Crash Analytics — Full Project Documentation

Data Cleaning → Python EDA → Power BI Modeling → 6 Dashboard Pages

1. Data Cleaning Report

1) Accidents Dataset

Purpose:

This dataset contains the essential information for each accident, such as date, location, severity, cause, number of vehicles involved, and casualties. It is the foundation for analysis, so it is critical to ensure the dataset is clean and consistent**.**

Data Cleaning Steps:

- Promoted Headers: Ensured the first row is used as column headers.
- Changed Data Types: Text columns: Country, Month, Day of Week, Time of Day, Urban/Rural, Road Type, Region
- Numeric columns: Emergency Response Time, Traffic Volume, Year, Number of Vehicles, Number of Injuries, Number of Fatalities
- Text Columns Cleaning: Removed invalid characters (e.g., "@")
- Trimmed extra spaces
- Capitalized the first letter of each word (Text.Proper)
- most logical or common value for that column: Country → "Canada"
- Month → "May"
- Day of Week → "Tuesday"
- Time of Day → "Night"
- Urban/Rural → "Rural"
- Road Type → "Main Road"
- Region → "Australia"

Numeric Columns Cleaning:

-Converted to Number type for calculations

-Replaced null with logical or mean values:

Emergency Response Time → 32.506557277086813

Traffic Volume → 5032.8236089941693

Year → 2012

Month converted from text to numeric format (1–12)

Region column cleaned and standardized

Added a new Date column created by combining Day Name, Day, Month and Year into a single date field
Reason for replacing Nan / null:

Missing text values were replaced with the most common or logical values to avoid blank entries that could affect grouping or analysis.

Missing numeric values were replaced with median values to maintain the natural distribution and avoid distortion in averages or visualizations.

---

2) Casualties Dataset Adjusted data types for text and numeric columns:

Purpose:

Provides details about casualties related to each accident, such as the number of injuries, fatalities, pedestrians, or cyclists involved.

Data Cleaning Steps:

Driver Gender: Removed invalid characters ("@")

Trimmed extra spaces

Capitalized first letters

Replaced Nan → "Unknown"

Pedestrians Involved / Cyclists Involved:

Removed invalid text values ("nan")

Converted to Number type

Rounded up numbers

Custom Logic Applied:

If Number of Fatalities = 0 and Number of Injuries = 0, then Pedestrians Involved and Cyclists Involved were set to 0, as incidents with no casualties logically have no involved pedestrians or cyclists.

Remaining missing values were filled using statistical imputation.

• Mean and median were evaluated and both returned 1, so the median (1) was used to replace missing values.

Reasoning Behind Missing Value Handling:

Text-based missing values were replaced with “Unknown” to maintain grouping consistency and avoid misclassification.

For numeric fields, imputation using logical conditions + median ensured realistic

values and improved statistical accuracy without introducing bias.

---

3) Weather Dataset :

Purpose:

Provides information about weather and environmental conditions related to accidents.

Data Cleaning Steps:

Promoted headers and set proper data types.

• Weather Conditions: cleaned text, removed "@", capitalized words, Nan → "Windy".

• Visibility Level: cleaned, converted to Number, rounded to 3 decimals, null → 274.797 (median).

• Road Condition: cleaned text, Nan → "Icy".

• Medians and common values ensure balance in visualizations and modeling.

Reason for replacing Nan / null:

Medians and common values ensure balance in visualizations and modeling

---

2. Python EDA & Data Exploration :

Python Exploratory Data Analysis Steps :

1. Loaded the cleaned datasets

2. Merged datasets on Accidents ID

•

•

- Checked: Missing values
- Duplicates
- Data types
- Summary statistics
-
- **Insights from EDA:**
- Evening and night: highest accident frequency
- Poor visibility increases severity
- Weather conditions (Rain, Snow, Fog) directly increase injuries/fatalities
- Speeding and distracted driving dominant causes
- Rural areas show more severe accidents
- Response time correlates with severity
- Years 2005–2010 show notable increase Urban vs Rural distribution matches 54.28% vs 45.72%
-

- **3.Modeling and Dashboard Development :**
- **Introduction :**
- This document provides a detailed explanation of the modeling approach and the five dashboard pages used to visualize accident-related data.

Modeling Approach :

The modeling framework focuses on:
- Severity Segmentation: Minor, Moderate, Severe accidents with fatalities and injuries.
- Driver Analysis: Gender-based accident and injury distribution; accident causes (Distracted, Drunk, Mechanical, Weather, Speeding).
- Environmental Factors: Weather Impact Index, Road Condition Risk Index, visibility levels, poor road condition percentages.
- Predictive Indicators: Emergency response time correlation with severity

---

Dashboard Pages :

The dashboard is designed to:

- Provide a clear overview of total accidents, injuries, and fatalities
- Show year-to-year and region-to-region variations
- Highlight the influence of driver behavior
- Examine the severity of accidents and the most common accident types
- Visualize geographic hotspots
- Study the impact of weather and road conditions on accident outcomes

Each section contributes to understanding the bigger picture and supports better decision-making.

*****Dashboard Structure**

The report is divided into six main pages:

1. Home Page
2. Overview Analysis
3. Casualties & Driver Insights
4. Accident Severity & Types
5. Geographic Analysis
6. Weather & Road Condition Analysis

---

1 Home Page

The Home Page acts as an entry point.
Its purpose is simply to organize the report and help users move easily between the different sections of the dashboard.
It improves navigation and gives the project a clean, professional flow.

---

2 Overview Page

This page provides a high-level summary of the dataset. It presents the main KPIs and general trends that describe the overall situation.

Main KPIs:

- Total Accidents
- Total Injuries
- Total Fatalities
- Average Severity
- Total Cost
- Accidents by Region
- Accidents by Road Type
- Accidents by Year
- Accidents by Time of Day

Why these visuals were chosen:

- **KPI Cards:** make the most important numbers stand out and give a quick impression before diving into details.
- **Line Chart (Accidents by Year):** best way to show changes over time and spot trends.
- **Bar Chart (Accidents by Region):** easy to compare regions side-by-side.
- **Donut Chart (Road Type):** clearly shows percentage distribution of accidents by type of road.

Page Purpose:
To give the user a complete summary at a glance. This page acts as an executive overview.

---

3 Casualties & Driver Analysis

This section examines how driver-related factors and behaviors contribute to accidents.

Key visuals:

- Accidents by Driver Gender
- Injuries by Gender
- Accidents by Cause
- Pedestrians Involved

Reason for chart types:

- **Bar Charts:** provide a straightforward way to compare categories like male vs. female drivers or different accident causes.
- **Donut Chart:** used where percentage distribution is more meaningful, such as injuries by gender.

Page Purpose:
To understand which groups of drivers are most involved, identify the main causes of accidents, and highlight pedestrian involvement.

---

4 Accident Severity & Type

This page focuses on how serious the accidents are and what types are most frequent.

Main visuals:

- Total Fatalities
- Average Severity
- Total Injuries
- List of Accident Types (Table)
- Fatalities by Time of Day
- Severity by Region

Why these charts were used:

- **Table:** best option for detailed accident-type information because it shows exact values.
- **Donut Chart:** used to show how fatalities are distributed across different times of the day.
- **Area Chart:** effective for illustrating the overall intensity of accident severity across regions.

Page Purpose:
To highlight the most dangerous accident categories and identify regions with more severe incidents.

---

5 Geographic Analysis

This page centers on the spatial distribution of accidents.

Visuals include:

- Fatalities by Country (Map)
- Injuries by Country
- Region vs. Road Type
- Urban vs. Rural Comparison

Why these visuals:

- **Map Chart:** the natural choice for geographic data—it instantly shows hotspots and high-risk zones.
- **Bar/Column Charts:** work well for comparisons between countries or regions.
- **Stacked Column:** helps show the relationship between two dimensions (region + road type).

Page Purpose:
To pinpoint which areas require more attention, showing where accidents cluster and whether urban or rural zones face more risks.

---

6 Weather & Road Conditions

This page studies the environmental side of the dataset.

Metrics analyzed:

- Weather Impact Index
- Accidents by Weather Condition
- Road Condition Risk Levels
- Fatalities by Road Condition

Choice of visuals:

- **Line Chart (Weather Impact):** because it represents changes in impact levels clearly.
- **Column Charts:** best way to compare different weather or road conditions.

Page Purpose:
To understand how the environment influences accident risk and where improvements in road infrastructure might be needed.

*** Conclusion

The dashboard presents a complete analysis of road accidents from multiple angles.
By combining statistical trends, geographic insights, driver behavior, and environmental conditions, the report provides a strong foundation for understanding accident patterns and identifying areas that need intervention.

The structure of the dashboard keeps the information organized, while the chosen visualizations make the insights clear, intuitive, and easy to communicate.