Prediction of accidents in Catalonia through Machine learning

between 2010-2018

Table with the data

	Any	zona	dat	via	pk	nomMun	nomCom	nomDem	F_MORTS	F_FERITS_GREUS	F_FERITS_LLEUS	F_VICTIMES
0	2010	Zona urbana	25/01/2010	SE	999999.0	CANOVES I SAMALUS	Valles Oriental	Barcelona	0	1	0	1
1	2010	Carretera	31/10/2010	N-240	999.0	LLEIDA	Segria	Lleida	0	1	3	4
2	2010	Carretera	17/05/2010	N-II	7087.0	FORNELLS DE LA SELVA	Girones	Girona	1	0	2	3
3	2010	Zona urbana	21/08/2010	SE	999999.0	BARCELONA	Barcelones	Barcelona	0	2	7	9

Description of the attributes (1/3)

Year - Year of the accident

zone - Type of urban or interurban road

date - Date of the accident

road - Name of the road where the accident took place.

pk - Number of kilometers where the accident occurred.

nameMun - Municipality where the accident occurred

comName - County where the accident occurred

nameDem - Demarcation where the accident occurred

F_MORTS - Number of fatalities

F_HAPTER INJURIES - Number of seriously injured

F_LIGHT_WOUND - Number of minor injuries

D_INFLUIT_CARACT_ENTORN - Indicates whether the presence of fog, at the discretion of the agent, may have influenced the accident.

D_INFLUIT_CIRCULATION - Indicates whether the amount of traffic, at the discretion of the agent, may have influenced the accident.

D_INFLUIT_CLIMATE_STATE - Indicates whether the weather condition, at the discretion of the agent, may have influenced the accident.

D_INFLUIT_INTEN_VENT - Indicates whether wind intensity, at the discretion of the agent, may have influenced the accident.

D_INFLUIT_LIGHT BRIGHTNESS - Indicates whether the brightness, at the discretion of the agent, may have influenced the accident.

D_INFLUIT_MESU_ESP - Indicates whether driving under special measures, at the discretion of the agent, may have influenced the accident.

D_INFLUIT_OBJ_CALCADA - Indicates if there are any objects on the road that, in the opinion of the agent, may have influenced the accident.

D_INFLUIT_SOLCS_RASES - Indicates whether there are furrows or ditches on the road that, in the opinion of the agent, may have influenced the accident.

Description of the attributes (2/3)

C_WAY SPEED - Track speed

D_ACC_AMB_FUGA - Leak accident

D_FOG - Presence of fog

D_CARACT_ENTORN - Influence characteristics of the environment

D_SPECIAL ROAD - Special traffic lane

D_CIRCULACIO_MESURES_ESP - Circulation under special measures

D_CLIMATOLOGY - Firm conditions according to the climatology

D_FUNC_ESP_VIA - Special function of the road as a variant, crossroads, ring road or roundabout

D_GREVITY - Severity of the accident

D_INFLUIT_BOIRA - Indicates whether the characteristics of the environment, at the discretion of the agent, may have influenced the accident.

F_VICTIMES - Total number of victims

F_UNITS INVOLVED - Number of units involved in the accident

F_VIANANTS_IMPLICATED - Number of pedestrians involved in the accident

F_BICYCLES INVOLVED - Number of bicycles involved in the accident

F_CICLOMOTORS_IMPLICATED - Number of mopeds involved in the accident

F_MOTORCYCLES INVOLVED - Number of motorcycles involved in the accident

F_VEH_LIGHT_IMPLICATED - Number of light vehicles involved in the accident

F_VEH_PESANTS_IMPLICADES - Number of heavy vehicles involved in the accident

F_ALTRES_UNIT_IMPLICADES - Number of other types of units involved in the accident

F_UNIT_DESC_IMPLICADES - Number of units of unknown type involved in the accident

Description of the attributes (3/3)

D_INFLUIT_VISIBILITY - Indicates if there are any visibility restrictions that, in the opinion of the agent, may have influenced the accident.

D_INTER_SECTION - Whether the accident occurred in section or intersection

D_LIMIT_ SPEED - Speed limit of the road at the scene of the accident

BRIGHTNESS - Brightness that existed at the time of the accident

D_REGULATION_PRIORITY - Signage that appears at an intersection.

D_SENTITS_VIA - Indicates whether the road has a direction of traffic or whether it has a two-way street

D_SUBTYPE_ACCIDENT - Type of accident

D_SUBTYPE_TRAM - Type of intersection

D_SUBZONE - Subclassification of the accident area

D_SUPERFICIE - Conditions in which the road surface is located at the time of the accident

D_TYPE_VIA - Type of route

ROAD OWNERSHIP - Owner of the track

D_ALTIMETRIC TRACING - Altimetric tracing

D_VENT - State of the wind at the time of the accident

grupDiaLab - Type of day of the accident: working day, weekend ...

hor - Time of the accident

grupHor - Type of time of the accident: morning, afternoon, night

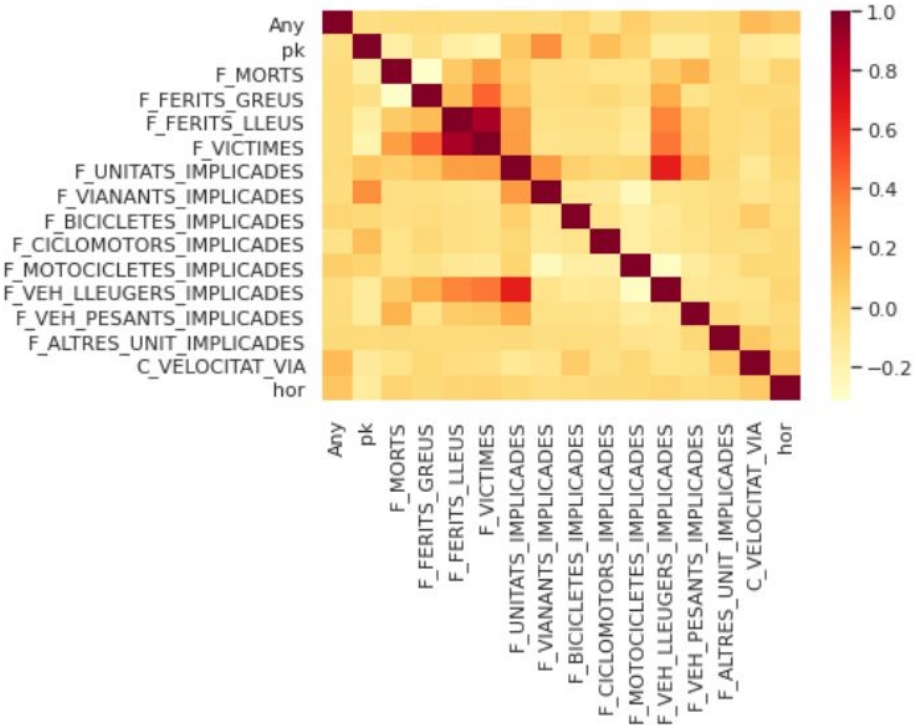
tipAcc - Type of accident

tipDay - Type of day: Monday to Thursday, Friday ...

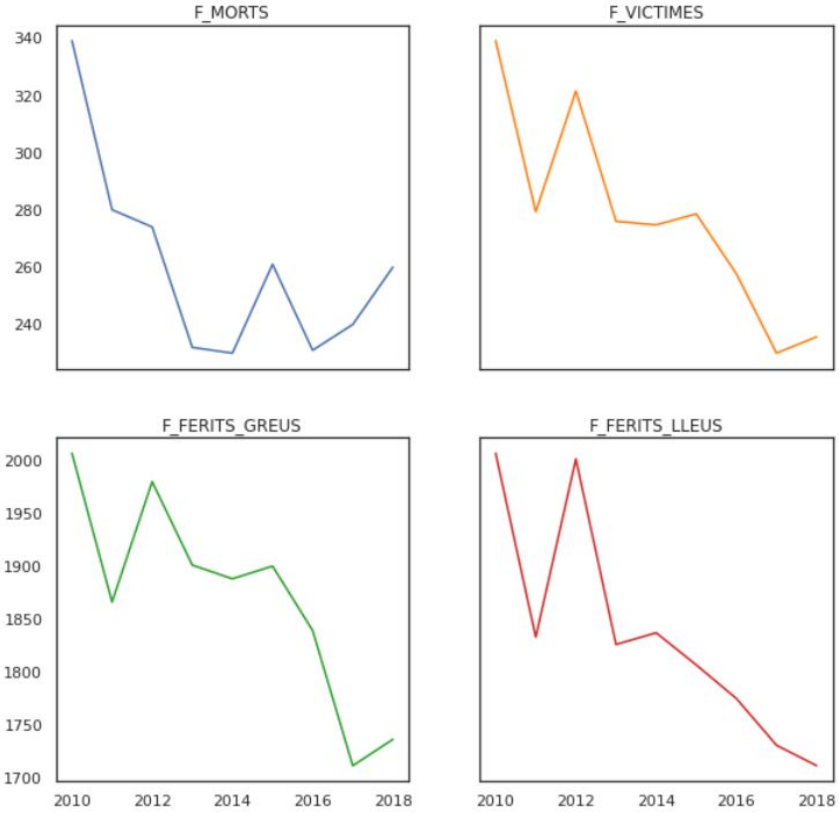
Dependent variable: F_VICTIMES

	Any	zona	dat	via	pk	nomMun	nomCom	nomDem	F_MORTS	F_FERITS_GREUS	F_FERITS_LLEUS	F_VICTIMES
0	2010	Zona urbana	25/01/2010	SE	999999.0	CANOVES I SAMALUS	Valles Oriental	Barcelona	0	1	0	1
1	2010	Carretera	31/10/2010	N-240	999.0	LLEIDA	Segria	Lleida	0	1	3	4
2	2010	Carretera	17/05/2010	N-II	7087.0	FORNELLS DE LA SELVA	Girones	Girona	1	0	2	3
3	2010	Zona urbana	21/08/2010	SE	999999.0	BARCELONA	Barcelones	Barcelona	0	2	7	9

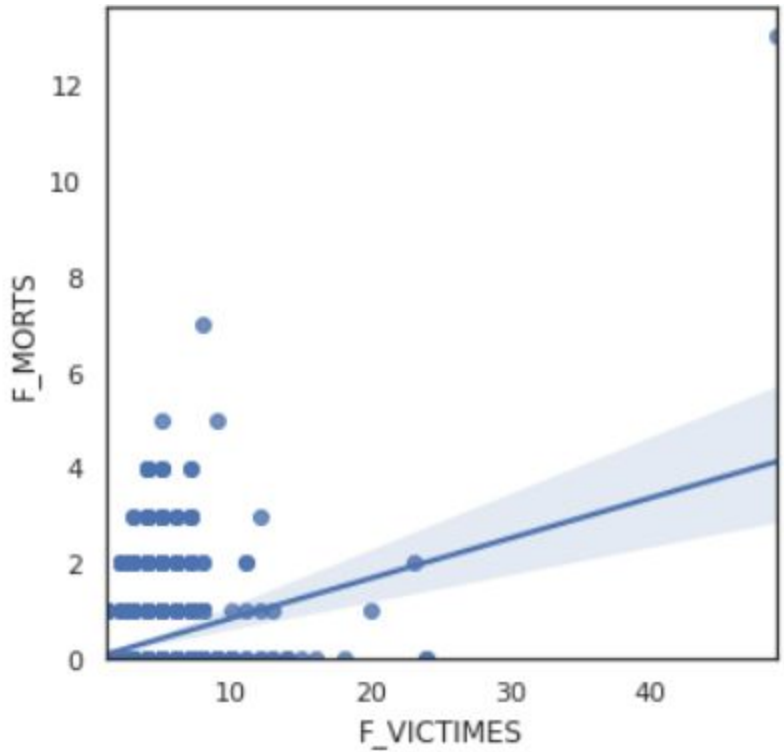
Correlation matrix of all numerical variables



Dead and injured vs time



Linear regression F_MORTS vs F_VICTIMES

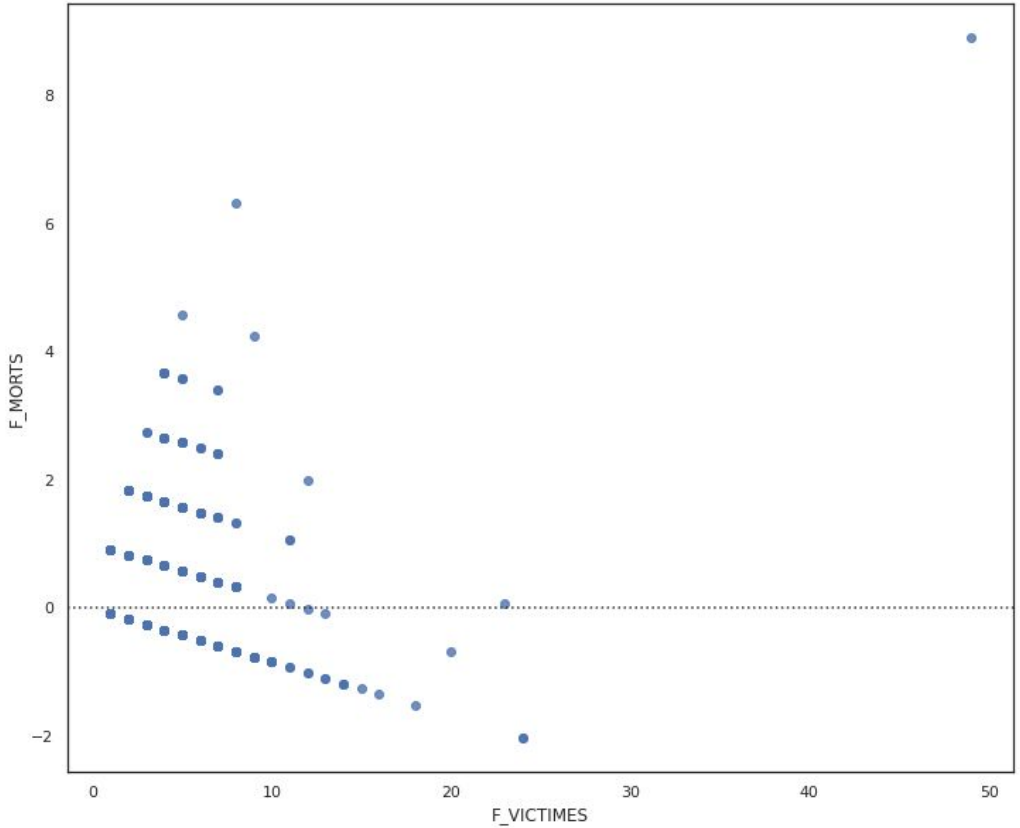



```
1 lm.score(X, Y)
```

```
0.06373222047399818
```

MORTS = 0.083844473330723009 * VICTIMES + 0.0099035057115558777

Residual plot

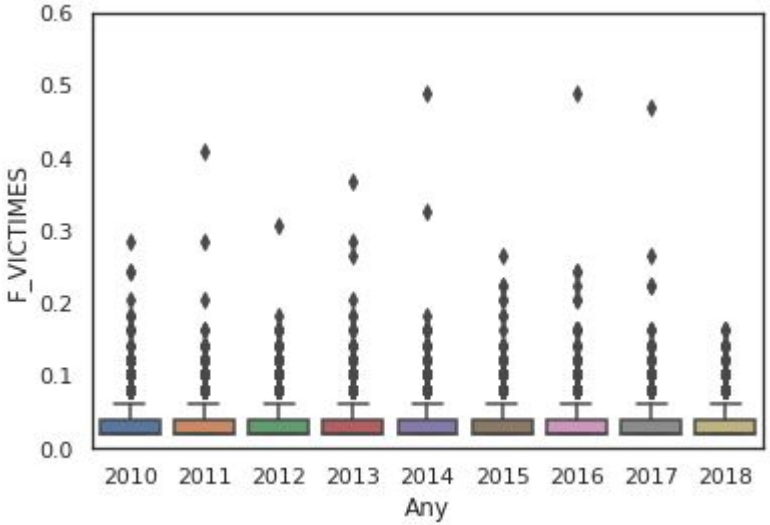


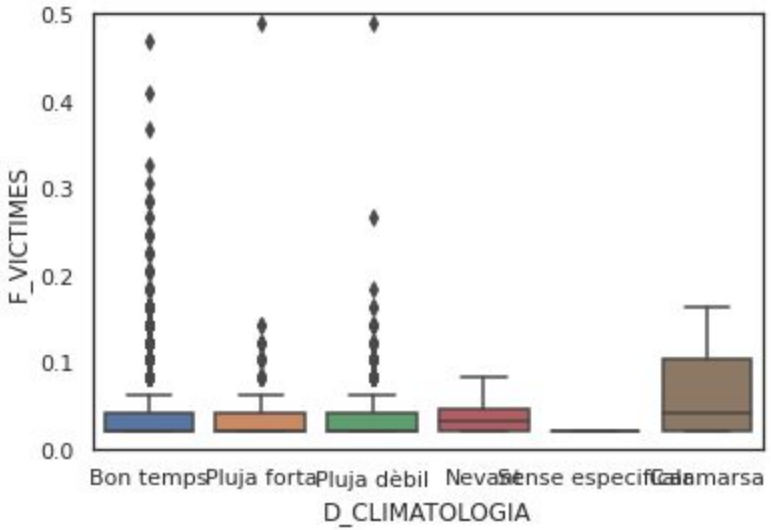
Determinacion de Puntos Negros Catalunya

via	pk	
CR	0.0	41
	1.0	57
	2.0	18
	4.0	13
	9.0	26
	22.0	11

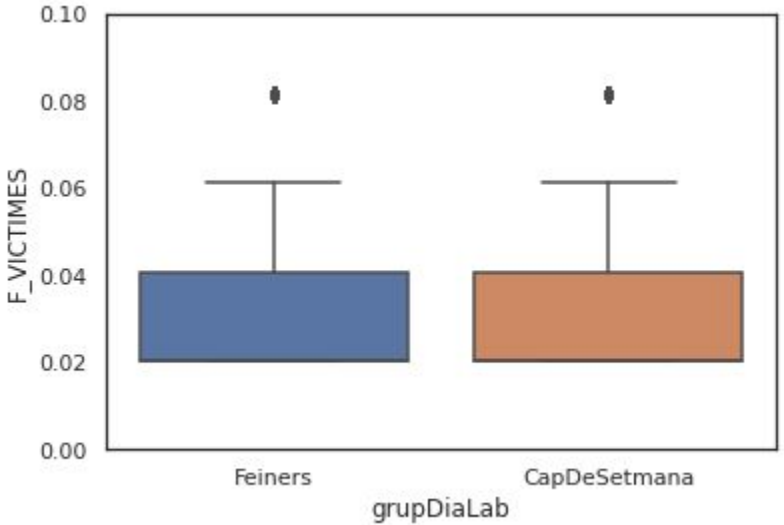
Conclusion: 5 black spots have been found where more than 10 accidents have occurred.

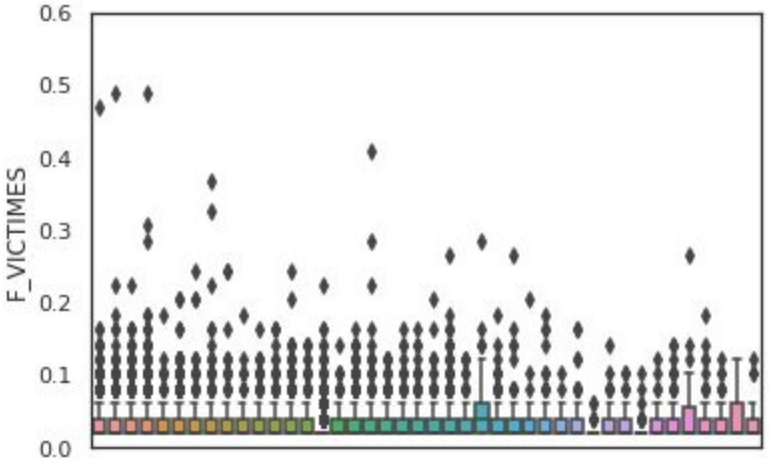
ANOVA (Boxplots)



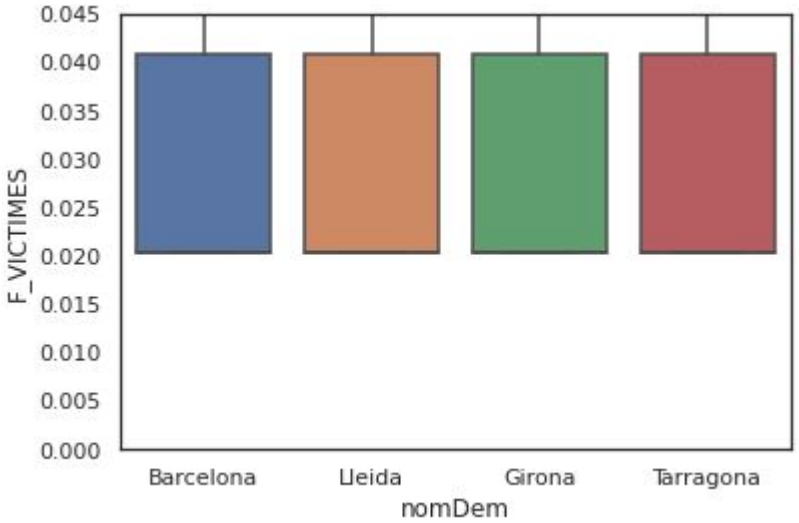


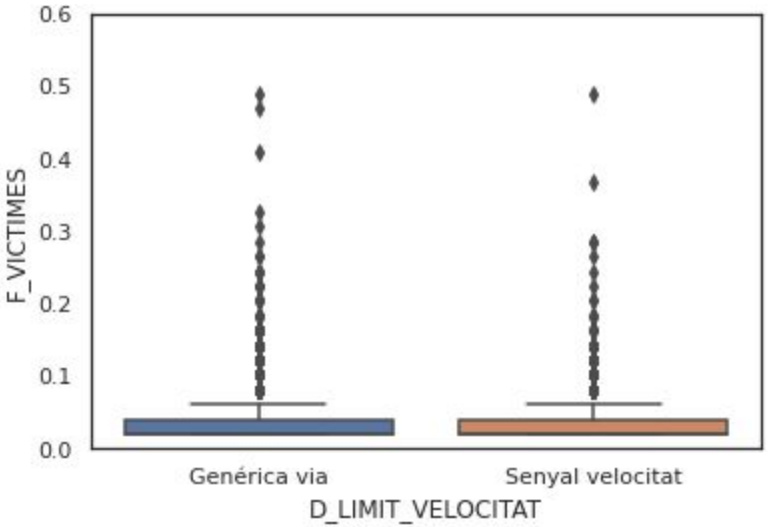
ANOVA (Boxplots)



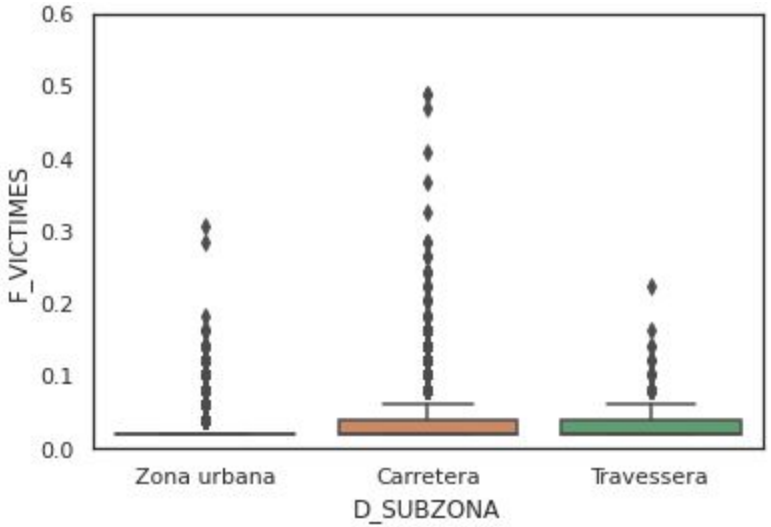


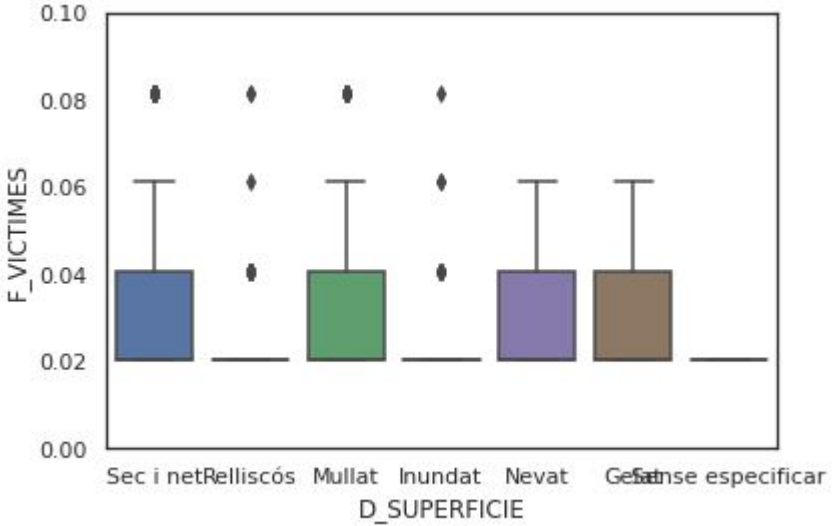
ANOVA (Boxplots)



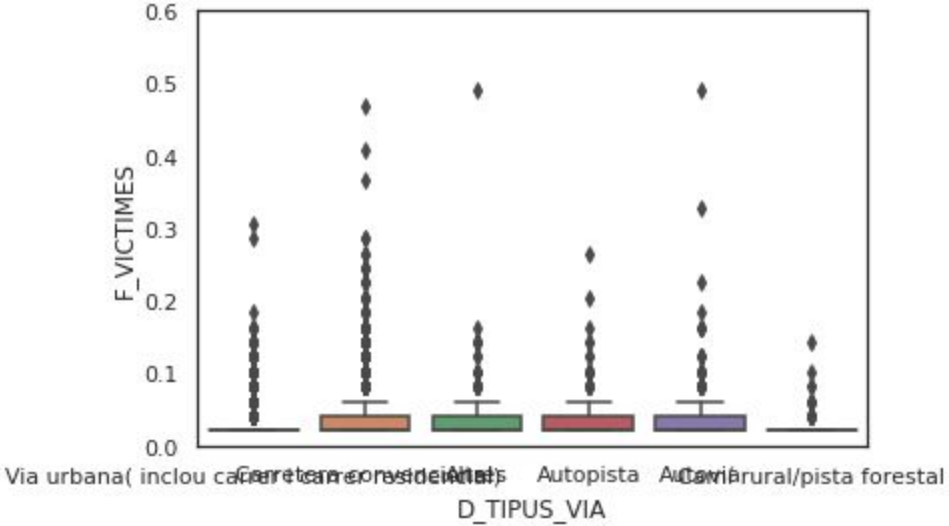


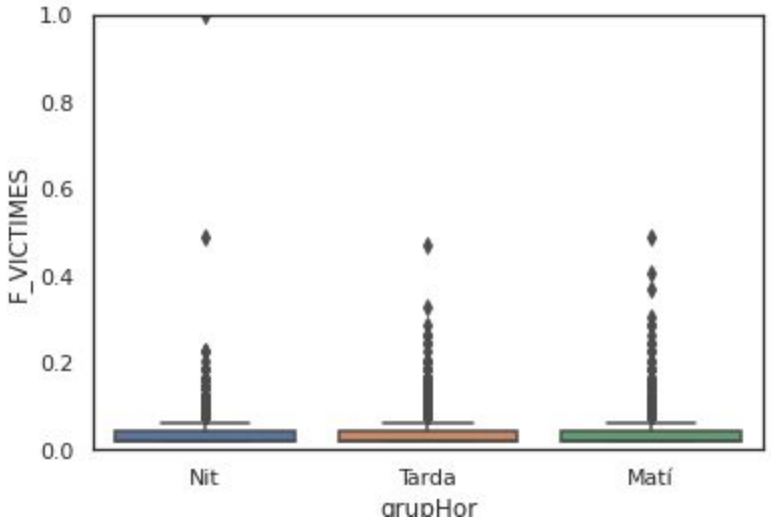
ANOVA (Boxplots)



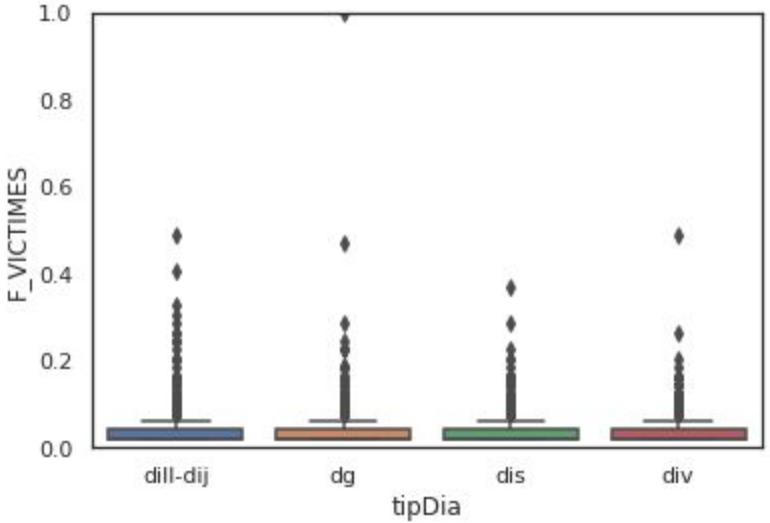


ANOVA (Boxplots)



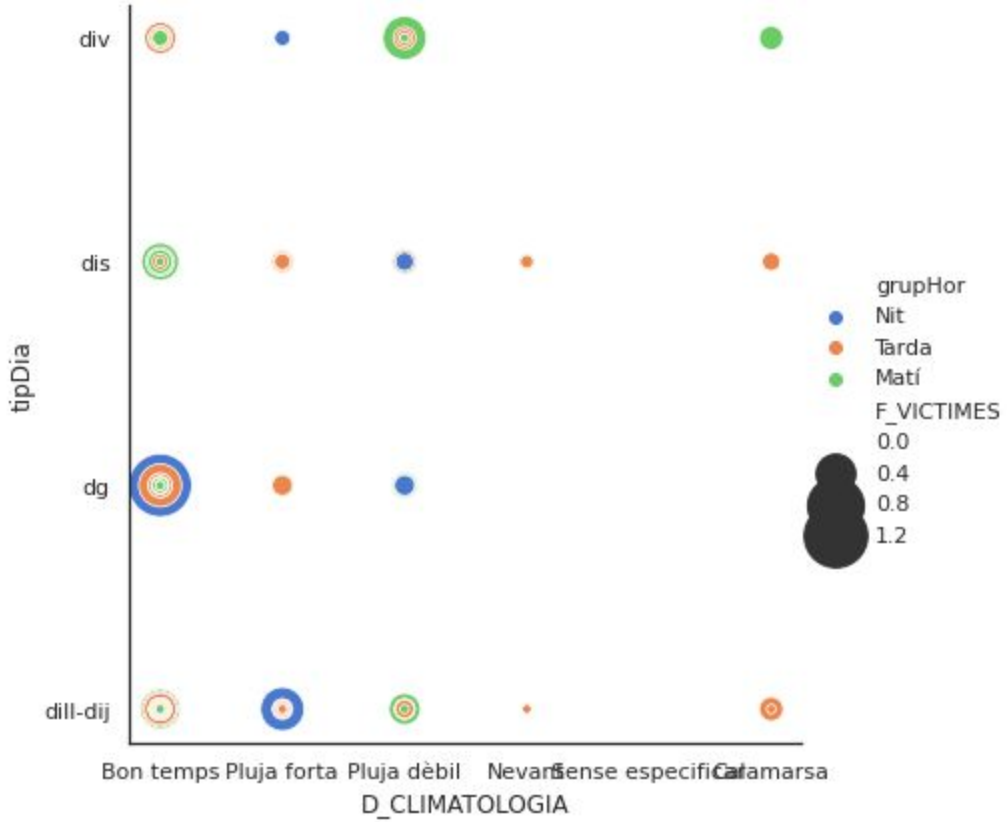


ANOVA (Boxplots)



It does not seem that any of the categorical variables can explain our target variable

4 variables in a 2-D graphic representation



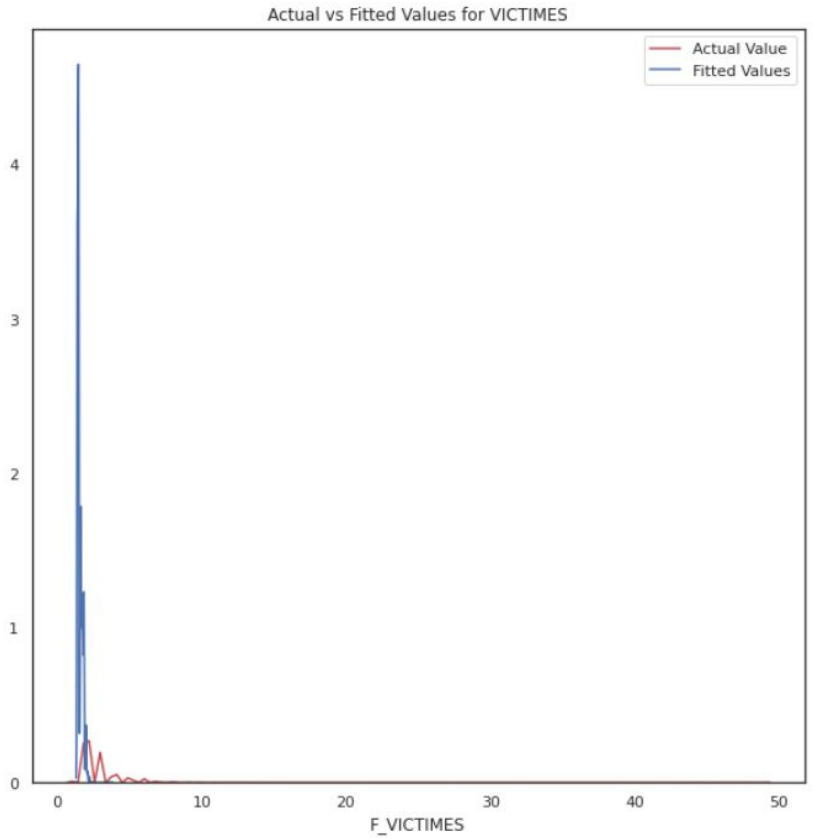
1. D_CLIMATOLOGIA
2. tipDia
3. grupHor
4. F_VICTIMES

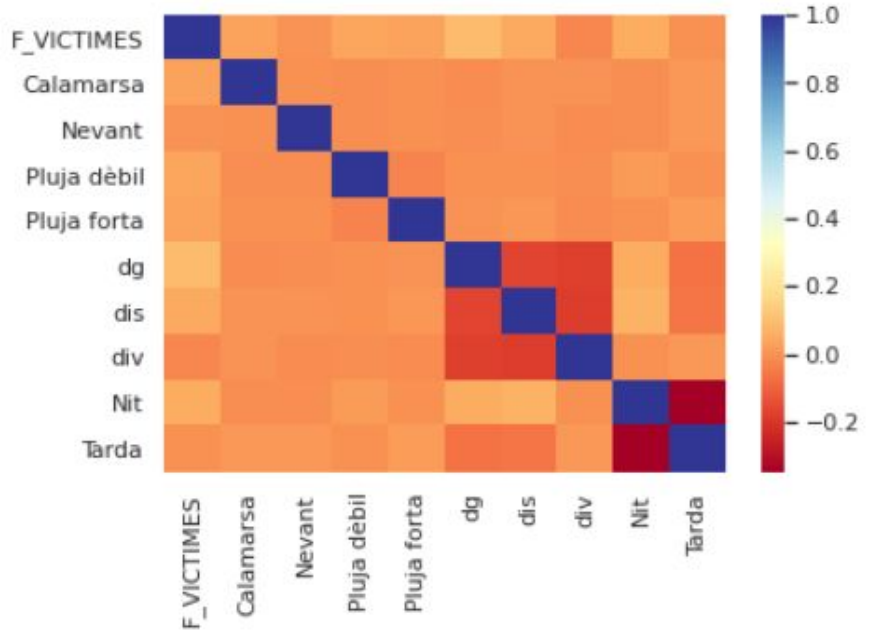
Dummy variables will be created for some categorical variables

	Calamarsa	Nevant	Pluja dèbil	Pluja forta	dg	dis	div	Nit	Tarda
0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	1	0	0	1	0
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	1	0	1	0
4	0	0	0	0	0	0	1	0	1

```
1 lm.score(Z, y)
```

```
0.021490257325099 35
```


Actual vs Fitted Values for VICTIMES
Actual Value
Fitted Values
0
1
2
3
4
0
10
20
30
40
50
F_VICTIMES


F_VICTIMES
Calamarsa
Nevant
Pluja dèbil
Pluja forta
dg
dis
div
Nit
Tarda
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Conclusion

None of the variables, neither categorical nor numerical, has a sufficiently large correlation for an adjustment to be made that can explain the variable "F_VICTIMAS". I believe that accidents in Catalonia with the data I have at my disposal occur in a random manner.

I point out different problems:

1) "F_VICTIMAS" is a discrete variable and has few possible different values.

2) "F_VICTIMAS" cannot be considered a normal distribution.

3) Insufficient data