FIT5147 – Data Exploration & Visualisation

Programming Exercise 1: Tableau (5%)

Student Name: Sanchita Reddy

Student ID: 33483086

Tutor: Mohit Gupta

Applied class: 06.

Loading of the data in Tableau:

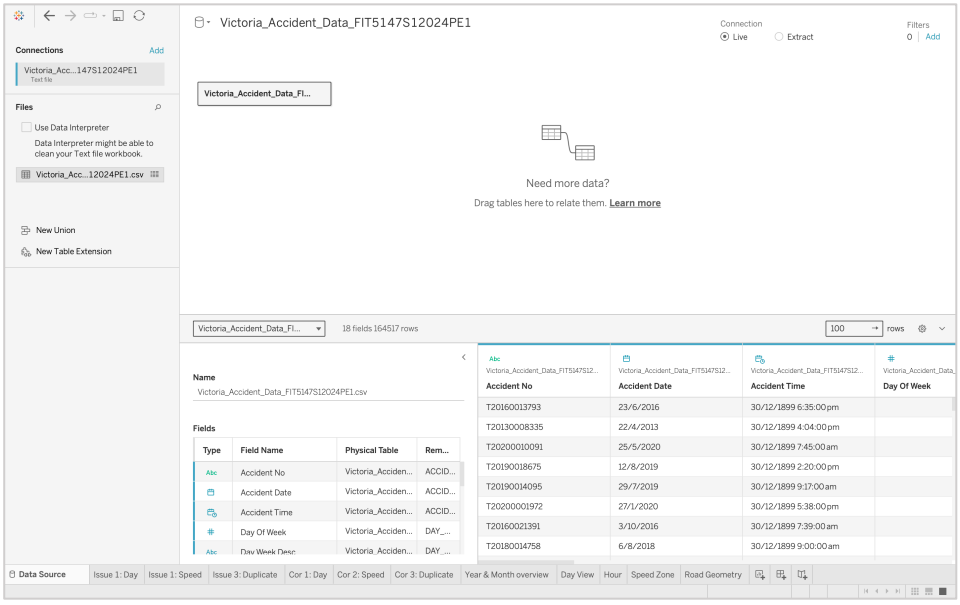

Image1: Data (Victoria_Accident_Data_FIT5147S12024PE1 data set) loaded in Tableau.

The image displayed above depicts that the provided dataset has been successfully uploaded to Tableau. To ensure the accuracy of the data, crucial steps were followed, such as **adjusting the data types**, for instance, the "Accident No." attribute was converted to a string, "Accident Time" was loaded as Date and Time type, etc.

Task 1 – Data checking and cleaning

Error 1: Inconsistency with 'DAY_OF_WEEK' and 'DAY_WEEK_DESC'

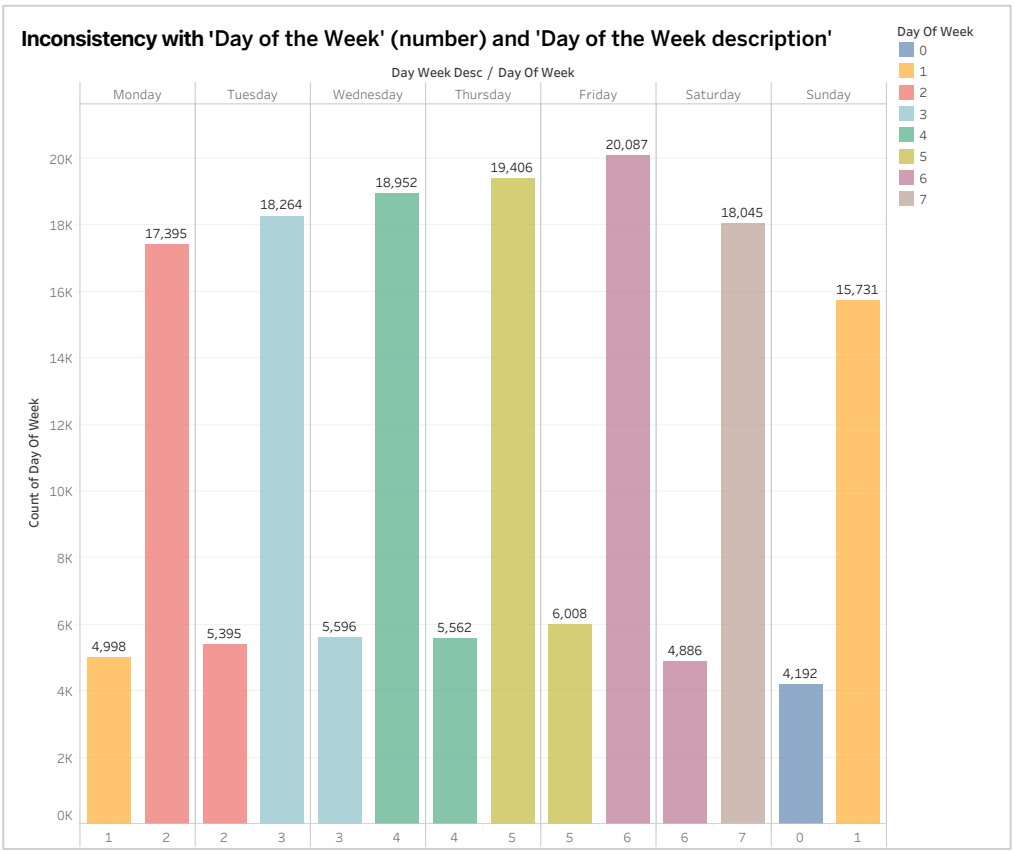

Image 2: Error 1 – Comparison between the Day of the week (description) and the day of the week (number)

In the visualisation above, the x-axis represents the days of the week, and each day has two numbers assigned to it. This indicates that **a day has multiple numerical identifiers**, which violates the principle that each day should have a unique identifier. Additionally, the numbers range from 0 to 7, suggesting that there are eight days in a week. This issue was addressed in Excel and a unique number was allocated each day starting from Sunday (*The Seven-Day Week*, 2019).

Error 2: Inconsistency with 'DAY_OF_WEEK' and 'DAY_WEEK_DESC'

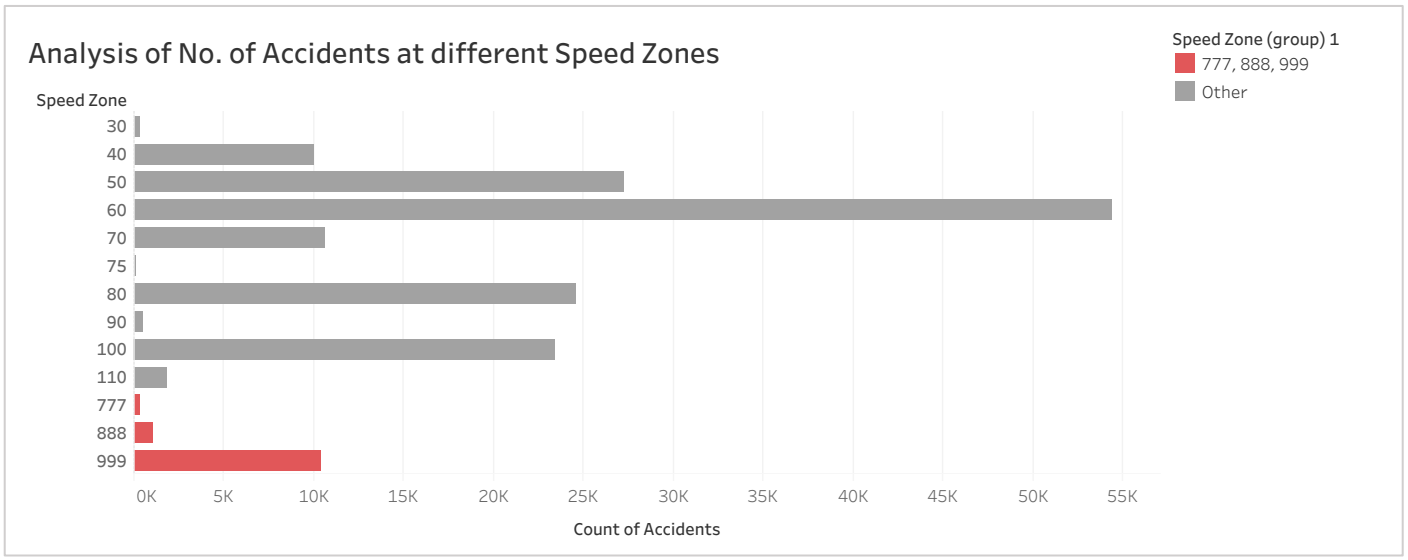

Image 3: Error 2 – Analysing the speed zones used as bins for Victoria.

The dataset has a second error caused by **illogical speed zones**. The y-axis displays the speed zones, and the x-axis shows the number of accidents. The image above highlights three speed zones, 777, 888, and 999, which fall outside the normal range. These values are unrealistic and not prescribed by the Victorian government (Speed Limits - City of Melbourne, n.d.). To fix the error, the bucketing feature in Excel was used. All accidents that occurred at high speeds were adjusted to indicate that they happened at the highest speed zone, i.e., 110 km/hr.

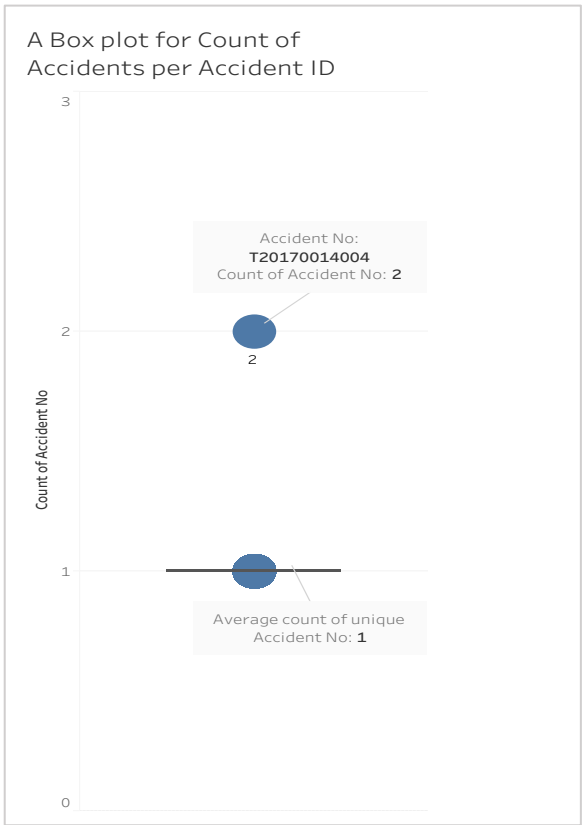

Image 4: A box plot to identify duplicate key values in the data.

Error 3: The database key, 'Accident Number' is not unique.

The image 4 adjacent displays a record of accidents with their unique identification numbers. It is important to note that **each ID is distinct and cannot be duplicated** as per the metadata. However, upon closer inspection, it was discovered that one of the accident IDs had been repeated, which violates the integrity of the database. Upon further investigation, we found that all the data fields related to this accident record were identical and redundant. Consequently, we removed the duplicated row using the "delete duplicates" feature in Excel.

Task 2 – Data Exploration using Tableau

Q1 – Exploring when and what types of accidents occur over different periods.

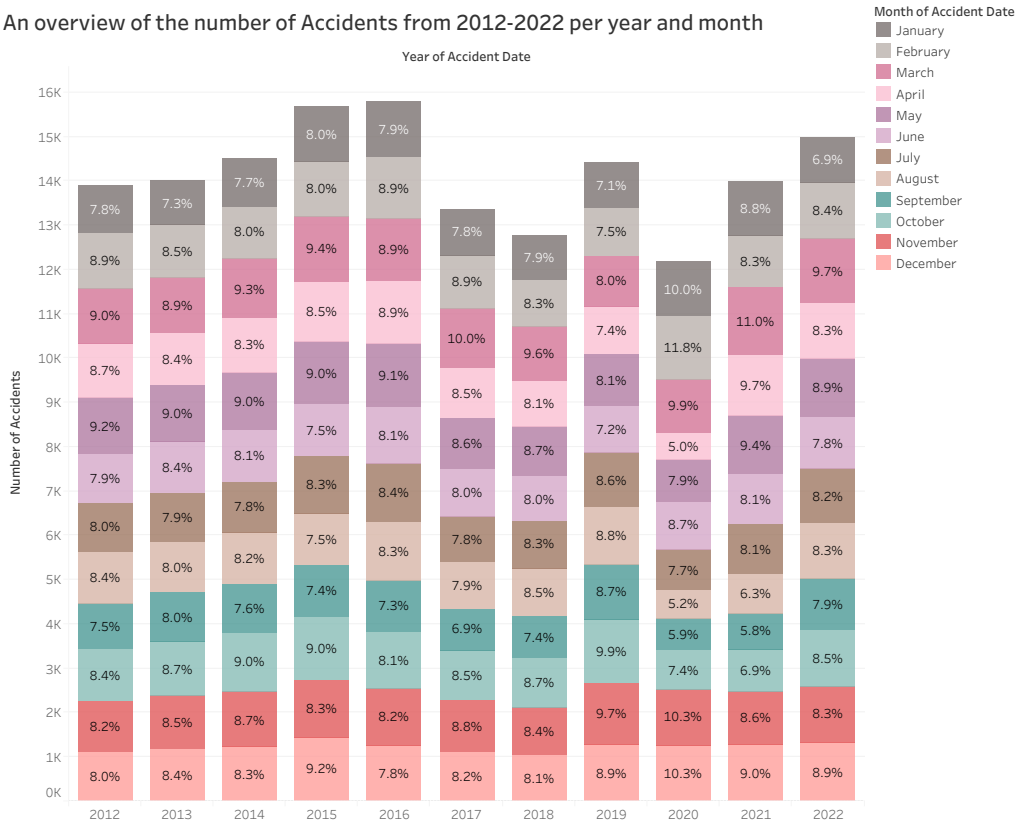

Image 5: A stacked bar chart indicating the number of accidents by year from 2012-2022 and by month.

The bar chart shows the number of accidents from 2012 to 2022. The partial data in the dataset for the year 2023 was excluded to avoid incomplete data. From the visualisation, we observe that the average number of accidents a year is ~13k, and the highest number of accidents was noticed in 2016. The year 2020 had the lowest number of accidents (~12k), but it was not too distant from the average. The decline in 2020 could be attributed to lesser populations outdoors due to a pandemic. Recently, the frequency of accidents has increased, being about 15k in 2022. The months of **April to June** see the highest number of accidents in a year. In these months, Victoria has warm weather and high rainfall (Victoria, n.d.), resulting in high traffic and low visibility on roads.

A stacked bar chart is a useful tool for **comparing data**, which was the goal in the above case. The visualisation helped to compare the number of accidents over the year and highlight the trend of which month sees the most accidents.

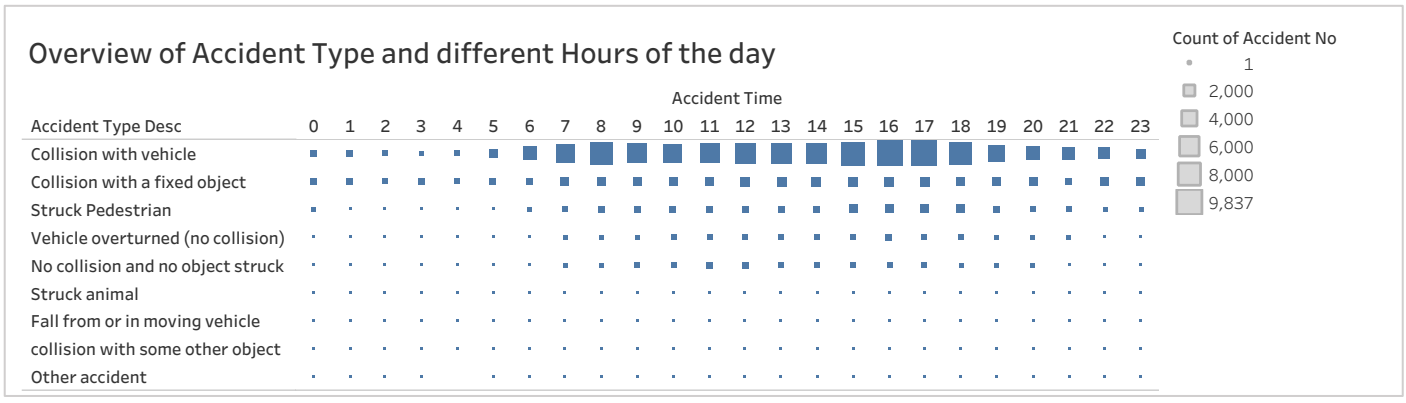

Image 6: A heat map depicting the different types of Accidents and their frequencies by hour.

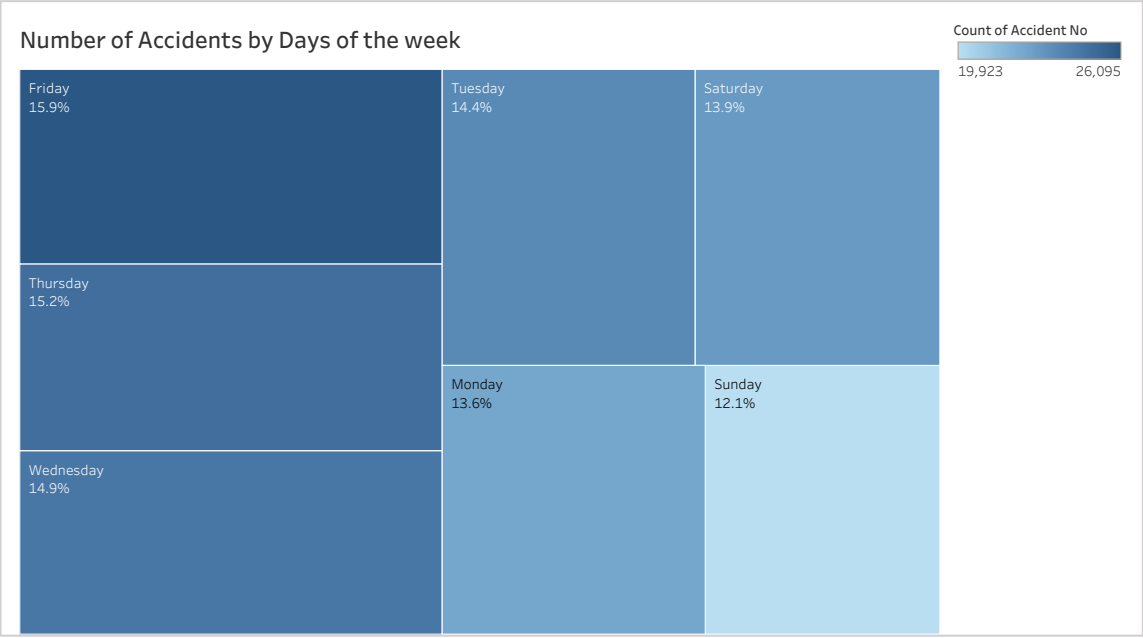

Image 7: A tree map for the number of accidents occurring on different days of the week.

Images 6 and 7 reveal key accident patterns. The heatmap (Image 6) highlights peak hours (7 am to 7 pm) as accident hotspots, likely due to more vehicles on the road and rushed drivers. It can also be noticed that Collision is the main driver of these accidents, followed by other issues. The tree map (Image 7) shows weekdays, especially Fridays, have the most accidents, likely due to higher traffic volume. Research conducted by the traffic institution indicates that the eagerness to start the weekend leads to more accidents on weekdays (Motorists Crash More on Fridays, 2016).

For Image 6, a heat map was chosen since it excels at revealing variations and patterns in data across 2 dimensions. The variation in size helps identify high-concentration and low-concentration areas. They are also great at effectively visualising relationships between two categorical variables. On the other hand, a tree map is ideal for hierarchical data like the frequency of accidents in this case. The colour gradient and nested triangles allow for efficient comparison of the proportion between different categories.

Q2 – Exploring the Geometry of Accidents.

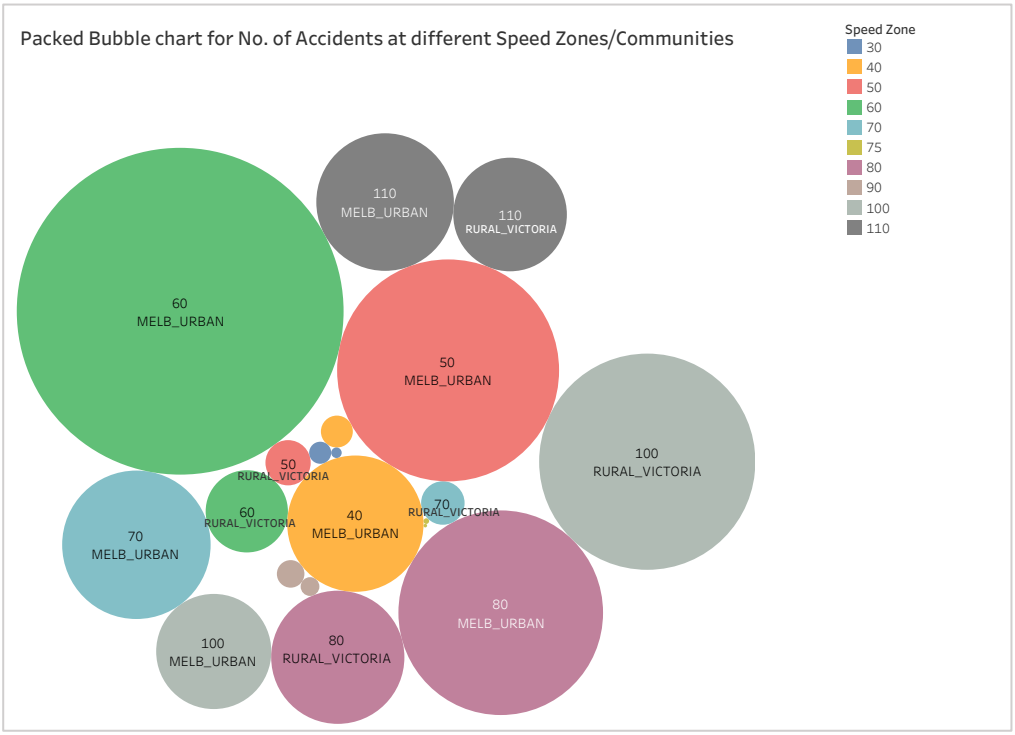

Image 8: A bubble map to show the number of accidents at different speed zones and communities (Urban/Rural)

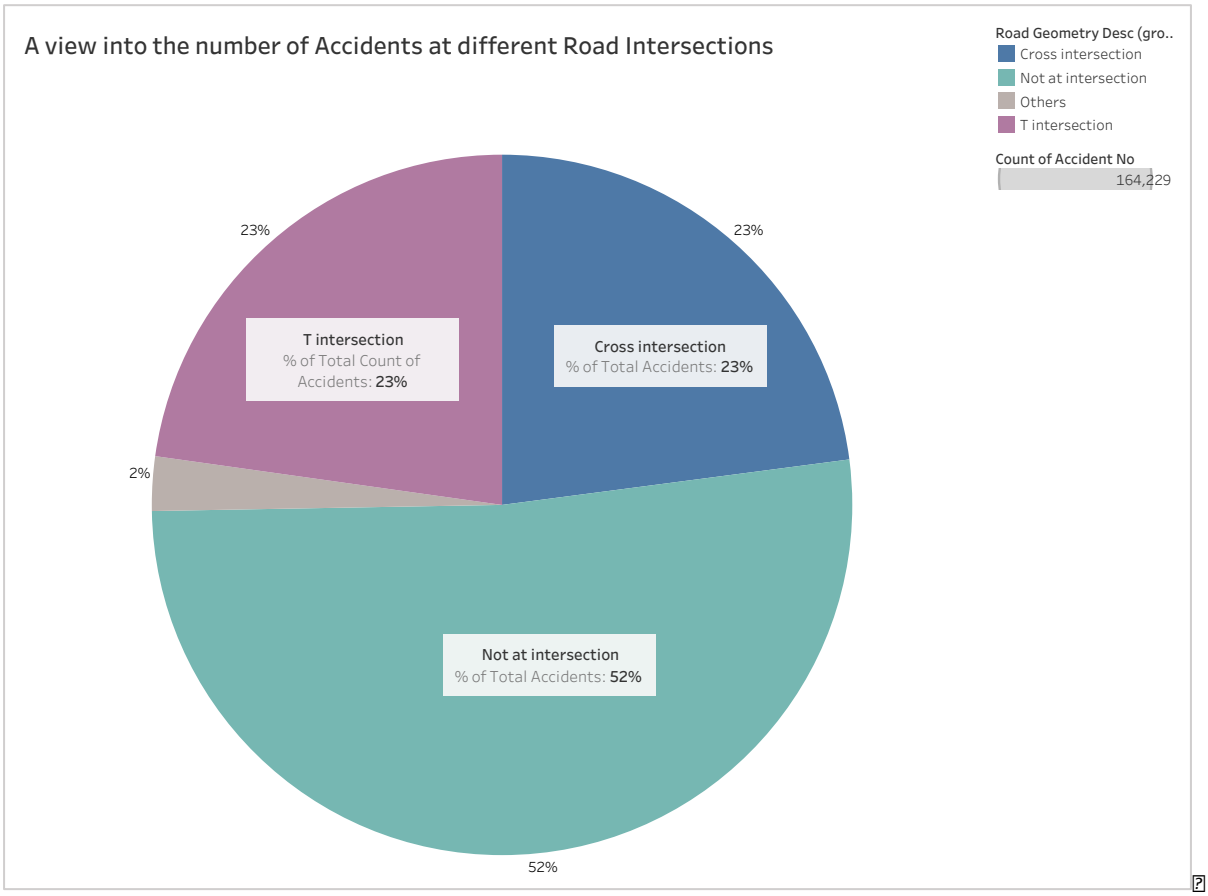

Image 9: A pie chart representing the proportion of accidents at different kinds of intersections.

Images 8 & 9 show **accident locations**. The bubble map has been filtered to understand accident patterns between the rural and urban areas of Victoria. The image highlights urban areas have high accidents, especially in **60km/h zones**. Interestingly, the bubble chart reveals **11% of accidents in rural 100km/h zones**, suggesting potential weekend travel risks. From this visualisation, we understand that more accidents are occurring in mediocre to high-speed zones (50 – 100 km/hr) and mainly in the urban regions of Victoria.

Finally, 52% of accidents occur outside intersections, i.e., on straight roads. This emphasises the need for broader data like car breakdowns and a focus on safe driving habits (since these are mostly collisions). However, 48% of the accidents are still happening at an intersection (T Intersection & Cross Intersection) and this drives a need for organised traffic and safety at intersections.

Unveiling accident trends requires the right tools. Packed bubble charts excel at analysing multiple factors simultaneously. By using both size and colour, they effectively convey the impact of speed zones and communities on accident rates. Pie charts, on the other hand, are effective in visualising proportions and enabling clear comparisons between different accident categories. In both cases, using different colours and not using colour hue highlights different categories, while size reflects the proportions.

In conclusion, a significant number of accidents occur in urban areas, particularly at intersections and within speed zones commonly found in Victoria's cities. This data supports the initial theory that accidents are mainly caused by high volumes of traffic during peak times in urban areas. This highlights the need for improved traffic management strategies and safety protocols to prevent vehicle collisions.

References:

Motorists crash more on Fridays. (2016, October 2). Drive. https://www.drive.com.au/news/motorists-crash-more-on-fridays-20120510-1yf58/

Speed limits - City of Melbourne. (n.d.). Www.melbourne.vic.gov.au. https://www.melbourne.vic.gov.au/parking-and-transport/roads/road-safety/Pages/city-speed-limits.aspx

The Seven-Day Week. (2019). Timeanddate.com. https://www.timeanddate.com/calendar/days/

Victoria, c=au; o=Visit. (n.d.). *Weather, Travel information, Victoria, Australia*. Visit Victoria Home. https://www.visitmelbourne.com/practical-information/melbourne-weather#:~:text=Melbourne%20enjoys%20warm%20summers%2C%20glorious