Table of Contents

1. Introduction

Road traffic accidents have emerged to be an important global health crisis and cities are investing on multi-disciplinary approaches to tackle the issue. With an exponentially growing population in a city like Melbourne, it becomes a priority for us to investigate on road crashes to identify patterns and possible reason for the events.

An article from World Bank reported how road crashes can impact financial stability of an individual and affects the poverty in a city than we probably thought. This motivated me to analyze road traffic accidents in Melbourne in the past 5 years to identify patterns and deduce possible reasons for the event to occur. I will be exploring the data to answer the following questions:

- How time/location of the road accidents is co-related to the location of pubs/bars.
- Find trends with the volume of vehicles, speed zone, road geometry, lighting, the day of the week and road crashes.
- Trends with "Hit and run" cases and the speed-limit of that road.
- Pattern in the type of accident over months of a year.
- Trends on road geometry and accidents to determine dangerous road geometry.
- Determine the correlation between accident type, alcohol time, road geometry and casualties.

This report takes you on a journey to understand better about road traffic accidents and investigate the patterns observed to determine the possible events causing the pattern in Melbourne road accident data.

2. Data Wrangling

This section shared the detailed information of how I extracted, formatted and wrangled the data for the best use during exploration and analysis for trends. I have used 4 different data sets for this task, 2 of which is scraped from a webpage to extract data using R and the rest is obtained from converting the JSON format data to tabular format(dataframe). The tabular format data can be easy to inspect for data errors, explore and perform analysis using R and tableau.

2.1 I have extracted Melbourne road accidents data for past 5 years from vicroads website.

- The GeoJSON format data was downloaded from vicroads.
- Used 'pandas' and 'ast' library in python to iterate through the JSON objects, remove the tags and append it to a dataframe with 74908 records and 65 attributes.
- Converting the JSON into dataframe format created a lot of NULL values which was addressed by deleting the rows with more than 50% NULL attributes.
- Used mean substitution to impute quantitative attributes with null values.

2.2 I gathered data about Bars and pubs in Melbourne from City of Melbourne website.

- THE JSON format data was gathered from City of Melbourne site.
- Used "DT" library in R to iterate through the JSON objects, remove the tags and convert it to a dataframe with 3339 records and 11 attributes.
- Converting the JSON into dataframe format created a lot of NULL values which was addressed by deleting the rows NULL values in coordinates and location attributes.

2.3 Performed webpage scraping to extract weather data from Australian Government Bureau of Meteorology.

- I have used "xml2", "rvest" and "stringr" libraries in R to extract only the average temperature and rainfall data of Melbourne for the year 2014-2018.
- Made complete use of regular expression to discard HTML tags and extract the data into a dataframe of 2176 records and 3 attributes.

2.4 Performed web scraping to extract working days/public holiday data from Business Victoria.

- I have used "xml2", "rvest" and "stringr" libraries in R to extract only the Public holiday data of Melbourne for the year 2014-2018.
- Made complete use of regular expression to discard HTML tags and extract the data into a dataframe of 146 records and 2 attributes.
- I now had to add weekends to the dataset to arrive at the final data for off-working days in Melbourne.

3. Data Checking

Check for Null values and correlated attributes. I used R for imputing and inspecting data in this section.

3.1 Melbourne Road accident data(2014-2018)

- Dataset consists of 74908 records and 65 qualitative and quantitative attributes.
- Eliminated the rows with more than 50% NULL entry attributes.
- Employed mean substitution to impute quantitative attributes with null values.
- Performed chi-square test using the library "vcd" in R to determine correlated attributes to reduce dataset by eliminating unwanted attributes.
- I found attributes like "male" and "female" were correlating attributes which I eliminated and constructed a single attribute named "gender".
- I also found attributes "VICGRID_X" and "VICGRID_Y" which was co-related to generated "longitude" and "latitude", where I decided to eliminate the VICGRID from the data-frame.
- Similarly, "DEG_URBAL", "LGA_NAME_ALL" and "OTHER_INJURY" were also deleted to avoid highly correlated attributes in the dataset.

3.2 Pubs and Bars data

- Dataset consist of 3339 records and 11 qualitative and quantitative attributes
- Deleted all the rows NULL entries under coordinates and location attributes. As they couldn't be imputed and is a vital feature to analyze.
- I deleted location attribute as I already had x and y coordinates in my data-frame.
- Also employed regression imputation to determine the NULL values in "Number of patrons" attribute which was required in our analysis.

3.3 Melbourne Weather data (2014-2018)

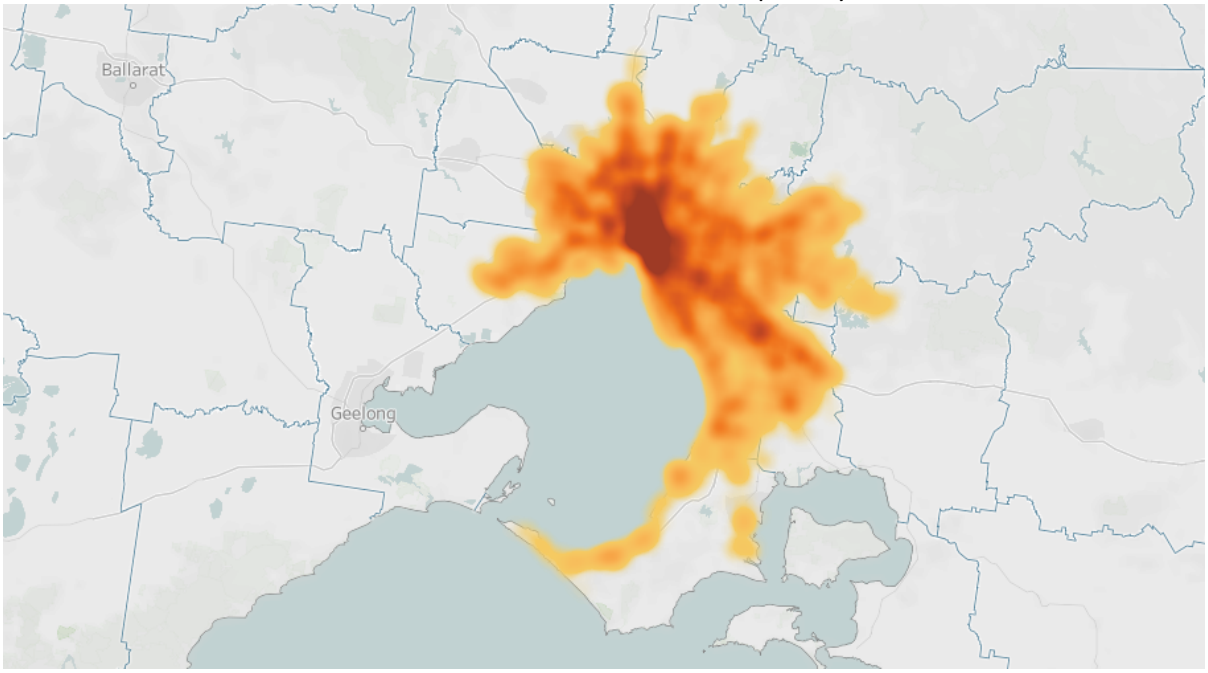
- Dataset has 2176 records and 3 attributes (temperature, rainfall and date).
- Performed mean imputation for 6 NULL values in "temperature" and "rainfall" attributes from the dataset.
- As the data was extracted with regular expression for the attributes of interest, we do not have to bother about irrelevant attributes.

3.4 Melbourne Holiday data(2014-2018)

- Primary public holiday dataset was 146 records and 2 attributes (Name of the holiday and date).
- This data was then combined with generated dataset of weekend dates.
- Final dataset is found to be free from NULL values and duplicate entries.

4. Data Exploration

Victoria has encountered over 74908 road crashes during the span of June 2013 to March 2019. Out of which 18400 road crash reports are contributed by Melbourne alone. Analyzing the road crash coordinates in Melbourne could unfold a picture of where the crash reports are concentrated and how the crashes are distributed over the past 5 years.

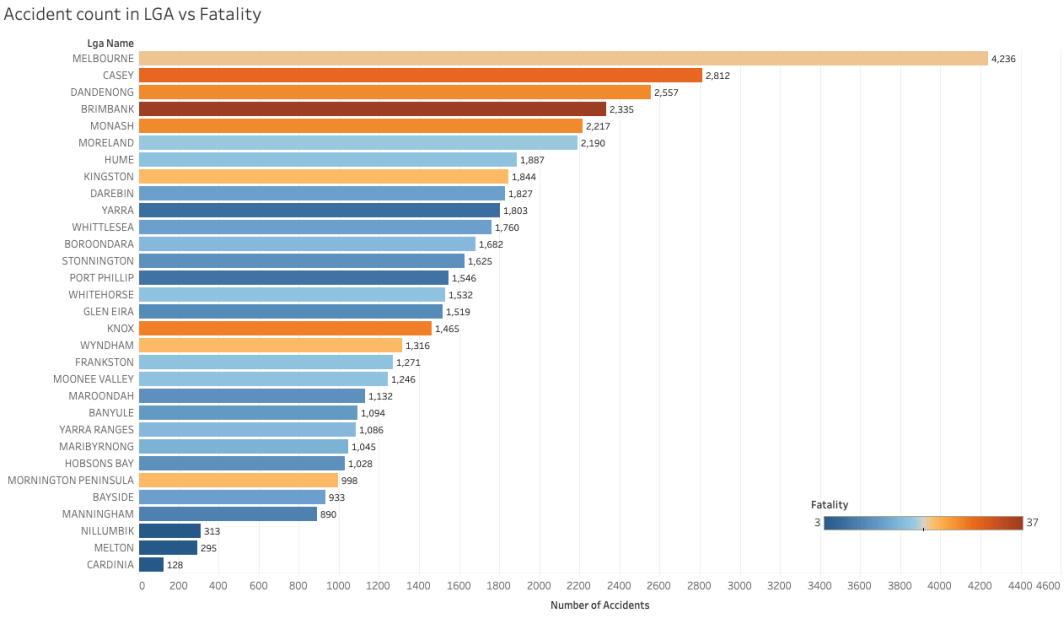


By observing the density plot above we can notice the following:

- Melbourne city seems to have the higher concentration of road traffic accidents over the span of 2014-2018.
- Road crashes are evenly distributed over the years.

Investigating further on the parts of Melbourne:

Melbourne is again subdivided into 30 LGAs(Local Government areas) which could be broken down for our analysis. We would also be interested to know about fatality rate of each region under Melbourne. Below is a plot of number of crashes in each LGA under Melbourne. Color of the bar represents the fatality in each region.



By observing the bar plot above, we can notice the following:

- **Melbourne city** has the highest number of road crashes. However, it has a tolerable fatality count at Melbourne.
- **Brimbank**, the western part of the Melbourne city has the highest fatality count among all the LGAs.
- **Casey**, the eastern suburb of Melbourne has the second highest reported road crashes with a slightly higher fatality count.
- **Brimbank and Melbourne city** are considered most accident prone regions in Melbourne. On the other hand, **Cardinia and Melton** are the safest regions with low road crash and fatality count in the whole of Melbourne.

Exploring the relation between road crashes and the day of a week:

We would be interested to know the general trend in road crashes over the days of the week. I have used a heat map to understand the trend in accidents on a particular day of a week over the span of 5 years (2014-2018).

Accident count VS days of the week

Number of Records
1,293 — 2,302

Accident Date

Year of Acci..	Sunday	Monday	Tuesday	Wednes..	Thursday	Friday	Saturday
2014	1,837	1,881	1,966	2,161	2,141	2,173	2,084
2015	1,817	1,907	2,053	2,111	2,216	2,302	2,021
2016	1,817	1,960	2,040	2,114	2,116	2,287	2,015
2017	1,524	1,765	1,781	1,848	1,795	1,922	1,646
2018	1,293	1,560	1,666	1,672	1,715	1,753	1,518

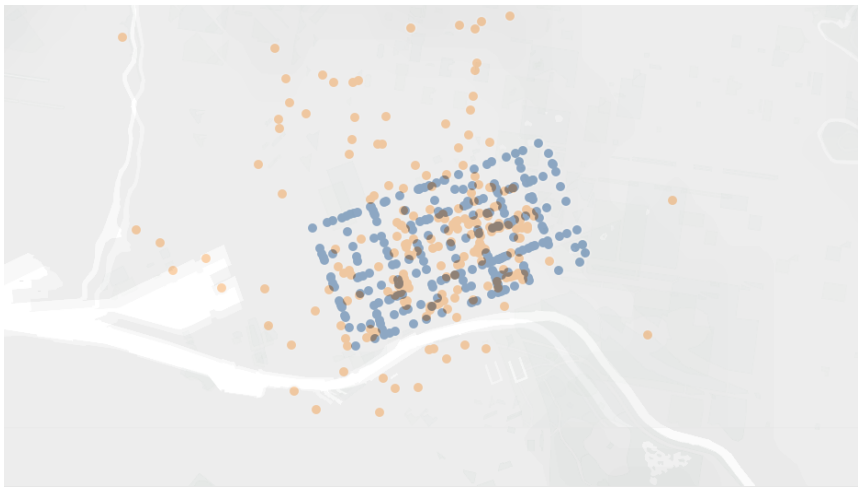
By analyzing the Heat map above, we can conclude that:

- Number of accidents have decreased significantly over the years from 2014 to 2018.
- **Fridays** have the highest count of road crashes followed by **Saturdays** being the second highest day prone to accidents over the years.
- **Sunday** being the safest day on the road over the years.

Possible reason for a high accident count during Fridays and Saturdays could be due to drunk drivers?

Hypothesis: location of pubs/bars are related accident location:

I have concentrated on the Melbourne CBD location to understand the correlation between the location of pubs/bars and the location of the accidents on a weekend. I have extracted the location of pubs/bar in Melbourne from another source as stated above.

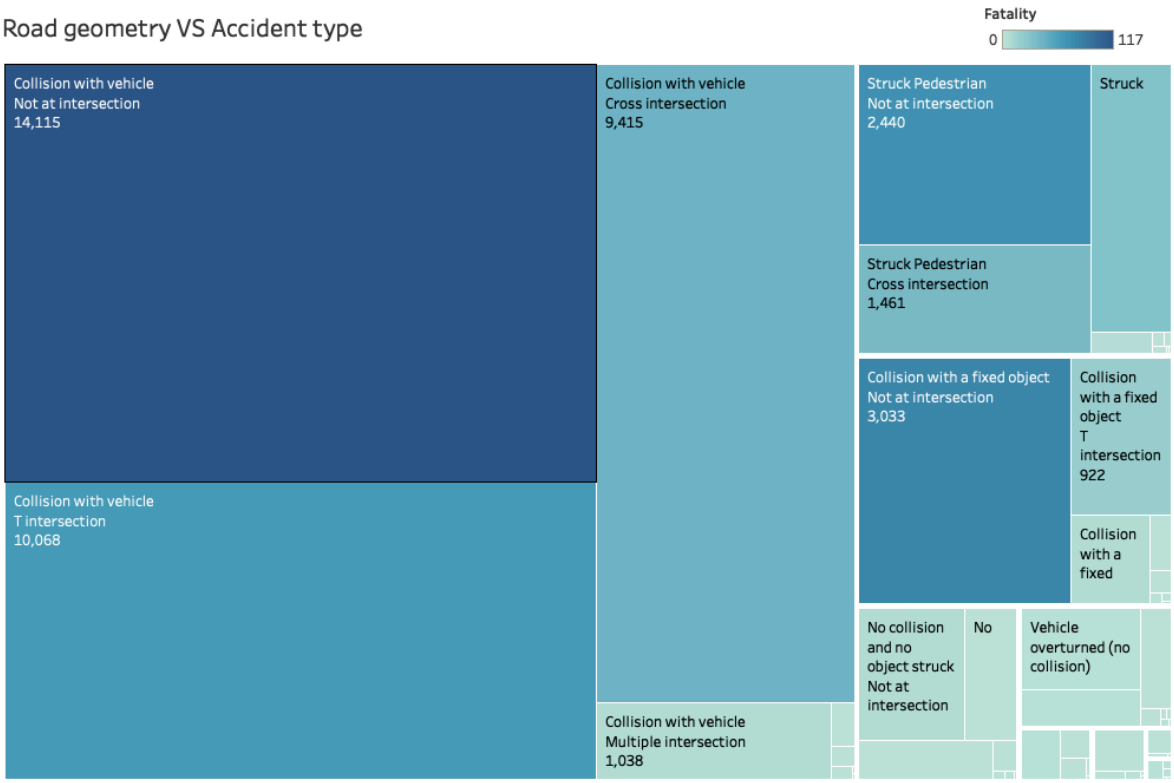


Here the yellow dots show the location of pubs and blue dots depict the accident location.

Observing the plot, we can conclude that our hypothesis isn't true. The plot clearly shows that the location of pubs/bars are not related to the accident spots in Melbourne CBD.

Inspecting the orientation of accident type over road geometry:

The road crash data set is classified into 9 Accident types and 9 Road geometry. I have explored the occurrence of a certain combination of accident type and road geometry that has higher road crash count and fatality.



The above tree map illustrates the following:

- Road accidents caused by **collision with a vehicle on a road without intersection** has the highest count of crashes with 14,115. Which is again rated as the **highest fatality** among all the combinations of accident type and road geometry.
- Second highest count for road crashes are **collision with a vehicle at T intersection.** However, this category possesses a tolerable fatality count.
- Accidents by **colliding with a fixed object on a straight road** and **accidents stuck by pedestrian on a straight road** has the highest fatality count per crash.

To conclude, **Collision with vehicle, fixed object and pedestrian** is the most common type of accident which is fatal. A **non-intersection road** is the most dangerous road attribute with highest fatality count.

Inspecting "hit and run" cases at the non-intersecting roads and their speed limits:

Since we found that non-intersecting roads are the most dangerous and prone to road crashes. Further exploration on the speed limits of the non-intersecting roads, the accident count and the fatality will reveal the dangerous roads and their speed limits.

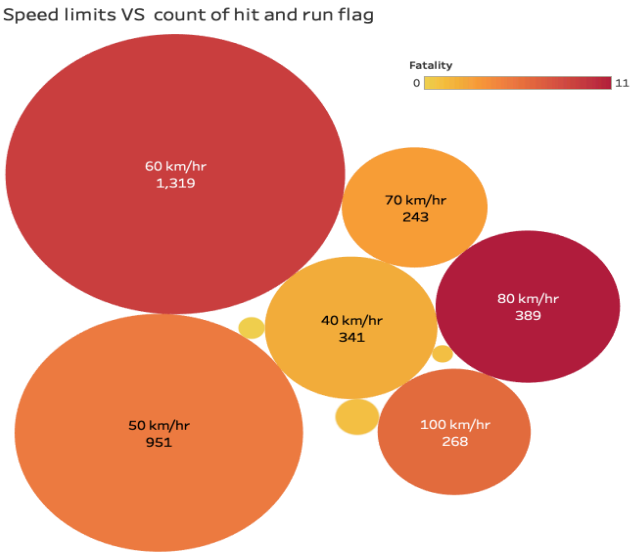
I have filtered only the cases which are flagged as "hit and run". Below is a bubble plot to depict the speed limit in each bubble, the size of the bubble shows the number of road crashes on that speed limit and the colour of the bubble signifies the fatality count.

The bubble graph illustrates:

- **60Km/h roads** have the highest number of road crashes with the second highest fatality count.
- **50Km/h roads** have the second highest number of road accidents with tolerable fatality rate.

- **80Km/h roads** have the highest fatality count while it just has 389 reported road crashes. With fatality count of 11 per 389 road crashes, 80Km/h roads are concluded to be the most dangerous roads with the highest fatality rate in Melbourne.

Thus, we have now concluded that **Collision with vehicle, fixed object and pedestrian** is the most common and dangerous type of accident while **60km/h and 80km/h non-intersecting roads** one of the most dangerous roads in Melbourne.



Exploring the relation between road crashes and the month of a year:

Investigating the correlation with number of road crashes and the month of a year unveiled a strange yet interesting pattern. I have used heat map to visualise the relation between the number of crashes over months.

Accident count VS months

Number of Records
818 1,350

Accident Date

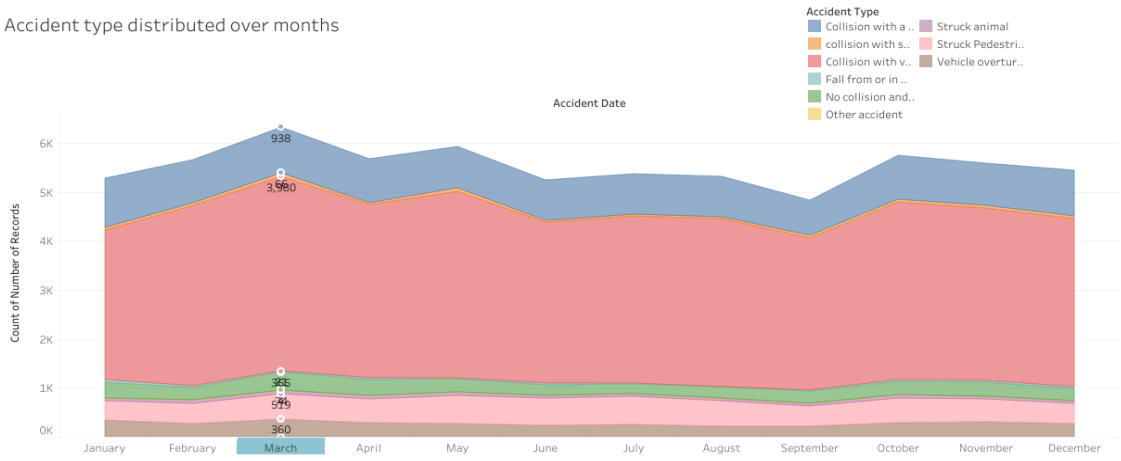
Year of Acci..	January	February	March	April	May	June	July	August	Septemb..	October	November	December
2014	1,091	1,140	1,335	1,176	1,285	1,156	1,111	1,167	1,080	1,277	1,245	1,180
2015	1,169	1,149	1,350	1,243	1,296	1,071	1,202	1,064	1,054	1,290	1,190	1,349
2016	1,127	1,290	1,295	1,284	1,311	1,153	1,184	1,191	1,026	1,175	1,177	1,136
2017	975	1,105	1,249	1,036	1,050	980	932	963	851	1,038	1,090	1,012
2018	921	964	1,098	941	986	889	944	933	818	963	890	830

Above Heat map unveiled the following:

- **March** has constantly been the month with most number of road crashes over the years.
- Number of accidents have decreased significantly over the years from 2014 to 2018.
- **September** has constantly been the safest month on roads.

This strange pattern impels us to investigate further on accident type, road geometry, light condition, types of drivers, weather and traffic volume during the month of March which is contributing to the peak in accident count.

Inspecting the change in number of records under accident types over months:

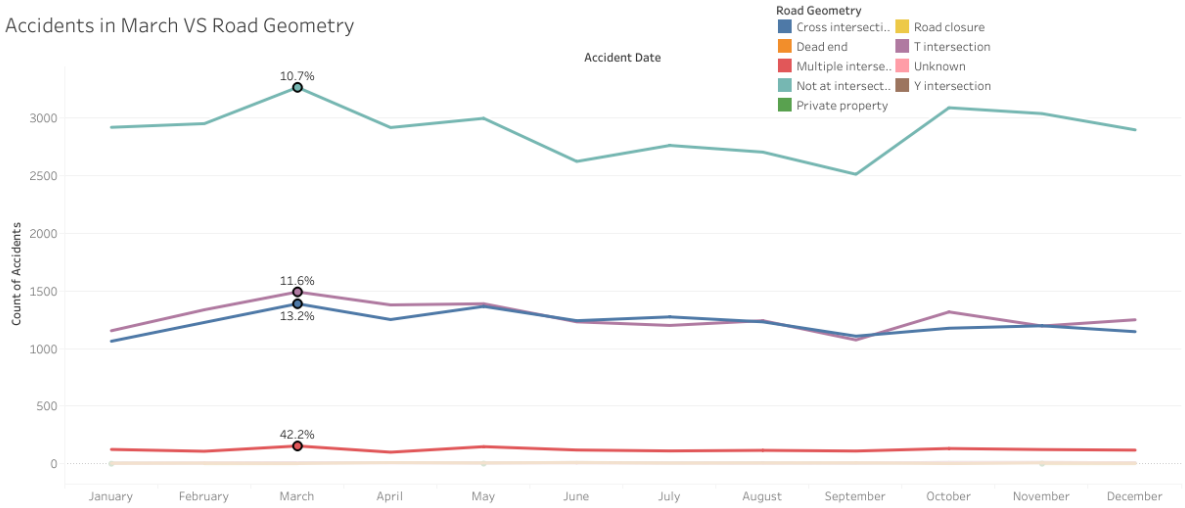


Area chart would help us determine the fine changes in the records over the months in past 5 years. Carefully analysing the graph, we can notice the following:

- There is a small increase in the number of cases reported by **struck pedestrian**.
- We can notice a significant peak in accident type – **Collision with vehicle** during the month of march.
- Similar pattern but with a lesser magnitude can be observed during October too.

Inspecting the change in number of records under road geometry over months:

Exploring the changes in road crashes at a particular road geometry could be the second important aspect to investigate to notice the possible pattern over the months.

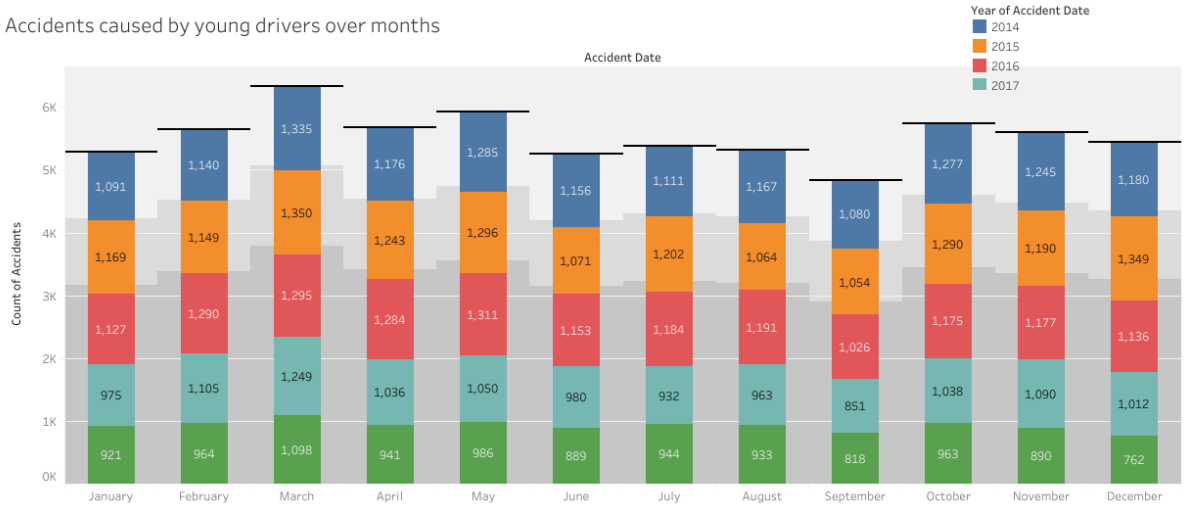


Line graph discloses the following observations:

- Significant increase in the number of crashes on a **non-intersecting roads** during March.
- Gradual peaking occurs with **T intersected roads and cross intersected** roads during the month of March.
- Similar pattern but with a lesser magnitude can be observed during **October** too.

Inspecting the road crashes caused by young drivers over months:

We have identified young drivers to be aged between 18 and 25. Considering the general hypothesis of young drivers more likely to encounter accidents over others.

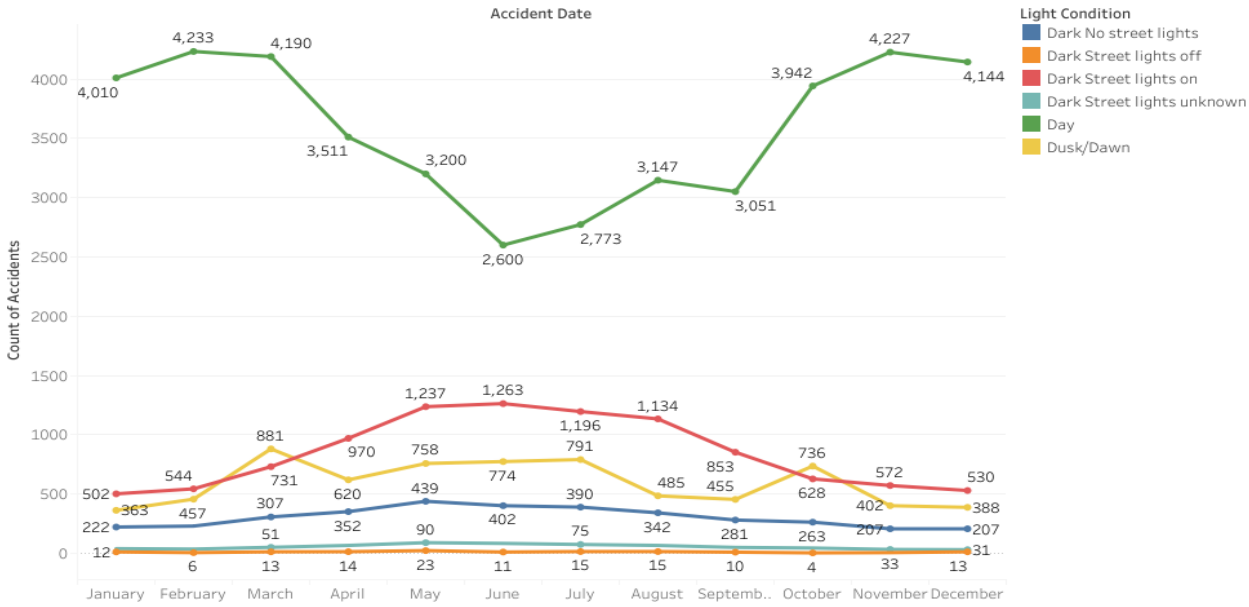


Bullet graph has helped us identify the change in road crashes by young drivers over time. We can notice a **peak during the month of March which correlates with our hypothesis**.

Inspecting the road crashes caused by light condition over months:

Though there is a presence of better natural light in march over most of winter months, I have performed this analysis to identify different light conditions and time of the day the accidents occur. I have again used a line graph to explain the trend.

Light conditions VS months



From the above line graph, we can notice the following:

- An expected high day accidents during summer and dark street accidents during winter.
- However, we also notice a strange peak in road crashes at **Dusk/Dawn** during march and October.

This helps us conclude that the number of accidents at dawn/dusk peak during March.

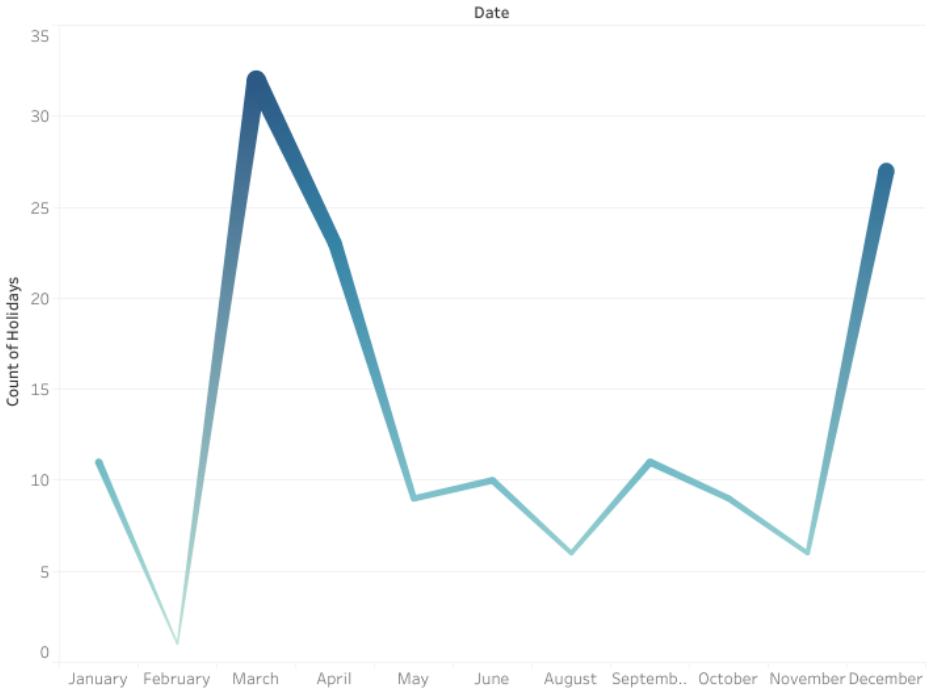
Exploring more about the Number of public holidays over the months:

Hypothesis is more the number of public holidays/weekends in a month, higher vehicle volume on the roads. I have used a line graph to aggregate the total number of non-working days in a month over 5 years.

This graph shows a clear evidence about number of holidays being the highest during March.

This seconds my hypothesis about **number of vehicles on the road would increased causing higher road crash count during March due to holiday.**

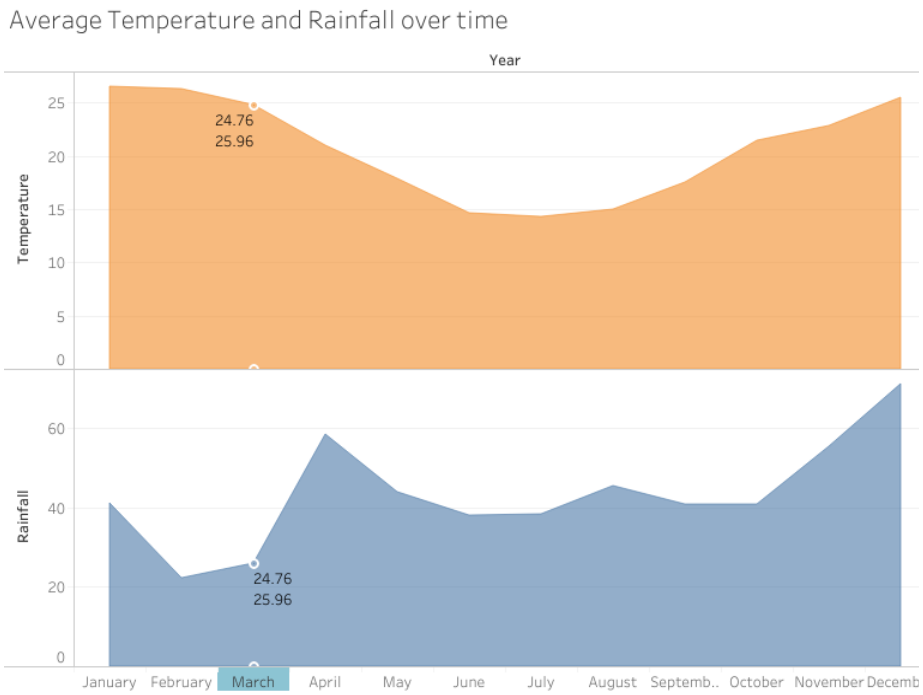
Number of Holidays over months(2014-2018)



Exploring the weather conditions over months (between 2014-2018):

I'm exploring temperature and rainfall over the months to help support a general hypothesis of a pleasant day with optimal temperature and low rainfall will support higher number of vehicles on roads due to excursion plans.

This graph shows an average, optimal temperature of 24.76°C and the rainfall is at 25.96mm. This makes a perfect weather for an excursion and supports my hypothesis of **increased number of vehicles on roads heading to an excursion due to perfect weather and public holidays.**

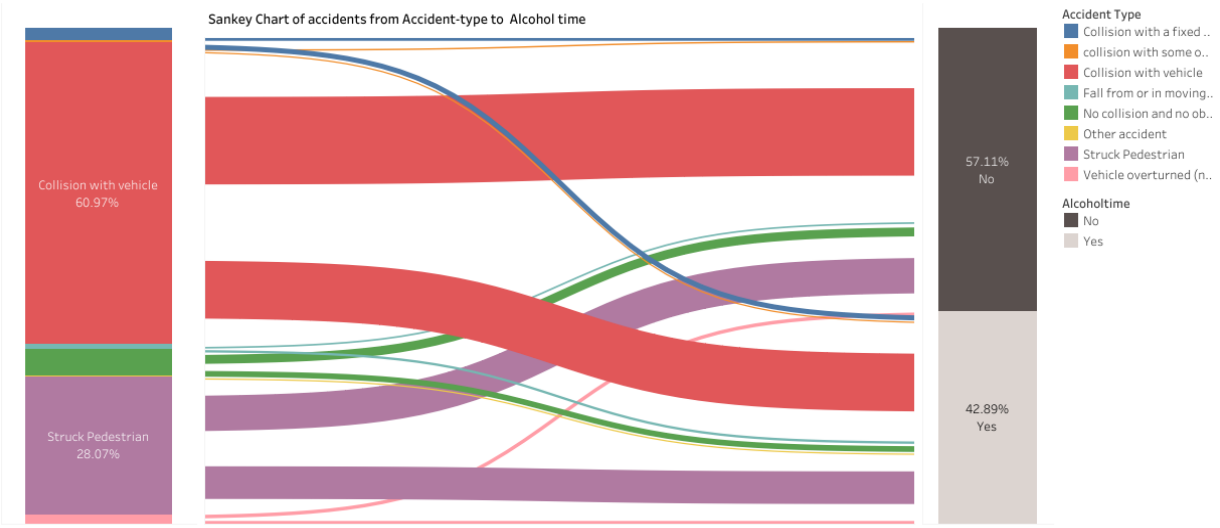


Investigating the flow between accident type, alcohol time, road geometry and fatality:

Inspecting further on the flow of type of accidents and alcohol time during the month of March and all Fridays over the period of 5 years (2014-2018). Alcohol Times are defined as the following:

- Monday – Thursday 00:00-06:00 hours & 18:00-23:59 hours
- Friday 00:00-06:00 hours & 16:00-23:59 hours
- Saturday 00:00-08:00 hours & 14:00-23:59 hours
- Sunday 00:00-10:00 hours & 16:00-23:59 hours

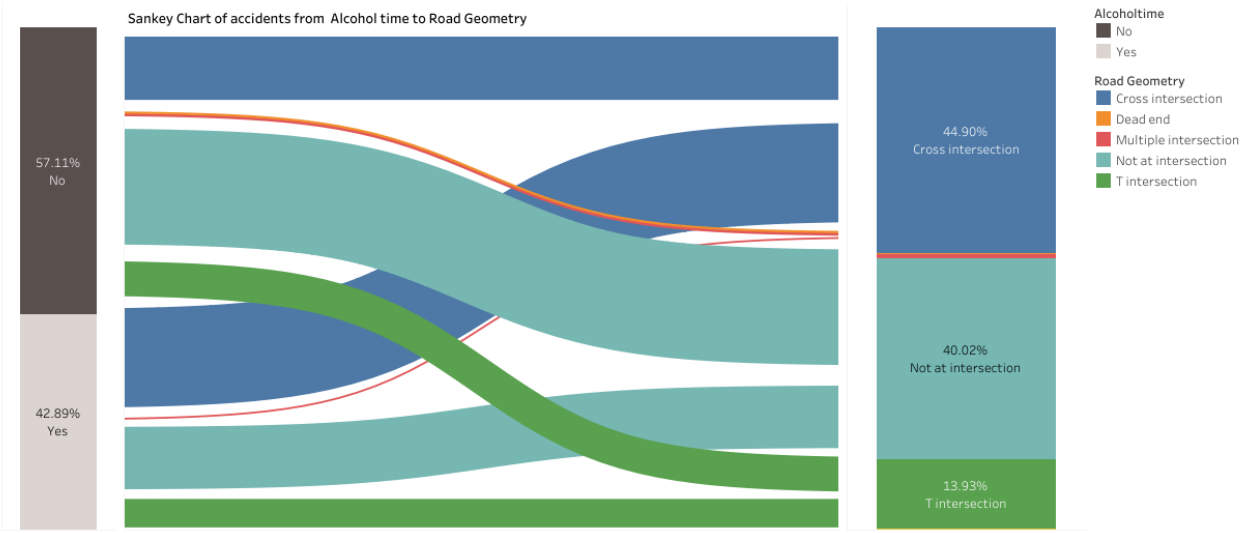
Sankey diagrams emphasize on major transfer flow within a system. I have implemented Sankey diagrams to understand the transfer between the types of accidents and time of accident.



From the above Sankey diagram, we can clearly interpret the following:

- Collision with vehicle (61%) is the most common type of accident, followed by crashes which struck pedestrians (28%).
- 43% of the accidents happen during the defined alcohol time.
- A large chunk of **collision with vehicle in not during the alcohol time.**

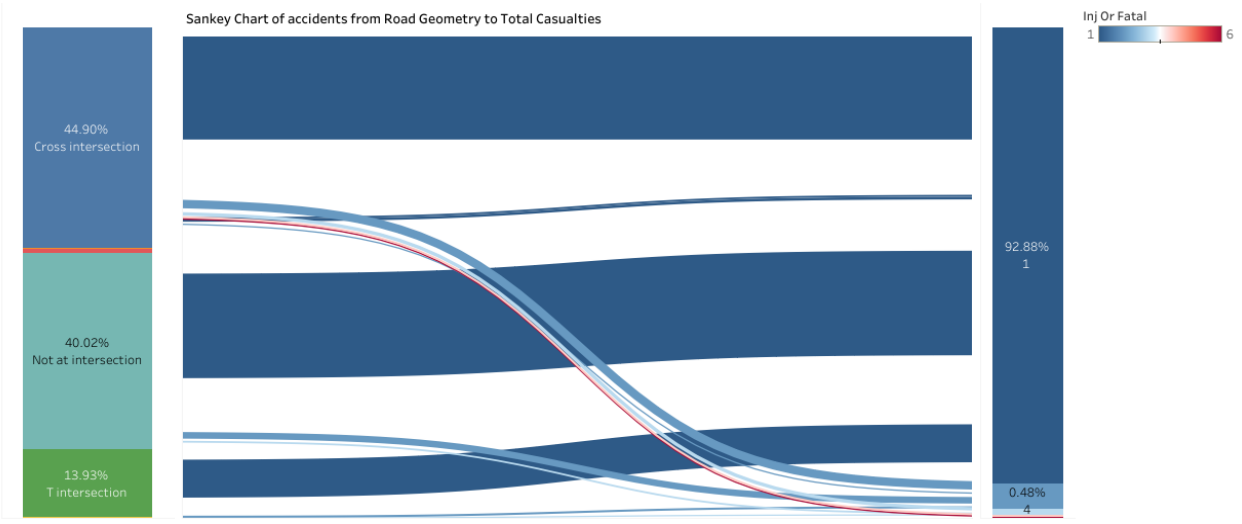
Now, Let's understand more about how accidents during alcohol time are related to the road geometry. We again use Sankey diagram representation to understand the flow between crashes reported during alcohol time and the road geometry of that crash.



From the above Sankey diagram, we can notice that:

- About 45% of the crashes are at the **cross intersection**, which is followed by 40% of the crashes on a non-intersecting roads.
- We can notice that a large portion of accidents that occur during **alcohol time** directs towards crashes at cross intersection.
- However, large section of crashes not during alcohol time in classified under crashes at **non-intersecting roads**.
- **Crashes at the T intersection** which accounts to 13.93% has a balanced share from crashes during alcohol time.

We can critically analyse the road geometry and help us understand the most dangerous road geometry during the month of March and all Fridays of past 5 years (2014-2018).



From the Sankey diagram between road geometry and fatality, we can interpret:

- About 0.48% account to a higher fatality count of 4 – 6, while over 93% of the cases face a fatality count of 1.
- We can also conclude **that highest fatality rate is contributed by roads with cross intersection** making it the most dangerous road geometry in Melbourne.

When we compare all the three Sankey diagrams, we unveil that less than half of the accident type crash reports contribute towards records under alcohol time, while majority of the records under alcohol time contribute towards accidents on a cross-intersection road. We already know that accidents on a cross-intersection roads are more likely to be fatal. Thus, we conclude that **accidents which involves collision with vehicle during the alcohol time on a cross intersection is the most dangerous accident with the highest risk on fatality.**

5. Conclusion

In consideration of global health crisis caused by road traffic accidents, I have performed deep dive exploratory analysis on the road crashes in Melbourne for the past 5 years (2014-2018). Initial investigation on the location of accidents revealed that **Brimbank and Melbourne city** are considered most accident prone regions in Melbourne. On the other hand, **Cardinia and Melton** are the safest regions with low road crash and fatality count in the whole of Melbourne.

We also learnt that **Friday** constantly turned out to be the day with highest road crashes over the years and **Sunday** to be lowest. However, we failed to relate the number of accidents on Friday to the location of pub/bar. Never the less, we discovered that **collision with vehicle, fixed object and pedestrian** is the most common and dangerous type of accident while **60km/h and 80km/h non-intersecting roads** one of the most dangerous roads in Melbourne.

Investigating the correlation between months of a year to the number of road accidents, we discovered that **March** was constantly prone to accidents while **September** contributed least towards the road crashes over the past 5 years (2014-2018). In search of patterns that are causing the rice of accidents during march revealed the following:

- A peak in number of **collision with vehicle** reports during the month of march (Accident type).
- Significant increase in the number of crashes on a **non-intersecting roads** (Road geometry)**.**
- Road crashes by young drivers (age 18-25) is highest during the month of March.
- A strange peak in road crashes at **Dusk/Dawn** during march and October.
- Number of Public holidays are known to be the highest during march.
- Perfect Temperature and lowest rainfall supports the hypothesis of **increased number of vehicles on roads heading to an excursion causing the rise in road crashes.**

We also conclude that **accidents which involves collision with vehicle during the alcohol time on a cross intersection is the most dangerous accident with the highest risk on fatality.**

6. Reflection

An article from World Bank explained how cities in the world are undergoing global health crisis due to road accidents and this inspired me explore and investigate on road traffic crashes in Melbourne. I found trusted source for my main dataset, which I had to extract tabular data from JSON format. I learnt scraping the webpage using libraries like "rvest" in R to extract all the information on the webpage. I further learnt to use regular expression in R to filter out the data of my interest.

Post the wrangling process, I performed data checking for all my data sets to find errors and impute the data. I also performed chi-square test to determine the correlation between attributes I identified to look similar and eliminated those attributes.

Initial exploration of the dataset helped me understand the importance of visualisation to explore larger datasets. I used Tableau to explore the road crash coordinates and figured out that Melbourne city had the most number of road crashes in the past 5 years. I then tried investigating on the days of the week and months of a year which surprised me and motivated me to explore further to identify the reasons and events causing certain weekday(Fridays) and certain month(march) to have the most number of accidents.

Further investigation on this revealed the possible events causing the peak in the road crashes and the reasons are listed in the conclusion above. I also learnt to use tableau to generate Sankey diagrams using the help of "superdatascience" website and materials. These diagrams helped me visualise the flow and relation between import attributes in the investigation.

I identified Friday to be the most dangerous day and March to be the most dangerous month on roads. I also conclude that accidents which involves collision with vehicle during the alcohol time on a cross intersection is the most dangerous accident with the highest risk on fatality.

7. References

- http://blogs.worldbank.org/transport/miga/road-crashes-have-more-impact-poverty-you-probably-thought
- https://opendata.arcgis.com/datasets/c2a69622ebad42e7baaa8167daa72127_0.geojson
- https://data.melbourne.vic.gov.au/api/views/mffim9yn/rows.json?accessType=DOWNLOAD
- http://www.bom.gov.au/?ref=logo
- https://www.business.vic.gov.au/
- https://medium.freecodecamp.org/an-introduction-to-web-scraping-using-r-40284110c848
- https://www.superdatascience.com/pages/yt-tableau-custom-charts-series