US Accidents (2016 - 2021)

MULTIVARIATE ANALYSIS

Dataset URLs

Initial dataset https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Updated dataset
https://drive.google.com/file/d/1tv_um3xnRHhO04-qO29VtPt21GY5BfB4/view?usp=sharing

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Group 8:
Ander Barrio Campos
Biel Caballero Vergés
Miona Dimic
Gerard Gómez Jorba
Odysseas Kyparissis
Cecilia Pérez Pérez

Table of Contents

Summary of Report

This project focuses on accidents that happened in the US between the years of 2016 and 2021 and aims to make a deep analysis of the data to fully understand the nature of the dataset and of course, the numerous road accidents that took place in the past years in several states of the US.

The project will be carried out in different steps according to the different types of analysis that need to be carried out due to the different variables the dataset includes. This paper will start by introducing the dataset and the metadata and will be followed by a concise description of the preprocessing steps that the data has undergone in order to be clear, exempt from missing values and treatable for future analysis.

Next up, taking the preprocessed variables, a univariate and bivariate statistics are presented followed by the different analysis depending on the type of variables. These analyses include Principal Component Analysis for numerical variables, MCA analysis of multiple qualitative variables and Multiple Factorial Analysis for both numerical and qualitative variables. To finish off, the association rules are presented for this dataset.

The last part makes reference to the appendix. Plots which were generated during the analysis process of the dataset but did not provide quality levels of insights are placed.

Motivation and General Description

Nowadays, traffic accidents are one of the most frequent problems surrounding society. Road safety and minimization of accidents have to play a crucial role in the development of the road network of cities. Besides that, regularly, weather conditions and road formation characteristics are the most impactful factors for the appearance of road accidents. Nowadays, with the help of data and the appropriate analysis, regulation bodies which are responsible for the security of the road networks can make more sophisticated data-driven decisions.

The selected database is of great interest in order to tackle different problems concerning road accidents, as mentioned before. The country of origin of these accidents is the United States. The large number of recorded traffic accidents makes it easier to draw conclusions on the topic, which will lead to better decision-making in the future. The number of attributes in the database is large and satisfies all the restrictions imposed by the faculty in charge of the project. However, not only are there a large number of them, but these attributes contain intriguing information which offer great insights for the occurrence of accidents.

Turning to the usefulness of this data, as mentioned above, these accidents can be very useful for predicting real-time car accidents, studying car accidents hotspot locations, casualty analysis and extracting predicting effect rules. Finally, studying the impact of precipitation, humidity or other environmental stimuli on accident occurrence can be accomplished.

Data Source Presentation

The dataset used as the basis of our project assignment is the US Accidents from 2016 until 2021. It covers 49 states of the USA and it uses multiple APIs that provide streaming traffic incident (or event) data. There are about 2.8 million accident records in the initial dataset with 48 different variables describing them. For the scope of the specific assignment, around 1 million observations were randomly selected. Finally, a selection on the variables of the dataset has taken place as well, resulting in the attributes of the final dataset which are presented in the following table:

Variables of the Dataset					
ID	State	City	Weather_ Condition	Year	County
Crossing	Bump	Stop	Traffic_ Signal	Severity	Wind_ Chill(F)
Visibility(mi)	Distance(mi)	Duration(min)	Precipitation (in)	Pressure(in)	Humidity(%)
Temperature (F)	Start_Time	End_Time			

Table 1: Variables of the dataset

Additionally, information of only five different states was maintained in order to reduce the massive amount of data.The chosen US states are illustrated below:

Chosen States of the Dataset				
California (CA)	Pennsylvania (PA)	Tennessee (TN)	South Carolina (SC)	Connecticut (CT)

Table 2: States used in the dataset

These states have been chosen randomly with R, and they also form one qualitative variable.

Description of Data Structure and Metadata

Each row of the dataset contains information about the chosen variables mentioned above, for each one of the accidents.

The dataset used has data related to the location of the accident (state, city, county, etc.), when it occurred (start time, end time and year), values on weather conditions (pressure, humidity, temperature, precipitation etc.) and other data which can be visualized in the metadata table shown in Figure 1. Ultimately, there is one qualitative attribute (Severity) describing the severity of the accident.

ALL VARIABLES	UNITS	% NA	TYPE	DESCRIPTION
ID	Null	0,00%	Null	Unique identifier of the accident record
State	Null	0,00%	Qualitative	State where the accident took place
City	Null	0,00%	Qualitative	City where the accident took place
Weather_Condition	Null	0,00%	Qualitative	Weather condition at the time of the accident
Year	Year	0,00%	Qualitative	Year of the accident
County	Null	0,00%	Qualitative	County where the accident took place
Crossing	Null	0,00%	Binary	Was there a crossing nearby where the accident took place
Bump	Null	0,00%	Binary	Was there a bump nearby where the accident took place
Stop	Null	0,00%	Binary	Was there a stopsign nearby where the accident took place
Traffic_Signal	Null	0,00%	Binary	Was there a traffic signal nearby where the accident took place
Severity	Null	0,00%	Qualitative	Severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay)
Wind_Chill(F)	Fahrenheit	14,61%	Quantitative	Wind chill at the time of the accident. Wind chill or windchill is the lowering of body temperature due to the passing-flow of lower-temperature air. Wind chill numbers are always lower than the air temperature for values where the formula is valid. When the apparent temperature is higher than the air temperature, the heat index is used instead
Visibility(mi)	miles	2,13%	Quantitative	Visibility (in miles) in the scene of the accident
Distance(mi)	miles	0,00%	Quantitative	The length of the road extent affected by the accident.
Precipitacion(in)	inches	18,47%	Quantitative	Precipitation at the time of the accident
Pressure(in)	inches	1,95%	Quantitative	Air pressure at the time of the accident
Humidity(%)	%	2,58%	Quantitative	Humidity at the time of the accident
Temperature(F)	Fahrenheit	2,45%	Quantitative	Temperature at the time of the accident
Start_Time	Time	0,00%	Null	Shows start time of the accident in local time zone.
End_Time	Time	0,00%	Null	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow

QUALITATIVE	MODALITIES	MEANING
State	CA, PA, TN, SC, CT	5 different states
City	Livermore, Ontario, Arvin ...	Different cities from the selected states
Weather_Condition	Fog, Haze, Heavy Snow ...	Different type of weather conditions
Year	2016, 2017, 2018, 2019, 2020, 2021	5 different years
County	Alameda, Knox, Kern ...	Different counties from the selected states
Severity	1, 2, 3, 4	1 indicates the least impact on traffic whereas 4 means a huge impact

QUANTITATIVE	RANGE
Wind_Chill(F)	[-22, 120]
Visibility(mi)	[0, 140]
Distance(mi)	[0, 137.6180]
Precipitacion(in)	[0, 24]
Pressure(in)	[0.30, 58.90]
Humidity(%)	[1, 100]
Temperature(F)	[-22, 120]

Table 3: Description of the variables

The final scope of this study is to find a relationship between the accidents and the weather and physical conditions of the environment. We believe that, with this dataset and the pertinent analysis, we'll be able to justify and determine the causes of the numerous traffic accidents.

Therefore the variables that we have decided that would best describe our case of study are those related to weather metrics and conditions. In that way we can analyze the relationship between weather variables and the severity of the accidents, therefore our target value is severity. The selected features also include the physical aspects of the environment so as to find a correlation between presence and absence of traffic signals, stop signs, crossroads, etc., at the place where the accidents take place. Consequently, some variables that we have taken into account are the severity of the accidents and also the distance (the length of the road extent affected by the accident) and the time of the accident.

Variables such as coordinates (latitude and longitude) of the accident, street number, zip code and side were not taken into account as we already have location variables that define in a more general way where the accident took place. The variables airport code and time zone were also omitted due to the fact that they would be redundant to our chosen location and start-end time variables.

At last, other variables that defined more transit elements were not taken into account because we decided that presence of bumps, stop signs, traffic signals and crossing were the minimum required binary elements that would describe the environment enough.

Preprocessing

In order to continue with the analysis of our dataset, the first step has been to clean the data and then impute for any possible missing value. The total amount of NAs found in the dataset is 473565, meaning that an imputation must be undertaken.

In the following paragraph, a list with all the preprocessing steps is shown, as well as the justification of all the decisions taken during these steps.

1. This amount of NAs is also a consequence of the numerous variables, so the number of levels in the Weather Condition dimensions have been reduced:
   Regarding the weather conditions, as there are a lot of low proportionality ones, it had been decided to group weather conditions with less than a 1% representation into a new level called “Other”. These are the resulting levels, after the grouping:
   - Clear
   - Cloudy
   - Fair
   - Fog
   - Haze
   - Light Rain
   - Mostly cloudy
   - Other
   - Overcast
   - Partly cloudy
   - Rain
2. The second step followed is the change of qualitative variables into factors: To do this, the as.factor function was used for every qualitative dimension (State, City, Weather Condition, Year, County, and Severity).

3. The final step undertaken in the preprocessing process has been to treat the missing data treatment, these have been the substeps that make up this step.

   3.1. Regarding the missing data, first, a little’s test was carried out to see what kind of missing values we had. The resulting p-value of this test was 0 therefore, we reject the null hypothesis, so our data is not MCAR (Missing completely at random). Then, it was taken into account which were the variables that had the most missing values, so they could be tackled and tried to be reduced in the amount of missing values using logical imputation.

   3.2. It was found that the column with the most missing values was “Precipitation.in.”, with 197195 missing values. After some data exploration it was noticeable that the majority of missing values appeared when the weather was “good” (e.g "Clear", "Cloudy", "Fair", "Overcast", "Partly Cloudy"). This implies that there was no precipitation, and in that case sometimes it was recorded as 0 and others as NA, so it was decided to impute these remaining values as 0.

   3.3. The second column with the most missing values was “Wind_Chill.F.”, with 155984 missing values. A similar scenario appeared as before, in some “good” weather conditions some values were recorded as NA, so we decided to impute

the parameters with the mean of its temperature on its weather condition, when the weather condition was “good”.

3.4. For the missing values regarding the city variable, we imputed the capital city of the corresponding county.

3.5. For the rest of the data imputation, a Knn imputation was presented as a good way of solving the issue, but this wasn’t able to be done because there were resulting ties. Afterwards, the number of neighbors taken into account was increased and still found to have ties. It was supposed it was due to the missing data in all the numerical variables, and this didn’t allow the data to separate correctly.

3.6. Afterwards, normal variables were looked at and the only one that was normal was the temperature. By randomly selecting variables from a normal distribution with the same mean and standard deviation, the imputation was done.

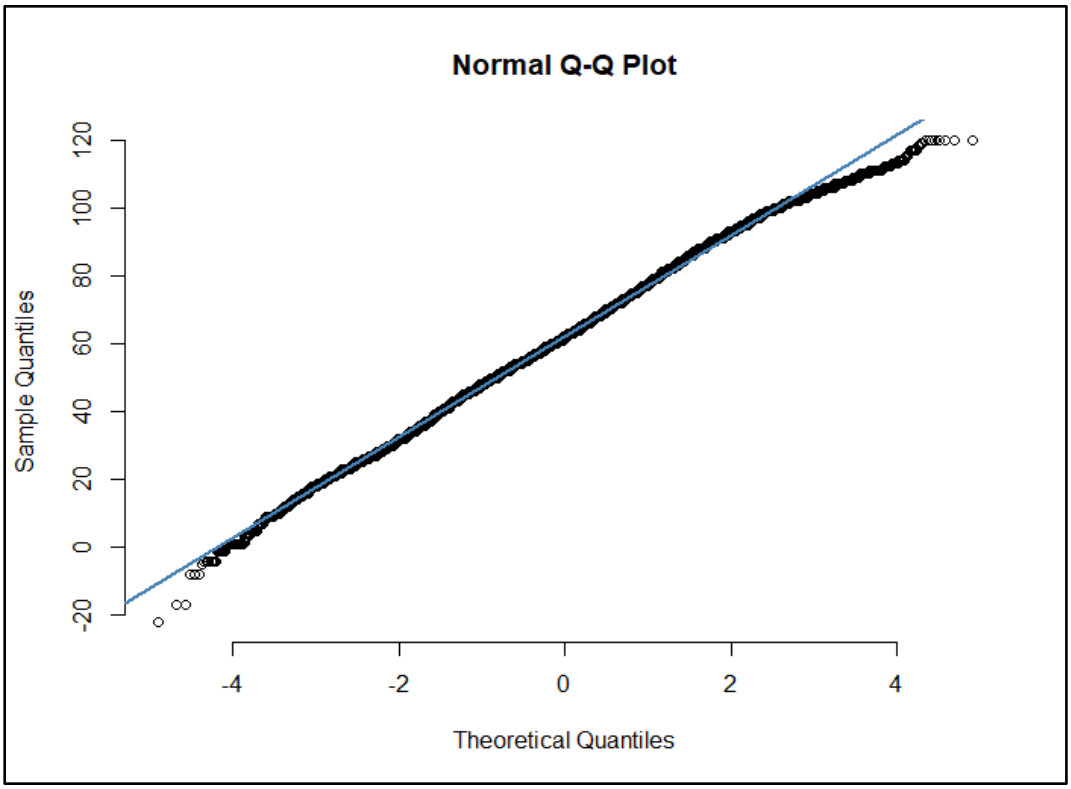

Figure 1: Normal Q-Q plot of the dataframe

The following graph (Fig. 1) shows the qqplot of the temperature, and although the data being skewed, it can be considered that it follows a normal distribution.

3.7. The next step was to check for the variables that have NA, these are pressure, visibility, humidity, precipitation and windchill. These variables were grouped

by its Weather condition, State and Season, and so the imputation for the missing values used the mean of the group.

3.8. Lastly, the remaining NA were eliminated due to the fact that the mean cannot be imputed, as there is no mean to be calculated, due to some groups having just NA.

To conclude the preprocessing steps, it must be mentioned that imputation has been tried by means of the mice function, but it couldn't be completed due to space allocation constraints.

Univariate and Bivariate Descriptive Statistics of Preprocessed Variables

In order to ensure good descriptive statistics we firstly observed correlation between numerical variables of our preprocessed data. We concluded that variables “Temperature” and “Wind_Chill” were highly correlated. Indeed, for large values of “Wind_Chill” we notice large values for “Temperature”, as well as that for small values of one we also detect small values of the other. In the preprocessed dataset we decided to exclude the variable “Wind_Chill”.

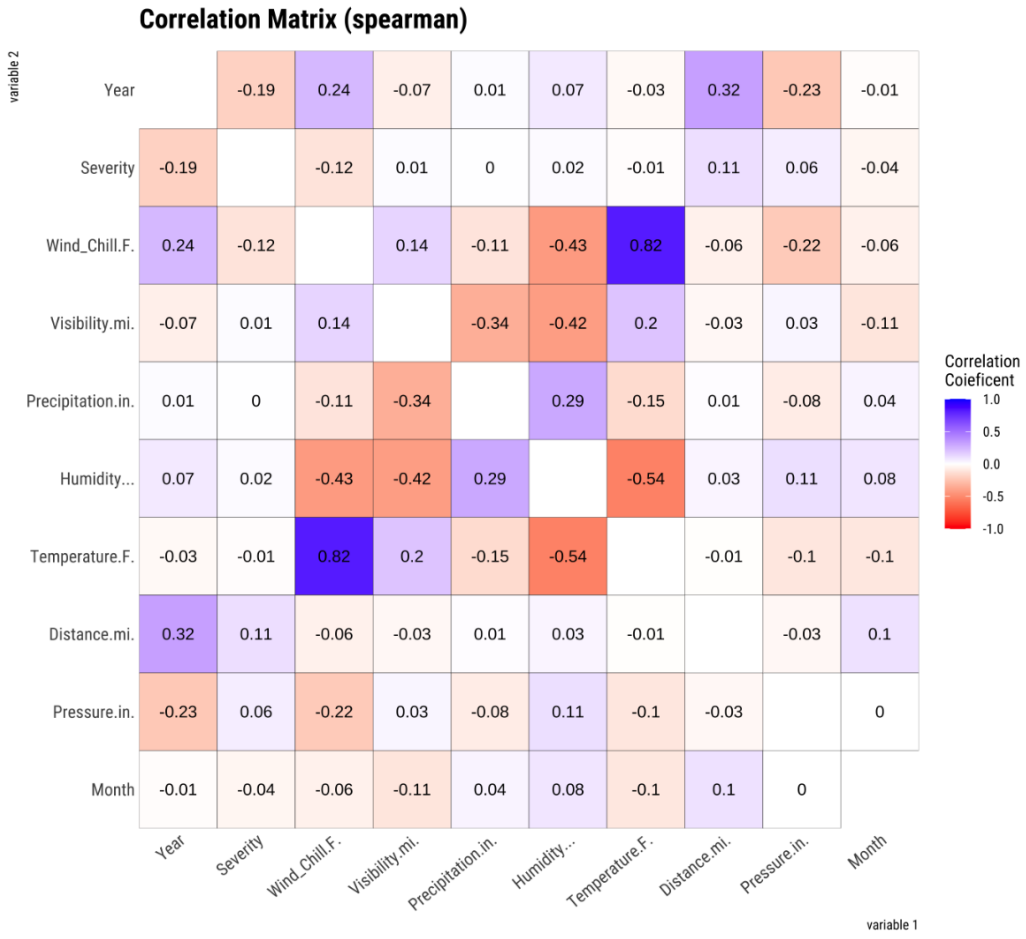

Figure 2: Correlation matrix between variables (method Spearman)

This graph (Fig. 2) shows the correlation using the Spearman’s rank correlation coefficient, where blue values indicate high positive correlation between “Temperature” and “Wind_Chill”.

Univariate Descriptive Analysis

To gain insight into preprocessed data compared to the original, summary() function was performed for both datasets. All NAs in variables from the original dataset were replaced by

imputed values so there are no missing values in preprocessed data. Another difference is that there are two additional variables Season and Month in the preprocessed dataset as well as removal of the Wind Chill variable because of above stated reasons.

For all quantitative variables, visualization of the distribution of each variable and the possible existence of outliers has been done by creating histograms and boxplots.

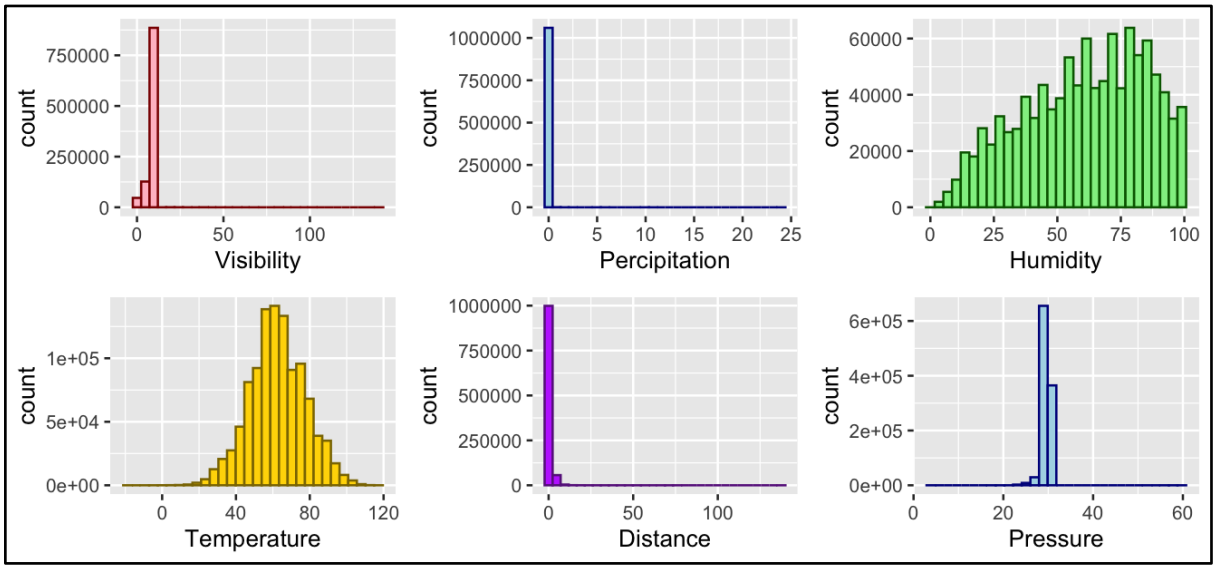

Figure 3:Distribution of numeric variables

Left-skewed distribution is observed for variable Humidity and right-skewed for variables Visibility, Precipitation and Distance. Temperature is following the normal distribution similar to variable Pressure.

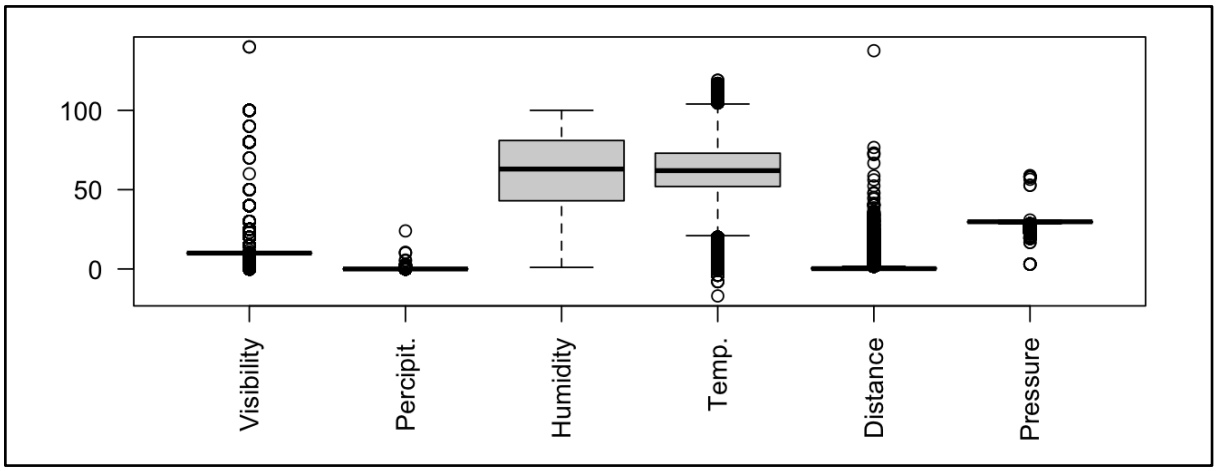

Figure 4: Boxplot for the distribution of the numeric variables

IQR range tends to be zero for variables Visibility, Precipitation, Distance and Pressure, so it is shown only for variables Humidity and Temperature.

There are no outliers detected for variable Humidity compared to all other numerical variables.

By using function boxplot.stats()$out we gathered outlier values of a particular variable, took the indexes of those within the observed variable and counted the number of those observations that we consider as potential outliers.

- Visibility = 242295
- Precipitation = 62406
- Humidity = 0
- Temperature = 3419
- Distance = 99332
- Pressure = 57222

Distribution of qualitative variables has been determined using barplots and pie charts.

The following barplots shows that out of all observed accidents close to 50% happened during Fair Weather Condition, as well that most two frequent seasons in which accidents occur are Winter and Autumn.

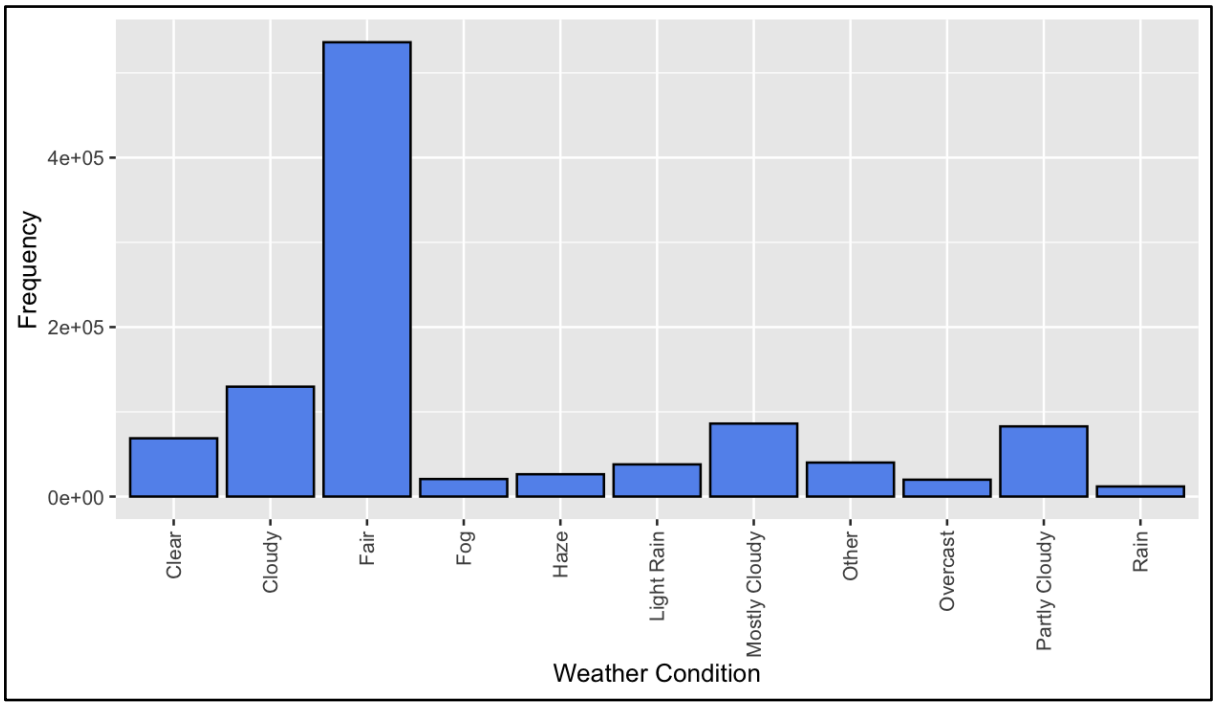

Figure 5: Barplots of the accidents depending on the weather condition

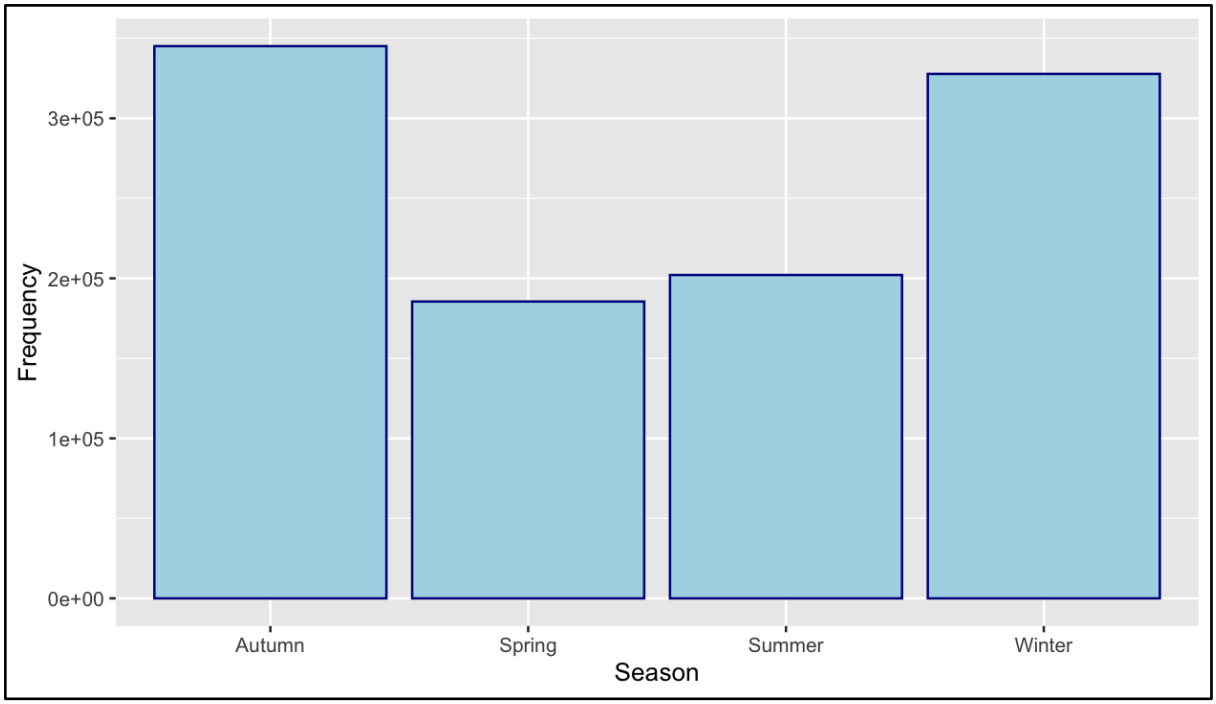

Figure 6: Barplots of the accidents distributed according to the season

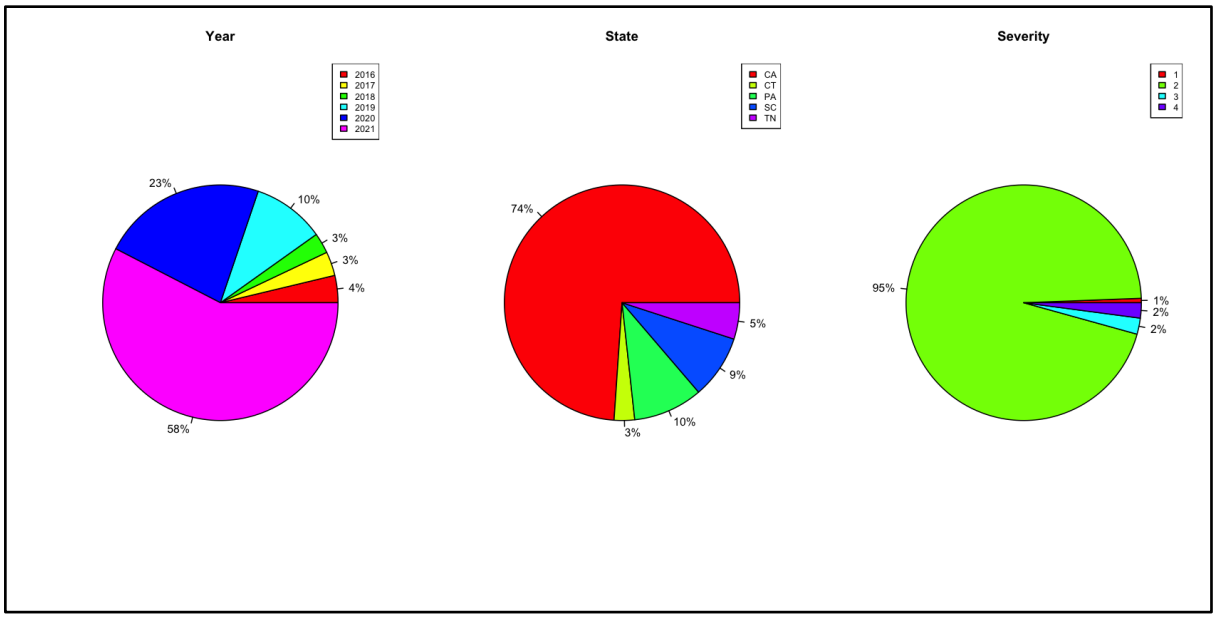

Figure 7: Percentage of the accidents according to the year, state and severity

Pie charts describing distribution of categorical variables (Fig. 7) point out that 2021 is the year with the highest number of recorded accidents, in the state of California happened majority of recorded accidents compared to other 4 observed states, as well as that most accidents are described with severity level 2.

For qualitative variables with a high number of levels only most frequent 5 are taken into consideration for visualizing distribution. Out of all observations, 54% are recorded in county Los Angeles and 40% in the City of Los Angeles.

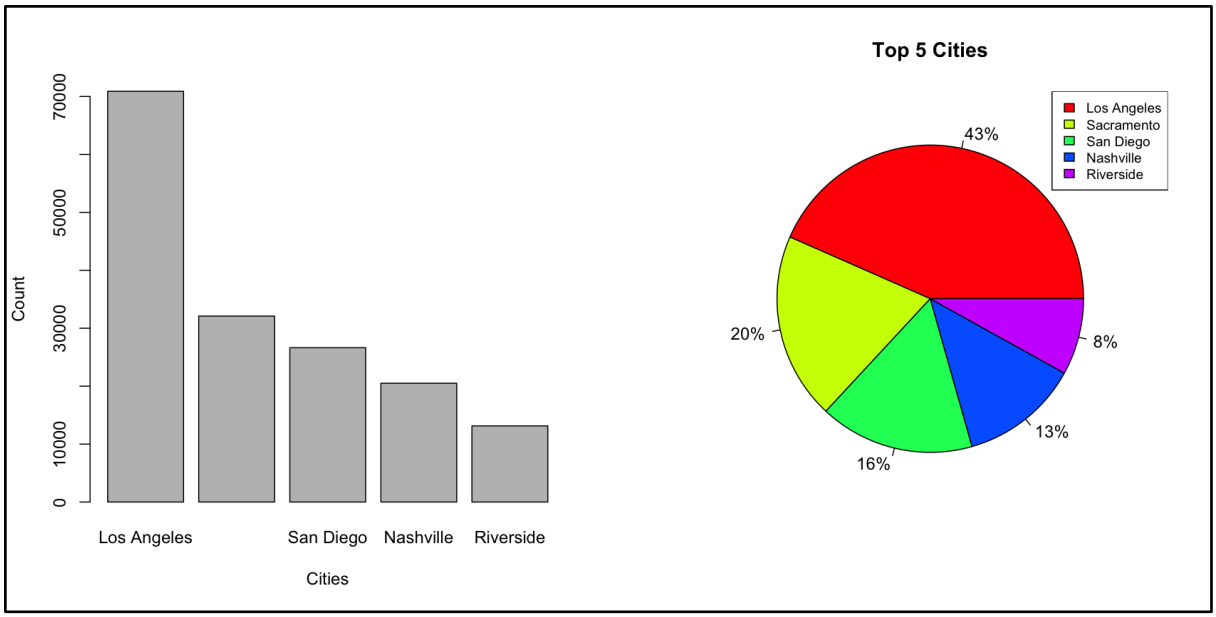

Figure 8: Top 5 cities with the most accidents

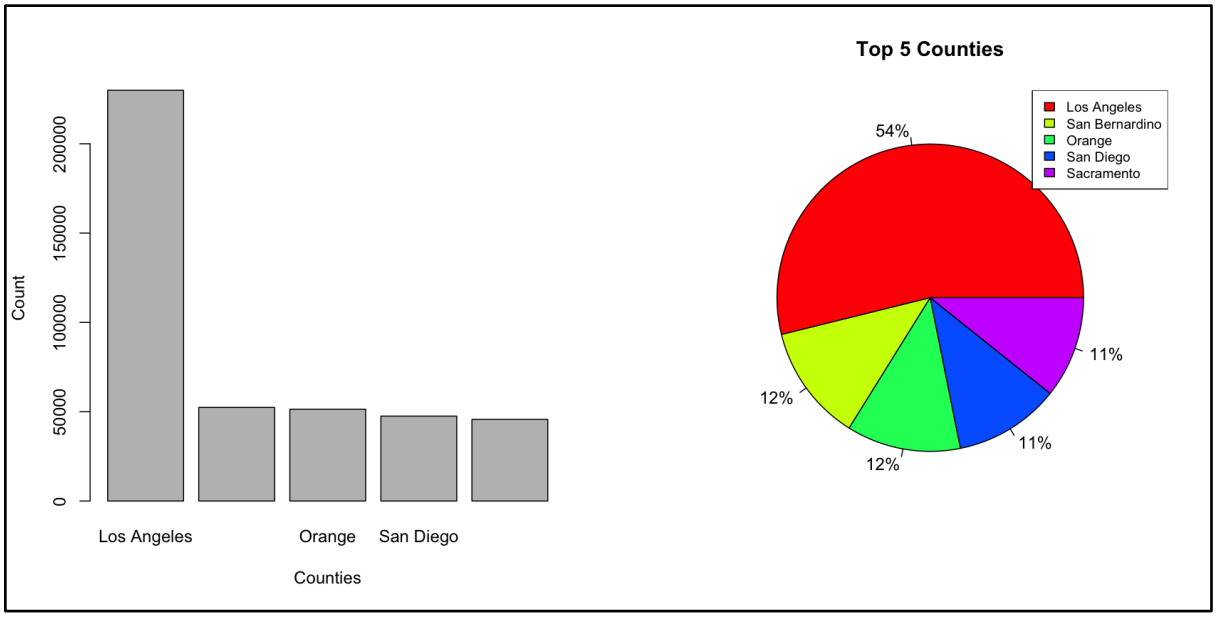

Figure 9: Top 5 counties with the most accidents

Bivariate Descriptive Analysis

The analysis of the variables of our dataset is an important step in the understanding of it. Equally important is to analyze two or more variables at the same time, in order to examine their relationship and the depth of this relationship. This type of analysis is made to determine the discrepancies between the variables analyzed and to find any cause of this difference.

The first step in the bivariate descriptive analysis, and complementing the outlier detection made for each numerical variable previously described, is to make the multivariate outlier detection in order to find unusual scores or numbers on two variables or more. The multivariate outlier detection was made using the Mahalanobis distance from the center of the data. The following plot (Fig. 10) shows the individuals that are far from the center of the data, thus showing them as outliers.

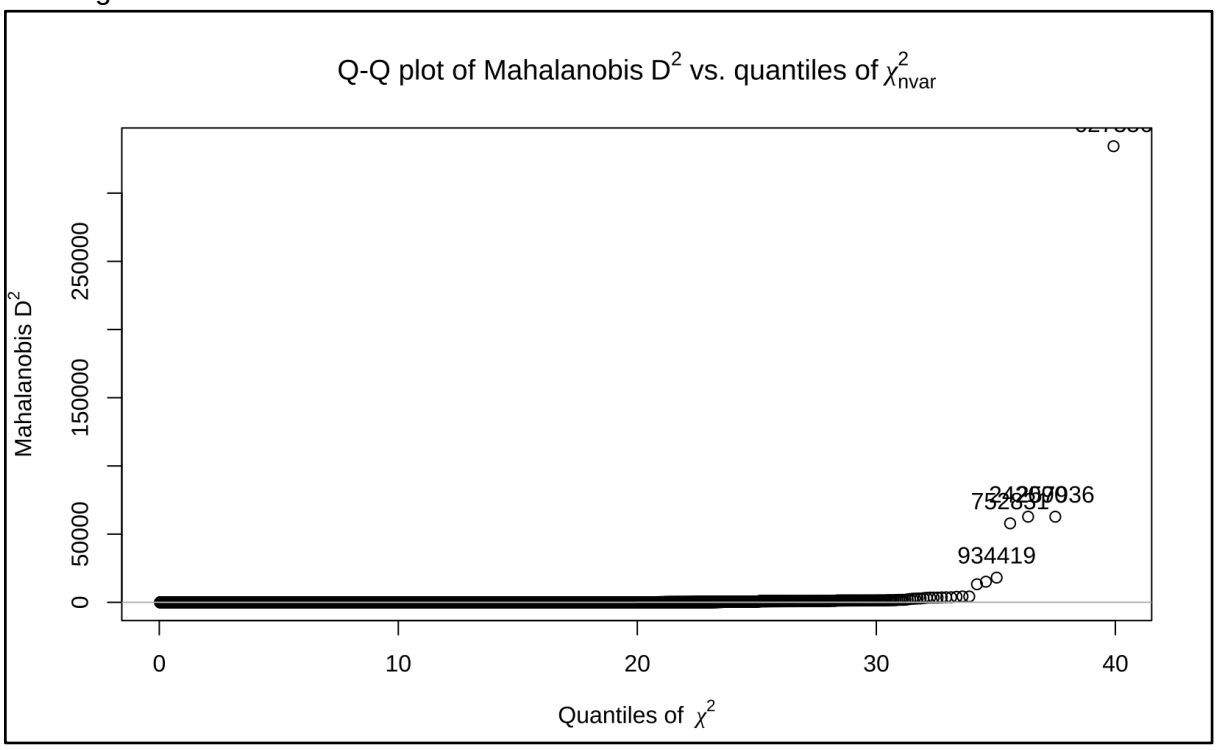

Figure 10: Multivariate outlier detection Q-Q plot.

The number of multivariate outliers detected for our dataset was 40272 multivariate outliers.

The second step was to make some comparisons regarding the type of variables used. The three comparisons made were using two numerical variables, one numerical variable and one categorical variable and finally two categorical variables.

1. Two numerical variables:

Following scatterplot (Fig. 11) shows the relation between numerical variables Distance and Visibility. Length of the road extent affected by the accident is higher in conditions of low visibility. Most recorded accidents occur in visibility conditions lower than 25 miles and affect road distance up to 25 miles.

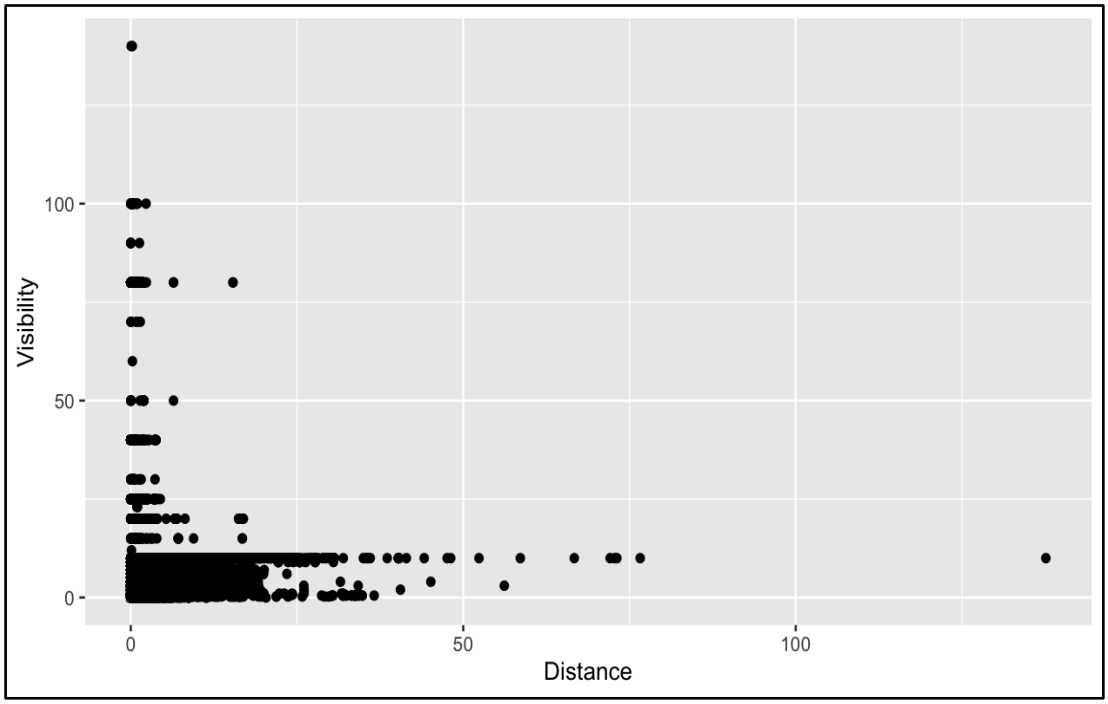

Figure 11: Scatterploot of relationship between distance (numeric) and visibility (numeric)

One numerical and one categorical variables:

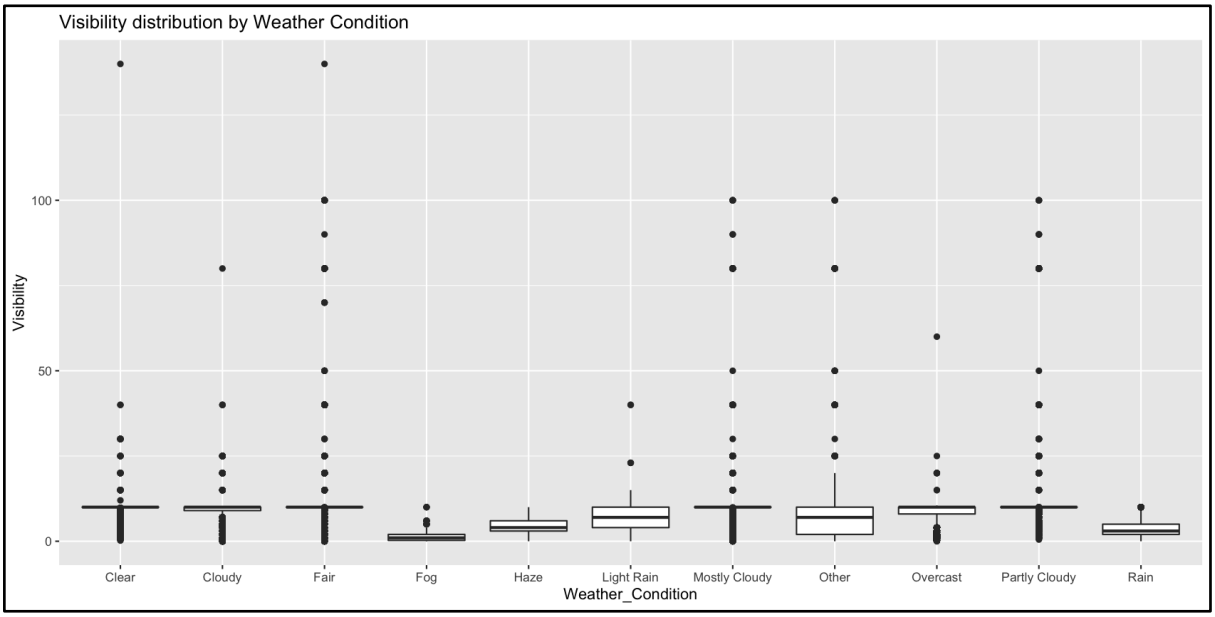

Figure 12: Scatterplot of relationship between visibility (numeric) and weather conditions (categorical)

Multiple boxplots (Fig. 12) are used for visualizing the relationship between Weather Condition and Visibility. It is noticed that observations with weather conditions different from Fog, Rain and Haze have almost the same mean visibility value, whereas we observe specifically smaller mean values for condtions Fog, Rain and Haze.

2. Two categorical variables:
   The following plots (Fig. 13) show the comparison between the number of accidents that happened for each severity (on a scale of 1 to 4) and out of the accidents of each severity, the number that happened with a bump, a stop, a crossing and a traffic signal present, respectively.

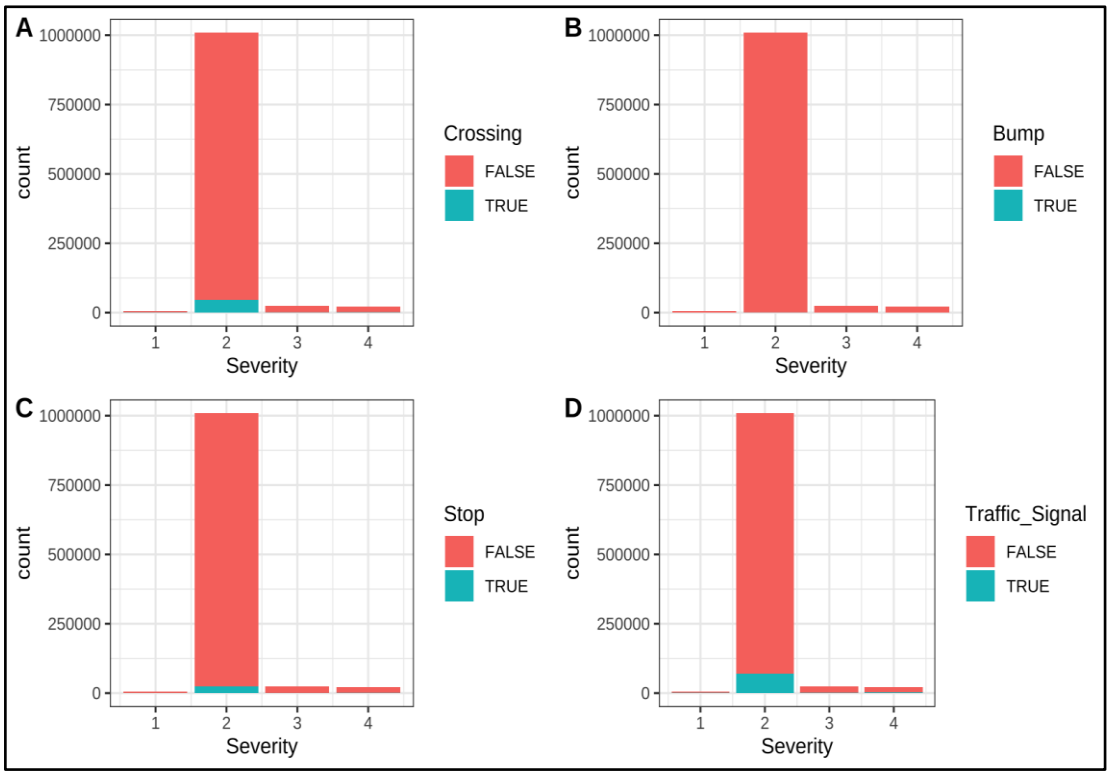

Figure 13: Distribution of accidents according to the presence of traffic element

These plots show a high number of accidents that happened in the United States to be of severity 2 and the majority of accidents happened without each one of the previously described traffic effects. Another interesting point is that there are very few accidents where road elements are present, and as normal, the majority is concentrated in the accidents of severity 2.

Finally, the last step in the comparison of variables is the correlation between the “Severity” variable and all the numerical variables in our dataset. The following plot (Fig. 14) presents the Spearman’s rank correlation coefficient comparing “Severity” with each numerical variable in our dataset.

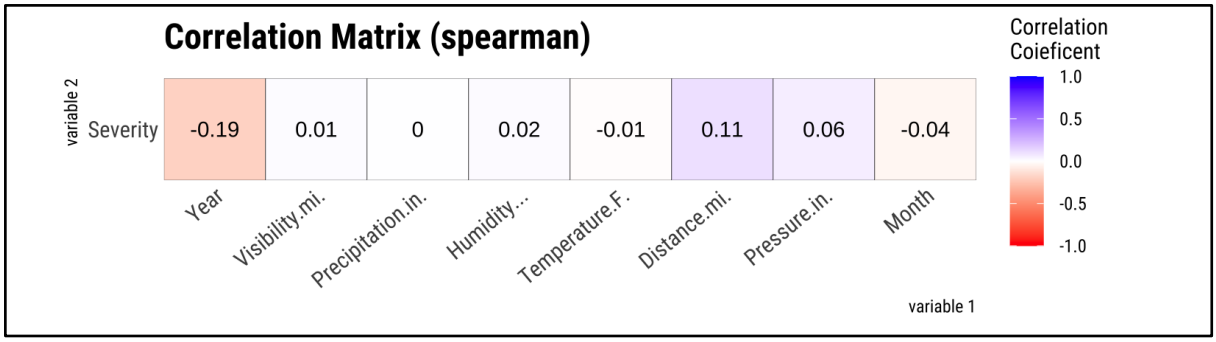

Figure 14: Correlation Matrix (spearman) between severity and the variables depicted above

The result of the correlation shows that the severity variable, that is the one that we are interested in, has weak correlation with “Year” and “Distance”. For the other variables there is no correlation between the variables analyzed.

Conclusions of the Univariate and Bivariate Statistics

From univariate and bivariate descriptive analysis performed, it can be concluded that most of the recorded accidents are highly linked with deficiency of control indicators such as bump, traffic light, stop sign or crossing.

Because these signals are not present, it can be assumed that velocity and whether factors are the main influences for the presence of an accident. However univariate analysis for variable Weather_Condition clearly showed that most accidents occur in Fair weather conditions, so we exclude this from the previous assumption.

Autumn and Winter are seasons with most recorded accidents and what is also noticed is that throughout years recorded, accidents tend to happen more and more often, but low negative correlation coefficient between variables Severity and Year shows that level of severity slightly tends to decrease in more recent years. Out of five states chosen for this analysis, the State of California is the state in which 74% of recorded US accidents occur.

Principal Component Analysis - PCA

To begin with, the dataset which was produced after the completion of the preprocessing and univariate and bivariate analysis steps contained 6 numerical variables in total. Those variables are named as follows: “Temperature(F)”, “Humidity(%)”, “Pressure(in)”, “Precipitation(in)”, “Distance(mi)” and “Visibility(mi)”. After applying the PCA algorithm from the FactoMineR[1] library we concluded that the number of principal components need to be selected is 4. The reason behind this decision is that the cumulative sum of the variance percentage reaches more than 80% with the first four components. Whatsmore, this assumption can be seen as well in the cumulative sum of variance plot and the scree plot shown in figures 15 and 16.

[1] Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*. **25(1)**. pp. 1-18.

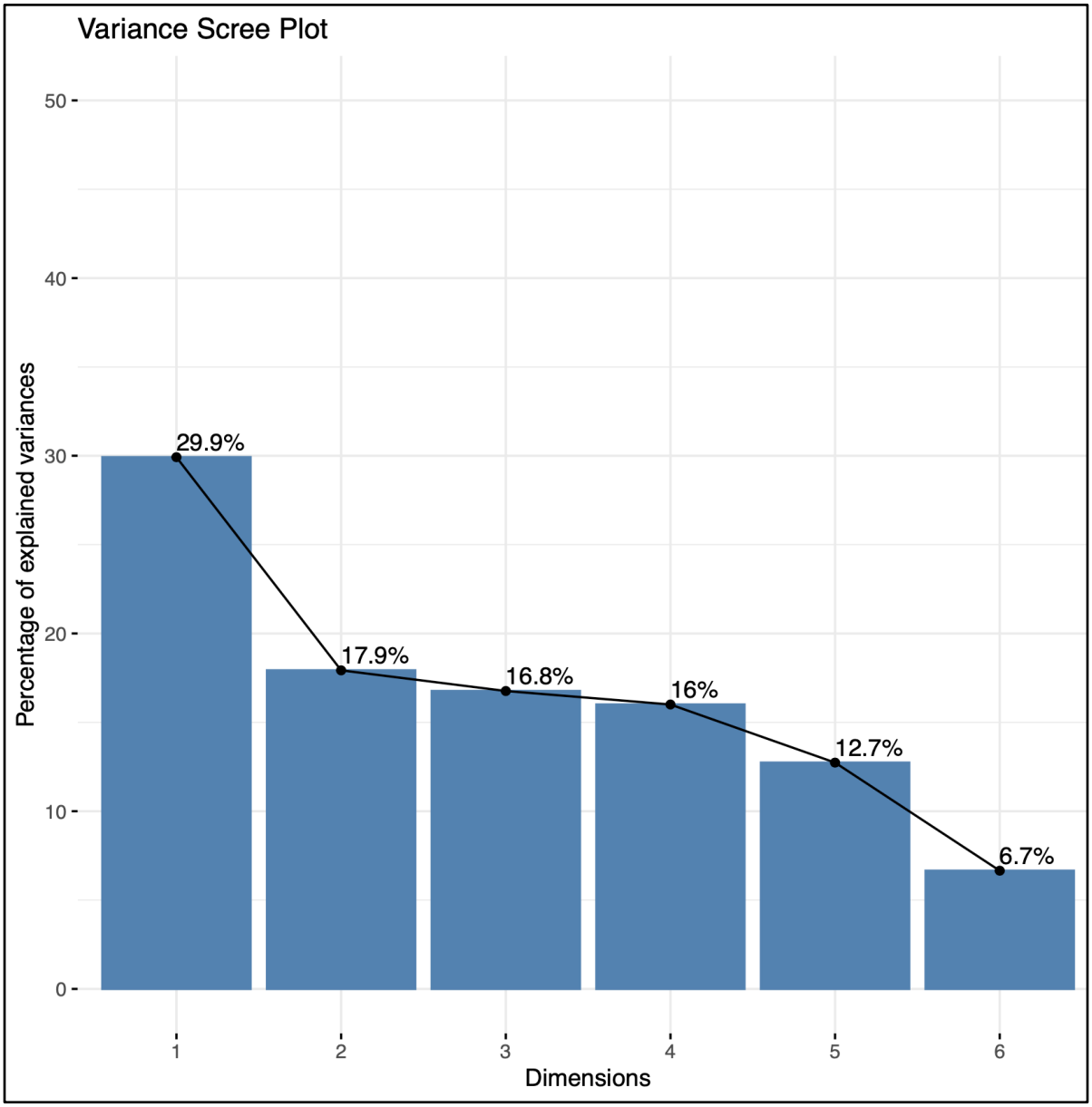

Figure 15: Scree plot for the percentage of explained variance for the Principal Components

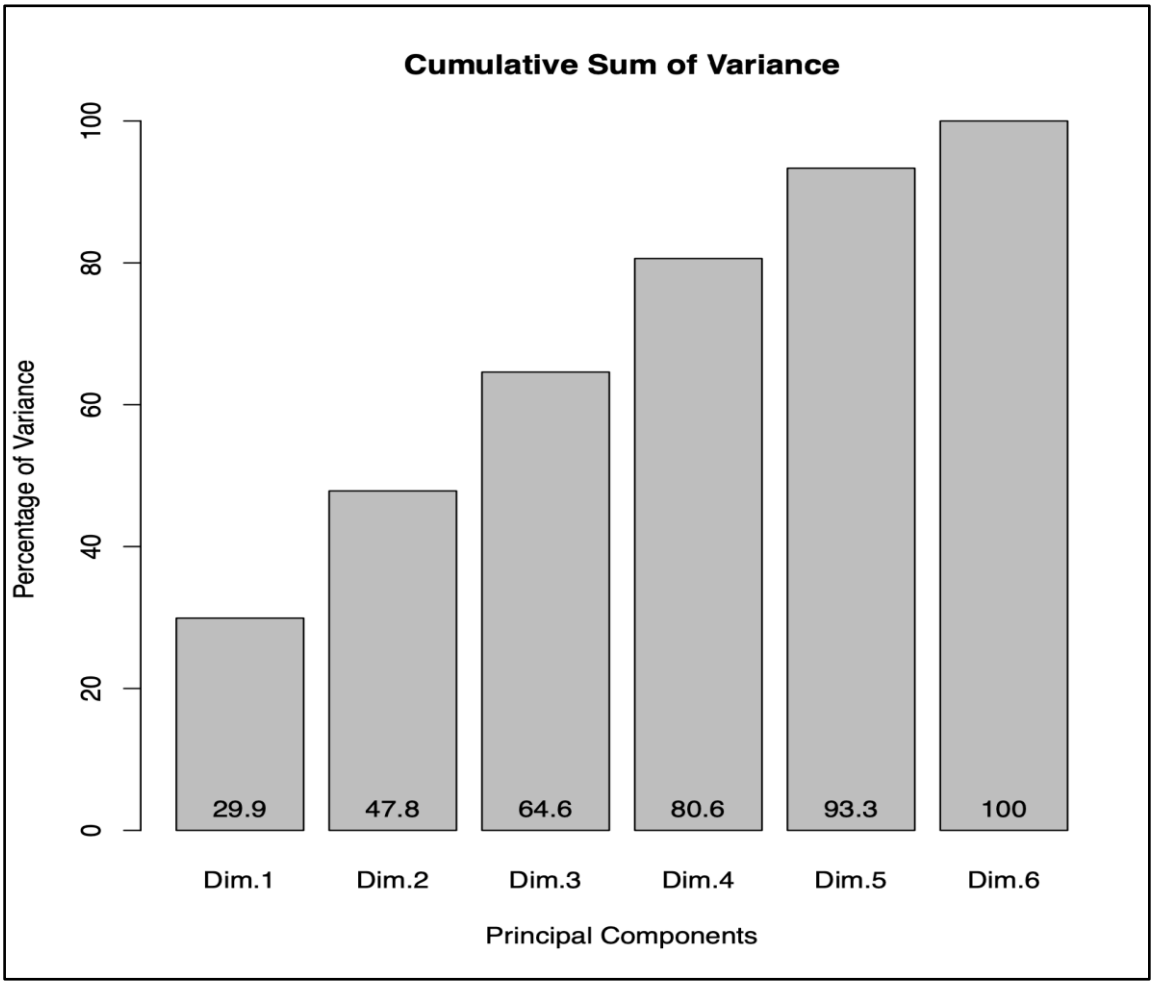

Figure 16: Barplot depicting the cumulative sum of variance for the Principal Components.

Factorial Maps

In this section factorial map visualizations occur containing graphs for: individual projections, common projection of the numerical variables and the modalities of the target qualitative variable (Severity) and finally, interpretation of relationships among observed variables. In some specific cases - when it is possible - interpretation of the latent variable associated with the principal axis takes place as well.

In the first of the following graphs (Fig. 17), the individual projections based on the target variable (Severity) are illustrated. It is clear that the Principal Component 1 (PC1) and the Principal Component 2 (PC2) were used as the axis of the projection, which describe the 47.8% of the total inertia of the numerical variables of the dataset. In addition, each distinct individual projection is colored based on its “contrib” value, which describes the contribution of an individual to the corresponding dimensions. For this specific graph, due to the large number of observations in the dataset (1m. approximately), in order to get a clearer picture of

the individual projections we only plotted the first 100.000 observations (10% of the total data) with the highest contribution and also zoomed it to have a more clear vision about it. From this plot can be noticed that projections are all close to the center of the axes, which means that they are mostly correlated with each other, so we do not have a high percentage of variance explained in those two axes, because groups are united and it's very difficult to separate them and make conclusions. Same thing is happening when projecting the individuals to other combinations of Principal Components (like PC1 and PC3, plots can be found in Appendix).

The second plot (Fig. 18) of this section describes the interpretation of relationships among the variables observed and the PCs. Once again, PC1 and PC2 have been the selected axes for the projection. From this graph it can be deduced that variables whose correlation is positive are grouped together and that variables whose correlation value is negative are positioned on the opposite side of the graph (opposite quadrants). The quality of the variables within the factor map is measured by the distance and the angle between them and the origin. On the one hand, the length of the vector of a variable describes its contribution to the axes. On the other hand, the angle of the vector of a variable describes the correlation between this variable and the other variables or the axes. After checking the correlation between every variable with the other ones, and between every variable and the axes, the results show that “Humidity(%)” and “Temperature(F)” have the highest correlation with PC1. Whatsmore, as these two variables are negatively correlated, a high value for “Temperature(F)” means that a low value for “Humidity(%)” will be represented. The same analysis can be made with “Pressure(in)” and “Distance(mi)”. For PC2, “Pressure(in)” is the variable with more correlation and “Precipitation(in)” is the one for PC3.

In the third plot (Fig. 19) the common projection of the numerical variables and the modalities of the qualitative target variable (Severity) is depicted. Here the PC1 and the PC2 were used as the axes of projection. Moreover, the observations were grouped based on their value of severity. The groups are represented by the colored ellipses of different colors. Furthermore, the contribution of every origin variable is depicted by the blue arrows of the plot. The length and the angle of the vectors describe their contribution and their relationship with the PCs respectively.

We can see here the combination of the two graphs previously presented, where individuals are located close to the center of the axes, when at the same time, the contribution of every variable can be seen by the length and the angles of the axes. This graph has not been zoomed because the variables could not be seen. However, a better vision is presented in Fig. 17.

This biplot can lead us to the same conclusions as before, indicating that PCA is not so insightful for the reduction of the dimensions of numerical variables for describing the target variable.

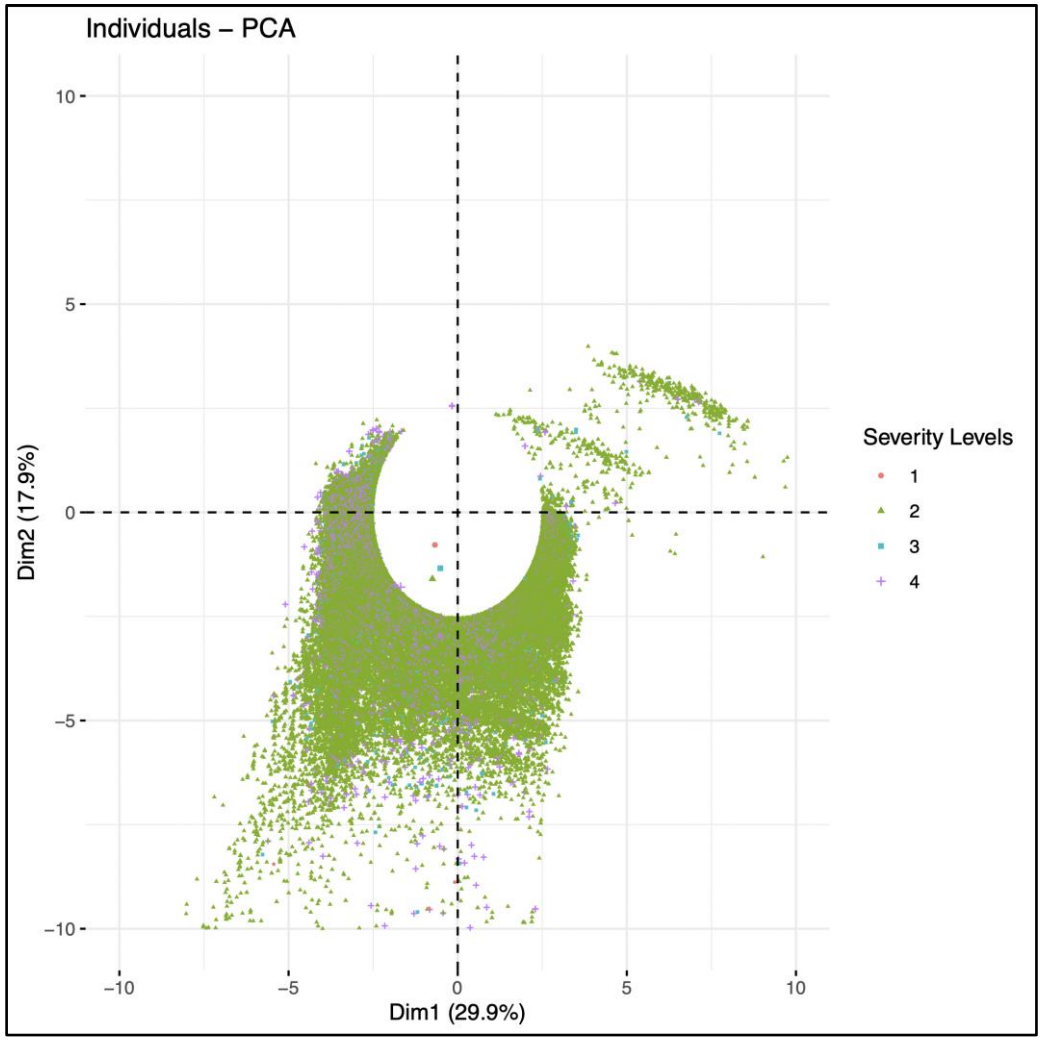

Figure 17: Individuals projections with coloring based on Severity,

taking a sample of 100000 units due to a large number of observations in dimensions 1 and 2 (zoomed)

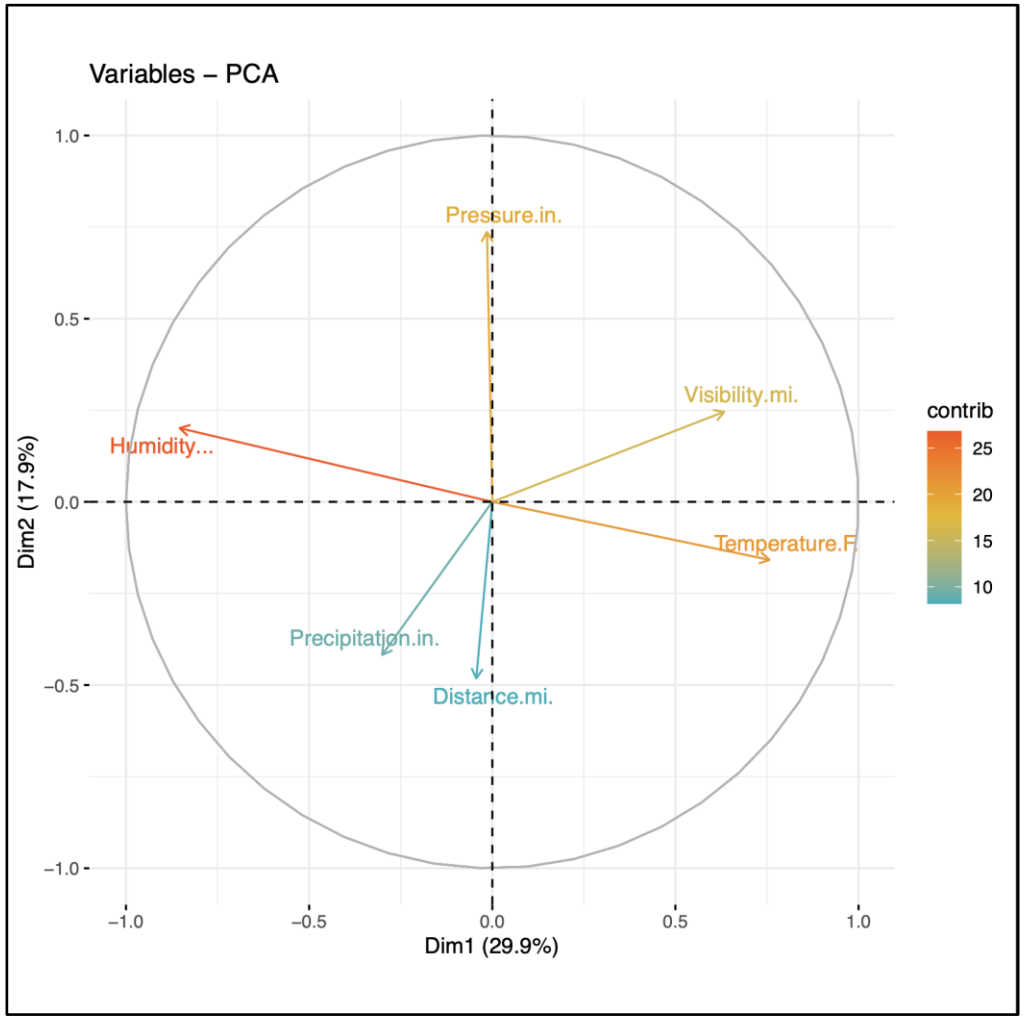

Figure 18: Variables factor graph in dimensions 1 and 2

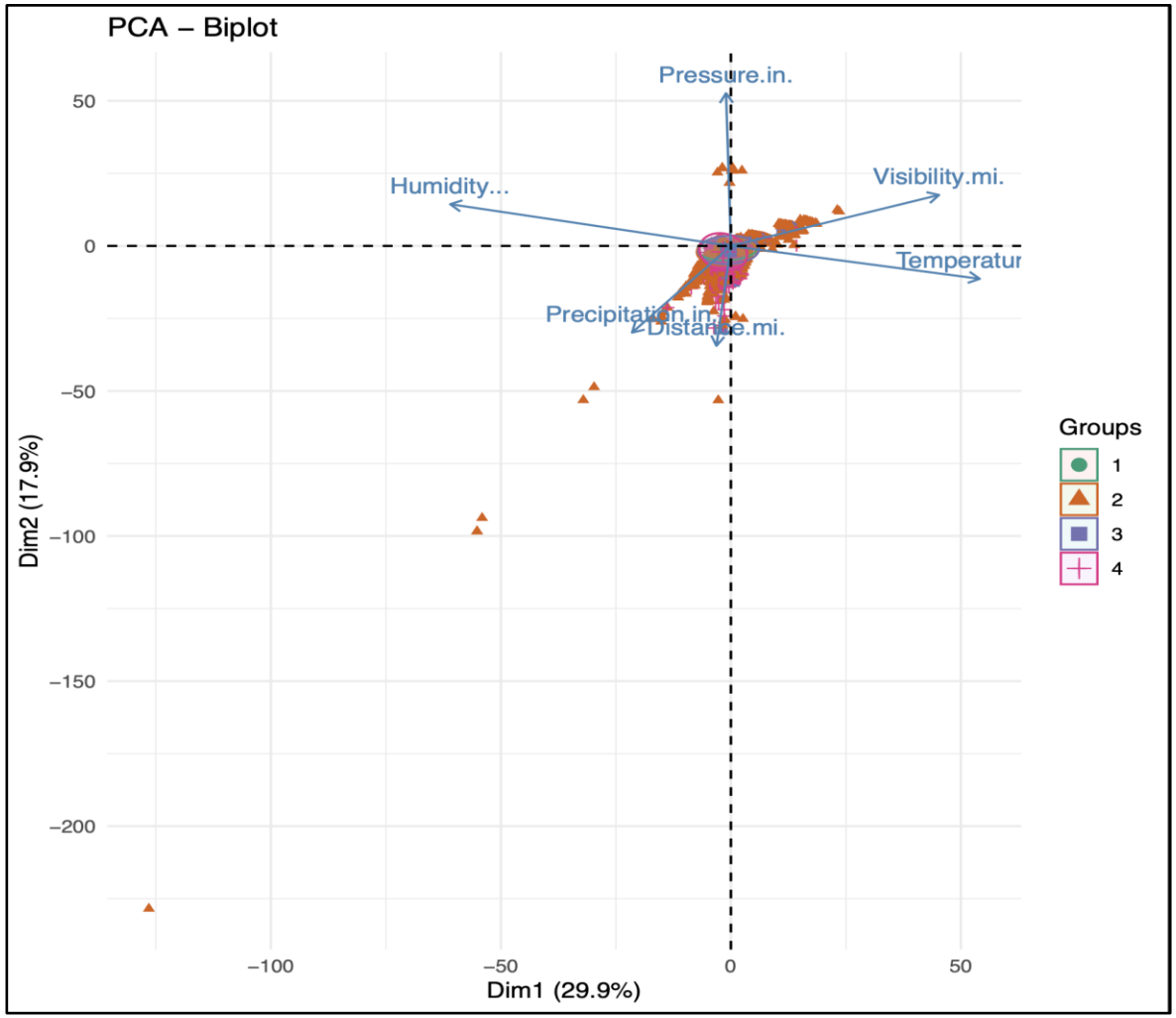

Figure 19: Biplots Containing Individuals and Variable Factor Maps

colored and grouped based on Severity levels in dimensions 1 and 2.

Conclusions on PCA

The number of principal components selected has been 4. The reason behind this decision is that the cumulative sum of the inertia percentage reaches more than 80% with the first four components.

The general conclusions that can be drawn after analyzing the results and interpreting the graphs dictate that the PCA is not very valid in providing a good overview.

This makes a lot of sense since out of the 6 numerical variables that have been used for the analysis, 5 of them are related to weather conditions. So their combination in our case is not so good to describe the severity level of the accident. Better results are achieved when performing MFA.

Multiple Correspondence Analysis - MCA

The dataset obtained from the preprocessing and descriptive statistics analysis steps, had 9 categorical variables including: “State”, “Year”, “Weather_Condition”, “Crossing”, “Bump”, “Stop”, “Traffic_Signal”, “Season” and “Severity”. The variable regarding the severity of the accident was used as a supplementary qualitative variable.

The first step was to apply MCA to our dataset of only categorical variables. After applying the technique, we decided to analyze the results using a Logic Table, meaning that the number of components that are selected is given by the following formula: 1/p, and p being the number of active categorical variables. The number of active categorical variables in our dataset was 8, so dimensions having eigenvalues greater than 1/8 = 0.125 were selected. The number of selected dimensions was 12, explaining 55.1% of the total variance of the data, but the main analysis was performed using dimension 1 and dimension 2 as they have the highest percentage of explained variances, and in combination they only explain 12,8% of the total variance of our data. In addition we can state that as we have a low percentage of explained variance of our data, the conclusions made from this analysis cannot be taken as a representation of all the dataset.

The Scree plot (Fig. 20) shows for each dimension the percentage of explained variances and the Cumulative Sum of Variance plot shows the percentage of data explained as the number of dimensions increases (Fig. 21).

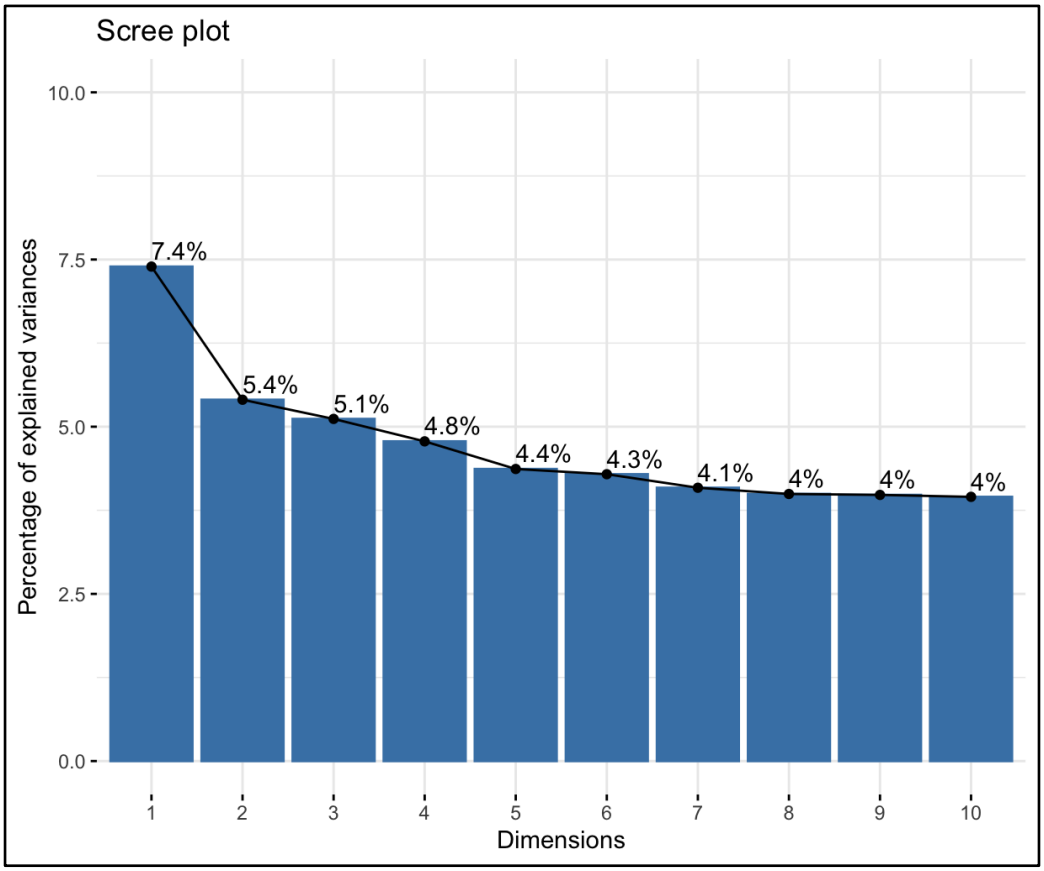

Figure 20: Scree plot for the percentage of explained variances for each dimension

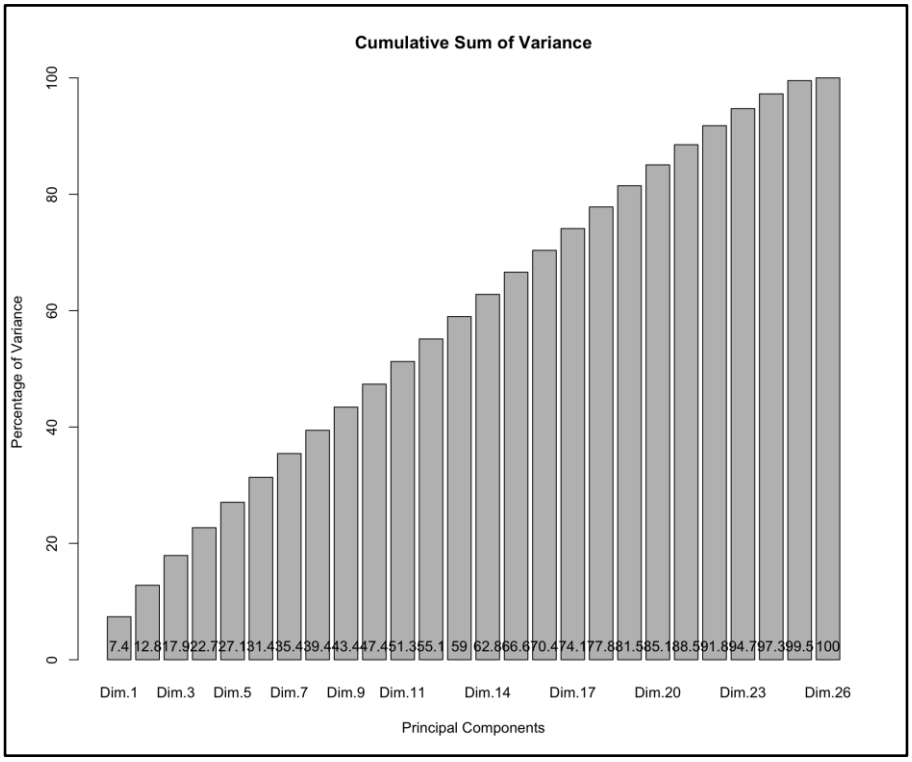

Figure 21: The percentage of data explained as the number of dimensions increases

The first plot (Variables - MCA) shows the correlation between categorical variables and MCA principal dimensions. This plot displays the coordinates of each variable and how close they are from the origin. The conclusion that we can take from this plot is that, on one hand, variables “Year” and “Weather_Condition” are the most correlated with dimension 1. On the other hand variables “Crossing” and “Traffic” signals are the most correlated with dimension 2.

The second plot (Fig. 22) shows the correlation between variable categories and the first two dimensions. From this type of plots we can extract information about the groups of categories, the correlation between categories and the quality of the variable category on the factor map. This plot shows that two groups are made in total. For dimension 1 we have a group formed by “Year” variable categories (2016, 2017 and 2018) and “Weather_Condition” variable categories (clear and Overcast). For dimension two, there is one group and this group combines variable categories regarding “Crossing” and “Traffic_Signal”, we select these variable categories in the same group because they are the representation of the presence of road elements. The variable category “Bump_TRUE” cannot be included in this group as in the following analysis, it is shown that there is no contribution of this variable category for dimension 2. Labeling dimension 1 as weather condition without precipitation and the year and for dimension 2 the appearance of road elements. Another analysis that can be made is the correlation between variable categories of the same variable. Unfortunately this plot (Fig. 23) shows that the previously explained categories do not have negative correlation with their complementary variables, because all other variable categories are grouped together around the origin. The main conclusion to this is that there is a weak correlation between the variable categories that are far from the origin with those that are close to it. In addition, variable categories being far from the origin mean that they are well represented on the factor map.

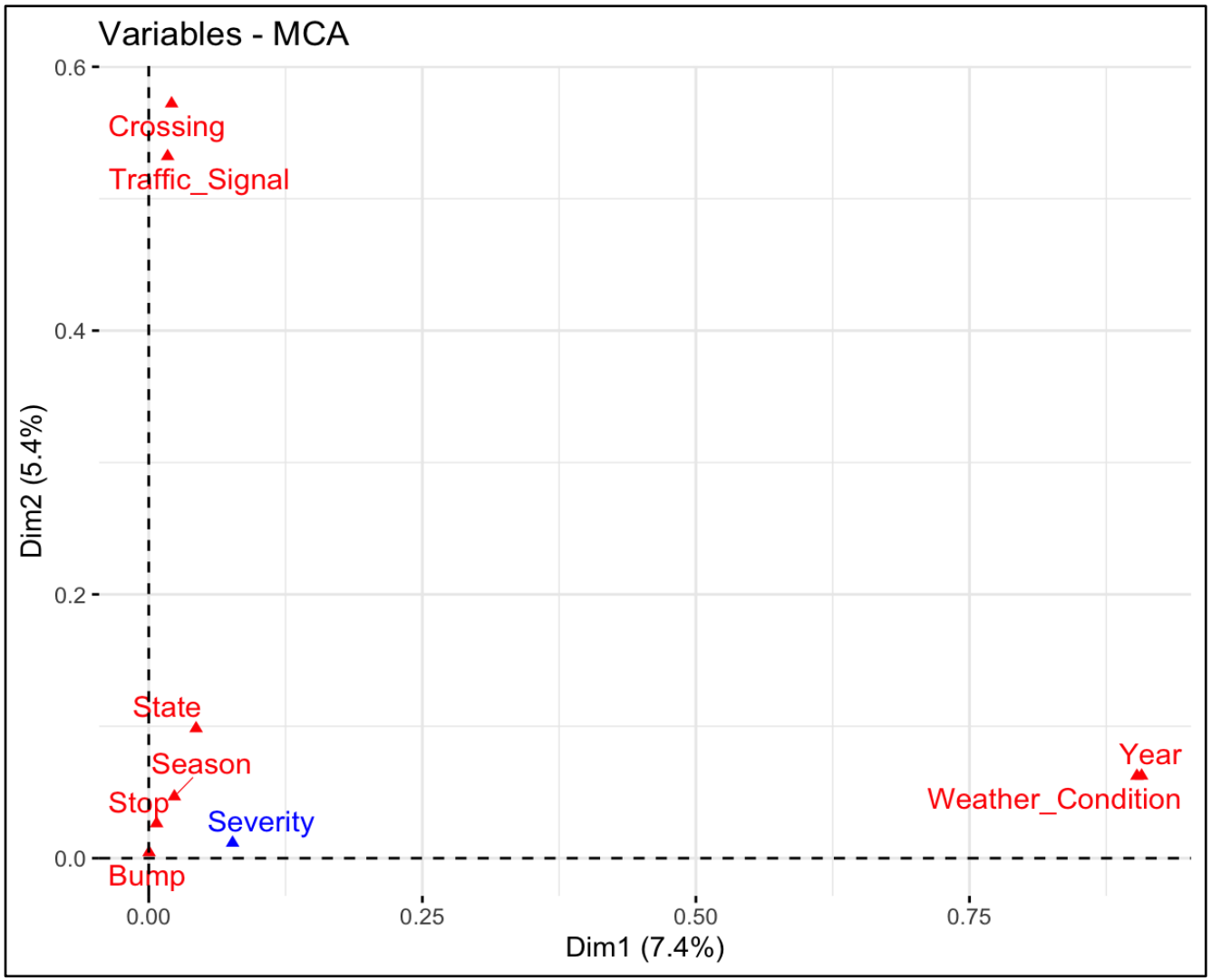

Figure 22: Plot of correlation between categorical variables and MCA principal dimensions

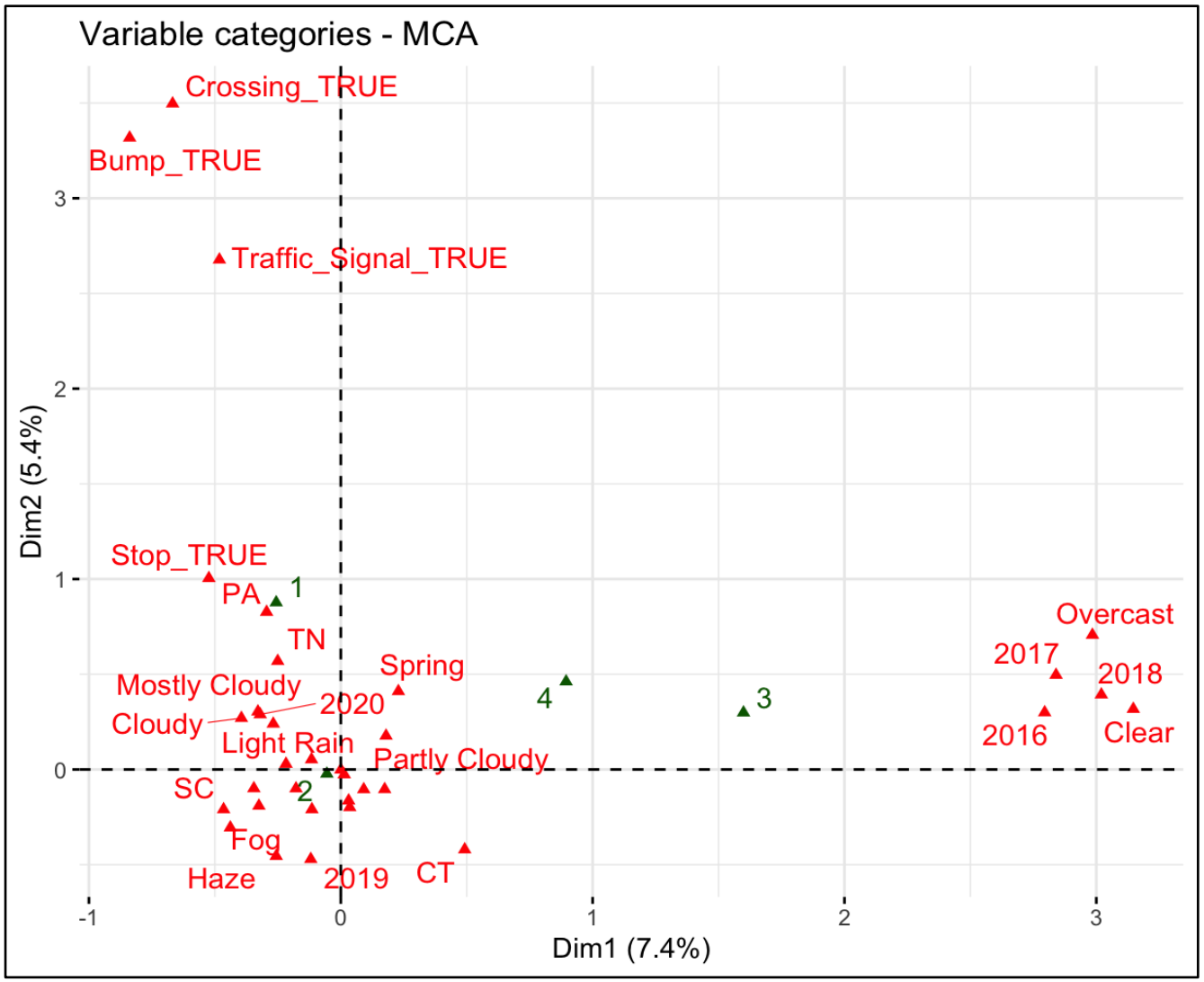

Figure 23: Plot of correlation between variable categories and the first two dimensions

The following four graphs (Fig. 24, Fig. 25, Fig. 26, Fig. 27) represent the contribution of variable categories for each dimension. For dimension 1 we have that the categories that contribute the most are “Clear”, “2016”, “2017”, “2018” and “Overcast”. For dimension 2, the categories that contribute the most are “Crossing_TRUE”, “Traffic_Signal_TRUE” and “PA”. The next plot shows that when looking at the contribution for the categories of our data, we have that the ones that contribute the most to both dimensions are “Clear”, “Crossing_TRUE”, “Traffic_Signal_TRUE”, “2016”, “2017”, “2018” and “Overcast”. The last plot of the four (Fig. 26), shows the level of contribution of each variable category when plotted for the first two dimensions.

The final plot (Fig. 28) shows the correlation between variable categories and the first two dimensions and in addition there is a 1% of individuals of our data represented in blue. With this plot, the main conclusion is that as well as with variable categories, the closer they are from other individuals the more similar they are. In this way, there are three formed groups for our subset of the data.

Applying MCA and analyzing the results for our dataset has been a challenge regarding the visualization of the last plot. Because of the large number of individuals that our data has, we had to only provide the visualization of a 1% of our total data (approximately 100000 individuals). This has been the main drawback regarding the MCA analysis.

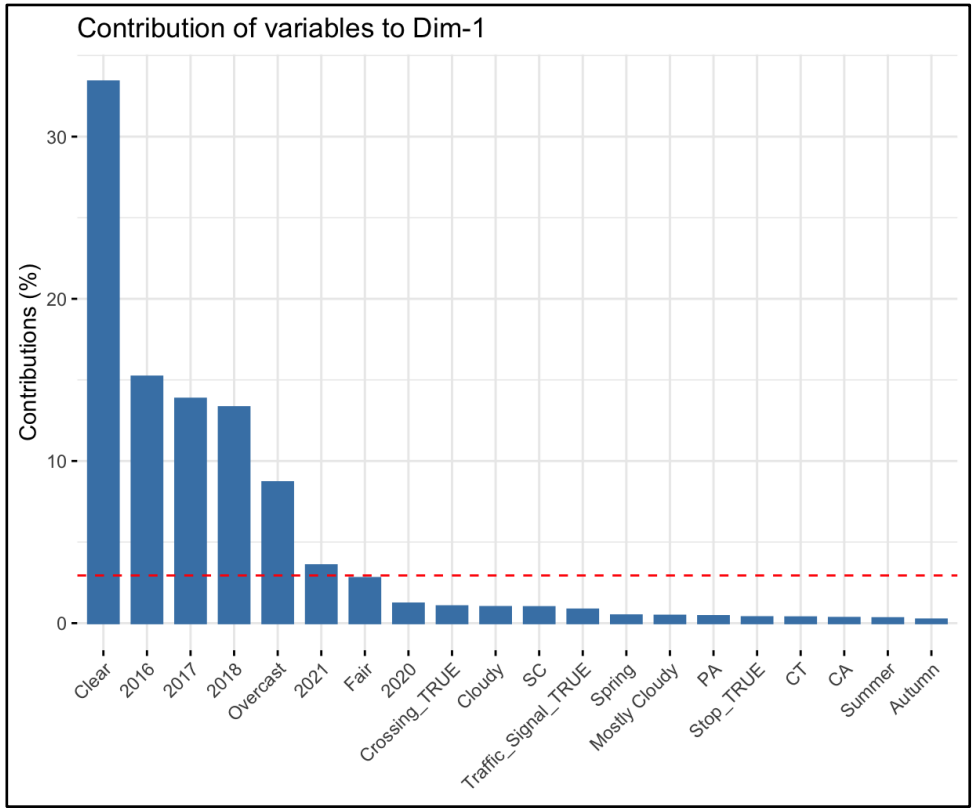

Figure 24: Contribution of variable categories for Dim1

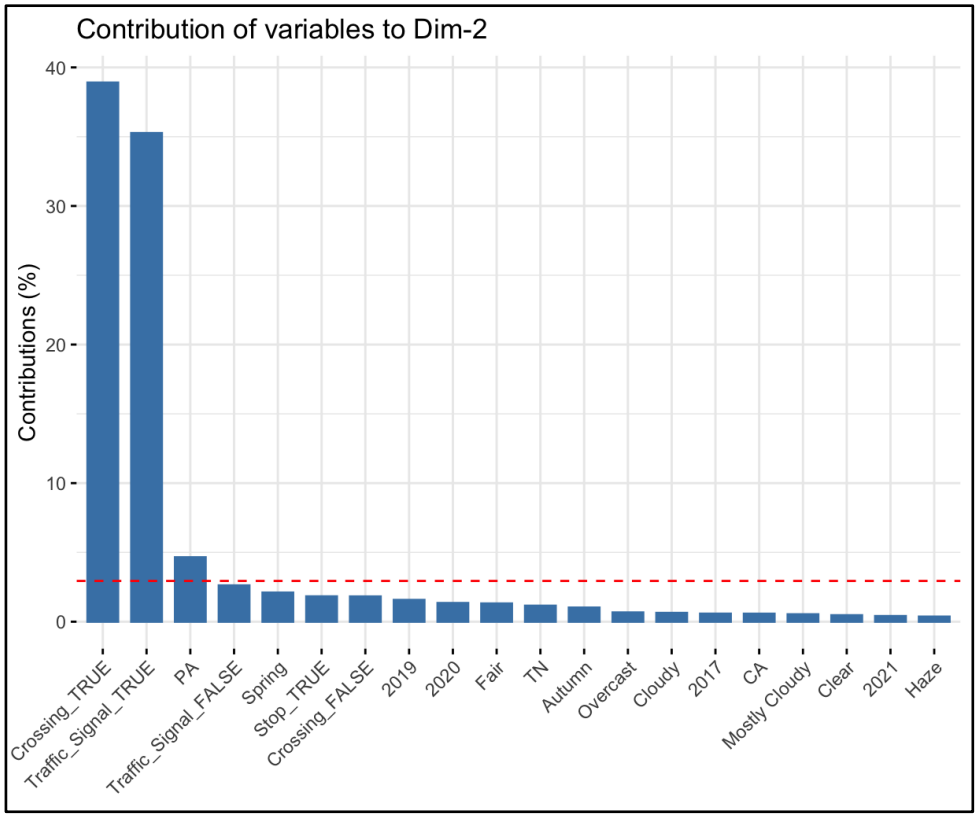

Figure 25: Contribution of variable categories for Dim2

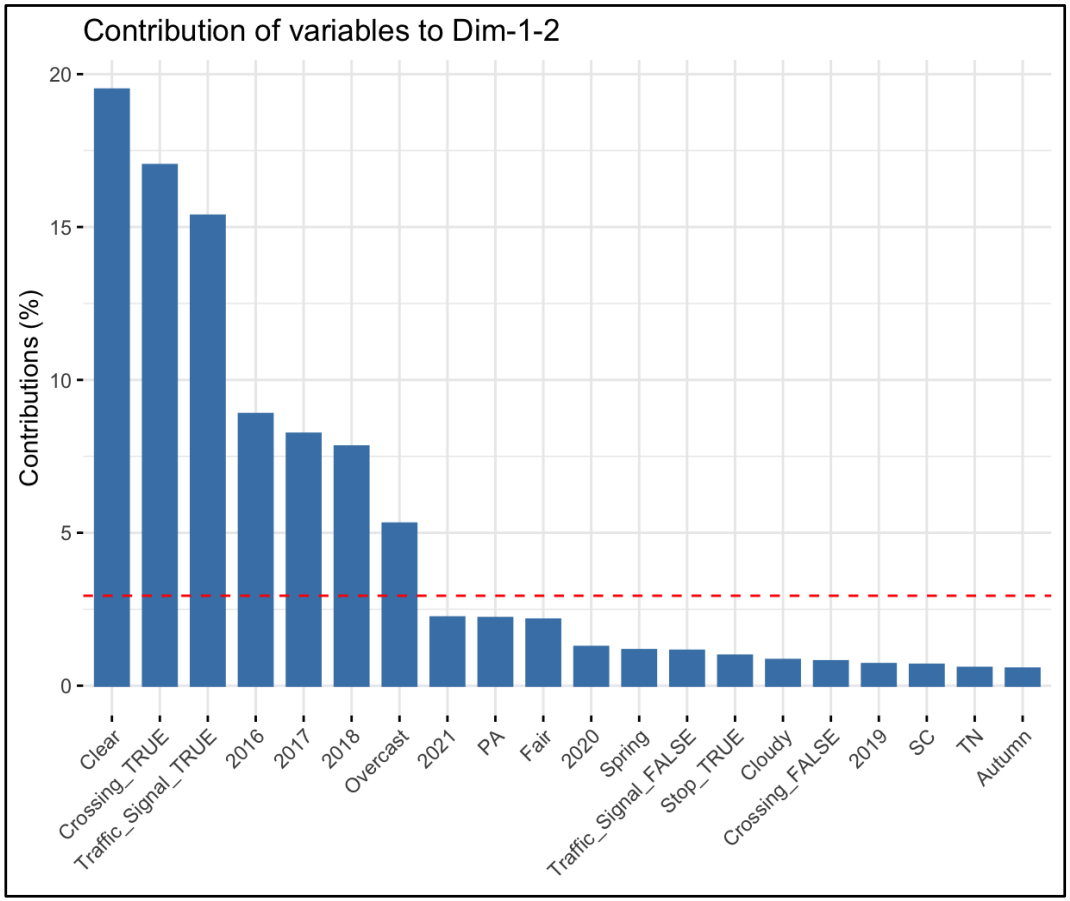

Figure 27: Contribution of variable categories for Dim1 and Dim2

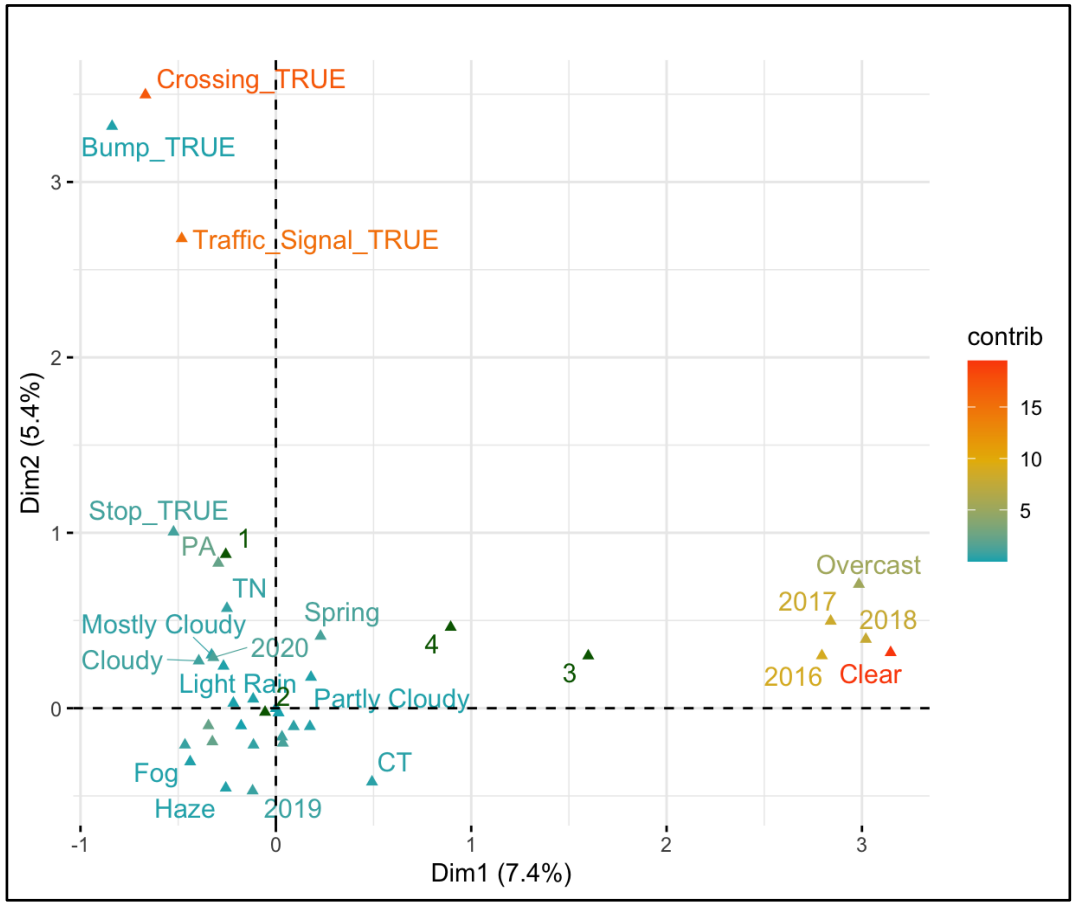

Figure 26: Level of contribution of each variable category plotted for the first two dimensions

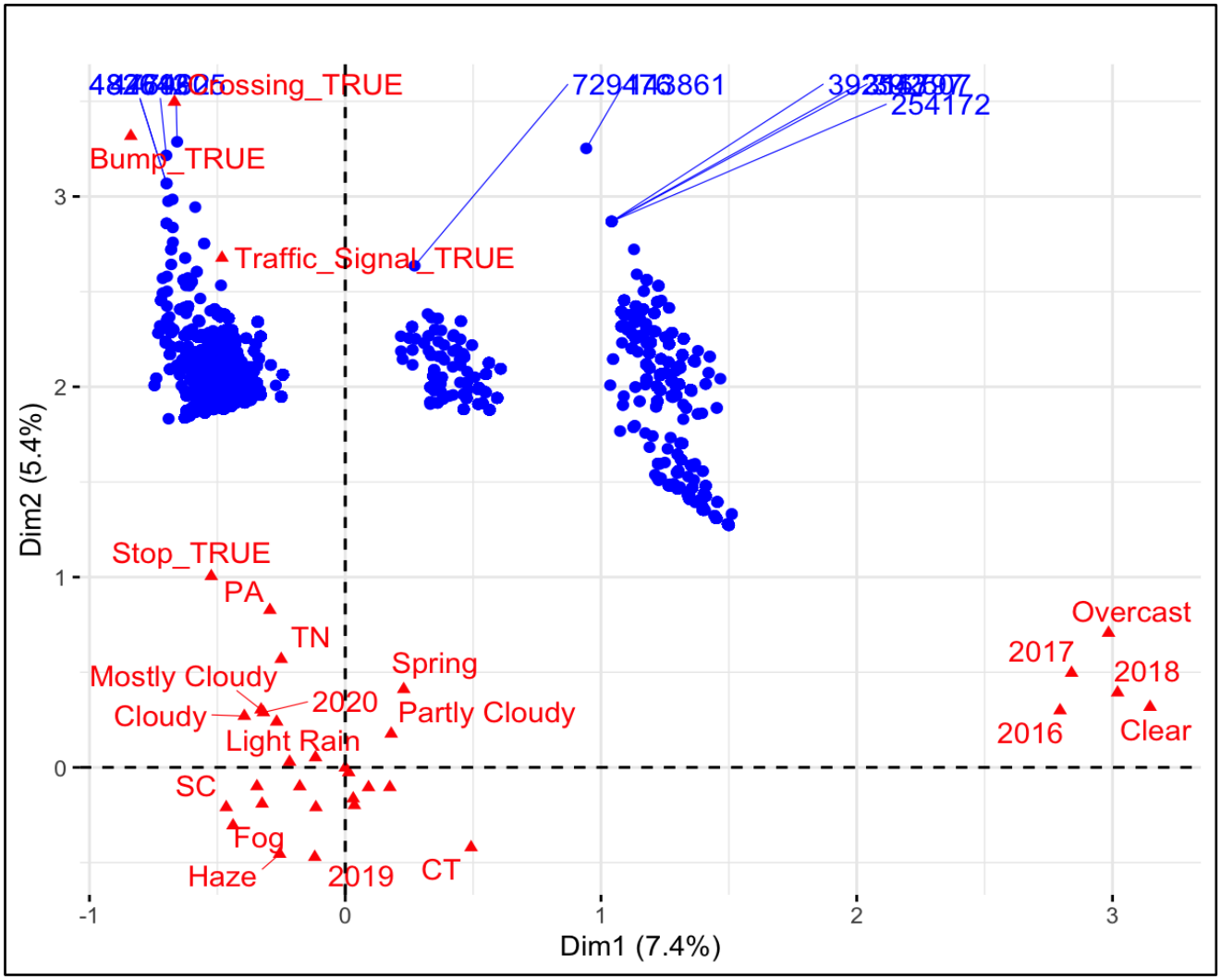

Figure 28: Correlation between variable categories and the first two dimensions

Multiple Factor Analysis - MFA

The used dataset for this project is complex due to the fact that there is a mixed type of data. The solution to deal with this is to apply a multi factor analysis. We decided to group the data into 6 different groups, 4 categorical and 2 numerical. To set a proper scenario, the grouping that has been done in the MFA analysis has been the following:

- Geographic group (taking into account the state to lower dimensionality, it is a qualitative variable)
- Distance group (Quantitative variable)
- Scenario group (taking into account the physical components that compose the scenario of the accidents, it is a qualitative binary variable).
- Weather condition group (qualitative variable)
- Time group (taking into account the season variable, therefore a qualitative variable)
- Atmospheric conditions group (taking into account all numerical variables that described the atmospheric conditions at the time of the accident).

The following scree plot (Fig. 29) has been performed in order to explain the percentage of representation each dimension has on the dataset. We can conclude that a low percentage of data is represented by a few numbers of dimensions. Dimensions 1 and 2 only represent around 13% of the total variance of data, meaning that we cannot take the respective conclusions as a general representation of the whole dataset. Let's look at the following graphs.

In order to choose the number of dimensions to be represented we had to look into the ones that had an inertia above 1/p, which in this case was 0.14. We were only able to get the inertia for the first 5 dimensions and all of them are above 0.14 the last one having an inertia of 0.1923565. So we need to get all 5 dimensions.

For this next graphic representation (Fig. 30), we can see the contributions of each group in the dimensions 1 and 2. The atmospheric conditions followed by the weather condition grouping and the time group have a high contribution to dimension 1, so this first dimension could be labeled as weather through time, whereas the weather condition group followed by the geographic group have a high contribution to dimension 2, this dimension could be labeled as weather depending on the geography, in this case, the states. Finally, distance and scenario groupings have a low or no representation in Dimensions 1 and 2. It must be said that due to the large distance between the variables, there is no grouping between them so no positive correlation can be concluded, nevertheless, the points that are away from the origin are well represented on the factor map.

We did the same for all the other dimensions, and we can conclude that dimension 3 is explained mainly by the distance, scenario, and weather condition, therefore showing the distance of the accident depending on the weather condition and the scenario. Dimension 4 has mainly the contribution of weather conditions, time_df, and geography, this dimension looks like a combination of dimensions 1 and 2, so it shows how the weather, the season and

the state can affect the accident. Lastly, dimension 5 is mainly explained by the groups' weather condition and geography, showing weather conditions depending on the state.

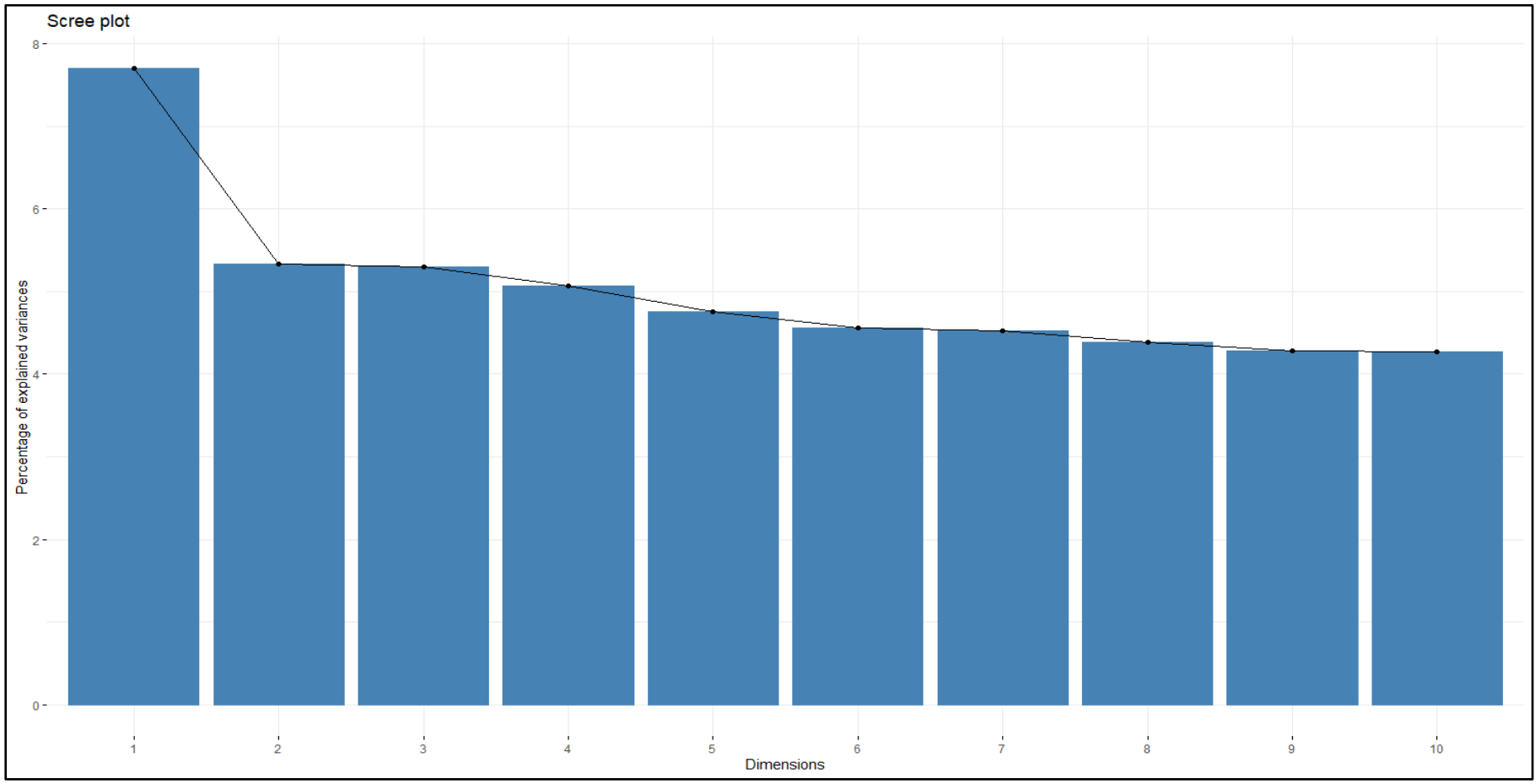

Figure 29: Variance explained by each dimension

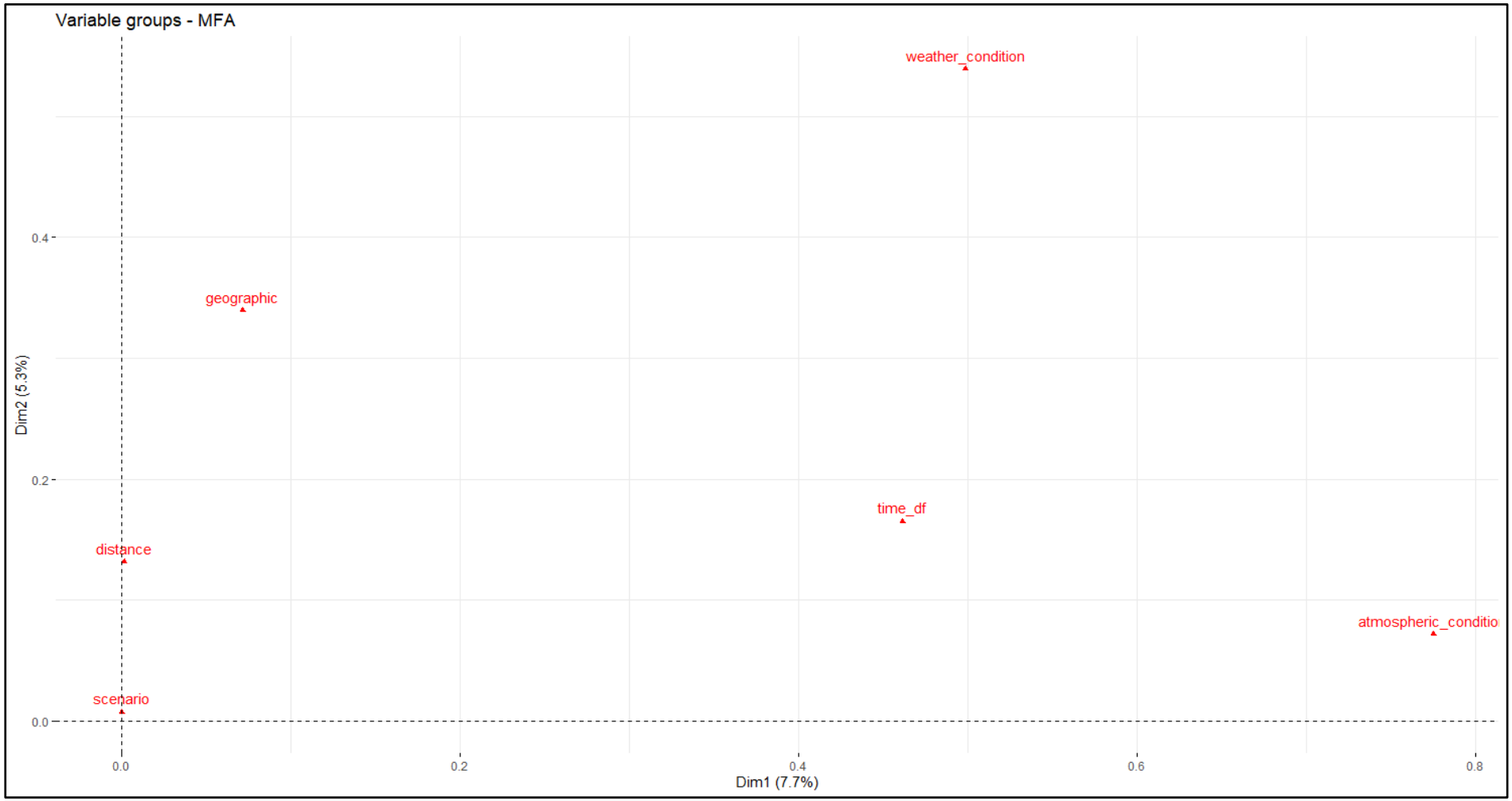

Figure 30: MFA groups dimension 1 and dimension 2

The third graph that is represented in this project regarding MFA (Fig. 31) is the following regarding the quantitative. For this graph, the variable arrows for precipitation and humidity are both pointing towards the positive side of d1, meaning they have a big correlation with MFA's first dimension. Out of all these arrows, just visibility appears in the positive side of d2, meaning its correlation with the MFA second dimension. Dimension 1 mostly reflects the inference that atmospheric conditions have in the traffic accidents dataset. This aligns also for the time variable. While the second dimension reflects that the geographic information and the distance have a relationship with this dimension, scenario elements have weak relationship with dimension 2.

Then we looked into the quantitative contributions between dimensions 1 and 3 (Fig. 32). Regarding dimension 1, as expected, we get the same results. But when looking into dimension 3 we can see almost nothing is explained by quantitative variables, only visibility has a little contribution having a contribution of around 0.4.

Moving on into the qualitative variables (Fig. 33), the first graph represents the contribution of the qualitative levels in the dimension 1 and 2. Looking at the first dimension, we can see that the highest contributions are only positive contributions, this being light rain, rain, and fog. This first dimension, regarding the qualitative, is mostly explained by weather conditions, having a lower contribution, we can also see winter in the positive section and summer in the negative, this being part of the time_df group. Now looking at the second dimension, we found haze and fog are contributing in a negative manner, while the state of Tennessee is contributing in a positive way. The third dimension is built of Cincinnati, Other and Rain having a positive contribution, and on the other hand, we can see all the scenarios true having a negative contribution. This being, from highest to lowest contribution, bump_true, crossing_true, traffic_signal_true, and lastly having a contribution under 1, stop_true.

Finally, and to end this MFA part, the last graph (Fig. 34) represents how the plotting of the accidents distribute themselves within both dimensions 1 and 2 taking into account our target variable which is severity (on a scale from 1 to 4). For this graph, the whole dataset was taken to plot, but the graph turned out to be unreadable as the plotting was 'squished' in one big cloud of dots. Therefore, a sample of 1% of the registers has been taken to plot this graph. We can conclude that the vast majority of the accidents can be classified as severity number 2 and also the majority of this sample have a big correlation with MFA 's first dimension.

For the Multi Factor Analysis it must be noticed the following problems encountered and how they have been dealt with:

A big game changing problem we have faced has been that for this analysis we found that once we had included the respective code, approx. 700 dimensions explained 60% of data. Apart from this being a massive amount of dimensions, we also encountered memory problems when running the code. Part of the problem was caused by the numerous variables that the dataset includes, specifically for the weather condition levels and the number of cities per each county. The way to deal with this and therefore solve the previous problem, has been to group all the weather condition levels that make up less or equal 1% into a group called "Other". This turned out to have 11 levels in weather conditions in the end. Now when running

the MF analysis, we include all of the variables except city and county, and the total amount of dimensions that explain 80% of the data are 18.

Briefly, for qualitative variables:

- Most of the scenario characteristic variables are positively related to each other.
- Those variables regarding rain, fog and other high humidity weather conditions are correlated with Dim1, whereas cloudy weather conditions are more correlated to Dim2. Those variables that inform of a ‘Good weather condition’ including the season of summer, are found in the negative side of Dim1.
- Four out of the 5 states selected find themselves in the positive area of both dimensions, whereas California is found in the opposite direction meaning to be negatively correlated.
- Once more, humidity and precipitation take a lot of importance in this plot being those variables that are most correlated with Dimension1.

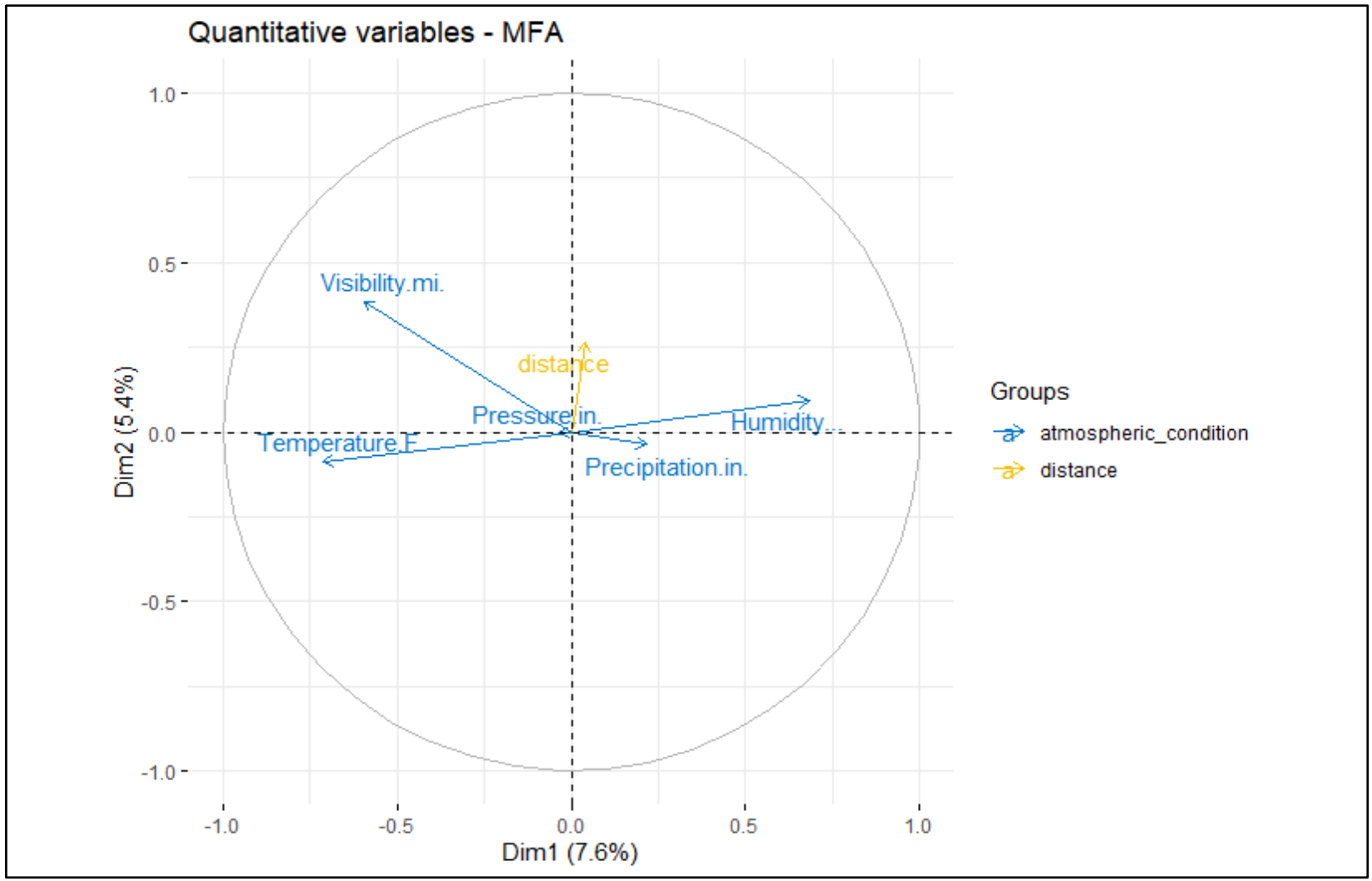

Figure 31: Dimension 1 and dimension 2 of the quantitative variables

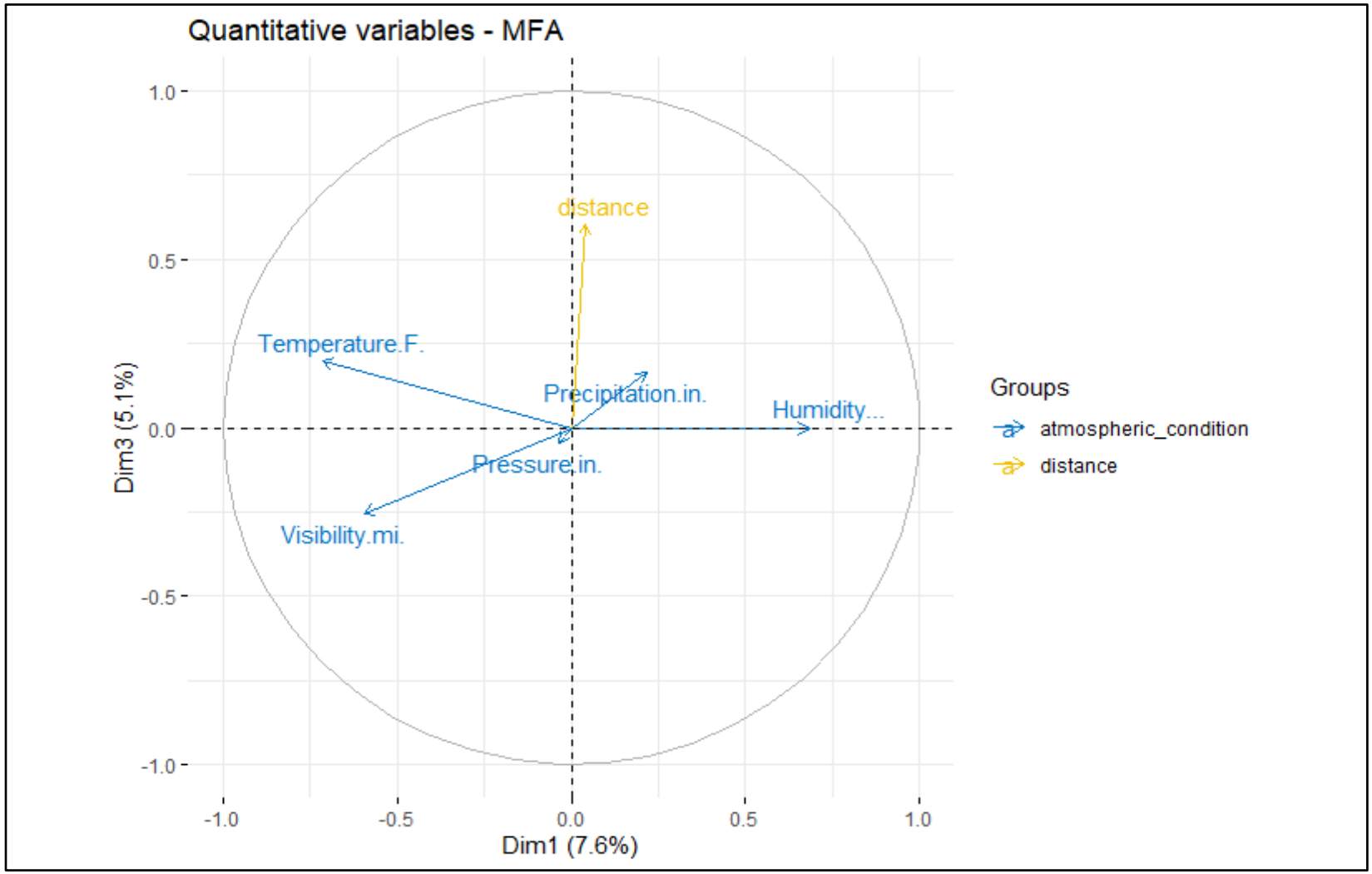

Figure 32: Dimension 1 and 3 of the quantitative variables

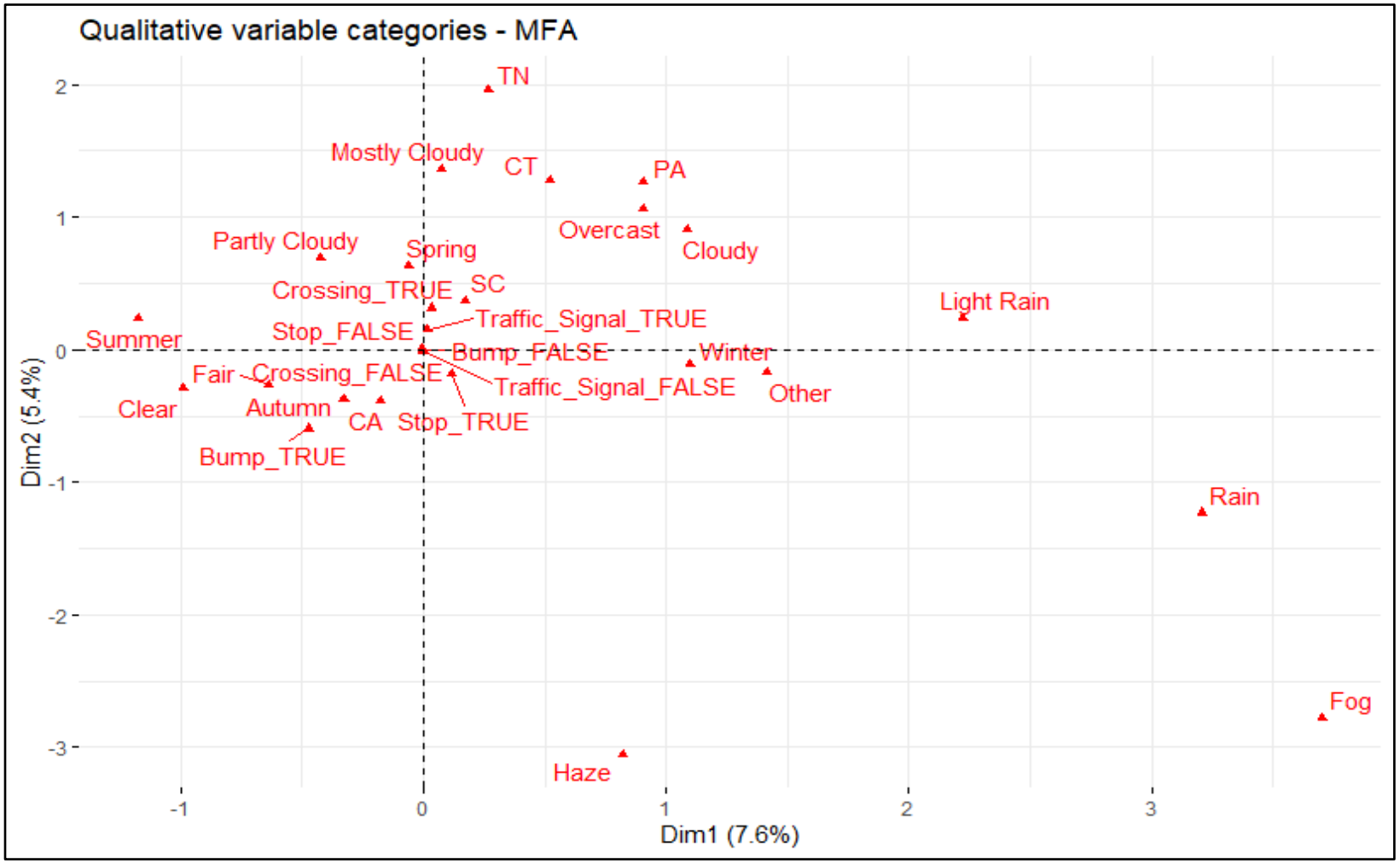

Figure 33: Dimension 1 and 2 of the qualitative variables

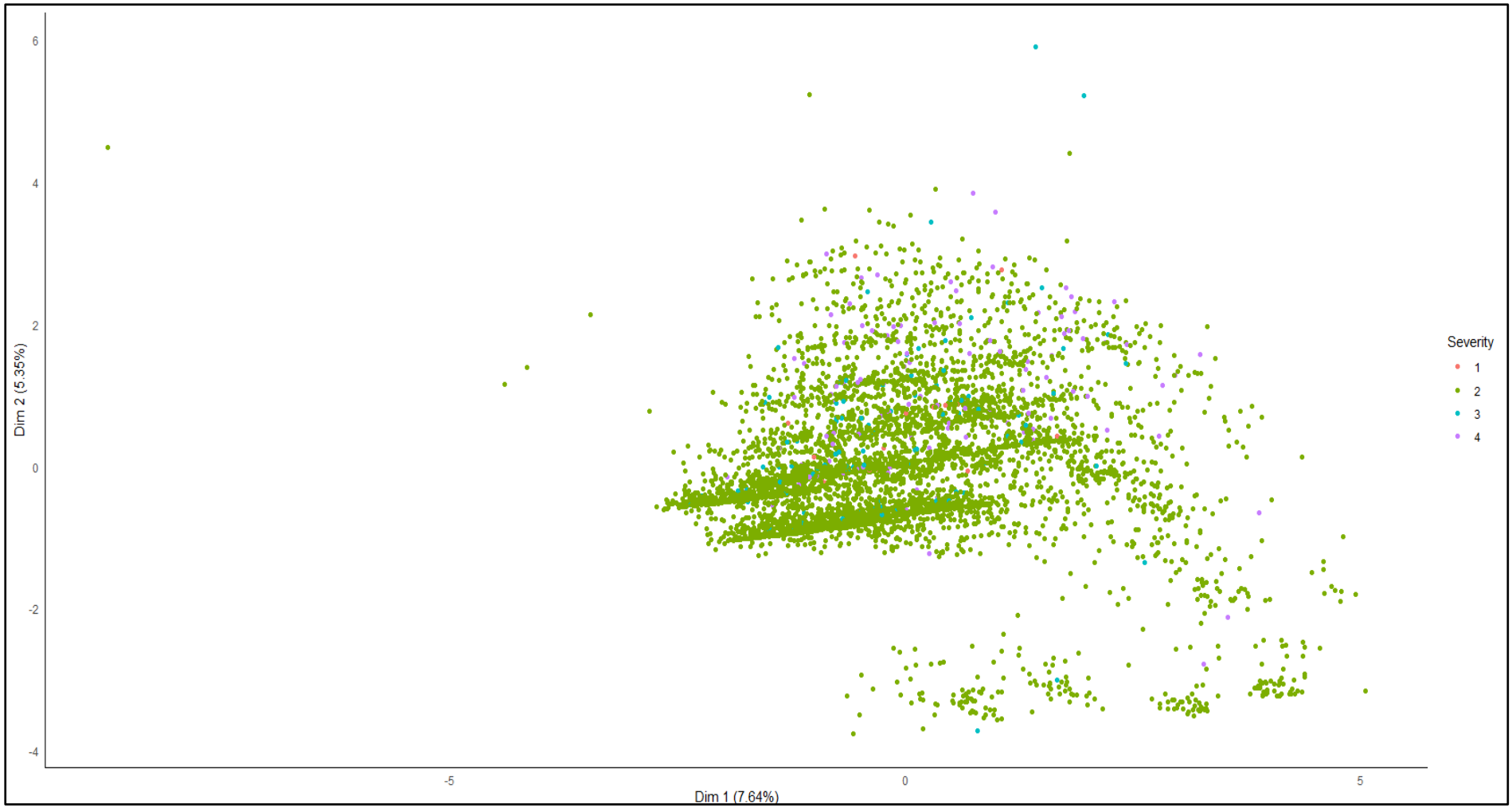

Figure 34: Plot of distribution of the severity of the accidents according to Dim 1 and 2.

Association Rules Mining Analysis

Data Preparation

In this section of the project the association rules mining analysis is taking place. The objective of this stage is to mine association rules between the different values of all the variables of the dataset, and consequently create insights from the most interesting and frequent ones. In order to achieve that, firstly, it was necessary to convert all the variables of the dataset into categorical variables (factors).

As mentioned before, after the preprocessing and univariate-bivariate analysis stage of the project the final dataset consists of 6 numerical and 15 categorical features. The variables “ID”, “Start_time”, “End_time”, “State”, “City” and “Month” were not taken into account. The reasons behind ignoring those columns of the dataset are explained here. For the execution of the association rules algorithms the “ID” column is not necessary. Moreover, “Start_time” and “End_time” columns are timestamps, which in most of the cases are unique. Thus, trying to convert them into factors would create a total number of different levels equal to the total number of rows of the dataset, which finally does not provide any insight in this specific analysis. Finally, concerning the “State” and “City” variables, as it is logical, there is always an association between the values of “State”, “City” and “County”. For that reason, using the overall combination of those variables as inputs for the association rules algorithms would lead to redundant results of bad quality. For that reason, it was decided to use only the variable County of the dataset, for this specific analytical purpose. Exactly the same situation is observed for the features “Season” and “Month” which are logically connected. As a consequence, it was decided to use only the “Season” variable between the two.

Additionally, it is crucial to mention that some changes applied, as well, for the variable “Severity” which is the target of the analysis. Severity levels which were represented with integers from 1 to 4, were replaced by the strings “Severity1”, “Severity2”, “Severity3” and “Severity4”, respectively. In that way, the results of Eclat[2] and Apriori[3] algorithms from the arules package are easier to understand.

To continue with, the numerical features are named as follows: “Temperature(F)”, “Humidity(%)”, “Pressure(in)”, “Precipitation(in)”, “Distance(mi)”, and “Visibility(mi)”. In order to apply association rules mining algorithms, it was mandatory to find a way of translating the numerical values of those variables into categories. Those transformations can be found in the table below. More specifically, for every variable there is a mapping from numerical sets to categorical levels.

[2] https://www.rdocumentation.org/packages/arules/versions/1.6-8/topics/eclat
[3] https://www.rdocumentation.org/packages/arules/versions/1.6-1/topics/apriori

Variable	Numerical Sets	Categorical Levels
Visibility(mi)	-	-
	[0, 0.57539)	Very Poor
	[0.57539, 2.30156)	Poor
	[2.30156, 5.7539]	Moderate
	[5.7539, inf)	Good
Precipitation(in)	-	-
	(-inf, 0]	None
	(0, 0.1]	Light
	(0.1, 0.3]	Moderate
	(0.3, 24)	Heavy
Humidity(%)	-	-
	[0,55]	Low
	(55,65]	Medium
	(65,101]	High
Temperature	-	-
	[-19, 32]	Very Cold
	(32,50]	Cold
	(50,59]	Cool
	(59,68]	Warm
	(68,77]	Warm to Hot

	(77,86]	Hot
	(86,120]	Very Hot
Pressure(in.)	-	-
	[0, 29.81]	Low
	(29.81, 30.21]	Normal
	(30.21, 59]	High
Distance.mi.	-	-
	[0,0.1960]	Small
	(0.1960,2.54]	Medium
	(2.54,137]	High

Table 4: Mapping of Numerical Sets of Categorical Levels

The mapping of the numerical sets to categorical values was completed in most of the cases by combining information from weather-related websites such as the National Weather Service[4] (NWS). What's more, k-means grouping was accomplished for the mapping of "Distance" variable numerical values. In order to give a clearer view of the cardinalities of the dataset's variables to the reader, a table including them is presented below(Table 5).

Variable	**Cardinalities**
Weather_Condition	{Fair, Cloudy, Mostly Cloudy, Partly Cloudy, Clear, etc.}
Year	{2016, 2017, 2018, 2019, 2020, 2021}
County	{Los Angeles, San Bernardino, Orange, San Diego, etc.}
Crossing	{TRUE, FALSE}
Bump	{TRUE, FALSE}
Stop	{TRUE, FALSE}
Traffic_Signal	{TRUE, FALSE}
Severity	{Severity1, Severity2, Severity3, Severity4}

[4] https://www.weather.gov/

Season	{Autumn, Spring, Summer, Winter}
Visibility.mi.	{Very Poor, Poor, Moderate, Good}
Precipitation.in.	{None, Light, Moderate, Heavy}
Humidity...	{Low, Medium, High}
Temperature.F.	{Very Cold, Cold, Cool, Warm, Warm to Hot, Hot, Very Hot}
Distance.mi.	{Small, Medium, High}
Pressure.in.	{Low, Normal, High}

Table 5: Variables Cardinalities

Relative Item Frequency in the Transactions Object

After transforming the latest version of the dataset into a transactions object, it is observed that there are 1060500 transactions, in total, containing different combinations of 304 items. In the following graph, the relative item frequency of all the items that take part in the transactions object of the dataset is presented. For this specific projection only the top 20 items based on their support value are considered. Support mainly defines the frequency of occurrence of an item or an itemset in the total number of transactions divided by the total number of transactions.

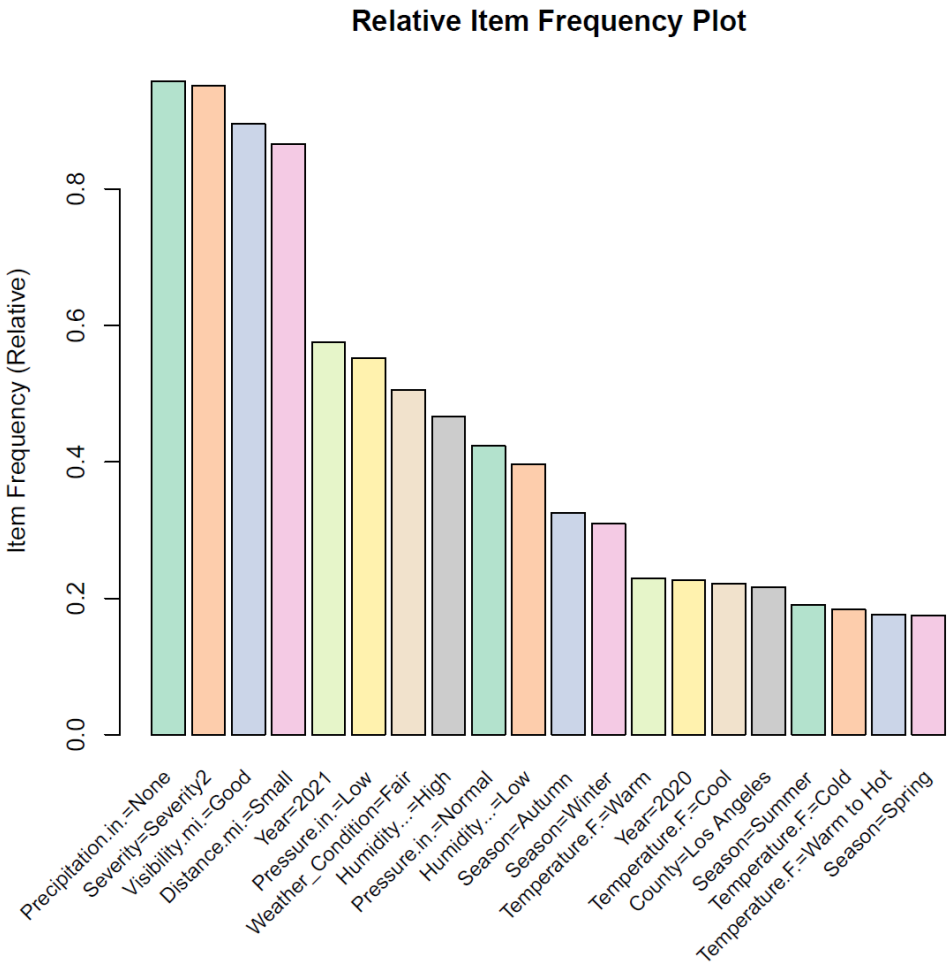

Figure 35: Top 20 items based on the relative item frequency

From this graph (Fig. 35) one can understand that the most frequent items of the dataset are the following ones: Precipitation.in =”None”, Severity = “Severity2”, Visibility.mi = “Good”, Distance.mi = “Small”, Year = “2021” and Pressure.in = “Low”.

In more depth, the itemset length distribution is described below. The following table (Table 6) presents the amount of transactions or itemsets containing a specific number of items.

Number of Items in Transaction	Amount of Transactions
10	1846
11	940307
12	89249
13	28690

14	407
15	1

Table 6: Numer of Transactions per Number of Items in Transactions

It is clear that the most frequent length of the itemsets in the transactions is 11 by far.

Identification of Frequent Itemsets

Eclat Algorithm

In this subsection, the most frequent itemsets of the transaction object are presented and explained. An itemset is composed of specific values for specific variables that occur together. In order to mine those itemsets the Eclat algorithm has been used. The thresholds used as parameters for the execution of the algorithm are:

- **minSupp** = 0.1 and **minlen** = 3

By using the mentioned parameter values, the output of the algorithm will hold in itemsets with a minimum length of 3 items that achieve a support value greater than 0.1. The following table (Table 7) contains the top 5 itemsets, sorted by their support value. The whole table containing the top 30 itemsets can be found in the Appendix.

Itemsets	**Support**	**Count**
Severity=Severity2,Visibility.mi.=Good,Precipitation.in.=None	0.834	884825
Severity=Severity2,Precipitation.in.=None,Distance.mi.=Small	0.791	839372
Visibility.mi.=Good,Precipitation.in.=None,Distance.mi.=Small	0.763	808775
Severity=Severity2,Visibility.mi.=Good,Distance.mi.=Small	0.740	784784
Severity=Severity2,Visibility.mi.=Good,Precipitation.in.=None, Distance.mi.=Small	0.726	770217

Table 7: Top 5 Most frequent itemsets

By taking a look in the previous table, one can conclude that the most frequent itemsets consist of different combinations of the most frequent items projected in the relative item frequency plot, which totally makes sense. In most of them, it is readily apparent that “Severity= Severity2”, “Visibility.mi.= Good”, ”Precipitation.in.= None”, ”Distance.mi.= Small”, and ”Pressure.in.=Low” are presented.

Identification of Association Rules

Apriori Algorithm

After taking a look at the most frequent itemsets in the transactions of the dataset, the generation of the association rules is followed. Apriori algorithm has been used for this specific step, with its parameters being set as:

- **minSupport** = 0.1, **minConfidence** = 0.25 and **minlen** = 2

In that way, the results will only contain rules that surpass a minimum support threshold of 0.1, as well as, a minimum support confidence threshold of 0.25. The confidence of a rule defines the conditional probability of the right-hand side (RHS) of a rule given the left-hand side (LHS) of it. Finally, the minimum number of items contained in a rule is set to be equal to 3. After the execution of the algorithm, a set of 3148 rules is generated in total. The total set of rules and their quality metrics (support, confidence, coverage, lift and count) can be found in the file "total_rules.csv" which exists in the archived file of this deliverable.

To continue with, in the following graphs (Fig. 36) the illustration of the rules is accomplished, by taking into account their quality values of: support, confidence and lift. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the rule body and the rule head divided by the support of the rule body. First of all, in the first plot, the rules are presented on the support and lift axes. It is conspicuous that the rules with high confidence values are mainly presented in the level where lift measure is approximately one. The support level does not play a significant role there. To conclude, it seems that most of the rules with high confidence have a lift value close to 1 and that means that most of the rules do not provide a good association between LHS and RHS. A value of the measure lift close to one assumes independence between the two sides of a rule. Additionally, there are many rules with a value of lift which is less than 1 and in this case they are not considered important. On the other hand, there are several rules that have a lift value greater than 1 (some of them reach around 2.2) and some of them have a good level of confidence as well. Those cases will be described in more detail in next subsections.

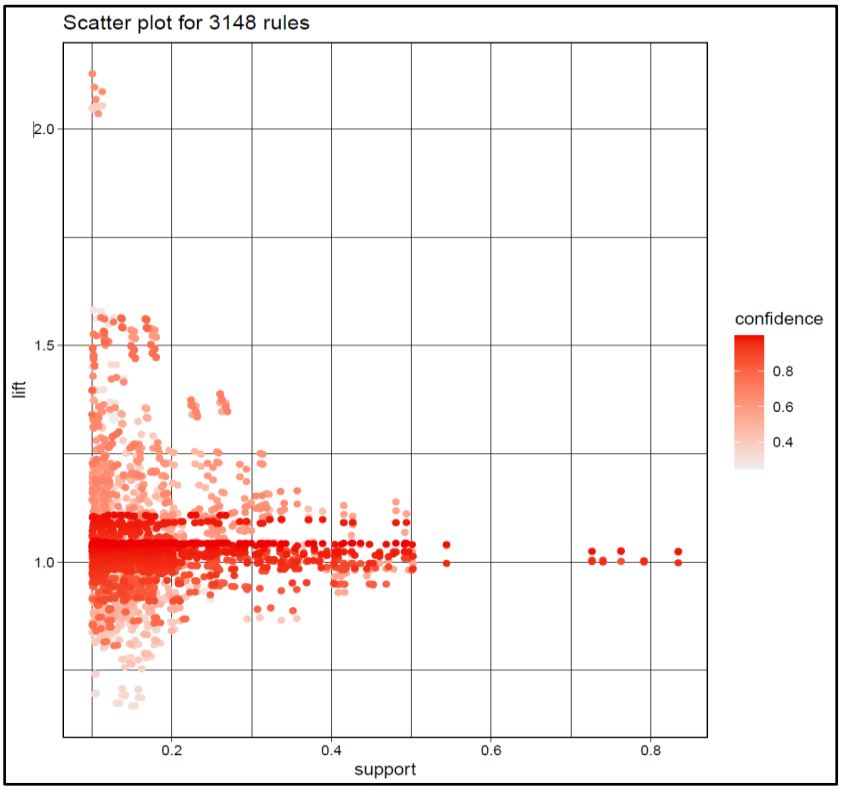

Figure 36: Support-Lift-Confidence of the 3148 Rules

The same conclusions can be derived from taking a look in the following chart (Fig. 37) which depicts the rules again, but this time by using as axes the quality measures support and confidence. One addition could be that the rules with the highest values of lift have a value of support around 0.1 and confidence approximately 0.4 or 0.6.

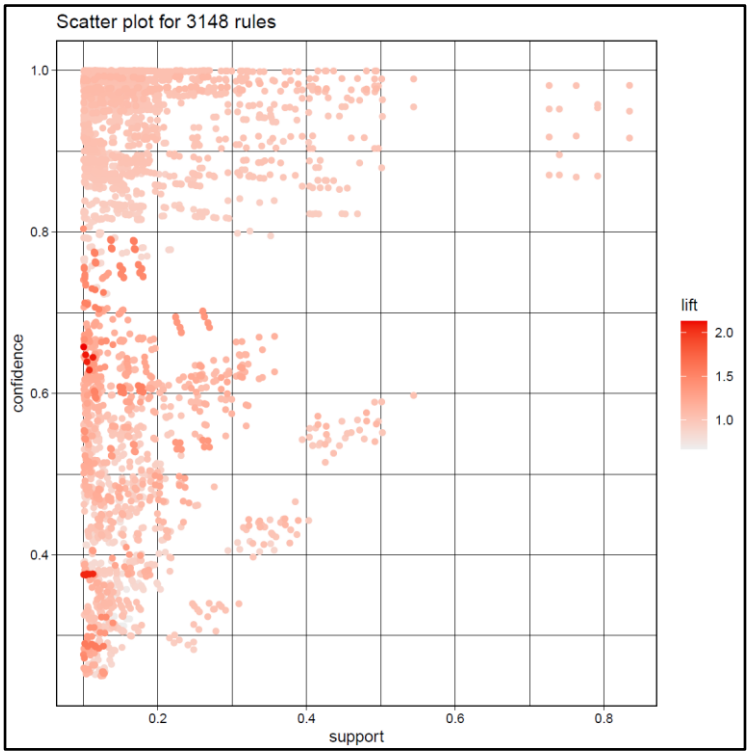

Figure 37: Support-Confidence-Lift of the 3148 Rules

Here in the report, only the top rules based on their values of lift and confidence will be displayed. The next table (Table 8) presents the count of rules based on the number of items they contain.

Number of Items in a Rule	Count of Rules
3	768
4	1172
5	868
6	305
7	35

Table 8: Number of Rules per Number of Items in a Rules per Number of Items in a Rule

In most of the cases, it is discernible that it is more frequent for a rule to be composed of 4 or 5 items. In the following subsection, a closer look will be taken into the top 20 association rules produced.

Top 20 Association Rules

Additionally, the top 20 rules, sorted by their lift value are included in the next table (Table 25: Top 20 rules of the Dataset).Then the confidence measure was used as a second sorting step of the top rules. A rule as mentioned before consists of a LHS and a RHS. In general, the meaning of a rule is the following: If the variables of the LHS take specific values then there is a big probability (depending on the lift, confidence and support measures respectively) to lead to the RHS of the rule. In general, in the RHS of a rule a specific variable is equal to a specific value from its cardinalities. The top 20 rules of the dataset are presented in detail at the table(Table 25: Top 20 rules of the Dataset) located at the Appendix.

It is clear that the top association rules from our dataset - the ones with the highest values of lift and confidence - are deriving conclusions about weather conditions. For example, it is explicit that in all of the cases above the RHS of the rule either contain the value Temperature.F.= “Cold” or Season= “Winter”. Additionally, the LHS principally holds values such as Severity = “Severity2”, Precipitation.in. = “None”, Distance.mi .= “Small” or Temperature.F. = “Cold”. If the appearance of Severity2 is not taken into account, the rules would describe in most of the cases associations between weather conditions and seasons. In the next subsection, a detailed look is taken for the specific rules which contain severity level “Severity2” in the RHS, but for the other severity levels as well. A graphical representation of the table presented is illustrated with the next two plots. The first one (Fig. 38) is a vector graph pointing the connection between different values of the dataset’s variables in the association rules based on their support and lift values.

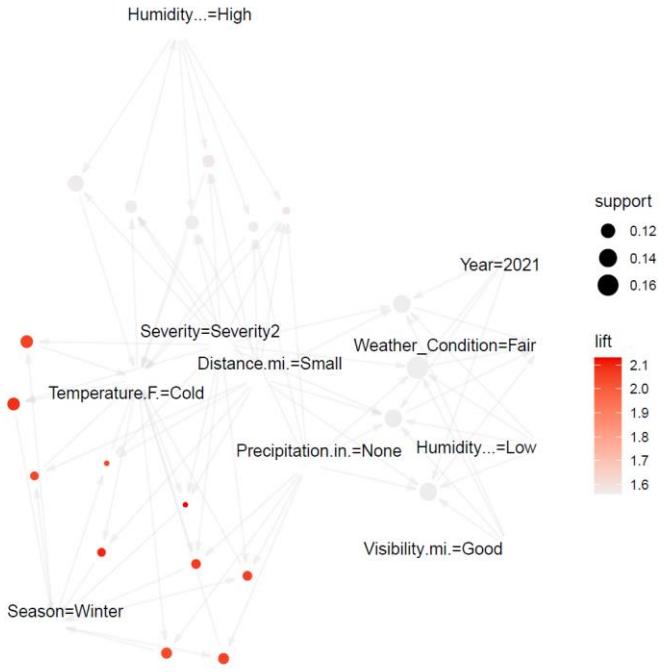

Figure 38: Graph of Tope 20 Rules Based on Support and Lift

The second plot (Fig. 39) depicts the same thing, but with parallel coordinates which make it easier to understand the connection of the variables and their values from the LHS to RHS of the rules.

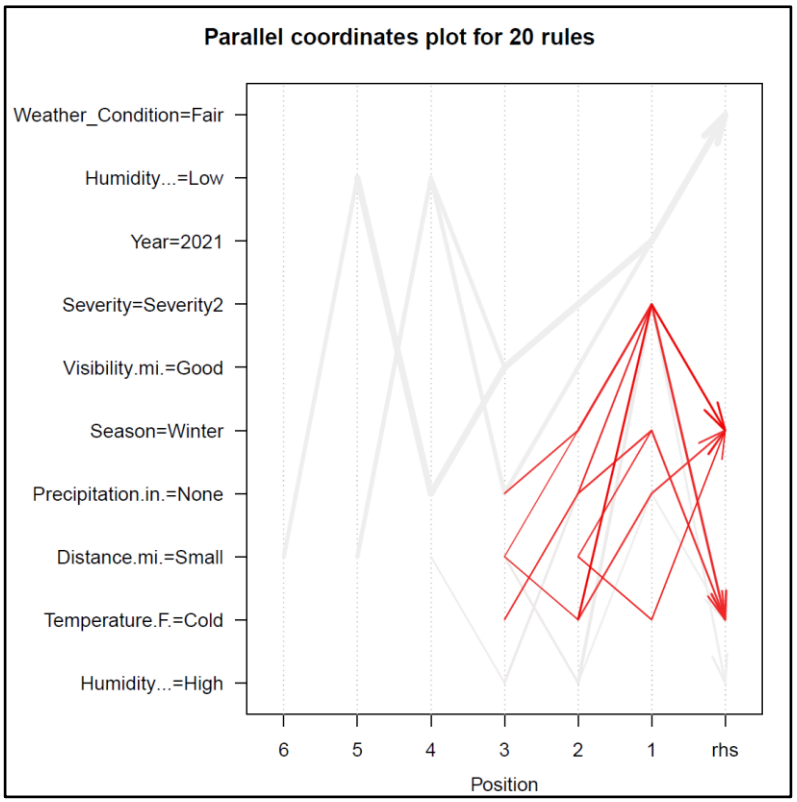

Figure 39: Parallel coordinates plot for the 20 rules

Rules Containing Severity Levels on the RHS

After generating the total set of rules which describe the transactions of the dataset, they are divided into 4 different subsets. Each of those subsets hold the 4 different levels of the "Severity" factor in the RHS part. Based on the available data, the only level of the Severity variable existing on the RHS of a rule was "Severity2". The subsets for the other levels of Severity are empty. The top 10 rules containing "Severity2" value in their RHS are presented here(Table 9), but the full set of those rules is contained in the "total_rules_severity2.csv" file, which has been archived into the project's deliverable. For space saving, only a column indicating the LFS of the rules is added to the table (RHS is always "Severity=Severity2" in those cases).

LHS of Rules	Support	Confidence	Lift
Year=2021, County=Los Angeles, Precipitation.in.=None	0.102	0.998	1.049
Year=2021, County=Los Angeles	0.106	0.998	1.049
Weather_Condition=Fair,	0.115	0.995	1.046

LHS of Rules	Support	Confidence	Lift
Year=2021, Visibility.mi.=Good, Precipitation.in.=None, Pressure.in.=Normal			
Weather_Condition=Fair, Year=2021, Visibility.mi.=Good, Pressure.in.=Normal	0.115	0.995	1.046
Weather_Condition=Fair, Year=2021, Precipitation.in.=None, Pressure.in.=Normal	0.119	0.995	1.046
Weather_Condition=Fair, Year=2021, Pressure.in.=Normal	0.119	0.995	1.046
Year=2021, Visibility.mi.=Good, Pressure.in.=Normal	0.180	0.994	1.046
Year=2021, Visibility.mi.=Good, Precipitation.in.=None Pressure.in.=Normal	0.178	0.994	1.046
Year=2021, Precipitation.in.=None Pressure.in.=Normal	0.199	0.994	1.046
Year=2021, Pressure.in.=Normal	0.204	0.994	1.046

Table 9: Top 10 Rules Leading to Severity Level at RHS

One can see that all of the rules have a value of lift around the value 1. This, again, means that the LHS and RHS are independent. Also, the most observed values on the LHS of the rules pointing to “Severity2” level of the Severity variable are: Year= “2021”, County= “Los Angeles”, Precipitation.in.= “None”, Weather_Condition= “Fair” and Pressure.in.= “Normal”. From an analytical perspective, Year and County variables’ values are really subjective, so they don’t provide much information for the severity of the accident. On the other hand, all the weather condition variables have values which indicate a good weather for the severity level 2 accidents (like fair weather condition, normal atmospheric pressure and no precipitation). A graphical representation of the above-mentioned insights for the top 10 rules of accidents with severity level equal to 2 can be seen in figures (Figure 71: Graph of Top 10 Severity2 Rules Based on Support and Lift and Figure 72: Parallel Coordinates for Top 20 Rules with Severity2 in RHS) located at the Appendix.

Hierarchical Clustering

Hierarchical Clustering is an algorithm that groups objects that are similar into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other, and the objects within each cluster are quite similar to each other. This means that the intercluster distance (distance between two objects belonging to two different clusters) has to be maximized whereas the intracluster distance (distance between two objects belonging the same cluster) has to be minimized.

Precise Description of Data Used

However, before starting to test different strategies to address this chapter, it is necessary to redo a preprocessing of the data. The reason leading to this necessity is the total number of individuals which exceeds one million and due to limitations of the RAM memory in our devices, the following steps have been conducted:

- Create factors for specific variables (Year, Severity and Month).
- Scaling numerical variables before clustering.
- Checking associations and variance of severity (target variable), on different groups of categorical variables (State, Season, Year).
- Visualize tables only containing the variables Year, State and Severity.
- Feature selection before clustering

Checking the tables of Year, State and Severity we have been able to see that for California in 2020 there are several accidents characterized by all different levels of severity. For that reason our analysis for the rest of the project will only focus on accidents that took place in California in 2020, and we will sample randomly 20K observations from all 4 different levels of Severity (20K due to RAM memory limitation for the calculation of the distance matrices for the clustering).

Concerning the feature selection part, before clustering, the following variables of the dataset were not taken into account: ID, City, County, Start_Time, End_Time, Month, State, Year, Severity. To begin with, it is obvious that the target variable (Severity) was not included in the clustering analysis. In addition, Year and State variables were not included as well since the sample used for this procedure only refers to accidents in California in 2020. At the same time, the ID variable was discarded, because including identification columns in clustering algorithms is not a correct approach. Moreover, variables Start_Time and End_Time provide information about the time that the accident took place and without further processing of their values, they do not provide any similarity or dissimilarity information to the process, because most of the accidents have unique values for those variables. Also, City and County variables were excluded, for the reason that, as it was concluded from the Univariate and Bivariate analysis of the project, those variables are highly correlated with the State variable, which in our case is California for all observations. Thus, our clustering approach is not focusing on separating accidents based on the place they happened (City or County) but mainly on the conditions that existed during the accident (the rest of the variables of the dataset). Finally, Month was not taken into account, since the variable Season was used instead, for providing time information to the clustering procedure.

Clustering method used: metrics and aggregation criteria used

As we know, K-Means, DBSCAN and hierarchical clustering all have one thing in common: they are all distance-based clustering algorithms. Those algorithms, in their default mode, use Euclidean distance function for the calculation of dissimilarity between observations and clusters. Thus, since euclidean distance is not appropriate for clustering categorical or mixed data, in this case, it was necessary to use other distance functions. One of the most famous distance functions that can be applied on categorical or mixed data is Gower Distance function.

In this project, daisy has been used. Daisy is a built-in function of the package “cluster” in R, which implements the calculation of the Dissimilarity Matrix by computing all the pairwise dissimilarities (distances) between observations in the data set. The original variables may be of mixed types. The metric that has been used is the Gower Distance. This is a measure that can be used to calculate distance between two entities whose attribute has a mixed of categorical and numerical values. The dissimilarity between two rows is calculated by taking into account the weighted mean of the contributions of each variable. Specifically, the daisy distance with the use of Gower distance is calculated as follows:

$$d\_ij = d(i,j) = sum(k=1:p;\ w\_k\ delta(ij;k)\ d(ij,k)) / sum(k=1:p;\ w\_k\ delta(ij;k))$$

where, d_ij is a weighted mean of d(ij,k) with weights w_k delta(ij;k), where w_k= weights[k], delta(ij;k) is 0 or 1, and d(ij,k), the k-th variable contribution to the total distance, is a distance between x[i,k] and x[j,k]. Additionally, as the individual contributions *d(ij,k)* are taking values in the interval *[0,1]*, the dissimilarity *d_ij* will remain in the same range. Finally, due to the fact that the project’s dataset is unbalanced and it consists of several binary, numerical and categorical variables (with a lot of different categories), the use of the daisy distance in a squared form leads to a more robust clustering result.

On the other hand, one of the essential prerequisites for hierarchical clustering is the linkage function. Algorithms for hierarchical clustering normally differ by the linkage function they are using. The most common type of linkage functions are: single, complete or average linkage clustering functions. However, it has been decided to make use of the Ward’s linkage. The idea has much in common with analysis of variance. The linkage function specifying the distance between two clusters is computed as the increase in the “error sum of squares” (ESS) after fusing two clusters into a single cluster. Ward’s method seeks to choose the successive clustering steps so as to minimize the increase in ESS at each step. The ESS of a set is the sum of squares of the deviations from the mean value or the mean vector (centroid).

Resulting Dendrogram

The basic idea of how agglomerative hierarchical clustering works is starting by treating each observation as a separate cluster. Then, it repeatedly executes the following steps. First of all, identify the two clusters that are closest to each other and after this step, merge the two most similar clusters. This is an iterative process that continues until all the clusters are merged together. The main output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters. The output obtained from our data will be

available using a single page for it, as it can be seen in Figure 42: Resulting dendrogram at the end of this chapter.

Selection of Optimal Number of Clusters

The selection of the number of clusters is one of the most important tasks to perform in this subsection of the project. It is also one of the most difficult ones. There are several methods that can be used for this analysis. The explanation of some of them is taking place below in order to understand how they work.

- Elbow method: it is the most popular method for determining the optimal number of clusters. The method is based on calculating the Within-Cluster-Sum of Squared Errors (WSS) for different numbers of clusters (k) and selecting the k for which the change in WSS first starts to diminish. The idea behind the elbow method is that the explained variation changes rapidly for a small number of clusters and then it slows down leading to an elbow formation in the curve. The elbow point is the number of clusters we can use for our clustering algorithm.
- Dendrogram heights: this is a specific technique for hierarchical clustering. Hierarchical clustering starts by considering each point as a separate cluster and starts joining points to clusters in a hierarchical fashion, based on their distances. The dendrogram creates an illustration of the procedure just described. By taking a look at the heights (distances) of the different levels of the dendrogram, one can understand if there is a “big” or “small” distance between different clusters, before or after merging them with another cluster.
- Silhouette Coefficient: this metric is used to calculate the goodness of a clustering technique. Its values range from -1 to 1. 1 means clusters are well apart from each other and clearly distinguished. 0 means clusters are indifferent, or we can say that the distance between clusters is not significant. Finally, -1 means clusters are assigned in the wrong way. This coefficient is computed in the following way:

$$Silhouette\ Coefficient\ = (b - a)/max(a, b) \quad ,$$

  where “a” is the average intra-cluster distance (i.e. the average distance between each point within a cluster) and “b” is the average inter-cluster distance (i.e. the average distance between all clusters).

In this specific solution, the silhouette coefficient was taken into account for selecting the number of clusters. Thus, having a look at the plot obtained (Fig. 40Figure 40: Silhouette coefficients regarding the different numbers of clusters) it can be seen that the optimal number of clusters following the Silhouette Coefficient method is k=4 as we obtain the value closer to 1.

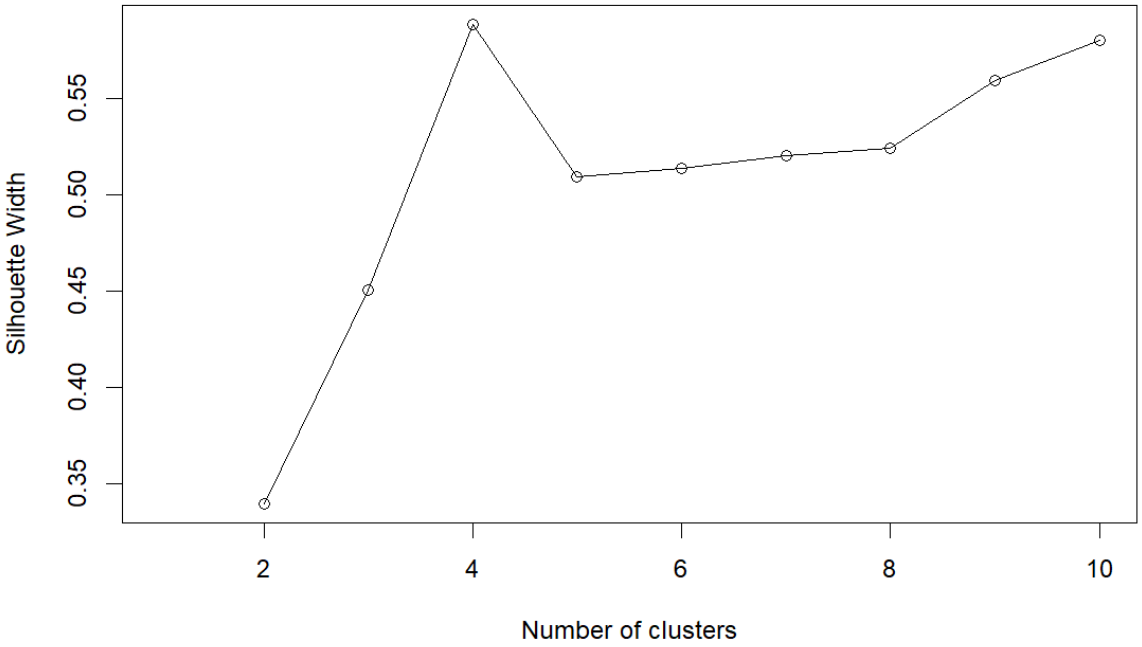

Figure 40: Silhouette coefficients regarding the different numbers of clusters

Regarding the Fig. 41 ,it can be seen that we have in total a 0.45 of average silhouette. Making an order from the maximum to the minimum values, we have a value of 0.65 for cluster 1, 0.47 for cluster 2, 0.40 for cluster 3 and 0.12 for cluster 4.

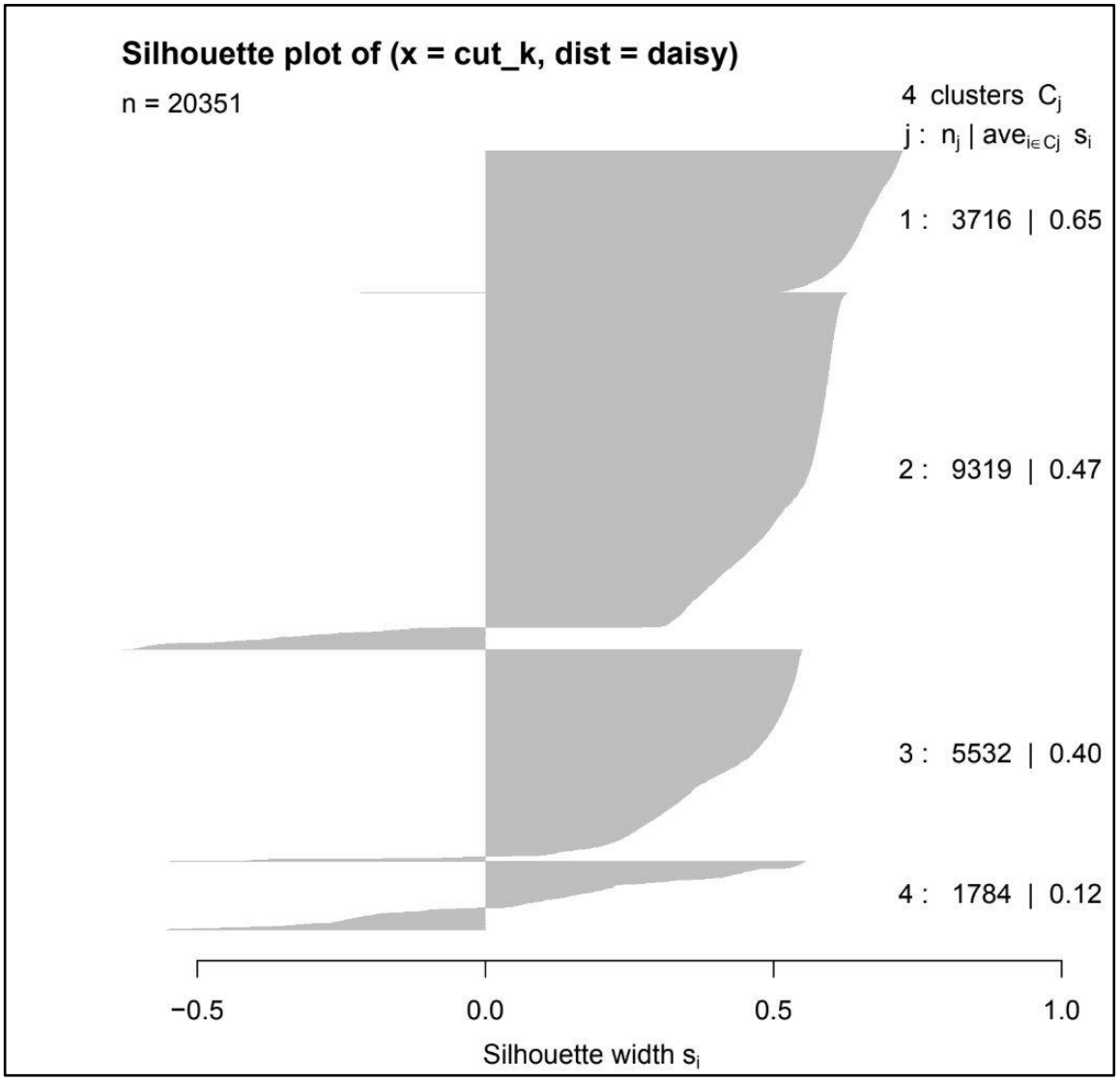

Figure 41: Silhouette average plot

Table with a description of the clusters size

To finish with this analysis of hierarchical clustering, a table with the size of each of the clusters will be displayed.

	Clusters Size
Cluster 1	3716
Cluster 2	9319
Cluster 3	5532
Cluster 4	1784

Table 10: Cluster Sizes

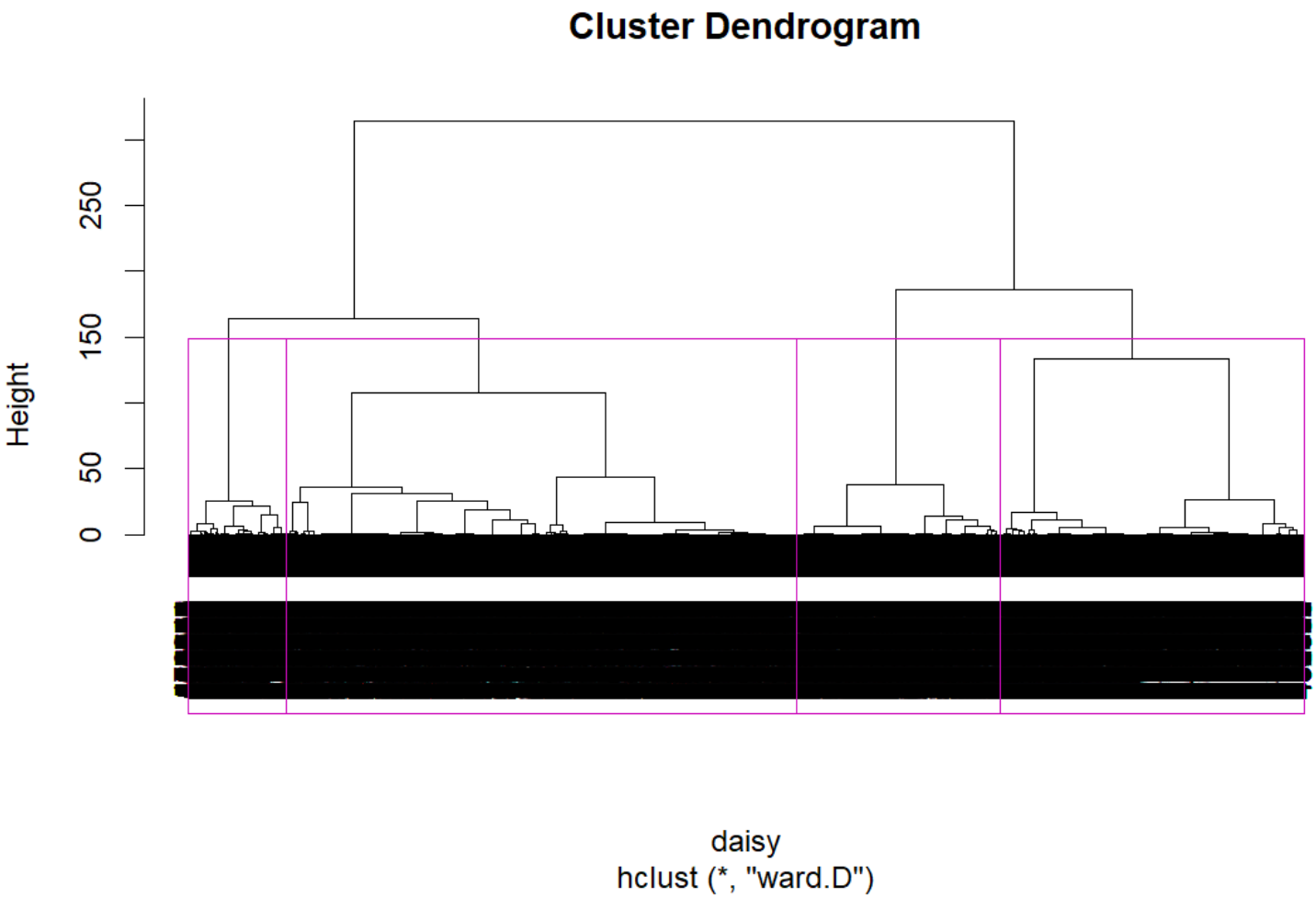

Figure 42: Resulting dendrogram

Profiling of Clusters

The following step after performing hierarchical clustering was to apply profiling. In the previous step a resulting dataset was created containing a variable cluster which labels each observation in one of the four existing clusters.

Features that have been used for profiling are Weather_Condition, Crossing, Bump, Stop, Traffic_Signal, Visibility, Precipitation, Humidity, Temperature, Distance, Pressure, Season, Severity, City, Country, Start_Time, End_Time and Month.
Variables Year and State have not been taken into account, since it has been decided to inspect only accidents recorded in 2020 in the state of California.
This resulted in a total of 18 features to describe by each cluster obtained.

For each of the six numerical variables, multiple boxplots were created to describe the distribution by each cluster. Based on the median detected within each cluster, features Humidity and Temperature are considered to be important, having the median values that clearly differentiate between clusters.

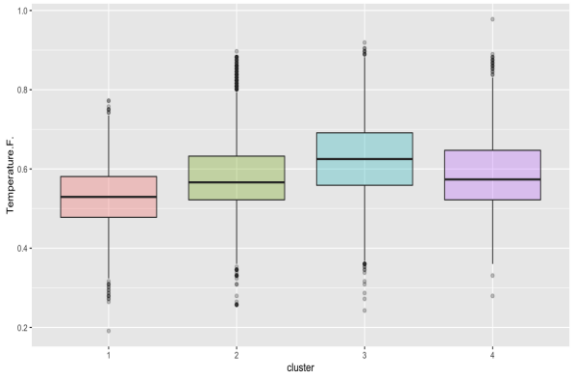

Figure 44: Temperature distribution for each cluster

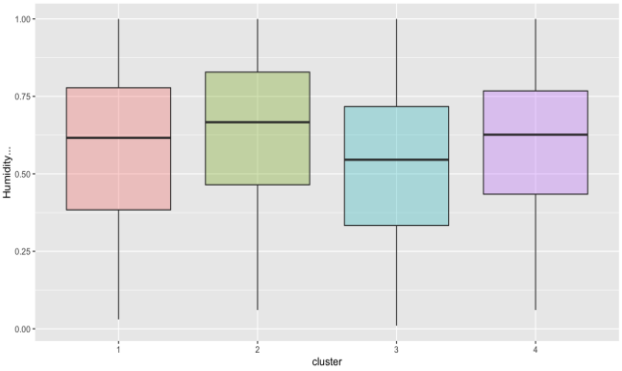

Figure 43: Humidity distribution for each cluster

Mean values have been compared between each cluster and with the global mean of that numeric variable. Created barplots for features Precipitation, as well as Distance show the most noticeable mean difference between clusters.

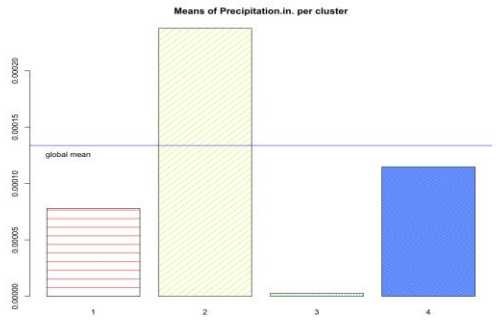

Figure 46: Meand of Precipitation in inches per cluster

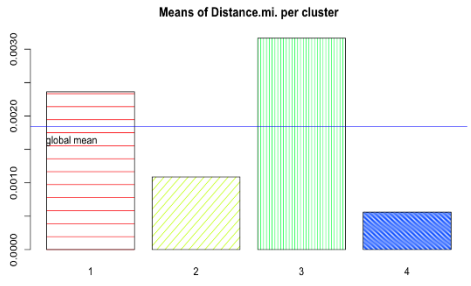

Figure 45: Means of Distance in miles per cluster

Histograms have been created for Start_Time and End_Time variables by each cluster, but the plot obtained has not shown clear differences between classes. It has been decided that these two features should be combined into one called duration, that would represent the difference between End_Time and Start_Time in minutes.

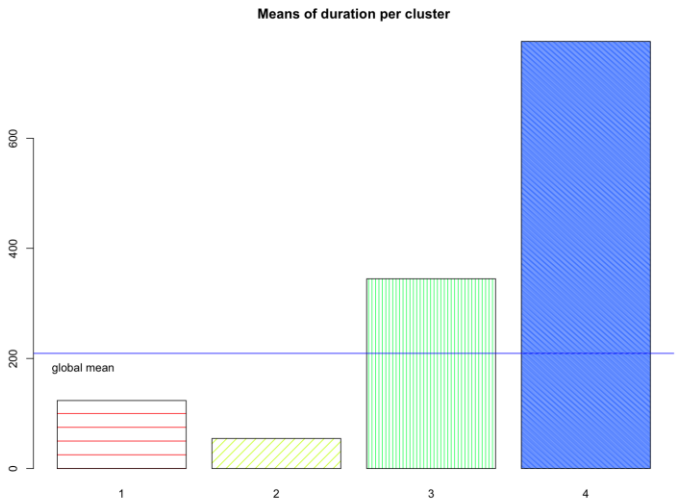

Figure 47: Means of duration per cluster

It is concluded that mean duration clearly differentiates between clusters based on the barplot obtained.

For the remaining categorical features, firstly frequency tables representing the number of occurrences of each level within each cluster have been made. Proportionally values are taken for all levels in each cluster, as well as for the same level in all clusters in order to obtain different perspectives out of snake plots.

Useful insights have been detected for variables Season, Severity and Weather Condition.

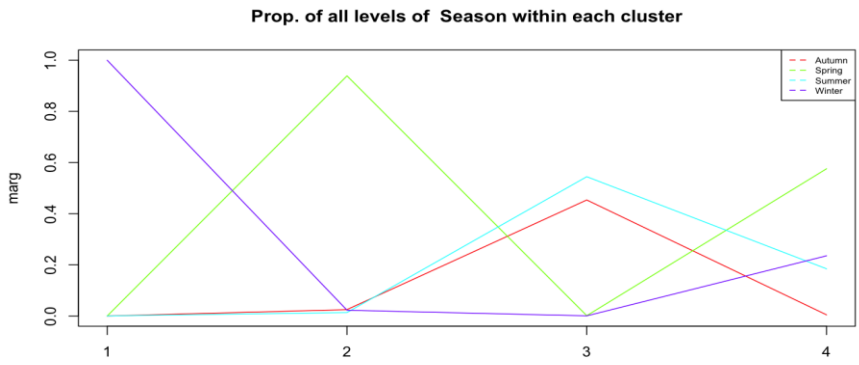

Figure 48: Proportionality of all levels of Season within each cluster

Distributions of levels in each cluster clearly show that Winter corresponds to cluster 1, as well as that level Spring is dominant in cluster 2 and 4. Cluster 3 contains levels Summer and Autumn.
Same conclusions are confirmed based on the distribution of each level of variable Season between clusters.

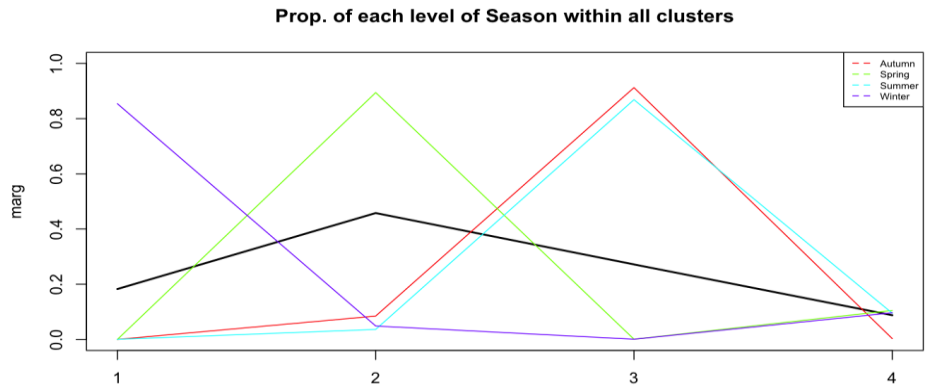

Figure 49: Proportionality of all levels of Season with all clusters

It can be also concluded that cluster 1 is the one corresponding to Severity level 2. This level is also dominant in cluster 3 and 4, whereas most of the observations in cluster 2 are labeled with severity level 1.

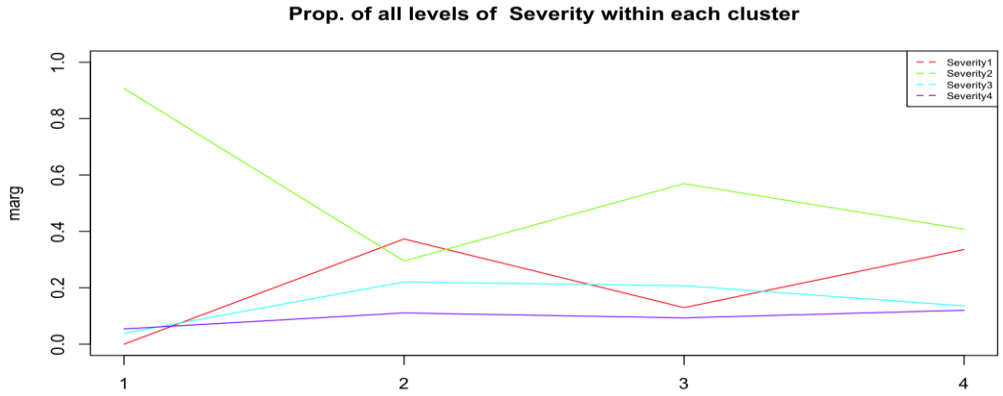

Figure 50: Proportionality of all levels of Severity within each cluster

Distribution of each Severity level through all clusters confirms dominance of Severity level 2 only in cluster 1.

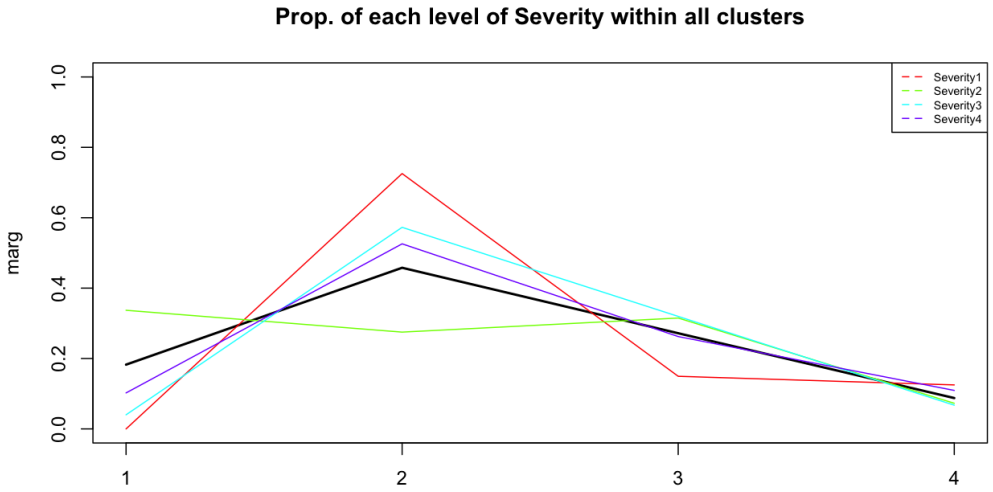

Figure 51: Proportionality of each level of Severity with all clusters

The distribution of all levels of the *Weather_Condition* variable are shown in the plot below. In this plot it can be clearly seen that among all clusters, there is one level ("Fair") that dominates all of them.

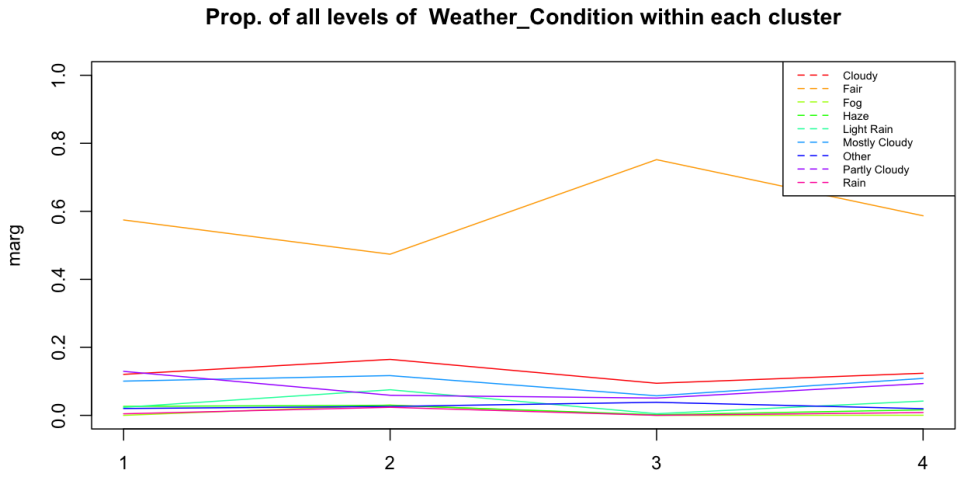

Figure 52: Proportionality of all levels of Weather Condition within each cluster

Based on the plot referring to the distribution of single level among all clusters (Fig. 53), it can be concluded that conditions Fog, Rain and Light Rain are corresponding to cluster 2.

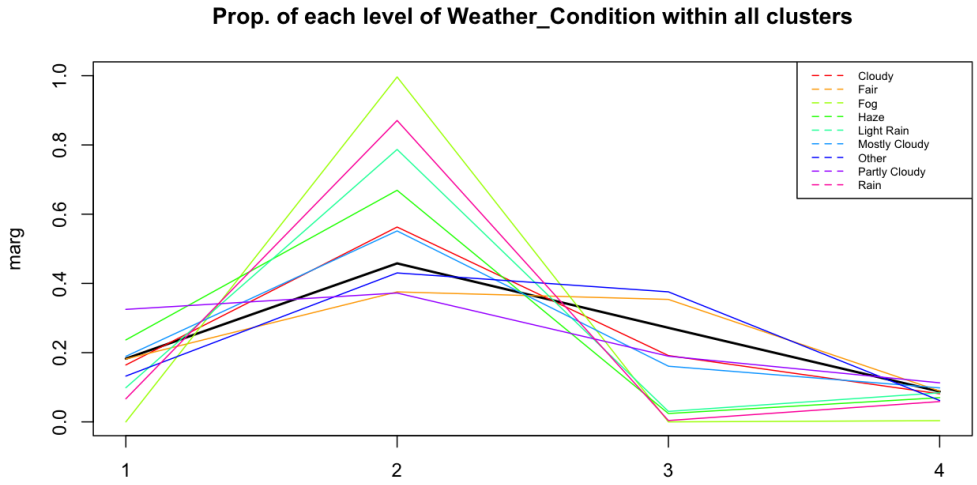

Figure 53: Proportionality of each level of Weather Condition within all clusters

Previously mentioned conclusions can also be validated with other types of plots such as barplots, but showing results in a grouped and stacked format.

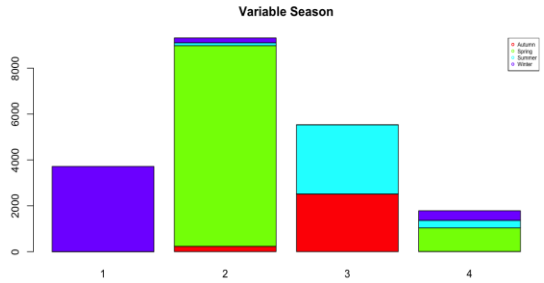

Figure 55: Contributions of Season levels per cluster

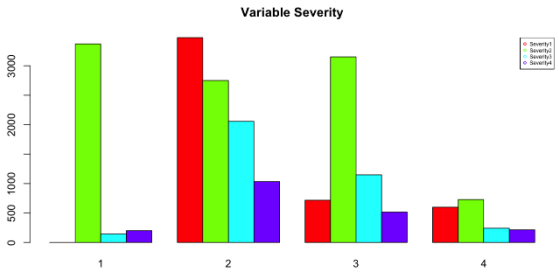

Figure 54: Contributions of Severity levels per cluster

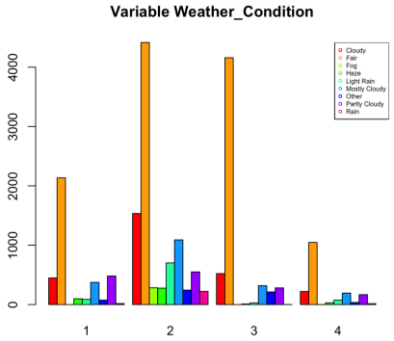

Figure 56: Contributions of Weather Condition levels per cluster

Looking at the plots below (Fig. 57), of binary variables, we can clearly see that the one that dominates the majority of the clusters is the absence of any road elements. The only clusters where this is changed is in cluster 4 for the existence of a crossing on the road and also for cluster 4, for the appearance of a traffic signal. The main difference that can be appreciated

is the fact that for crossing the dominance is still in favor of the absence of there being a crossing on the road while for traffic signals the dominance in the mentioned cluster changed in favor of the appearance of a traffic signal on the road.

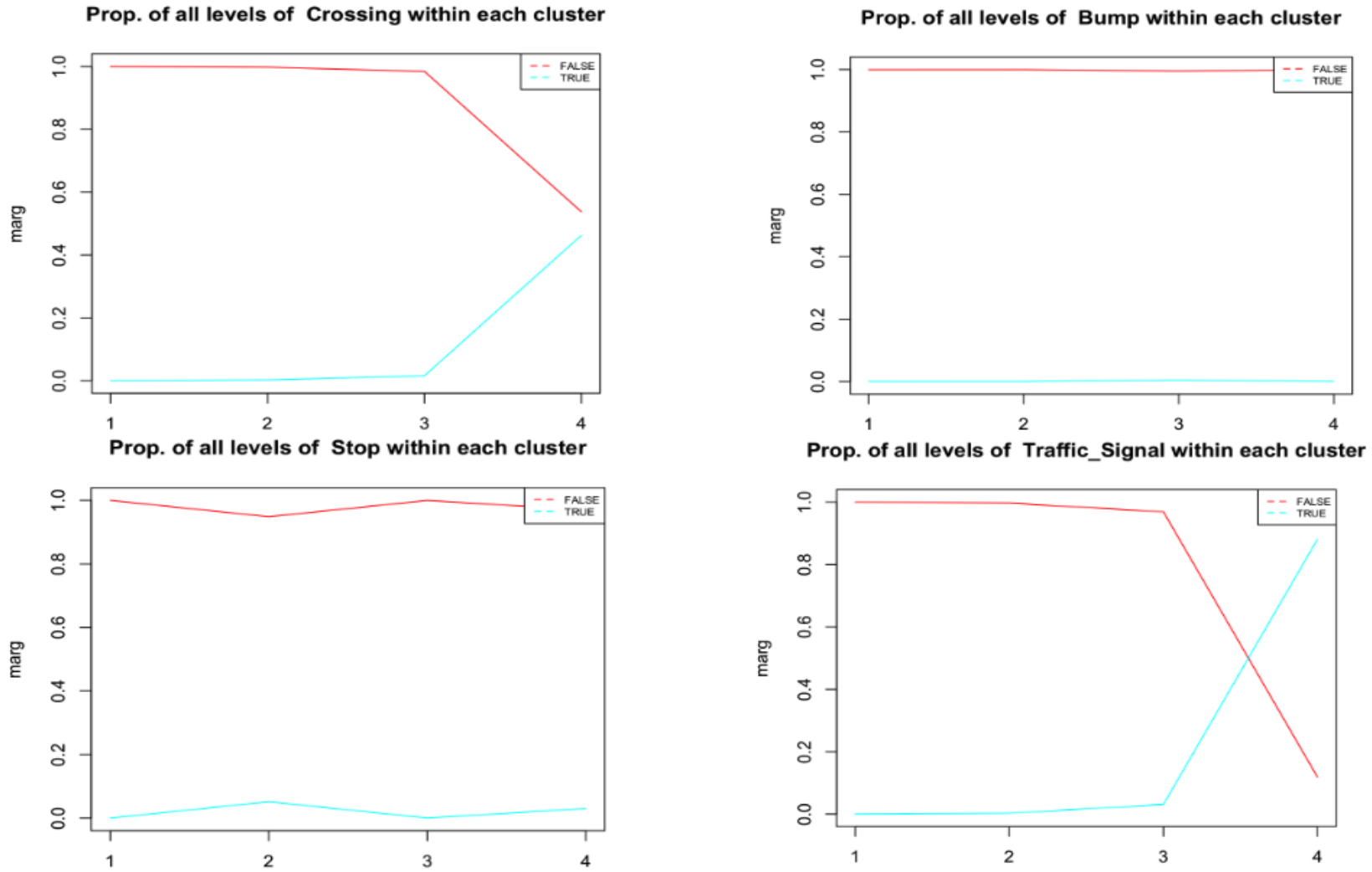

Figure 57: Proporionality of all levels of the traffic elements within each cluster

Once all graphical explanations about the most important variables for each cluster had been made, we also performed some test statistics in each of the clusters in order to obtain a ranking of the most significant variables and to have an idea of which variables are the most important for each cluster.

So as it is shown in the table 11, p-values of each of the variables for the 4 clusters obtained using our data are shown. From this we can rank the variables from most significant to less.

- Cluster 1:
  The first cluster has Distance and Precipitation as the most significant variables while Humidity, Pressure and Visibility are the less significant.
- Cluster 2:
  The second cluster has Humidity and Precipitation as the most significant variables while Temperature and Pressure are the least significant.
- Cluster 3:
  The third cluster has Distance and Visibility as the most significant variables.
- Cluster 4:
  The fourth cluster has Pressure, Distance and Visibility as the most significant variables while Temperature, Precipitation and Humidity are the least significant.

Table 12, shows the p-values of each class of each categorical variable for each cluster. The most relevant classes for each cluster are:

- Cluster 1:
  The classes that are the most relevant (the most significant) are *Partly Cloudy* (Weather Condition), *FALSE* (Crossing), *TRUE and FALSE* (Bump), *FALSE* (Stop) and *FALSE* (Traffic Signal).
- Cluster 2:
  The classes that are the most relevant (the most significant) are *Light Rain* (Weather Condition), *FALSE* (Crossing), *TRUE and FALSE* (Bump), *TRUE* (Stop) and *Severity3* (Severity).
- Cluster 3:
  The classes that are the most relevant (the most significant) are *Fair* (Weather Condition), *FALSE* (Crossing), *TRUE* and *FALSE* (Bump), *FALSE* (Stop), *FALSE* (Traffic Signal), and *Severity2* (Severity).
- Cluster 4:
  The classes that are the most relevant (the most significant) are *Fog* (Weather Condition), *TRUE* and *FALSE* (Bump), *TRUE* and *FALSE* (Stop), *Spring* (Season) and *Severity1* (Severity).

```
[1] "P.values per class: 1"
Weather_Condition          Crossing    Traffic_Signal    Temperature.F.            Season          Severity              City        Start_Time
         0.00e+00          0.00e+00         0.00e+00          0.00e+00         0.00e+00          0.00e+00          0.00e+00          0.00e+00
         End_Time             Month            County              Stop             Bump      Distance.mi. Precipitation.in.       Humidity...
         0.00e+00          0.00e+00       6.02e-295        1.55e-103         2.80e-10          2.49e-06          6.84e-06          4.52e-03
     Pressure.in.    Visibility.mi.
         1.35e-02          3.83e-02
[1] "P.values per class: 2"
Weather_Condition          Crossing    Traffic_Signal    Visibility.mi.      Distance.mi.            Season          Severity              City
         0.00e+00          0.00e+00         0.00e+00          0.00e+00         0.00e+00          0.00e+00          0.00e+00          0.00e+00
       Start_Time          End_Time            Month            County       Humidity...              Stop Precipitation.in.              Bump
         0.00e+00          0.00e+00         0.00e+00       6.02e-295         4.38e-125        1.55e-103          1.21e-55          2.80e-10
   Temperature.F.      Pressure.in.
         3.20e-09          1.86e-08
[1] "P.values per class: 3"
Weather_Condition          Crossing    Traffic_Signal Precipitation.in.       Humidity...    Temperature.F.      Pressure.in.            Season
         0.00e+00          0.00e+00         0.00e+00          0.00e+00         0.00e+00          0.00e+00          0.00e+00          0.00e+00
         Severity              City       Start_Time          End_Time            Month            County              Stop      Distance.mi.
         0.00e+00          0.00e+00         0.00e+00          0.00e+00         0.00e+00        6.02e-295        1.55e-103          1.82e-50
   Visibility.mi.              Bump
         4.30e-33          2.80e-10
[1] "P.values per class: 4"
Weather_Condition          Crossing    Traffic_Signal            Season          Severity              City        Start_Time          End_Time
         0.00e+00          0.00e+00         0.00e+00          0.00e+00         0.00e+00          0.00e+00          0.00e+00          0.00e+00
            Month            County              Stop      Pressure.in.      Distance.mi.              Bump    Visibility.mi.    Temperature.F.
         0.00e+00         6.02e-295        1.55e-103          1.24e-20         1.46e-13          2.80e-10          5.27e-08          1.48e-01
Precipitation.in.       Humidity...
         1.65e-01          3.84e-01
```

Table 11: P-values for the four clusters

Weather Condition:

Cloudy	4.129931e-03	1.471755e-32	0.000000e+00	9.658593e-02
Fair	3.302360e-01	0.000000e+00	6.903332e-208	2.070252e-01
Fog	5.551115e-16	1.802039e-75	0.000000e+00	2.216924e-07
Haze	1.990002e-03	1.480062e-18	0.000000e+00	1.002157e-01
Light Rain	1.608547e-11	8.276285e-91	0.000000e+00	3.494756e-01
Mostly Cloudy	2.094374e-01	6.927147e-19	0.000000e+00	4.095901e-02
Other	8.286560e-04	9.064103e-02	8.687382e-09	1.339294e-02
Partly Cloudy	2.192610e-49	3.318568e-12	9.181544e-14	1.878394e-04
Rain	7.102845e-07	1.007507e-40	0.000000e+00	5.064123e-02

Crossing:

	1	2	3	4
FALSE	9.941007e-50	1.000509e-164	1.944388e-36	0.000000e+00
TRUE	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

Bump:

	1	2	3	4
FALSE	3.921521e-02	1.751969e-05	2.993827e-12	1.293896e-01
TRUE	3.921521e-02	1.751969e-05	2.993879e-12	1.293896e-01

Stop:

	1	2	3	4
FALSE	8.963224e-29	0.000000e+00	4.008821e-46	1.324810e-01
TRUE	0.000000e+00	6.228552e-95	0.000000e+00	1.324810e-01

Traffic Signal:

	1	2	3	4
FALSE	1.174883e-96	0.000000e+00	1.752945e-67	0.000000e+00
TRUE	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

Season:

	1	2	3	4
Autumn	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
Spring	0.000000e+00	0.000000e+00	0.000000e+00	2.070465e-17
Summer	0.000000e+00	0.000000e+00	0.000000e+00	5.035011e-02
Winter	0.000000e+00	0.000000e+00	0.000000e+00	9.686972e-03

Severity:

	1	2	3	4
Severity1	0.000000e+00	0.000000e+00	0.000000e+00	8.047572e-26
Severity2	0.000000e+00	0.000000e+00	1.200680e-42	8.626433e-14
Severity3	0.000000e+00	1.126773e-52	6.090868e-13	1.105160e-06
Severity4	0.000000e+00	9.477560e-11	1.629930e-01	1.723493e-04

Table 12: P-values of each class of each categorical variable for each cluster

Decision Trees

In the following part of the project the decision trees for the data will be introduced as well as conclusions to which is the best decision tree.
The following parameters have been used in order to build the models:
Response variable: Severity (going from values 1 to 4, with being 4 the most severe).
Explanatory variables:

- Regarding the weather: Weather_Condition, Visibility.mi. Precipitation.in., Humidity, Temperature.F, Pressure.in, Season
- Regarding scenario of the accident: Crossing, Bump, Stop, Traffic_Signal, Distance.mi
- Regarding the place of the accident: district. It must be noted that this is a new variable added specifically for this part due to having many categories inside 'County' and this caused computation problems. This new variable groups the counties according to which district they belong[5].

The data used for this part only regards accidents that happened in year 2020 for California.
In addition, using these variables four decision trees have been created using different cp values and minsplit for the tree branches and nodes.

Tree	cp	minsplit	maxdepth
Tree 1	0.01	20	30
Tree 2	0.005	200	20
Tree 3	0.005	1000	20
Tree 4	0.0006	1000	20
Tree5	0.05	200	20
Tree6	0.5	200	20
Tree7	0.001	1000	20

Table 13: Trees' parameters

Tree	Error rate	Severity taken for the tree
Tree 1	30.26036	Severity 1, Severity 2 and Severity 3
Tree 2	29.69564	Severity 1, Severity 2, Severity 3 and Severity 4

[5] https://cwwp2.dot.ca.gov/documentation/district-map-county-chart.htm

Tree 3	30.26036	Severity 1, Severity 2 and Severity 3
Tree 4	29.49028	Severity 1, Severity 2 and Severity 3
Tree 5	34.05574	Severity 1 and Severity 2
Tree 6	76.67033	Severity 2
Tree 7	29.49028	Severity 1, Severity 2 and Severity 3

Table 14: Trees' characteristics

The same trees were gotten using the parameters on tree 7 and tree 4.

Tree 4 & Tree 7

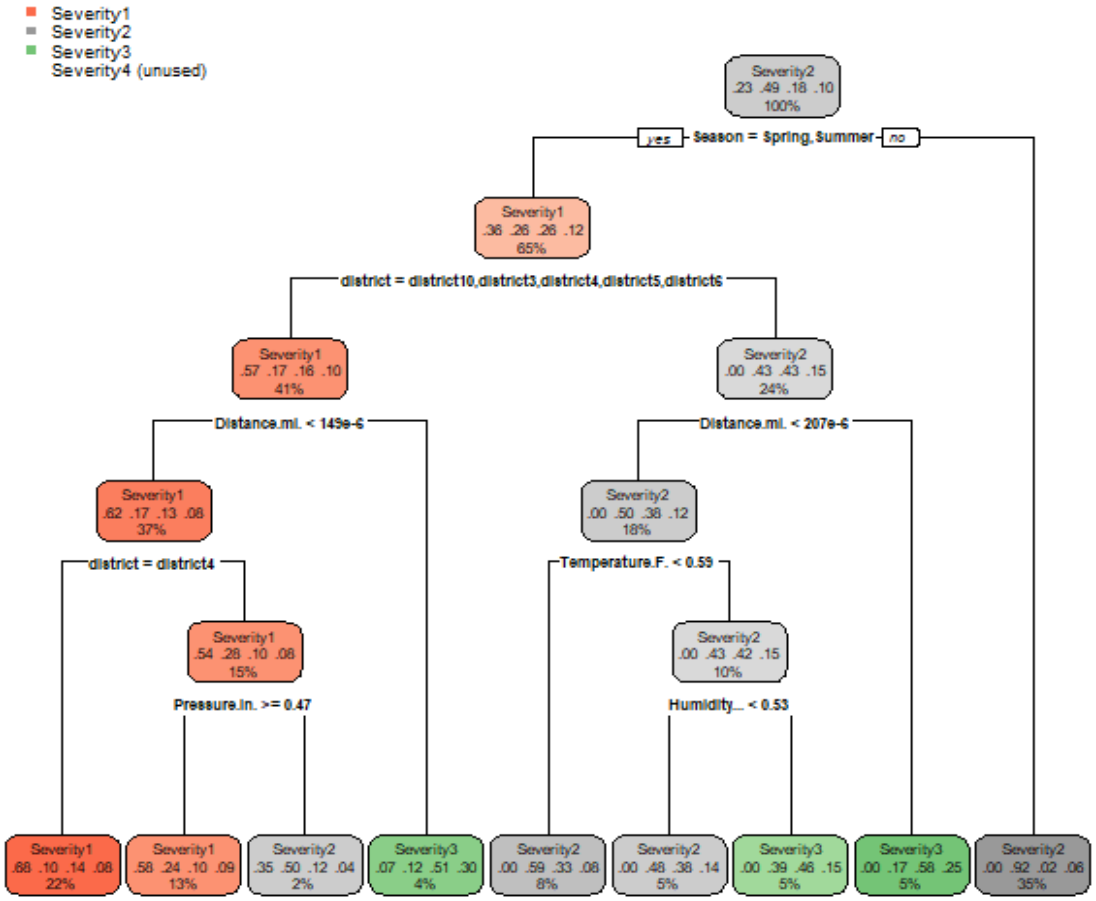

Figure 58: Resulting decision tree 4 and 7

PredClass tree 4 and 7	1	2	3
Severity 1	3029	115	37
Severity 2	705	5564	474
Severity 3	599	778	1021
Severity 4	398	466	449

Table 15: Confusion matrix for tree 4 and tree 7

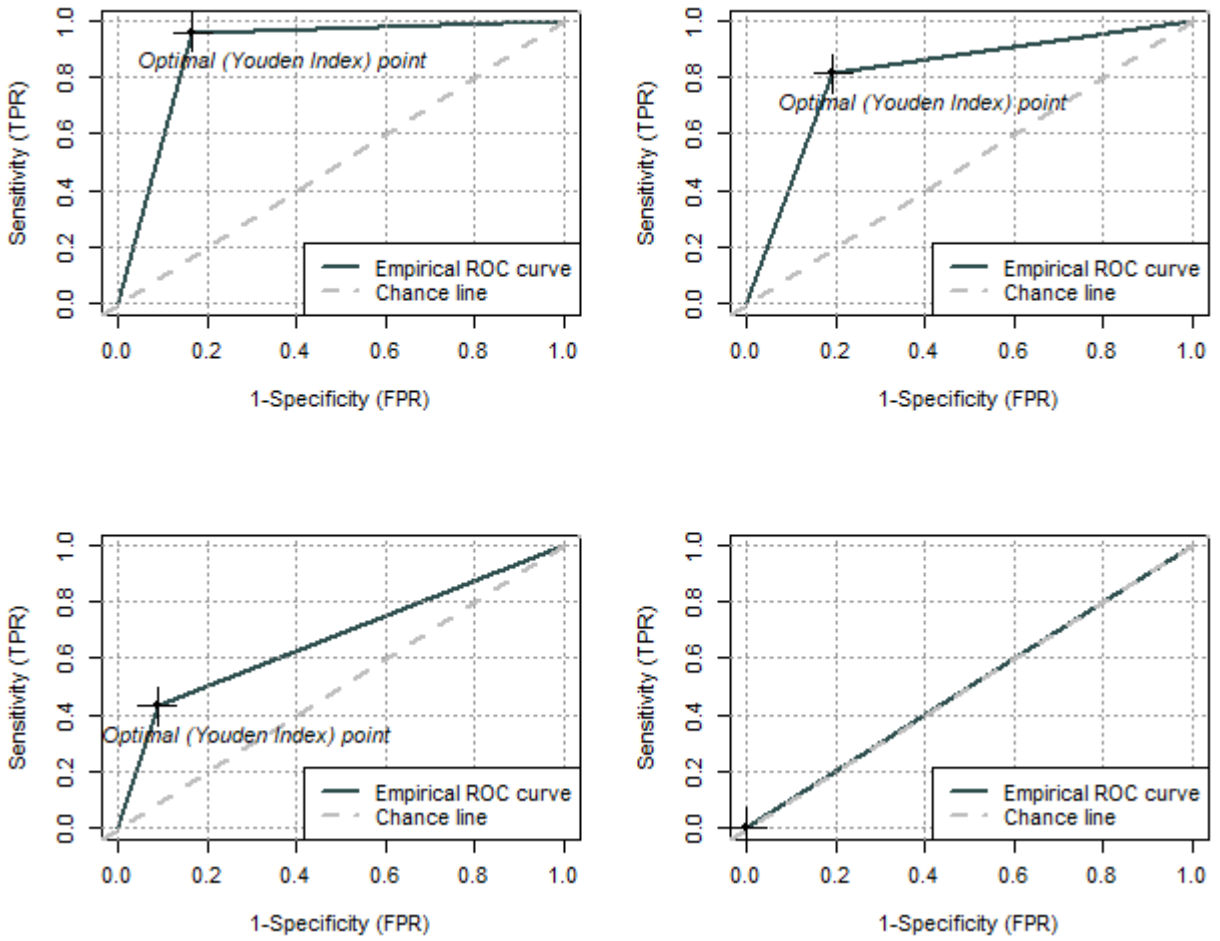

Figure 59: ROC curve for four severity values according to tree 4 and 7. Top left: ROC curve for severity 1, Top right: ROC curve for severity 2, Bottom left: ROC curve for severity 3, Bottom right: ROC curve for severity 4

Tree 2

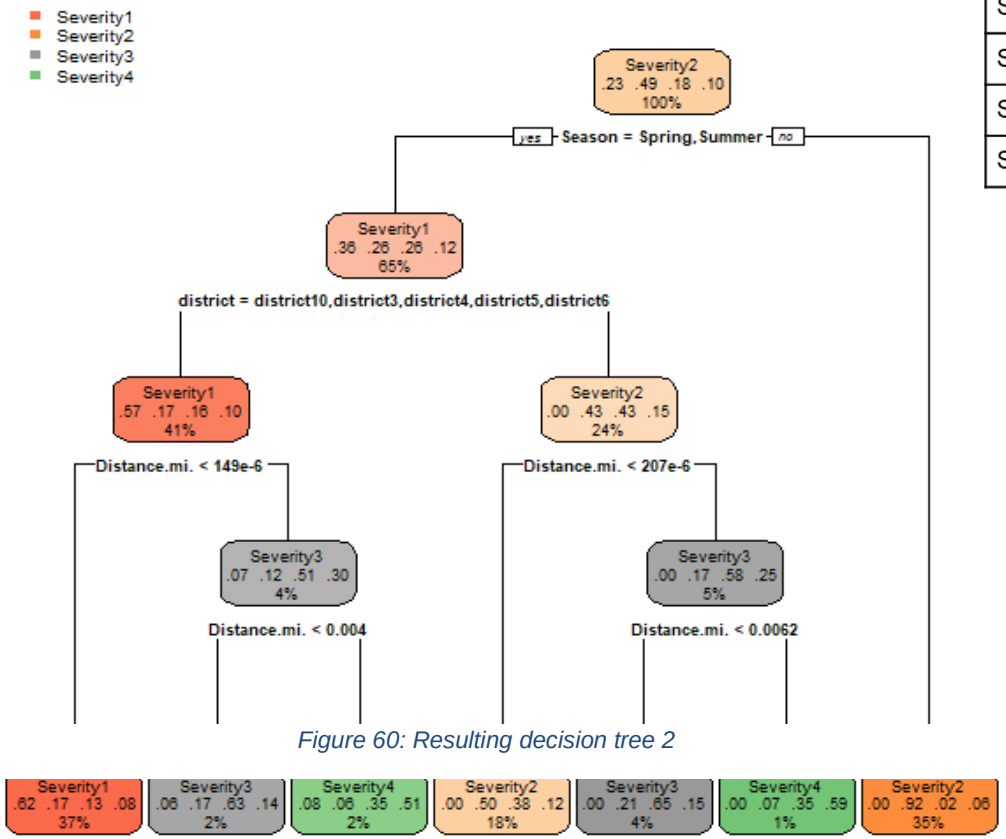

Figure 60: Resulting decision tree 2

PredClass tree 2	1	2	3	4
Severity 1	3144	0	19	18
Severity 2	870	5683	165	25
Severity 3	638	1078	547	135
Severity 4	412	565	124	212

Table 16: Confusion matrix for tree 2

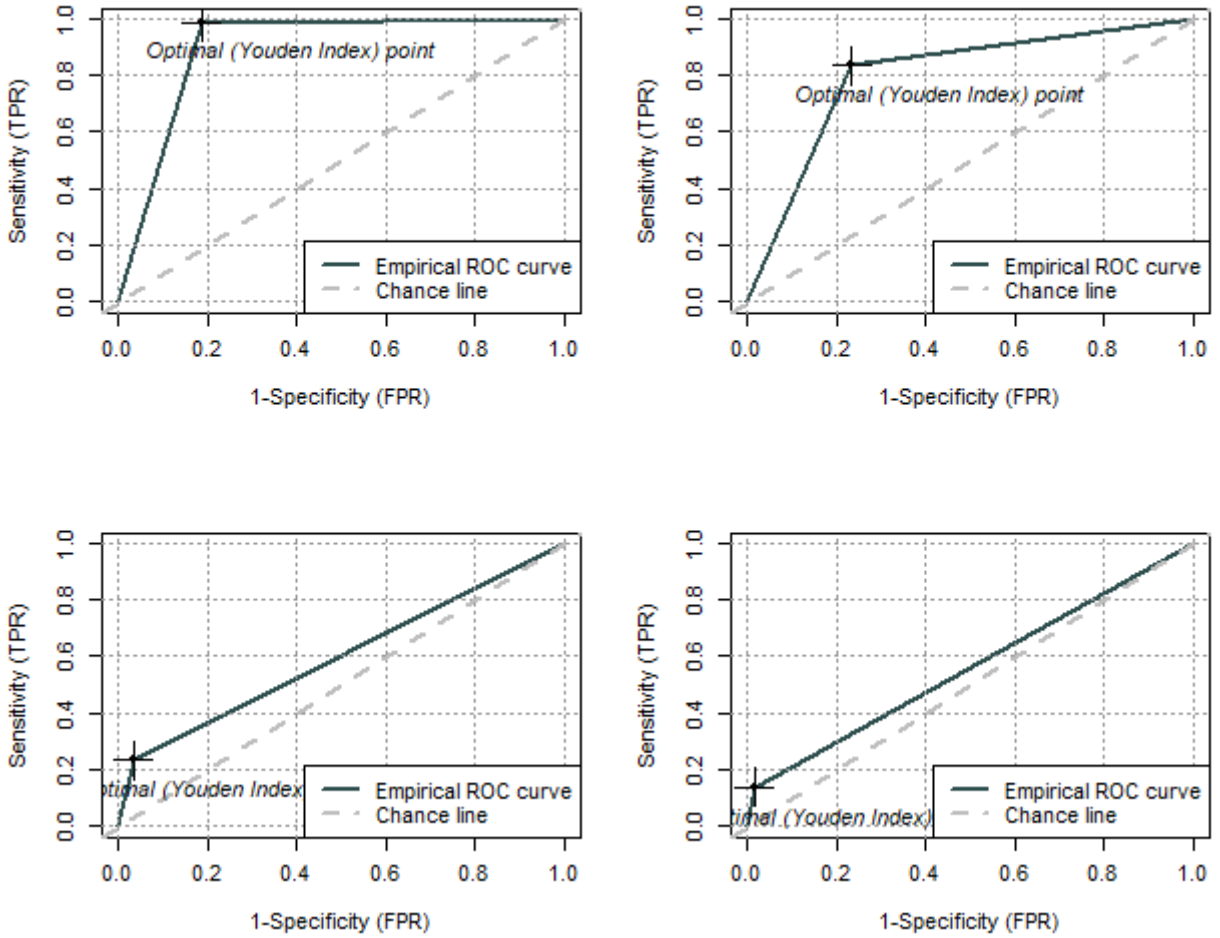

Figure 61: ROC curve for four severity values according to tree 2. Top left: ROC curve for severity 1, Top right: ROC curve for severity 2, Bottom left: ROC curve for severity 3, Bottom right: ROC curve for severity 4

Regarding decision tree 4 and tree 7 (Fig.58) , it can be seen that 35% of the accidents are directly classified as of severity 2 if the season when it's taking place is not spring nor summer. In the case of spring or summer, the rest of 65% of the accidents are classified depending on the district where they took place, distance affected by the accident and several weather conditions. In case of severity 3 accidents, which are those accidents that have had the worst impact in this tree, it can be concluded the following. When accidents affect a distance bigger than 149*10^-6 mi and happen in districts 3,4,5,6 and 10, these are classified as severity 3. Meanwhile if the distance affected is bigger than 207*10^-6 mi and has happened in districts 1,2,6,7,8,9,11 and 12 the severity value is 3. If a certain weather condition is added to the equation, it can be found that those accidents belonging to the previous mentioned districts, that have a distance affection of less than 207 *10^-6 and that the Temperature at the time of the accident is above 0,59 F and Humidity more than 0,53, severity is recorded as 3. It must be noted that Temperatures above 0,69 F (-17 ºC) in spring or summer are very likely.

Regarding the ROC curves (Fig. 59) , good AUC values can be seen regarding the severity 1 and severity 2, having AUC values of 0.89 and 0.81. Severity 3 has lower AUC values than severities 1 and 2, with a value of 0.67 and last severity 4 has an AUC value of 0.5, as it is not predicted in the tree.

Regarding decision tree 2 (Fig. 60), it can be seen that 35% of the accidents are directly classified as of severity 2 if the season when it's taking place is not spring nor summer. In the case of spring or summer, the rest of 65% of the accidents are classified depending on the district where they took place, distance affected by the accident and several weather conditions. In case of severity 4 accidents, which are those accidents that have had the worst impact, it can be concluded the following. When accidents affect a distance greater than 0.004 mi and happen in districts 3,4,5,6 and 10, these are classified as severity 4. Meanwhile if the distance affected is bigger than 0.0062 mi and has happened in districts 1,2,6,7,8,9,11 and 12 the severity value is 4. This tree does not take weather conditions into consideration.

Looking at the ROC curves (Fig 61) for the second tree we can see that the AUC values for Severities 1 and 2 are similar to the ones we had before, the ones being 0.9 and 0.8 respectively. Regarding severity 3 an AUC value of 0.6 can be seen and lastly now, for severity there is an AUC value of 0.55.

Checking for prediction powers of both trees we find the following values. Tree number 2 has an error rate for the prediction power of 30.24122 and tree number 4 and 7 have an error rate of 29.73496, which is slightly lower than for tree 2.

In conclusion, for this project, tree number 2 is selected due to being capable of classifying all four types of severity values, instead of tree 4/7 which only classifies up until severity 3. It is true that the error rate for the predicting power for tree 4/7 is lower than for tree 2, but it is preferable to be able to classify all four severities.

In addition, the decision trees do not allow us to greatly predict the severity for the accidents that surround certain characteristics of the environment. Meaning, that the weather conditions are not one of the main conditions causing the severity of an accident to increase.

It also must be added that both tree 4/7 and tree 2 take the same classification values for districts, weather conditions and instances, meaning these classifications are most likely to be the best classifiers out of the total.

Linear Discriminant Analysis - LDA

In this section of the report the implementation of the linear discriminant analysis (LDA) is taking place. The ultimate goals of this analysis are: the dimensionality reduction of the dataset's numerical explanatory variables and finally, the prediction of new observation's class, meaning Severity levels (1,2,3 or 4). LDA or discriminant function analysis is a generalization of Fisher's linear discriminant, which is used in statistics and other domains to determine a linear combination of features that describes or distinguishes between two or more classes of objects or events. In general, the resulting mixture of the algorithm can be applied as a linear classifier or, more frequently, to reduce the dimensionality prior to a subsequent classification.

Regression analysis and analysis of variance (ANOVA), which both aim to express one dependent variable as a linear mixture of other traits or measures, are closely connected to LDA. Discriminant analysis employs continuous independent variables and a categorical dependent variable, whereas ANOVA uses categorical independent variables and a continuous dependent variable (i.e. the class label). Due to the fact that similar techniques use the values of continuous independent variables to explain a categorical variable, logistic regression and probit regression are more comparable to LDA than ANOVA. In applications where it is not fair to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA approach, these other methods are preferred.

Moreover, when groups are known beforehand, discriminant analysis is employed (unlike in cluster analysis). Both a value for a group measure (Severity) and a value on one or more quantitative predictor measures are required for each case. Discriminant function analysis can be defined as the classification of entities into groups, classes, or categories of the same kind.

In this specific case, as we have previously seen during the chapters Univariate and Bivariate Descriptive Statistics of Preprocessed Variables and Profiling of Clusters , most of the numerical variables/features of the dataset do not follow a normal-distribution. At the same time, some of the assumptions that need to be satisfied in order to apply LDA to the data set are not fulfilled as well. In more detail, homoscedasticity is not present between the different groups of Severity, when only the numerical variables are taken into account, leading to the conclusion that some absent numerical variables could help the analysis and could lead to the generation of more robust predictions and results. On the other hand, for the available numerical features of the dataset, multicollinearity does not exist. Besides the fact that not all the assumptions are fulfilled, for the sake of the specific assignment instead of implementing a logistic regression model or a probit regression one, in the following subsections the implementation of LDA is presented.

Generation of the Discriminant Functions

Finding the linear combinations of the initial quantitative variables - in this case: "Temperature(F)", "Humidity(%)", "Pressure(in)", "Precipitation(in)", "Distance(mi)" and "Visibility(mi)" - that provide the best possible separation between the groups (the accident's Severity levels) is the goal of LDA.

The dataset is composed of 4 different Severity levels (1,2,3 or 4), hence there are 4 groups (G), and 6 numerical variables describing them, as it was just mentioned above (p=6). The minimum of G-1 and p, or in this case the minimum of 3 and 6, which is 3, determines the maximum number of useful discriminant functions that can distinguish the accidents by severity. Therefore, utilizing the 6 available numerical variables in the data, it is possible to discover a maximum of 3 valid discriminant functions to differentiate the accidents by severity level.

The same approach was followed in this part of the project, as it was completed for the Decision Trees as well. The data used for this part only regards accidents that happened in the year 2020 for California (20351 observations in total). This decision was taken in order to enable the comparison of the LDA model with the decision tree models implemented beforehand. Those observations, before applying any change or processing technique, are described by the a-priori probabilities of the Table 17.

Severity Levels	**Severity1**	**Severity2**	**Severity3**	**Severity4**
A-priori Probabilities	0.23	0.5	0.17	0.1

Table 17: A-priori Probabilities of Severity Groups

Those probabilities describe the number of occurrences for each Severity level in the initial dataset (California 2020 dataset). As one can understand, most of the cases are derived from accidents of severity level 2 followed by accidents of severity levels 1,3 and 4 respectively. In addition, Table 18 describes the mean values of all the 6 numerical explanatory variables used in this part, for each Severity group of the dataset.

Group Means	**Visibility**	**Precipitation**	**Humidity**	**Temperature**	**Distance**	**Pressure**
Severity1	0.086	-0.032	0.097	0.035	-0.221	0.317
Severity2	-0.073	0.010	-0.015	-0.182	0.018	-0.110
Severity3	0.059	0.019	-0.061	0.399	-0.036	-0.045
Severity4	0.058	-0.009	-0.043	0.120	0.509	-0.120

Table 18: Severity Group Means of Numerical Variables

From the presentation of the mean values of the numerical variables of the dataset, it is clear that there are not big distinctions between the values of different groups. It is important to state here that the numerical variables were normalized before being used for this part of the analysis.

To continue with, the first necessary step was to split the available data set into training (13635 observations) and testing sets (6716 observations). The algorithm was trained by using the training set and tested with the testing set. Firstly by applying the LDA algorithm over the

numerical variables of the training dataset 3 discriminant functions are returned as a result. After the implementation of the algorithm, the existing separation of the different severity groups of the dataset is described by 62% from LD1, 30% from LD2 and 8% from LD3. For this reason, the first two discriminant functions are able to describe a total of 92% of the separation existing in the 4 groups of severity levels, and the analysis following will be focused only on those two. In order to make conclusions about which numerical variable plays the most significant role in each linear function, the standardization of the numerical variables took place before applying LDA. In that way, the loadings of the linear functions produced are easier to interpret. The loadings of the functions are presented in Table 26: Loadings of linear discriminant functions produced by LDA in the Appendix.

For convenience, the value for each discriminant function (eg. the first discriminant function) are scaled so that their mean value is zero. Note that these loadings are calculated so that the within-group variance of each discriminant function for each group is equal to 1. From the loadings of the linear discriminant functions it is obvious that, for the 1st linear discriminant (LD) Humidity and Temperature play the most significant roles by contributing positively to the result of the function, followed by Visibility with a big difference. For the 2nd LD, Distance is the variable contributing the most (with a positive value) followed by Pressure and Temperature respectively, when the absolute value is taken into account. In this case, pressure is contributing negatively to the final result, in comparison to the Temperature loading. Finally, for the 3rd LD, again, Pressure and Distance achieve the highest loadings and that's why LD3 does not provide sufficient separation between the groups (only 8% of the total separation).

Therefore, it seems that LD1 could represent an axis indicating weather conditions with similar temperature and humidity levels and slightly smaller visibility levels. On the contrary, LD2 could represent an axis describing cases where the distance of the accident plays a huge role, while the temperatures have quite similar values with the distance, but slightly smaller, and pressure values contribute contrarily. Providing a name to the produced LD axes is not so easy, while the combination of variables describing the separation in each one of them, do not form a logical situation which can lead to the generation of a descriptive condition.

A nice way of displaying the results of LDA is to make a stacked histogram of the values of the discriminant function for the different groups. In Figure 62 stacked histogram for LD1 is presented.

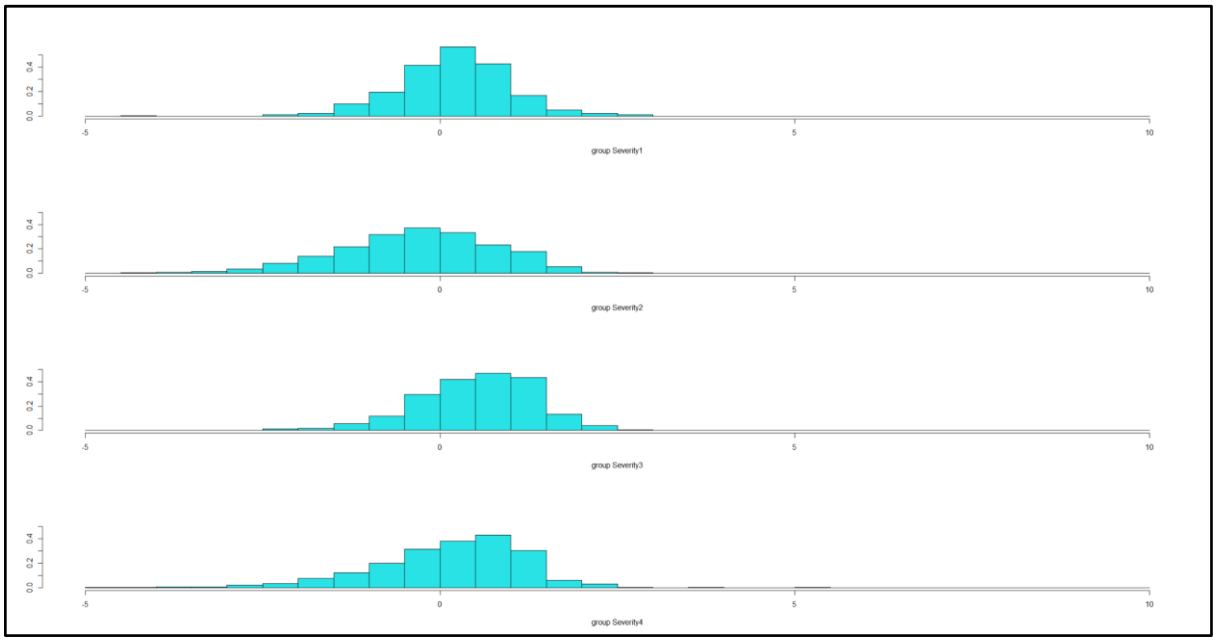

Figure 62: Stacked Histogram of the 1st Discriminant Function's Values

As it is illustrated, besides the fact that the first discriminant function describes the 62% percent of the total separation for the severity groups based on the numerical data, it is clear that the explanatory variables cannot achieve a great separation between the different target groups. The same results are true for LD2 and LD3, for which, their stacked histograms are located in the Appendix.

In order to make the conclusion more clear, the projection of the initial data to the LD1 and LD2 axis has been conducted and is presented in Figure 63.

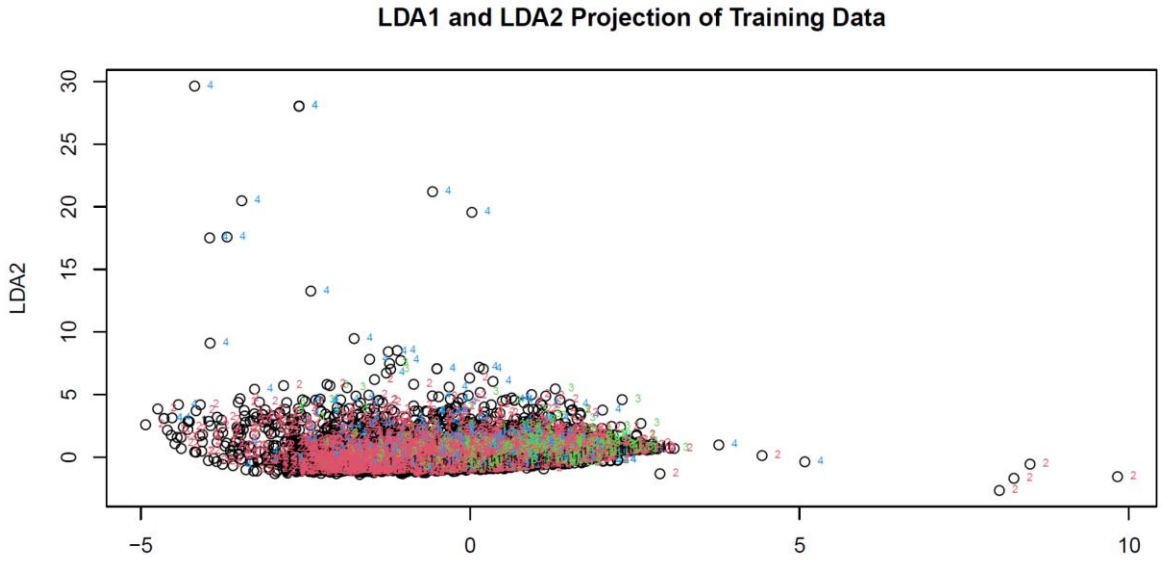

Figure 63: LDA1 and LDA2 Projection of Training Data

From the illustration, it is obvious that the discriminant functions do not achieve a clear separation between the groups of severity levels. The cloud of points close to the two LD axes consist of mixed data (from all severity levels) indicating the poor separation achieved by the algorithm. The reason behind that result, as it was concluded during the previous steps of the analysis, is the low explanatory power of the independent variables of the dataset and the unbalanced initial data (containing approximately only Severity 2 accidents). The projection of the training data to all the combinations of axes based on the 3 LDs produced can be found in the Appendix.

Classification

In this section, instead of using the mean values of the discriminant functions, or the group centroids, in order to create rules for classifying new observations, the posterior probability approach was used. During this approach an estimated probability is computed for each observation indicating to which particular class it is more possible to belong. The calculation of the posterior probabilities, as they are called, use the Gaussian distribution. In order to achieve that, conditional probabilities, as shown below, are used to generate the estimates for class probabilities.

$$\hat{P}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{c=1}^{K} e^{\hat{\delta}_c(x)}}$$

Thus, in order to classify a new observation, a closer look is taken at the posterior probabilities of the observation and the class with the highest probability is assigned to it.

Confusion Matrix and AUC for Training Set

Furthermore, after implementing the above technique, the predictions of the training set are computed. After the generation of the predictions the construction of the training confusion matrix is conducted. The training confusion matrix is presented in Table 19.

	Observed			
Predicted	**Severity1**	**Severity2**	**Severity3**	**Severity4**
Severity1	21	33	21	3
Severity2	2982	6264	2007	1160
Severity3	174	398	345	97
Severity4	4	48	25	53

Table 19: Confusion Matrix of Training Set

By using the confusion matrix the calculation of validation measures for the quality of the predictions is enabled. The accuracy level of the LDA model described before is 0.49, for the training set. For the calculation of the accuracy the sum of the true positive values of the confusion matrix is used divided by the total number of the data set. The accuracy indicates

that a random classifier may achieve better results in this specific case. At the same time, the misclassification rate is computed to be 0.51 (1-accuracy), leading to the conclusion that the model predictions are really poor. All measures, depicted in Table 19 were calculated for each distinct class of the training test, including sensitivity, specificity, precision, recall, f1-score. Most of them indicate poor results, except from sensitivity and recall for class Severity 2. By checking on the other hand the f1-score for Severity 2, one can understand that even for that level the predictions are not of a good quality.

	Class: Severity1	Class: Severity2	Class: Severity3	Class: Severity4
Sensitivity	0.006602	0.9290	0.14387	0.040366
Specificity	0.994548	0.1078	0.94046	0.993751
Pos Pred Value	0.269231	0.5046	0.34024	0.407692
Neg Pred Value	0.766910	0.6080	0.83733	0.906701
Precision	0.269231	0.5046	0.34024	0.407692
Recall	0.006602	0.9290	0.14387	0.040366
F1	0.012887	0.6540	0.20223	0.073458
Prevalence	0.233297	0.4945	0.17587	0.096296
Detection Rate	0.001540	0.4594	0.02530	0.003887
Detection Prevalence	0.005721	0.9104	0.07437	0.009534
Balanced Accuracy	0.500575	0.5184	0.54217	0.517058

Table 20: LDA Model Measures for Training Set

For illustration purposes only, in Figure 64 the ROC curve for each combination of severity levels is taking place. In order to illustrate the ROC curves in this specific case, in which we have 4 different levels of Severity, the calculation of the measures (true positive - TP, true negative - TN, false positive - FP, false-negative FN) which lead to the AUC numbers need to be calculated in groups of Severity levels. During this implementation the usage of sensitivity and specificity was conducted. In order to make it more clear, sensitivity is equal to recall or true positive rate (TPR), and is given by the division of TP cases and the sum of TP and FN cases. On the other hand, specificity is the true negative rate (TNR) and is calculated by dividing TN cases by the number of TN and FP cases.

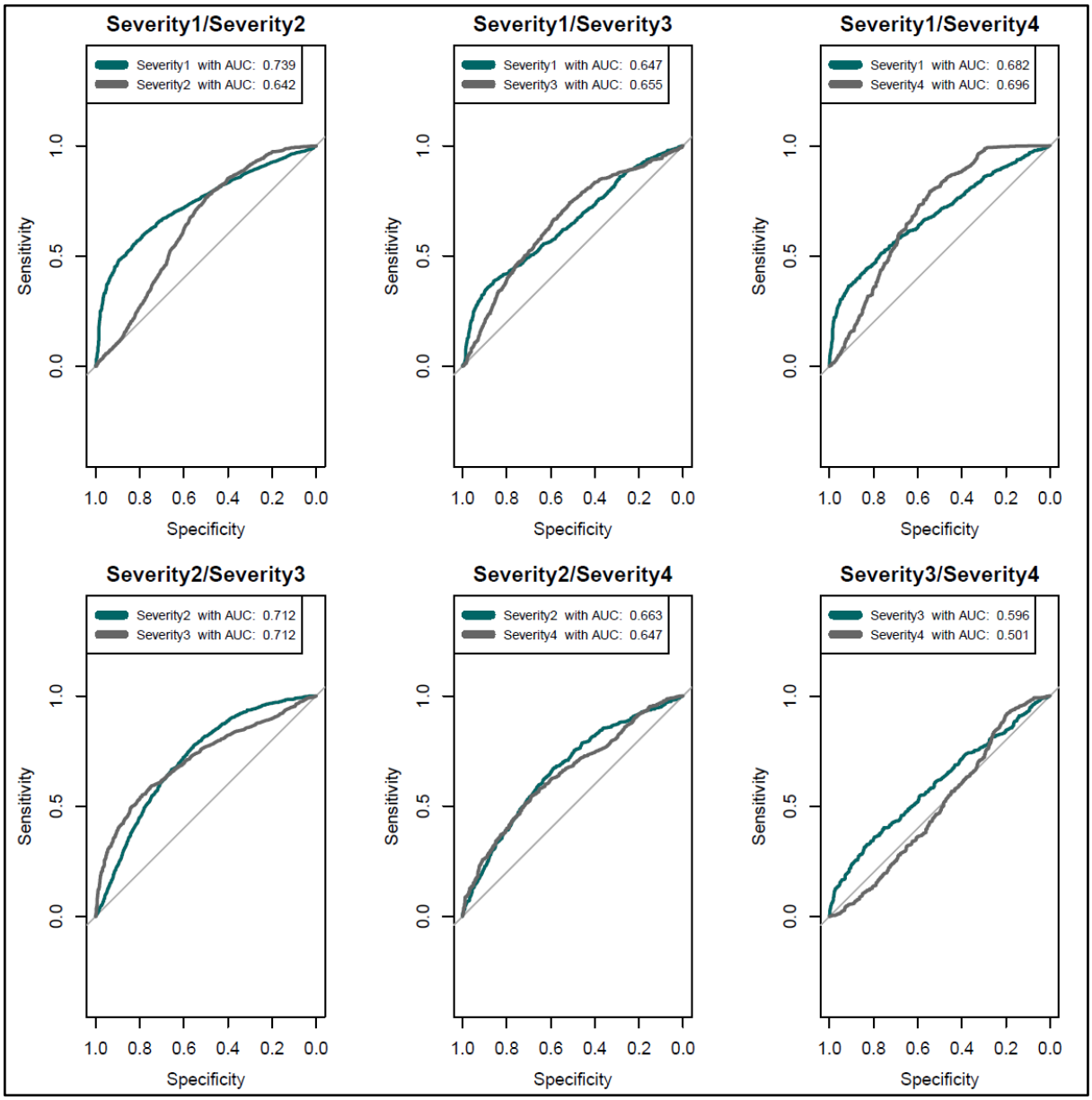

Figure 64: ROC Curves for All Severity Levels of the Training Set

From the plot, it is obvious that the best values for the AUC measure take place only when Severity 2 is presented, followed by Severity 3 and Severity 1. At the same time, the poorness of the predictions is illustrated while the AUC for some severity levels is really close to 0.5.

Confusion Matrix and AUC for Testing Set

In order to check the performance of the model, the same approach as before was followed. In more detail, the calculation of the confusion matrix and the AUC took place for the testing set (6716 observations). Table 21 presents the confusion matrix of the testing set.

	Observed			
Predicted	**Severity1**	**Severity2**	**Severity3**	**Severity4**
Severity1	18	22	10	3
Severity2	1487	3012	1019	550
Severity3	107	190	151	59
Severity4	1	33	13	41

Table 21: Confusion Matrix of Testing Set

The accuracy level of the LDA model for the test set reaches only 0.48. At the same time, the misclassification rate is computed to be 0.52, leading to the conclusion, again, that the model predictions are really poor. All the measures were calculated again for each distinct class of the testing set, and they are presented in Table 22.

	Class: Severity1	Class: Severity2	Class: Severity3	Class: Severity4
Sensitivity	0.011159	0.9248	0.12657	0.062787
Specificity	0.993141	0.1165	0.93554	0.992248
Pos Pred Value	0.339623	0.4964	0.29783	0.465909
Neg Pred Value	0.760618	0.6219	0.83218	0.907664
Precision	0.339623	0.4964	0.29783	0.465909
Recall	0.011159	0.9248	0.12657	0.062787
F1	0.021609	0.6460	0.17765	0.110661
Prevalence	0.240173	0.4850	0.17764	0.097230
Detection Rate	0.002680	0.4485	0.02248	0.006105
Detection Prevalence	0.007892	0.9035	0.07549	0.013103
Balanced Accuracy	0.502150	0.5206	0.53106	0.527518

Table 22: LDA Measures for Testing Set

By taking a closer look at the results of the performance indicators, the same conclusions can be drawn for the testing set, as for the training set. Some indicators (sensitivity and recall) achieve great values for group Severity2 but in general the model is not performing well.

The same conclusion can be derived from the Figure 65 presented below, containing the ROC curves per Severity levels combination, as it was completed for the training set before.

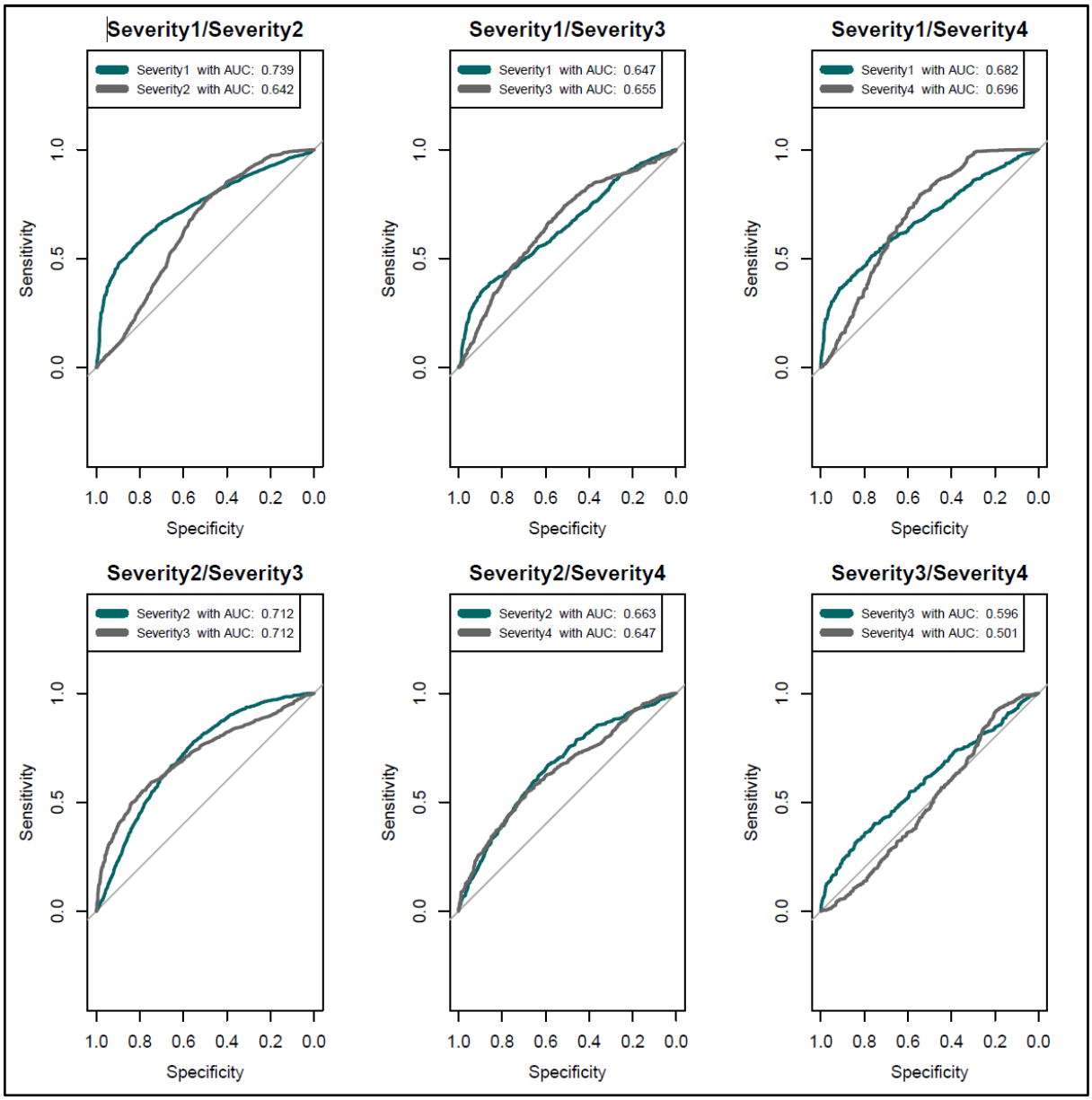

Figure 65: ROC Curves for All Severity Levels of the Test

Discussion and Conclusions

Regardless of the big quantity of records in the data, and the conclusions that have been presented for each part of the project stating the poor decisive results on the data, these are the following conclusions and discussion guidelines for the project. On the one hand, the dataset is a heavy dataset with a lot of records, that is the cause of reducing our dataframe for most of the analysis and some preprocessing steps, such as deciding to cut on the number of states used, deciding to cut on the year of occurrence used as well as only selecting a single state for further clustering and classification. Weather condition variables and counties have also suffered from grouping to reduce dimensionality. In addition, regardless of the variety of variables in the dataset, the variance of the data is poorly explained with these variables. Once these decisions have been taken and the preprocessing steps have been done, the various analyses show that the majority of accidents in the USA and in particular in California in the last year that appears on this dataset's records, are recorded as of severity 2. Also, the most recorded accidents are highly linked with deficiency of control of transit indicators, but this does not mean that the absence produces the accident. It only states that in the place of the accident, there were mostly no traffic signs. Furthermore, and surprisingly, despite the vast amount of the data, the most frequent weather condition in all the accidents has been 'Fair', despite having different seasons and states recorded in the dataset. In addition, regarding dimensions and their correlations with different variables, fog, rain and humidity were found to be correlated between each other. On the other hand, it can also be concluded that the best variables to classify the data into a prediction are distance and district in the state of California in the last year, but again, due to error rates in predictor powers, this is not concluding enough.

To summarize, this dataset has been challenging to manage and to extract firm conclusions as most analyses have not been clear enough on the results they offered. An important point has been concluded and that is that weather conditions and traffic signs do not condition in an important way the severity of the accident. It is true that depending on the location more accidents appear, but the reasoning for this can not be analyzed with this dataset. In addition, and according to studies, the reasoning behind an accident is strongly correlated with human actions and characteristics before any weather condition factor or road variable. For future analyses this dataset would be more concludent if some human variables were added to it, such as speed of the vehicle at the time or state of the driver.

Gantt Diagram and Task Distribution Grid

Initial

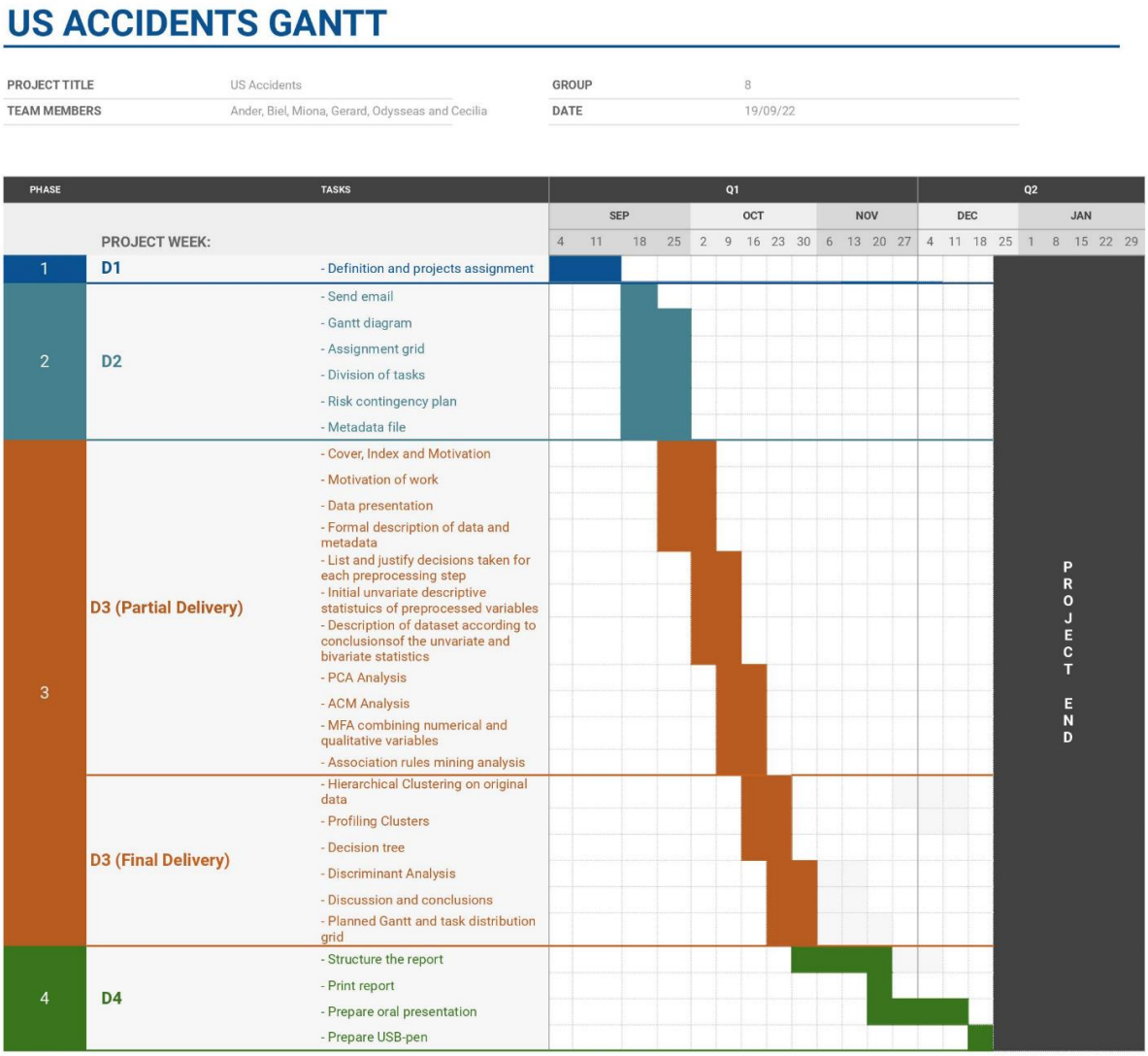

Figure 66: Gantt Initial Diagram of the project's tasks' time consumption

Final Gantt Diagram

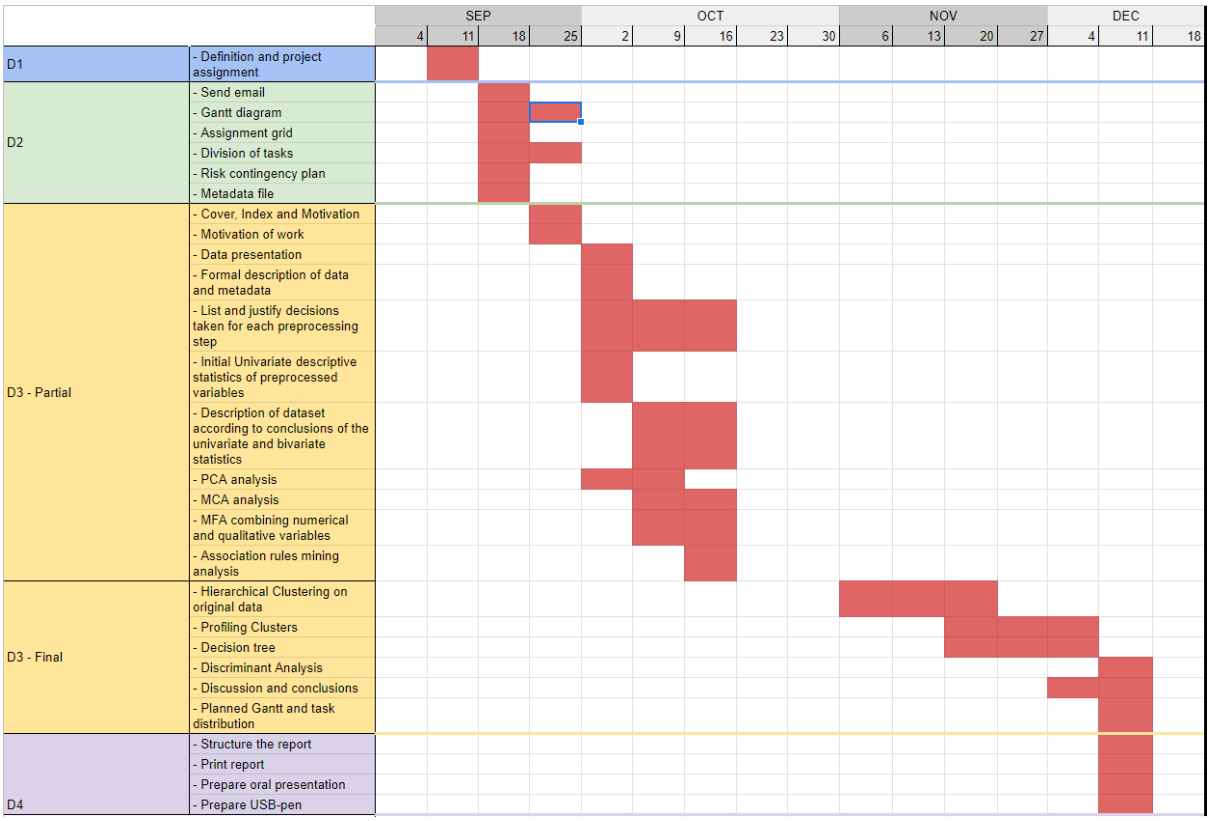

Figure 67: Final Gantt Diagram

Assignment of tasks Grid

	Particiant	Ander	Odysseas	Miona	Biel	Cecilia	Gerard
D1	Definition and project assignment	X	X	X	X	X	X
D2	Send email					**X**	
	Gantt diagram		X	**X**	X		
	Assignment grid	X	X	X	X	**X**	X
	Division of tasks	X	X	X	X	X	**X**
	Risk contingency plan	X	**X**	X	X	X	X
	Metadata file	**X**				X	
D3_Partial	Cover / Index / Motivation	**X**	X				
	Motivation Work	X	**X**				
	Formal Description of Metadata					X	**X**
	List and justify all the steps taken for the preprocessing step					X	**X**
	Descriptive univariate descriptive statistics of preprocessed variables			**X**	X		
	Description of the dataset acording to the main conclusions acording to the univariate and bivariate statistics			**X**	X		
	PCA Analysis	X	**X**				
	ACM Analysis			X	**X**		
	MFA combining numerical and qualitative					X	**X**
	Association rules mining analysis	X	X				
D3_Final	Hierarchical Clustering on original data	**X**	X				
	Profiling Clusters			**X**	X		
	Decision tree					X	**X**
	Discriminant Analysis	X	**X**				
	Discussion and conclusions			X	**X**		
	Planned Gantt and task distribution grid					**X**	X
D4	Structure the report	X	X	X	**X**	X	X
	Print report						X
	Prepare oral presentation	X	**X**	X	X	X	X
	Prepare USB-pen	**X**					X

Table 23: Assignment of tasks

Appendix

Plots Produced with Reduced Level of Insights

In this section of the appendix, plots which were generated during the analysis process of the dataset but did not provide quality levels of insights are placed. The different redundant plots are categorized based on the step of the analysis where they occurred.

PCA

All the plots added in this section have PC1 and PC3 as the selected axes for the projection. First plot (Fig. 68) will make reference to individual projections, second one (Fig. 69) to the interpretation of relationships among the variables observed and the PCs and the last one to biplot (Fig 70)

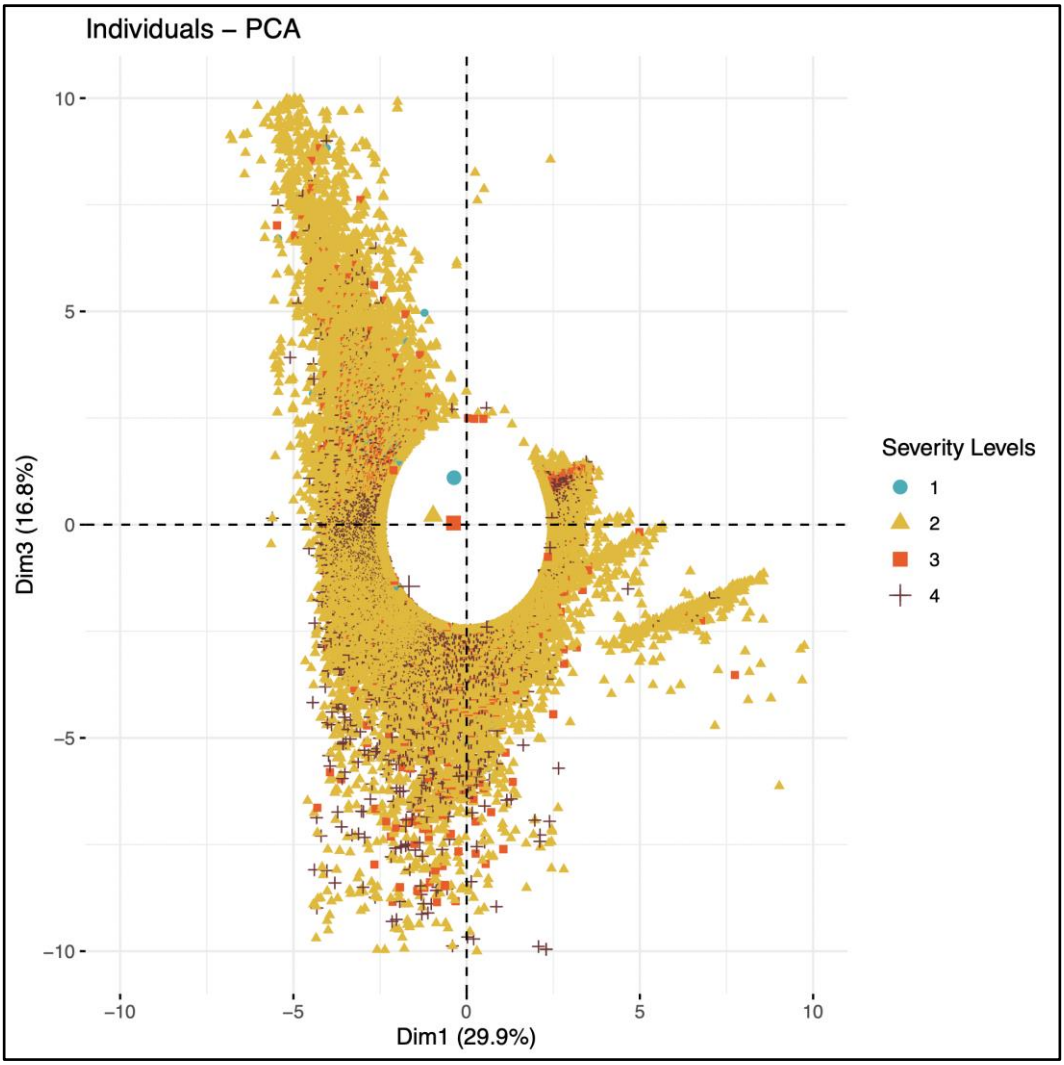

Figure 68: Individuals projections with coloring based on Severity taking a sample of 100000 units due to a large number of observations in dimensions 1 and 3 (zoomed)

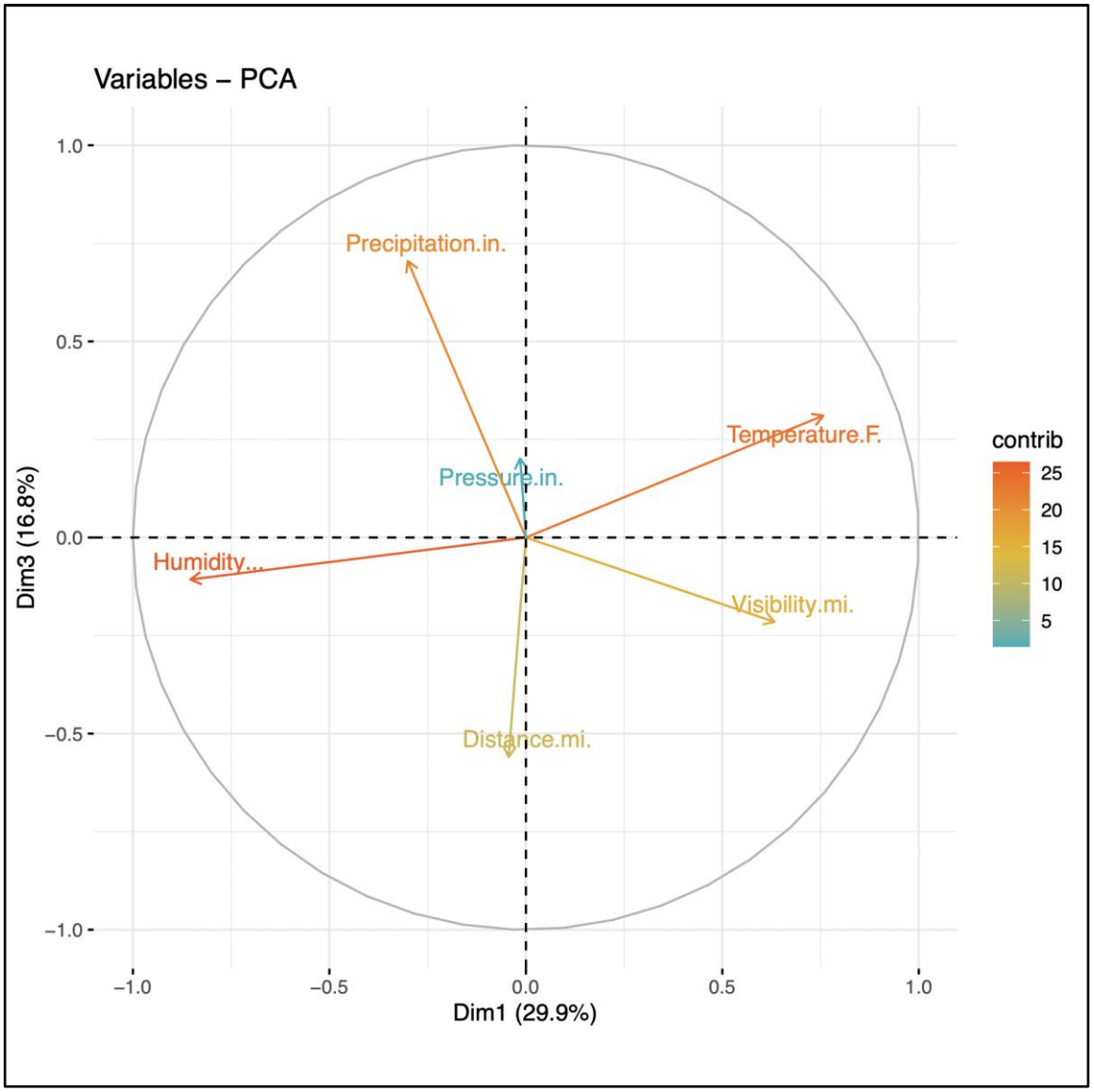

Figure 69: Variables factor graph in dimensions 1 and 2

Figure 70: Biplots Containing Individuals and Variable Factor Maps colored and grouped based on Severity levels in dimensions 1 and 3

Association Rules

In this section, tables containing frequent itemsets and rules are presented, as well as some plots illustrating rules based on different measures like support, confidence or lift. The tables and the plots are placed here in order to provide more details to the reader if necessary.

Itemsets	Support	Count
{Severity=Severity2,Visibility.mi.=Good,Precipitation.in.=None}	0.834	884825
{Severity=Severity2,Precipitation.in.=None,Distance.mi.=Small}	0.791	839372
{Visibility.mi.=Good,Precipitation.in.=None,Distance.mi.=Small}	0.763	808775

Itemsets	Support	Count
{Severity=Severity2,Visibility.mi.=Good,Distance.mi.=Small}	0.740	784784
{Severity=Severity2,Visibility.mi.=Good,Precipitation.in.=None, Distance.mi.=Small}	0.726	770217
{Year=2021,Severity=Severity2,Precipitation.in.=None}	0.544	576958
{Severity=Severity2,Precipitation.in.=None,Pressure.in.=Low}	0.502	532478
{Year=2021,Severity=Severity2,Visibility.mi.=Good}	0.501	531424
{Year=2021,Visibility.mi.=Good,Precipitation.in.=None}	0.497	526800
{Weather_Condition=Fair,Visibility.mi.=Good,Precipitation.in.=None}	0.494	523734
{Weather_Condition=Fair,Severity=Severity2,Precipitation.in.=None }	0.492	521718
{Year=2021,Severity=Severity2,Visibility.mi.=Good, Precipitation.in.=None}	0.492	521264
{Weather_Condition=Fair,Severity=Severity2,Visibility.mi.=Good}	0.481	509942
{Weather_Condition=Fair,Severity=Severity2,Visibility.mi.=Good, Precipitation.in.=None}	0.481	509690
{Visibility.mi.=Good,Precipitation.in.=None,Pressure.in.=Low}	0.478	506534
{Severity=Severity2,Visibility.mi.=Good,Pressure.in.=Low}	0.471	499935
{Year=2021,Severity=Severity2,Distance.mi.=Small}	0.469	496919
{Severity=Severity2,Visibility.mi.=Good,Precipitation.in.=None, Pressure.in.=Low}	0.460	487980
{Severity=Severity2,Distance.mi.=Small,Pressure.in.=Low}	0.455	482337
{Year=2021,Precipitation.in.=None,Distance.mi.=Small}	0.452	479420
{Year=2021,Severity=Severity2,Precipitation.in.=None, Distance.mi.=Small}	0.448	474649
{Precipitation.in.=None,Distance.mi.=Small,Pressure.in.=Low}	0.444	471083
{Weather_Condition=Fair,Precipitation.in.=None,Distance.mi.=Small }	0.436	462794
{Severity=Severity2,Precipitation.in.=None,Distance.mi.=Small, Pressure.in.=Low}	0.429	455140
{Weather_Condition=Fair,Visibility.mi.=Good,Distance.mi.=Small}	0.427	452581
{Weather_Condition=Fair,Visibility.mi.=Good,Precipitation.in.=None,	0.427	452350

Itemsets	Support	Count
Distance.mi.=Small}		
{Weather_Condition=Fair,Severity=Severity2,Distance.mi.=Small}	0.425	450962
{Weather_Condition=Fair,Severity=Severity2,Precipitation.in.=None, Distance.mi.=Small}	0.425	450719
{Visibility.mi.=Good,Distance.mi.=Small,Pressure.in.=Low}	0.418	443697
{Year=2021,Visibility.mi.=Good,Distance.mi.=Small}	0.416	441618

Table 24: Top 30 most frequent itemsets

LHS	RHS	Support	Confidence	Lift
Severity=Severity2, Temperature.F.=Cold, Distance.mi.=Small	Season=Winter	0.100	0.657	2.127
Temperature.F.=Cold, Distance.mi.=Small	Season=Winter	0.103	0.648	2.096
Severity=Severity2, Temperature.F.=Cold	Season=Winter	0.113	0.645	2.086
Severity=Severity2, Precipitation.in.=None, Temperature.F.=Cold	Season=Winter	0.105	0.639	2.068
Severity=Severity2, Season=Winter	Temperature.F.=Cold	0.113	0.377	2.053
Severity=Severity2, Season=Winter, Precipitation.in.=None	Temperature.F.=Cold	0.105	0.376	2.053
Season=Winter, Precipitation.in.=None	Temperature.F.=Cold	0.108	0.376	2.051
Severity=Severity2, Season=Winter, Distance.mi.=Small	Temperature.F.=Cold	0.100	0.376	2.048
Season=Winter, Distance.mi.=Small	Temperature.F.=Cold	0.103	0.375	2.044
Precipitation.in.=None, Temperature.F.=Cold	Season=Winter	0.108	0.629	2.035
Severity=Severity2, Precipitation.in.=None, Humidity...=High,Distance	Temperature.F.=Cold	0.102	0.290	1.582

LHS	RHS	Support	Confidence	Lift
.mi.=Small				
Severity=Severity2, Humidity...=High, Distance.mi.=Small	Temperature.F.=Cold	0.111	0.289	1.578
Severity=Severity2, Precipitation.in.=None, Humidity...=High	Temperature.F.=Cold	0.116	0.287	1.567
Severity=Severity2, Humidity...=High	Temperature.F.=Cold	0.127	0.287	1.565
Severity=Severity2, Temperature.F.=Cold, Distance.mi.=Small	Humidity...=High	0.111	0.730	1.564

Table 25: Top 20 rules of the Dataset

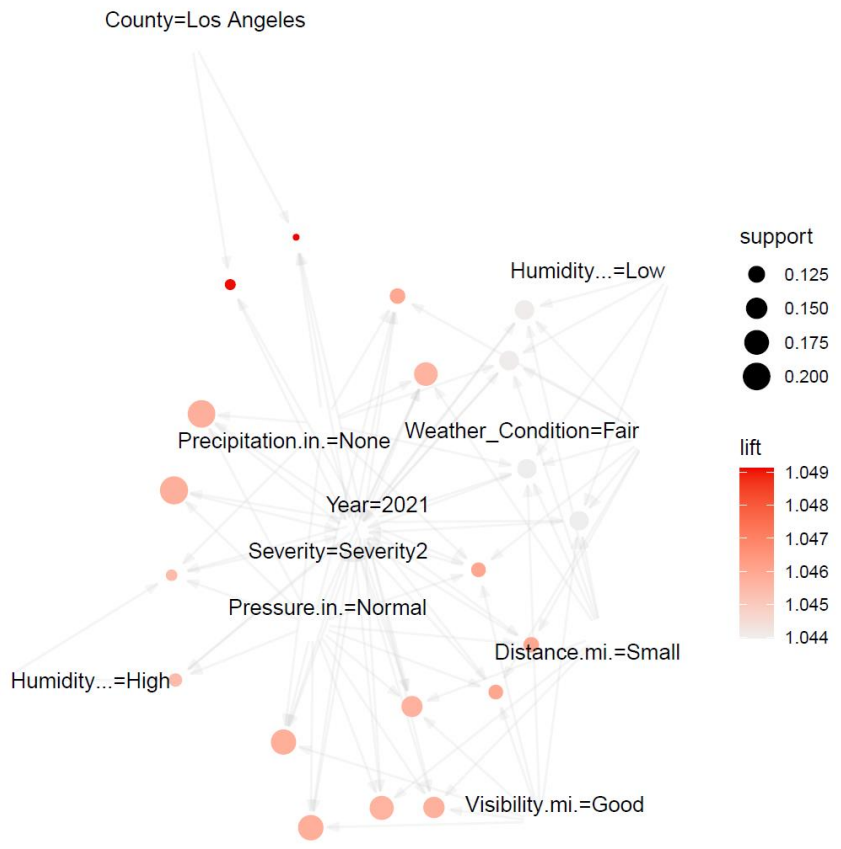

Figure 71: Graph of Top 10 Severity2 Rules Based on Support and Lift

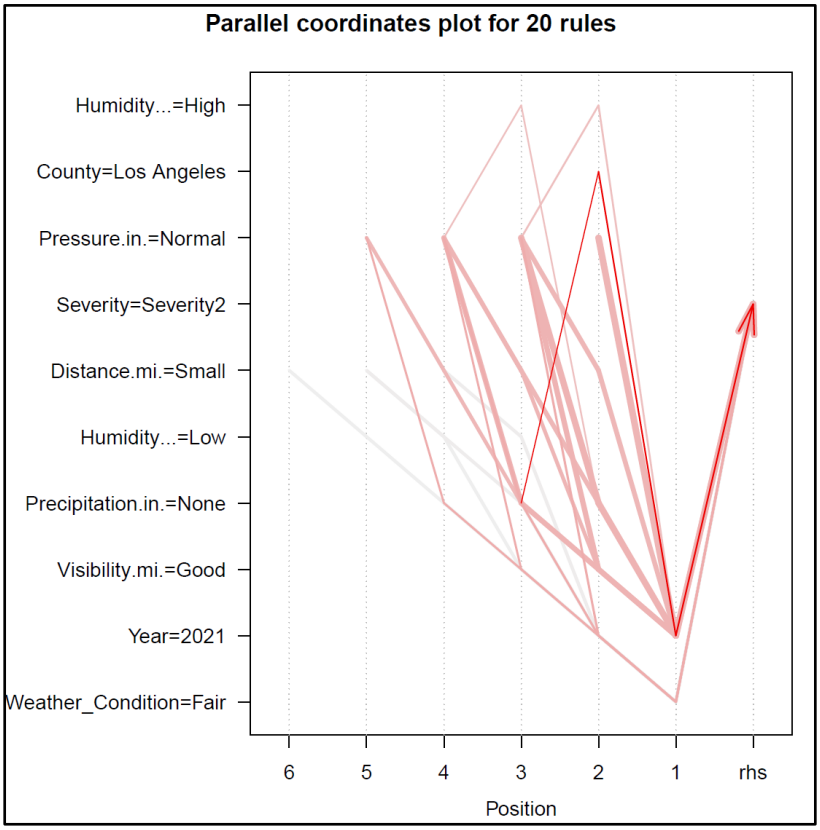

Figure 72: Parallel Coordinates for Top 20 Rules with Severity2 in RHS

LDA

In this section of the appendix, more technical information about the implementation of LDA is provided. Moreover, some plots generated during the application of LDA are added here which do not provide so much insightful information for the specific analysis.

Loadings	LD1	LD2	LD3
Visibility	0.313	-0.054	0.286
Precipitation	0.037	0.072	-0.072
Humidity	1.135	0.070	0.092
Temperature	1.423	0.342	-0.256
Distance	-0.025	0.727	0.686
Pressure	0.0137	-0.558	0.702

Table 26: Loadings of linear discriminant functions produced by LDA

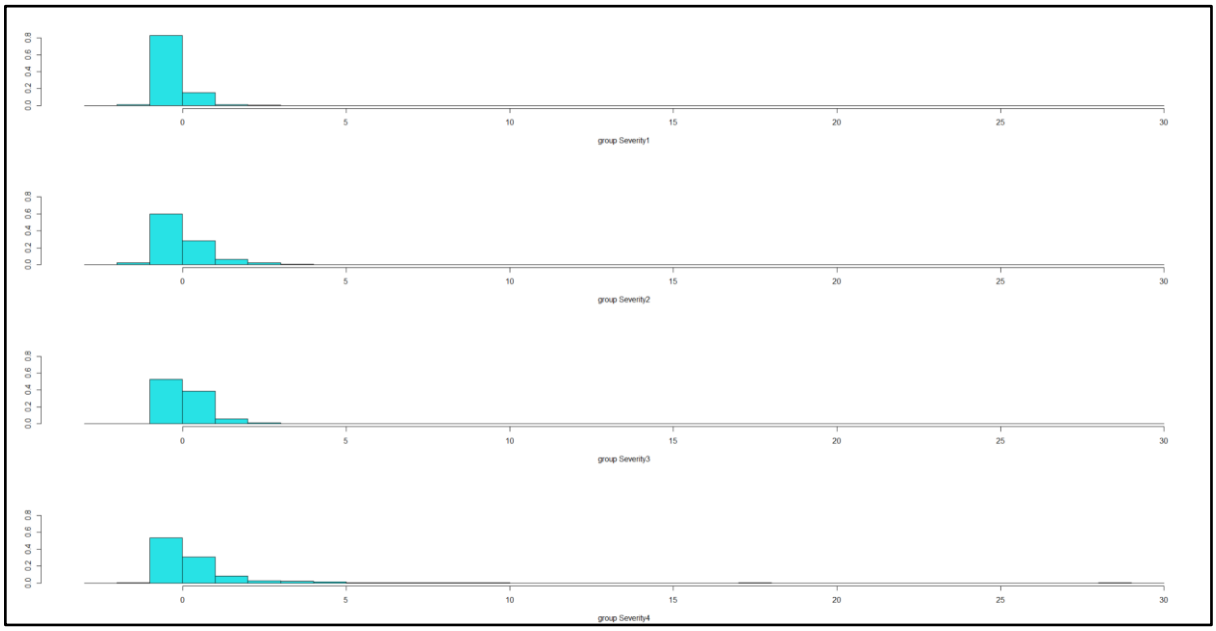

Figure 73: Stacked Histogram of the 2nd Discriminant Function's Values

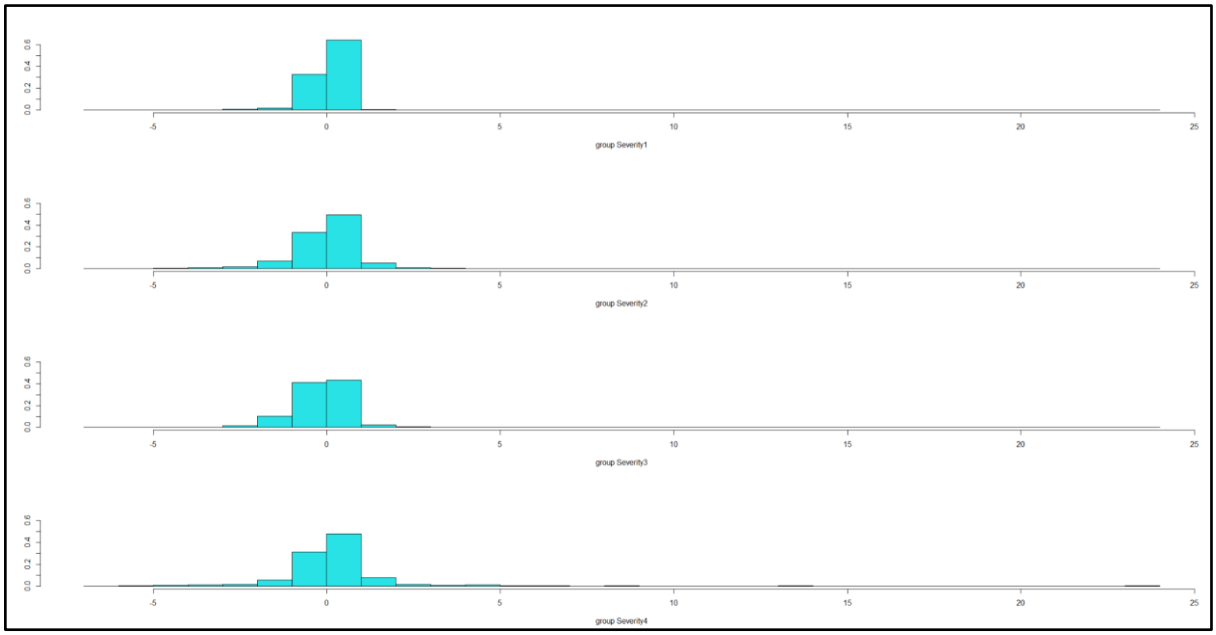

Figure 74: Stacked Histogram of the 3rd Discriminant Function's Values

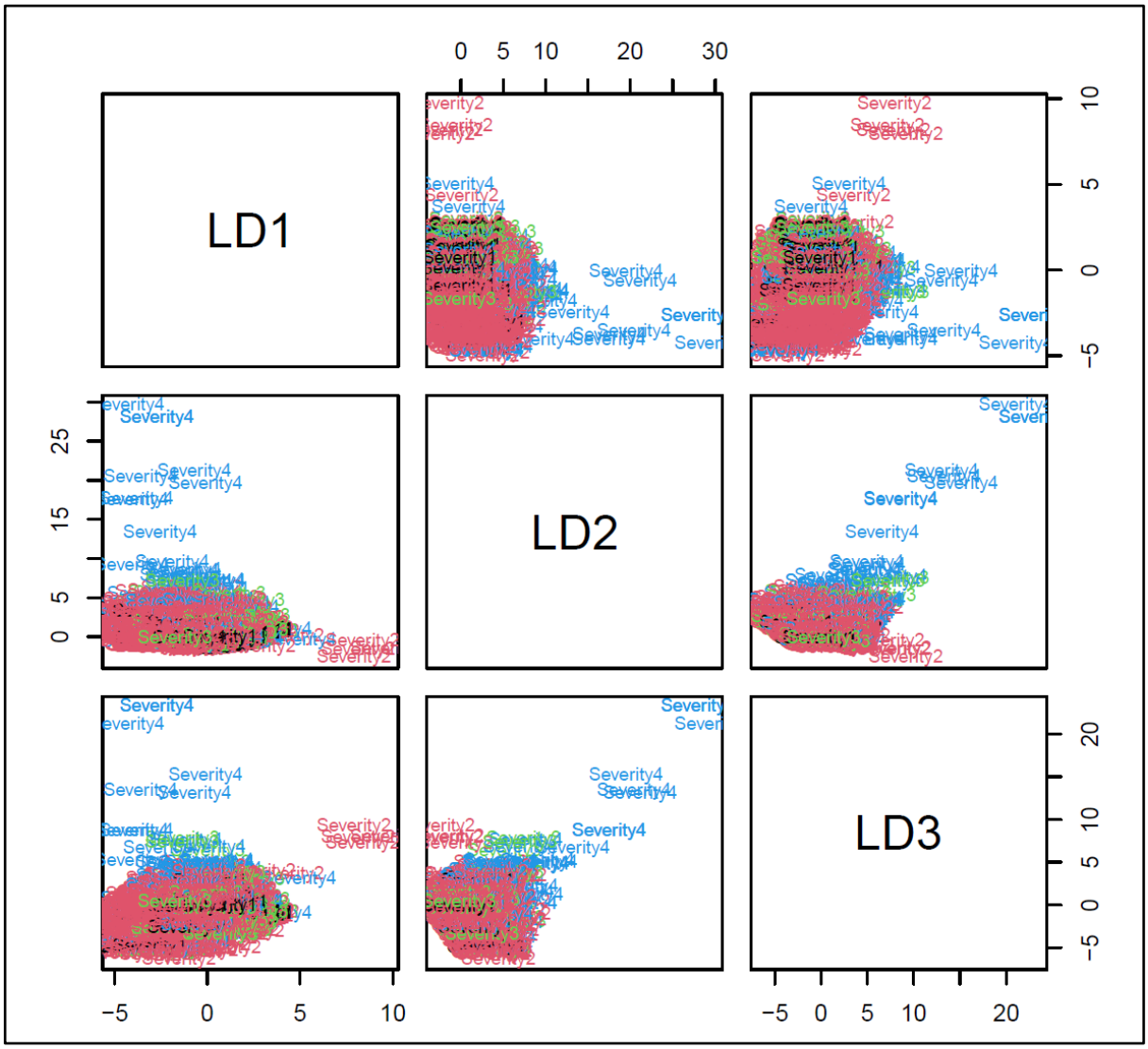

Figure 75: Projection of Training Data to all Combinations of LDs