ANALYSIS OF US-ACCIDENTS DATASET

Reza Mosavi ,400222100

Abstract

This data science exercise centers around a comprehensive countrywide car accident dataset spanning 49 states of the USA. Collected from February 2016 to March 2023 through multiple APIs, this dataset comprises approximately 7.7 million accident records. These records originate from diverse sources, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and road network traffic sensors. The dataset's versatility and extensive temporal coverage make it invaluable for various applications such as real-time accident prediction, hotspot location analysis, casualty assessment, causal inference for accident prediction, and the study of environmental influences like precipitation on accident occurrence. Moreover, the dataset's temporal span allows for the investigation of the impact of external factors such as COVID-19 on traffic behavior and accident rates. It is crucial to note that while this dataset may have occasional missing data due to network issues during collection, it remains a valuable resource for research purposes under a Creative Commons license, and proper citation is required when using it for non-commercial, research, or academic applications.

Introduction

In our endeavor to work with this extensive dataset, we are presented with nearly 7 million records encompassing 46 distinct features. A key aspect of our analysis is grappling with the challenge of missing data and determining appropriate strategies for their management.

The dataset, notably, contains features with missing coordinate data, and addressing this issue warrants careful consideration. Imputing these missing values is a complex task, as a simplistic approach may lead to significant inaccuracies. Therefore, our focus is on conducting a thorough examination of this dataset to better comprehend the nuances of the missing data and develop effective solutions.

We recognize the delicate nature of handling missing coordinates. A careful approach is essential since these coordinates are integral for location-based analysis, including hotspot identification and geospatial insights. Blindly filling in the gaps could compromise the dataset's integrity and the validity of subsequent analyses.

Moving forward, we aim to manage these challenges adeptly. Our strategy is rooted in data quality preservation and meaningful analysis. We will explore diverse techniques to address missing data, all while ensuring our actions align with the dataset's context and our research objectives.

To summarize, our data science practice revolves around addressing the complexity of missing data, especially concerning coordinates. By approaching this task systematically, we strive to enhance the dataset's reliability and facilitate insightful analysis.

EXPLORATORY DATA ANALYSIS

Missing data

Working with a dataset that boasts nearly 7 million records, we face a pressing issue - a considerable amount of missing data, particularly in columns related to accident coordinates. Filling in these gaps accurately is imperative since simplistic approaches can lead to substantial inaccuracies. This is primarily due to the substantial data loss in these specific columns.

Our approach commences by addressing columns with fewer missing values, encompassing over 1 million records. In this initial phase of the review, the majority of missing values fall below 100,000. To tackle this challenge, we employ the following strategies:

Most missing values in these columns are well below the 100,000 mark. For categorical data, we use the mode as our imputation technique. This method replaces missing values with the most frequently occurring category, ensuring data consistency. On the other hand, for numerical data, we adopt mean imputation, filling missing values with the average. This approach offers a balanced representation of the data.

By taking these measures, we ensure systematic handling of the data, considering both categorical and numerical data types. Our objective is to initiate a robust analysis while minimizing the risk of introducing erroneous information.

In conclusion, our data science practice is grounded in a methodical approach to address the challenges of missing data. We prioritize columns with fewer missing values and apply customized imputation techniques, tailored to the nature of the data. This sets the stage for a comprehensive dataset examination and meaningful data analysis.

During our data science exercise, we identified columns with missing values exceeding 1 million entries. To address this issue, we made a deliberate decision to remove these columns. This choice was motivated by the need to maintain data integrity and ensure the accuracy of our analysis, given the significant volume of missing data in these columns. By removing them, we streamlined the dataset, creating a more reliable foundation for our subsequent data analysis. This approach aligns with our commitment to data quality and the exercise's objectives. 1-End-Lat
2-End-Lng
3-Wind-Chill(F)
4-Precipitation(in)

Numerical columns

In this section, we shift our focus to the examination of numerical columns, despite the predominantly categorical nature of the dataset. Specifically, we will investigate a crucial feature: the start time of accidents. Our primary

objectives in this analysis are to determine when accidents are most likely to occur and the distribution of accident times.1

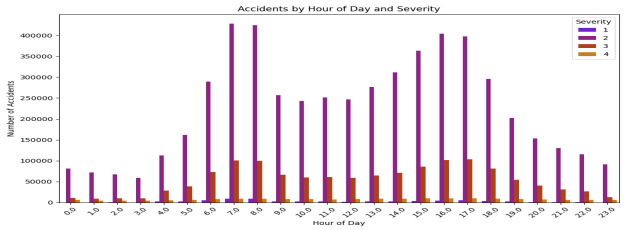

Figure 1: Accident severity based on time of day

As a general trend, we often find that the majority of accidents fall within the category of normal injury level. This pattern is typically observed in accidents that occur during the early morning as people begin their workday and, conversely, in the late afternoon and evening when work and office hours conclude.

Furthermore, our analysis indicates that the severity of accidents tends to be consistent across different times of the day. This consistency suggests that the level of injury in accidents is not significantly influenced by the time of occurrence.These observations contribute to our understanding of accident patterns, highlighting the importance of temporal factors, such as the start and end of the workday, while also emphasizing the relatively stable nature of accident severity throughout the day.

In this section, our attention turns to the examination of accidents based on the day of the week. Our objective is to uncover the temporal patterns and discern which days experience the highest frequency of accidents.By exploring this aspect, we aim to identify the days when accidents are most prevalent, shedding light on potential trends or factors that influence accident occurrence throughout the week.2

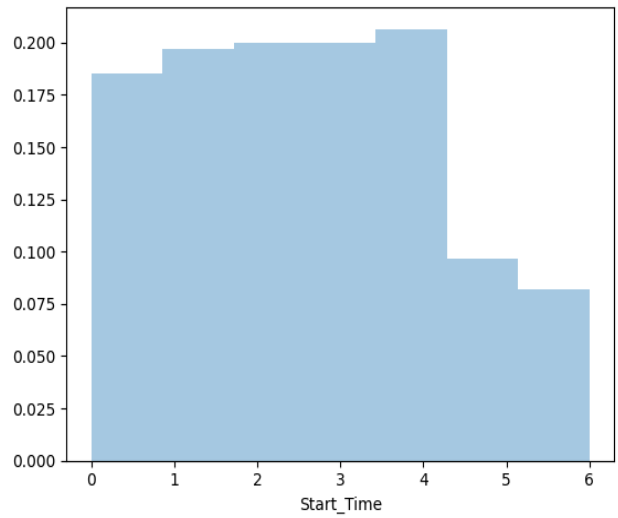

Figure 2: Distribution of accidents based on days of the week

Upon examining accidents based on the day of the week and time, a clear trend emerges. The data shows that the majority of accidents occur during working days of the week, with a notably smaller portion happening on holidays.Moreover, our analysis underscores the significance of specific time intervals, particularly the commencement and conclusion of typical working hours, in influencing accident occurrence. Accidents are more frequent during these periods, indicating the impact of busy working hours on weekdays.In essence, our findings suggest that accident patterns are closely linked to both the day of the week and specific times of the day. This insight highlights the importance of considering working hours and weekdays when assessing accident volume and timing.2

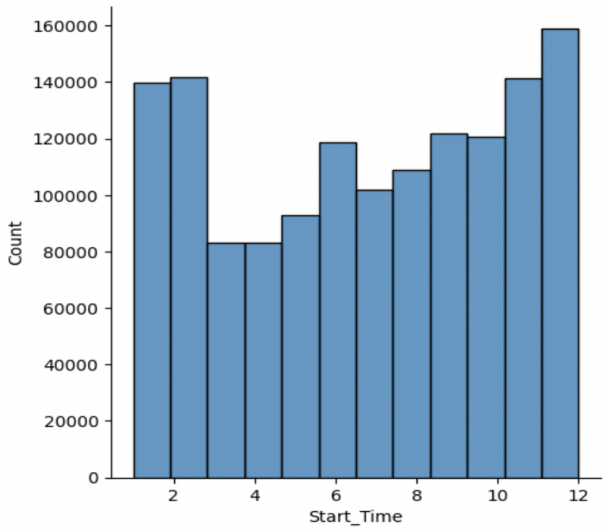

Figure 3: The volume of accidents based on different months of the year

Our examination of accident frequency by month reveals a distinct pattern. During the initial two months of the year, there is a notable increase in accidents. Subsequently, from the third month onward, the number of accidents begins to decline, reaching its lowest point. This trend persists until the end of the year, where the number of accidents gradually increases once more.In the latter months of the year, we observe a resurgence in accident numbers, ultimately surpassing the levels recorded during the initial two months. This cyclical pattern underscores the seasonality of accidents, with the year's start and end marking the highest accident activity.Understanding these seasonal trends is critical for informed decision-making and resource allocation, as it allows for better preparation during peak accident periods and underscores the need for enhanced safety measures during these times.3

In this segment, we shift our focus towards analyzing the volume of accidents spanning the period from 2017 to 2022. Our primary goal is to gain insights into how accident frequency has evolved over these years.

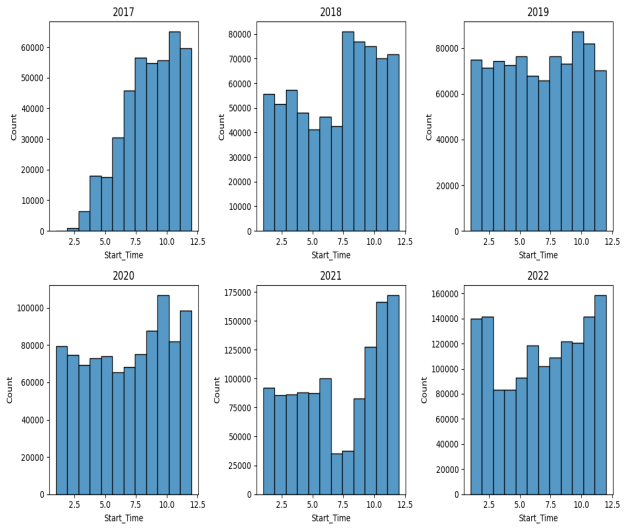

Figure 4: The volume of accidents based on different years

Our examination of accident volume from 2016 to 2022 reveals a significant insight: each year exhibits its unique pattern. While we may have observed specific patterns in the distribution of accidents by month in 2022, it's crucial to recognize that these patterns don't hold true for the previous years.In essence, our analysis suggests that accident trends can vary significantly from year to year. Understanding these yearly variations is vital for accurate risk assessment and the development of effective safety measures, as it underscores the dynamic nature of accident occurrence.4

Now we will examine some categorical and numerical features in this plot:

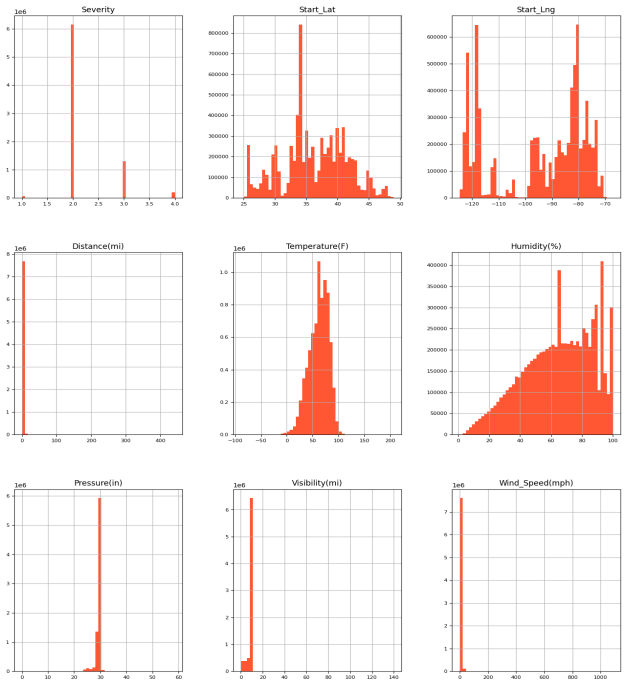

Figure 5: Histogram diagram of some numerical and categorical features.

The prominent observation from the graph is the prevalence of accidents with a severity level of 2. These accidents typically fall in the middle range, neither too minor nor extremely severe. Additionally, the conditions, such as air pressure and temperature, exhibit remarkable consistency in their distributions across most of the samples.While normalization could potentially yield more precise insights, it's apparent that a substantial portion of accidents occurred under similar weather conditions, including wind patterns and other atmospheric factors. This consistency suggests that a majority of accidents were influenced by comparable environmental conditions, making it a significant factor in accident occurrences.Understanding these commonalities in accident conditions is valuable for identifying risk factors and implementing preventative measures in locations or during times when such conditions prevail.5

Numerical columns

Now we are going to check the categorical columns, our main focus is checking them based on feature Severity.

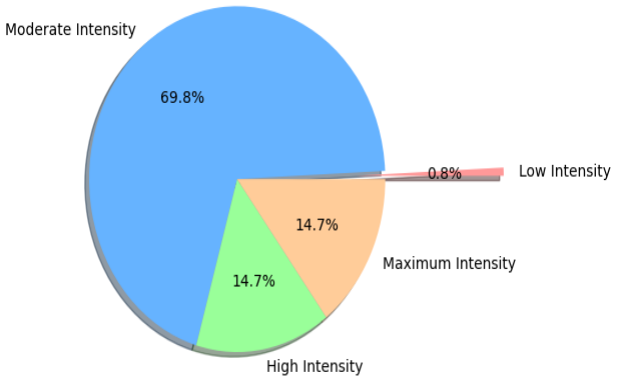

Figure 6: Severity of different accidents

Our analysis underscores that the majority of accidents fall within the categories of moderate and, to a somewhat lesser extent, low severity. Notably, approximately 28 percent of accidents are classified as high severity.In the next phase, we will delve into the geographic distribution of these high severity accidents across the United States. By examining their spatial patterns, we aim to gain insights into the regions or areas where high-severity accidents are more prevalent. This geographic perspective will provide valuable context for safety and infrastructure planning and implementation.(figure6)

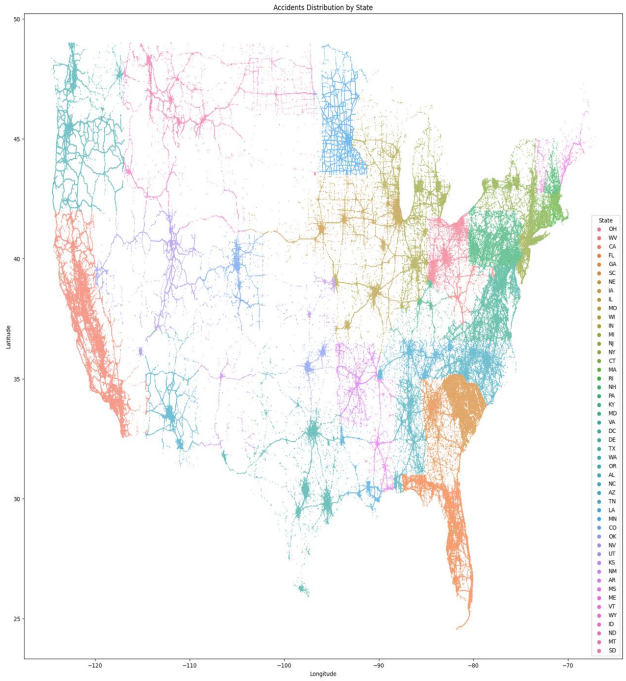

Figure 7: Amount of accidents by different states

The graphical representation of accident statistics distinctly highlights that the majority of accidents are concentrated in the eastern and western regions of the United States. Conversely, the central region exhibits a notably lower occurrence of accidents. This pattern is strongly associated with population density, as it aligns with the presence of more people in these eastern and western areas.Building upon our earlier analyses, we can infer that the higher incidence of accidents in these regions may be attributed to the impact of working hours. The greater population in these areas contributes to increased traffic congestion and, consequently, a higher likelihood of accidents. These insights underscore the complex interplay of demographics, geography, and temporal factors in influencing accident patterns. Understanding these dynamics is vital for targeted safety and traffic management strategies.(figure7)

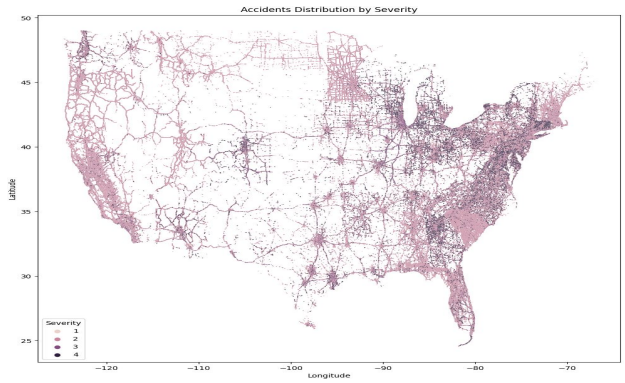

Figure 8: Severity and volume of accidents by state

Our analysis reveals an interesting trend in the severity of accidents concerning regional distribution. While the majority of accidents are concentrated in the eastern and western parts of the United States, the severity of accidents is notably higher in the eastern region. Following the east, the western region experiences a relatively higher incidence of severe accidents.This finding underscores the regional variation in accident severity, with the eastern region standing out as an area where accidents are more likely to result in higher severity levels. Understanding these regional variations in accident severity is crucial for tailoring safety measures and response protocols to address the unique challenges presented by different parts of the country.(figure8)

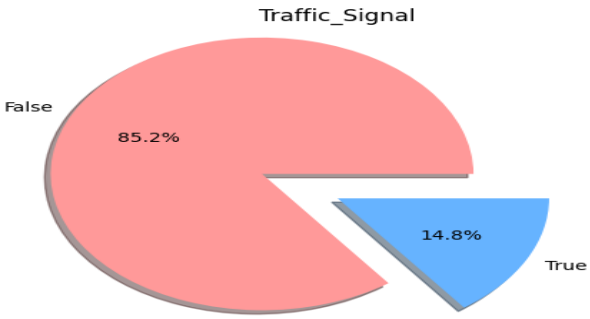

Figure 9: The effect of traffic signs on accidents

In our final analysis, we investigate the influence of road signs on accident occurrence. The data clearly indicates that a significant proportion of accidents have taken place in areas that lack these types of signs.This observation suggests a potential correlation between the absence of road signs and accident frequency. It raises important questions about the role of road signage in accident prevention and safety. Understanding this relationship can inform decisions regarding the installation and maintenance of road signs to enhance road safety and reduce the risk of accidents.(figure9)

Hypothesis Test

The confidence interval for all tests is equal to 95

Is the severity of accidents the same at different times of the day?

ANOVA Test

Null Hypothesis (H0): The severity of accidents is the same at different times of the day. In other words, there is no statistically significant difference in accident severity across different time periods.

Alternative Hypothesis (H1): The severity of accidents is not the same at different times of the day. This means that there is a statistically significant difference in accident severity across at least some of the time periods.

statistic : 48.19759522332919

P-value : 3.867073532151914e-31

There is a statistically significant difference in the severity

of accidents across different times of day.

is that the severity of the accident has no relationship with the presence or absence of a stop

Chi2-Distribution

Null Hypothesis (H0): There is no relationship between the severity of accidents and the presence or absence of a stop sign. In other words, the two variables are independent.

Alternative Hypothesis (H1): There is a relationship between the severity of accidents and the presence or absence of a stop sign. The two variables are dependent or related.

Chi-squared statistic: 50.38897580812082
P-value: 1.26100608870048e-12
There is a statistically significant relationship between the presence or absence of a stop sign and accident severity.

Is the average value in the entire community for A equal to 0.58? (0.02 more than the value we have in the data)

T-Distribution

$$H_0 : \mu_{Distance(mi)} = 0.58$$
$$H_1 : \mu_{Distance(mi)} \neq 0.58$$

statistic : -28.409574521283574
p_value : $1.572976008944932e - 177$
$Rejectthenullhypothesis$:
$Theaveragedistanceisnotequalto$0.58.

Does being day and night have an effect on the severity of the accident?

Chi2-Distribution

Null Hypothesis (H0): There is no relationship between the day or night and accident severity. In other words, the two variables are independent.

Alternative Hypothesis (H1): There is a relationship between the day or night and accident severity. The two variables are dependent or related.

Chi-squared statistic: 50.38897580812082
P-value: 1.26100608870048e-12
There is a statistically significant relationship between the presence or absence of a Sunrise-Sunset and accident severity.

Does timezone affect the severity of the accident?

Chi2-Distribution

Null Hypothesis (H0): There is no association between the timezone and the severity of accidents.

Alternative Hypothesis (H1): There is an association between the timezone and the severity of accidents.

Chi-squared statistic: 42.5060812082
P-value: 2.24109909970043e-13
There is a statistically significant relationship between the presence or absence of a Timezone and accident severity.

Reference

1 : scikit-learn.org
2 : www.kaggle.com