DATA SET:

While searching for a meaningful dataset for my capstone project, I came across data.govt.nz*;* a repository of open datasets.
This related to all kind of activities throughout New Zealand and published by the central government. The chosen dataset comes directly from the *New Zealand Transport Agency.*
It contains data from car crashes since January 1st, 2000 until the present day and it's automatically updated on a quarterly basis.
The dataset comes with a pdf file containing a clear definition for each of the available features.

The dataset has an initial total of 655,697 samples. Each with 89 different features, both numerical and categorical.Simply by reading the definition of each feature from the pdf mentioned above, we can see that there are some features whose values are derived from other features.

These *derived* features don't add intrinsic value to the dataset, since they are used, for example, to summarize the characteristics of the crash in a more concise manner and, as the pdf describe, as used for administrative purposes.

Thus, we will remove them to prevent unnecessary feature correlation. This at the same time will help in keeping the size of the dataset to a minimum, which will help in the training time of the models.

On the other hand, there are a few features whose values are provided after the crash has been reported and processed, based on its severity. Specifically, the features *minorInjuryCount*, *seriousInjuryCount,* and *fatalCount* (which are self-explanatory) are provided in such way. Therefore, they create data leakage into our target variable and we must remove them. Plus, these features are not available at the time when the crash is first reported –which is when our model would be making live predictions– and emergency agencies need to dispatch a proper response as quickly as possible.

Besides these particular features, most of the rest are categorical. Indicating different condition at the time of the crash. Like *weatherA* and *weatherB,* both used to denote two different aspects of weather. Or *light, roadCurvature, trafficControl* and *numberOfLanes* which are used to describe the different natural light conditions, curvature level of the road, if there were traffic signals and which ones and the number of lanes the road had; respectively.

Finally, most of the numerical features represent counters that indicate the number of objects involved in the crash.

After understanding what each feature represents, we need to explore their respective values to understand their distribution and if there are any inconsistencies from their definition.