Data dictionaries :

NSW_crash_clean.csv:

Column Name	Description	Data Type
Crash ID	Unique identifier for a car crash	Integer
Degree of crash	Severity classification of crash : Fatal - Injury - Non casualty (towaway)	Text
Degree of crash - detailed	The severity classification (or degree) of the crash, incorporating injury severity.	Text
Year of crash	Year when the crash was reported.	Integer
Month of crash	Month when the crash was reported	Text
Day of week of crash	Day of the week when the crash occurred	Text
Latitude	The latitude of the crash location.	Float
Longitude	The longitude of the crash location.	Float
LGA	The Local Government Area where the crash occurred.	Text
Urbanisation	The urbanization where the crash occurred. Sydney metropolitan - all Sydney metropolitan LGAs Newcastle metropolitan - Newcastle City and Lake Macquarie City Wollongong metropolitan - Wollongong City and Shellharbour City Country urban - Other LGAs, where speed limit is up to and including 80 km/h. Country non-urban - Other LGAs, where speed limit is more than 80 km/h Country unknown - Other LGAs, where speed limit is unknown	Text
Conurbation 1	Distinguishes the Sydney-Newcastle-Wollongong greater conurbation from other urban and rural areas.	Text
Surface condition	The condition of the road surface at the crash location. Wet - Dry - Snow or Ice - Unknown	Text
Weather	The weather conditions at the time of the crash. Fine - Overcast - Snowing - Raining - Fog or Mist	Text
Speed limit	The maximum speed limit applicable at the crash location.	Integer
Road classification (admin)	The administrative classification of the type of road on which the crash occurred. Local - Regional - State	Text
Alignment	The road alignment of the road at the location of the crash. Straigth - Curved - Unknown	Text
hour	Hour when the crash occurred (more or less one hour)	Integer
Injuries	Binary value that told if accident lead to injury : 0 : Non casualty (towaway) 1: Injury or fatality	Integer
Date of crash	Month and year when the crash occurred.	Date

Regional_stat_by_LGA_2016_2020.csv:

Column Name	Description	Data Type
Region	Local Government Area	Text
2016	Population of LGA estimated on the 30th of June 2016	Integer
2017	Population of LGA estimated on the 30th of June 2017	Integer
2018	Population of LGA estimated on the 30th of June 2018	Integer
2019	Population of LGA estimated on the 30th of June 2019	Integer
2020	Population of LGA estimated on the 30th of June 2020	Integer

Sources:

NSW_crash_clean.csv:

Result of the cleaning of this file: NSW_Road_Crash_Data_2016-2020_CRASH.csv

The original file was obtained here: Open Data Transport NSW: Car Crash

You can access here to the complete raw dataset, and to the original dictionary.

Regional_stat_by_LGA_2016_2020.csv:

The data is obtained from the Australian Bureau of Statistics:

The entire dataset can be obtained here: Regional Statistics by LGA 2020, 2011-2020 (abs.gov.au)

I used the following filters:

- Data Item: Estimated resident population (n0.)
- Time Period:
  - Start: 2016
  - Stop: 2020
- Region: I checked all the regions in NSW

Regional Statistics by LGA 2020, 2011-2020 (abs.gov.au) - FILTERED

Data Cleaning:

NSW_crash_clean.csv: [1]

Column Name	Action to take	Taken Action	Result
'Time of crash - Two-hour intervals'	Handle null values	Hour is the core of my scenario. I dropped null values	Dropped 8 rows
'Time of crash - Two-hour intervals'	Change data type	Convert an interval to the median hour '10:00 - 11:59' -> 11	Replace by column 'hour'
Weather	Handle 900 null values	Weather is the core of my scenario I dropped the rows where the weather was 'Unknown'.	Dropped 900 rows
Weather	Handle outliers	Weather is the core of my scenario I dropped the rows where the weather was 'Other' (167), because it was undefined, and I treat them as 'Unknown.'	Dropped 167 rows
Urbanisation	Handle 27 null values	Replace null value by the value of the nearest crash	
Conurbation 1	Handle 26 null values	Replace null value by the value of the nearest crash	
Alignment	Handle 2 null values	I let the values until looking impact of these features on the target. Alignment doesn't have impact on the rate of injury	Dropped the column
Speed limit	Handle 35 missing values	Replace the null value by the most common value, for a fixed urbanisation	
Speed limit	Change data type	Keep only the numerical value of speed numbers '90 km/h' -> 90	Text to Integer
Year	Handle outliers	Some crashes reported in 2016 happened in 2015. To keep regular data, I deleted car crashes that happened in 2015.	Dropped 67 rows
Year and months	Create a date	Combine month and year to combine data into a date The day of the month is put at one. Format DD/MM/YYYY	Added a column, 'Date of the crash'
Surface Condition	Handle 14 missing values	Let the values until looking impact of this feature on the target. Surface condition has high impact on the target, but is highly correlated to weather, I keep only the weather.	Dropped the column

[1] Look at *Data_Cleaning.ipynb* to replicate the cleaning.

Regional_stat_by_LGA_2016_2020.csv:

Column Name	Action to take	Taken Action	Result
Region	Make it the same name as on the other file.	1 - On all names, I deleted the suffixes '(A)' or '(C)' or '(NSW)' 2 – Deleted the word 'Shire' when necessary 3 – Change 'Unincorporated NSW' to 'Unincorporated'	Both table can be joined by LGA.

Note:

Cleaning was made manually for LGA that ended by 'Shire' and for 'Unincorporated NSW'

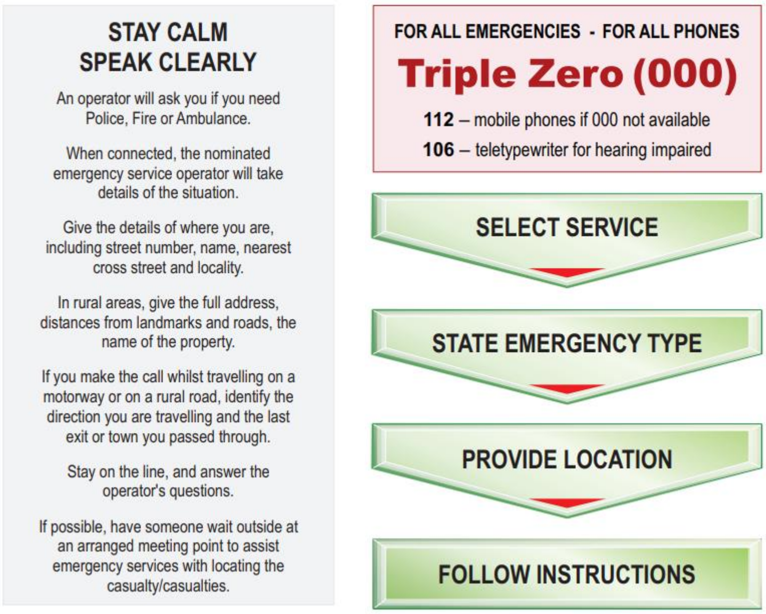
Cleaning was made with "replace" function otherwise.

Report: Car crashes in NSW Australia

Between 2016 and 2020, in NSW, every 24 minutes, a crash is reported to the police. And about 2/3 of them lead to injury.

When it comes to car crashes, there are many ways to talk about it. Seeing ads, I was thinking about looking at prevention and road safety, but it is difficult to summary the prevention around Australia, as a lot of distinct association drive local actions about it. To stay on this axis, I decided to look at the causes of crashes (alcohol, high speed, …) but it is difficult to have data on it.

I then begin to think, not about the cause of accidents, but about consequences. What happens after an accident when someone calls an emergency number? I know that there are a lot of charts explaining how to make a to call 000 and give the information:



STAY CALM
SPEAK CLEARLY

An operator will ask you if you need Police, Fire or Ambulance.

When connected, the nominated emergency service operator will take details of the situation.

Give the details of where you are, including street number, name, nearest cross street and locality.

In rural areas, give the full address, distances from landmarks and roads, the name of the property.

If you make the call whilst travelling on a motorway or on a rural road, identify the direction you are travelling and the last exit or town you passed through.

Stay on the line, and answer the operator's questions.

If possible, have someone wait outside at an arranged meeting point to assist emergency services with locating the casualty/casualties.

FOR ALL EMERGENCIES - FOR ALL PHONES

Triple Zero (000)

112 – mobile phones if 000 not available

106 – teletypewriter for hearing impaired

SELECT SERVICE

STATE EMERGENCY TYPE

PROVIDE LOCATION

FOLLOW INSTRUCTIONS

But the witness's talk can be irrelevant by a lot of things: his misestimation of danger, his shock, some technical problem on the channel of communication, …

This problem about the question of car crashes leads me to look at three problems:

Problem 1: Location of accidents

Can we have a global representation of the number of accidents in NSW and are there some locations where the concentration of accidents is greater than average.

Problem 2: The number of accidents.

The goal is to find a pattern in two scales. Firstly, I want to see if there is one daily pattern to see the distribution of the accident throughout the day. Secondly, I want to look at seasonality is the number of accidents, to forecast the number of car crashes for the next few months.

Problem 3: Can we help an emergency operator to predict the severity of an accident with the minimum of information.

I estimated that the only sure information an emergency operator can access for is location (and what can be deduced from it), weather, and time of the crash.

My goal is to conduct a knn-model on these features, to classify crashes in 2 categories : injury or not.

Location of accidents:

By looking at the locations of car crashes on the map, we can see that there are concentrated around Sydney greater area, News Castle greater area, and in general, around cities.

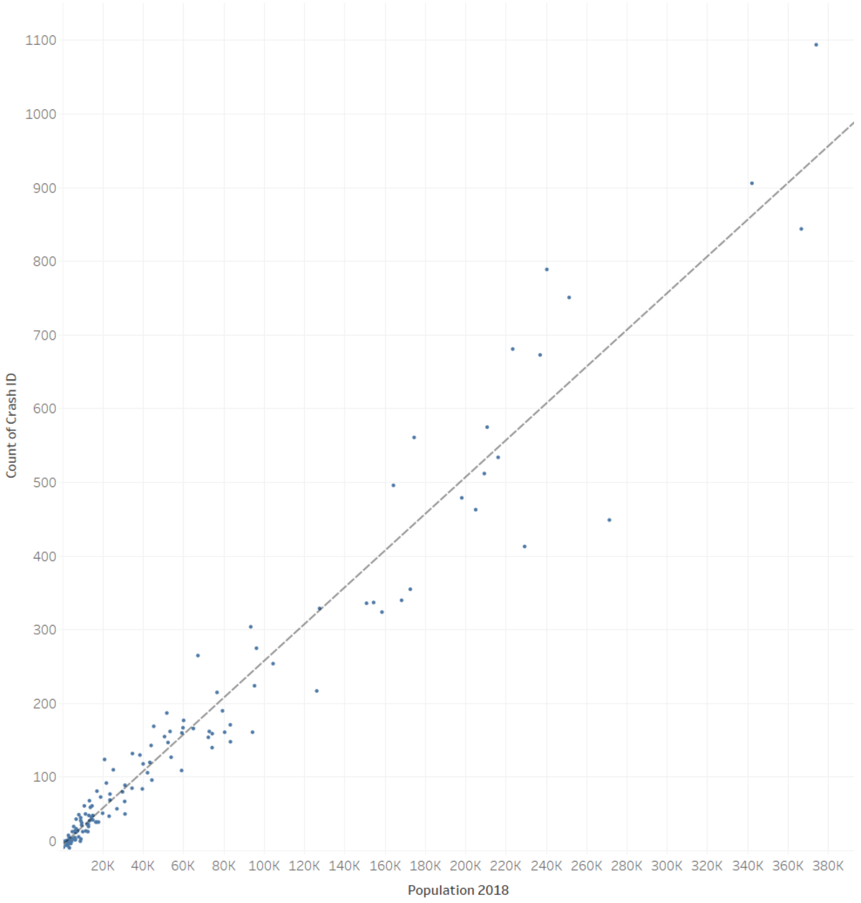
More than half of accidents happen in Sydney metropolitan.

It was expected that car crashes are more frequent when population density is higher. I looked at information on populations' density in each Local Government Area.

But the density is not reliable. Indeed, the median density is only 7.07/km^2, while it varies 0.01/km^2 to 8334.96/km^2. And in the largest LGA, people (like car crashes) are concentrated in the main city. So, the surface of the LGA decreases the population's density without being relevant. I choose finally to focus on the total population of each LGA and look for correlation with the number of accidents.

For each year, we obtain a p-value <0.001, that means that the values are highly correlated. Moreover, we can see that the number of crashes decreases in all counties of NSW, while population increases.

Correlation between population of LGA and number of crashes that happened in 2018. Both are highly correlated.

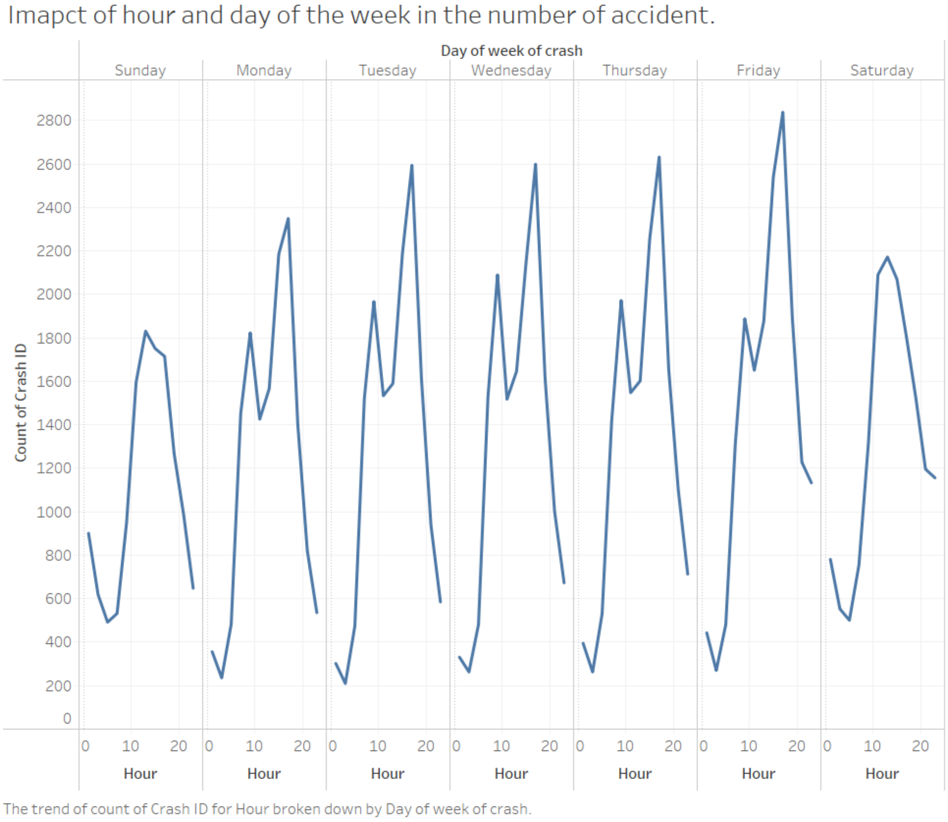


The trend of count of Crash ID for Population 2018. Details are shown for LGA.The data is filtered on Year of crash, which keeps 2018.

1 Example of plot for data in 2018.

Number of accidents:

On one hand, I look at the distribution of accidents along the day.



2 Observations of daily patterns of distribution of accidents throughout the day.

We can see a pattern in the number of accidents that happen every two hours interval. I distinguished two distinct patterns between the weekend and weekday.

On weekends, there are more car crashes between midnight and 4:00 a.m., probably because of people who come back from a party.

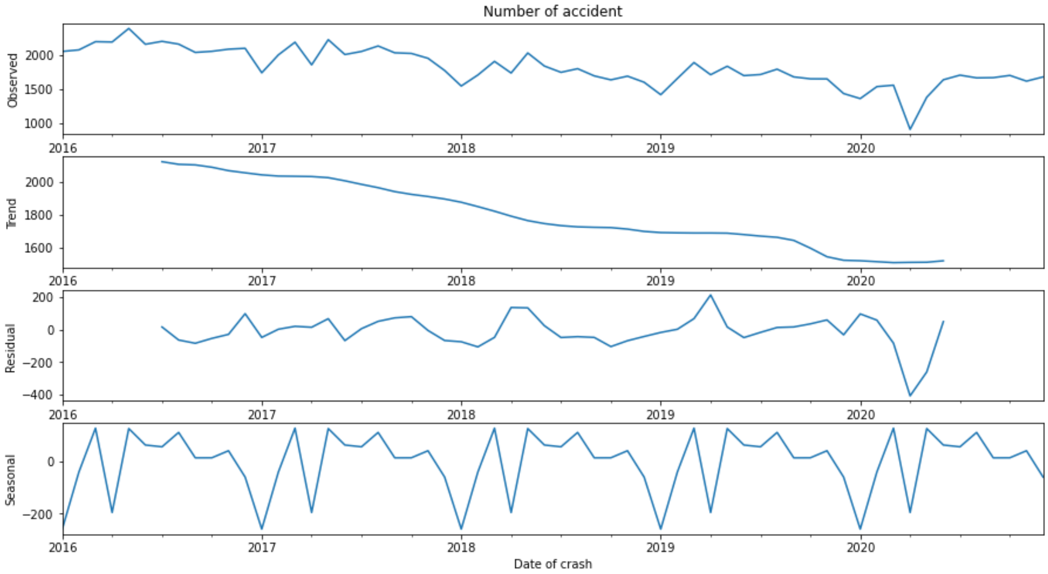
On weekdays, there are more car crashes during the day, and there are two peaks around 9 a.m. and 5 p.m. that coincide with the hour when people commute between work and home. The peak around 5 p.m. is more important, and it increases every day from Monday to Friday. That confirms precedent analysis around the world that says that rush hour is the most dangerous hour[2]. A study showed that “tiring work” and “uncomfortable position” is a risk factor of accident [3] and can explain the high number of accidents.

[2] “Crashes by Time of Day and Day of Week”, in Motor Vehicles, NSC INJURY FACTS, Car Crashes by Time of Day and Day of Week - Injury Facts (nsc.org), (consulted Nov 10, 2020)

[3] CHIRON Mireille, “Tiring jobs and work-related injury road crashes in the GAZEL cohort”, *Accident Analysis and Prevention,* May 2008

On the other hand, I looked at the monthly crashes and the number of accidents decreases over the year.

I observed an annual pattern: The number of accidents is always lower in January, and present at least two peaks in March and May. By decomposing the number of accidents, I was able to better understand the seasonality.[4] (I used additive seasonality.)



Note: the low number of accidents in April 2020 is not only explained by seasonality, but mostly by the restrictions given by the NSW government to limit the COVID crisis. Indeed, the first restrictive measures (including limitations on travel only when it is essential) took place on the 18th of March 2020 and were eased on the 10th of May 2020.

Since most accidents happen during rush hour, the low number of accidents in January and April can be explained by school holidays, that encourages parents to have some holidays and then not commute from work to home. That can equally explain the low number of crashes in September and October because Spring holidays cover the last week of September and first week of October. The same phenomenon happens in June and July, but it is less significant.

[4] See Seasonality.ipynb to reproduce the decomposition

Prediction of accident severity:

My main problem was to see if a model could predict the severity of an accident. I aggregated the “degree of crash” in a binary value: 0 if there is no injury, 1 otherwise. Indeed, the goal is not to predict exactly the degree of potential injury, but only to determine if the victims need some medical assistance.

Particularly, my goal was to focus on information that can be immediately verified by an emergency operator. I keep at first all this information:

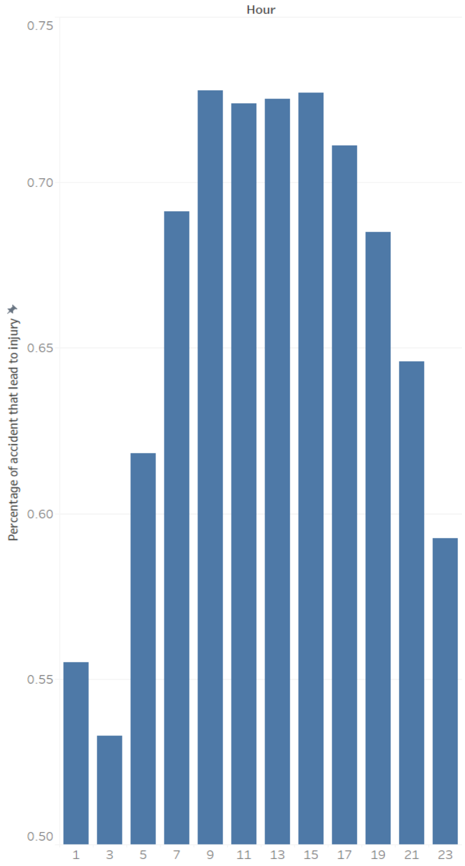
- The time of the crash:
  - Month of crash
  - Day of the week of the crash
  - Hour
- Weather:
  - Weather
  - Surface condition
- Location (since emergency services can locate a phone call)
  - Latitude and Longitude
  - Speed Limit
  - Road Classification
  - County
  - Urbanisation
  - Conurbation 1

(All the last features can be deduced from Latitude and Longitude using a GPS or a map).

A) Choice of features for the model [5]
   1) The impact of each feature on the target[6]:

[5] See Knn.ipynb to have the process of the features with appropriate charts.

[6] See Car_Crash_NSW.twbx to have access to all the charts.



3 Example of bar chart to see the impact of a feature: Hour of the day

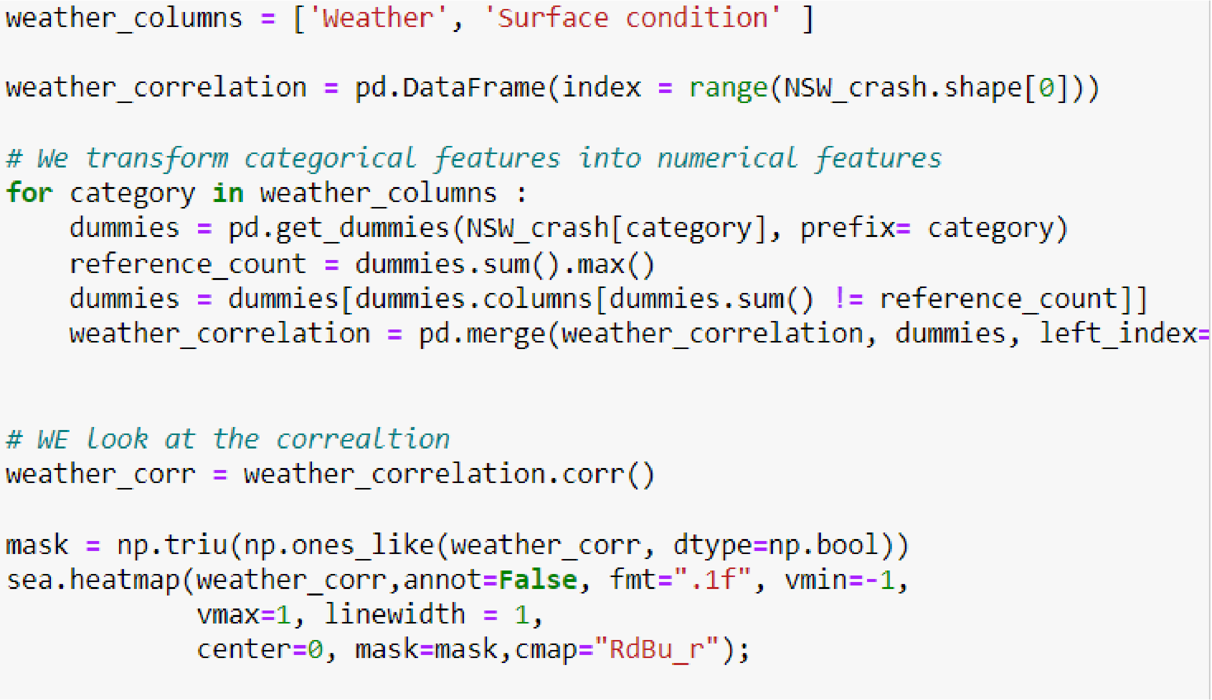
By looking at the percentage of crashes that cause injury, I can estimate the impact of a feature on my target. A feature was estimated important if the range of the percentage was greater than 5%.

2) Correlation of features.

I assumed that features under the same group (as defined before) are more likely to be correlated but independent of features in other groups : the speed limit can be dependent on road classification but is assumed independent of weather.

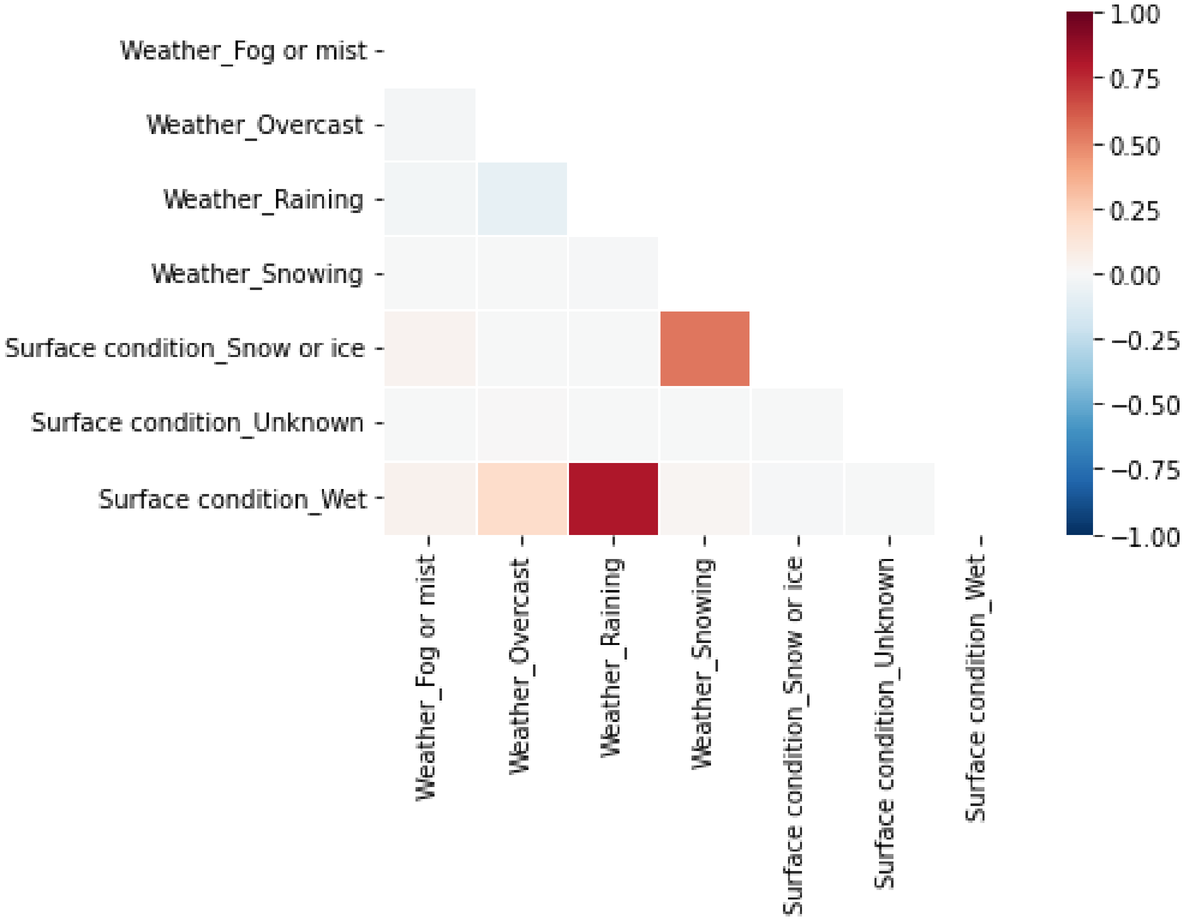
That was made to continue the selection of features and eliminate correlation that can induce an error in the model.

Weather and Surface condition were correlated, I keep only the Weather.



```
weather_columns = ['Weather', 'Surface condition' ]

weather_correlation = pd.DataFrame(index = range(NSW_crash.shape[0]))

# We transform categorical features into numerical features
for category in weather_columns :
    dummies = pd.get_dummies(NSW_crash[category], prefix= category)
    reference_count = dummies.sum().max()
    dummies = dummies[dummies.columns[dummies.sum() != reference_count]]
    weather_correlation = pd.merge(weather_correlation, dummies, left_index=


# WE look at the correaltion
weather_corr = weather_correlation.corr()

mask = np.triu(np.ones_like(weather_corr, dtype=np.bool))
sea.heatmap(weather_corr,annot=False, fmt=".1f", vmin=-1,
            vmax=1, linewidth = 1,
            center=0, mask=mask,cmap="RdBu_r");
```

4 How to see the correlation between features on the same group



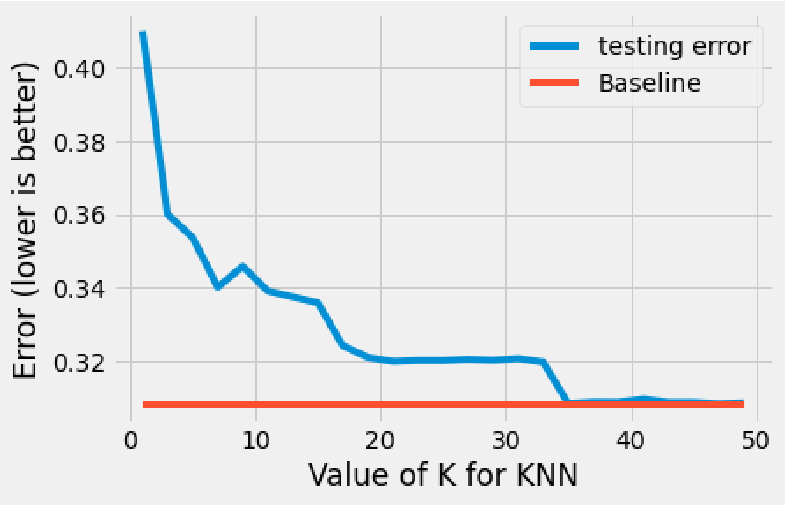
5 Correlation heatmap between weather and surface condition.

B) Knn-model

Our baseline was to say that any crashes cause injury.

That gives us accuracy equal to 0.692

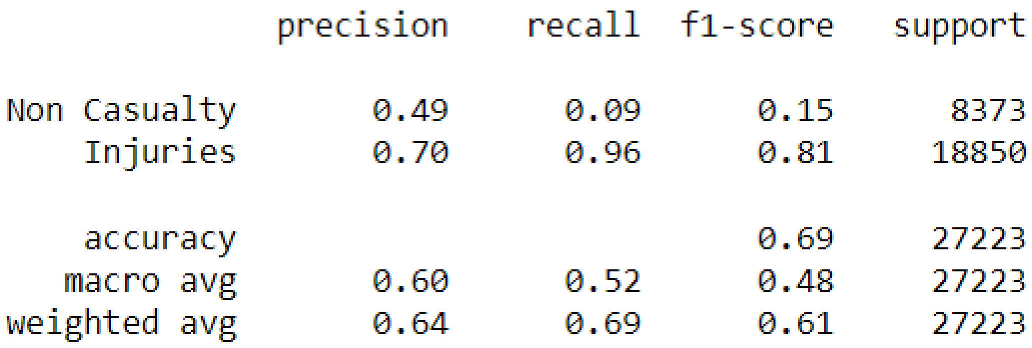
I made my first model using the following features: hour, weather, speed limit and road classification.



6 Looking for the best value of k for my k-NN model.

Finally, I obtained a model that is better than the baseline but when k is big enough (k> 30), and at this point, it acts likely like my baseline model.

Here is the classification report for k = 35:



```
               precision    recall  f1-score   support

Non Casualty        0.49      0.09      0.15      8373
    Injuries        0.70      0.96      0.81     18850

    accuracy                            0.69     27223
   macro avg        0.60      0.52      0.48     27223
weighted avg        0.64      0.69      0.61     27223
```

7 Classification reports of k-NN model

Here is the predictive table:

		Real degree of crash	
		Non-casualty	Injury
Predicted Degree	Non-casualty	488	514
	Injury	7885	18,336

We see that 97% of injuries are detected, the sensitivity is very high. But the sensitivity is very low. That is because 96.3% of accidents are predicted to cause injury, while the real percentage is only 69.2%.

For a second model, I used Hour, Weather, Longitude and Latitude.

Latitude and Longitude are correlated since most accidents happened in Sydney metropolitan. That's why I didn't use it on my first model.

I obtained mostly the same result, with a knn model that is less efficient than our baseline model.

So we can finally say that we can predict the severity of an accident without witness's information, and we have to trust the emergency operator experience and ability to ask the right questions and combine the answers.

CONCLUSION:

The amount of data was not overwhelming by itself. Coming from an official source, the data (for the part that interested me) was mostly clean, and the dictionary very detailed. But it was the first time I use knn model and I still have. But it is impossible to predict crashes' severity with so little information. The data we need are data that has to be collected directly from the witness. By having a dataset with this information, we could build a decision tree to help emergency to ask more relevant questions.

I would like to go back over other data. I discovered some patterns that could be more precise if I managed to find more precise data. (Daily data to really see the difference between holidays and workdays for example). Maybe, if I consider the witness's information, I would be able to develop a more efficient model. (I know for example that if a pedestrian or a bike is in an accident, the accident always caused injury.)

In terms of skills, I would like to progress in making Time series forecasts. I use seasonality in Python but forecast in Tableau. But I would like to explore more models, like ARIMA model, in Python. In Tableau, I still miss some skill to compose a dashboard.