UNIVERSITAT POLITÈCNICA DE CATALUNYA

Final Report

CATALONIAN ROAD ACCIDENTS

MD

Authors

MARINA ALAPONT VIDAL
DANIEL PULIDO GÁLVEZ
SIMÓN HELMUTH OLIVA STARK
JOEL CARDONA SAUS
DAVID LATORRE ROMERO

2022-23 Q1

Table of Contents

1. Motivation of the work

The main motivation for choosing this dataset has been the rich content in data that this one has, since this content has enough categorical (inside this, also boolean) and numerical variables. This allows the team to extract interesting information from the relation between these variables and go through multiple stages of study that are relevant.

The team also has thought this study will be relevant, because of its topic. Due to the fact that learning about accidents can help the society to prevent them by knowing how these were caused. This could be done informing drivers and pedestrians of what situations have to be avoided, and also, giving this knowledge to civil engineers in order to better build and signal the road.

Finally, this information is important for everyone, because accidents affect the whole society. Every person in the world walks in the street, drives, or travels in public transport. And all of this could be involved in an accident.

2. Data source presentation

We have obtained our dataset from the Catalonia Government open data page named *Dades Obertes Catalunya*. This is a public website with open and transparent data from the Catalonia Government. There are a lot of datasets such as traffic accidents, Covid-19 incidence in Catalonia, Catalonia's Government Investment, etc.

The dataset we have chosen consists of information about traffic accidents in the Catalonia region between 2010 and 2021. In this dataset we can get such information as the number of mortal victims, if there was fog when the accident happened, the maximum velocity of that road, where the accident happened, etc. The data set has 21161 rows and 58 columns.

The dataset can be found in the following link:
https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb

3. Formal description of data structure and metadata

Data structure

The data set selected for this study has 21.161 rows and 58 columns. Each one of the rows represents the data compiled from a grave or mortal traffic accident that occurred in Catalonia. These accidents can be of different types like a hit between vehicles, a run-over or a roll over, for example. For each accident, the data set has information like, the type of accident, the date when it occurred, the region and zone where the accident happened, under what situations it occurred and what influenced that accident, what was the number of victims, the number of entities involved, in what type of day and at what hour it all happened, among others. In total, the data set has 13 quantitative variables and 45 qualitative variables.

Metadata file

In order to help understand and identify the information that is provided for each accident on the data set, a metadata table has been created. In this metadata table it can be found, for each variable, the name of it's modalities (in case of qualitative variables), the meaning of the variable, the type of variable, the measuring unit (for numerical variables), the missing code in case the variable has any missings, the range (for numerical variables) and the role the variable has in the dataset.

Source: https://analisi.transparenciacatalunya.cat/Transport/Accidents-de-tr-nsit-amb-morts-o-ferits-greus-a-Ca/rmgc-ncpb

After data selection the dimensions are

n: 5000

k: 23

VARIABLE	MODALITIES	MEANING	TYPE	MEASURING UNIT	MISSING CODE	RANGE	ROLE
zona		kind of zone where the accident has happened	Qualitative	-	-	-	Explanatory
	Zona urbana	Urban zone					
	Carretera	In road					
dat		date of the accident	Qualitative	-	-	-	Explanatory
nomCom	***	Region where the accident occurred	Qualitative	-	-	-	Explanatory
nomDem		Province of the region where the accident has happened	Qualitative	-	-	-	Explanatory
	Barcelona						
	Lleida						

	Tarragona						
	Girona						
F_MORTS		Number of mortal victims	Numeric		-	[0, ∞]	Response
F_FERITS_GREUS		Number of grave injured victims	Numeric		-	[0, ∞]	Response
F_FERITS_LLEUS		Number of minor injured victims	Numeric		-	[0, ∞]	Response
F_UNITATS_IMPLICADES		Number of involved entities (vehicles, bikes, pedestrians...)	Numeric		-	[0, ∞]	Explanatory
F_VIANANTS_IMPLICADES		Number of involved pedestrian	Numeric		-	[0, ∞]	Explanatory
F_BICICLETES_IMPLICADES		Number of bikes involved	Numeric		-	[0, ∞]	Explanatory
VEHICLES_MOTOR		Number of motor vehicles involved	Numeric		-	[0, ∞]	Explanatory
C_VELOCITAT_VIA		Maximum velocity pemrited in the zone	Qualitative	km/h	NA / 999	[10,120]	Explanatory
D_ACC_AMB_FUGA		Accident with escapist	Boolean	-		-	Explanatory
D_CLIMATOLOGIA		Details of the weather	Qualitative	-	Sense especificar	-	Explanatory
	Bon temps	Good weather					
	Nevant	It was snowing					
	Pluja forta	It was raining strongly					
	Pluja dèbil	It was raining weakly					
D_INFLUIT_CIRCULACIO		Accident influenced by the traffic	Boolean	-	Sense especificar	-	Explanatory
D_INFLUIT_ESTAT_CLIMA		Accident influenced by the weather	Boolean	-	Sense especificar	-	Explanatory

D_INFLUIT_LLUMINOSITAT		Accident influenced by the light	Boolean	-	Sense especificar	-	Explanatory
D_INFLUIT_VISIBILITAT		Accident influenced by the vision	Boolean	-	Sense especificar	-	Explanatory
D_INTER_SECCIO		Accident occurred in an intersection	Qualitative	-	-	-	Explanatory
	Arribant o eixint intersecció fins 50m	The accident occurred when arriving into an intersection or exiting (50 m precision)					
	En secció	The accident occurred in section					
	Dintre intersecció	The accident occurred inside an intersection					
D_SUPERFICIE		State of the roadway surface	Qualitative	-	Sense especificar	-	Explanatory
	Gelat	The roadway was frozen					
	Inundat	The roadway was flooded					
	Mullat	The roadway was wet					
	Nevat	The roadway had snow					
	Relliscós	The roadway was slippery					
	Sec i net	The roadway was dry and clean					
grupDiaLab		Work day or non working day	Qualitative	-	-	-	Explanatory
	Feiners	Week day, working day					
	CapDeSetmana	Weekend or non working day (holidays)					
grupHor		Period of the day in which the accident occurred	Qualitative	-	-	-	Explanatory
	Matí	Morning					

	Tarda	Afternoon					
	Nit	Night					
TipAcc		Type of the accident	Qualitative	-	-	-	Explanatory
	Altres	Others					
	Atropellament	Run-over					
	Bolcada a la calçada	The car knocked over					
	Col.lisió d'un vehicle contra un obstacle de la calcada	Collision with an object that was on the roadway					
	Col.lisió de vehicles en marxa	Collision between two or more vehicles					
	Sortida de la calçada sense especificar	Unspecified roadway exit					

Table 3.1. Metadata table of the selected variables.

*** The modalities of the variable “NomCom” are all the different regions (“comarques”) of Catalonia, the original name of each modality corresponds to the real region name. For space reasons, these modalities' names can be found in the Annex section

Final scope

Out of all the variables of the original data set, we have kept 7 numerical variables, 11 categorical variables and 5 boolean/binary variables. Also the number of individuals (accidents) has been reduced to 5000. The variables kept in the analysis are:
F_MORTS, F_FERITS_GREUS, F_FERITS_LLEUS, F_UNITATS_IMPLICADES, F_VIANANTS_IMPLICADES , F_BICICLETES_IMPLICADES, VEHICLES_MOTOR*, zona, dat, nomCom, nomDem, C_VELOCITAT_VIA, D_CLIMATOLOGIA, D_INTER_SECCIO, D_SUPERFICIE, grupDiaLab, grupHor, TipAcc, D_ACC_AMB_FUGA, D_INFLUIT_CIRCULACIO, D_INFLUIT_ESTAT_CLIMA, D_INFLUIT_LLUMINOSITAT, D_INFLUIT_VISIBILITAT.

In order to choose which variables to keep for analysis, the following criteria has been taken into account:

- Elimination of variables that are redundant with other variables.
- Elimination of variables that the team doesn't find interesting for the study.
- Elimination of variables with a not clear meaning.
- Elimination of variables too specific or concrete for our study.

Also, we have created a new variable called "VECHICLES_MOTOR" that is an aggregation of the variables F_CICLOMOTORS_IMPLICADES, F_MOTOCICLETES_IMPLICADES, F_VEH_LLEUGERS_IMPLICADES, F_VEH_PESANTS_IMPLICADES (variables that the team found too specific).

In terms of selection of individuals (rows) a random algorithm has been used in order to eliminate random rows, but maintaining the ratio of accidents per year (to not distort any possible study). The algorithm used to make this selection can be found in the R material submitted with this document.

4. Complete data mining process

The first step has been the data collection, where we have taken a look at different datasets and have discussed which one was of our interest. Once decided, we have downloaded the .csv file. Secondly, we have done the data preparation, where we have joined a few columns as a new one and selected a subset of columns we have found more interesting for our study. Also, we have developed an algorithm to generate a subset of 5000 random rows of the original dataset in order to work with a significant sample of the original dataset. Thirdly, once we had generated the dataset, we have stepped up with the preprocessing process, all this step has been made using RStudio. Here, we have factorized the categorical variables, defined levels of categorical variables and changed names (making them short and translating them), and treated the missing values of the variables.

Once we have finished with the data cleaning and preprocessing the data, we have done the PCA (Principal Components Analysis), where we have used the dataset obtained after executing the preprocessing process. In order to complete the PCA, we have executed an R script, which, first of all, takes all the numerical variables (because it only works with numerical variables) of the dataset and generates the inertia plots of which principal components are more representative in order to reduce the dimensions. Then, we have generated plots to project the individuals in each factorial map obtained and plots projecting the qualitative variables to observe the relationship between these and the numerical ones.

Following this step, we have obtained the clusters with the hierarchical clustering, where we have obtained a dendrogram, using Ward's D2 method, in order to decide the number of clusters. Also, to make a good decision on the number of clusters, we have calculated different KPI's using the R *NbClust* function, and we have created a table in order to compare different numbers of clusters and decide which of them to use.
Finally, we have completed our study doing a cluster profiling, using the two clusters resulting from the clustering process. In this final step, we have executed an R script which generates a set of plots comparing each feature of the dataset between clusters, with all those plots we have extracted conclusions of our clusters.

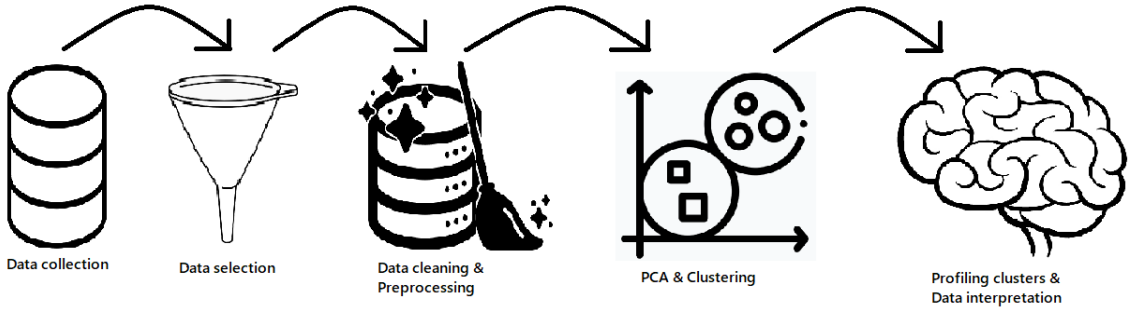


Image 4.1. Workflow of the complete data mining process.

5. Detailed description of preprocessing and data preparation

Once we had our data set selected and we reduced the number of instances and variables, we proceeded to do the preprocessing.

1. **Variables, factorization, levels, and sorting**

   One of the first things we have to do once we have the dataset with the variables we are going to work with, is to give RStudio some context about the variables, that is, tell RStudio what type they are.

   If we do not specify the correct type for every variable respectively, then we are going to end up making a bad analysis of that variable, or even not being able to make any.

   Let's read our principal csv, which contains the raw data of the dataset, with R, and let's execute the command (where dd contains our csv read by R)

```
sapply(dd, class)
```

   If we take a look at the output of this command, we can observe that while there are some variables interpreted correctly (the numerical (quantitative) ones), others are not (the categorical ones, or the Date of the accidents).

   For example, nMortal (note that the abbreviation to the name variables, which is explained next, is already applied here), a numerical variable, is interpreted correctly:

```
   nMortal
"integer"
```

   But, on the other hand, Weather, a categorical one, is not:

```
   Weather
"character"
```

   Therefore, to interpret correctly the categorical variables, we have to declare them as factors.

In addition, there is another variable in which we have to change its type, but it won't be a factor. This variable is the one that represents the date of the accidents. R has a special type for dates.

However, when we do the PCA analysis and clustering on the next sections we won't be able to work with a variable of class Date, that's why we've made two new factor variables from this one: Year, and Month

After we have done this, and according to the METADATA file above, if we re-execute the sapply() command, we have the next output:

Variable name (after renaming, see next section)	**Class**
Zone	Factor
Date	Date
Region	Factor
Prov	Factor
nMortal	Integer
nGraveInj	Integer
nMinorInj	Integer
nInvolv	Integer
nPedest	Integer
nBikes	Integer
nMotor	Integer
Vel	Factor
Escaped	Factor
Weather	Factor
TrafficInf	Factor

WeatherInf	Factor
LightInf	Factor
VisionInf	Factor
Intersect	Factor
Surface	Factor
DayGroup	Factor
HourGroup	Factor
AccType	Factor
Month	Factor
Year	Factor

Table 5.1. Table with the class of each feature of the dataset.

When declaring the factors, we also have to process its levels for every one of them. That includes:

- Changing the level names, if necessary. We will talk about this later.
- Consider what we do with those levels that represent “Other”, “Unknown” or “NA” in the Variable. We’ll also talk about this later.
- Ordering the levels for each factor. This, although it is not specially necessary, is very useful to visualize the data in an easier and comfortable way.

  For example, it's better to visualize the bar plot of the factor HouGroup like this:

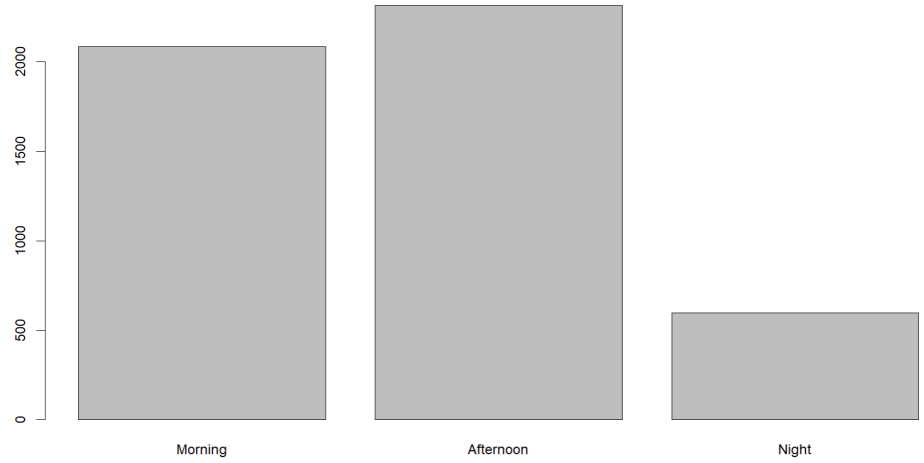


Image 5.1. Barplot with levels not sorted correctly.

Than like this:

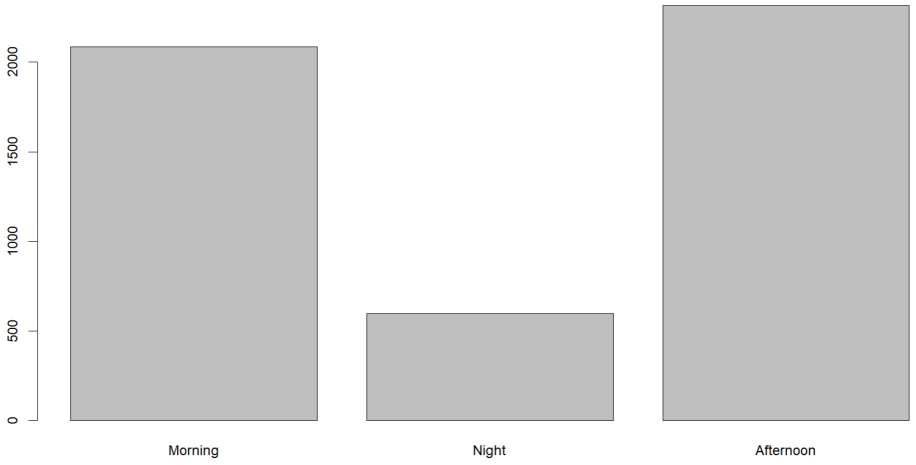


Image 5.2. Barplot with levels sorted correctly.

2. Renaming variables and levels

Another step that we have done during our preprocessing phase, along the declaration of factors and their levels, has been renaming all the variables and their modalities. Although it may seem an unnecessary step, when the data set has too many columns, that is, when there are a lot of variables to study, having long and complex names can make the study difficult in terms of, per example, charts analysis and interpretation (see figure 1 and 2). Therefore, we have changed the names for a

better chart and diagram interpretation, a good understanding and also to adjust the language.

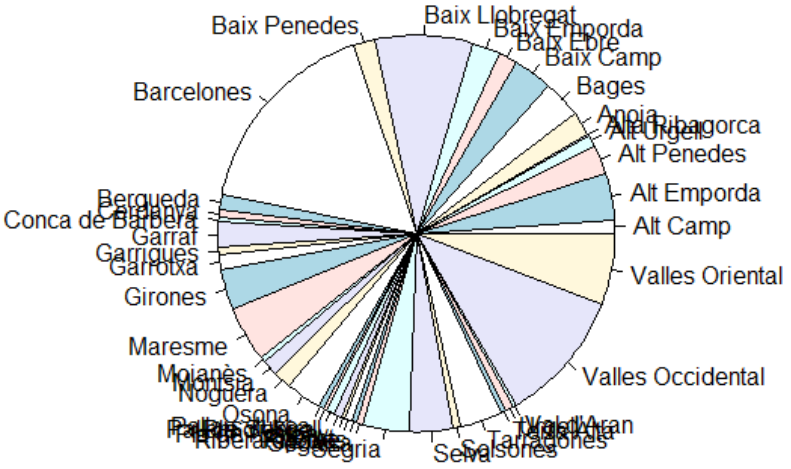


Image 5.3. Pie plot of the variable “nomCom”.

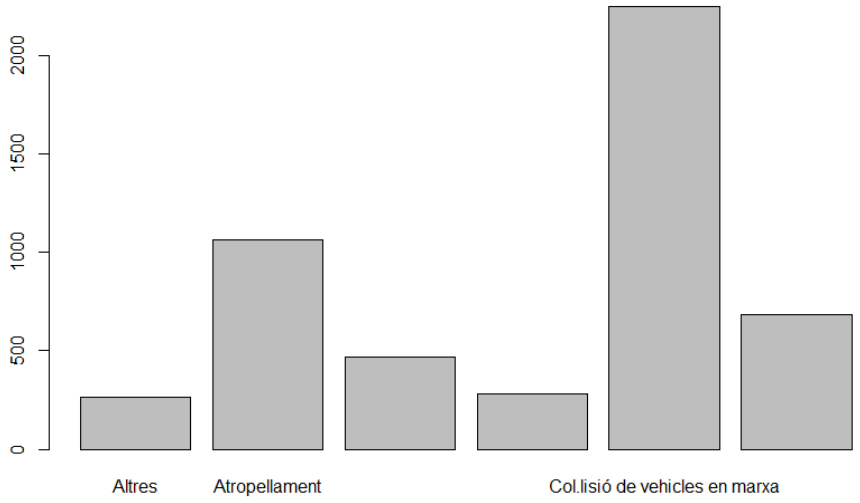


Image 5.4. Bar plot of the variable “TipAcc”.

As it can be seen in the images avobe, when having bad modalities names, in this case, the analysis of the plots can be even impossible and also, it may happen that the names may not fit in inside the plot, generating a lack of information as it happens in figure 2. That’s two reasons why (among others), in our specific case, we had to change and abbreviate several modalities and variable names.

When changing the names, we have tried to keep the meaning, and we have procured to maintain an easy interpretation. The modifications made are reflected in the following table, where variable names are marked in bolt:

ORIGINAL NAME	**AFTER RENAMING**
zona	**Zone**
Zona urbana	Urban
Carretera	Road
dat	**Date**
nomCom *	**Region**
nomDem	**Prov**
Barcelona	Barcelona
Girona	Girona
Tarragona	Tarragona
Lleida	Lleida
F_MORTS	**nMortal**
F_FERITS_GREUS	**nGravelnj**
F_FERITS_LLEUS	**nMinorInj**
F_UNITATS_IMPLICADES	**nInvolv**
F_VIANANTS_IMPLICADES	**nPedest**
F_BICICLETES_IMPLICADES	**nBikes**
VEHICLES_MOTOR	**nMotor**
C_VELOCITAT_VIA	**Vel**

D_ACC_AMB_FUGA	**Escaped**
D_CLIMATOLOGIA	**Weather**
Bon temps	Good
Nevant	Snow
Pluja forta	WeakRain
Pluja dèbil	StrongRain
D_INFLUIT_CIRCULACIO	**TrafficInf**
D_INFLUIT_ESTAT_CLIMA	**WeatherInf**
D_INFLUIT_LLUMINOSITAT	**LightInf**
D_INFLUIT_VISIBILITAT	**VisionInf**
D_INTER_SECCIO	**Intersect**
Arribant o eixint intersecció fins 50m	Arriving
En secció	InSection
Dintre intersecció	Inside
D_SUPERFICIE	**Surface**
Gelat	Icy
Inundat	Flooded
Mullat	Wet
Nevat	Snowy
Relliscós	Slippery
Sec i net	Dry&Clean
grupDiaLab	**DayGroup**
Feiners	Weekday

CapDeSetmana	Weekend
grupHor	**HourGroup**
Matí	Morning
Tarda	Afternoon
Nit	Night
TipAcc	**AccType**
Altres	Other
Atropellament	RunOver
Bolcada a la calçada	Rollover
Col.lisió d'un vehicle contra un obstacle de la calcada	HitObstacle
Col.lisió de vehicles en marxa	HitVehicle/s
Sortida de la calçada sense especificar	NARoadExit

Table 5.2. Table with the original and new names of all the variables and their modalities.

*The original names and the new ones of the different regions of catalonia can be found on the Annex section.

3. Missings and its randomness

We have done a study about the basic descriptive of our variables before treating them and we have seen that 8 of these have missing values. In order to quantify how important are, we have calculated the proportion of missing per variable and we have obtained these values:

Variable name	Proportion of missing values
Vel	18,56%
Escaped	0,82%
Weather	0,02%

TrafficInf	0,02%
WeatherInf	0,02%
LightInf	0,02%
VisionInf	6,68%
Surface	0,02%

Table 5.3. Table with the percentage of missing values for each variable that has some.

We have also analysed manually the randomness about this missings, and we conclude that, apparently, there is no pattern or relation between the variables that is causing this absence of data.

Also, to mention that in the renaming step, the missing values were renamed to “NA " instead of “Sense especificar” or similar.

4. **Missings treatment**

All of our missing values belong to categorical variables, and in this case, we can not apply algorithms such as KNN. The only treatment we have done here, is to rename the missings values to Unk<AbreviatedvarName>. For example, missings of Vel, were renamed to “UnkVel”.

5. **Outliers**

The univariate descriptive study that we have carried out in order to see the initial data quality (study that is going to be explained later in this document), has helped us to identify outliers in certain variables.

The following table shows what variables have outliers and which type of outliers these are:

VARIABLE NAME	TYPE OF OUTLIERS
nMortal	Extreme value of the population
nGraveInj	Extreme value of the population

nMinorInj	Extreme value of the population
nInvolv	Extreme value of the population
nPedest	Extreme value of the population
nBikes	Extreme value of the population
nMotor	Extreme value of the population
Vel	Missing code and mistake

Table 5.4. Table with the type of outlier for each variable that has outliers.

If we take a closer look at our data and the descriptive statistics obtained (that can be found in the next section of this document), we can observe that the outliers of the numerical variables are due to extreme values that we can not eliminate or ignore, and that can become crucial in our study. For example, it is not very normal to have a big number of dead people in a traffic accident but, it's not impossible, so in our data the number of dead people tends to be 0, but there are cases where it is higher and, when drawing box plots, that cases can look like outliers that can be eliminated, but they can not.

With regard to the “Vel" variable, the outliers are due to two reasons. The first reason is a possible error, as there is one traffic accident where the maximum velocity of the road is said to be 0, which is impossible. The second and last reason is that the value is a missing code; in the variable “Vel" the missings are specified originally as “NA” or “999” and taking into account that the maximum velocity in Catalonia is of 120 km/h, a velocity of 999 is clearly an outlier. The treatment of the “Vel" outliers has been done along with the missing treatment of this variable: the 999 value and the 0 value have been replaced with a missing represented by the label “UnknownVel”, as it’s a qualitative variable.

6. Basic statistical descriptive analysis

In order to obtain the basic **univariate statistics**, we have used an RMarkdown script to automatically generate the descriptive plots and tables for each variable of the accidents dataset. In addition, if some variable has been affected after preprocessing, plots are shown before and after the changes. Different information is generated depending on the type of each variable:

1. For numerical variables, a histogram and a box plot have been generated. In addition, a table shows the minimum and maximum values for that variable; together with the 1st, 2nd (Median), and 3rd quartiles; as well as the mean, standard deviation, coefficient of variation and number of unknown values.

2. For dates, a histogram showing the density of accidents grouped by years has been created. This plot is followed by the exact minimum (oldest) and maximum (most recent) date values for that variable. The number of unknown values has also been provided.

3. For categorical variables, a pie chart and a bar plot have been computed. Next, the number of different modalities is shown, followed by a table containing the count for each modality, together with the relative frequency of each factor as a proportion (sorted in decreasing order).

Regarding the **bivariate statistics**, we have also used the RMarkdown script to generate plots confronting the numeric variables *nMortal*, *nGraveInj* and *nMinorInj* with all the other variables in our dataset. After the generation, we have chosen some plots that we considered added more information to the analysis. In addition, for numerical variables, we also have calculated the correlation with the three variables mentioned above.

Dimensions of the dataset

Number of variables (columns)	25
Number of instances (rows)	5000

Table 6.1. Summary of the dataset dimensions.

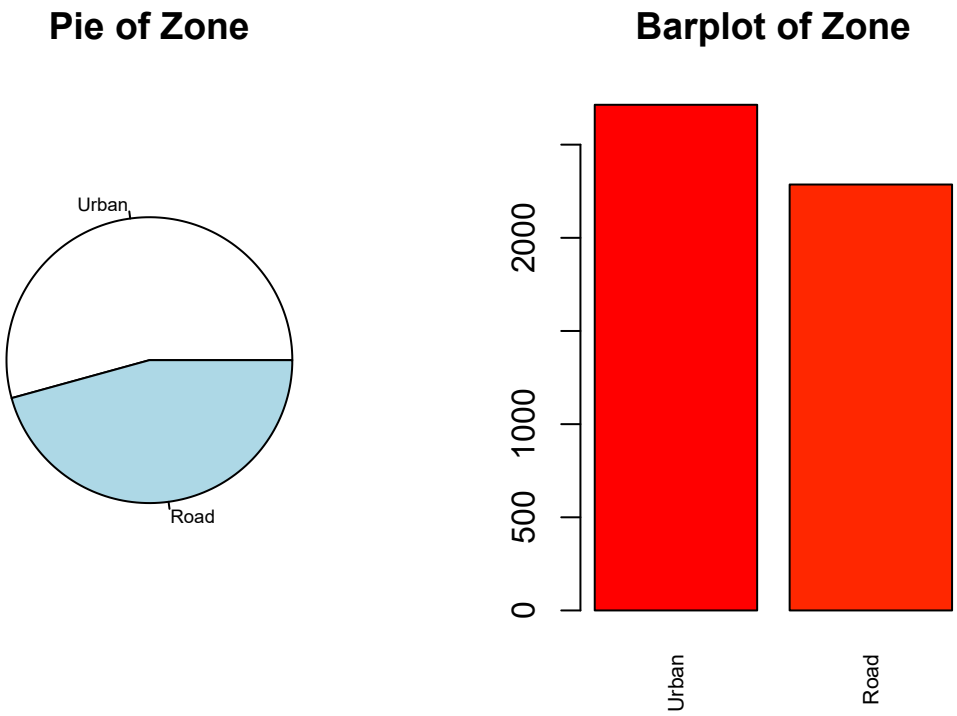
Variables' names

Zone	Date	Region	Prov	nMortal
nGraveInj	nMinorInj	nInvolv	nPedest	nBikes
nMotor	Vel	Escaped	Weather	TrafficInf
WeatherInf	LightInf	VisionInf	Intersect	Surface
DayGroup	HourGroup	AccType	Month	Year

Table 6.2. Enumeration of all the variables' names of the accidents dataset.

Univariate analysis

Variable 1 : Zone

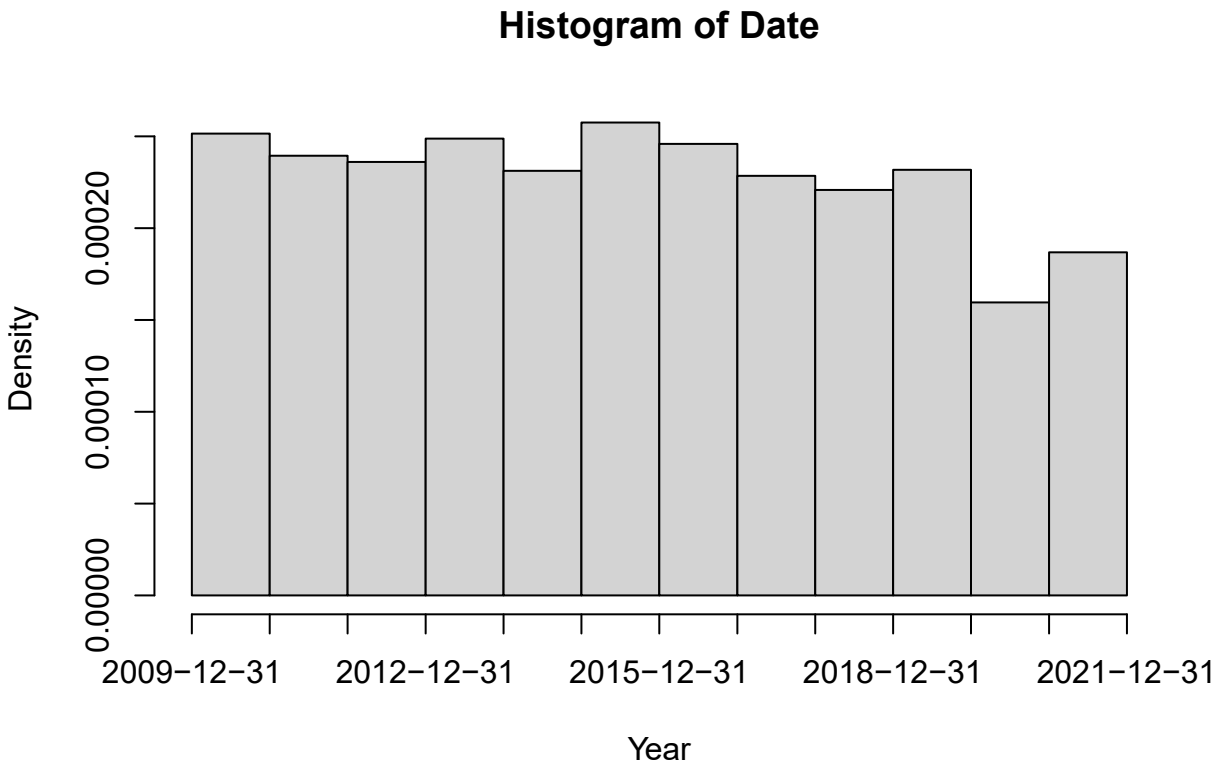


Number of modalities: 2

Zone	Frequency	Proportion
Urban	2714	0.5428
Road	2286	0.4572

Table 6.3. Zone frequency and proportion table (sorted).

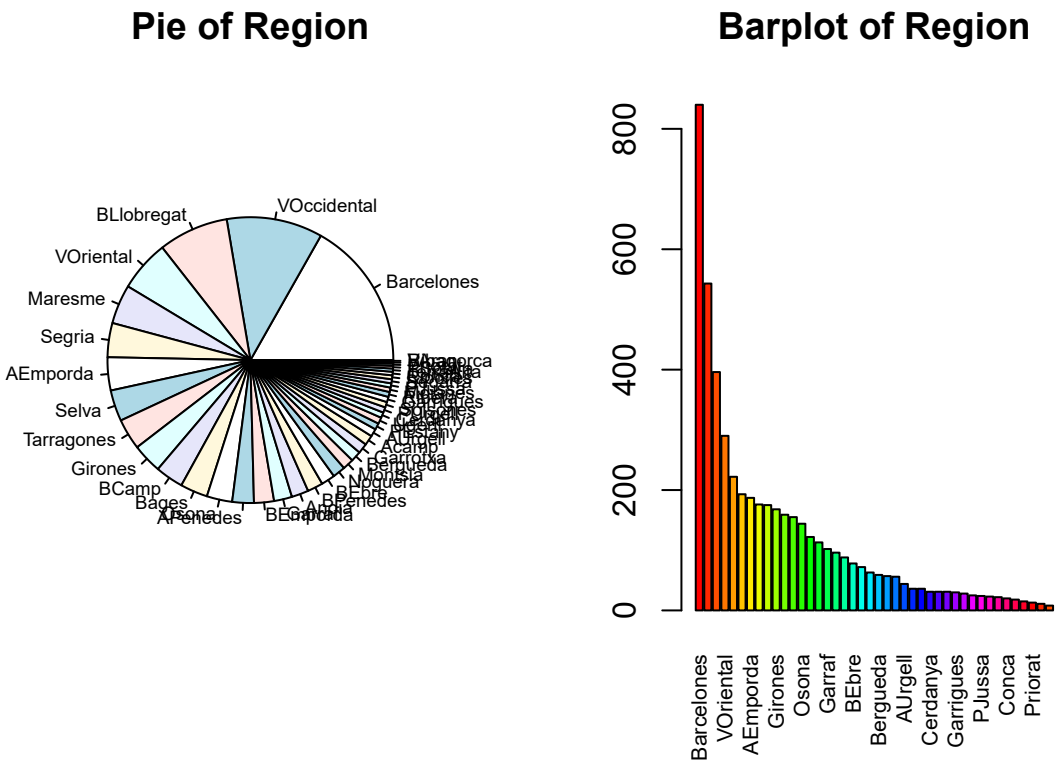
Variable 2 : Date



Min.	Max.	Missing values
2010-01-01	2021-12-30	0

Table 6.4. Date summary.

Variable 3 : Region

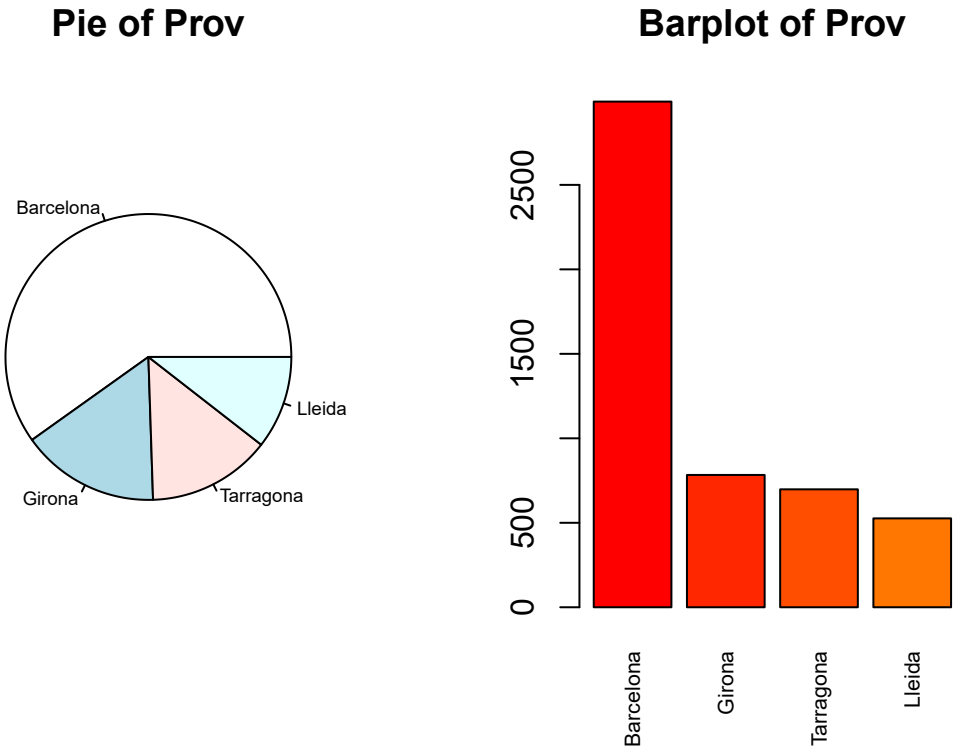


Number of modalities: 42

Region	Frequency	Proportion
Barcelones	840	0.1680
VOccidental	543	0.1086
BLlobregat	396	0.0792
VOriental	290	0.0580
Maresme	222	0.0444
Segria	193	0.0386
AEmporda	187	0.0374
Selva	176	0.0352
Tarragones	175	0.0350
Girones	168	0.0336
BCamp	159	0.0318
Bages	155	0.0310
Osona	144	0.0288
APenedes	122	0.0244
BEmporda	113	0.0226
Garraf	102	0.0204
Anoia	96	0.0192
BPenedes	88	0.0176
BEbre	78	0.0156
Noguera	72	0.0144
Montsia	63	0.0126
Bergueda	59	0.0118
Garrotxa	57	0.0114
Acamp	56	0.0112
AUrgell	44	0.0088
PEstany	36	0.0072
Ugerll	36	0.0072
Cerdanya	31	0.0062
PUrgell	31	0.0062
Solsones	31	0.0062
Garrigues	30	0.0060
Ribera	28	0.0056
Moianes	25	0.0050
PJussa	24	0.0048
Segarra	23	0.0046
Ripolles	22	0.0044
Conca	20	0.0040
TerraAlta	18	0.0036
PSobira	15	0.0030
Priorat	13	0.0026
Ribagorca	11	0.0022
VAran	8	0.0016

Table 6.5. Region frequency and proportion table (sorted).

Variable 4 : Prov

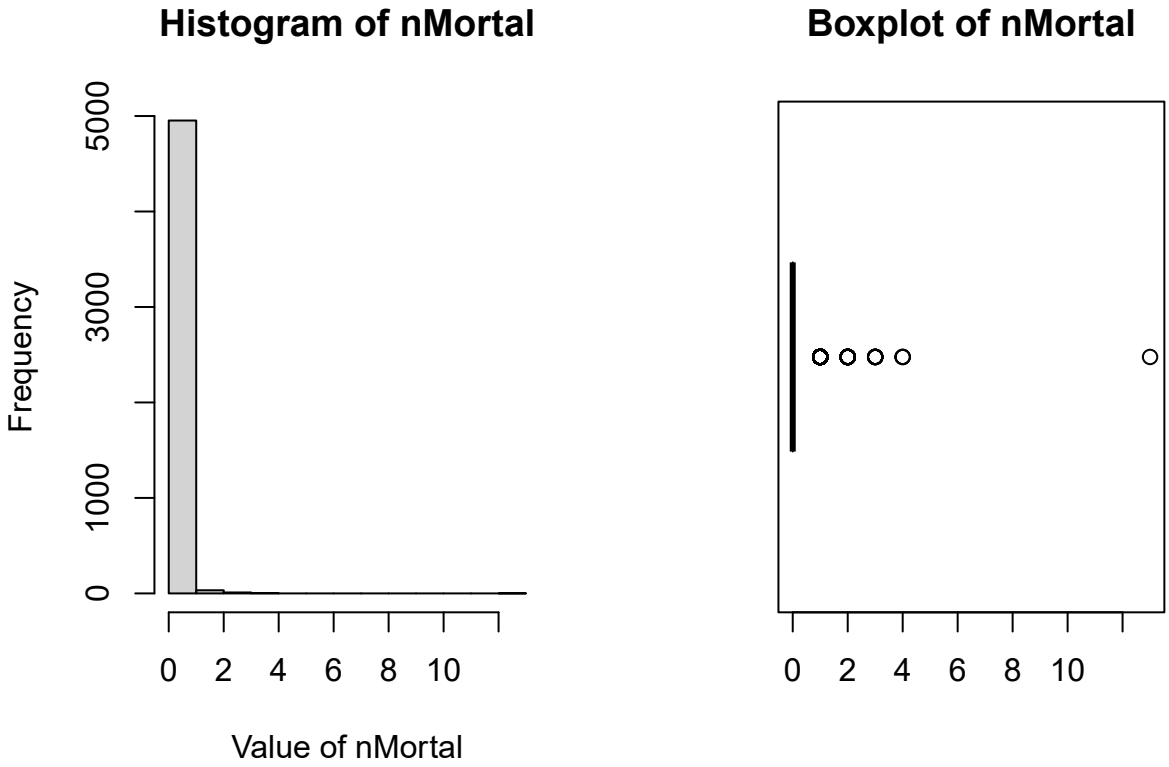


Number of modalities: 4

Prov	Frequency	Proportion
Barcelona	2993	0.5986
Girona	783	0.1566
Tarragona	698	0.1396
Lleida	526	0.1052

Table 6.6. Prov frequency and proportion table (sorted).

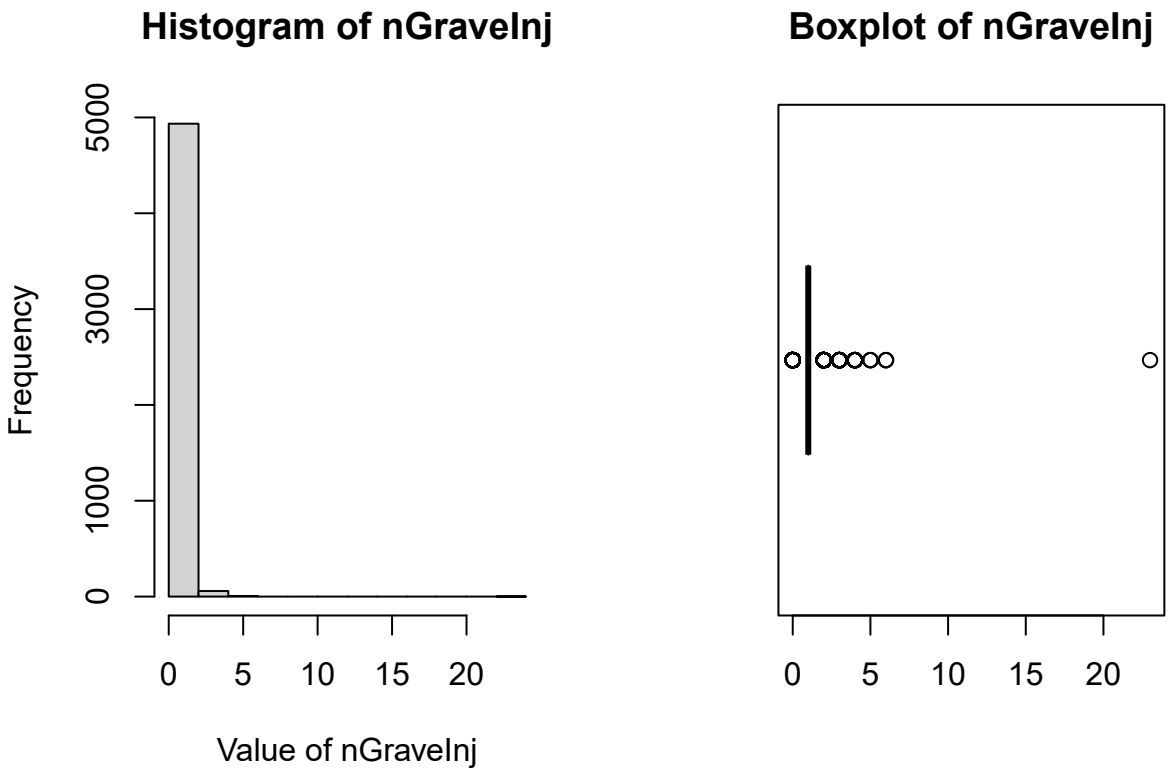
Variable 5 : nMortal



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	0.14	0	13	0.4289951	3.064251	0

Table 6.7. nMortal extended Summary Statistics.

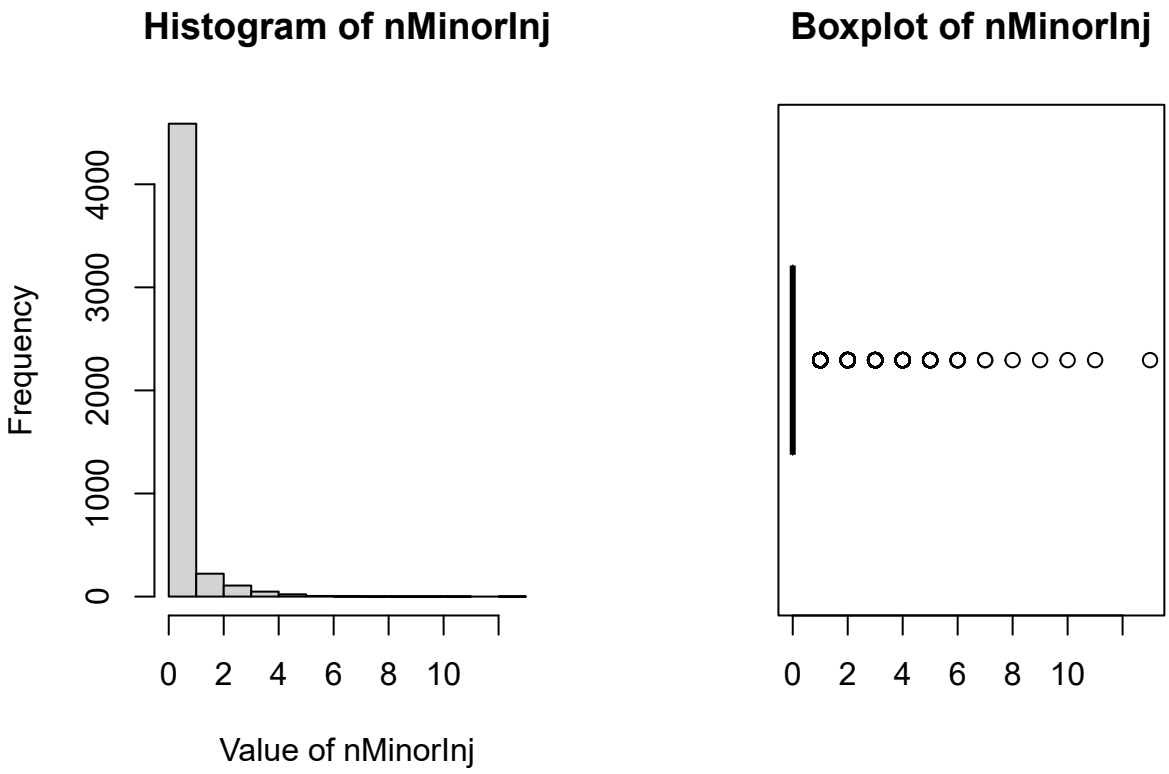
Variable 6 : nGraveInj



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	1	1	1.0034	1	23	0.5904728	0.588472	0

Table 6.8. nGraveInj extended Summary Statistics.

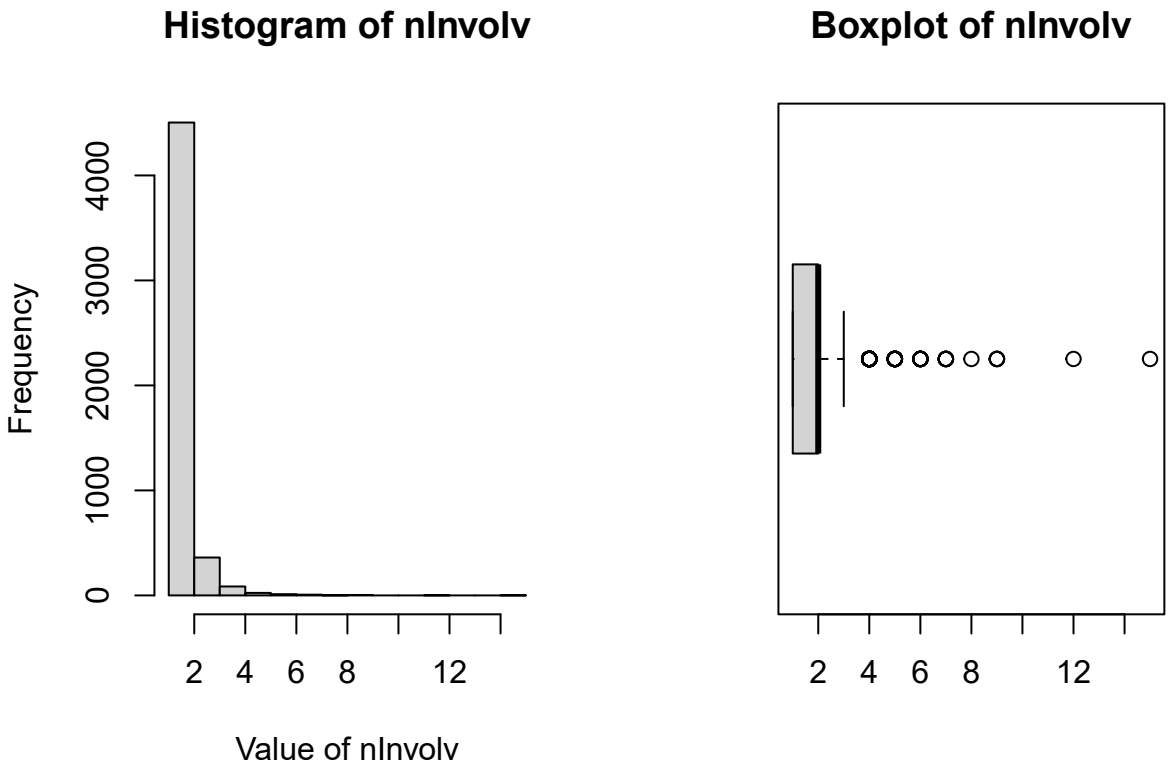
Variable 7 : nMinorInj



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	0.3924	0	13	0.899213	2.291572	0

Table 6.9. nMinorInj extended Summary Statistics.

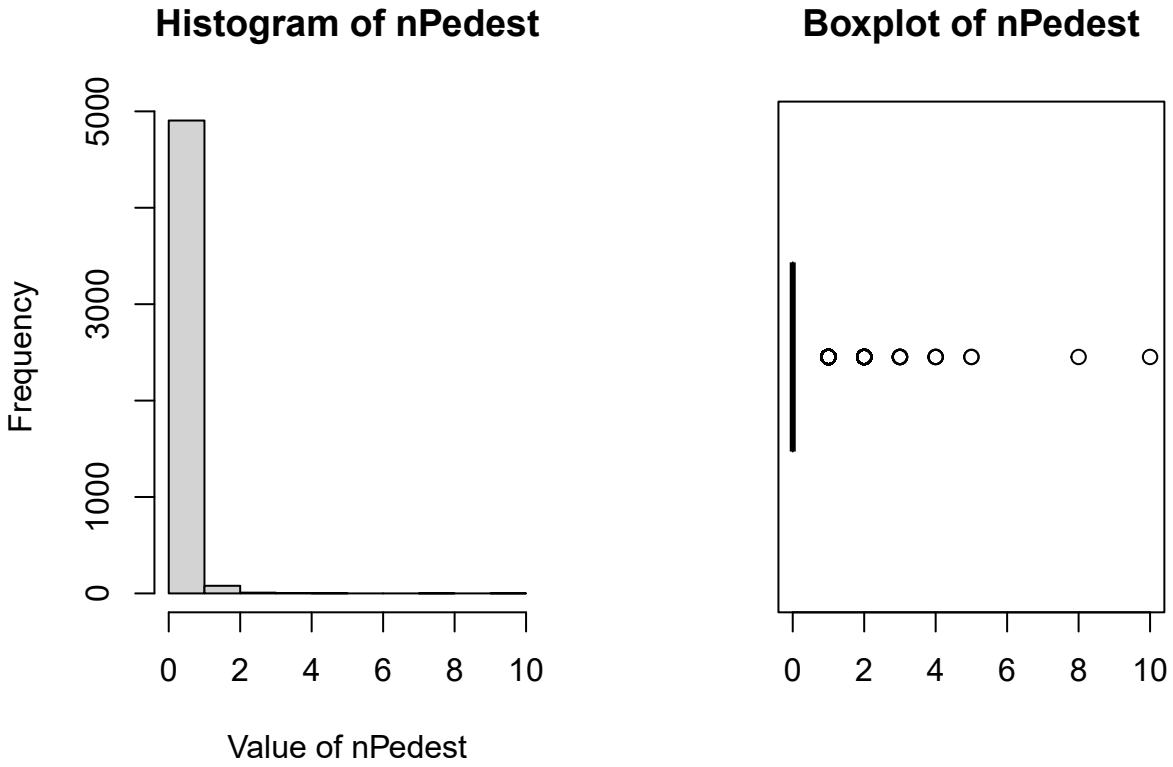
Variable 8 : nInvolv



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
1	1	2	1.8806	2	15	0.7815791	0.4156009	0

Table 6.10. nInvolv extended Summary Statistics.

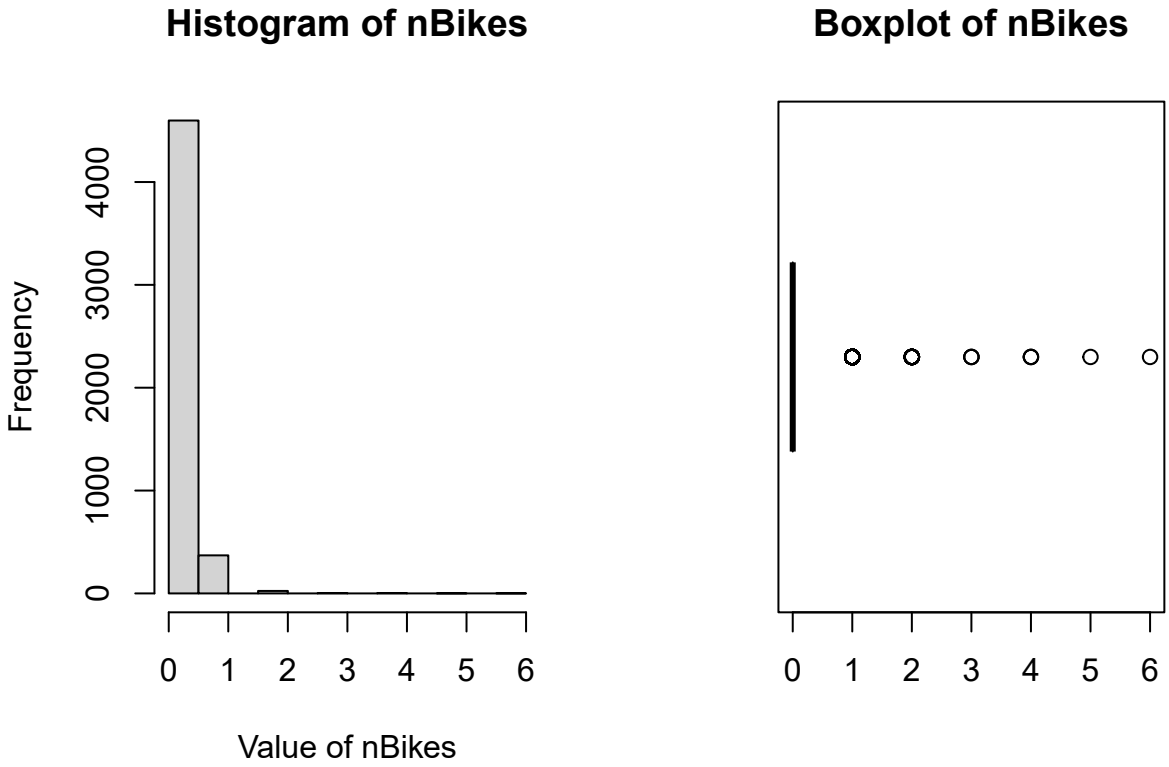
Variable 9 : nPedest



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	0.2456	0	10	0.5228148	2.128725	0

Table 6.11. nPedest extended Summary Statistics.

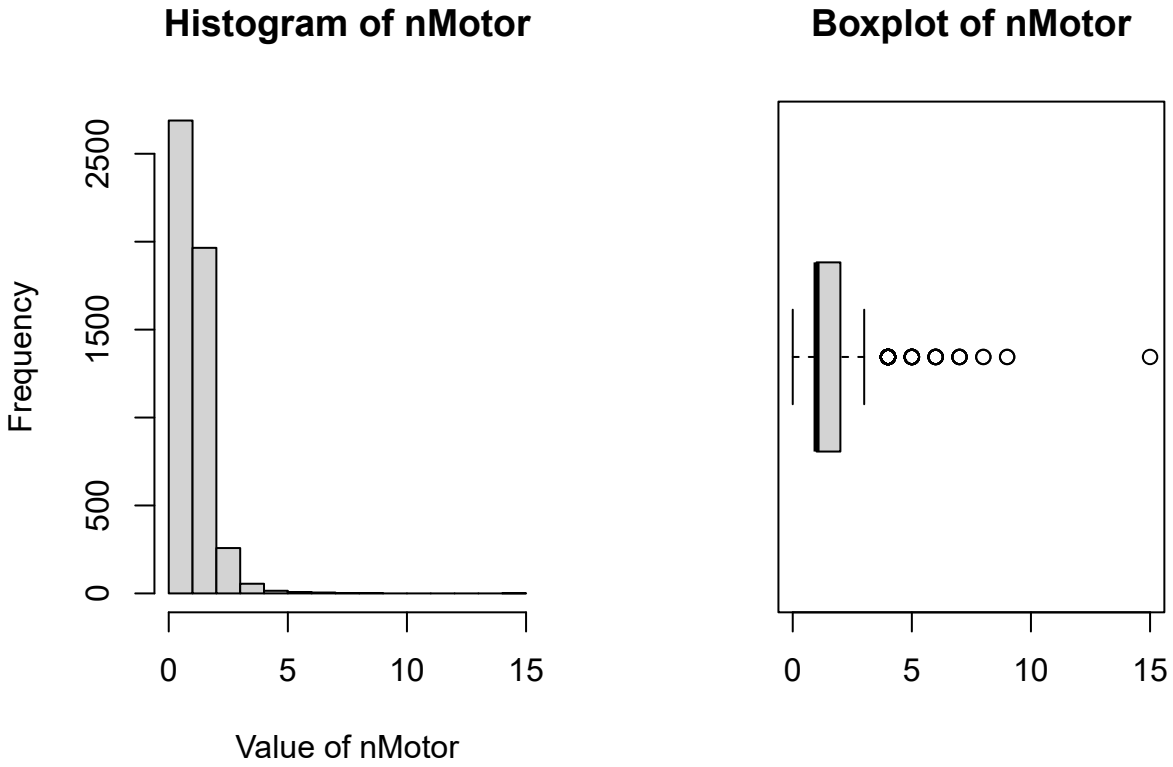
Variable 10 : nBikes



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	0.09	0	6	0.3351454	3.723838	0

Table 6.12. nBikes extended Summary Statistics.

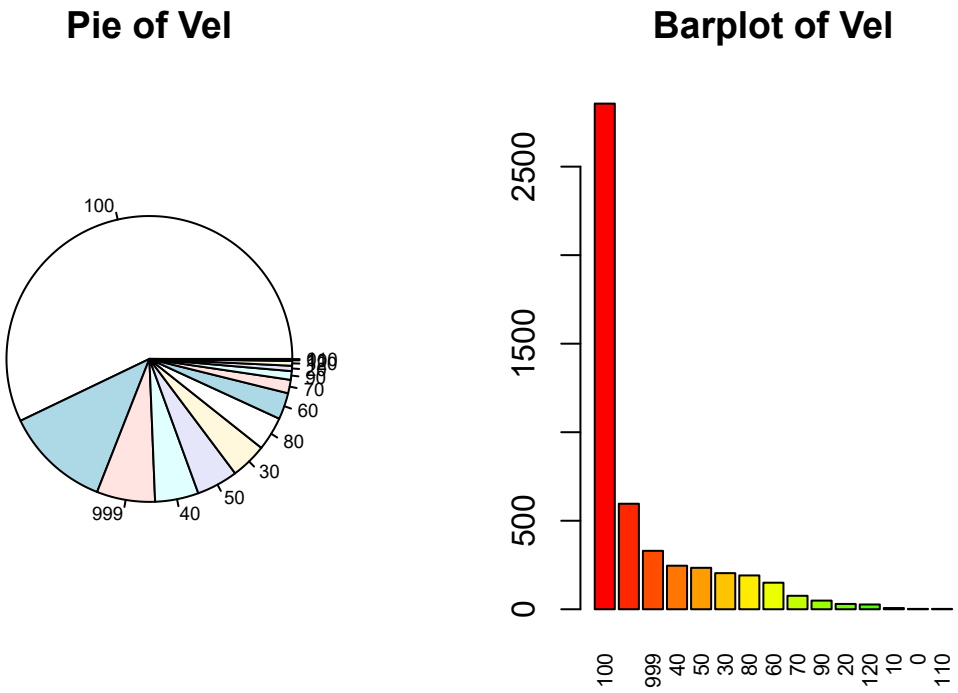
Variable 11 : nMotor



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	1	1	1.5214	2	15	0.8232117	0.5410883	0

Table 6.13. nMotor extended Summary Statistics.

Variable 12 : Vel

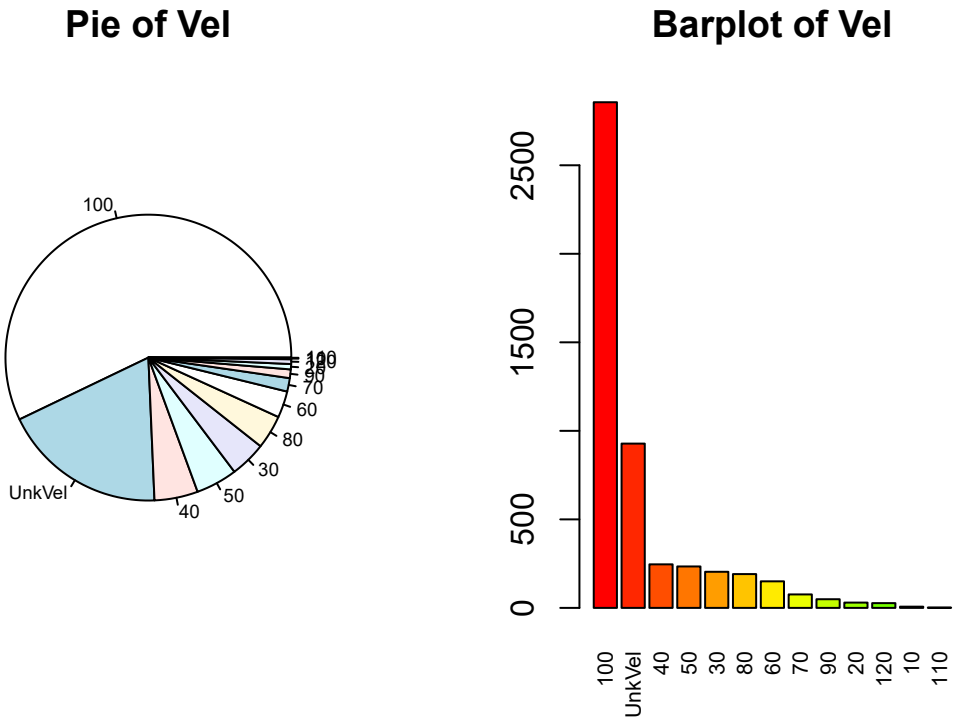


Number of modalities: 15

Vel	Frequency	Proportion
100	2856	0.5712
NA	596	0.1192
999	330	0.0660
40	246	0.0492
50	234	0.0468
30	204	0.0408
80	191	0.0382
60	150	0.0300
70	76	0.0152
90	49	0.0098
20	30	0.0060
120	27	0.0054
10	7	0.0014
0	2	0.0004
110	2	0.0004

Table 6.14. Vel frequency and proportion table (sorted).

Variable 12 : Vel (CHANGED in preprocessing)

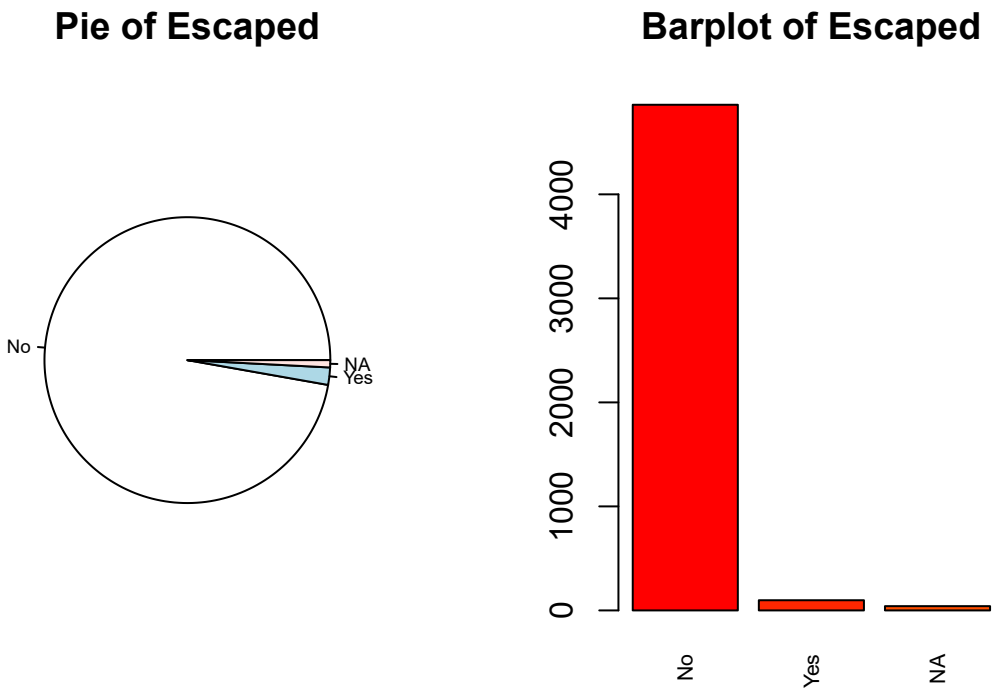


Number of modalities: 13

Vel	Frequency	Proportion
100	2856	0.5712
UnkVel	928	0.1856
40	246	0.0492
50	234	0.0468
30	204	0.0408
80	191	0.0382
60	150	0.0300
70	76	0.0152
90	49	0.0098
20	30	0.0060
120	27	0.0054
10	7	0.0014
110	2	0.0004

Table 6.15. Vel frequency and proportion table (sorted).

Variable 13 : Escaped

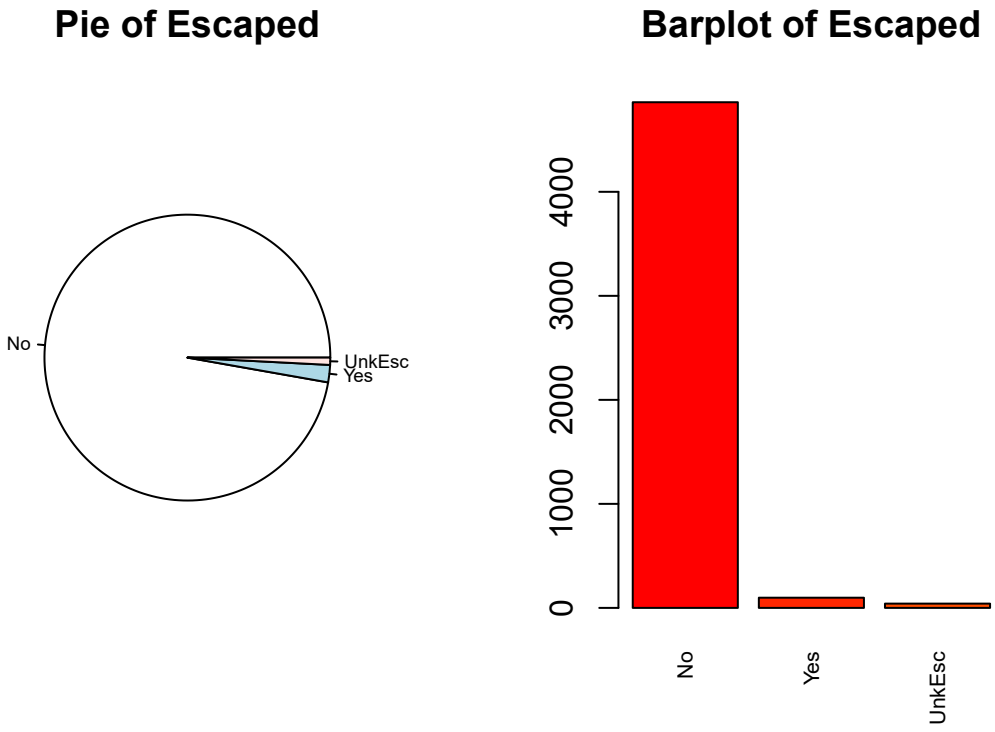


Number of modalities: 3

Escaped	Frequency	Proportion
No	4861	0.9722
Yes	98	0.0196
NA	41	0.0082

Table 6.16. Escaped frequency and proportion table (sorted).

Variable 13 : Escaped (CHANGED in preprocessing)

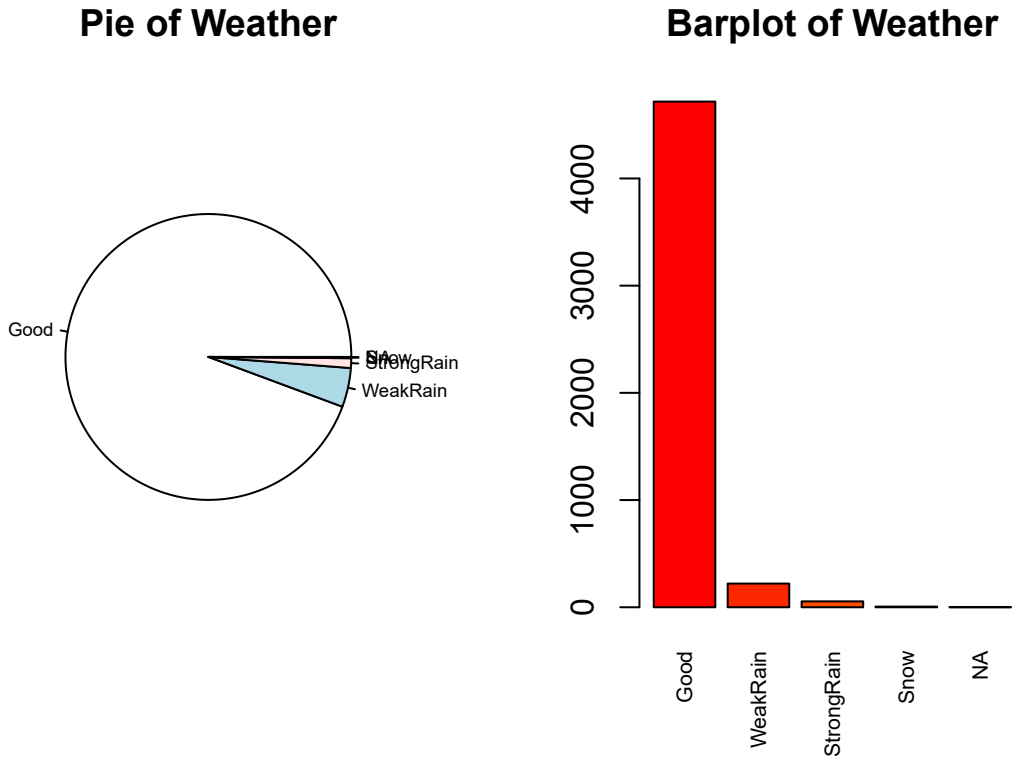


Number of modalities: 3

Escaped	Frequency	Proportion
No	4861	0.9722
Yes	98	0.0196
UnkEsc	41	0.0082

Table 6.17. Escaped frequency and proportion table (sorted).

Variable 14 : Weather

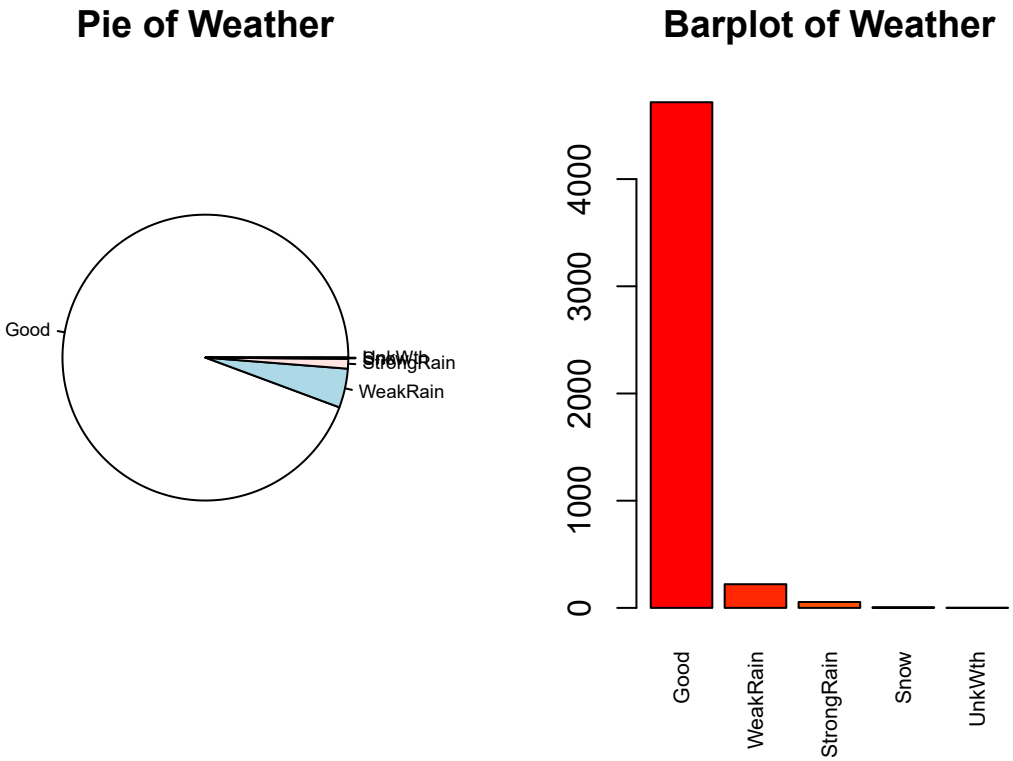


Number of modalities: 5

Weather	Frequency	Proportion
Good	4717	0.9434
WeakRain	221	0.0442
StrongRain	55	0.0110
Snow	6	0.0012
NA	1	0.0002

Table 6.18. Weather frequency and proportion table (sorted).

Variable 14 : Weather (CHANGED in preprocessing)

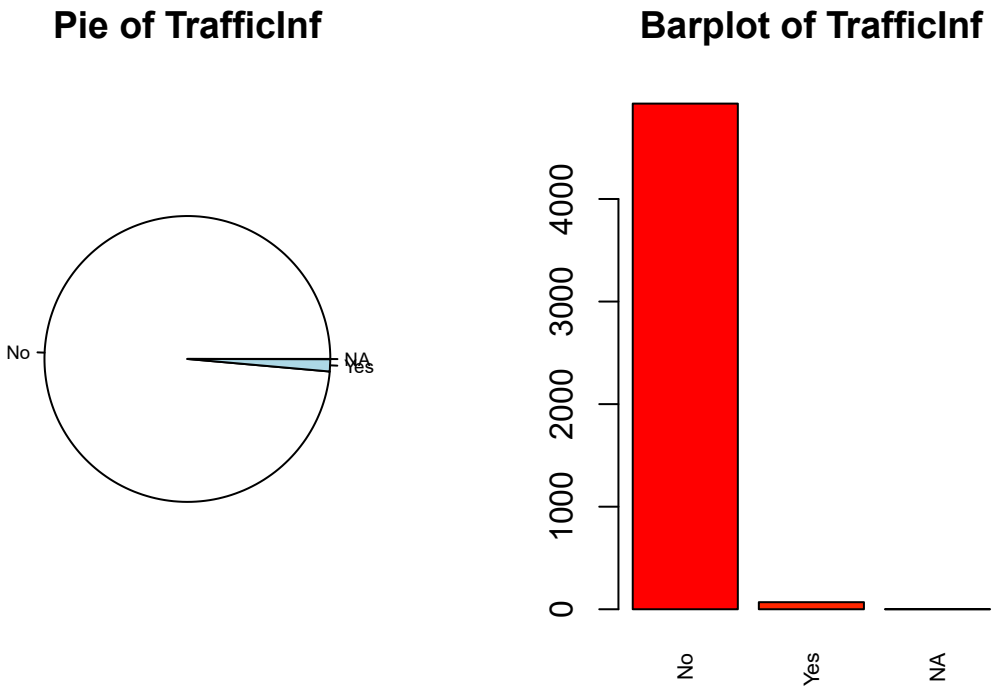


Number of modalities: 5

Weather	Frequency	Proportion
Good	4717	0.9434
WeakRain	221	0.0442
StrongRain	55	0.0110
Snow	6	0.0012
UnkWth	1	0.0002

Table 6.19. Weather frequency and proportion table (sorted).

Variable 15 : TrafficInf

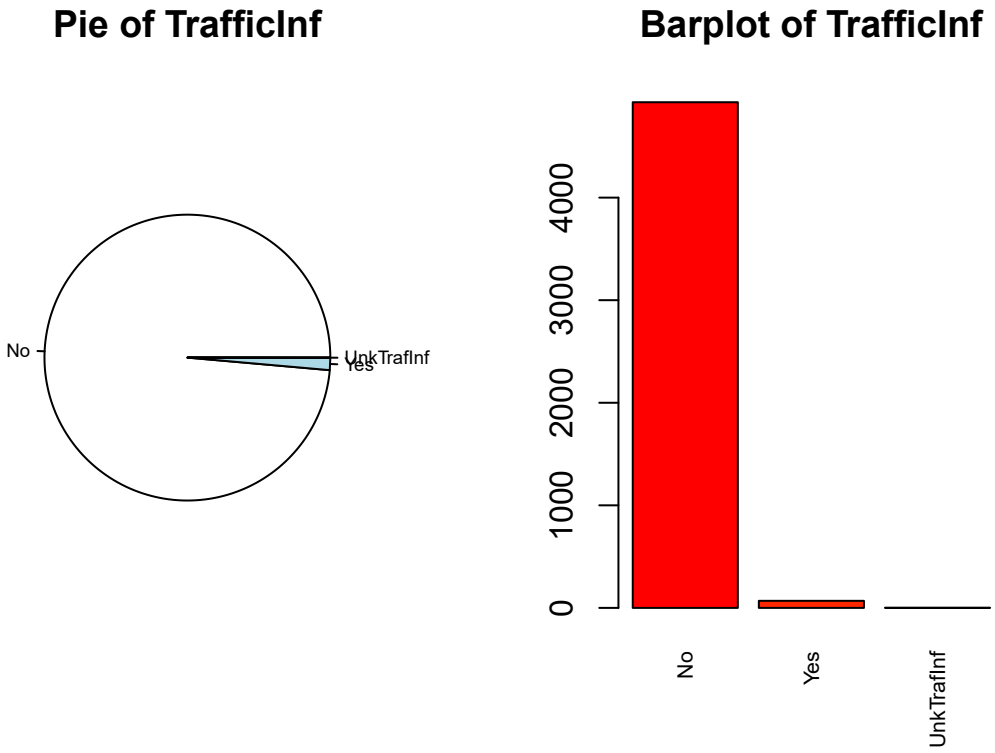


Number of modalities: 3

TrafficInf	Frequency	Proportion
No	4930	0.9860
Yes	69	0.0138
NA	1	0.0002

Table 6.20. TrafficInf frequency and proportion table (sorted).

Variable 15 : TrafficInf (CHANGED in preprocessing)

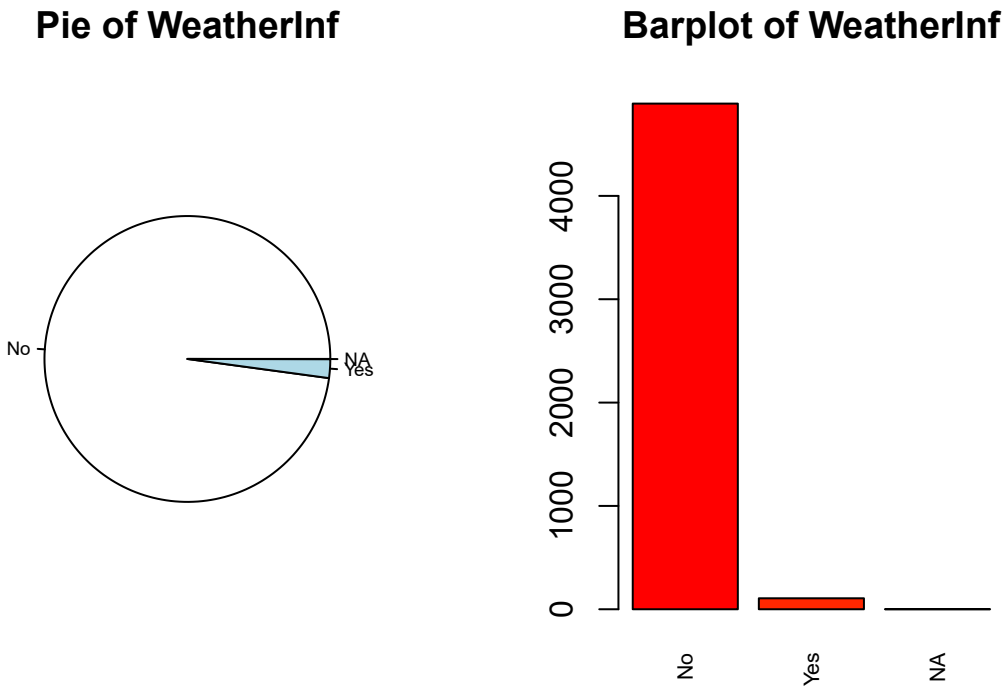


Number of modalities: 3

TrafficInf	Frequency	Proportion
No	4930	0.9860
Yes	69	0.0138
UnkTrafInf	1	0.0002

Table 6.21. TrafficInf frequency and proportion table (sorted).

Variable 16 : WeatherInf

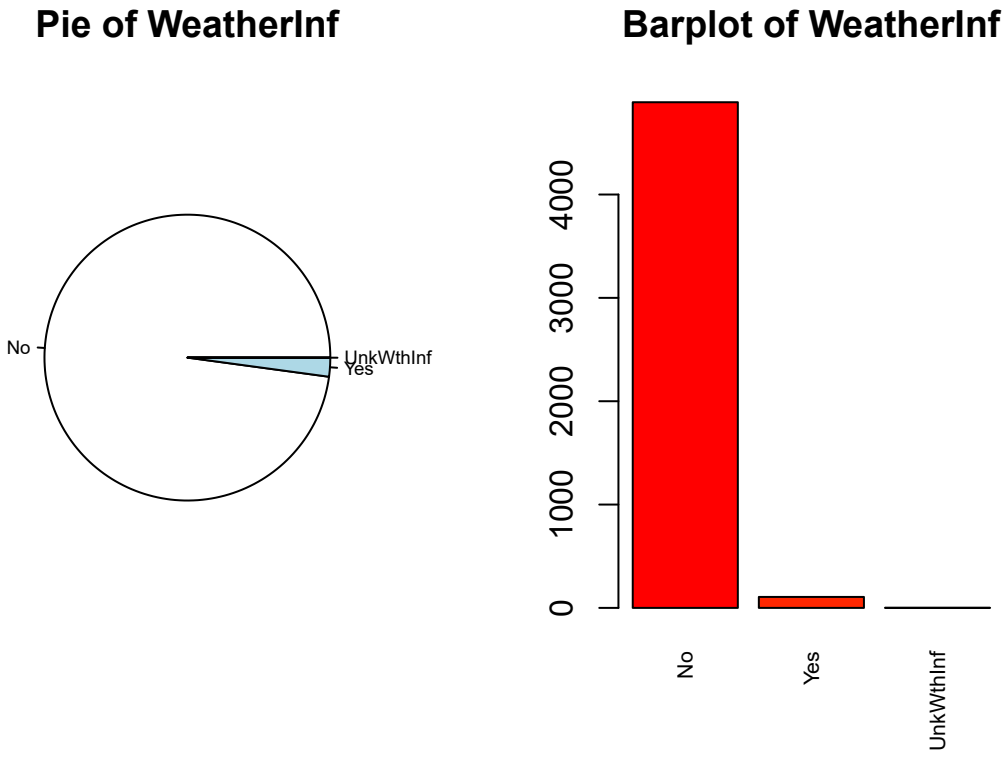


Number of modalities: 3

WeatherInf	Frequency	Proportion
No	4893	0.9786
Yes	106	0.0212
NA	1	0.0002

Table 6.22. WeatherInf frequency and proportion table (sorted).

Variable 16 : WeatherInf (CHANGED in preprocessing)

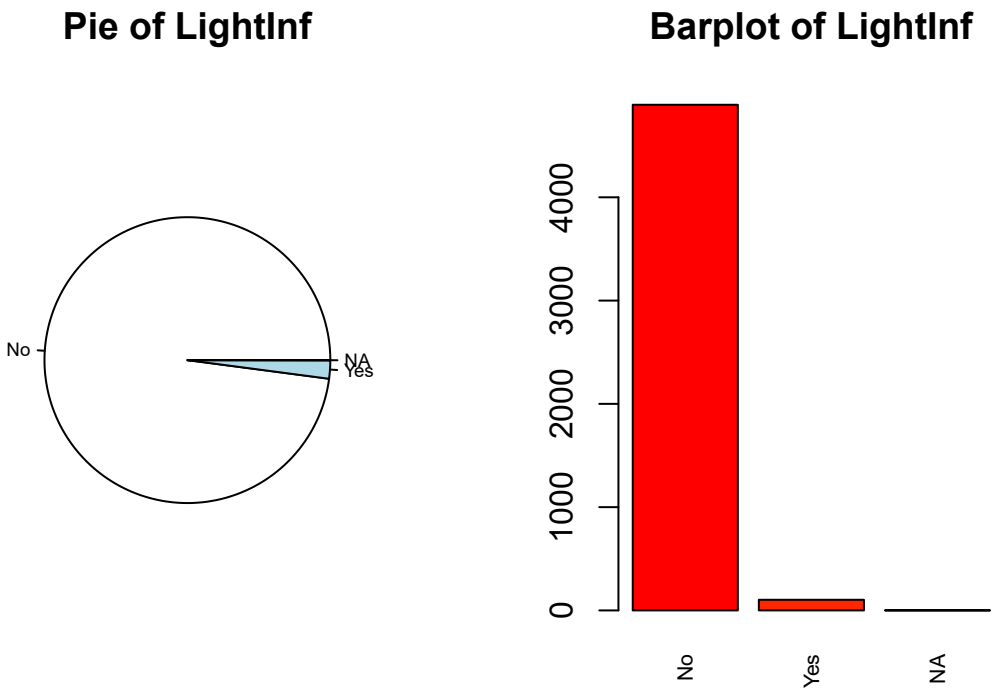


Number of modalities: 3

WeatherInf	Frequency	Proportion
No	4893	0.9786
Yes	106	0.0212
UnkWthInf	1	0.0002

Table 6.23. WeatherInf frequency and proportion table (sorted).

Variable 17 : LightInf

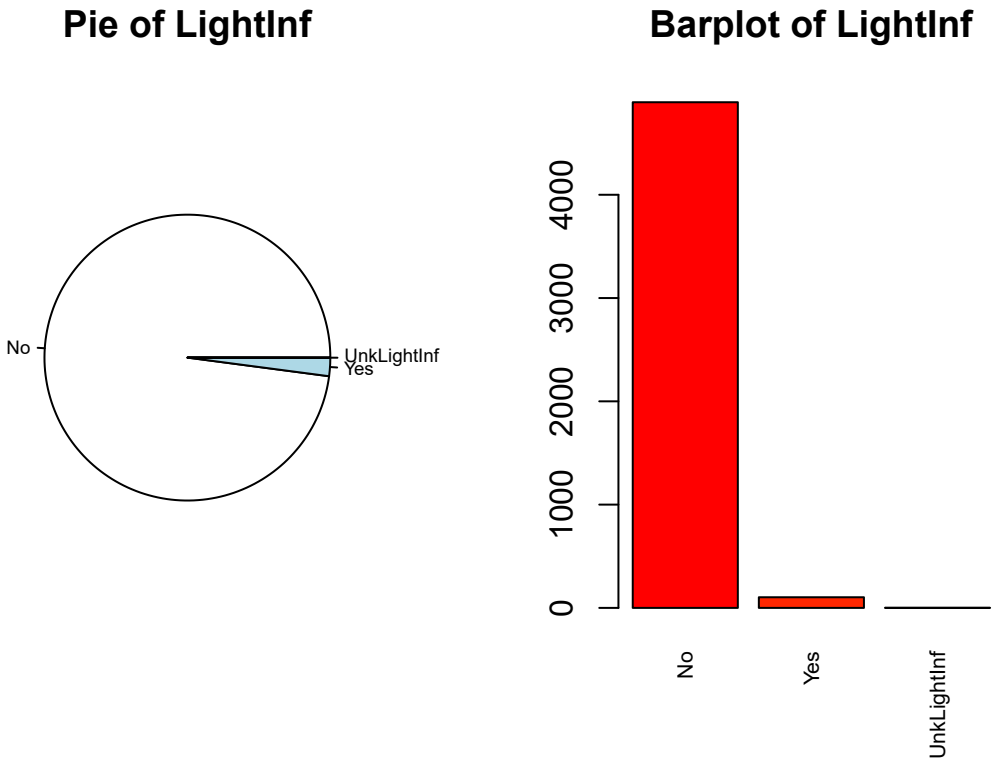


Number of modalities: 3

LightInf	Frequency	Proportion
No	4896	0.9792
Yes	103	0.0206
NA	1	0.0002

Table 6.24. LightInf frequency and proportion table (sorted).

Variable 17 : LightInf (CHANGED in preprocessing)

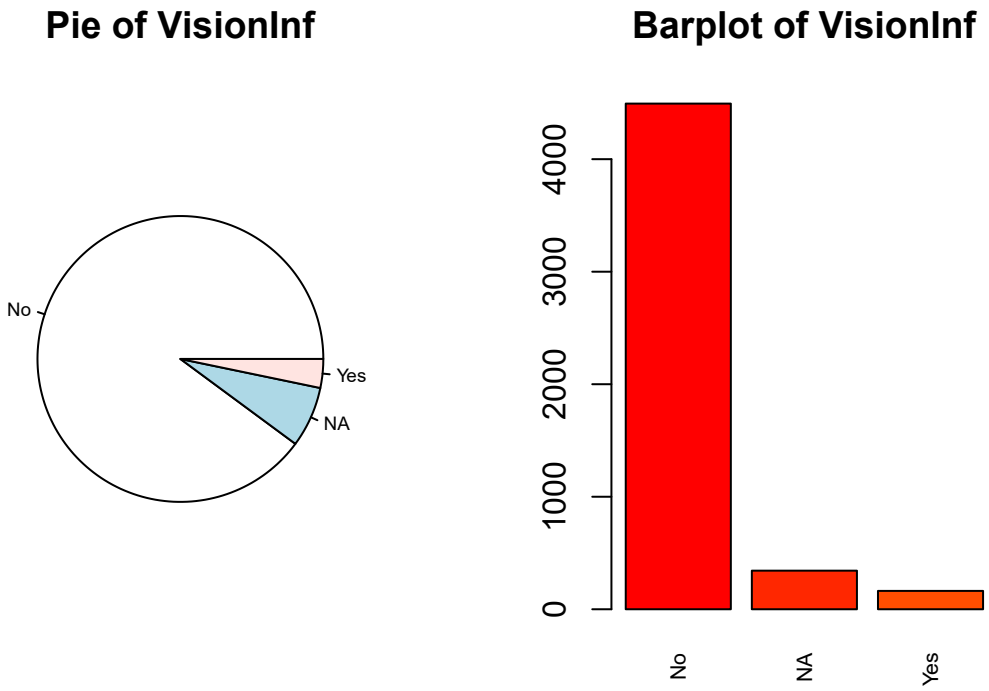


Number of modalities: 3

LightInf	Frequency	Proportion
No	4896	0.9792
Yes	103	0.0206
UnkLightInf	1	0.0002

Table 6.25. LightInf frequency and proportion table (sorted).

Variable 18 : VisionInf

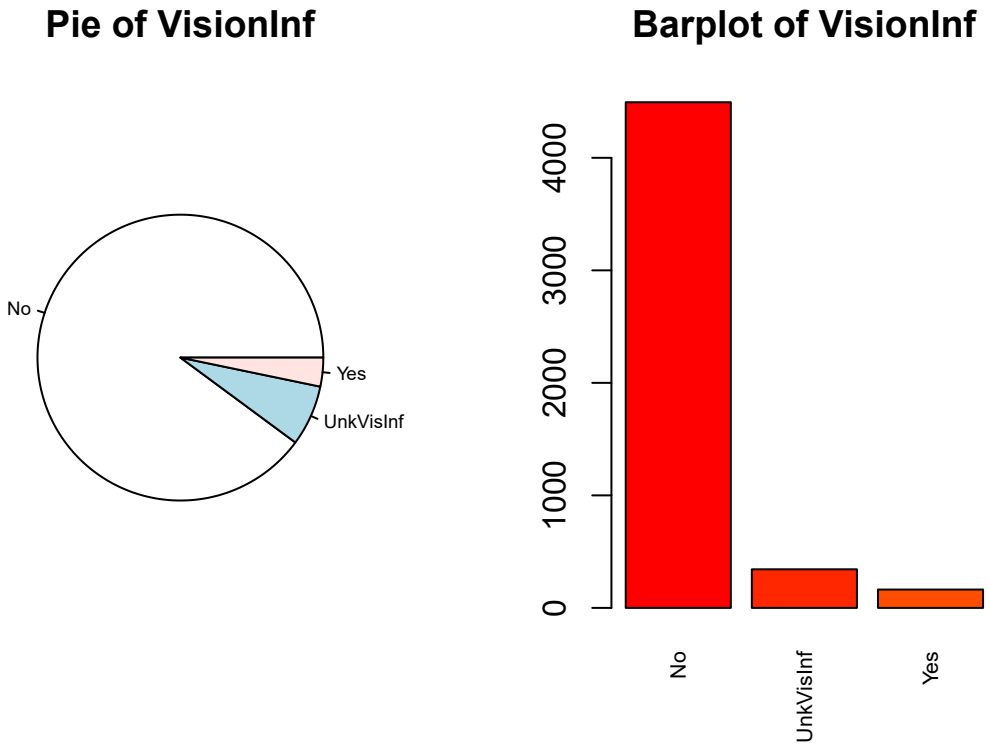


Number of modalities: 3

VisionInf	Frequency	Proportion
No	4494	0.8988
NA	343	0.0686
Yes	163	0.0326

Table 6.26. VisionInf frequency and proportion table (sorted).

Variable 18 : VisionInf (CHANGED in preprocessing)

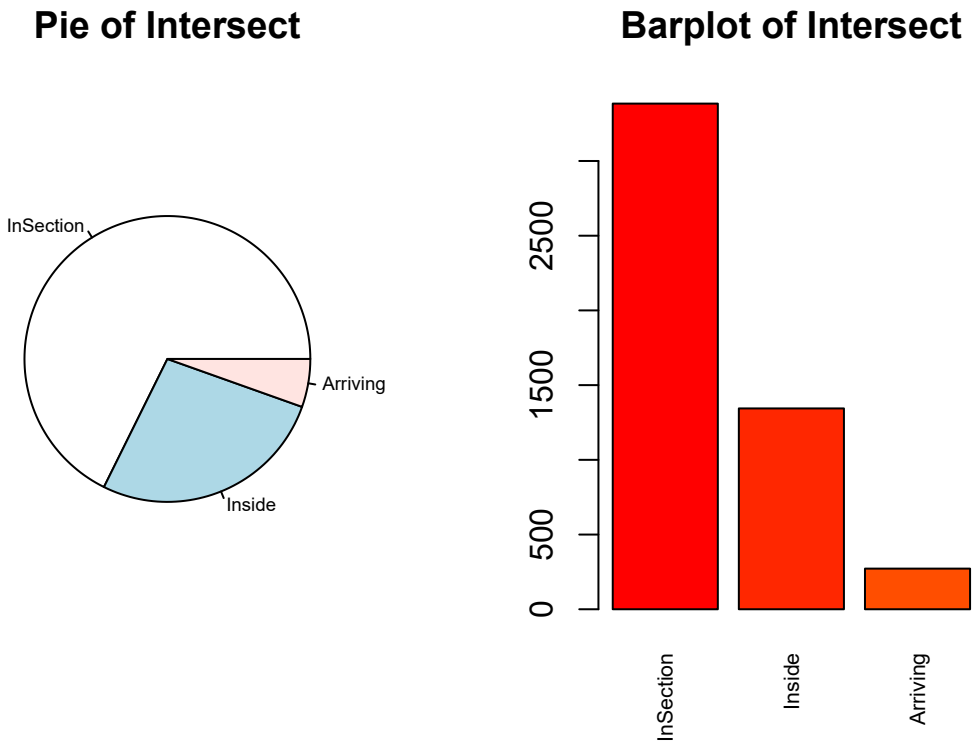


Number of modalities: 3

VisionInf	Frequency	Proportion
No	4494	0.8988
UnkVisInf	343	0.0686
Yes	163	0.0326

Table 6.27. VisionInf frequency and proportion table (sorted).

Variable 19 : Intersect

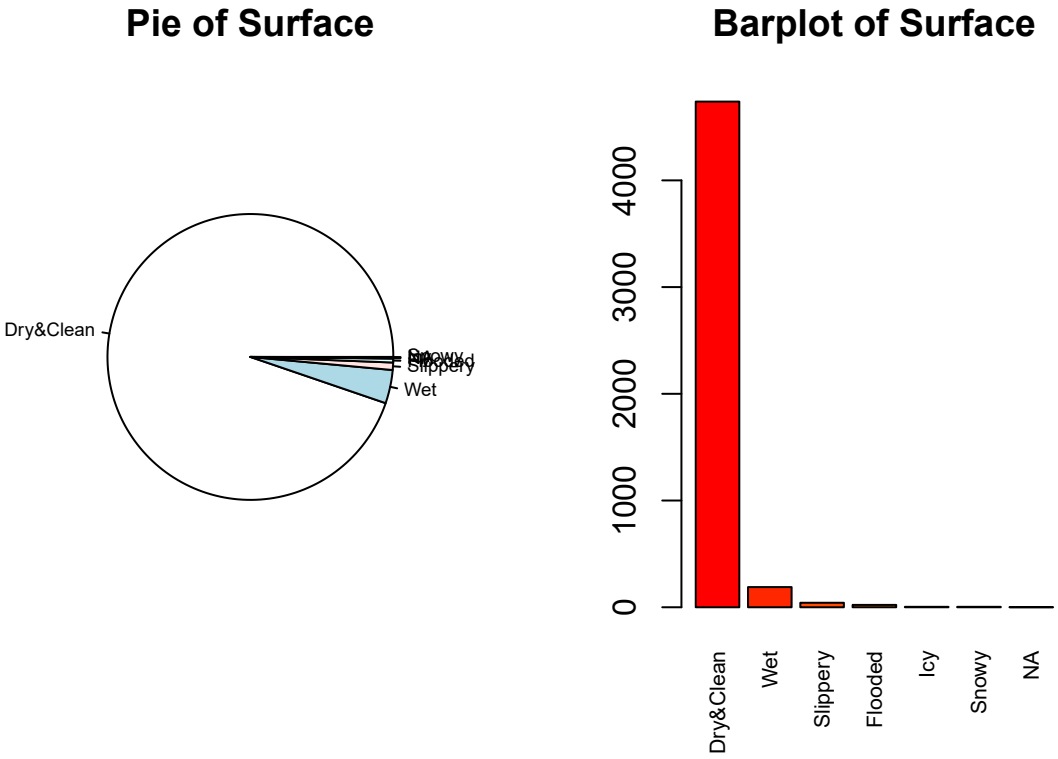


Number of modalities: 3

Intersect	Frequency	Proportion
InSection	3384	0.6768
Inside	1344	0.2688
Arriving	272	0.0544

Table 6.28. Intersect frequency and proportion table (sorted).

Variable 20 : Surface

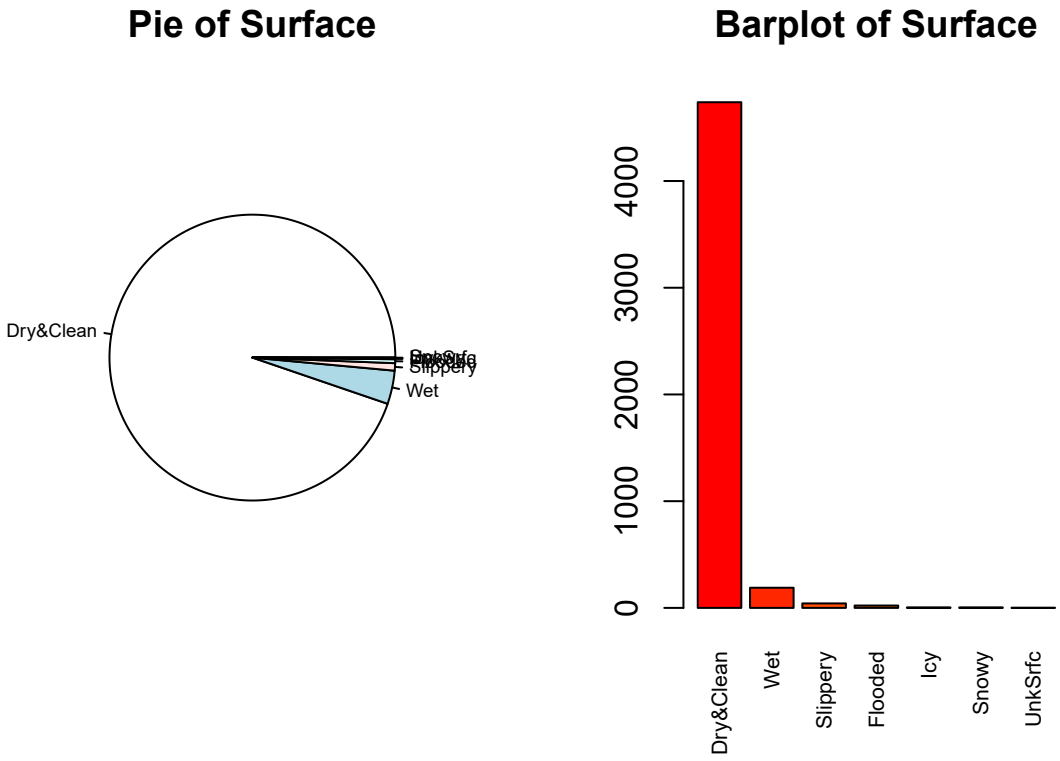


Number of modalities: 7

Surface	Frequency	Proportion
Dry&Clean	4738	0.9476
Wet	189	0.0378
Slippery	42	0.0084
Flooded	22	0.0044
Icy	4	0.0008
Snowy	4	0.0008
NA	1	0.0002

Table 6.29. Surface frequency and proportion table (sorted).

Variable 20 : Surface (CHANGED in preprocessing)

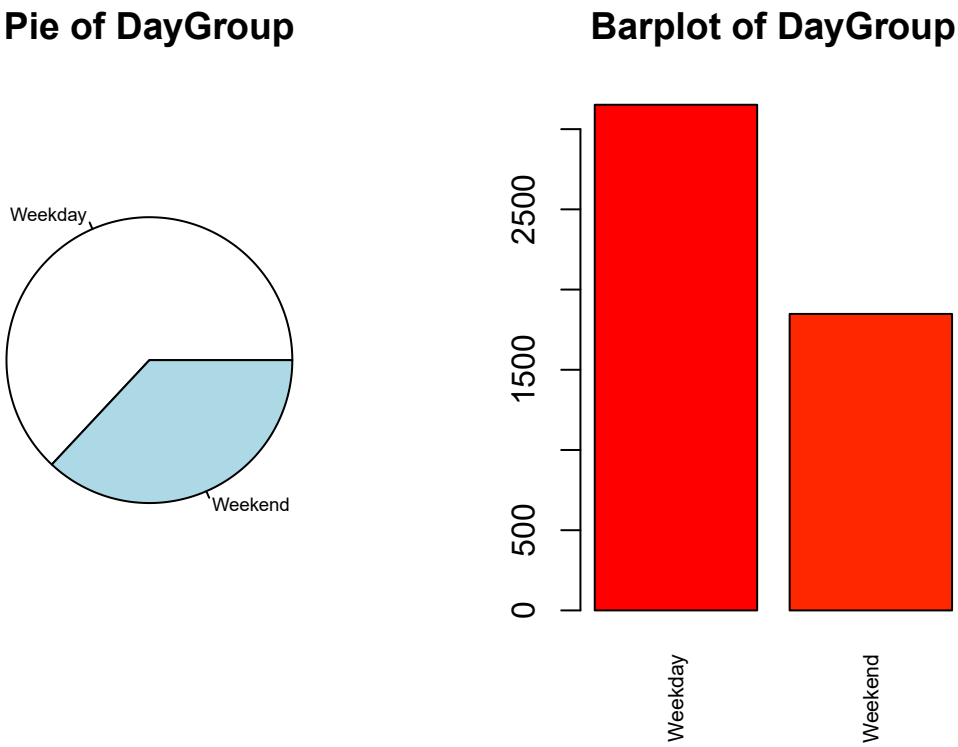


Number of modalities: 7

Surface	Frequency	Proportion
Dry&Clean	4738	0.9476
Wet	189	0.0378
Slippery	42	0.0084
Flooded	22	0.0044
Icy	4	0.0008
Snowy	4	0.0008
UnkSrfc	1	0.0002

Table 6.30. Surface frequency and proportion table (sorted).

Variable 21 : DayGroup

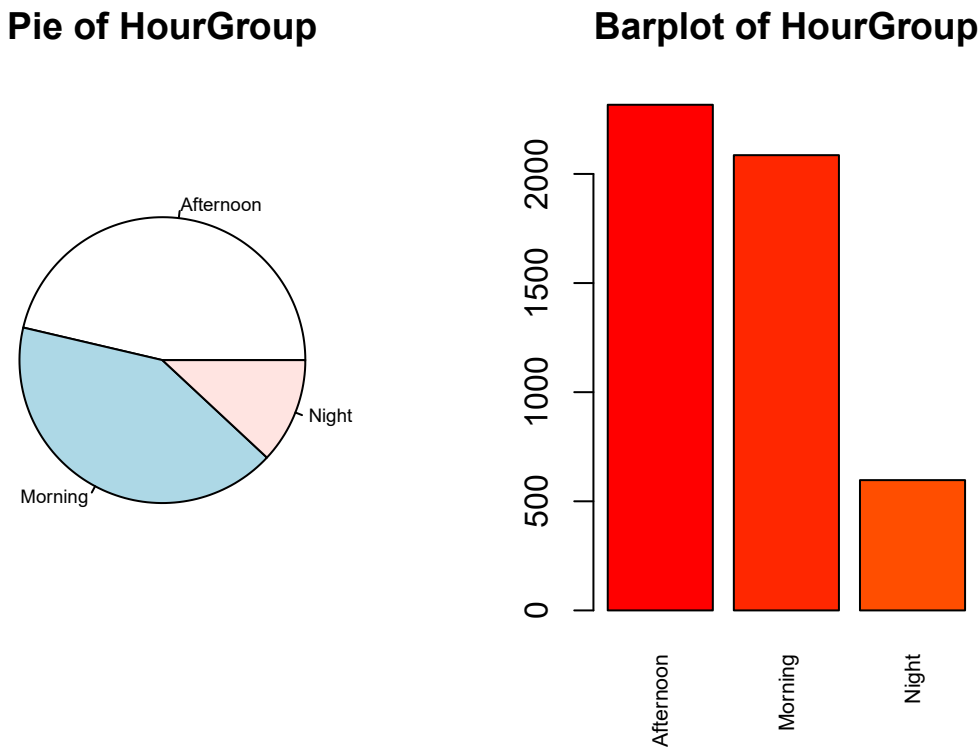


Number of modalities: 2

DayGroup	Frequency	Proportion
Weekday	3152	0.6304
Weekend	1848	0.3696

Table 6.31. DayGroup frequency and proportion table (sorted).

Variable 22 : HourGroup

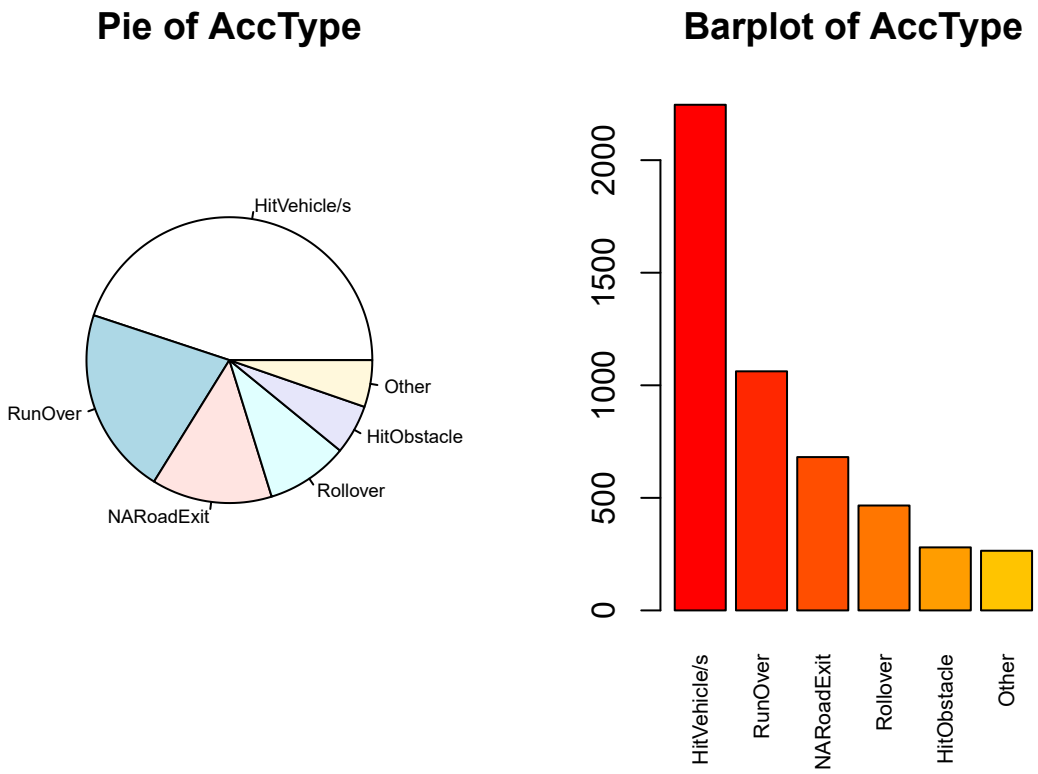


Number of modalities: 3

HourGroup	Frequency	Proportion
Afternoon	2317	0.4634
Morning	2086	0.4172
Night	597	0.1194

Table 6.32. HourGroup frequency and proportion table (sorted).

Variable 23 : AccType

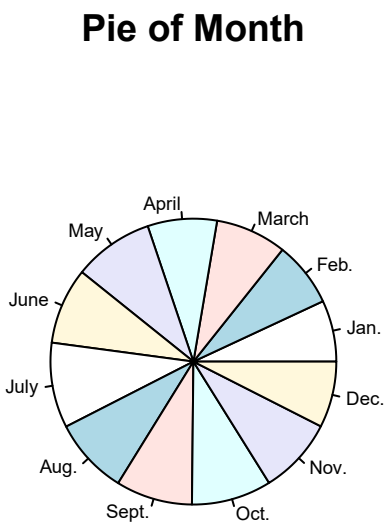


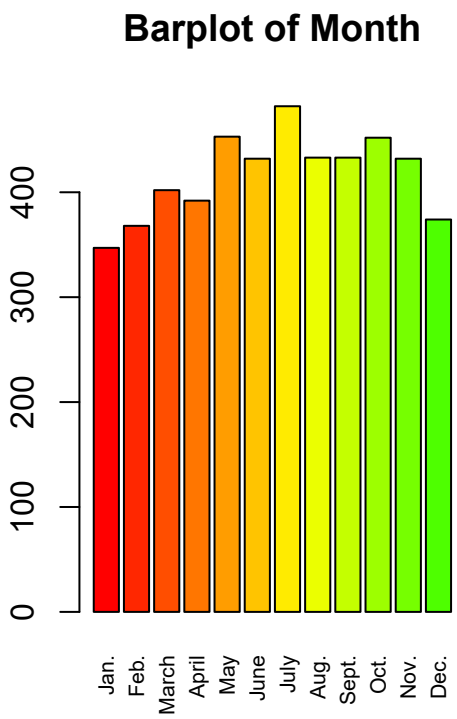
Number of modalities: 6

AccType	Frequency	Proportion
HitVehicle/s	2246	0.4492
RunOver	1062	0.2124
NARoadExit	681	0.1362
Rollover	466	0.0932
HitObstacle	280	0.0560
Other	265	0.0530

Table 6.33. AccType frequency and proportion table (sorted).

Variable 24 : Month (ADDED in preprocessing)



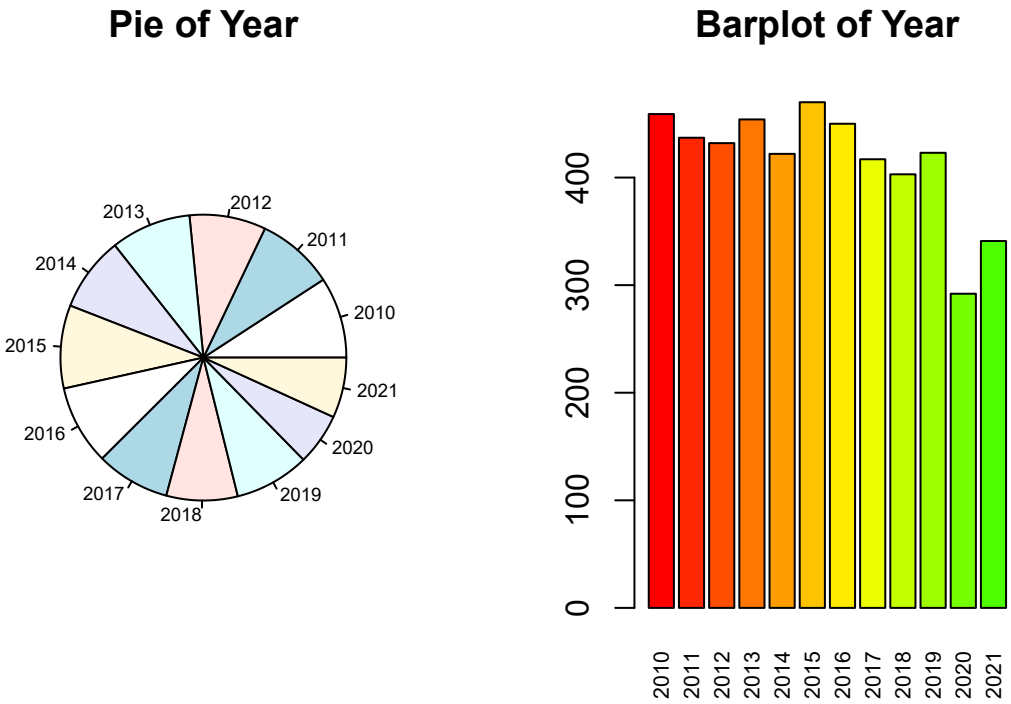


Number of modalities: 12

Month	Frequency	Proportion
Jan.	347	0.0694
Feb.	368	0.0736
March	402	0.0804
April	392	0.0784
May	453	0.0906
June	432	0.0864
July	482	0.0964
Aug.	433	0.0866
Sept.	433	0.0866
Oct.	452	0.0904
Nov.	432	0.0864
Dec.	374	0.0748

Table 6.34. Month frequency and proportion table (sorted chronologically).

Variable 25 : Year (ADDED in preprocessing)

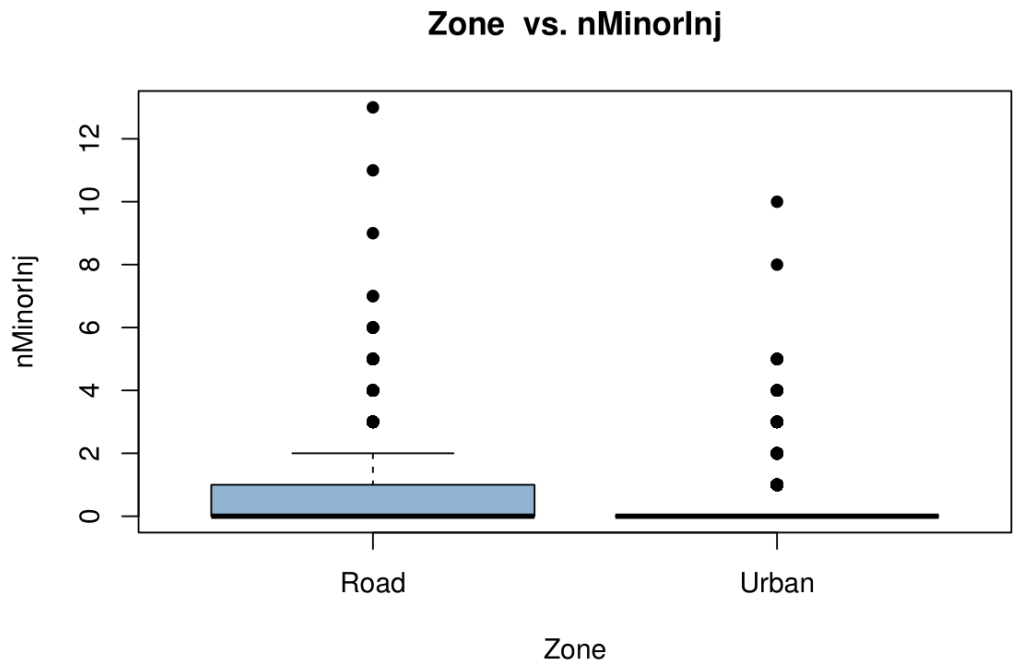


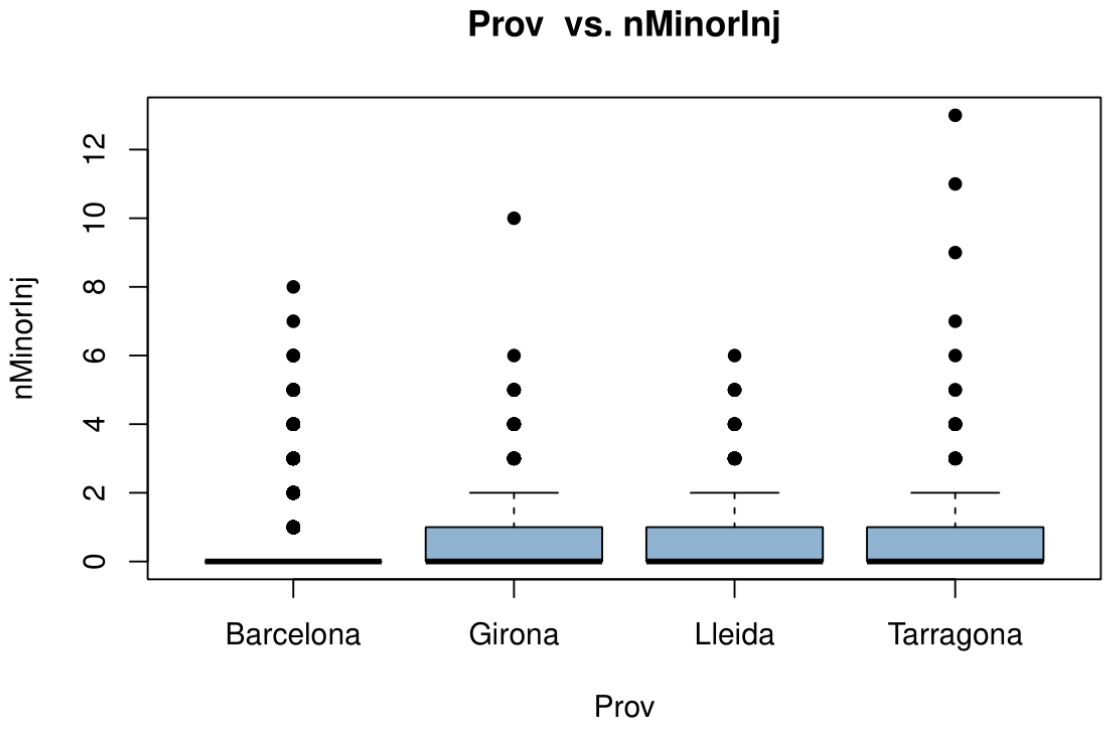
Number of modalities: 12

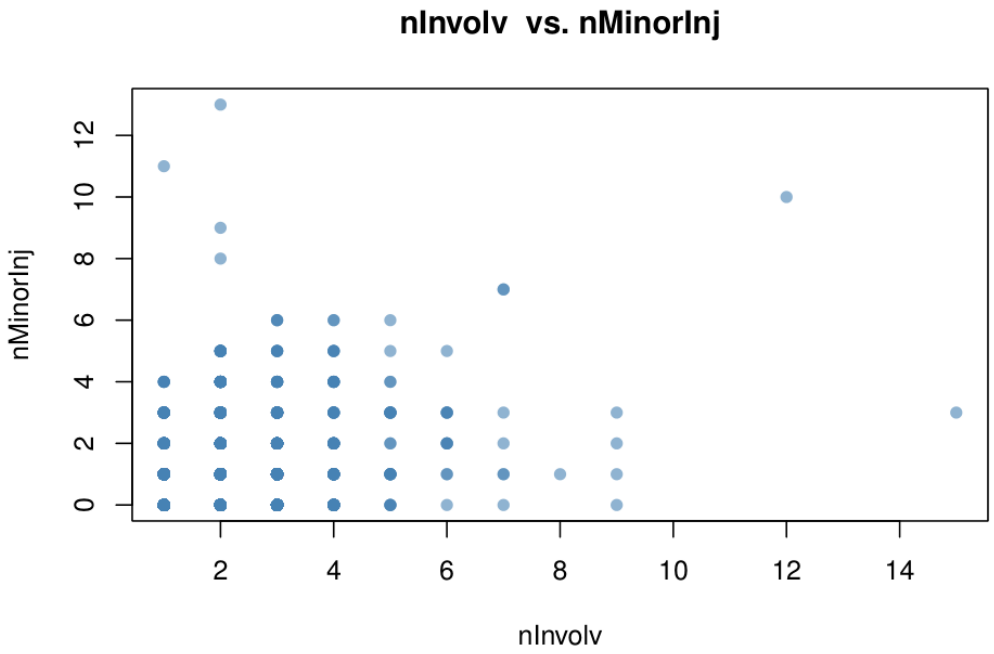
Year	Frequency	Proportion
2010	459	0.0918
2011	437	0.0874
2012	432	0.0864
2013	454	0.0908
2014	422	0.0844
2015	470	0.0940
2016	450	0.0900
2017	417	0.0834
2018	403	0.0806
2019	423	0.0846
2020	292	0.0584
2021	341	0.0682

Table 6.35. Year frequency and proportion table (sorted chronologically).

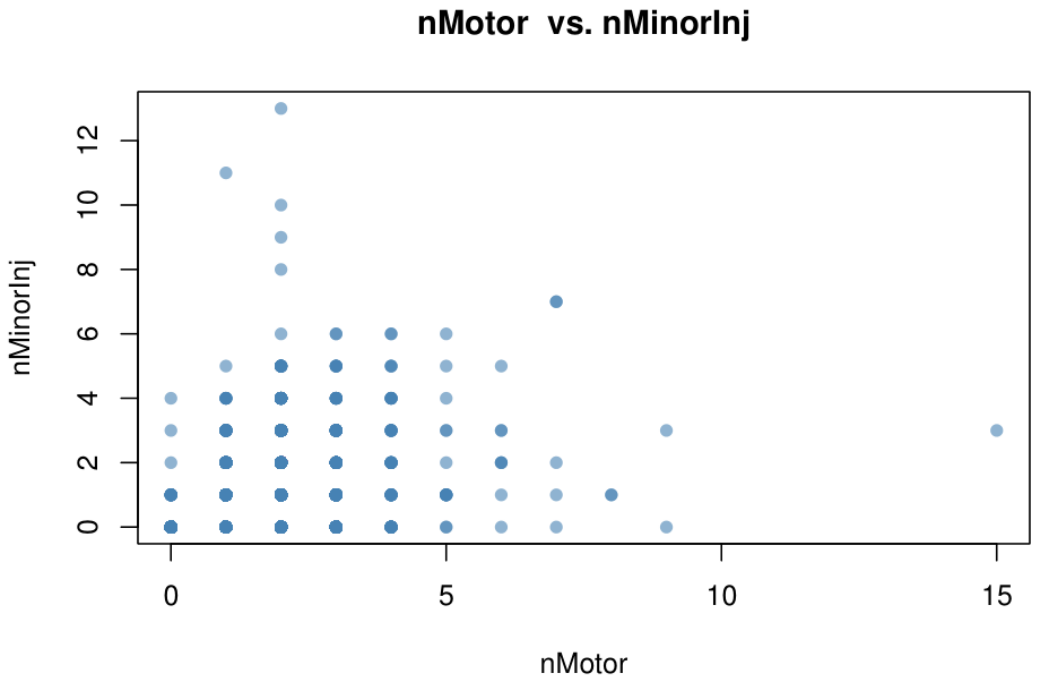
Additional bivariate plots



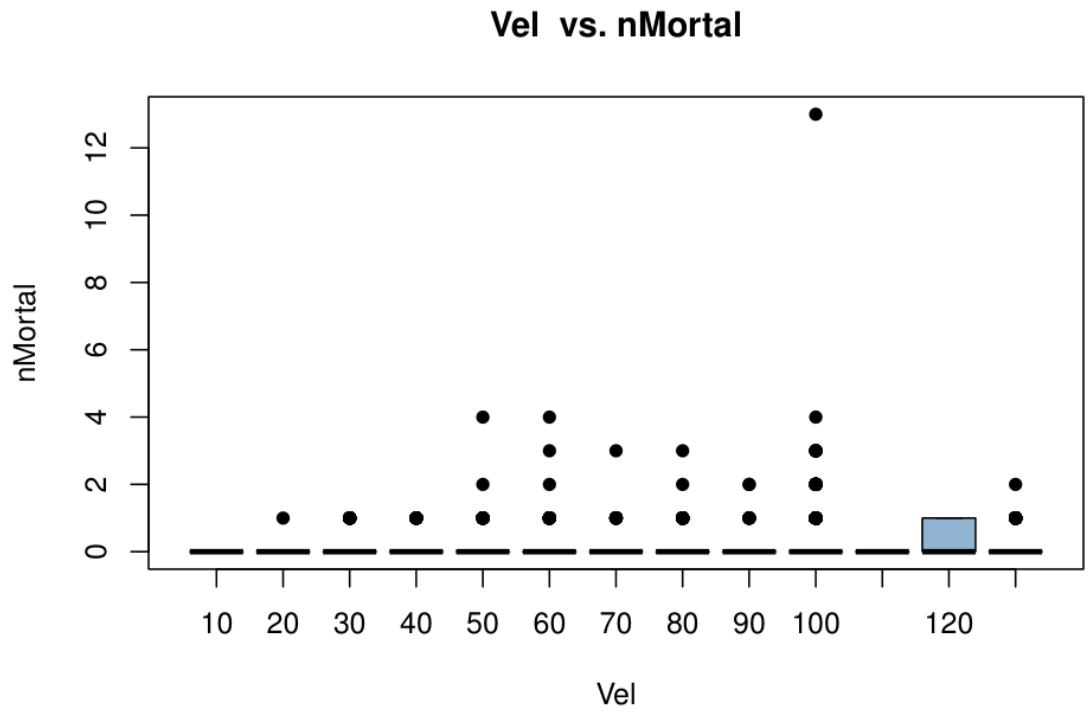


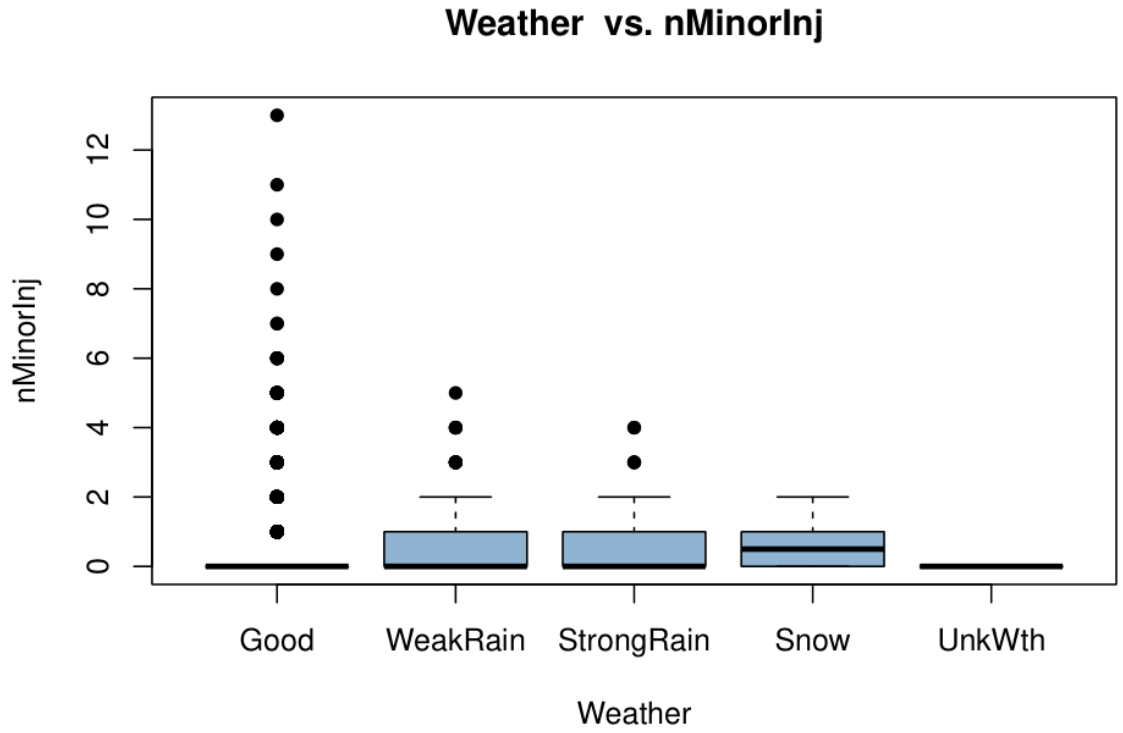


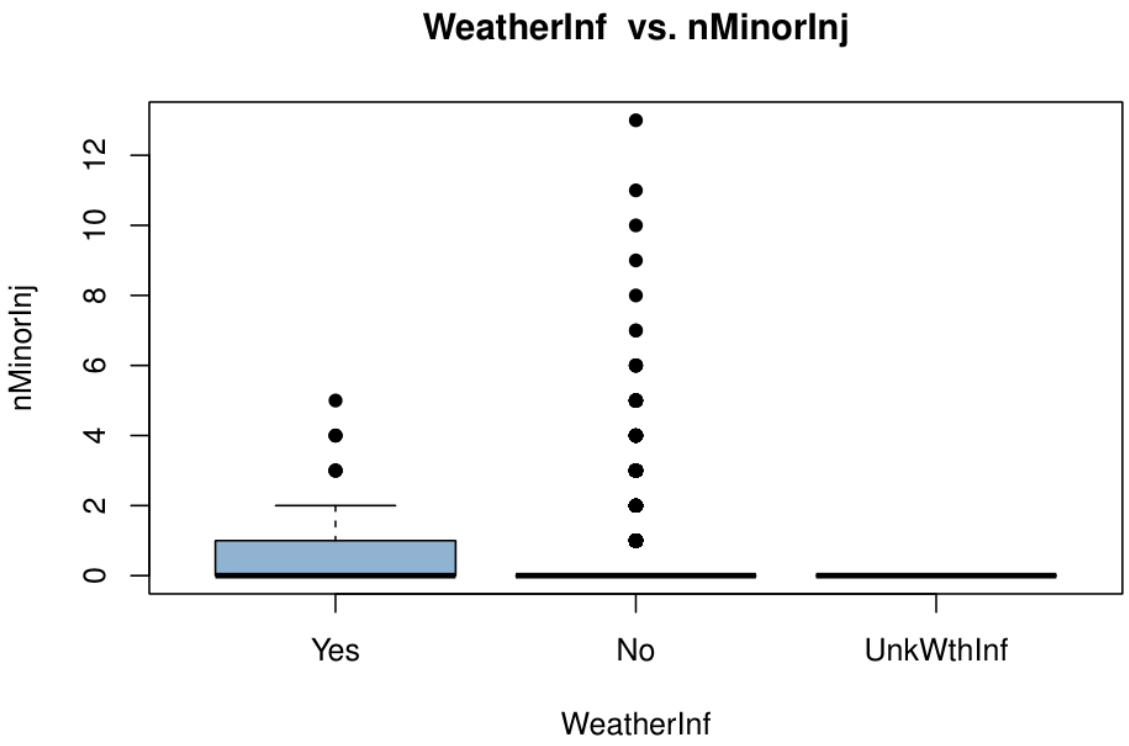
Correlation between nInvolv and nMinorInj: ***0.308329601929549***

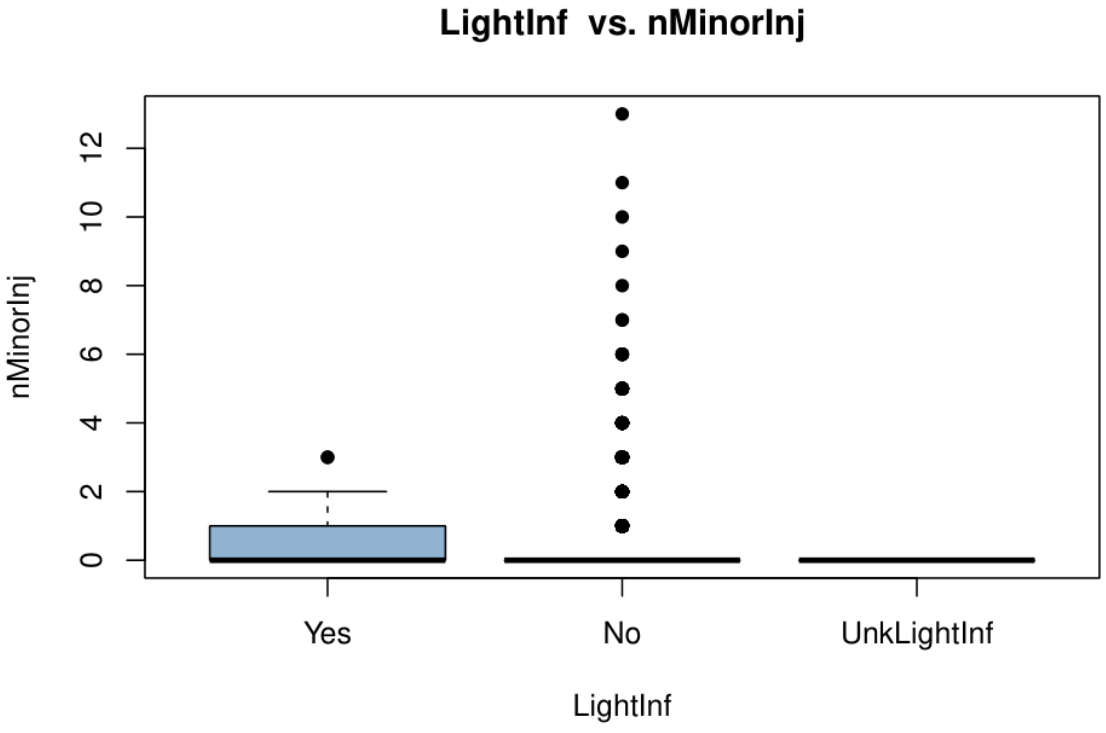


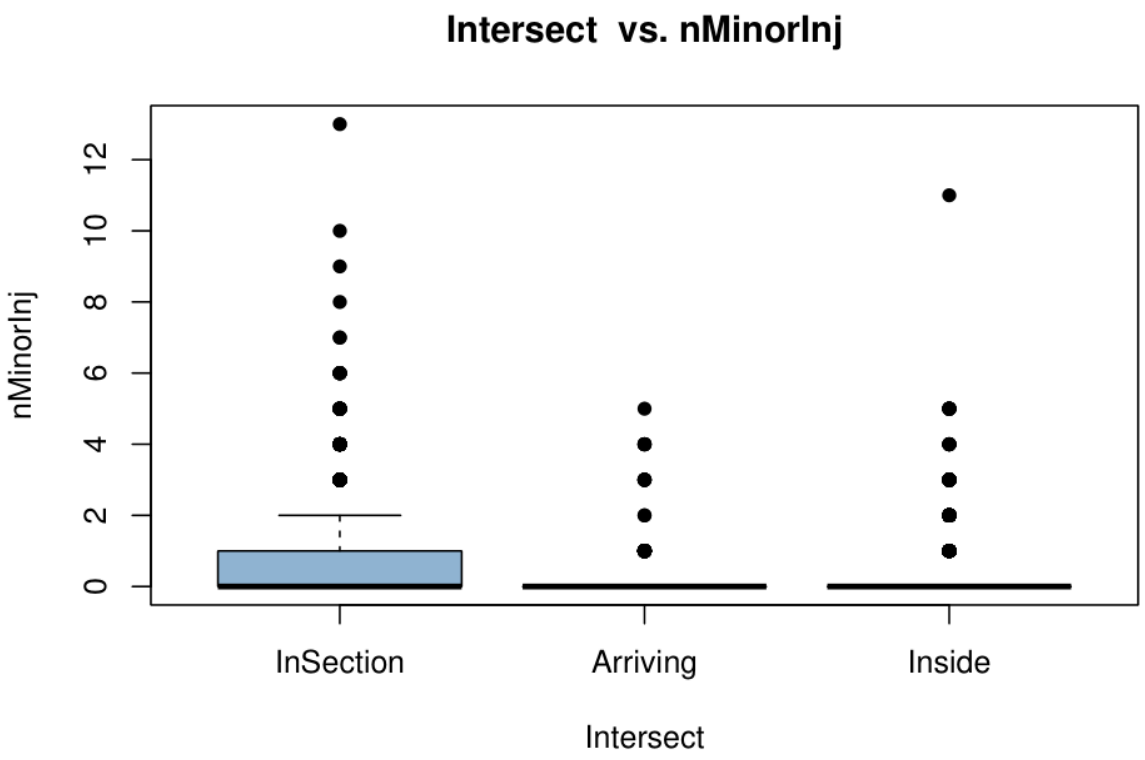
Correlation between nMotor and nMinorInj: ***0.325637603518631***

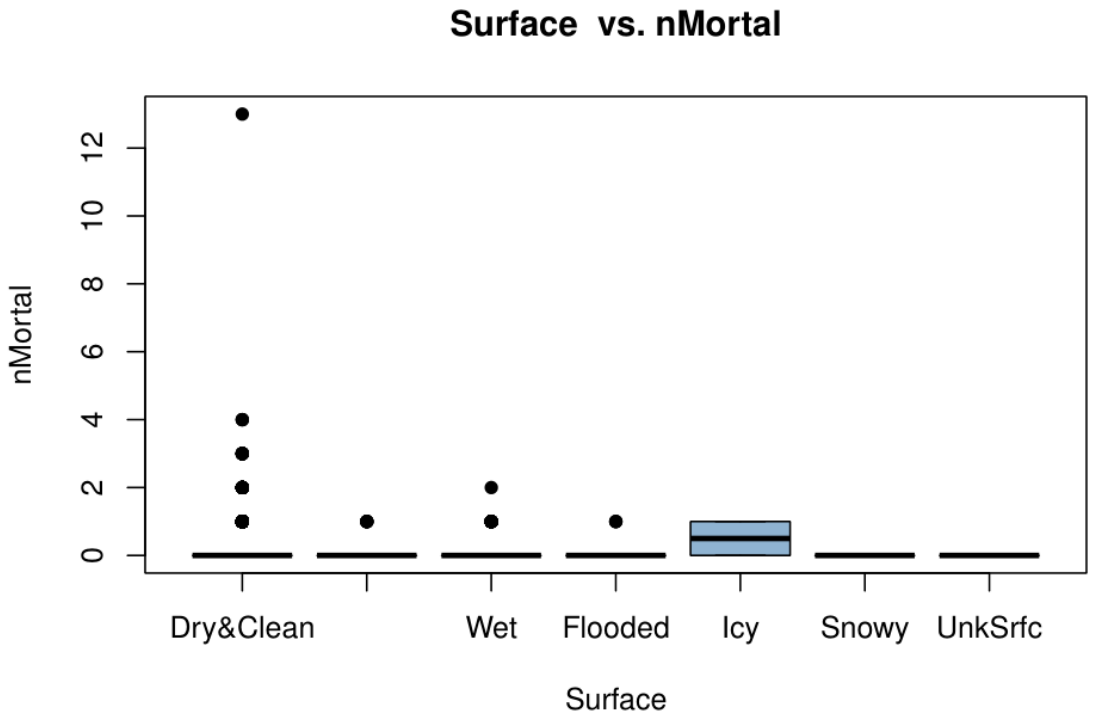

Vel vs. nMortal
nMortal
0
2
4
6
8
10
12
10
20
30
40
50
60
70
80
90
100
120
Vel

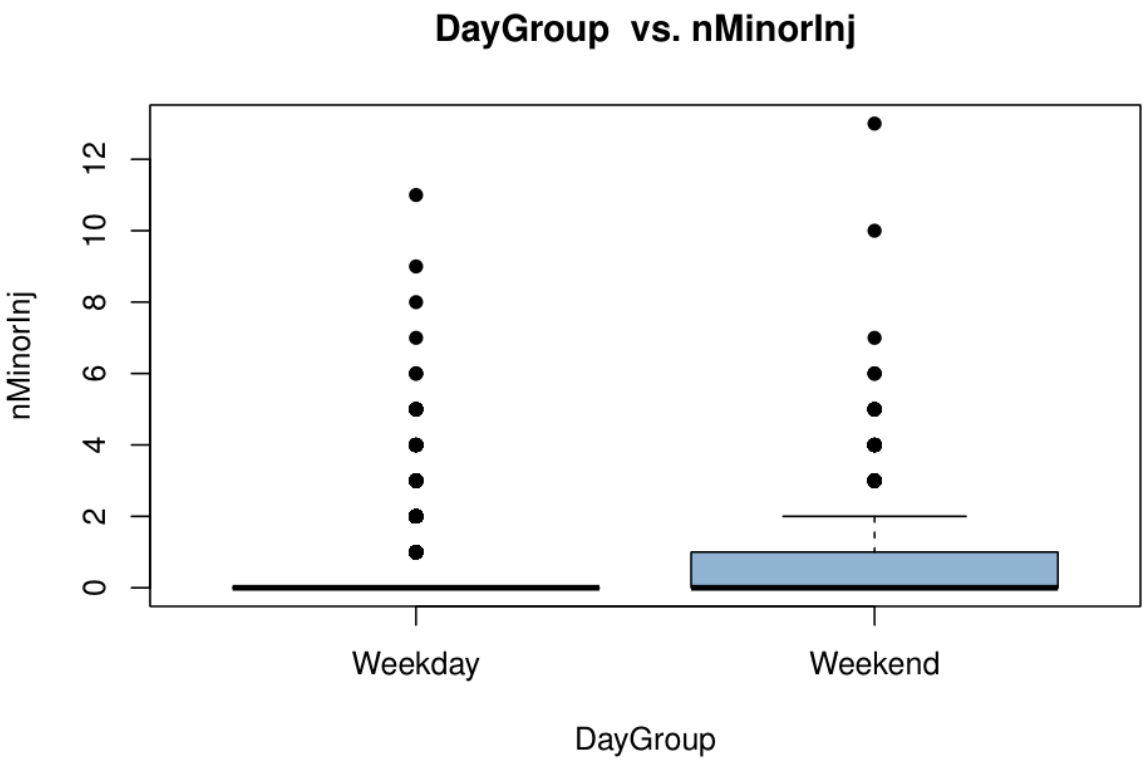

Weather vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
Good
WeakRain
StrongRain
Snow
UnkWth
Weather

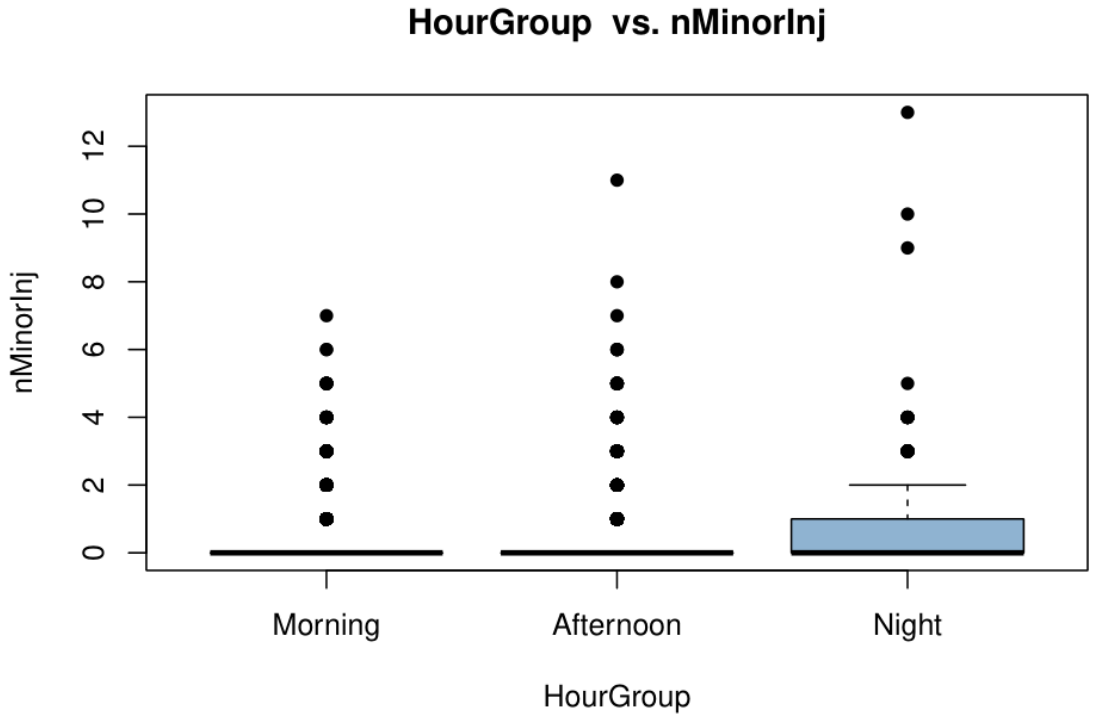

WeatherInf vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
Yes
No
UnkWthInf
WeatherInf

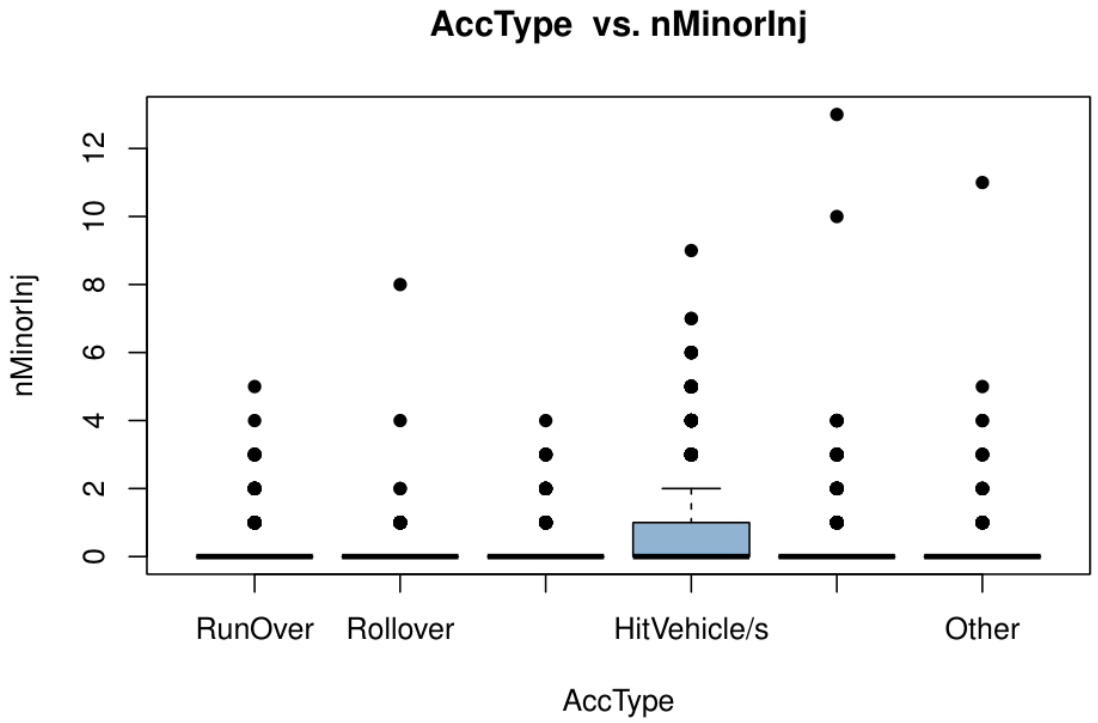

LightInf vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
Yes
No
UnkLightInf
LightInf

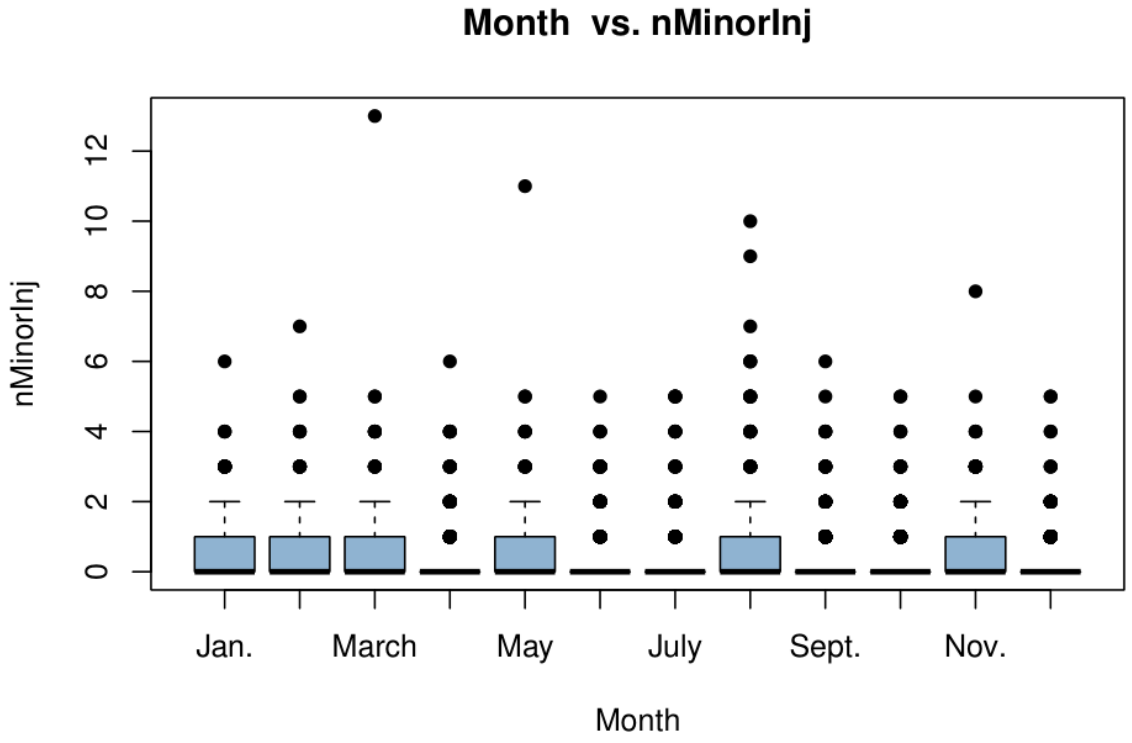

Intersect vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
InSection
Arriving
Inside
Intersect


Surface vs. nMortal
nMortal
0
2
4
6
8
10
12
Dry&Clean
Wet
Flooded
Icy
Snowy
UnkSrfc
Surface


DayGroup vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
Weekday
Weekend
DayGroup


HourGroup vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
Morning
Afternoon
Night
HourGroup


AccType vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
RunOver
Rollover
HitVehicle/s
Other
AccType


Month vs. nMinorInj
nMinorInj
0
2
4
6
8
10
12
Jan.
March
May
July
Sept.
Nov.
Month

Conclusions

From the descriptive analysis, we have extracted some conclusions about the accidents dataset. Firstly, we could remark that there is a low variability in numerical variables, which can clearly be observed in their respective histograms and boxplots. Next, for categorical variables, although they have shown a bit more variability, typically one category predominates over the others (often with more than a 50%, and some are even close to 100%). Examples of this last statement could be all the variables about influence (WeatherInf, TrafficInf, LightInf, etc.). Moreover, from dates, we can conclude that the accident density has decreased in the last 10 years (especially in 2020 and 2021, probably as a consequence of covid-19). Finally, if we complement the univariate analysis with the bivariate, we can also induce an increase in the variability of mainly minor-injured persons for example during the weekend, with bad weather, or in inter-urban (road) zones.

Other interesting facts:

- The most common type of accident has been a collision between two vehicles.
- Around 57% of accidents in Catalonia have taken place in zones where the maximum allowed velocity was 100 km/h, but there were more possibilities of mortal accidents when it was 120 km/h.
- Nearly 36% of accidents in Catalonia have happened on weekends, and only 12% of all the accidents were during night hours.
- Almost the 60% of the accidents in Catalonia in the last 10 years have occurred in the province of Barcelona
- There have been slightly more accidents during the Summer months.

7. PCA analysis

The principal component analysis is a method of dimensional reduction that, as the name says, is used to reduce the number of variables used in a data mining process. PCA reduces the number of variables, but preserves as much information provided by the original variables as possible.

We have applied this method to our data set by following the PCA script provided by the professor of the course. We have modified this r-script in order to add some extra graphs that are going to help to do this analysis and also to create more visual and understandable plots. This script can be found in the folder of scripts attached to the submission of this study.

The first thing that we have done has been setting a table with all the rows but only the numerical variables, because PCA can only be applied to numerical variables. Once we had that, and we have verified that all variables were well declared, we have computed the PCA with the prcomp() r function.

Scree plot

Once we had the pca computed, in order to decide the number of dimensions to use for our study, we have represented the percentage of variance explained by each principal component by using a scree plot.

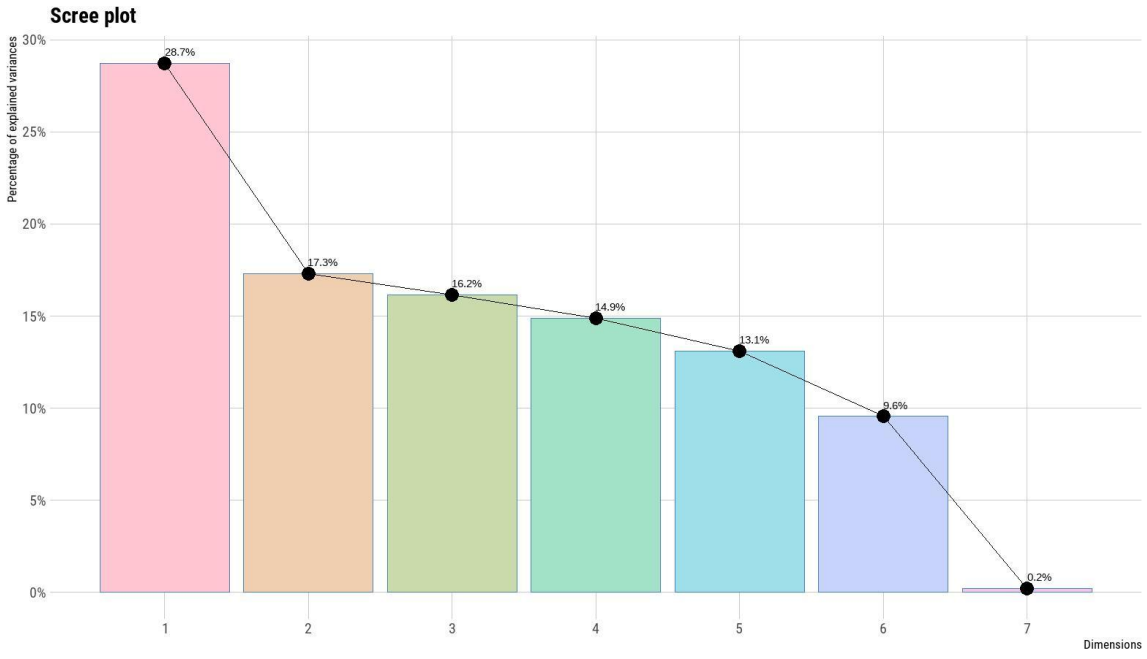


Image 7.1. Scree plot that represents the percentage of variance that each dmension explains.

In the scree plot represented above, we can observe that the first dimension is the one with higher percentage of variance explained and that there is a 'jump' between this dimension and the following ones. Also, we can observe that dimension 7 is not representative as it only explains 0.2% of the variance. At first sight we can conclude that, on one hand, dimension 1 has to be used in the study as it holds an important percentage of the inertia and, on the other hand, dimensions 6 and 7 can be discarded.

To support this decision and to decide the number of dimensions to use in our study, we have represented a cumulated scree plot, which represents the accumulation of inertia among the subspaces. In order to decide how many subspaces to keep, we have to take into account that 80% of the total inertia has to be kept but also the number of dimensions selected has to be minimum.

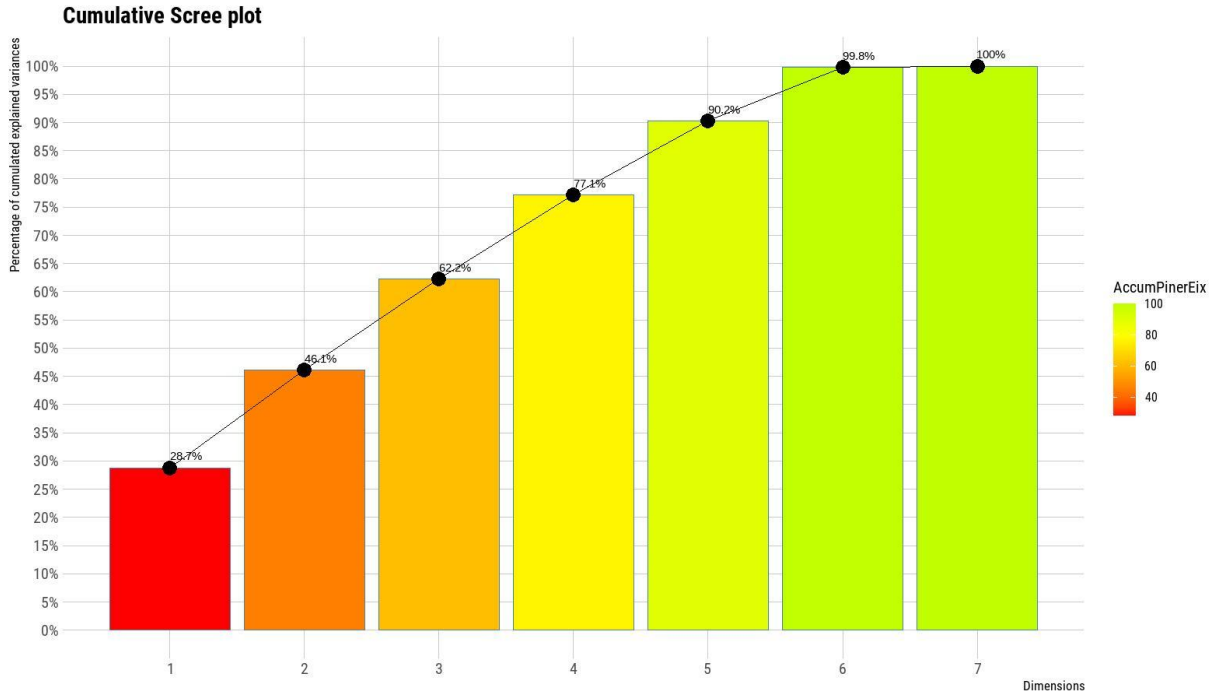


Image 7.2. Cumulative scree plot that represents the accumulation of variance explained along the dimensions.

In our case, we have decided to use 4 dimensions, as they hold 77% of total inertia and that it is enough.

Quality variables plot

Before analyzing in depth the different factorial maps, we have considered it important to make a simple corrplot, that is, a matrix graph, in which we display the quality of each numerical variable for each PCA.

By quality, we mean the square cosine of the correlation of a variable to a PCA. This way, with values from 0 to 1 we can evaluate how much a numerical variable is correlated with a PCA, with a dimension.

Let's see the corrplot:

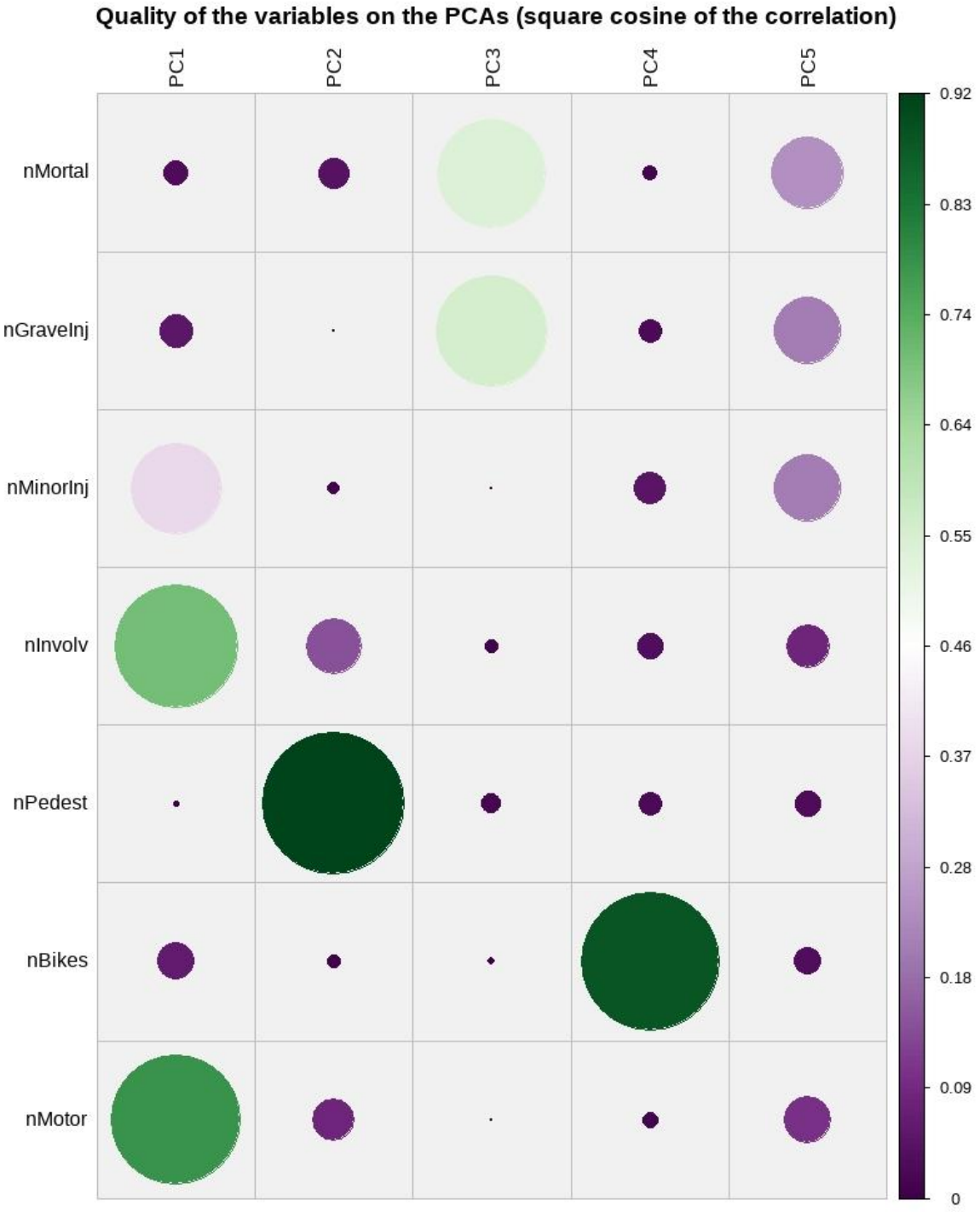


Image 7.3. Corrplot: Quality of the variables on the PCAs.

One thing to note is that, because the dimensions are perpendicular to each other, the sum of all qualities of a numerical variable (all PCA qualities of a variable), is always 1.

In the previous section we've said we were gonna work only with the first 4 dimensions. In this corrplot we've also included the fifth dimension to reinforce this idea. As we can see, although with only 4 dimensions we cannot reach an 80% of the total inertia, including the fifth dimension wouldn't do much, because the PCA5 doesn't represent any variable with a good enough quality.

On the other hand, we can observe how each of the first 4 dimensions has a high correlation with at least 1 numeric variable.

Before analyzing the factorial maps, it's important to remark that in this PCA analysis we are not gonna analyze all the different projections of categorical variables for each combination of dimensions. Due to the large number of different variables we have, we're only going to study those representations that we consider most interesting.

All the different graphs for each combination of the first 4 dimensions are produced by our R script, and are saved in a subfolder on our deliverable, though.

FACTORIAL MAPS

Plot of individuals

For each possible combination between dimensions, we have created the Individuals Plot. Each plot has projected the 5000 individuals of our data set, but in all of our six possible combinations of dimensions, some of the individuals are overlapped. So, in order to observe the real individual concentration in one point of the plot, we have used a system of point transparency. If the point is darker, it means that more individuals are projected in the same point (which means that those individuals are similar in terms of the variables explained by that pca combination).

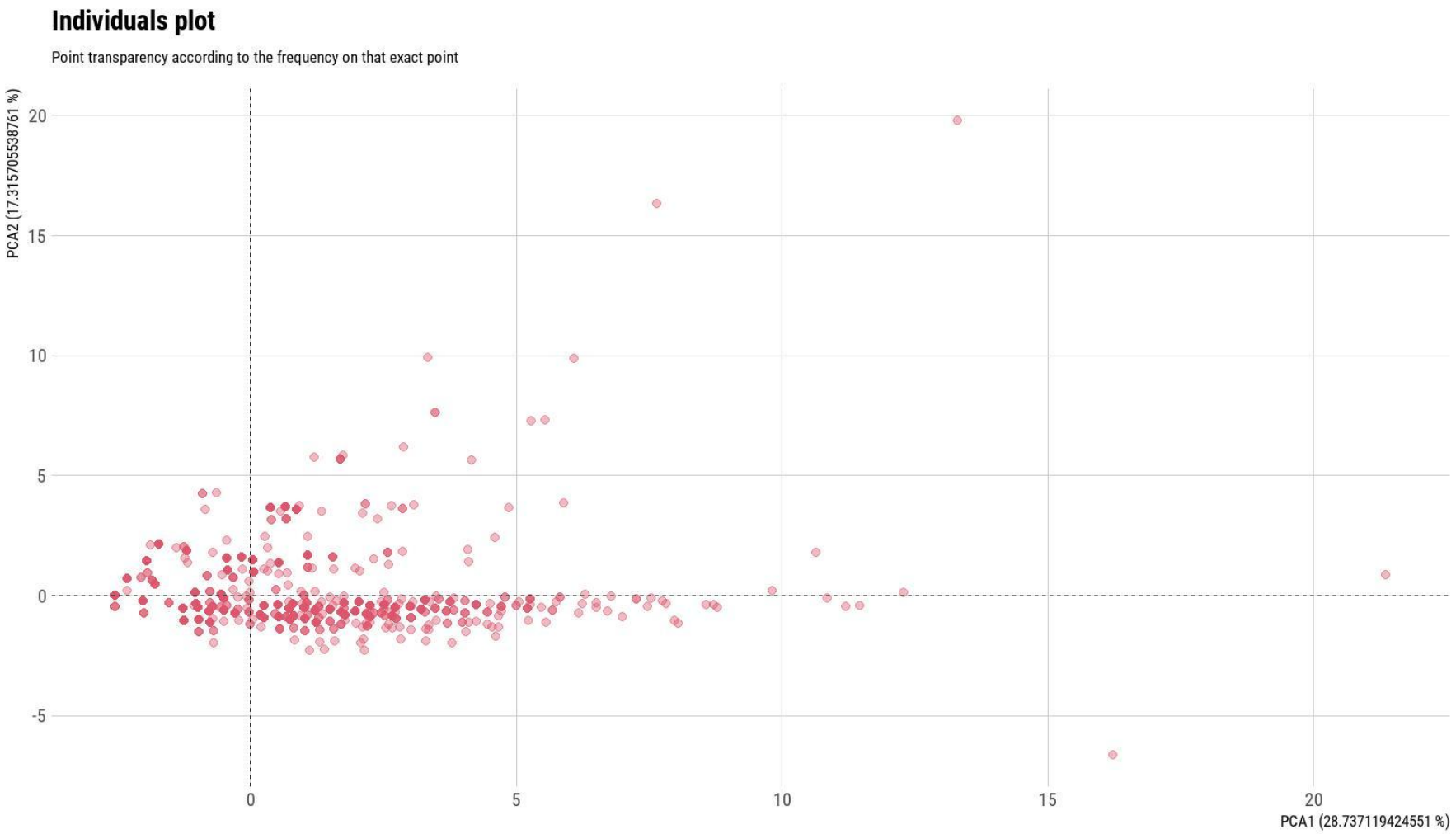


Image 7.4. Plot with the projection of the individuals on the factorial map with PCA1 and PCA2 on the axis.

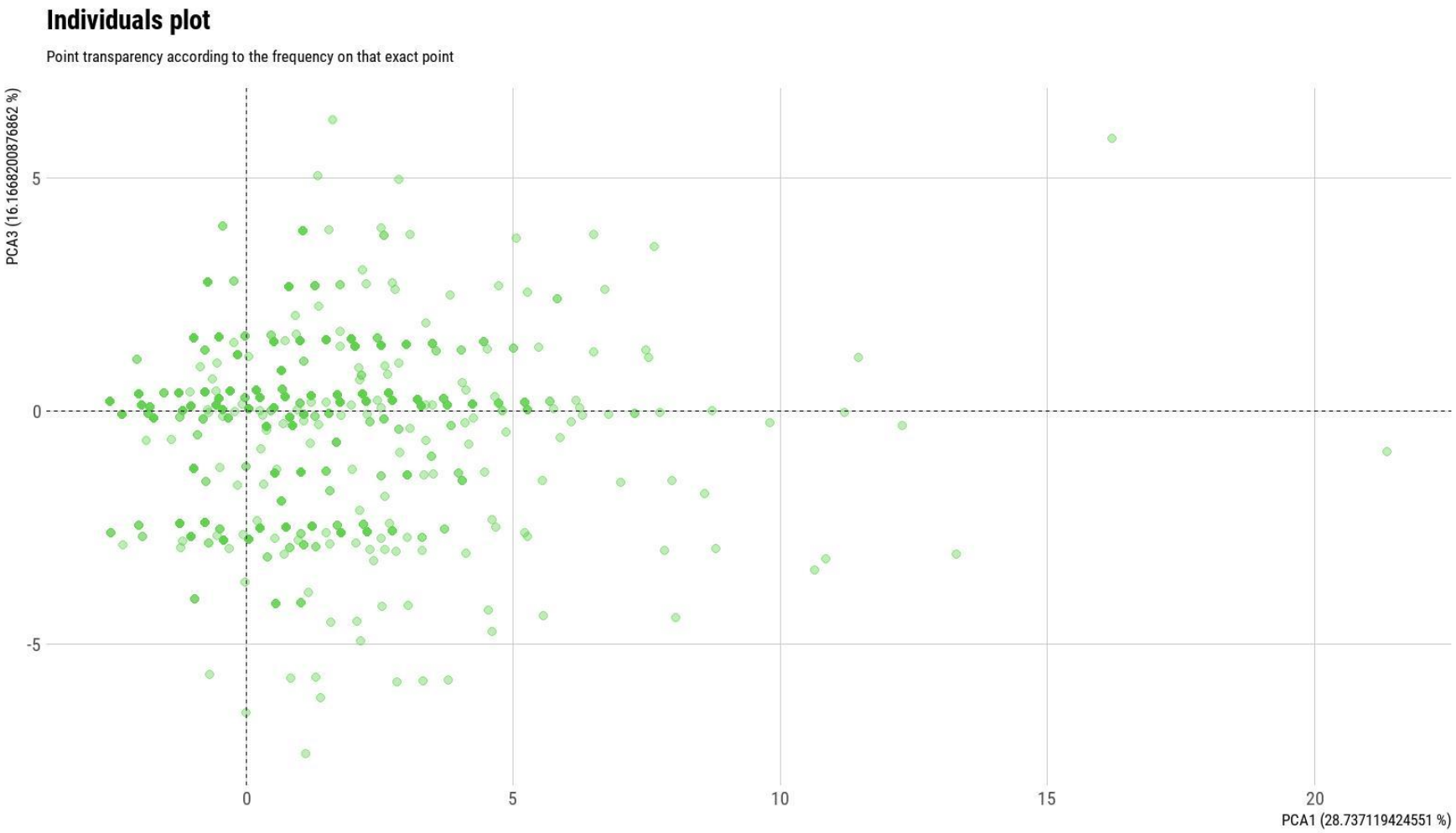


Image 7.5. Plot with the projection of the individuals on the factorial map with PCA1 and PCA3 on the axis.

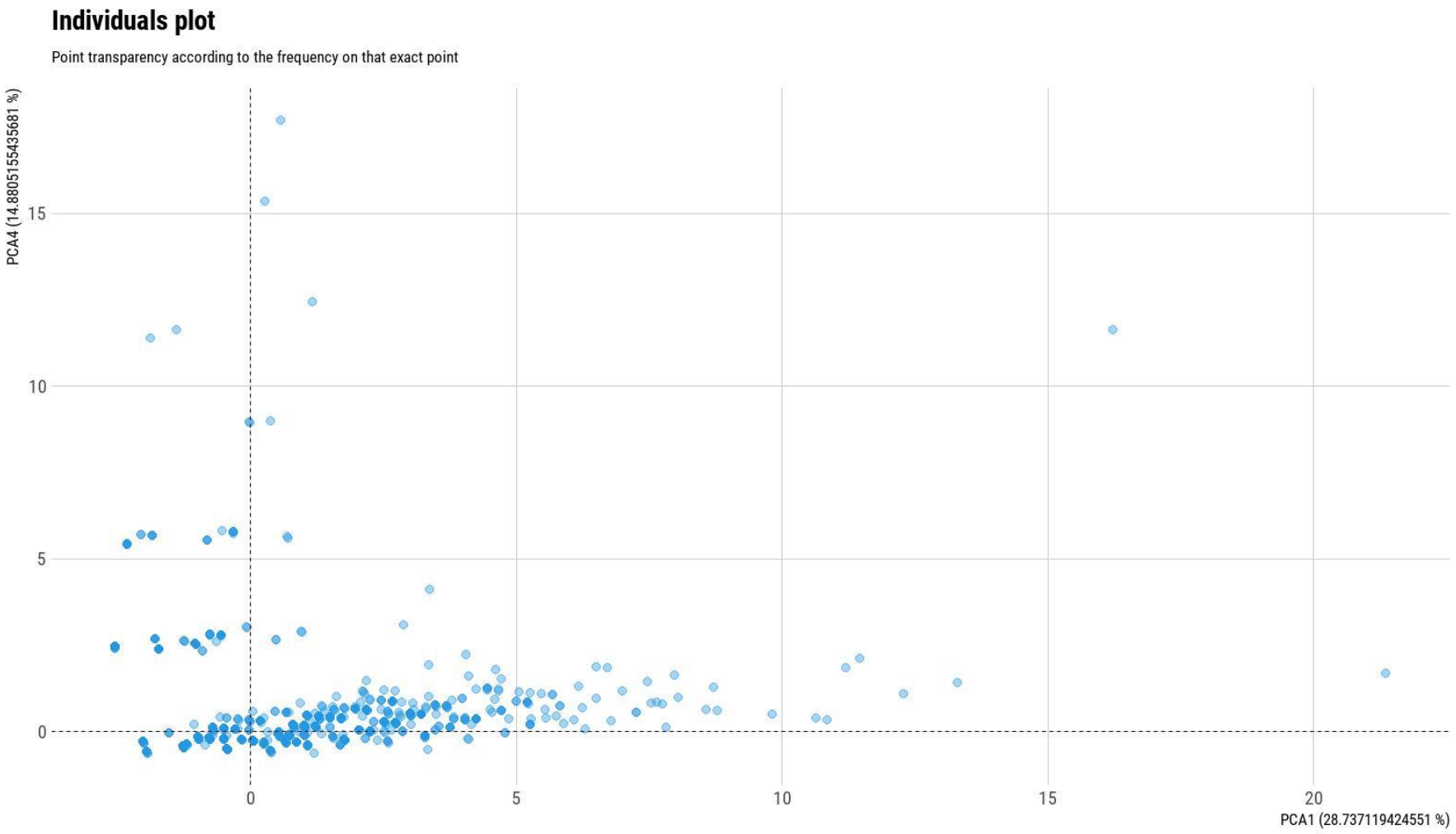


Image 7.6. Plot with the projection of the individuals on the factorial map with PCA1 and PCA4 on the axis.

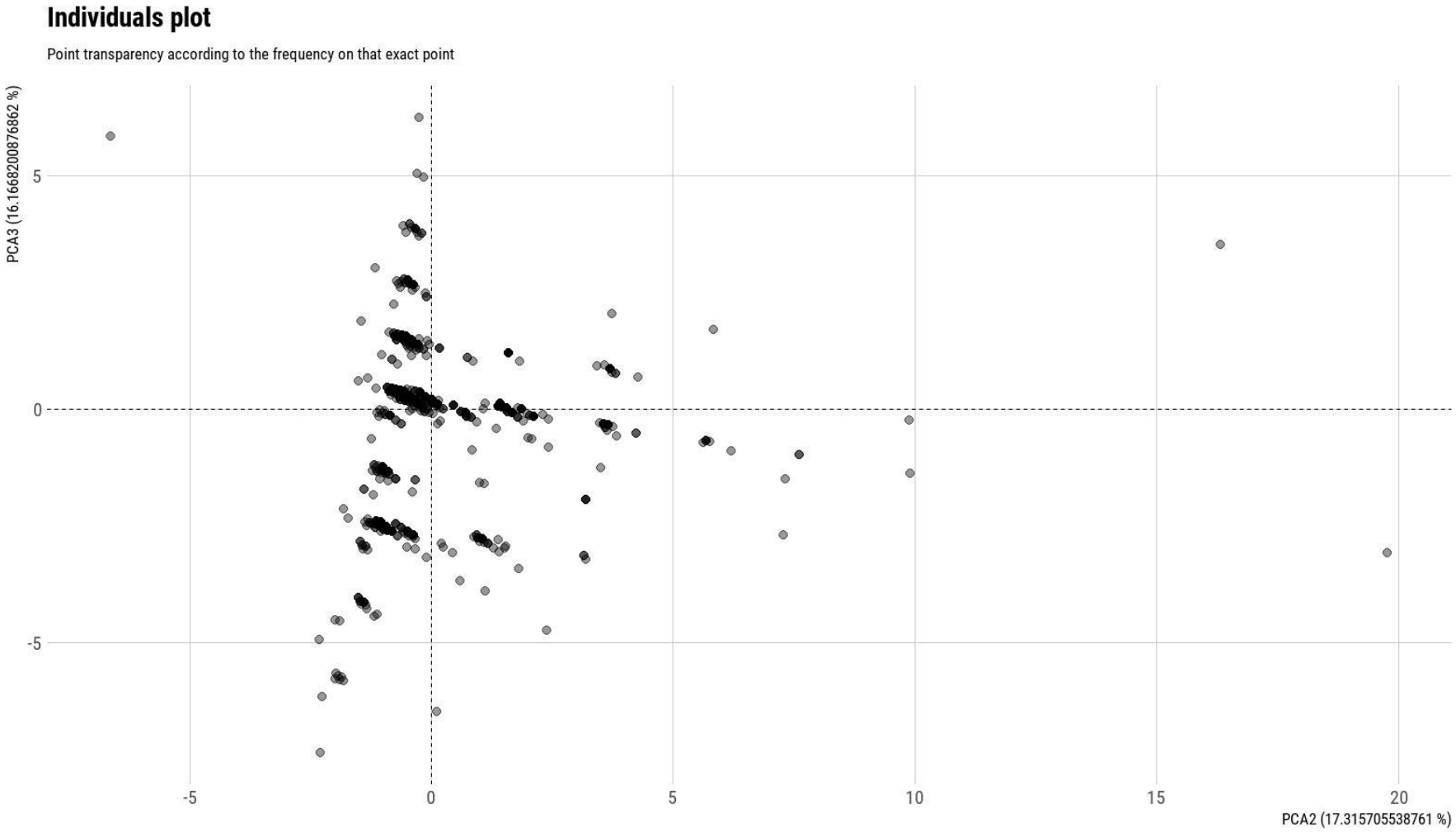


Image 7.7. Plot with the projection of the individuals on the factorial map with PCA2 and PCA3 on the axis.

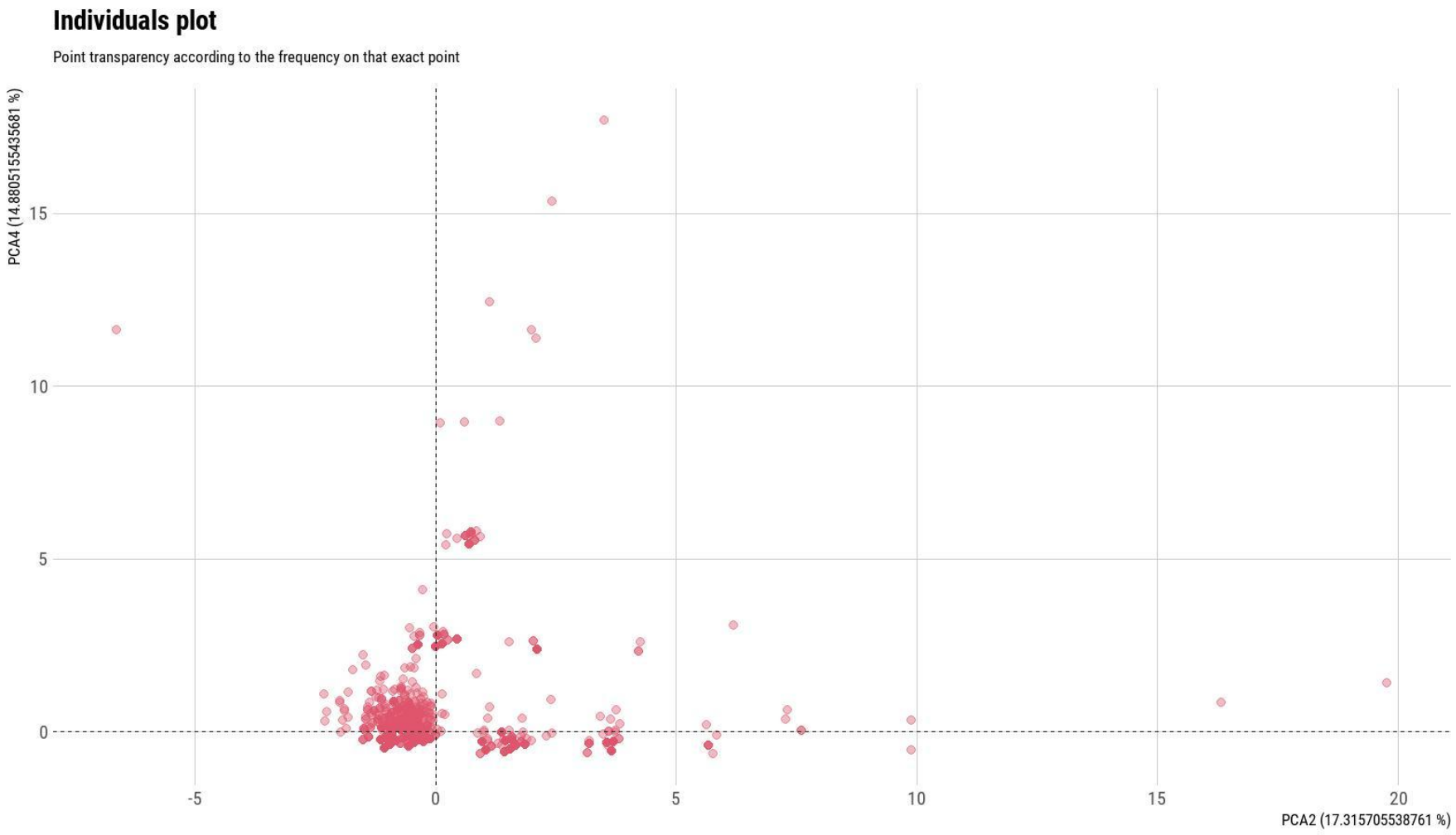


Image 7.8. Plot with the projection of the individuals on the factorial map with PCA2 and PCA4 on the axis.

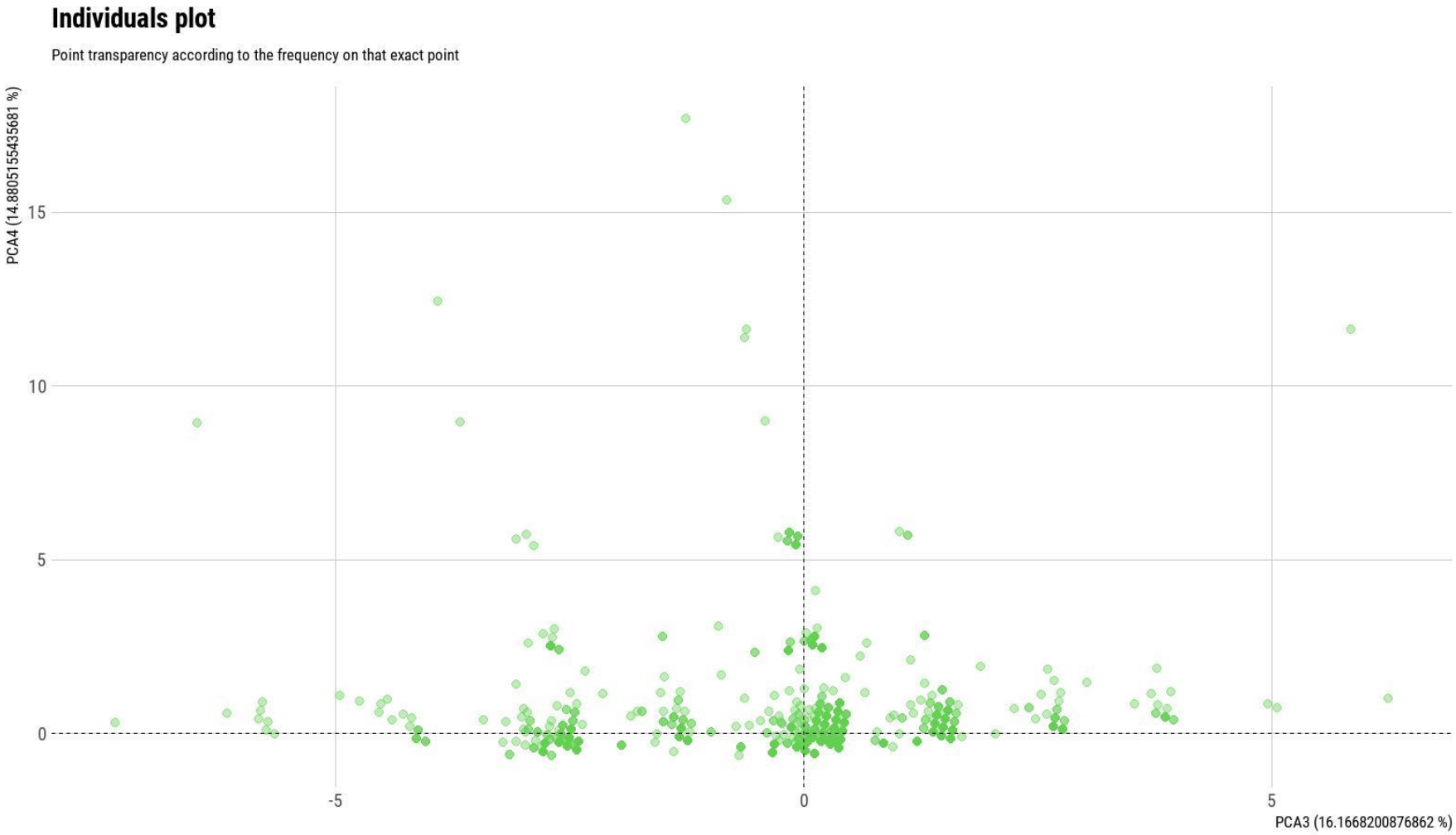


Image 7.9. Plot with the projection of the individuals on the factorial map with PCA3 and PCA4 on the axis.

After analyzing all these plots, we can observe that PC1 has always a less centralized distribution on its axis than the other principal components, that is due to the percentage of variance this component explains.

Also, we can see that the plot of individuals for the components PC1 and PC2 , and PC1 and PC3 are the plots with less centered distribution. That is due to the fact that they are the principal components with more percentage of variance explained.

The plot of individuals of the factorial map with PC2 and PC3 is very similar to the individuals plot on dimensions PC1 and PC3, but has a bit more centralized distribution on the x-axis.

On the other hand, the plots for the factorial maps PC2+PC4 and PC3+PC4 are the ones with less variance. Although there are some individuals far from the origin, the big concentrations of individuals remain centered, mostly on the PC4 axis. It is something that we would expect, as the PC4 component is the one with less variance.

In some of these graphs we can distinguish several groups of individuals, in some more clearly than in others.

After visualizing all the individuals plots, we have decided to study just some of these factorial maps. We have kept those factorial maps with more variance explained and less centralized distribution. Those are: PC1+PC2, PC1+PC3 and PC2+PC3.

We have also decided not to study the PCA4 deeper because there are really not so many samples that are expanded on it except certain outliers. Also, if we see the different correlation circles of the factorial maps that include the PCA4, we can observe that there is no relation between the variable nBikes (which has a strong correlation on the PCA4) and the other variables:

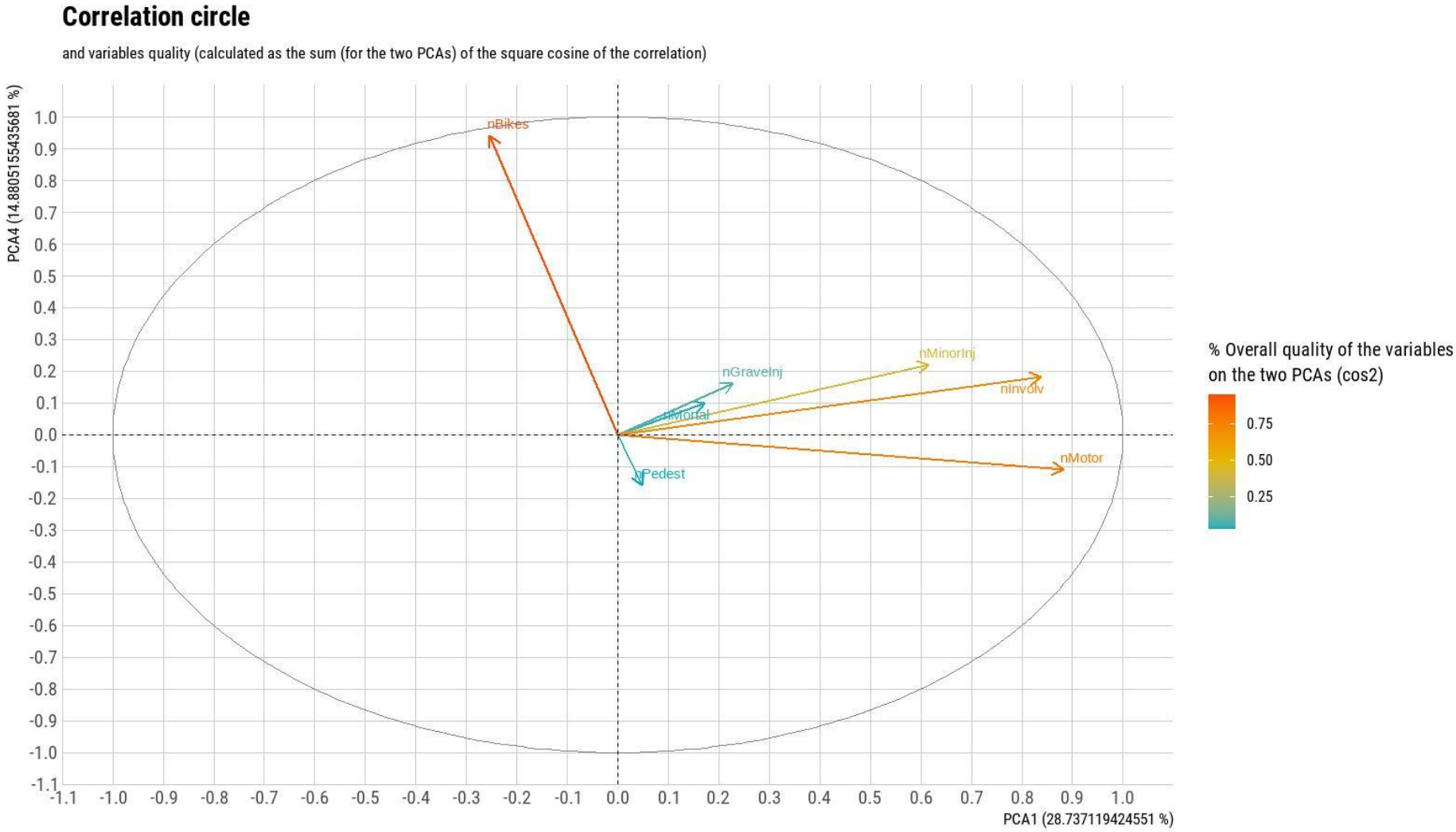


Image 7.10. Correlation circle: PCA1+PCA4.

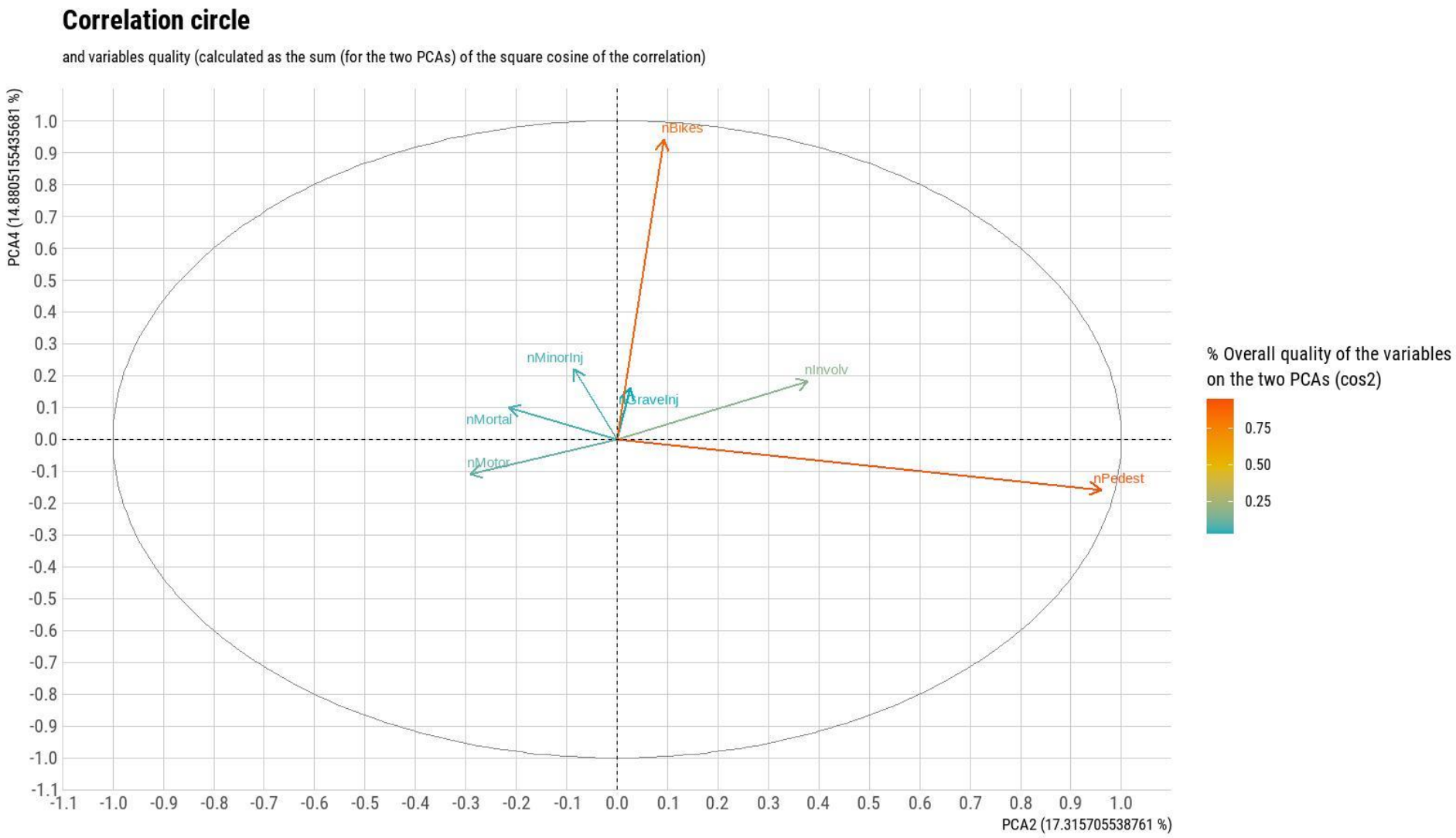


Image 7.11. Correlation circle: PCA2+PCA4.

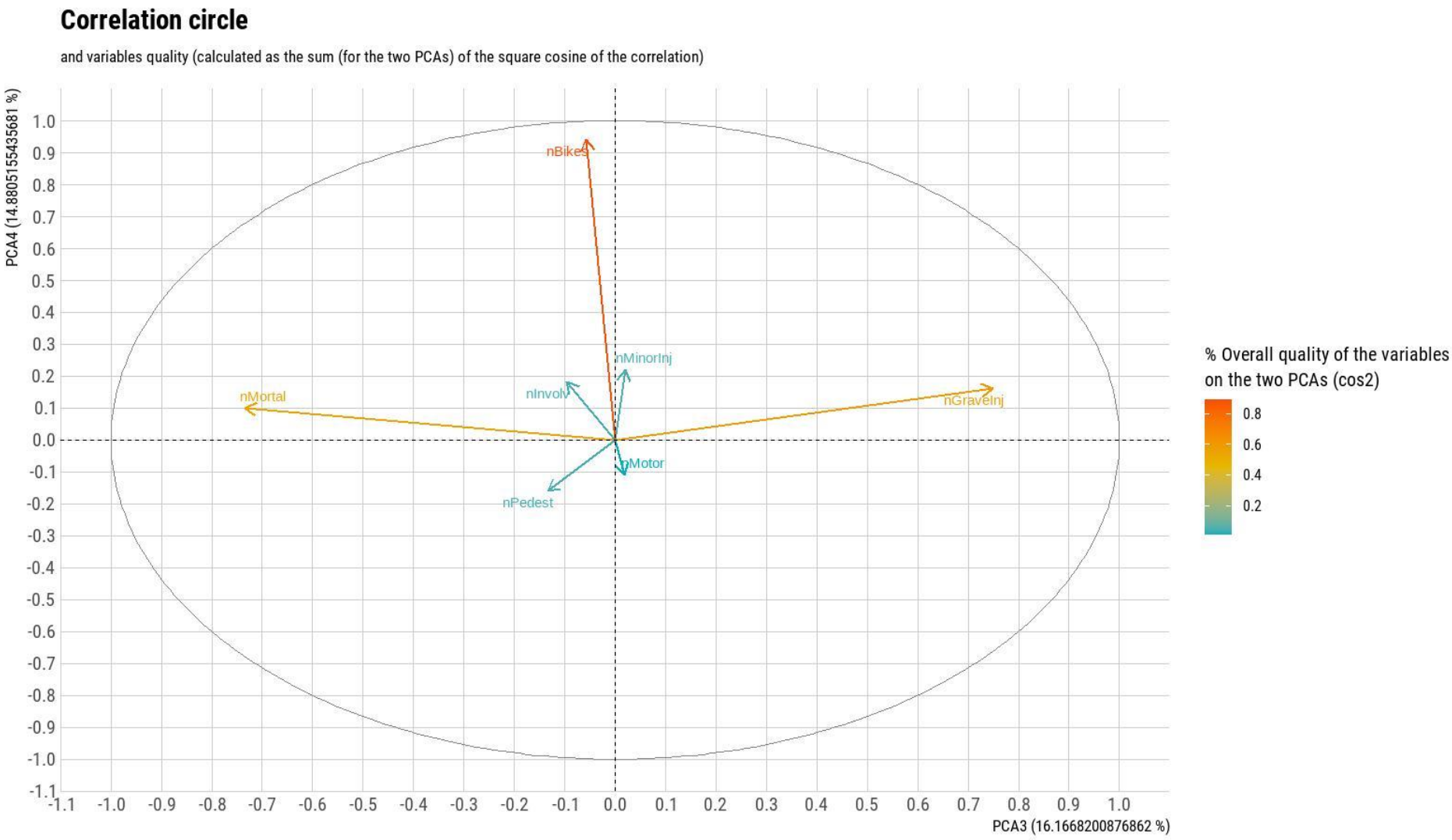


Image 7.12. Correlation circle: PCA3+PCA4.

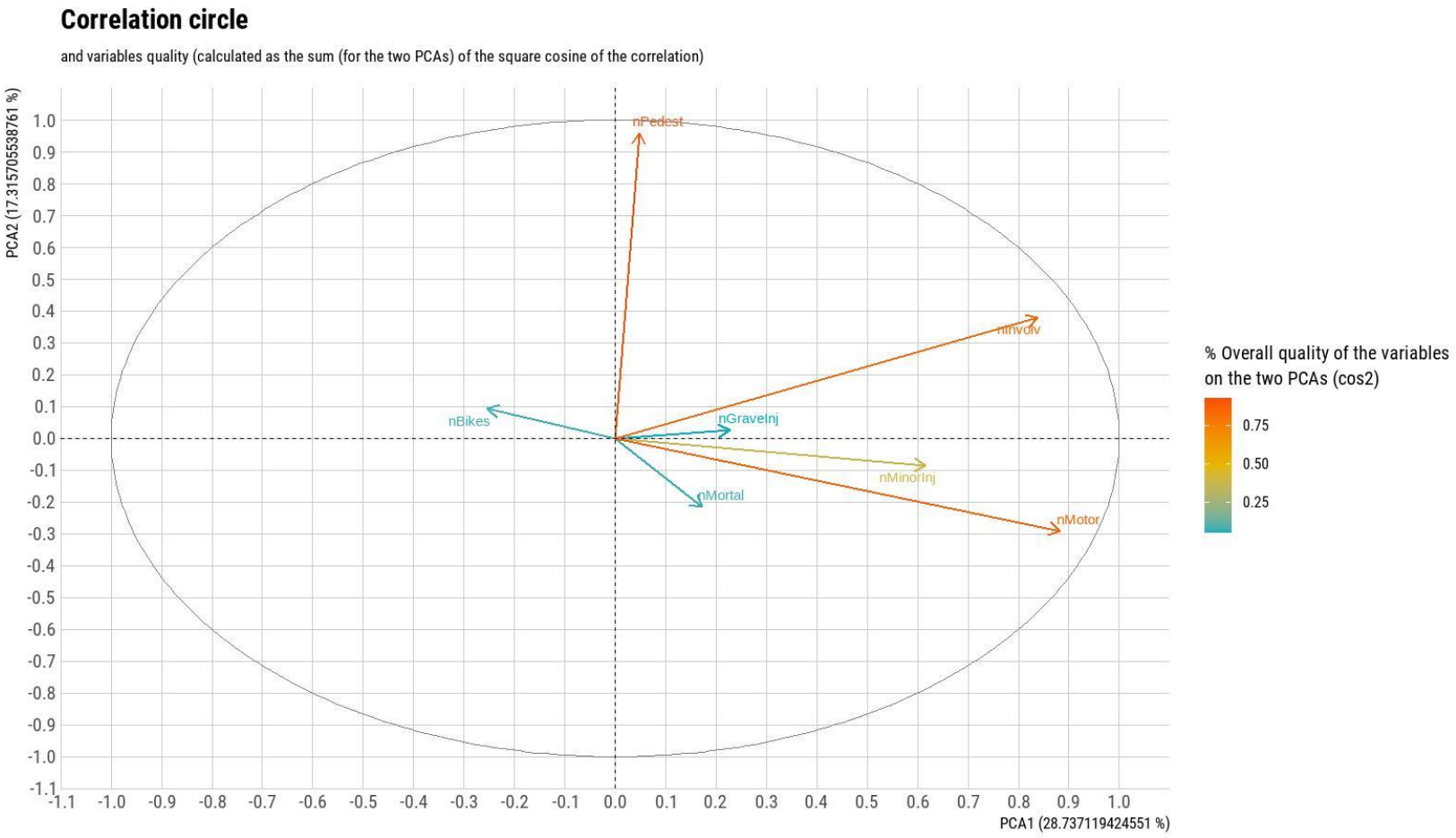


Image 7.13. Correlation circle: PCA1+PCA2.

Diving deeper on the factorial map with the two first dimensions, if we observe the previous image which has the correlation between the numerical variables and these dimensions, the first thing that catches our attention is the big direct correlation between nInvolv, nMotor, and nMinorInj.

This is logical, because as more vehicles with engines are in the accident, more overall units we'll end up having in the accident.

On the other hand, there doesn't seem to be any relation between the number of vehicles with engine and the number of pedestrians in the accident.

We also observe a small inverse correlation between nBikes and nMotor, which makes us think that the majority of accidents that have bikes implicated, it's just a car hitting a bicycle.

Let's now check where the centroids of all the modalities of all our categorical variables are on this factor map. For the sake of easily understanding the data, we have not displayed the modalities of Region on the next graph, because there are a lot. Instead, we show them on a different graph next to the following one:

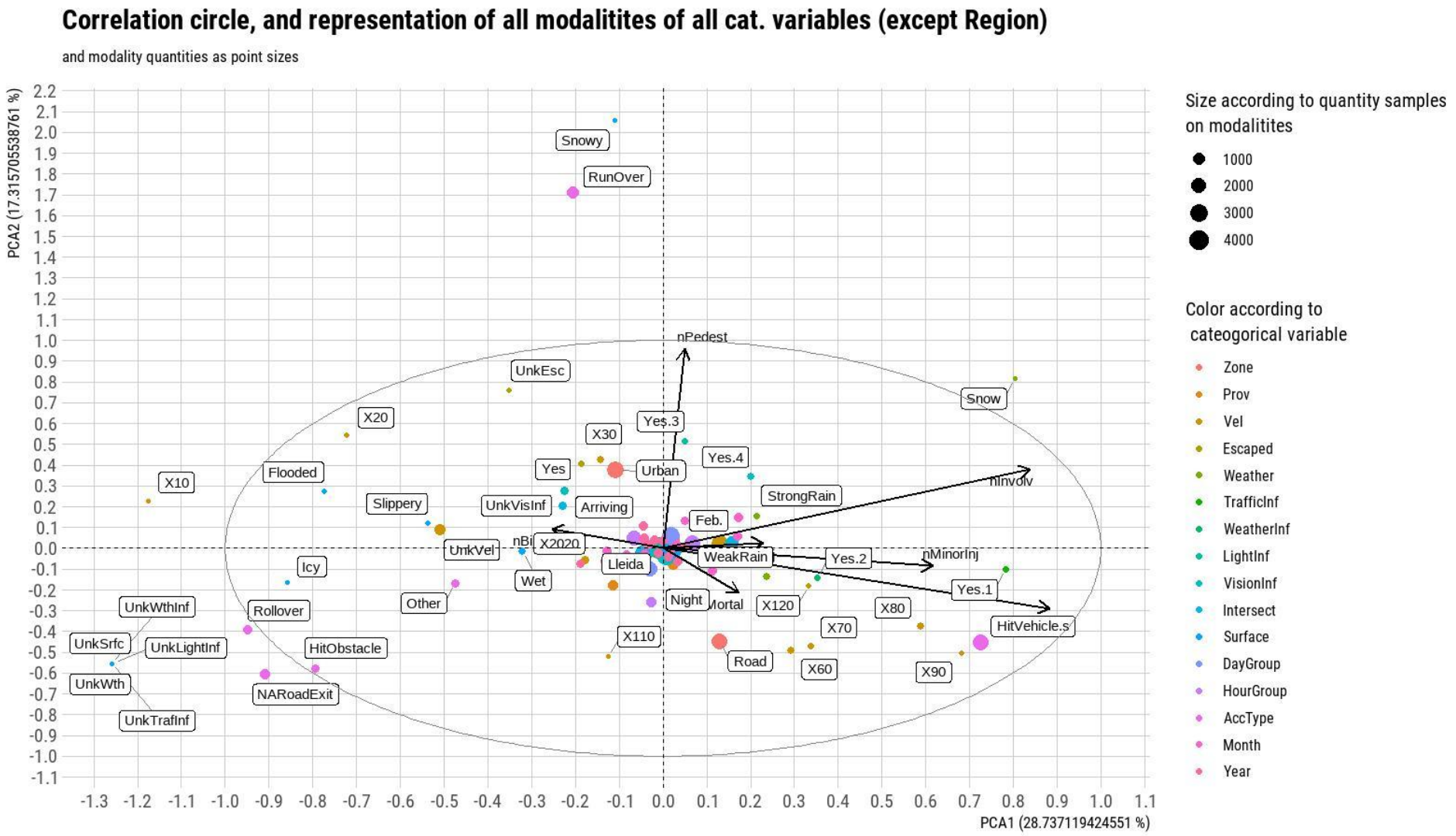


Image 7.14. Correlation circle, and representation of all modalities of all categorical variables (except Region): PCA1+PCA2.

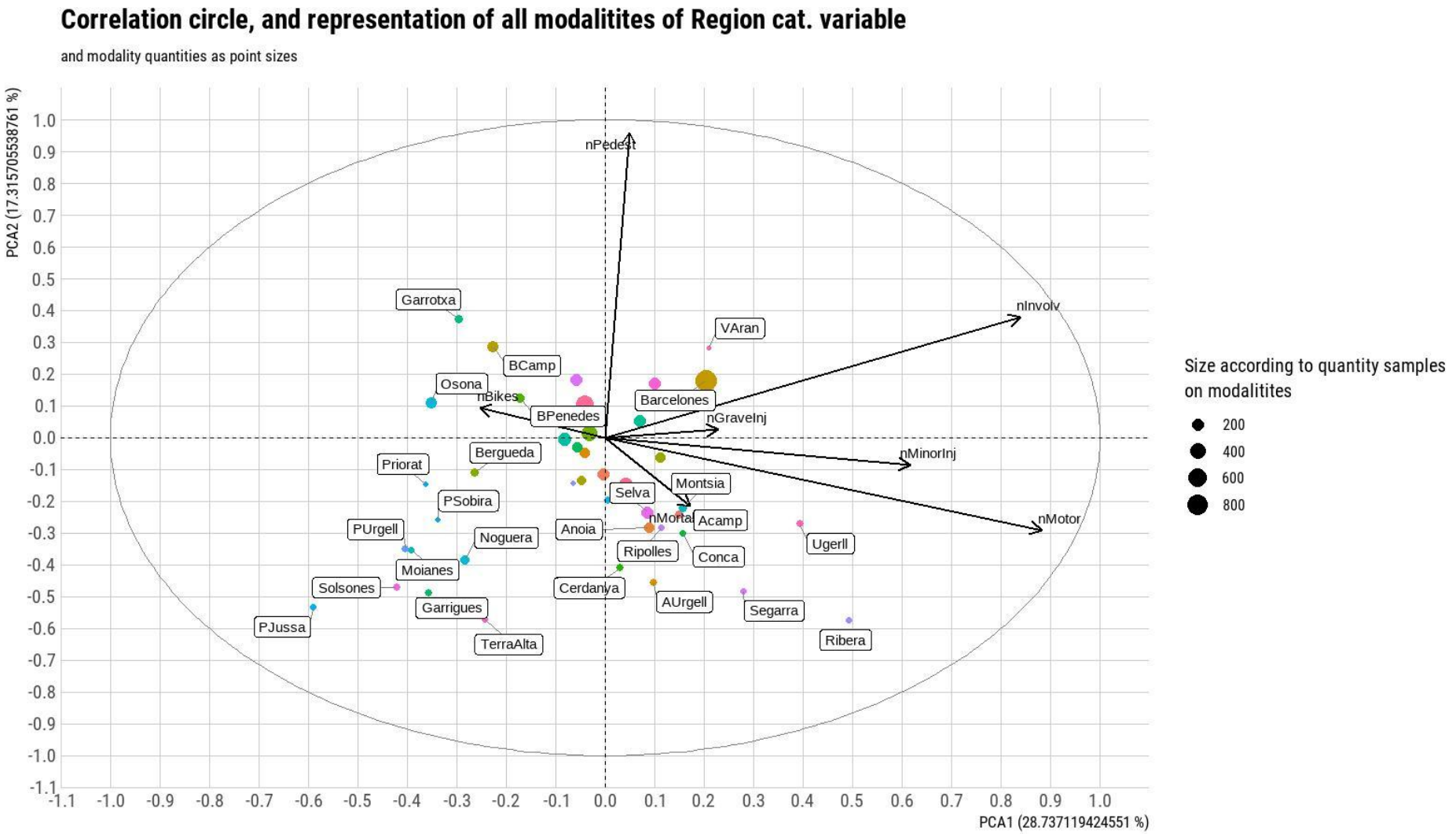


Image 7.15. Correlation circle, and representation of all modalities of Region categorical variable: PCA1+PCA2.

On the first graphic each modality is colored depending on which categorical variable it is from. Also its size will depend on the number of samples it has. On the second graphic the colors are just to better visualize the different modalities of the Region variable.

One important thing that clearly shows the first graphic is that there are a lot of modalities on the center of the factor map, because the majority of our data doesn't have a lot of variability. And, in addition, the modalities which are most separated from the center are the ones that are most uncommon, the ones that have less samples.

Apart from this, we observe different things which are interesting to note:

First off, it's curious how close snowy (surface) is from RunOver. They both are in a notably high position on the y-axis, meaning that they are related with big nPedest. We expected that from RunOver, but not from snowy (surface). Let's watch this better showing only the graphics that contain modalities of surface and AccType:

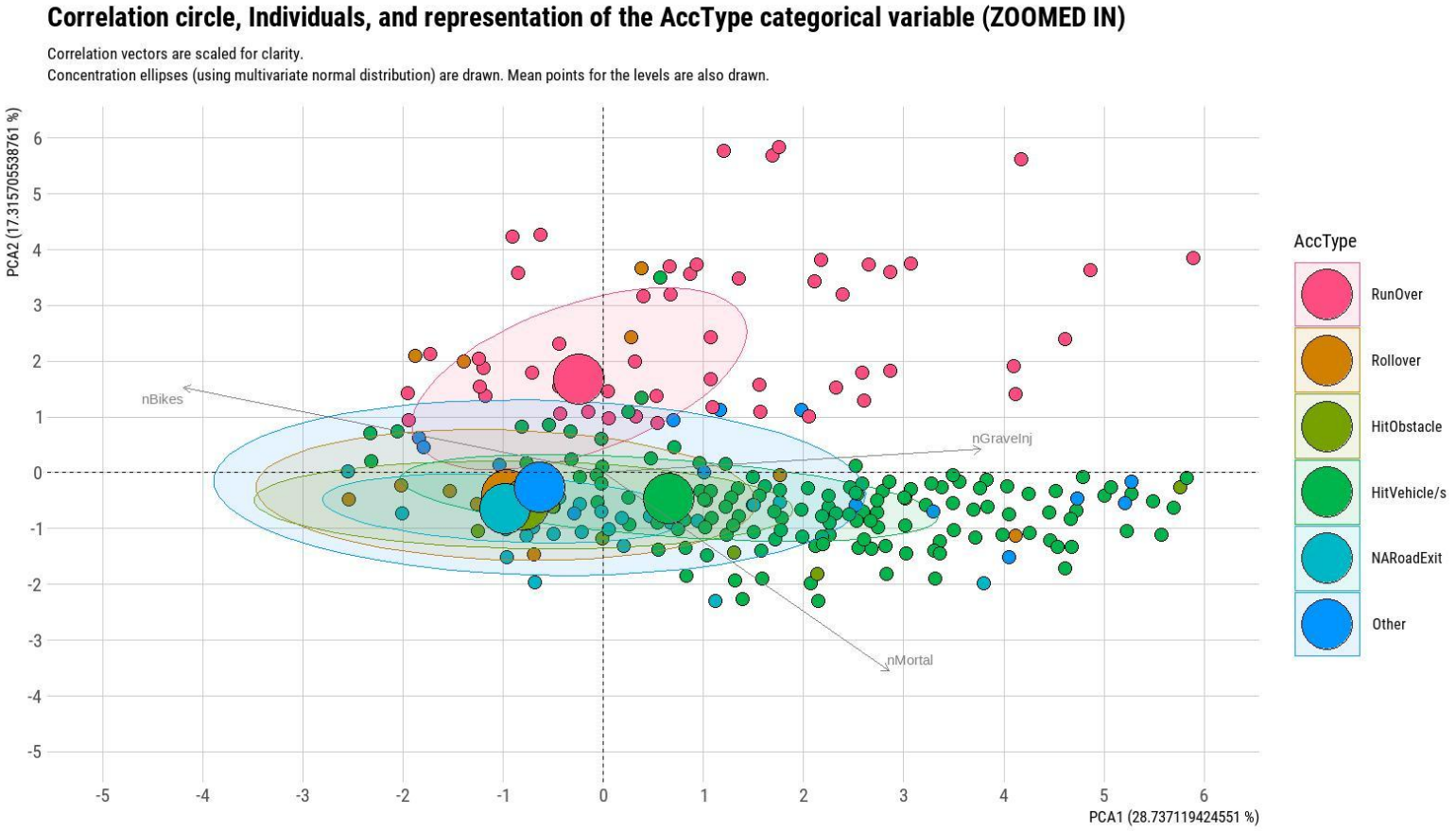


Image 7.16. Correlation circle, individuals, and representation of the AccType categorical variable (ZOOMED IN): PCA1+PCA2.

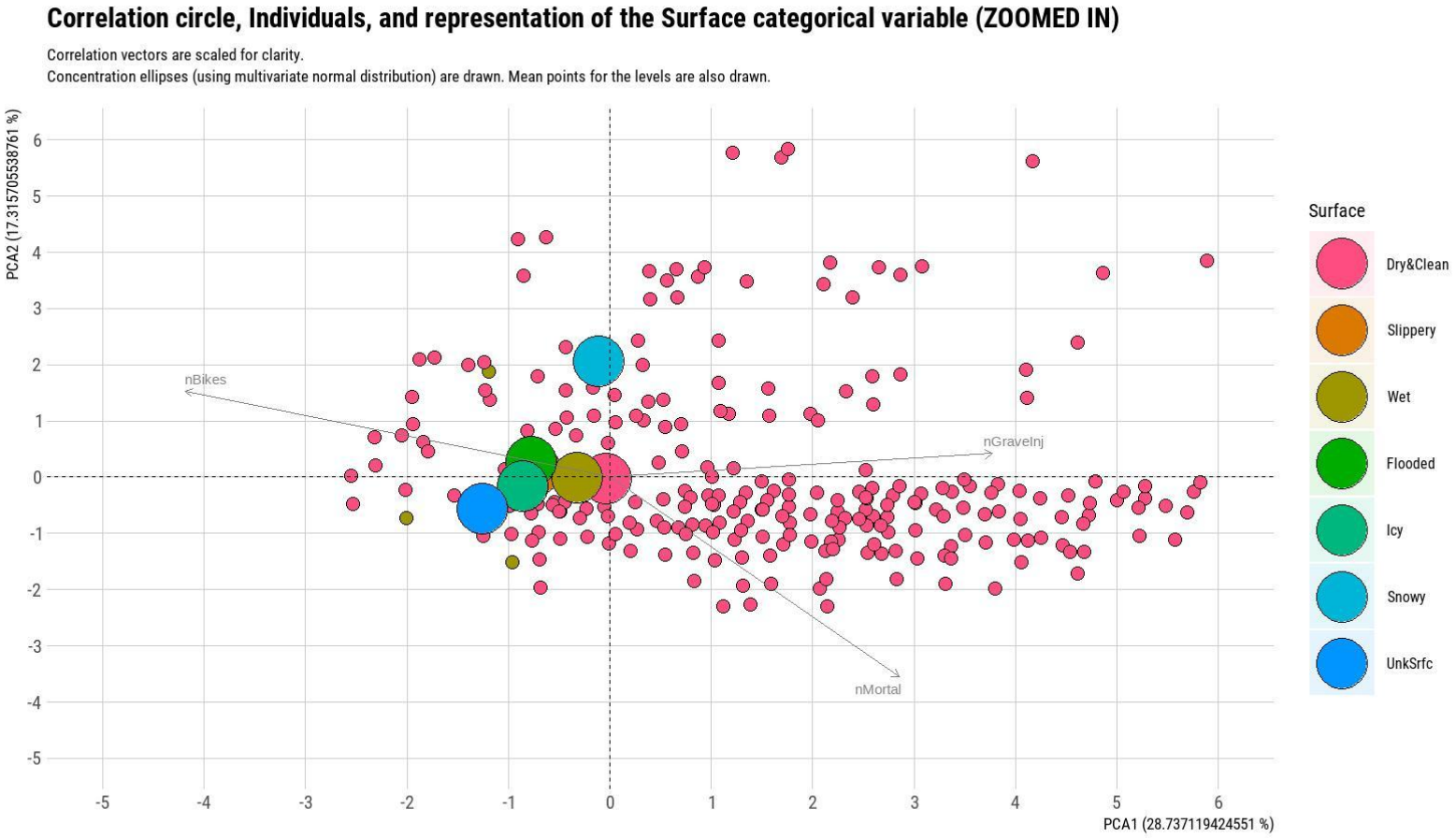


Image 7.17. Correlation circle, individuals, and representation of the Surface categorical variable (ZOOMED IN): PCA1+PCA2.

From this two graphics, we can also see that when the accident type is HitVehicle/s, the accident tends to have more units implicated, more nMortal and more injured people, and that when the accident is of type RollOver, HitObstacle or a Road exit, they are close with a wet, flooded or icy surface, and there are not so many units implicated or injured/dead people.

It is also interesting to see what happens when the Zone is of type Road or Urban:

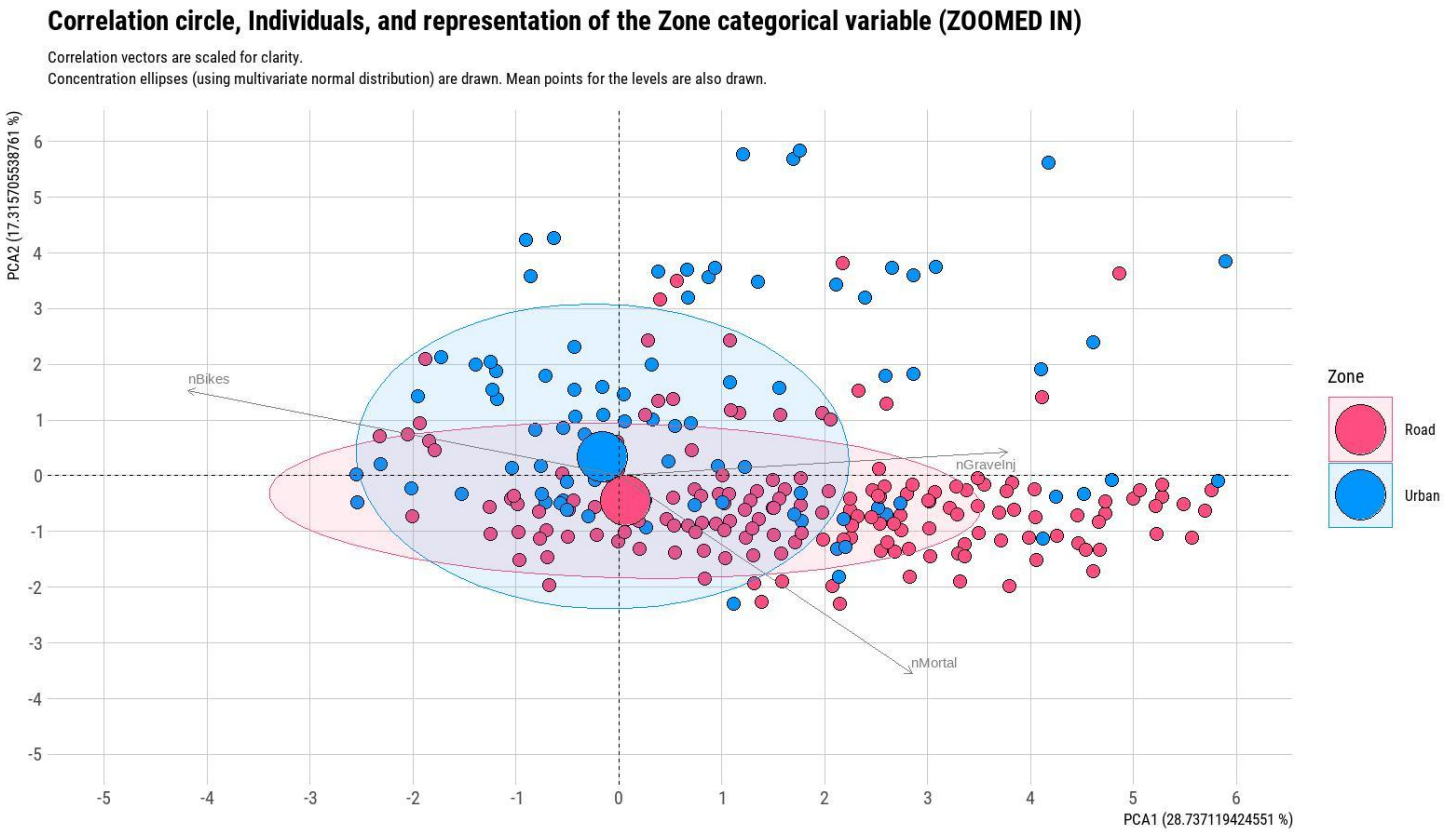


Image 7.18. Correlation circle, individuals, and representation of the Zone categorical variable (ZOOMED IN): PCA1+PCA2.

When the Zone is a road, there tends to be less pedestrians and more mortal people. The opposite happens when the zone is of type Urban.

Finally, going back to the graphic that represents the Region categorical variable, we have to remark that when the accident happens in Barcelona, it tends to be associated with pedestrians and the number of units involved. And when the accident happens in more rural regions, it tends to be associated with a low number (or zero) of pedestrians.

Next, let's cover the study of the factorial map PC1 + PC3:

FACTORIAL MAP WITH PCA1 AND PCA3

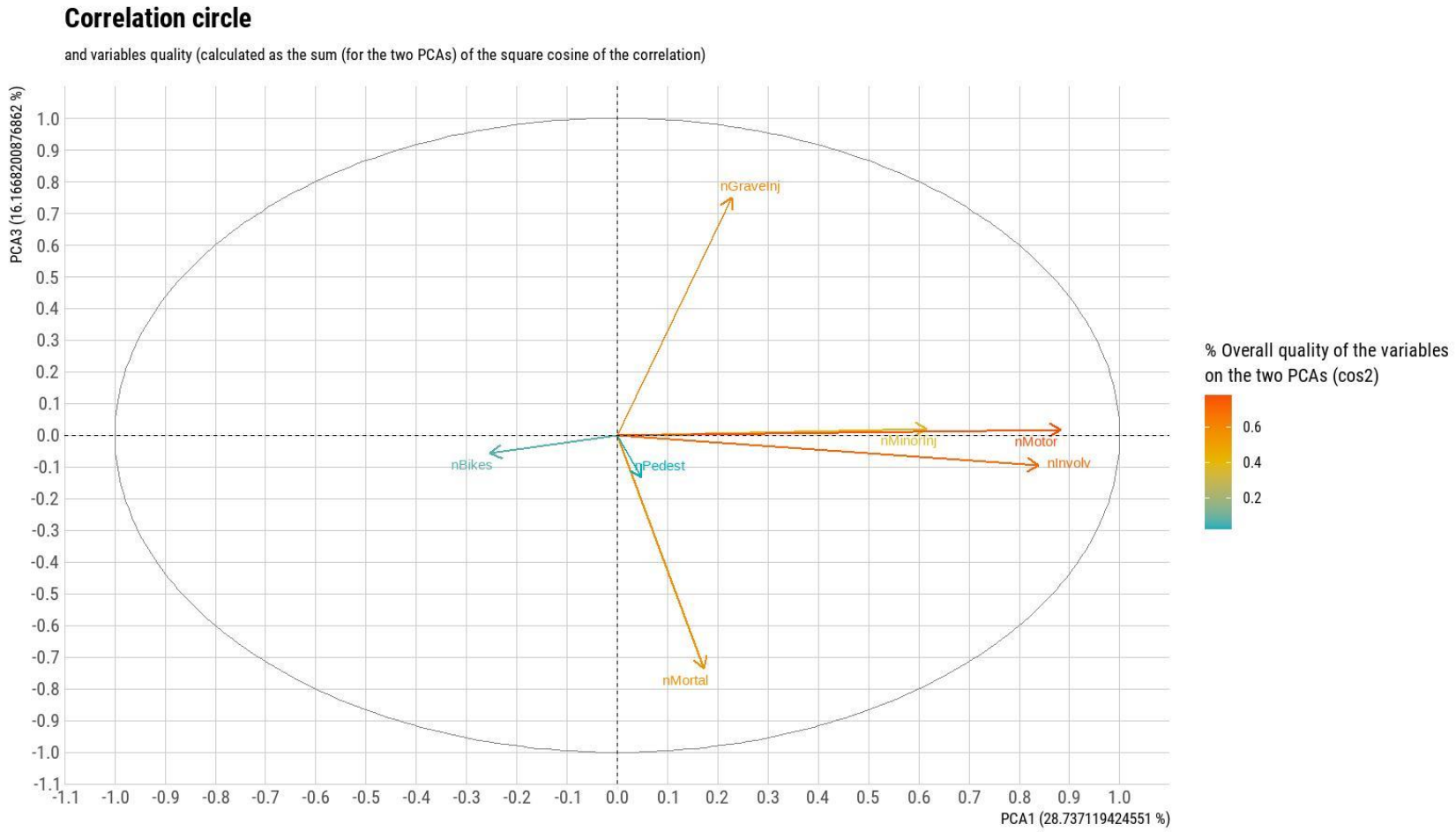


Image 7.19. Correlation circle: PCA1+PCA2

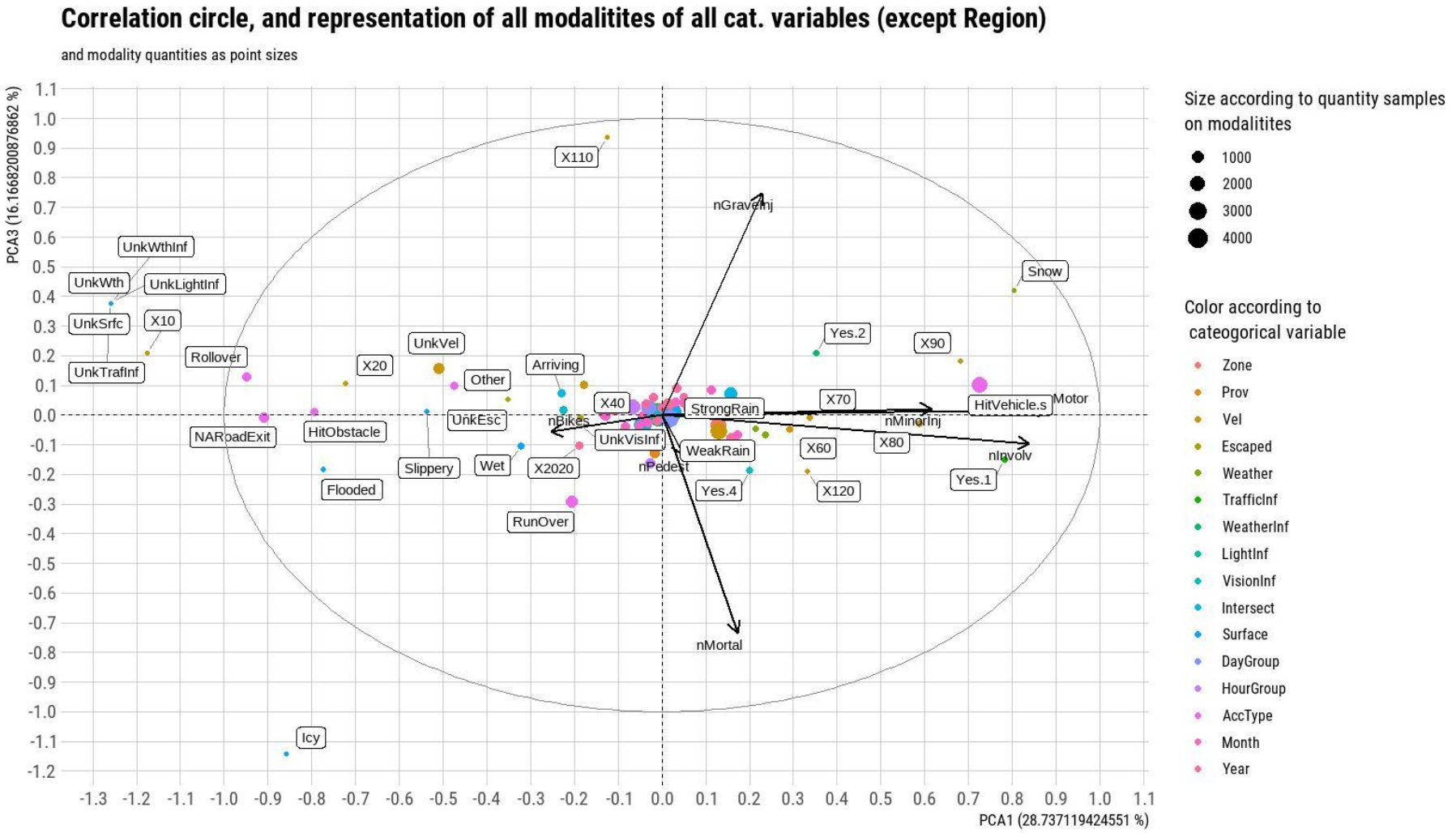


Image 7.20. Correlation circle, and representation of all modalities of all categorical variables (except Region): PCA1+PCA3.

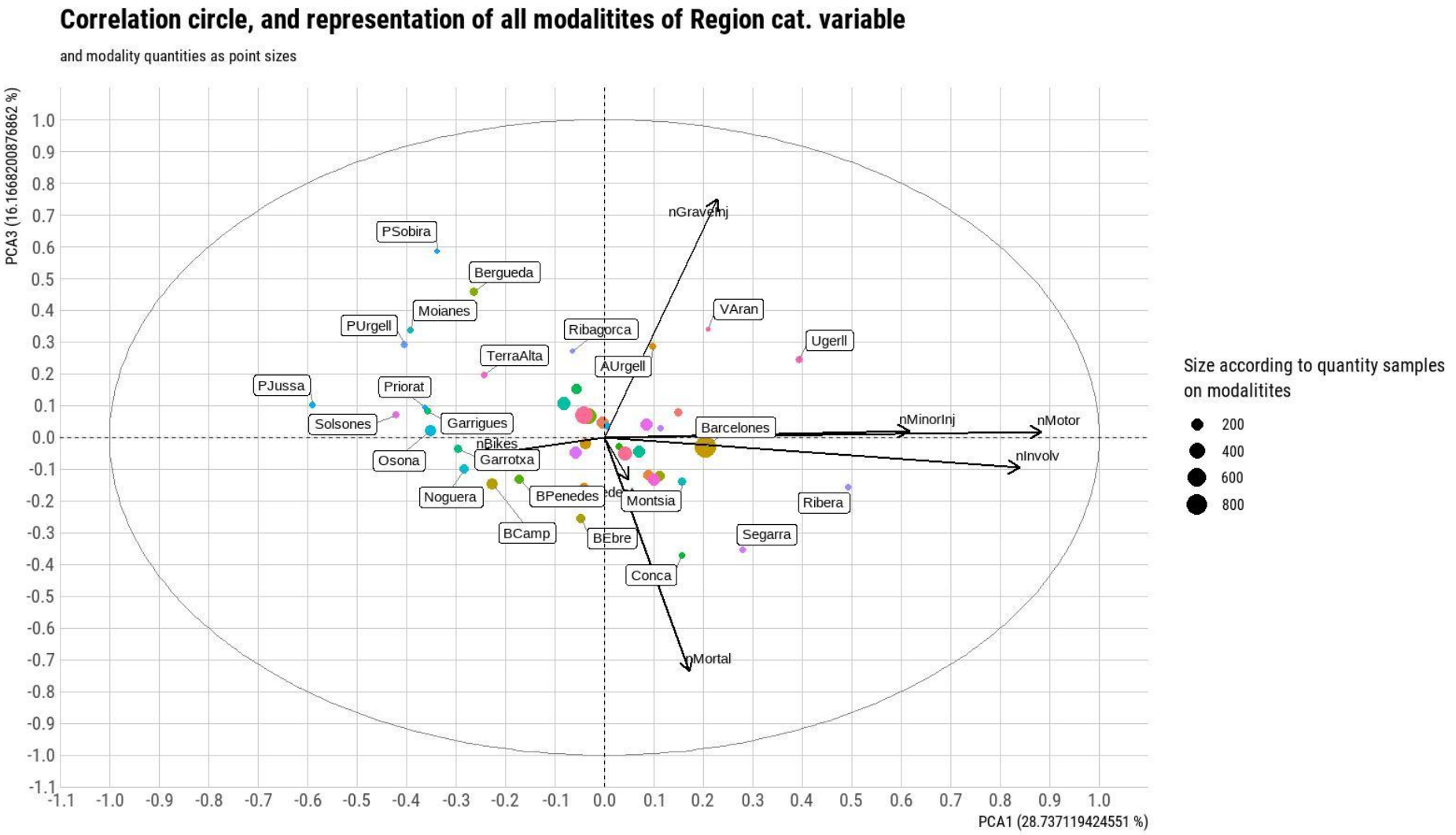


Image 7.21. Correlation circle, and representation of all modalities of Region categorical variable: PCA1+PCA3.

Analyzing the Correlation circle for this factorial map, it doesn't give us much more useful information than the one we extracted from the previous factorial map.

The strong direct correlation between units involved, vehicles with motor and number of people injured is still present and also the small inverse correlation between the number of units involved and the number of bicycles.

But now, PCA3 tells us there seems to be an inverse relationship between the number of serious injuries and the number of mortal people, which we consider strange, because an accident which causes mortal people should also cause seriously injured people.

Apart from this, we have to remark that when the surface is icy, the modality seems to have a strong relationship with mortal people.

Another interesting conclusion we can make from this factorial map is by seeing the next graphic:

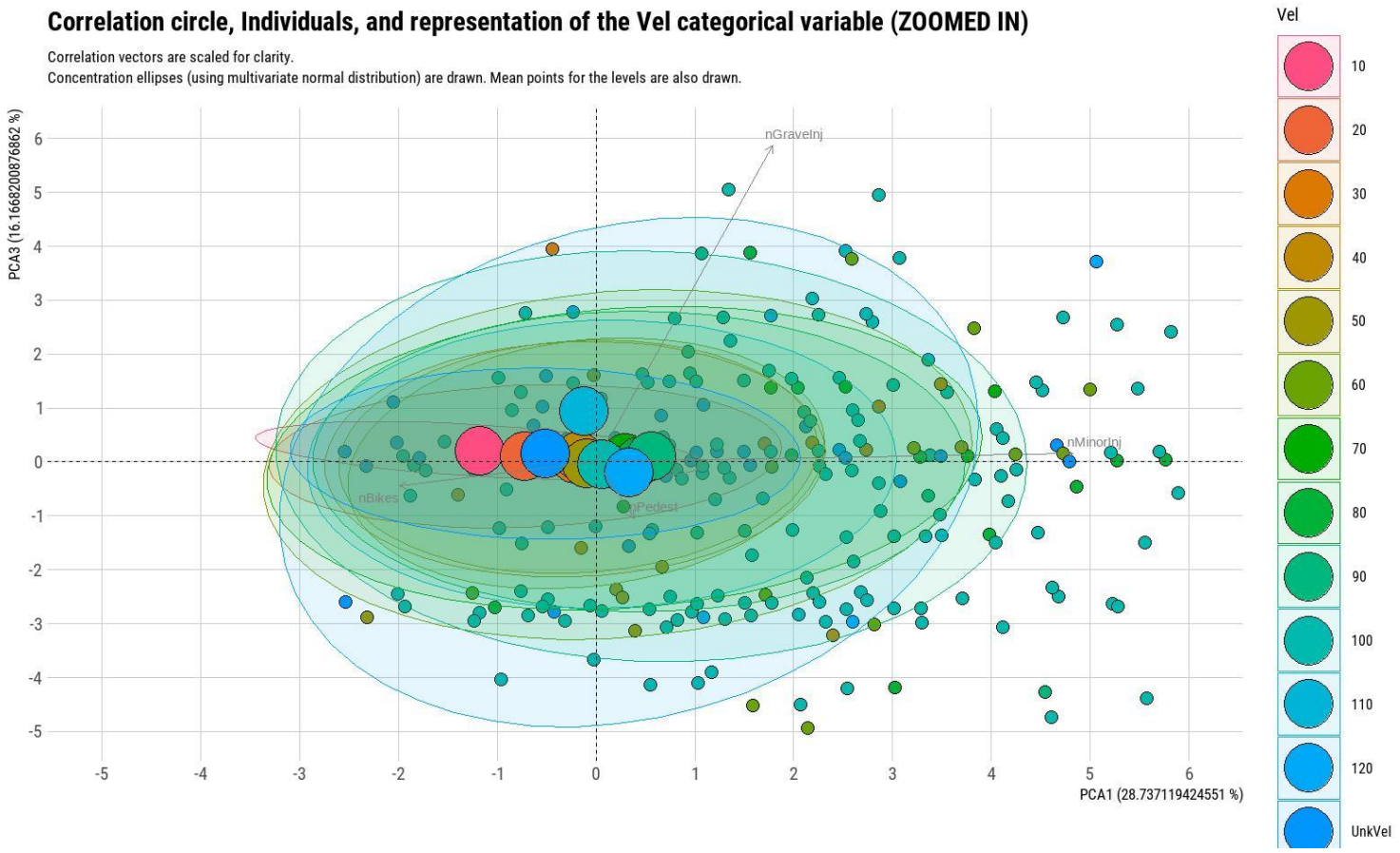


Image 7.22. Correlation circle, individuals, and representation of the Vel categorical variable (ZOOMED IN): PCA1+PCA3

From this graph we can observe that when the accident happens in roads where there's a high maximum speed (closer to 100 kmh) the accident tends to have more units implicated and to be more serious (more injured/dead people) than when it occurs in roads with a lower maximum speed.

Finally, let's study the factorial map PC2 + PC3

FACTORIAL MAP WITH PC2 AND PC3

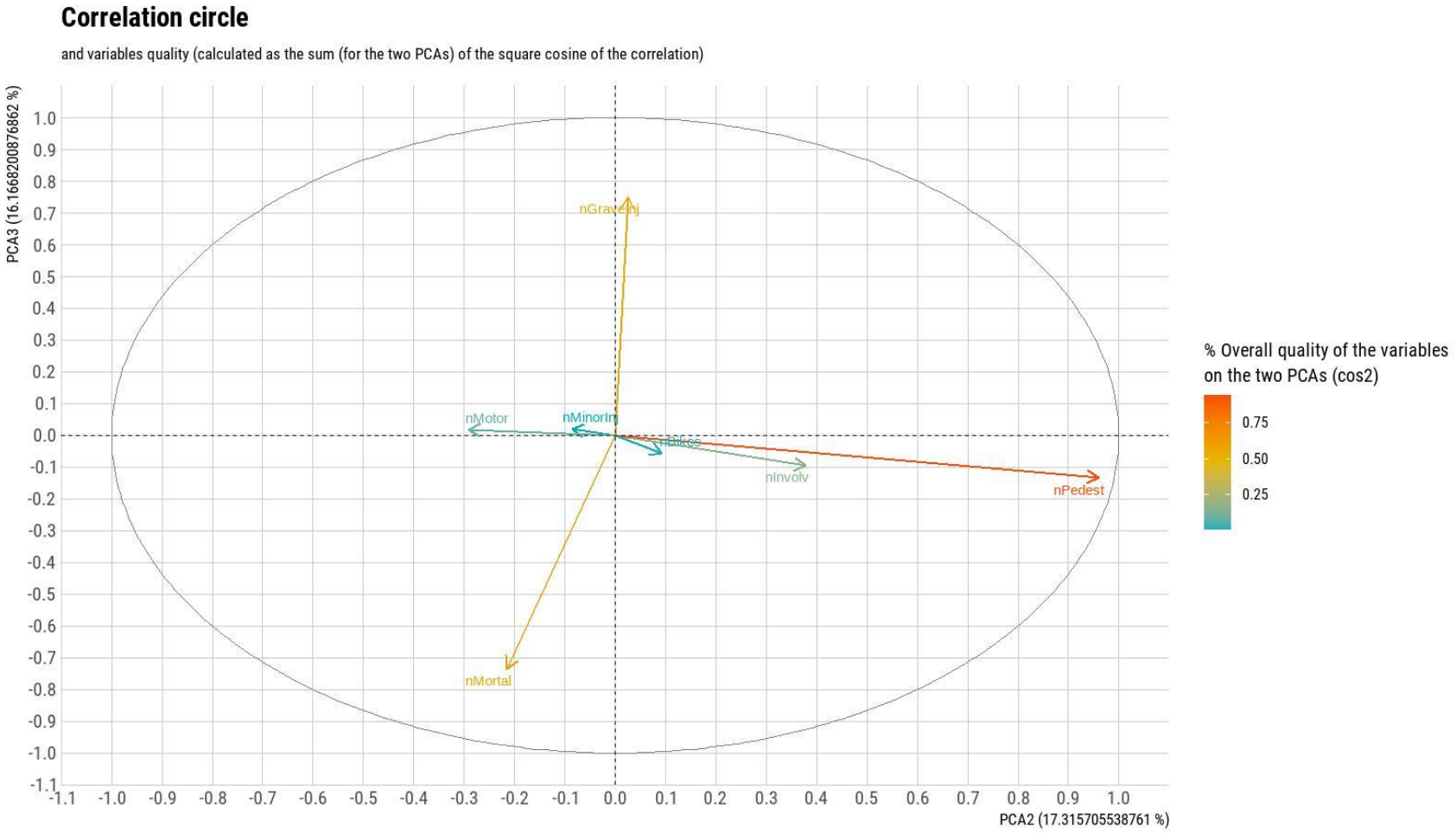


Image 7.23. Correlation circle: PCA2+PCA3

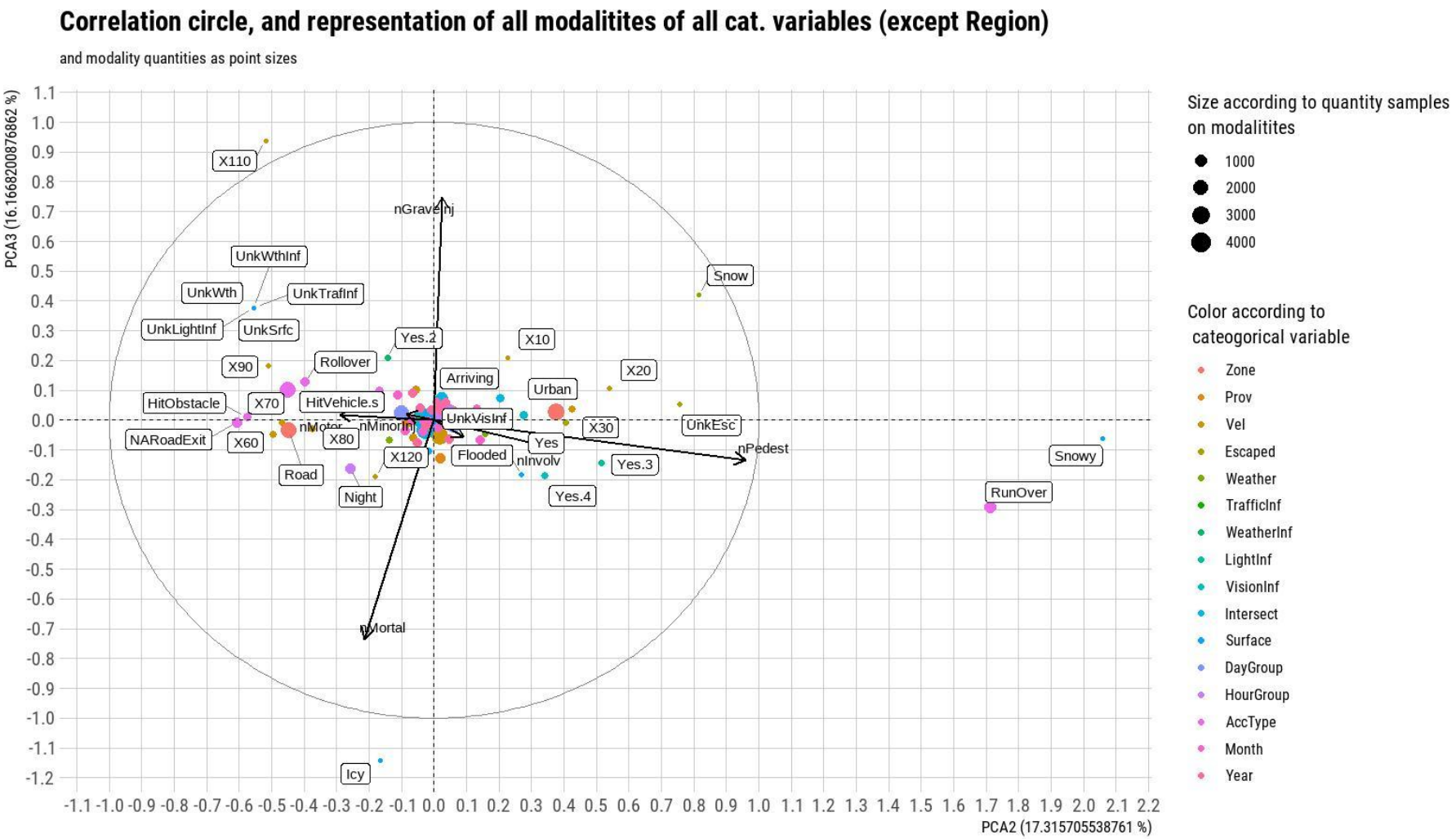


Image 7.24. Correlation circle, and representation of all modalities of all categorical variables (except Region): PCA2+PCA3.

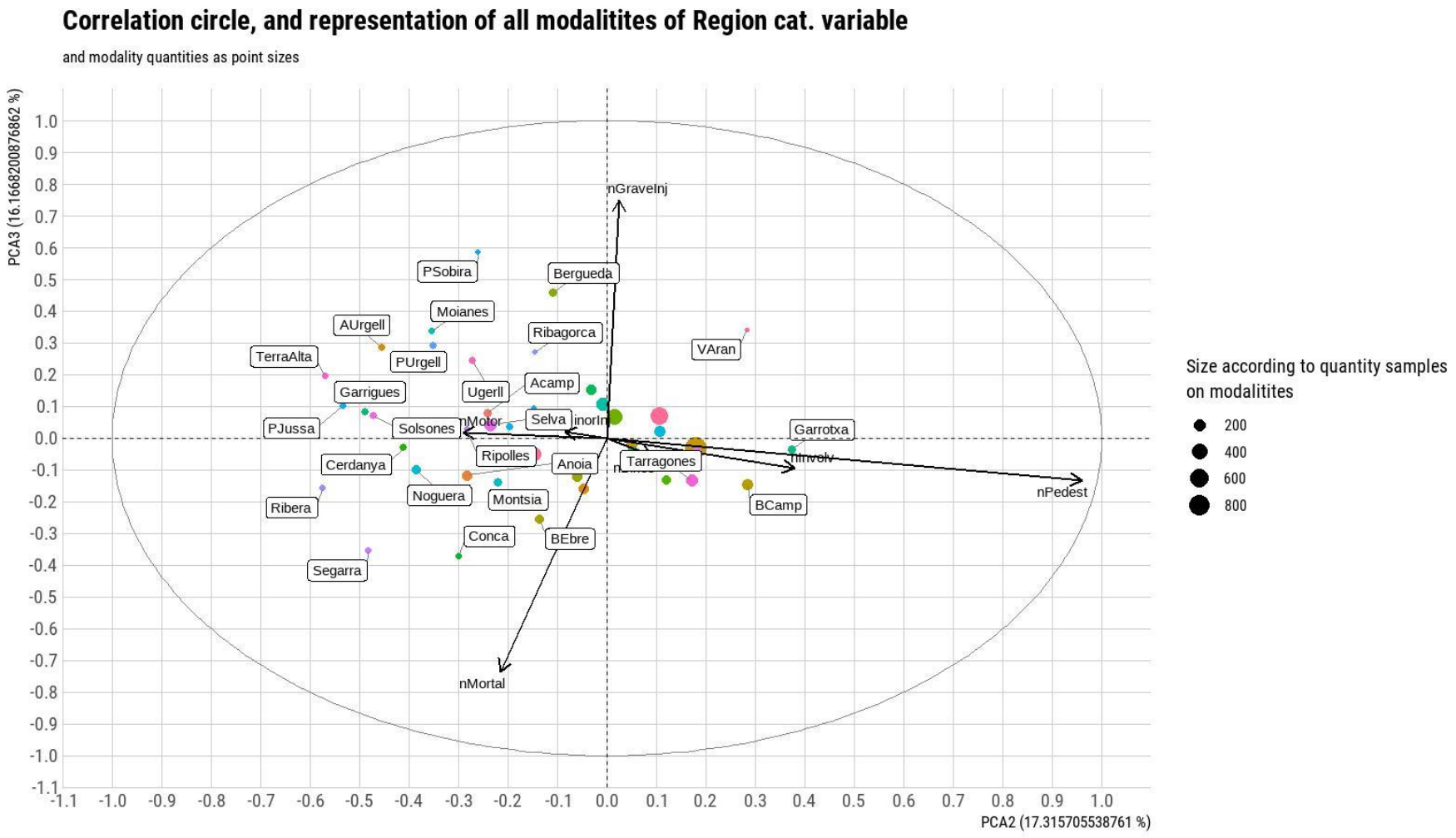


Image 7.25. Correlation circle, and representation of all modalities of Region categorical variable: PCA2+PCA3

Again, looking at the correlation circle of this factorial map doesn't seem to give us any more useful information than the first factorial map (PCA1 + PCA2) gives us.

The variable nGraveInj seems to continue having a strange inverse correlation with the nMortal.

On the representation of the modalities of the categorical variables, we keep seeing that snowy is close with Runover and that this two are associated with a high number of nPedestrian; that icy is associated with a high number of nMortal, and that rural Regions have a lower number of nPedest on their accidents that urban regions, like Tarragones o Barcelones.

PCA CONCLUSIONS

For the conclusion of this PCA analysis, it's important to start off by saying that in general there's very little variability between samples on the factorial maps, as we have seen on the individuals representation for the different factorial maps.

This leads to a small and centered distribution of the modalities of the categorical variables. Those samples whose representation on the factor maps is more away from the center tend to be outliers, and so the centroids that represent modalities which are away from the center have a small number of samples.

All of this, principally occurs because our numeric variables have very low ranges between the lowest value and the highest value, resulting in a lot of samples with the same numeric values on the numeric variables.

Apart from this, we can also extract some interesting conclusions, for example:

- On the factorial map PCA1 + PCA2, we observe two principal behaviors on the samples, one that seems to be strongly correlated with the PCA1, and another one that is both correlated with PCA1 and PCA2.
- Accidents that happen in big populated regions (like Barcelonès, or Tarragonès), tend to be associated with a high number of pedestrians, and rural regions tend to be associated with a small number of pedestrians (or 0).
- Similar to the previous conclusion, when the Zone where the accident happens is Urban, there tends to be more pedestrians and be less serious than when it happens on a Road.
- When an accident happens on roads that have a high maximum speed limit, there usually are more units implicated and be more serious than when the speed limit is lower.
- There doesn't seem to be any relationship between the number of bicycles involved and the other numerical variables. Only a small inverse correlation between nBikes and nMotor that should tell us that when there is a bicycle involved in an accident, it usually happens because of a hit with a car.
- There is a high correlation with a high number of pedestrians on an accident, a snowy surface, and an accident of type Run over.
- There is a high correlation with a high number of mortal people and an icy surface.

- When the accident is of type HtiVehicle/s (vehicles colliding to each other) the accident has more units implicated and is more serious; and when is of type RollOver, HitObstacle or a Road exit it seems to be associated with a wet, slippery, flooded or icy surface and be less serious.

8. Hierarchical Clustering on original data

Precise description of the data

The data used for the clustering process includes all instances of the preprocessed accidents database; and only excludes the *Date* column, since the columns *Year* and *Month* have been introduced and it would be redundant. In addition, the “Date” type of data is not compatible with the function used in R to build the distance matrix.

Clustering method and aggregation criteria

Ward’s (D2) method has been the one used in order to calculate the classes by hierarchical clustering. The metric used has been Gower mixed distance, to simultaneously deal with numerical and qualitative data. Finally, the aggregation criteria has been to minimize the inter-class inertia loss (with Ward’s method).

Resulting dendrogram

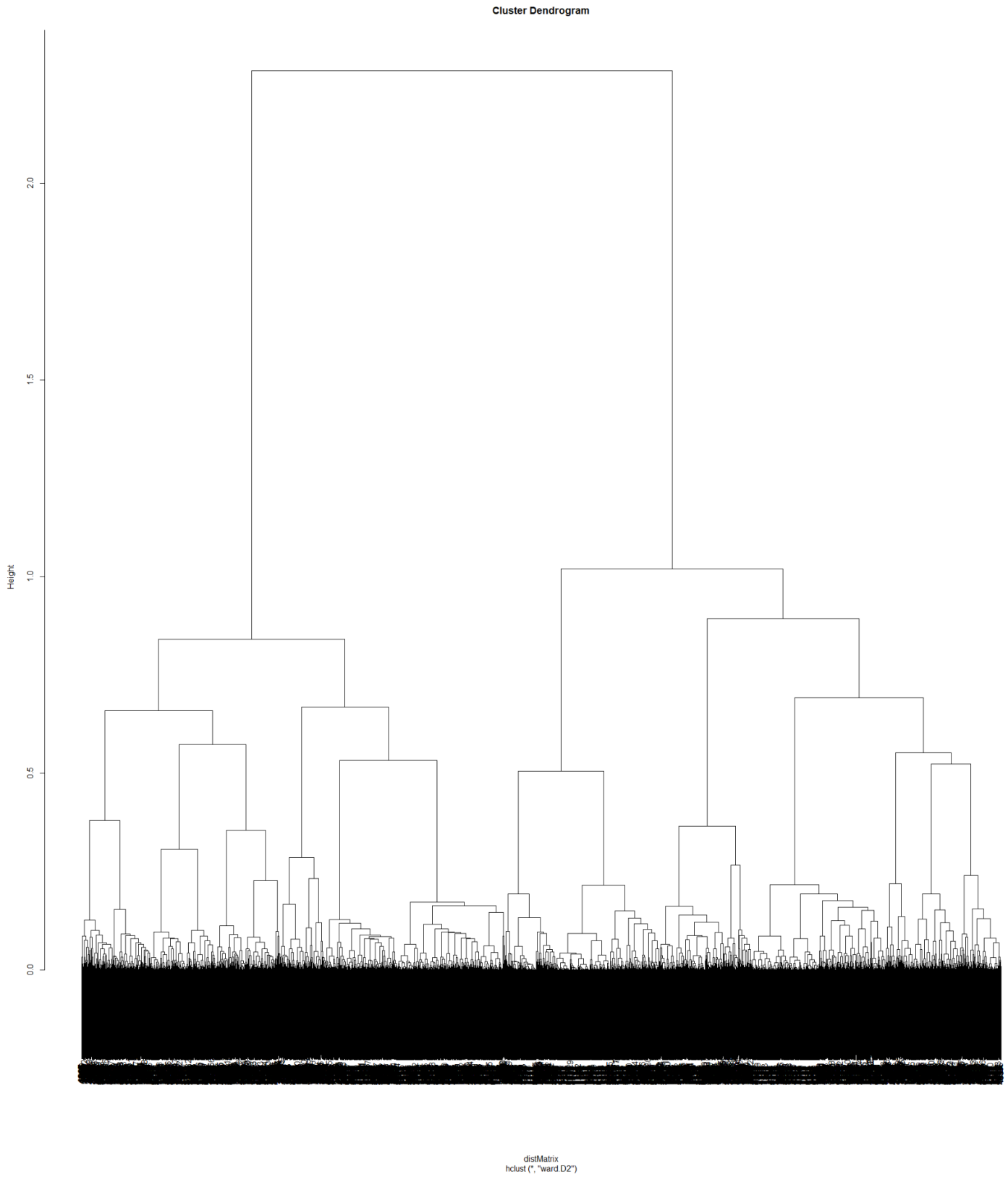


Image 8.1. Dendrogram obtained from performing the hierarchical clustering using Ward's D2 method in RStudio.

Discussion about the final number of clusters

To obtain the final number of clusters, the table of KPI's has been calculated using the *NbClust* function in R, resulting in the following index table:

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Cindex	0.3788	0.3510	0.3312	0.3232	0.3465	0.3329	0.3285	0.3229	0.3162
Mcclain	0.7592	1.5408	1.9721	2.7166	3.0248	3.4005	4.3610	5.0876	5.3953
Silhouette	0.1582	0.1388	0.1684	0.1309	0.1574	0.1662	0.1391	0.1254	0.1407
Dunn	0.0815	0.0815	0.0815	0.0815	0.0915	0.0791	0.0458	0.0458	0.0460

Table 8.1. Summary of the values for each k (from 2 to 10) according to 4 different indexes.

In the previous summary, it can be seen that the candidates for each one of the four indexes considered are k=10, k=2, k=4 and k=6 from top to bottom.

When considering Cindex, k=10 has been discarded because a big number of clusters is difficult to analyze and also because Cindex benefits bigger values of k. Having that in mind, k=5 was the second option, since it is the second lowest value and it is of a manageable size.

Regarding the Mcclain index, k=2 has been clearly the winner, with a huge difference if compared to the other values of k.

For Silhouette, k=4 obtained the best results, but if the corresponding clusters are highlighted in the dendrogram using the *rect.hclust* function in R, it can be seen that they are not very homogeneous. Thus, the second best option (k=2) was the one considered.

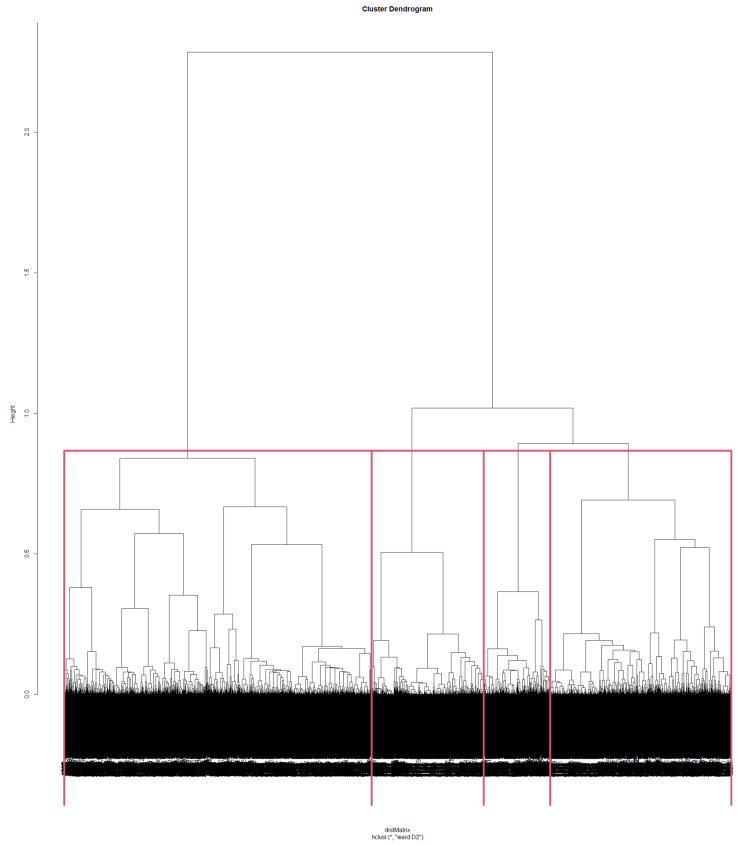


Image 8.2. Highlight of the dendrogram division into 4 clusters.

Finally, for Dunn, k=6 has shown a better result, but k=2, k=3, k=4 and k=5 were also not very far from it.

Also taking into account the PCA resulting plots, we have seen that k=2 was an appropriate candidate, since 2 groups could be appreciated.

The arguments exposed before were consulted with Dante Conti (researcher specialized in Data Mining among other fields), in order to obtain an expert's opinion. And we discussed which number of clusters would be better for this data set, k=2 or k=5. Finally, with Dante's opinion and with the points mentioned above, in the end, we decided to choose **k=2**.

Table describing the clusters size

	Cluster 1	Cluster 2
Number of individuals	2696	2304

Table 8.2. Number of instances of the data set in each cluster.

9. Cluster profiling

Cluster profiling is the final step of this study, where descriptions of the different clusters are generated by analyzing profiling plots and statistics. By profiling, a clear profile of individuals can be obtained that can help to make good decisions and understand the dataset.

In order to obtain the profiles of our clusters, we have used an R-script that generates for each variable different profiling plots where the values for this variable can be compared between clusters.

We have analyzed all the profiling plots for all the features of our dataset (which you can find in a subfolder of this project) and we have selected those that are significant to create our profiles.

Profiling plots and statistics

Zone

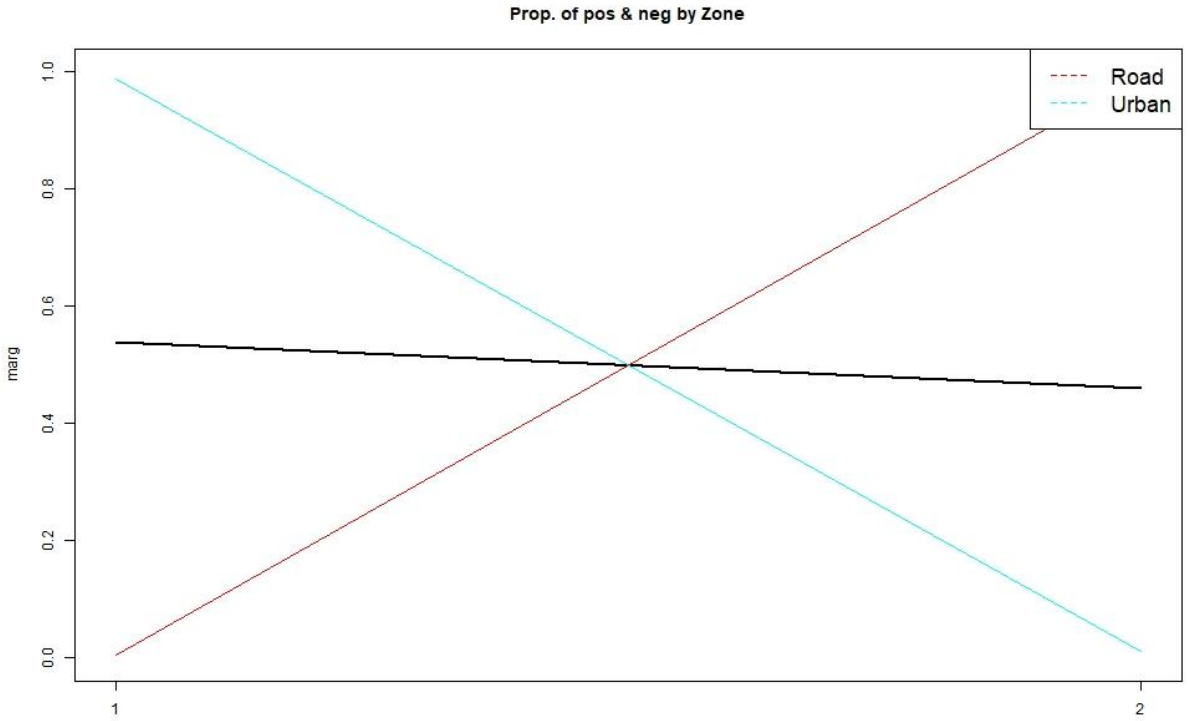


Image 9.1. Prop of Zone feature between classes.

```
[1] "Distribucions condicionades a columnes:"

P          Road       Urban
  1 0.005249344 0.988946205
  2 0.994750656 0.011053795
```

Image 9.2. Distribution conditioned to columns of Zone feature by classes.

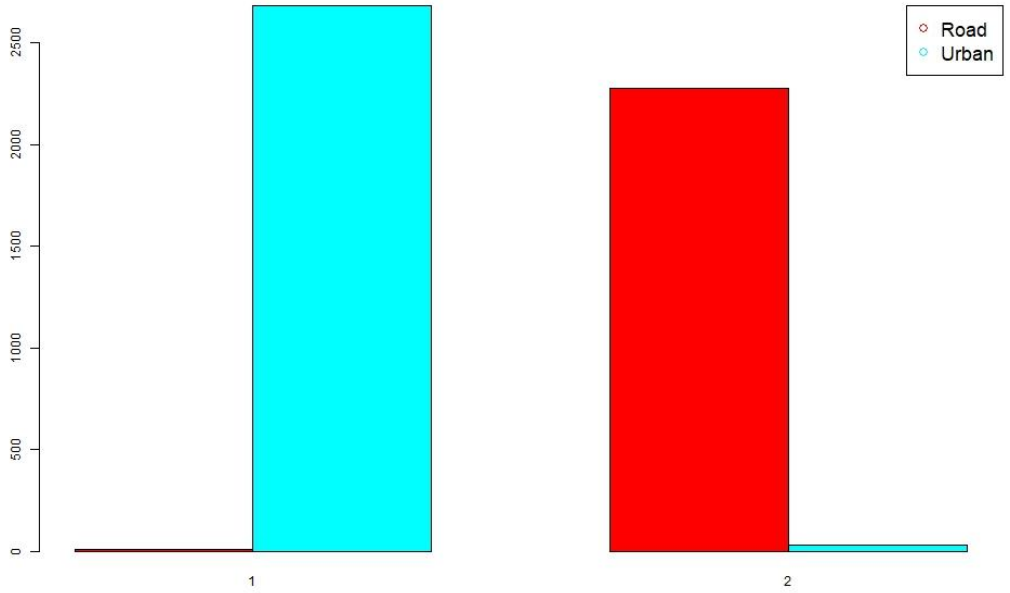


Image 9.3. Barplot of Zone feature for each class.

As it can be seen in the three previous images, the feature zone is a very significant one because each cluster has almost the 100% of one zone level. In cluster 1 the 98.89% of individuals have the urban value for the variable zone, that means that 2684 accidents in cluster one have occurred in an urban zone and only 12 in a road zone. On the other hand, in cluster 2 the 99.47% of individuals have the road value for the zone variable, which means that 2274 accidents in this cluster have occurred in roads and only 30 in urban zone.

Region and province

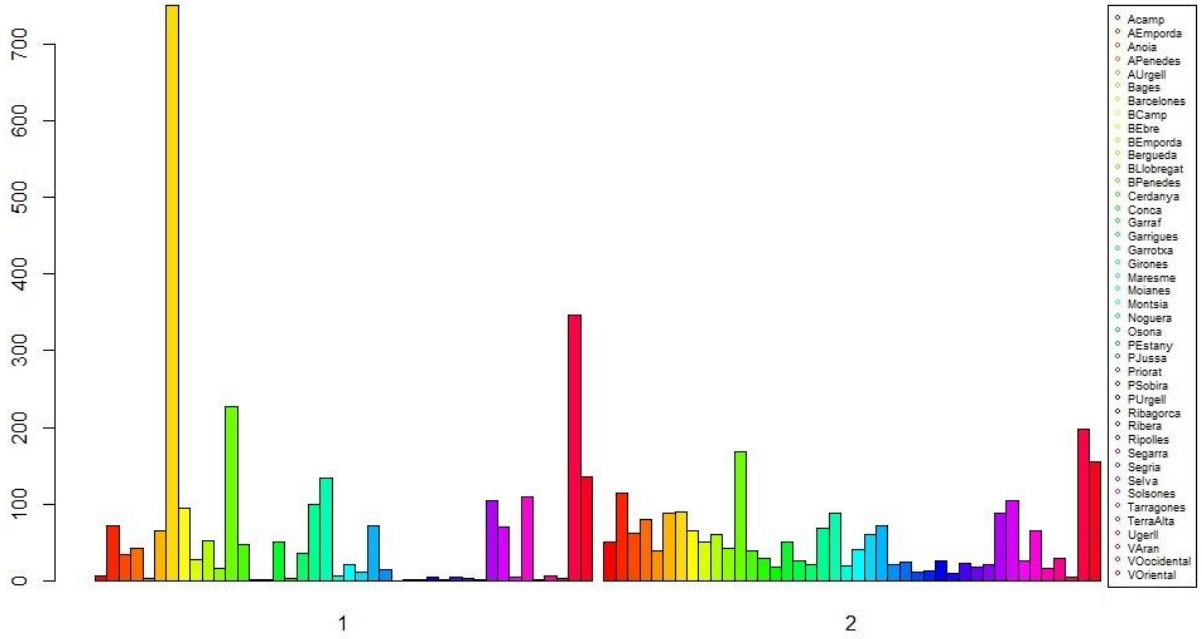


Image 9.4. Barplot of Region feature for each class.

```
[1] "Distribucions condicionades a columnes:"

P         Acamp    AEmporda      Anoia   APenedes     AUrgell       Bages Barcelones       BCamp       BEbre   BEmporda   Bergueda BLlobregat
  1 0.10714286 0.38502674 0.35416667 0.34426230 0.09090909 0.42580645 0.89285714 0.59119497 0.34615385 0.46017699 0.27118644 0.57323232
  2 0.89285714 0.61497326 0.64583333 0.65573770 0.90909091 0.57419355 0.10714286 0.40880503 0.65384615 0.53982301 0.72881356 0.42676768

P      BPenedes    Cerdanya      Conca      Garraf  Garrigues   Garrotxa    Girones    Maresme    Moianes    Montsia    Noguera      Osona
  1 0.54545455 0.06451613 0.10000000 0.50000000 0.13333333 0.63157895 0.58928571 0.60360360 0.24000000 0.34920635 0.16666667 0.50000000
  2 0.45454545 0.93548387 0.90000000 0.50000000 0.86666667 0.36842105 0.41071429 0.39639640 0.76000000 0.65079365 0.83333333 0.50000000

P       PEstany     PJussa    Priorat    PSobira    PUrgell  Ribagorca     Ribera   Ripolles    Segarra     Segria      Selva   Solsones
  1 0.38888889 0.00000000 0.15384615 0.13333333 0.16129032 0.18181818 0.17857143 0.18181818 0.04347826 0.54404145 0.40340909 0.16129032
  2 0.61111111 1.00000000 0.84615385 0.86666667 0.83870968 0.81818182 0.82142857 0.81818182 0.95652174 0.45595855 0.59659091 0.83870968

P   Tarragones  TerraAlta     Ugerll      VAran VOccidental  VOriental
  1 0.62857143 0.05555556 0.19444444 0.37500000  0.63720074 0.46551724
  2 0.37142857 0.94444444 0.80555556 0.62500000  0.36279926 0.53448276
```

Image 9.5. Distribution conditioned to columns of Region feature by classes.

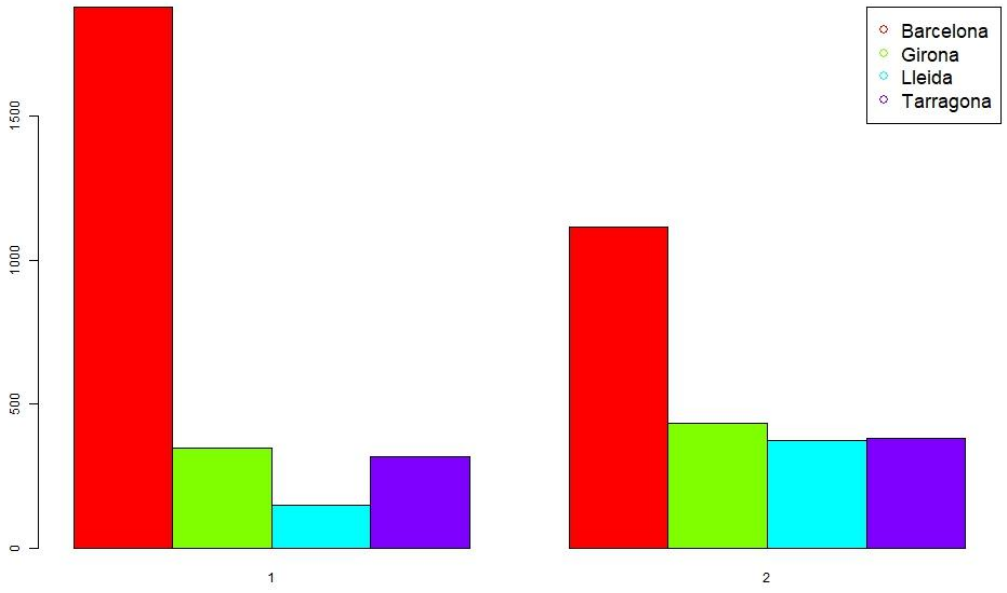


Image 9.6. Barplot of Province feature for each classes

```
[1] "Distribucions condicionades a columnes:"

P   Barcelona    Girona    Lleida Tarragona
  1 0.6277982 0.4457216 0.2870722 0.4541547
  2 0.3722018 0.5542784 0.7129278 0.5458453
```

Image 9.7. Distribution conditioned to columns of Province feature by classes.

Although Region and Province are not the most relevant and significant features, it is interesting to analyze them because it can be observed that cluster 1 has more accidents in the Barcelona province and cluster 2 has more accidents in Lleida.

Also, it can be noticed that although the two clusters have more or less the same distributions for the Province levels, if we take a look at images 9.6 and 9.7, we can see that there are regions that are not that equally distributed between clusters. For example, cluster 1 has 89.28% of the accidents that occurred in the Barcelonés region. Cluster 2 has the totality of the Pallars Jussà accidents, the 90.9% of accidents that occurred in Alt Urgell, etc.

It is interesting to see that, although there are huge differences between the clusters in terms of regions, in the province feature (except for the Lleida and Barcelona provinces) the distributions end up by equalling.

nMortals

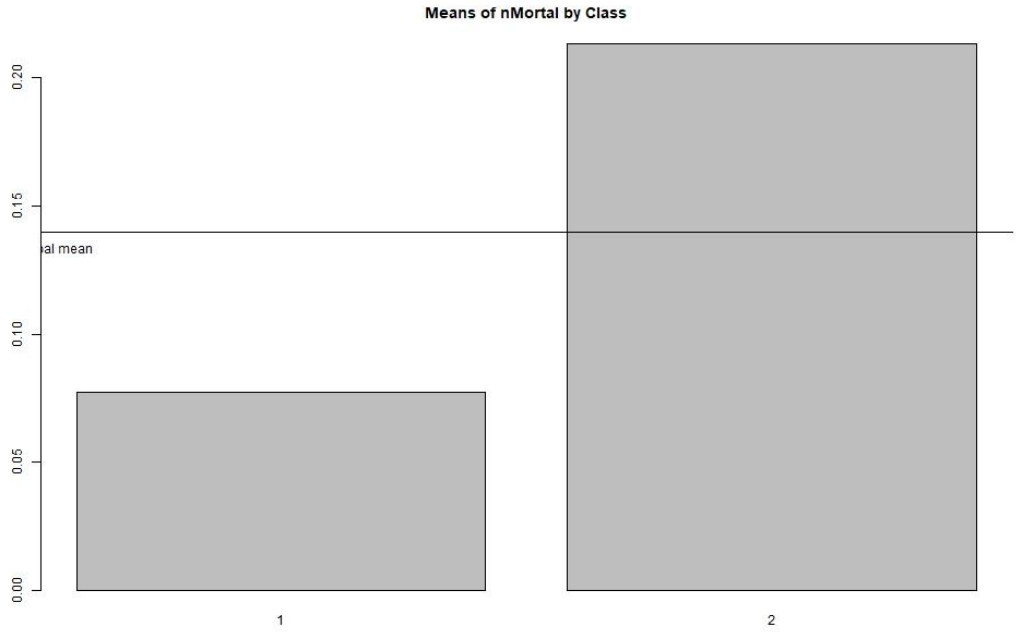


Image 9.8. Barplot of means of nMortal feature by class.

```
[1] "Anàlisi per classes de la variable: nMortal"
[1] "Estadístics per groups:"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.07715 0.00000 4.00000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2135  0.0000 13.0000
```

Image 9.9. Statistics of nMortal feature by classes.

In these images it can be observed that cluster one has less mortal victims than cluster 2. Cluster 1 has a mean of 0.077 with a maximum value of 4 mortal victims in one accident. Cluster 2 instead, has a mean of 0.2135 and a maximum value of 13 deaths in one single accident.

Note that we haven't added the boxplot image because, due to the little variance of our numerical variables and the "outliers" (which are not discardable outliers, just high values), some of the resulting boxplots are not useful to analyze the variables.

nMinorInj

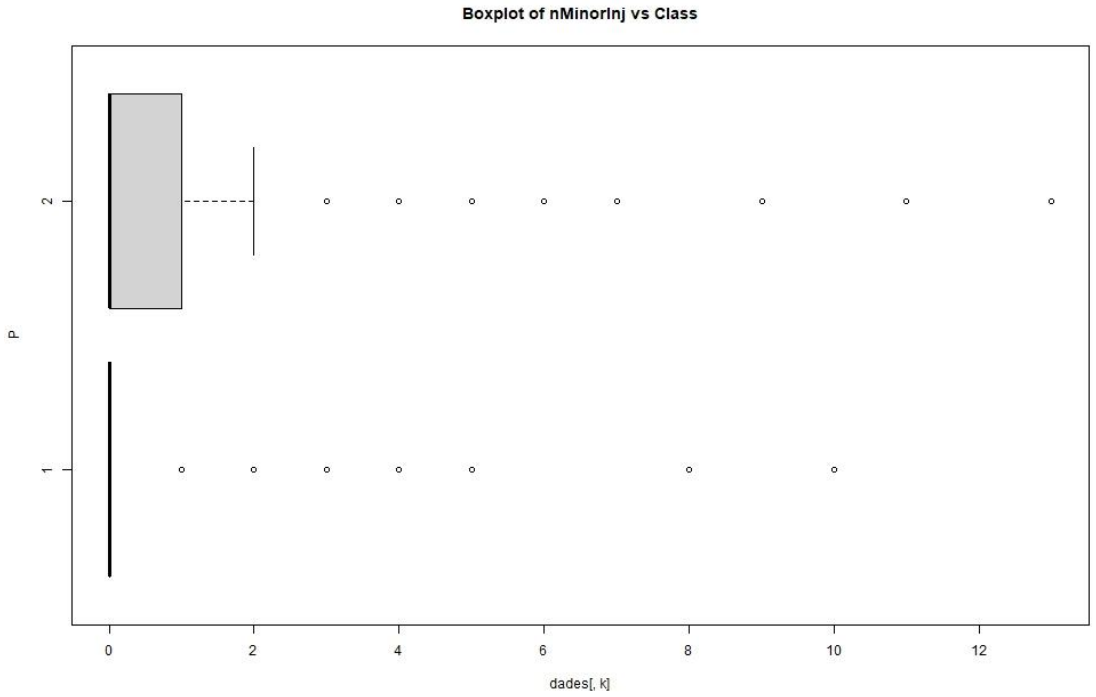


Image 9.10. Boxplot of nMinorInj feature by class.

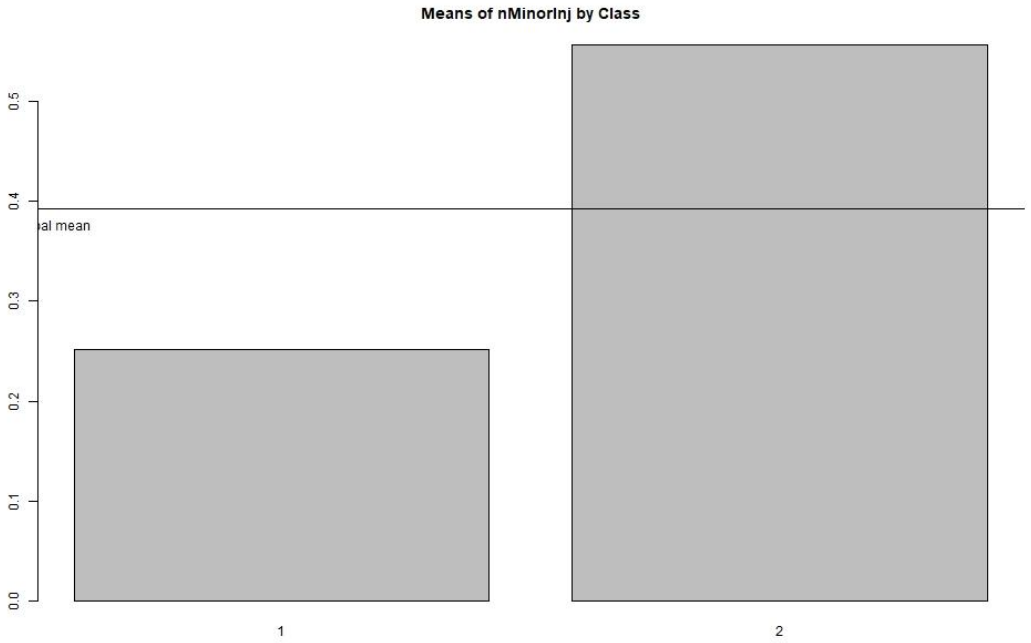


Image 9.11. Barplot of means of nMinorInj feature by class.

```
[1] "Anàlisi per classes de la variable: nMinorInj"
[1] "Estadístics per groups:"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2519  0.0000 10.0000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.5569  1.0000 13.0000
```

Image 9.12. Statistics of nMinorInj feature by classes.

In terms of minor injured victims, cluster 1 has a minor mean of minor injured victims (0.2519) and a maximum value of 10 minor injured in a single accident. On the other hand, cluster 2 has a bigger mean (0.5569) and a maximum value of 13 minor injured victims.

nPedest

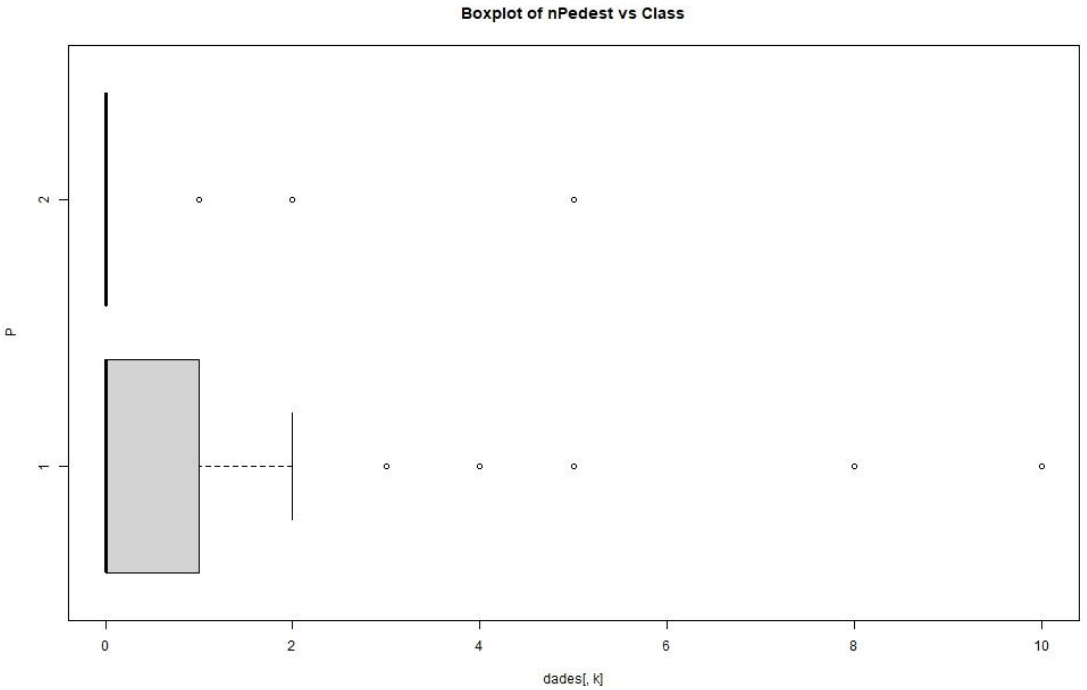


Image 9.13. Boxplot of nPedest feature by class.

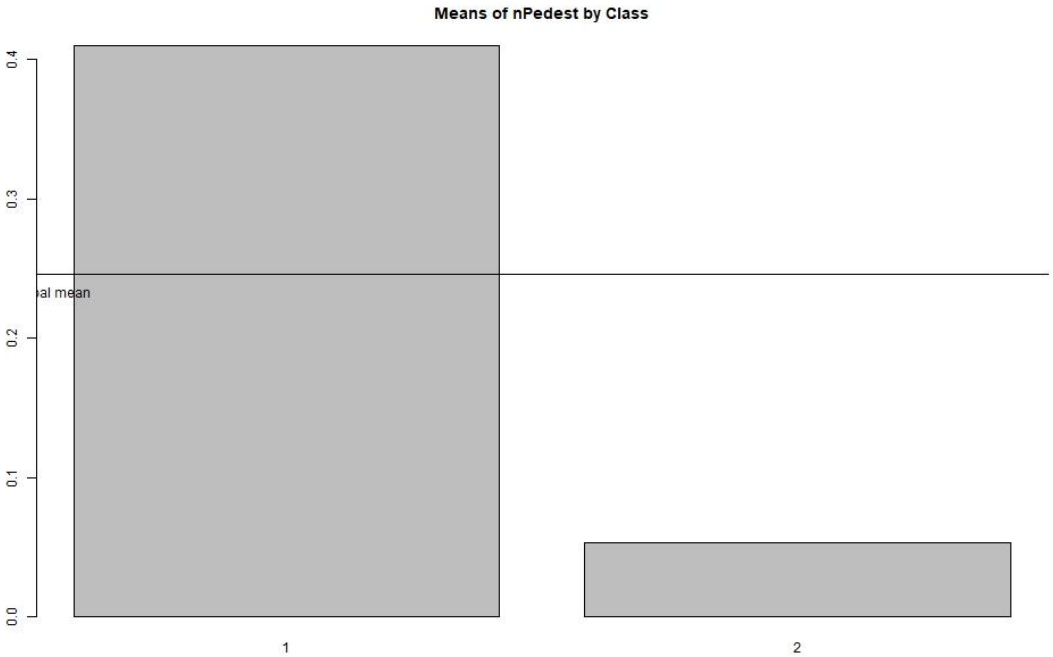


Image 9.14. Barplot of means of nPedest feature by class.

```
[1] "Anàlisi per classes de la variable: nPedest"
[1] "Estadistics per groups:"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.4102  1.0000 10.0000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.05295 0.00000 5.00000
```

Image 9.15. Statistics of nMinorInj feature by classes.

The feature nPedest is a very significant one, as it can be seen that cluster 1 has a mean of 0.4102 pedestrians involved in his accidents and cluster 2 has a mean of 0.05295 pedestrians. Also, in cluster 1 the maximum number of pedestrians involved in one same accident is 10 and, on the other hand, cluster 2 has a maximum value of 5.

Vel

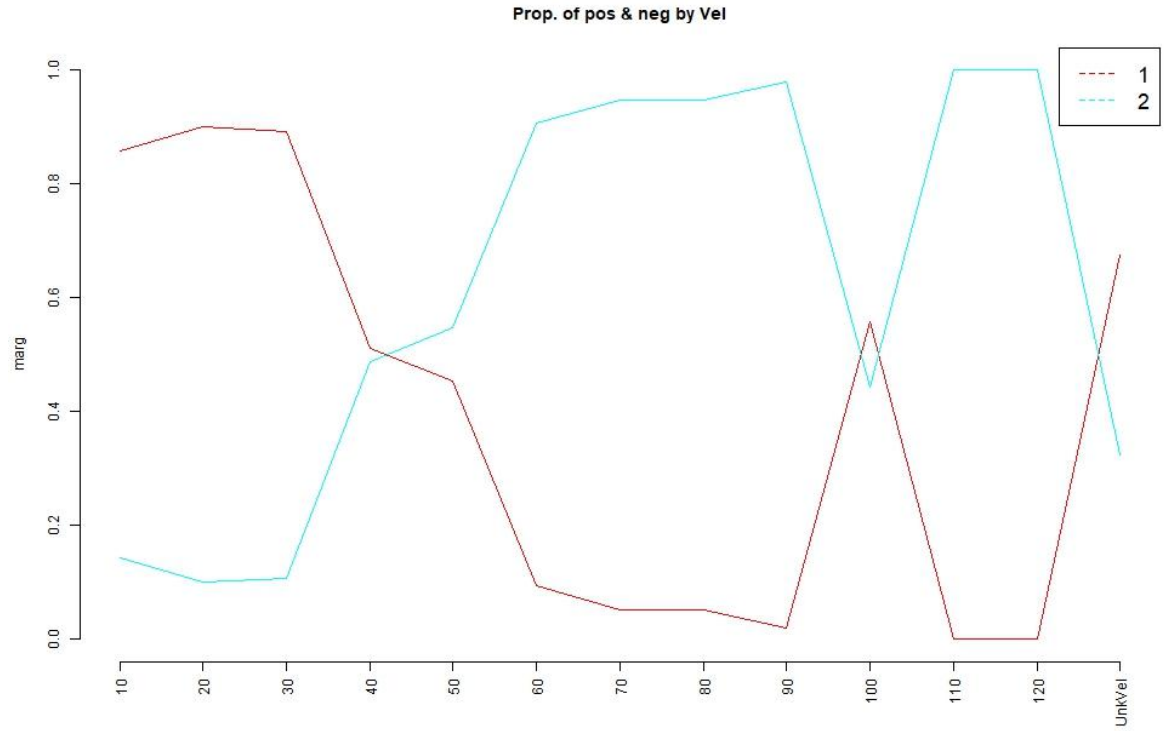


Image 9.16. Prop of Vel feature between classes.

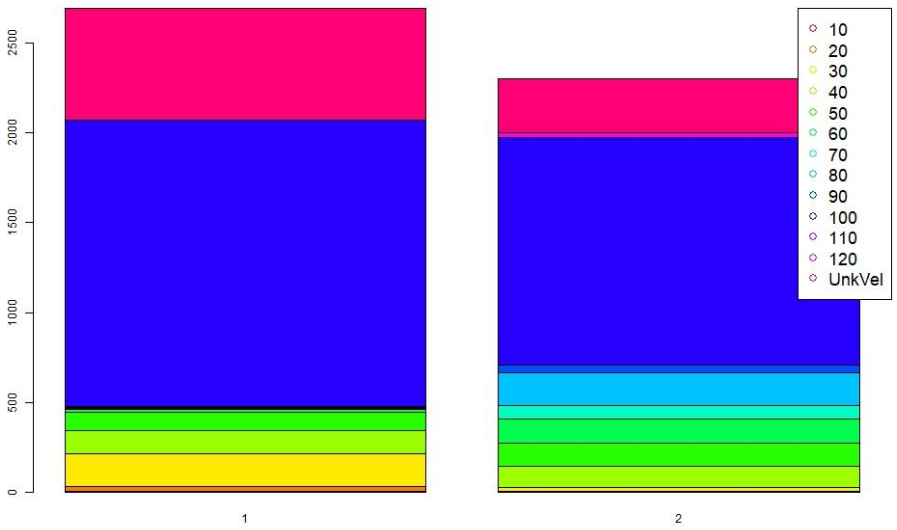


Image 9.17. Barplot of Vel feature for each classes

```
[1] "Distribucions condicionades a columnes:"

P           10         20         30         40         50         60
  1 0.85714286 0.90000000 0.89215686 0.51219512 0.45299145 0.09333333
  2 0.14285714 0.10000000 0.10784314 0.48780488 0.54700855 0.90666667

         70         80         90        100        110        120     Unkvel
0.05263158 0.05235602 0.02040816 0.55777311 0.00000000 0.00000000 0.67564655
0.94736842 0.94764398 0.97959184 0.44222689 1.00000000 1.00000000 0.32435345
```

Image 9.18. Distribution conditioned to columns of Province feature by classes.

In terms of velocity, the plot above shows that cluster 1, except for the 100 km/h value, has the majority of the accidents with low values of maximum velocity permitted, as it can be seen that cluster 1 holds the 85.7%, 90% and 89.2% of accidents that occurred with a maximum velocity value of 10 km/h, 20 km/h and 30 km/h respectively. Then, each cluster has an equal number of accidents that have the velocity value at 40 km/h, 50 km/h and 100 km/h. For the rest, cluster 2 has the majority of accidents with high velocity values.

AccType

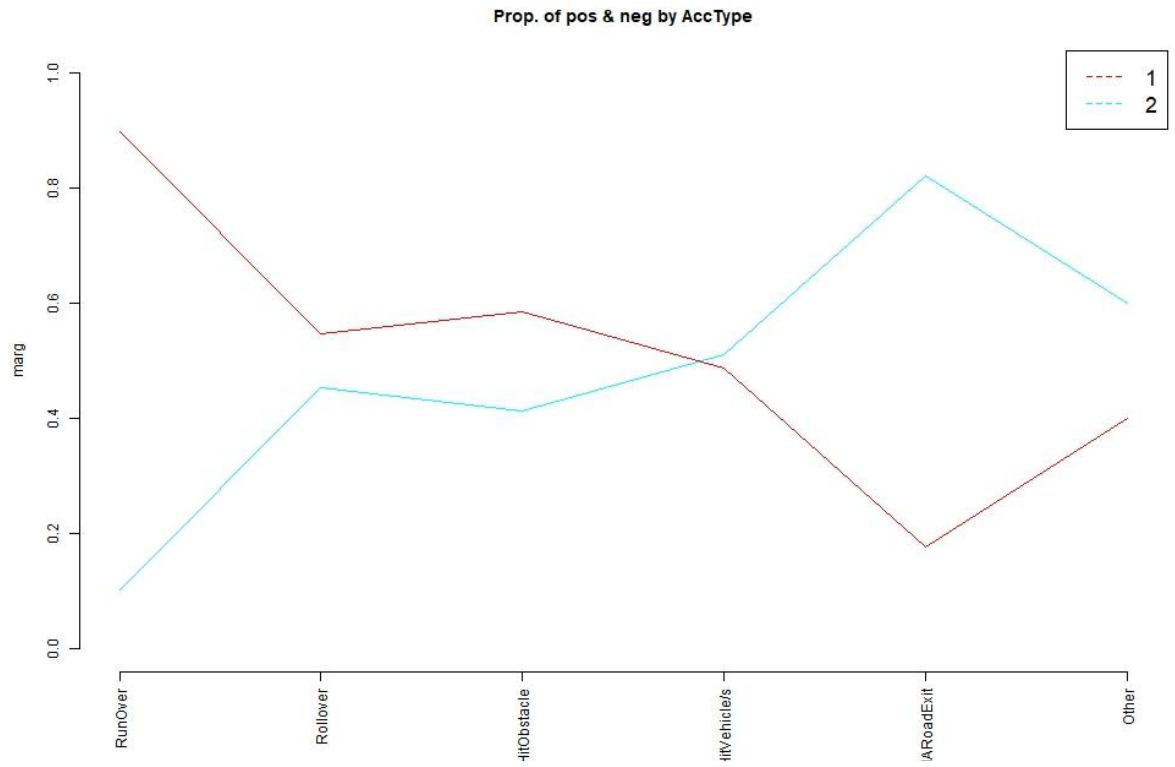


Image 9.19. Prop of AccType feature between classes.

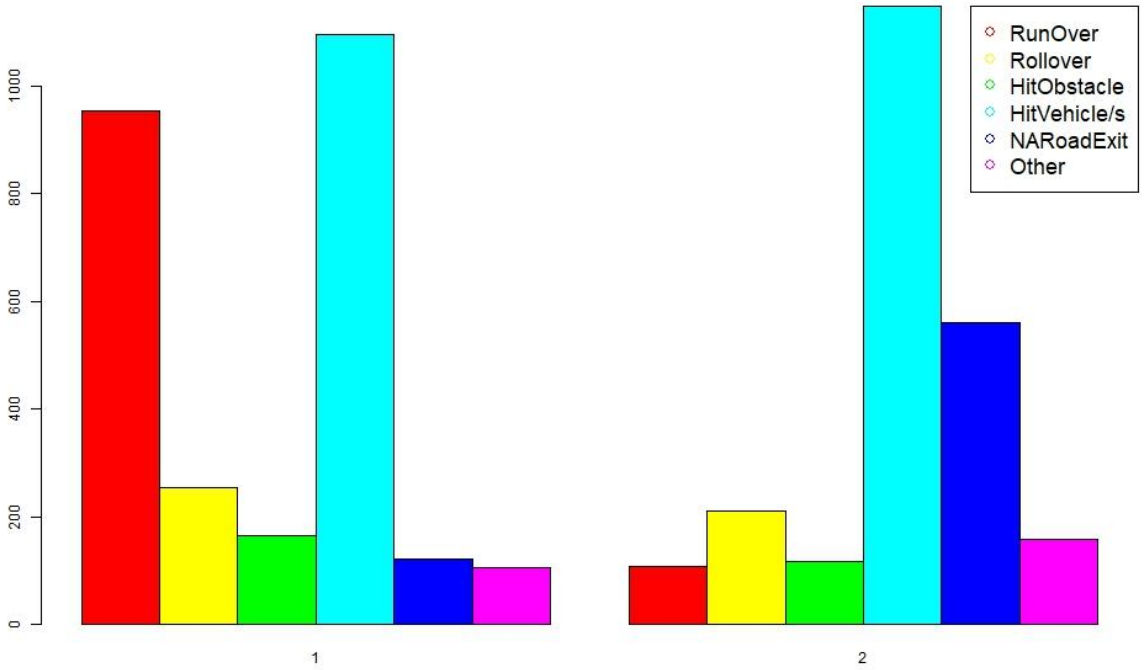


Image 9.20. Barplot of AccType feature for each class.

```
[1] "Distribucions condicionades a columnes:"

P     RunOver  Rollover HitObstacle HitVehicle/s NARoadExit     Other
  1 0.8983051 0.5472103   0.5857143    0.4879786  0.1776799 0.4000000
  2 0.1016949 0.4527897   0.4142857    0.5120214  0.8223201 0.6000000
```

Image 9.21. Distribution conditioned to columns of AccType variable by classes.

Finally, for the accident type it is clearly noticeable that cluster 1 has almost all the run-over accidents, with 89.83%, and that cluster 2 holds almost all the unknown road exit accidents (NARoadExit), with 82.23%. For all the other types of accidents, each cluster has approximately the same number of accidents of each accident type.

P-values

```
[1] "P.values per class: 1"
      Zone     Region       Prov    nMortal  nMinorInj     nMotor        Vel
  0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00
   Escaped    Weather TrafficInf WeatherInf   LightInf  VisionInf  Intersect
  0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00
   Surface   DayGroup  HourGroup    AccType      Month       Year    nPedest
  0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00  1.11e-121
   nInvolv  nGraveInj     nBikes
  1.42e-12   6.11e-03   1.59e-02
```

Image 9.22. P-values of the different variables for class 1.

```
[1] "P.values per class: 2"
      Zone     Region       Prov    nPedest        Vel    Escaped    Weather
  0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00
TrafficInf WeatherInf   LightInf  VisionInf  Intersect    Surface   DayGroup
  0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00   0.00e+00
 HourGroup    AccType      Month       Year  nMinorInj    nMortal     nMotor
  0.00e+00   0.00e+00   0.00e+00   0.00e+00   8.43e-33   4.19e-29   5.85e-22
   nInvolv  nGraveInj     nBikes
  1.42e-12   6.11e-03   1.59e-02
```

Image 9.23. P-values of the different variables for class 2.

These p-values represent, for each cluster, which variables are significant to describe it. As minor p-value has the variable, more independent it is and more significant to describe the cluster.

Many of the variables that we have found significant by analyzing the profiling plots are appearing in these p-values also significant and the same happens with the variables we have found less or not significant. But it must be noted that there are variables that have a p-value of 0 that we have not found significant. If we take a closer look at these variables, we can see that, in most of the cases, the difference between clusters remains on the unknown values (this happens for example in the escaped variable where cluster 1 has all the unknown values but approximately the same percentage of yeses and no values as cluster 2).
Also, there are other variables that seem significant to define a cluster when observing some plots, but when analyzing other types of plots and statistics are not relevant enough. For example, there are variables that some of their modalities only appear in one of the clusters, which can make us think that it would be definitory enough, but when looking the statistics of that modality we observe that there are very few individuals that have this modality, so they end up being a little fraction of the real individuals that are on that cluster and are not definitory.

Conclusions

Once all the profiling plots have been analyzed, we have been able to generate a description for each cluster.

Cluster 1 is formed by those accidents that occurred in an urban zone. This cluster holds the majority of the accidents that happened in zones with low velocities (10 km/h to 30 km/h) and half the accidents that occurred in a zone with medium velocity (40-50 km/h) and with a velocity of 100 km/h. Also in this cluster there are the big majority of the run over accidents

of the data set, and in relation to that, this cluster has more number of pedestrians involved(which makes sense), less mortality and less minor injured victims.

Cluster 2 is formed by those accidents that occurred in a road zone. This cluster holds almost all the accidents that occurred with a velocity from 60 km/h to 120 km/h (except for the case of 100 km/h) and half the accidents that occurred in a zone with medium velocity. This cluster has more mortality and more minor injured victims. Also, in this cluster can be found the majority of the accidents with an unknown exit of the road.

It is important to take into account, before diving into conclusions, that although we have seen that variables like vel, region and province have noticeable differences between clusters, any of these differences are sufficient to define our clusters. For example, we have observed that cluster 1 has all of the accidents that occurred in places where the maximum velocity permitted has a low value and cluster 2 has all the accidents that occurred in places where the maximum velocity permitted has a high value, but there are not much individuals (accidents) with that values, a big percentage of the accidents in both clusters have a value of vel of 100 km/h. So we can not use velocity to define the clusters and say that the accidents of cluster 1 occurred in zones with a low velocity, as the number of accidents in cluster 1 with vel of 100 km/h is way much bigger than the sum of velocities from 10 km/h to 30 km/h. A similar thing happens with region and prov.

In conclusion we can say that:

(1) Cluster 1 is formed by accidents that have occurred in an urban zone, accidents that are mostly run overs and hits between vehicles and that have less mortality, less minor injured victims but more pedestrians involved.
(2) Cluster 2 is formed by accidents that have occurred in a road zone, accidents that are mostly unknown exits of the road and hits between vehicles and accidents that have more mortality, more minor injured victims and less number of pedestrians involved.

Other considerations

As commented below in the Hierarchical Clustering section, when cutting the dendrogram tree and when observing the kpi's results we doubted on the number of clusters to pick and, after analyzing it with an expert (the professor Dante) we reduced the options to k=2 and k=5. We decided to use profiling as the last method to discard one of the two options (by doing a quick analysis of the possible profiles we could extract and the quality of these) and

we have finally kept option of k=2. But we find it interesting to explain why the profiling with 5 clusters is not good or worse than the profiling for 2 clusters.

When executing the profiling R-script with 5 clusters, we can observe that we could relate the clusters obtained with k=5 to the clusters obtained with k=2. Cluster 1 from the k=2 clustering seems to be divided into clusters 1, 4 and 5 of the k=5 clustering. In addition, cluster 2 from the k=2 clustering seems to be divided into clusters 2 and 3 of the k=5 version. The reasons are the following:

(On the k=5 clustering version)

(1) Clusters 1,4 and 5 are formed by accidents that occurred in urban zones, and clusters 2 and 3 are formed by accidents that occurred in road zones.
(2) Clusters 2, 3 have more mortality. Clusters 1,4,5 have less mortality.
(3) Clusters 2, 3 have more minor injured victims. Clusters 1,4 and 5 have less minor injured victims.
(4) Clusters 2 and 3 have much less number of pedestrians involved in accidents. Clusters 1, 4 and 5 have more number of pedestrians involved.

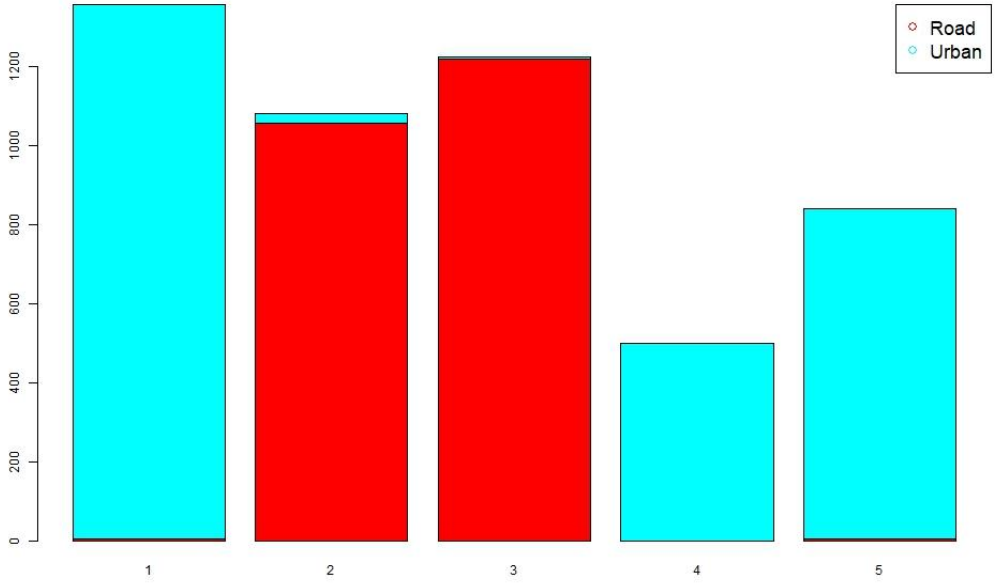


Image 9.24. Barplot of Zone feature for each class in the case of 5 clusters.

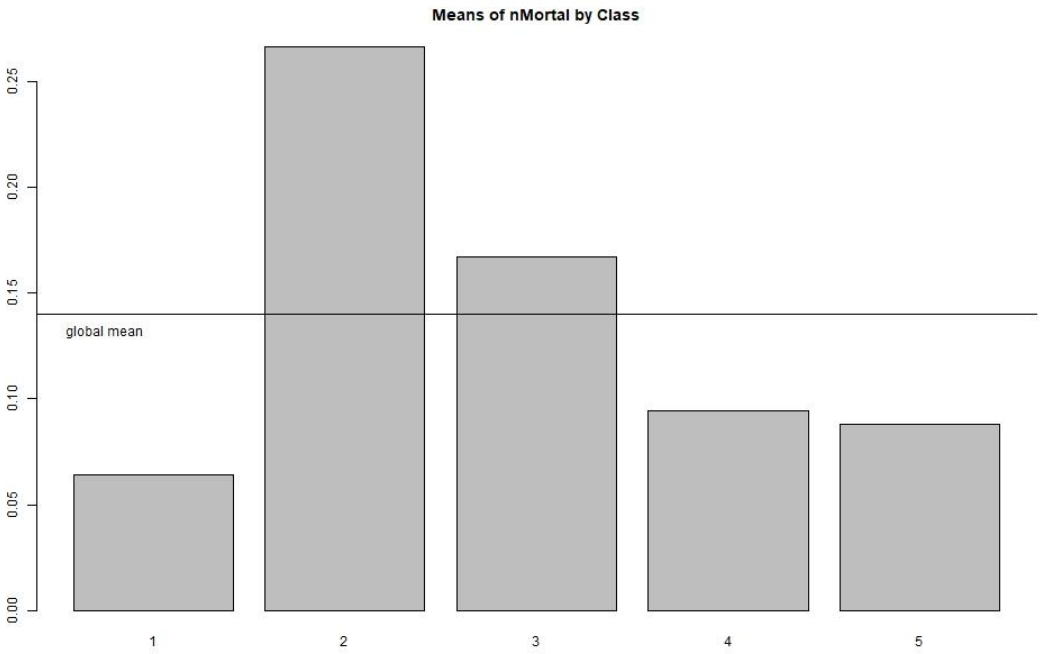


Image 9.25. Barplot of means of the nMortal variable for each class in the case of 5 clusters.

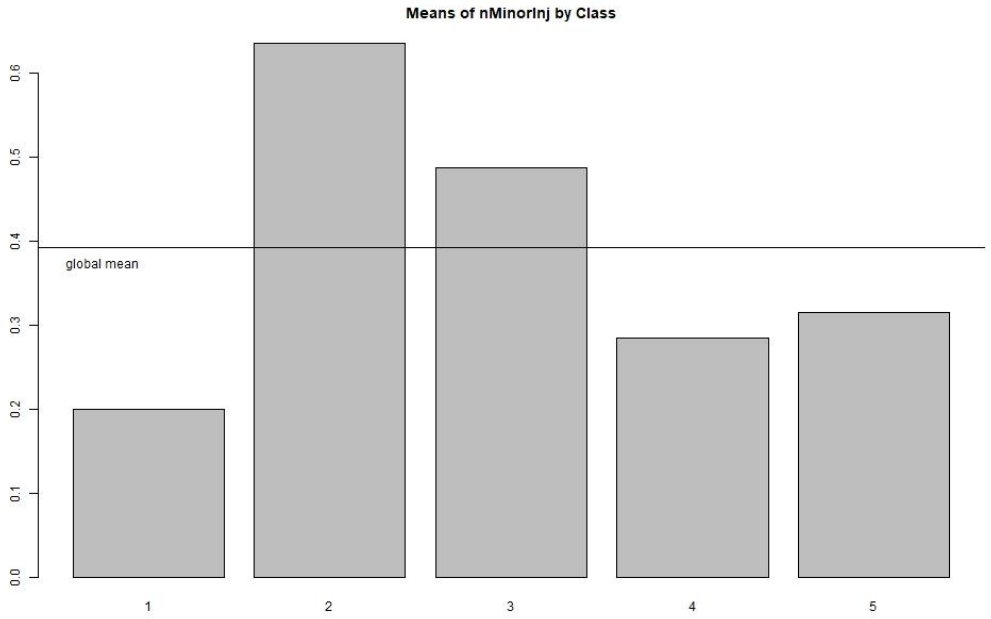


Image 9.26. Barplot of means of the nMinorInj variable for each class in the case of 5 clusters.

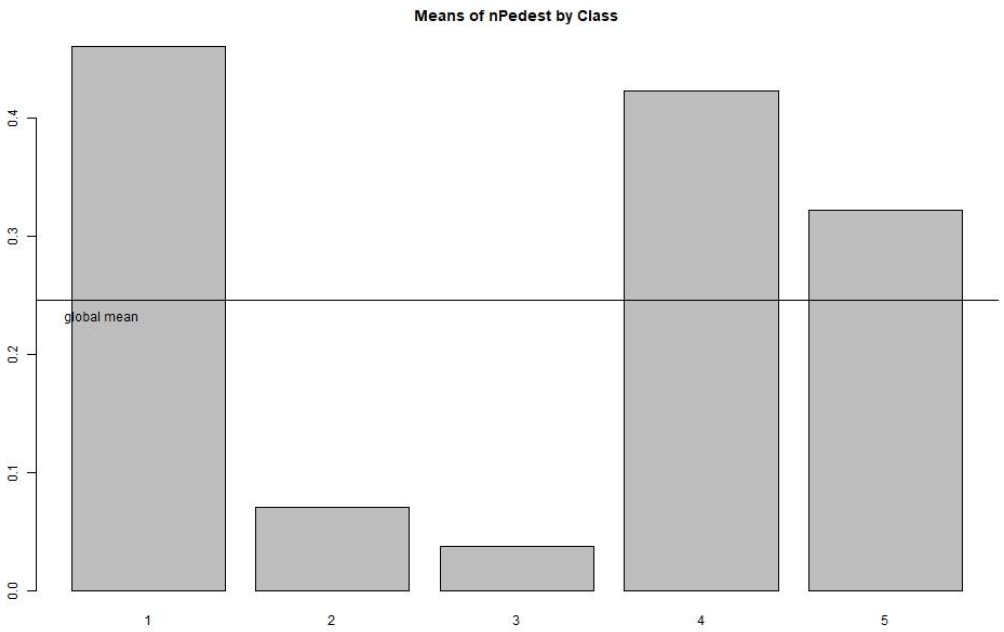


Image 9.27. Barplot of means of the nPedest variable for each class in the case of 5 clusters.

If we observe the previous profiling realized for the k=2 version, we can observe the similarity mentioned: cluster 1 is the one with all the accidents in urban zones, with less mortality and minor injured but with more number of pedestrians involved and cluster 2 is the one with all the road zones accidents, with more mortality and more minor injuries and with low value of pedestrians.

On the version with five clusters, there are other variables that have differences between clusters, for example, Intersection, DayGroup and Prov. For the intersection variable, cluster 4 has almost no accidents that have happened inside an intersection, cluster 5 has no accidents that happened in section and majoritary has accidents that happened inside intersections. The other clusters have approximately the same proportion of individuals with the different values for Intersection feature. A similar thing happens with DayGroup where cluster 4 has only weekend accidents, cluster 1 has much more accidents that happened on weekdays than on weekends and the rest of clusters have a similar proportion. Also in prov feature we can observe that cluster 2 has no Barcelona accidents and cluster 3 has almost only Barcelona accidents and no Tarragona, Girona and Lleida. The rest of the clusters have no difference between them in terms of prov.
These differences are not sufficient to differentiate the clusters.

(Reminder that you can extract this conclusions by running the Clustering and Profiling r script with a number of clusters equal to 5).

With this, we can conclude that with k=5 the profiling is poor, there are not enough differences between clusters, each one is very similar to at least another one, which means that 5 is not a good k value. (A good 5 value minimizes the distance of individuals inside the clusters and maximizes the differences between clusters). That similarity between two or more clusters helps us to see that maybe a minor number of clusters would be better.

10. Conclusions (PCA vs CLUSTERING)

Finally, we have to say that, although we firstly thought with the PCA analysis that there were in general very low variances and differences between distributions, and that those distributions more pronounced away from the center were mainly because of outliers, at the end we can extract surprisingly similar conclusions between the PCA analysis and the clustering/profiling.

First off, if we look again at the factor map of individuals for the PCA1 + PC2, the two principal behaviors on the samples that we observe, one that seems to be strongly correlated with the PCA1, and another one that is both correlated with PCA1 and PCA2, seem to correlate with the two clusters we've analyzed.

Also, among the things we've concluded in the PCA analysis, we've said that: Accidents that happen in big populated regions (like Barcelonès, or Tarragonès), tend to be associated with a high number of pedestrians, and rural regions tend to be associated with a small number of pedestrians (or 0). When the Zone where the accident happens is Urban, there tends to be more pedestrians and be less serious than when it happens on a Road, and when an accident happens on roads that have a high maximum speed limit, there usually are more units implicated and be more serious than when the speed limit is lower. These differences between modalities seem to totally correlate with the two clusters we have. On one hand, we have a cluster with not so serious accidents, a high number of pedestrians involved in the accidents, accidents mainly involved in Urban and populated Regions (like Barcelones) and with roads that have low speed limits; and on the other hand, another cluster with these opposite properties.

With that said, measures we could apply to try to avoid accidents in Catalonia could be:

- On accidents that happen on Urban Roads, populated cities, and on roads with low speed limits, we should focus on trying to avoid runovers, and protect pedestrians more.

- To avoid accidents that happen in non-urban roads, and with roads that have a high speed limit, we should add more section radars, because these are the most serious accidents. Also, because these accidents usually happen with a lot of units implicated, we should increase restrictions when there's a lot of traffic.

- We also should increase the minimum distance between a car and a bicycle, because when bicycles are involved in an accident it is usually because a car hit a bicycle.
- We should take extreme safety measures when the surface is icy (because this type of surface is correlated with mortal accidents) or snowy (because this type of accident is correlated with run overs).
- We also should take extreme safety measures when the surface is wet, slippery, or flooded, to avoid rollovers, hitting obstacles or having road exits.

11. Working plan

This section has the final working plan we have followed during this project. Also, we will compare this section with the initial working plan in order to see the difference between the original working plan and the real one. First of all, we are going to take a look at the final divisions of tasks, which do not have any modifications from the original one. Secondly, we will see the original designed Gantt chart with the one we have really followed. Thirdly, we will take a look at the original contingency plan. Finally, we will discuss deviances of final scheduling with respect to the original designed one and discuss risks avoided by the initial contention plan.

Final divisions of tasks

As we have already commented before, there was not any modification in the divisions of tasks. In image 11.1, we can see the original and final division of tasks.

Participant	Daniel	David	Joel	Marina	Simón
D3					
Identify tasks			X	x	x
Task assignments	x	x	X	x	x
Gantt's diagram			X	x	
Risk contingency plan			x	x	X
Metadata file			x	X	
Declaring factors		X			x
Initial descriptive statistics	x				X
Treating missings	X	x			
Additional descriptive statistics		x			X
Document D3	x	x	x	X	x
D4					
Cover + index		X	x		
Formal description of data structure and metadada			x	X	
Motivation of the work	X				x
Data source presentation			X	x	
Complete data mining process performed		x	X		
Detailed description of preprocessing and data preparation	x	X			
Review / modify basic statistical descriptive analysis				x	X
PCA analysis for numerical variables		X		x	
Hirerchical clustering on original data	X				x
Profiling clusters			x	X	
Global discussion and general conclusions of the work	x	X			
Review working plan			X		x
Prepare & submit R scripts	x			X	
Prepare oral presentation	x	x			X

Coordinator
X

Image 11.1. Final division of tasks table.

Original Gantt chart

Activity name	Initial date	End date
D3		
Identify tasks	16/09/2022	16/09/2022
Task assignments	16/09/2022	16/09/2022
Gantt's diagram	16/09/2022	23/09/2022
Risk contingency plan	23/09/2022	24/09/2022
Metadata file	16/09/2022	21/09/2022
Declaring factors	16/09/2022	21/09/2022
Initial descriptive statistics	21/09/2022	23/09/2022
Treating missings	21/09/2022	23/09/2022
Additional descriptive statistics	25/09/2022	28/09/2022
Document D3	25/09/2022	29/09/2022
D4 and final submission		
Cover + index	17/10/2022	17/10/2022
Formal description of data structure and metadada	30/09/2022	6/10/2022
Motivation of the work	30/09/2022	6/10/2022
Data source presentation	30/09/2022	6/10/2022
Complete data mining process performed	30/09/2022	6/10/2022
Detailed description of preprocessing and data preparation	30/09/2022	12/10/2022
Review / modify basic statistical descriptive analysis	30/09/2022	12/10/2022
PCA analysis for numerical variables	30/09/2022	12/10/2022
Hirerchical clustering on original data	8/10/2022	13/10/2022
Profiling clusters	14/10/2022	16/10/2022
Global discussion and general conclusions of the work	17/10/2022	19/10/2022
Review working plan	17/10/2022	17/10/2022
Prepare & submit R scripts	17/10/2022	18/10/2022
Prepare oral presentation	17/10/2022	21/10/2022

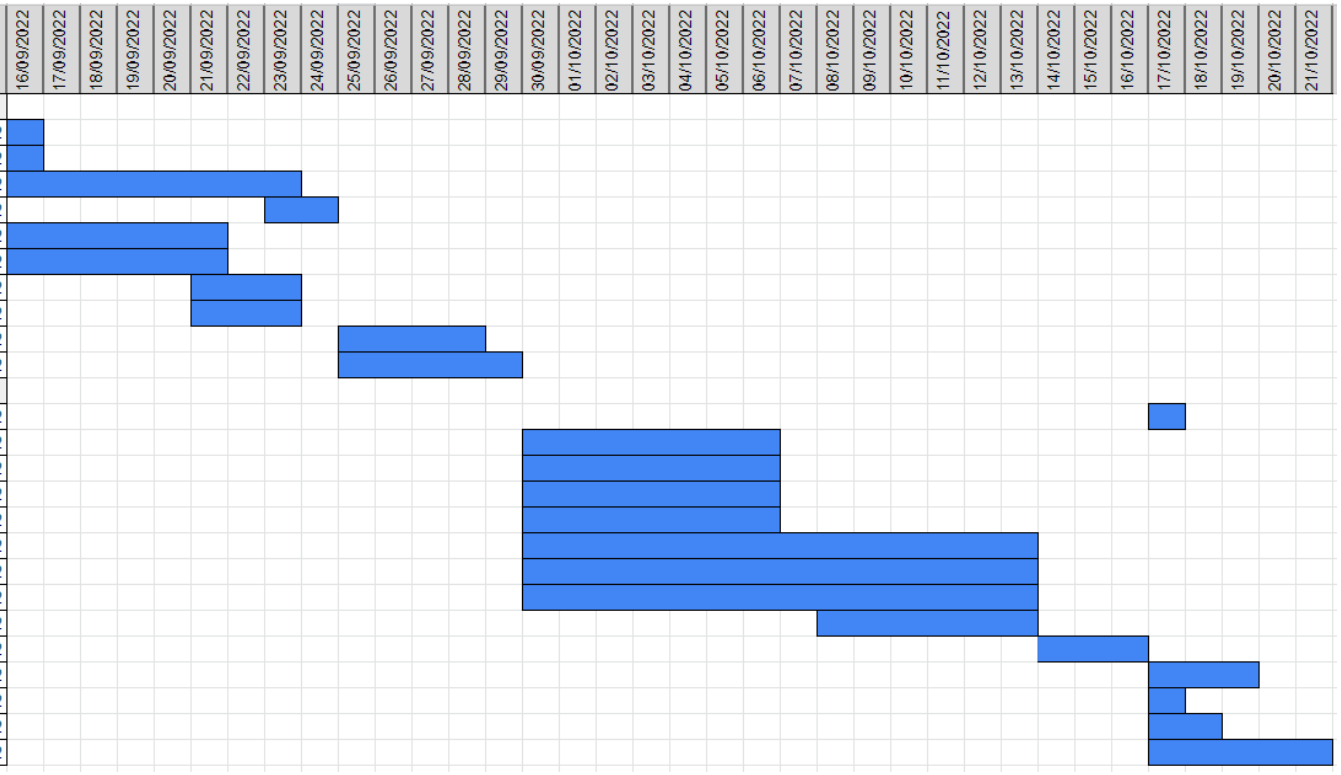


Image 11.2. Original Gantt chart

Final Gantt chart

Activity name	Initial date	End date
D3		
Identify tasks	16/09/2022	16/09/2022
Task assignments	16/09/2022	16/09/2022
Gantt's diagram	16/09/2022	23/09/2022
Risk contingency plan	23/09/2022	24/09/2022
Metadata file	16/09/2022	21/09/2022
Declaring factors	16/09/2022	21/09/2022
Initial descriptive statistics	21/09/2022	23/09/2022
Treating missings	21/09/2022	23/09/2022
Additional descriptive statistics	25/09/2022	28/09/2022
Document D3	25/09/2022	29/09/2022
D4 and final submission		
Cover + index	17/10/2022	17/10/2022
Formal description of data structure and metadada	30/09/2022	6/10/2022
Motivation of the work	30/09/2022	3/10/2022
Data source presentation	03/10/2022	6/10/2022
Complete data mining process performed	30/09/2022	16/10/2022
Detailed description of preprocessing and data preparation	16/10/2022	17/10/2022
Review / modify basic statistical descriptive analysis	14/10/2022	18/10/2022
PCA analysis for numerical variables	30/09/2022	16/10/2022
Hirerchical clustering on original data	7/10/2022	15/10/2022
Profiling clusters	15/10/2022	16/10/2022
Global discussion and general conclusions of the work	17/10/2022	19/10/2022
Review working plan	18/10/2022	18/10/2022
Prepare & submit R scripts	19/10/2022	19/10/2022
Prepare oral presentation	17/10/2022	21/10/2022

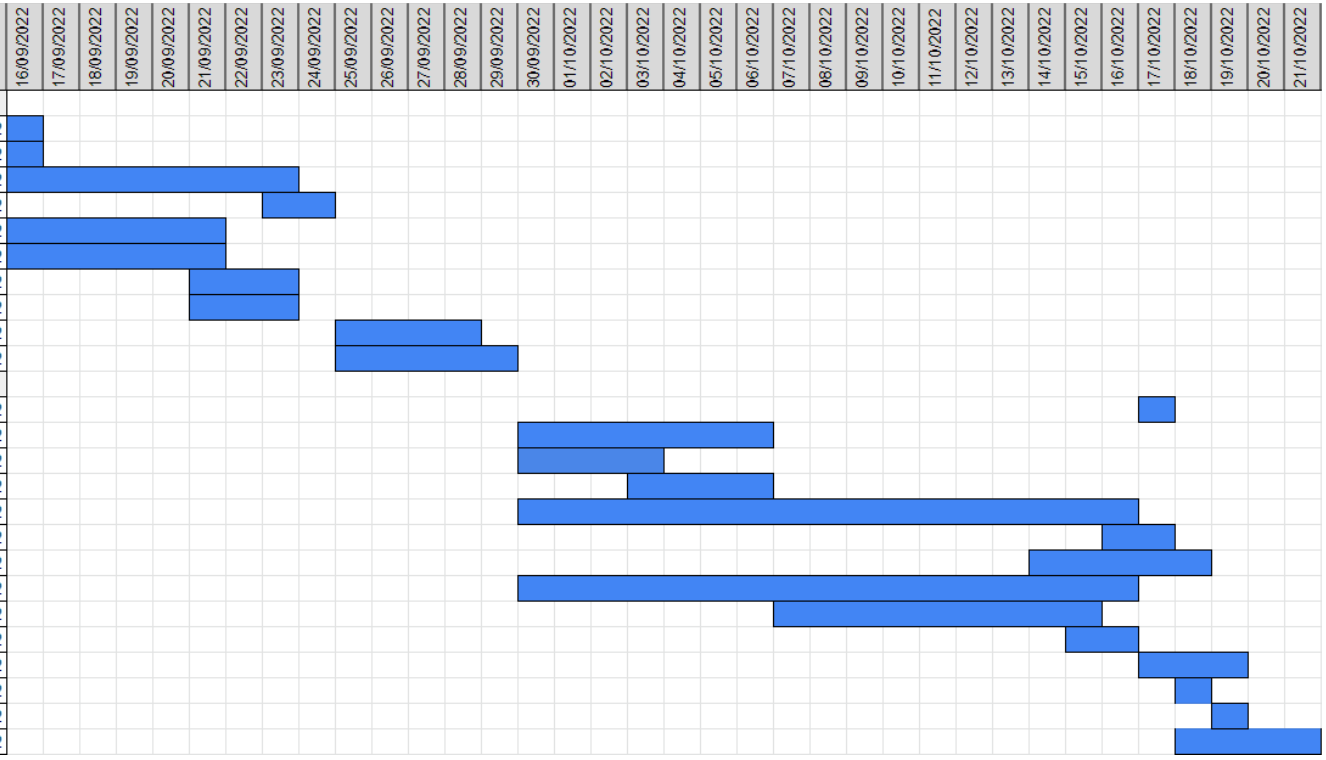


Image 11.3. Final Gantt chart

Contingency plan

Risk	How to prevent	How to manage
Lockdown	Store all the documents and scripts on the cloud	Work remotely on the cloud documents. Stay in contact with the group and teachers.
Bad time estimation	Good planification (establishing priority and dependencies among tasks), and periodical revisions	Taking into account priorities and reducing efforts on non-essential tasks
Group member leaves	Tasks are divided up into more than one member	Replanification of work distribution among members
Cannot access the cloud	Having preventive local backups of documents and scripts (each member)	Meeting in some place and merge the work of each member
Bad preprocessing performed	Follow the instructions and advice of the teachers. Use the scripts provided as a base from which to start and pay close attention to possible errors. Document and justify all steps.	Take a look at the documentation and justification of the steps taken in order to identify where an error has been done. Repeat the preprocessing steps looking for errors.

Table 11.1. Contingency plan table

Conclusions

Although we have planned the Gantt chart at the beginning of the project, we can see in image 11.3 that it is very different from the original designed one (see image 11.2). One of the reasons why the real Gantt chart is not as same as the ideal one, is because when we designed the original one we did not know if there were dependencies between some tasks. For example, the *Complete data mining process performed* task was scheduled before the finalization of some of the data mining tasks, so we could not finish writing this section without finishing all the other tasks that are part of this data mining process. The other reason is that it is hard to estimate the effort of each task and when the members responsible for the tasks have enough time to finish them or start them in the time scheduled. Actually, it is common for projects to have some deviations on schedule terms.

On the other hand, fortunately, we did not have any unexpected problems during the project. However, we did not follow the prevention plan of bad time estimation, which could have been a drastic problem.

ANNEX

Table that contains the original and the new names of the different regions of Catalonia:

Original Name	Short Name
Alt Camp	Acamp
Alt Emporda	AEmporda
Alt Penedes	APenedes
Alt Urgell	AUrgell
Alta Ribagorca	Ribagorca
Anoia	Anoia
Bages	Bages
Baix Camp	BCamp
Baix Ebre	BEbre
Baix Emporda	BEmporda
Baix Llobregat	BLlobregat
Baix Penedes	BPenedes
Barcelones	Barcelones
Bergueda	Bergueda
Cerdanya	Cerdanya
Conca de Barbera	Conca
Garraf	Garraf
Garrigues	Garrigues
Garrotxa	Garrotxa

Girones	Girones
Maresme	Maresme
Moianes	Moianes
Montsia	Montsia
Noguera	Noguera
Osona	Osona
Pallars Jussa	PJussa
Pallars Sobira	PSobira
Pla d’Urgell	PUrgell
Pla de l’Estany	PEstany
Priorat	Priorat
Ribera d’Ebre	Ribera
Ripolles	Ripolles
Segarra	Segarra
Segria	Segria
Selva	Selva
Solsones	Solsones
Tarragones	Tarragones
Terra Alta	TerraAlta
Urgell	Urgell
Val d’Aran	VAran
Valles Occidental	VOccidental
Valles Oriental	VOriental