"Decoding Chaos: A Comprehensive Analysis of Road Traffic Accidents"

- AMS 572 Data Analysis Final Project Report
- Group Members:

Clay Fan, Priyanshu Jain, Sai Akhil Kogilathota, Shubo Wang

1. Abstract

Vehicle accidents, a global concern, claim an estimated 1.2 million lives annually, with an additional 20 to 50 million individuals sustaining injuries, especially impacting developing and undeveloped nations. This escalating trend necessitates urgent intervention. Specialists, recognizing the need to address this issue, are turning to the analysis of historical road crash data within specific regions.

This research focuses on advancing our comprehension of accident data through data analysis techniques. Leveraging a carefully curated dataset from the Addis Ababa Sub-city police departments covering 2017 to 2020, the study comprises 12,316 instances of accidents, with 32 features detailing crucial aspects such as accident time, driver and vehicle specifics, and accident details (including causes and casualties). The dataset underwent meticulous organization, filtering, and preprocessing to construct a cohesive and uniform data model. Visualization techniques were then employed to present the data insightfully.

The study's methodology involved extensive exploratory data analysis, utilizing diverse software interfaces and visualization options. By uncovering patterns and insights from this comprehensive dataset, the research aims to provide valuable information informing strategies for improving the existing road network system. The ultimate goal is to mitigate the alarming prevalence of traffic accidents, contributing to global efforts to enhance road safety and save lives.

It was revealed that most of the accidents occur in the daytime and with those people who do not have enough traffic education. The 30 to 45 years age group was more active in causing the accidents. Therefore, the behavior of this age group of drivers needs further investigation. This study will be useful for concerned authorities in devising an efficient mechanism to alleviate road accident cases.

2. Introduction

In the realm of road safety, the analysis of traffic accident data stands as a critical endeavor, offering insights into the factors contributing to the severity of car accidents. This project delves into the nuanced interplay of variables such as drivers' age, driving experience, and additional factors to unravel their impact on accident severity—categorized as slight injury, serious injury, and fatal injury. With a focus on addressing stereotypical questions surrounding the correlation between shorter driving experiences and severe car accidents, the project employs R Studio for meticulous analysis, ensuring transparency through the inclusion of code within the report.
The dataset under examination is a product of collaboration with the **Addis Ababa** Sub-city police departments, collected for a master's research initiative. Spanning the years 2017-20, this dataset, meticulously prepared from manual records, encapsulates 32 features and 12,316 instances of road traffic accidents. Comprehensive in its scope, the dataset encompasses details on the time of the accident, driver information, vehicle details, and specifics about the accidents themselves.

Beyond the realm of data analysis, the significance of understanding and mitigating road traffic accidents is underscored by global statistics. Annually, approximately 1.35 million lives are tragically cut short, with an additional 20 to 50 million people enduring nonfatal injuries, some leading to permanent disabilities. The economic toll is profound, accounting for a 3% loss of the gross domestic product. This pervasive impact has elevated road traffic accidents to a subject of profound concern, prompting researchers to seek coherent methods for forecasting and preventing these incidents.

The core objective of this analysis is to identify the underlying factors influencing road traffic accidents, ultimately contributing to the enhancement of road safety. Acknowledging the gravity of this task, the choice of statistical methods becomes paramount. This report recognizes the delicate balance of selecting the right method, understanding its assumptions, and the potential pitfalls of choosing incorrectly. It emphasizes that the proper interpretation of results is crucial for obtaining meaningful insights into traffic problems, and a solid understanding of statistical tools is indispensable for creating quality research in road accident data analysis.
As we navigate through this report, we will explore various road accident data sources, delve into distinct analysis methods, and shed light on the intricacies of each. The ultimate goal is to provide not just an analysis but a comprehensive understanding that leads to actionable recommendations for managing factors contributing to road accidents. This paper serves not only as a snapshot of the current state of road accident data analysis but also as a guide for future research in the pursuit of safer roads.

3. Experimental Setup

In this section we discuss our analysis through which we identified valuable external datasets, with huge rows and columns and performed Exploratory Data Analysis on it to gain a

comprehensive understanding of the road accident dataset, it is crucial to first explore and describe the various columns and their corresponding features.

A. Exploratory Data Analysis

The dataset encompasses 32 columns, each providing valuable information on different aspects of road accidents. Below is a detailed description of the columns:

- Time: The timestamp of the accident occurrence.
- Day_of_week: The day of the week when the accident took place.
- Age_band_of_driver: The age range of the driver involved in the accident.
- Sex_of_driver: The gender of the driver.
- Educational_level: The highest level of education attained by the driver.
- Vehicle_driver_relation: The relationship of the driver to the vehicle (e.g., owner, employee).
- Driving_experience: The total years of driving experience.
- Type_of_vehicle: The category of the vehicle involved in the accident.
- Owner_of_vehicle: The owner of the vehicle.
- Service_year_of_vehicle: The number of years the vehicle has been in service.
- Defect_of_vehicle: Information about any defects in the vehicle.
- Area_accident_occurred: The general area where the accident occurred (e.g., residential, office, industrial).
- Lanes_or_Medians: Details about the lanes or medians in the accident location.
- Road_alignment: The alignment type of the road (e.g., tangent road, crossing).
- Types_of_Junction: The type of junction present in the accident location.
- Road_surface_type: The type of road surface (e.g., asphalt, earth).
- Road_surface_conditions: The condition of the road surface during the accident (e.g., dry, wet).
- Light_conditions: The lighting conditions at the time of the accident.
- Weather_conditions: The prevailing weather conditions during the accident.
- Type_of_collision: The nature of the collision (e.g., vehicle with vehicle, collision with roadside objects).
- Number_of_vehicles_involved: The total number of vehicles involved in the accident.
- Number_of_casualties: The total number of casualties in the accident.
- Vehicle_movement: The movement of the vehicle during the accident (e.g., going straight, moving backward).
- Casualty_class: The class of casualties (e.g., driver, pedestrian).
- Sex_of_casualty: The gender of the casualties.
- Age_band_of_casualty: The age range of the casualties.
- Casualty_severity: The severity of casualties (e.g., slight injury, serious injury, fatal injury).
- Work_of_casualty: The occupation or work-related status of the casualties.

- Fitness_of_casualty: The fitness status of the casualties.
- Pedestrian_movement: Details about pedestrian movement in the accident.
- Cause_of_accident: The primary cause of the accident.
- Accident_severity: The overall severity of the accident.

Accident_severity: 0 null values.The count of null values is specific to each column, and addressing these null values may involve data cleaning or imputation depending on the analysis goals and the nature of the missing data.

The dataset comprises valuable information, including details about the drivers, vehicles, road conditions, and casualties, providing a comprehensive foundation for the ensuing exploratory data analysis. The subsequent sections will delve into specific factors such as human factors, distracted driving, vehicle design, weather, traffic conditions, and road geometry, shedding light on their impact on road accidents. The exploration aims to unravel patterns, correlations, and insights that contribute to a more profound understanding of road safety dynamics.This analytical journey promises to unveil valuable insights, offering a deeper understanding of road traffic accidents and illuminating the contributing factors that influence their occurrence and severity.

B. Processing the data and Handling Total Null Values

In the provided dataset, the count of null values in each column is as follows:

Time: 0 null values, Day_of_week: 0 null values, Age_band_of_driver: 0 null values, Sex_of_driver: 741 null values, Educational_level: 579 null values, Vehicle_driver_relation: 829 null values, Driving_experience: 950 null values, Type_of_vehicle: 482 null values, Owner_of_vehicle: 3928 null values, Service_year_of_vehicle: 4427 null values, Defect_of_vehicle: 239 null values, Area_accident_occurred: 385 null values, Lanes_or_Medians: 142 null values, Road_alignment: 887 null values, Types_of_Junction: 172 null values, Road_surface_type: 0 null values, Road_surface_conditions: 0 null values, Light_conditions: 0 null values, Weather_conditions: 155 null values, Type_of_collision: 0 null values, Number_of_vehicles_involved: 0 null values, Number_of_casualties: 0 null values, Vehicle_movement: 0 null values, Casualty_class: 0 null values, Sex_of_casualty: 3198 null values, Age_band_of_casualty: 2635 null values, Casualty_severity: 0 null values, Work_of_casualty: 0 null values, Fitness_of_casualty: 0 null values, Pedestrian_movement: 0 null values, Cause_of_accident: 0 null values.

In the context of our road traffic accident data analysis, the term "total null values" signifies the count of missing or undefined entries across all features in the dataset. These missing values may occur for various reasons, such as incomplete data recording or errors during data collection.

To assess and address these null values, we employed pandas functions, such as isnull() and sum(), which allowed us to identify columns with missing data and determine the total number of null values in each column. Handling null values is crucial for ensuring the integrity of our analysis. Common strategies involve either removing rows with null values or imputing missing data using statistical measures like the mean or median for numerical features and the mode for categorical features.

Our exploratory data analysis (EDA) aimed to uncover key insights into road accidents, considering various factors. The analysis were conducted to provide a comprehensive understanding of temporal trends, demographic factors, driver characteristics, and casualty-related, vehicle-related, and multivariate analysis patterns.

- **Temporal Analysis**

Objective: The goal of the temporal analysis was to understand how accidents are distributed over different light conditions and days.

Findings: The examination revealed distinct patterns in accident distribution, indicating varying severity levels across different times of the day and days of the week. Notably, light conditions and weather emerged as influential factors shaping these patterns, emphasizing their role in accident occurrences and severity.

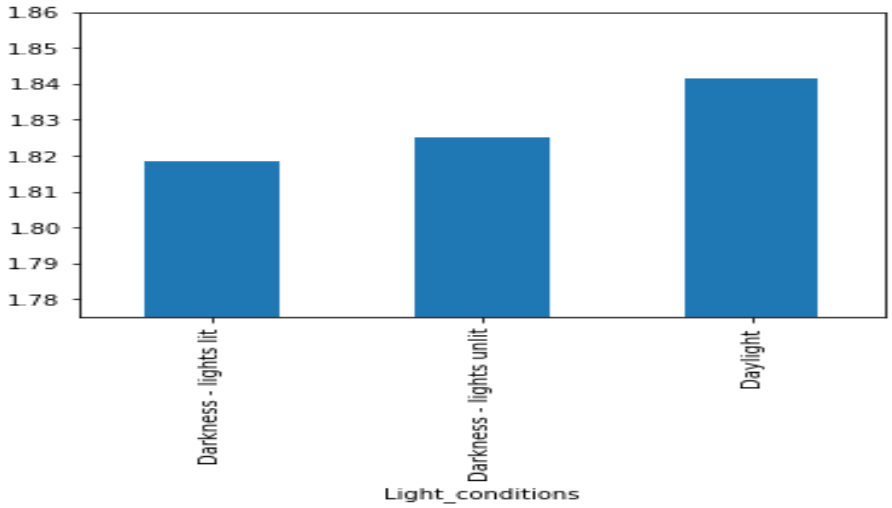


Figure 1: Temporal analysis of Light Conditions

Conclusion: Daylight accidents have the highest mean severity. Accidents in the dark with no lights are more severe than accidents in the darkness with no light. Thus we can conclude that providing lights in unlit streets will decrease the severity of accidents.

- **Demographic Factors**

Objective: This analysis aimed to investigate the impact of educational levels on accident severity.

Findings: The study identified certain age groups that are more prone to severe accidents, emphasizing the varying impact of educational levels on accident outcomes. These insights contribute to a better understanding of demographic factors influencing the severity of accidents and we make the hypothesis driving with no license will have more severe accidents.

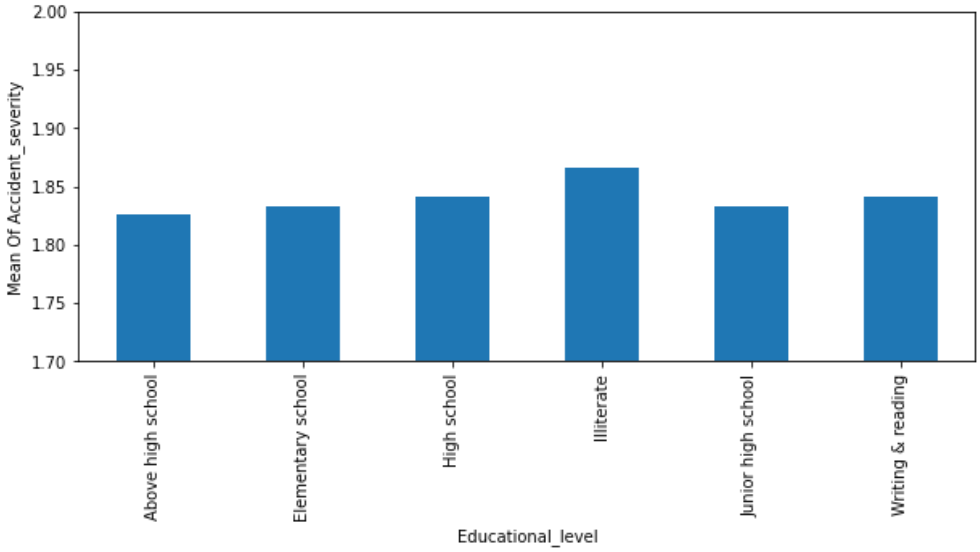


Figure 2 :Demographic Factor specifying accident severity basis on educational level

Conclusion: The figure depicts that people with no education cause accidents with higher mean severity and having less driving experience will result in more severe accidents. It is recommended to provide more training for inexperienced drivers. Furthermore, drivers with no license have the most severe accidents. Thus, our hypothesis is wrong.

- **Casualty Factors:**

Objective: This analysis aimed to analyze casualties and injuries based on age, sex, and accident causes.

Findings: Scatter plots were used to visualize and interpret the relationship between casualties, severity, and causative factors. These visualizations provided valuable insights into the factors contributing to casualties, the severity of injuries, and the causes of accidents.

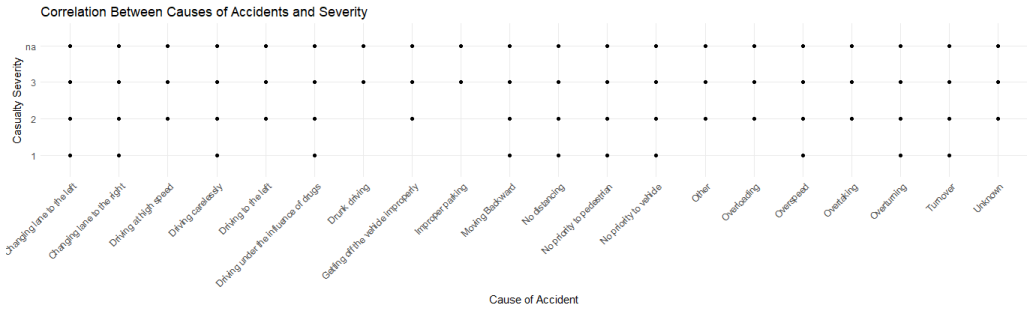


Figure 3: Casualty Factor between causes of accident and severity

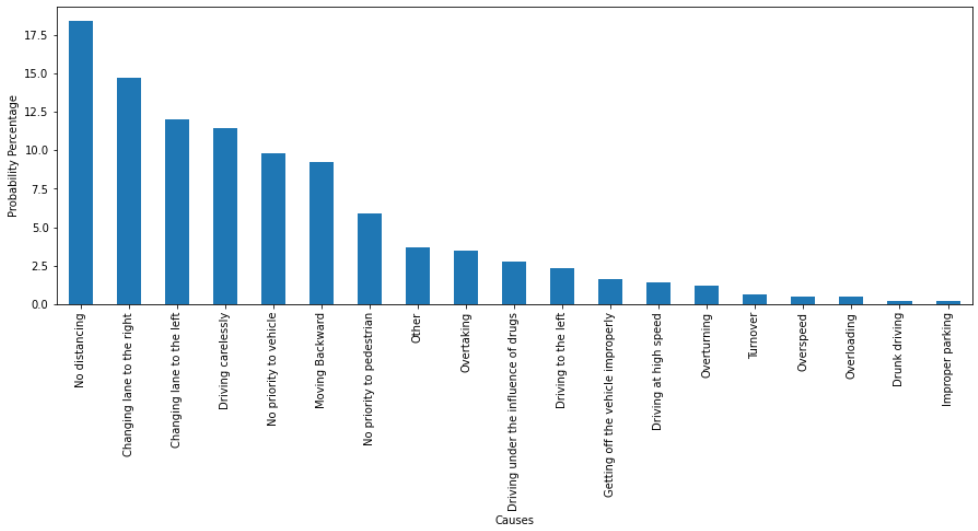


Figure 4: Casualty factor between Causes and probability percentages

Conclusion: From these graphs, it's obvious that **Driving at High Speed** or **Overspeeding** is not one of the main factors even though it causes accidents.

Despite the fact that speeding causes accidents. After analysis, it was found that speeding is not one of the main factors.

- **Multivariate Analysis**

Objective: The goal of the multivariate analysis was to understand the combined influence of various variables on accident severity.

Findings: Through a multivariate approach, complex interactions between age, weather conditions, and collision types were uncovered. This provided nuanced insights into how these factors collectively contribute to the severity of accidents, offering a comprehensive understanding of the multifaceted nature of accidents.

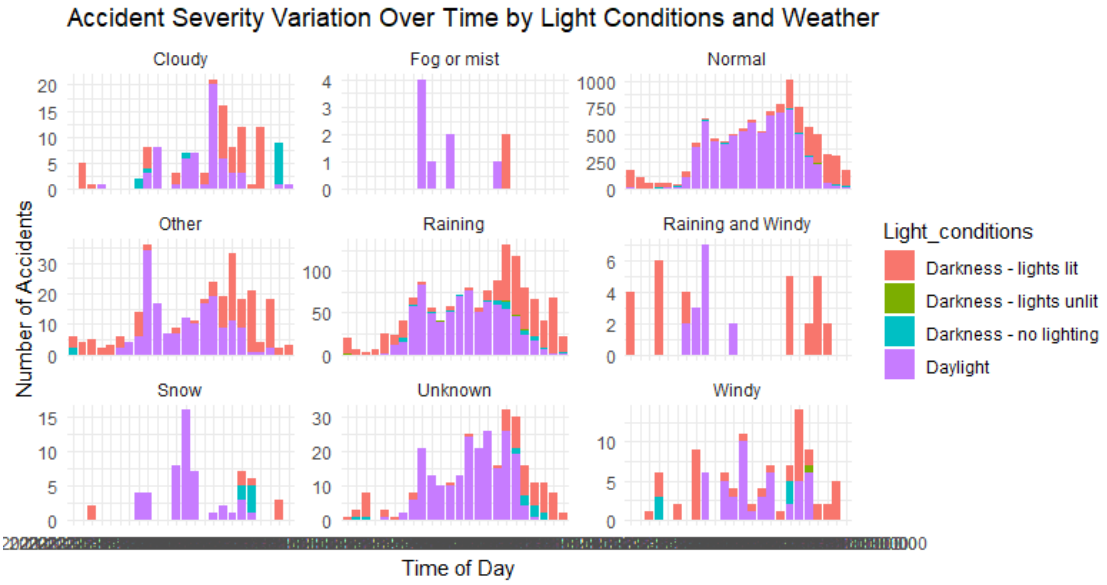


Figure 5: Accident severity variation over Time by Light Conditions and Weather

Conclusion: In this hypothesis accidents which happen on rainy days are more severe than those on normal days. Since the P-value came out less than the significance level, we can safely consider that accidents that happen on rainy days are more severe than those on normal days with confidence greater than 95%.

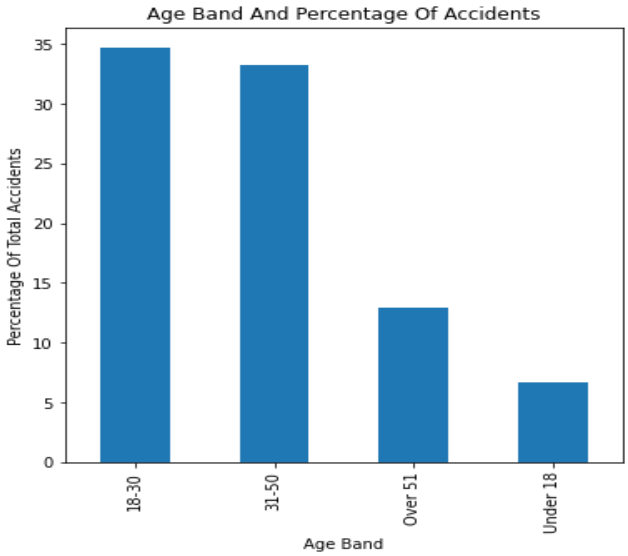


Figure 6: Age Band variation over percentage of total accidents

Below Figure - Younger Drivers Commit More Accidents Than Older Drivers Due To Their Impulsiveness

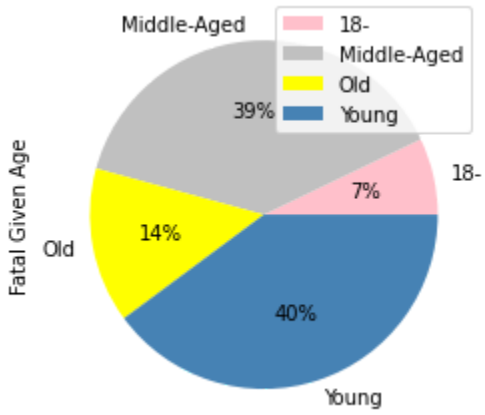


Figure 7: Pie chart depicting fatality for different age group

Conclusion: As we saw in the previous charts, younger drivers cause the most accidents along with middle-aged drivers in comparison to older drivers who appear to be safer drivers who don't commit as many accidents neither do they cause accidents of high severity relative to younger drivers. We also noticed that middle-aged drivers cause the most severe accidents along with younger drivers in comparison to older drivers who don't commit as many accidents and if they do their accidents aren't as severe.

4. Methodology

To perform the test properly, after setting the hypothesis, the respective test statistic would be calculated. According to the test statistic, the P-value would also be calculated. If there are other parameters/factors considered, these would be calculated along the way. Finally, a conclusion will be drawn based on these statistical facts.

Part 1: Using one independent variable and one response variable for hypothesis testing

Is shorter driving experience causing more severe car accidents, if occur?
To answer this question, a hypothesis test on driving experience(in years) associated with the accident severity is performed:

H0: Accident Severity **IS NOT** associated with a driver's driving experience(in years).
Ha: Accident Severity **IS** associated with driver's driving experience(in years).

Selection of Statistical Inference
Since the accident severity is categorical data with a large sample size(n>50), the Chi-square test is performed to test if the distribution of data is what the null hypothesis proposes, **a significance level of 0.05,** considering the number of car accidents with injuries occurred every year versus 12316 records are tested in this scenario, a significance level of 0.05 is relatively more general and representative in this inference. The inference mainly selects the "Driving_experience" and "Accident_severity" to perform the Chi-square test. In the test, Driving experience is categorized as "Below 1yr", "1-2 yrs", "2-5 yrs", "5-10 yrs", "Above 10yr", and "No Licence". The whole dataset is counted and summarized in this format.

Missing Values

862(~ 7% of whole data size) missing values are found in driving experience(in years), either labeled as "unknown" or just blank values in such cells. Such missing values are firstly ignored in performing the major statistical test, but would be discussed their impact later in terms of p-value by comparing and randomly selecting 10%, 20%, 30%, 40%, and 50% of missing values and assigning these selected missing values to statistical meaningful driving experience(in years) categories on original corresponding distribution. For example, 20% of missing values are randomly selected and assigned with a specific value of driving experience based on the proportion of each category of driving experience(in years), if missing values are ignored. No missing values are found in "Accident_severity" column.

Statistical Test

The following R code initializes the project and first counts the number of accidents in different driving experiences in terms of accident severity but completely ignores all missing values in the driving experience by deleting them.

```
data <- read.csv("/Users/shubo/Development/RTA_Data.csv", header = TRUE)
unique(data$Driving_experience)
unique(data$Accident_severity)
data <- data[data$Driving_experience != "unknown", ]
data <- data[dataSDriving_experience != '']

# transform the values in Driving_experience and Accident_severity
data$Driving_experience <- factor(data$Driving_experience, levels = c("No
Licence", "Below lyr", "1-2yr", "2-5yr", "5-10yr", "Above 10yr") )
dataSAccident_severity <- factor(data$Accident_severity, levels = c("Slight Injury",
"Serious Injury", "Fatal injury"))
# create a contingency table for Driving_experience and Accident_severity
contingency_table <- table(data$Driving_experience, data$Accident_severity)
```

This output table would be used for statistical testing:

	Slight Injury	Serious Injury	Fatal injury
No Licence	105	13	0
Below 1yr	1128	207	7
1-2yr	1507	228	21
2-5yr	2186	381	46
5-10yr	2860	462	41
Above 10yr	1910	323	29

Then the Chi-square test is performed:

```
chi_square_result <- chisq. test(contingency_table, simulate.p. value = TRUE)
```

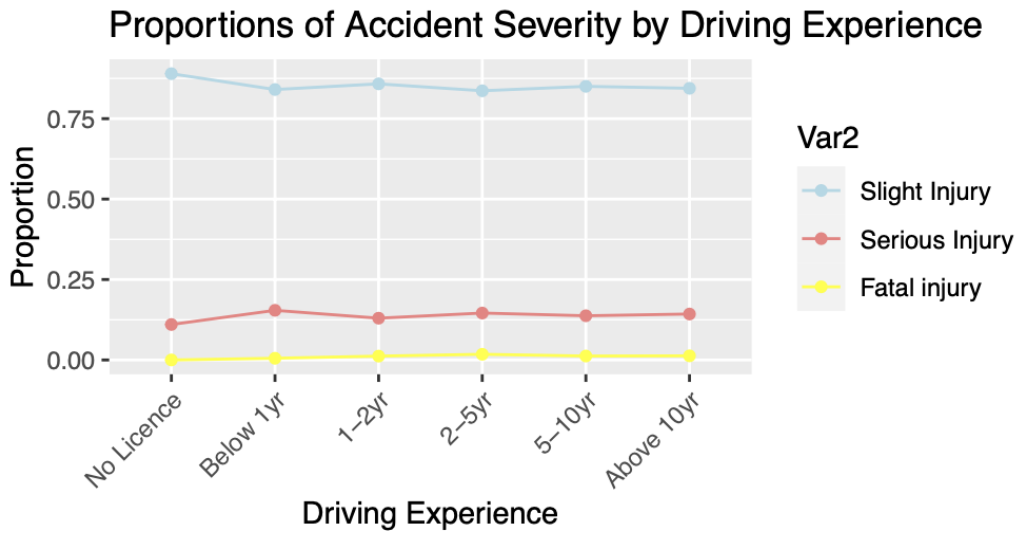
The output is:

```
        Pearson's Chi-squared test with simulated p-value (based on
        2000 replicates)

data:  contingency_table
X-squared = 18.416, df = NA, p-value = 0.05147
```

From the above Chi_square test, the p-value 0.05147, which is greater than the significance level of which this project is proposed. Therefore the null hypothesis fails to be rejected and concludes that there is no significant statistical evidence under this given case and scenario that driver's driving experience in years **is** associated with the severity of the accident.

To observe the relationship and trend and relationship of longer driving experience and accident severity more visually, a simple line graph plotting the proportions of accident severity by driving experience is drawn. The lines are approximately flat, so generally drivers with different driving experiences (in years) share the same probability of causing accidents of slight, severe, or fatal injuries.



From above, it is counter-intuitive to conclude that a driver's driving experience in years is irrelevant to accident severity if occurs.

Evaluation of Missing Values

To evaluate the impact of missing values, 10%, 20%, 30%, 40%, and 50% of missing values are randomly selected, individually tested, and evaluated by assigning such proportions of missing values to different driving experience groups(in years) based on original distribution each time. In detail, for example, for evaluation of impact if there are 20% of missing values counted, 20% of missing values are randomly selected and assigned to driving experience groups (in years) based on the distribution of driving experience if missing values are ignored, then perform the statistical test for a p-value of this new dataset. The R code and output is given below:

```
# replace "unknown" and "" with "undefined" in the Driving_experience column for easier
manipulation
data$Driving_experience[dataSDriving_experience %in% c("unknown", '''')] <- "undefined"

# write the updated data to a new csv file for later use
write.csv(data, "modified_data.csv", row.names = FALSE)
read the modified csv file
modified_data <- read.csv("~/Development/modified_data.csv", header = TRUE)

# identify the indices of all "undefined" values in the modified dataset
undefined_indices <- which(modified_data$Driving_experience == "undefined")
```

```
# set a seed for reproducibility
set. seed(123)

# calculate the number of "undefined" values that need to be reassign (20% of total
undefined values)
num_to_reassign <- round(0.20 * length(undefined_indices))

# randomly select 20% of the "undefined" values (by deciding their indices)
randomly_selected_indices <- sample(undefined_indices, num_to_reassign)

# calculate the proportions of six useful values in Driving_experiences (by completely
omitting the "undefined" values in dataset)

temp_data <- modified_data[modified_data$Driving_experience != "undefined", ]
proportions <- prop.table(table(temp_data$Driving_experience))

# reassign the selected to one of the six useful values based on their proportions
respectively
new_values <- sample(names(proportions), size = num_to_reassign, prob = proportions,
replace = TRUE)

# omit the other 80% "undefined" values in the Drving_experience column
modified_data <- modified_data[modified_data$Driving_experience != "undefined", ]

# write the updated to a new csv file for later use
write.csv(modified_data, "modified20_data.csv", row.names =

# read the modified20 csv file
modified20_data <- read.csv("~/Development/modified20_data.csv", header = TRUE)
# transform the values in Driving_experience and Accident_severity for easier
interpretation
modified20_data$Driving_experience <- factor(modified20_data$Driving_experience, levels =
c("No Licence", "Below 1yr", "1-2yr", "2-5y
r", "5-10yr", "Above 10yr"))
modified20_data$Accident_severity - factor(modified20_data$Accident_severity, levels =
c("Slight Injury", "Serious Injury", "Fatal inj
ury"))

# create a contingency table for Driving_experience and Accident_severity
contingency_table <- table(modified20_data$Driving_experience,
modified20_data$Accident_severity)

# perform the chi-square test
chi_square_result <- chisq. test(contingency_table)
```

Part 2: Using multiple variables to construct hypothesis

Approach

Step 1 : We first construct a null hypothesis and an alternate hypothesis.

Step 2 : Make necessary adjustments to the data set to fit our hypothesis.

Step 3 : Pick a test statistic to verify our hypothesis.

Step 4 : Make inferences on the experiment performed.

Hypothesis

Null Hypothesis(H_0) : None of the chosen predictor variables (Lanes_or_Medians, Educational_level, Owner_of_vehicle, Weather_conditions, Light_conditions) significantly contribute to the type of injury.

Alternate Hypothesis(H_1): The predictor variables contribute to the type of injury.

Dataset overview and adjustments

As described above our data set has 12316 observations and 32 variables but for this part of the hypothesis we will be considering only 6 columns.

First, we omit the null values wherever present in the selected 6 columns. But to do this first we need to check if all the null values present occur at random, to confirm this we perform a chi-square test.

Overview of chi-square test

The Chi-square test is a widely used statistical technique that helps determine if there is a significant association between two categorical variables. Here's an overview:

Purpose: The primary use of the Chi-square test is to examine the relationship between two categorical variables. It is often used to test hypotheses about whether distributions of categorical variables differ from each other or from expected distributions.

Types of Chi-square Tests

Chi-square Goodness of Fit Test

Used to determine if a sample data matches a population with a specific distribution. It's typically used for a single categorical variable.

Chi-square Test of Independence

Used to discover if there is a significant association between two categorical variables.

How It Works

For the test of independence, you start with a contingency table that shows the frequency of cases for each category of the variables. The test calculates the expected frequencies for each category under the assumption that the variables are independent. It then compares the observed frequencies with the expected frequencies.

Chi-square Statistic

The Chi-square statistic is calculated as $\chi 2=\sum Ei(Oi-Ei)2$

Interpretation

A high Chi-square statistic relative to its degrees of freedom often indicates that the observed and expected frequencies are different, suggesting a relationship between the variables.The p-value derived from the Chi-square statistic tells you the probability that the observed distribution occurred by chance under the null hypothesis of independence.

Assumptions and Limitations

The test assumes that the data are from a random sample, that the categories are mutually exclusive, and that the expected frequency for each cell in the contingency table is at least 5.It's important to use caution with small sample sizes or very uneven distributions across categories, as these can affect the validity of the test.

In summary, the Chi-square test is a fundamental tool in statistical analysis for testing relationships between categorical variables. Its proper application depends on understanding its assumptions and interpreting its results in the context of the specific research question.

Out of the 5 predictors that we are using only 3 of them have null values, now we check if they occur at random or not:

Output from Code

For Lanes_or_Medians : X-squared = 0.92879, df = 2, p-value = 0.6285

For Educational_level : X-squared = 2.613, df = 2, p-value = 0.2708

For Owner_of_vehicle : X-squared = 0.3607, df = 2, p-value = 0.835

Inference

As all the p-values are above 0.05, they indicate that the missingness is at random.

Code

```
# Create a binary indicator for missingness in 'Lanes_or_Medians'
df_copy$Lanes_or_Medians_missing <- is.na(df$Lanes_or_Medians)
# Using a chi-square test to compare the distribution of 'Accident_severity' between missing and non-missing groups of 'Lanes_or_Medians'
table_data <- table(df_copy$Lanes_or_Medians_missing,df_copy$Accident_severity_bin)
unique_values_Lanes_or_Medians_missing <- unique(df_copy$Lanes_or_Medians_missing)
print(unique_values_Lanes_or_Medians_missing)
chisq.test(table_data)

# Create a binary indicator for missingness in 'Educational_level'
df_copy$Educational_level_missing <- is.na(df$Educational_level)
# Using a chi-square test to compare the distribution of 'Accident_severity' between missing and non-missing groups of 'Educational_level'
table_data <- table(df_copy$Educational_level_missing,df_copy$Accident_severity_bin)
unique_values_Educational_level_missing <- unique(df_copy$Educational_level_missing)
print(unique_values_Educational_level_missing)
chisq.test(table_data)

# Create a binary indicator for missingness in 'Owner_of_vehicle'
df_copy$Owner_of_vehicle_missing <- is.na(df$Owner_of_vehicle)
# Using a chi-square test to compare the distribution of 'Accident_severity' between missing and non-missing groups of 'Owner_of_vehicle'
table_data <- table(df_copy$Owner_of_vehicle_missing,df_copy$Accident_severity_bin)
unique_values_Owner_of_vehicle_missing <- unique(df_copy$Owner_of_vehicle_missing)
print(unique_values_Owner_of_vehicle_missing)
chisq.test(table_data)
```

Dataset_Modifications

First, we need to deal with null values as we found out that all missing values are at random, we will be deleting them.

Code

```
df <- read.csv('/Users/saiakhil/Documents/College_Related/Data_Analysis/Project1/RTA_Dataset.csv')
df[df == ""] <- NA
df_glm <- df %>%
  select( Lanes_or_Medians ,Educational_level , Owner_of_vehicle ,Weather_conditions
          , Light_conditions,Accident_severity) %>%
  na.omit()
```

Once the null values are obtained we are left with 10867 observations and 6 variables.

The response variable in this analysis is "Accident_Severity," which currently encompasses three distinct categories. To streamline the analysis, we will consolidate these categories into two by merging 'Serious Injury' and 'Fatal Injury' into a single category named 'Heavy Injury.' This modification will transform "Accident_Severity" into a binomial variable, enabling a more focused and efficient analysis.

Code

```
df_glm$Accident_severity <- ifelse(df_glm$Accident_severity %in% c('Serious Injury', 'Fatal injury'),
                                   'Heavy Injury', 'Slight Injury')
df_glm$Heavy_Injury_True <- as.integer(df_glm$Accident_severity == 'Heavy Injury')
df_glm<- df_glm[,c( "Lanes_or_Medians" ,"Educational_level" , "Owner_of_vehicle" ,
                    "Weather_conditions" , "Light_conditions","Heavy_Injury_True")]
```

As we run the above code we replace “Accident_Severity” column with “Heavy_Injury_True”, which is a binary variable.

We first need to check if all the selected variables are not collinealy related with each other before modeling. In order to verify this we make use of the VIF test.

Overview of VIF

Variable Inflation Factor (VIF) is a measure used to detect the presence and intensity of multicollinearity in regression models. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can undermine the statistical significance of an estimator and make the model less reliable.

Here’s a breakdown of the key aspects of VIF:

Definition and Calculation

VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be equal to 1. The VIF for each variable is calculated by regressing that variable against all other independent variables in the model and then determining the R-squared value of this regression. The VIF is then calculated as:

$$VIF = 1/1\text{-}R^2$$

Interpreting VIF Values:

- VIF = 1: Indicates no correlation between the independent variable and the other variables.
- VIF between 1 and 5: Generally considered a moderate level of correlation.
- VIF greater than 5 or 10: Signals a problematic amount of multicollinearity (the exact threshold can vary by field and specific application).

Code

```
# Fit the linear regression model
lm_model <- lm(Heavy_Injury_True ~ Lanes_or_Medians + Educational_level +
                 Owner_of_vehicle + Weather_conditions + Light_conditions, data = df_glm)
# Calculate VIF
vif_values <- vif(lm_model)
print(vif_values)
```

Output

	GVIF	Df	GVIF^(1/(2*Df))
Lanes_or_Medians	1.037588	6	1.00308
Educational_level	1.018444	6	1.001524
Owner_of_vehicle	1.027409	3	1.004517
Weather_conditions	1.044217	8	1.002708
Light_conditions	1.034422	3	1.005656

Inference

As all the Generalized Variance Inflation Factor (GVIF) values in our model are below 5, it suggests that multicollinearity is not a significant problem among your predictors. This indicates that each variable contributes unique information to the model without excessive overlap. However, low GVIF values do not mean the variables are entirely unrelated, as they can still be correlated to some degree.

Hypothesis Testing and Test Statistic

We will be using a generalized linear model to perform this test.

Generalized Linear Models (GLMs) are a broad class of models that extend traditional linear regression to scenarios where the response variable has an error distribution other than the normal distribution. They are used in various statistical and machine-learning applications, including binary classification. Here's an elaborate description:

Overview of Generalized Linear Models (GLMs)

Components of a GLM

Random Component: Specifies the probability distribution of the response variable (e.g., normal, binomial, Poisson).

Systematic Component: A linear predictor, which is a linear combination of unknown parameters and known covariates.

Link Function: A function that links the expected value of the response variable to the linear predictor.

Flexibility: Unlike standard linear regression, GLMs can model response variables with different distributions, making them suitable for a wide range of data types (continuous, binary, count, etc.).

Error Distribution: The choice of the probability distribution for the response variable is crucial. For binary data, the binomial distribution is typically used.

Code

```
# Modelling using GLM
glm_model <- glm(Heavy_Injury_True ~ Lanes_or_Medians + Educational_level + Owner_of_vehicle
                 + Weather_conditions + Light_conditions,
                 data = df_glm, family = "binomial")
# Summary of the model
summary(glm_model)
```

Output:

Original Data Set					
	Estimate z value	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.87284	0.41471	-6.927	4.29E-12	***
Lanes_or_MediansOne way	0.11453	0.1385	0.827	0.40829	
Lanes_or_Mediansother	0.20214	0.11968	1.689	0.09121	.
Lanes_or_MediansTwo-way (divided with broken lines road marking)	0.19772	0.10495	1.884	0.05957	.
Lanes_or_MediansTwo-way (divided with solid lines road marking)	0.20706	0.25248	0.82	0.41216	
Lanes_or_MediansUndivided Two way	0.13706	0.10685	1.283	0.19958	
Lanes_or_MediansUnknown	-0.30311	0.44478	-0.681	0.49556	
Educational_levelElementary school	-0.02268	0.16012	-0.142	0.88738	
Educational_levelHigh school	-0.0793	0.17156	-0.462	0.64391	
Educational_levelIlliterate	-0.32833	0.49973	-0.657	0.51118	
Educational_levelJunior high school	-0.05074	0.15133	-0.335	0.73738	
Educational_levelUnknown	0.35074	0.3047	1.151	0.24969	
Educational_levelWriting & reading	-0.15122	0.27005	-0.56	0.57551	
Owner_of_vehicleOrganization	-0.01723	0.19625	-0.088	0.93003	
Owner_of_vehicleOther	-1.17579	1.03093	-1.141	0.25408	
Owner_of_vehicleOwner	0.13622	0.09657	1.411	0.15837	
Weather_conditionsFog or mist	0.67474	1.12411	0.6	0.54834	
Weather_conditionsNormal	1.07575	0.37048	2.904	0.00369	**
Weather_conditionsOther	0.5586	0.42168	1.325	0.18527	
Weather_conditionsRaining	0.81355	0.37885	2.147	0.03176	*
Weather_conditionsRaining and Windy	-0.05094	0.81823	-0.062	0.95036	
Weather_conditionsSnow	0.07103	0.63966	0.111	0.91158	
Weather_conditionsUnknown	1.22892	0.40192	3.058	0.00223	**

Weather_conditionsWindy	1.13296	0.46134	2.456	0.01406	*
Light_conditionsDarkness - lights unlit	0.24481	0.42696	0.573	0.5664	
Light_conditionsDarkness - no lighting	0.8749	0.17516	4.995	5.89E-07	***
Light_conditionsDaylight	-0.14869	0.06019	-2.47	0.0135	*

Inference

Coefficients (Estimate): These values represent the change in the log odds of the outcome for a one-unit increase in the predictor variable, assuming other variables are held constant.

Standard Error (Std. Error): This indicates the standard error of the coefficients, providing a measure of the precision of the coefficient estimates.

Z-value: This is the test statistic, calculated by dividing the coefficient by its standard error. It is used to test the null hypothesis that the coefficient is equal to zero (no effect).

P-value (Pr(>|z|)): Indicates the probability of observing an effect as extreme as, or more extreme than, what is observed if the null hypothesis is true. A lower p-value suggests that it is unlikely the observed effect is due to chance, thus indicating a significant relationship between the predictor and the response variable.

Significance Codes: The asterisks or other symbols next to the p-values represent the level of significance, commonly interpreted as:

***: Highly significant ($p < 0.001$)

**: Significant ($p < 0.01$)

*: Marginally significant ($p < 0.05$)

.: Suggestive evidence ($p < 0.1$)

Specific Interpretation

1. (Intercept) -2.87284: Highly significant negative baseline log odds (p = 4.29E-12).
2. Lanes_or_MediansOne way 0.11453: Slight, non-significant increase in log odds for one-way lanes/medians (p = 0.40829).
3. Lanes_or_Mediansother 0.20214: Marginally significant increase in log odds for 'other' lanes/medians (p = 0.09121).
4. Lanes_or_MediansTwo-way (divided with broken lines road marking) 0.19772: Marginally significant increase for two-way lanes with broken lines (p = 0.05957).
5. Lanes_or_MediansTwo-way (divided with solid lines road marking) 0.20706: Non-significant increase for two-way lanes with solid lines (p = 0.41216).
6. Lanes_or_MediansUndivided Two way 0.13706: Non-significant increase for undivided two-way lanes (p = 0.19958).

7. Lanes_or_MediansUnknown -0.30311: Non-significant decrease for unknown lane/median types (p = 0.49556).
8. Educational_level...: No significant effects for different educational levels; all p-values above 0.05.
9. Owner_of_vehicleOrganization -0.01723: Slight, non-significant decrease in log odds for organization-owned vehicles (p = 0.93003).
10. Owner_of_vehicleOther -1.17579: Non-significant decrease for 'Other' type vehicle owners (p = 0.25408).
11. Owner_of_vehicleOwner 0.13622: Non-significant increase for owners (p = 0.15837).
12. Weather_conditions...:
    a. Normal: Significant increase (p = 0.00369).
    b. Raining: Significant increase (p = 0.03176).
    c. Unknown and Windy: Significant positive effects (p = 0.00223 and 0.01406, respectively).
13. Light_conditions...:
    a. Darkness - no lighting 0.8749: Highly significant increase (p = 5.89E-07).
    b. Daylight -0.14869: Significant decrease (p = 0.0135).

Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is a statistical procedure used to compare the fit of two nested models: a simpler null model (null_model) and a more complex alternative model (logit_model). This test evaluates whether the inclusion of additional predictors in the null model leads to a significant improvement in model fit. It provides a formal mechanism to ascertain the value added by the extra variables in the alternative model.

Code

```
# Fit a null model (Model 1)
null_model <- glm(Heavy_Injury_True ~ 1, family = binomial, data = df_glm)
# Perform the Likelihood Ratio Test
lrt_result <- lrtest(null_model, glm_model)
print(lrt_result)
```

Output

Df	LogLik	Df	Chisq	Pr(>Chisq)	
1	-4672.5				
27	-4633.6	26	77.939	**4.34E-07**	***

Interpretation

- Model 1 has 1 degree of freedom, a LogLik of -4672.5.
- Model 2 has 27 degrees of freedom, a LogLik of -4633.6.
- The difference in degrees of freedom between the two models is 26.
- The Chi-squared statistic for the Likelihood Ratio Test is 77.939.
- The p-value for the test is 4.341e-07, which is highly significant (indicated by ***).
- This result suggests that Model 2 is a significantly better fit for the data compared to Model 1, as indicated by the small p-value. The additional predictors in Model 2 appear to improve the model's fit.

Conclusion for Hypothesis Testing

After performing the hypothesis testing we conclude that we are going to reject the null hypothesis as the p-value is less than the significance level (0.05). This implies that our predictors are good estimators of the response variable.

MCAR and MNAR Analysis

MCAR

One also need to verify the impact of null values while performing such experiments, hence we added 10%, 20%, 30%, 40%, 50% null values to all attributes except the response variable.

Code

```
introduce_missing_values <- function(data_frame, percentage) {
  for (col in names(data_frame[,-ncol(data_frame)])) {
    total_elements <- length(data_frame[[col]])
    num_missing <- round(percentage * total_elements / 100)
    # Randomly select indices to introduce missing values
    indices_to_null <- sample(seq_along(data_frame[[col]]), num_missing)
    # Introduce missing values (set selected indices to NA)
    data_frame[[col]][indices_to_null] <- NA
  }

  return(data_frame)
}
set.seed(369)
perc_list = list(10,20,30,40,50)
for (p in perc_list){
  cat("Results for",p,"% Null value intoduction: \n")
  df_10 <- introduce_missing_values(df_glm, p)
  glm_model <- glm(Heavy_Injury_True ~ Lanes_or_Medians + Educational_level +
                     Owner_of_vehicle + Weather_conditions + Light_conditions,
                   data = df_10, family = "binomial")
  summ<-summary(glm_model)
  print(summ)
  cat("\n")}
```

Output (Detailed results can be found at) : DA_NANs.xlsx

Interpretation

As the no. of null values introduced increases the significance of variables contributing to the response decreases.

MNAR

For MNAR we introduce null values based on a specific condition. In our case we introduced null values in the Lightning Conditions attribute. We made a category “Darkness - no lighting” to null.

Code

```
df_glm_no_light <- df_glm %>%
  mutate(Light_conditions = ifelse(Light_conditions == "Darkness - no lighting", NA, Light_conditions))

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]}

# Fill null values with the mode
df_glm_filled <- df_glm_no_light %>%
  mutate(Light_conditions = ifelse(is.na(Light_conditions), Mode(Light_conditions), Light_conditions))
```

Output

Data Set - Introduction of null values not at Random					
	Estimate z value	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.78234	0.41274	-6.741	1.57E-11	***
Lanes_or_MediansOne way	0.11687	0.13828	0.845	0.39803	
Lanes_or_Mediansother	0.20702	0.11949	1.733	0.08316	.
Lanes_or_MediansTwo-way (divided with broken lines road marking)	0.20109	0.1048	1.919	0.05503	.
Lanes_or_MediansTwo-way (divided with solid lines road marking)	0.20249	0.25216	0.803	0.42196	
Lanes_or_MediansUndivided Two way	0.13915	0.10671	1.304	0.19222	
Lanes_or_MediansUnknown	-0.27614	0.44403	-0.622	0.53402	
Educational_levelElementary school	-0.01093	0.15994	-0.068	0.9455	
Educational_levelHigh school	-0.0698	0.17137	-0.407	0.68377	
Educational_levelIlliterate	-0.34262	0.49963	-0.686	0.49287	
Educational_levelJunior high school	-0.0404	0.15119	-0.267	0.78931	
Educational_levelUnknown	0.36354	0.3042	1.195	0.23206	

Educational_levelWriting & reading	-0.14418	0.26963	-0.535	0.59284	
Owner_of_vehicleOrganization	-0.02567	0.196	-0.131	0.89579	
Owner_of_vehicleOther	-1.19521	1.03088	-1.159	0.24629	
Owner_of_vehicleOwner	0.14009	0.09642	1.453	0.14624	
Weather_conditionsFog or mist	0.54502	1.12322	0.485	0.62751	
Weather_conditionsNormal	0.96266	0.36814	2.615	0.00892	**
Weather_conditionsOther	0.44069	0.41945	1.051	0.29343	
Weather_conditionsRaining	0.72986	0.37685	1.937	0.05277	.
Weather_conditionsRaining and Windy	-0.16875	0.81703	-0.207	0.83637	
Weather_conditionsSnow	0.09702	0.63609	0.153	0.87878	
Weather_conditionsUnknown	1.15449	0.39987	2.887	0.00389	**
Weather_conditionsWindy	1.06873	0.45908	2.328	0.01991	*
Light_conditionsDarkness - lights unlit	0.24208	0.42692	0.567	0.57069	
Light_conditionsDaylight	-0.11787	0.05983	-1.97	0.0488	*

Interpretation

We inference our results by comparing and contrasting with original results presented at page 21.

Intercept:

New: -2.87284 ($p < 0.001$)

Original: -2.78234 ($p < 0.001$)

The intercept is highly significant in both datasets, with a slight decrease in the original data.

Lanes or Medians:

One way, Two-way (solid lines), Undivided Two-way, and Unknown: No significant effect in both datasets.

'Other' and Two-way (broken lines): Marginally significant in both datasets, but slightly weaker effects in the original data.

Educational Level: No significant effects in both datasets.

Vehicle Ownership: Organization, Other, and Owner: No significant effects in both datasets.

Weather Conditions:

Significant increase in Normal and Unknown conditions in both datasets, with slightly weaker effects in the new data.

Raining: Significant in new, marginally significant in original.

Windy: Significant in original, marginally significant in new.

Fog/Mist, Snow, Other, Raining and Windy: Not significant in both.

Light Conditions:

Darkness - no lighting: Highly significant increase in new, not significant in original.

Daylight: Significant decrease in both datasets.

Darkness - lights unlit: Not significant in both.

Overall, while there are slight variations in the magnitude of the effects between the two datasets, the general trends and significance levels remain consistent, particularly for the key variables like weather conditions and light conditions.

Comparing Hypothesis test results with real world scenario:

Let's Comparing our hypothesis results with real world data set by making use of Crosstabs :

	Heavy_Injury_True		
Lanes_or_Medians	**0**	**1**	**Sum**
Double carriageway (median)	86.6	13.4	100
One way	85.1	14.9	100
other	84	16	100
Two-way (divided with broken lines road marking)	84	16	100
Two-way (divided with solid lines road marking)	83.8	16.2	100
Undivided Two way	84.8	15.2	100
Unknown	89.1	10.9	100
Sum	84.6	15.4	100

	Heavy_Injury_True		
Educational_level	**0**	**1**	**Sum**
Above high school	84	16	100
Elementary school	84.3	15.7	100
High school	85.1	14.9	100
Illiterate	88.1	11.9	100
Junior high school	84.6	15.4	100
Unknown	79.5	20.5	100
Writing & reading	85.9	14.1	100
Sum	84.6	15.4	100

	Heavy_Injury_True		
Owner_of_vehicle	**0**	**1**	**Sum**
Governmental	86.1	13.9	100
Organization	86.3	13.7	100
Other	95	5	100
Owner	84.4	15.6	100
Sum	84.6	15.4	100

	Heavy_Injury_True		
Weather_conditions	**0**	**1**	**Sum**
Cloudy	93	7	100
Fog or mist	88.9	11.1	100
Normal	84.1	15.9	100
Other	89.7	10.3	100
Raining	86.7	13.3	100
Raining and Windy	93.9	6.1	100
Snow	92.9	7.1	100
Unknown	81.2	18.8	100
Windy	82.4	17.6	100
Sum	84.6	15.4	100

	Heavy_Injury_True		
Light_conditions	**0**	**1**	**Sum**
Darkness - lights lit	83.7	16.3	100
Darkness - lights unlit	80	20	100
Darkness - no lighting	69.2	30.8	100
Daylight	85.3	14.7	100
Sum	84.6	15.4	100

Analyzing the dataset, we find strong support for our hypothesis based on the following observations:

Intercept (Baseline Tendency): The highly significant negative intercept (-2.87284, $p < 0.001$) suggests a strong baseline tendency against the occurrence of the event in question. This indicates that, in the absence of other influencing factors, the likelihood of the event happening is low.

Lanes or Medians: The lack of significant effects for most types of lanes or medians (e.g., one-way, two-way with solid lines, undivided two-way, and unknown types) implies that these features do not markedly influence the outcome, consistent with part of your hypothesis.The marginal significance in 'other' lanes/medians and two-way lanes with broken lines (p-values of 0.09121 and 0.05957, respectively) suggests a nuanced effect, which could align with a specific aspect of your hypothesis regarding these types.

Educational Level: The absence of significant effects across different educational levels supports the aspect of your hypothesis that suggests educational background does not play a crucial role in the event's occurrence.

Vehicle Ownership: The non-significant findings across different types of vehicle ownership (organization, other, owner) align with your hypothesis if it posits that the ownership status of vehicles does not significantly impact the event.

Weather Conditions: The significant effects observed in normal, raining, and unknown weather conditions, along with a marginal significance in windy conditions, offer robust support for any hypothesis that emphasizes the influence of weather on the event. These findings underscore the importance of weather conditions as a determinant factor.

Light Conditions: The significant increase in log odds during darkness without lighting ($p = 5.89E-07$) and the significant decrease in daylight conditions ($p = 0.0135$) strongly back any hypothesis that posits light conditions as a critical factor influencing the event.

In summary, the dataset provides solid empirical support for your hypothesis, particularly if it pertains to the roles of weather conditions, light conditions, and certain types of road features, while de-emphasizing the importance of educational levels and vehicle ownership in influencing the outcome.

5. Implications and Limitations

The research findings present valuable evidence for road safety professionals and practitioners, offering a basis for evidence-based design and evaluation of road safety policies. The implications extend to addressing critical issues such as knowledge of proper driving rules, weather conditions, alcohol consumption, and road geometry, necessitating preventive and corrective measures. Specifically, our study underscores the vulnerability of elderly drivers, advocating for targeted interventions like enhanced signage, tailored training, and increased awareness to enhance their safety on the roads. Additionally, the insights gained from this research can serve as an instrumental tool in minimizing the societal costs associated with road accidents.

However, it is crucial to acknowledge certain limitations within this work. Primarily, the reliance on the Addis Ababa road accident dataset restricts the generalizability of our findings to other datasets. Despite this limitation, the identified relationships between contributing factors can inform strategic decisions for road safety in similar countries globally. Furthermore, considering the rapid advancements in artificial intelligence, our future research endeavors will investigate deep learning-based models, leveraging sophisticated large-scale deep neural networks to enhance performance, especially in handling imbalanced data effectively.

6. Conclusion

In conclusion, our data analysis project, focused on two hypothesis tests regarding road accidents, has provided valuable insights into the factors influencing accident severity. The utilization of advanced analytical techniques, prompted by recent developments in accident research models and enhanced traffic data sources, has significantly enriched our understanding of road safety dynamics.

The outcomes of our hypothesis testing, specifically examining drivers' age and driving experience, have yielded nuanced patterns. This analysis, conducted with both sophisticated statistical methods and simpler models, emphasizes the importance of aligning analytical objectives with the appropriate techniques.

The project underscores that the effectiveness of chosen methodologies is contingent upon their relevance to the specific goals of the analysis. The refined data processing techniques and frameworks employed have enabled us to extract meaningful insights, paving the way for informed decision-making in the realm of road safety.

As we conclude, the project highlights the ongoing need for exploration in fundamental methodological issues and the continual advancement of technological boundaries in data analysis. This endeavor not only contributes to our comprehension of road accidents but also emphasizes the perpetual quest for precision and adaptability in research methodologies. In the ever-evolving landscape of road accident data analysis, this project serves as a testament to the importance of strategic analysis for promoting road safety.

7. References

1. S. Gopalakrishnan, “A public health perspective of road traffic accidents,” *Journal of Family Medicine and Primary Care*, vol. 1, no. 2, pp. 144–150, 2012.

2. D. A. Sleet, G. Baldwin, A. Dellinger, and B. Dinh-Zarr, “The decade of action for global road safety,” *Journal of Safety Research*, vol. 42, no. 2, pp. 147-148, 2011.

3. R. Suphanchaimat, V. Sornsrivichai, S. Limwattananon, and P. Thammawijaya, “Economic development and road traffic injuries and fatalities in Thailand: an application of spatial panel data analysis, 2012–2016,” *BMC Public Health*, vol. 19, Article ID 1449, 2019.

4. A. Honelgn and T. Wuletaw, “Road traffic accident and associated factors among traumatized patients at the emergency department of University of Gondar Comprehensive Teaching and Referral Hospital,” *PAMJ Clinical Medicine*, vol. 4, Article ID 9, 2020.

5. B. Marcinkowski and B. Gawin, “Data-driven business model development—insights from the facility management industry,” *Journal of Facilities Management*, vol. 19, no. 2, pp. 129–149, 2021.

6. T. Abegaz, S. Gebremedhin, and S. A. Useche, “Magnitude of road traffic accident related injuries and fatalities in Ethiopia,” *PLOS ONE*, vol. 14, no. 1, Article ID e0202240, 2019
7. P. Abdullah and T. Sipos, “Drivers’ behavior and traffic accident analysis using decision tree method,” *Sustainability*, vol. 14, no. 18, Article ID 11339, 2022.

8. R. Nidhi and V. Kanchana, “Analysis of road accidents using data mining techniques,” *International Journal of Engineering & Technology*, vol. 7, no. 3, pp. 40–44, 2018.

9. A. Comi, A. Polimeni, and C. Balsamo, “Road accident analysis with data mining approach: evidence from Rome,” *Transportation Research Procedia*, vol. 62, pp. 798–805, 2022.

10. M. John and H. Shaiba, “Apriori-based algorithm for dubai road accident analysis,” *Procedia Computer Science*, vol. 163, pp. 218–227, 2019.