ROAD ACCIDENTS ANALYSIS OF INDIA

1. Akash Verma, 12208873, akashverma7703@gmail.com
2. Shivangi Agarwal, 12206738, shivangiagarwal0277@gmail.com

----------***----------

Abstract

Road traffic accidents create massive loss of life and affect to the economic development in India. The present study focuses on the trend of road accidents, potential causes, and severity by incorporating data from five years of both urban and rural regions. This approach involves descriptive analytics and various statistical methods ANOVA, regression analysis, correlation analysis, classification, and clustering to identify patterns and high-risk zones. Machine learning models also forecast accident severity to comprehend the reason of accidents. Driver behavior, quality of roads, and traffic contraventions are studied, as well as distribution-based models for probability estimates of accidents. The insights derived should lead to recommendations on the formulation of stricter traffic laws, improvement in infrastructure, and the development of various safety measures to ease or even eliminate accidents and save lives. The worth of this research reflects in combining technology, data analysis, and policy to address challenges in road safety in India.

Keywords- Road accidents, Mortality & Economy, Machine learning algorithms, Accident probabilities

1. Introduction

Road traffic accidents are the major public health and economic problem in India with 150,000+ deaths annually and many more injured. India still is one of the worst countries in world because even though the motor vehicle safety and road system has improved over time, but road accidents have become frequent with increasing with the same rate. Such things are caused by several reasons, ranging from careless driving, bad road conditions to improper traffic law implementations and low knowledge about the safety on roads.

The study provides an analysis of road accidents data over a period of 5 years in urban and rural areas across India, It identifies factors that cause crashes and explores how these affect severity using ANOVA, regression, classification and clustering methods.

The research is focus to provide data-driven recommendations on how to reduce road accidents by examining factors such as driver behaviour, quality of roads and vehicle types. These findings can inform policy-makers on how to develop better traffic laws, road safety and public awareness campaigns to prevent deaths. Deeply alarming road crash fatalities are projected to increase substantially if there is no intervention and this study hammers home that data and technology will be key in tackling one of India's biggest challenges.

Research Questions:

1) How Are Accidents During Different Weather influenced by the distribution of data?
2) What is the distribution of accidents by vehicle type, and which vehicle type has the highest number of accidents?
3) How does alcohol consumption influence the likelihood of road accidents involving different vehicle types?
4) How do accident trends change across different times of the day, and what role does driver behavior play during peak vs. off-peak hours?

This type of research is important because road accidents in India result in large loss of life and economy. Data analysis can tell us where the accident hotspots are and when high risk conditions for accidents will happen. The findings from the study will assist in shaping road safety policies and interventions to reduce accidental deaths. This study will help in the use of machine learning to make strategies by authorities so that accidents can be stopped, and it will save both lives and resources

2. Literature Review

In the past years, there have been many studies conducted to find out what is triggering road accidents in Indian roads. In this context, these studies work on evolving models for predicting accidents with significant factors like road and driver nature, vehicle fitness specifications etc., along with location wise accident patterns. In the next section, we review these studies to inform our analysis and provide us with a set of insights and a methodological guide in this research.

- **Paper 1: [Sanjay Kumar Singh, (2017) "Road Traffic Accidents in India: Issues and Challenges"]**
  Summary: The paper examines road crashes for India, which close corresponding to death but from a public health perspective also cause injury and suffering. Males 30-59 years are at highest risk of being killed in such crashes--much the same is known for death itself.
  Relevance: Relevant in the context of causes and demography which is the most important part to analyze the road accident gist trends.

- **Paper 2: [Akansha Jadhav, Shruti Jadhav, Kirti Suryanshi, (2020) "Road accident analysis and prediction of accident severity using machinery"]**
  Summary: The following paper is based on machine learning algorithms to predict road accident severity with using AdaBoost classifier and it has achieved 80% accuracy.
  Relevance: It consists of valuable functions to predict accident severity in our project.

- **Paper 3: [Mayank Mahla, (2020) "Road accidents and safety in India"]**
  Summary: The paper investigates rising road accidents in India, attributing driver faults and bad road conditions as prime culprits.
  Relevance: It gives perspectives into why accidents take place and what safety issues were hit which is very useful in dining up ourcesis to analyze.

- **Paper 4: [Atul Yadav, Shivam Singh Patel, (2012) "Analysis of road accident in India using machine learning and deep learning"]**
  Summary: Zero The paper discusses something about road accidents in India, but only basically deals with wrong doings by drivers and deficient physical road infrastructure.
  Relevance: A review the research paper discusses the number of reasons for road accidents in India focusing on high driver errors and substandard road infrastructure that are termed as major cause.

- **Paper 5: [Jinzi Zheng, Jun liu, Yanan Chai, (2023). "Research on Road Traffic Accident Scene Investigation Technology Based on 3D Real Scene Reconstruction]**
  Summary: This paper presents the use of UAV-based 3D scene reconstruction to assist road traffic accident investigations, offering greater accuracy and efficiency
  Relevance: For us with accident analysis it also provides some interesting technological insights on how to reconstruct accidents scenes.

3. Methodology

It describes systematic process for analyzing the road traffic accident in India. It serves as a guide for gathering, processing and analytical assessment of information to identify trends and reasons why accidents occur. The data we use in this study include 5 years of accident records in urban and rural areas. Using statistical techniques (ANOVA, regression, classification and clustering) to analyze high-risk areas and relevant factors such as road quality, driver behavior and types of vehicles.

3.1 Data collection

This study makes use of a five-year dataset on road accidents in India obtained from Kaggle. The data set covers accident severity, location, road conditions, driver details (ages, experience), weather conditions and other features. Attribute Information the CSV file contains the accident-related attributes such as Date_of_Accident, Time_of_Accident, Location, Vehicle_Type and Severity etc. It is an exhaustive dataset to gain insight on journey, causes and severity of road crashes in India.

```
> summary(Road_Accidents_data)
 Accident_ID        Date_of_Accident               Time_of_Accident     Location
 Length:1000        Min.   :2019-01-02 00:00:00.0   Length:1000        Length:1000
 Class :character   1st Qu.:2020-03-29 06:00:00.0   Class :character   Class :character
 Mode  :character   Median :2021-08-13 12:00:00.0   Mode  :character   Mode  :character
                    Mean   :2021-07-14 15:24:28.7
                    3rd Qu.:2022-10-25 18:00:00.0
                    Max.   :2023-12-31 00:00:00.0
  Road_Type          Weather_Condition  Vehicle_Type          Driver_Age    Driver_Gender
 Length:1000        Length:1000        Length:1000        Min.   :18.00   Length:1000
 Class :character   Class :character   Class :character   1st Qu.:30.75   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median :44.50   Mode  :character
                                                          Mean   :44.01
                                                          3rd Qu.:58.00
                                                          Max.   :70.00
   Severity           Casualties     Cause_of_Accident  Alcohol_Involved   Police_Action_Taken
 Length:1000        Min.   : 0.000   Length:1000        Length:1000        Length:1000
 Class :character   1st Qu.: 2.000   Class :character   Class :character   Class :character
 Mode  :character   Median : 5.000   Mode  :character   Mode  :character   Mode  :character
                    Mean   : 4.907
                    3rd Qu.: 8.000
                    Max.   :10.000
 Day_of_Week        Vehicle_Speed    Road_Condition     Cluster
 Length:1000        Min.   : 30.00   Length:1000        1:369
 Class :character   1st Qu.: 52.00   Class :character   2:274
 Mode  :character   Median : 68.00   Mode  :character   3:357
                    Mean   : 71.43
                    3rd Qu.: 86.00
                    Max.   :140.00
```

Figure 1 Summary

Figure 1 shows the summary of the accident dataset. It is providing an overview of data such as accident reason, Road type, Alcohol involved etc by calculating their mean mode median and Standard deviation.

3.2 Descriptive Analysis Techniques

Statistical measures like mean, mode, median and standard deviation shows the importance of these tools in terms of gaining insight from accident-related datasets. It helps us to analyze the pattern and trend in an efficient way. Works like average could help us to visualize the overall trends, the typical daily number of accidents or how many vehicles are involved in each accident as a mean, this might guide us for some load data on a daily basis. In contrast, the bar chart presents the mode of a variable which can possibly be the vehicle type that appears in accidents most. Perhaps it is found that motorcyclists, being the type of vehicle, most often involved in a crash, are at greater fault.
Since median, or middle value, is a balance point in the data it works great when you have a set of data containing outliers. Every average is affected by some extreme values, for example, high casualty counts Average gives a very imprecise number and may not represent what a typical accident looks like; the median can be used to estimate of a more average event. Finally, standard deviation is just a metric on how dispersed the values are around the average. Once again, a high standard deviation of vehicle speeds at crash locations would indicate considerable variability in driving behavior and possibly variable enforcement of speed limits, further suggesting situational risk factors. Collectively, these measures allow for a multi-faceted understanding of accident data and an enhancement of predictive analytics and action planning for safety.

```
> source("~/.active-rstudio-document")
[1] "Mean values:"
       Driver_Age        Casualties     Response_Time
           44.008             4.907            32.531
   Vehicle_Speed Number_of_Vehicles_Involved
           70.571             2.873
[1] "Median values:"
       Driver_Age        Casualties     Response_Time
             44.5               5.0              32.0
   Vehicle_Speed Number_of_Vehicles_Involved
            71.0               3.0
[1] "Standard Deviation values:"
       Driver_Age        Casualties     Response_Time
         15.735296          3.065901         16.379479
   Vehicle_Speed Number_of_Vehicles_Involved
         29.204312          1.397435
[1] "Mode values:"
       Driver_Age        Casualties     Response_Time
               22                 6                9
   Vehicle_Speed Number_of_Vehicles_Involved
               41                 2
```

Figure 2 Descriptive Statistics Summary

Figure 2 shows information about Descriptive Statistics Summary of the data set, in which we are calculating the mean, mode and median by which we can easily determine the Reason of the accident.

3.3 Probability distribution techniques

Poisson Distribution: This models the probability of observing a certain number of accidents during a fixed period (such as one month) using a Poisson distribution. It is perfect for generating events that casually arrive out with a constant average rate per unit time, in this case, road accidents.

a) Calculate the probability of exactly 5 accidents in a month.
b) What is the probability of having 2 or less accidents in a month?
c) Find the 90th. The 90th percentile count of injury: Which is the number of injuries that occur 90% of the time or less.
d) Performed a simulation for the next 12 months for the number of accidents?

```
> source("C:/Users/akash/OneDrive/AppData/Desktop/poisson.R")
1. Probability of exactly 5 accidents in a month: 0.000619186
2. Probability of having 2 or fewer accidents in a month: 9.045386e-06
3. 90th percentile number of accidents: 22
4. Simulated accidents for the next 12 months: 22 22 13 20 9 24 29 13 21 17 19 12
```

Figure 3. Poisson distribution

Figure 3. showing the probability of exactly 5 accidents is 0.062%, and the likelihood that 2 or fewer accidents is close to zero. The 90th percentile predicts up to 22 accidents in most months, and simulated monthly accidents range from 9 to 29. These analyses thus provide useful indications of what accident patterns may be predicted.

3.4 Statical and ML technique

1) Regression:
Regression analysis will help in finding insight and, finally, predictive understanding of the causality rate based on variables such as speed of vehicle and driver's age. It may indicate how various variables are related to accidents and serve for the projection of the number of casualties that could be expected under a given condition. The above analyzes have assisted me in making a data-driven decision regarding the implementation of countermeasures in order to improve road safety.

```
Call:
lm(formula = Casualties ~ Vehicle_Speed + Driver_Age, data = Road_Accidents_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3059 -2.7316  0.1015  2.5000  5.5932

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.803458   0.390007  12.316   <2e-16 ***
Vehicle_Speed 0.007587   0.003864   1.963   0.0499 *
Driver_Age   -0.009962   0.006158  -1.618   0.1061
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.059 on 997 degrees of freedom
Multiple R-squared:  0.006172,  Adjusted R-squared:  0.004179
F-statistic: 3.096 on 2 and 997 DF,  p-value: 0.04567
```

Figure 4 Linear Regression Analysis

Figure 4 shows impact of Vehicle_Speed and Driver_Age on Casualties respectively. Since the coefficients for intercept (4.803) and Vehicle_Speed (0.0499) are all statistically significant, this indicates a small positive relationship between speed and casualties.

2) Correlation:
Correlation is used to evaluate the strength and direction of relationships between numerical variables, such as Driver_Age, Vehicle_Speed, and Casualties. It identifies which variables are closely related, helping to answer questions like whether higher speeds result in more casualties. Additionally, it provides insights to prioritize intervention areas, such as focusing on speed monitoring in accident-prone regions.

	Driver_Age	Casualties	Vehicle_Speed
Driver_Age	1.00000000	-0.04826796	0.04606308
Casualties	-0.04826796	1.00000000	0.05969798
Vehicle_Speed	0.04606308	0.05969798	1.00000000

Figure 5. Correlation Matrix

Figure 5 shows Driver Age, Casualties and Vehicle Speed how are they related. Driver Age vs Casualties is a very low (negative) value correlation, which means it hardly affects each other. In the same manner, there is a small positive correlation between Driver Age and Vehicle Speed, as well as Vehicle Speed and Casualties

3) Annova:
ANOVA is used to compare the means of casualties across multiple categories, such as Road_Type and Severity, to assess if the differences are statistically significant. It helps determine whether specific road types or severity levels are associated with higher casualties, providing valuable insights for implementing targeted safety measures based on road characteristics or accident severity.

```
> summary(anova_model)
              Df Sum Sq Mean Sq F value Pr(>F)
Road_Type      2 179645   89823 199.910 <2e-16 ***
Severity       2   1479     740   1.646  0.193
Residuals    995 447069     449
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Figure 6 Annova summary model

Figure 6 shows that Road Type has a big impact on accidents and is very important, while Severity does not have much effect. Most of the other differences are due to random factors. In simple terms, the type of road plays a major role, but the severity of the accidents doesn't matter much in this analysis.

4) Clustering (K mean):
Clustering is used to group similar data points, like Driver Age and Vehicle Speed, into different groups or clusters. It helps us find patterns in the data. For

example, it can group young drivers who drive fast into one group and older drivers who drive slower into another. This makes it easier to understand the behavior of different types of drivers. By knowing this, we can create better road safety campaigns or rules that match the needs of each group.

4. Tools and software

- **R programming:** Data Analysis and Visualizations: R is extensively used for data analysis, statistical computing, as well as to generate visualizations. It has strong inbuilt tools for basic processing, cleaning and performing analysis on big datasets. Libraries- dplyr, ggplot2, readr,

- **RStudio:** An IDE (Integrated Development Environment) for R, which allows you to write code much more quickly and changes in an instant, run the code with very few clicks. These tools and libraries are designed to use which will help you in the project of your road accident analysis to carry out the workflow smoothly for data analysis, visualization, and modeling.

5. Result

The analysis of road accident data in India over a five-year period presented interesting insights into the nature, causes, and contributory factors for deaths and injuries. The findings showed that highways, being the fastest routes, have the highest rates of accident severity, while city streets exhibited lower fatality rates. Accidents tend to increase in severity due to adverse conditions like rain and fog, which reduce visibility and make roads slippery. Additionally, larger vehicles such as trucks and buses are linked to more severe injuries, whereas motorcycles show higher casualty rates. These multilayered results highlight several areas that could be targeted for interventions to improve road safety, particularly on highways and during adverse weather conditions.

Research Questions output –

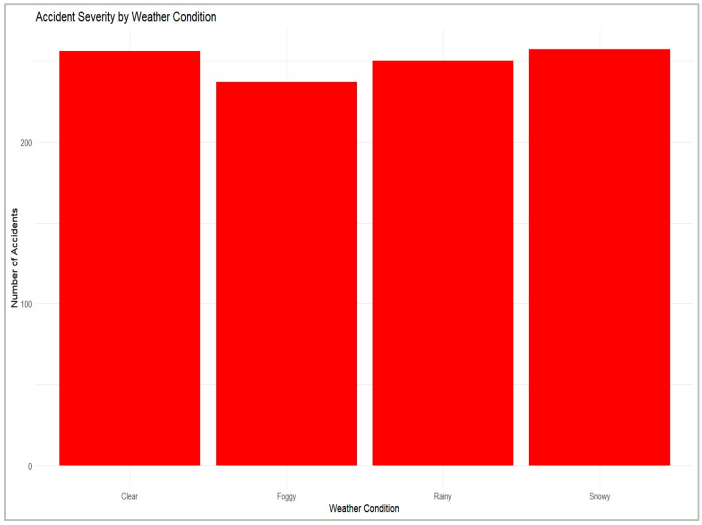

Figure 7. Accidents severity by Weather condition

Fig 7 shows the distribution of road accidents under four weather conditions across the year: Clear, Foggy, Rainy, and Snowy. The highest number of accidents in clear and snowy conditions was recorded as just over 200 each. The rest of the conditions like foggy and rainy had low numbers but were approximately the same as one another.

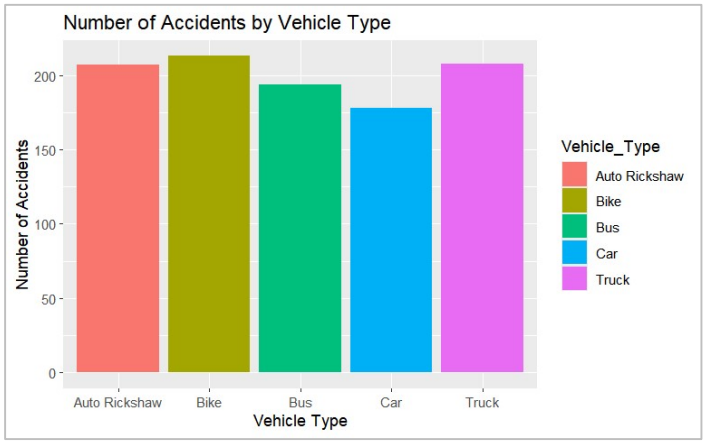

Figure 8 Road accident by vehicle type

Fig. 8 shows the number of accidents by vehicle type. The bar graph reflects a direct comparison, with bikes topping the list at 213 accidents, followed closely by auto rickshaws, while trucks and cars have the lowest at 178. The pie chart, however, illustrates each vehicle type's contribution to the total, showing that bikes, auto rickshaws, and trucks account for a larger portion of accidents, while cars and buses contribute less.

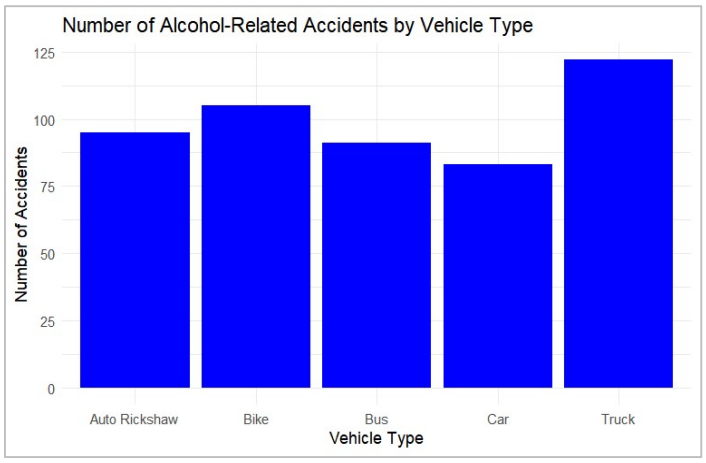

Figure 9 Effect of alcohol on accidents

Fig 9 depicts alcohol-related accidents in various types of vehicles. Trucks have the maximum number of incidents: bikes, auto rickshaws, and buses follow in that order. Cars report the minimum number of cases of alcohol-related accidents. The chart depicts larger vehicles and bikes on the road most often regarding alcohol-related accidents.

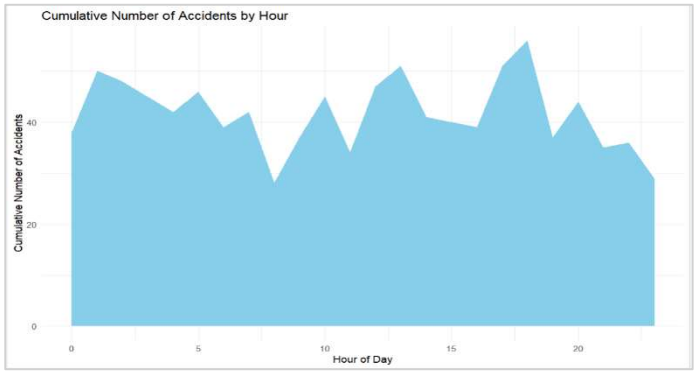

Figure 10 Number of accident Hourly

Figure 10 shows how accidents are distributed over a 24-hour period, with fluctuations throughout the day. Accident peaks occur around 14:00 and 18:00, likely due to higher traffic. The number of accidents starts high, remains steady in the early hours, then decreases, reaching the lowest point by the end of the day. This indicates that accidents are more frequent during midday and early evening hours.

Probability distribution output-

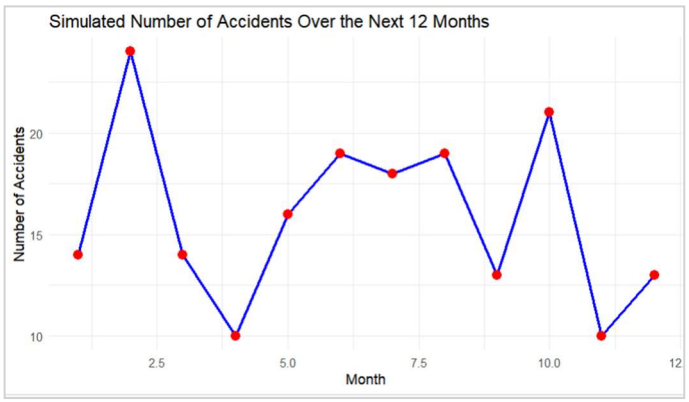

Figure 11 Number of accident Hourly

Fig 11 shows the total number of accidents over the next 12 months based on a Poisson distribution. The data shows fluctuations, with accident counts ranging between 10 and 29. Peaks are observed in months 2 and 10, while the lowest count occurs in month 11. This variability highlights the expected monthly variations in accident occurrences.

Statical and ML technique output-

1) Regression model

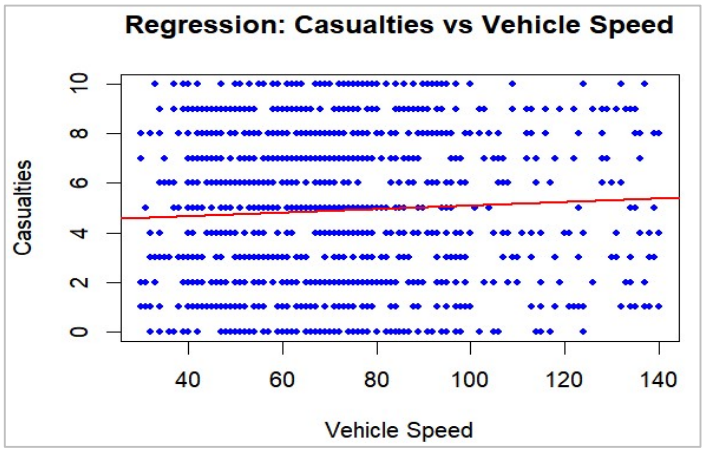

Figure 12. Regression Analysis of Casualties Based on Vehicle Speed and Driver Age

Figure 12 shows the relationship between vehicle speed and the number of casualties in accidents. The blue dots represent individual incidents, showing a wide range of casualties across different speeds. The red line represents the trend, indicating a slight increase in the number of casualties as vehicle speed increases.

2) Correlation Matrix

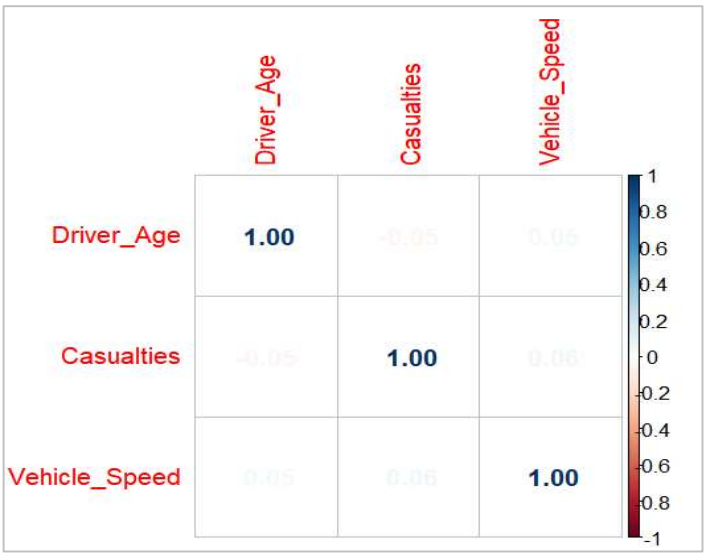

Figure 13. Correlation Matrix of Driver Age, Vehicle Speed, and Casualties

Figure 13 shows the weak relationships between driver age, casualties, and vehicle speed. Driver age has almost no impact on casualties or speed, and there is only a very slight link between speed and casualties. This suggests other factors are more important in influencing accidents.

3) Annova model

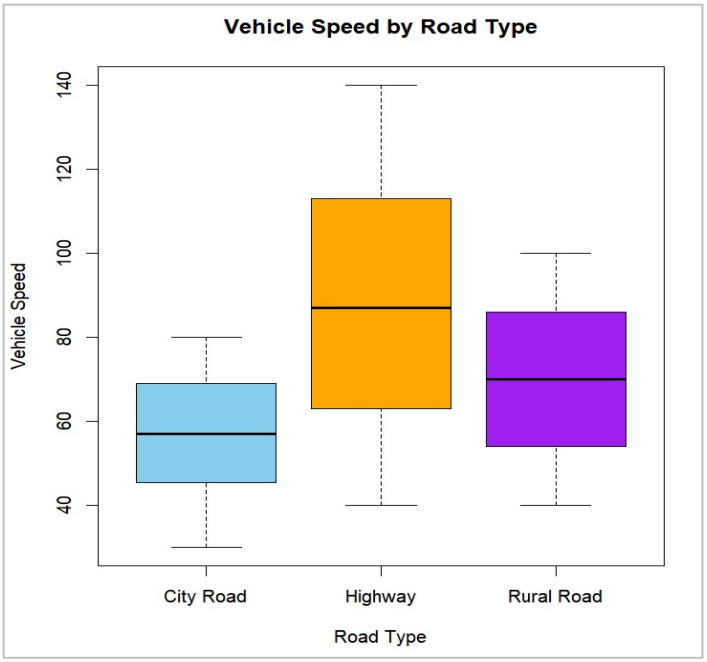

Figure 14(a). Vehicle Speed by Road Type

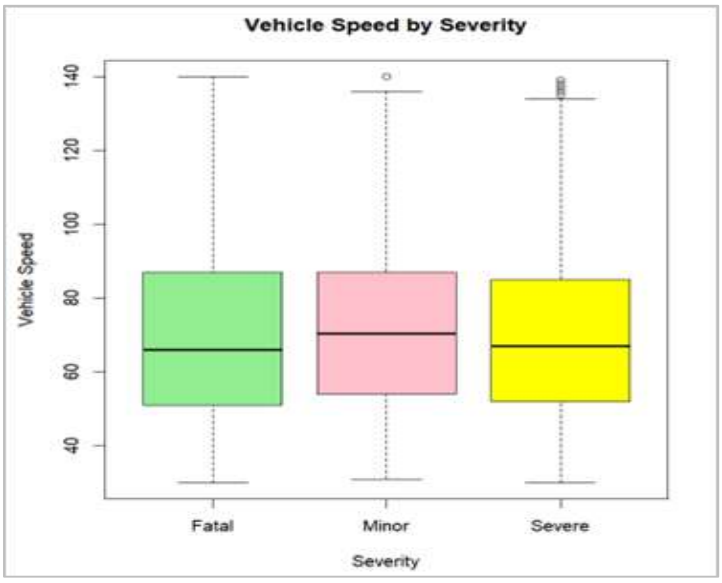

Figure 14(b). Vehicle speed by severity

Figure 14 shows the relationship between vehicle speed, road type, and accident severity. The first chart shows that vehicle speed varies based on road type: cars drive the slowest on city roads, the fastest on highways, and at moderate speeds on rural roads. This indicates that the type of road significantly influences how fast cars travel. The second chart examines vehicle speed across different accident severities- fatal, minor, and severe and reveals that speeds are quite similar across all categories, with no clear pattern. This suggests that while road type affects speed, the severity of an accident might not always depend on speed alone.

4) Clustering

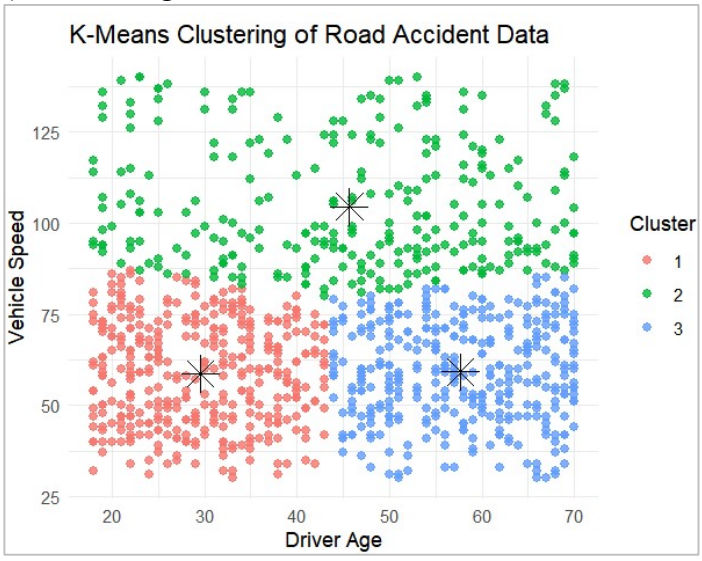

Figure 15 K-Means Clustering of Road Accident Data Based on Driver Age and Vehicle Speed

Figure 15 shows how drivers are grouped based on their age and speed. The red group has younger drivers who drive slower, the blue group has older drivers who also drive slower, and the green group includes drivers of all ages who drive faster. The black stars mark the center of each group, showing the average age and speed.

6. Discussion and Next steps

- **Road Type and Traffic Condition**: Highways often see more severe accidents due to higher speeds, while city roads, despite frequent accidents, are generally safer due to better surfacing. Rural and underdeveloped roads, particularly during construction or repairs, contribute significantly to accidents.
  Future research: Exploring advanced road technologies like intelligent traffic systems and automated road management could help dynamically control speed limits and provide real-time warnings about traffic, weather, or congestion, reducing accident severity.

- **Weather Effect:** Adverse weather like rain, snow, and fog increases accident severity by reducing visibility and delaying driver reactions. Over-speeding is more common in mild weather, while risky driving behaviors are prevalent in clear conditions.
  Future research: Developing connected vehicle systems and smart weather-responsive traffic management using machine learning can help predict accidents, recommend weather-based speed limits, and provide real-time alerts to reduce risks.

- **Driver Behavior**: Driver behaviors like drunk driving and speeding greatly impact accident severity, with older drivers (above 50) contributing significantly. ML-based telematics systems can monitor driver behavior, provide warnings, and identify patterns linking age, speed, and accident outcomes.
- **Future research:** Focus on telematics systems for real-time feedback, especially for elderly and high-risk groups. ML models could predict risky behaviors and suggest preventive actions to reduce accidents.

7. Conclusion

This study provides valuable insights into accident trends, contributing factors, and potential solutions for road safety in India. Statistical methods, such as clustering and regression analysis applied to the dataset, show that speed, weather, road types, and driver behavior are key factors in accident severity. Highways are associated with higher speeds, contributing to more severe collisions, with an average speed of 70.57 km/h during accidents. Adverse weather conditions like fog, snow, and rain further increase accident risks due to reduced visibility and slippery roads.

Driver demographics and behaviors are crucial, with the average driver age being 44 and alcohol contributing significantly to fatal crashes. Machine learning models could predict accident risk based on these variables, enabling interventions like real-time warnings or adaptive speed limits. Emergency response time (32.5 minutes on average) and the timing of accidents (more frequent during late hours) also contribute to fatality rates.

The evidence highlights the need for a multi-faceted approach involving stricter traffic laws, advanced road infrastructure, and road safety campaigns. Integrating statistical and ML techniques can enhance vehicle safety systems, intelligent traffic management, and telematics-based monitoring solutions. These technologies can adapt to road, speed, and weather conditions, helping policymakers address the economic and human cost of road crashes in India.

8. Reference

1) Sanjay Kumar Singh, (2017) “Road Traffic Accidents in India: Issues and Challenges,” Transportation Research Procedia.

2) Akansha Jadhav, Shruti Jadhav, Kirti Suryanshi, (2020) “Road accident analysis and prediction of accident severity using machinery”.

3) World Health Organization, (2018) “Global Status Report on Road Safety 2018”.

4) Mayank Mahla, (2020) “Road accidents and safety in India”.

5) Atul Yadav, Shivam Singh Patel, (2012) “Analysis of road accident in India using machine learning and deep learning”.

6) Jinzi Zheng, Jun liu, Yanan Chai, (2023) “Research on Road Traffic Accident Scene Investigation Technology Based on 3D Real Scene Reconstruction”.