# City of Yarra: Which month can you expect a heavy volume of traffic and witness hoons?

Author: Taemine Song

## ABSTRACT

Traffic counts data provides valuable insights on many accounts. These include but not limited to road planning, monitoring of critical infrastructures such as bridges, and ensuring safety of the community. This report focuses on investigating the volume of traffic and 85$^{th}$ percentile speed recorded in each suburb of Yarra municipality in Victoria. With this investigation, more efficient measures can be placed on streets where speeding is likely to occur and at specific times of the year. The data set was recorded over the period of August 2009 to May 2019 and includes road names, related sections for that road, average volume of vehicles in each direction. A various number of tools is used to clean and process the data *(is.na(), select ())* before wrangling and manipulating *(group_by(), summarise(), filter())* the dataset in order to analyse and visualize the insights. As the dataset contains missing observations for particular months and have already been removed, there needed a strategy around what to do about the missing records. Whilst imputing for these missing values is possible, this report will simply investigate suburbs with the least amount of missing data. The initial hypothesis with this study is that the holiday periods (December, January) will display lower number of traffic but greater average speed on the roads. This analysis relies on visualizing the traffic and average speed information to deliver its targeted insights (line plots). With the analyses performed in this report, it was found that drivers of the city of Yarra are more likely to speed around May and August. On the other hand, this community can expect heavy number of traffic around October and November. It is also interesting to note the relationship between the volume of traffic and the average speed recorded as they show almost a negative correlation, which will not be discussed in this report.

## INTRODUCTION

Traffic data is often neglected more so in the modern society with the pandemic keeping us all at home. However, it is still a priceless set of information and must be taken into consideration when regulating the roads and setting necessary measures to digest heavy traffic. The roads still must be made safe and speeding and other regulations must be maintained. With this set of data containing over a decade of information, useful information can be obtained around how much traffic we can expect and when and where the drivers are likely to speed. This report aims to help initiate projects that monitor and implement road safety measures.

# DATA DESCRIPTION

This dataset contains 1521 observations of traffic counts with 11 variables within different suburbs of the Yarra city in Victoria, which is a home to a diverse community of about 94,000 people and an area of 19.5 square kilometres. The count was done monitoring the start and end of different roads, noting the direction and speed of the traffic for the duration of a month. This dataset was published in August 2019 and is publicly available on data.gov.au. According to Yarra City Council, they declare that all due care has been taken to ensure the dataset is accurate but they do not warrant that this data is definitive nor free of error. The table below (Table 1) was obtained from the data dictionary on data.gov.au that shows details of the data before any pre-processing was applied.

| Column | Type | Label | Description |
|---|---|---|---|
| date_captured | text | Date captured | Date the traffic count was done |
| road_name | text | Road name | Name of the road |
| section_start | text | Section start | |
| section_end | text | Section end | |
| suburb | text | Suburb | Suburb where the traffic count was done |
| direction_1 | text | Direction 1 | One of: North, East, South, West |
| vehicles_1 | text | Vehicles 1 | Number of vehicles counted for direction 1 |
| direction_2 | text | Direction 2 | One of: North, East, South, West |
| vehicles_2 | text | Vehicles 2 | Number of vehicles counted for direction 2 |
| volume_per_day | text | Total volume | Combined volume for both directions |
| 85th_percentile_speed | text | Speed | 85th percentile speed |

Table 1: table of description of each variable

As can be seen in the table, the type of every data recorded is originally text. During the import of this dataset in R, variables including 85th_percentile_speed, volume_per_day, vehicles_1, and vehicles_2 were converted into numeric data. Inspecting the raw data, it is noted that some months are not recorded for some suburbs in particular years. These values will not be imputed and be simply ignored.

# METHODS

A series of data pre-processing, data wrangling, and exploratory visualisation methods were applied in quest of finding out on which month of the year at which section of the road is the busiest and on which road are the drivers speeding more than usual. The R codes of version *R-4.1.0* are referenced with the brackets in this section which can be looked up in *Appendix A*.

### PART A: DATA PRE-PROCESSING (A of Appendix A):

The data set, *yarra-traffic-count.csv*, was imported with *read.csv()*, allowing the header to be recognized. The dataset is then named 'data' and *str()* is run to check the type of 11 variables. It was found that the variables whose values were numeric were imported as numeric values and characters as characters. Any blank spaces were also trimmed with *str_trim* from *stringr* library.

The next part of the pre-processing was to check for missing observations and implement a strategy to impute for the missing values. *Summary()* was used to check for missing values for each column. There were no missing values for each column.

Next, the *data_captured* variable was required to be split into two variables *'data_captured_year'* and *'data_captured_month'* to enable more robust data wrangling exercises. This process was performed with several data manipulation techniques. Firstly, the original type of the variable 'YEAR-Month' (e.g 2009-Aug) was a character and made into a numeric value using *AsDate()* and *substr()* from the *'flipTime'* library (refer X). Then, *mutate()* function from '*dplyr'* package created these two variables. These two steps yield in two variables as intended but their types were charactered. They were coerced into the numeric type using *as.numeric()*.

### PART B: DATA WRANGLING & MANIPULATION (B of Appendix A)

After cleaning the data, there needed a grouping of particular variables to better visualize and compute insightful information. Firstly, to investigate the dataset by each month, *data_grouped* was created using *group_by* function, followed by *summarise* function to compute and display the mean value of both *volume_per_day* and *$85^{th}$_percentile_speed*. As the dataset has missing observations in some months for particular suburbs, observations in Collingwood, Fitzroy North, Richmond, and Carlton North have been chosen for this study as they contain least amount of missing data.

### PART C: VISUALISATION (C of Appendix A)

The processed and filtered dataset is visualized using *ggplot* from *ggplot2* library as a line graph. The line graph was chosen due to the discrete nature of the variables in discussion with *geom_line().* As can be seen in both figure 1 and 2, one plot was dedicated to each suburb collating each year's data. This is useful in glancing out the coordinates of the maximum and minimum points. Plotting multiple graphs was made possible by binding each year's data using *bind_rows()*. The axis and title were also appropriately named using *ggplot*'s feature *labs().* The x-axis was coerced into a discrete scale of numbers between 1 – 12 to represent the months in a year using *scale_x_discrete()*.

# RESULTS AND DISCUSSION

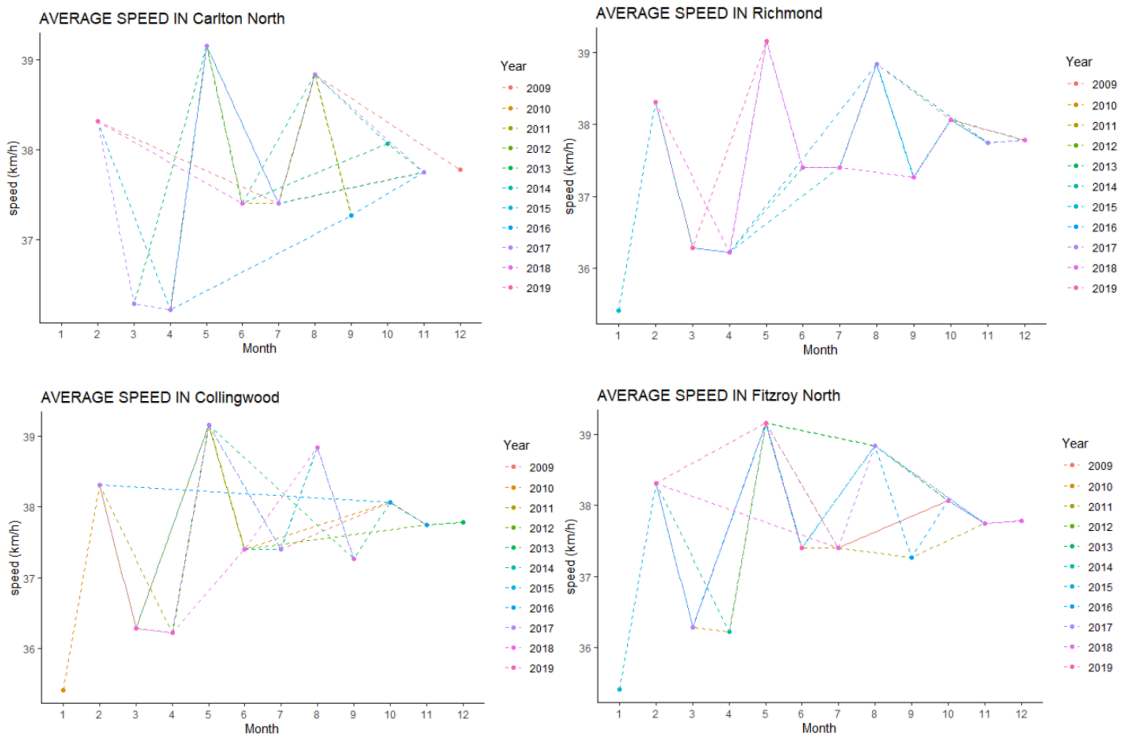


Figure 1: Average Speed recorded in each suburb by year

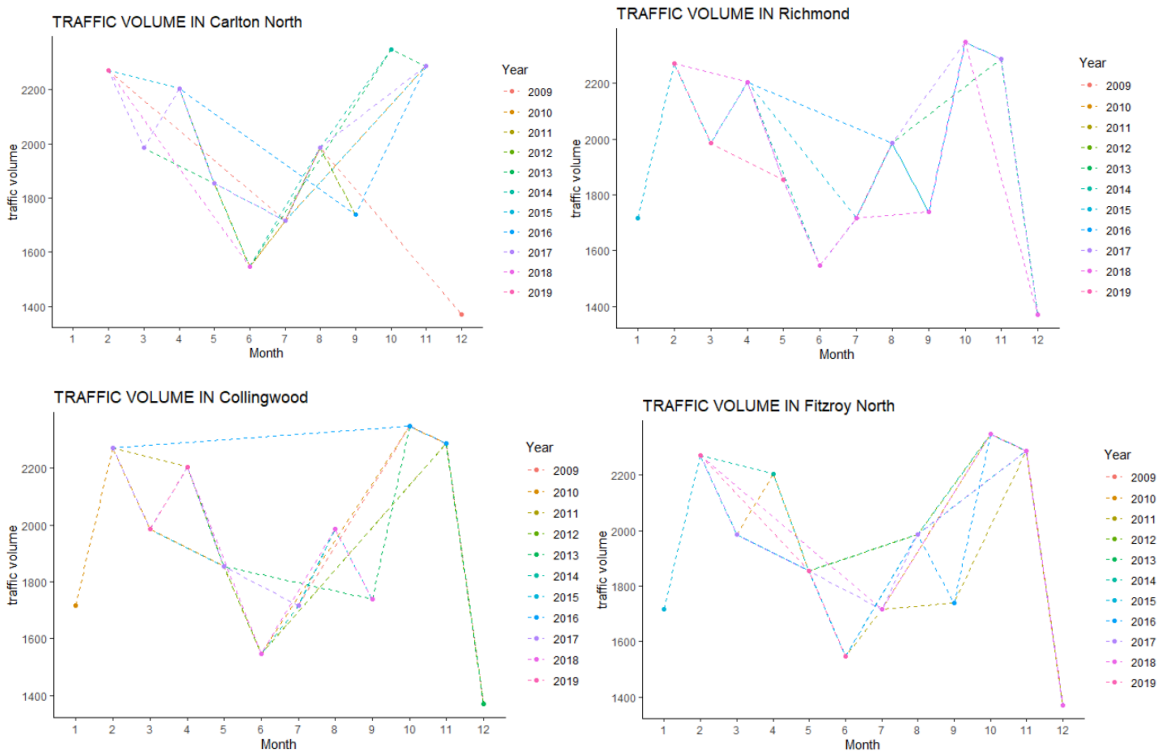


Figure 2: Average traffic volume recorded in each suburb by year

The results above show visual information regarding the average speed and the volume of traffic in each of the 4 suburbs, Figure 1 and Figure 2, respectively. In Figure 1, it can be seen that these 4 suburbs have similar patterns in how the average speed varies throughout the year. January, March, and April display relatively low average speeds whereas the drivers are speeding more in May and August. These findings can be contributed to a number of factors such as volume of traffic on the road, driving under influence, and holiday period.

Similarly, Figure 2 illustrates that these four suburbs show similar patterns in how much traffic they have experienced throughout the year. People in the city of Yarra have seen lower number of traffic in the months of June and December but experienced higher number in February, October, and November. These finding could again be explained with a number of factors such as school holidays, sickness, and roadworks.

In a closer inspection, some graphs were not plotted over every month which is due to the nature of this dataset where there were no observations in some months. A prediction imputation of these missing values is believed to enhance the quality of the insight included in this report.

Comparing two figures, it is interesting to note that they have a negative correlation to some extent which can be justified using correlation test. By intuition, it does make sense that most drivers will speed with less traffic on the road rather than with heavy traffic.

# CONCLUSIONS

In conclusion, traffic data although specific to one municipal has proven to contain many valuable insights which can be taken into account when making initiatives or decisions for the roads in the city of Yarra. This report conducted a series of techniques to process and manipulate the data to comprehensively visualize the data and draw conclusions from it. The initial hypothesis that December and January will show less number of traffic but greater speeds on the road was partially proven correct as they do show relatively lower volume of traffic and greater speed. However, further studies and analyses are required to compare the statistical data between each month and to also investigate the correlation between volume of traffic and the average speed.

# REFERENCES


- STHDA: Statistical tools for high-throughput data analysis (http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles)
- Data .gov.au: City of yarra traffic counts (https://data.gov.au/data/dataset/yarra-traffic-counts/resource/9e26683b-6b30-424e-ace7-59047d811d1c?view_id=5abed666-4533-4622-b5d3-67f0271f3b10)
- R for Data Science by Hadley Wickham & Garrett Grolemund

# APPENDICES A – R CODE

A:

```
install.packages('lubridate')
library(lubridate)
install.packages('devtools')
require(devtools)
install_github("Displayr/flipTime")
install.packages('flipTime')
library('flipTime')
library('stringr')


#Mutate data_capture into two variables, 'date_captured_month' and 'date_captured_year'
data_dates_mutated <- mutate(data,
            date_captured_month = as.integer(substr(AsDate(癤ate_captured), start = 6, stop = 7)),
            date_captured_year  = as.integer(substr(AsDate(癤ate_captured), start = 1, stop = 4))
                             )

#trim any white spaces
data[sapply(data, str_trim)]
```

B:

```
#Group data
data_grouped <- data_dates_mutated %>%
  group_by(date_captured_month) %>%
  summarise(avg_volume = mean(volume_per_day), avg_speed = mean(X85th_percentile_speed),
            suburb, year = date_captured_year)
```

C:

```
Abbotsford_speed = ggplot(bind_rows(filter(data_grouped, year ==2009 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2010 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2011 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2012 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2013 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2014 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2015 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2016 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2017 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2018 & suburb == 'Abbotsford'),
                                    filter(data_grouped, year ==2019 & suburb == 'Abbotsford'),
                                     ), value,
                          mapping = aes(x = date_captured_month, y = avg_speed,
                                        colour = factor(year))) + labs(x="Month", y="speed (km/h)",
                                                                       colour="Year") +
                                                ggtitle("AVERAGE SPEED IN ABBOTSFORD")

Abbotsford_speed = Abbotsford_speed + geom_line(linetype = "dashed") + geom_point() + theme_classic()
Abbotsford_speed = Abbotsford_speed + scale_x_discrete(limit = factor(c(1,2,3,4,5,6,7,8,9,10,11,12)))
Abbotsford_speed
```

```
Fitzroy_North_traffic = ggplot(bind_rows(filter(data_grouped, year ==2009 & suburb == 'Fitzroy North')
                                    filter(data_grouped, year ==2010 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2011 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2012 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2013 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2014 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2015 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2016 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2017 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2018 & suburb == 'Fitzroy North'),
                                    filter(data_grouped, year ==2019 & suburb == 'Fitzroy North'),
                ), value,
mapping = aes(x = date_captured_month, y = avg_volume,
              colour = factor(year))) + labs(x="Month", y="traffic volume",
                                             colour="Year") +
  ggtitle("TRAFFIC VOLUME IN Fitzroy North")

Fitzroy_North_traffic = Fitzroy_North_traffic + geom_line(linetype = "dashed") + geom_point() + theme_
Fitzroy_North_traffic = Fitzroy_North_traffic + scale_x_discrete(limit = factor(c(1,2,3,4,5,6,7,8,9,10
Fitzroy_North_traffic
```