جامعة مصر للمعلوماتية
EGYPT UNIVERSITY
OF INFORMATICS

Egypt University of Informatics
Computer and Information Systems
Data Analysis Course

The Analysis of Vehicle Accidents

Submitted by: Radical Analysts

26/5/2024

Introduction:

In India, car crashes are a serious public health concern that result in large financial losses and the loss of many lives. The purpose of this project is to test two major hypotheses using crash data from automobiles: first, that wearing a seatbelt lessens the severity of crashes, emphasising the importance of seatbelts for occupant safety; and second, that alcohol-related crashes are more common on weekends than on weekdays, indicating the impact of social activities on traffic incidents. The ultimate goal of this study is to lower the frequency and severity of auto accidents in India by using thorough data analysis to unearth insights that can direct efficient road safety initiatives and policy decisions.

Research Question:

Question 1: Does seat belt usage reduce the severity of car crashes in India?

Question 2: Is alcohol involvement in car crashes higher on weekends compared to weekdays in India?

Hypothesis:

Hypothesis I:

Seatbelt usage reduces the severity of car crashes in India.

Hypothesis II:

Alcohol involvement in car crashes is higher on weekends compared to weekdays in India.

Population of Interest:

This study examines car crashes in India, focusing on drivers, passengers, and pedestrians, examining seat belt usage and alcohol involvement.

Data Set:

Vehicle Accident

Data Set description:

This dataset contains comprehensive information on vehicle accidents in India from 2009 to 2022. It includes data on various aspects of each accident, such as the location, date, time, vehicles involved, weather conditions, road conditions, and details of the casualties. The dataset spans a decade, providing valuable insights for analysis and research on road safety, accident causes, and prevention strategies.

Data Set Cleaning:

The data cleaning process for the vehicle accidents dataset involves several steps to ensure the data is clean, consistent, and ready for analysis.

I.Dropping Irrelevant Columns:

Several columns deemed unnecessary for the analysis are removed from the dataset. These columns include:

II.Handling Missing Values:

Categorical Columns: Missing values in categorical columns are filled with the mode (the most frequent value) of each respective column.

Numerical Columns: Missing values in numerical columns are filled with the mean of each respective column.

III.Dropping Remaining Missing Values:

Any remaining rows with missing values are dropped to ensure the dataset is complete and does not contain any null values.

These cleaning steps ensure that the dataset is free from irrelevant columns and missing values, making it ready for further analysis and modelling.

Hypothesis I :

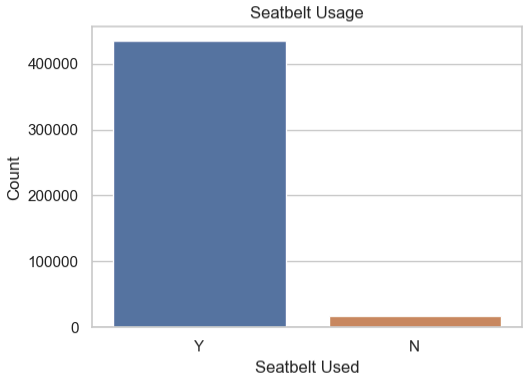
Seatbelt usage reduces the severity of car crashes in India.

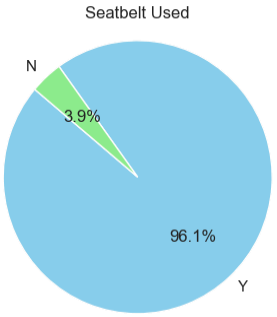
Description:

Seat belt usage is hypothesised to mitigate the severity of car crashes, potentially reducing the risk of injuries and fatalities among vehicle occupants in India.

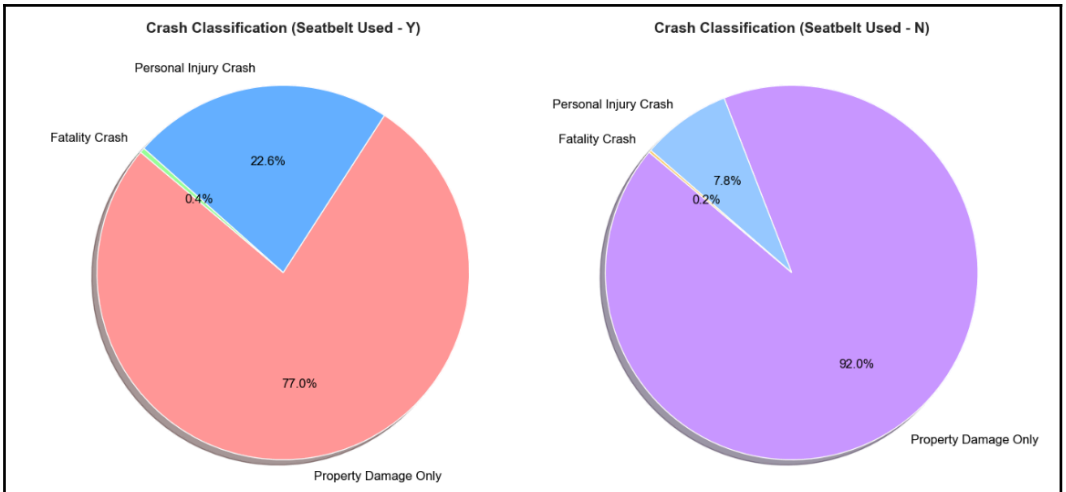
Analysis:

Pie charts were used to analyse the dataset in order to investigate the link between two categorical variables: "SEAT BELT USED", which indicates whether the driver wore a seatbelt, and "CRASH CLASSIFICATION DESCRIPTION", which describes the accident's damage. In cases where "SEATBELT USED" was specified as 'Y' or 'N', separate pie charts were created to show the corresponding proportions of crash types. We also used pie and bar charts to create a ratio that made it easier to compare various categories.





The aforementioned graphs highlight a noteworthy disparity in the percentage of drivers who wore seatbelts ('Y' value) compared to those who did not ('N' value), indicating a considerable variation in seatbelt usage among drivers involved in accidents. This significant disparity indicates a strong likelihood that the marked difference will produce different results.



These two graphs suggest that drivers who wore seatbelts were involved in more deadly collisions than those who did not. The large difference in seatbelt usage between the two groups, however, indicates that this proportional difference has a major impact on crash severity. Therefore, it makes sense to draw the conclusion that wearing a seatbelt lessens the severity of crashes, especially those that result in fatalities and bodily injuries, based on this striking disparity.

Hypothesis Testing Steps using Chi-squared test:

1. Null hypothesis : No significant association between seatbelt usage and crash severity
2. Conducting the p-value. which is equal to 7.67578479722609e-203.
3. Reject the null hypothesis: Seatbelt usage is associated with the severity of crashes.

```
Chi-squared: 930.7734066716706
P-value: 7.67578479722609e-203
Reject the null hypothesis: Seatbelt usage is associated with the severity of crashes.
```

Conclusion Hypothesis :

The data reveals a noteworthy variation in the use of seatbelts by drivers who are involved in collisions. Despite this variance, there were fewer fatal collisions involving drivers who were using seat belts. This emphasises how crucial seatbelt use is in lessening the severity of crashes, especially in terms of fatalities and injuries. The idea that wearing a seatbelt reduces the severity of crashes is supported by these results. Encouraging and enforcing the use of seatbelts is still essential for improving road safety and averting negative consequences from collisions

Hypothesis II :

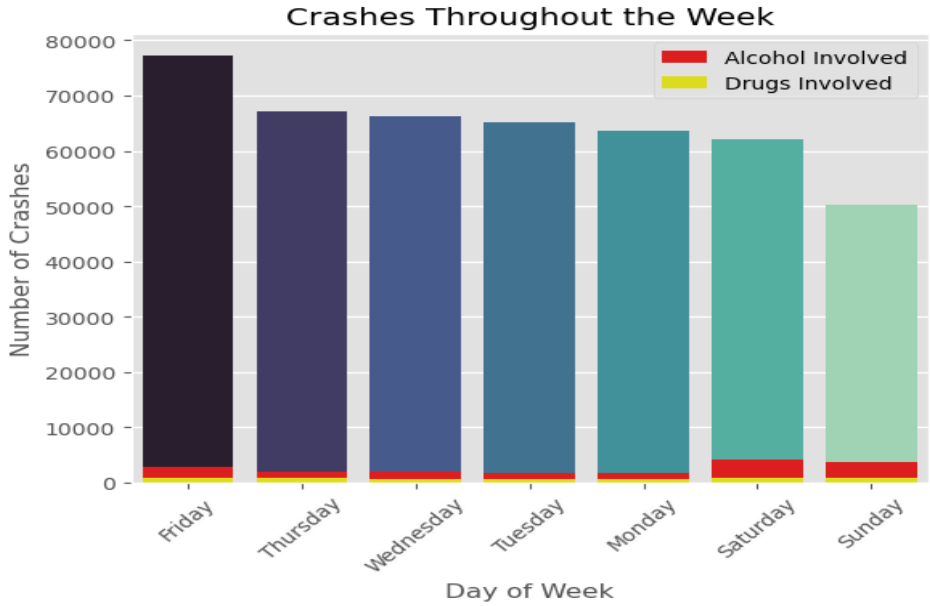
Alcohol involvement is higher in crashes occurring during weekends compared to weekdays.

Description:

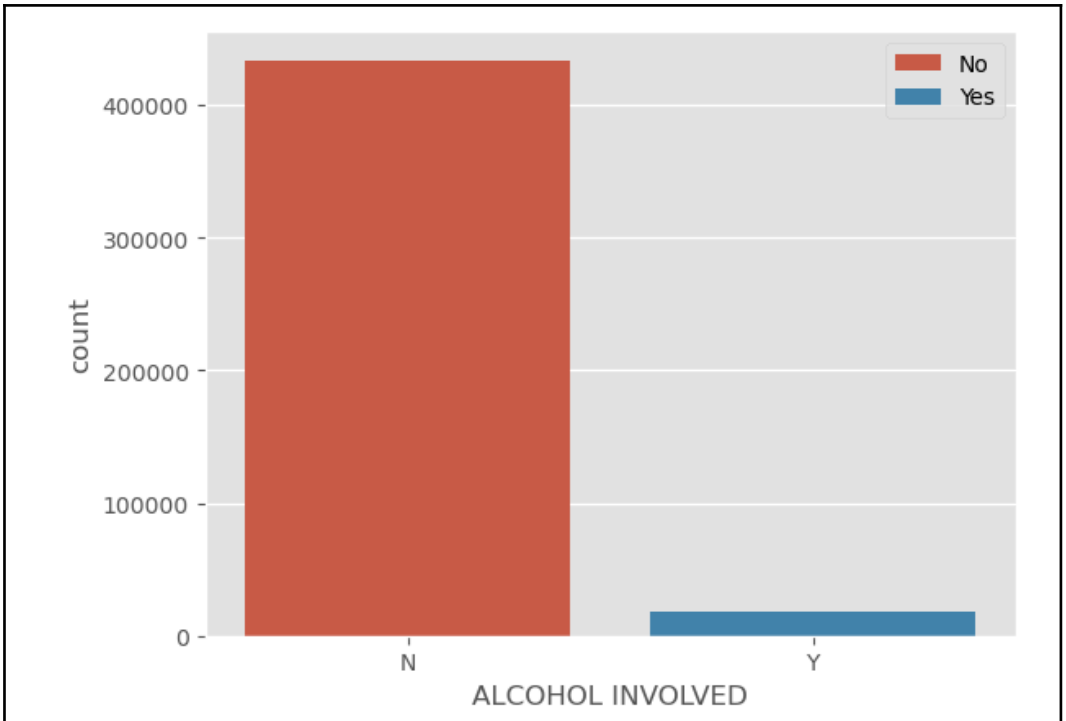
The hypothesis suggests that alcohol-related crashes are more frequent on weekends (Saturday or Sunday) compared to weekdays (Monday through Friday).

Analysis:

We used the "DAY OF WEEK DESCRIPTION" variable, which contains the days of the week each accident happened, combined with the "ALCOHOL INVOLVED" variable, which indicates whether or not alcohol was a factor in the accident (Y) in order to analyse the needed data about the involvement of alcohol in car accidents (N). To show how much alcohol is consumed on weekdays (Monday through Friday) as opposed to weekends (Saturday and Sunday), a stacked bar chart was used. This method makes it possible to thoroughly investigate alcohol-related occurrences on various days of the week, offering insightful information about alcohol consumption trends and how they affect traffic safety.



Analysis of the stacked graph indicates a clear trend: alcohol involvement in accidents is notably higher on weekends (Saturday and Sunday) compared to weekdays. This initial evidence supports our hypothesis regarding elevated alcohol involvement during weekend crashes. However, further investigation, including a p-value test and graphical analysis of 'Y' and 'N' values in the "ALCOHOL INVOLVED" variable, is needed to confirm the strength of this relationship.



A significant difference between the 'Y' and 'N' values in the 'ALCOHOL INVOLVED' variable is shown in the bar graph. Nevertheless, the final validation of our hypothesis depends on doing a p-value test to determine its acceptance or rejection.

Hypothesis Testing Steps using Chi-squared test and T-statistic test:

1. Null Hypothesis : No significant difference in alcohol involvement between weekends and weekdays.
2. Conducting the p-value :which is equal to zero.
3. We are unable to reject or accept the null hypothesis since the p-value is zero, indicating insufficient data for conclusive inference.

T-statistic test:

```
T-statistic: 59.853710781820986
P-value: 0.0
```

Chi-squared test:

```
Chi-squared: 3553.2624328894644
P-value: 0.0
```

Hypothesis Conclusion:

In conclusion, our research demonstrates a clear pattern of higher alcohol participation in accidents on weekends compared to weekdays, which validates our theory. Further validation via a p-value test and graphical analysis is necessary to confirm the strength of this link. The variable 'ALCOHOL INVOLVED' has a noteworthy variation; nonetheless, the p-value of 0 suggests inadequate data to draw a definitive conclusion. Consequently, further investigation is required to get a clear conclusion and locate further evidence, even though the preliminary data is congruent with our theory.

Conclusion:

In summary, our data analysis project explored two hypotheses related to car crashes in India. Firstly, our investigation into seatbelt usage revealed a significant difference in crash severity between drivers who used seatbelts and those who did not. The data strongly suggests that seatbelt usage reduces crash severity, particularly in terms of fatalities and personal injuries. Secondly, our examination of alcohol involvement in accidents showed a clear trend: higher alcohol involvement during weekends compared to weekdays. While initial evidence supports our hypotheses, further validation through p-value tests and graphical analysis is necessary for conclusive findings. Overall, this project underscores the critical importance of seatbelt usage and highlights the need for targeted interventions to reduce alcohol-related accidents, ultimately contributing to enhanced road safety measures in India.

Any potential issues :

In analysing accident data, challenges arose due to low percentages of individuals not wearing seatbelts or testing positive for alcohol, introducing bias. Despite non-zero p-values from statistical tests, such as t-tests and chi-square analyses, hypotheses couldn't be definitively accepted or rejected. For seatbelt usage, although the p-value wasn't zero, indicating potential significance, the small proportion of non-users required a shift to probability analysis. Similarly, in alcohol involvement, where positive tests were more frequent on weekends, the zero p-values suggested insufficient data to confirm or deny the hypothesis. These findings highlight the need to consider statistical significance alongside practical implications when dealing with skewed or limited datasets.