Assessment 4 Report

Bingkun Chen(992113), Zhihang Niu(1016294)

June 15, 2020

Abstract

This report will implement the spatial data extraction, export the analysis-ready data, and perform some spatial analysis on the provided dataset.

1 Task 1

1.1 Text output

- The average number of accidents per year is 12211.
- The most common type of accident in all the recorded years is Collision with vehicle(equal to 62.98% accidents), and the second most common is Collision with a fixed object(15.77%).

This report also provide an interactive pie chart [pieChart] to show the changes in accident types in 2013 and 2018. It can be found that collision with vehicle is the most common type, but there is a significant increase in collision with a fixed object. (If the link cannot be opened, please open *pieChart.html* in the *output* folder)

1.2 Tabular output

1.2.1 Table 1

Table 1 shows the number of accidents by vehicle type (rows) by year (columns). Based on the definition from Wikipedia[2]: Vehicles include wagons, bicycles, motor vehicles (motorcycles, cars, trucks, buses), railed vehicles (trains, trams), watercraft (ships, boats), amphibious vehicles (screw-propelled vehicle, hovercraft), aircraft (airplanes, helicopters) and spacecraft. Therefore, this report included bicycle as one of vehicle types for this task. For the table 1, it can be found that Passenger vehicle has the largest proportion.

Table 1: *Number of accidents by vehicle type by year*

Year *Type*	2013	2014	2015	2016	2017	2018
Publicvehicle	86	173	209	153	146	160
Heavyvehicle	283	583	669	580	559	516
Bicycle	666	1586	1534	1357	1288	1183
Motorcycle	918	2149	2175	2178	1837	1696
Passengervehicle	3687	8192	8376	8478	7082	6251

1.2.2 Table 2

Table 2 shows the top 10 LGAs (Local government areas) that have the highest number of accidents in 2013, sorted in decreasing order. Then, compute the changes of the numbers of accidents in 2014 compared to the previous year, and so on, up until 2018 for these 10 LGAs. These differences should be computed both in absolute numbers and as percentage change. The data from table 2 demonstrate that the biggest difference occurred between 2013 and 2014.

Table 2: The accidents change of top 10 LGAs of 2013

Year *LGA*	2013	2014			2015		
	No.	*No.*	*Diff.*	*Change*	*No.*	*Diff.*	*Change*
MELBOURNE	380	837	457	120.26%	853	16	1.91%
GEELONG	241	505	264	109.54%	489	-16	-3.17%
CASEY	234	532	298	127.35%	600	68	12.78%
BRIMBANK	232	475	243	104.74%	454	-21	-4.42%
DANDENONG	220	468	248	112.73%	503	35	7.48%
MONASH	208	442	234	112.50%	435	-7	-1.58%
HUME	191	447	256	134.03%	448	1	0.22%
MORELAND	185	405	220	118.92%	435	30	7.41%
KINGSTON	177	369	192	108.47%	338	-31	-8.40%
DAREBIN	170	330	160	94.12%	365	35	10.61%

Year *LGA*	2016			2017			2018		
	No.	*Diff.*	*Change*	*No.*	*Diff.*	*Change*	*No.*	*Diff.*	*Change*
MELBOURNE	746	-107	-12.54%	683	-63	-8.45%	642	-41	-6.00%
GEELONG	533	44	9.00%	519	-14	-2.63%	454	-65	-12.52%
CASEY	640	40	6.67%	498	-142	-22.19%	474	-24	-4.82%
BRIMBANK	431	-23	-5.07%	361	-70	-16.24%	353	-8	-2.22%
DANDENONG	541	38	7.55%	417	-124	-22.92%	355	-62	-14.87%
MONASH	420	-15	-3.45%	353	-67	-15.95%	314	-39	-11.05%
HUME	478	30	6.70%	387	-91	-19.04%	390	3	0.78%
MORELAND	405	-30	-6.90%	352	-53	-13.09%	362	10	2.84%
KINGSTON	320	-18	-5.33%	308	-12	-3.75%	282	-26	-8.44%
DAREBIN	344	-21	-5.75%	281	-63	-18.31%	291	10	3.56%

1.3 Task 1.3 Charts and Maps

1.3.1 Figure 1

Figure 1 is a bar chart of the accident numbers in 2013 and 2018 by days of the week. It can been see from the figure below that the number of accidents on weekends is less than weekdays. Besides, the accidents numbers in 2018 are greater that year 2013. This report also provide a interactive bar chart [plotly_fig1] if you want to further research the details. (If the link cannot be opened, please open *plotly_fig1.html* in the *output* folder)

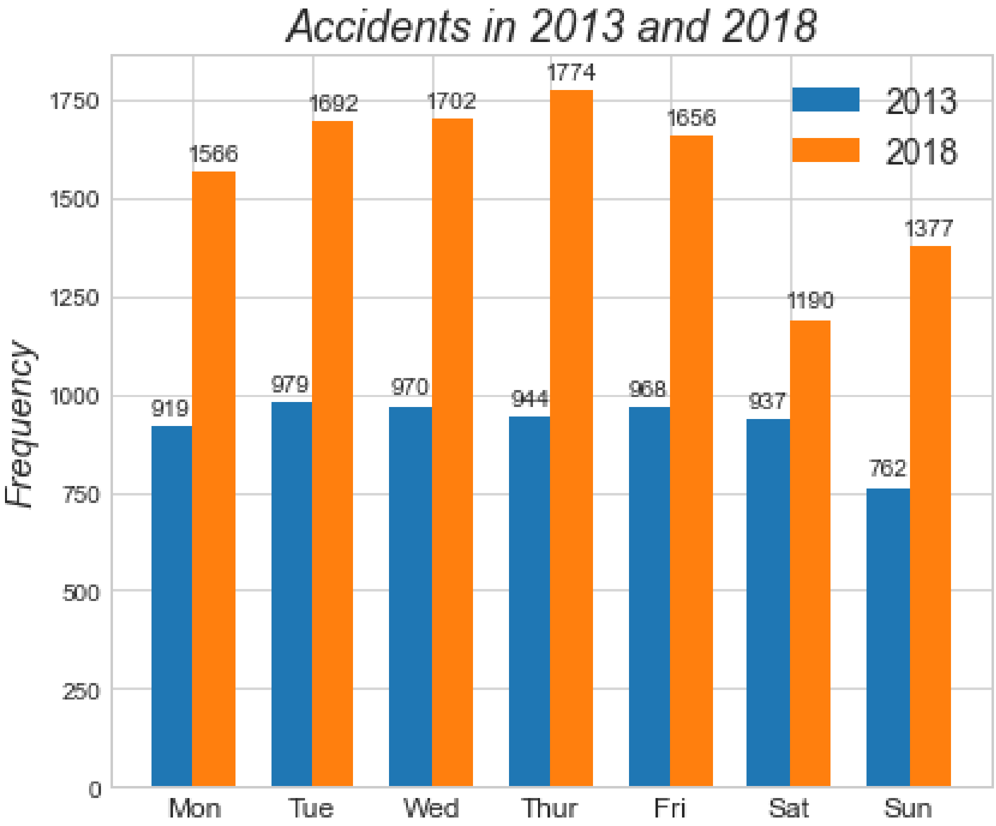


Figure 1: *Accident numbers in 2013 and 2018*

1.3.2 Figure 2

Figure 2 is a line chart of the yearly change of the total number of accidents from all year between 2013 and 2018, for each severity category. The figure shows that all types have similar trends, peaking in 2015 and 2016, then falling in 2017 and 2018. For the types Serious injury and non injury, since the accident numbers is so small that hard to explore the trend, this report provide a interactive line chart [plotly_fig2] for better research the details. (If the link cannot be opened, please open *plotly_fig2.html* in the *output* folder)

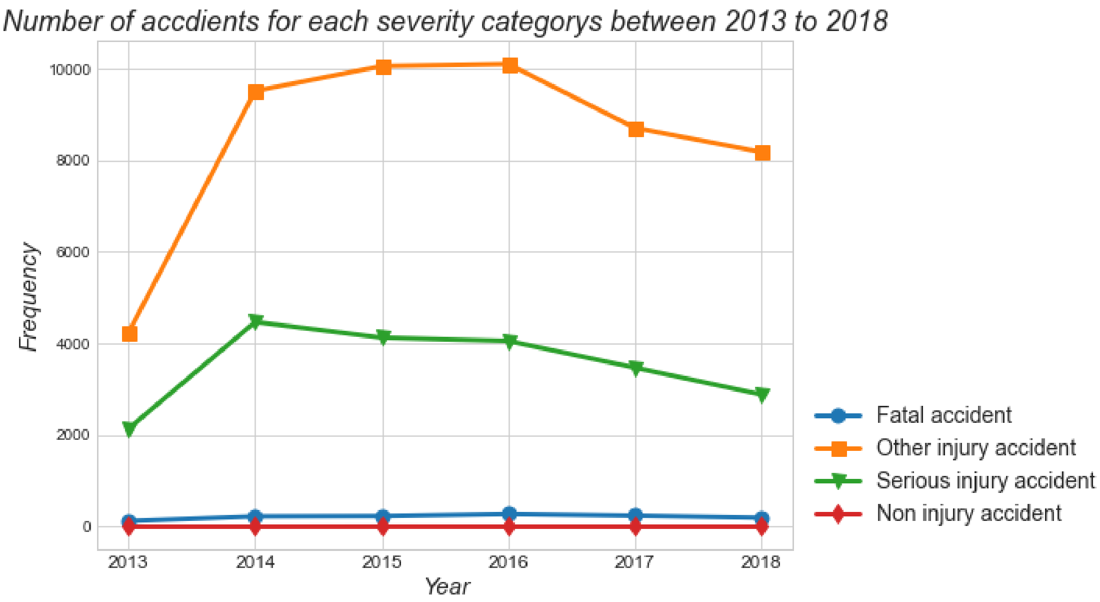


Figure 2: Number of accidents for each severity categorys (2013-2018)

1.3.3 Figure 3

Figure 3 are choropleth maps of showing the fraction of the most common severity category of accidents (in all the data, for all of Victoria) per LGA for 2013 and 2018. This report also provide two interactive choropleth maps [plotly_fig3_2013] [plotly_fig3_2018] if you wan to further research the details. (If the link cannot be opened, please open *plotly_fig3_2013.html* and *plotly_fig3_2018.html* in the *output* folder)

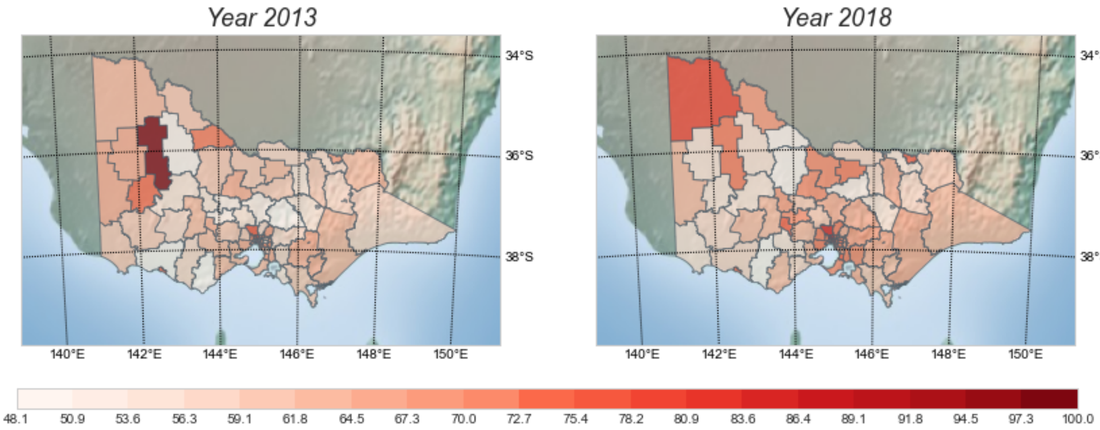


Figure 3: Choropleth maps for Year 2013 and 2018

2 Task 2

2.1 Task 2.1

This part export the content captured by the two tables into a single, geopackage dataset (Task2.gpkg), writing layers of “AccidentsByYear” and “AccidentsByLGA”, respectively, holding the data from the two tables above. For the layer “AccidentsByYear”, this report use the centroid of the whole State Of Victoria as the geometry.

2.2 Task 2.2

This part create a new layer called “AccidentLocations” for the geopackage(Task2.gpkg).

2.3 Task 2.3

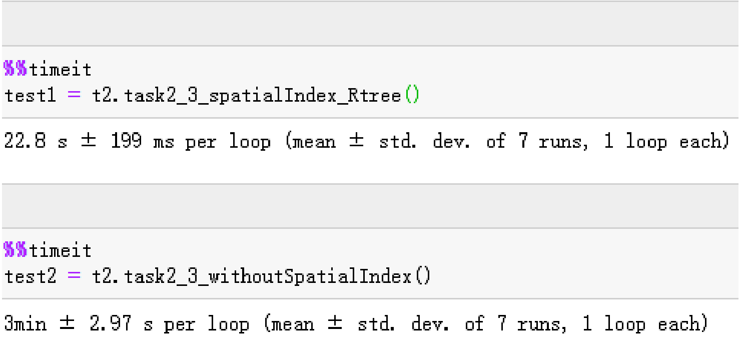
This part added a new field “SA2” to “AccidentLocations”. The value of the field will be set to the Statistical Area 2 name the accident happened within (this report consider “within” as “intersect” since only using within will ingore the accidents happended on the boundry of SA2 area). The output of this part will update the layer “AccidentLocations” for the geopackage (Task2.gpkg).

2.4 Analysis of spatial index

In order to identify the effect of spatial index for spatial query, thie report implemented an analysis of spatial index (R-tree).

Firstly, this report create two function to implement task2.3 with and without using spatial index. Then make sure the output of these functions same to the task2_3_spatialJoin(). Finally, this report compared the running time of these functions and explore the relative acceleration.

The results shows that the mean running time for the function without spatial index is 3 min ± 2.97s, while the function using spatial index (R-tree) only cost 22.8s ± 199 ms. Therefore, using spatial index can improve the efficiency of spatial query and reduce the running time.



```
%%timeit
test1 = t2.task2_3_spatialIndex_Rtree()
```

```
22.8 s ± 199 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

```
%%timeit
test2 = t2.task2_3_withoutSpatialIndex()
```

```
3min ± 2.97 s per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

Figure 4: The result of Mean Running time

2.5 Task 2.4

This task provided a function to split the entries in the created layer “AccidentLocations” into two layers: “SevereAccidentsWeekday” and “SevereAccidentsWeekend”.

3 Task 3

3.1 Analysis 1 : Spatio-temporal visual analysis

3.1.1 Introduction

Based on the previous data processing and analysis, the group believes that the number, type and reason of accidents may be changed with the timeline. Therefore, the group wants to confirm that by the following result.

- The number and distribution of accidents per SA2 between 2013 and 2018;
- The change of different years and working days/weekends may impact on the number of accidents;
- The number of different vehicle types by different year;
- Numbers of young people and old people involved in accidents within a week;
- Numbers of causes of common accidents (eg. drunk driving and unlicensed driving) affected by accidents within a week.

3.1.2 Dataset

All the original data used in this part is as follows:

- SA2 region in 2016
- The accidents data between 2013 to 2018
- The Victoria SA2 geojson file imported by the whole SA2 data

3.1.3 Result and Discussion

In response to the above conjecture, this report first drew a choropleth map [plotly_fig4] to show the number of accidents that occurred in each statistical area. This is a map with a timeline. Drag the timeline to easily get the number of accidents in each statistical area between 2013 to 2018. These numbers indicate that the distribution of accidents in the past 6 years has not changed much, but there has been a significant increase in the number from 2013 to 2014, while there has been a steady fluctuation between 2014 and 2018. This is a cluster distribution. It is obvious that the accidents are concentrated in the central area of Melbourne, while the number of accidents in the suburbs of Victoria is relatively small. The characteristics of this may be due to data loss in 2013. (If the link cannot be opened, please open *plotly_fig4.html* and in the *output* folder)

Also, from the line chart [plotly_fig5] of the change of vehicle type in the following accident, we can clearly see the number of every vehicle type involved has increase from 2013 to 2014, and after 2016, there has been a significant decline in the number. (If the link cannot be opened, please open *plotly_fig5.html* and in the *output* folder)

In addition, this report selected unlicensed driving and drunk driving as common causes of accidents for analysis. Because people need to stay awake and work efficiently on weekdays and choose to relax or go to parties on weekends, the group speculates that accidents due to the above two reasons will occur in large numbers on weekends. As shown in the figure [plotly_fig6], the number of drunk driving has increased a lot on Saturdays and Sundays, while it has remained at a low level from Monday to Thursday, but the number of unlicensed driving is keeping a lower level in weekends,this may be because the perpetrators of unlicensed driving may be forced to drive on weekdays. (If the link cannot be opened, please open *plotly_fig6.html* and in the *output* folder)

For the elderly and young people [plotly_fig7], whether they are drivers or passengers, it can be clearly seen from the line chart that the number is larger from Wednesday to Friday, and there will be a downward trend on the weekend, and the younger the number less. This is consistent with the results we have obtained before. (If the link cannot be opened, please open *plotly_fig7.html* and in the *output* folder)

Finally, the group produced a heat map [plotly_fig8] of the accident according to the week and month. It can be clearly found that the accident is significantly reduced on the weekend, and it is more concentrated on Tuesday to Friday from March to December. (If the link cannot be opened, please open *plotly_fig8.html* and in the *output* folder)

3.1.4 Conclusion

According to the above results, the group found that probability of accidents on weekends in Victoria is lower than on weekdays, and the number of drunk drivers does increase significantly on weekends, which all meet the expectations of the group. In addition, the number of accidents has increased significantly from 2013 to 2014. This may be due to the loss of accident data in 2013, or caused by the sharp increase in vehicles in 2014. After 2016, the number of accidents decreased slowly, indicating that the government or the driver effectively controlled the accidents.

3.2 Analysis 2 : Predictive modelling

3.2.1 Introduction

The result of the first analysis shows that the accidents is significantly reduced on the weekend, and concentrated on Tuesday to Friday from March to December, so the group inferred that the date may affect the number of accidents. In addition, weather conditions may also affect the number of accidents because severe weather conditions are more likely to cause traffic accidents [1]. This report will construct three machine-learning models (LinearRegression, Support vector machine and Random Forest) to predict the numbers of accidents in the future based on the date and weather conditions.

3.2.2 Dataset

All the added data [Source] used in this part is as follows:

- The daily weather data from 2013 to 2018: *Weather_Condition.csv*
- The forcasted weather data from 16/6/2020 to 25/6/2020: *future_weather.csv*

3.2.3 Workflows

Step 1: Data-Preprocessing Firstly, the group checked the collected dataset and find there are some missing value, thus we filled the missing value with the median. Next, for the float variables from attributes temperature, rainfall and solar exposure, the group applied the normalization on them by Max-Min Scale. Finally, the variables from attributes months and days are discontinuous variable which need to be encoded. There are two common encode methods: Ordinal Encode and One-Hot Encode, therefore, this report create two function to implement different encode methods for these discontinuous variables and will compare their effect in the validation part.

Step 2: Cross-Validation In this Part, the group utilize 10 folds cross-validation to evaluate the models with different encoded dataset, we set the mean R-square as our evaluation standard to judge the goodness of fit for a model.

Table 3: *Results for Cross-validation*

Model	LinearRegression		Support vector machine		Random Forest	
	Ordinal	OneHot	Ordinal	OneHot	Ordinal	OneHot
R-square	0.097567	0.103812	0.022964	1.192134	0.160295	0.1604

The result shows that dataset with One-Hot Encode perform better than ordinal, and the Random Forest model get best R-square value among the tested models.

Step 3: Train and Predict The group selected the dataset with One-Hot Encode as the final training dataset, then trained the models respectively. After training phase, the group used these trained models to predict the accident number for the future days (16/6/2020-25/6/2020).

3.2.4 Result and Discussion

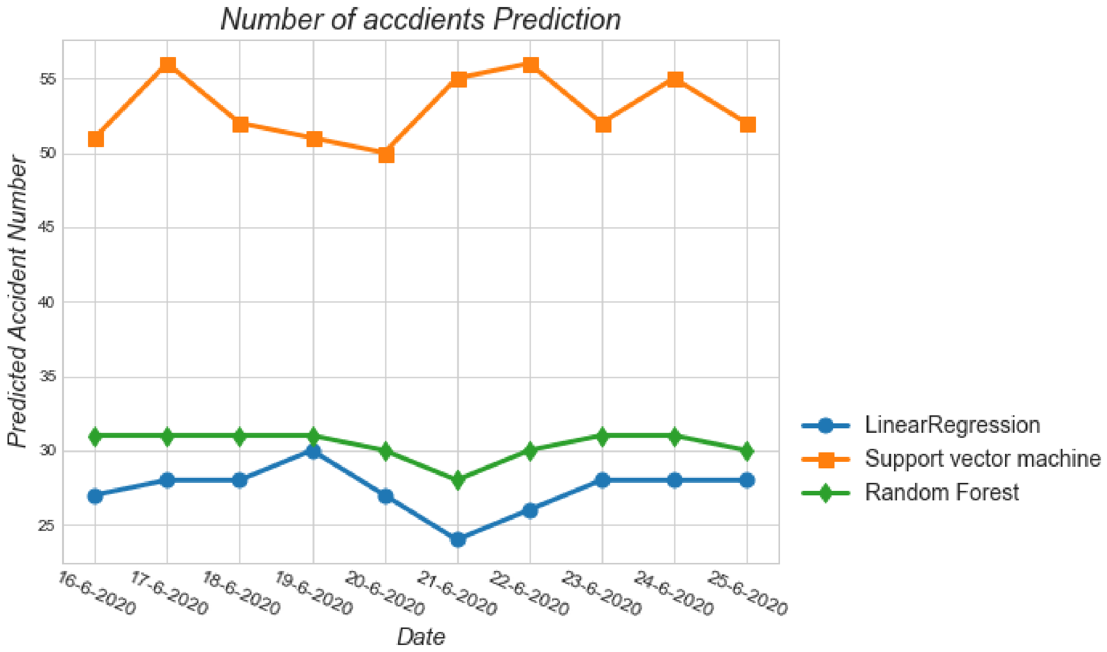


Figure 5: *The result of Prediction*

Figure 4 shows the predicted accident number for different models, Linear Regression model and Random Forest model have relatively close results, while support vector machine has very different predictions. From the analysis before, the accident number of weekend usually lower than weekdays. In the result of this analysis, the predicted number from LR model and Random forest model demonstrate the same trend, peaking on 19/6/2020 (Friday), then falling on 20/6/2020 and 21/6/2020 (Weekends). Therefore, this report finally chosen the result of LR model and Random forest model, and got the average of their prediction as the final result for this analysis.

Table 4: *Predicted Accident Number*

Date	LinearRegression	Random Forest	Mean
16-6-2020	27	31	29
17-6-2020	28	31	29
18-6-2020	28	31	29
19-6-2020	30	31	30
20-6-2020	27	30	28
21-6-2020	24	28	26
22-6-2020	26	30	28
23-6-2020	28	31	29
24-6-2020	28	31	29
25-6-2020	28	30	29

3.2.5 Conclusion

Through this analysis, the group successfully predicted the number of accidents in the next 10 days. However, the goodness of fit for the predictive models in this project is not good (only 0.16), which may be caused by the following reasons:

- The attributes of the train-data are not enough. Since the whole weather data is not free, the group only included four weather data to illustrate the weather conditions in this report, in the further research we may expand the attributes of train-data to get better prediction.
- The group did not use any processes of the feature engineering since the number of features (i.e., attributes) is small.

References

[1] J. B. Edwards. Weather-related road accidents in england and wales: a spatial analysis. *Journal of Transport Geography*, 4(3):201–212, 1996.

[2] W. D. Halsey. *Macmillan contemporary dictionary*. Macmillan Pub Co, 1979.