A Comprehensive Statistical Analysis on Car Crashes in the United States 2021

Anuj Hundia, Pranit Katwe, Shruti Wakchoure

December 19, 2023

Abstract

An overview of motor vehicle traffic crashes in 2021 is given in this report. In the United States, motor vehicle traffic crashes claimed the lives of 42,939 persons in 2021. This represents a 10-percent increase from the 39,007 fatalities in 2020 or 3,932 more lives lost in traffic crashes. From 5.25 million reported traffic crashes by the police in 2020 to 6.10 million in 2021—a statistically significant 16 percent increase—the estimated number of crashes rose. The paper delves into the significant role of diverse road elements, including vehicle involvement, type of area, type of crashes, and monthly data in shaping the occurrence and severity of car crashes. By employing rigorous statistical methods, and hypothesis testing, the research quantitatively measures the impact of specific conditions on accidents, isolating these effects from other contributing factors.

Contents

1 Introduction

Vehicle accidents remain a major problem in the US, harming people's lives, property, and general well-being. Amidst these worries, it is imperative to comprehend the complex interactions that exist between weather and automobile accidents. 42,939 people lost their lives in motor vehicle traffic crashes in 2021, an alarming 10% more than the year before. This increase in fatalities, together with a significant increase in reported crash rates, calls for a rigorous analysis of the influence that various factors have on the frequency and intensity of these incidents.

The impact of accident conditions nationwide in 2021 is thoroughly examined in this report. It explores different conditions, influences of seasons, types of crashes, vehicles involved in the crash, type of crash, etc. Through the use of strong statistical techniques, rigorous hypothesis testing, and a careful examination of data from 2021, this study seeks to quantify the unique impact of various conditions on accidents while separating these impacts from the numerous other elements that contribute to these incidences.

To better understand the relationship between other elements and auto accidents, this report focuses on three main research questions. It examines the constant nature of temperature as a factor in crashes in various places, the variable influence of dew point on crash frequency at various times, and the relationship between seasonal variations and fatalities. With motor vehicle crash fatalities expected to rise by a startling 10% to 42,939 in 2021, this report aims to quantify and isolate the unique impact of different conditions on accidents while separating these impacts from other contributory factors.

The aim is to investigate the impact of seasonal variation on the total number of fatalities in car crashes across states. Additionally, we seek to identify specific weather patterns that significantly influence the proportion of severe crashes in different regions. By analyzing statistical data obtained from the National Highway Traffic Safety Administration, we endeavor to uncover the correlation between seasonal changes and the occurrence of fatal car crashes, shedding light on potential trends and variations across the fifty states. Furthermore, the report examines how diverse weather conditions, such as "Clear" and "Cloudy" conditions, play a role in the severity of crashes, particularly those involving six or more individuals.

2 Data

2.1 Source

This report provides data on all police-reported traffic crashes including fatalities and people injured from the 2021 Fatality Analysis Reporting System (FARS). FARS contains data on every fatal motor vehicle traffic crash within the 50 States, the District of Columbia, and Puerto Rico. To be included in FARS, a traffic crash must involve a motor vehicle traveling on a public traffic-way that results in the death of a vehicle occupant or a non-occupant within 30 days of the crash.

2.2 Data Description

This dataset appears to be comprehensive, encompassing various factors such as crash details, location, time, environmental conditions, and additional details related to emergency response. It would be valuable for conducting in-depth analyses and understanding the patterns and factors contributing to car crashes. Crash Details such as the number of vehicles, persons, and fatalities involved are present. With that position of the crash i.e. Longitude and Latitude are also there. Environmental Details such as Light Conditions and Atmospheric Conditions (Weather). Time Details such as Month, Day, and Hour of crash are also included.

2.3 Possible Bias

Geographical bias in car crash data arises when the dataset is not well-balanced across different geographical regions. This imbalance can lead to skewed conclusions about the relationship between seasonal conditions and crashes throughout the year. Certain areas may experience specific seasonal conditions more frequently, and if the dataset is disproportionately concentrated in those regions, it may not accurately represent the diversity of climates, road conditions, and driving habits. This lack of representation compromises the external validity of research findings, limiting the ability to make accurate and generalized conclusions about the impact of weather on car crashes.

3 Exploratory Data Analysis

This section will provide the data analysis based on the visualizations and tables. Each visualization is unique and provides interesting distributions between various aspects of the data.

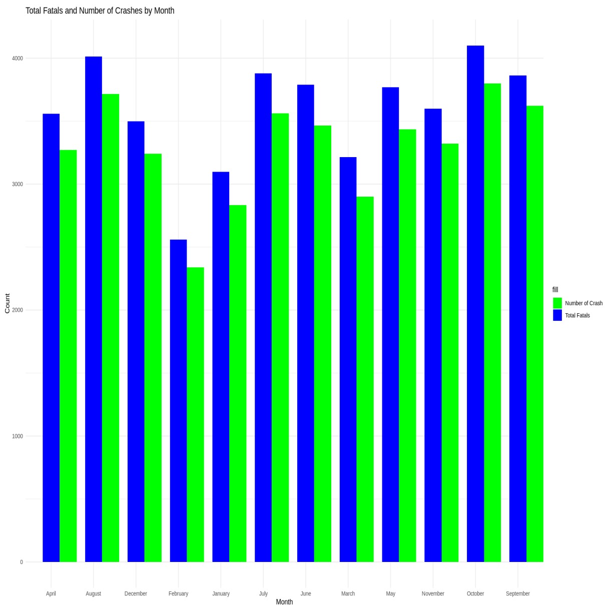

Figure 1: Total Fatalities and Number of Crashes by Month

Figure 1 shows the comparison between the Number of Fatalities per Crash and the Total Number of Crashes by each month in 2021. October and August have the highest Number of Crashes and the highest Number of Fatalities. While January and February have the lowest number of crashes and fatalities.

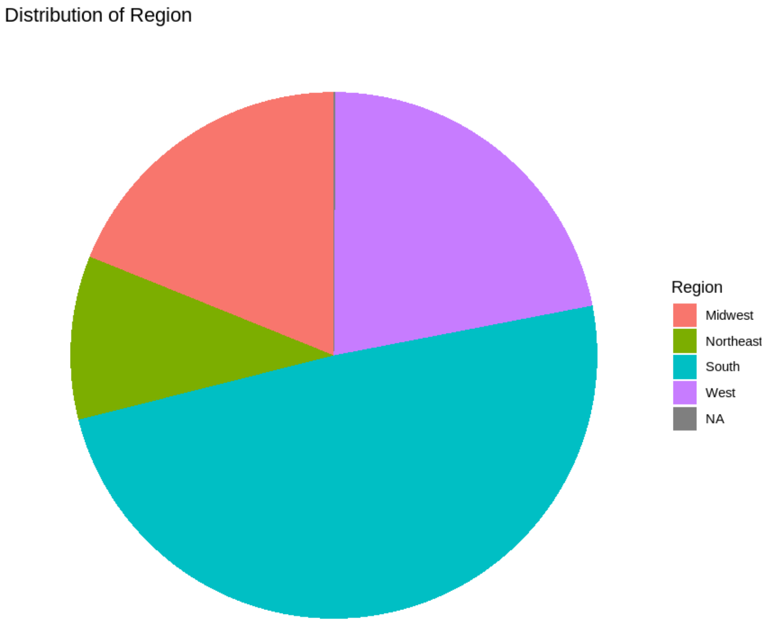

Figure 2: Pie Chart of Regions in the USA

Figure 2 shows the Distribution of the Total Number of Crashes by Region such as Midwest, Northeast, South, and West. South has the Highest Number of Crashes by Percentage.

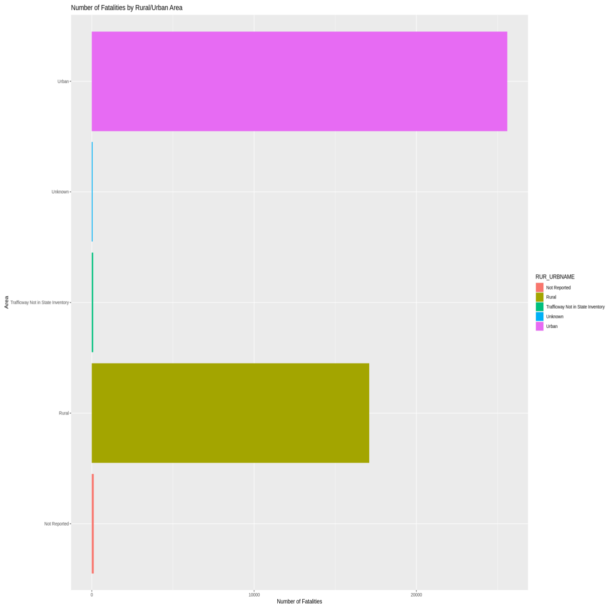

Figure 3: Urban vs Rural

Figure 3 provides a comparison between the Total number of crashes in Urban and Rural areas. There are some crashes recorded where Traffic Way was not in state inventory and some of the crashes were not being recorded.

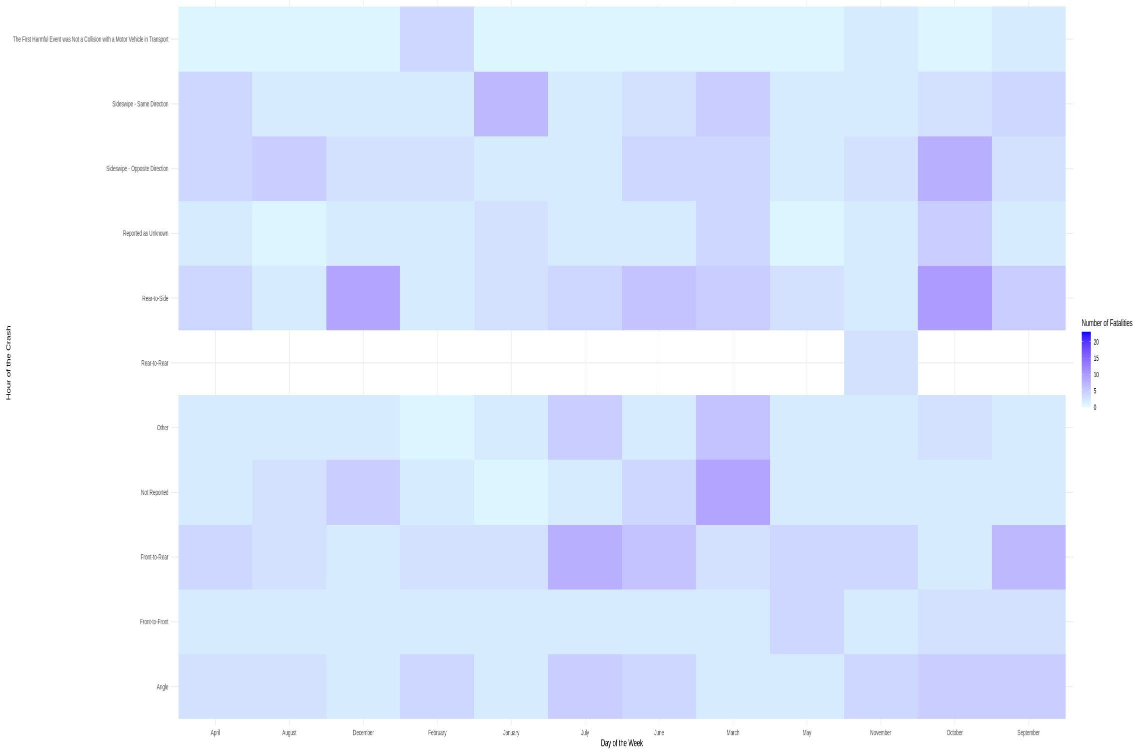

Figure 4: Heat-map of Type of Crashes According to each Month

This figure represents a heatmap that shows the relation of Persons involved in Crashes where the Type of crashes happened vs each month. Some of the highlighted portion was Rear to Side type during October and December. Similarly, during the month of July, the maximum type of crash was Front to Rear.

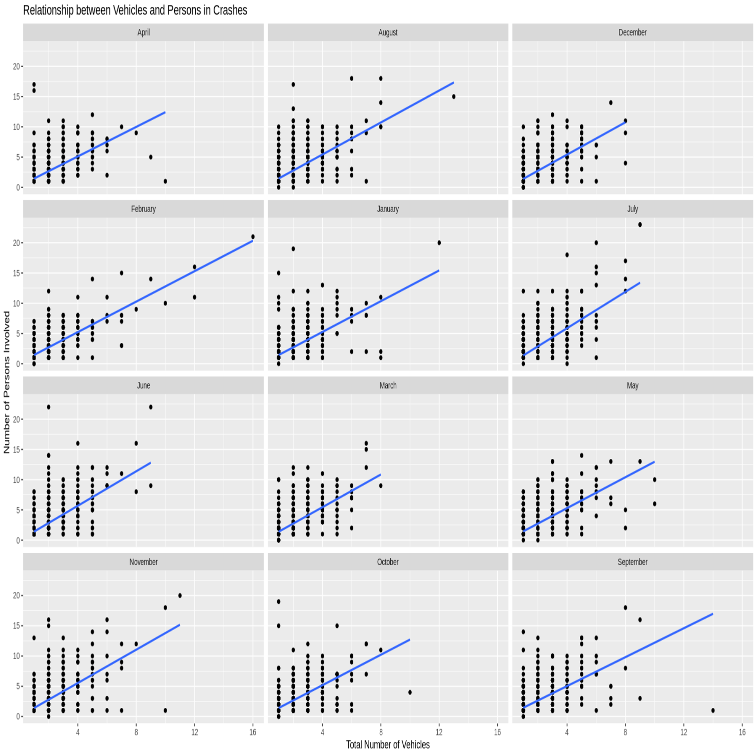

Figure 5: Scatter Plot

	mean persons	**mean vehicles**
Januray	2.17	1.56
February	2.26	1.65
March	2.19	1.59
April	2.18	1.60
May	2.26	1.62
June	2.29	1.61
July	2.25	1.59
August	2.26	1.60
September	2.18	1.60
October	2.11	1.56
November	2.28	1.61
December	2.17	1.58

Table 1: Summary Statistics for Total Vehicles and Persons Involved

There is a positive correlation between the number of vehicles involved in a car crash and the number of people involved. This means that as the number of vehicles involved in a crash increases, the number of people involved in the crash also tends to increase. This is likely because more vehicles mean more people are at risk of being injured in a crash.

There is a lot of variation in the number of people involved in crashes, even for crashes involving the same number of vehicles. This suggests that other factors contribute to the number of people involved in a crash, such as the type of vehicles involved, the speed of the vehicles, and the road conditions.

The data points are somewhat clustered around the lower left corner of the plot. This suggests that there are more crashes involving a small number of vehicles and people than there are crashes involving a large number of vehicles and people.

Overall, the scatterplot suggests that there is a positive correlation between the number of vehicles involved in a car crash and the number of people involved. However, there is also a lot of variation in the number of people involved in crashes, even for crashes involving the same number of vehicles. This suggests that other factors contribute to the number of people involved in a crash.

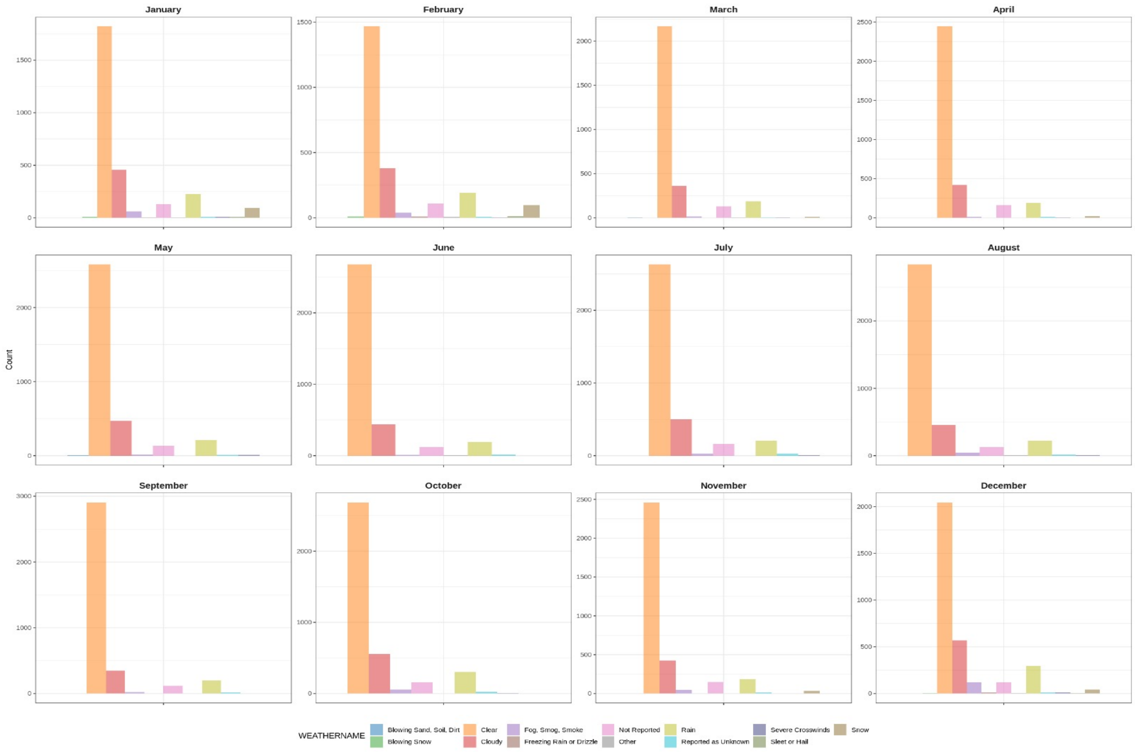

Figure 6: Count of Accidents Based on Weather Conditions

Month	Clear	Cloudy	Rain	Fog,Smog,Smoke
January	1820	456	225	61
February	1466	382	192	40
March	2165	361	186	18
April	2442	421	187	11
May	2582	473	208	14
June	2678	438	194	13
July	2631	500	205	28
August	2839	452	224	42
September	2908	347	200	24
October	2687	558	306	54
November	2462	424	186	45
December	2046	567	296	118

Table 2: Summary of the count of accidents based on the reported weather conditions

The count of Clear Weather is consistently higher Each Month. This might indicate that there is a higher probability of human error happening during the crash. With that most second consistent weather is Cloudy weather type.

4 Methods - Hypothesis Testing

Hypothesis testing is a statistical method used to determine if there is enough evidence in sample data to conclude population parameters, it is sometimes called significance testing. Hypothesis testing is used to assess the plausibility of a hypothesis by using a specific set of sample data. This involves assessing two types of hypotheses about the population parameters.

- Null Hypothesis (H_0): Hypothesis testing is performed to test the validity of a claim or assumption that is made about a population this is also known as a Null Hypothesis. It is denoted by H_0
- Alternative Hypothesis(H_1): The alternative hypothesis is considered valid if the null hypothesis is rejected. Primarily, the data and the statistical calculations that go along with it serve as the evidence in the sample. It is denoted

Several hypothesis tests can be performed on this data. As there is qualitative as well as quantitative data present. There are two two-tailed and one one-tailed tests performed based on the research questions this report is trying to answer.

- One-Tailed Z Test: A one-tailed z-test for two samples is a statistical test that compares the means of two independent samples to see if there is a significant difference in one direction or the other. The alternative hypothesis specifies the area of rejection in a one-tailed test, which is limited to one direction. This indicates that, depending on the direction indicated in the alternative hypothesis, the test is intended to demonstrate whether the sample mean is considerably higher or lower than the population mean.
- Welch two-sample t-test: The Welch two-sample t-test is a statistical test used to evaluate whether the means of two independent groups differ by a significant amount. This test is a variant of the conventional two-sample t-test, but it performs better when the variances and/or sample sizes of the two groups are different.

4.1 Hypothesis Test Results

4.1.1 Welch Two Sample Test for comparing summer and winter fatalities

- Null Hypothesis (H_0): The mean number of total fatalities in summer months is less than or equal to the mean number of total fatalities in winter months.
- Alternate Hypothesis (H_1): The mean number of total fatalities in summer months is greater than the mean number of total fatalities in winter months.

The p-value is 0.08301 for this test hence there is weak evidence against the null hypothesis. and we can say that the no of total fatalities in summer months is greater than the no of total fatalities in winter months.

4.1.2 Prop Test

- Null Hypothesis (H_0): There is no difference in the proportions of severe crashes (defined as involving 6 or more people) between "Clear" and "Cloudy" weather conditions.
- Alternate Hypothesis (H_1): There is a significant difference in the proportions of severe crashes between "Clear" and "Cloudy" weather conditions.

Result of test: P-value ¡ 0.05: Since the p-value (0.04066) is less than the common significance level of 0.05, you would reject the null hypothesis. There is enough evidence to conclude that there is a significant difference in the proportions of severe crashes between "Clear" and "Cloudy" weather conditions.

5 Conclusion

Seasonal patterns appear influential in car crashes across regions, as indicated by the hypothesis. Severe crashes show varying proportions under different weather conditions, suggesting condition-based differences. While weak evidence challenges the null hypothesis, more comprehensive studies in controlled environments are vital for stronger causal conclusions Based on the hypothesis performed seasonal patterns might play a role in car crashes across different regions. Considering severe crashes, the proportion of accidents under different conditions changes. The weak evidence against the null hypothesis in some cases indicates that further studies with more robust methodologies and controlled environments are necessary to establish a stronger causal relationship.

References

[1] Sheykhfard, A., Haghighi, F., Papadimitriou, E., Van Gelder, P. (2021). Review and assessment of different perspectives of vehicle-pedestrian conflicts and crashes: Passive and active analysis approaches. Journal of Traffic and Transportation Engineering (English Edition), 8(5), 681–702. https://doi.org/10.1016/j.jtte.2021.08.001

[2] Blum, J. J., Scullion, P., Morgan, R. M., Digges, K., Kan, C.-D., Park, S., Bae, H. (2008). Vehicle-Related Factors that Influence Injury Outcome in Head-On Collisions. Annals of Advances in Automotive Medicine / Annual Scientific Conference, 52, 131–140. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3256768/Vehicle-Related

[3] Thomas, P., Frampton, R. (1999). Large and Small Cars in Real-World Crashes -Patterns of Use, Collision Types and Injury Outcomes. Annual Proceedings / Association for the Advancement of Automotive Medicine, 43, 101–118. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3400214/

[4] Thomas, P., Frampton, R. (1999). Large and Small Cars in Real-World Crashes -Patterns of Use, Collision Types and Injury Outcomes. Annual Proceedings / Association for the Advancement of Automotive Medicine, 43, 101–118. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3400214/

[5] Reddy, S. S., Chao, Y. L., Kotikalapudi, L. P., Ceesay, E. (2022, June 1). Accident analysis and severity prediction of road accidents in United States using machine learning algorithms. IEEE Xplore. https://doi.org/10.1109/IEMTRONICS55184.2022.9795852