Anàlisi descriptiva

Anna Salazar

2022-07-28

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE
BARCELONA

Índex

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)
HOUR	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 = sobreviu, 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 regitrats, 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	Categòrica	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.

Las 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 tranformar-les, ja que les variables que presenten dades mancants les tenen codificades.

	NA	Percentatge de NA
DAY	0	0.00
HOUR	33	1.19
MINUTE	34	1.22
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
HIHAMORTS	0	0.00

	NA	Percentatge de NA
DAY	0	0.00
HOUR	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 (HOUR), 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 `HOUR` i `MINUTES`, mentre que per a la base de dades de persones, s'hi troben missings per a les variables `HOUR`, `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
HOUR	0	0
MINUTE	0	0

	NA	Percentatge de NA
HOUR	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:

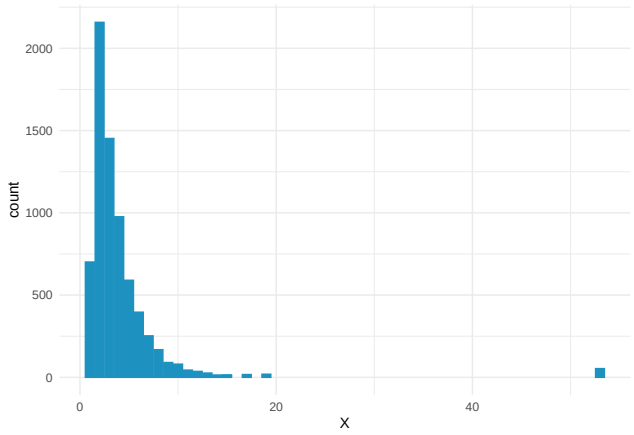


Figura 1. Histograma de la variable Nombre de persones

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 baeses de dades.

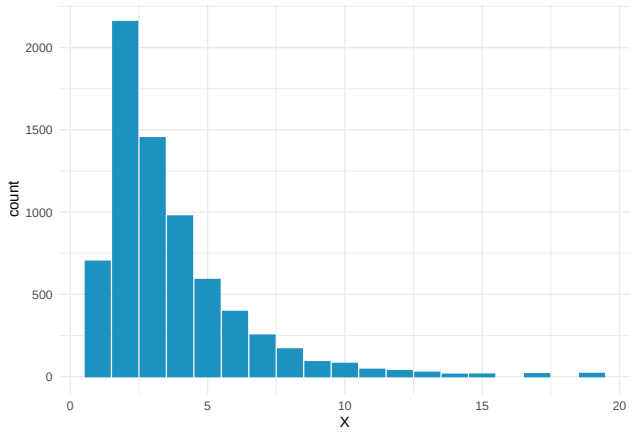


Figura 2. Histograma de la variable Nombre de persones després d'eliminar l'outlier

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

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 = sobreviu, 7 = mort a l'accident, 8 = mort al trasllat, 9 = Desconegut)

- Abans: 0, 7, 8, 9
- Després: Sobreviu, Mor, Desconegut

Per aquesta última variable, `DOA`, hi ha una categoria anomenada "Desconegut", que representa aquelles persones que no se sap si sobreviuen a l'accident o no. Ja que en aquest estudi el fet de sobreviure 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.

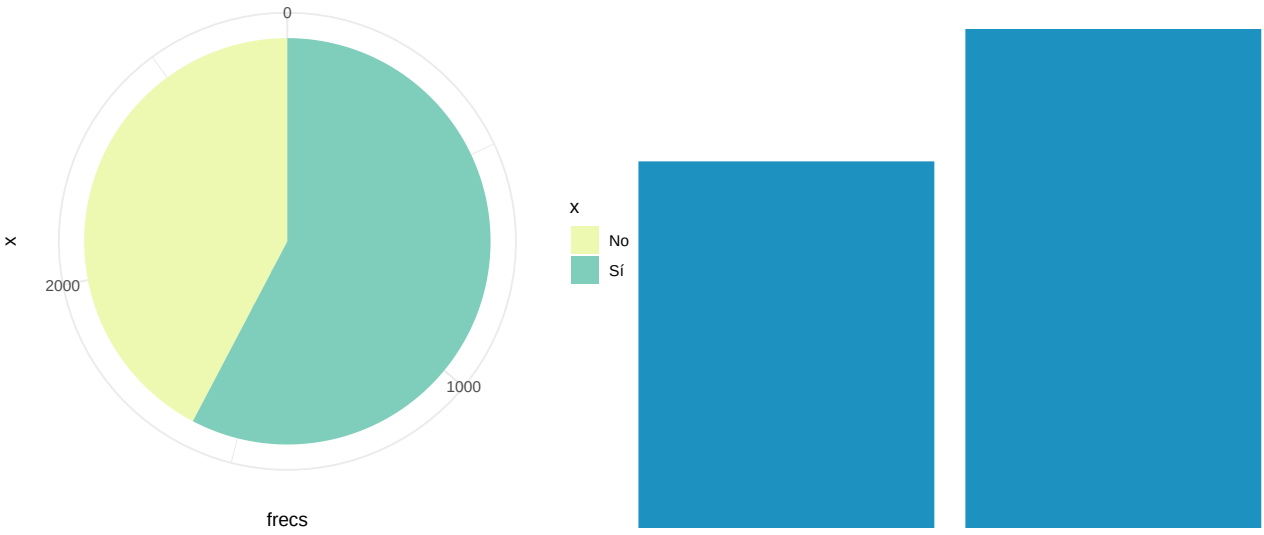
Variable resposta

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	1. No	1176 (42.3%)	0
[factor]	2. Sí	1604 (57.7%)	(0.0%)

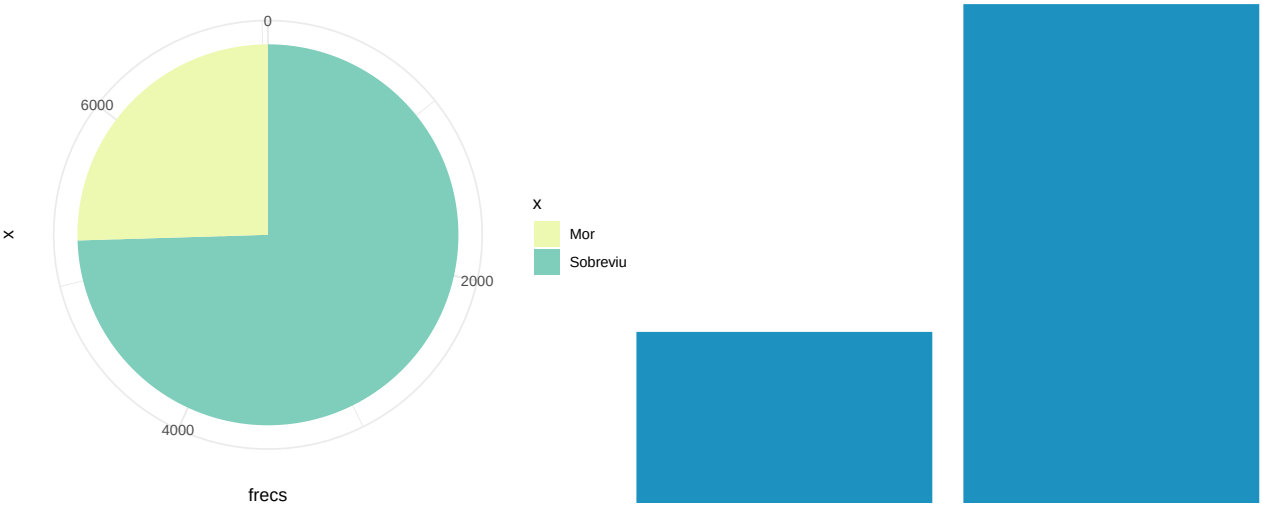


(a) Pie plot **(b)** Bar plot

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 figuara següent.

Variable	Stats / Values	Freqs (% of Valid)	Missing
Tipus de víctima	1. Sobreviu	5240 (74.5%)	0
[factor]	2. Mor	1791 (25.5%)	(0.0%)



(a) Pie plot **(b)** Bar plot

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
HOUR	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
HOUR	2780	0	7	15	13.0079137	6.7981176	18	23	11
MINUTE	2780	0	13	28	28.0140288	17.2367931	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	1	1	2	2.5302158	1.6805367	3	19	2
MORTS	2780	0	0	1	0.6442446	0.6346020	1	5	1
NO_VEHICLE	2780	1	1	1	1.5079137	0.6997727	2	6	1

Taula 13. Resum de les variables numèriques vinculades als accidents

Variables vinculades a les persones

Taula 14. Variables numèriques vinculades a les persones

Variable	Tipus	Descripció
HOUR	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
HOUR	7031	0	8	15	13.4528517	6.5492784	18	23	10
MINUTE	7031	0	13	28	27.8627507	17.2472133	43	59	30
FATALS	7031	1	1	1	1.1674015	0.5253148	1	5	0
DRUNK_DR	7031	0	0	0	0.2191722	0.4325411	0	2	0
NO_PER	7031	1	2	3	3.6468497	2.5212645	4	19	2
MORTS	7031	0	0	1	0.6935002	0.7307991	1	5	1
NO_VEHICLE	7031	1	1	2	1.7757076	0.8776740	2	6	1
NO_FUGITS	7031	0	0	0	0.0688380	0.2859146	0	3	0
EDAT	7031	0	24	37	39.9839283	20.3557940	55	98	31

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

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	1. 1	72 (2.6%)	2780	0
[factor]	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	1. Rural	1174 (42.2%)	2780	0
[factor]	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	1. Diumenge	356 (12.8%)	2780	0
[factor]	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	1. No	1176 (42.3%)	2780	0
[factor]	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 (65.0%)	7031	0
[factor]	2. Dona	2335 (33.2%)	(100.0%)	(0.0%)
	3. Desconegut	127 (1.8%)		

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	4158 (59.1%)	7031	0
[factor]	2. Ocupant	2117 (30.1%)	(100.0%)	(0.0%)
	3. Altres	756 (10.8%)		

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	5240 (74.5%)	7031	0
[factor]	2. Mor	1791 (25.5%)	(100.0%)	(0.0%)

Anàlisi descriptiva bivariant