Project Report on

ACCIDENT PREDICTION IN INDIA ANALYSIS

In partial fulfillment for the award

Of

Professional Certification in Data Analysis and Visualization

Year 2024-2025

Submitted By

Sreelakshmi.R

Tools & Technologies Used:

Python, R Programming, Tableau

Duration:

15/07/2025 – 30/07/2025

Submission Date:

31/07/2025

G - TEC CENTRE OF EXCELLENCE PERINTHALMANNA

Abstract

Road accidents are a major public health and infrastructure issue in India, causing significant fatalities and economic losses each year. This project aims to predict and analyze accident severity and related patterns using real-world data collected from Kaggle. The dataset includes features like weather conditions, helmet/seatbelt use, alcohol involvement, and vehicle type.

The project involves data preprocessing using Python, statistical hypothesis testing using R (T-test, ANOVA, Z-test, F-test, Chi-Square), and interactive dashboard creation using Tableau. The analysis revealed statistically significant relationships between helmet use, weather, and accident outcomes. These insights help inform road safety policies and interventions, contributing to better decision-making and public safety.

Introduction

India records one of the highest numbers of road traffic accidents globally, with annual fatalities exceeding 150,000. This alarming statistic represents not just a loss of human life but also a massive socio-economic burden. Road traffic incidents are attributed to a variety of contributing factors, including human negligence, inadequate use of safety gear, alcohol consumption, poor road infrastructure, and adverse weather conditions. These causes reflect systemic challenges in traffic law enforcement, public awareness, and infrastructure planning.

Despite numerous efforts by governmental agencies and non-governmental organizations, the current strategies for reducing accidents are often reactive and lack a robust, data-driven foundation. One significant challenge is the underutilization of the vast amount of accident data collected every year. While data is being gathered, it is seldom used effectively for in-depth analysis that can lead to proactive safety interventions.

This project addresses this critical gap by conducting a comprehensive statistical analysis of road accident data in India. The dataset, sourced from Kaggle, encompasses approximately 5,000 accident records from different Indian states. By leveraging Python for data cleaning and preprocessing, R for statistical hypothesis testing, and Tableau for visualizing key trends, the project aims to uncover actionable insights into the underlying causes and severity of road accidents.

The primary objectives of this study are threefold: (1) To assess key risk factors such as helmet and seatbelt usage, weather conditions, and alcohol involvement that contribute to the severity of road accidents; (2) To analyze the statistical significance of these factors using well-established inferential techniques; and (3) To create an interactive dashboard that provides intuitive access to the findings for stakeholders including policymakers, law enforcement, and public health officials.

The scope of this project is limited to the static dataset available on Kaggle. While the dataset offers valuable insights, it lacks fine-grained features such as GPS-level location data and real-time accident reporting. Furthermore, the current phase does not incorporate

predictive machine learning models; rather, it focuses on establishing statistically significant relationships that can guide future modeling efforts.

Nevertheless, the project holds significant business value. Insights derived from the analysis can be instrumental for traffic law enforcement, enabling them to identify high-risk zones and times for targeted interventions. Urban planners and infrastructure authorities can use the results to improve road design and signage based on accident hotspots and conditions. Moreover, the Tableau dashboard created as part of this project makes the findings accessible to a broader audience, fostering greater public awareness and encouraging safer road usage behaviour.

By integrating data preprocessing, statistical analysis, and dynamic visualization, this project serves as a foundational step toward more intelligent and responsive road safety strategies in India.

The project objectives are as follows:

- To identify and assess the primary factors contributing to accident severity.
- To apply statistical hypothesis testing techniques in R (such as T-tests, ANOVA, and Chi-Square tests) to establish the significance of relationships between variables.
- To construct an interactive Tableau dashboard for intuitive, real-time visual exploration of key insights.

The scope of the project is limited to exploratory and inferential analysis. Real-time data feeds, GPS-level granularity, and predictive modeling are excluded from this phase due to data constraints. Additionally, while the study offers insight into patterns and relationships, it does not establish causality or simulate intervention effects.

Nonetheless, the potential impact is considerable. The results can support data-informed decision-making in public safety, enabling traffic law enforcers to identify high-risk behaviors and accident-prone conditions. Urban planners can leverage location and time-based trends

to optimize traffic flow, signage, and infrastructure safety features. Public safety campaigns can utilize the insights to promote greater compliance with safety measures like helmets and seatbelts.

Furthermore, the Tableau dashboard democratizes access to complex findings by offering a user-friendly interface for policymakers, researchers, and the general public. It serves as a communication bridge between technical data analysis and practical implementation, thereby enhancing the societal impact of the project.

In summary, this project contributes to road safety in India by combining statistical rigor, computational efficiency, and visual storytelling to uncover meaningful patterns in accident data and offer scalable solutions for future traffic management and policy interventions.

Literature Review

1. Data Visualization and Dashboarding Techniques
   Visualization-based studies have become increasingly popular for accident monitoring and reporting. Roy et al. (2021) created Tableau dashboards showing accident hotspots and temporal trends, improving public awareness. However, most dashboards are descriptive and fail to incorporate hypothesis testing or inferential statistics
2. Human Factors and Road Accident Severity
   Studies such as those by Mohan et al. (2019) have emphasized the importance of human behavior—including speeding, alcohol consumption, and failure to wear helmets or seatbelts—in influencing accident severity. In particular, Sharma and Singh (2020) found a strong correlation between non-helmet usage and fatal accidents in urban Indian settings.
3. Global Perspective on Accident Analysis: Several international studies have employed machine learning and statistical methods to identify risk factors contributing to road accidents. For instance, Zhang et al. (2018) used decision trees and logistic regression to predict crash severity based on weather, driver behaviour, and vehicle conditions. Jiang and Zhang (2020) emphasized the importance of multi-factor data fusion, combining sensor data with external variables such as traffic signals, to improve prediction accuracy.
4. Impact of Environmental Conditions
   Research by Das and Maurya (2018) indicated that adverse weather conditions like rain, fog, and poor visibility significantly contribute to accident risk. The findings were further supported by Mehta et al. (2021), who used weather-index-based modeling to demonstrate that rainy and foggy conditions increase the probability of fatal crashes by 40% in India.
5. Statistical vs. Machine Learning Approaches
   While traditional statistical methods such as Chi-Square and ANOVA (as used by Iyer et al., 2017) help identify significant relationships between categorical variables, modern studies have employed machine learning (ML) models like Decision Trees, Random Forests, and SVMs for predicting accident severity. For

example, Kumar and Bansal (2022) used logistic regression to classify accident outcomes and achieved 73% accuracy. However, most ML approaches lack interpretability and are not commonly adopted in policy decisions due to their "black-box" nature.

Research Gap

Although multiple studies have investigated accident risk factors in India, most of them adopt either descriptive analytics or predictive machine learning without validating relationships statistically. For instance, while ML models may classify accident severity, they often do not explain *why* certain features are significant. Similarly, descriptive dashboards improve accessibility but do not offer statistical rigor.

Moreover, few studies combine multiple tools—like Python for cleaning and EDA, R for formal statistical testing, and Tableau for interactive dashboards—to present a holistic view. Many existing works focus on limited regions, lack real-world datasets, or omit the integration of multiple influencing variables (like weather, alcohol use, and safety gear) in a unified framework.

Data Collection and Pre-processing

Data Source and Collection Methods

The dataset used for this project, titled accident_prediction_india.csv, was obtained from Kaggle and contains 5,000 records on road accidents reported across various Indian states. The dataset includes crucial information such as:

- Time of day, day of the week
- Type of vehicle involved
- Number of casualties and fatalities
- Use of helmets or seatbelts
- Weather conditions at the time of the accident
- Alcohol involvement
- Accident severity

The data was collected from multiple official sources and traffic surveillance records before being published on Kaggle. It was chosen for its richness in variables that contribute directly to accident outcomes and severity levels.

Data Quality Assessment and Cleaning Procedures

The cleaning process was performed using Python (with pandas, numpy, and seaborn libraries). Key steps included:

- Missing Values: Handled missing or null values by imputing with appropriate substitutes (e.g., using mode for categorical values).
- Data Types**:** Categorical fields like Weather_Condition, Helmet_Seatbelt_Used, and Alcohol_Involved were converted to categorical types, and numeric fields were verified for consistency.
- Inconsistencies: Standardized inconsistent entries (e.g., "Helmet used" vs. "Helmet Used") to ensure uniformity in group-based analyses, and applied logical corrections where necessary (e.g., ensuring fatality counts align with accident severity, or seatbelt usage isn't marked 'Yes' for two-wheelers).

Feature Engineering and Selection Techniques

To support effective visualization and statistical testing, new derived features were created and irrelevant columns were dropped:

- **Time_Category:** Transformed exact times into categories like Morning, Afternoon, Evening, and Late Night for better trend analysis.
- **Age_Group:** Grouped individual ages into logical bins (e.g., 18–30, 31–45, etc.).
- **Severity_Level_Numeric:** Converted Accident_Severity (Minor, Major, Fatal) into numeric codes for correlation and regression analysis.
- **High_Risk_Flag:** A binary field created to indicate accidents involving high-risk conditions (e.g., no helmet + night time + rainy weather).

Columns Selected for Analysis

Accident_Severity,Helme_,Seatbelt_Used,Alcohol_Involved,Weather_Condition,Time_Category,Vehicle_Type, Number_of_Fatalities, Age_Group, Traffic_Control_Presence, Number_of_Casualties

Methodology

The research adopts a multi-layered analytical framework integrating programming, statistical reasoning, and visual storytelling to explore and predict accident patterns across Indian states. By combining Python for data preparation, R for statistical validation, and Tableau for interactive visualization, the methodology ensures both technical robustness and communicative clarity.

Technologies and Tools Applied

1. Python was the primary tool for data import, preprocessing, transformation, and initial exploration. Libraries such as pandas, numpy, seaborn, matplotlib, and plotly allowed detailed inspection of variables influencing accident outcomes.
2. R Programming supported statistical examination through hypothesis testing and variance analysis. Core functions included chisq.test() for categorical independence, t.test() for mean comparisons, and aov() for multi-group comparisons.
3. Tableau facilitated the design of a user-focused, visual analytics dashboard. It enabled seamless communication of results to stakeholders, decision-makers, and the general public.

Exploratory Data Analysis – Python

- Data Cleaning:
  - Removal of nulls, duplicates, and logically inconsistent entries (e.g., 0 fatalities with severe damage).
  - Recoding categorical fields like “Helmet_Seatbelt_Used” and “Traffic_Control” for clarity.
- Univariate Analysis:
  - Frequency distributions of accident severity, weather conditions, time zones, etc.
- Bivariate Analysis:
  - Pairwise relationships visualized using stacked bar plots and Pie Chart.
- Grouping & Aggregation:

- Used groupby() to aggregate and analyse means, counts, and distributions based on different features.
- The crosstab() function was applied to understand relationships between two categorical variables helping in determining the association between them.
- Used. groupby() to calculate total accidents per state/year/month, average casualties per vehicle type, etc
- Employed pd.crosstab() to identify high-risk intersections between factors (e.g., Foggy Weather vs. Night Time).

Statistical Testing (R Programming)

1. Data Import

The cleaned dataset, exported from Python, was imported into R for further analysis. Column names were standardized, and categorical variables were converted into factors for appropriate statistical testing.

T-Test (Independent Samples T-Test)

The T-test was used to compare the mean number of fatalities between two groups: those who used helmets/seatbelts and those who did not.

F-Test (Variance Comparison)

The F-test was applied to check whether the variance in fatalities differs based on alcohol involvement

Z-Test (Proportion Testing)

A Z-test was used to compare the sample mean number of fatalities to a hypothetical population mean (e.g., 5 fatalities per incident).

ANOVA (Analysis of Variance)

ANOVA was conducted to examine if fatalities significantly vary across different weather conditions.

The Chi-square test was employed to test the association between helmet/seatbelt use and accident severity (Fatal, Major, Minor)

Data Visualization and Dashboarding - Tableau

After completing the data preparation and analysis, Tableau was utilized to develop an interactive and intuitive dashboard. The choice of Tableau was driven by its capability to transform complex data into a visually appealing format, thereby facilitating stakeholders' understanding and decision-making based on the insights derived.

Key Features of the Dashboard:

- Interactive Filters: The inclusion of filters such as vehicle type, Current Year and Accident Severity enabled users to delve into various data subsets, enhancing the exploratory analysis.

- Simple Layout and Clear Visuals: The dashboard was crafted with a straightforward layout, leveraging color coding and labels to accentuate key findings. This design ensured that users could readily discern trends and patterns within the data.

Benefits of the Dashboard:

1. Enhanced Data Interpretation: The visual representation of data facilitated a deeper understanding of road accident trends and patterns.

2. Data-Driven Decision Making: By providing stakeholders with actionable insights, the dashboard supported informed decision-making regarding road safety measures.

3. User Engagement: The interactive nature of the dashboard encouraged users to explore the data in greater detail, fostering a more comprehensive understanding of the factors influencing road accidents.

RESULTS AND ANALYSIS

Python- Based Results

1. Year-wise Accident Trends: clear upward trend in accidents from 2020 (611 cases) to 2025 (1,099 cases), with the highest spike in 2025. This steady rise indicates a growing road safety concern. The sharp increase after 2022 suggests the need for stronger preventive measures and policy interventions.
2. Top 10 states with highest fatal accidents: Uttar Pradesh, Goa, an d Tamil Nadu top the list for fatal accidents, indicating critical h otspots for road safety issues. This insight helps policymakers an d transport authorities prioritize these states for stricter enforcem ent, awareness campaigns, and infrastructure improvements.
3. Signals record the highest number of fatal and major accidents, indicating that mere presence of signals may not ensure safety. This insight helps authorities focus on enforcing traffic rules and improving signal management**.**
4. Foggy, rainy, and stormy weather show high fatal and major accident counts, with no minor cases—highlighting severe outcomes in poor weather. This emphasizes the need for weather-specific safety measures and public awareness during adverse conditions.
5. Late night and evening show the highest accident counts, suggesting increased risk during darker or high-traffic hours. This insight can help authorities prioritize patrol and safety campaigns during these critical time windows.
6. Two-wheelers (bikes) are the most involved in accidents, followed by autos, cars, vans, and buses. This highlights the vulnerability of bike riders and the need for stricter safety measures and awareness campaigns targeted at two-wheeler users.
7. Accidents involving alcohol show significantly higher fatal and major cases (2,362 fatal) compared to non-alcohol cases (1,116 fatal). This emphasizes the critical need for stricter enforcement of drink-and-drive laws and public awareness initiatives to reduce alcohol-related accidents.

8. Fatal accidents are predominantly associated with male drivers, accounting for over 80% of the cases. This insight highlights the need for targeted road safety education and behavioral interventions focusing on male driving behavior.
9. The 15–24 age group recorded the highest number of accidents, followed by the 25–44 and 45–64 groups. This indicates that young and middle-aged drivers are most at risk, suggesting the importance of focused awareness programs and stricter enforcement for these age groups.
10. Insight: Fatal accidents account for over 73% of all reported cases, while major accidents make up the remaining 26.7%. This highlights the critical severity of accidents in India, emphasizing the urgent need for preventive measures, stricter enforcement, and better emergency response systems to reduce fatal outcomes
11. Insight: Among fatal accidents, male drivers (2,792) were involved far more frequently than female drivers (686) .This highlights a gender disparity in fatal accidents, suggesting that male drivers may be at a higher risk or more exposed to risky driving conditions. This can guide targeted road safety policies, awareness campaigns, and driver training programs focused on high-risk groups.

R-Based Statistical Results

1. Insight: The T-test shows a statistically significant difference in the average number of fatalities between individuals who did not use helmets/seatbelts (mean = 1.83) and those who did (mean = 1.50), with a p-value < 0.05. This result strongly supports that helmet and seatbelt usage significantly reduces fatalities in accidents, highlighting the critical role of safety gear in saving lives and guiding policymakers in enforcing protective gear laws.
2. Insight: The F-test reveals a statistically significant difference in the variance of fatalities between accidents with and without alcohol involvement (F = 1.21, p-value < 0.001). This indicates that alcohol involvement leads to more variability in fatality

outcomes, underscoring the unpredictable and heightened danger of drinking and driving. This insight can aid in developing stricter DUI enforcement policies.

3. Insight: The ANOVA test shows a statistically significant difference in the number of fatalities across different weather conditions ($F = 57.68$, $p < 0.001$). This finding suggests that weather plays a crucial role in accident fatality rates. Policymakers and traffic authorities can use this insight to issue weather-based safety alerts and implement preventive measures during high-risk weather conditions like rain, fog, or storms.
4. Insight: The Chi-Square test reveals a statistically significant association between helmet/seatbelt use and accident severity ($\chi^2 = 9.95$, $p = 0.0069$). This indicates that protective gear usage like helmets and seatbelts is linked to reduced accident severity, supporting the enforcement of strict safety regulations and awareness campaigns to promote their usage.
5. Insight: The Z-Test shows that the sample mean of fatalities significantly differs from the hypothetical population mean of 5 (Z = value, $p < 0.05$).This indicates that the actual fatality rate is statistically lower or higher than expected, emphasizing the need to revisit assumptions used in safety planning, policy decisions, or national fatality benchmarks.

Tableau- Based Results

- Total Records Analyzed: 5,000 road accident cases included.
- Year Comparison: Current year shows more accidents than the previous year.
- Fatal Accidents Increased: Fatal cases are higher in the current year.
- Weather Impact: Most accidents occur in foggy and rainy weather.
- Road Type Risk: State highways have the highest number of accidents.

- Traffic Control: Signals have the most accidents; police check-posts have the fewest.
- Helmet/Seatbelt Use: Accidents are more severe when helmets or seatbelts are not used.
- Alcohol Involvement: Male drivers under alcohol influence are more involved in accidents than females.
- Vehicle Type: Two-wheelers and autorickshaws are the top vehicles in accident cases.
- Driver Gender: Male drivers are involved in more fatal accidents than female drivers

Conclusion

This analysis forms a critical component of our academic project aimed at understanding and mitigating the causes of road accidents in India. By leveraging a dataset of 5,000 records, we applied an integrated analytical approach using Python for data preprocessing and exploratory data analysis, R for statistical validation, and Tableau for intuitive and interactive data visualization.

The study demonstrated the value of combining multidisciplinary tools to extract meaningful patterns from raw accident data. Our objective was not only to fulfill academic requirements but also to develop a solution-oriented understanding of public safety concerns. The comprehensive analysis provided insights into the relationships between accident severity and contributing factors such as environmental conditions, road type, human behavior, and safety compliance.

This project reinforced key academic concepts like hypothesis testing, data visualization, and multivariate analysis in a real-world context. The outcome also highlights the growing importance of data-driven policymaking, especially in areas like transportation safety and urban planning.

Future Work

While this study provided significant insights, it also opens up several opportunities for future exploration:

- Incorporating real-time sensor or GPS data from vehicles for predictive modeling.
- Expanding the dataset to include longitudinal trends across multiple years and more granular regional data.
- Using machine learning algorithms to build a predictive model for accident severity or hotspot detection.
- Collaborating with traffic departments to integrate accident analytics into city-level transport safety dashboards.

References

1. Ministry of Road Transport & Highways, Government of India. (2023). *Road Accidents in India – Annual Report*.
2. Singh, S. (2017). *Critical Factors in Road Accidents: An Indian Perspective*. Indian Journal of Transport.
3. World Health Organization. (2022). *Global Status Report on Road Safety*.
4. Kumar, R., & Mehta, V. (2020). *Predictive Analytics for Road Accidents Using Machine Learning Techniques*. Journal of Data Science & Applications.
5. Gupta, A., & Bansal, M. (2019). *Impact of Weather and Road Infrastructure on Road Accidents in India*. International Journal of Engineering Research.

SUPPORTING FILES

Python :

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("C:\\Users\\Sreelakshmi\\Desktop\\accident_prediction_india.csv")
df.head()
```

```
df.shape
df.info()
df.isnull().sum()
df.describe()
```

```
#Convert to hour and make a new column Time category
```

```
df['Hour'] = pd.to_datetime(df['Time of Day'], errors='coerce').dt.hour
df.head()
```

```
def categorize_hour(hour):
    if pd.isnull(hour):
        return "Unknown"
    elif 0 <= hour < 6:
        return "Late Night"
    elif 6 <= hour < 12:
        return "Morning"
    elif 12 <= hour < 18:
        return "Afternoon"
    else:
        return "Evening"
df['Time Category'] = df['Hour'].apply(categorize_hour)
```

```
df['Hour'] = pd.to_datetime(df['Time of Day'], errors='coerce').dt.hour
df.head()
```

```
def categorize_hour(hour):
    if pd.isnull(hour):
        return "Unknown"
    elif 0 <= hour < 6:
        return "Late Night"
    elif 6 <= hour < 12:
        return "Morning"
    elif 12 <= hour < 18:
        return "Afternoon"
    else:
        return "Evening"
df['Time Category'] = df['Hour'].apply(categorize_hour)
df.head()
```

```
yearly_counts = df['Year'].value_counts().sort_values(ascending=False)
print("Total accidents per year (descending):")
print(yearly_counts)
```

```
import matplotlib.pyplot as plt

# Step 1: Count total accidents per year and sort by year (ascending)
yearly_counts = df['Year'].value_counts().sort_index()

# Step 2: Plot the bar chart
plt.figure(figsize=(10, 6))
bars = plt.bar(yearly_counts.index.astype(str), yearly_counts.values, color='skyblue')
```

```
# Step 3: Add value labels on top of each bar
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval + 10,  # slight offset above bar
             f'{yval}', ha='center', va='bottom', fontsize=9)

# Step 4: Add titles and formatting
plt.title("Total Number of Accidents Per Year (Ascending)", fontsize=14)
plt.xlabel("Year")
plt.ylabel("Number of Accidents")
plt.xticks(rotation=0)
plt.tight_layout()

# Step 5: Show the chart
plt.show()
```

```
fatal_by_state = df[df["Accident Severity"] == "Fatal"]['State Name'].value_counts().head(10)
print("Top 10 states with highest fatal accidents:")
print(fatal_by_state)
```

```
import matplotlib.pyplot as plt

# Step 1: Get top 10 states by fatal accidents
fatal_by_state = df[df['Accident Severity'] == 'Fatal']['State Name'].value_counts().head(10)

# Step 2: Plot horizontal bar chart
plt.figure(figsize=(10, 6))
bars = plt.barh(fatal_by_state.index, fatal_by_state.values, color='skyblue')
```

```
# Step 3: Add value labels INSIDE each bar (to the right)
for bar in bars:
    width = bar.get_width()
    plt.text(width - 10,                    # Move inside the bar
             bar.get_y() + bar.get_height()/2,
             f'{width}',
             va='center', ha='right',        # Align to right inside the bar
             color='black', fontsize=9)

# Step 4: Invert y-axis so highest is on top
plt.gca().invert_yaxis()

# Step 5: Titles and formatting
plt.title("Top 10 States with Fatal Accidents", fontsize=14)
plt.xlabel("Number of Fatal Accidents")
plt.ylabel("State")
plt.tight_layout()

# Step 6: Show chart
plt.show()
```

```
print(pd.crosstab(df['Traffic Control Presence'], df['Accident Severity']))
```

```
# Crosstab between Weather_Condition and Accident Severity
result = pd.crosstab(df['Weather_Condition'], df['Accident Severity'])

# Sort by Weather_Condition alphabetically
print(result.sort_index())
```

```
df['Time Category'] = df['Hour'].apply(categorize_hour)
print(df['Time Category'].value_counts())
```

```
time_counts = df['Time Category'].value_counts().reindex(['Late Night', 'Morning', 'Afternoon', 'Evening'])

plt.figure(figsize=(8, 5))
plt.pie(time_counts.values, labels=time_counts.index, autopct='%1.1f%%', colors=['skyblue', 'lightgreen', 'lightcoral', 'gold'])
plt.title("Accidents by Time of Day", fontsize=14)
plt.tight_layout()
plt.show()
```

```
vehicle_counts = df['Vehicle_Type'].value_counts().head(10)
plt.figure(figsize=(10, 6))
bars = plt.barh(vehicle_counts.index, vehicle_counts.values, color='skyblue')
for bar in bars:
    width = bar.get_width()
    plt.text(width - 10, bar.get_y() + bar.get_height()/2, f'{width}', ha='right', va='center')
plt.gca().invert_yaxis()
plt.title("Top Vehicle Types Involved in Accidents", fontsize=14)
plt.xlabel("Number of Accidents")
plt.tight_layout()
plt.show()
```

```
print(pd.crosstab(df['Alcohol Involvement'], df['Accident Severity']))
```

```
# Frequency table to check seatbelt use against severity
seatbelt_vs_severity = pd.crosstab(df['Helmet_or_Seatbelt'], df['Accident Severity'])
print(seatbelt_vs_severity)
```

```
#GENDER WISE ACCIDENT SEVERITY
fatal_gender = df[df['Accident Severity'] == 'Fatal']['Driver Gender'].value_counts()
print(fatal_gender)
```

```
age_counts = df['Age_Group'].value_counts().sort_index()
print(age_counts)
```

```
plt.figure(figsize=(10, 6))
ax = sns.countplot(data=df, x='Age_Group', hue='Alcohol Involvement', palette='magma')

plt.title('Alcohol Involvement by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.legend(title='Alcohol Involvement')

# Add data labels on each bar
for container in ax.containers:
    ax.bar_label(container, fmt='%d', label_type='edge', padding=3)

plt.tight_layout()
plt.show()
```

```
# Grouping and plotting
traffic_severity = df.groupby(['Traffic Control Presence', 'Accident Severity']).size().unstack().fillna(0)
```

```python
# Grouping and plotting
traffic_severity = df.groupby(['Traffic Control Presence', 'Accident Severity']).size().unstack().fillna(0)

ax = traffic_severity.plot(kind='bar', stacked=True, colormap='Set2', figsize=(10, 6))

plt.title('Accident Severity by Traffic Control Presence')
plt.ylabel('Number of Accidents')
plt.xlabel('Traffic Control Type')
plt.xticks(rotation=45)
plt.legend(title='Accident Severity')
plt.tight_layout()

# Add labels to each segment in the stacked bars
for i, bar_group in enumerate(ax.containers):
    ax.bar_label(bar_group, fmt='%d', label_type='center', padding=2)

plt.show()
```

```python
severity_counts = df['Accident Severity'].value_counts()
severity_percent = df['Accident Severity'].value_counts(normalize=True) * 100

severity_summary = pd.DataFrame({
    'Count': severity_counts,
    'Percentage (%)': severity_percent.round(2)
})

print(" Distribution of Accident Severity Types:")
print(severity_summary)
```

```python
# Load the dataset
df = pd.read_csv("accident_prediction_india.csv")
```

```python
# Load the dataset
df = pd.read_csv("accident_prediction_india.csv")

# Count values
severity_counts = df['Accident Severity'].value_counts()

# Donut chart
colors = ['#66c2a5', '#fc8d62', '#8da0cb']  # Optional: customize colors
plt.figure(figsize=(6, 6))
plt.pie(severity_counts, labels=severity_counts.index, autopct='%1.1f%%',
        startangle=90, colors=colors, wedgeprops={'width': 0.4})

plt.title("Distribution of Accident Severity Types", fontsize=14)
plt.axis('equal')  # Equal aspect ratio ensures the pie is circular
plt.show()
```

```python
pd.crosstab(df['Road Type'], df['Alcohol Involvement'])
```

```python
df.groupby(['Road Type', 'Weather_Condition', 'Alcohol Involvement'])['Number of Casualties']\
.sum().sort_values(ascending=False).head(10)
```

```python
# Define high-risk group: Foggy + Night (or Late Night) + No Helmet/Seatbelt
df_risky = df[
    (df['Weather_Condition'] == 'rainy') &
    (df['Time Category'].isin(['night', 'late night'])) &
    (df['Helmet_or_Seatbelt'] == 'no')
]
```

```
    (df['Helmet_or_Seatbelt'] == 'no')
]

# Summarize risky group
total_risky = len(df_risky)
fatal_risky = df_risky['Fatal'].sum()
fatality_rate_risky = round((fatal_risky / total_risky) * 100, 2)

print(f"High-Risk Group: {total_risky} accidents, {fatal_risky} fatal, Fatality Rate = {fatality_rate_risky}%")
```

```
labels = ['Fatal', 'Non-Fatal']
counts = [fatal_risky, total_risky - fatal_risky]

plt.figure(figsize=(6, 4))
bars = plt.bar(
    labels,
    counts,
    color='lightblue',
    edgecolor='black',   # Thin black border
    linewidth=1.5        # Reduce thickness
)

# Add count labels inside the bars with dynamic placement
for bar in bars:
    yval = bar.get_height()
    position = yval - 10 if yval > 20 else yval + 5  # place inside if tall, above if small
    va = 'top' if yval > 20 else 'bottom'

    plt.text(
        bar.get_x() + bar.get_width() / 2,
        position,
        int(yval),
        ha='center',
```

```
    plt.text(
        bar.get_x() + bar.get_width() / 2,
        position,
        int(yval),
        ha='center',
        va=va,
        fontsize=12,
        fontweight='bold',
        color='black'
    )

# Chart layout
plt.title('Accident Outcomes in High-Risk Group\n(Rainy + Night + No Helmet/Seatbelt)', fontsize=14)
plt.ylabel('Number of Accidents')
plt.ylim(0, max(counts) + 30)  # Add buffer space
plt.tight_layout()
plt.show()
```

```
 #Define custom color palette: dark color for 'Fatal'
severity_order = ['Minor', 'Major', 'Fatal']
custom_colors = ['#B2DF8A', '#FDBF6F', '#1F78B4']  # light green, orange, dark blue for Fatal

fig, axes = plt.subplots(2, 2, figsize=(16, 10))
fig.suptitle("Impact of Key Factors on Accident Severity", fontsize=18)

# 1 Road Type vs Severity
road = df.groupby(['Road Type', 'Accident Severity']).size().unstack().fillna(0)[severity_order]
road.plot(kind='bar', stacked=True, color=custom_colors, ax=axes[0, 0])
axes[0, 0].set_title("Road Type vs Accident Severity")
axes[0, 0].set_ylabel("Accident Count")
axes[0, 0].set_xlabel("Road Type")
```

```
axes[0, 0].set_ylabel("Accident Count")
axes[0, 0].set_xlabel("Road Type")
axes[0, 0].legend(title="Severity")

# 2 Alcohol Involvement vs Severity
alcohol = df.groupby(['Alcohol Involvement', 'Accident Severity']).size().unstack().fillna(0)[severity_order]
alcohol.plot(kind='bar', stacked=True, color=custom_colors, ax=axes[0, 1])
axes[0, 1].set_title("Alcohol Involvement vs Accident Severity")
axes[0, 1].set_ylabel("Accident Count")
axes[0, 1].set_xlabel("Alcohol Involvement")
axes[0, 1].legend(title="Severity")

# 3 Vehicle Count (binned) vs Severity
vehicle = df.groupby(['Vehicle_Category', 'Accident Severity']).size().unstack().fillna(0)[severity_order]
vehicle.plot(kind='bar', stacked=True, color=custom_colors, ax=axes[1, 0])
axes[1, 0].set_title("Number of Vehicles vs Accident Severity")
axes[1, 0].set_ylabel("Accident Count")
axes[1, 0].set_xlabel("Vehicle Category")
axes[1, 0].legend(title="Severity")

axes[1, 1].axis('off')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

]:
```
# Cleaned DataFrame
df.to_csv("cleaned_accident_data.csv", index=False)
```

R programming:

```
# Load required libraries
install.packages("dplyr")
library(dplyr)


# Load the dataset
df <- read.csv("C:\\Users\\Sreelakshmi\\Desktop\\accident_prediction_india.csv")

# View structure
str(df)
summary(df)
# Convert Helmet/Seatbelt to factor if needed
df$Helmet_Seatbelt_Used <- as.factor(df$Helmet_or_Seatbelt)
# OPTIONAL: Clean column names (replace spaces with underscores)
names(df) <- gsub(" ", "_", names(df))
names(df)
# STEP 1: Ensure grouping variable is categorical (factor)
df$Helmet_Seatbelt_Used <- as.factor(df$Helmet_Seatbelt_Used)

# STEP 2: Run T-Test comparing mean fatalities
result <- t.test(Number.of.Fatalities ~ df$Helmet_Seatbelt_Used, data = df)

# STEP 3: Print result
print(result)

# STEP 4: Simple interpretation
if(result$p.value < 0.05) {
  print("Statistically significant: Fatalities differ based on helmet/seatbelt use.")
} else {
  print("Not statistically significant: No clear difference in fatalities based on helmet/seatbelt use.")
}
# Load required packages
library(dplyr)


2#. F-Test
# Compare variance in fatalities by Alcohol involvement
df$Alcohol.Involvement <- as.factor(df$Alcohol.Involvement)

f_test_result <- var.test(Number.of.Fatalities ~ df$Alcohol.Involvement, data = df)
print(f_test_result)

if(f_test_result$p.value < 0.05) {
  print("-Test: Variance in fatalities differs by alcohol involvement.")
} else {
  print("F-Test: No significant difference in variance by alcohol involvement.")
}
```

```r
3#. Z-Test
# Z-test: Compare sample mean of fatalities to hypothetical population mean (e.g., 5)
sample_mean <- mean(df$Number.of.Fatalities, na.rm = TRUE)
sample_sd <- sd(df$Number.of.Fatalities, na.rm = TRUE)
n <- sum(!is.na(df$Number.of.Fatalities))
pop_mean <- 5
pop_sd <- 2
# Step 3: Calculate Z-score and p-value
z_score <- (sample_mean - pop_mean) / (pop_sd / sqrt(n))
p_value_z <- 2 * (1 - pnorm(abs(z_score)))

# Step 4: Print output
cat("Z-Score:", z_score, "\nP-Value:", p_value_z, "\n")

# Step 5: Interpretation
if(p_value_z < 0.05) {
  print("-Test: Sample mean fatalities significantly differ from population mean.")
} else {
  print("Test: No significant difference from population mean.")
}


4#. ANOVA
# ANOVA: Fatalities across different weather conditions
df$Weather_Condition <- as.factor(df$Weather_Condition)
df$Weather_Condition <- as.factor(df$Weather_Condition)

anova_result <- aov(Number.of.Fatalities ~ df$Weather_Condition, data = df)
summary(anova_result)
if(summary(anova_result)[[1]]$`Pr(>F)`[1] < 0.05) {
  print("ANOVA: Statistically significant difference in fatalities across weather conditions.")
} else {
  print("ANOVA: No significant difference in fatalities across weather conditions.")
}



5. #Chi-Square Test
# Chi-square test: Helmet use vs Accident severity
df$Accident_Severity <- as.factor(df$Accident.Severity)

chisq_data <- table(df$Helmet_Seatbelt_Used, df$AccidentSeverity)
chisq_result <- chisq.test(chisq_data)
print(chisq_result)

if(chisq_result$p.value < 0.05) {
  print("Chi-Square: Significant association between helmet use and accident severity.")
} else {
  print("Chi-Square: No significant association between helmet use and accident severity.")
}
```

Tableau: