Catalonian Road Accidents 2010-21

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

Presentation: 21/10/2022

Outline

1. Goals and Database overview
2. Data Mining process
3. Descriptive analysis
4. Preprocessing
5. PCA and Clustering
6. Profiling
7. Conclusions
8. Task Scheduling

Topics and Goals

Topics

- Factors that can affect accidents
  - Velocity
  - Weather
  - State of the road
- Consequences of the accident
  - Number of injuries
  - Entities involved

Goals

- What kinds of situations facilitate accidents
- Help the society studying accidents
- Accidents affects everyone

Database Overview

ORIGINAL DIMENSIONS

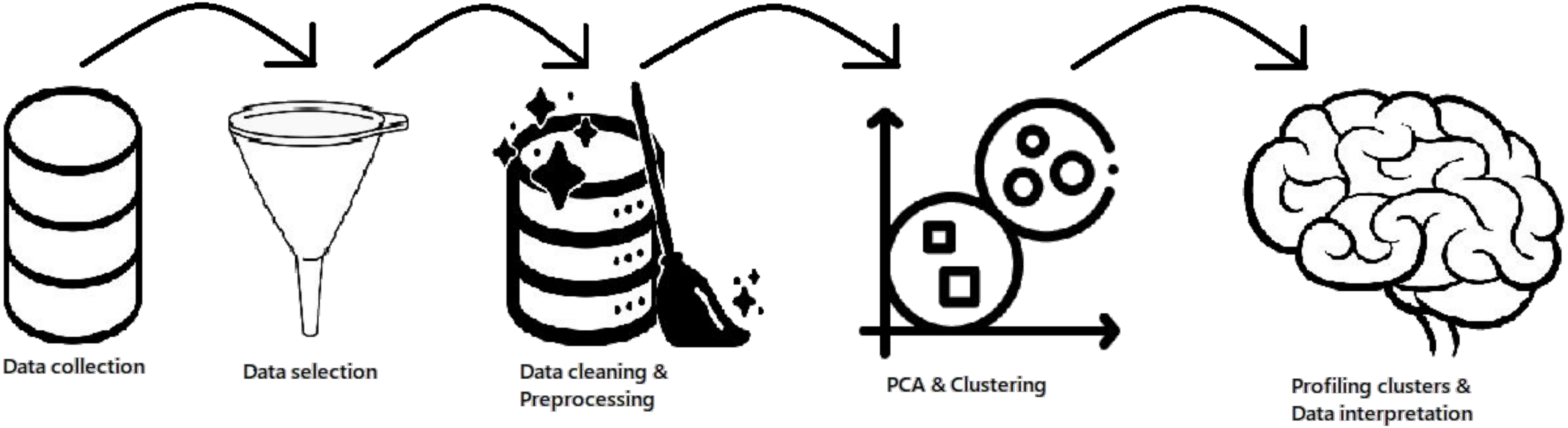
- 21.161 individuals
- 58 features

AFTER DATA SELECTION

- 5.000 individuals
- 23 features:
  - 7 numerical
  - 11 categorical
  - 5 boolean

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

Data Mining Process



Data collection

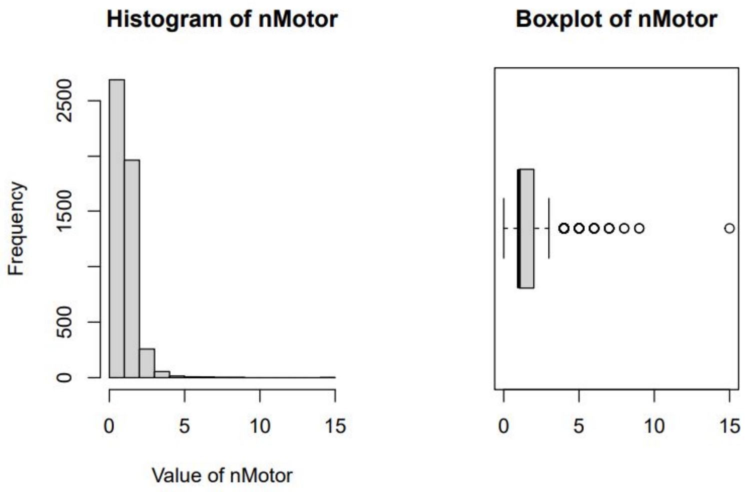
Data selection

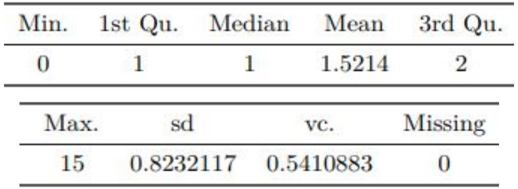
Data cleaning &
Preprocessing

PCA & Clustering

Profiling clusters &
Data interpretation

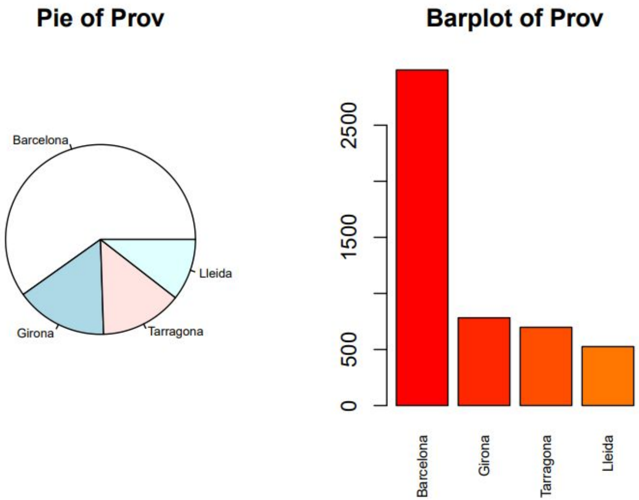
Descriptive Analysis

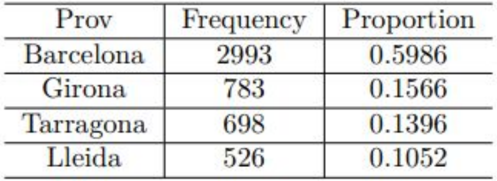




Min.	1st Qu.	Median	Mean	3rd Qu.
0	1	1	1.5214	2

Max.	sd	vc.	Missing
15	0.8232117	0.5410883	0





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

Preprocessing Steps (1)

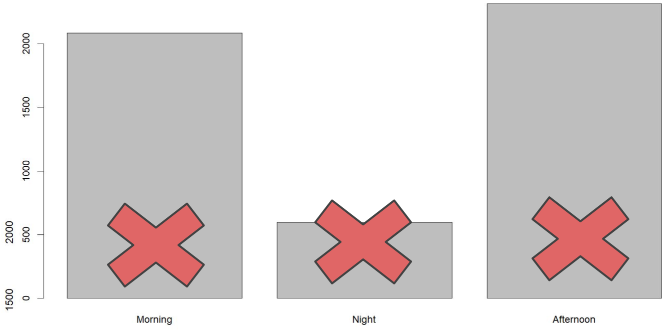
➔ **Factorization**, levels, sorting

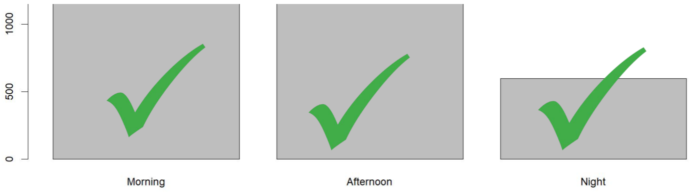
~~Character~~

Factor -> 17

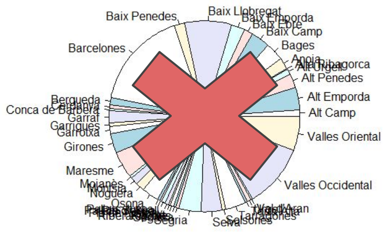
Date -> 1

Integer -> 7





➔ **Renaming**: Variables, levels



~~D_INFLUIT_CIRCULACIO~~ -> TrafficInf

~~D_INFLUIT_ESTAT_CLIMA~~ -> WeatherInf

etc...

~~Catalan~~ -> English

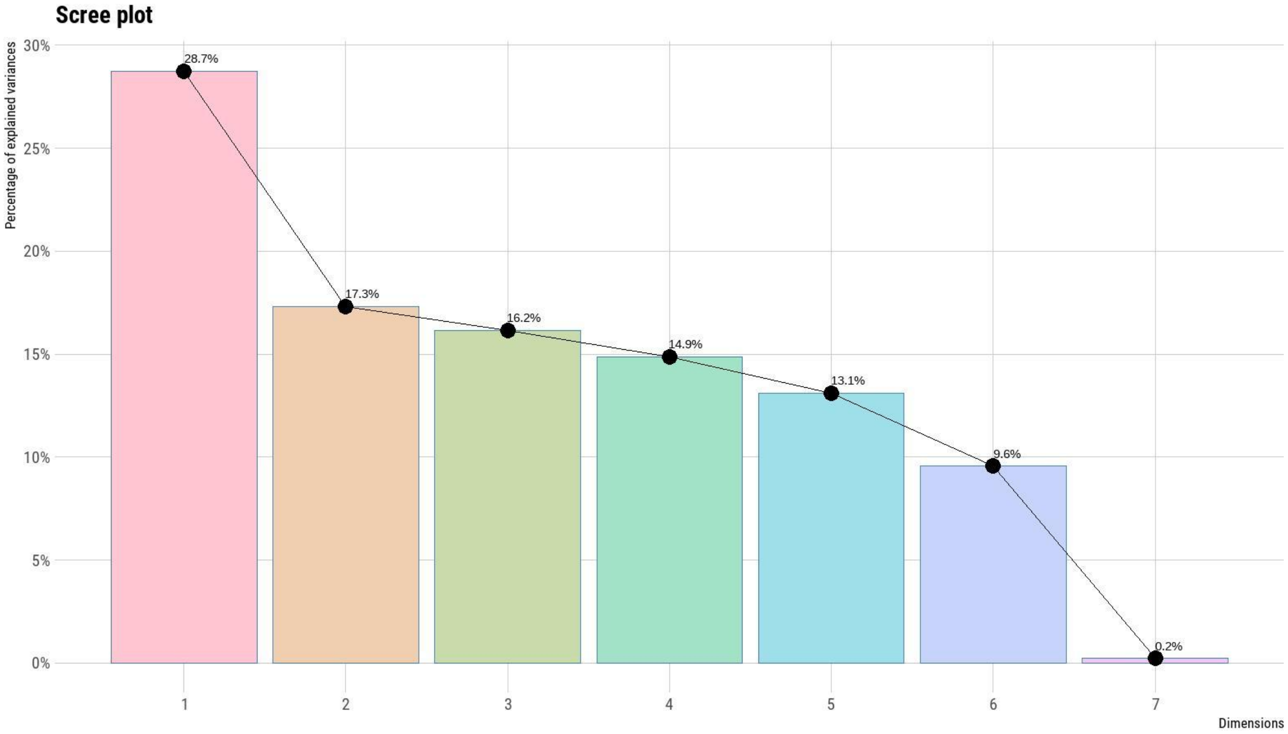
Preprocessing Steps (2)

➔ MISSINGS

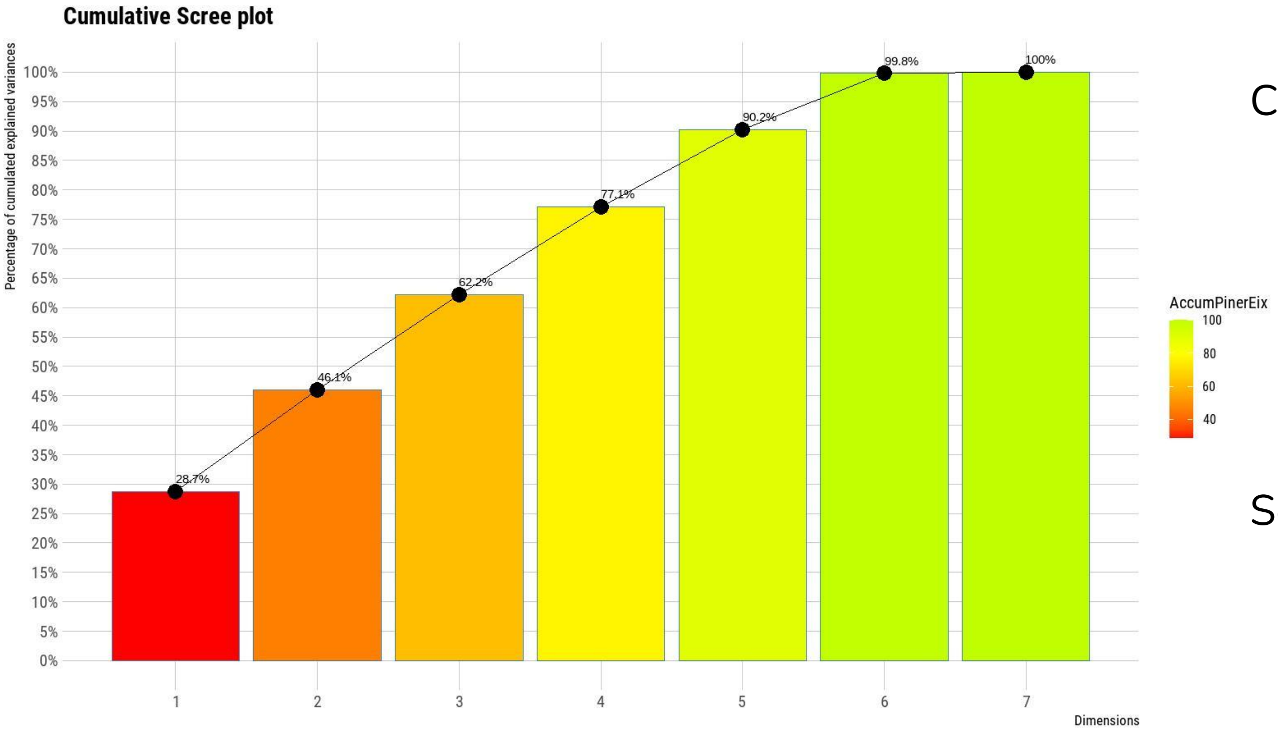
Random, no pattern

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%

PCA: Specifications



PCA: Specifications



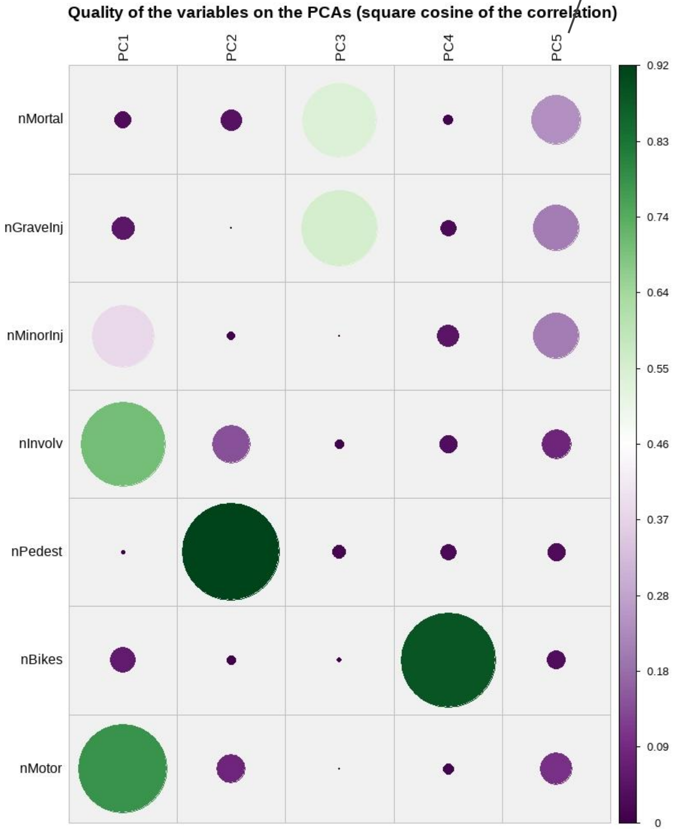
Criteria: keep 80% of the inertia

Selected dimensions: 1 to 4

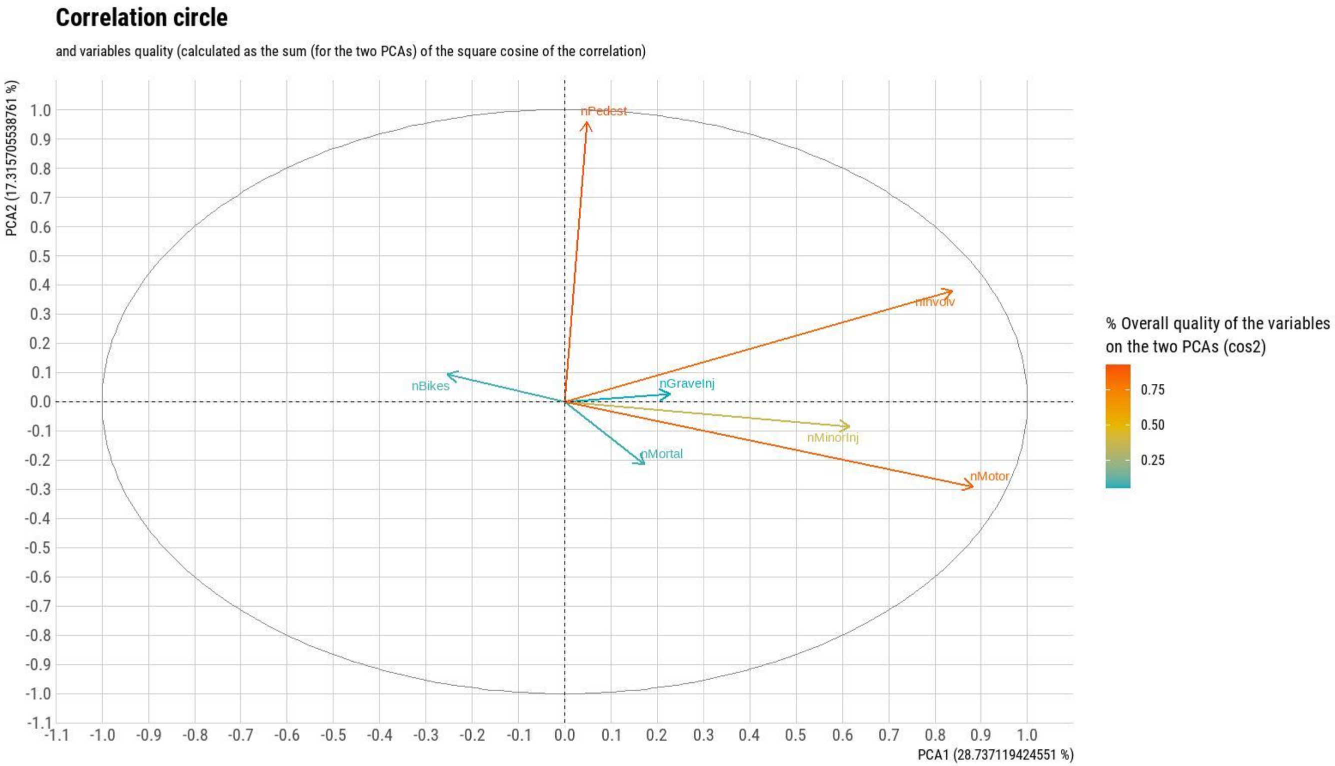
PCA: Specifications (2)

QUALITY -> ^2 cosine of **correlation**

[0, 1]



PCA: First Factorial Plane



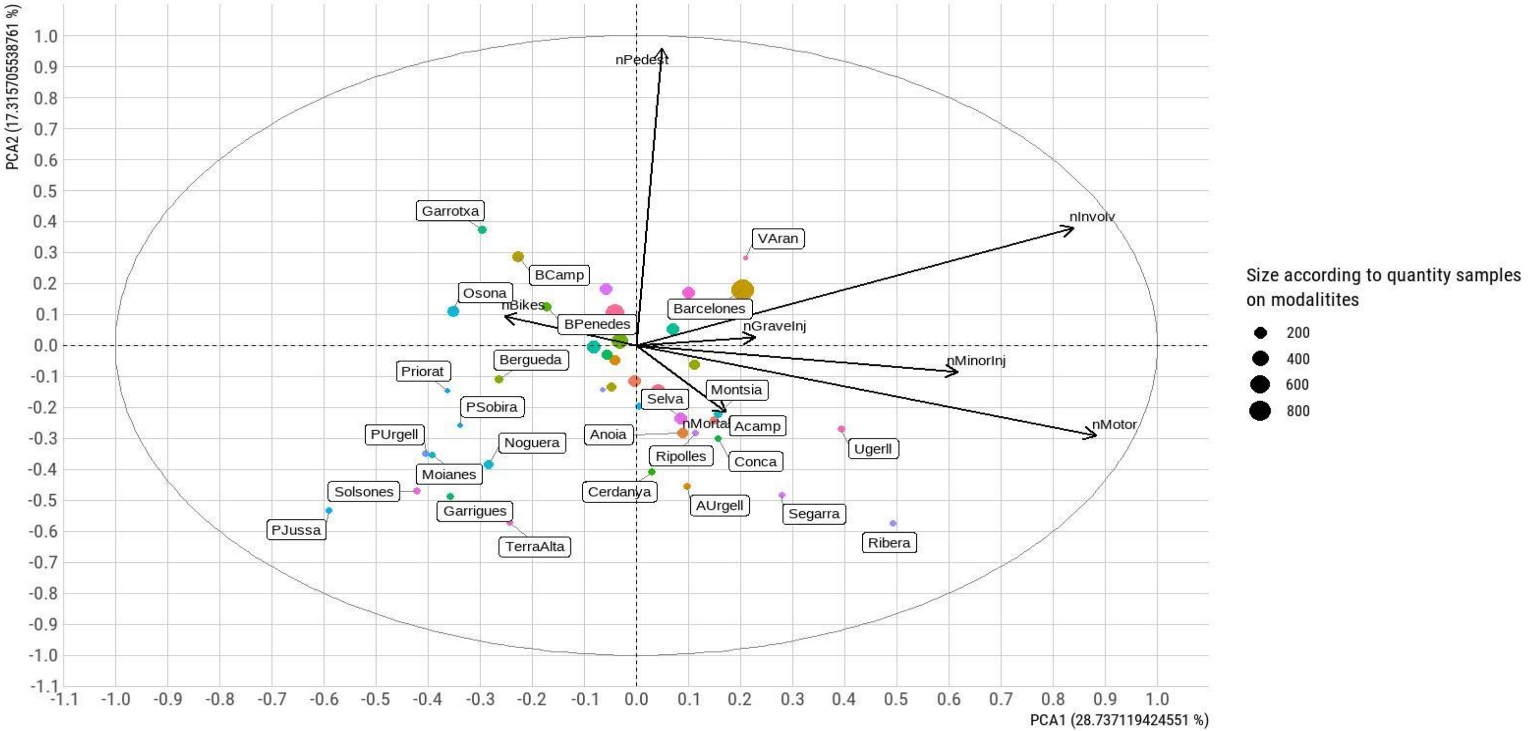
Correlation circle

and variables quality (calculated as the sum (for the two PCAs) of the square cosine of the correlation)

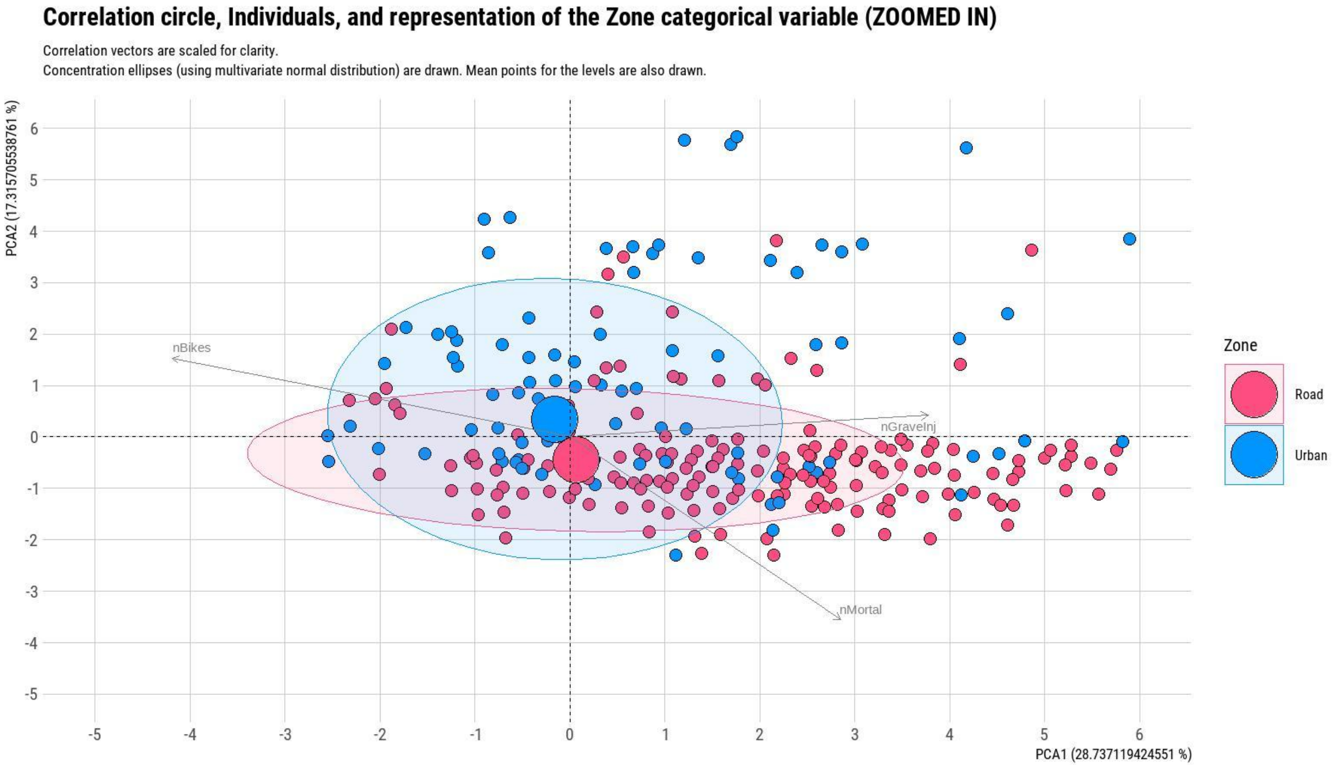
PCA: First Factorial Plane (2)

Correlation circle, and representation of all modalitites of Region cat. variable

and modality quantities as point sizes



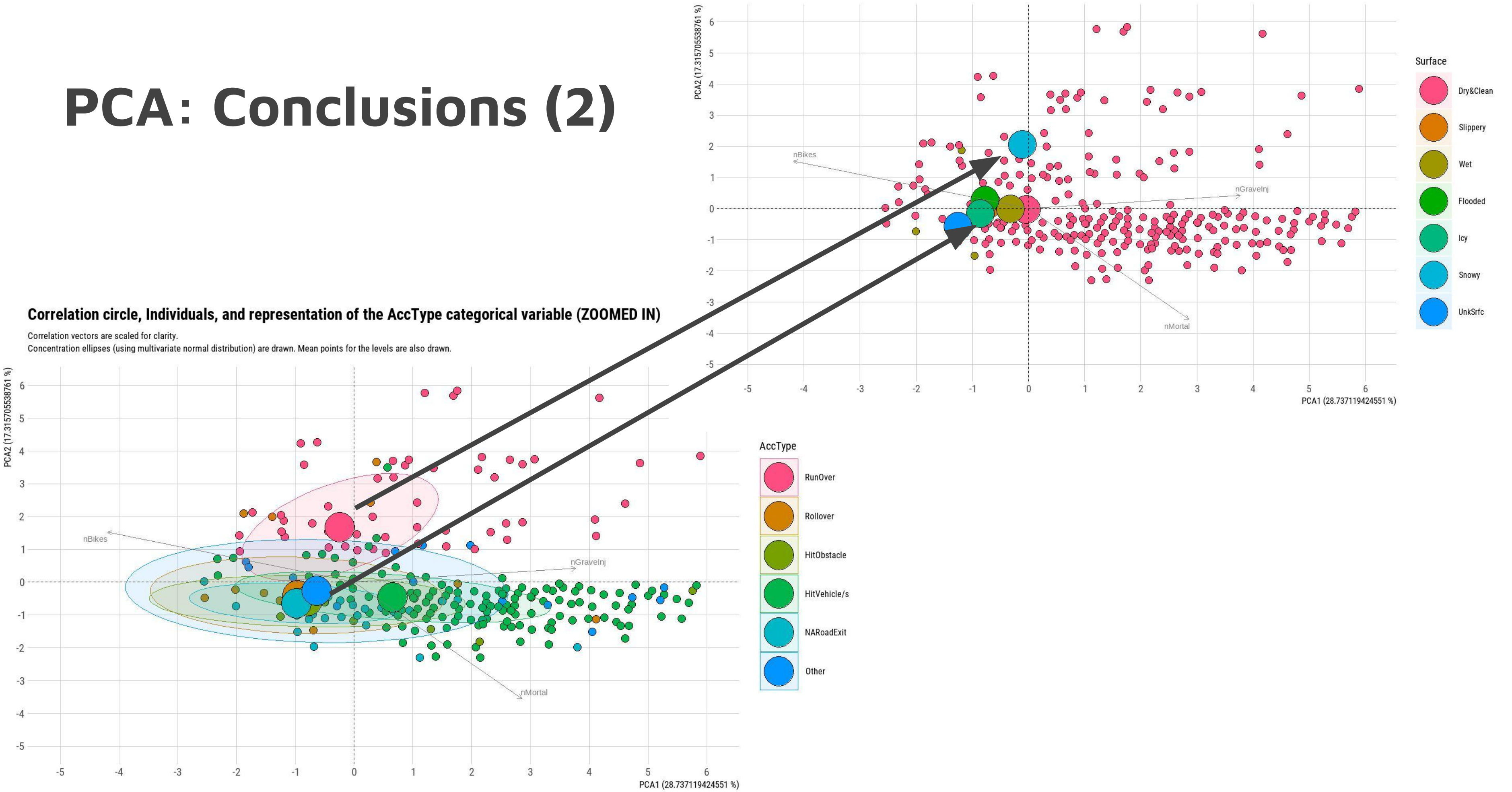
PCA: Conclusions (1)



Correlation circle, Individuals, and representation of the Zone categorical variable (ZOOMED IN)

Correlation vectors are scaled for clarity.
Concentration ellipses (using multivariate normal distribution) are drawn. Mean points for the levels are also drawn.

PCA: Conclusions (2)



Correlation circle, Individuals, and representation of the Surface categorical variable (ZOOMED IN)

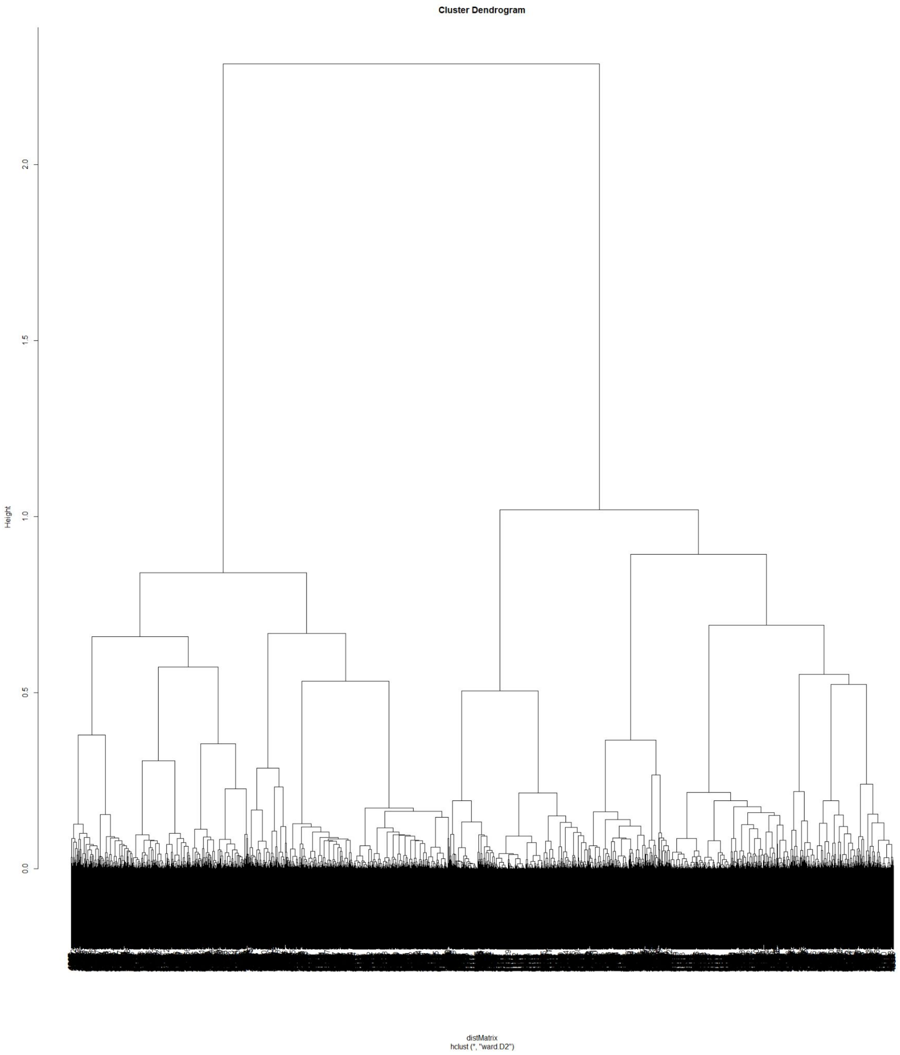
Correlation vectors are scaled for clarity.
Concentration ellipses (using multivariate normal distribution) are drawn. Mean points for the levels are also drawn.

Correlation circle, Individuals, and representation of the AccType categorical variable (ZOOMED IN)

Correlation vectors are scaled for clarity.
Concentration ellipses (using multivariate normal distribution) are drawn. Mean points for the levels are also drawn.

Clustering process

- Date decomposed into Year and Month
- Ward’s D2 method
- Gower mixed distance
- Minimize inter-class inertia loss



Cluster Number Discussion

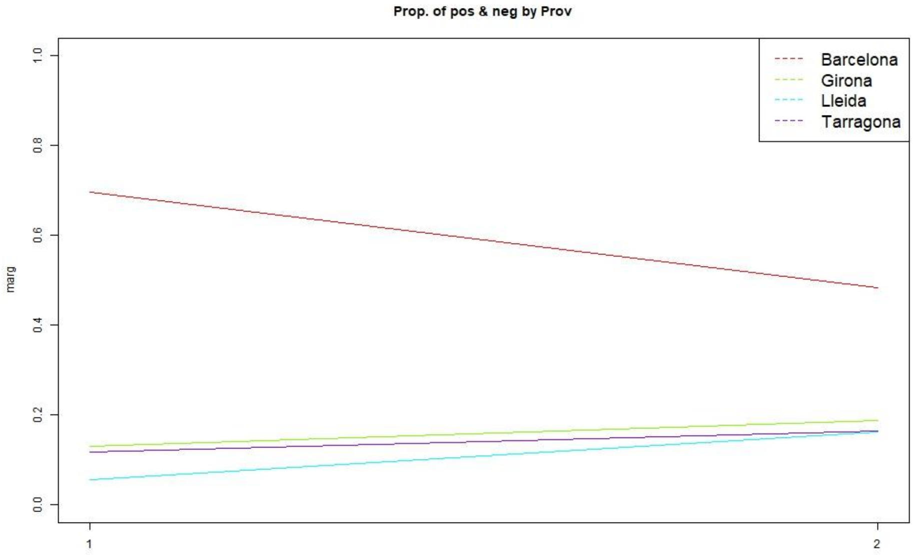
- KPI's
- PCA results
- Expert's opinion

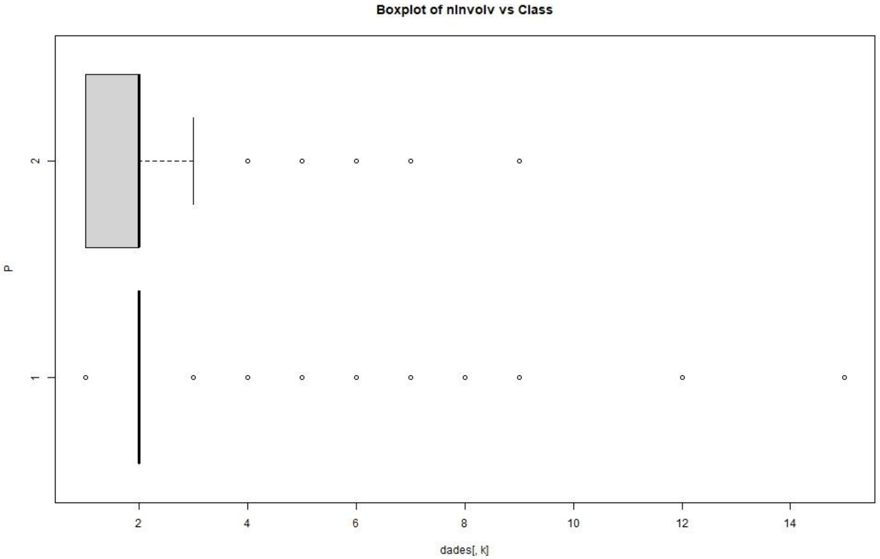
	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
Cindex	0.3788	0.3510	0.3312	0.3232	0.3465	0.3329	0.3285	0.3229	0.3162
Mcclain	0.7592	1.5408	1.9721	2.7166	3.0248	3.4005	4.3610	5.0876	5.3953
Silhouette	0.1582	0.1388	0.1684	0.1309	0.1574	0.1662	0.1391	0.1254	0.1407
Dunn	0.0815	0.0815	0.0815	0.0815	0.0915	0.0791	0.0458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

	Cluster 1	Cluster 2
Number of individuals	2696	2304

Number of clusters = 2

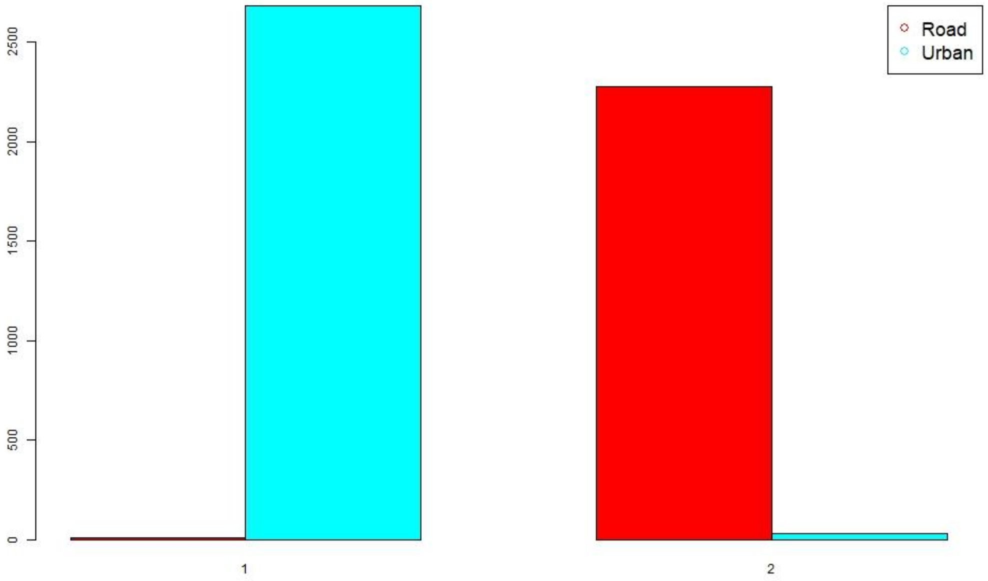
Class Interpretation Tools

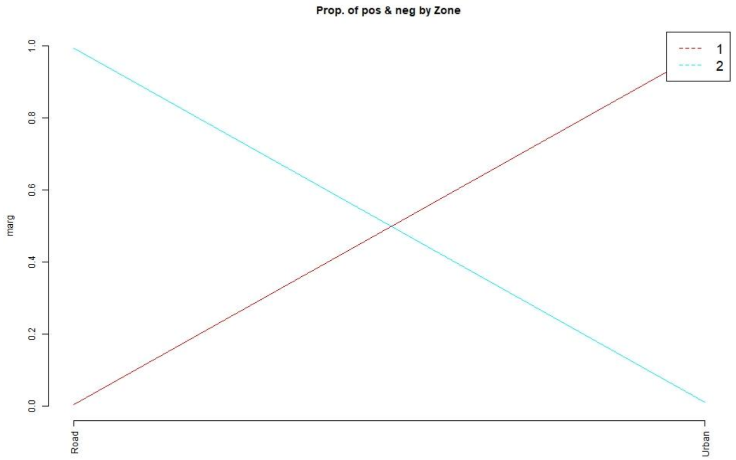


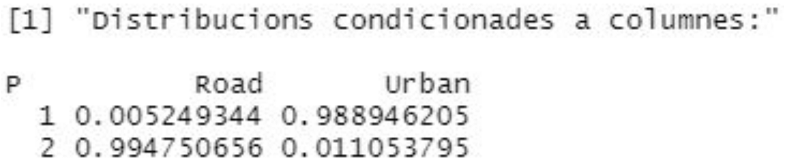


Profiling plots + statistics

Profiling graphs or numerical information about our clusters to be highlighted



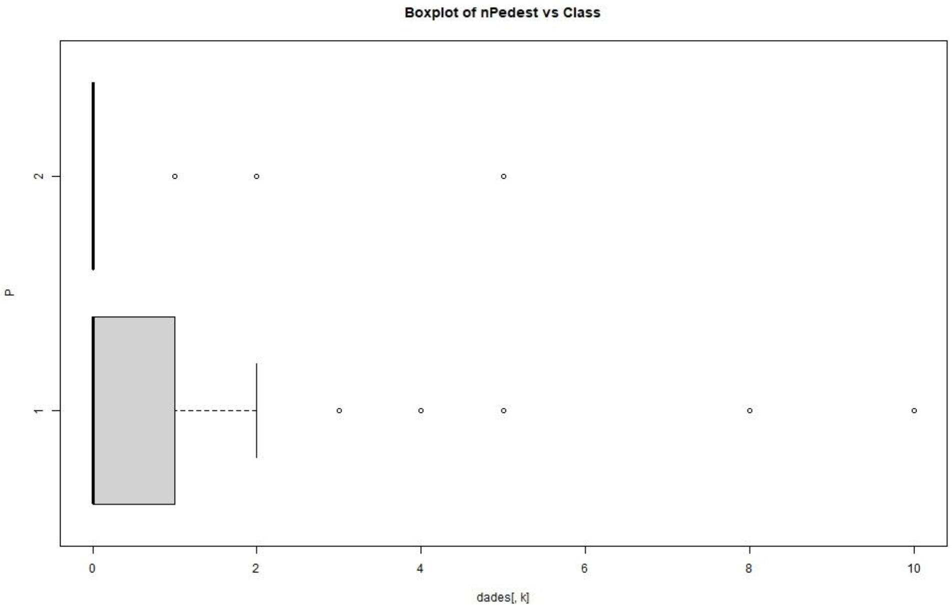


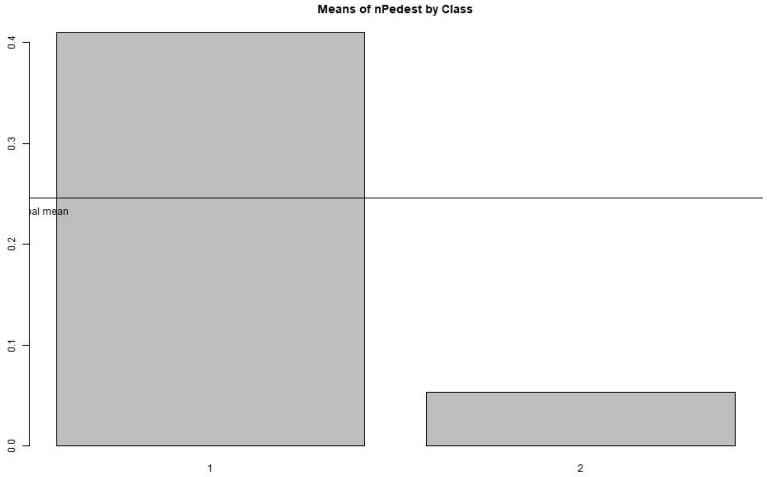


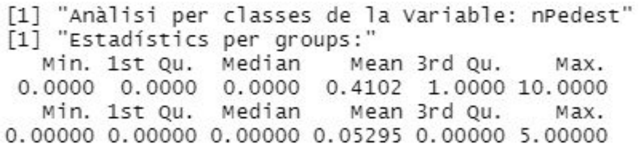
```
[1] "Distribucions condicionades a columnes:"

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

Profiling graphs or numerical information about our clusters to be highlighted







```
[1] "Anàlisi per classes de la Variable: nPedest"
[1] "Estadístics 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 
```

Final Class Profiling

CLUSTER 1:

Urban zones

Run overs

↓ mortality

↓ minor injured victims

↑ number of pedestrians

CLUSTER 2:

Road zones

Unknown exits from the road

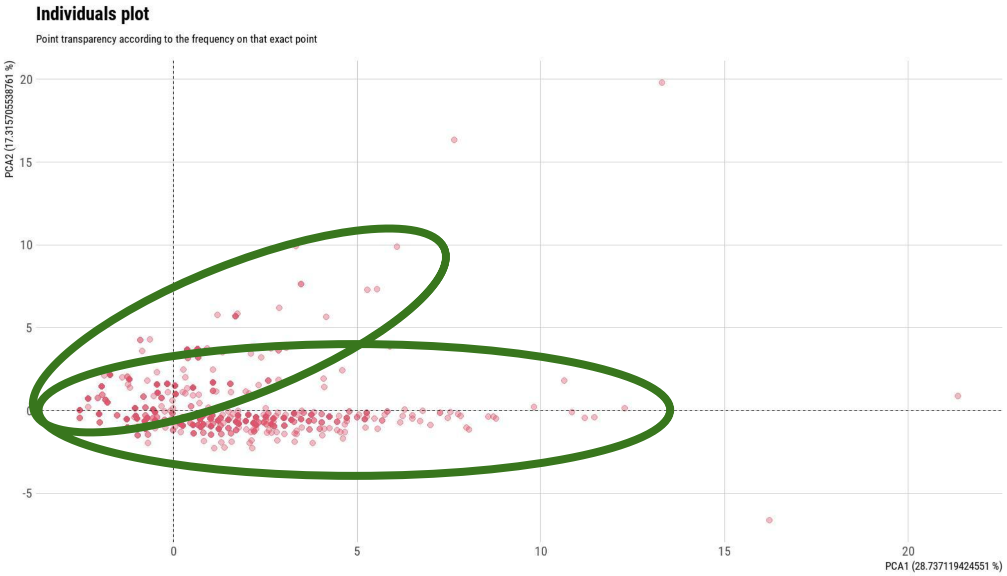
↑ mortality

↑ minor injured victims

↓ number of pedestrians

PCA and Clustering Conclusions

➔ SIMILAR



CLUSTER 1	CLUSTER 2
not so s. accidents	serious accidents
low speeds	high speeds
urban roads	rural roads
populated regions	rural regions
not so many un. imp.	many un. implicated

↓

similar conclusions on PCA!

General Conclusions

Urban, populated, low speeds → focus **protect pedestrians**

rural roads, high speed → most dangerous, most units implicated. **More sector radars**

car, bicycle → **Increase distance**

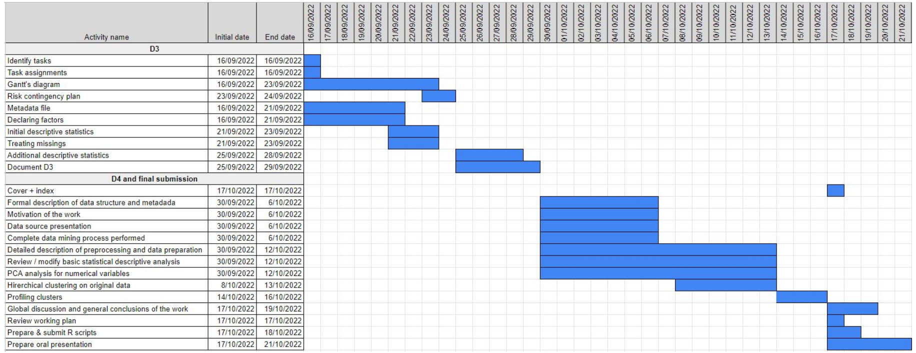
icy surface → Very mortal. Extreme safety measures

snowy surface → Related with Runovers

Wet, slippery, flooded → Related with Rollowers, hitting obstacles, road exits.

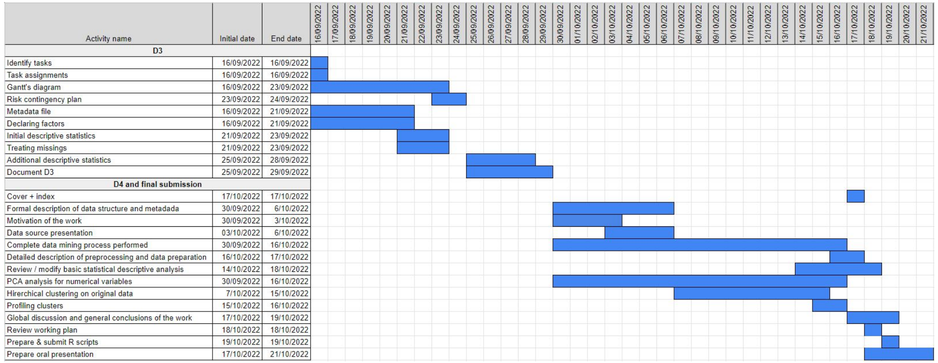
Task Scheduling

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

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

THANKS

Any questions?