

Anàlisi descriptiva

Anna Salazar

2022-11-17



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística



UNIVERSITAT DE
BARCELONA

Índex

Descripció de la base de dades	1
Objectius del projecte	2
Lectura de les dades	3
Preprocessament	3
Missings	3
Outliers	5
Categoritzar	6
Variable resposta	7
Anàlisi descriptiva univariant	9
Variables numèriques	9
Variables vinculades als accidents	9
Variables vinculades a les persones	9
Variables categòriques	11
Variables vinculades als accidents	11
Variables vinculades a les persones	12
Anàlisi descriptiva bivariant	13
Variables vinculades als accidents	13
Variables vinculades a les persones	14
Anàlisi per Components Principals	16
Projecció variables numèriques	20
Projecció variables categòriques	21

Descripció de la base de dades

Les bases de dades que seran utilitzades al llarg de l'estudi provenen de l'agència estatal de trànsit dels Estats Units i contenen tres taules, entre les quals s'hi troba un llistat d'accidents de tràfic ocorreguts al desembre de 2015 als Estats Units, juntament amb un recompte de totes les persones (conductors, passatgers o vianants) involucrades als accidents i, finalment, un inventari de tots els vehicles involucrats als accidents.

L'enllaç a la base esmentada és el següent:

<https://www.transportation.gov/briefing-room/traffic-fatalities-sharply-2015>

Més concretament, en cada taula es poden trobar les variables següents:

Accident és un llistat d'accidents de trànsit ocorreguts al desembre de 2015 als Estats Units.

Taula 1. Llistat de variables de la taula Accident

Variable	Tipus	Descripció
ST_CASE	Categòrica	Codi de l'accident
DAY	Categòrica	Dia de l'accident (de l'1 al 31)
HOURL	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident (99 = desconegut)
RUR_URB	Categòrica	Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut)
DAY_WEEK	Categòrica	Dia de la setmana (1 = Diumenge, 2 = Dilluns, ..., 7 = Dissabte)
FATALS	Numèrica	Nombre de ferits a l'accident
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident

Person és un llistat de totes les persones (conductors, passatgers o vianants) involucrades als accidents.

Taula 2. Llistat de variables de la taula Person

Variable	Tipus	Descripció
ST_CASE	Categòrica	Codi de l'accident al qual està involucrada la persona
PER_NO	Categòrica	Nombre de la persona dins de cada accident
AGE	Numèrica	Edat de la persona (998 = No registrada, 999 = Desconeguda)
SEX	Categòrica	Sexe de la persona (1 = home, 2 = dona, 8 = No registrat, 9 = Desconegut)
PER_TYP	Categòrica	Tipus de persona (1 = conductor, 2 = ocupant, resta de codis = altres)
DOA	Categòrica	Tipus de víctima (0 = sobrevis, 7 = mort a l'accident, 8 = mort al trasllat, 9 = Desconegut)

Vehicle és un llistat de tots els vehicles involucrats als accidents.

Taula 3. Llistat de variables de la taula Vehicle

Variable	Tipus	Descripció
ST_CASE	Categòrica	Codi de l'accident al qual està involucrat el vehicle
NO_VEH	Numèrica	Nombre de vehicles implicats en l'accident
HIT_RUN	Categòrica	Identificador de vehicle fugit (0 = No, 1 = Sí, 9 = Desconegut)
TRAV_SP	Numèrica	Velocitat estimada (mph) del vehicle quan va tenir l'accident (997,998 i 999 = Desconegut)

Variable	Tipus	Descripció
PREV_SP	Categòrica	Indicador d'existència de límit de velocitat permesa just abans de l'accident (997,998 i 999 = Desconegut)

Objectius del projecte

Estudiant aquesta base de dades sobre persones que s'han vist implicades, de forma directa o indirecta, en accidents de trànsit es preten:

- Descriure els tipus d'accidents que estan registrats
- Analitzar els diferents perfils de persones que pateixen accidents de trànsit
- Desenvolupar un model de predicció que ens permeti establir el tipus de víctima que serà cada persona depenent les característiques de l'accident i els vehicles.
- Estudiar les relacions de dependència entre variables

Lectura de les dades

A partir d'aquestes tres taules, s'extreuran dues bases de dades a partir de les quals es treballarà al llarg del projecte.

En primer lloc, es tindrà en compte la informació dels accidents. D'aquesta manera es podrà estudiar les característiques dels diferents accidents registrats, així com es podran fer prediccions sobre els nous accidents en funció de les seves característiques. S'ha anomenat aquesta base **accident**, i està conformada per les variables següents: DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, MORTS, NO_VEHICLE, HIHAMORTS.

Les variables **MORTS**, **NO_PER**, **NO_VEHICLE** i **HIHAMORTS** han sigut creades a posteriori a partir de les taules de les que es disposava, i es defineixen a continuació:

Taula 4. Llistat de variables definides a posteriori per a la taula Accidents

Variable	Tipus	Descripció
MORTS	Numèrica	Nombre de morts en l'accident
NO_PER	Numèrica	Nombre de persones implicades en l'accident
NO_VEHICLE	Numèrica	Nombre de vehicles implicats en l'accident
HIHAMORTS	Catègorica	Variable identificadora dels accidents mortals (0: no hi ha morts en l'accident, 1: hi ha morts en l'accident)

D'altra banda, s'estudiarà la informació sobre les persones implicades en aquests accidents. D'aquesta manera es podrà perfilar el tipus de conductors en els casos en que hi hagi morts en l'accident, així com en els que no hi hagi. Aquesta informació també ens facilitarà l'elaboració de possibles models per predir el tipus de víctima que serà una persona involucrada en un accident de trànsit en base a les seves característiques. en aquest cas, s'ha anomenat aquesta base **persones**, i està conformada per les variables següents: DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, MORTS, NO_VEHICLE, NO_FUGITS, AGE, SEX, PERTYP i DOA.

La variable **NO_FUGITS** ha sigut creada a posteriori a partir de les taules de les que es disposava, i es defineix a continuació:

Taula 5. Llistat de variables definides a posteriori per a la taula Persones

Variable	Tipus	Descripció
NO_FUGITS	Numèrica	Nombre de vehicles fugits implicats en l'accident

Preprocessament

La base de dades d'accidents està formada per 2781 casos (accidents) i 11 variables. En canvi, la base de dades de persones la conformen 7087 individus (files) i 15 variables (columnes).

Les variables que tenim són DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, MORTS, NO_VEHICLE, HIHAMORTS, NO_FUGITS, AGE, SEX, PERTYP i DOA.

Missings

Per a poder tractar les dades mancants de la base de dades, en primer lloc haurem de transformar-les, ja que les variables que presenten dades mancants les tenen codificades.

	NA	Percentatge de NA		NA	Percentatge de NA
			DAY	0	0.00
			HOURL	55	0.78
			MINUTE	58	0.82
			RUR_URB	0	0.00
			DAY_WEEK	0	0.00
			FATALS	0	0.00
			DRUNK_DR	0	0.00
			NO_PER	0	0.00
			MORTS	0	0.00
			NO_VEHICLE	0	0.00
			NO_FUGITS	0	0.00
			AGE	222	3.13
			SEX	0	0.00
			PER_TYP	0	0.00
			DOA	0	0.00

Taula 6. Percentatge de missings per variable

En el cas de les variables numèriques amb *missings*, que són l'edat (AGE), l'hora (HOURL), el minut (MINUTE) i la velocitat estimada del vehicle quan va tenir l'accident (TRAV_SP), les codificacions per aquestes dades són 99, 997, 998 o 999, depenent de cada cas.

Un cop transformades aquestes dades, podem visualitzar a la taula següent els *missings* per cada variable numèrica, tant en terme absolut com relatiu. A la taula següent s'hi poden trobar les variables de les bases de dades d'accidents i de persones, respectivament, juntament amb el nombre de dades mancants que presenten, i el tant per cent que aquestes suposen al total de la informació de la variable.

Tal i com es pot observar, a la base de dades d'accidents s'hi troben *missings* per a les variables HOURL i MINUTES, mentre que per a la base de dades de persones, s'hi troben missings per a les variables HOURL, MINUTES i AGE. En ambdós casos, totes les variables són numèriques i, per aquest motiu es pot usar l'algoritme KNN per a la imputació de valors a les dades mancants.

K-nearest neighbors (KNN) és un tipus d'algoritme d'aprenentatge supervisat que s'utilitza tant per a la regressió com per a la classificació. La seva funció és intentar predir la classe correcta per a unes dades de prova (que, en el nostre cas, seran les variables que presenten dades mancants) en base a la seva similitut amb altres mostres de dades conegudes (en el nostre cas, les variables completes). Tot això es fa assumint que les dades amb trets similars es troben juntess, i utilitza mesures de distància en el seu nucli.

Un cop s'ha aplicat l'algoritme per a les variables corresponents, es pot veure, a continuació, com cap de les dues bases de dades presenta cap missing a les variables conflictives.

Recordem que la taula mostra les bases de dades d'accidents i de les persones implicades en els accidents, respectivament:

	NA	Percentatge de NA		NA	Percentatge de NA
HOURL	0	0	HOURL	0	0
MINUTE	0	0	MINUTE	0	0
			AGE	0	0

Taula 7. Percentatge de missings per variable després del KNN

Outliers

Pel que fa a les dades atípiques, en destaca el nombre de persones implicades a l'accident. Més específicament, hi ha un cas en que 53 persones estan involucrades en un accident. A priori, res ens fa pensar que aquesta dada, tot i ser atípica, sigui certa. Això no obstant, a l'hora de la segmentació les dades es podrien veure afectades per aquest valor, ja que alguns algoritmes són molt sensibles a les dades atípiques.

A la següent figura es representa la variable nombre de persones (NO_PER), on es poden identificar de forma clara aquests valors atípics:

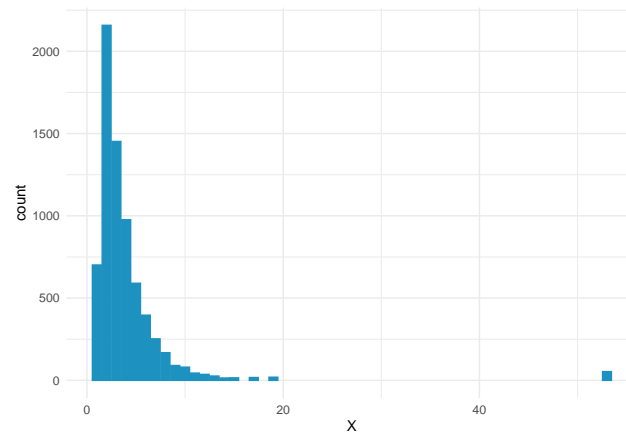


Figura 1. Histograma de la variable Nombre de persones

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
7087	1	2	3	4.015098	4.938707	5	53	3

Taula 8. Resum numèric de la variable Nombre de persones

Per tal d'assegurar-nos que aquesta dada no afecta al nostre anàlisi, i tenint en compte que disposem d'una base de dades molt gran, treurem aquests casos d'ambdues bases de dades.

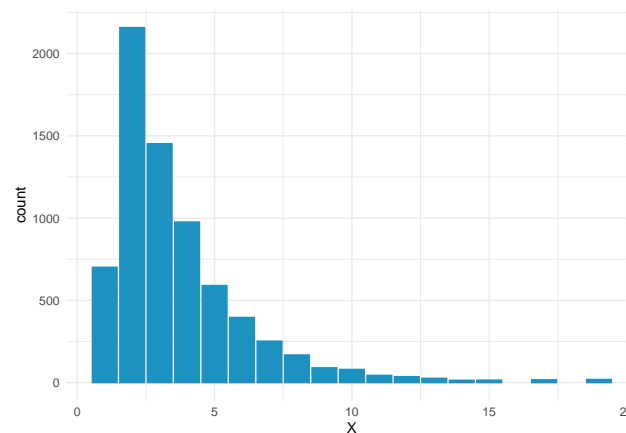


Figura 2. Histograma de la variable Nombre de persones després d'eliminar l'outlier

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
7034	1	2	3	3.646005	2.521077	4	19	2

Taula 9. Resum numèric de la variable Nombre de persones

Categoritzar

En el cas de les dades mancants que es troben en variables categòriques, el que es farà serà factoritzar-les i, seguidament, definir els nivells que presenta el factor. Així, per exemple, la variable **PER_TYP** presenta 8 nivells que s'han d'agrupar en 3 (*Conductor*, *Ocupant* i *Altres*).

A continuació es mostren els canvis realitzats a algunes de les variables categòriques de la base de dades:

PER_TYP: Tipus de persona (1 = conductor, 2 = ocupant, resta de codis = altres).

- Abans: 1, 2, 3, 4, 5, 6, 8, 9
- Després: Conductor, Ocupant, Altres

DAY_WEEK: Dia de la setmana (1 = Diumenge, 2 = Dilluns, . . . , 7 = Dissabte).

- Abans: 1, 2, 3, 4, 5, 6, 7
- Després: Diumenge, Dilluns, Dimarts, Dimecres, Dijous, Divendres, Dissabte

SEX: Sexe de la persona (1 = home, 2 = dona, 8 = No registrat, 9 = Desconegut).

- Abans: 1, 2, 8, 9
- Després: Home, Dona, Desconegut

Per la variable variable **SEX** hi ha una categoria anomenada “Desconegut”, que representa aquelles persones de les quals no tenim informació del seu sexe. Com aquesta categoria no ens aporta informació d'utilitat a l'hora de realitzar l'estudi ni per a relitzar models predictius, prescindirem dels individus que corresponguin aquesta categoria per a realitzar el nostre anàlisi.

RUR_URB: Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut).

- Abans: 1, 2, 6, 8, 9
- Després: Rural, Urbà, Desconegut

HI HA MORTS: Variable identificadora dels accidents mortals (0: no hi ha morts en l'accident, 1: hi ha morts en l'accident).

- Abans: 0, 1
- Després: No, Sí

DOA: Tipus de víctima (0 = sobrevis, 7 = mort a l'accident, 8 = mort al trasllat, 9 = Desconegut)

- Abans: 0, 7, 8, 9

- Després: Sobrevisu, Mor, Desconegut

Per aquesta última variable, DOA, hi ha una categoria anomenada “Desconegut”, que representa aquelles persones que no se sap si sobrevisuen a l’accident o no. Ja que en aquest estudi el fet de sobrevisure o no a l’accident és de gran interès, i aquesta categoria no ens aporta informació útil, prescindirem dels individus enmarcats en aquesta categoria per a realitzar el nostre anàlisi.

Per últim, es crearan dues variables noves a partir de HOURS i DAY, que ja són presents a ambdós conjunts de dades. Aquest pas es realitza perquè les variables HOURS i DAY tenen un rang de valors molt elevat que ens aporta poca informació.

HOURS_agrupat

En el cas de la variable HOURS, es tindrà en compte que, comunament, es considera que el dia està format per 5 intervals de temps segons la posició del sol. Aquest són la matinada (de les 0 a les 5 h incloses), el matí (de les 6 a les 11 h incloses), el migdia (de les 12 a les 14 h incloses), la tarda (de les 15 a les 19 h incloses) i la nit (de les 20 a les 23 h incloses). S’han fet servir aquests intervals per definir la nova variable HOURS_agrupat

- Abans: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23
- Després: Matinada, Matí, Migdia, Tarda, Nit

SETMANA

Pel que fa a la variable SETMANA, s’han definit les setmanes del mes en que es va realitzar el seguiment que presenten les dades. S’ha considerat el primer dia de la setmana el dilluns i l’últim el diumenge, tenint en compte que el dia 1 del mes era un dimarts. Per aquest motiu les setmanes 1 i 5 són les més curtes, especialment la cinquena, ja que el dia 31 va caure en dijous.

- Abans: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31
- Després: Setmana 1, Setmana 2, Setmana 3, Setmana 4, Setmana 5

Variable resposta

Per últim, definirem les variables resposta per a cada base de dades, és a dir, aquelles característiques que ens interessa poder predir tant en els futurs accidents com en les pròximes persones que es vegin involucrades en aquests.

Per una banda, és d’interès classificar els accidents segons si aquests han ocasionat morts o bé no ha sigut el cas. D’aquesta manera, es podria crear un model de predicció que permeti establir si un accident serà mortal o no en el futur en funció de les característiques que presenti.

Per tant, la variable d’interès és HIHAMORTS, que es mostra a la següent figura.

Variable	Stats / Values	Freqs (% of Valid)	Missing
Hi ha morts [factor]	1. No	1176 (42.3%)	0
	2. Sí	1604 (57.7%)	(0.0%)

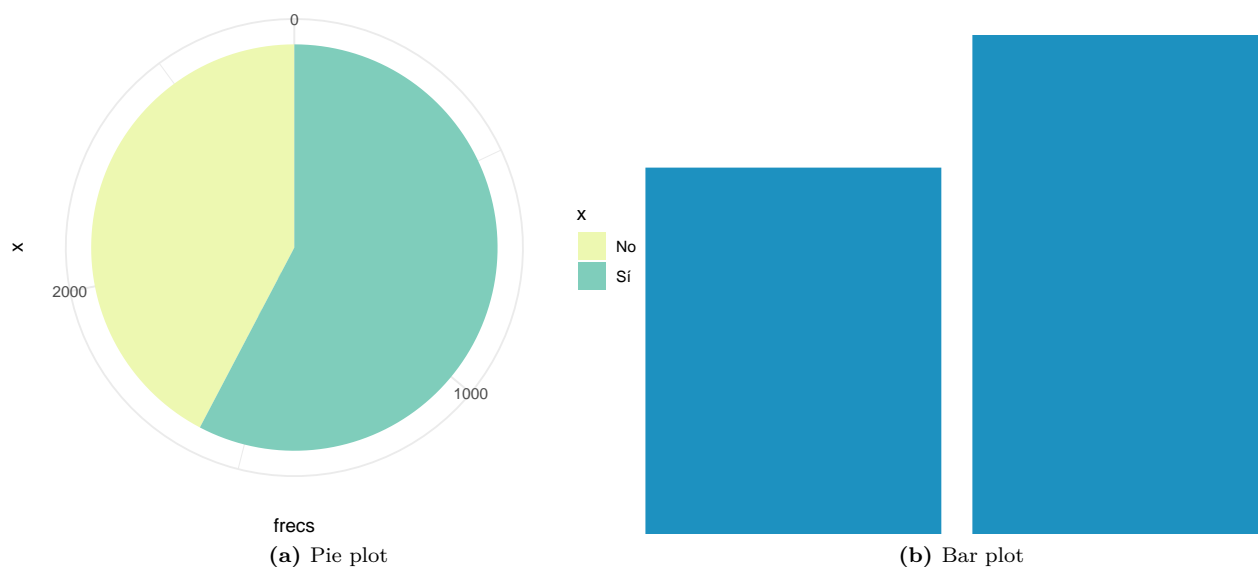


Figura 3. Anàlisi descriptiu de la variable Hi ha morts

Seguint aquesta línia, serà també de gran importància el tipus de víctima que esdevindran cadascuna de les persones implicades en un accident. En aquest cas, la variable d'interès serà **DOA**, de la qual es pot trobar un breu anàlisi descriptiu a la figura següent.

Variable	Stats / Values	Freqs (% of Valid)	Missing
Tipus de víctima [factor]	1. Sobreviu	5113 (74.1%)	0
	2. Mor	1791 (25.9%)	(0.0%)

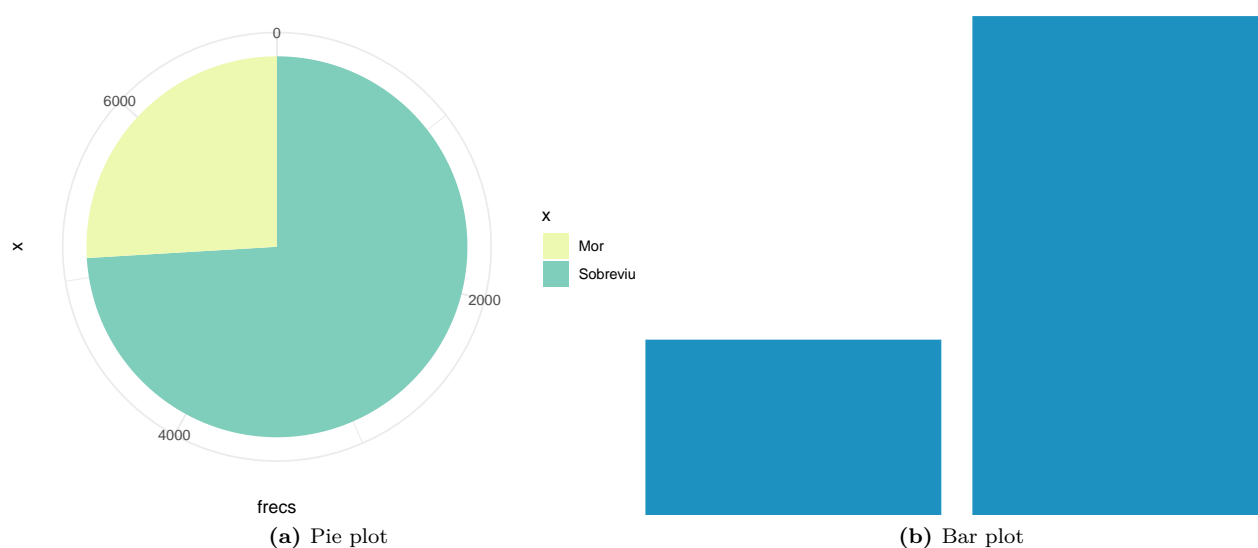


Figura 4. Anàlisi descriptiu de la variable Tipus de víctima

Anàlisi descriptiva univariant

Variables numèriques

Variables vinculades als accidents

Taula 12. Variables numèriques vinculades als accidents

Variable	Tipus	Descripció
HOURL	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident (99 = desconegut)
FATALS	Numèrica	Nombre de ferits a l'accident
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident
MORTS	Numèrica	Nombre de morts en l'accident
NO_PER	Numèrica	Nombre de persones implicades en l'accident
NO_VEHICLE	Numèrica	Nombre de vehicles implicats en l'accident

Variable	N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
HOURL	2780	0	7	15	13.0125899	6.7953310	18	23	11
MINUTE	2780	0	13	28	27.9665468	17.2769275	43	59	30
FATALS	2780	1	1	1	1.1000000	0.3832589	1	5	0
DRUNK_DR	2780	0	0	0	0.2456835	0.4429278	0	2	0
NO_PER	2780	0	0	1	0.6442446	0.6346020	1	5	1
MORTS	2780	1	1	2	2.5302158	1.6805367	3	19	2
NO_VEHICLE	2780	1	1	1	1.5079137	0.6997727	2	6	1

Taula 13. Resum de les variables numèriques vinculades als accidents

Per a comprovar que les dades són correctes i que les dades s'han tractat bé en termes de *missings* i *outliers*, farem ús del resum numèric de cada variable que es mostra a la taula superior (taula 13).

Les variables HOURL i MINUTE prenen valors entre 0 i 23 i entre 0 i 59 respectivament, els quals són rangs esperats i no donen peu a cap dada atípica en la seva distribució. A la tercera fila de la taula es troba el nombre de ferits a l'accident, FATALS, que senyala que no hi ha cap accident en què no hi hagi, com a mínim, un ferit. També veiem com el nombre màxim que ferits és 5, però en termes generals i en base als quartils 1 i 2 i la mitjana s'observa que la gran majoria dels accidents només presenten una persona ferida. Així mateix, pel que fa al nombre de conductors beguts involucrats a l'accident, DRUNK_DR, no involucren a cap conductor begut, és a dir, aquesta variable pren el valor 0 en la majoria dels casos (el 75% com a mínim, com indica el tercer quartil). Això no obstant, hi ha casos en que hi ha fins a 2 conductors beguts involucrats en un mateix accident.

-Mirar variables NO_PER i MORTS-

Per últim, en tots els accidents de trànsit hi ha un mínim d'un vehicle, i així ho mostra la variable NO_VEHICLE. L'accident amb més vehicles implicats en té 6.

En quant als *missings*, la primera columna de la taula és un recompte dels casos vàlids per a cada variable, és a dir, aquells que no presenten valors mancants per a aquella variable concreta. Es pot comprovar com el mètode de imputació s'ha realitzat correctament, perquè totes les variables presenten el mateix recompte de casos vàlids, que coincideix amb el nombre d'accidents total a la base de dades.

Variables vinculades a les persones

Taula 14. Variables numèriques vinculades a les persones

Variable	Tipus	Descripció
HOURL	Numèrica	Hora de l'accident (99 = desconeguda)
MINUTE	Numèrica	Minut de l'accident
FATALS	Numèrica	Nombre de ferits a l'accident
DRUNK_DR	Numèrica	Nombre de conductors beguts involucrats a l'accident
MORTS	Numèrica	Nombre de morts en l'accident
NO_PER	Numèrica	Nombre de persones implicades en l'accident
NO_VEHICLE	Numèrica	Nombre de vehicles implicats en l'accident
NO_FUGITS	Numèrica	Nombre de vehicles fugits implicats en l'accident
AGE	Numèrica	Edat de la persona

variable	N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
HOURL	6904	0	8	15	13.4419177	6.5218596	18	23	10
MINUTE	6904	0	13	28	27.8825319	17.2525855	43	59	30
FATALS	6904	1	1	1	1.1704809	0.5296297	1	5	0
DRUNK_DR	6904	0	0	0	0.2219003	0.4346383	0	2	0
NO_PER	6904	0	0	1	0.6965527	0.7340242	1	5	1
MORTS	6904	1	2	3	3.6548378	2.5291142	4	19	2
NO_VEHICLE	6904	1	1	2	1.7783893	0.8766397	2	6	1
NO_FUGITS	6904	0	0	0	0.0528679	0.2477490	0	3	0
EDAT	6904	0	23	37	39.8973059	20.3925982	55	98	32

Taula 15. Resum de les variables numèriques vinculades a les persones

En quant a les variables vinculades a les persones, només varien les variables **NO_FUGITS** i **EDAT** respecte a la taula anterior. De fet, la distribució de les variables **HOURL**, **MINUTE**, **FATALS**, **DRUNK_DR**, **NO_PER**, **MORTS** i **NO_VEHICLE** no presenta canvis rellevants respecte a la distribució que es troba a la taula anterior. No es troba cap cas de dada atípica ni cap dada mancanta per aquestes variables. Per aquest motiu, ens centrarem ara en els resums de les variables **NO_FUGITS** i **EDAT**.

Pel que fa al nombre de vehicles fugits a l'accident, el tercer quartil indica que al 75% dels casos no hi ha cap vehicle fugit, i si en fixem en el valor màxim que pren aquesta variable, el cas amb més vehicles fugits en presenta 3. Finalment, en quant a l'edat de les persones, el rang de valors va des de 0 a 98 anys, és a dir, en alguns accidents hi ha implicats nadons, i en d'altres persones de mitjana edat.

Novament, ens trobem amb que la primera columna de la taula és un recompte dels casos vàlids per a cada variable, és a dir, aquells que no presenten valors mancants per a aquella variable concreta. Es pot comprovar com el mètode de imputació s'ha realitzat correctament, perquè totes les variables presenten el mateix recompte de casos vàlids, que coincideix amb el nombre d'accidents total a la base de dades.

Variables categòriques

Variables vinculades als accidents

DAY: Dia de l'accident (de l'1 al 31). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
DAY [factor]	1. 1	72 (2.6%)	2780	0
	2. 2	84 (3.0%)	(100.0%)	(0.0%)
	3. 3	108 (3.9%)		
	4. 4	96 (3.5%)		
	5. 5	117 (4.2%)		
	6. 6	112 (4.0%)		
	7. 7	78 (2.8%)		
	8. 8	81 (2.9%)		
	9. 9	88 (3.2%)		
	10. 10	85 (3.1%)		
	[21 others]	1859 (66.9%)		

RUR_URB: Informació sobre la localització (1 = Rural, 2 = Urbà, 6 = Via no classificada, 8 = No registrat, 9 = Desconegut). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
RUR_URB [factor]	1. Rural	1174 (42.2%)	2780	0
	2. Urbà	1288 (46.3%)	(100.0%)	(0.0%)
	3. Desconegut	318 (11.4%)		

DAY_WEEK: Dia de la setmana (1 = Diumenge, 2 = Dilluns, . . . , 7 = Dissabte). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
DAY_WEEK [factor]	1. Diumenge	356 (12.8%)	2780	0
	2. Dilluns	311 (11.2%)	(100.0%)	(0.0%)
	3. Dimarts	410 (14.7%)		
	4. Dimecres	436 (15.7%)		
	5. Dijous	460 (16.5%)		
	6. Divendres	387 (13.9%)		
	7. Dissabte	420 (15.1%)		

HIHAMORTS:

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
HIHAMORTS [factor]	1. No	1176 (42.3%)	2780	0
	2. Sí	1604 (57.7%)	(100.0%)	(0.0%)

Variables vinculades a les persones

SEX: Sexe de la persona (1 = home, 2 = dona, 8 = No registrat, 9 = Desconegut). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
SEX	1. Home	4569 (66.2%)	6904	0
[factor]	2. Dona	2335 (33.8%)	(100.0%)	(0.0%)

PER_TYP: Tipus de persona (1 = conductor, 2 = ocupant, resta de codis = altres). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
PER_TYP	1. Conductor	4057 (58.8%)	6904	0
[factor]	2. Ocupant	2096 (30.4%)	(100.0%)	(0.0%)
	3. Altres	751 (10.9%)		

DOA: Tipus de víctima (0 = sobreviu, 7 = mort a l'accident, 8 = mort al trasllat, 9 = Desconegut). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
DOA	1. Sobreviu	5113 (74.1%)	6904	0
[factor]	2. Mor	1791 (25.9%)	(100.0%)	(0.0%)

Anàlisi descriptiva bivariant

Per acabar l'anàlisi descriptiva de les dades, s'estudiarà la relació que existeix entre diferents parells de variables. Aquest tipus d'anàlisi ajudarà a esbrinar si existeix una associació entre les variables i, en cas afirmatiu, quina és la força d'aquesta.

Per a la visualització d'aquestes relacions entre variables s'ha fet d'ús d'una plataforma de Google anomenada Data Studio. Aquesta permet convertir les dades en panells o informes complets, fàcils de llegir i de compartir i totalment personalitzables. Algunes de les seves funcionalitats són:

- Descriure les dades amb gràfiques, que inclouen gràfics de línies, de barres i circulars, mapes geogràfics, gràfics d'àrea i de bombolles, taules de dades dinàmiques i molt més.
- Permet que els informes siguin interactius amb filtres de visualització.
- Inclou enllaços i imatges en les quals es pot clicar per crear catàlegs de productes, biblioteques de vídeo i altres continguts amb hipervincles.
- Facilita l'anotació i descripció dels informes amb text i imatges.

A més de presentar totes aquestes característiques, amb Data Studio es poden elaborar fàcilment informes sobre dades procedents d'una gran varietat de fonts, sense necessitat de programar. En tan sols uns instants, permet la connexió a grans conjunts de dades com els que es troben a BigQuery.

Variables vinculades als accidents

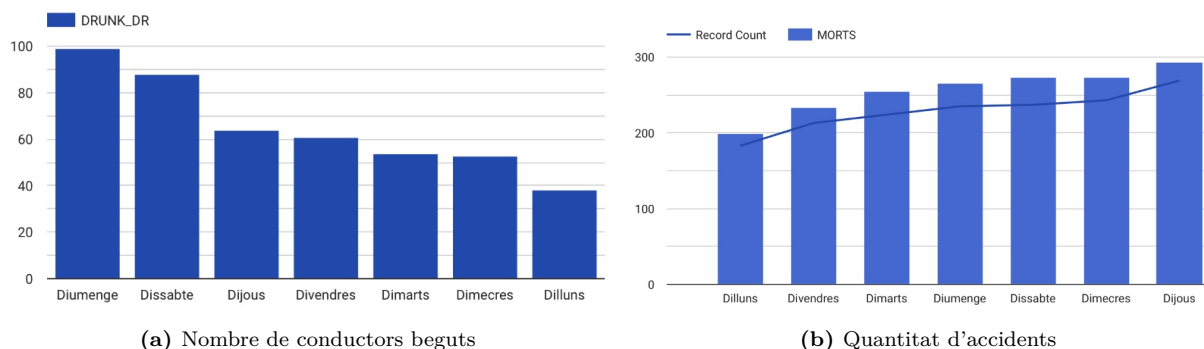


Figura 5. Segons el dia de la setmana

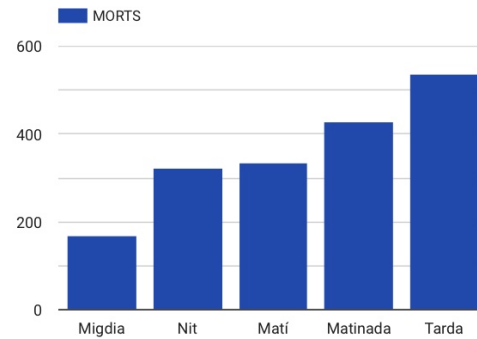
Pel que fa al nombre de conductors beguts segons el dia de la setmana en el qual succeeix l'accident, s'observa de forma clara a la gràfica esquerra de la figura 5 com la majoria dels conductors beguts es concentren al cap de setmana. Sembla que podria ser un patró perquè hi ha diferències notables entre la quantitat de conductor beguts a finals de la setmana, en comparació als dilluns, dimarts i dimecres. D'altra banda, si ens fixem, en aquest cas, en el nombre de morts segons el dia de la setmana, el dilluns es troba altra vegada en l'última posició, ja que és el dia en què es donen menys morts en accidents de trànsit. Paral·lelament, els últims dies de la setmana presenten un major nombre d'accidents mortals. Aquesta és la informació que presenta la gràfica dreta de la figura 5.

Si es té en compte la informació d'aquestes últimes gràfiques, es podria pensar que existeix una relació entre el nombre de conductors beguts i el nombre de ferits mortals als accidents de trànsit. S'haurà de tenir en compte aquesta hipòtesi per anàlisis posteriors de les dades.

Si centrem l'atenció en les hores del dia, a la figura 6 es poden veure les freqüències absolutes quant a la quantitat de morts en els diferents moments del dia. A l'esquerra, es divideix el dia en les seves hores, i



(a) Hores sense agrupar



(b) Hores agrupades

Figura 6. Nombre de morts en funció del moment del dia

s'observa com l'hora en què es produeixen més accidents és a les 6 de la tarda. Perquè la gràfica sigui més informativa, s'han agrupat les hores segons els moments del dia per crear la gràfica de la dreta.

En definitiva, de les dues gràfiques s'extreu que la majoria de les morts es produeixen a la tarda i a la matinada, mentre que el moment del dia on hi ha menys morts és el migdia.

Variables vinculades a les persones

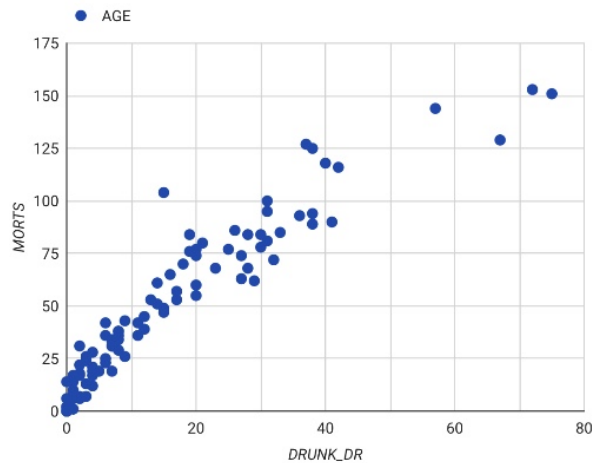


Figura 7. Nombre de morts segons el nombre de conductor beguts a les diferents edats

Pel que fa a la informació que tenim registrada sobre les persones involucrades en els accidents, i en la línia de les anàlisis anteriors, la figura 7 presenta una gràfica que mostra una clara relació entre el nombre de conductors beguts i el nombre de morts en l'accident a les diferents edats de les persones implicades. Els punts més extrems d'aquesta gràfica (els que presenten major nombre de conductors beguts i, alhora, major nombre de morts) són les edats 22 i 23. Aquestes dades indueixen a pensar que hi ha més perill d'accidents mortals per a la gent jove a la carretera, si hi ha conductors beguts. Una altra manera d'interpretar aquesta gràfica podria ser que hi ha més conductors joves que agafen el cotxe beguts i, en conseqüència, aquest grup d'edat pateix més accidents mortals.

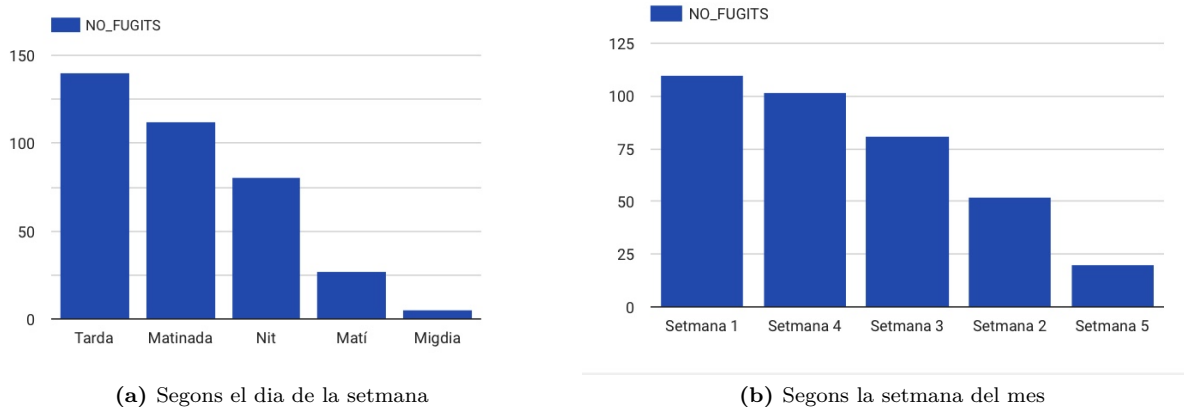


Figura 8. Nombre de vehicles fugits

El nombre de vehicles fugits en els accidents de trànsit varia en funció del moment del dia en que es produeix l'accident, així com en funció de la setmana del mes. A la figura 8 es troba, a l'esquerra, la quantitat de vehicles fugits en els diferents moments del dia, i aquests es concentren sobretot a la tarda (de 15 a 19h incloses) i la matinada (de les 0 a les 5h incloses). En canvi, pel que fa a les diferents setmanes del mes es veu com la majoria de cotxes fugits es troben a la primera i la quarta setmana, i la distribució no és uniforme durant totes les setmanes del mes, com s'esperaria si no hi hagués cap relació entre ambdues variables.

SEX	DOA / Record Count	
	Sobreviu	Mor
Home	3.296	1.273
Dona	1.817	518

Figura 9. Quantitat de supervivents segons el sexe

Pel últim, la figura 9 mostra una taula de contingència entre les variables SEX i DOA (és a dir, el sexe de la persona i el tipus de víctima, si sobreviu o no a l'accident). Aquesta taula marca com a categoria més abundant a la base de dades els homes que sobreviuen a l'accident, i com a menys abundant les dones que no sobreviuen al mateix.

Si es considera aquesta informació de forma relativa, aproximadament el 39% dels homes implicats en els accidents han siguts ferits mortals, mentre que d'entre les dones ho ha sigut aproximadament el 29%. S'hauria d'investigar si aquestes diferències són estadísticament significatives o, d'altra banda, no existeix relació entre el sexe i el tipus de víctima en els accidents de trànsit.

Anàlisi per Components Principals

La nostra base de dades depurada té un total de 7 variables numèriques. Per tant, l'anàlisi de components principals tindrà un total de 7 components. Després de realitzar els càlculs corresponents, obtenim que l'ACP de les variables numèriques és el següent:

```
## Standard deviations (1, ..., p=7):
## [1] 1.3664438 1.1888053 1.0064675 0.9950990 0.9377479 0.6779800 0.6142853
##
## Rotation (n x k) = (7 x 7):
##
##          PC1      PC2      PC3      PC4      PC5
## HOUR      0.019447174  0.36060003 -0.40650132  0.5760074  0.59452512
## MINUTE    -0.006351470 -0.01678656  0.89311808  0.3578663  0.27063661
## FATALS     0.526628413 -0.35520064 -0.08824177  0.2920504 -0.03536658
## DRUNK_DR   0.007473372 -0.41783466 -0.02765416 -0.5129940  0.74630364
## NO_PER     0.539705253  0.38704433  0.05638312 -0.1641500  0.04884373
## MORTS      0.450059841 -0.49739155 -0.07997349  0.2115722 -0.10206688
## NO_VEHICLE 0.477862437  0.41440423  0.13773378 -0.3463854  0.04767007
##
##          PC6      PC7
## HOUR     -0.13793603  0.008666826
## MINUTE    -0.01719641 -0.020243912
## FATALS     0.50414273  0.498000267
## DRUNK_DR   0.06336543 -0.021128405
## NO_PER     0.40916482 -0.599156992
## MORTS     -0.58191318 -0.387102848
## NO_VEHICLE -0.46525699  0.492165297
```

Sabem que cada component representa una inèrcia concreta. Ho podem veure gràficament en els següent gràfic de barres.

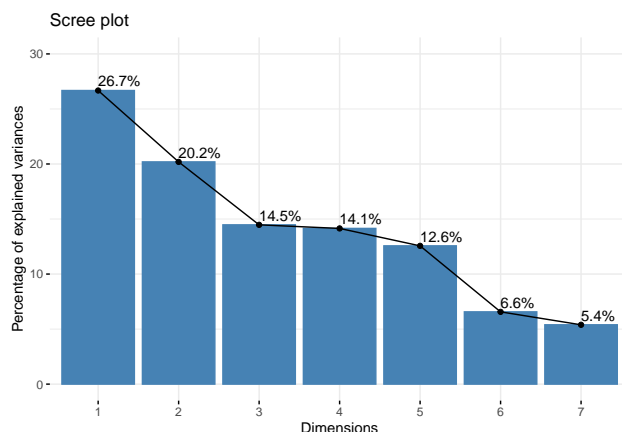


Figura 10. Barplot de la inèrcia de cada component

```
## [1] 26.67384 46.86324 61.33434 75.48037 88.04281 94.60934 100.00000
```

Tenint en compte que la inèrcia equival a la proporció de la variabilitat de les dades, sabem que amb un 80% d'inèrcia, podem obtenir gairebé tota la informació. Mirant el gràfic de la inèrcia acumulada, es pot veure que amb les 4 primeres components ja obtenim gairebé el 80% de la inèrcia, així que ens podem servir d'aquestes per al nostre anàlisi.

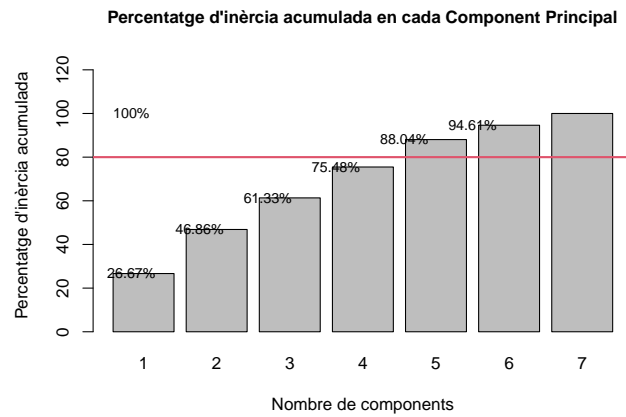


Figura 11. Barplot de la inèrcia acumulada

A continuació realitzem un gràfic de dispersió per a totes les combinacions possibles.

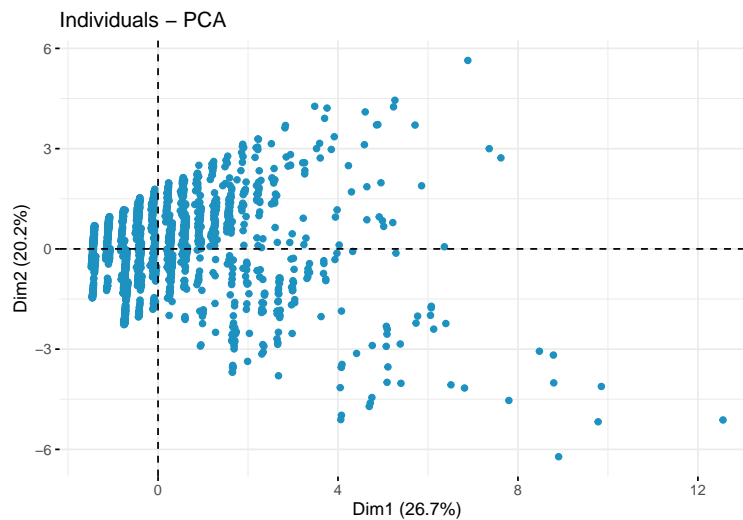
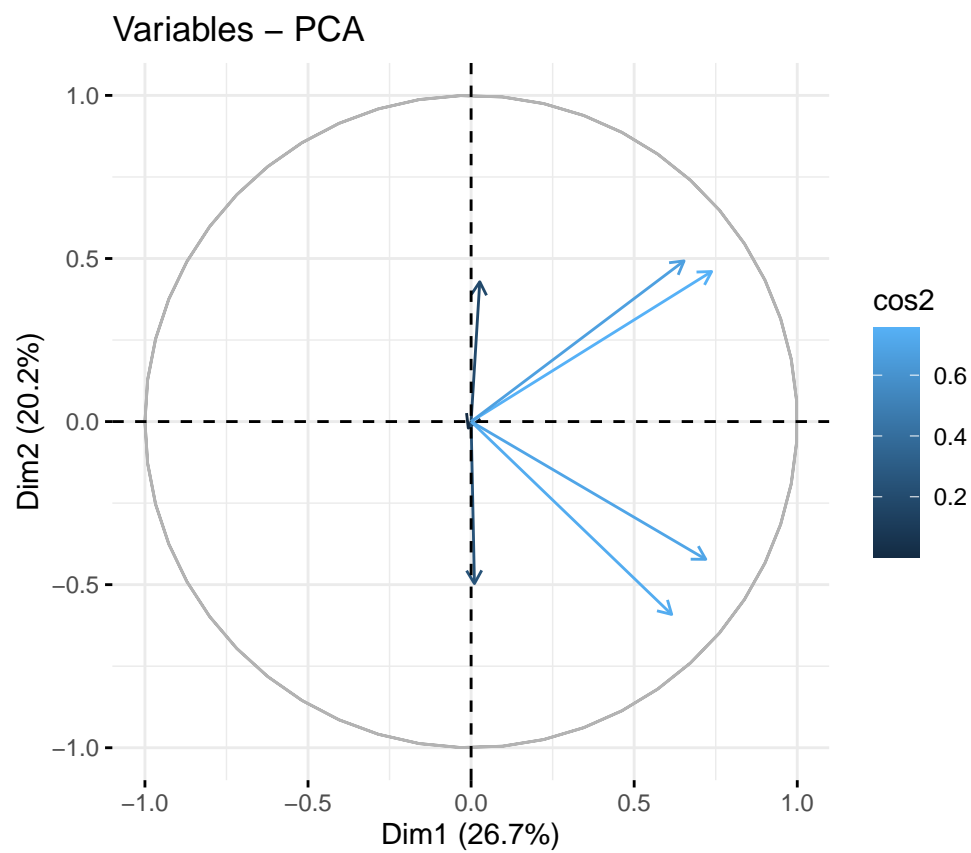
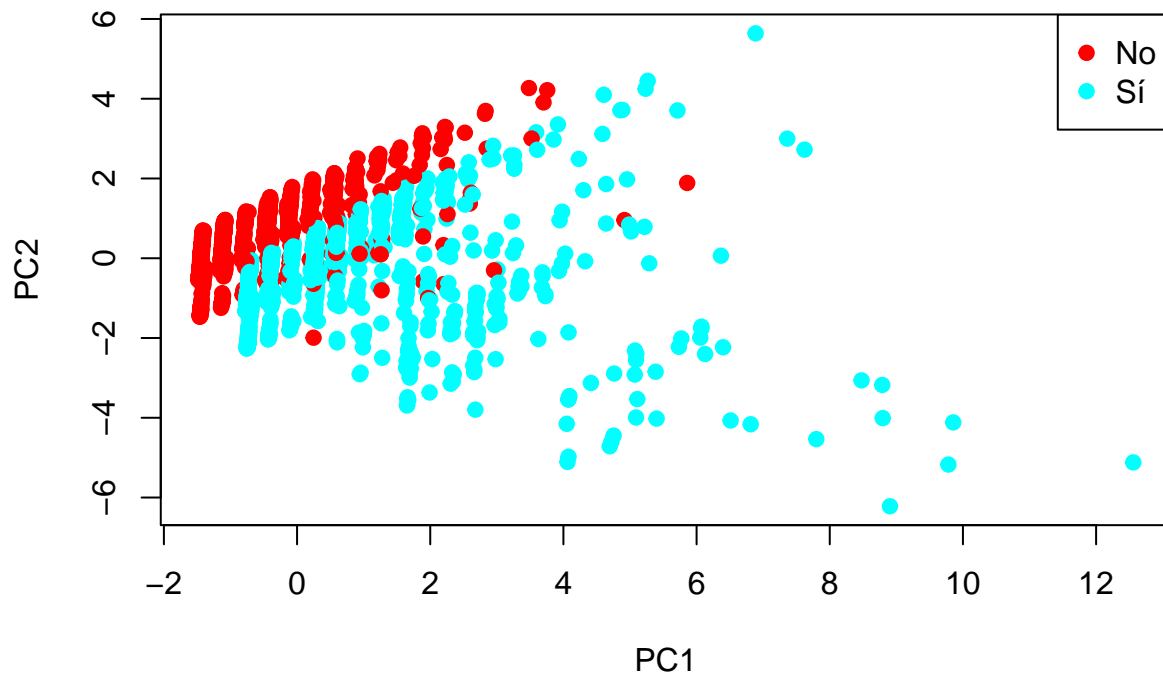


Figura 12. Gràfica de la projecció dels individus



Para los vectores (variables), nos fijamos en su longitud y en el ángulo con respecto a los ejes de las componentes principales y entre ellos mismos:

Ángulo: cuanto más paralelo es un vector al eje de una componente, más ha contribuido a la creación de la misma. Con ello obtenes información sobre qué variable(s) ha sido más determinante para crear cada componente, y si entre las variables (y cuales) hay correlaciones. Ángulos pequeños entre vectores representa alta correlación entre las variables implicadas (observaciones con valores altos en una de esas variables tendrá valores altos en la variable o variables correlacionadas); ángulos rectos representan falta de correlación, y ángulos opuestos representan correlación negativa (una observación con valores altos en una de las variables irá acompañado de valores bajos en la otra).

Longitud: cuanto mayor la longitud de un vector relacionado con x variable (en un rango normalizado de 0 a 1), mayor variabilidad de dicha variable está contenida en la representación de las dos componentes del biplot, es decir, mejor está representada su información en el gráfico.

Para los scores (observaciones), nos fijamos en los posibles agrupamientos. Puntuaciones próximas representan observaciones de similares características. Puntuaciones con valores de las variables próximas a la media se sitúan más cerca del centro del biplot (0, 0). El resto representan variabilidades normales o extremas (outliers). Por otro lado, la relación de las observaciones con las variables se puede estudiar proyectando las observaciones sobre la dirección de los vectores.

Projecció variables numèriques

En aquest gràfic es pot veure totes les variables numèriques representades en la primera i segona component. Veiem que la majoria de variables estan representades sobre l'eix horitzontal, que correspon a la primera component. Aquestes variables són `NO_VEHICLE`, `MORTS`, `FATALS` i `NO_PERSONES`. A més, aquestes dues últimes prenen un valor de gairebé 0.8 i per tant, són les que expliquen amb més precisió la primera component. Ens fixem que totes aquestes variables tenen relació amb el nombre de persones, i per tant, a la primera component li assignarem l'etiqueta de Nombre de persones involucrades. Pel que fa a l'eix vertical, només hi ha dues variables que estiguin una mica relacionades amb la segona component. Aquestes són `HOURL` i `DRUNK_DR`, que prenen un valor prop del 0.5. Com que a priori, aquestes dues variables no tenen gaire relació l'una amb l'altra, assignarem a la segona component l'etiqueta de Condicions en què es dona l'accident, ja que a simple vista cap de les dues destaca sobre l'altre en la seva aportació a la segona component.

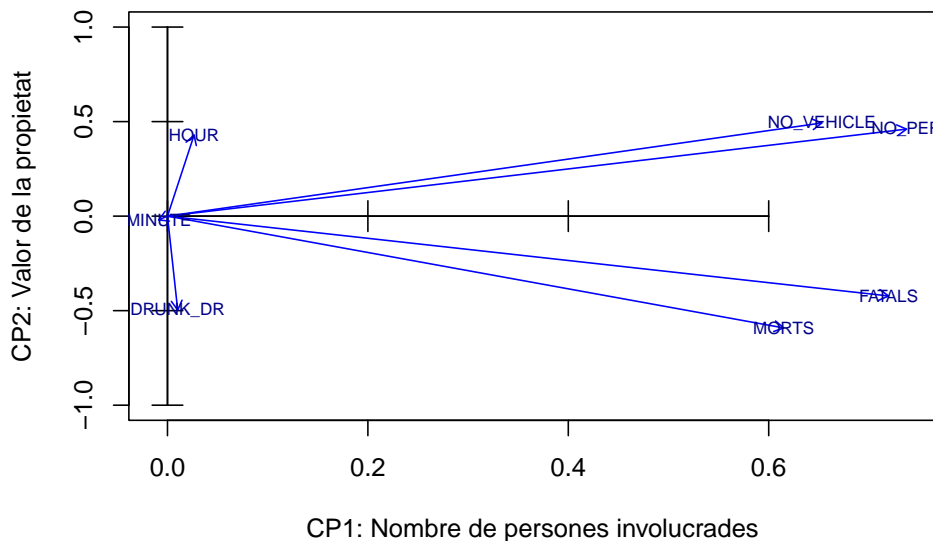
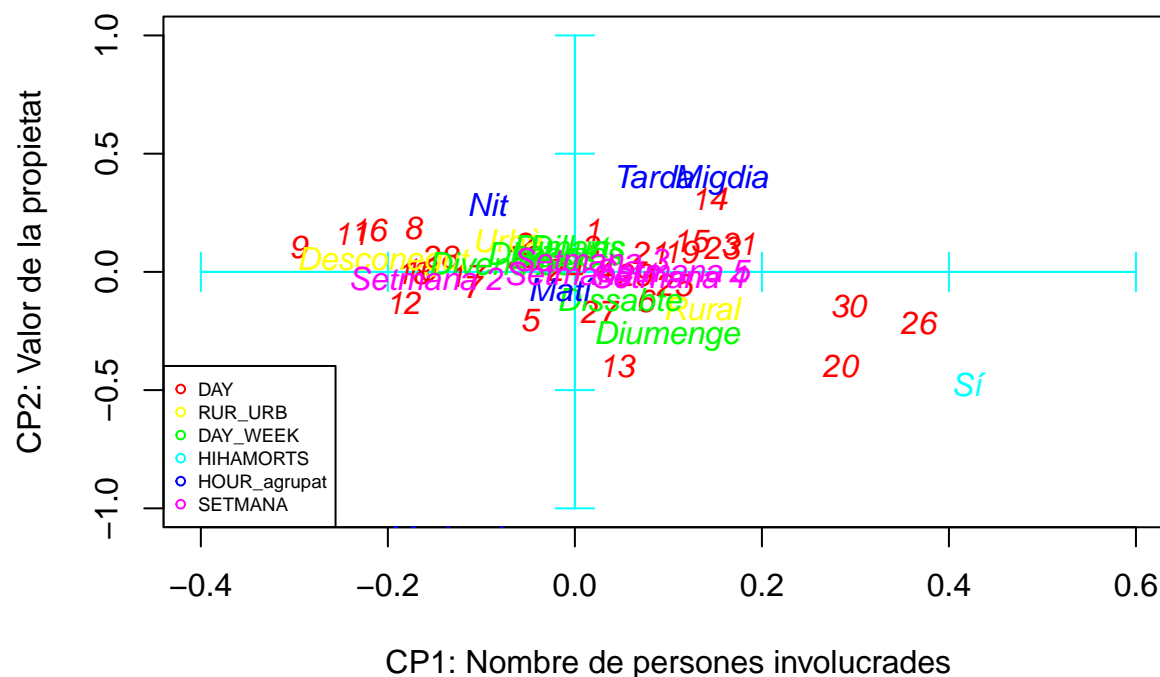


Figura 13. Gràfica de la projecció de variables numèriques

Projecció variables categòriques

Gràfics de projecció de les variables categòriques

Si en el mateix gràfic, hi afegim totes les categories de totes les variables categòriques, obtenim el següent:



Clarament aquest gràfic no es pot interpretar, ja que tenim tantes categories que no es poden distingir. Per tant, el que farem serà crear un gràfic per a cada variable categòrica, a veure si així ens és més fàcil interpretar-los.

Variable DAY

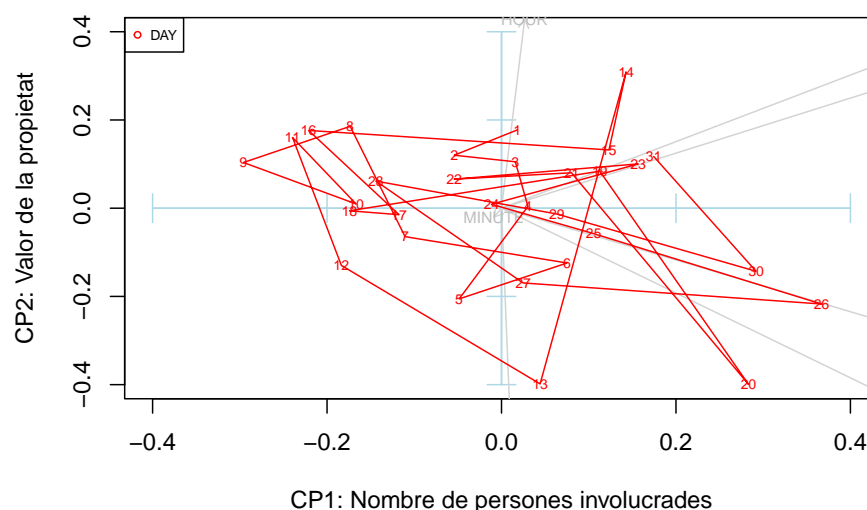


Figura 14. Gràfica de la projecció de la variable regió

Amb la variable **regió** veiem que les cinc categories es troben al centre del gràfic. Tot i així, veiem que la categoria **Oceania** es troba un pèl més paral·lela a la segona component, a diferència de les altres quatre, que estan més a prop de la primera component. Així doncs, encara que agafat molt amb pinces, podríem dir que la categoria **Oceania** explica millor la segona component i les altres quatre expliquen millor la primera.

Variable RUR_URB

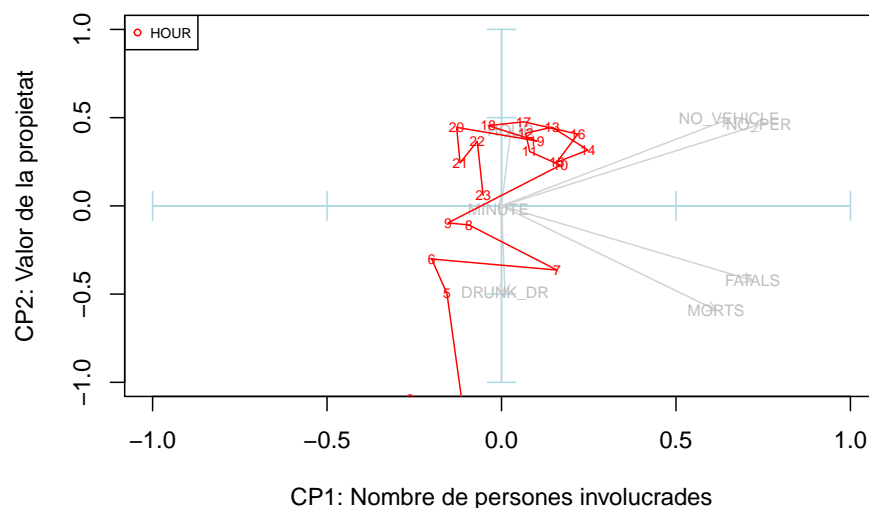


Figura 15. Gràfica de la projecció de la variable Èxit atac

Per la variable Èxit_atac podem veure clarament que la categoria No explica la segona component, ja que es troba exactament a sobre el seu eix. Tot i així, per la categoria Sí, podem dir que explica d'igual manera la primera i la segona component, ja que es troba just al centre dels dos eixos.

Variable Dia de la setmana

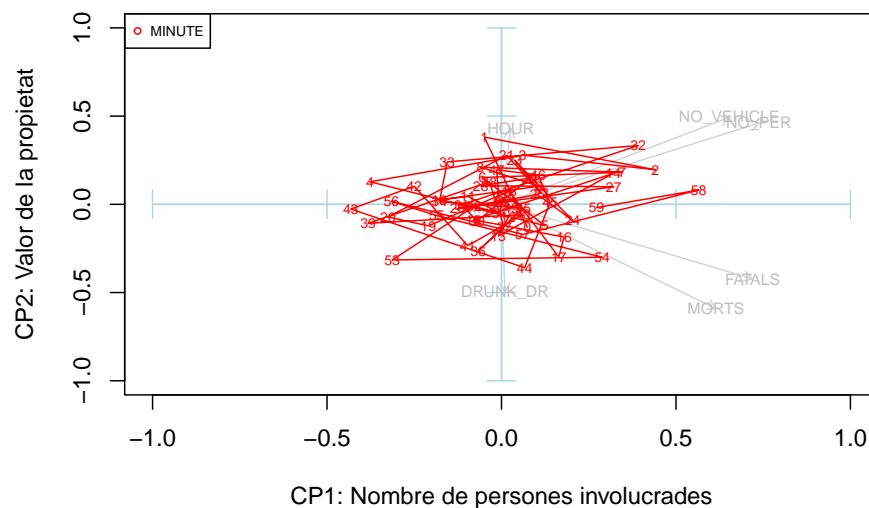


Figura 16. Gràfica de la projecció de la variable suïcidi

Per la variable Suïcidi ens trobem en una situació semblant a l'anterior. Aquí però, les dues categories expliquen millor la segona component, ja que es troben completament a sobre d'aquest eix. També podem dir que tenen una correlació semblant, és a dir, que expliquen la segona component de la mateixa manera, amb l'única diferència que la categoria No l'explica de forma negativa, mentre que la categoria Sí l'explica de forma positiva.

Variable HIHAMORTS

Per la variable Tipus_atac, tenim moltes categories, i la gran majoria es troben al centre de la gràfica, de manera que significa que expliquen d'igual manera la primera i segona component. Podem destacar tres categories, Assalt (armat o no), Atac a instal·lacions i Atac no armat, que destaquen sobre la resta.

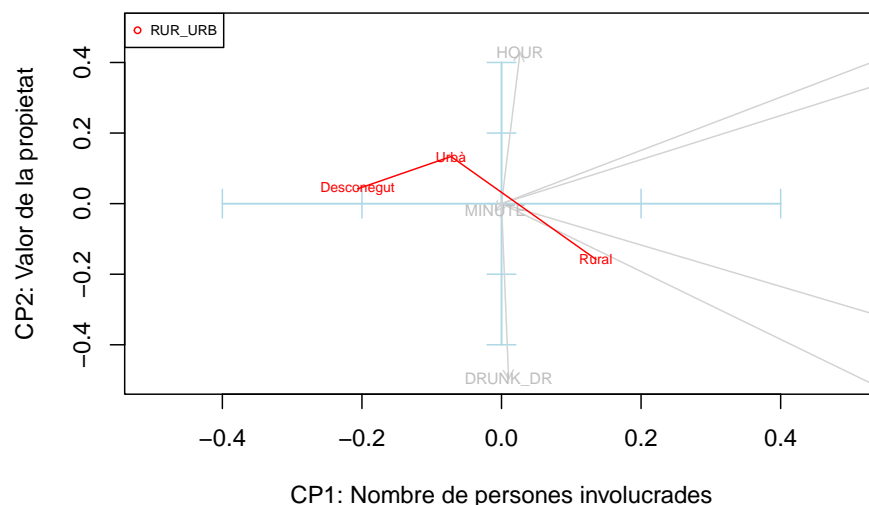


Figura 17. Gràfica de la projecció de la variable Tipus atac

Si ens fixem amb l'angle que formen entre les dues components, veiem que formen un angle més petit amb la primera, de forma que podem dir que aquestes tres categories expliquen millor la primera component.

Variable HOUR_agrupat

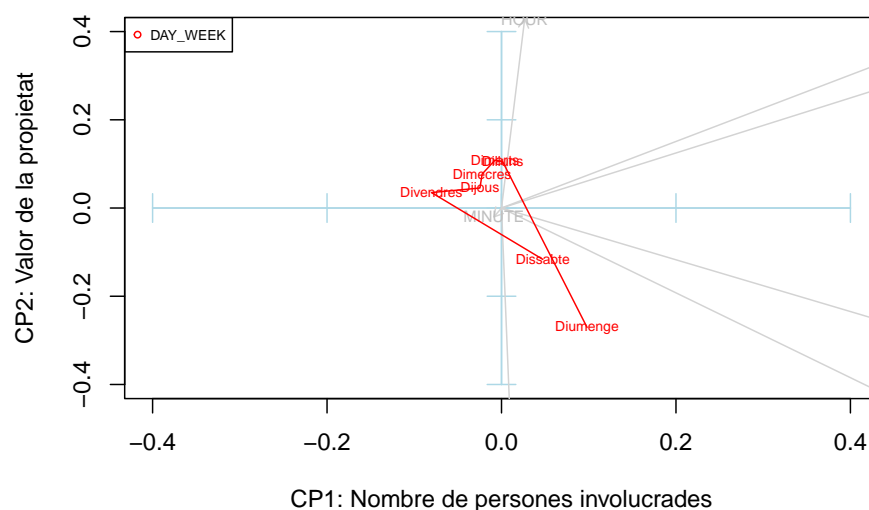


Figura 18. Gràfica de la projecció de la variable Tipus objectiu atac

Per la variable `Tipus_objectiu_atac` tenim encara més categories que la variable anterior, cosa que ens dificulta veure quines són les que expliquen millor una component o altra. Tot i així, podem destacar la categoria `Utilitats`, que seria la que millor explica la segona component, i les categories `Transport` i `Desconegut`, que serien les que millor expliquen la primera component.

Variable Setmana

Per la variable `Nacionalitat_objectiu` veiem com la gran majoria de categories expliquen millor la primera component, i aproximadament de la mateixa mesura. Tot i així, la categoria `desconegut` i `oceania` expliquen millor la segona, ja que es troben just o gairebé just a sobre l'eix de les ordenades.

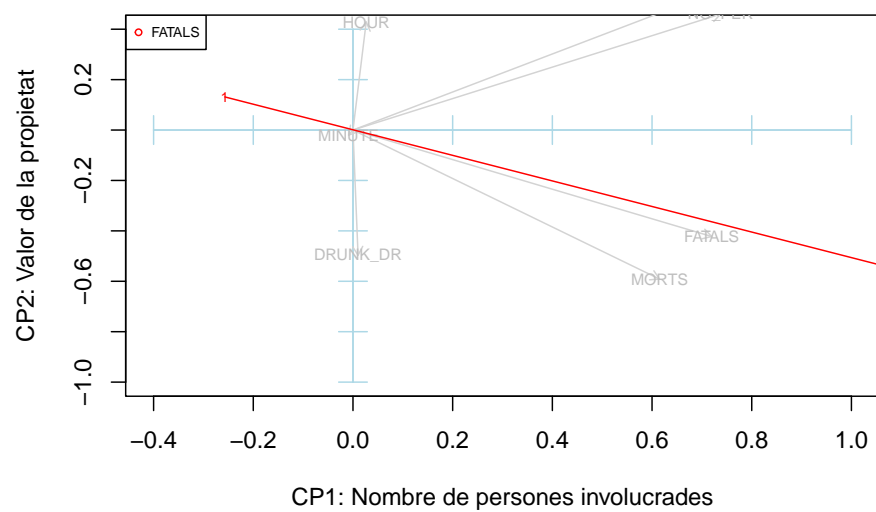


Figura 19. Gràfica de la projecció de la variable Nacionalitat objectiu