Title : Data Exploration Project

Student name - Dishi Jain

Student ID - 30759307

Tutorial Number - 24

Tutor Name - Pratik

Table Of Contents

1. Introduction

This data exploration project focuses on the information about the crashes on the roads of Victoria for the last couple of years. The data is spread across different categories giving details about the seriousness of the crashes i.e. whether it leads to a fatality or not. The data also contains the dates, times and location of the accidents, hence providing us with an option to explore the geographical aspects too.

The questions that we will focus on are -

a. Number of accidents on the basis of the location(using latitude and longitude), year,days of the week and region wise individually.
b. Crashes based on the severity and the geometry of the road hence giving a relationship between the two.
c. Number of fatalities according to the region of the crashes.

These questions will allow us to explore the data well and answer appropriately using the different tools and graphs. While answering these questions, we'll not only limit ourselves to answering them but will also try to explore something out of the box.
The motivation of picking this topic came to me when I read around 4-5 road accident news in Victoria on the same day. It caught my eye and had me thinking about the reason for these accidents, any relationship between the accidents and so on. Hence I picked up this topic.

2. Data Wrangling

a. The data has been taken from the site - https://vicroadsopendata-vicroadsmaps.opendata.arcgis.com/datasets/crashes-last-five-years
The data has columns of different types including field types as text and number both. There are in total 65 columns and 74908 rows. The geographic extent of the data is for Victoria only. Many different locations are recorded with the help of latitude and longitude and geometry of the roads is also recorded.
b. Data wrangling includes the transformation and cleaning of the data. Using R as a tool we have removed the unwanted columns. Reading the dataset into R as a data frame using read.csv and then removing certain columns by their names using list(NULL). These columns were not needed to answer the questions and were irrelevant in data exploration. The columns that have been deleted are "OBJECTID", "ACCIDENNT_NO", "ABS_CODE", "NDE_ID", "SRNS", "SRNS_ALL", "RMA", "RMA_ALL", "DIVIDED", "DIVIDED_ALL". The data is in a tabular format and requires no transformation. After removing the columns the data is ready for further exploration.
c. Another issue with the dataset is that some columns are repeated twice, for example the latitude and longitude columns are given two times each as **X and Longitude columns**, **Y and Latitude columns.** Hence removing the columns X and Y and keeping only Longitude and Latitude is ideal. This has been done again by using list(NULL) just as before.
d. Another transformation that has been made is converting the data type of the field "ACCIDENT_TYPE" and changing it into Date format. This is useful as it is needed in

answering the further questions. This formatting has been done in R using format(as.Date()) and providing as arguments the column name and format needed. data$ACCIDENT_DATE <- format(as.Date(data$ACCIDENT_DATE, format = "%d/%m/%Y"), "%Y-%m-%d")

e. Two columns are also added in the dataframe after reading the data from the csv file, which are “YEAR_OF_ACCIDENT” and “MONTH_OF_ACCIDENT”. These two fields are required in further answering of questions and hence have been added by using the ACCIDENT_DATE column.

f. Another Wrangling that has been done was checking the data for the years. It was noticed that for the years 2013 and 2019 there was not sufficient data given. For these two years the data that is provided is only for a couple of months. Hence the rows for these years have been deleted to avoid incorrect visualisations for later on. The rows have been deleted using data <- data[!(data$YEAR_OF_ACCIDENT==2013),] AND , data <- data[!(data$YEAR_OF_ACCIDENT==2019),] in R.

3. Data Checking

While exploring the data columns and its values we’ve found some errors in the data.

a. **Error 1 -**

Missing values in the column “Day_Of_Week”.

While looking at the column Day_Of_Week, we’ve found there were some null values. Plotting the data using a bar graph in R the plot looks as shown below. It shows that there are 1361 Null Values in the column. This can be seen in Figure 1.

Correcting the error -

The error can be corrected by filling the days of the week based on the dates given of the accident. Using - data$DAY_OF_WEEK <- weekdays(as.Date(data$ACCIDENT_DATE)) we can fill the null values of the days. After correcting this error the bar graph now looks like as shown in Figure 2.

While creating this bar graph, it was also noticed that some entries of the column DAY_OF_WEEK were incorrect based on the accident date. These have also been corrected while filling the null values of the column DAY_OF_WEEK.

Figure 1 - Number of accidents by days of the week

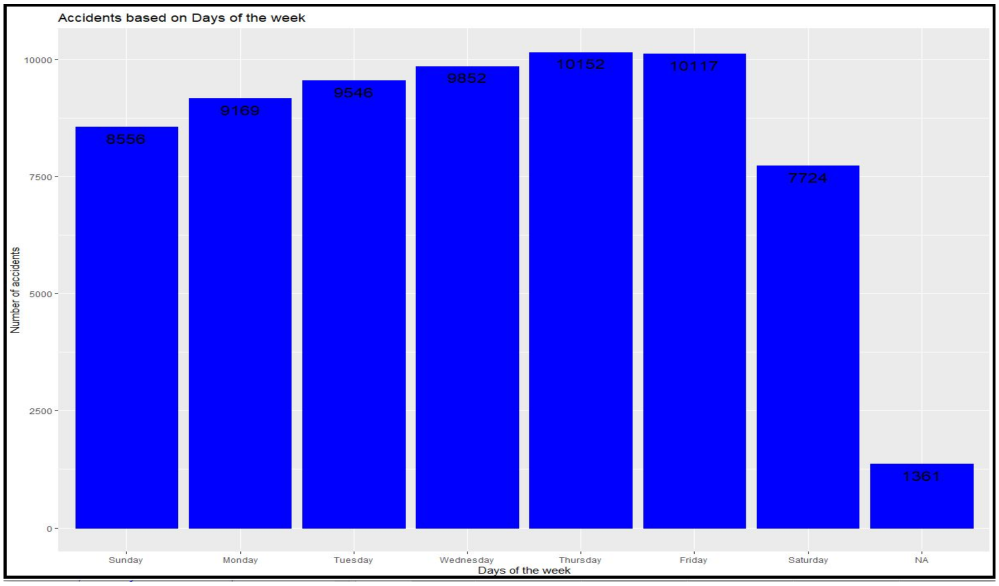
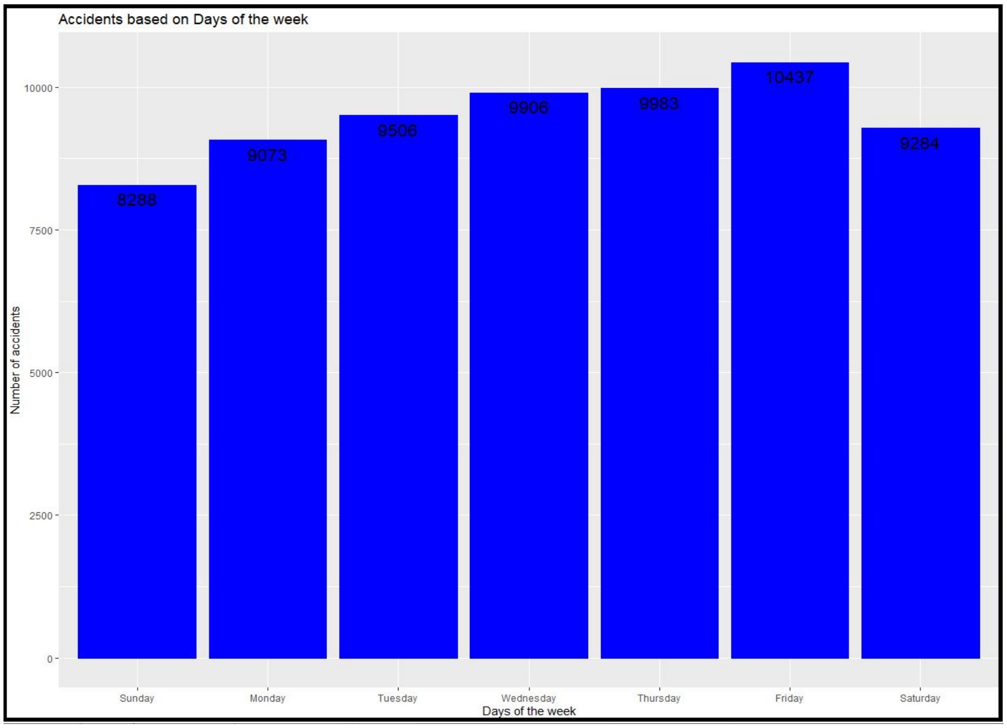


Figure 2 - Number of accidents by days of the week after correcting error

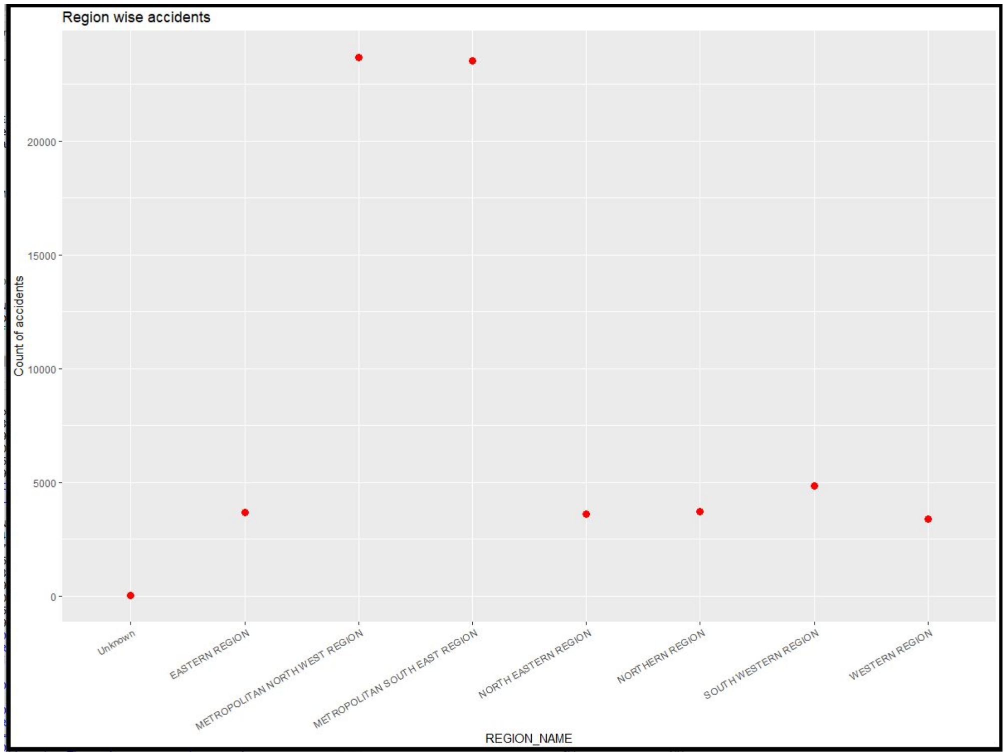


b. Error 2 -

Missing values in the column REGION_NAME.

While exploring the data for the column REGION_NAME it was seen that there are some blank values present. This is an error as Blank values should not be present in the dataset. The blank values are first replaced by Unknown to visualise it and then by using group_by function the count of crashes for each region is calculated. The plot looks like as shown below.

Figure 3 - Unknown Value as a Region Name



Correcting the Error -
The error has been corrected by removing the entries in the data that had REGION_NAME as Unknown. This is done by using data <- data[!(data$REGION_NAME=="Unknown"),]

4. Data Exploration

For the data exploration process we'll answer the questions and try to explore some new aspects of the dataset too.

Q1 - Number of accidents on the basis of the location(using latitude and longitude), year,days of the week and region wise individually.
A1 - Accidents based on location can be plotted using the latitude and longitude given for each accident. Using the leaflet map in R and the markerClusterOptions() the points can be marked on the map. The ClusterOptions() give us an opportunity to look at the points in a cluster manner and then zoom in/out according to the region. This way we can find in which region the maximum number of crashes happen. The visualisation for this is attached.

Figure 4 - Leaflet Map with markerClusterOptions to show location of accidents

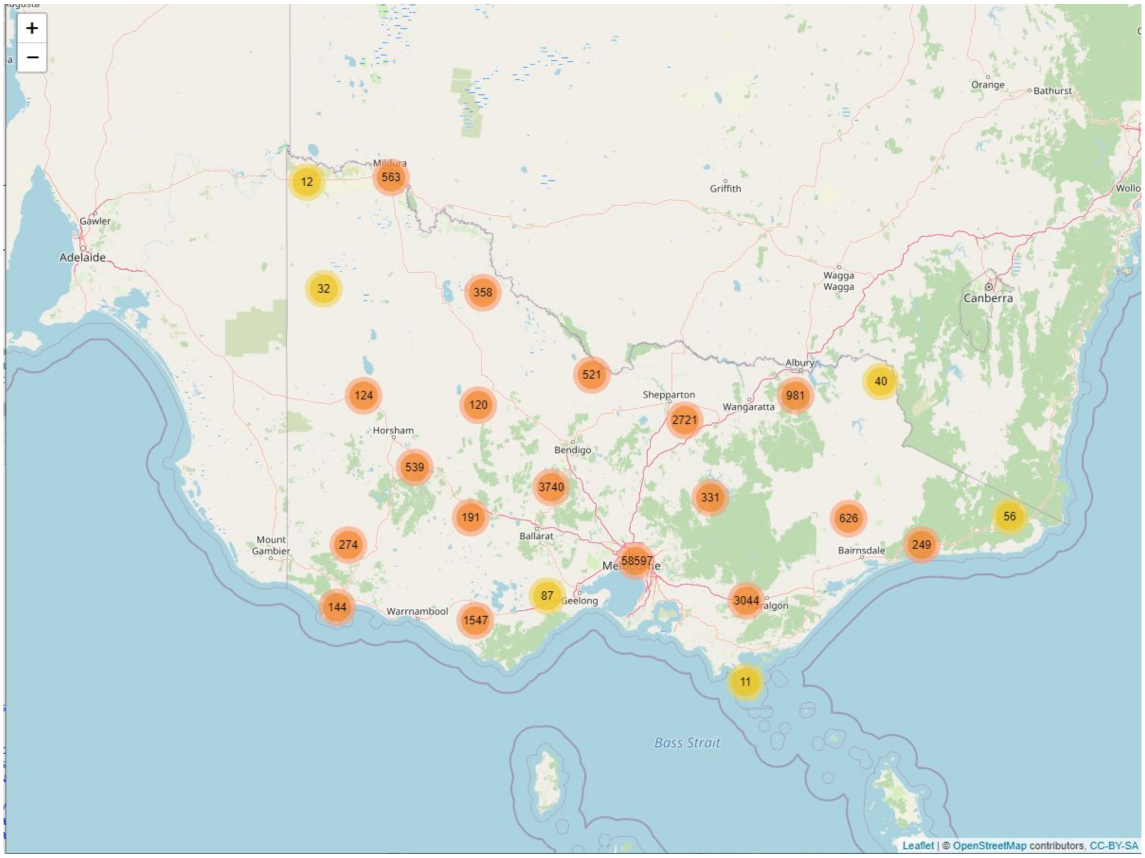
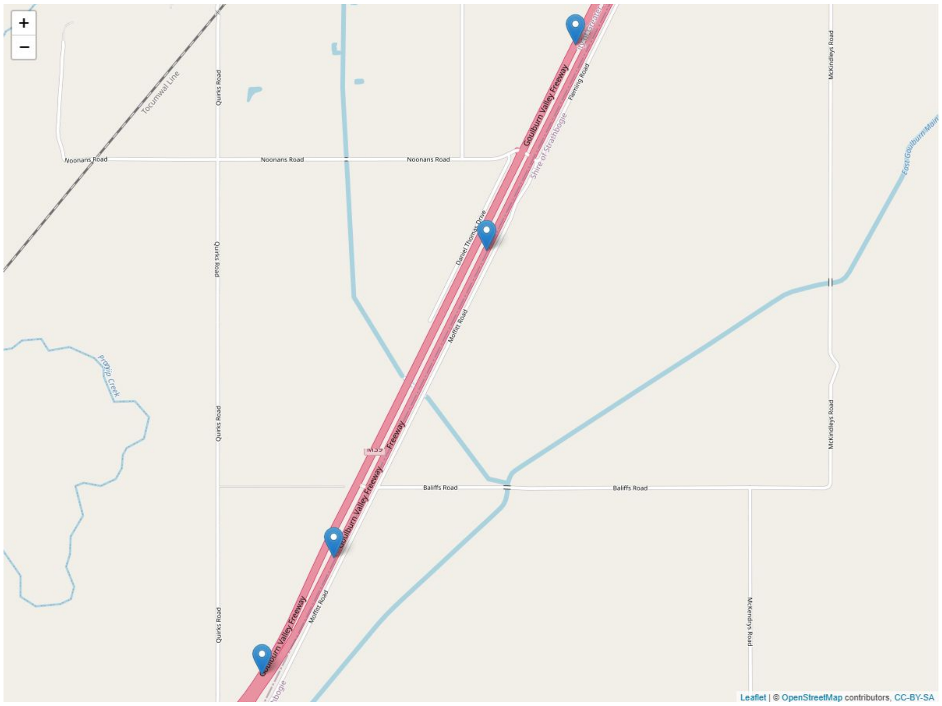
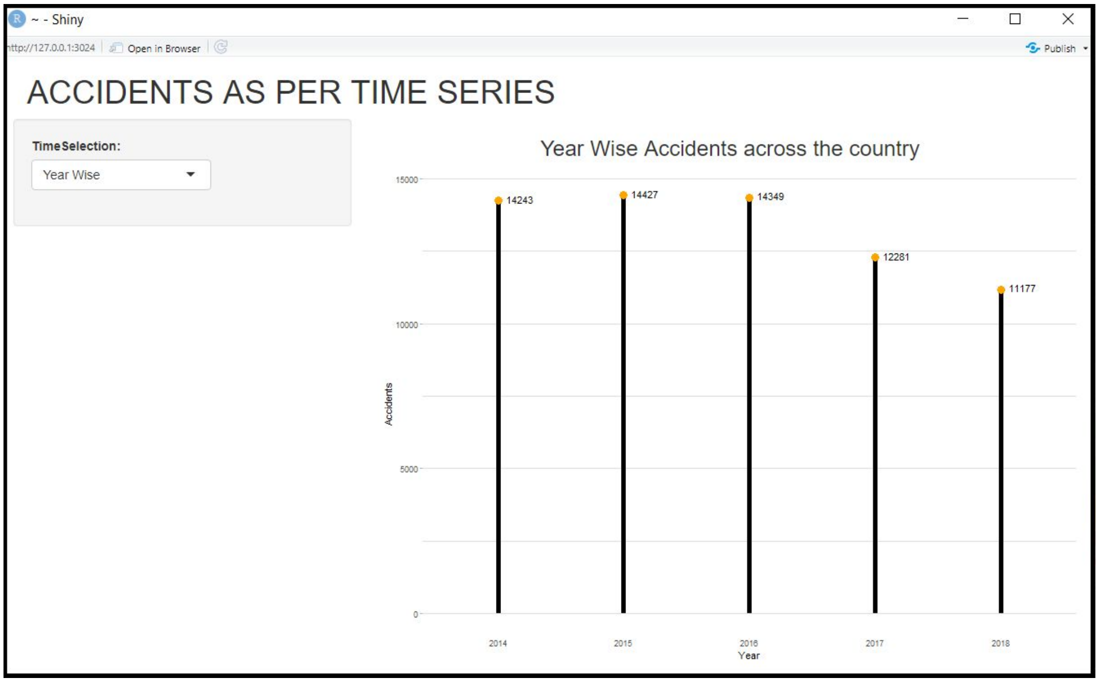


Figure 5 - Zoomed in version of leaflet map



While plotting the accidents based on Years, **another addition to the question has been made to plot accidents based on Month too.** For these two scenarios a Shiny App has been created which allows the user to select an option Year Wise or Month Wise from a drop down and the selected graph is displayed. For Year Wise we have plotted a Lollipop Chart. This depicts that the year 2019 has the most number of accidents with a value of 14427.

Figure 6 - Shiny Visualization for Year Wise Accidents

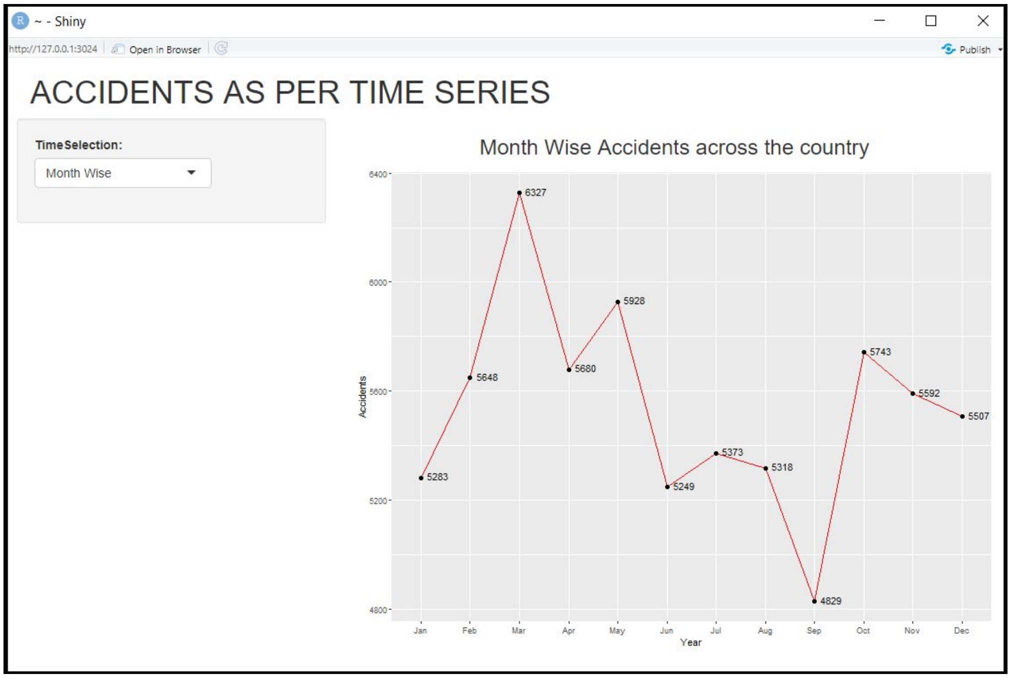


ACCIDENTS AS PER TIME SERIES

Time Selection: Year Wise

For the graph of Month wise accidents we have plotted a line graph which shows the point/value for each Month. This tells us that in the Month of March there are comparatively a huge number of accidents. The month of September however sees the least number of accidents at just a value of 4829.

Figure 7 - Shiny Visualization for Month Wise Accidents

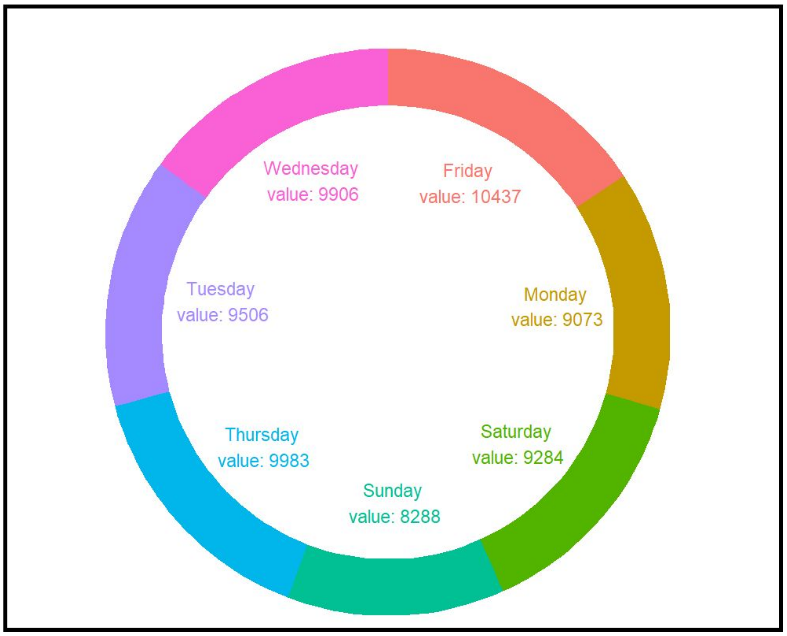


ACCIDENTS AS PER TIME SERIES

Time Selection: Month Wise

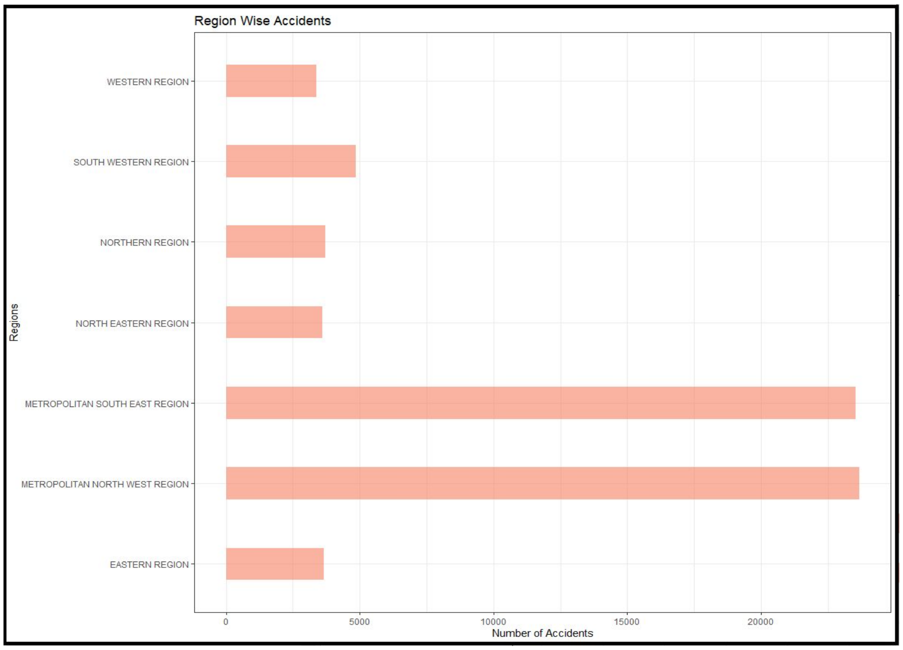
Accidents based on Days of the Week are calculated using the doughnut graph. The function aggregate() is used to calculate the count of accidents based on each day. The graph finally looks like the image shown below. This helps us to identify that almost each day has an equal distribution of the number of accidents. This is shown in Figure 8.

Figure 8 - Accidents based on Days of Week using Doughnut map



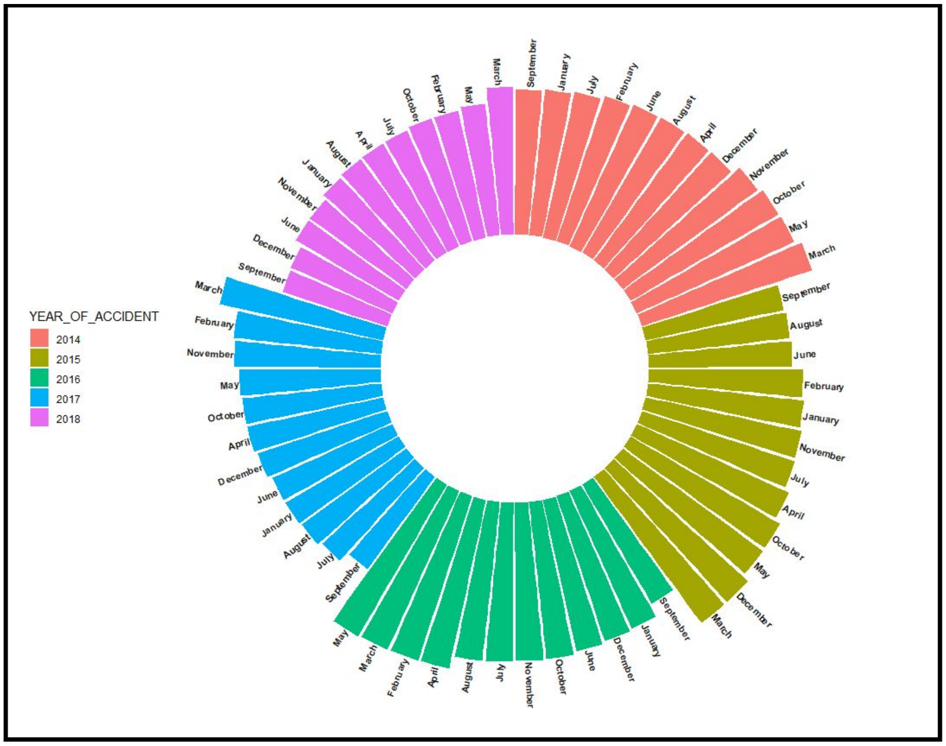
For region wise the plotting has been done using a bar chart. The different regions are taken from the column REGION_NAME and an aggregate data frame has been created to get count of accidents for each region. The final graph tells us that the regions Metropolitan South East and North West show a large number of accidents while the other regions have less number of accidents. Hence more concentration should be focused upon these two regions. This is shown in Figure 9

Figure 9 - Region Wise Accidents



Another additional visualisation has been done that gives us an idea collectively about the accidents for each Year in each Month. A circular bar plot has been created to depict this. The colors represent the years and each color range has bars ranging from Jan to Dec to represent each month. It shows that comparatively the year 2018 recorded less number of accidents. For 2015’s March,Dec and 2016’s Feb, March, Apil, May and 2017’s March we can see a high number of accidents.

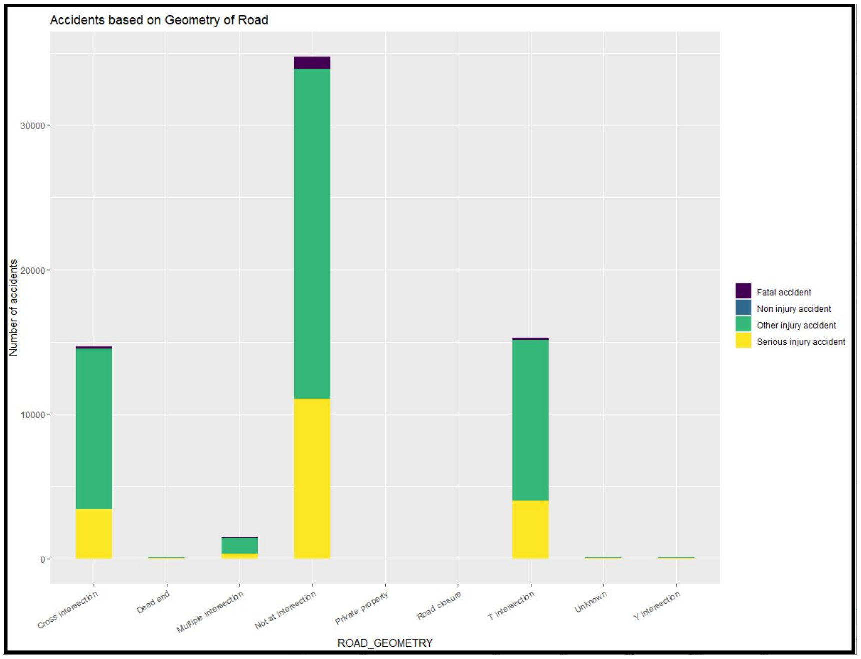
Figure 10 - Year and Month Wise segregated accidents



Q2 - Crashes based on the severity and the geometry of the road hence giving a relationship between the two.

A2 - For finding the relationship between the geometry of the road and severity of the crashes, we have used R to plot a stacked bar chart. Each geometry of the road is represented along the x axis and the severity is represented by the colors in the graph. It tells us that most of the accidents occur on roads with a geometry of Cross intersection , Not at intersection and a T intersection. Hence more care should be taken on such roads. In these categories only the injuries that took place were either other injuries or very serious injuries. The graph shows a relationship between the two.

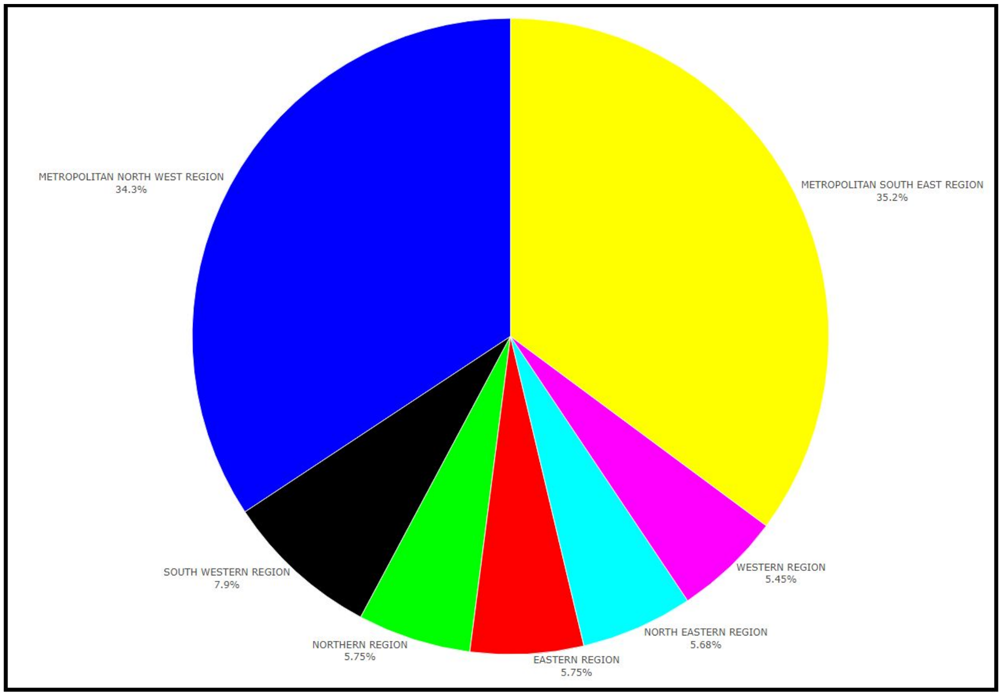
Figure 11 - Accidents based on road geometry



Q3 - Number of fatalities according to the region of the crashes.

A3 - The number of fatalities per region is plotted using a pie chart. The regions which show the most number of fatalities are Metropolitan North West and South East Regions with a percentage of 34.3% and 35.2% respectively.

Figure 12 - Fatalities based on Regions



5. Conclusion

After wrangling, cleaning and exploring the dataset we have learned many things about the crashes on Vic roads. After answering the questions a clear image is presented while comparing different aspects like Year, Month, Geometry, etc to the number of crashes. It can be concluded that 2019 was the year with maximum incidents and similarly when talking about month wise we can say that March is a month that recorded the most number of accidents. If we talk about the days of the week then we can say that almost each day has an equal proportion of accidents. Talking about regions we see a big difference in the values. For the regions Metropolitan North West and South East a huge number of accidents have occured. Hence special consideration and care should be focused on these regions to minimize the number of accidents. The cause should also be identified that are leading to such a high number. Finally if we talk about geometry of the roads again the Roads with not an intersection, cross intersections and T intersections have experienced the most number of accidents. Again focus should be on such roads and special measures like more red lights, less speed limit boards can be put up in these areas to reduce accidents. The exploration process hence answered all the questions that were initially put up.

6. Reflection

From this project I have learned how to load, wrangle and play with the dataset in R. I have also expanded my limits to learn new graphs like the doughnut graph and circular barplot. Different wrangling techniques and functions have also been used in R which helped me understand the data well and visualise it in a better manner. A shiny App has also been created in the visualisation that gave me much more confidence in creating more such interactive visualizations.

7. Bibliography

For wrangling and plotting purposes I have used various sites online.
www.stackoverflow.com
https://www.r-graph-gallery.com/
And documentation of functions in R