Algorithms for Data Mining

Data visualization

on Vehicle accidents dataset

Third delivery

C., Laura
G., Carles

June 30, 2018

Contents

1 Introduction

In our previous deliveries, we focused on preprocessing the data in order to make one large dataset performing a feature selection.

Mainly, we performed a feature elimination to reduce the dimensionality of the data to end up only the dimensions we are interested in. And then also added some new dimensions derived from the disjoint datasets by performing some aggregations, such as the resulting number of injured people from an accident.

Afterwards, in the next delivery we studied how *Hidden Naive Bayes* worked, made our own implementation in *Python* and the classification algorithm to our dataset to see how accurate it was on it.

We saw that, compared to Naive Bayes, the accuracy percentage of Hidden Naive Bayes heightened greatly, in some cases, sometimes increasing more than an 80%.

Now, in order to finish with our small study on this Vehicles' accident dataset, we now want to visualize the data. We will apply some techniques seen in this masters program and one seminar specialized in Data Visualization, which are complementary to the ones seen during this ADM course.

In other words, we will try to use data visualization techniques that we already know that will show us underlying data, as well as create new plots that we have not yet used and have been introduced recently to us to prove their usefulness.

Since we have been introduced to KNIME, we have used it to corroborate our results in our own implementation of Naive Bayes, for example, but we have not had the chance to put it in our deliveries so far. Nevertheless, we have seen that as the powerful tool it is, it also allows us to see the data with different visualization tools it has. So we take this last opportunity to include KNIME into one of our deliveries.

Additionally, we want to apply a couple of techniques called MCA and Clustering to not only plot the data *as is*, but to plot how the variables interact with each other.

So, this delivery will be two-fold and have two goals:

1. Visualize the data with multivariate plots using the dataset preprocessed in the first delivery.

2. Visualize the data by *processing* it to see how the variables interact and see if the accidents can be clustered.

2 Multivariate plots

In this section, we will plot our data using what we call multivariate plots.

This type of plots allow us to visualize as its name indicate two or more variables.

2.1 Correlation matrix

A correlation matrix allows us to see whether two variables are correlated and to which degree.

When two variables are positively correlated, it means that while one of them increases, the other does it as well. If they are negatively correlated, then when one increases the other decreases, so they are inversely correlated.

Although **correlation does not mean causality** they are useful to see possible relationships between those variables and it is a good first step on checking our data.

Looking at a matrix of NxN numbers, where N is the number of variables, may not always be the easy way to check correlations when N is large. However, humans are quick to interpret shapes and colors as well as colors intensity.

So instead at looking at plain numbers, we found a library to beautifully plot it using colors and ellipses.

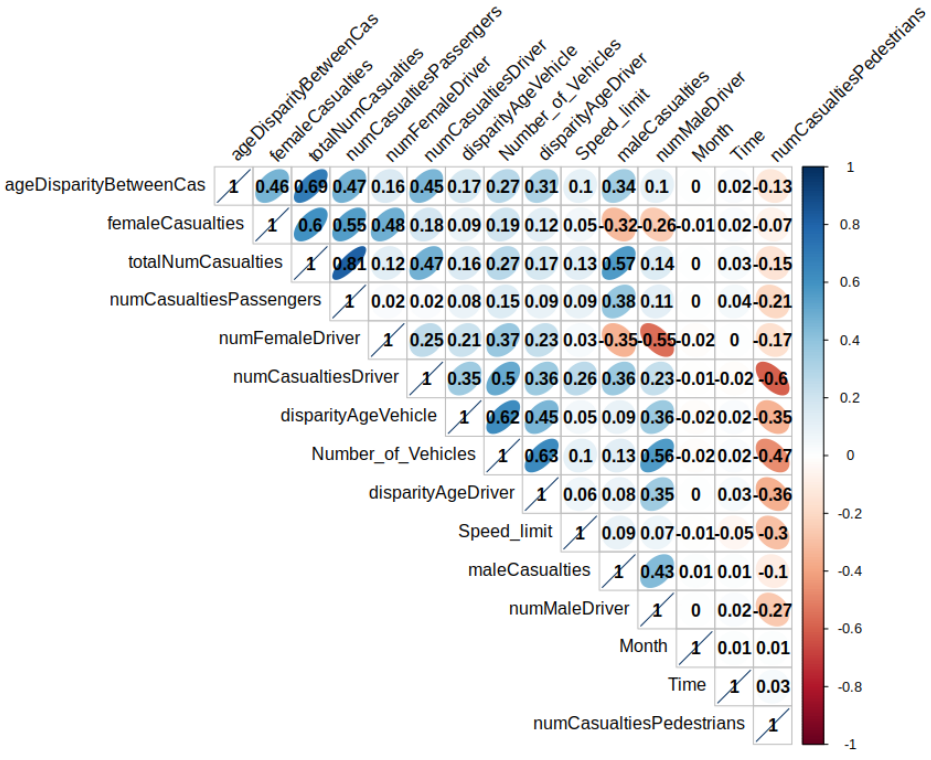


Figure 1: Correlation matrix

In it, we can see, for example, that:

1. **The number of female drivers is negatively correlated with the number of male drivers by -0.55.** We have checked the data and this correlation is due to the fact that, mainly, when there is an accident of a car and there is only one car, then the number of drivers is only one, which can only have one gender. So when *numFemaleDriver* is 1, *numMaleDriver* is usually 0. On all the other cases the number of drivers is sometimes mixed and sometimes not, but they do not happen so often.
2. **Number of casualties** correlates well with **age disparity**.
3. Passengers correlate by 0.81 to the total number of casualties, indicating that **many casualties are in fact the passengers of the cars**.
4. The number of casualties on pedestrians correlates negatively on the number of casualties of drivers, **meaning that mainly when drivers die, no pedestrians day and viceversa.**
5. Number of vehicles correlate with the disparity of age between vehicles, which of course is due to that **when there is more than one vehicle involved, the age disparity on the vehicles will usually be greater than 0**.
6. **Male casualties** with **number of male drivers**.

And many more, but the report would get too long. As we can see, the correlation matrix allows us a **first, easy and quite interesting glance at the data.**

2.2 Matrix plot

A matrix plot is a plot that contains multiple plots in a grid formation, such that the correlation and interaction between the two variables defined, one in the row and the other in the column, is explained visually.

In our case, since we had so many variables we had to choose only a subset of them: those that were numerical and seemed important for us, so that we could be able to plot boxplots, scatter plots and histograms in one big plot.

The result can be seen in the next page. Since the plot is so big, we recommend taking a look at it in the original image extracted from it attached to this report.

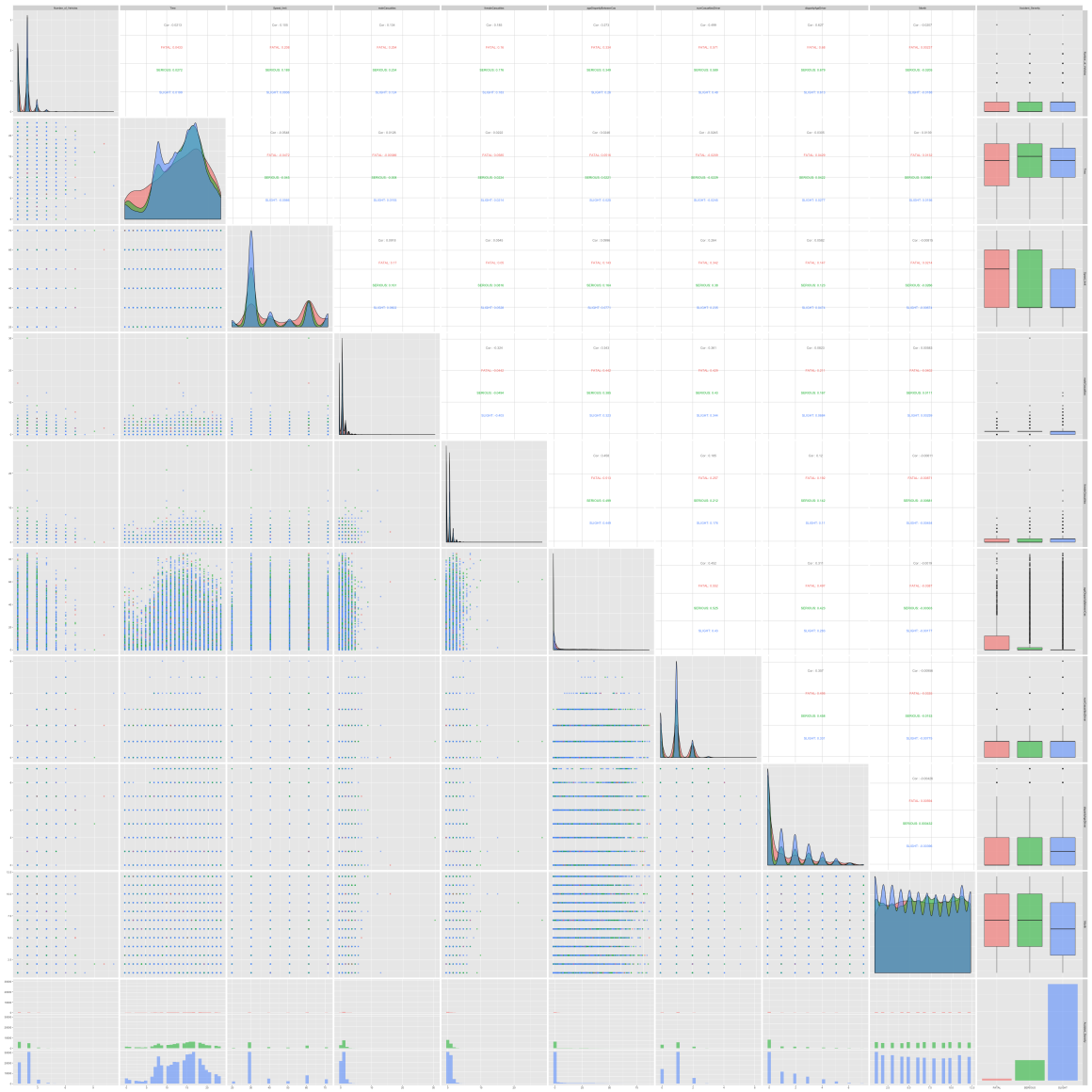
The matrix is split in five parts: the upper section, the lower section, and the diagonal, which separates the former two. Additionally, on the right of the plot we can see a column of boxplots for each variable and, in the bottom, histograms.

And even more, an underlying feature is depicted in the plot not specifically in none of the cells, but in all of them altogether by its color: the gravity of the accident.

The upper section depicts the correlation between the two variables (the one selected on the row and the one on the column), as well as the correlation on the different categorical values selected. In our case, we selected the colour to be determined by the accident severity.

The lower section, on the same way as the upper section, depicts a scatter plot between the two variables.

The diagonal contains density plots for the variable selected (since the row is equal to the column, so the variable is the same). The density plots show the distribution of the accidents' severity for that variable.



2.3 Heatmap

Now, in order to visually represent the accident associated to a certain coordinate, we have decided to use a heat map as the preferred visualization method. The visualization of the accidents using a heat map will provide insights on where the accidents happen timelessly, so we can locate black areas where accidents are more prone to happen.

The heatmap is computed using the following procedure:

1. Extracting the 2D Binned Kernel Density Estimation of the coordinates.
2. Drawing the contourLines of the previous density.
3. For each contour line, transform it into a polygon in order to be displayed into the map.

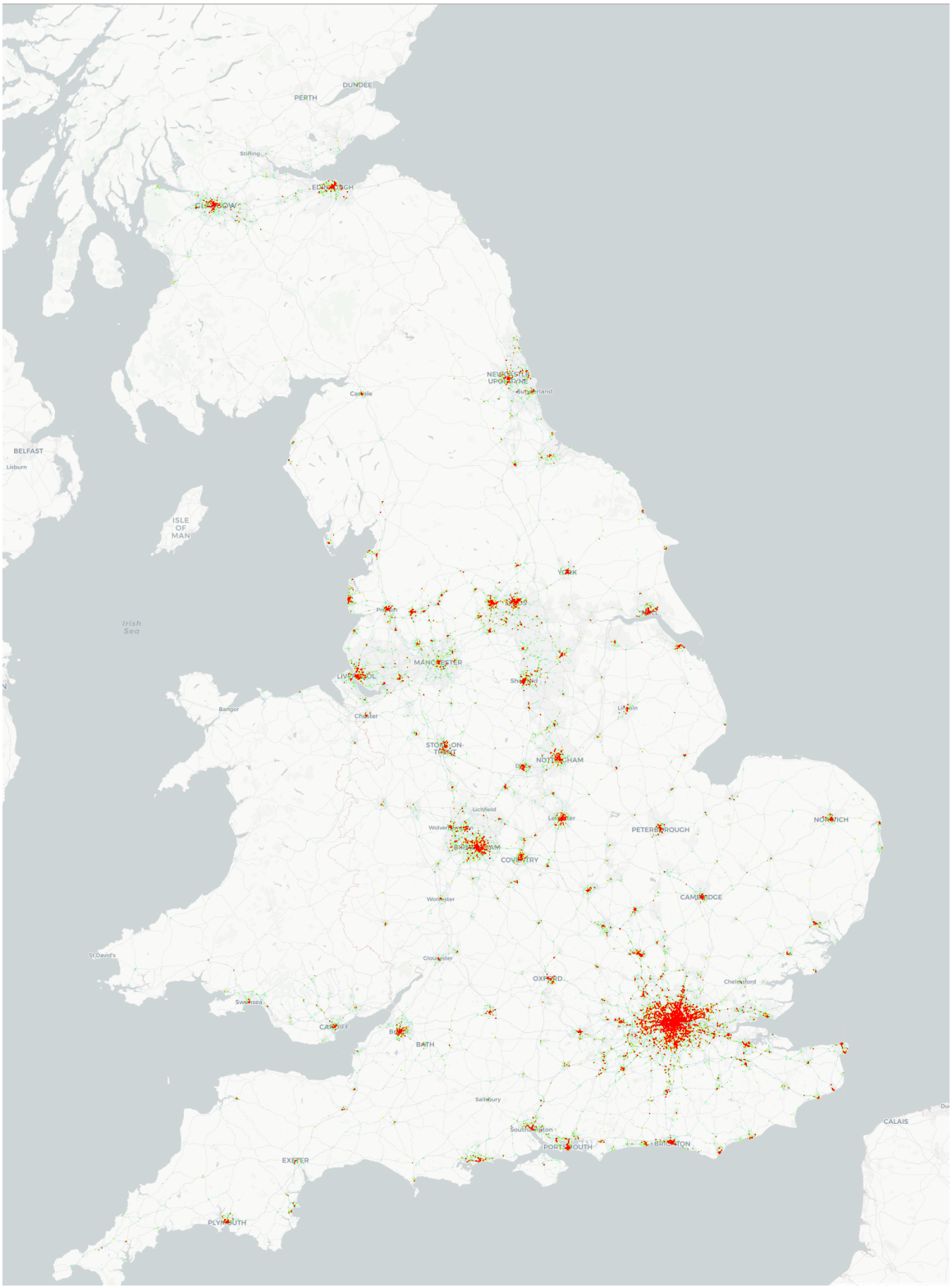


Figure 2: Accidents Heatmap

Observing the heatmap, properly created in order to reflect the major concentration of accidents allows us to say that (as expected) the majority of the accidents happen in the

important cities or nearby them. All the incoming crossroads, as well as city roads of London get an important share of accidents. However, the highest density of accidents is given in the city centers.

2.4 Parallel sets on KNIME

2.4.1 Why on KNIME

After having seen different visualizations in R, using new libraries and kinds of plot we did not know about prior to this delivery, now we want to see at least one of them on KNIME.

Although we have used KNIME on the previous deliveries to, for example, validate the results from Naive Bayes with those obtained with our implementation, we did not include this usage on our reports for it was not relevant on it. Rather it was only for our own verification.

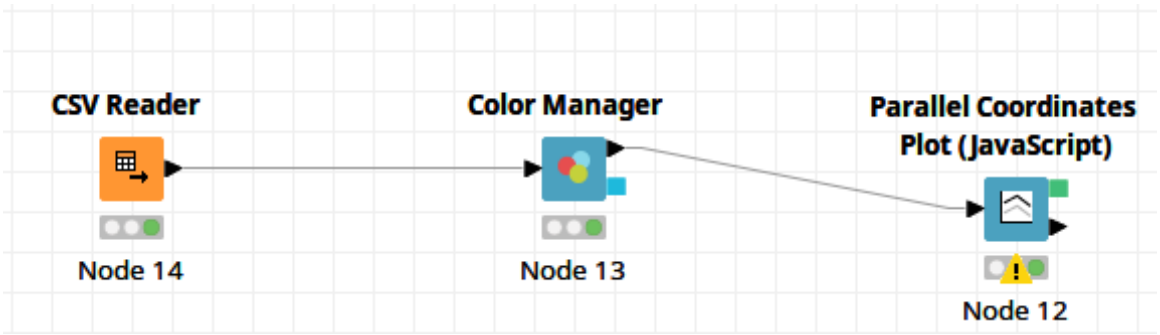
Nevertheless, we have found KNIME to be a useful tool with lots of potential and as such, we wanted to include its usage by taking a look at, at least, one visualization tool that it provides.

2.4.2 Parallel sets

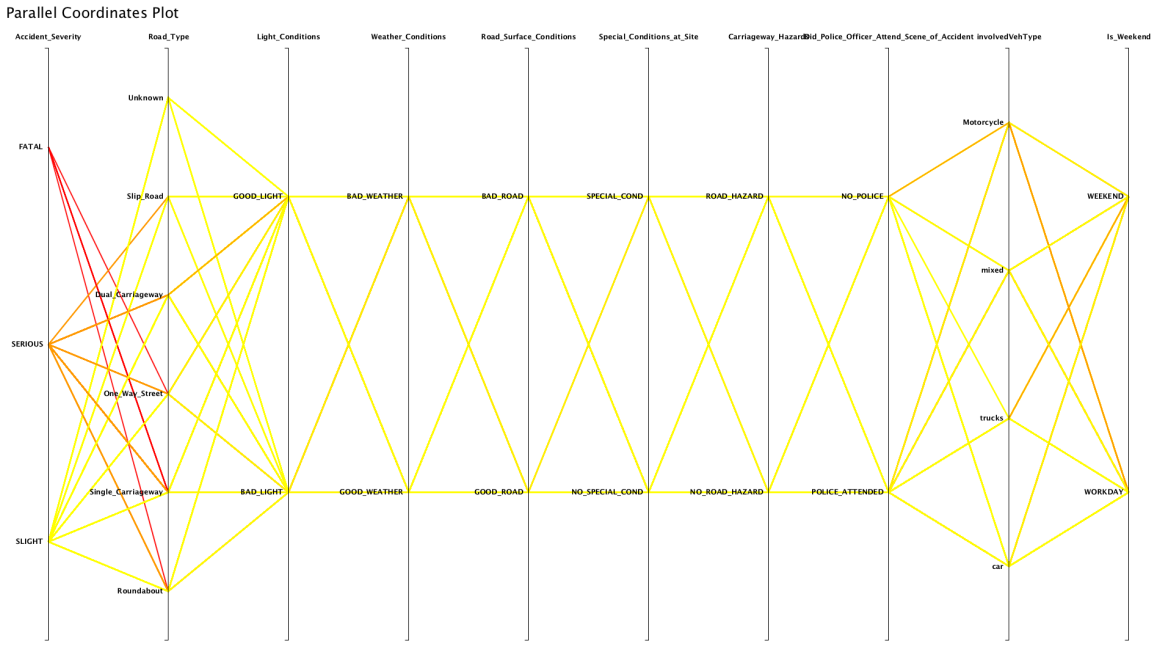
A parallel sets plot, also known as Parallel coordinates, is a type of plot very useful for multivariate, categorical data and see how it is distributed.

Since our data is mainly categorical, we have found the idea to be very appealing and interesting for our problem at hand and so we used it to study how our categories are related.

The schema of our Knime Workflow can be found attached to our report and is as follows:



The result plotted is the one below.



Each vertical line represents one variable, and the points that are in each variable are the possible values.

For example, the first vertical line starting from the left is the Accident severity, and it can have three possible values: fatal, serious and slight.

The colored lines between each variable are the relationships between vars.

However, we have **greatly missed** that the line between each variable has not **thicken or slimmed** in **proportion of how well related a variable value is related to another one**. We deem this feature as essential, and thus KNIME not having it has really shocked us.

On the other hand, we have used a Color Manager to at least colorize the lines depending on which type of accidents move between two variables. The fatal accidents are plotted red, the serious ones are orange and the slight ones in yellow.

As we can see, yellow predominates on our plot, with some orange between Dual Carriage way and Good Light (meaning that most accidents with those two conditions were of the serious type, but not fatal).

Taking a closer look at the involved vehicles' type, we can see that accidents where only motorcycles were involved and no police attended on site were much related to those of serious gravity. We have found this relationship strange, for if an accident is serious then police should be expected, specially if the involved vehicles are motorcycles.

So we looked at the dataset and found the following:

# of motorcycles accident with no police attending on site

Accident severity	**# motorcycles**
Fatal	6
Serious	123
Slight	222

We can see that mainly these accidents were of slight gravity or serious gravity, so it makes sense for us that the color between those two variables is a yellow-orangish.

However, this does not explain why Police did not attend the accident on site and, with the data on site, there is not a very logical explanation for us other than police attended the victims on the hospital after they were taken care of.

On the other hand, we found that most motorcycle accidents that happened on a work day were also of Serious gravity.

The same applies on trucks on weekends: the accident was of serious gravity.

As we can see after having taken a closer look to this type of plot, the **main disadvantage** of using it is that **you can only see how each variable interacts with the adjacent ones**. So, for example, we can only see how weather conditions interacts with Light conditions and road surface conditions.

3 Beyond simple plotting

3.1 Multiple Correspondence Analysis

A method to see relationships between categorical data, which our dataset mainly is, is Multiple Correspondence Analysis.

This method, also known as **MCA**, is a technique in Data Analysis applied to data which is nominal categorical. That is, not numerical. It is used to detect and represent structures in a data set by showing each individual in the data set as a point in a low-dimensional Euclidean Space.

It is considered an extension of Correspondence Analysis, but applied (as its name indicates) to variables' categories that are large.

Its homologous on continuous data would be Principal Component Analysis.

This method treats rows and columns equivalently, so an individual cannot have a field empty. It creates a factor score for each individual from weighting the rows and columns. There are three phases to apply in MCA.

1. **Preprocess the table** creating a contingency one of the same size: $C = m * n$ where the weights for each cell is computed. The row weight is computed such that $w_m = \frac{1}{n_C} CV$ and the column weight such that: $w_m = \frac{1}{n_C} V_1^T C.$, where each cell $n_C = \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}$.
   In other words, w_m and w_n compute the marginal probabilities of the rows and columns classes.

   V_1 is a column vector of 1 with the dimension of the column.

   Then it creates table S where each cell of the table C is divided by n_C, which is the total sum of C table: $S = \frac{1}{n_C} C$

This computes the joint probability distribution of the rows and columns.

And finally, it creates table M applying to the table S the weights of each row and column computed previously: $M = S - w_m w_n$, which means that it computes the deviations from independence.

2. **Create orthogonal components** from the computed table M, where M is decomposed applying the generalized singular value decomposition, resulting in $M = U \sum V^*$, where $U_m^W U = V^* W_n V = I$
3. Compute the **factor scores** for both the row and columns items: $F_m = W_m U \sum$, $F_n = W_n V \sum$

Having applied MCA to our dataset, we obtained the results below.

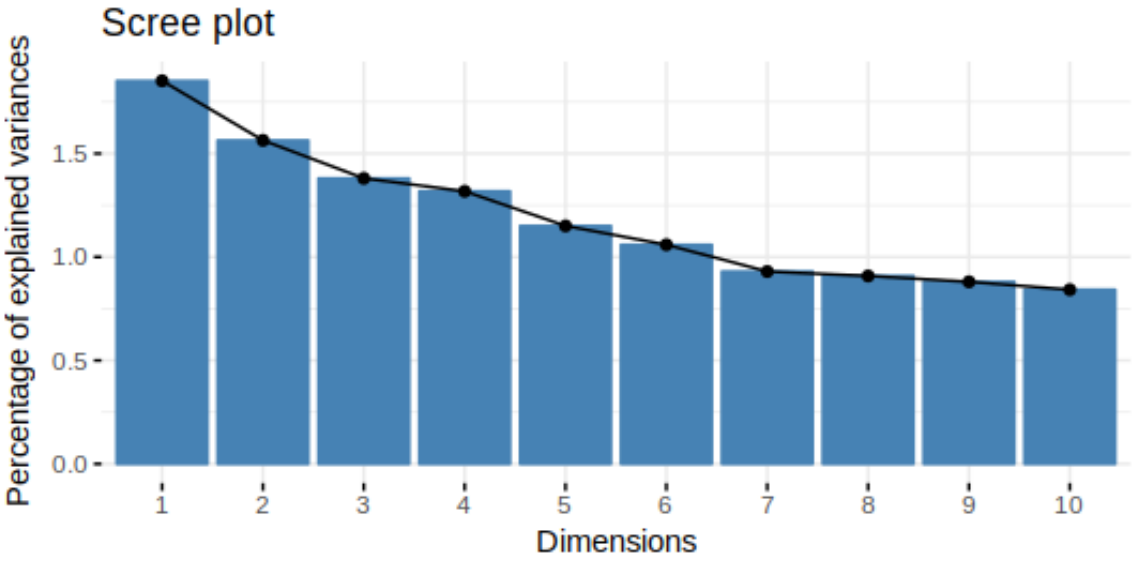
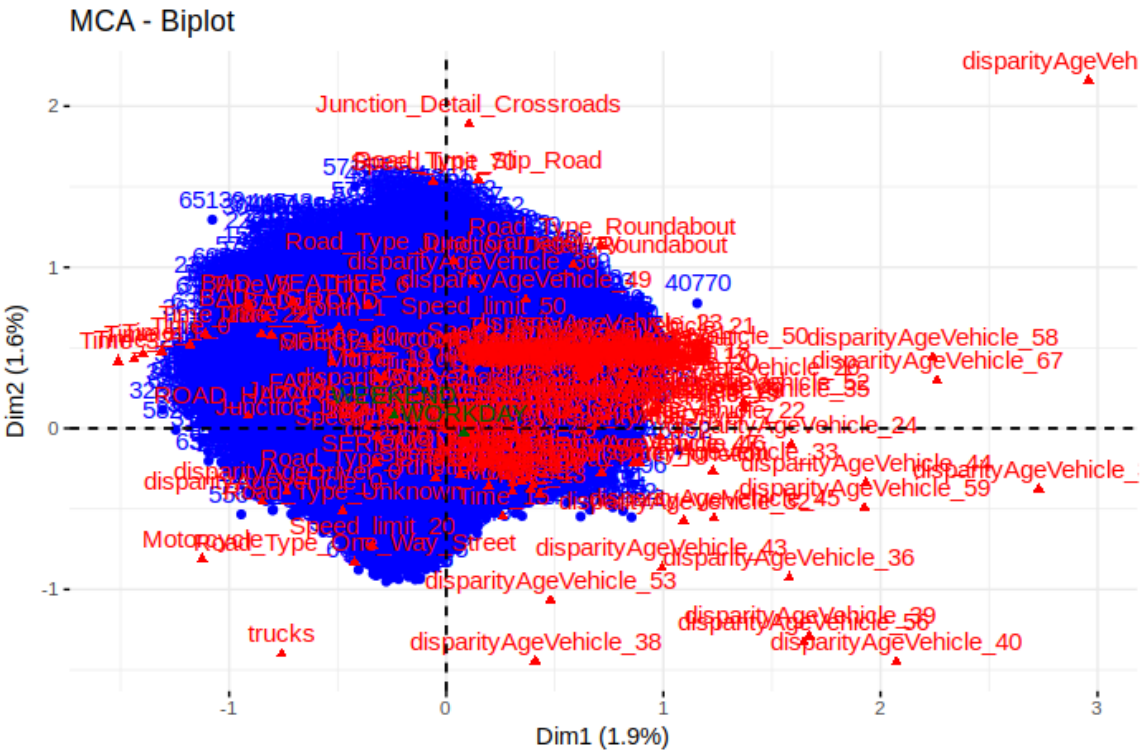


Figure 3: MCA Screeplot

First of all, we want to look at the information retained on each dimension. That is, how much information can be explained by looking at the results produced by MCA.

By looking at the screeplot, we can see that not much information is explained in the first 3 dimensions. The sum of them is less than the 5% of the total information (and the sum of of the first two dimensions, which conform the first factorial plane, is 3.41%), so we cannot take these plots as very explanatory, if at all.

However, we proceed to analyze the result to see how MCA applies.



Looking at the biplot, we can see that all the individuals gather in a round cloud. Thus, meaning that there really is no structure or pattern followed, so we cannot relate the individuals.

However, if we take a look at the categories represented, some of them seem somewhat correlated. Again, taking into account that less of 5% of the information is retained, we have to take these observations with a grain of salt.

1. The first dimension seems to explain and differentiate between different disparities of age of the vehicles.
2. The second dimension, on the other hand, seems to differentiate between the type of vehicles involved in an accident.
3. Vehicles whose disparity of age is large seem to relate well.

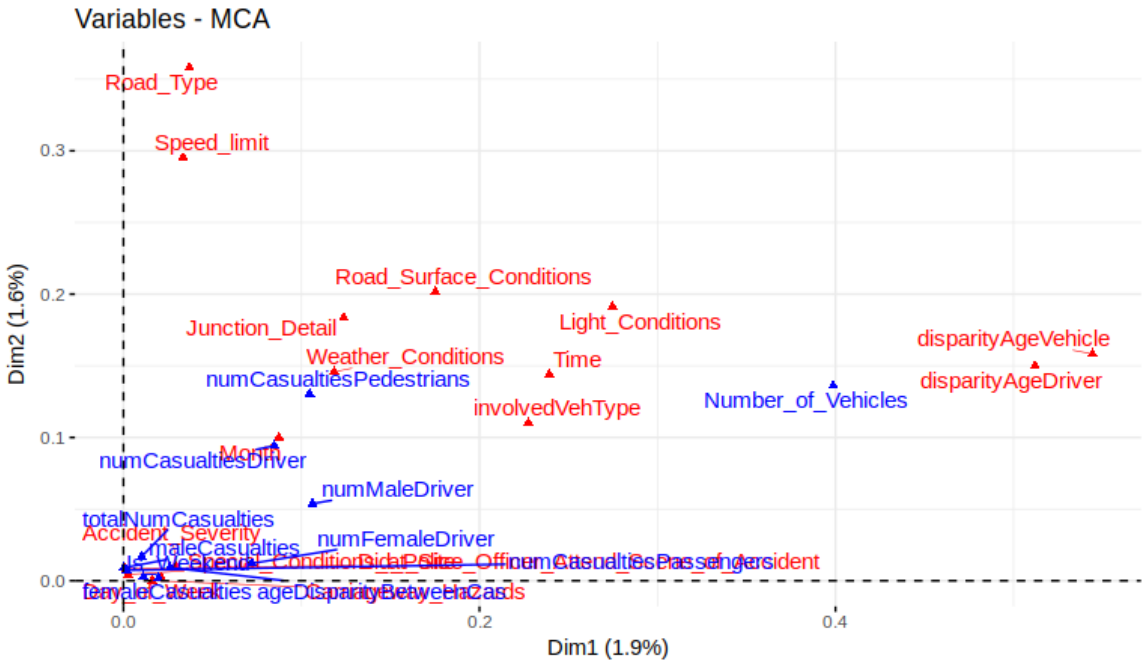
Other than that, no clearly conclusions can be extracted.

As for the variables represented in the plot below, we can see that the disparity of ages between vehicles somewhat relate to the disparity of ages of the drivers.

The road type also relates well to the speed limit on the road, which makes sense, for different speed limits apply on different types of road.

The time at which occurred the accident seems somewhat related to the Light conditions, which may relate because of the position of the sun.

All the other variables, although somewhat related if looking at the plot, do not have really logical interactions on the real world, so we think that since only 3% of the information is explained on this factorial plane, we cannot deem these proximities as important or representative.



In conclusion, although MCA has not given us results as good as we wanted, it still served us on visualizing possible relationships between variables.

3.2 Clustering

After having computed the MCA, we wanted to know if we could differentiate different types of accident using different clustering techniques.

3.2.1 Dimensionality reduction

To perform the clustering and given that we have a big number of dimensions, we have taken advantage of the results from MCA to perform a dimensionality reduction.

This has been done in three main steps:

1. First, by selecting only the dimensions in the dataset that their **eigenvalues are higher or equal than the average eigenvalues**.
2. Later on, the **average value of the eigenvalues** is **subtracted** to the previous eigenvalues that fulfilled the condition.
3. Finally, **only** the **dimensions** that their **eigenvalues provide up to a cumulative inertia up to 90%** are taken.

3.2.2 Approach

The first idea of how the clustering was going to be applied was to perform a hierarchical clustering in order to explore the feasible number of clusters that the data could be partitioned.

After that, and using the centroids calculated with the MCA and the clusters given by the dendogram resulted from it, a consolidation would be achieved by applying K-means.

However, **the large amount of instances as well as dimensions has made impossible to compute a dendogram** for our dataset. This has brought us the idea of exploring **sub optimal** clustering using only **iterative K-means** execution. Despite having time series data, we didn't want to partition the data into several set according to the time the accident happened in order to be able to generalize to better global clusters.

Having tried several clustering cuttings by varying the numbers of clusters, we have found that the **best number of clusters was three and two**. Being K-means an algorithm that finds a sub optimal solution, we have employed multiple iterations trying to find a suitable result by examining the resulting clusters in the first factorial plane of the MCA.

Once the clusters have been computed, we need to explain the characteristics that define each one of them.

3.2.3 Three main groups

The **first cluster**, with around 21k individuals and marked as **red**, has the following characteristics:

- Has accidents involving trucks or involving motorcycles.
- There was no difference in age of the drivers or in the age of the vehicle.
- The conditions at the moment of the accident were good.
- The accidents are considered serious and happened in low speed zones (20 - 30 mph).
- The roads were the accident happened had only one lane for each direction.

The **second cluster**, with 17k individuals and marked as **green**, is characterized for:

- Having occurred in a cold month (November, December or January).
- The conditions at the moment of the accident were bad.
- The accidents happened in a high speed roads (60 - 70 mph).

The **third** and last cluster, being the biggest with 30k individuals and marked as **blue**, is characterized by:

- Having mixed type of vehicles involved.
- The conditions at the moment of the accident were good.
- The severity of the accident was slight.
- There was a junction with 3 or more arms where the accident happened.
- There was some disparity of age between vehicles.

Knowing the characteristics of each clusters, we can display the clusters into their actual position (determined by the longitude and latitude of each individual) colored by the specified color markers.

We can observe in the map that both blue and red marked clusters are more prone to take place in a city, in a town or in a secondary roads where the speed limits are lower than a crossroad. Those clusters represent the majority of the accidents, compared to accidents that happen in high speed roads (denoted as green).

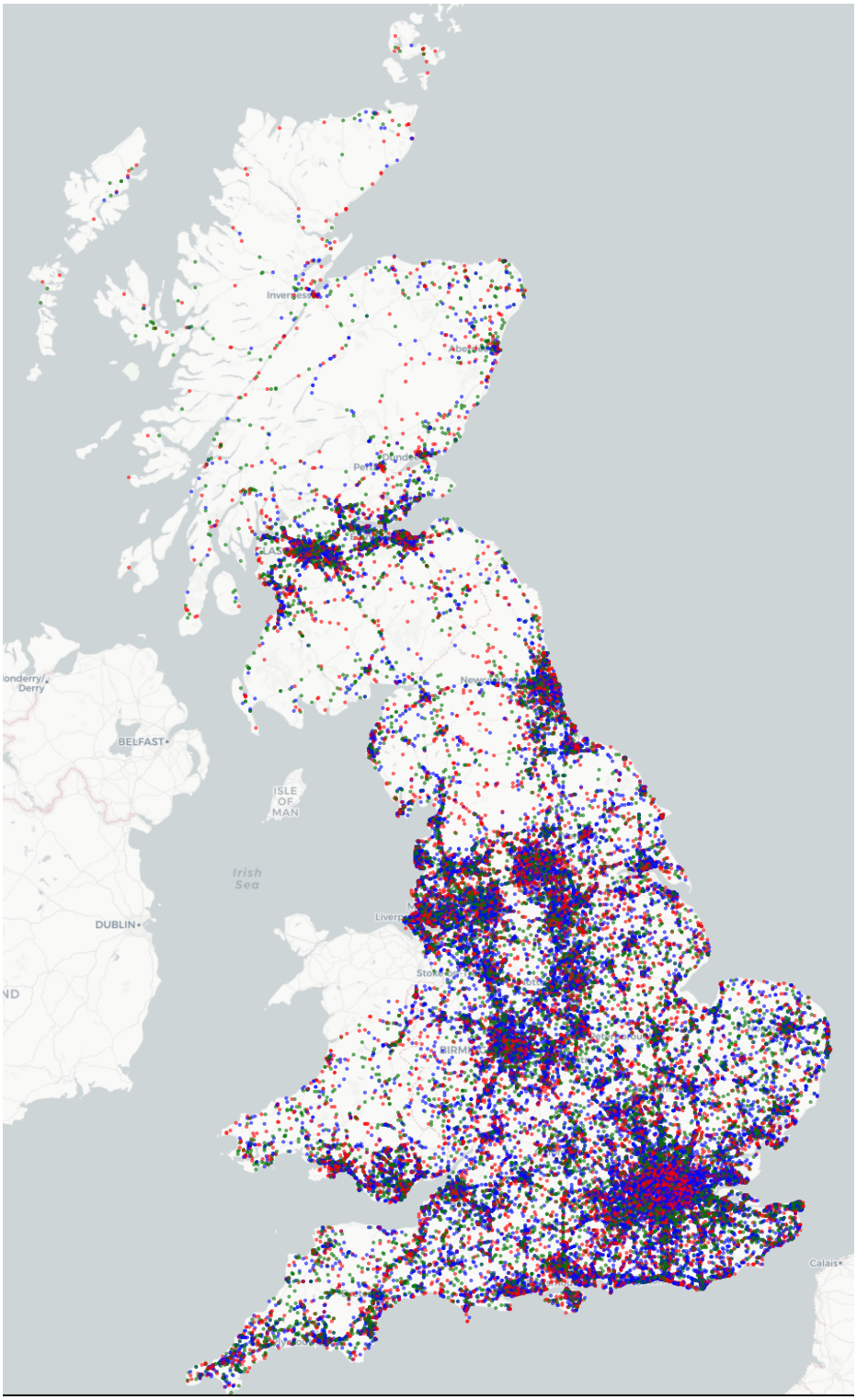


Figure 4: Kmeans with 3 clusters

3.2.4 Two main groups

As explained before, two main clusters instead of three, were also considered.

The **first cluster**, with nearly 40k individuals and marked as **red** are the ones with the following characteristics:

- Involved different types of vehicle.
- No special conditions were given.
- The age of the vehicles was different.
- The accident happened in some type of junction.

The **second cluster**, with around 30k individuals and marked as **green**, is characterized for:

- Having accidents involving trucks or involving motorcycles.
- There was no difference in age of the drivers or in the age of the vehicle.
- The conditions at the moment of the accident were good.
- The accidents are considered serious and happened in low speed zones (20 - 30 mph).
- The roads were the accident happened had only one lane for each direction.
- The time when the accident happened was around midnight.

Knowing the characteristics of each clusters, we can display the clusters into their actual position (determined by the longitude and latitude of each individual) colored by the specified color markers.

Looking at the map below with the markers, we cannot affirm that the location of those accidents is what may have affected the accident (or how it was clustered).

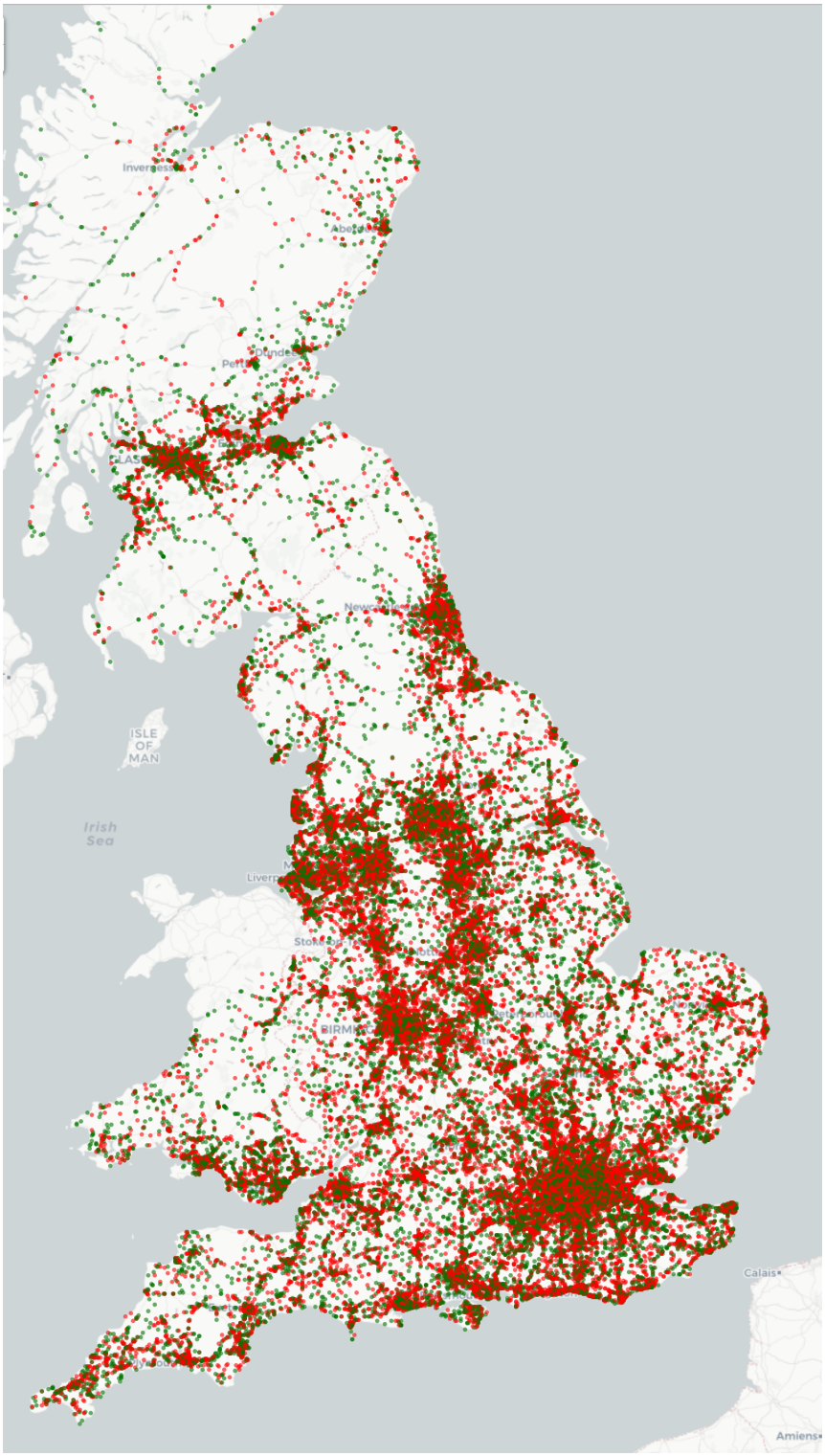


Figure 5: Kmeans with 2 clusters

4 Conclusions

The conclusions derived from this delivery are many.

1. A correlation matrix can be a good first step on seeing correlations/interaction between two variables at first glance and very interesting libraries exist to automatize the plotting, producing beautiful and easily understandable results.

2. In the same way, a matrix plot is also very useful to plot many variables, being each cell a representation of how two variables interact. Additionally, if one uses color you can represent in each cell **three variables**, which we have taken into our advantage to plot the severity of each accident within the interaction of two variables.

3. Computing the heatmap and displaying it into a map using the coordinates has helped us both to identify where the accidents are prone to happen, as well as the the magnitude for each zone or city.

4. Parallel sets have been interesting to use to look at our data, although we have been a little bit **disappointed** at how low versatile the **KNIME** library was, for we could not thicken the lines between variables depending on the weight of individuals on each relationship. Nevertheless, it has helped us to see something very interesting: not in all somewhat serious accidents between motorcycles did the police attend, which we find strange and have not a good explanation for this happening.

5. Although we know that MCA is a very good technique to see how the variables are related to others, it has **not proved to be useful for this particular case**, for by taking the first 3 dimensions only less than the 5% of the information is explained/retained. Which means that the **results are not meaningful or representative**.

6. Using clustering methods we expected to find different groups the accidents are classified in, in order to get to know the characteristics of each group. However, the large amount of data and dimensionality, has forbidden us to be able take a global picture of how the data could be partitioned into using hierarchical clustering. On the other hand, **Iterative k-means** has brought us the opportunity to be able to extract different clusters from the data. The displaying of the clusters in an actual map, using the clusters marker as the color the accidents are represented has shown us how the different types of accidents are represented in the map which are the properties of each type. However as expected, we have found that the types of accident in the map are quite tied to the characteristics of the road. Also, several time and vehicle related patterns has been discovered in the clusters.

References

[1] *Example for JS Parallel Coordinates*. URL: https://www.knime.com/nodeguide/visualization/javascript/example-for-js-parallel-coordinates.

[2] Liangxiao Jiang, Harry Zhang, and Zhihua Cai. "A novel bayes model: Hidden naive bayes". In: *IEEE Transactions on knowledge and data engineering* 21.10 (2009), pp. 1361–1371.

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