US Accidents (2016-2021)

MVA Assignment D4

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

Table of contents

1. Introduction

Dataset

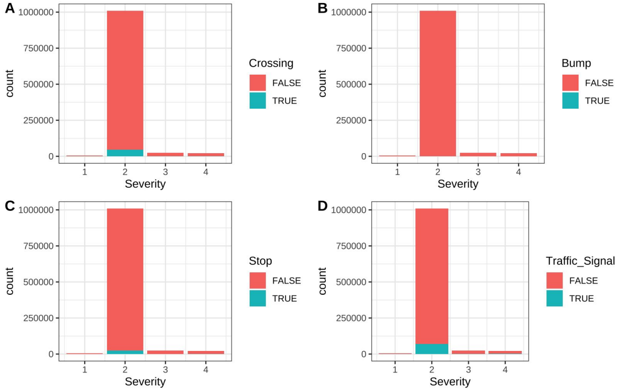
US Accidents that took place from 2016-2021

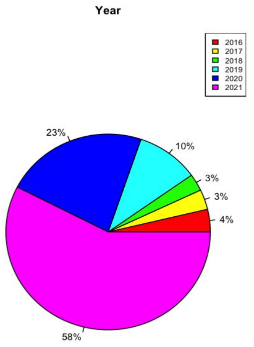
Data source

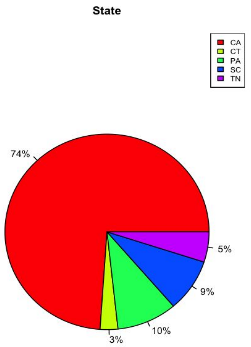
https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

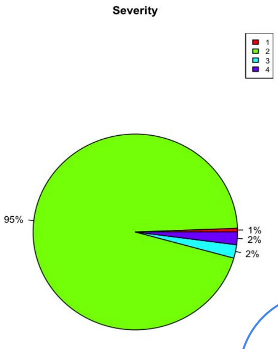
- Motivation
  - Tackle the problematic surrounding road accidents using data.
- General description
  - Large dataset
  - Qualitative, quantitative and binary variables are included.
  - Dataset includes information regarding:
    - Where the accident happened?
    - When did it happen?
    - How it affected the road?
    - What severity did it imply?
    - What where the conditions surrounding it?

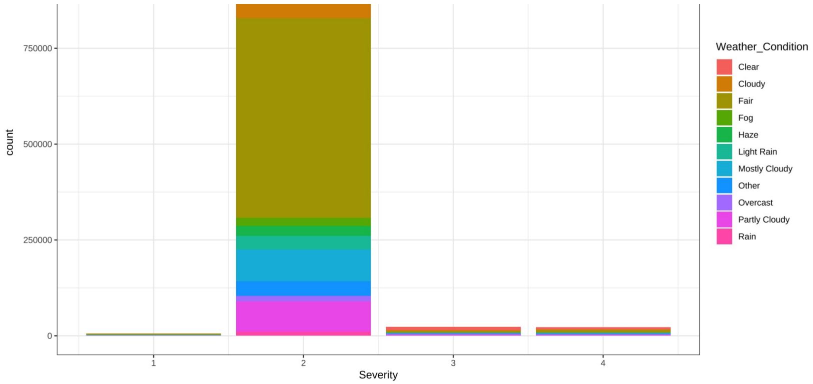
2. Univariate and Bivariate Descriptive Statistics











➢ Absence of traffic control indicators

➢ More frequent accident occurrence over years

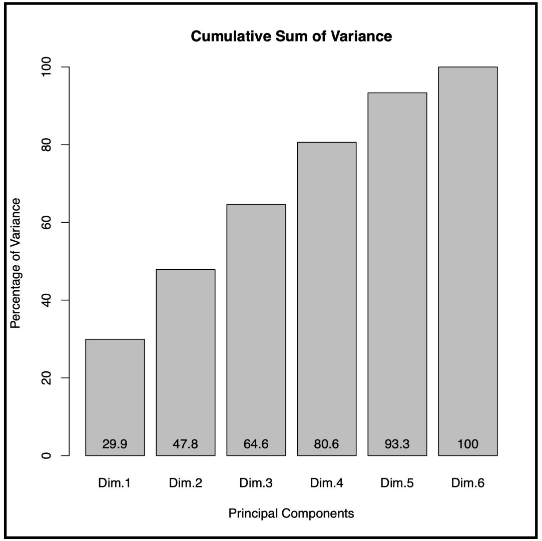
➢ Majority of accidents with Severity level 2

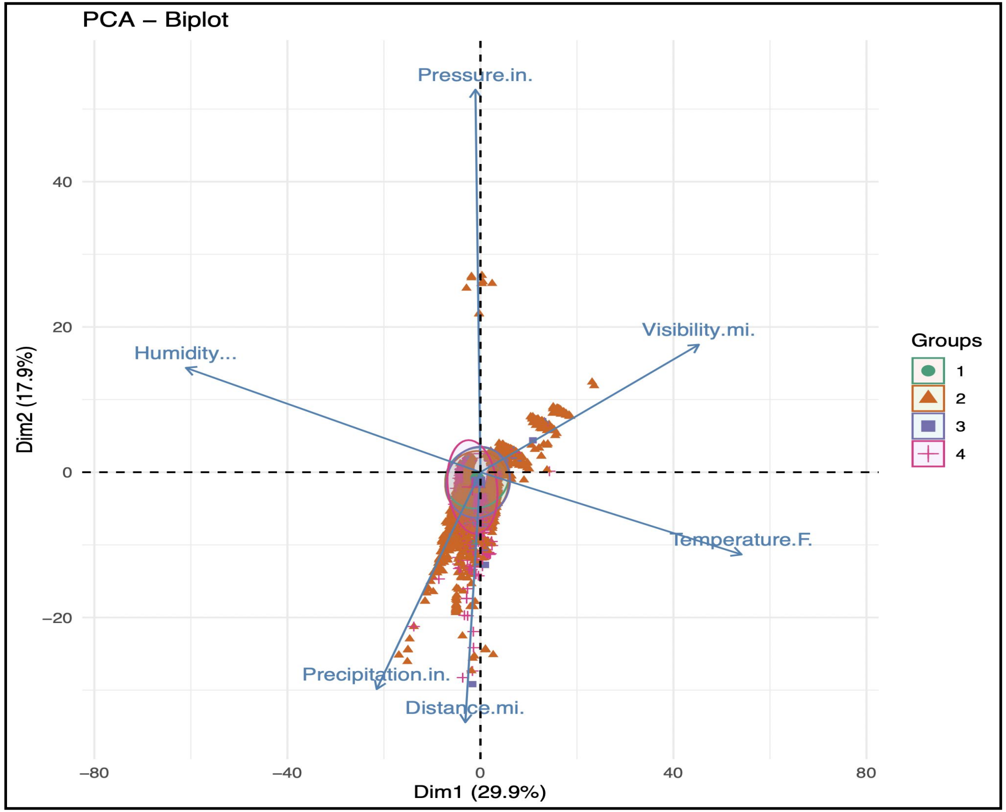
➢ Minor influence of weather conditions

➢ 74% of recorded US accidents are in the State of California

3. PCA

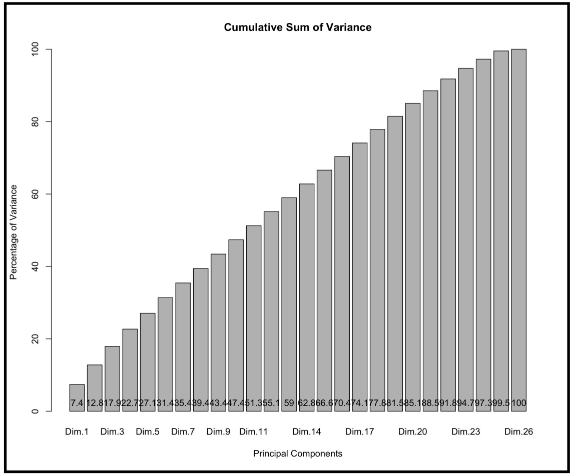
- Number of selected PCs
- General conclusion ➔ not very valid
- Weather condition associated variables

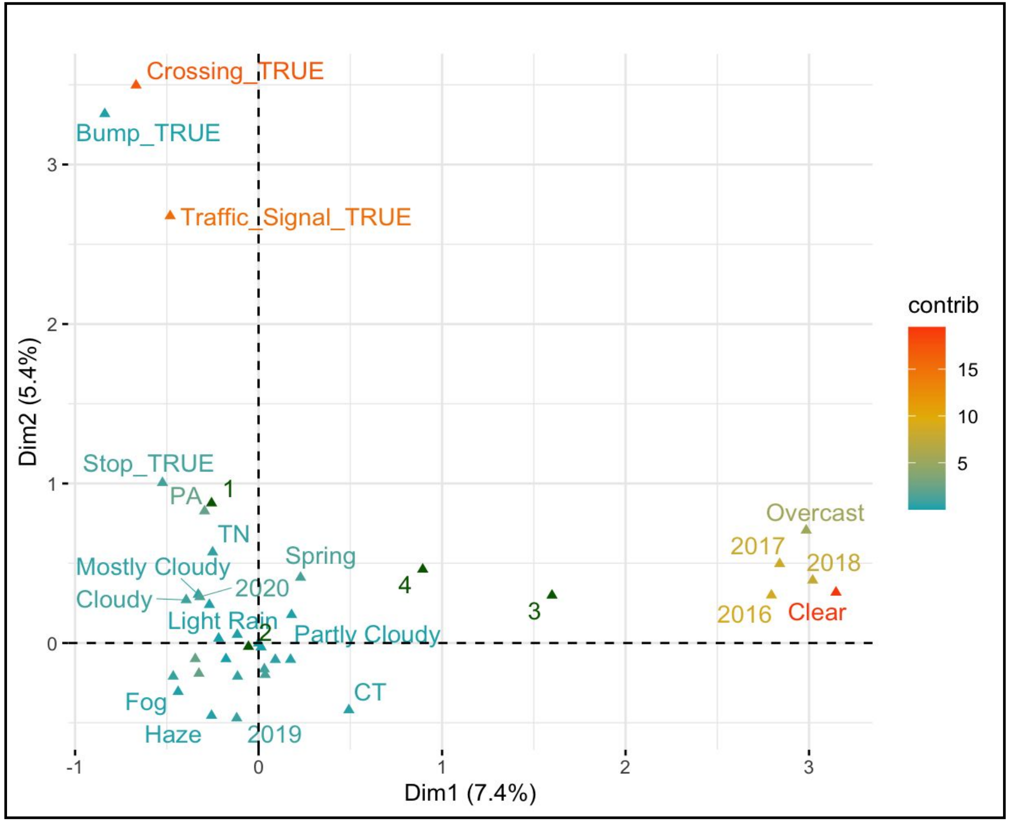




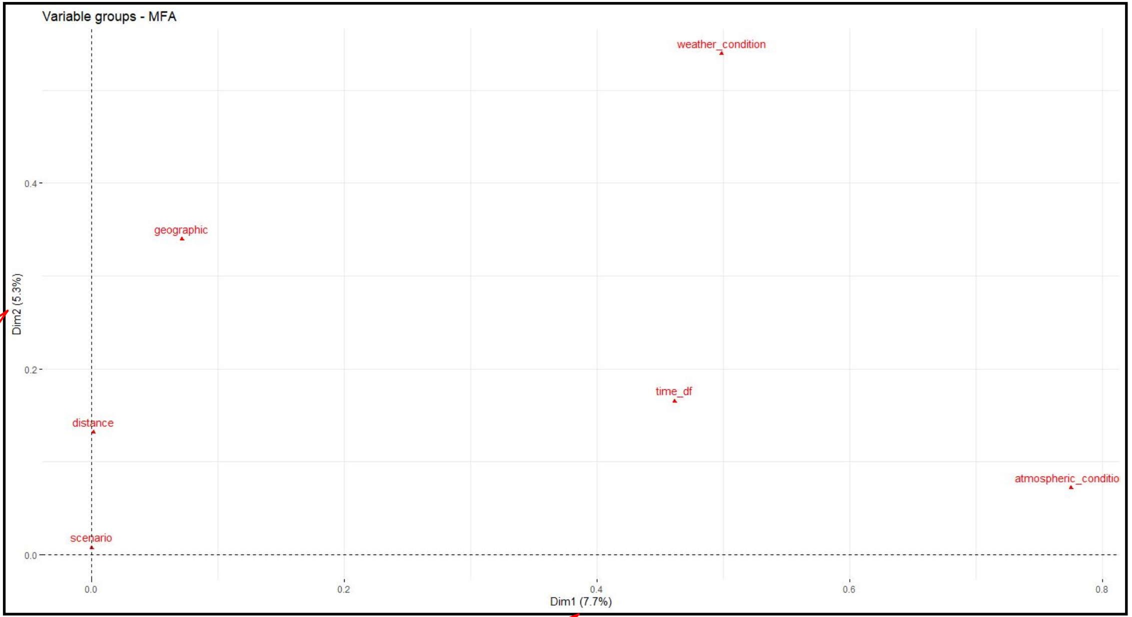
4. MCA

- Number of selected dimensions is 12
- General conclusion ➔ two clear groups are formed
- Weather_Condition (without precipitation) and Year explain dimension 1 and Road Elements for dimension 2



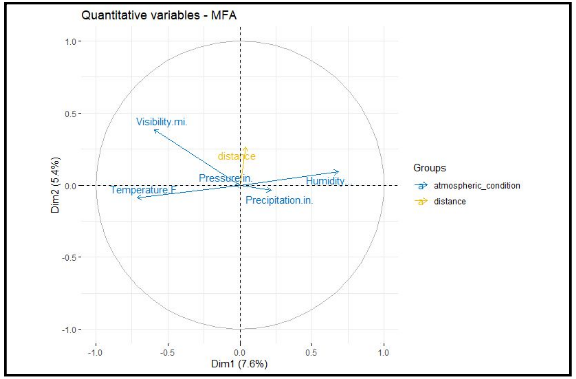


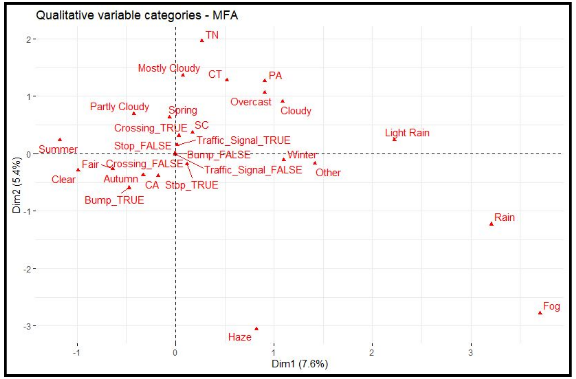
5. MFA


Variable groups - MFA

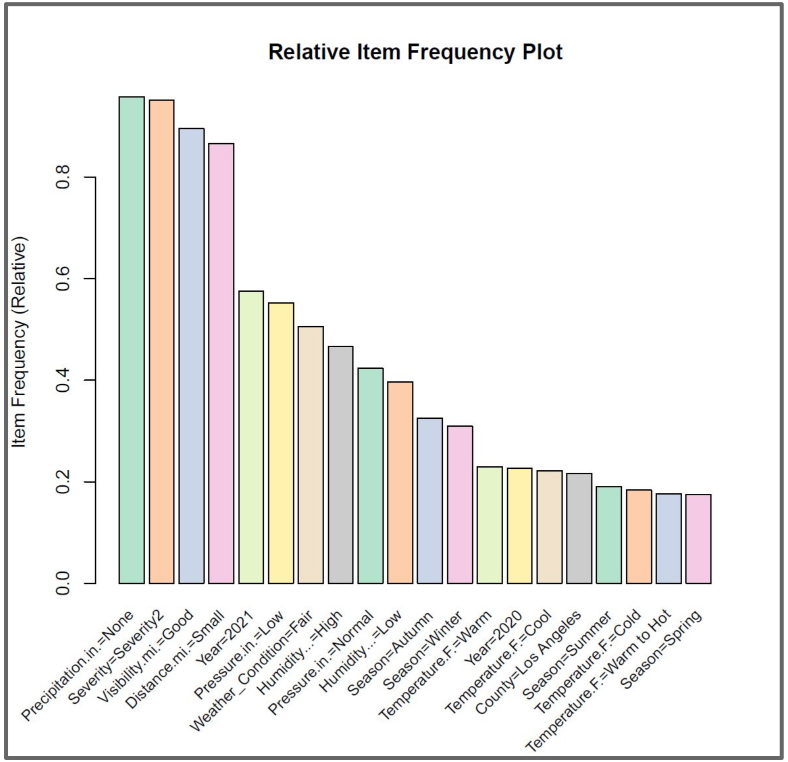
Weather condition + location

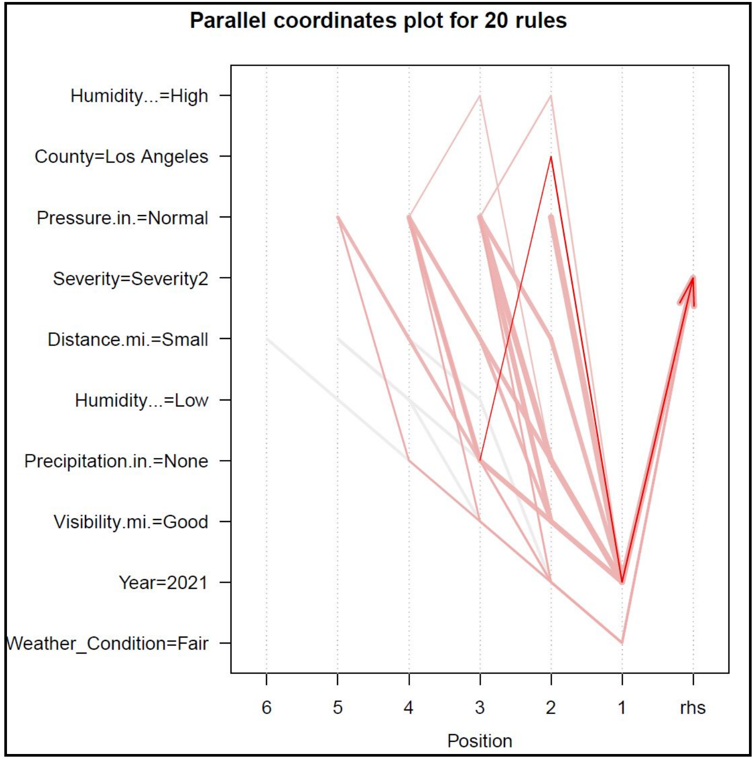
Weather condition + time


Quantitative variables - MFA


Qualitative variable categories - MFA

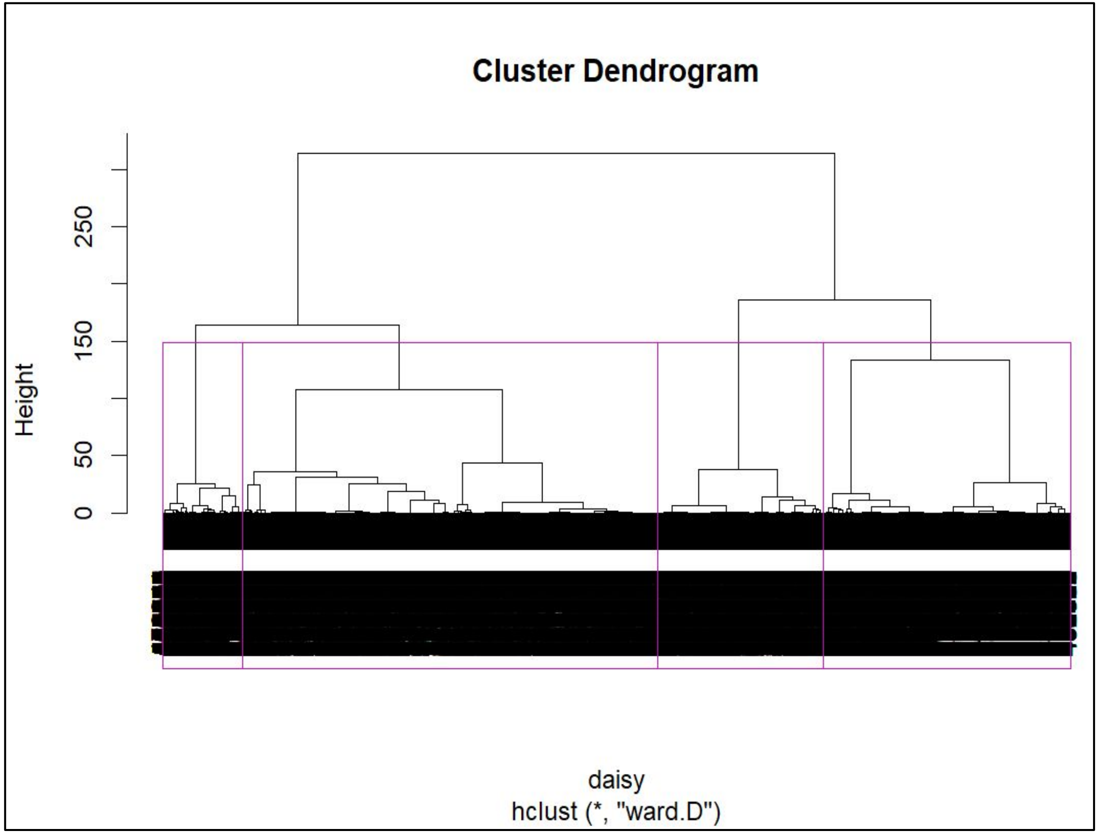
6. Association Rules Mining Analysis





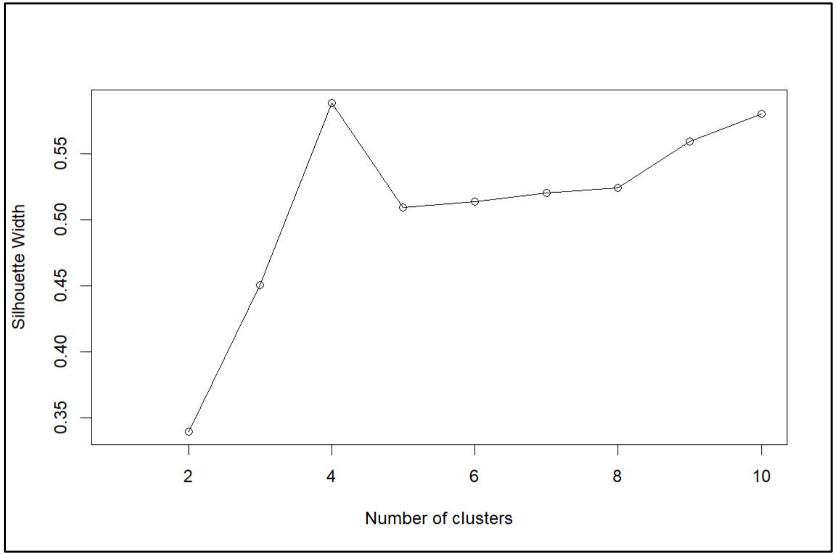
7. Hierchical Clustering

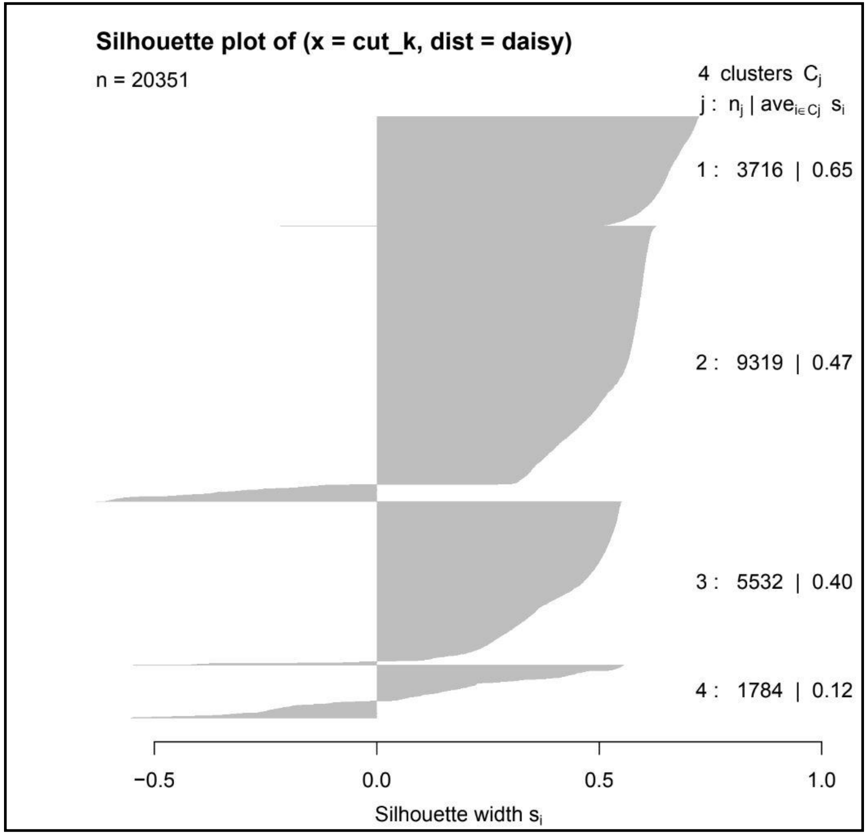
- Preprocessing
  - Selecting Accidents of California 2020
  - ~20K observations
- Clustering method
  - Agglomerative HC
  - Gower distance
- Metrics
  - Daisy Squared
- Aggregation criteria
  - Ward method



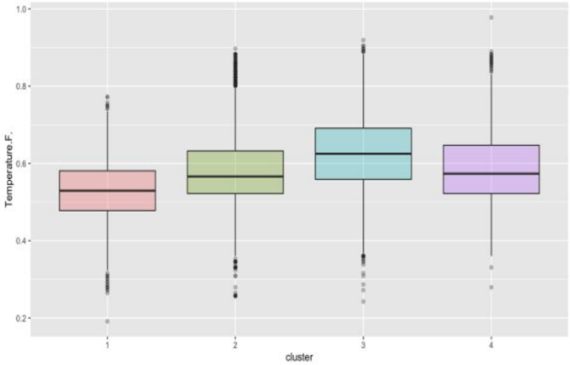
7. Hierchical Clustering

- Selection of optimal number of clusters
  - NbClust
    - cindex, dunn, frey, mcclain
  - Silhouette coefficient

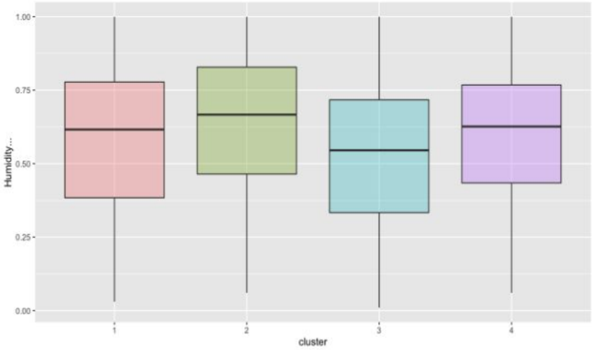




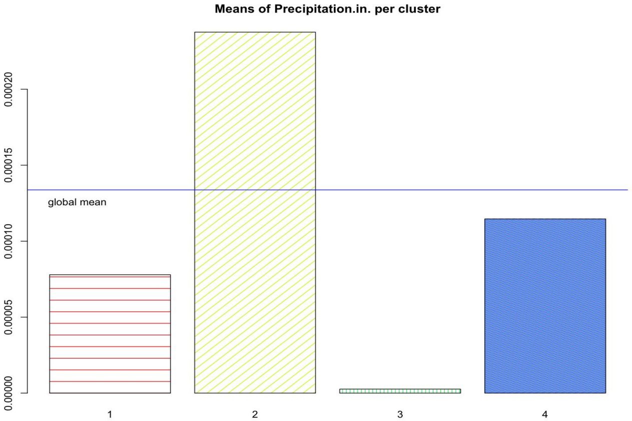
8. Profiling of Clusters

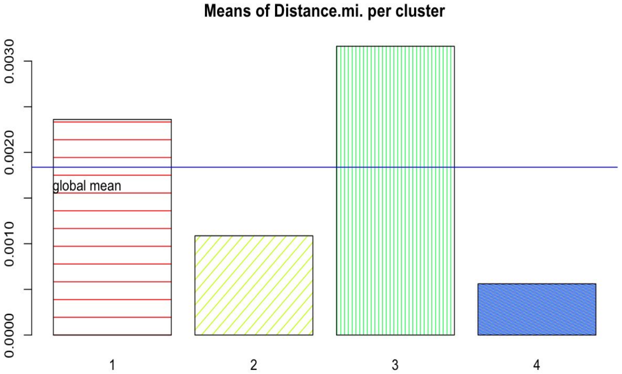


Temperature distribution for each cluster



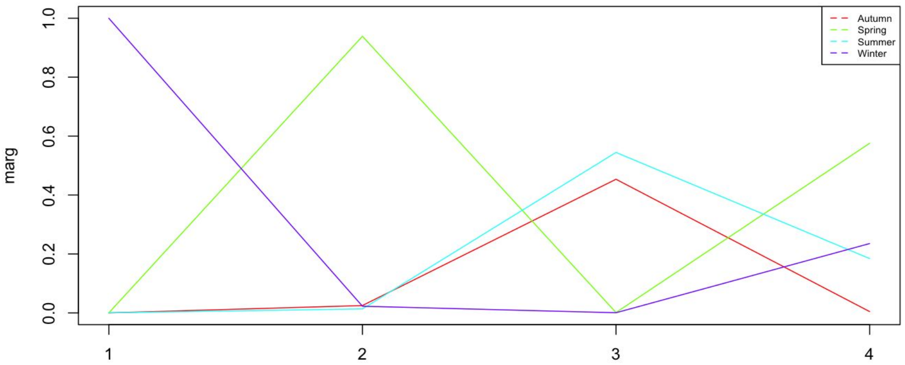
Humidity distribution for each cluster



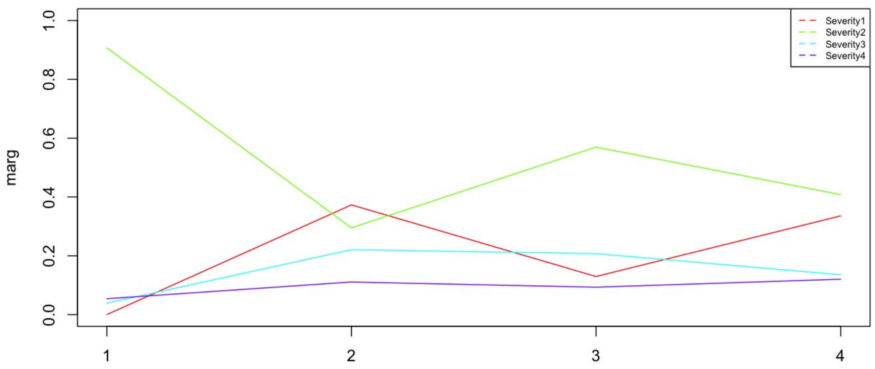


8. Profiling of Clusters

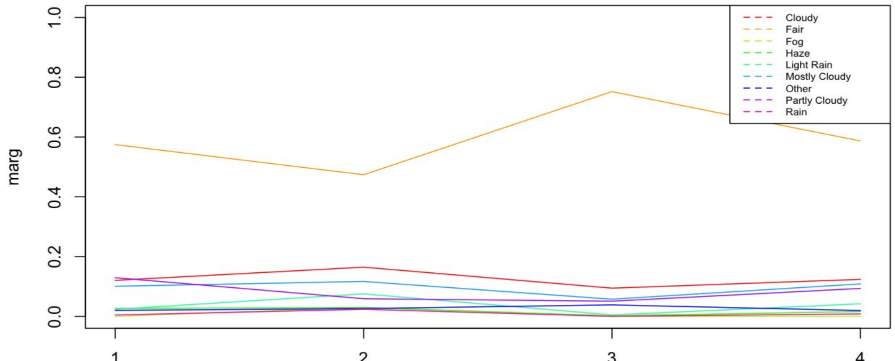
Prop. of all levels of Season within each cluster



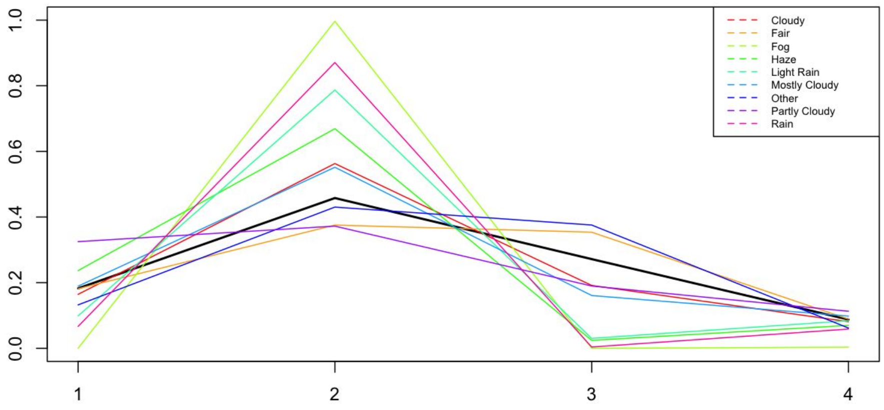
Prop. of all levels of Severity within each cluster



Prop. of all levels of Weather_Condition within each cluster

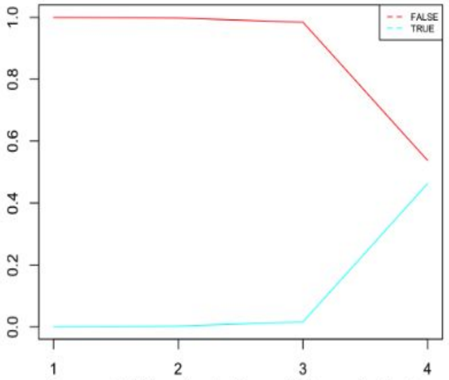


Prop. of each level of Weather_Condition within all clusters

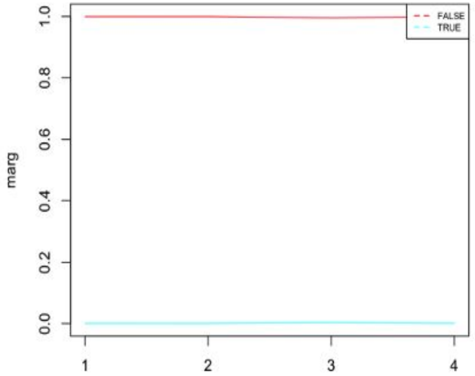


8. Profiling of Clusters

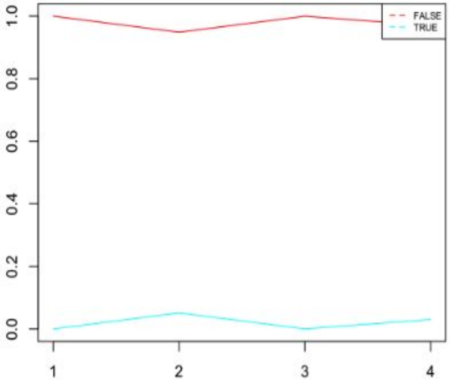
Prop. of all levels of Crossing within each cluster



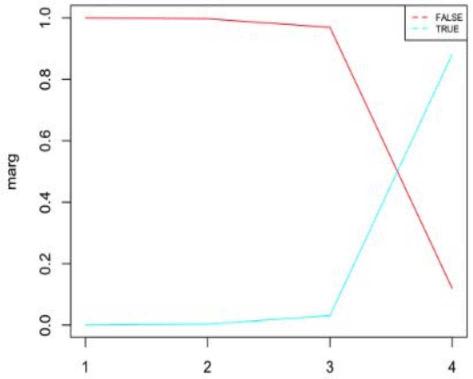
Prop. of all levels of Bump within each cluster



Prop. of all levels of Stop within each cluster



Prop. of all levels of Traffic_Signal within each cluster



Significant features per cluster

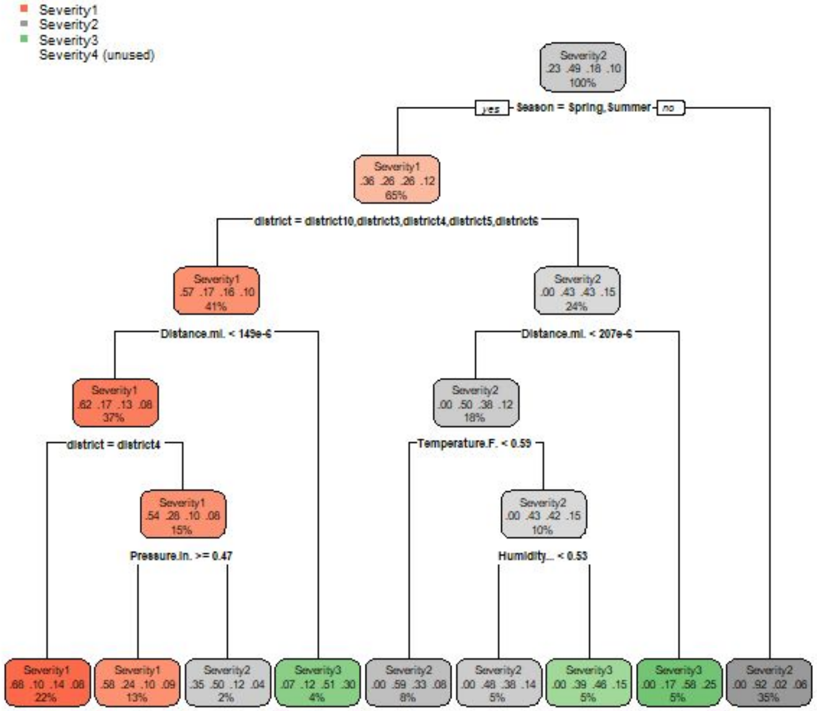
Cluster 1- Distance, Precipitation, Crossing and Traffic Signal

Cluster 2 - Humidity and Precipitation, Stop, Weather Condition, Crossing and Severity.

Cluster 3 - Distance, Crossing, Traffic Signal and Severity

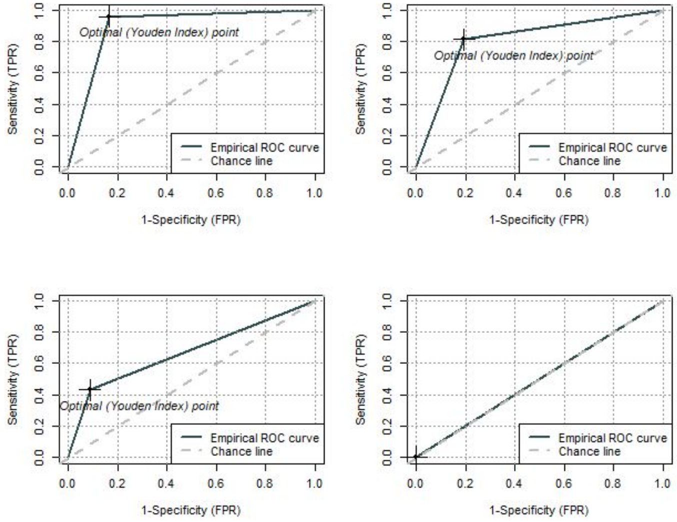
Cluster 4 - Season and Severity

9. Decision Trees

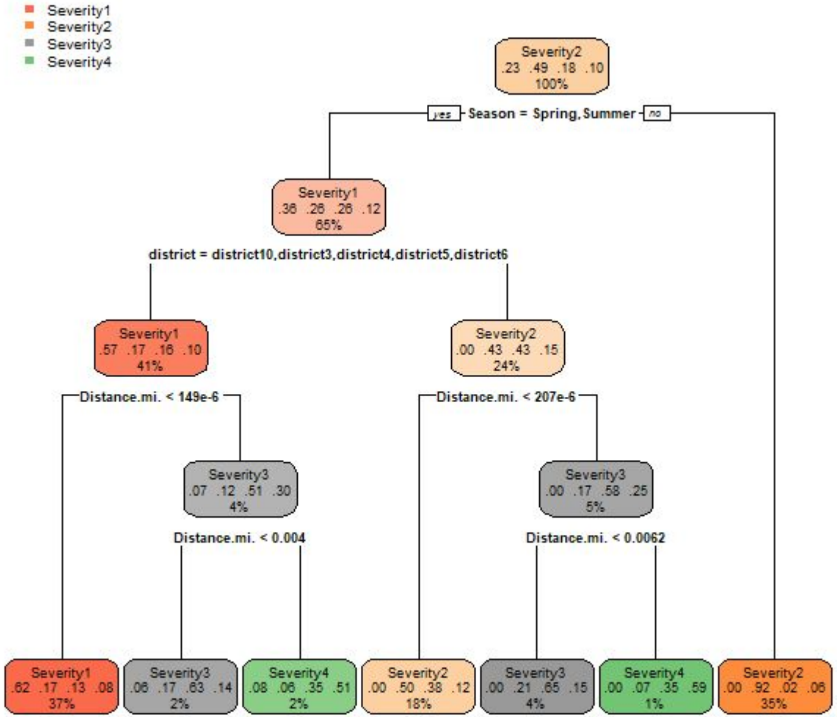

🟥 Severity1
▨ Severity2
🟩 Severity3
Severity4 (unused)
Severity2 .23 .49 .18 .10 100%
yes
Season = Spring,Summer
no
Severity1 .36 .26 .26 .12 65%
district = district10,district3,district4,district5,district6
Severity1 .57 .17 .16 .10 41%
Severity2 .00 .43 .43 .15 24%
Distance.mi. < 149e-6
Distance.mi. < 207e-6
Severity1 .62 .17 .13 .08 37%
Severity2 .00 .50 .38 .12 18%
district = district4
Temperature.F. < 0.59
Severity1 .54 .28 .10 .08 15%
Severity2 .00 .43 .42 .15 10%
Pressure.in. >= 0.47
Humidity... < 0.53
Severity1 .68 .10 .14 .08 22%
Severity1 .58 .24 .10 .09 13%
Severity2 .35 .50 .12 .04 2%
Severity3 .07 .12 .51 .30 4%
Severity2 .00 .59 .33 .08 8%
Severity2 .00 .48 .38 .14 5%
Severity3 .00 .39 .46 .15 5%
Severity3 .00 .17 .58 .25 5%
Severity2 .00 .92 .02 .06 35%

Tree 4 and 7

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

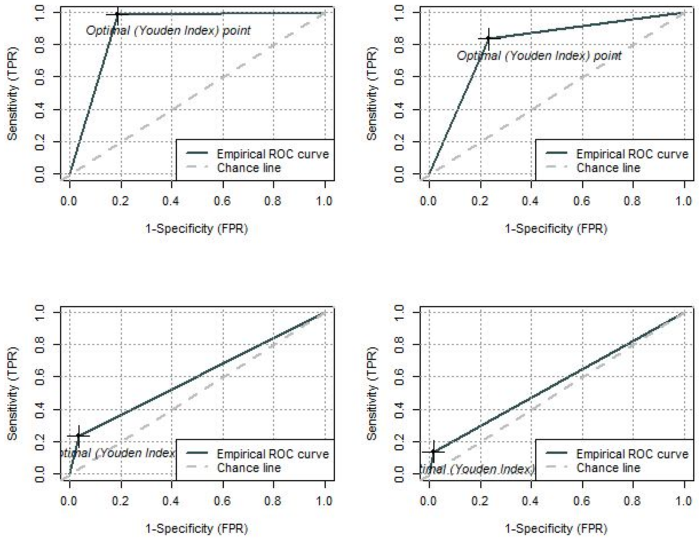


9. Decision Trees


Severity1
Severity2
Severity3
Severity4
Severity2 .23 .49 .18 .10 100%
yes
Season = Spring, Summer
no
Severity1 .36 .26 .26 .12 65%
district = district10,district3,district4,district5,district6
Severity1 .57 .17 .16 .10 41%
Severity2 .00 .43 .43 .15 24%
Distance.mi. < 149e-6
Distance.mi. < 207e-6
Severity3 .07 .12 .51 .30 4%
Severity3 .00 .17 .58 .25 5%
Distance.mi. < 0.004
Distance.mi. < 0.0062
Severity1 .62 .17 .13 .08 37%
Severity3 .06 .17 .63 .14 2%
Severity4 .08 .06 .35 .51 2%
Severity2 .00 .50 .38 .12 18%
Severity3 .00 .21 .65 .15 4%
Severity4 .00 .07 .35 .59 1%
Severity2 .00 .92 .02 .06 35%

Tree 2

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

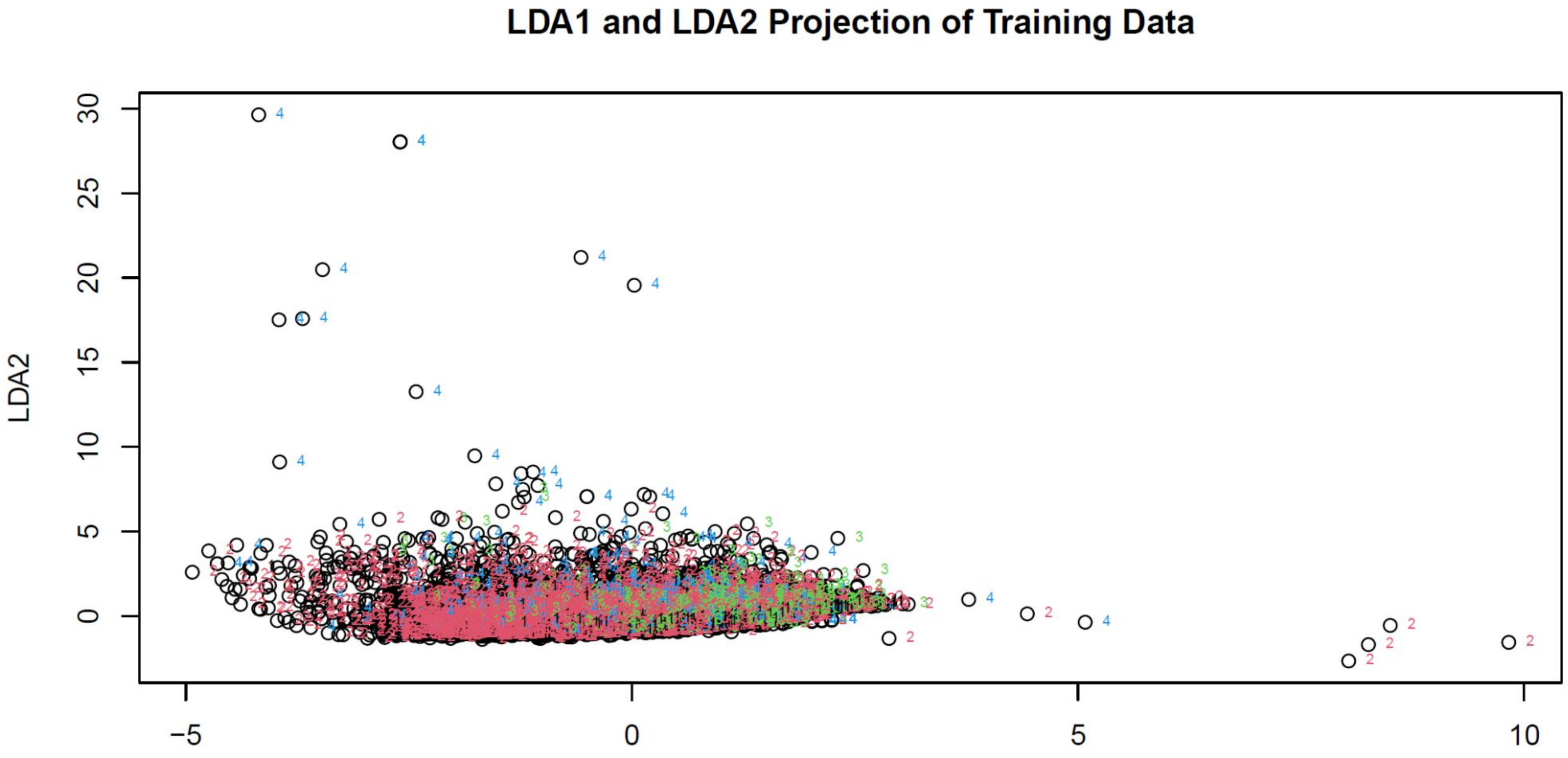


10. LDA

- Explanatory Continuous Variables Used:
  - Temperature(F), Humidity(%), Pressure(in), Precipitation(in), Distance(mi) and Visibility(mi)
  - Number of variables: 6
- Qualitative Target Variable:
  - Severity
  - 4 Cardinality Levels: Severity1, Severity2, Severity3, Severity4
- Splitting Dataset into Training - Testing Sets:
  - 66% Training Set (13635 observations)
  - 33% Test Set (6716 observations)
- Training LDA on the Training Set
  - Separation explained:

	Discriminant Functions		
Separation Explained	**LD1**	**LD2**	**LD3**
	62%	30%	8%

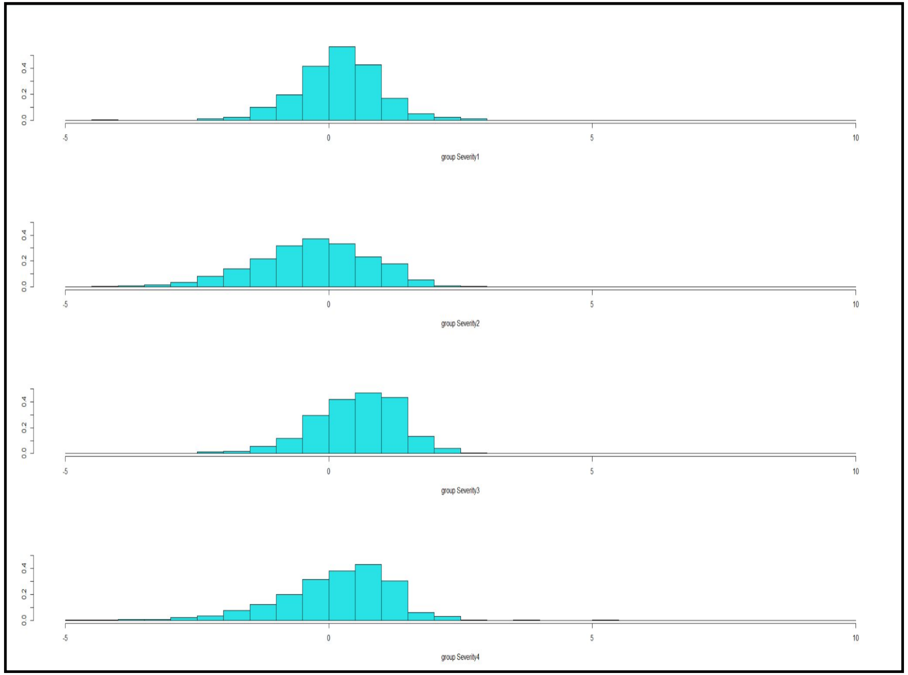
10. LDA



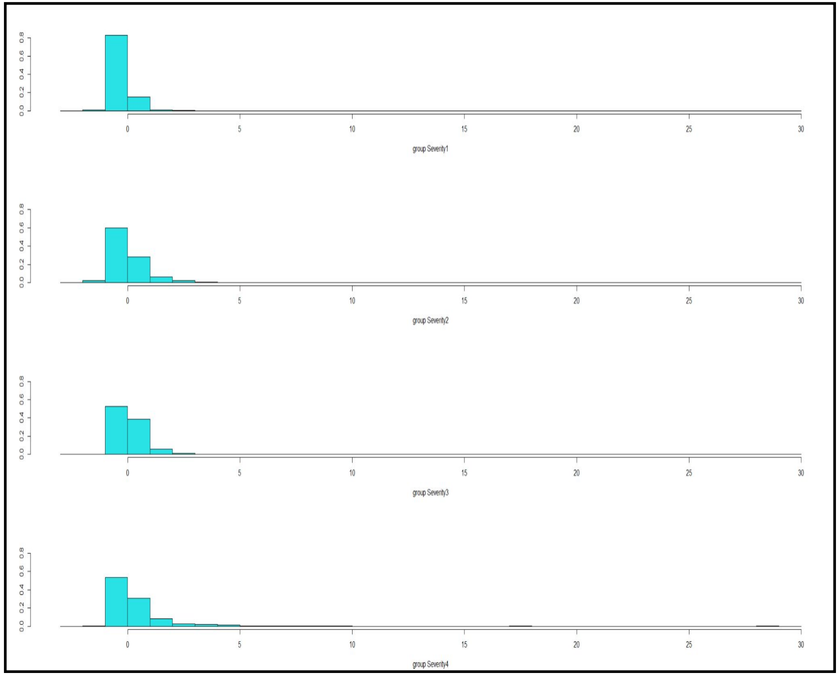
LDA1 and LDA2 Projection of Training Data

10. LDA

Stacked Histogram of the 1st Discriminant Function Values



Stacked Histogram of the 2nd Discriminant Function Values

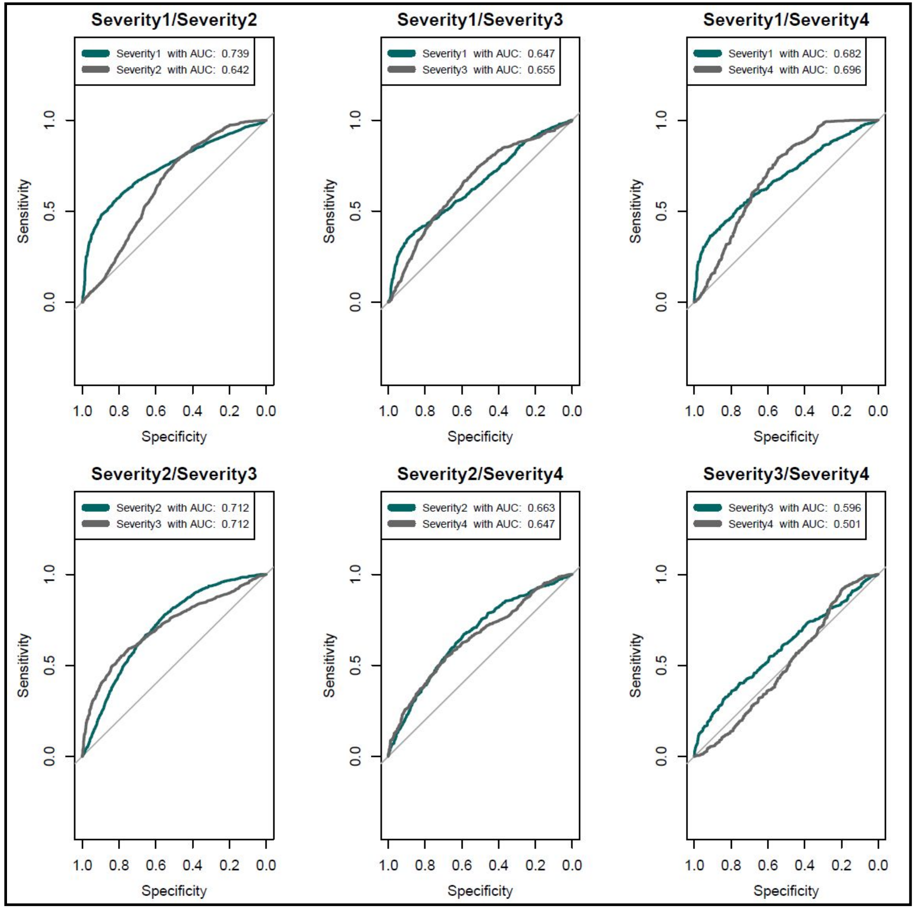


10. LDA

Training Set - Validation Metrics

- Accuracy: 0.49
- MCR: 0.51
- Precision-Recall-F1 Table

	Severity1	Severity2	Severity3	Severity4
Precision	0.269	0.5046	0.340	0.407
Recall	0.006	**0.929**	0.143	0.040
F1	0.012	**0.6540**	0.2022	0.073

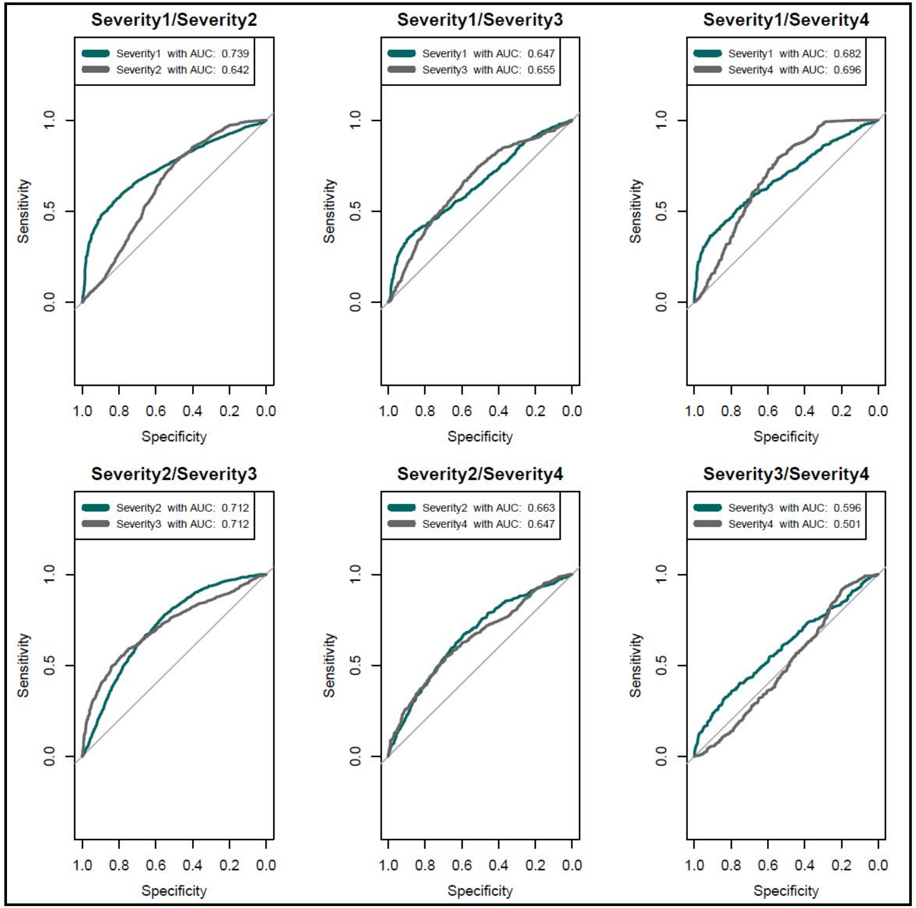


10. LDA

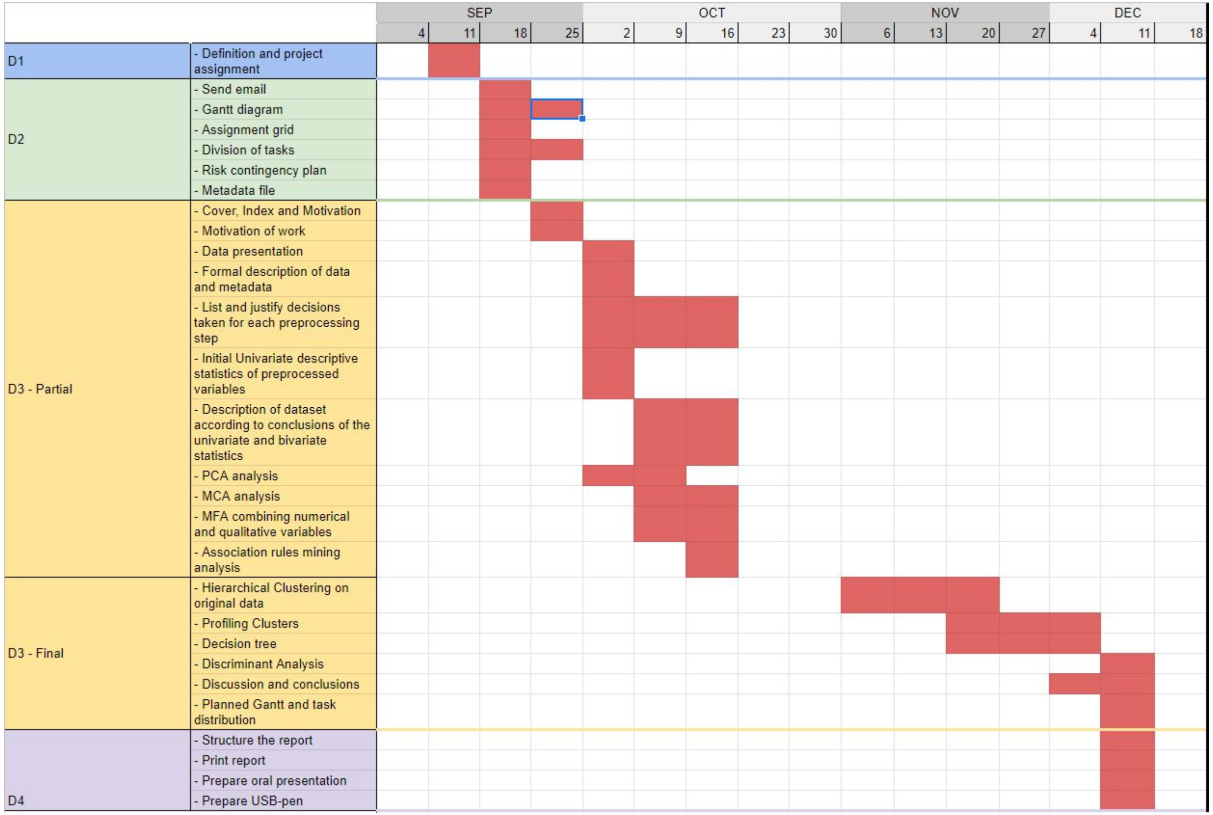
Testing Set - Validation Metrics

- Accuracy: 0.48
- MCR: 0.52
- Precision-Recall-F1 Table

	Severity1	Severity2	Severity3	Severity4
Precision	0.339	0.496	0.298	0.4659
Recall	0.011	**0.9248**	0.1265	0.0627
F1	0.0216	**0.6460**	0.1776	0.1106



11. Gantt Diagram



		SEP				OCT					NOV				DEC		
		4	11	18	25	2	9	16	23	30	6	13	20	27	4	11	18
D1	- Definition and project assignment		🟥														
D2	- Send email			🟥													
	- Gantt diagram			🟥	🟥												
	- Assignment grid			🟥													
	- Division of tasks			🟥	🟥												
	- Risk contingency plan			🟥													
	- Metadata file			🟥													
D3 - Partial	- Cover, Index and Motivation				🟥												
	- Motivation of work				🟥												
	- Data presentation					🟥											
	- Formal description of data and metadata					🟥											
	- List and justify decisions taken for each preprocessing step					🟥	🟥	🟥									
	- Initial Univariate descriptive statistics of preprocessed variables					🟥											
	- Description of dataset according to conclusions of the univariate and bivariate statistics						🟥	🟥									
	- PCA analysis					🟥	🟥										
	- MCA analysis						🟥	🟥									
	- MFA combining numerical and qualitative variables						🟥	🟥									
	- Association rules mining analysis							🟥									
D3 - Final	- Hierarchical Clustering on original data										🟥	🟥	🟥				
	- Profiling Clusters												🟥	🟥	🟥		
	- Decision tree												🟥	🟥	🟥		
	- Discriminant Analysis															🟥	
	- Discussion and conclusions														🟥	🟥	
	- Planned Gantt and task distribution															🟥	
D4	- Structure the report															🟥	
	- Print report															🟥	
	- Prepare oral presentation															🟥	
	- Prepare USB-pen															🟥	

12. Final Conclusions

- Data was challenging to manage and to extract firm conclusions

- Weather Condition and Traffic Signs do not influence severity of the accident

- Location makes a difference in the number of accidents

- Occurrence of accident is strongly correlated with human actions

- To improve the analysis human factors (speed of the vehicle, time of accident, …) should be added

Thank you