Algorithmics for Data Mining

Vehicle accidents preprocessing

First delivery

C., Laura
G., Carles

June 30, 2018

Contents

1 Introduction

For many years, people have been using different types of vehicles in order to transport themselves from one place to another. As technology evolved so did the ways of transportation and thus automobile vehicles were introduced in conjunction to the usage of animals as means of transportation. Nowadays they have almost fully substituted them.

In the later decades, the creation of highways as well as the road's improvement have helped this transition from animals to engine powered vehicles such as cars. However, a major drawback associated to it is the occurrence of multiple accidents, which are both unpredictable and unavoidable and are caused mostly by human cause.

Until fully automated automobile vehicles come eventually around (if they do), where no human interaction is needed while they drive themselves, this occurrence of accidents can be diminished but not fully eradicated.

This situation seems to be going around for many more years and, in conjunction with its gravity, it has picked our interest. That is why we have chosen it as the topic to study, checking which variables can affect and increase the probabilities of an accident occurring.

We have found three datasets related between them published by the UK Government where all accidents from 2016 are provided.

The purpose of this first project is to preprocess vehicles' accidents data to be able to perform afterwards a study on it to see if there are conditions that favor the occurrence of an accident.

To summarize, **the preprocessing is threefold** and its objectives are:

1. **Read the datasets** and choose the interesting features, leaving out the rest.
2. **Merge** these three datasets, aggregating two of them to be able to merge by appending them to the main dataset.
3. **Briefly explore them** to get familiar with the data.

2 Dataset

The government published a total of three files in *Comma Separated Values* (*.csv*) extension:

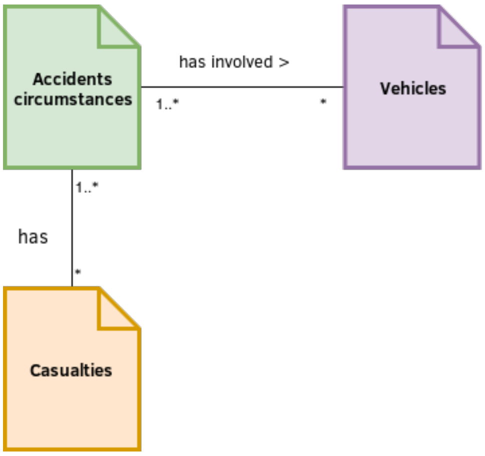
- **Accident circumstances**, where the general circumstances and main information related to an accident are found.
- **Vehicles involved**, where all the data related to the vehicles involved in an accident is found.
- **Casualties**, containing all the victims data.

In addition to those three datasets a guide stored in a spreadsheet too is provided in order to understand what each variable and label represents.

All three datasets contain a lot of information:

	# **Individuals**	# **Variables**
Accident	136.621	32
Vehicles	252.500	24
Casualties	181.384	16
Total	**570505**	**72**

They are related as seen in this diagram:



The number of individuals for each dataset is large which is great for our future study since this means that the sample of a whole year can be representative.

However, the number of variables is also somewhat large (a total of 72 variables) and we will need to cut them down and prioritize the most interesting ones.

Furthermore, lots of variables are not continuous numeric values but categorical labels with many ranging options, with up to 10 labels, which can add a lot of complexity in the future when studying the dataset.

That is why a simplification of the labels in top of the reduction of the number of variables is required in our eyes.

3 Preprocessing

In this section we are going to explain what features we have kept and how, followed by an explanation of the simplification of variables that we have performed in order to decrease the complexity a little of the dataset in the next report.

Our implementation can be found in an .R file attached to this document, as well as all the datasets (provided in a Google Drive link).

We need to make sure that the final dataset meets certain **requirements**:

- **Lack of missing values**: After taking a lot at the datasets, we've encountered several cells where missing data was present. Therefore, we want to obtain a final dataset as "clean" as possible.
- **Dimensionality reduction**: Having 3 datasets, both the complexity and the amount of data would make impossible use any data mining method. Therefore, Our first priority (and goal) is going to be the reduction of factors and its future merging, in order to ensure that only one dataset remains. Along with the merging and for the same implicit reasons, a feature selection will be done. The process of the selection of relevant dimensions will be discussed.
- **Level simplification**: Several features of the dataset contain a fine graded level of information. Although it may seem good, some of the categories can be directly reduced to two simple classes, indicating a binary variable. Thus, it will result in a simpler dataset, providing faster computational time (for any given data mining algorithm) as well as having more clearer values.

3.1 Dimensionality reduction

First of all, to reduce the dataset greatly so it can be easily computed afterwards in the next operations, we have reduced the number of dimensions of it. In other words, we have performed a feature selection.

This selection has been performed both objectively and subjectively.

First and foremost, we have removed all those variables that were repeated and/or were not interesting for us. That is, we did not find them as interesting as those we kept and thus we removed them.

Afterwards, we performed a small check on how many missing values there were in the resulting datasets.

In the case of the accident dataset, **we found that ∼57k of rows had missing data after having performed a feature selection**, out of a total of ∼136k rows. A deeper check confirmed us that ∼56k of those rows were missing data from one of our selected *interesting* features.

We decided that more than 1/3 of data was too much to be imputed in the next steps (since the ratio of *predicted data* versus the real data would be too high), and sacrificing a third of our data was out of question.

So our decision was to remove this feature that contained so much missing data. That is, removing yet another dimension.

We also removed the rows that contained missing Latitude and Longitude values, since there were only 7 cases of them and the coordinates are somewhat precise and not continuous, thus we do not consider them as imputable. This is because we don't want to deal in the future with coordinates that somehow point to in the middle of a building or someplace innaccessible by a vehicle.

Finally, the same unimputability rule has been applied to the *Time* and *Speed Limit* variables and we have removed a couple of rows containing missing data on that field.

The resulting datasets from this dimensionality reduction have left us with way less variables and around 100 less rows in the accident dataset. In the table below we have also explicitly computed the % of information retained from the original datasets in the new *clean* ones, as well as its counterpart: the percentage of information lost.

	# Individuals	**# Variables**	**% Info. retained**	**% Info. lost**
Accidents	136.575	17	53,125	46,875
Vehicles	252.500	5	20,83	79,17
Casualties	181.384	5	31,25	68,75

3.1.1 Accidents kept features

The accident dataset features that we have kept are the following:

- **Accident Index.** The id of the accident.
- **Longitude.** Longitude of the coordinates of the accident.
- **Latitude.** Latitude of the coordinates of the accident.
- **Accident severity.** Severity of the accident.
- **Number of vehicles.** Number of vehicles implicated in the accident.
- **Number of casualties.** Number of causalities resulting of the accident.
- **Date.** Day, month and year of the accident.
- **Time.** Hour and minute of the accident.
- **Road Type.** The type of road where the accident happened (Highway, urban, rural...)
- **Speed limit.** Speed limit of that road in *mph*.
- **Junction Detail.** If there was any junction near the accident.
- **Light conditions.** Light conditions by the time the accident happened.
- **Weather conditions.** Weather conditions by the time the accident happened.
- **Road Surface conditions.** Road Surface conditions by the time the accident happened, which can be wet surface or the presence of another substance on site, such as mud or oil, for example.
- **Special conditions at site.** Special conditions of the road where the accident happened.
- **Carriageway hazards.** Specifies if any object was in the road by the time the accident happened.
- **Did police officer attend scene of accident.** Specifies if police came after the accident.

3.1.2 Casualties kept features

The casualties dataset features that we have kept are the following:

- **Accident Index.** The id of the accident where this casualty was involved.
- **Casualty class.** Whether the casualty was the driver, a passenger or a pedestrian.
- **Sex of casualty.** The casualty's gender. Male or female.
- **Age of casualty.** The age of the casualty.
- **Casualty severity.** The severity of the casualty's injures. Ranges from Fatal to serious and slight.

3.1.3 Vehicles kept features

The vehicles dataset features that we have kept are the following:

- **Accident Index.** The id of the accident where this vehicle was involved.
- **Sex of driver.** The driver's gender. Male or female.
- **Age band of driver.** Age band of the driver: [0 - 10), [10 - 20)... etc.
- **Age of vehicle.** How many years old the vehicle was in the moment of the accident.
- **Vehicle type.** Its type: a motorcycle, a car, a van...

3.2 Simplification of variable levels

After feature selection, several kept variables had a large amount of categories that could be simplified. In order to make sure that the levels are correctly identified, we make sure that no category is mislabeled by having discussed which categories would fall into the corresponding binary value.

However, for the rest of the categorical values, we have decided not to transform its values into the actual string, since the size of its value would greatly increase the dataset size as well as the computational time of the data mining algorithms.

From the accidents dataset, the following variables have been simplified:

- **Light Conditions.** Now the light conditions can be either *Good* or *Bad* for proper driving.
- **Weather Conditions.** Now the weather conditions can be either *Good* or *Bad* for proper driving.
- **Road Surface Conditions.** Now the road surface conditions can be either *Good* or *Bad* for proper driving.
- **Carriageway hazards**. Specifies whether there was a hazard in the carriageway or not: value is either *TRUE* or *FALSE*.
- **Special Conditions at site**. Specifies whether there were extra conditions at the site at the moment of the accident. Value is either *TRUE* or *FALSE*.

As we can see, all of the previous variables represent on their levels any given conditions that could increase the chances of an accident to happen. Therefore, we have decided to simplify those categories in simple binary ones.

3.3 Imputation

Now that the three datasets have been reduced by feature selection and simplified we can impute the missing values that we did not remove on the *Dimensionality reduction* section.

This imputation will only be done to the accidents dataset. We don't want to impute values in both the casualties and vehicles datasets because the number of missing values is very high in the case of vehicles as can be seen in the table below, and we do not want to bias the obtained results by imputing so many values. It is not imputed in the casualties dataset either because we want to apply the same rules to both datasets.

We also take into account that we are going to aggregate the datasets of casualties and vehicles in the next section, thus not wanting to impute values that will count afterwards which are not real at all.

Since the rows with missing data in the accidents circumstances dataset is less than 1% we impute the values there.

	# rows with NA	**# fields with NA**	**% rows with NA**
Accidents	1083	1939	0.79
Vehicles	81748	97971	32.37
Casualties	2862	2898	1.57

The method chosen to impute the values has been K Nearest Neighbours, also known as **KNN**, which we have seen in class.

Instead of implementing it ourselves, we have used a method provided by the *DMwR* package in R.

We decided the ***k* value** (the number of neighbours on which to look to impute the data) to be **11**.

We wanted it to be somewhat large and found that taking the average of the 10 first neighbours was a reasonable solution. Since we wanted to avoid ties we chose to put an odd number, thus selecting 11.

3.4 Merging

We have merged the datasets by aggregating the casualties and vehicles ones.

The aggregation is done because for each accident there can be multiple casualties and vehicles involved (or none at all). Which means that there was no correspondence of 1-on-1 rows between the datasets.

To be able to merge the datasets for an easier computation afterwards in future studies, we have decided on **performing an aggregation of vehicles and casualties**.

The computation of the aggregation of these new data frames has taken more than $\tilde{2}$h due to the dimensions of them.

This aggregation has been done following these steps:

1. Create a new empty dataset.
2. Taking the accidents' id.
3. Iterating by those ids one by one and finding the related casualties and vehicles.
4. Extracting the information related to that accident (if present) and added a new entry with the accident id as its key.

3.4.1 Vehicles aggregation

The information extracted in the aggregation of the vehicles with its possible values is the following.

- **Accident ID.** The id of the accident where this vehicle was involved.
- **Involved vehicle type.** The type of vehicle involved in the accident. The variable can represent an accident involving only motorcycles (1), cars (2), trucks (3), animals (4) or a mix of those (5).
- **Number of male drivers** involved in the accident.
- **Number of female drivers** involved in the accident.
- **Disparity age driver**. The disparity of ages between the drivers of the involved vehicles.
- **Disparity age vehicles**. The disparity of ages of the vehicles involved.

3.4.2 Casualties aggregation

The information extracted in the aggregation of the casualties with its possible values is the following.

- **Accident ID.** The id of the accident where this vehicle was involved.
- **Male casualties.** The number of male casualties.
- **Female casualties.** The number of female casualties.
- **Age disparity casualties.** The disparity of age between all the casualties.
- **Worst casualty severity.** The worst level of severity in the injuries of the involved casualties.
- **Num. casualties driver.** The number of casualties involved that were the driver.
- **Num. casualties passenger.** The number of casualties involved that were a passenger.
- **Num. casualties pedestrian.** The number of casualties involved that were a pedestrian.

Now that the three datasets have been preprocessed and merged into a single large file we can proceed onto a brief study of the resulting dataset.

4 Final dataset exploration

In this section we are going to explore briefly the new dataset, product of our aggregations, feature selections and merging of all three datasets.

4.1 Dimensions of the resulting dataset

After having merged the files we have as a result the following one:

	# Individuals	# Variables
Merged dataset	136.575	31

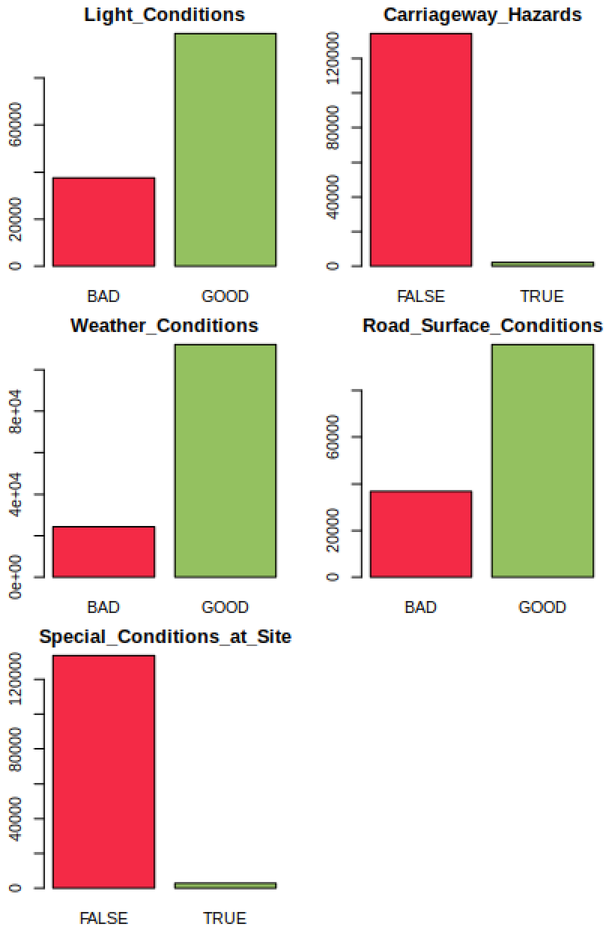
It can be seen that there are still a lot of dimensions despite having removed a lot of them which we deemed unnecessary.

Nevertheless, the great number of individuals which have been imputed means that the dataset is representative enough for the number of accidents in a year.

On top of that, we have reduced the number of possible labels a lot of variables/factor can have, thus reducing the complexity of the dataset despite the large number of variables.

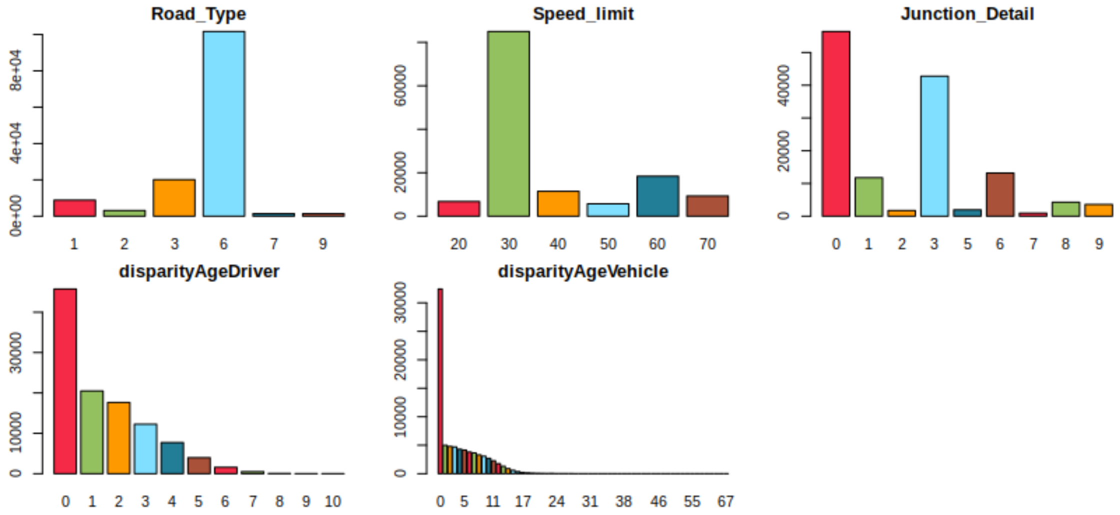
4.2 Brief study

Now that we have the dataset cleaned and merged, we have made a small exploration of the resulting dataset with plots to see how our data is distributed in different attributes that we are specially interested on. **Not all variables have been plotted in the report since plotting all of them would have taken a lot of pages, but they can easily be plotted with R.**



From these plots we can conclude a number of things:

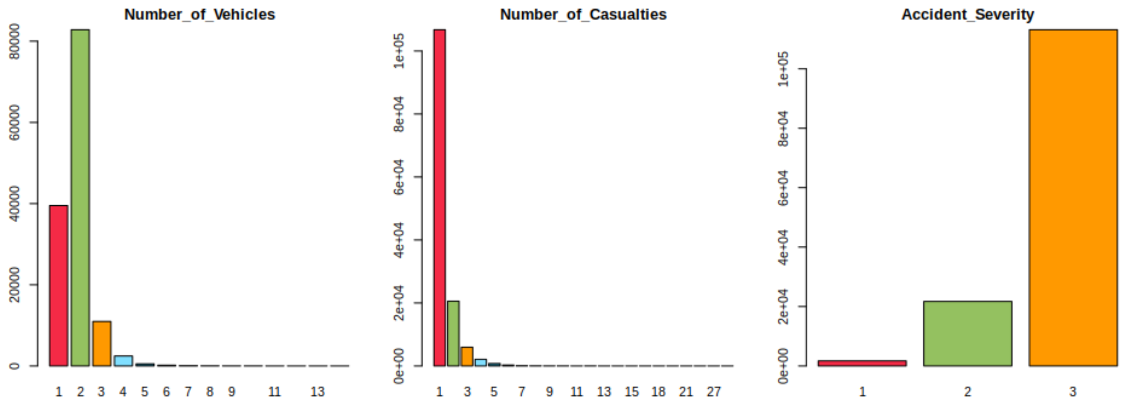
- Taking a look at the light conditions we can extract that most of the accidents happened during daylight.
- The majority of the accidents didn't involve any hazardous object.
- Being a low amount, the number of accidents with bad weather conditions surprise us the most, provided that UK is known for having a quite rainy weather.
- The total number of accidents with bad road surface conditions could explain a lack of maintenance.
- However, seeing the favorable amount of no special conditions at site (damaged traffic lights, damaged plates, etc.) could make us believe that the roads are less taken care of because of how cheap is to repair a traffic light compared to having to pave a part of the road.



Looking at the histograms we can also conclude a number of things:

- The road type with most accidents is by far the one labeled by a 6, which is the single carriageway.
- Most accidents have occurred where the speed limit is 30mph (which is around 50km/h). It has surprised us a little. Our bias without looking at the data is that most accidents would occur because of speeding, but 50km/h is not that much of a speed, so it seems clear that other variables influence in accidents, speed not being much influencing (assuming the accidents did not occur more than 20km/h above the limit, which is a naive assumption, but we do not have the data about the speed the vehicles had at time of the accident).
- Most accidents happened with not a junction nearby (Junction detail with label 0) followed closely by accident in staggered junctions, meaning in this case that probably most accidents involved more than one vehicle, probably by accessing the lane at the same time.
- The disparity of the age of the driver is very curious. Since the disparity of ages is computed using the ranges provided (meaning that a disparity of 1 equals to a disparity ranging from 10 to 19 years between the oldest and youngest drivers), we can see that there is usually not that much of a difference between the age of drivers in most cases. And accidents where there is a lot of difference between the drivers are not commonplace.
- Finally, the disparity between the age of vehicles is mostly 0. Since this age is computed using the exact age and not the range, this can be caused by two possibilities:
  1. Only one car was involved in the accident, meaning that there is no disparity with itself. This is mostly the case.
  2. The involved cars were from the same year. Highly improbable but not impossible.

  So in this case we assume that most accidents involved only one vehicle, and when this was not the case, the difference of age between vehicles ranged from 1 to 10.



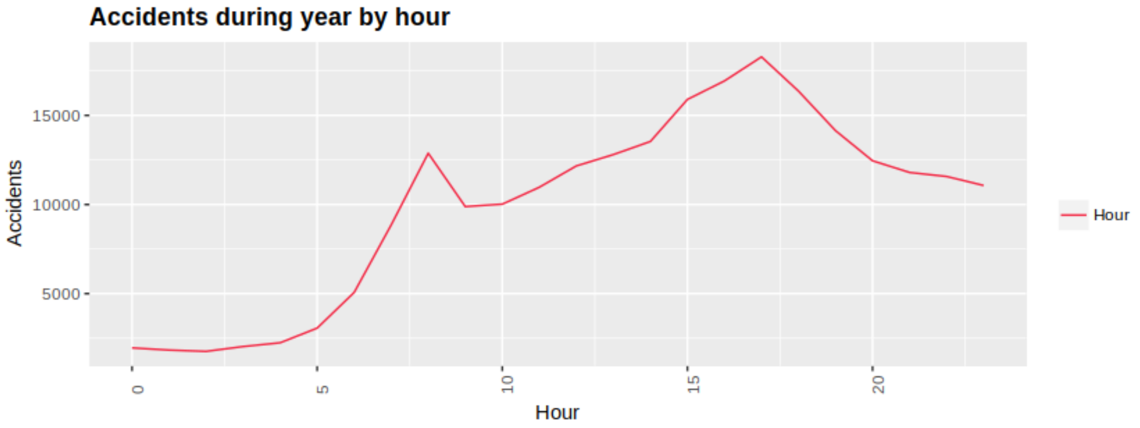
In order to detect several extreme cases, we have plotted three extra barplots for certain variables:

- The majority of accidents involve two vehicles or only one, at most 3. However, accidents with 4 or more vehicles are extremely uncommon.
- The majority of accidents result in a tremendous number of casualties. There are some extreme cases with even 27 casualties, which we imagine are from multi-collision accidents or with buses transporting many people. These, however, are pretty uncommon.
- Being the security one of first concerns while driving a car, a non negligible but definitely **not** the major part of the accidents still end up with fatal injuries. That is, deaths, which are marked with a label of 1. Most casualties only suffer minor injuries (label 3).

Although not being representative, the accidents that are clearly outlines are worth of study since they provide the extreme cases and all could point to a variable (or a combination of them) causing them.

4.2.1 Accidents by hour

We have made a small study on how many accidents there are by hour through the year by grouping the accidents by hour.



As it can be seen in the plot above, the number of accidents by hour spikes at the evening, around the 19 hours and also around 8 o'clock in the morning.

It did not surprise us since it makes sense that most accidents happen at rush hours where people are taking the car or bus to commute to work and throughout the day and that they decline greatly at midnight when people go to sleep.

5 Conclusions

The conclusions we have arrived at after this first study are many.

First of all and the most important of them is that **a good and thorough preprocessing of a dataset this large takes a lot of time**. It is the first time we have preprocessed a largish dataset by ourselves and it has taken more time than we expected at first. In fact, the increase of real time spent on the preprocessing is because the dataset is splitted in three different files, which leads us to the second most important point.

Working with data that is found in different sources of data is difficult, even if it is in the same format, which in this case is in three files stored in a *.csv* extension format.

Since each file had a lot of dimensions we had to do a lot of checking with the documentation provided about the data to know what exactly contained each variable. We had to discern between useful and not useful data, as well as deciding which data was interesting or which we had to sacrifice in order to not impute too much data and lack real data.

In fact, **the missing of data has been the third big problem we have encountered**. By removing the rows that did not contain all the information instead of imputing and *generating* data based on the other complete cases would have lead us to having around half of the original dataset, meaning the final dataset would not have been as significative.

The fact that there was no correspondence at all of 1 row to 1 row between two datasets also implied that we had to somehow **merge the files** with more steps than just copying and pasting the columns. Our solution of **aggregation** had lots of discussion between us and led us to sacrifice some information in favor of having more specific data aggregated, such as the number of drivers based on their gender.

All in all, our conclusions about the preprocessing can be summarized by: ***The greater the number of variables, the more complex the decisions to take about it.***

Additionally, after our brief study of the final dataset by plotting a few variables, we can see that there are outliers that should be studied more closely since they can provide a lot of information about extreme cases.

We can conclude **that most accidents happen throughout the day, specially during rush hour when people are commuting to work**, they only involve 1 or 2 cars, light and weather conditions are usually favorable and there are no objects in the road.

All in all, **there does not usually seem to be special conditions which induces us to think that most accident happen because of human errors**.

References

[1] *Road Safety Data*. URL: `https://data.gov.uk/dataset/road-accidents-safety-data`.