Climate Change and Road Safety (2012-2017)

Leonardo Linardi, 855915

Section 1. Introduction

The domain of this study is a combination of transport and environment.

The Transport Accident Commission (TAC), a Victorian Government-owned organization, teams up with Victoria Police, the Department of Justice and VicRoads to set up to be the world's leading social insurer by aiming for zero deaths and serious injuries on our roads. They are ambitious to find out what mainly cause accidents on roads and how could they prevent them through road safety strategies and campaigns. Now, a question arises, **does climate change affects road safety in Melbourne?**

Melbourne, city of 4-seasons-in-1-day, has unpredictable weather throughout the year. And with Melbourne being Australia's fastest growing capital city with up to 40% of people born outside Australia (Australian Bureau of Statistics, 2016), this means people have different backgrounds in driving vehicles. Given these circumstances, Melbourne is somewhat unique and must be treated differently. With the government investing more than $1 billion in road safety strategies (Towards Zero, 2016), the question now becomes non-trivial, as all aspects affecting road safety needs to be assessed, and innovative, as there hasn't been a clear explanation what exact climate feature that affects road safety, particularly in Melbourne. A deeper question is, **what specific climate features and to what extend does it impacts road safety?**

Section 2. Datasets

Data for Greater Melbourne is used, as proposed to the topic of this study. But should be relevant elsewhere in Victoria, as it could be abstracted away from the results. The datasets used are:

- Crashes Last Five Years – csv
  This comprehensive dataset contains detailed fatal and injury crashes on Victorian roads during the latest five-year reporting period (2012-2017). This crash dataset is used as a measure of road safety in Melbourne, indicating how safe and secure are the roads are. A snippet of the important features of the dataset that is going to be assessed is on Table 2.1. URI: https://www.data.vic.gov.au/data/dataset/crashes-last-five-years

- Climate Data Online, Monthly Climate Statistics – csv
  Contains data of different features of a climate change, from several weather stations in Victoria Park, VIC. The two main stations to be observed are from: Bundoora (086351), Mornington (086079). And other supplementary stations are from: Moorabin Airport (086077), Scoresby R.I. (086104) and Essendon Airport (086038). Those stations are chosen specifically as they cover major areas on Greater Melbourne. A more detailed reason for choosing these stations is explained on Section 3. All the datasets have the same general format, with some of the important features explained on Table 2.1. URI: http://www.bom.gov.au/climate/data/

Dataset Name	Important Features	Feature Type	Range (if any)	Data Values
Crashes Last Five Years	ACCIDENT_DATE	Continuous	Year 2012-2017	%d/%m/%y
	ACCIDENT_TIME	Continuous	24 hours	%H.%M.%S
	ACCIDENT_TYPE	Categorical	-	Strings e.g. Collision with vehicle, Struck Pedestrian, etc.
	REGION_NAME_ALL	Categorical	-	Strings in CAPS, e.g. METROPOLITAN SOUTH EAST REGION, etc.
Climate Data Online	Mean rainfall (mm)	Discrete	42.7 - 71.5 mm	Float64
	Mean daily sunshine (hours)	Discrete	24 hours	Float64
	Mean number of clear days	Discrete	0-31 days	Float64
	Mean 9am cloud cover (oktas)	Discrete	4.0 - 5.3 oktas	Float64

Table 2.1. Details of the datasets to be used in this report.

Section 3. Pre-Processing

Detailed steps on approaches to pre-process the data are explained.

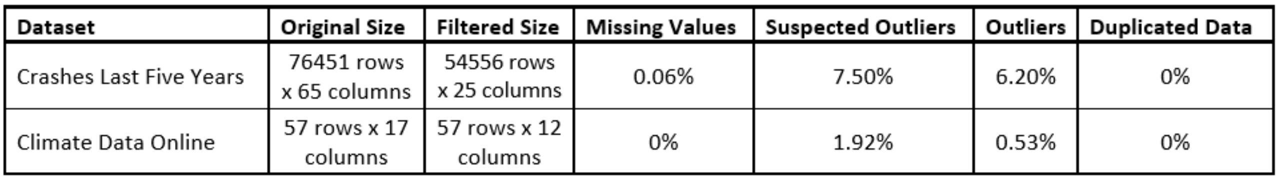
1. Datasets are imported from reliable sources mentioned in Section 2. Both data are in csv format, so no data transformations are held. The climate data, however, has file headers which are deleted for it to be imported and accessed through Jupyter Notebook.
2. Datasets are filtered by removing rows and columns with features that are not relevant to the objective of this study. Columns such as *'X', 'Y', 'ABS_CODE'*, are removed from the crash dataset, and columns such as *'Annual', 'Number of Years', 'Start Year'* are removed from the climate datasets. Comparisons between the size of the original and modified datasets are in Table 3.1. The filtered datasets are used from now on.

The two main climate datasets to be examined are from Bundoora and Mornington stations. They contain the most complete weather data suited on the best locations, with the Bundoora station representing North West (NW) Metropolitan Region and the Mornington station representing South Eastern (SE) Metropolitan Region of Greater Melbourne, the two regions where we want to observe our crashes from.

3. A quick skim of the climate dataset is made, as the dataset is relatively small and structured, to check if there are any missing values. Then, datasets are iterated over to search for missing values. The results are represented in Table 3.1.

The Bundoora dataset has missing values on the feature *'Mean 3pm wind speed (km/h)'* while the Mornington dataset has missing values on *'Mean daily sunshine (hours)'* and *'Mean 3pm relative humidity (%)'*, which are the important climate features going to be observed.

Instead, the Essendon Airport station, located in NW Metropolitan Region, the Moorabbin station, located in SE Metropolitan Region, datasets are used to supplement the missing feature values. However, the feature 'Mean daily sunshine (hours)' is replaced by the Scoresby dataset, located in Eastern Metropolitan Region (the next closest region), since there are no other stations in SE Metropolitan Region that could supplement the data.



Dataset	Original Size	Filtered Size	Missing Values	Suspected Outliers	Outliers	Duplicated Data
Crashes Last Five Years	76451 rows x 65 columns	54556 rows x 25 columns	0.06%	7.50%	6.20%	0%
Climate Data Online	57 rows x 17 columns	57 rows x 12 columns	0%	1.92%	0.53%	0%

Table 3.1. Results of the pre-processing made to all the datasets.

The crash dataset contains 873 null values all in the feature *'DAY_OF_WEEK'* and *'REGION_NAME'.* These features are removed as *'DAY_OF_WEEK'* can be derived from *'ACCIDENT_DATE'* and the *'REGION_NAME_ALL'* is used instead of *'REGION_NAME'*.

4. Outliers are checked by iterating over discrete data of both datasets, with data above or below 3*IQR as outliers, and data above or below 1.5*IQR as suspected outliers. Results are in Table 3.1.

Outliers are not removed from all datasets as they represent important findings and gives details towards each crash or climate instances.

5. Lastly, duplicated records on all datasets are searched. Results are in Table 3.1.

Section 4. Integration

Before integrating the crash and climate datasets, some classifications and new features are introduced. In the crash dataset, they are,

- *'ACCIDENT_MONTH'* to store what month each crash happens. Data are classified by months as the climate records are captured by months.
- *'REGION_NS'* to store in what region each crash happens. Either NE or SE Metropolitan Region.
- *'ACCIDENT_PERIOD'* to classify which period a crash happens. Either on early day (before 6AM), mid-day (between 6AM and 6PM) or night time (after 6PM).
- *'ACCIDENT_TYPE_NEW'* to simplify the classification of the type of crash. Either colliding with vehicles, objects, struck pedestrians or the vehicle overturned.

The crash dataset is then classified per month per region, to match the data provided by the climate dataset. In other words, both datasets are structured from January to December in NE Region followed by January to December in SE Region. Creating a total of 24 rows at both (crash and climate) datasets and integrated. The results of the integration are explained in Section 5.

The new features, such as *'REGION_NS'* and *'ACCIDENT_TYPE_NEW'*, are introduced due to inconsistent string values in the data records. For example, crashes that struck pedestrians can be written as *'Struck Pedestrian'* and *'Struck Pedestrians'*. The column *'ACCIDENT_DATE'* in the original dataset also contain formatting errors, where data can't be passed directly into pd.to_datetime built-in function by python. A rather naïve approach is used, splitting on every string to get the month value, which is rather time consuming. These are some challenges that demands extra work, whereas it will aid in processing the data later.

Section 5. Visualization and Results

Each feature of the crash and climate dataset are integrated and the Pearson correlation within each of them are computed and examined, to get a better insight on which climate feature change that affects the different aspects of a crash. The results are represented with heatmap in Figure 5.1.

The target features of the crash dataset are:

- Collision with vehicle
- Struck Pedestrian

As they are the aspects of the crash where it involves someone else. The other climate features are then ranked, to find which have the most significant impact towards the target features. The top 5 most significant feature is shown in Table 5.1.

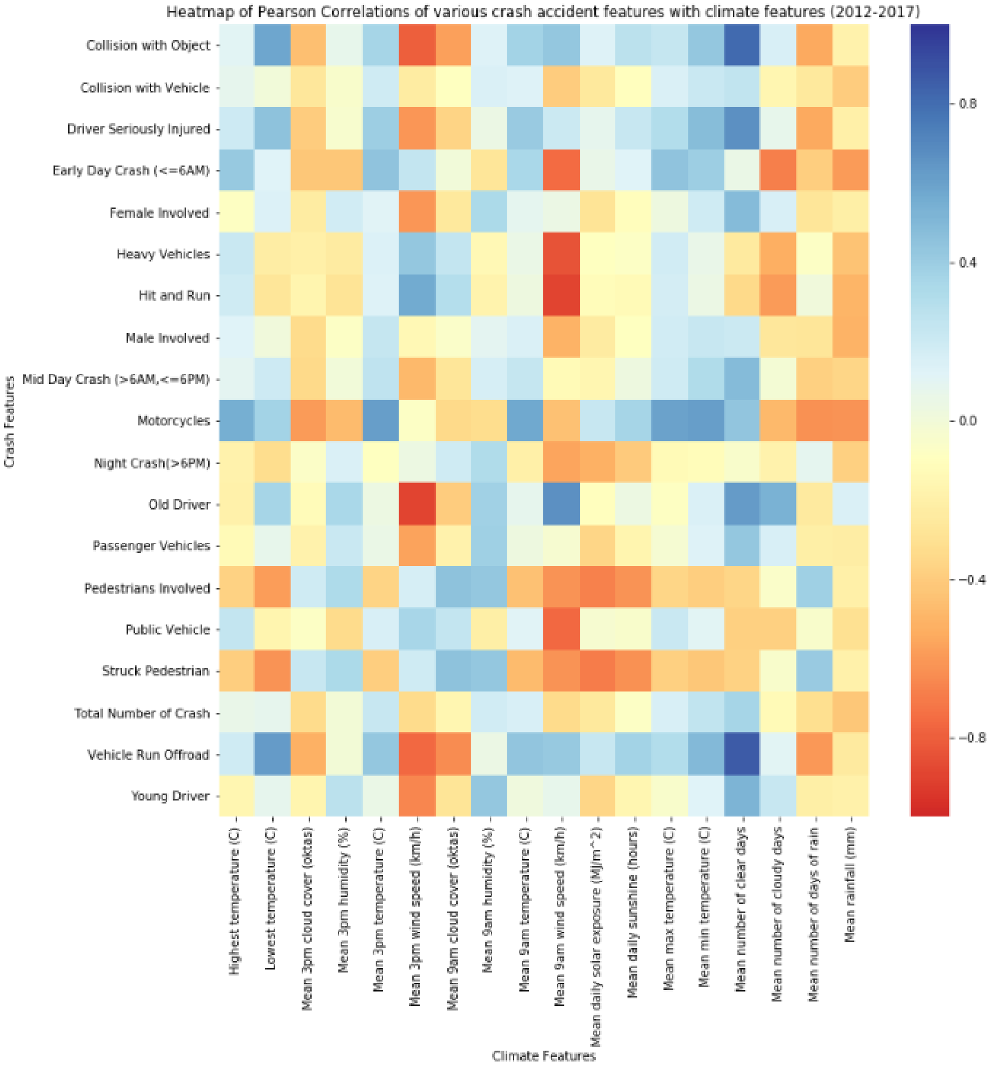


Figure 5.1. Heatmap of Pearson Correlation of every feature in all dataset

Climate Feature	Pearson Correlation
Collision with Vehicle	
Mean 9am wind speed (km/h)	-0.395309
Mean rainfall (mm)	-0.391593
Mean 3pm cloud cover (oktas)	-0.25891
Mean daily solar exposure (MJ/m^2)	-0.257057
Mean number of clear days	0.252664
Struck Pedestrians	
Mean daily solar exposure (MJ/m^2)	-0.692857
Mean daily sunshine (hours)	-0.629418
Lowest temperature (C)	-0.622285
Mean 9am wind speed (km/h)	-0.621794
Mean 9am temperature (C)	-0.474459

Table 5.1. Top 5 most significant climate feature.

As we can see, there is a mild negative correlation between various climate features towards number of crashes that collides with another vehicle. And quite a strong negative correlation on crashes that struck pedestrians.

A further processing of all the climate features are required, to find out which feature does gives impact towards crashes of different types.

Principal component analysis (PCA) is performed on each climate feature and represented on a scatter plot, shown in Figure 5.2. To visualize better on how the climate features are related towards each other.

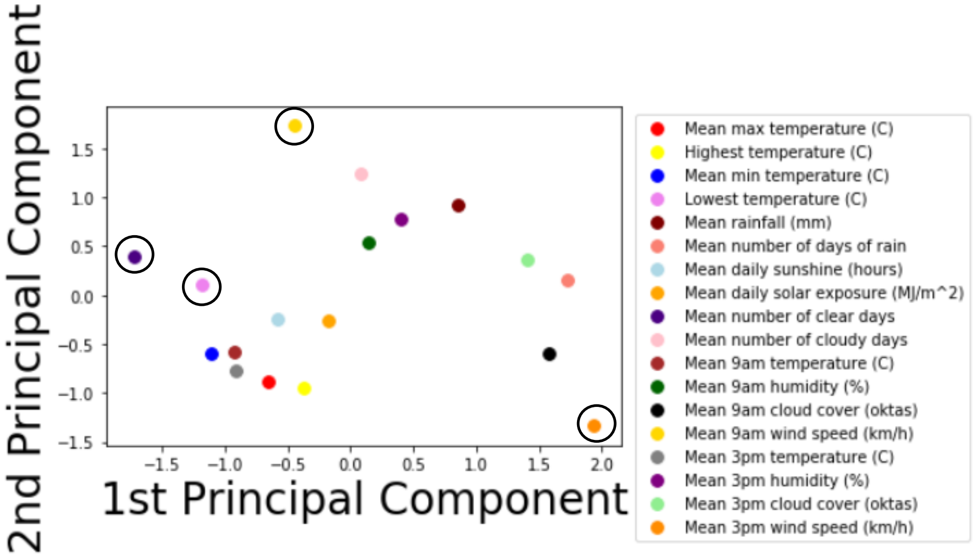
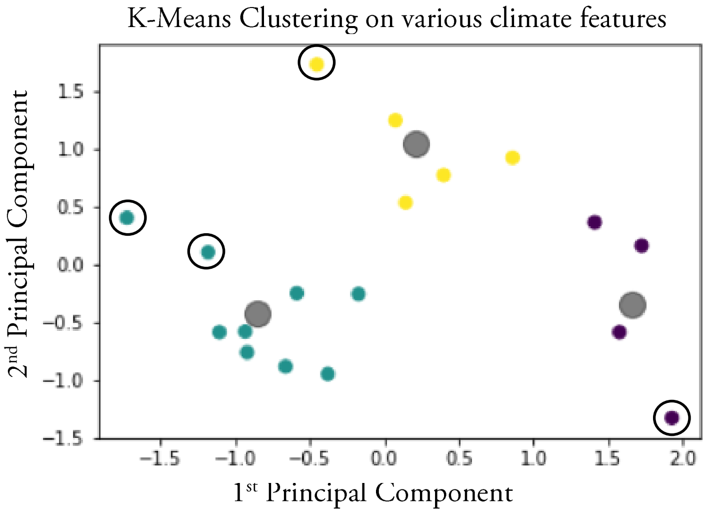


Figure 5.2. PCA on every climate feature.

The climate features are spread out with variance 0.53714228 on the 1st Principal Component (PC) and 0.30828713 on the 2nd PC. Furthermore, to give a better understanding on how the climate features are related, K-means clustering method, with k=3, are computed on every features. Results are on Figure 5.3.



The top most positively and negatively correlated climate towards crash features (not with respect to our target features) are computed, and they are:

- Mean number of clear days
- Mean 9am wind speed (km/h)
- Mean 3pm wind speed (km/h)
- Lowest Temperature

And could be visually seen in Figure 5.1.

Figure 5.3. K-Means Clustering on PCA of every climate feature.

Surprisingly, these top climate features that gave significant impacts are mostly located at the extremes of each clusters. The plots that has circles around it, in Figure 5.2 and 5.3, represent those features. As they are located at the furthest distance from the centroids of each clusters, those 4 features are least related towards each other, and could be the most unbiased, reliable climate feature to predict future crashes.

Those 4 features are extracted and computed in a K-Nearest Neighbor (K-NN) algorithm, to predict future crashes that involves other people (defined as 'People-Involved Crash' crash type). This is when a crash either collides with another vehicle or struck pedestrians, our target features. The monthly number of people-involved crash are classified as either Low (1505-1755.3), Medium (1755.3-2005.60) and High (2005.6-2256), divided through 3 equal-width bins. The K-NN algorithm analysis for the best K is in Figure 5.4 visualized with a line plot.

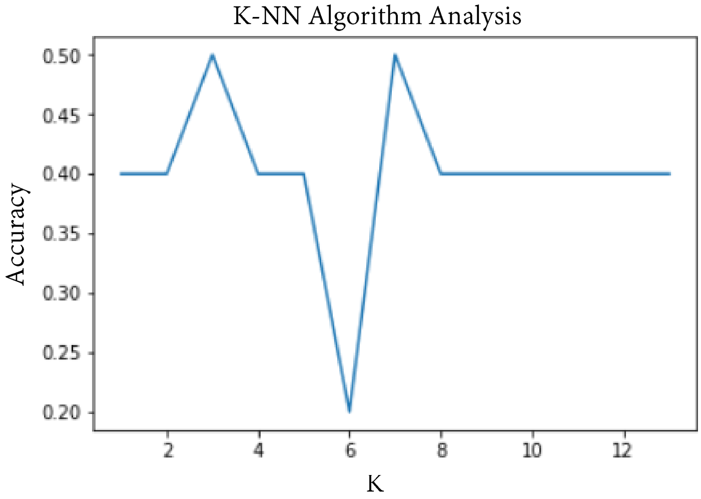


Figure 5.4. K-NN Algorithm Analysis

K=3 and K=7 yields the best accuracy for future predictions, with the best accuracy of around 50%. This concludes that there is not much chance to be able to predict how many number of monthly crashes in the future from a given climate data.

In overall, all the pre-processing, integrations and visualizations methods, helps us to find out which climate feature gave the most significant impact towards various features of a crash. And gave more clarity in terms of how distinct or related these significant climate features are.

Section 6. Conclusion

Several challenges are encountered, some mentioned in Section 4, and several limitations towards answering the question are,

- The climate datasets are categorized monthly, which could be more accurate if the data represented weekly, as this increases the accuracy towards assessing the correlations and making future predictions.
- The climate datasets features are also an average value of tenths or hundreds of years collected from the Australian Bureau of Meteorology, which decreases the accuracy of making predictions and finding correlations on the crashes, as the crashes can't be evaluated yearly (2012-2017), rather than on an average basis too.

In conclusion, the results help answering the question. And yes, climate change does affects road safety and there are lots of climate features strongly correlated, mentioned in Section 5, towards various crash types. Though it's is hard to make future predictions on how many crashes are going to happen the next month, it still gives an insight on which areas of road safety that the government should improve. For instance, the number of clear days gives a negatively strong impact on vehicle running off road (seen in Figure 5.1.), TAC could prevent this by adding extra street lights on street borders and edges.

Around 80% of the codes are written from scratch, from either the knowledge gained during workshops or through videos and tutorials publicly available in the browser.
The major python library used are pandas, matplotlib, numpy, seaborn, datetime and several others such as scipy and sklearn.

References

Towards Zero 2016-2020 Road Safety Strategy and Plan, 2012.
Australian Bureau of Statistics, 2016 Census QuickStats, 2016.