Jonas Olausson
Student ID: 751462

COMP20008 Project Phase 3 Final Report

An Investigation of Factors That Affect Car Crash Outcomes

Domain: Transport & Health

The purpose of this report is to analyze data pertaining to automobile accidents in Victoria to bring insight into the conditions that affect these accidents. Furthermore, in the context of the goal proposed by VicRoads to lower car crash deaths close to 0 by 2020, which LGA's transport safety need improvement in order to bring this goal to fruition.

Specifically, this report aims to answer the following questions:

1. What are the prevalent factors in determining the fatality outcome of an accident?
2. What are current areas that LGA's can improve ambulance response times to decrease fatalities arising from automobile accidents

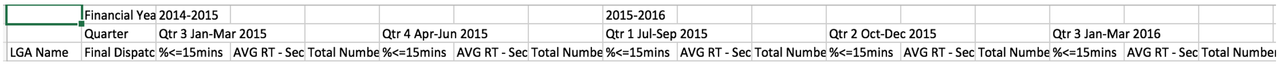
Datasets:

Ambulance Victoria LGA Response Time Performance Quarterly (csv)**:** Ambulance response times as reported by Ambulance Victoria. Breaking up the data into either code 1 responses and code 2 responses, information is given about response times in seconds and the proportion of responses under 15 minutes. The data only spans quarter 3 and 4 of the 2014-2015 financial year and quarter 1 to 3 of the 2015-2016 financial year. URI: https://www.data.vic.gov.au/data/dataset/ambulance-victoria-lga-response-time-performance-quarterly

Crashes Last Five Years (csv): A dataset provided by VicRoads that describes injury and fatality related car accidents. The data spans 2012 to 2016It details specific information about each crash event . URI: https://www.data.vic.gov.au/data/dataset/crashes-last-five-years

Preprocessing:

Firstly, excel was used to wrangle with **Ambulance Victoria LGA Response Time Performance Quarterly**. The data was presented in two blocks, code 1's and code 2's. I separated each code into a separate csv file to allow for ease of processing. The data had been organized in terms of financial year quarter as shown below:



	Financial Yea	2014-2015						2015-2016								
	Quarter	Qtr 3 Jan-Mar 2015			Qtr 4 Apr-Jun 2015			Qtr 1 Jul-Sep 2015			Qtr 2 Oct-Dec 2015			Qtr 3 Jan-Mar 2016		
LGA Name	Final Dispatc	%<=15mins	AVG RT - Sec	Total Numbe	%<=15mins	AVG RT - Sec	Total Numbe	%<=15mins	AVG RT - Sec	Total Numbe	%<=15mins	AVG RT - Sec	Total Numbe	%<=15mins	AVG RT - Sec	Total Numbe

As each quarter available was inconsistent, I simply averaged a single value for each unique column. The result is as shown below:

	LGA	average_percent_under_15	average_response_time_seconds	average_response_time_minutes

This was a perfect solution as I was only interested in a single metric for each LGA.

The cleaning and preprocessing phase for this dataset took place on excel due to duplicated column names making it hard for code to be written with pandas.

Moving on, the next dataset **Crashes Last Five Years** was cleaned and preprocessed through manipulating **DataFrames** in pandas. First the speed column contained string values indicating either 'no speed data was available' or 'accident took place on camping grounds. Mean imputation was considered but rather I was mostly interested in cases where accidents took place on roads with available speed limits. This caused 5.6% (4084 out of 72482) of rows to be deleted. Furthermore, accident dates were converted into date time objects for ease of use later in the report (to determine day of week).

/* Analysis and Visualizations:
I. Ambulance times analysis – which LGA needs improvement
*/

General Statistics:

Top 6 LGA's for fatal crash occurrences:

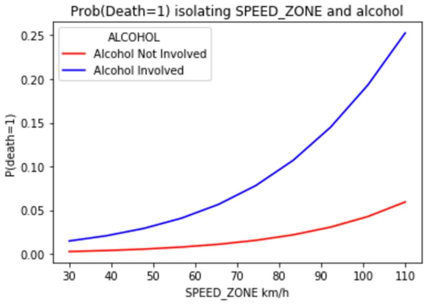
Rank	Local Government Area
1	Geelong
2	Wyndham
3	Yarra Ranges
4	Brimbank
5	Mornington Peninsula
6	Shepparton

Top 6 LGA's for total occurrences of crashes:

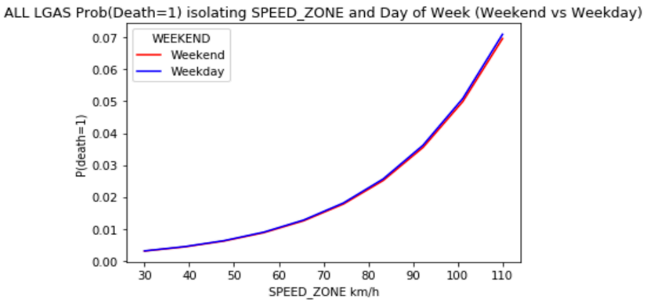
Rank	Local Government Area
1	Melbourne
2	Casey
3	Geelong
4	Dandenong
5	Hume
6	Brimbank

Factors That Affect Fatality - Logistic Regression:

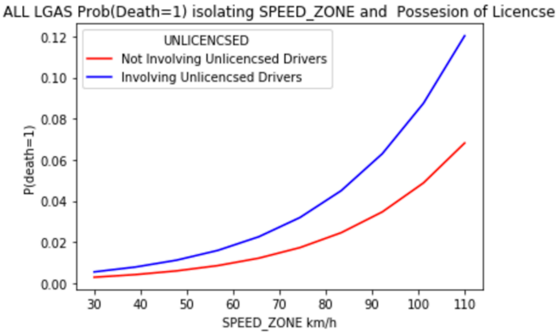
To demonstrate how certain factors interact with each other to determine the outcome of a car crash, logistic regression analysis is needed. In our case, the dependent variable is a binary value/dummy variable (1 = Fatal, 0 = Non-Fatal), and thus normal linear regression/scatter plotting will not work well. For this analysis I modeled how changing speed zones effects the odds of fatality when paired with different dummy variables taking on different factors in a car crash. Below are notable examples of results obtained from regressing speed zone against probability of fatality for all LGAs.


Prob(Death=1) isolating SPEED_ZONE and alcohol

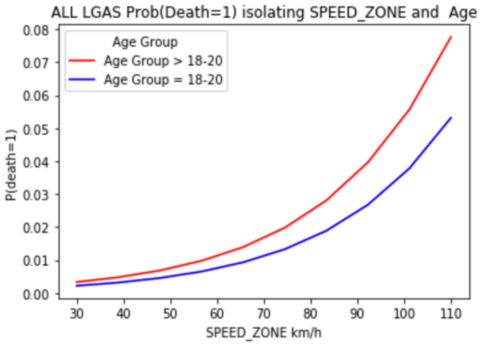
```
SPEED_ZONE    1.039615
ALCOHOL_1     5.346936
intercept     0.000880
dtype: float64
```


ALL LGAS Prob(Death=1) isolating SPEED_ZONE and Day of Week (Weekend vs Weekday)

```
SPEED_ZONE    1.040738
WEEKEND_1     1.020513
intercept     0.000927
dtype: float64
```

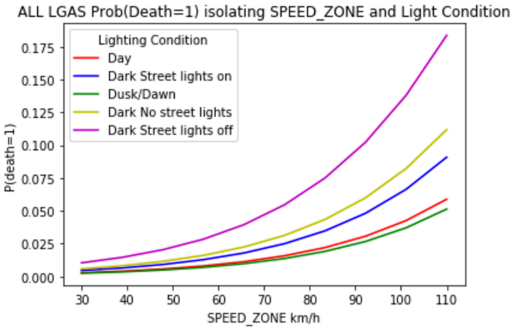

ALL LGAS Prob(Death=1) isolating SPEED_ZONE and Possesion of Licencse

```
SPEED_ZONE       1.040742
UNLICENCSED_1    1.868646
intercept        0.000906
```


ALL LGAS Prob(Death=1) isolating SPEED_ZONE and Age

```
SPEED_ZONE    1.040928
YOUNG_1       0.667052
intercept     0.001021
dtype: float64
```

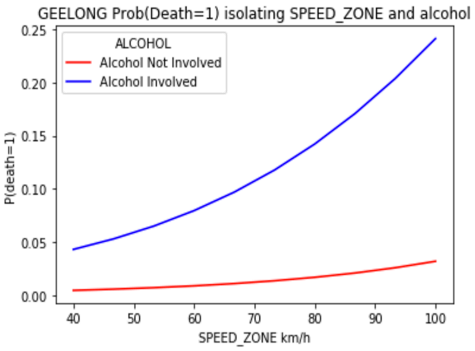
Interpreting the results are as such:
The given odds values for each variable are its marginal effect on the probability of a fatality. This means for the first figure increasing speed by one kilometer per hour increases the probability of death by 104%.
Firstly, alcohol has a large effect in increasing probability of a fatality. Throughout all LGA, it seems to have a huge presence in the data. The factor of weekend and week day does not seem to have any effect on the odds of a fatality. Notably, incidents labeled with a ‘no license’ feature seem to have an increasing probability on the fatality of an accident. Interestingly, the results show that if the driver was in the age group between 18 and 20 this has a decreasing effect on the odds of a fatality. Lastly, as we would expect, poor lighting conditions are shown to increase the odds of death.

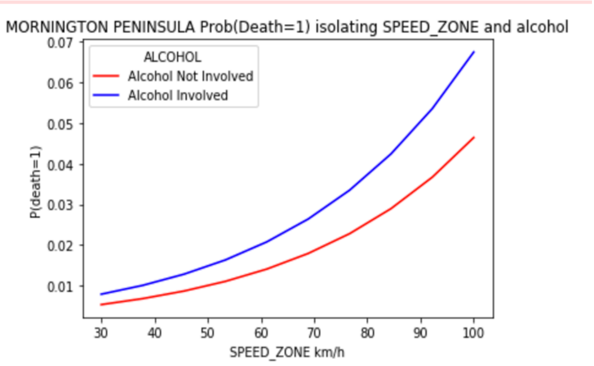

ALL LGAS Prob(Death=1) isolating SPEED_ZONE and Light Condition

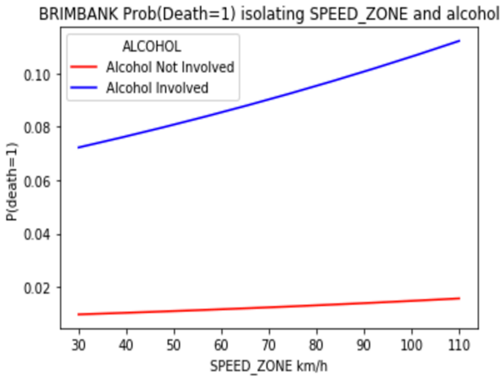
```
SPEED_ZONE      1.039000
LIGHT_RANK_1    1.598431
LIGHT_RANK_2    0.866012
LIGHT_RANK_3    2.014021
LIGHT_RANK_4    3.597673
intercept       0.000930
dtype: float64
```

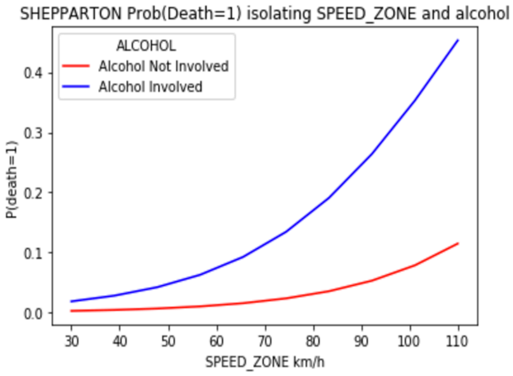
Jonas Olausson
Student ID: 751462

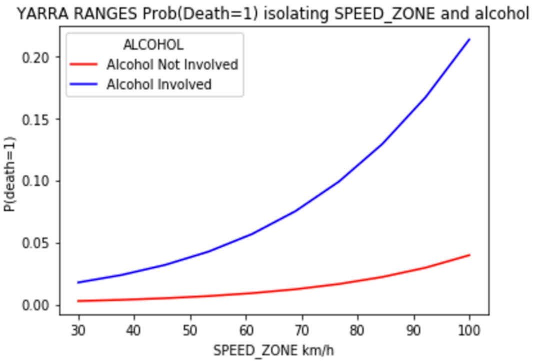
I felt that further investigation was needed to see if some of these results were consistent throughout specific LGAs. Below are logit models applied to various LGA's, for this report I decided to focus on the top 6 most fatal LGA's as these would be of most concern to policy makers intending to decrease automobile deaths.

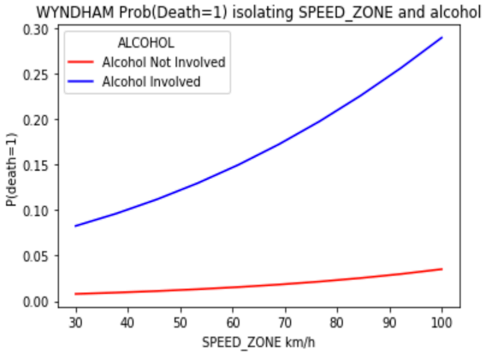




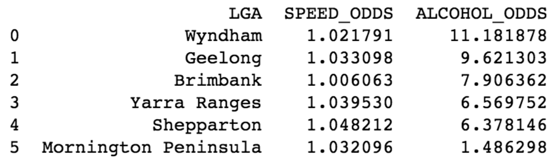




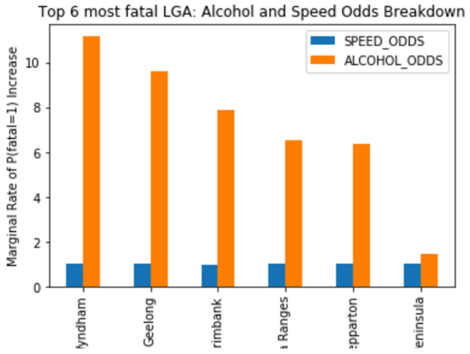




Our results show that as intuition would predict, alcohol and speed has an increasing effect on the probability of an accident being fatal. Comparing each result in terms of LGA, the trend remains. However, the magnitude of the effect of alcohol on the probability of death changes drastically. To further investigate this difference a table and plot containing the logit results are shown below:



```
                     LGA  SPEED_ODDS  ALCOHOL_ODDS
0                Wyndham    1.021791     11.181878
1                Geelong    1.033098      9.621303
2               Brimbank    1.006063      7.906362
3           Yarra Ranges    1.039530      6.569752
4             Shepparton    1.048212      6.378146
5  Mornington Peninsula    1.032096      1.486298
```



The logistic regression results show the marginal effect on P(death = 1) from a single unit increase in speed. For the LGA Wyndham, we can see that the odds of dying from a car accident when speed is increased a single unit increases tenfold when alcohol is involved. In contrast the LGA Mornington Peninsula shows little increase in odds when alcohol is involved. For remaining LGA's alcohol increases odds of a fatality by about 6 to 9 times. This indicates that alcohol related deaths in Wyndham and Geelong are largely prevalent.

Another aspect that critically determines the survival of a crash victim is ambulance response times. With serious trauma it follows that immediate attention is required as soon as possible. Comparing the response times against crash frequencies will assess whether an LGA needs to focus resources into decreasing response times in order to accommodate these crashes. The following data from Ambulance Data was used here and Crashes Last Five Years was filtered to only show years 2014-2016. This was done as ambulance data was only available during these times.

Top 5 Worst Average Response Times:

Local Government Area	**Average Response Time in Minutes**	**Crash Count for 2014-2016**
Buloke	25.62	55
Towong	24.69	100
Mansfield	24.51	184
West Wimmera	23.84	33
Yarriambiack	23.58	30

Our data shows that these LGA's have high response times which could negatively impact the survival rate of crash victims. Although these times are high, these LGAs do not account for much crash frequency over the years 2014-2016.

Average Ambulance Response Times for the Top 6 Most Fatal LGA's:

Rank	**Local Government Area**	**Response Time (Minutes)**	**Crash Count 2014-2016**
1	Geelong	13.27	1507
2	Wyndham	13.25	912
3	Yarra Ranges	15.27	1284
4	Brimbank	12.56	1362
5	Mornington Peninsula	13.30	841
6	Shepparton	12.56	499

Conclusion and Limitations:

The results above show interactions between factors that come into play into affecting the outcome of a crash. Through analyzing the change in speed zone against different dummy variables we get a picture of which factors are prevalent. Alcohol is shown to have a major effect on the outcome of a crash. In particular one LGA which demonstrated a large prevalence of this was Wyndham. Furthermore poor lighting conditions and a lack of license showed an increasing impact on probability of death. This confirms intuition and reinforces the need for street lights to improve visibility and for penalties to those who drive without a license. Interestingly, the data shows that given a certain speed zone, younger drivers are less likely to die in an accident. This could possibly be from less fatalities in the data set paired with that certain age group. Finally, when comparing crash counts with ambulance response times, we see very high response times paired with smaller crash counts in general and vice versa with fast response times. LGAs such as Yarra Ranges and Mansfield should improve their response times to match their crash count frequencies. In regards to VicRoad's goal of lowering car crash deaths to zero by 2020, this information would be useful to policy makers in specific LGAs to enact target policies to address the factors discussed above.

The report only discusses a few factors that could affect car crash outcomes, possible further investigation could include factors such as, seat position, how long the driver has been driving for, type of vehicle, instantaneous speed at collision. Furthermore the time horizon of the ambulance data was extremely short and limited my time span when cross analyzing crash frequency. This report is also limited in its scope as the population density was not discussed.

9. Challenges and Reflections

1. **Wrangling:**

In the previous stage I mainly used the data set Crash Stats Data Extract, a folder which contained csv files and entries from 2006-2017. While the data was rich and informative, the fragmented structure coupled with the immense number of entries made my primitive data integration and cleaning processes very slow and difficult to deal with. I decided to pivot to another data set (**Crashes Last Five Years)** that had a simpler structure and contained the same information with the added benefit of having alcohol dummy variable.

2. **Logistic Regression Model:**

Building this model proved to be an extremely difficult task. The specifications of this model were not taught in this class but were introduced to me in "Introductory Econometrics", therefore examples and documentation were hard to come by. I managed to find an online example[1] that performed the tasks I needed. Adapting my data to this example was difficult and took many attempts before I successfully managed to run a visualization of the model.

[1] Logistic Regression: "http://blog.yhat.com/posts/logistic-regression-python-rodeo.html"

Jonas Olausson
Student ID: 751462

Code:
I made use of http://blog.yhat.com/posts/logistic-regression-python-rodeo.html to create and visualize logistic regression models. This was modified and made modular for ease of repetition. Cleaning and preprocessing code was written entirely from scratch. Libraries used: pandas, math, matplotplit, statsmodels.api, pylab, numpy, datetime, calendar.

Biography:

1. Logistic Regression: "http://blog.yhat.com/posts/logistic-regression-python-rodeo.html"