COMP20008 Elements of Data Processing
Semester 1 2016 Project – Phase 4
Ivan Ken Weng Chee 736901

Project Title

Traffic Congestion and Efficiency in Victoria

Domain

Transport

Question

Why and where do traffic congestions and bottlenecks happen and how can we improve its efficiency?

Datasets

- **Road Width and Number of Lanes (VicRoads Open Data)**
  - Shows the road and shoulder widths and number of traffic lanes on freeways and arterial roads
  - *http://vicroadsopendata.vicroadsmaps.opendata.arcgis.com/datasets/24ccad5c745e4addabfcfb32c400ee83_0*
  This dataset contained the number of lanes and present in Victorian roads, as well as their relative widths, shoulder widths and seal widths. Also present is the direction of the road (forwards and / or backwards).

- **Victorian Road Traffic Volumes (AURIN)**
  - Contains road traffic volumes for freeways and arterial roads
  - *http://data.aurin.org.au/dataset/vic-govt-vicroads-vicroads-evol-mar13-na*
  Traffic related information such as average annual vehicle and truck count, growth rate, year of measurement, traffic flow direction, and gps coordinates of the midpoint of Vicorian roads are present in this dataset.

- **VicRoads Speed Data by Road Segment (Victorian Government Open Data Repository)**
  - Records the latest typical hourly speed data by road segment in kilometres per hour
  - *https://www.data.vic.gov.au/data/dataset/vicroads-speed-data-by-road-segment*
  24-hour average vehicle speed data on Victorian roads can be found in this dataset, as well as their speed limits, gps coordinates, number of traffic surveys conducted and the year in which it was measured.

Pre-processing

- Inspecting:

I first inspected the data by randomly looking around the dataset to determine what methods needed to be applied. For example, whether there were missing data, how naming conventions were used, and what format it was in.

- Parsing:

There were several illegal characters present due to different encodings, and I had to convert these datasets into the standard UTF-8. Traditional csv methods were not sufficient in processing the data due to the sheer size of the datasets (15k – 200k rows), so I used the pandas library which converted the data into hash tables for better efficiency.

- Cleaning:

The three datasets had a common column (Road Name) to be merged by, but different naming conventions and abbreviations used caused difficulties in matching them. I stripped non-relevant characters and converted abbreviations to their full names in order to normalise these datasets for string matching.

- Combining:

Some datasets had duplicate entries and multiple instances of each road (separated by different segments and timestamps). I decided to combine these into single instances by applying the mean function on their numerical data.

- Aggregating:

Depending on the state of the data, I saved datasets upon different levels of processing so as to prevent lost of important data. For example, calculating the mean of column such as (Year) would simply render that column useless. I've also separated the data into two versions, one containing GPS coordinates and another the numerical data, since it would not make sense to group them together.

Integration

- Normalising:

I've normalised the data by applying a standard naming convention across datasets and disregarded columns that would not be feasible to combining. I've chose the column (Road Name) to be the pivot in concatenating the data, and therefore ensuring that the casing and whitespace content of this column matches across all datasets.

- Concatenating:

Based on the column (Road Name), I've combined the matching roads into a larger dataset to work with. Matching using the pandas merge function produced a dataset containing all columns from the original dataset. Columns deemed unsuitable for merging were disregarded.

- Expanding:

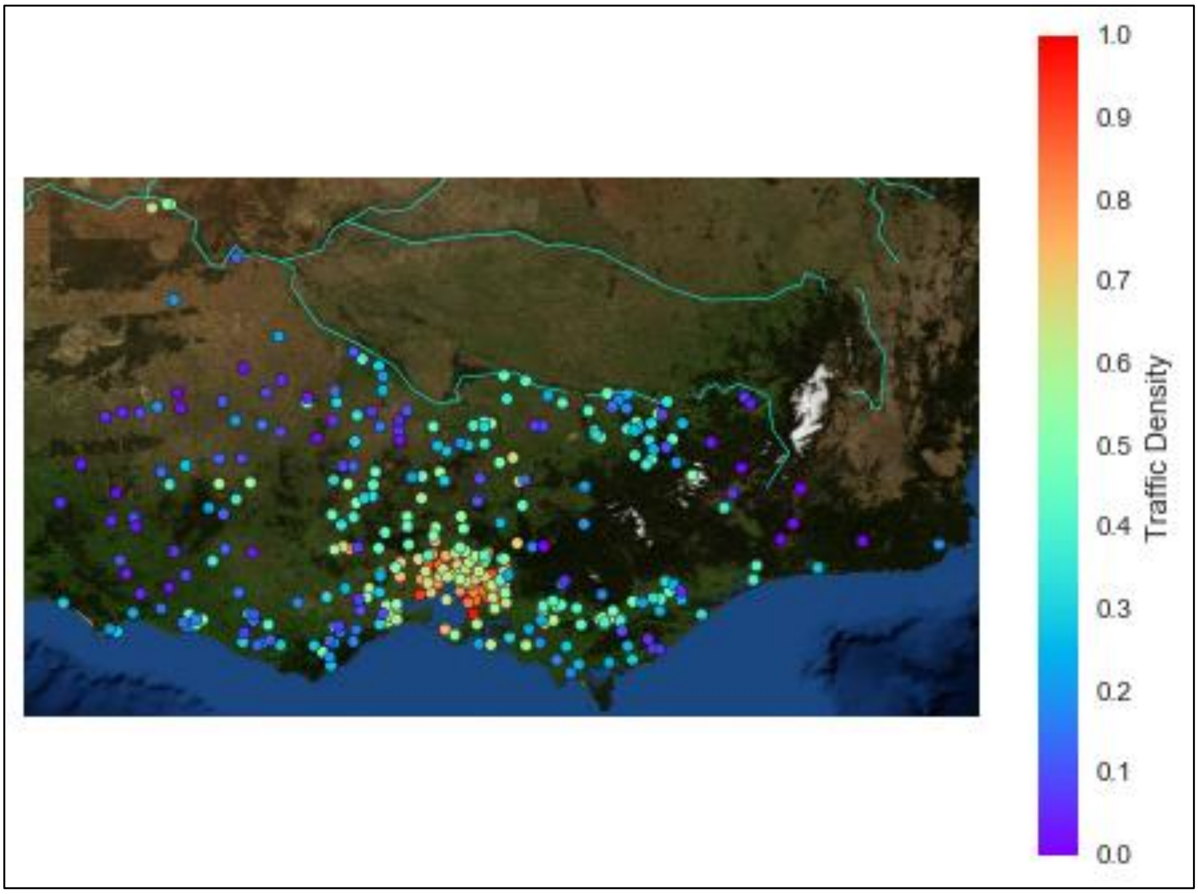
New columns which are relevant to the produced dataset were added to give a better overall picture. In some instances it made more sense to multiply or divide certain columns and store them instead to better fit the new data. These extra columns can help in categorising the data according to different attributes.

- Limitations:

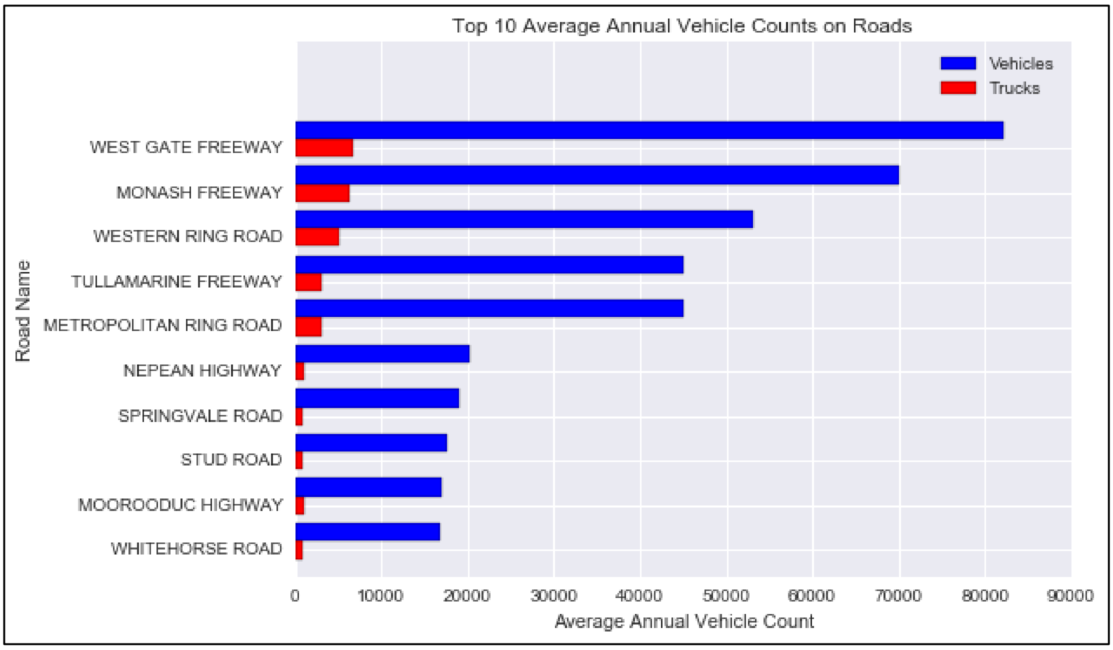
However, I am aware that a large number of roads may not match correctly depending on how well I pre-processed the data, and certain trade-offs between precision and relevancy had to be made. There is most probably some form redundancy that would occur when I cleaned and normalised the data since these datasets did not have a proper description and I was forced to decide by inspection on how to wrangle the data. Data lost through cleaning and merging could possibly cause skewness in visualisations which may not accurately represent the actual data.

Results

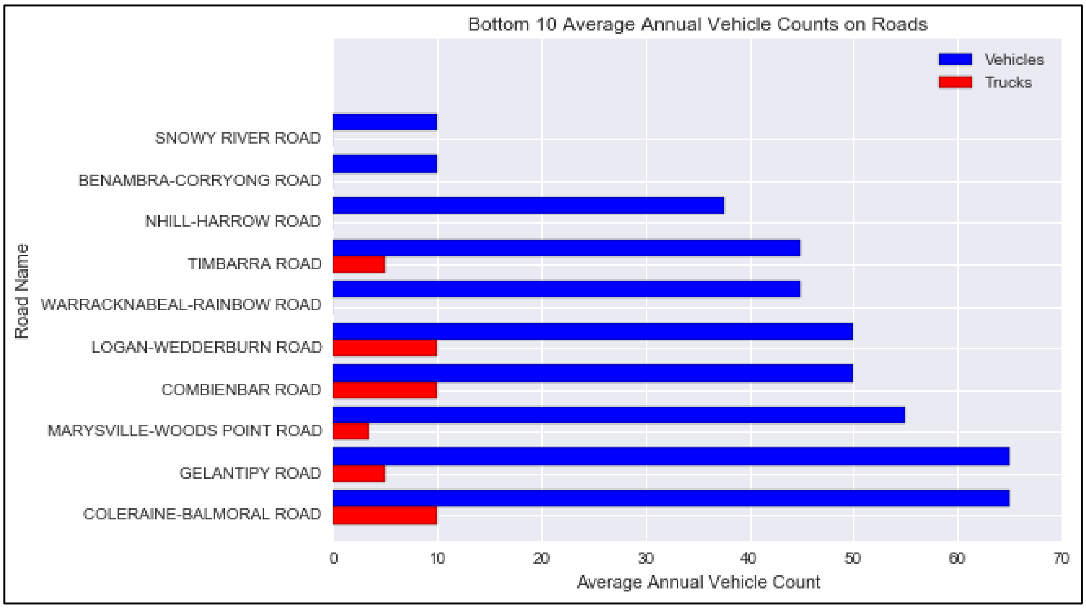
This map of Victoria was plotted using Basemap and the corresponding coordinates found in the datasets. The colours indicate the vehicle density. As expected, density in areas around the Melbourne CBD are is at its peak, as opposed to that of regional Victoria. Naturally, congestion frequently occur in areas of high traffic density, so efforts to monitor and ease traffic flow should be focused mainly on those areas.



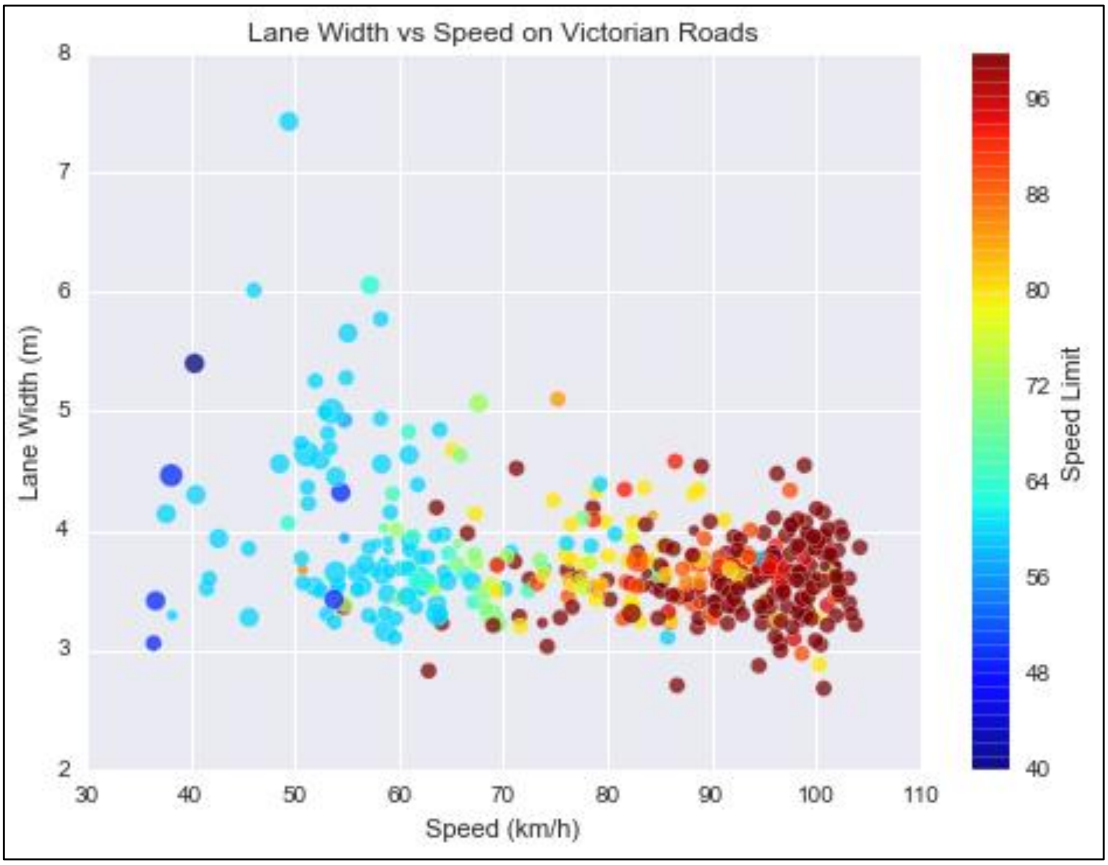
Massive amounts of traffic flow through Victoria each day. High-traffic volume roads are highly susceptible to traffic congestions. They also require more maintenance to function normally, especially with large amounts of trucks on the move. These are the top 10 highest traffic volume roads:



On the other hand, roads with low-traffic volume represent areas which drivers do not travel through often, possibly due to their location. However, these roads would be useful for traffic redirection, if the main roads are unable to accommodate traffic during peak hour or maintenance. These are the top 10 lowest traffic volume roads:

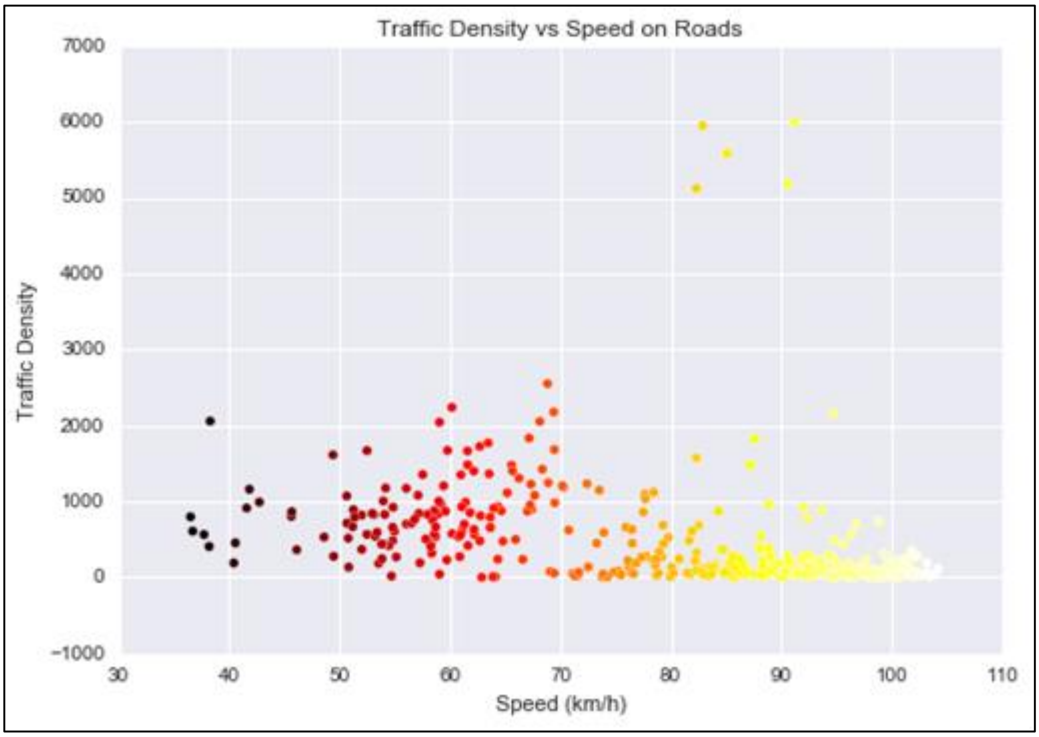


The number of traffic lanes often reflect how much traffic we would predict that flows through them. Lanes on a road are usually decided prior to construction since expanding roads is not easy and convenient due to factors such as the terrain and limitation of space due to development. Highways and freeways tend to have a higher number of lanes, as well as a higher speed limit. The following is a plot on the average lane width against speed:



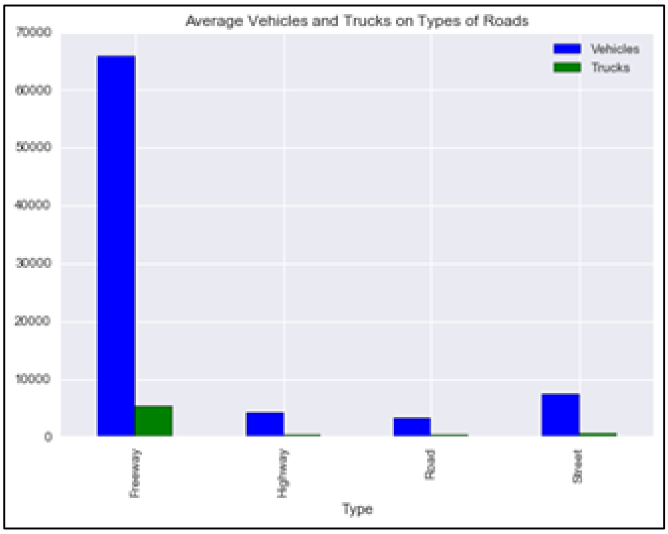
Here we see that vehicles tend to travel faster on narrower roads, and slower on roads which are wide. In general, higher speed limits are applied to highways and freeways, and lower limits on arterial roads. This would mean that highways and freeways are narrower than the other roads, and they should be wider to maintain safety since vehicles are travelling faster.

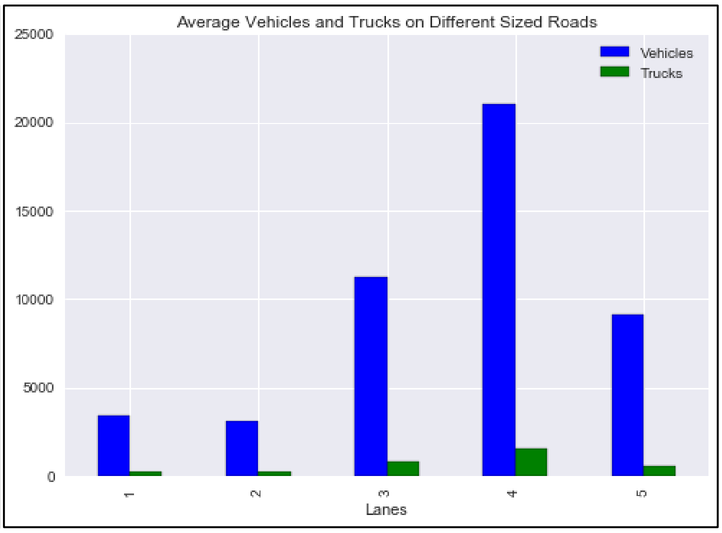
Traffic density is the average vehicle count over the number of lanes of a road. This scatterplot shows the lane density of traffic on Victorian roads. From this we can infer that denser roads with lower speed are more likely to be congested. Most of the roads have a similar amount of density, but there is a group of outliers which are extremely dense and also allow vehicles to travel at a high speed. While they are efficient, these roads should be monitored since an increase in traffic would be likely to bog them down.



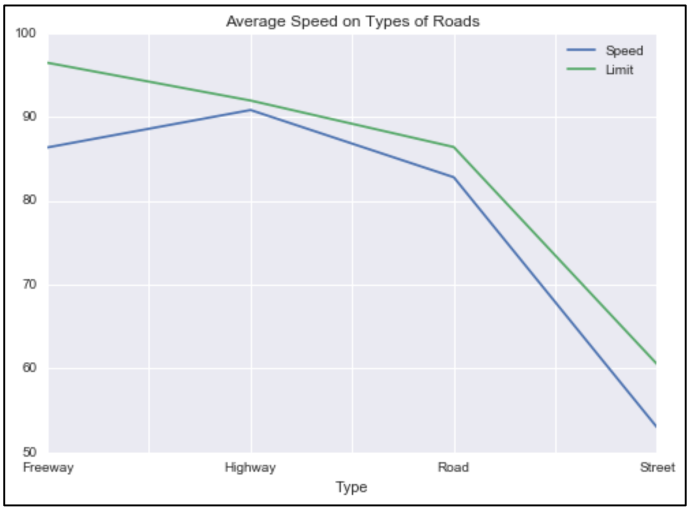
Different types of roads have different traffic patterns. These plots show that drivers like to travel on freeways as opposed to highways, possibly to avoid toll fees.

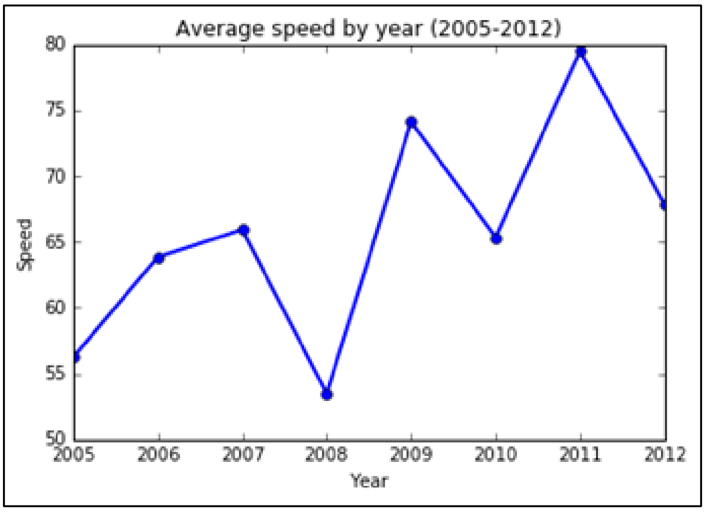
Congestion can occur when drivers have preferable routes, as some might prefer to save money over time. In this case, highways are not functioning at their optimum due to lesser convenient placement compared to freeways.





Speed limits are indicative of how fast cars are predicted to travel on these roads. Drivers tend to drive close to speed limits to save time, and therefore we can see that there is a large gap with speed travelled and speed limit on freeways. This would mean traffic jams occur more often on freeways. The graph on the right shows the average speed travelled from the year 2005-2012. There is an increasing trend, possibly due to more and more highways and freeways being built. However, there is an outlier in 2008 where speed was at its minimum. Higher average speed is a faint indicator that traffic is progressively becoming more efficient.





This is a pivot table indexed on the different types of roads and number of lanes.

type	lanes	limit	speed	trucks	vehicles	width
Road	1	85.95146	80.2383	255.2798	3126.997	7.263461
	2	88.46459	85.02274	215.0143	2562.859	8.13919
	3	64.10606	59.97144	734.1252	10244.59	12.85417
	4	63.33333	63.73175	551.9444	10136.11	13.93009
	5	60	51.15625	797.5	13925	23.2
Street	1	61.66667	50.52754	403.75	7050	7.511515
	2	61.19216	53.66811	546.1688	7503.86	9.919183
	3	60	53.56334	604.8084	7577.772	14.3274
	4	55	46.04971	516.5126	6871.569	16.62955
	5	60	53.5138	420	4383	25
Freeway	1	0	0	0	0	0
	2	100	85.06724	3034.545	45125.71	13.83362
	3	100	91.18153	6237.647	69988.24	16.98249
	4	89.41176	82.8596	6680	82057.14	18.89255
	5	0	0	0	0	0
Highway	1	98.4466	97.54631	216.8444	1333.133	8.29217
	2	92.29617	91.17917	405.6702	3711.664	9.719586
	3	70	68.11319	1037.065	20146.74	10.90746
	4	0	0	0	0	0
	5	0	0	0	0	0

Value

The wrangling done on these datasets would be beneficial in tracking down key areas that lead to traffic slowdown. Data from all three datasets can now be compared against and different inferences can be made about traffic. The raw data has data columns only pertaining to its type of data, as compared to the concatenated data which contains all columns and columns derived from the originals. In a way, the concatenated data is about viewing traffic data from a higher perspective. Since us humans are visual creatures, having visualisations make it easier to grasp the concept of the data and to what extent it might affect certain things.

Challenges and Reflections

Initially, I reached a dead end in trying to process the data by iterating through in csv form, due to the sheer size of the dataset. I chose to use pandas instead for the quickness and power which its DataFrame and Series had to offer, even though I had to spend a significant amount of time learning to use this library. The process of combining these datasets proved to be unfruitful as well, as the data in these datasets had mismatching naming conventions and column names were not self-explanatory. I also found difficulty in visualising the data in terms of my question, possibly due to my selection of datasets which were too general.

Question Resolution

The Victorian Government would probably find interest in my results as this would help ensure Victoria would be able to continuously deal with the increase in the driving population, and reduce the carbon footprint of the state by lessening the amount of time vehicles stay running on the road.

Also, transport organisations like VicRoads might find my results useful since they are the ones managing traffic directly. Information obtained could help them determine where and how to employ their new SmartRoads concept, which prioritises different types of vehicles during different times of the day to ensure maximum efficiency of roads. Above all, citizens of Victoria who find themselves losing time due to slow traffic would benefit as the traffic conditions of Victoria improve and become more manageable.

Code

The length of code I wrote is around 400 lines, including comments. Python proved to be an excellent tool for this task, due to the vast collection of libraries available.

The libraries I used include:

• Codecs	(Stripping unwanted characters contained in the datasets and formatting them correctly)
• Matlotlib	(Visualising graphs and plotting most of the results)
• Basemap	(Import map of Victoria as background for gps coordinate plotting)
• Numpy	(Assisting matlotlib in visualisation and generating specialised number types)
• Pandas	(Powerful csv processing tool to convert datasets into hashtables for quick access)
• Pylab	(Another powerful visualising tool for visualising clusters)
• Scipy	(Aids in standardising data)
• Seaborn	(Produced a heatmap of the data)
• Sklearn	(Helps with clustering and fitting data)
• Visual	(Taken from Week 6 workshop for generating VAT)

Bibliography

- Practical Business Python
  http://pbpython.com/pandas-pivot-table-explained.html
- Visualization: Mapping Global Earthquake Activity
  http://introtopython.org/visualization_earthquakes.html
- Trends in Melbourne Traffic
  https://chartingtransport.com/2010/10/31/trends-in-melbourne-traffic/