Crash Data of Victoria state Exploration Project

Xianlin Feng
28847458
xfen0007@student.monash.edu

April 20, 2019

1 Introduction

The road traffic accident is a threat to all people in their daily life. According to the WHO's statistics in 2018, road traffic accidents are the eighth of the top 10 causes of death, which is the only reason regarding injuries, and all the remaining reason are diseases. In worldwide, road injuries took 140 million lives in 2016, in which 74% are men and boys (WHO, 2019). A serious situation happened in Victoria too. There were 58 lives lost because of the road accident every day in 2018 just in Victoria. This number increased to 88 since 7 April 2019. According to the TAC report, the lives lost of drivers takes nearly half (48.2%) of the daily lives lost. The age of death concentrate in 30-69 years old. The most lives lost in rural roads(58%) (TAC, 2019). The analysis of past traffic accidents can provide a basis for future road construction, accident prevention, and accident rescue. The following three questions will be answered in this report:

1. What is the trend in the last ten years?
2. What is the main cause of road crashes in Victoria.
3. What suggestion we can provide for the people in different areas.

To answer those three questions, I will try to find one or more datasets, then perform data wrangling, data cleaning and data checking before data exploration. During the data exploration, I will use different statistic test methods and visualisation methods to explore the dataset and obtain insight. At last, I will provide some suggestions base on the result of the data exploration. The main structure of this report is as follow: the data wrangling will be processed in section 2, then the data will be checking in section 3. The data exploration will be carried out in section 4, followed by conclusion in section 5 . In the last section, there will be a reflection of this assessment.

2 Data Wrangling

Data wrangling is a process to manipulate data to make the data directly usable for analysis. According to "2016 Data Science Salary Survey", data scientists spend 53% of their time for data cleaning and data wrangling (Tomar, 2019). After data wrangling, the raw data is transformed into the data that can be analysed directly to generate insights and accurate results. Data wrangling is vital for a data science project, which not only improves the efficiency of data analysis but also reduces the error caused by incorrect data. In this section, I will divide the data wrangling process into the following small tasks base on the characteristics of the dataset:

1. Introduce the dataset
2. Filter the data

3. Time series data handling
4. Drop missing or null values in the dataset
5. deal with outliers or "illegal" values

2.1 the Dataset

The dataset I found for this project named "CrashStats data", which could be downloaded on the Victoria government open data website: https://www.data.vic.gov.au. VicRoads provided the dataset for educational purposes, and it includes the crash data of time, location, conditions and so on since 2000. The dataset is consist of 12 tables, the main tables which will be used for this report are in the following list:

1. **accident**: contain the basic information about the accident, such as date, time, location, environment condition and severity.
2. **vehicle**: vehicle information, such as make, body type, year of manufacture, fuel type, vehicle capacity and so on.
3. **person**: personal details, such as age, sex, sitting position, passengers or driver, license state and so on.
4. **accident_location**: the location information of the accident.
5. **road_surface_cond**: the condition of the road: wet, dry, or icy.
6. **atmospheric_cond**: weather condition: clear, wind, dust and so on.
7. **accident_node**: more detailed location about the crash.

Before the data wrangling and data cleaning, we need to understand the characteristics of each data table. For example, in the table of "accident", each accident has exact one record. However, as one accident may involve two or more people, one accident may have two or more records in the table "person". The same situation happens in table "vehicle", "accident_event" and other tables too, this is another reason why this dataset separate to 12 data tables. Due to this reason, when we process data wrangling, data cleaning, or data exploration, extra care should be paid with multiple records for one accident.

2.2 Time series data handling

In the original dataset, the time information is stored in the format "**hh.mm.ss**", which is not the common time format, so that I changed them to the format "**hh:mm:ss**" first. This step could easily perform in **Microsoft Excel** by just replacing every "." by ":". Then, the date and time information is separated into two columns, which is not very friendly to Tableau Public, so we need to combine them first. In order to complete this task, I need to insert a new column "**dateTime**" and in the cell "D2" use an Excel formulation:

"**TEXT(B2,"dmmyyyy")&" "&TEXT(C2,"hh:mm:ss"**)"

This formulation could combine the date information in cell "B2" and time information in cell "C2" without lose any information. By applying this formulation to all rows, The new column could contain both date and time information could be generated.

2.3 Filter data and Joint tables

The data of 2019 and before 2009 will not be considered in this report, so that, before we perform any data checking and data analysis, they need to be filtered. In the table "ACCIDENT", this step could easily be done in Tableau Public, since we have generated the new column named "DateTime". Then select data

with speed zone under 120, and most analysis will base on the records which "NO_PERSONS_KILLED" is greater than 0. For the other tables, as the date information is only stored in the "ACCIDENT" table, this action could be more complicated, but every record in other tables contains "accident_ID" field. So, an inner join is essential for each table and table "ACCIDENT".

2.4 Identify key attributes

Another critical step after selecting data is to identify key attributes. Each data record contains lots of information related to the accident. However, some of them are not related to our project. So, select the useful data and information will help us for data checking and data analysis. Then data checking and correcting will be performed for the key attributes.

3 Data Checking

Tableau Public is a powerful tool for data wrangling, as well as data checking. In this section, I will demonstrate the foremost step I performed for data checking. More specific, checking the critical attributes of each table base on the property of the attribute.

3.1 Check geographic data

In the table "NODE", the attributes "lat" and "long" indicate the latitude and longitude coordinates for each accident. The attribute "Postcode_No" demonstrate the suburb information. To check the attributes of "lat", "long" and "Postcode", just put them in the Tableau Public map, the result shows no record is outside Victoria state. So the geographic data is considered of complete and consist.

3.2 Check duplicate record

Eliminate duplicate records is important for data analysis. To complete this task, I used Tableau Public to identify the duplicate records in the key attributes. For example, for the "ACCIDENT_No" in the table "ACCIDENT". The following command will calculate the time of one record appears in the field:"{fixed[Accident_No:SUM([Number of Records])}". The result shows there is no duplicate records in "ACCIDENT_No" attributes.

3.3 Check the missing value

In Tableau Public, we can identify the missing value by summarise the table or each filed. The picture on the right shows the missing value of attribute "Road Type" in table "ACCIDENT_LOCATION". As it showing in figure 1, there are 2964 missing value for attribute "Road Type", but the missing values are not critical for the data analysis, so that I just ignore them. Another method to deal with missing value is replacing the missing value with another value, such as 0. this could be done with function *ZN()*.

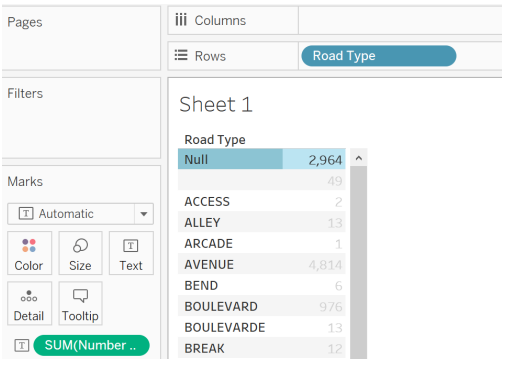


Figure 1: Identify missing value

3.4 Check outlier

To check outlier in the field, we should identify the range first. For example, when checking the data range of the attribute "Node_ID" in both table "ACCIDENT" and table "NODE" in Tableau Public, Two range could be found: [4, 342850] and [−10, 342850]. Then I check the meta-data on the website of VicRoads. For the attribute of "Node_ID", it explains that "The node id of the accident. It starts with 1 and incremented by one when a new accident location is identified" (VicRoads, 2019). Therefore, the values blew 1 in the table "ACCIDENT" are outliers, and should not be used for analysis and exploration. Since the correct node id information is not known, the node id information cannot be used for analysing. Fortunately, the table "NODE" contains both accident information and node-id information, and it will be used in further data exploration.

4 Data Exploration

After data wrangling, cleaning, and checking, we have a clean and well-formatted dataset to analysis. In this section, I will perform data exploration and data analysis with Tableau Public and R. To answer the three questions proposed in section 1, different statistical methods and visualisation methods will be used. I will mainly focus on traffic crashes which cause life loses.

4.1 Trend of road crashes

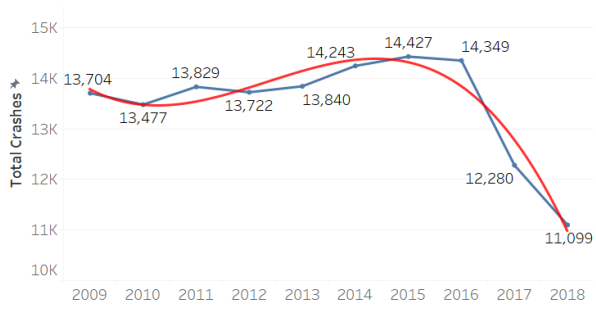


Figure 2: total crashes over 10 years

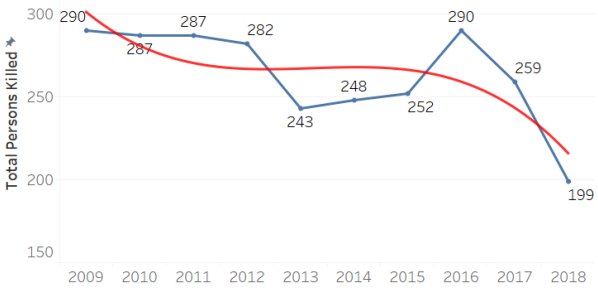


Figure 3: Total person killed over 10 years

The figure 2 indicate that between 2010 and 2016, the total number of road crash increase slowly, but the number decreased rapidly in the year 2017 and 2018. A similar situation could also found in figure 3 in 2017 and 2018, which is a significant improve compare to those years from 2009 to 2016.
Every person with a driver license knows that more accidents happened on Friday, as well as the rush hours in the morning and afternoon. So that, to save the paper, I will skip those analyses, but move on to the reasons for the accidents. First, the analysis of different types of accident.

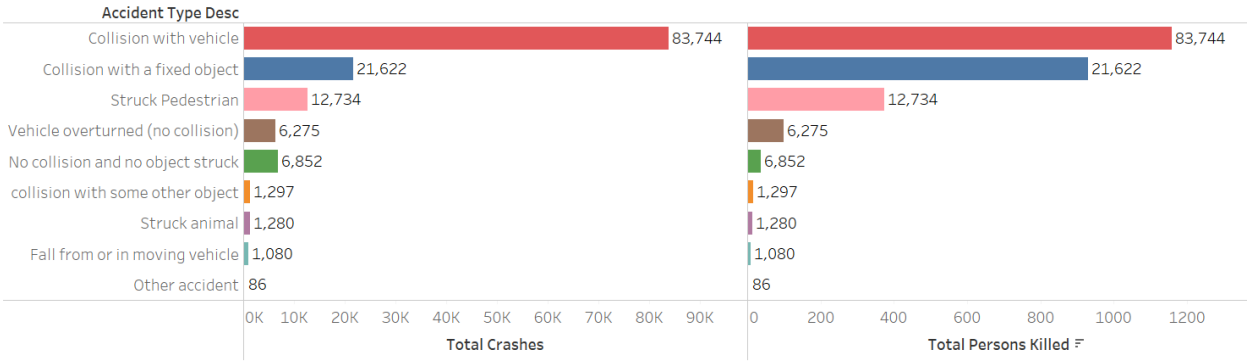


Figure 4: Total crashes and person killed by different accident type

According to figure 4, most of the accidents and life loses are caused by vehicles collision. However, collisions with fixed objects accounted for only 16.02% of the total number of accidents but accounted for 35.27% of all deaths in the dataset.

The figure 5 shows that the trends of life lose for the top three accident types are decrease over the ten years. The peak record appeared in 2016, then drop rapidly in the last two years.

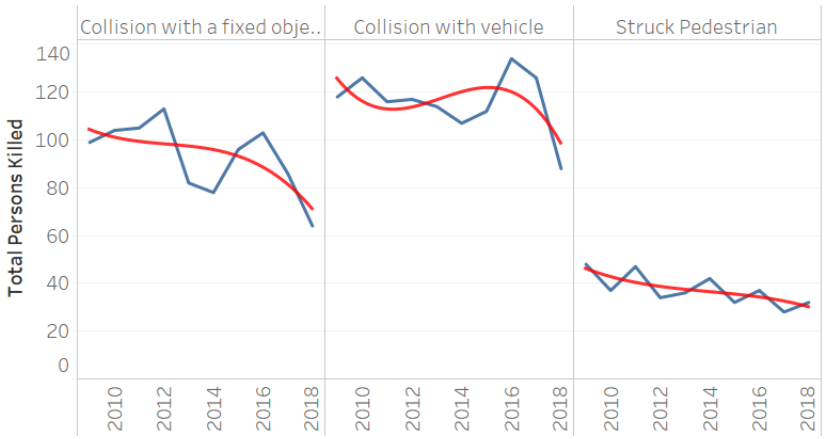
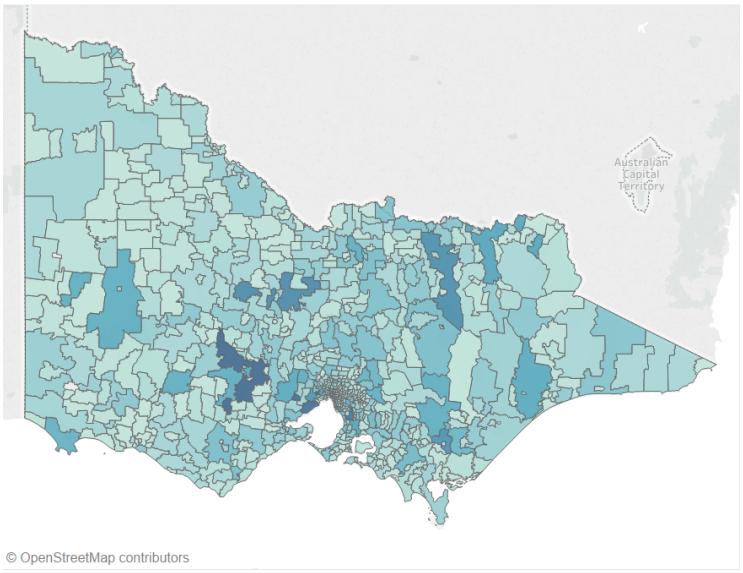


Figure 5: Trends of top 3 accident types

From the above, the statistics test indicate that not only the number of accidents but also the number of people get killed are decreased in the past ten years. To save space, more analysis details will not be shown. Instead, I will focus on the analysis of the cause of the accident, mainly on the accidents that caused the death of a person.

4.2 main cause of road crashes in Victoria

In this section, more data exploration will be performed to find the leading cause of the person get killed in an accident in Victoria state, such as light condition, atmosphere condition, different accident type at the different road surface and so on.
First of all, to perform data analysis base on geographic information, the table "ACCIDENT" and the table "NODE" need to be joined together. An inner join base on the attribute of "Accident_ID" could is the most reasonable choice for this task, because both tables use this attribute as their identity key.



It is clearly shown in figure 6, that **Bendigo**, **Cocoroc** and **Bangholme** take the top 3 number of life loses amount all suburbs in Victoria state.

Figure 6: Total live loses in each suburb

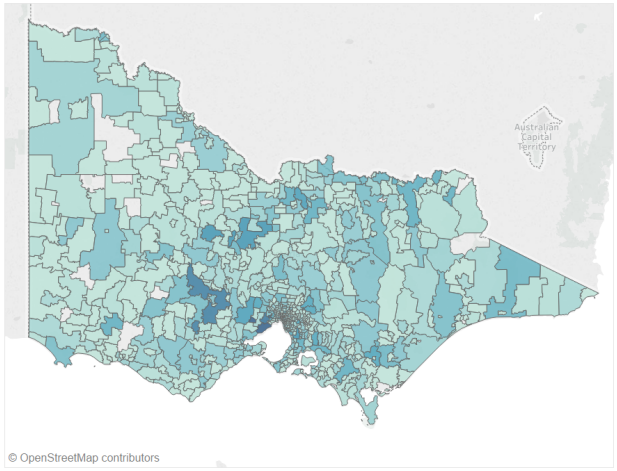


Figure 7: Total person killed by vehicles collision

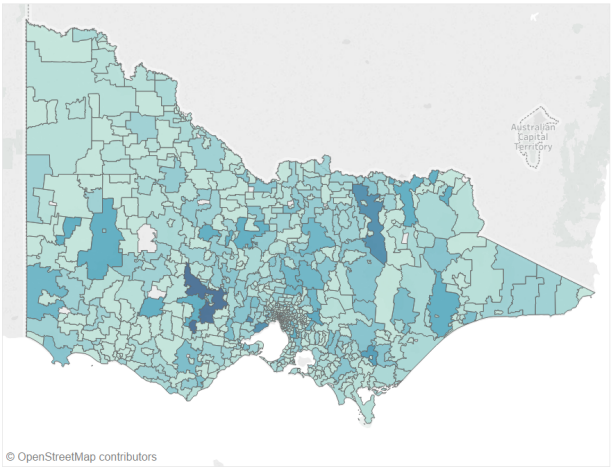


Figure 8: Total person killed by collision with a fixed object

As we can see from figure 8, the suburb with postcode 3552 (**BENDIGO**) always have the most number of the person killed for the top 2 accident types.

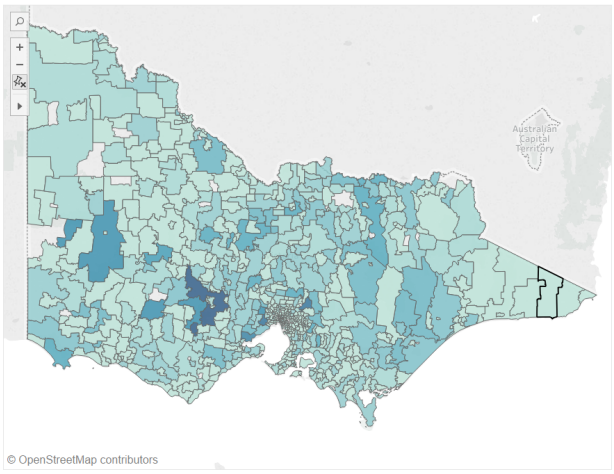


Figure 9: Dark street with NO light

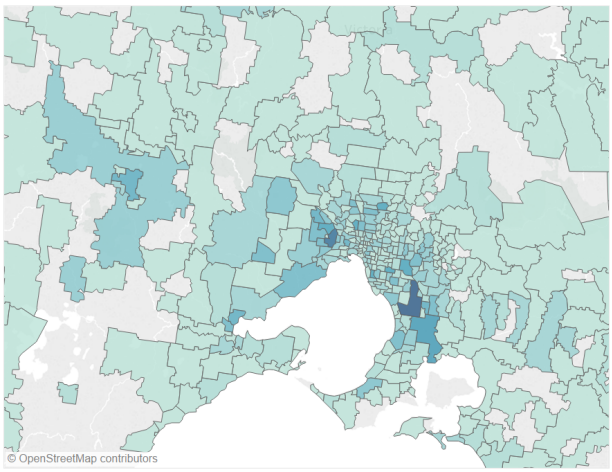


Figure 10: Dark street with light ON

The figure 9 indicates in the dark street the suburb with postcode 3552 (**BENDIGO**) got 15 people killed by the car accident, which is the highest record amount all. And figure 10 shows that the suburb of postcode 3175 (**BANGHOLME**) had 16 people killed. This result also reflects the construction of community street lights in different suburbs.

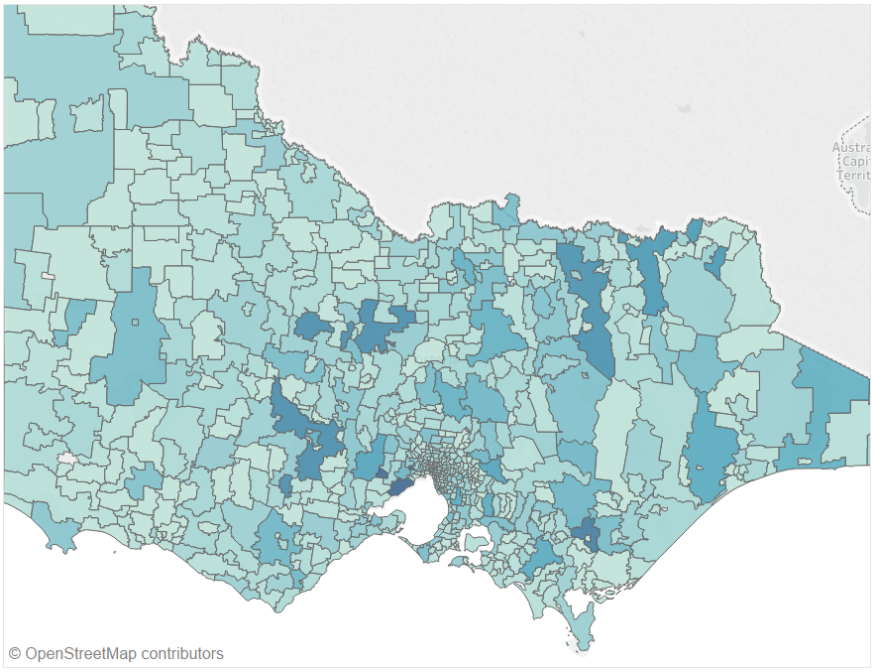


Figure 11: No.person killed in daytime

The statistic in figure 11 illustrates that in the middle and north-east Victoria state occurred more deadly accident than other places. Unfortunately, one more time, the suburb of postcode 3552 still in the list of top risk suburbs.

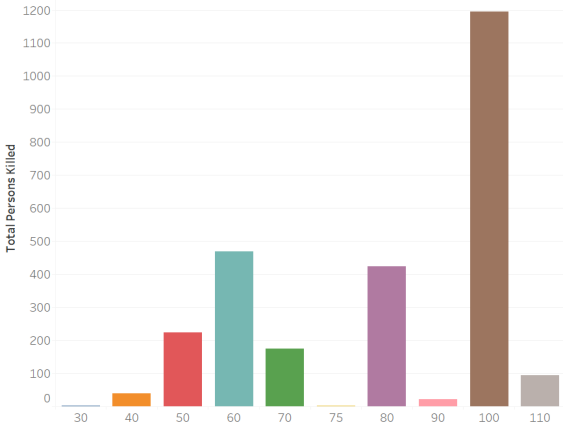


Figure 12: Total life loses in each speed zone

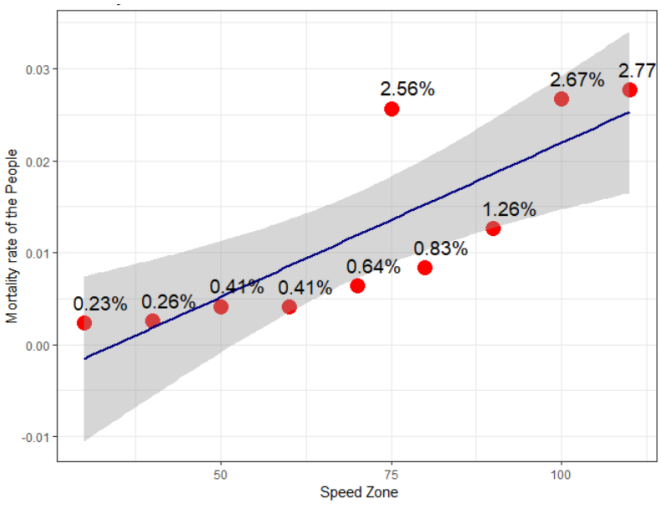


Figure 13: Percentage of people get killed in each speed zone

From figure 12 and figure 4.2, we could see that with the speed zone increase, people can easier lose their life in a car accident. Besides, most of life loses appeared in the speed zone of 100. The same result could be found in the accidents which are in the atmospheric condition is "**strong winds**", and surface condition are "**Dry, Wet, Muddy**". The analysis process is similar to the analysis of "Speed Zone", to save the paper, the detailed analysis process will not be shown in the report. Now let us move on to the cluster.

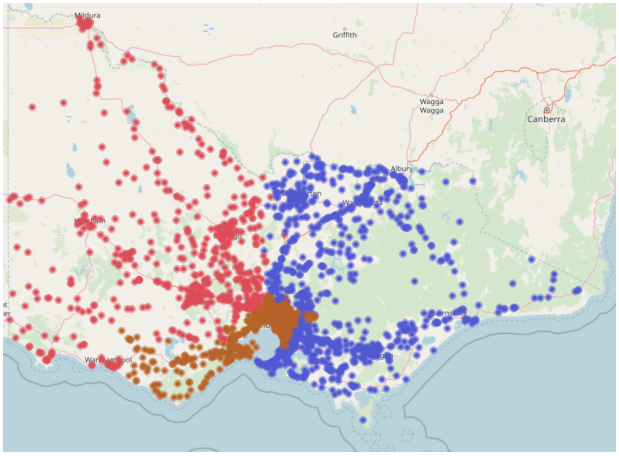


Figure 14: Clustering with postcode

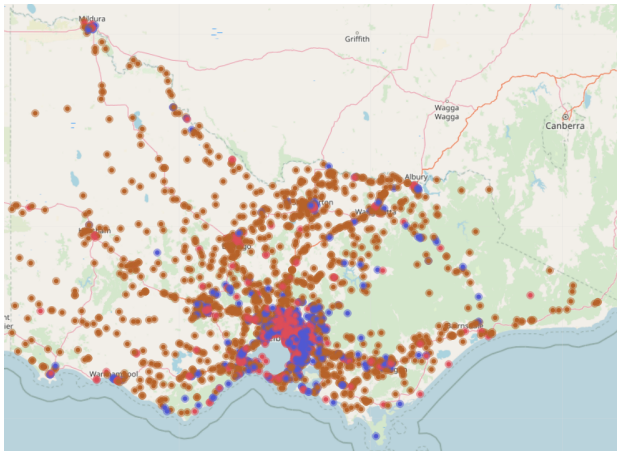


Figure 15: Clustering without postcode

I use R to cluster the dataset, first select the data with the year of 2009 - 2018, the number of people killed is greater than zero, speed zone is less than 120. Afterwards, select the clustering attributes: accident type, light condition, surface condition, and atmospheric conditions. Then the dataset is clustered into 3 clusters. The figure 14 shows that if we consider postcode as a clustering attribute, each cluster is classified by region. If the postcode is not considered, new clusters are classified by geographic location: near the Melbourne city and outside the city. By analysis of the cluster, I found most of the accidents outside the Melbourne city are:

- Collision with a fixed object in the day time
- In the speed zone of 100
- On the dry road surface of a clear day

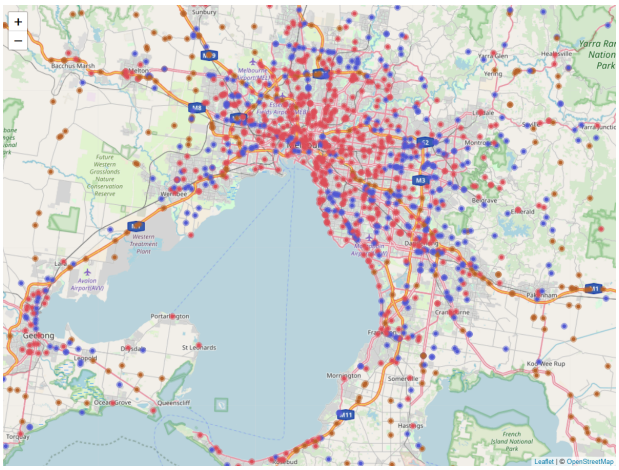


Figure 16: Clustering in Melbourne CBD

The figure 16 shows that the cluster type is highly consistent near the CBD of the Melbourne city. The analysis shows that most accidents which caused people killed happened in the speed zone of 60 where the vehicles collide with another one.

If we extract text words from "Accident.Type.Desc", "Day.Week.Description", "Day.Week.Description", "Light.Condition.Desc", "Surface.Cond.Desc" and "Atmosph.Cond.Desc" from the filtered dataset, and perform text analysis and visualisation, the result is shown in figure 4.2 on the right.

Figure 17: Word Cloud of deadly accidents

5 Conclusion

According to the analyses above, the following recommendations can be provided for the drivers and local councils in Victoria state:

- Bendigo, Cocoroc and Bangholme are the top 3 suburbs of people killed in accidents. Be more careful when driving through those suburbs.
- Compared to other suburbs, Bendigo has more number of people killed in various statistics, and the local council need to do more analysis and try to make some improvements.
- Collision with a fixed object is more deadly compared to other accident types.
- Strong winds can make driving more insecure, stop when encountering a strong wind.
- Road surface conditions of wet and muddy are more dangerous too, and people need to be more careful.
- Speed kills, pay more attention when driving at high speed.
- When driving in the Melbourne city, beware with other vehicles.
- When driving in the outskirt, beware of the road condition and ensure adequate rest. Inadequate attention when driving can easily cause the vehicle to deviate from the road, it is a fatal mistake on a high-speed road.

Due to the length limitation of the report, more detailed statistics and analyses are not carried on. However, for the future work, there still are lots of interesting topics, for example, how the sequence of event related to the person killed in a car accident, or the seating position could affect the probability of death. They are all worth to explore.

6 Reflection

I learned a lot in this project, from the topic selection to the dataset searching, then I tried to perform data wrangling, data cleaning and data checking with the dataset. Those are all new to me, however, at the end of those processes, I learned how to make the data more usable and how to extract useful information for data analysis in an extensive amount dataset. In the data analysis process, choose the right method is critical for the result, for example, The death number in the speed zone of 90 is quite low, which is because there are few roads with speed limit of 90. However, the percentage of people get killed in each speed zone shows an accident in a speed zone of 90 is more lethal than the speed zone from 30 to 80. After that, choose which method to demonstrate the result is another challenge too. To sum up, this is a very memorable experience.

References

TAC. (2019). *Lives lost - year to date.* Retrieved 2019-4-20, from `http://www.tac.vic.gov.au/road-safety/statistics/lives-lost-year-to-date`

Tomar, S. S. (2019). *A comprehensive introduction to data wrangling.* Retrieved 2019-4-20, from `https://www.springboard.com/blog/data-wrangling/`

VicRoads. (2019). *Crash stats - data extract - open data.* Retrieved 2019-4-20, from `http://data.vicroads.vic.gov.au/metadata/Crash%20Stats%20-%20Data%20Extract%20-%20Open%20Data.html`

WHO. (2019). *The top 10 causes of death.* Retrieved 2018-05-24, from `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death/`