ELEMENTS OF DATA PROCESSING ASSIGNMENT 2

By Harold Liu (1166132), Andy Ng (1170648), Elizabeth Wong (1166527) and Desmond Wong (1082634)

How is population growth impacting Victorian Roads?

Introduction

The last few decades have seen rapid growth in Victoria. Population growth, which contributes to economic development, could place a burden on the state's infrastructures - the road network. This project aims to evaluate the effect of population growth on Victoria's road network, which is essential to the normal functioning and future growth of Victorian towns and cities. By identifying such effects, stakeholders, such as state and local governments, may amend the road network accordingly, assisting with the improvement of liveability in Victoria and promoting sustainable state development.

Datasets

The following datasets are used:

- Motor Vehicle Census, Australia
    - https://www.abs.gov.au/statistics/industry/tourism-and-transport/motor-vehicle-census-australia
    - Motor vehicle registration data collected by ABS. Vehicle registration is given yearly sorted by different parameters. Registration by categories is used in section 2.
- National, state and territory population
    - https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/
    - Used %, which was renamed to Population_Growth. Local Government Area renamed to LGA_SHORT_NM. Creates Victorian_Population.csv.
- Standard Postcode File
    - https://postcode.auspost.com.au/product_display.html?id=1
    - Used in the initial stages as a trial.

- Traffic Signal Volume Data
    - https://discover.data.vic.gov.au/dataset/traffic-signal-volume-data
    - Vehicle volume collected by detectors in location areas in intervals of 15 minutes. Locations were identified through SCAT-site numbers, which refer to locations on a Melway map grid.
- Traffic Volume
    - https://discover.data.vic.gov.au/dataset/traffic-volume
    - Renamed Growth as TRAF_GROWTH and used LGA_SHORT_NM.

The datasets were downloaded in csv format and transformed into dataframes using pandas. The different datasets regarding variables within road networks were linked together by comparing the population growth, over multiple years, and an in-depth analysis of a singular year.

Analysis and Wrangling Methods

Section 1: (Traffic Signal Volume Data and Population Growth)

Due to inconsistency of file zipping, folders were manually extracted for their months and ordered by respective year. Due excessive amounts of data, 90 days were randomly selected as a sample of each year and concatenated into csv. Similarly, due to this issue, only the presumed “peak hours" and daily volume of different Victorian locations were kept, as they represent when there is most congestion. Throughout this process, locations with the 100 highest daily volumes were kept, as presumably those are vital areas which are the most visited. For top areas, the assumed peak hours of 8am to 9 am and 3pm to 6:30 pm (Butt, C., Jack, T., & Frederiksen, S., 2018) were calculated by adding up their corresponding 15-minute interval recordings, ignoring all NaN and negative values. When all years had been processed, they were concatenated into one csv, containing the filtered data for all years. (More details on specific file requirements in README).

Victorian Population data: Due to the lack of a singular dataset providing the required population data, manual collection from the ABS was completed and recorded in the csv.

Section 2: Motor Vehicle Registration vs. Population

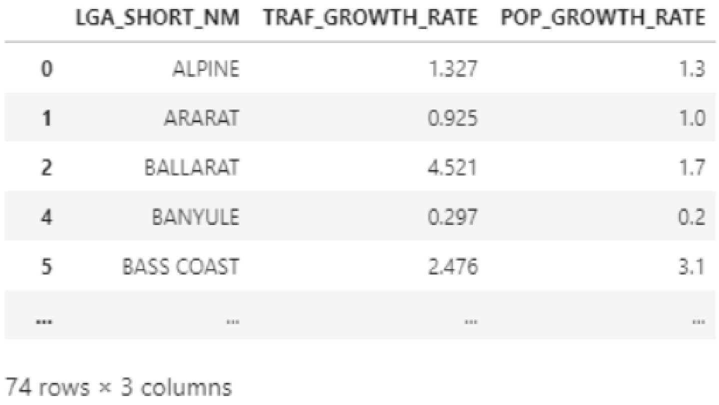
Motor Vehicle Census, Australia: Manual processed was used to merge registration data by vehicle category of the year 2014-2020 to a single csv file. Registration data by category was then processed to calculate growth by year and growth rate by year.

Victorian Population data: Same dataset of population is used for Section 1 and Section 2.

Section 3: Growth Rate Comparisons and Correlations

The raw csv files were pre-processed using the pandas library. All relevant fields were sliced out and renamed with .iloc(). Any empty NaN values due to the original csv file's format were removed with .dropna() and duplicates were dropped.

LGA_SHORT_NM was the common field used to merge Victorian_Population.csv and Traffic_Volume.csv with a right join. Due differences in each dataset's LGA_SHORT_NM, join selection was important to reduce data loss from joining.

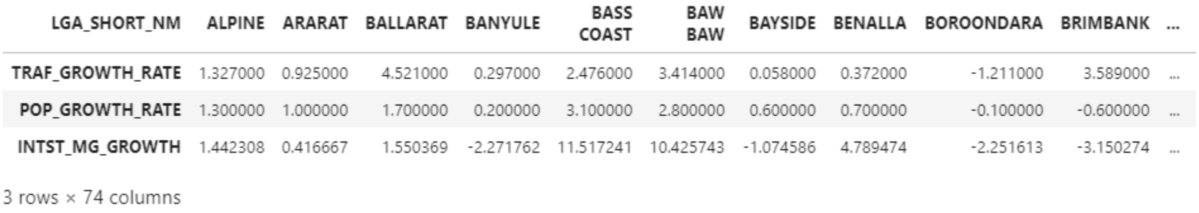


	LGA_SHORT_NM	TRAF_GROWTH_RATE	POP_GROWTH_RATE
0	ALPINE	1.327	1.3
1	ARARAT	0.925	1.0
2	BALLARAT	4.521	1.7
4	BANYULE	0.297	0.2
5	BASS COAST	2.476	3.1
...	...	...	...

74 rows × 3 columns

Traffic vs. population growth

Matplotlib graphed the Traffic_vs_Pop_Growth.png where traffic growth and population growth were displayed on the same graph. Migration growth was calculated between net internal migration no. and net overseas migration no. using the pct_change() function. Migration growth was the percentage of population in each LGA area that migrated outside or inside a Victorian suburb. A bar vs line graph was chosen to highlight the growth trend.



LGA_SHORT_NM	ALPINE	ARARAT	BALLARAT	BANYULE	BASS COAST	BAW BAW	BAYSIDE	BENALLA	BOROONDARA	BRIMBANK	...
TRAF_GROWTH_RATE	1.327000	0.925000	4.521000	0.297000	2.476000	3.414000	0.058000	0.372000	-1.211000	3.589000	...
POP_GROWTH_RATE	1.300000	1.000000	1.700000	0.200000	3.100000	2.800000	0.600000	0.700000	-0.100000	-0.600000	...
INTST_MG_GROWTH	1.442308	0.416667	1.550369	-2.271762	11.517241	10.425743	-1.074586	4.789474	-2.251613	-3.150274	...

3 rows × 74 columns

Traffic vs. population vs. migration growth rate

A heatmap was created using the seaborn library where a correlation matrix was formed with spearman’s correlation and visualised into a heatmap. The heatmap generated took a random sample of 15 LGA areas as using all 74 columns produced unclear results.

Population, traffic and migration growth were the variables examined. Spearman’s correlation was chosen as Pearson’s correlation was inappropriate as it was uncertain whether or not the relationship between the fields were linear. Additionally, the analysis data was suitable for Spearman’s as the variables had a scale, represented paired observations and there was a monotonic relationship between the growth rates (Statistics, L, 2018).

Key Results

Section 1: General Population to Vehicle Volume Growth:

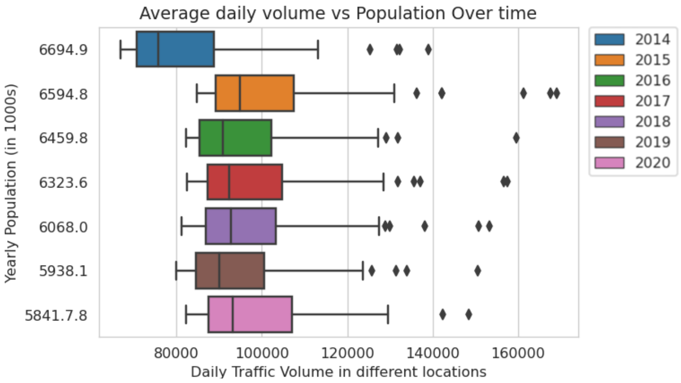


Figure 1: Average daily volume vs Population Over time

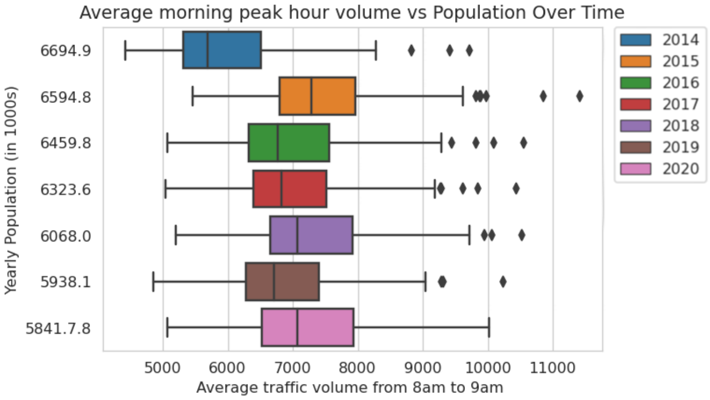


Figure 2: Average morning peak hour volume vs Population Over Time

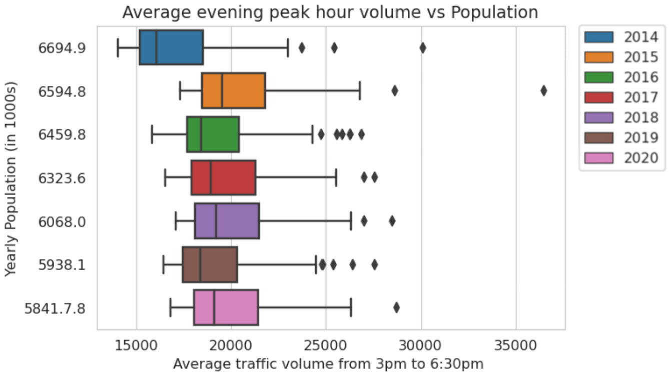


Figure 3: Average evening peak hour volumes vs Population

From Figure 1, despite a consistent increase in the general population within Victoria, it is not reflected in the median daily volume of the top 100 most visited roads, which merely fluctuated around 90,000. Similarly, the IQR (85,000-110,000) and whiskers (80,000-130,000) are nearly identical throughout the years. Notably, the top 2-3 locations' volumes, represented by the right-most diamonds, saw gradual increases over the years 2014-2019, possibly signifying an increase in traffic in these roads. 2020's traffic volumes are significantly lower, representative of the heavy restrictions throughout the year and are difficult to compare to other years.

For further investigation, Figure 2 and 3, the average volumes of peak hours throughout the sampled days were developed. Boxplot2 follows the trend of boxplot1, also showing the steady increase of volume within the top 2-3 busiest areas. Conversely, boxplot3 does not present any significant trends, with outlier locations varying throughout all years. These three boxplots together suggest that from 2014-2019; population increase may have influenced the overall daily traffic volume; with more of it being experienced in the morning peak hours. Nevertheless, with this dataset alone, a strong correlation has not yet been established.

Section 2: Motor Vehicle Registration vs. Population

Vehicle registration growth rate by vehicle type is plotted against Victorian population growth rate, as shown in figure #. In general, population growth can be positively associated with vehicle registration growth. Vehicle registration growth is largely proportional to that of population by a lesser percentage, potentially due to the availability of public transport which serves as an alternative, as well as population growth due to temporary residents (such as international students), who are less likely to own a vehicle. It is also worth noting that registration of Light Rigid vehicles (vehicles with GVM between 4.5 and 8 tons) has increased disproportionately, especially in the year of 2020, where the global pandemic of COVID-19 happened. This could indicate an increase in logistics demand of Victoria, especially during forced lockdown, where consumer orders were asked to be delivered.

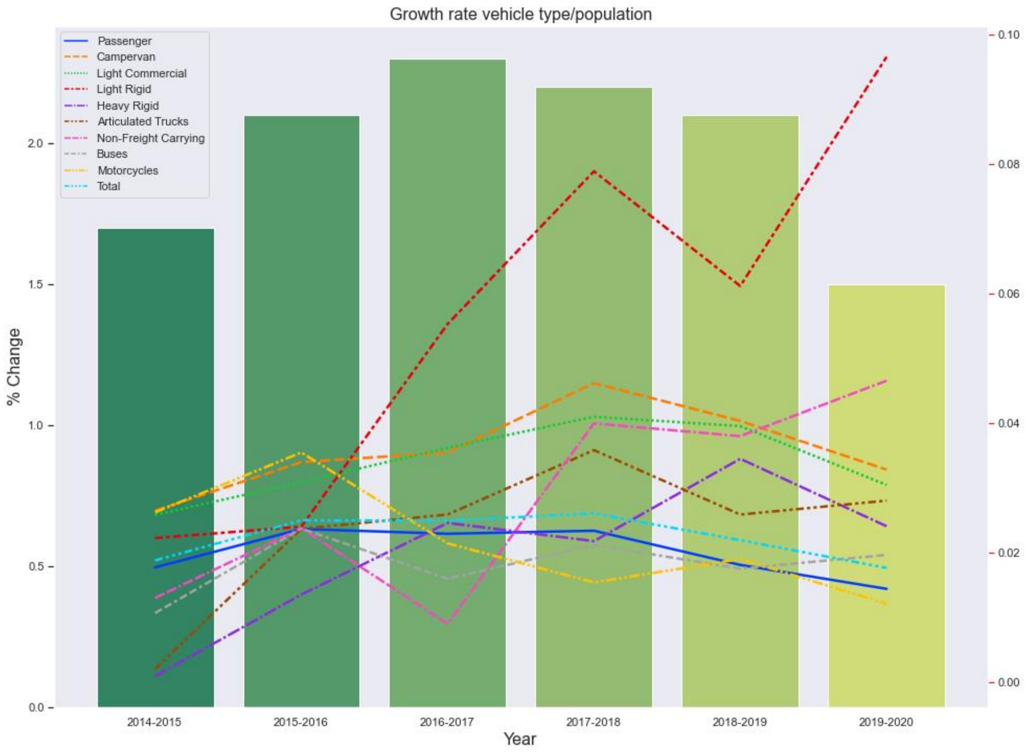


Figure 4: Growth rate vehicle type/population

Section 3: Growth Rate Comparisons and Correlations

Due to the small spread of data, the heatmap only shows a vague correlation between the population, traffic and migration growth. There are mostly strong correlations over the random sample of 15 LGA areas between the variables of population, traffic and migration growth was illustrated. As a random sample of 15 LGA areas are generated each time, results may vary, however, there seems to be a positive correlation between population increase and the increase of traffic.

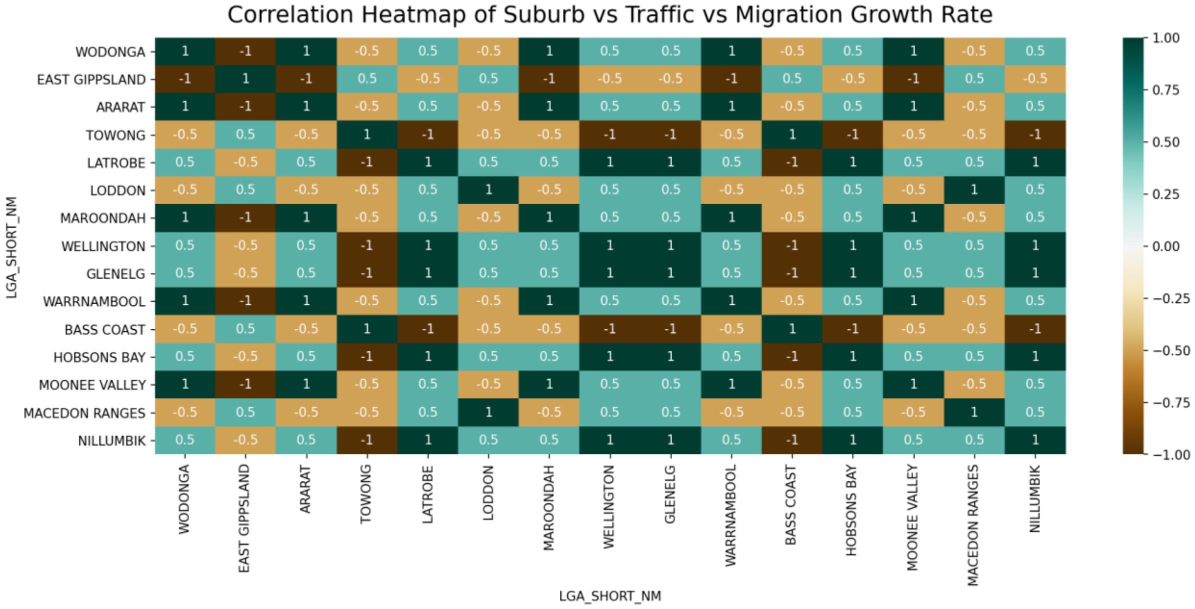


Figure 5: Correlation Heatmap of Suburb vs Traffic vs Migration Growth Rate

Traffic_vs_Pop_Growth.png was a line vs a bar graph where there mostly was an upwards trend in both population and traffic growth. This suggests that population growth is likely to have an impact towards Victoria’s road networks and it can be visually seen that there is a correlation between the two growth values. The LGA areas of increased traffic growth typically had spikes of population growth which demonstrates how there may be an increased demand for extended road networks within areas of high population.

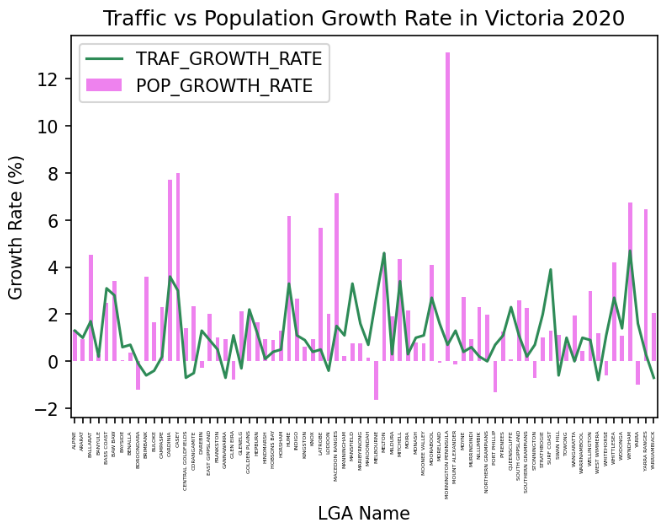


Figure 6: Traffic Population Growth Rate in Victoria 2020

Since the datasets used were recorded from June 2019 to June 2020, the growth rates for all variables would have been affected by the State-wide lockdown within 2020 due COVID-19.

Significance and Value of Results

Roads are essential for the entire society. By understanding how different variables such as the change in peak hour volumes change and the suburbs they occur in, resources can be used more efficiently to engineer roads and other infrastructures such as public transport. Different strategies can be developed for different areas for best results, such as infrastructural planning around areas with rapid population and vehicle growth to ensure that the road network in the vicinity is sufficient for the future population. The results could serve as an indicator for which areas these improvements could be considered for and should be maintained more regularly as they are the most visited roads.

Limitations and Future Improvements

The main challenges in the project stemmed from poor data availability, particularly the more specific data such as the speed and travel time of cars in different locations. Without these, it was difficult to determine the levels of congestion over time and find a strong correlation between population and traffic conditions. Similarly, as datasets did not share similar location measurements (SCAT-site, LGA, HMGS_FLOW_ID), traffic locations could not be directly linked. SCAT-site number, which represented a map grid on manual maps, was not suitable for its large dataset, which was why LGA, which had more identifiable suburbs, was chosen. However, this measurement had issues: it was only available for the outlier year 2020, and had some naming issues where locations were mismatched, e.g.: MONASH != MONASH UNIVERSITY.

An improvement that could be implemented would be to sort the data by their yearly quarters instead of the whole year. This would allow for more intervals of analysis between population growth and road traffic, possibly allowing for an easier recognition of correlation. However, a caveat is that given the sampling of 90 days, there could be certain quarters with fewer days included than others which would increase variability and decrease the accuracy of the results.

References

Butt, C., Jack, T., & Frederiksen, S. (2018). How to Beat Melbourne's Worsening Peak Hour Traffic. Retrieved 21 May 2021, from https://www.theage.com.au/interactive/2018/peak-hour-project/bespoke-features/melbourne-1.html?resizable=true&v=7&forceSubstituteMap=true

Statistics, L. (2018). Spearman's Rank Order Correlation using SPSS Statistics - A How-To Statistical Guide by Laerd Statistics. Retrieved 21 May 2021, from https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php