

Anàlisi descriptiva

Anna Salazar

2022-07-20



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística



UNIVERSITAT_{DE}
BARCELONA

Índex

Descripció de la base de dades	1
Objectius del projecte	1
Preprocessament	2
Missings	2
Outliers	2
Variable resposta	2
Categoritzar	3
Anàlisi univariant	5
Variables numèriques	5
Variables categòriques	5

Descripció de la base de dades

La base de dades que serà utilitzada al llarg de l'estudi prové d'una companyia d'assegurances per a vehicles i conté 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.

Més concretament, a la base de dades es poden trobar les variables següents:

Taula 1. Llistat de variables de la base de dades

Variable	Tipus	Descripció
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
NO_PER	Numèrica	Nombre de persones implicades en l'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)
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) ¹ del vehicle quan va tenir l'accident (997,998 i 999 = Desconegut)
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:

- Representar les dades de manera nítida
- 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

Preprocessament

La base de dades d'aquest projecte està formada per 16 variables (columnes) i 12594 individus (files).

Les variables que tenim són AGE, SEX, PER_TYP, DOA, DAY, HOUR, MINUTE, RUR_URB, DAY_WEEK, FATALS, DRUNK_DR, NO_PER, NO_VEHICLES, HIT_RUN, TRAV_SP i PREV_SP.

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.

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.

	NA	Percentatge de NA
AGE	380	3.02
HOUR	55	0.44
MINUTE	64	0.51
FATALS	0	0.00
DRUNK_DR	0	0.00
NO_PER	0	0.00
NO_VEHICLES	0	0.00
TRAV_SP	7669	60.89

Taula 2. Percentatge de missings per variable

¹mph vol dir milles per hora. 100 mph = 161 Km/h

Outliers

Per fer

Variable resposta

Ja que un dels objectius del treball és el desenvolupament d'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, la variable depenent o resposta del treball és DOA.

A continuació es pot veure un breu anàlisi descriptiu de la variable:

Variable	Stats / Values	Freqs (% of Valid)	Missing
Tipus de víctima [factor]	1. 0	9798 (77.8%)	0 (0.0%)
	2. 7	2752 (21.9%)	
	3. 8	41 (0.3%)	
	4. 9	3 (0.0%)	

No hi ha cap dada mancanta en aquesta variable, encara que hi ha 3 persones de que quals el seu estat després de l'accident és desconegut.

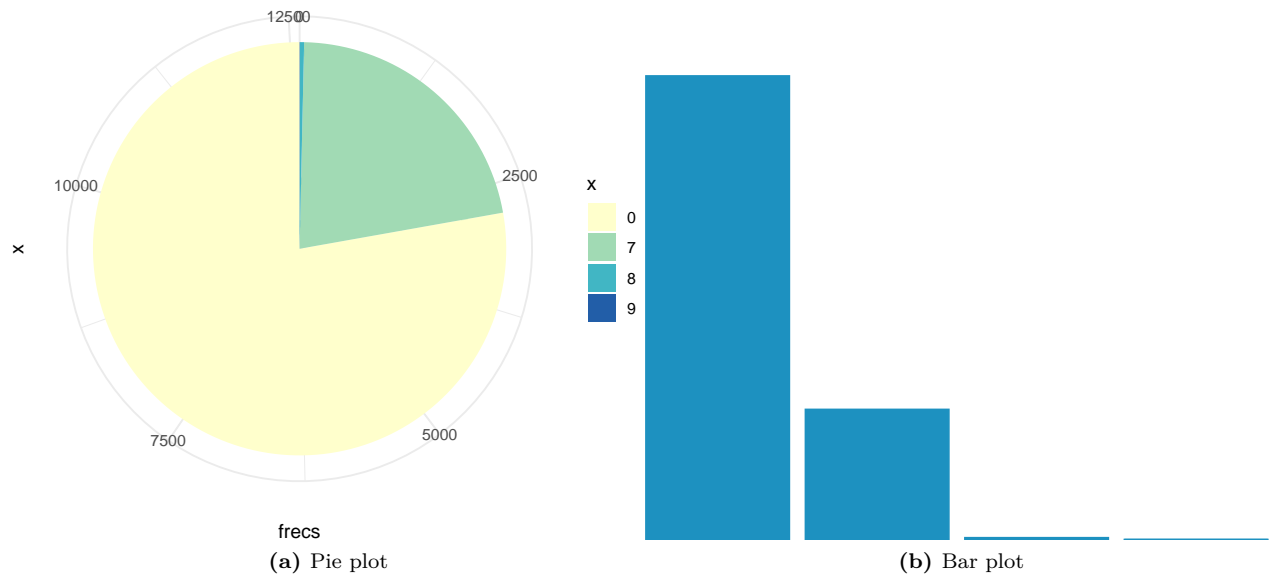


Figura 1. Anàlisi descriptiu de la variable Tipus de víctima

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] "1" "2" "3" "4" "5" "6" "8" "9"
```

Després

```
## [1] "Conductor" "Ocupant" "Altres"
```

HIT_RUN: Identificador de vehicle fugit (0 = No, 1 = Sí, 9 = Desconegut).

Abans

```
## [1] "0" "1" "9"
```

Després

```
## [1] "No" "Sí" "Desconegut"
```

PREV_SP: Indicador d'existència de límit de velocitat permesa just abans de l'accident (997,998 i 999 = Desconegut).

Abans

```
## [1] "0" "1" "2" "3" "4" "5" "7" "9" "99" "998"
```

Després

```
## [1] "0" "1" "2" "3" "4"
## [6] "5" "7" "9" "Desconegut"
```

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

Abans

```
## [1] "1" "2" "3" "4" "5" "6" "7"
```

Després

```
## [1] "Diumenge" "Dilluns" "Dimarts" "Dimecres" "Dijous" "Divendres"
## [7] "Dissabte"
```

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

Abans

```
## [1] "1" "2" "8" "9"
```

Després

```
## [1] "Home" "Dona" "No registrat" "Desconegut"
```

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

Abans

```
## [1] "1" "2" "6" "8" "9"
```

Després

```
## [1] "Rural" "Urbà" "Via no classificada"
## [4] "No registrat" "Desconegut"
```

Anàlisi univariant

Variables numèriques

HOURL: Hora de l'accident (99 = desconeguda). Tipus de variable: integer

MINUTE: Minut de l'accident (99 = desconegut). Tipus de variable: integer

FATALS: Nombre de ferits a l'accident. Tipus de variable: integer

DRUNK_DR: Nombre de conductors beguts involucrats a l'accident. Tipus de variable: integer

NO_PER: Nombre de persones implicades en l'accident. Tipus de variable: integer

AGE: Edat de la persona (998 = No registrada, 999 = Desconeguda). Tipus de variable: integer

NO_VEH: Nombre de vehicles implicats en l'accident. Tipus de variable: integer

TRAV_SP: Velocitat estimada (km) del vehicle quan va tenir l'accident (997,998 i 999 = Desconegut).
Tipus de variable: numeric

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12214	0	24.00	38.00	40.1055346	20.4484701	55.0	98.00	31.00
12594	0	0.00	0.00	0.2143878	0.4359339	0.0	2.00	0.00
12594	1	1.00	1.00	1.1928696	0.5653175	1.0	5.00	0.00
12539	0	9.00	15.00	13.5198182	6.2795629	18.0	23.00	9.00
12530	0	13.00	27.00	27.7267358	17.3687485	43.0	59.00	30.00
12594	1	2.00	4.00	4.6820708	5.2371006	5.0	53.00	3.00
12594	1	2.00	2.00	2.2074003	1.0811683	3.0	6.00	1.00
4925	0	40.25	74.06	68.4039147	40.9086117	96.6	215.74	56.35

Taula 4. Resum de les variables numèriques

Variables categòriques

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	339 (2.7%)	12594	0
[factor]	2. 2	413 (3.3%)	(100.0%)	(0.0%)
	3. 3	471 (3.7%)		
	4. 4	425 (3.4%)		
	5. 5	473 (3.8%)		
	6. 6	576 (4.6%)		
	7. 7	329 (2.6%)		
	8. 8	319 (2.5%)		
	9. 9	310 (2.5%)		
	10. 10	316 (2.5%)		
	[21 others]	8623 (68.5%)		

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	5138 (40.8%)	12594	0
	2. Urbà	6193 (49.2%)	(100.0%)	(0.0%)
	3. Via no classificada	37 (0.3%)		
	4. No registrat	1209 (9.6%)		
	5. Desconegut	17 (0.1%)		

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	1690 (13.4%)	12594	0
	2. Dilluns	1422 (11.3%)	(100.0%)	(0.0%)
	3. Dimarts	1848 (14.7%)		
	4. Dimecres	2001 (15.9%)		
	5. Dijous	1979 (15.7%)		
	6. Divendres	1680 (13.3%)		
	7. Dissabte	1974 (15.7%)		

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 [factor]	1. Home	7960 (63.2%)	12594	0
	2. Dona	4327 (34.4%)	(100.0%)	(0.0%)
	3. No registrat	149 (1.2%)		
	4. Desconegut	158 (1.3%)		

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 [factor]	1. Conductor	7620 (60.5%)	12594	0
	2. Ocupant	4083 (32.4%)	(100.0%)	(0.0%)
	3. Altres	891 (7.1%)		

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

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
DOA [factor]	1. 0	9798 (77.8%)	12594	0
	2. 7	2752 (21.9%)	(100.0%)	(0.0%)
	3. 8	41 (0.3%)		
	4. 9	3 (0.0%)		

HIT_RUN: Identificador de vehicle fugit (0 = No, 1 = Sí, 9 = Desconegut). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
HIT_RUN	1. No	12103 (96.1%)	12594	0
[factor]	2. Sí	485 (3.9%)	(100.0%)	(0.0%)
	3. Desconegut	6 (0.0%)		

PREV_SP: Indicador d'existència de límit de velocitat permesa just abans de l'accident (997,998 i 999 = Desconegut). Tipus de variable: factor

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
PREV_SP	1. 0	9415 (74.8%)	12594	0
[factor]	2. 1	1598 (12.7%)	(100.0%)	(0.0%)
	3. 2	534 (4.2%)		
	4. 3	167 (1.3%)		
	5. 4	62 (0.5%)		
	6. 5	14 (0.1%)		
	7. 7	3 (0.0%)		
	8. 9	1 (0.0%)		
	9. Desconegut	800 (6.4%)		