Multivariate Analysis
Exercise 1

Denaldo Lapi, Samy Chouiti, Francesco Aristei

July 16, 2022

1 Introduction

The goal of this report is to describe our solutions and intepretations of the three problems of Excercise 1 of the Multivariate Analysis course. We solved three problems each one related to a Principal Component Method by using the R software.

The problems are related to the following methods:

- Principal Component Analysis (PCA)
- Correspondence Analysis (MCA)
- Multiple Correspondence Analysis (MCA)

1.1 R scripts and packages

We built 3 R notebooks (*.Rmd* files), one per each problem, in which we describe in details all the steps of our analysis. For each problem we follow the same notebook struture:

- Data loading and exploration: we load the data of interest and we investigate it
- Data preprocessing: apply some preprocessing steps to the loaded data in order to make it compatible for the problem to solve and in order to improve the quality of the dataset
- Principal component method execution: we apply the algorithm of interest among the three mentioned before
- Results interpretation and plots: describe the results obtained after applying the algorithm

In particular, we mainly used 2 R packages to solve the excercise:

- built-in *R stats* package: used to apply PCA with the function *princomp()*
- *FactoMinerR*: used to apply both MCA and CA

We also used other R packages in order to obtain better visualizations, such as *factoextra* which allows to do plots about PCA, CA and MCA.

1.2 Some notes about the report

We would like to remark that the 3 provided notebooks contain a detailed analysis of each one of the problems, with all the steps that we followed to execute and interpret the FA methods. In this report we'll decribe only the most meaningful steps and results we obtained.
We also added an appendix in the end of the report containing the most important plots, so that this document could be able to provide a self-contained analysis of the 3 problems. However, we strongly suggest ro read the R notebooks for a more detailed and clear explanation of all the solutions.

2 Principal Component Analysis (PCA)

The first problem required us to apply PCA to the cars2004.csv file. This dataset is composed of 425 cars from the 2004 model year and 19 features. The first feature is the name of the car (variable Name). Apart from that, seven features are binary indicators used to determine the type of model of the car (Sports, Caravan etc) or to determine if it is a rear-wheel or an all-wheel car. The other 11 features are numerical, and describes several characteristics of the car (length, width, gas mileage etc).

2.1 Preprocessing

As first step we started doing some preprocessing on the dataset. First we checked the presence of NA values. We found out several missing values, both in the numerical and in the categorical variables, so we proceeded in removing them from the dataset, specifically, we substituted the NA values in the categorical variables with the most used value in that specific column (either 0 or 1) applying a sort of "majority voting". While for the numerical variable, we simply substituted the NA values with the mean of the column. After having removed all the missing values, we focused on the outliers. From the scatter plot (Figure 1) of the dataset, it is evident the presence of outliers. To delete such points we used a utility function, which for each numerical column it detects the points outside the interquartile range, defining them as NA points, using the boxplot.stats informations. After having applied such function, we plotted the box-plots of the variables, to show that the outliers were removed correctly, (Figure 2), (Figure 3), (Figure 4).

2.2 PCA Implementation

Once the pre-processing phase is completed, we started applying the PCA algorithm. The first thing we observed is that the columns work on very different scales, for example the variables 'Weigth', 'Retail' and 'Dealer' have values much greater than the other columns, therefore, it is necessary to extract the PCs from the correlation matrix R. So that we don't need to center the data set. When the PCA has been applied, via the princomp() function, we had to rescale the loadings, so that the coefficients of the most important components are larger than those of less important one. In order to rescale the loadings we used the standard deviations of the principal components.

2.2.1 Explained Variance

One important aspect to check, is how much variance each PCs explains. So we checked it, we observed that the first component explains alone a considerable proportion of the whole variability of the dataset (around 60%) and together with the second component the amount of

explained variability rises to 76%. This means that probably, we will be able to summarize the whole dataset using only the two obtained components.

2.2.2 Chosing the Number of Components

Now we should choose the number of components with which we want to explain our data. To do so it is useful to understand which are the components that retain the majority of the variability. A good graphical indicator is the screeplot, which plots the variance explaind by each component (Figure 5). To do such plot we load the factoextra library.

As stated before, the first and the second component alone explain almost 80% of the variance, therefore, they are the main candidate to summarize the information of the other variables. When looking at the screeplot, it is important to look for elbows, which means a point after which the eigenvalues start decreasing more slowly. As we can see from the plot, the elbow appears passing from the second to the third dimension.

2.2.3 Loading Plot

Then, we studied the loading for the PCs. What we observed is that the first principal component has large positive association with almost every variables except for CityMPG and HighwayMPG, with which it has a large negative association. CityMPG and HighwayMPG represent the gas mileage respectively in the city and in the highway. So this component could measures cars with powerful engine and big sizes, like SUV, Minivan etc. Moreover this component tries to look at low consumption models given that it has negative association with the gas mileage attributes.

The second component has a large positive association with Retail and Dealer, which means that it describes the most expensive cars. It has also positive association with cylinders and horsepower, this, with the above results, may suggests that this component is associated with Sports car.

The third component has positive association with HighwayMPG and CityMPG so it measures models with high consumption of gas. It has also positive association with Wheelbase and Length and a slightly negative association with Cylinders and Engine, so it may describe models with low performances, having big size, which implies an high consumption of gas.

2.2.4 Correlation Circle

To better understand the association between PCs and variables, we plotted the correlation circle (Figure 6)

It can be interpreted as follow:

- Positively correlated variables are grouped together.
- Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants).

The distance between variables and the origin measures the quality of the variables. Variables that are away from the origin are well represented on the plot. The plot confirms the highly correlation observed before in the scatter plot.

2.2.5 Quality of Representation

Another useful indicator to understand how much each PCs represent the variables of the dataset is the cos2 (quality of representation) Specifically:

- High cos2 indicates a good representation of the variable on the principal component. In this case the variable is positioned close to the circumference of the correlation circle.
- Low cos2 indicates that the variable is not perfectly represented by the PCs. In this case the variable is close to the center of the circle.

For a given variable, the sum of the cos2 on all the principal components is equal to one.

If a variable is perfectly represented by only two principal components (Dim.1 Dim.2), the sum of the cos2 on these two PCs is equal to one. In this case the variables will be positioned on the circle of correlations.

For some of the variables, more than 2 components might be required to perfectly represent the data. In this case the variables are positioned inside the circle of correlations.

Regarding our data, we can observe (Figure 7) from the correlation circle and the cos2 that almost every variable is perfectly represented with the first two components apart from CityMPG and HighwayMPG which seem to need also the third component.

2.2.6 Contributions of Variables to PCs

Next we looked for the contributions that each variables has in explaining the variability of the dataset. Such contributions are expressed in percentage.

- Variables that are correlated with PC1 (i.e., Dim.1) and PC2 (i.e., Dim.2) are the most important in explaining the variability in the data set.
- Variables that do not correlated with any PC or correlated with the last dimensions are variables with low contribution and might be removed to simplify the overall analysis.

As we can see (Figure 8) all the variables contributes more or less in the same way to the first dimension, while for the other dimensions there are unbalanced contributions in determining the PCs. So it seems like there aren't variables that can be discarded to simplify the overall analysis.

Next we used the function fviz_contrib() to draw a bar plot of variable contributions:

- (Figure 9)
- (Figure 10)

The red dashed line on the graph above indicates the expected average contribution. For a given component, a variable with a contribution larger than this cutoff could be considered as important in contributing to the component. For the first dimension, almost every variable has a contribution around the red dashed line, while for the second dimension the contribution given by Length is not comparable with the one of Cylinders or CityMPG.

2.2.7 PCs Scores

The PC scores are the values of each PCs (linear combination of the original variables) evaluated in the data set, (Figure 11) We checked that they are uncorrelated. Now, given the scores of each entry of the dataset, we visualized the Score Plot, which shows clusters of cars based on their similarity.

Due to the large amount of data points at disposal, it would be impossible to label each point in the graph. Therefore, we used a supplementary categorical variables, represented by the model type of the car, to color each point. Before doing so, it was necessary to collapse all the binary variables regarding the model type, in one labelling variable, which we called 'Model', having the following values: Sports, Minivan, Suv, Wagon and Others.

After this we analyzed the Score Plot, (Figure 12) As we can observe the data are quite spread along each directions. The most significant cluster that can be observed is the one representing the cars of the Minivan type in the bottom-right corner. Moreover we can observe that the majority of the SUV type cars lay in the first and fourth quadrant of the graph.

2.2.8 Biplot

Lastly, we drawn the biplot, which is the combination of the PCA score plot, together with the loading plot, (Figure 13)

What we observed is that the Minivan type of cars are depicted together with the Length, Wheelbase and Width variables, which make sense considering that this type of car is the one having the greatest values regarding these attributes. Also the SUV type of cars are partially together in a cluster, and are near the Engine, Cylinders and Horsepower variables, which is coherent given the kind of car. Even if not perfectly represented, the Sports car are more spreaded in the first quadrant, where Dealer and Retail variables are pointing, which is coherent with the fact that these kind of cars are usually the most expensive one.

To have a better understanding we plotted another version of the biplot in which we tried to color both the individuals and the groups. Specifically the groups are colored based on their contributions, (Figure 14)

As we observed also before, the less contributing variables are HighwayMPG and CityMPG.

3 Correspondence Analysis

The second problem required as to apply Correspondence Analysis to the "Mortality" dataset, already available in the package *FactoMineR*.

3.1 Dataset

As a first phase we just explored the dataset to get some precise insights about its main statistics. The "mortality" dataset is a data frame with 62 rows (the different causes of death) and 18 columns. Each column corresponds to an age interval (15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85-94, 95 and more) in a year. The first 9 columns correspond to data in 1979 and the 9 last columns to data in 2006. In each cell, the count of deaths for a cause of death in an age interval (in a year) is given.

The first 9 cols correspond to data in 1979 and the 9 last columns correspond to data in 2006.

Out goal was to perform CA only the data corresponding to 1979 (the first 9 columns)

3.2 CA and results

We applied CA to the first 9 columns of the mortality dataset by using the *CA* function of the *FactoMineR* package. All the detailed steps of the followed procedure are described in the attached notebook "CA.Rmd".

We then analyzed in details all the outputs of the perfomed CA and we gave our interpretation by answering the required questions.

We report below only the most representative parts of the R notebook, the ones useful to answer the list of the required questions and the way we answered them.

3.2.1 First factorial plane interpretation and study of the similarities between age groups and causes of death

In this section we try to answer to the following questions:

- Interpret the first factorial plane (dimensions 1 and 2)
- Study the similarities between age groups according to causes of death
- Study the similarities between causes of death according to age groups
- Which are the main associations between age groups and causes of death?

Indeed, by looking at the factor map of the first factorial plane we'll try to interpret it by taking into consideration also the similarities found among age groups, causes of death, and the main associations among them.

In order to get a useful interpretation of the 1st factorial plane we decided to plot the biplot (factor map) considering only the 10 row points contributing more to the first 2 dimensions, while we added some transparency to the other points. We did this because of the many overlappings that were present in the biplot, due to the many rows contained in the dataset. The obtained biplot is shown in the figure 23 in the appendix.

The first factorial plane shows that most of the deviation from independence (i.e. inertia of the cloud of points) comes from the separation of the young age ranges, i.e. 15-24, 25-34, 35-44 and the rest.

This plane is able to cluster together causes of death that are more specific to certain age ranges and it also create groups of "similar" age ranges on the basis of the causes of death that they share.

In particular, the 1st dimension of the plot separates young people (on the right side) from old people (on the left of the origin): it puts in ascending order the ages from the right to the left of the axes. We can see that it also creates a kind of separation between 2 kinds of diseases: on the right there are causes of death not related to illness conditions, while on the left there are illness realated casues of death.

The 2nd dimension separates the age range 15-24 (on the top) from the rest of the ranges: in particular it shows an opposition of that range with the range 45-54 and 55-64, that are the middle age ranges. This 2nd dimension also puts in opposition tumour related diseases and "Chronic liver disease" with "Road accidents" (on the top of the 2nd dimension) and other types of diseases, such as heart diseases, "Cerebrovascular disease", "Other ill-defined symptoms and conditions", "Other accidents", "Events of undetermined intention" and "Suicides".

The position of each cause of death in the first factorial plane shows the ranges of age it affects more. For instance, we can immediately see that causes of death in the first quadrant of the plot are mostly related to young age ranges: for instance "Road accidents" affects more people in the range 15-24 and then 25-34, as expected, since young people are typically more reckless when driving. "Events of undetermined intention", "Other accidents","Suicides" and "Events of undetermined intention" are close to the ranges 25-34 and 35-44. While "Other ill-defined symptoms and conditions", "Other heart disease", "Cerebrovascular disease" affects more the older people in the ranges 85-94 and 75-84: again this is an expected behavior. In the lower part of the graph we have all the diseases related to the the middle age ranges, i.e. 35-44, 45-54, 55,64 and also 65-75: the typical diseases for this categories are various types of tumours and "Chronic liver disease". So, overall, we can identify 3 main clusters of causes of death: on the 1st quadrant we have a group of causes of death regarding younger people, on the bottom we have the diseases regarding middle age people and on the 2nd quadrant we have another group of diseases regarding the older people.

While the position of each range of ages shows the similarities between the various age ranges, depending on the causes of death: for instance, we can say that the ranges "95 and more", 85-94 and 75-84 are very closed to each other since they share many causes of death; we can say the same for the ranges 55-64, 35-44, 45-54 and even 65-74 which represent the middle ages; the same reasoning can be also done for the ranges 25-34 and 15-24.

Overall, what we can say is that the 1st factorial plane is very able to cluster the age ranges based on the causes of death most frequent for each range and, similarly, it is able to group together causes of death that regards specific ages and, finally, it is able to create meaningful associations between age groups and the diseases that typically affect them.

3.2.2 Which percentage of variability is explained by the first two dimensions?

The first aspect we typically look at in FA methods are eigenvalues, i.e. we try to understand how the newly created axes/dimensions are able to capture the deviation from independence of the original contingency table, that is how the overall inertia of our cloud of points is explained by the new axes.

Besides printing the list of eigenvalues and their explained variance and cumulative variance, we graphically visualized the variance explained by each dimension by means of a Scree Plot, shown in figure 24.

From the obtained graph and from the displayed inertia values, we easily concluded that the first 2 dimensions explain 89.98% of the variability of the entire cloud of points, therefore we concluded that the first two dimensions are enough to describe the whole contingency table, since they capture a very large percentage of the deviation from independence of our original dataset.

We can finally conclude that an analysis of the plane consisting of the first 2 dimensions would be sufficient to extract conclusions from the performed CA.

3.3 What are the age intervals best represented in the first factorial plane?

In under to understand how good a column point is represented in the new dimensions obtained after applying CA, we need to analyze the so-called quality of representation of the columns in the new dimensions. To do this, besides visualizing the *cos2* for all the column variables, we built a bar plot showing the best represented columns (i.e. age intervals) in the first factorial plane, composed by the first 2 dimensions. The plot is illustrated if the figure 25. We then easily conlcuded that the best represented age ranges are, in order: 15-24 followed by 55-64, 85-94, and so on.

3.4 What are the three causes of death most influencing the formation of the first principal component?

In order to answer to this question we analyzed the contribution of row points (i.e. causes of death) to the newly created dimensions, i.e how much each row point contributes to the inertia captured by each new dimension. In particular, we plotted a bar plot showing the 15 most contributing rows to the first factorial plane; the plot can be seen in the figure 26.

The plot clearly shows that "Road accidents", "Suicides", "Other accidents" are the three most important causes of death in the definition of the first dimension.

4 Multiple Correspondence Analysis

The third problem required us to apply Multiple Correspondence Analysis to the "bicing" dataset, provided as a *csv* file.

4.1 Introduction to the dataset

By reading the column headers, we can understand that this dataset has probably been made from data gathered by the *Bicing* company, offering bike rental service in Barcelona (or a similar bike rental company). Each of the 4.877 rows (or individuals) of the dataset represents a rental with information such as:

- Rental start and end date ("Start.date", "Start.time", "End.date", "End.time", "Total.duration", ...)
- Information on the start and end stations (identification with "Start.station.number" or "End.station.number", position and altitude with "Altitude.E" or "End.lat" for example)
- Weather information, although we don't know if the data is related to the start or end station (or not any of those). We will assume that it is the weather in the city over the day.

4.2 Preprocessing

In order to solve this task we applied some fundamental preprocessing steps before applying the MCA algorithm: first, we built the X matrix containing the useful active and supplementary variables, and then we converted the variable types into the correct ones. One example of those transformations is the "Month" variable : it was read as numerical and interpreted as continuous, but because the values taken were not continuous, we made the choice to convert it into a categorical variable.

The first thing we did was to select only the variables of interest, i.e. the active categorical variables, the quantitative supplementary variable and the qualitative supplementary variable. After doing that, we applied a type casting to the variables that were not read as categorical. All the steps of this prepropressing phase and choices that we made are described in details in the notebook "MCA.rmd".

4.3 Dimensions

MCA generates 16 dimensions which all together explain 100% of the variance of the original data.

We know that the number of dimensions generated by MCA is given by $K - J$, where K represents the total number of categories of the active categorical variables, while J represents the number of active categorical variables.

In our case J is equal to 6, while K is equal to the total sum of the categories of each active categorical variable which is 26 (the computation of this number can be found in "MCA.rmd"). Therefore, we should obtain $K - J = 26 - 6 = 20$ dimensions/eigenvalues.

However, as we are applying ventilation (of 1%), 4 categories are removed for being rare which leaves us with 16 dimensions. For the rest of the exploitation, we will only keep the 10 first dimensions as we specified when computing MCA (through the *ncp* argument).

4.3.1 Keeping relevant dimensions

The problem with applying MCA on this dataset is that many dimensions are required to explain 100% of the variance of the data, as usually happens in the case of MCA, since we have many cateogories.

4.3.2 Which percentage of variability is explained by the first two dimensions?

As can be seen on Figure 15, the first dimension accounts only for 8.6% of the explained variance and the two first account for 16%.

4.3.3 Decide the number of significant dimensions that you retain

The main problem is: how can we keep significant dimensions without loosing on representation quality?

In order to decide the number of significant dimensions to retain, we know that we can use an empirical formula, which states that a good criteria is to interpret the axes of inertia above $1/J$, where J is equal to the number of active categorical variables, which is 6 in our case.

According to this empirical formula, we should retain the first 8 dimensions, as can be seen in the computation performed on the notebook.

4.3.4 Variable contribution to dimensions

With plotting the contribution of variables to the first and second dimensions (Figure 16 and Figure 17), we can explain:

- For the first dimension: we first note that variable categories have sparse rankings. It is mostly "Period" (with "Morning" 1st and "Night" 5th), "Account.type" (with Casual as 2nd) and "Start.weekday" (with "Saturday" 3rd and "Sunday" 4th) that are the top contributors to this dimension.
- For the second dimension: the variable categories are less spreaded than in the first dimension. Top contributors are Weather (with "Clear" 1st and "Cloudy" 2nd) then "Period" (with "Night" 3rd and "Afternoon" 4th) and "Start.weekday" (Friday 5th and Monday 6th).

We can also compare both dimension correlation. In overall, Period is the top contributor to both dimension (Morning/Night for the 1st, Night/Afternoon for the 2nd), followed by Start.weekday although contributing with different days (Saturday/Sunday for the 1st, Friday/Monday for the 2nd).

We computed a correlation score of -0.17. It is normal that there is not that much correlation beween those two dimensions as the goal of computing those dimensions is to maximize variance explaination thus reducing correlation at maximum (or having those two dimensions would be useless as giving the same insights).

4.4 Interpretation of MCA results

4.4.1 Interpret the first factorial plane (dimensions 1 and 2) by means of variables and categories

After applying MCA, we can now get insights on the data. To do so, we will first use the variable & categories plot on the principal dimensions (1st and 2nd).

4.4.2 Variable and categories plot

On the variable and categories plot (Figure 18), we can see how much each variable categories are correlated to each others. For example, we can see a strong negative correlation between 'Saturday' and 'Morning' as being opposed compared to the origin of the plot. This means that a very few profiles were combining those two variables, if not none.

Contextualization: A very few bike rentals were made on Satursday mornings, compared to other rentals.

Also, we can see that there is a relatively strong correlation between between 'Friday' and 'Wednesday' variables or between 'Showers' and 'Monday'.

Contextualization: Profiles of bike rentals of Friday and Wednesday are pretty similar, which could be due to populations going out at night and renting bikes to do so (as it is less frequent to go out at the beginning of the week).

4.4.3 Variable plot

On the variable-only coordinate plot, we can see how each variables are correlating with the 2 first dimensions. For example, we can see that 'Weather.main' is strongly correlated with the second dimension when 'Temperature' is having a low correlation with both dimensions, thus not having a strong relation on this dimension. Also, there are variables such as 'Period' or 'Start.weekday' which have a mixed correlation with both dimensions.

4.4.4 Variable representation

Above we interpreted the contribution of each variables in the dimensions using the eigenvalue decomposition and others. But how can we ensure that each variable are well represented in those dimensions ? This is where the 'cos2' score becomes relevant: it measures each variables' representation. A plot of this cos2 score for Dim1 and Dim2 can be found in Appendix (Figure 20). The more the cos2 score is higher (red), the more the variable is being represented in the dimension.

For the other variables, identify the dimensions where the categories of "Metro.station.E" and "Downtown.E" are best represented. After ranking the cos2 score of those two variables for each dimension, we can state:

- **Dim7 and Dim10 are the most representative** of 'Metro.station.E' with respectively 0.29 and 0.27 cos2 score when Dim4 (3rd position) only have 0.11, which makes it less representative than the two first dimensions.
- **Dim3 is clearly the most representative** of 'Downtown.E' with a 0.33 score. The second most representative dimension is the Dim1 with a 0.13 score which is way lower (and its even worse with the third most representative dimension Dim2 with 0.04 cos2 score).

4.4.5 Supplementary variables

Are the supplementary variables related to the principal axes? In which way?

The two principal dimensions are not so correlated to the "Temperature" variable (Figure 21). We also verified if other dimensions were more correlated but it was not the case as the highest correlation value is with Dim6 and is very similar to Dim2's correlation value (around 0.168).

Therefore we can conclude that the supplementary quantitative variable is not well correlated to the 10 first MCA dimensions.

The conclusions for the "Month" variable are relatively similar. On Figure 22, we can see the coordinates on the 1st factorial plane. While in figure 21 we can see the correlation of the "Month" variable with the 1st factorial plane (that we already explained before). The notebook contains the full results for whatr regards that supplementary qualitative variable.

For a conclusion about supplementary variables: whether it is quantitative or qualitative variables, we can see that they are not really related to principal dimensions Dim1 and Dim2 (but can be with other dimensions).

5 Conclusions

This first assignment allowed us to get in touch with factor analysis methods for dealing with multivariate data. We understood the importance of those methods for summarizing and visualizing the information contained in large multivariate datasets. We understood the differences of the three methods and especially in which cases they should be applied. During this first excercise we had the possibility to improve our R skills and to deepen our knowledge of the *FactoMineR* and *R stats* packages which provide a very simple and intuitive way to apply the various factor analysis methods.

In conclusion, we are very satisfied with what we learnt during this assignment, especially beacuse we know that principal component methods are a very useful preprocessing steps in data analysis, they are also important to get some useful insights about the relationships and associations among different variables, items, individuals in multivariate datasets.

6 Appendix

6.1 PCA

Main plots of the PCA analysis.

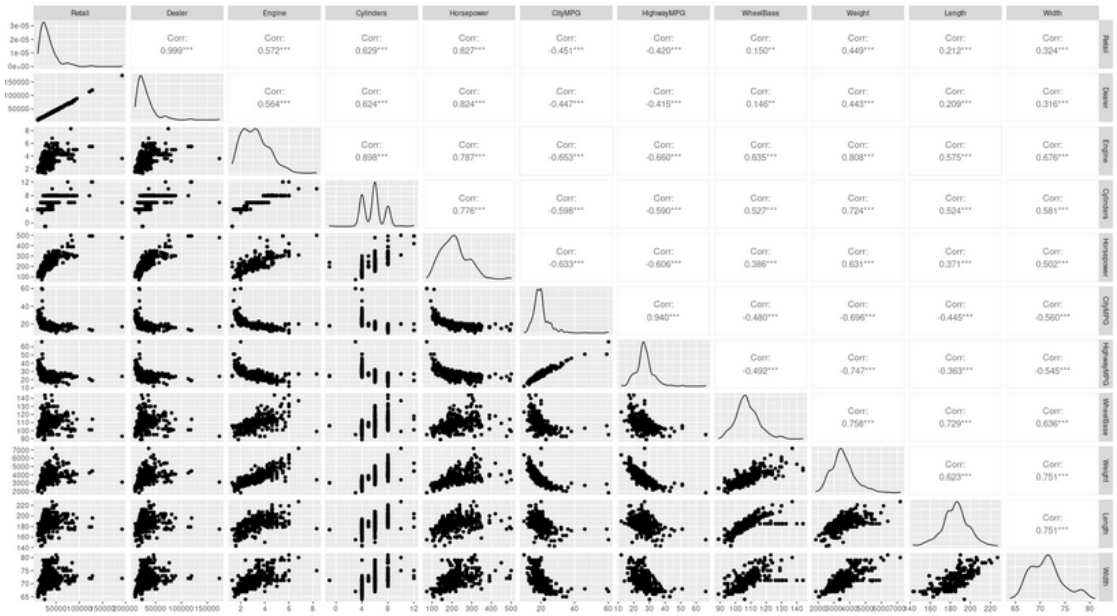

Figure 1: Scatterplot of the dataset.

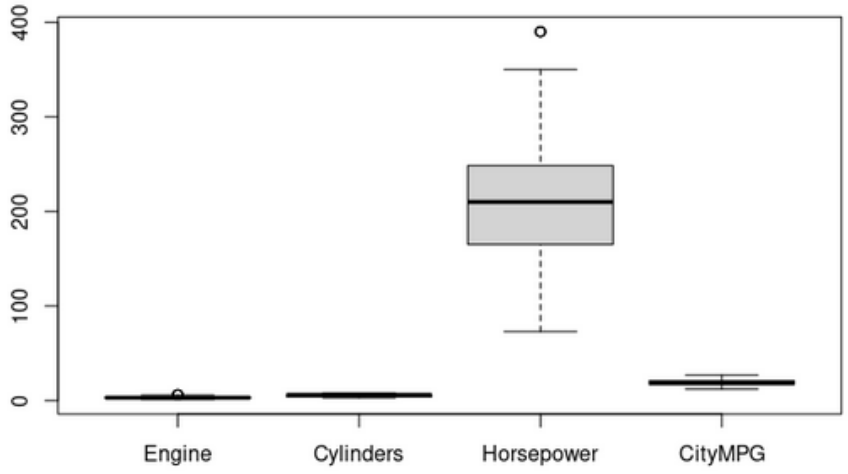

Figure 2: Boxplot of variables Engine, Cylinders, Horsepower and CityMPG.

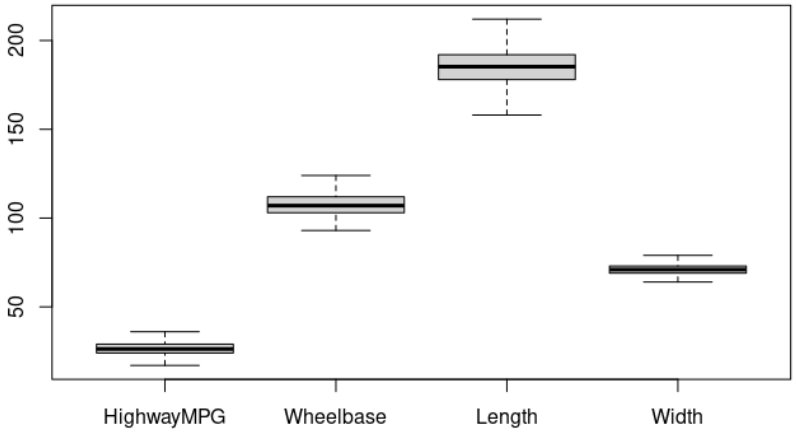

Figure 3: Boxplot of variables HighwayMPG, Wheelbase, Length and Width.

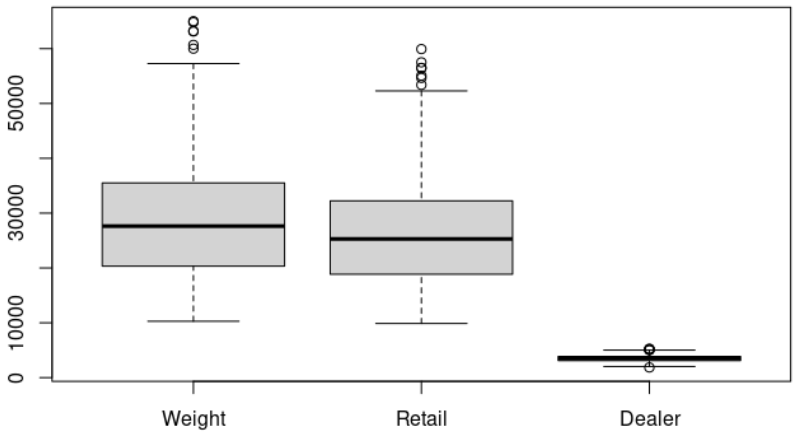

Figure 4: Boxplot of variables Weigth, Retail and Dealer.

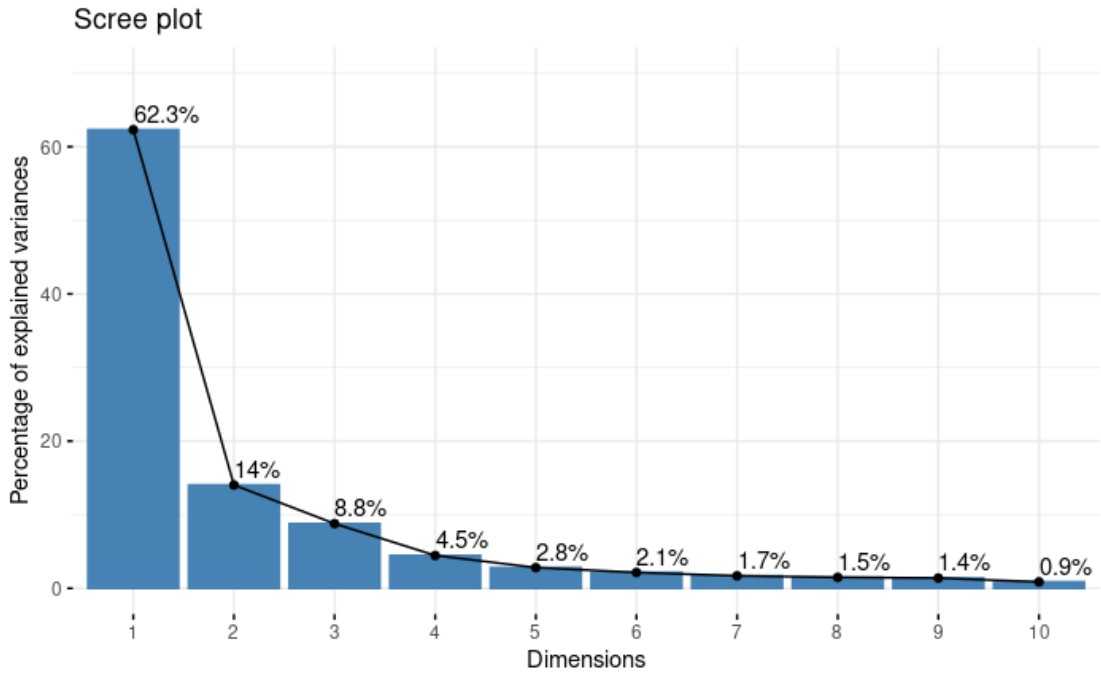

Figure 5: Proportion of variable explaination for each variable

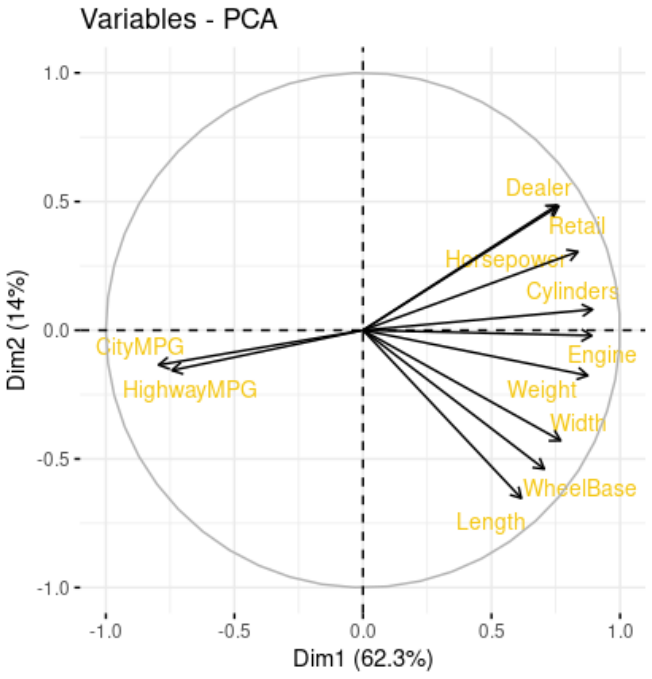

Figure 6: Relationship between variables

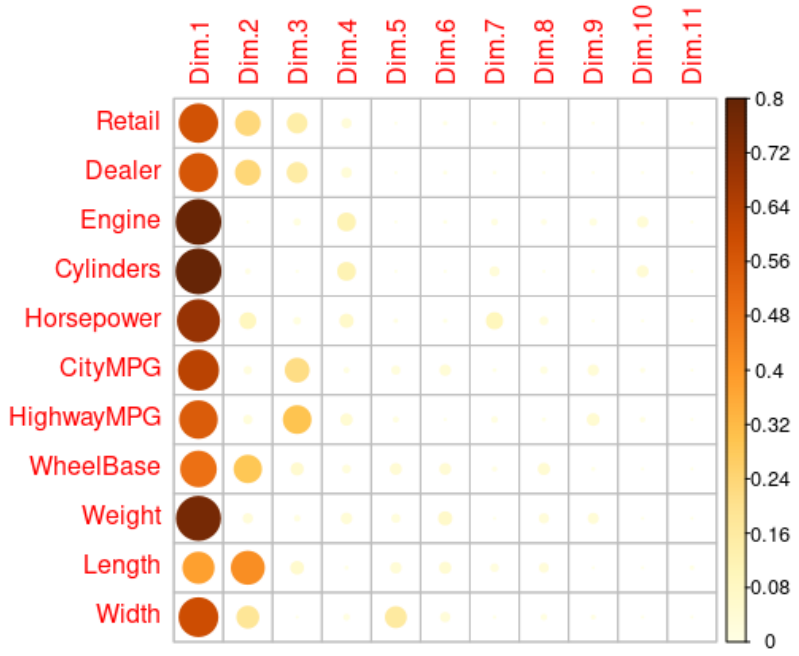

Figure 7: Cos2 score plot for variable and categories on Dim1 and Dim2

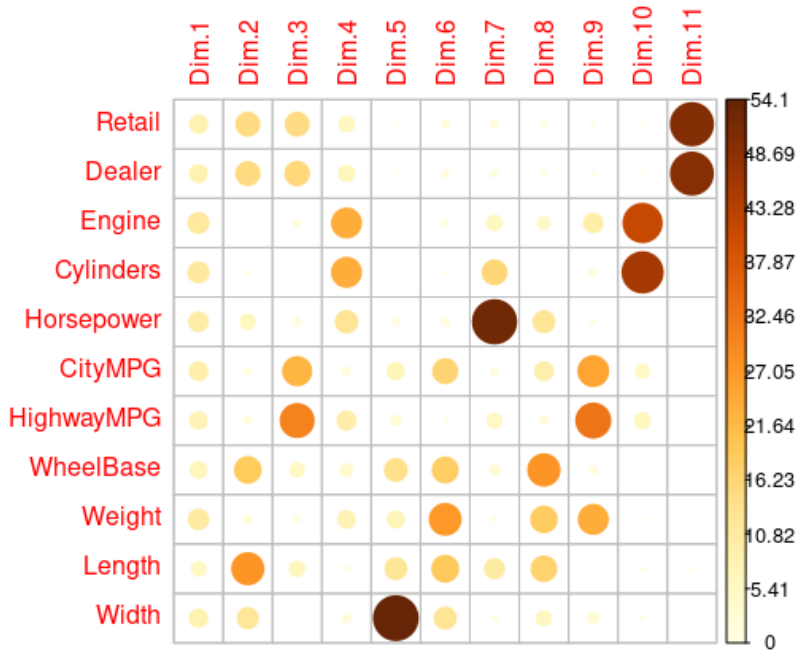

Figure 8: The contributions of variables in accounting for the variability in a given principal component

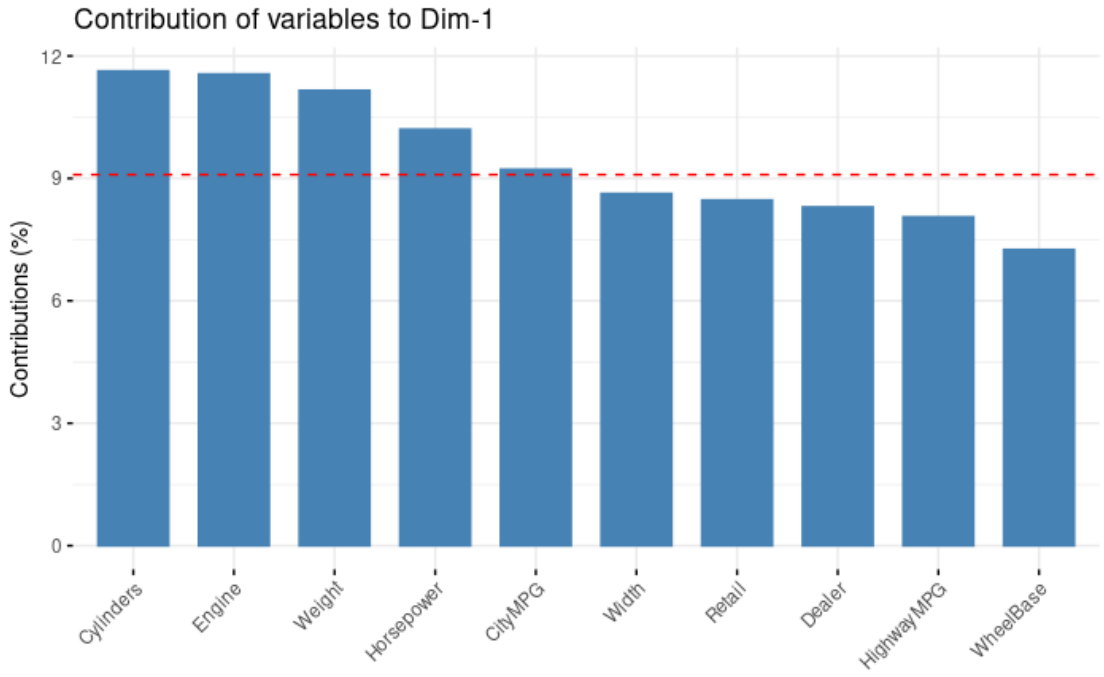

Figure 9: Contribution of variables to PC1

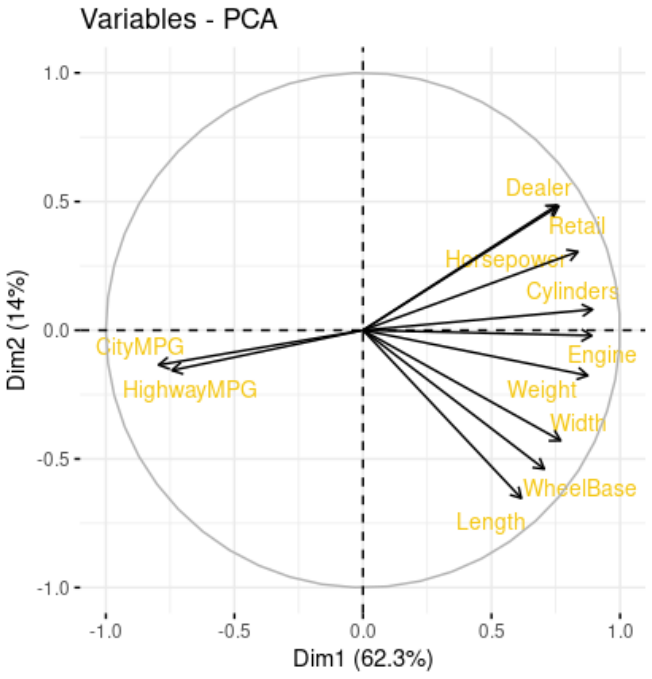

Figure 10: Contribution of variables to PC2

PCA scores

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
-4.745984	0.3994584	-0.7308411	-0.3395705	0.3223562	0.4184159	0.7437991	-0.6016117	-0.9797073	0.2977211	0.0103553
-4.319604	-0.4054732	-0.2550816	-0.4343770	0.9299903	1.1837491	0.2237446	0.0887011	-1.0048496	0.2036522	0.0071099
-3.380559	-0.8125141	0.0586200	-0.0779775	-0.0008481	-0.3420334	-0.4591238	0.3612503	1.4464502	-0.4149734	-0.0087511
-3.430932	-0.7033502	0.0636501	-0.0880839	0.2059852	-0.4989195	-0.4907921	0.4821711	1.3481792	-0.3900448	-0.0056795
-3.287393	-0.6870351	0.1851884	0.0041563	0.0118792	-0.3033762	-0.4198902	0.3898207	1.4673014	-0.4026706	-0.0127961
-3.862483	-0.3697403	0.0681276	-0.4343836	0.4992817	0.1569444	0.6434898	-0.9592688	-1.1435850	0.6725223	0.0110541

Figure 11: Values of each PCs

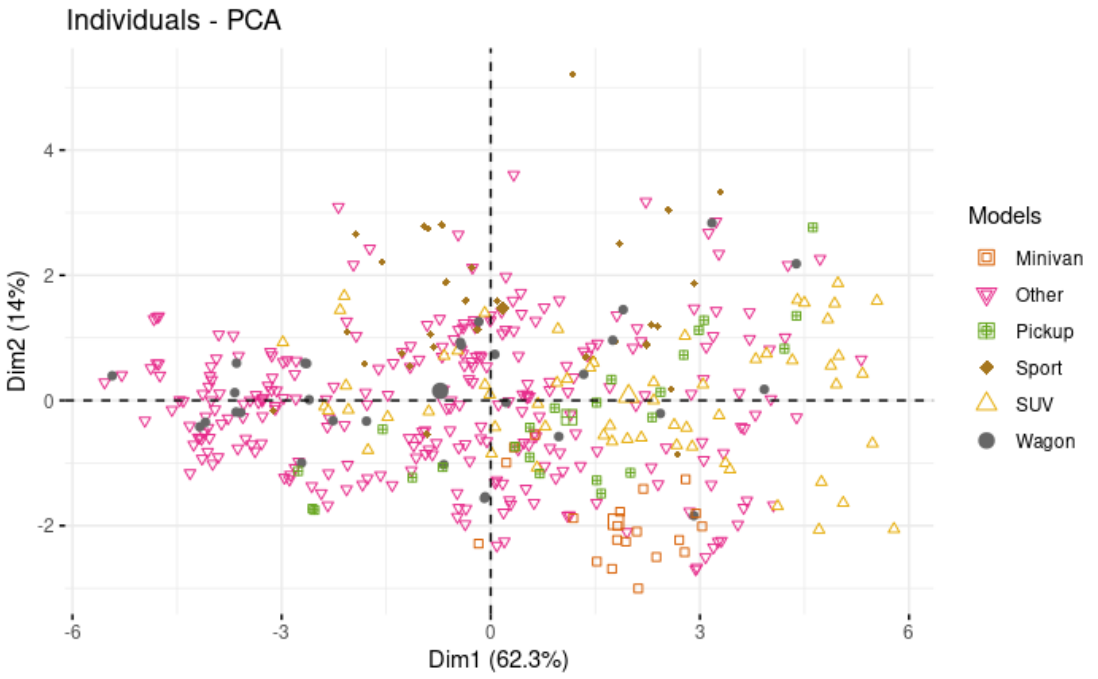

Figure 12: Clusters of cars based on their similarity.

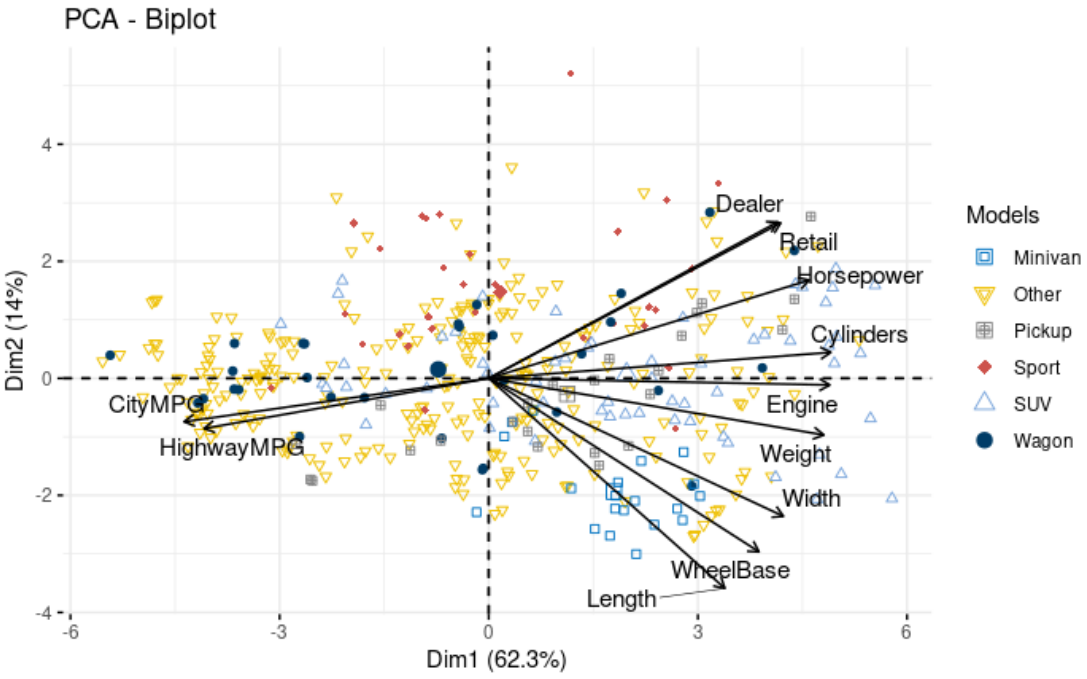

Figure 13: Combination of the PCA score plot, together with the loading plot.

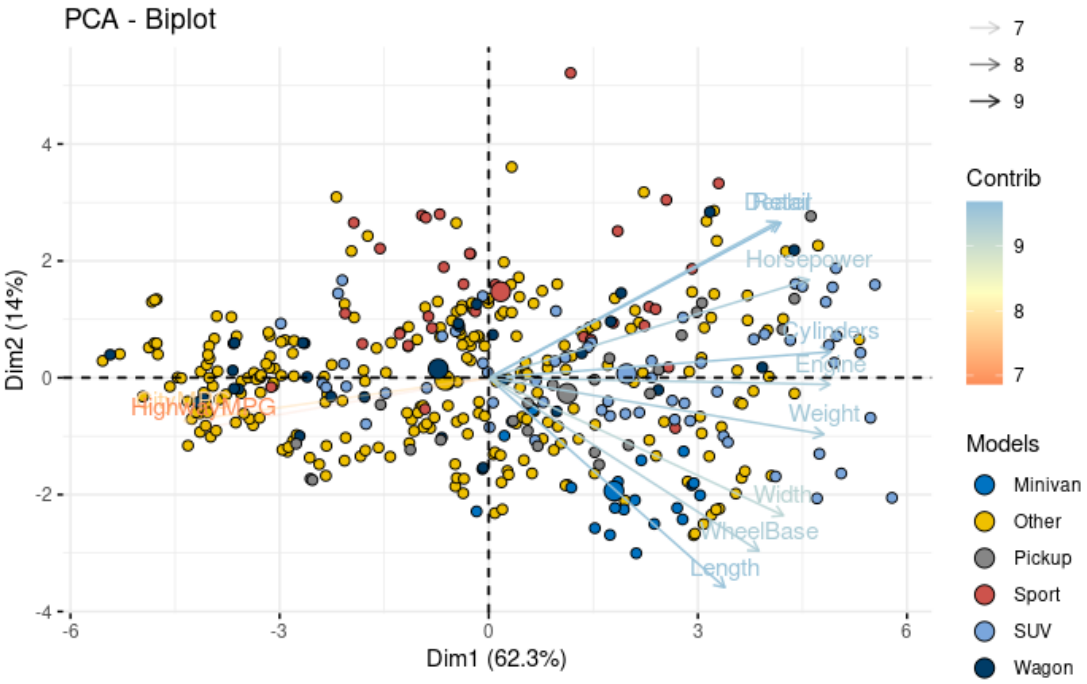

Figure 14: Colored biplot,the groups are colored based on their contributions.

6.2 MCA

Main plots of the MCA analysis.

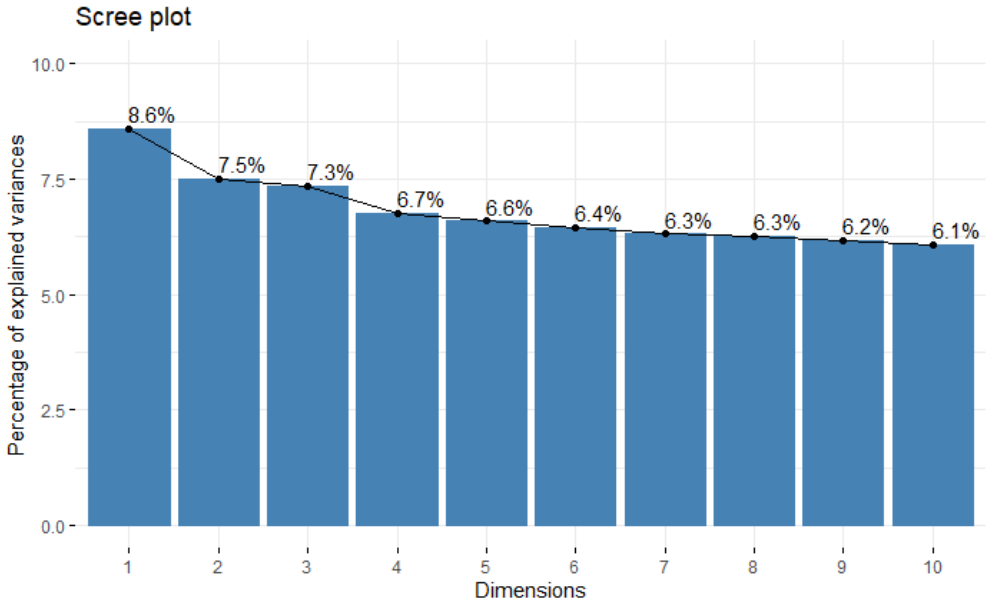

Figure 15: Proportion of variance explained by each dimension

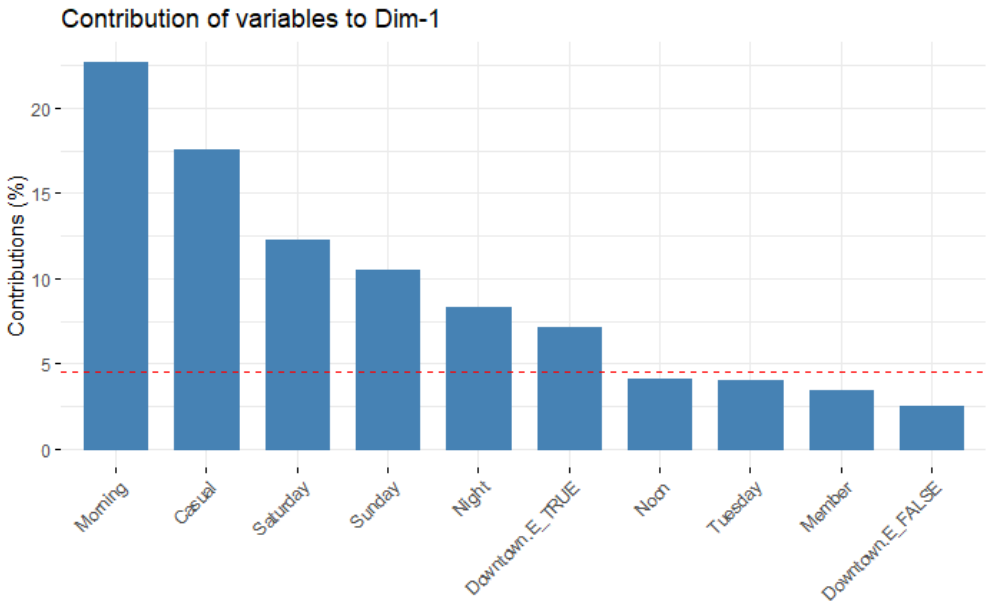

Figure 16: Contribution of variables to Dim1

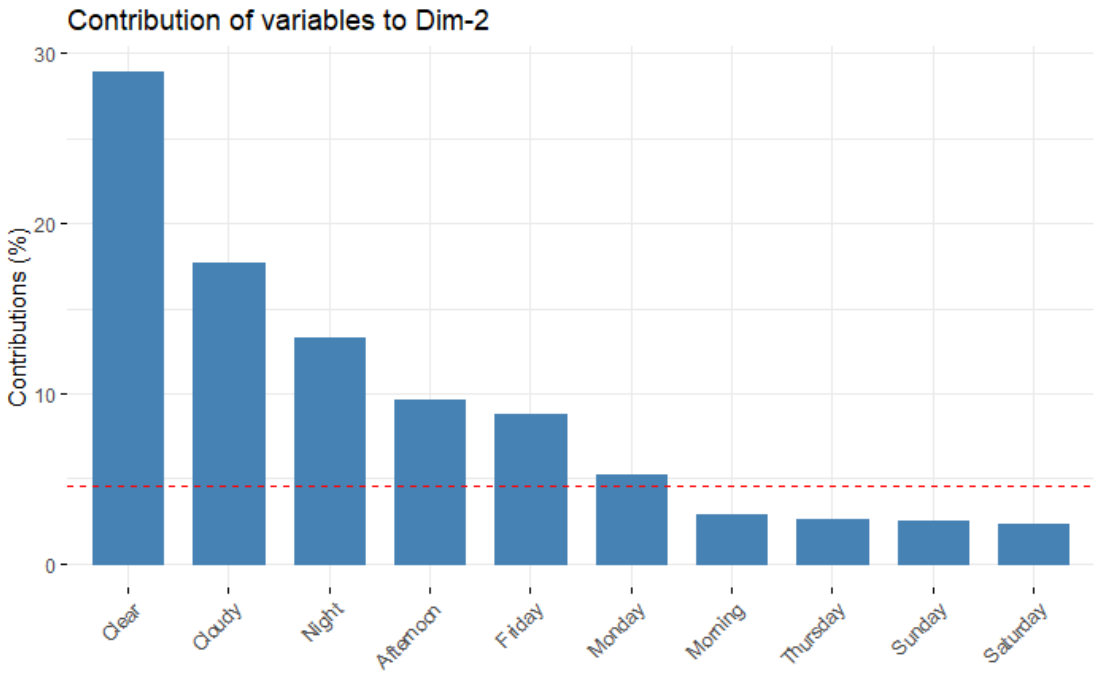

Figure 17: Contribution of variables to Dim2

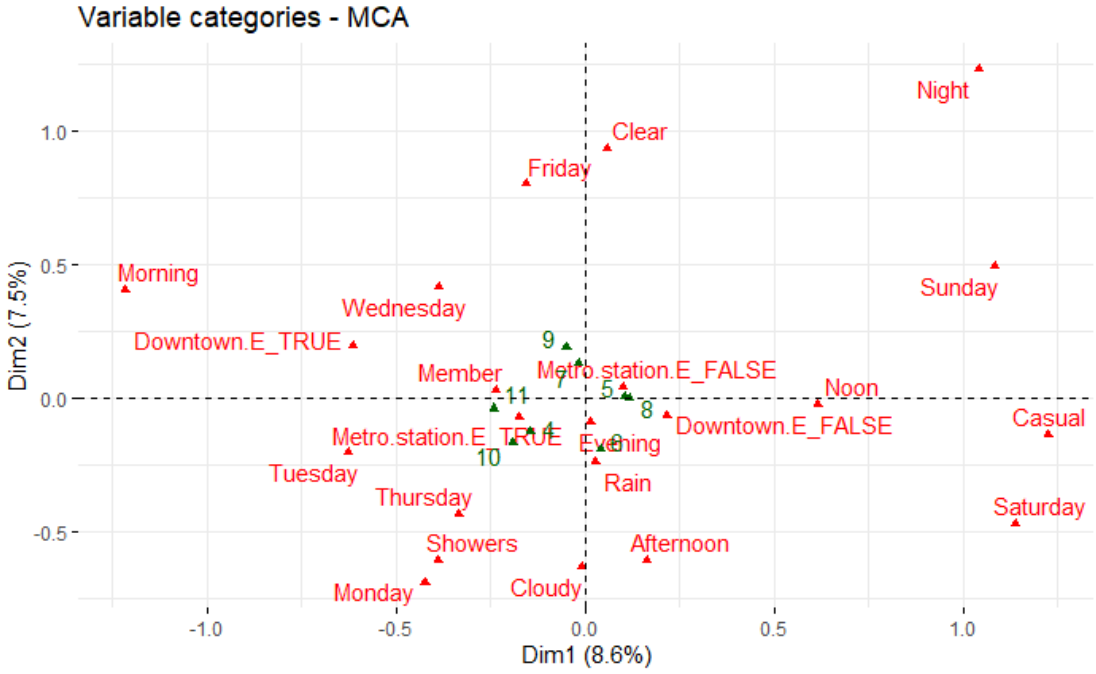

Figure 18: Variable and categories coordinates plot on Dim1 and Dim2

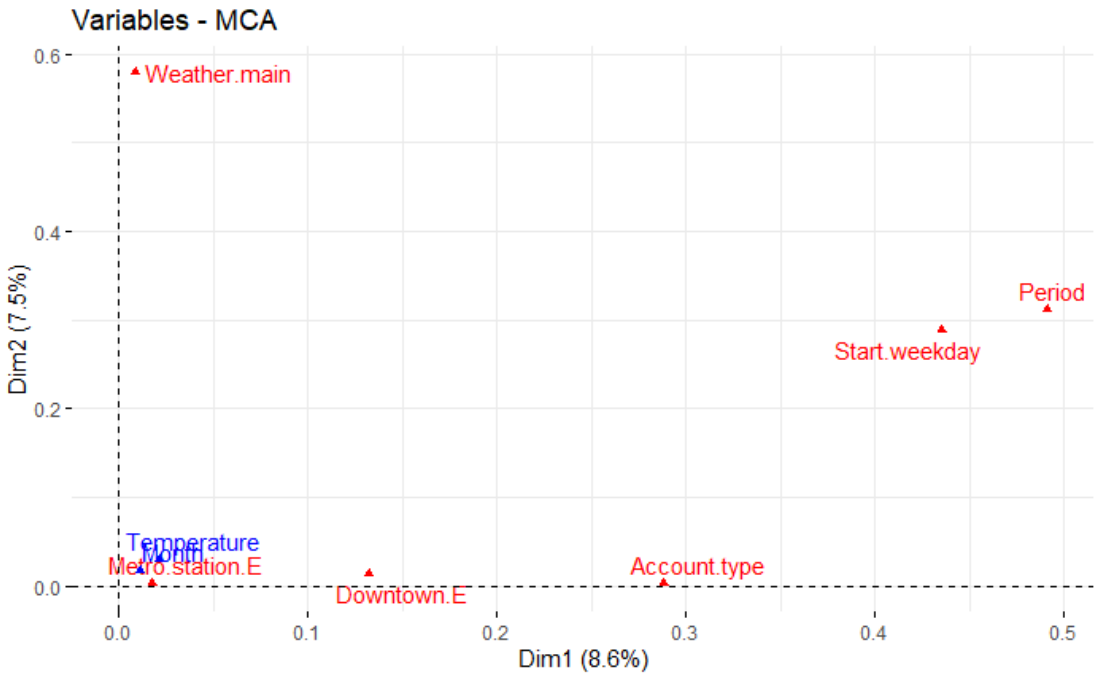

Figure 19: Variable only correlatio on Dim1 and Dim2

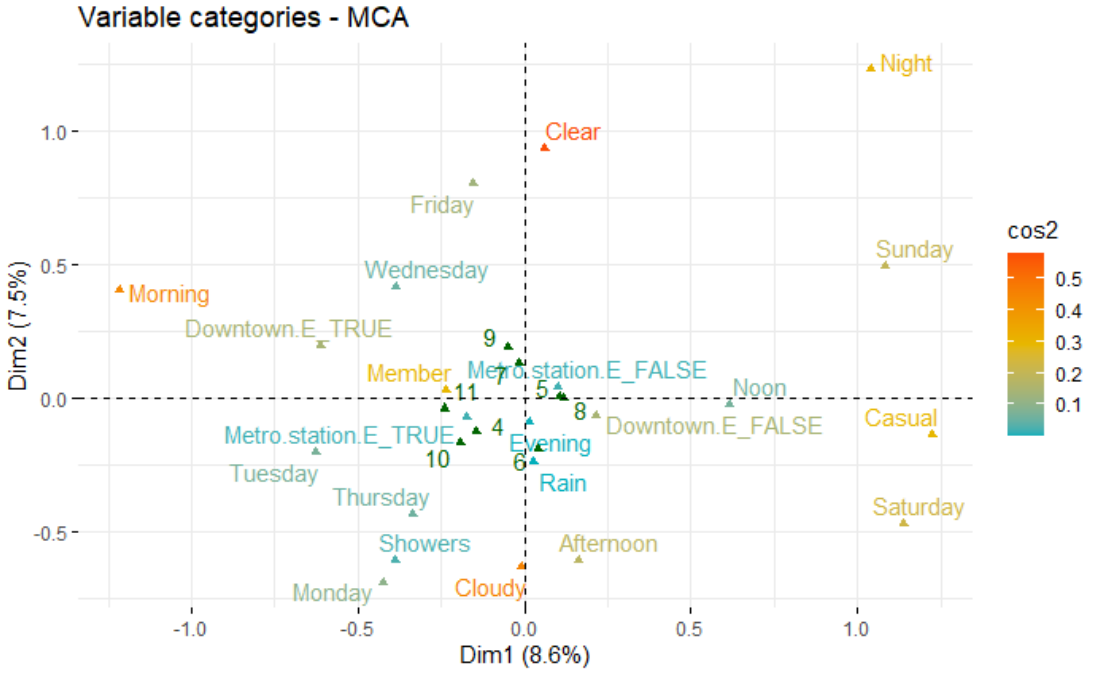

Figure 20: Cos2 score plot for variable and categories on Dim1 and Dim2

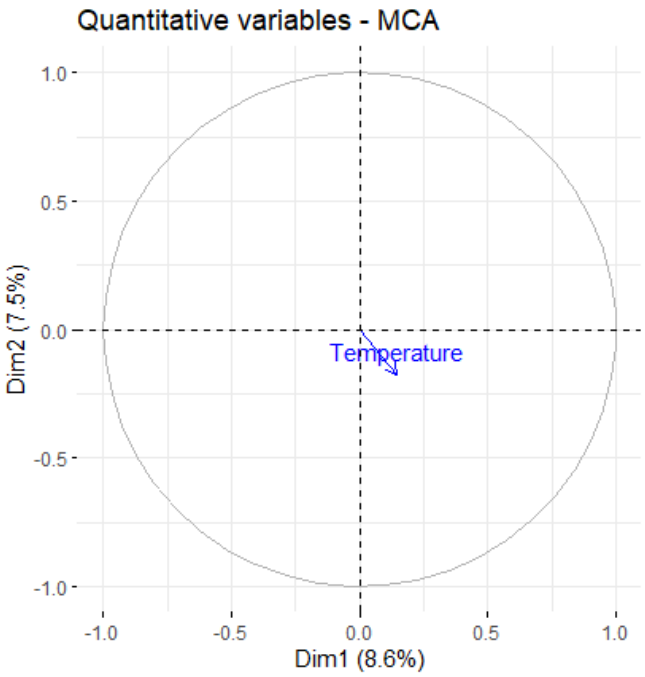

Figure 21: Correlation of the supplementary quantitative variable

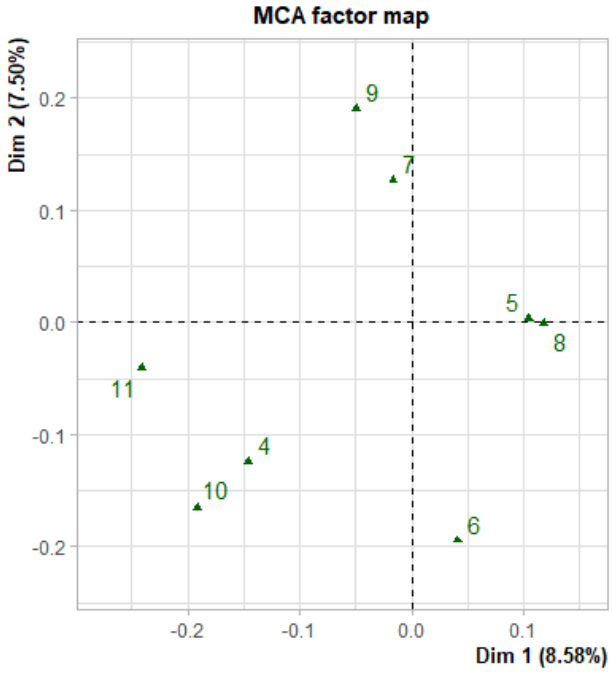

Figure 22: Coordinates of the supplementary qualitative variable

6.3 CA

Main plots of the CA analysis.

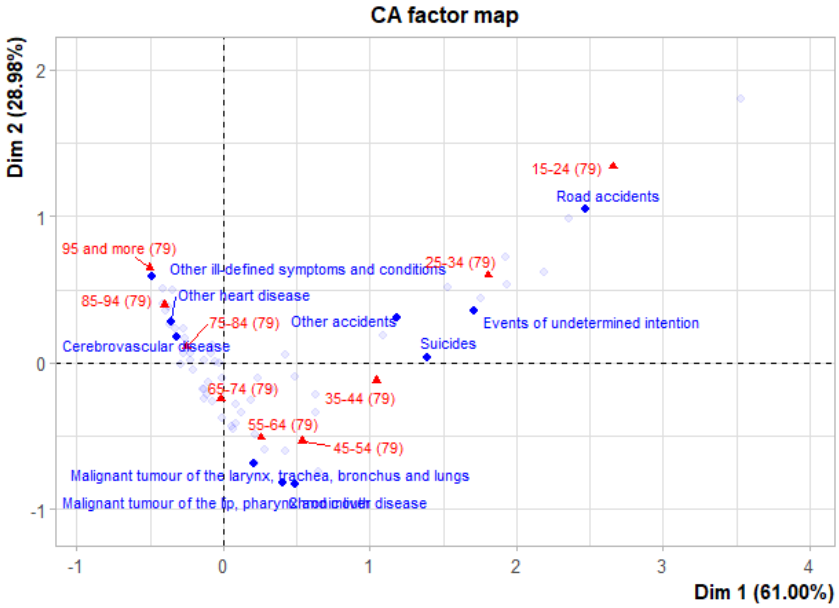

Figure 23: CA Biplot with 10 most contributing row points

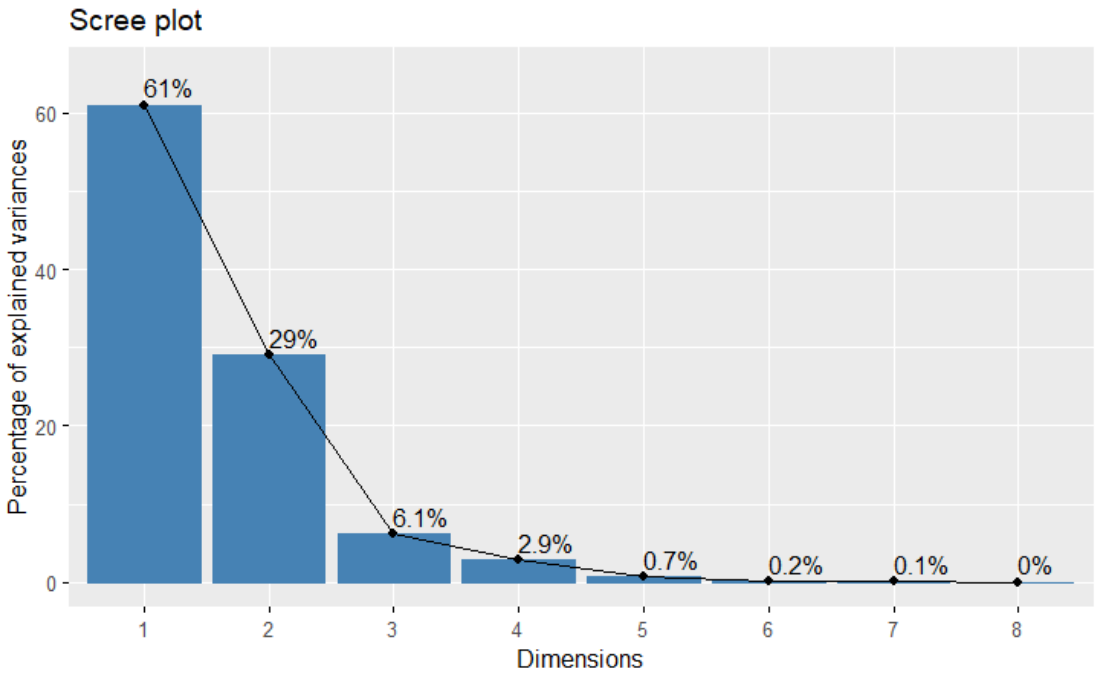

Figure 24: CA Scree Plot

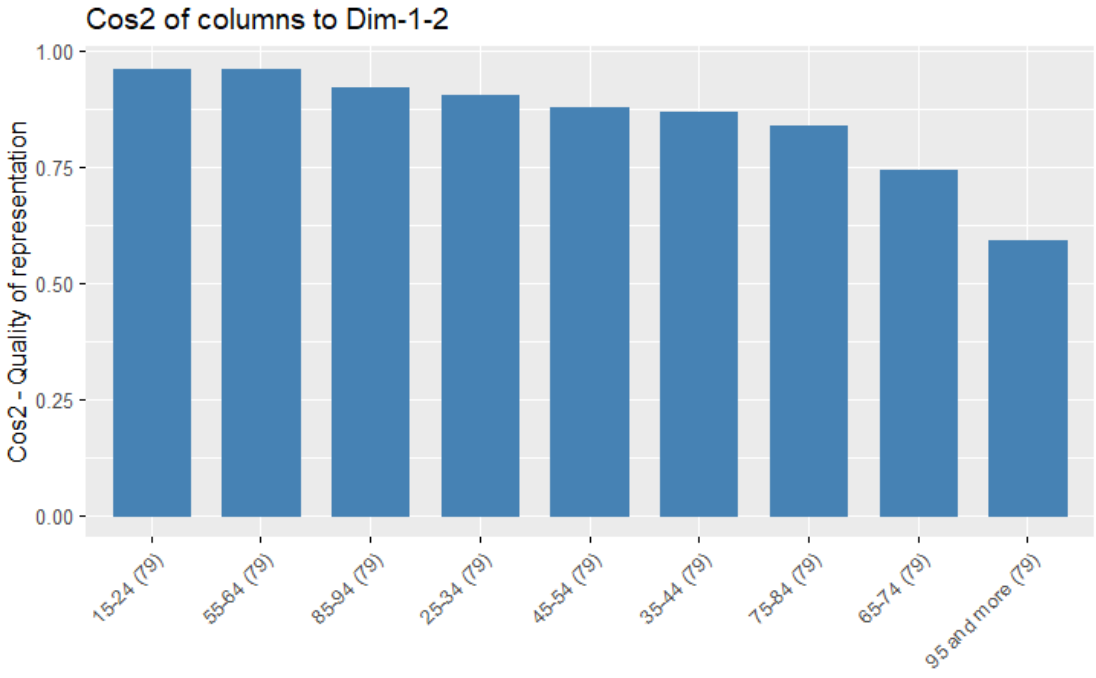

Figure 25: Quality of representation of age ranges on the 1st factorial plane

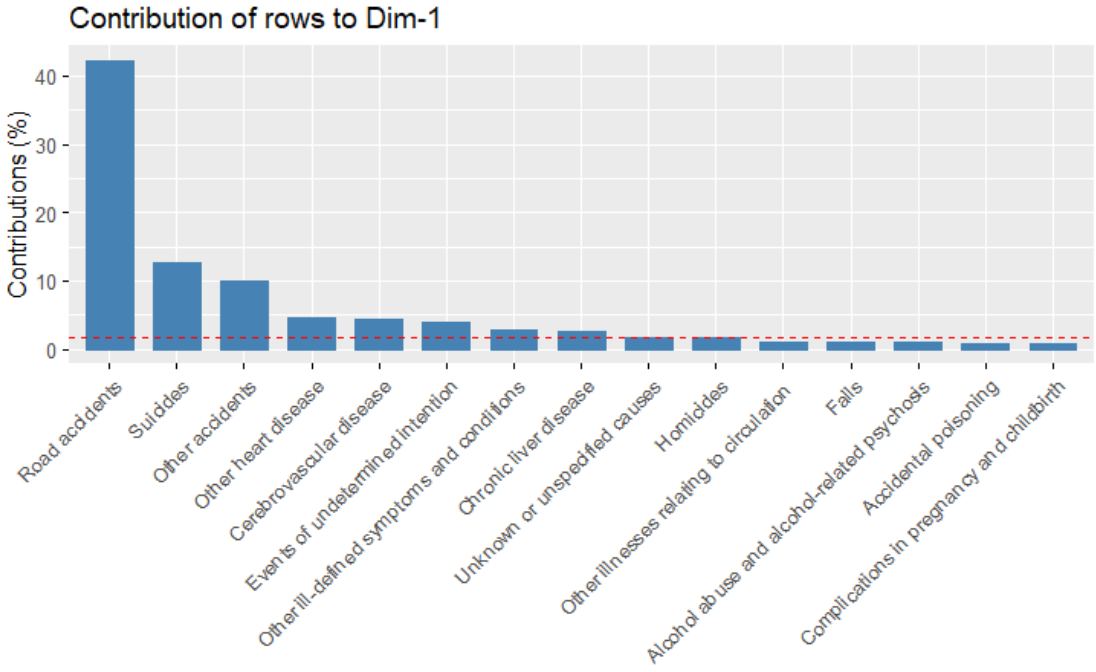

Figure 26: Causes of death most contributing to the 1st dimension