Fundamentals of Information Visualisation (COMP3042) Coursework

Student Name – Atifa Mohamed (20512872)

Chapter 1: Description of Data

The project utilizes the Leeds Road Traffic Accidents Dataset from 2019 which originally comprised 1,907 observations across 18 variables. Variable descriptions were sourced from the STATS20 Manual published by the UK Department for Transport. The table below provides details on the features included in the cleaned dataset only.

Column Name (Feature)	Data Type	Description
Reference Number	Character (Nominal)	The accident reference allocated by the police.
Grid Ref : Easting	Integer (Ordinal)	Represents the eastward coordinate.
Grid Ref : Northing	Integer (Ordinal)	Represents the northward coordinate.
Number of Vehicles	Integer (Ratio)	The number of vehicles involved in the accident.
Accident date	Character (Nominal)	The date of the Accident.
Time (24hr)	Character (Ordinal)	Time of the Accident.
1st Road Class	Character (Nominal)	Classification of the Roads.
Road Surface	Character (Nominal)	Road Surface condition at the time of the accident.
Lighting Conditions	Character (Nominal)	Represents the condition of the streetlights.
Weather Conditions	Character (Nominal)	Weather conditions at the time and location of the accident.
Type of Vehicle	Character (Nominal)	Types of vehicle involved in the accident.
Casualty Class	Character (Nominal)	Types of people involved in the accident.
Casualty Severity	Character (Nominal) , Measure	How severe did the casualty get injured.
Number of Males	Integer (Ratio)	Number of male casualties.
Number of Females	Integer (Ratio)	Number of female casualties.
Average Age	Integer (Ratio)	Average age of casualties at the time of the accident.

Table 1– *Description of the Features*

Chapter 2: Objective and Audience

Objective: To determine how factors such as road surface, lighting and weather conditions, vehicle types and counts, location, date, time, casualty class, and demographic details like gender impact casualty severity in traffic accidents.

Audience: Researchers and analysts interested in historical traffic patterns and accident data, or anyone seeking insights into factors influencing casualty severity, may find this report useful.

Access to Data – The dataset was obtained from data.gov.uk. Published by Leeds City Council under the Open Government Licence.

Chapter 3: Initial Pre-process

Initially, the dataset was explored using the **str()** and **summary()** functions to understand its structure. The **'Time (24hr)**' variable was then converted to a proper time format (POSIXct) for consistency. The **Age of Casualty** column was converted to numeric to ensure proper calculations.

Next, a summary dataset (**summary_data**) was created by grouping the data by **Reference Number**, summing the number of males and females, and calculating the average age of casualties for each reference number. The **Local Authority** column was removed from the dataset because it had the same value throughout, which was **E08000035**, representing Leeds. The **Vehicle Number** column was also removed as it would not impact the analysis; it is simply a vehicle reference number, and the dataset already includes the number of vehicles involved. No observations were removed.

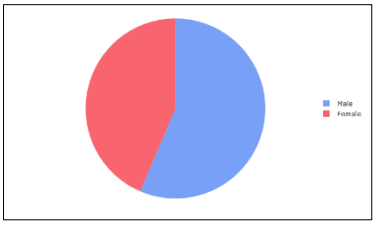
The cleaned data was joined with the summary data based on **Reference Numbe**r to include the aggregated information for each reference. Categorical variables such as **Road Surface**, **Lighting Conditions**, **Weather Conditions**, **Type of Vehicle**, **Casualty Class**, and **Casualty Severity** were recoded into more readable formats using the **dplyr::recode()** function.

Finally, the cleaned and transformed dataset was saved as a CSV file to be used for visualizing traffic accident trends and creating an interactive dashboard using Shiny.

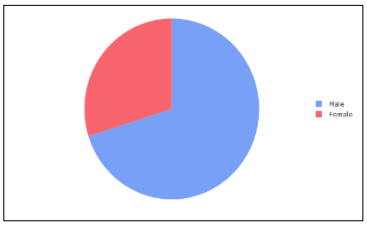
Chapter 4: Initial Questions

4.1 **Question**: What is the distribution of casualties by gender and their average age for each severity?

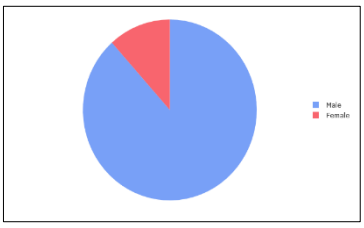
A subset of the **filtered dataset** was used to calculate the **total number of male** and **female casualties**. The counts were reshaped into a **long format** to categorize the data by gender. An **interactive pie chart** was created using. The chart includes **hover functionality** to display the respective **casualty count**, with distinct colors assigned for easy differentiation. Additionally, a **text element** was created to display the **average age** of casualties for males and females.



A) Slight



B) Serious



C) Fatal

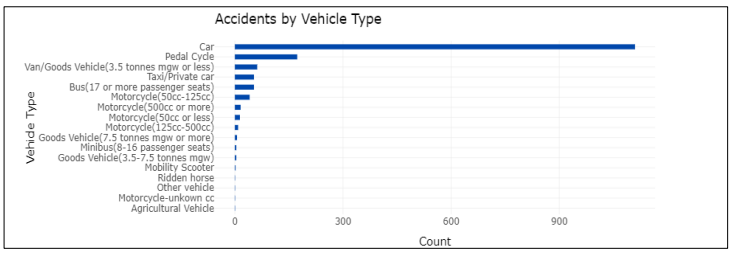
***Figure 1**: Pie Chart (Gender Distribution)*

Reason for visualization – The pie chart provides an intuitive way to compare the proportion of male and female casualties in traffic accidents. This type of visualization is ideal for showing **categorical data** as parts of a whole, making it easy to understand the gender distribution at a glance. Additionally, the HTML() function is used to display **avg_age_male** and **avg_age_female.**

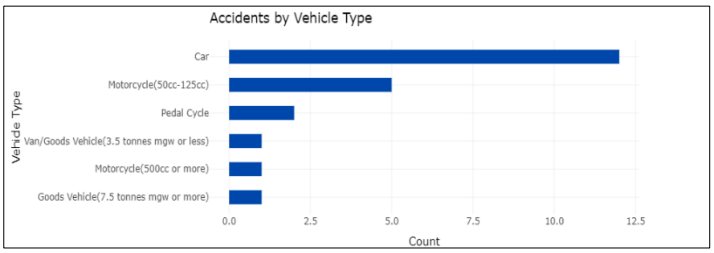
Answer - The gender distribution and average age of casualties across accident severities reveal key trends. In slight severity accidents, males (1,609) outnumber females (1,244), with similar average ages (36 for females, 35 for males), suggesting **younger adults** are more involved. In serious accidents, males (385) are more affected than females (164), with an **increase** in average age (41 for females, 37 for males). In fatal accidents, the **gender gap widens** significantly (38 males, 5 females), with females being older (66 years) than males (39 years). These findings show a **consistent male overrepresentation**, particularly in fatal accidents, with age influencing severity.

4.2 **Question**: Which vehicle types are most commonly involved in traffic accidents, and how does their involvement impact casualty severity?

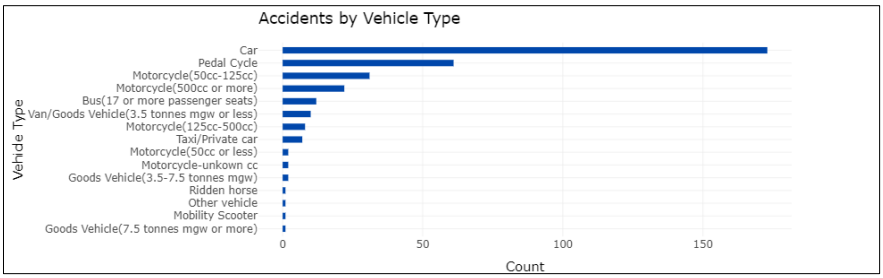
First, the data is grouped by the **Type of Vehicle** column, and the count of each type is calculated and sorted in descending order. A bar chart is then created using **ggplot2**, with each bar representing a vehicle type and its respective accident count.



A) Slight



B) Fatal



C) Serious

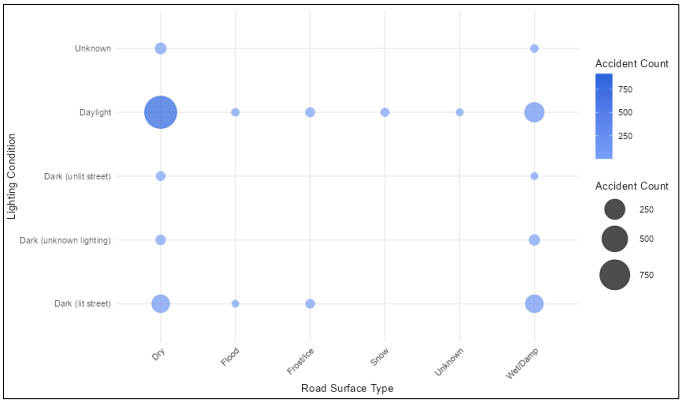
***Figure 2**: Bar Chart (Types of Vehicles)*

Reason for visualization – The bar chart provides a clear way to compare the **frequency of accidents** across various vehicle types. Its interactive nature allows users to explore the exact counts for each category using the **hover** functionality. It helps in uncovering patterns and prioritize **vehicle-specific safety** measures.

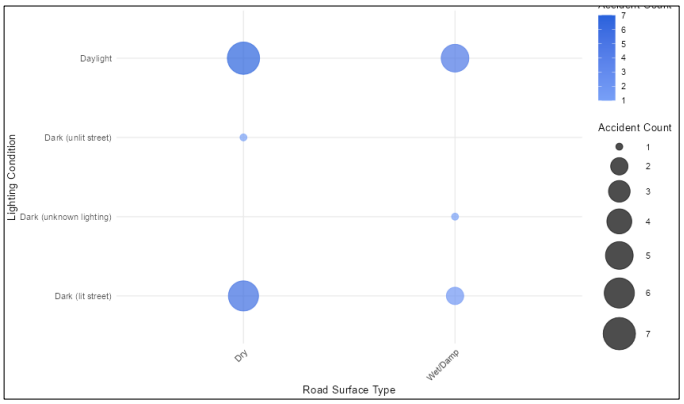
Answer - This visualization shows the distribution of accidents by vehicle type across severity levels. Cars are the most common vehicle involved in all severities, with **the highest counts** in slight severity accidents, followed by serious and fatal cases. In slight severity, cars are followed by pedal cycles and vans. In serious severity, pedal cycles and motorcycles (50cc–125cc) are the next **most common**. In fatal accidents, motorcycles (50cc–125cc) follow cars. Other vehicle types, such as buses, vans, and large motorcycles, contribute **fewer cases**. This highlights the dominance of **cars** and **two-wheelers** in accidents, with varying severity.

4.3 **Question**: How do road surface types and lighting conditions affect accident frequency, and how do they contribute to the impact on casualty severity?

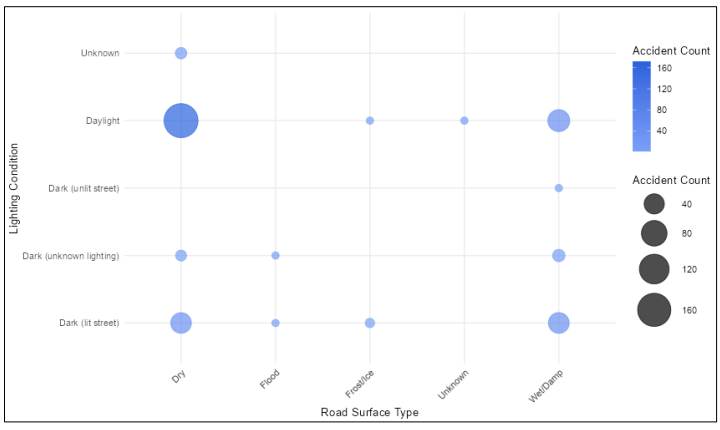
A bubble plot that visualizes the relationship between **road surface types and lighting conditions** is created to show how they affect accident frequency. It groups the filtered data by Road Surface and Lighting Conditions and calculates the **count of accidents** for each combination. Using **ggplot2**, it creates a **scatter plot** where bubble sizes and colors represent accident frequencies, allowing for easy comparison of high-risk scenarios.



A) Slight



B) Fatal



C) Serious

***Figure 3**: Bubble Plot (Road Surface Types and Lighting Conditions)*

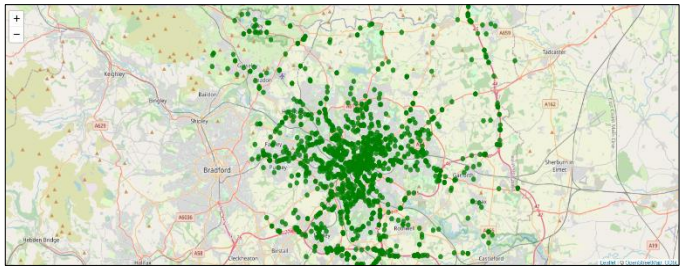
Reason for visualization – The bubble plot effectively visualizes the interplay between road surface types, lighting conditions, and accident frequencies. It allows viewers to identify **high-risk combinations**, such as specific lighting conditions contributing to more accidents on

certain road surfaces. The size and color of the bubbles provide clear visual cues for accident severity or frequency, making it easier to detect patterns and prioritize safety interventions.

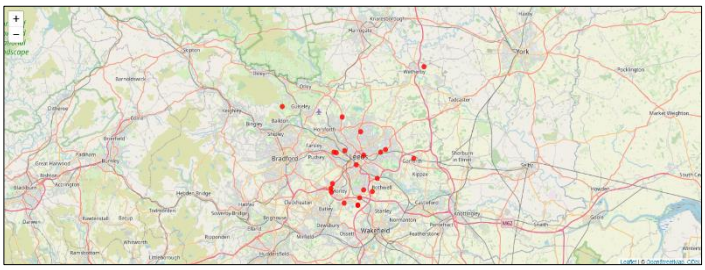
Answer - In all severities, **daylight and dry road conditions** account for the highest number of accidents, as indicated by the largest bubbles, reflecting the prevalence of high traffic volumes and normal driving conditions. **Wet/damp roads** under daylight consistently rank second, highlighting the increased risks associated with reduced traction and longer braking distances. **Dark (lit street)** conditions combined with **dry roads** also show notable accident frequencies across all severities, pointing to persistent visibility challenges even with artificial lighting. In contrast, adverse road surface conditions like **frost/ice**, **flood**, or **snow** contribute minimally to accidents of any severity. **Unknown road surface** or **lighting conditions** show moderate accident counts, potentially reflecting gaps in data reporting. Overall, the charts emphasize that most accidents daylight and dry roads.

4.4 **Question**: How does the location of accidents influence the severity of traffic accidents?

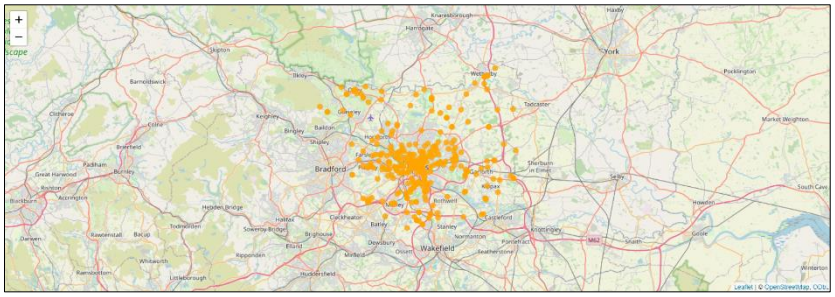
It processes UK Easting and Northing coordinates into **standard Latitude-Longitude** using spatial transformations, making them usable for mapping. The **transformed coordinates** are merged with the original dataset to include location data for each accident. A **Leaflet map** is rendered, displaying **accident locations**.



A) Slight



B) Fatal



C) Serious

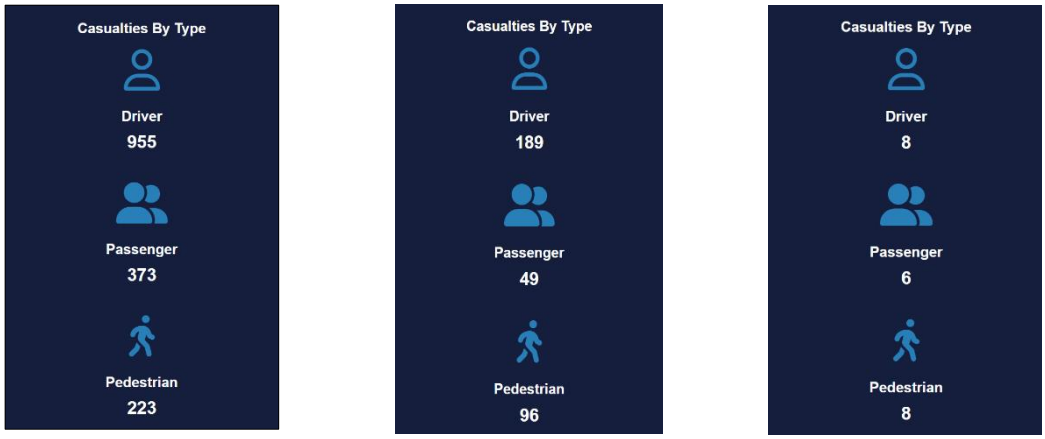
***Figure 4**: Location Plot*

Reason for visualization – By mapping accidents with distinct colors for severity levels, stakeholders can quickly pinpoint **high-risk areas**.

Answer - Slight-severity accidents show **dense clustering** in urban or high-traffic areas, particularly in central regions. Serious-severity accidents also tend to cluster in urban areas but with **less density** compared to slight-severity cases. Fatal accidents, being the **rarest**, are sparsely distributed across the region, with **no significant clustering** observed. **Urban areas** consistently emerge as hotspots for accidents of all severities, highlighting the impact of traffic density and infrastructure on accident occurrences.

4.5 **Question:** How does the distribution of casualties among drivers, passengers, and pedestrians vary across different accident severities?

Calculated and displayed the number of casualties for each category: **Driver/Rider**, **Vehicle Passenger**, and **Pedestrian** based on the filtered dataset. For each category, the dataset is filtered to match the respective casualty class, the rows are counted using nrow(), and the results are output as text in the UI.



A) Slight *B) Serious* *C) Fatal*

***Figure 5**: Casualties by Type*

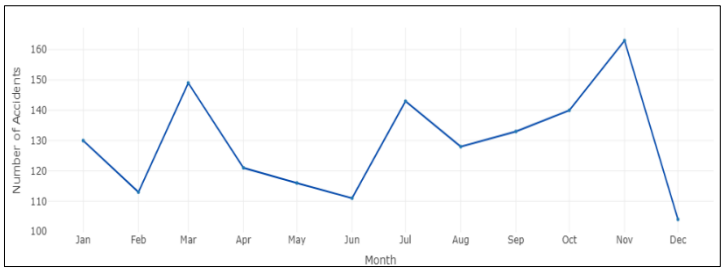
Reason for visualization – The visualization of casualty counts by category serves to highlight the distribution of those most affected in accidents. It provides critical insights into which groups are more frequently involved, allowing for targeted interventions and safety measures.

Answer - For slight-severity cases, drivers form **the largest group** due to constant road exposure, followed by passengers and pedestrians. In serious-severity cases, drivers remain the largest group, but pedestrians face a **higher proportion** of injuries due to their **direct exposure** in collisions, unlike the relatively protected passengers. For fatal-severity cases, drivers and pedestrians each account for 8 fatalities, highlighting their **vulnerability**, with passengers **slightly fewer**. These findings stress the need for improved safety measures.

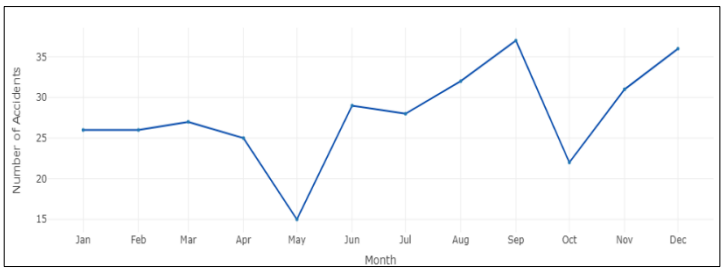
Chapter 5: Further Questions

5.1 **Question:** How do monthly variations in accident counts influence the severity of casualties?

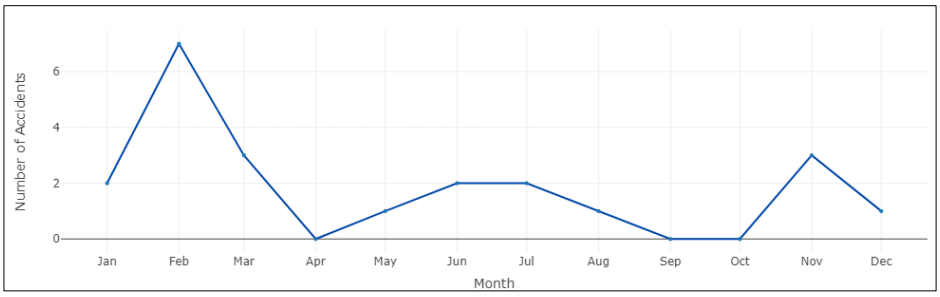
Generated an **interactive line chart** in Plotly to visualize the **monthly trend** of accident counts. It processes the dataset by converting the Accident Date column to a **date format** and **grouping** data by **year** and **month** to count the **number of accidents**. **Missing months** are handled by creating a complete sequence of months and **filling gaps** with zero counts. The processed data is then visualized as a line chart.



A)Slight



B)Serious



C) Fatal

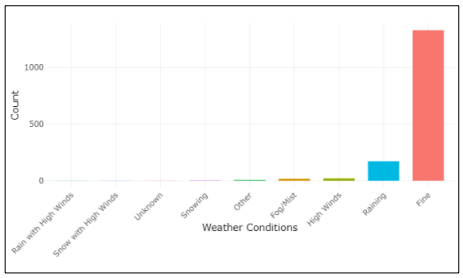
***Figure 6**: Line Chart (Monthly Trend)*

Reason for visualization – The purpose of this visualization is to analyze the trend of accident occurrences over time, helping to identify patterns in accident frequency by month. By displaying the data in a clear and interactive line chart, it enables stakeholders to pinpoint periods with **higher accident rates**, such as seasonal spikes, and investigate their potential impact on accident severity.

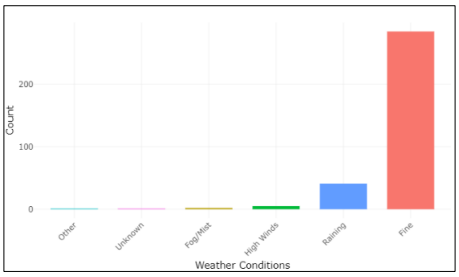
Answer - The monthly trends for accidents of different severities reveal distinct patterns that highlight seasonal variations and potential risk factors. Slight severity accidents are relatively consistent throughout the year, with **notable peaks** in **March** and **November**. Serious severity accidents show a gradual **increase** from **May to September**, with **peaks** in **late summer**. Fatal severity accidents, while **infrequent**, **spike dramatically** in **February**, then **decline** significantly through the **mid-year months**. Together, these trends underscore the importance of targeted safety measures during high-risk months, particularly February for fatal accidents and late summer and autumn for serious and slight accidents.

5.2 **Question:** How do weather conditions influence the severity of traffic accidents?

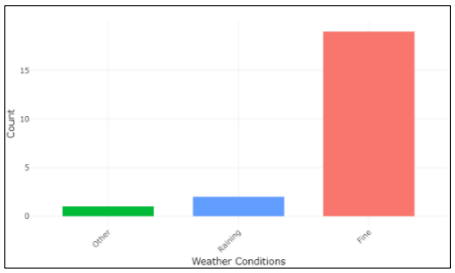
A **bar chart** showing accident counts by **weather conditions** is generated. It groups and summarizes the data, visualizes it using **ggplot** with styled axes and minimal design, and converts it into an interactive Plotly chart with tooltips for accident count



A) Slight



B) Serious



C) Fatal

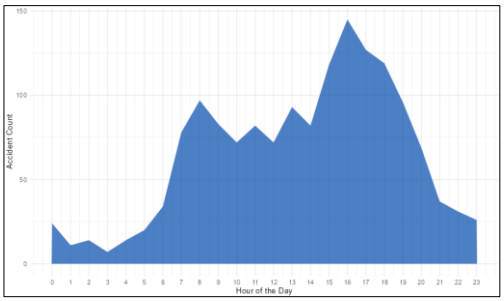
***Figure 7**: Bar Chart (Weather Conditions)*

Reason for visualization – This helps assess whether certain weather conditions contribute more to accident occurrences, providing insights for traffic safety measures and planning.

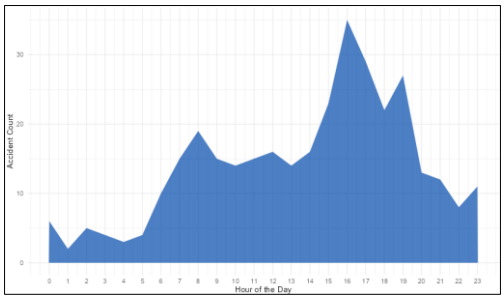
Answer : Slight-severity accidents occur most frequently during **fine weather**, with rain and adverse conditions playing a minimal role. Serious-severity accidents, though less common, show a more notable presence of **rainy conditions**, suggesting that wet roads may increase severity. Fatal accidents, while **rare**, also predominantly occur during fine weather, with minimal influence from adverse weather. Overall, while accidents are most common in fine weather, the frequency and severity tend to increase under rainy conditions. Other conditions are very rare across all severity cases.

5.3 **Question:** What are the peak times for severe accidents to occur, and how does severity vary by time of day?

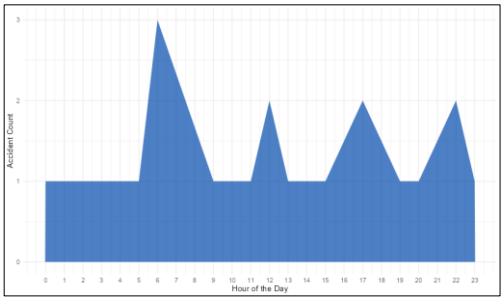
Generated an **area plot** showing the **count of accidents** by **hour of the day**. It first processes the filtered dataset, converting the 'Time (24hr)' column to a **character type**, extracting the hour as a **numeric value**, and **grouping** the **data** by the hour to calculate the accident count. The **ggplot** function then creates the plot with hours on the x-axis and accident counts on the y-axis, using **geom_area** to fill the area under the curve.



A) Slight



B) Serious



C) Fatal

***Figure 8**: Area Plot (Accidents by Hours)*

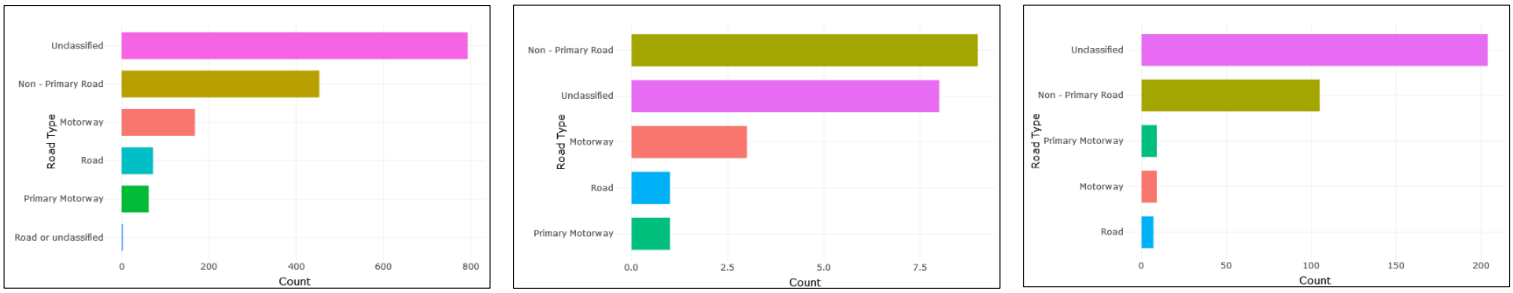
Reason for visualization – The visualization aims to provide insights into the distribution of traffic accidents across different hours of the day. By **summarizing** accident counts by hour, it helps identify **peak times** for accidents to guide traffic management and safety policies

Answer - Analyzing the distribution of traffic accidents by severity across different hours of the day reveals distinct patterns. For accidents with **slight severity**, the frequency is higher during daytime, peaking between 3 PM and 4 PM with a secondary rise during the morning rush hour

(7 AM - 8 AM). Accidents with **serious severity** follow a similar pattern, showing increased occurrences during morning (7 AM - 9 AM) and evening rush hours (4 PM - 5 PM. In contrast, **fatal accidents** occur less frequently overall but peak sharply early in the morning (6 AM - 7 AM) and again during the evening (5 PM - 7 PM. These observations highlight the need for targeted safety measures especially during high-risk hours.

5.4 **Question:** On which types of roads do accidents most frequently occur, and how do these road types influence the severity of the accidents?

Generated an **interactive bar chart** showing the number of accidents for each road type. The dataset is **grouped** by **1st Road Class**, and the accident counts (Count) are summarized and ordered in **descending order**. A horizontal bar chart is created using **ggplot2**, with **road types** on the y-axis and **accident counts** on the x-axis, styled minimally with hidden legends.



A) Slight *B) Serious* *C) Fatal*

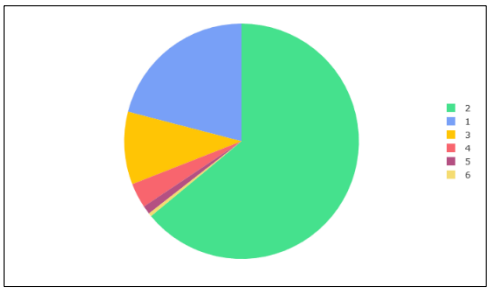
***Figure 9**: Bar Chart (Types of Road)*

Reason for visualization – The visualization helps identify which road types experience the **highest number of accidents**, providing insights into patterns for road safety concerns. By making the chart interactive with **hover functionality**, users can explore specific accident counts for each road type, enabling a more detailed analysis.

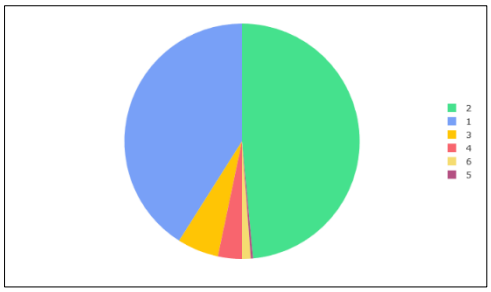
Answer – The analysis of accident severity across road types reveals distinct patterns. For slight severity, unclassified roads **dominate,** followed by non-primary roads, with **minimal contributions** from motorways. In serious severity, unclassified and non-primary roads are **most common**. However, for fatal accidents, non-primary roads **lead**, followed by unclassified roads, with motorways showing a higher contribution due to **higher speeds**. Primary motorways and standard roads have **minimal involvement** across all severities. These findings highlight the need for enhanced safety measures, especially on non-primary and unclassified roads.

5.5 **Question:** How does the number of vehicles involved vary with the severity of accidents?

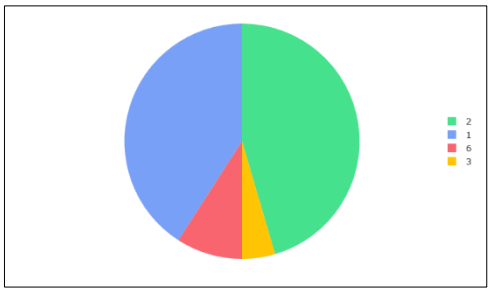
Generated an **interactive pie chart** to visualize the **number of vehicles** involved in accidents. It processes the filtered data by counting the occurrences of each vehicle count and **renaming columns** for clarity.



A) Slight



B) Serious



C) Fatal

***Figure 10**: Pie Chart (Number of Vehicles)*

Reason for visualization –By using a pie chart, the proportions of vehicle counts can be easily compared to understand whether severe accidents are more likely to involve single vehicles or multiple vehicles. The **hover functionality** provides precise counts for better insights.

Answer - The pie charts illustrate the distribution of vehicles involved in slight, serious, and fatal severity incidents across various categories. In all three cases, **two vehicles** consistently account for the **largest share**, indicating their predominant involvement across all severity levels. **Single vehicles** follow as the **second-largest segment** in each chart, representing a significant but smaller proportion compared to two vehicles. The **remaining categories** collectively occupy much **smaller portions**, highlighting their minimal contribution to these incidents. This highlights that the majority of incidents involve **fewer vehicles**.

Chapter 6: Reflection

Through this coursework, I gained a deeper understanding of R with the help of lectures and labs, building on the foundational knowledge I had previously acquired in my math modules. I was able to explore R in greater depth, particularly in the areas of data visualization and interactive dashboard creation.

Initially, finding a suitable dataset was challenging, but after successfully cleaning and analyzing it, I utilized the dplyr package for data manipulation and transformation. Posing the right questions to guide my analysis of the dataset was also a tricky task. I employed ggplot2 and Plotly to create various graphs, which were then combined on a single page to form an interactive dashboard using Shiny. I arranged the plots according to their importance, ensuring that users could quickly gain an understanding based on the selected severity level. Additionally, I incorporated CSS elements to enhance the aesthetics, making the dashboard both functional and visually appealing. The dataset provided most of the necessary features, though including additional factors such as speed, breath test results, or hazards involved could have further enriched the analysis.

Thanks to the support from my lecturers, Dr. Marina and Dr. Hafeez. I am grateful for the opportunity which I know will be valuable in my future career.