feature. Result: Approximately 47.75% of accidents belong to class US/Eastern, 27.03% belong to class US/Pacific, 21.12% belong to class US/Central and 4.10% belong to class US/Mountain*

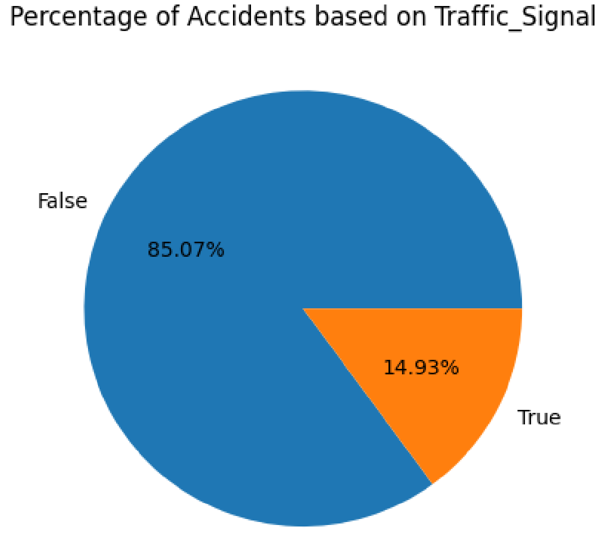

Figure 8: *PiePlot Of Traffic_Signal feature. Result: Almost 15% of accidents happened near traffic signals*

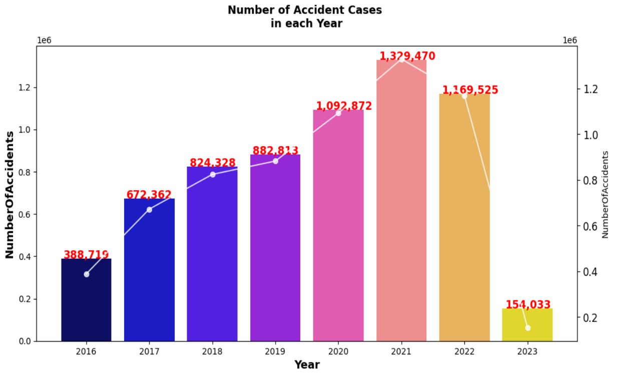

Figure 9: *The CountPlot Of Year Feature. Result: Most of the accidents happened between 2020 and 2022.*

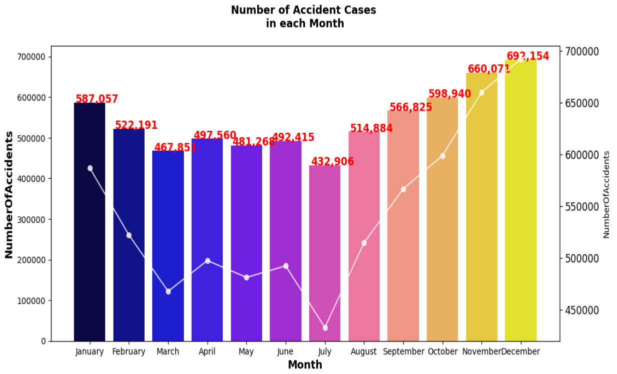

Figure 10: *The CountPlot Of Month Feature. Result: Most of the accidents happened in last three months of year.*

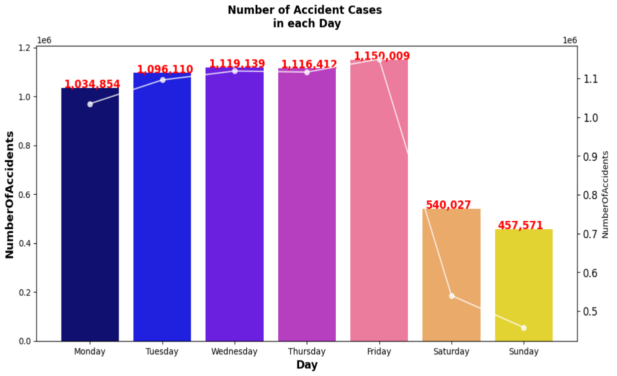

Figure 11: *The CountPlot Of Day Feature. Result: Most of the accidents happened between Wednesday and Friday.*

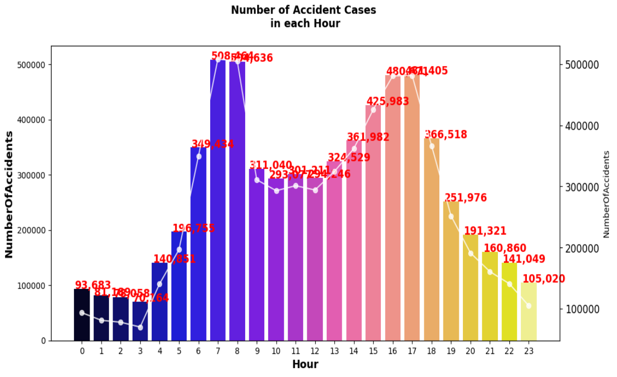

Figure 12: *The CountPlot Of Hour Feature. Result: Most of the accidents happened between 7 and 8 in the morning (starting time for office work) and also between 16 and 17 in the evening (ending time for office work).*

Results

The main goal of this project is to learn how to use statistical tests and exploratory data analysis to analyze the problem as much as possible. Now, the dataset selected for this work is US-Accidents(2016-2023) dataset, which provides us with information about the accidents that have occurred.

But finally, by using these techniques, we can partially understand the relationships and behavior between features so that we can work as an analyst in this field. To achieve better results using this data, we can go one step further and use machine learning models for creating predictive models.