Project and report developed by Mustafa Neguib with student number 922939

Question 1 -- Calculating the average number of accidents per year

Average number of accidents per year: 13788.55

Question 2 -- Number and proportion of the second most common type of accidents

The second most common type of accident in all the recorded years is 'Collision with a fixed object', and the percentage of the accidents that belong to this type is 16.56%

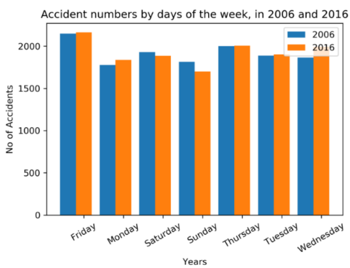
Question 3 -- Number of accidents by vehicle type by year

YEAR	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Vehicle Type Desc											
Bicycle	1256.0	1312.0	1384.0	1488.0	1496.0	1577.0	1427.0	1567.0	1615.0	1556.0	1334.0
Bus/Coach	106.0	133.0	125.0	153.0	123.0	168.0	133.0	137.0	127.0	136.0	99.0
Car	13395.0	13178.0	13640.0	13075.0	12846.0	13189.0	13088.0	13543.0	14672.0	12977.0	12204.0
Heavy Vehicle (Rigid) > 4.5 Tonnes	263.0	387.0	394.0	305.0	298.0	355.0	297.0	309.0	268.0	352.0	277.0
Horse (ridden or drawn)	1.0	1.0	1.0	2.0	1.0	6.0	2.0	0.0	1.0	0.0	0.0
Light Commercial Vehicle (Rigid) <= 4.5 Tonnes GVM	204.0	249.0	255.0	226.0	266.0	272.0	215.0	204.0	172.0	183.0	206.0
Mini Bus(9-13 seats)	23.0	16.0	29.0	27.0	15.0	21.0	27.0	13.0	12.0	23.0	13.0
Moped	12.0	18.0	16.0	14.0	14.0	19.0	15.0	16.0	20.0	16.0	25.0
Motor Cycle	1818.0	1951.0	2076.0	1963.0	1759.0	1817.0	1816.0	1969.0	2073.0	2077.0	1888.0
Motor Scooter	54.0	118.0	143.0	137.0	145.0	131.0	112.0	126.0	111.0	109.0	104.0
Not Applicable	2.0	2.0	1.0	2.0	5.0	0.0	2.0	1.0	0.0	0.0	0.0
Other Vehicle	73.0	85.0	79.0	70.0	54.0	59.0	62.0	71.0	58.0	22.0	47.0
Panel Van	621.0	533.0	526.0	530.0	508.0	499.0	581.0	541.0	564.0	542.0	540.0
Parked trailers	0.0	0.0	0.0	0.0	0.0	0.0	1.0	6.0	15.0	10.0	3.0
Plant machinery and Agricultural equipment	0.0	0.0	0.0	0.0	0.0	1.0	10.0	14.0	23.0	23.0	14.0
Prime Mover (No of Trailers Unknown)	342.0	292.0	4.0	3.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0
Prime Mover - Single Trailer	2.0	1.0	241.0	209.0	218.0	214.0	232.0	199.0	214.0	220.0	170.0
Prime Mover B-Double	0.0	1.0	76.0	78.0	96.0	96.0	102.0	94.0	65.0	92.0	75.0
Prime Mover B-Triple	0.0	0.0	2.0	0.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0
Prime Mover Only	11.0	13.0	21.0	26.0	32.0	43.0	40.0	51.0	51.0	39.0	56.0
Quad Bike	0.0	0.0	0.0	0.0	0.0	2.0	7.0	9.0	21.0	9.0	14.0
Rigid Truck(Weight Unknown)	93.0	33.0	20.0	20.0	12.0	0.0	0.0	0.0	0.0	0.0	0.0
Station Wagon	3105.0	3168.0	3331.0	3324.0	3541.0	3659.0	3795.0	3102.0	2776.0	4494.0	4517.0
Taxi	311.0	363.0	389.0	350.0	321.0	319.0	291.0	306.0	325.0	255.0	210.0
Train	15.0	5.0	9.0	5.0	3.0	6.0	12.0	6.0	5.0	3.0	5.0
Tram	53.0	74.0	82.0	76.0	80.0	69.0	53.0	54.0	40.0	71.0	37.0
Unknown	209.0	180.0	185.0	160.0	171.0	257.0	458.0	542.0	563.0	619.0	447.0
Utility	1487.0	1524.0	1565.0	1674.0	1796.0	1805.0	1914.0	1808.0	1794.0	2170.0	2043.0

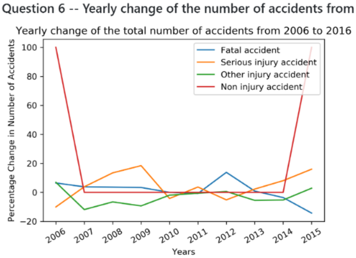
Question 4 -- Top 10 local government areas

	LGA_NAME	NUM_2006	NUM_2016	DIFFERENCE	CHANGE
0	MELBOURNE	862	763	-99	-11.484919
1	CASEY	516	606	90	17.441860
2	DANDENONG	478	516	38	7.949791
3	GEELONG	468	503	35	7.478632
4	YARRA RANGES	443	410	-33	-7.449210
5	MONASH	418	373	-45	-10.765550
6	KINGSTON	384	306	-78	-20.312500
7	MORELAND	368	382	14	3.804348
8	MORNINGTON PENINSULA	355	294	-61	-17.183099
9	BRIMBANK	348	416	68	19.540230

Question 5 -- Accident numbers by days of the week in 2006 and 2016



Question 6 -- Yearly change of the number of accidents from 2006 to 2016 for each severity category



Project and report developed by Mustafa Neguib with student number 922939

Spatial Temporal Visual Analysis (Basic)

This task has been developed and analysed from the view point of a Spatial Data Analyst who has been given a task for analysing data based on the following scenario, "The Government of Victoria has been informed by a third party that there has been an increase in the number of accidents through certain SA2 areas in Victoria and have been recommended that they employ policies which either restrict the vehicular access or make those areas pedestrian only (excluding of public transport). The government wants us to verify these claims and also to come up with a recommendation on what to do based on the data that the government has provided us. The government also wants us to find a reason that may be causing the rise in the number of accidents."

For this study we want to identify the SA2 areas that has the highest number of accidents in all of the SA2 areas. The first visualization shows the number of accidents that took place in the shown SA2 regions from 2006 to 2016. This is for all days (weekday and weekend) when the accidents took place. The second visualization shows the number of accidents that took place in the shown SA2 regions from 2006 to 2016 during the weekdays (Monday to Friday). The third visualization shows the number of accidents that took place in the shown SA2 regions from 2006 to 2016 during the weekends (Saturday and Sunday). The legend shows the bin ranges of the number of accidents that took place.

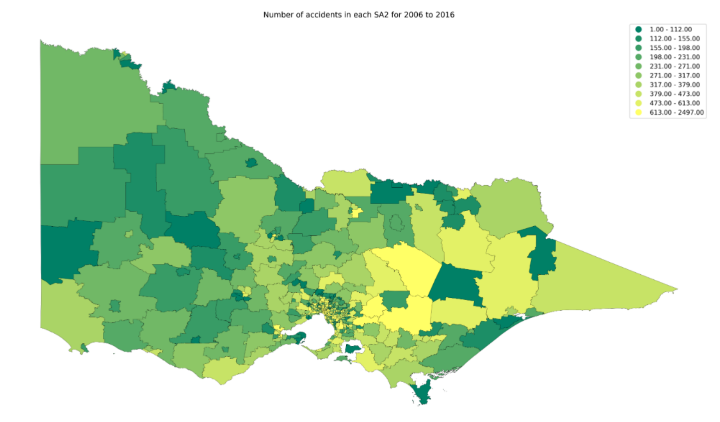


Figure 1: The figure shows the number of accidents that took place from 2006 to 2017 during both the weekdays and the weekends. Do note the bins in the legends.

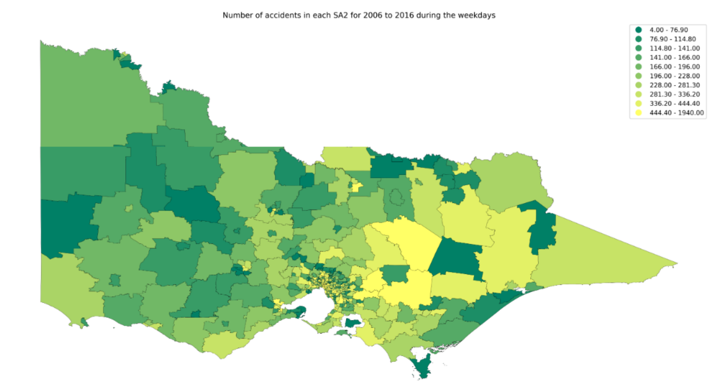


Figure 1: The figure shows the number of accidents that took place from 2006 to 2017 during both the weekdays and the weekends. Do note the bins in the legends.

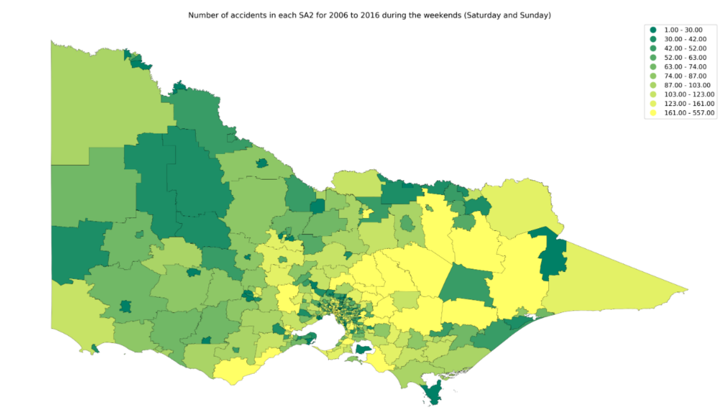


Figure 3: The figure shows the number of accidents that took place from 2006 to 2017 during the weekend. Do note the bins in the legends. The maximum number of accident is less than what we have in Figure 1. The bins tell that there were alot more fewer accidents on the weekends than on the weekdays and also as an overall total.

Spatial Autocorrelation Calculation (Advanced)

This section of the project required me to perform autocorrelation on the spatial data. What this means was that I was to find how similar the SA2 areas are to their neighbors and whether there are any clustering or not. Infact this is what autocorrelation tells me. To begin with we have a Null-hypothesis that the clusters are randomly distributed and there are no groups clustered together. We will show that this hypothesis is false, and that in reality there are clusters which are grouped together and are not randomized.

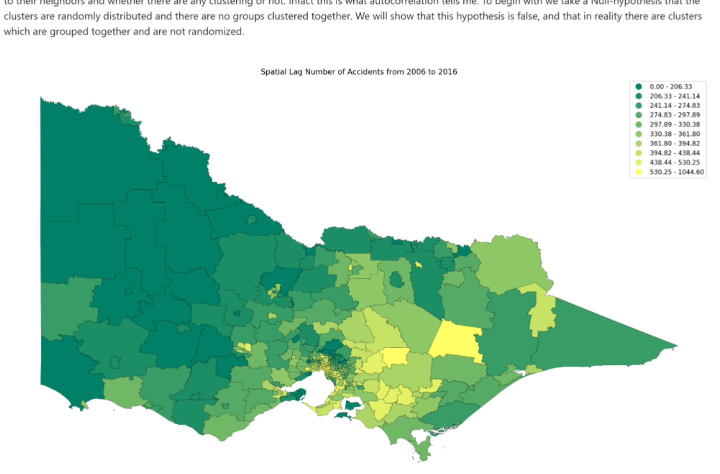


Figure 4: This figure shows the spatial lag that has been calculated using Queen contiguity statistic which builds an index where if an area is a neighbor is given a value of 1 and 0 if not. Then the spatial lag is calculated which tells how similar an area is similar to its neighbors.

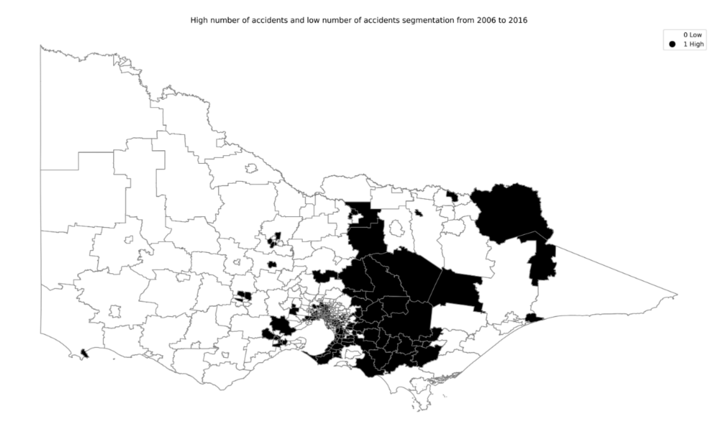


Figure 5: This figure shows the SA2 areas being categorized as being higher than the median value of the spatial lag (black in color) and lower than the spatial lag (white in color). We are regarding the black areas as having high number of accidents, while the white areas as having low number of accidents. We can already see that there are clusters grouped together. We will statistically prove this in the next figure.

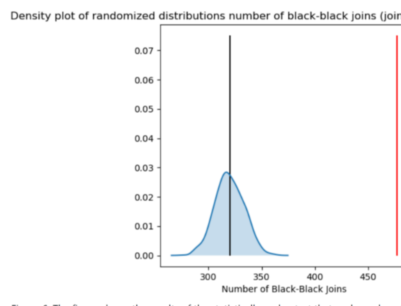
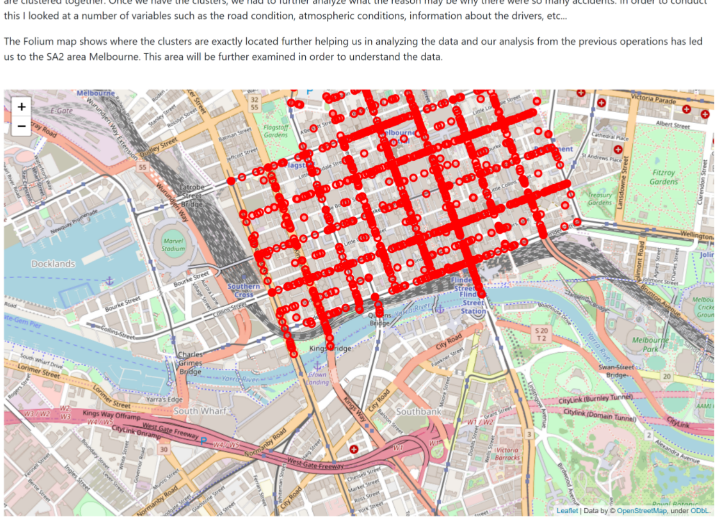


Figure 6: The figure shows the results of the statisticall p-value test that we have done in order to invalidate the null hypothesis. We have used Pysal, and it generates by default 999 randomized clusters and then finds their mean and other values. The density plot shown in the figure is of the randomized generations, and the black line shows the mean. However, our own actual value is an extreme value and this proves that the Null-hypotheses is wrong and that we can reject it, therefore it can be said that our data contains clusters that are closely grouped together, leading to the fact that those SA2 areas are similar to each other.

Clustering Analysis (More Advanced)

The results of the DBSCAN gives us 47 clusters. Since we are clustering on the longitude and latitude the accident locations that are spatially closer together are clustered together. Once we have the clusters, we had to further analyze what the reason may be why there were so many accidents. In order to conduct this I looked at a number of variables such as the road condition, atmospheric conditions, information about the drivers, etc...

The Folium map shows where the clusters are exactly located further helping us in analyzing the data and our analysis from the previous operations has led us to the SA2 area Melbourne. This area will be further examined in order to understand the data.



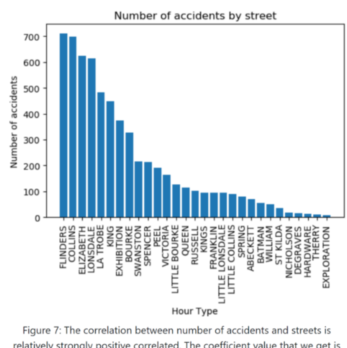


Figure 7: The correlation between number of accidents and streets is relatively strongly positive correlated. The coefficient value that we get is -0.899849793853915. There is significant negative correlation between the streets and the number of accidents. What this means is that the two variables of streets and the number of accidents are inverse of each other, however, due to the nature of the data and how it is plotted, infact it is actually showing positive correlation because, the streets where the majority of the accidents have taken placeshow that those streets have a higher propencity of having more accidents than the other streets where the number of accidents is significantly lower. The Government of Victoria should look at trying to make these streets more safer.

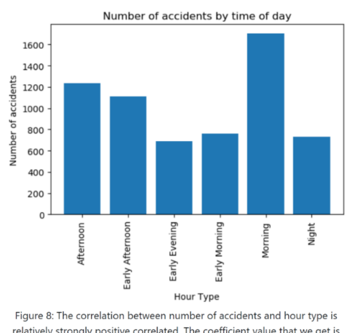


Figure 8: The correlation between number of accidents and hour type is relatively strongly positive correlated. The coefficient value that we get is -0.09062326443452029. There is no correlation between the time of day and the number of accidents. There is no specific time of day where most of the accidents took place. The accidents are pretty much spread out.

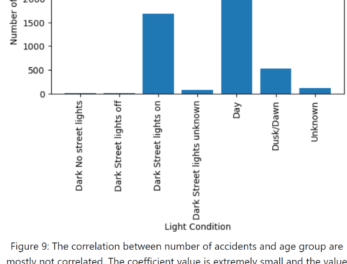


Figure 9: The correlation between number of accidents and age group are mostly not correlated. The coefficient value is extremely small and the value is 0.18816035355396343. There is no correlation at all. The r value is too small.

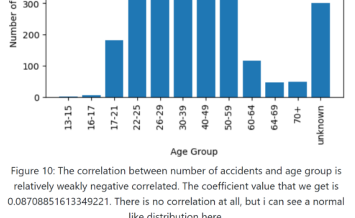


Figure 10: The correlation between number of accidents and age group is relatively weakly negative correlated. The coefficient value that we get is 0.08708851613349221. There is no correlation at all, but i can see a normal like distribution here.

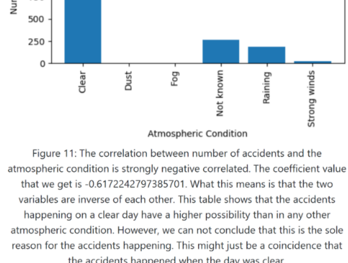


Figure 11: The correlation between number of accidents and the atmospheric condition is strongly negative correlated. The coefficient value that we get is -0.6172242797385701. What this means is that the two variables are inverse of each other. This table shows that the accidents happening on a clear day have a higher possibility than in any other atmospheric condition. However, we can not conclude that this is the sole reason for the accidents happening. This might just be a coincidence that the accidents happened when the day was clear.

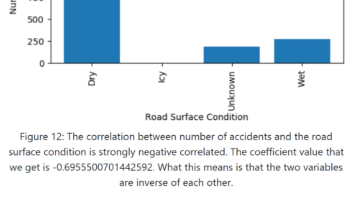


Figure 12: The correlation between number of accidents and the road surface condition is strongly negative correlated. The coefficient value that we get is -0.6955500701442592. What this means is that the two variables are inverse of each other.

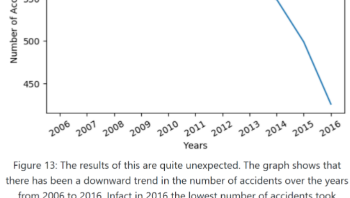


Figure 13: The results of this are quite unexpected. The graph shows that there has been a downward trend in the number of accidents over the years from 2006 to 2016. Infact in 2016 the lowest number of accidents took place. However, despite the downward trend, some actions do need to be taken by the Government of Victoria.

Conclusion

Figures 7, 11, and 12 all have negative correlation, but as discussed in the report earlier, we have to be careful in concluding based on the data because the data might mean something else altogether and we may get wrong results. Furthermore, Figure 13 shows that there has been a decline in the number of accidents, rather than an increase as suggested by the third party. However, there are high risk streets where the number of accidents is quite high and should be controlled and brought down. These streets are Flinders Street, Collins Street, Elizabeth Street and Londsdale Street. We recommend that the government take steps in order to further make the streets safer.