Anàlisi descriptiva

Anna Salazar

2022-07-14

Índex

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)
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
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 = sobreviu, 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 tranformar-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

[1]mph 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)	Valid	Missing
DOA	1. Sobreviu	9798 (77.8%)	12594	0
[factor]	2. Mort a l'accident	2752 (21.9%)	(100.0%)	(0.0%)
	3. Mort al trasllat	41 (0.3%)		
	4. Desconegut	3 (0.0%)		

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

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"
```

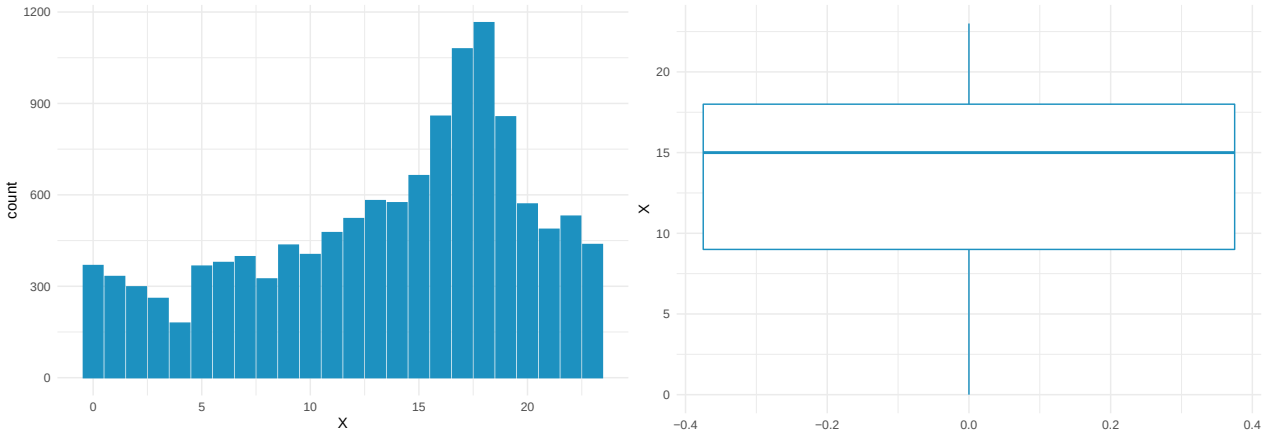
Anàlisi univariant

Variables numèriques

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12539	0	9	15	13.51982	6.279563	18	23	9

Taula 4. Resum numèric de la variable Hora de l'accident



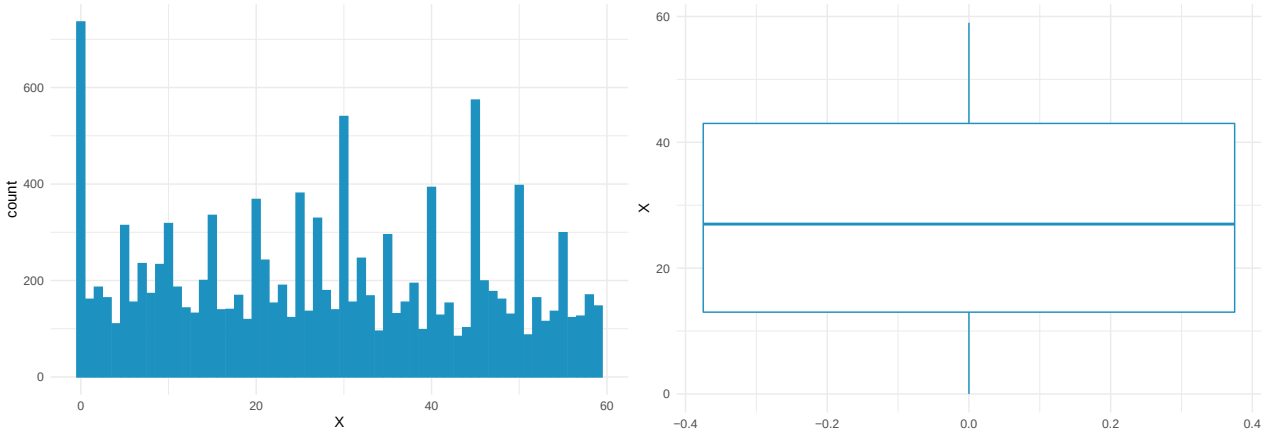
(a) Histograma **(b)** Boxplot

Figura 1. Anàlisi descriptiu de la variable Hora de l'accident

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12530	0	13	27	27.72674	17.36875	43	59	30

Taula 5. Resum numèric de la variable Minut de l'accident



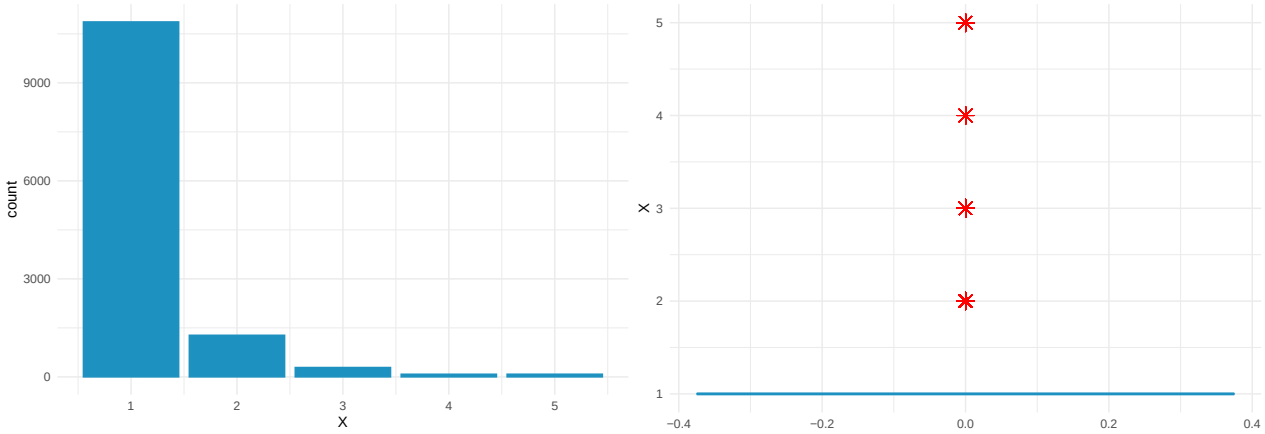
(a) Histograma **(b)** Boxplot

Figura 2. Anàlisi descriptiu de la variable Minut de l'accident

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12594	1	1	1	1.19287	0.5653175	1	5	0

Taula 6. Resum numèric de la variable Nombre de ferits a l'accident



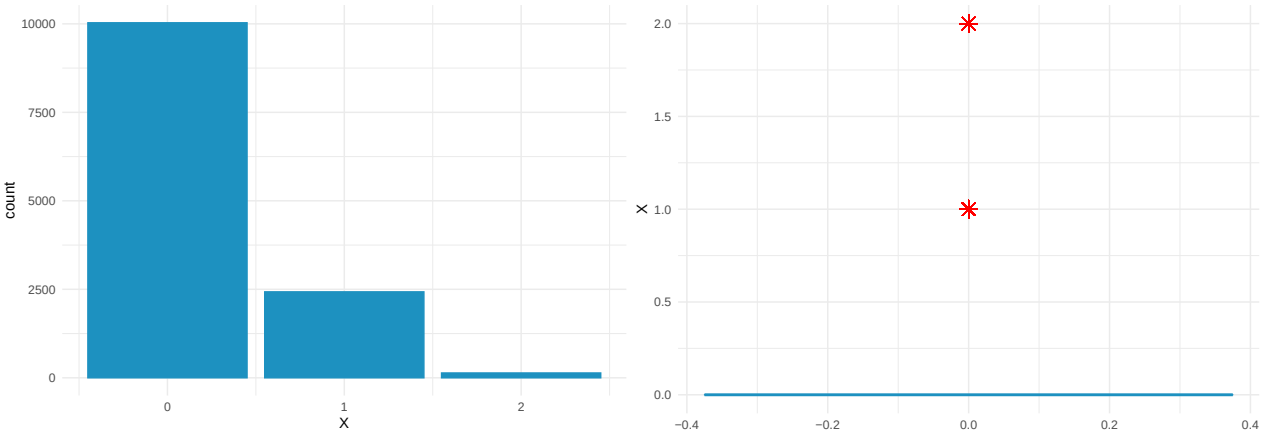
(a) Histograma **(b)** Boxplot

Figura 3. Anàlisi descriptiu de la variable Nombre de ferits a l'accident

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12594	0	0	0	0.2143878	0.4359339	0	2	0

Taula 7. Resum numèric de la variable Nombre de conductors beguts



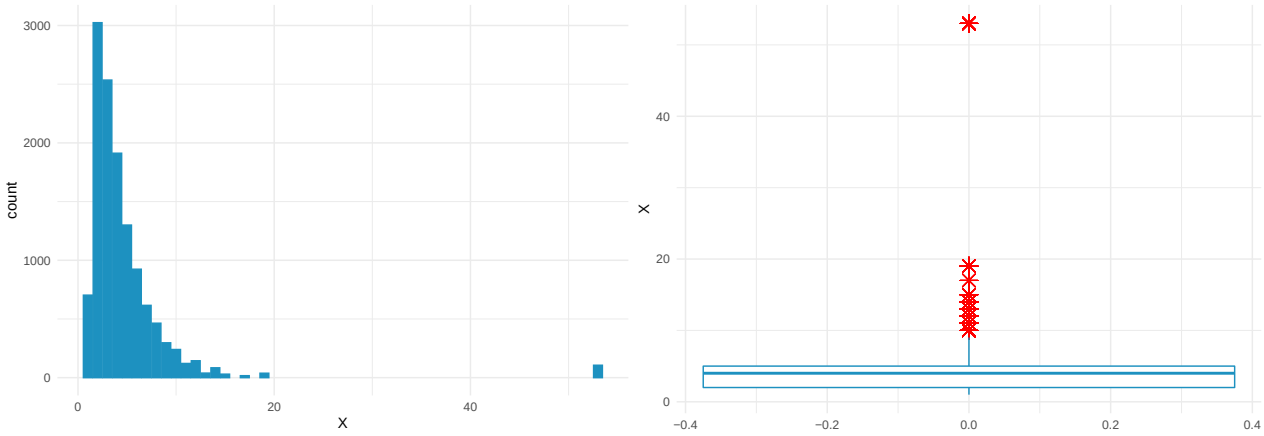
(a) Histograma **(b)** Boxplot

Figura 4. Anàlisi descriptiu de la variable Nombre de conductors beguts

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12594	1	2	4	4.682071	5.237101	5	53	3

Taula 8. Resum numèric de la variable Nombre de persones implicades en l'accident



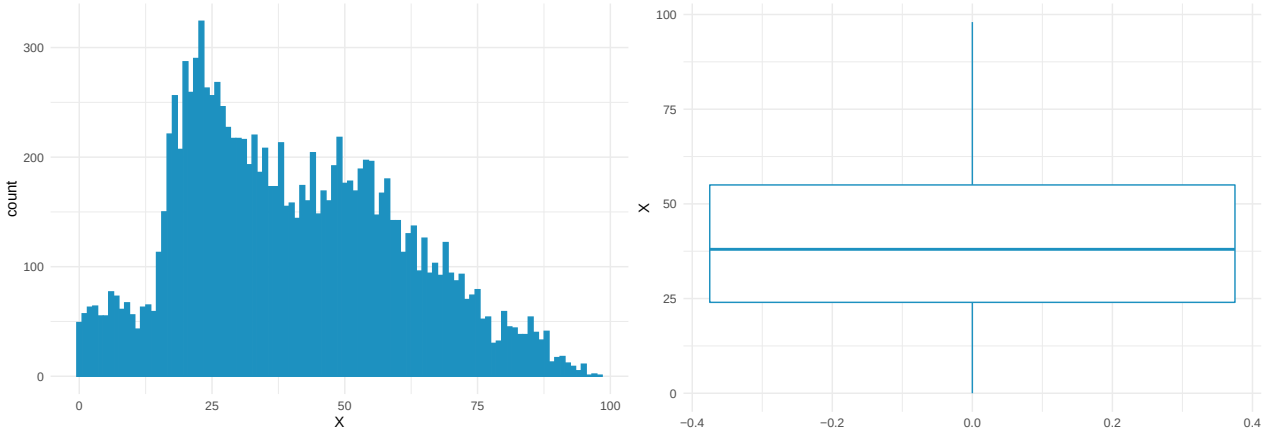
(a) Histograma **(b)** Boxplot

Figura 5. Anàlisi descriptiu de la variable Nombre de persones implicades en l'accident

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12214	0	24	38	40.10553	20.44847	55	98	31

Taula 9. Resum numèric de la variable Edat



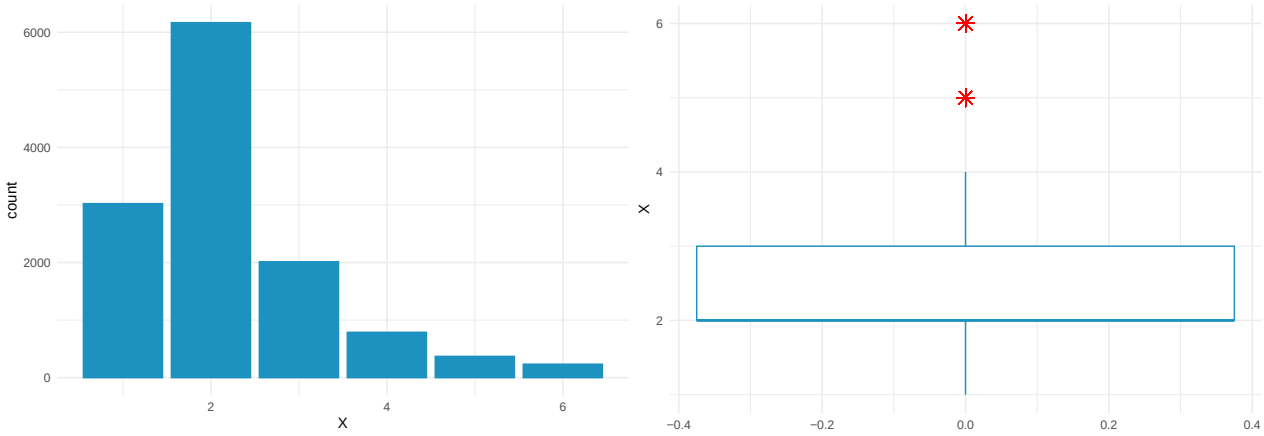
(a) Histograma **(b)** Boxplot

Figura 6. Anàlisi descriptiu de la variable Edat

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

N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
12594	1	2	2	2.2074	1.081168	3	6	1

Taula 10. Resum numèric de la variable Nombre de vehicles involucrats en l'accident



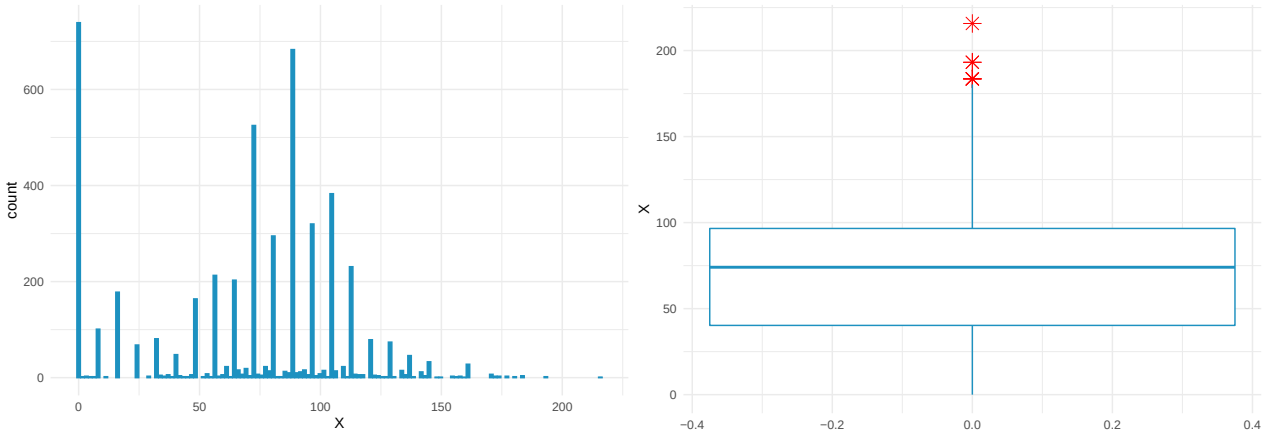
(a) Histograma **(b)** Boxplot

Figura 7. Anàlisi descriptiu de la variable Nombre de vehicles involucrats en l'accident

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
4925	0	40.25	74.06	68.40391	40.90861	96.6	215.74	56.35

Taula 11. Resum numèric de la variable Velocitat estimada



(a) Histograma **(b)** Boxplot

Figura 8. Anàlisi descriptiu de la variable Velocitat estimada

Variables categòriques

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

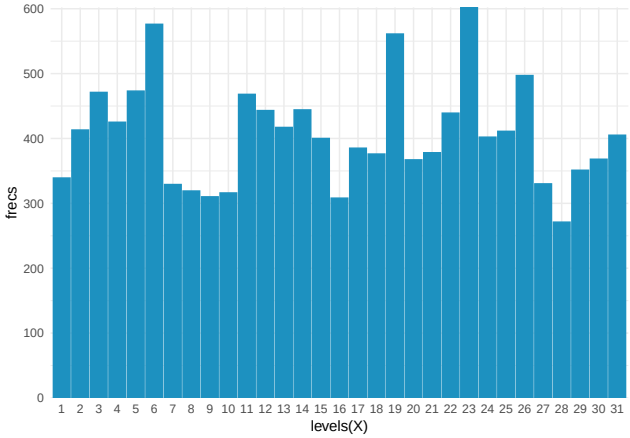
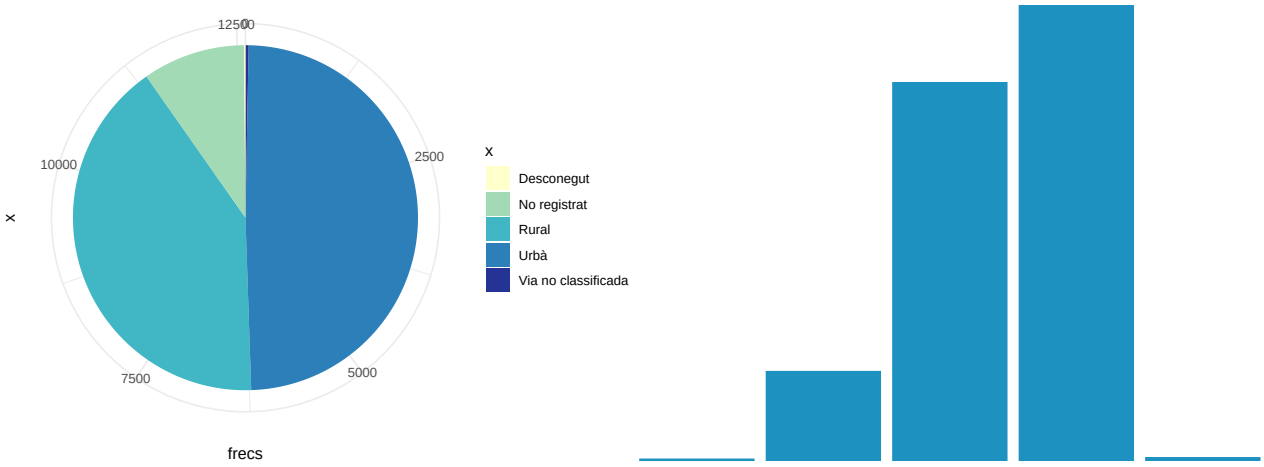


Figura 9. Anàlisi descriptiu de la variable Dia

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)	Missing
Localització [factor]	1. Rural	5138 (40.8%)	0 (0.0%)
	2. Urbà	6193 (49.2%)	
	3. Via no classificada	37 (0.3%)	
	4. No registrat	1209 (9.6%)	
	5. Desconegut	17 (0.1%)	

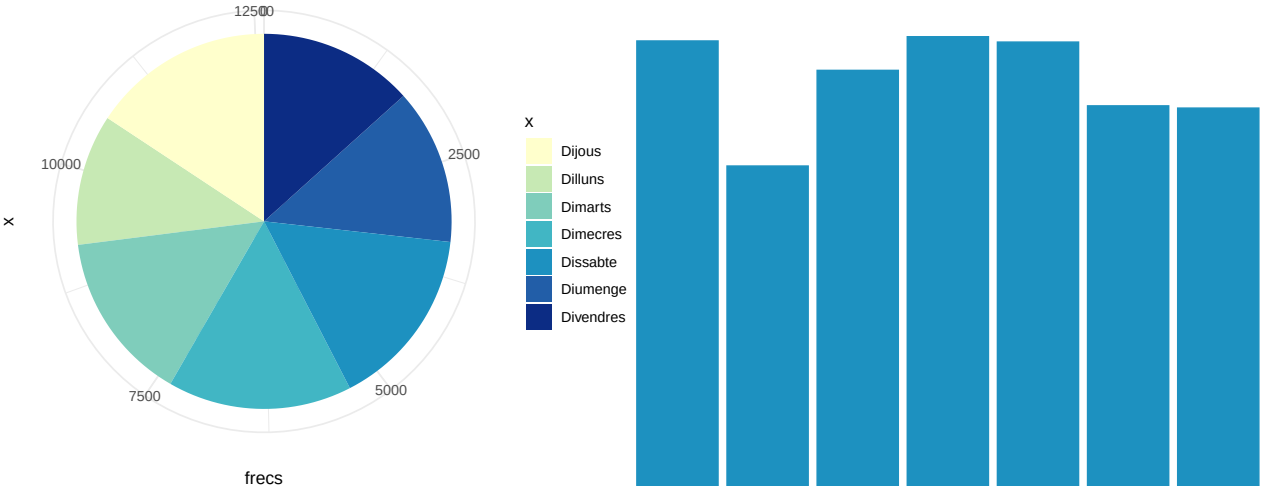


(a) Pie plot **(b)** Bar plot

Figura 10. Anàlisi descriptiu de la variable Localització

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

Variable	Stats / Values	Freqs (% of Valid)	Missing
Dia de la setmana [factor]	1. Diumenge	1690 (13.4%)	0 (0.0%)
	2. Dilluns	1422 (11.3%)	
	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%)	

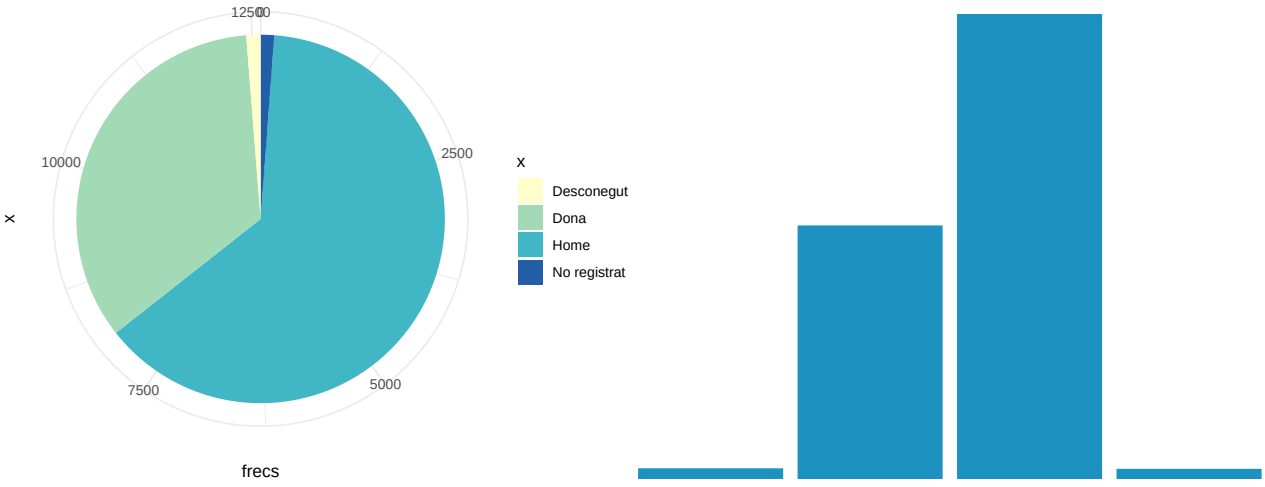


(a) Pie plot **(b)** Bar plot

Figura 11. Anàlisi descriptiu de la variable Dia de la setmana

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

Variable	Stats / Values	Freqs (% of Valid)	Missing
Sexe [factor]	1. Home	7960 (63.2%)	0 (0.0%)
	2. Dona	4327 (34.4%)	
	3. No registrat	149 (1.2%)	
	4. Desconegut	158 (1.3%)	

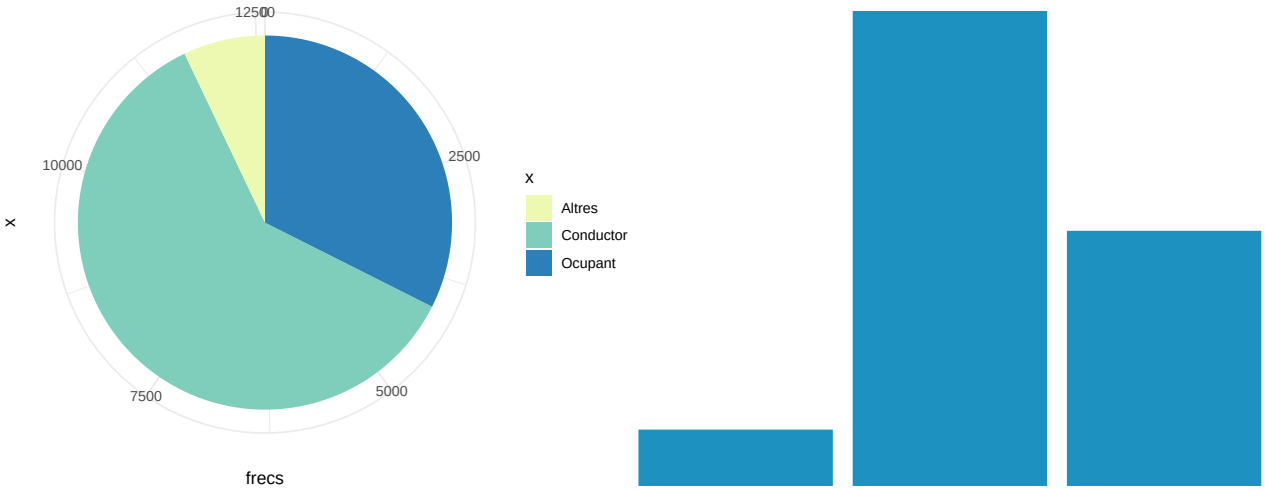


(a) Pie plot **(b)** Bar plot

Figura 12. Anàlisi descriptiu de la variable Sexe

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

Variable	Stats / Values	Freqs (% of Valid)	Missing
Tipus de persona [factor]	1. Conductor	7620 (60.5%)	0 (0.0%)
	2. Ocupant	4083 (32.4%)	
	3. Altres	891 (7.1%)	

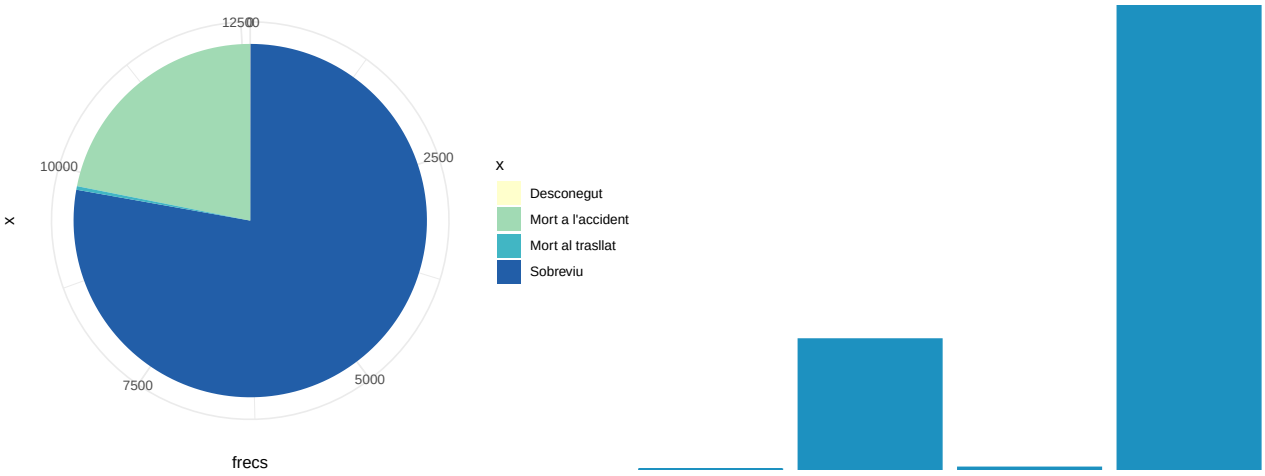


(a) Pie plot **(b)** Bar plot

Figura 13. Anàlisi descriptiu de la variable Tipus de persona

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)	Missing
Tipus de víctima [factor]	1. Sobreviu	9798 (77.8%)	0 (0.0%)
	2. Mort a l'accident	2752 (21.9%)	
	3. Mort al trasllat	41 (0.3%)	
	4. Desconegut	3 (0.0%)	

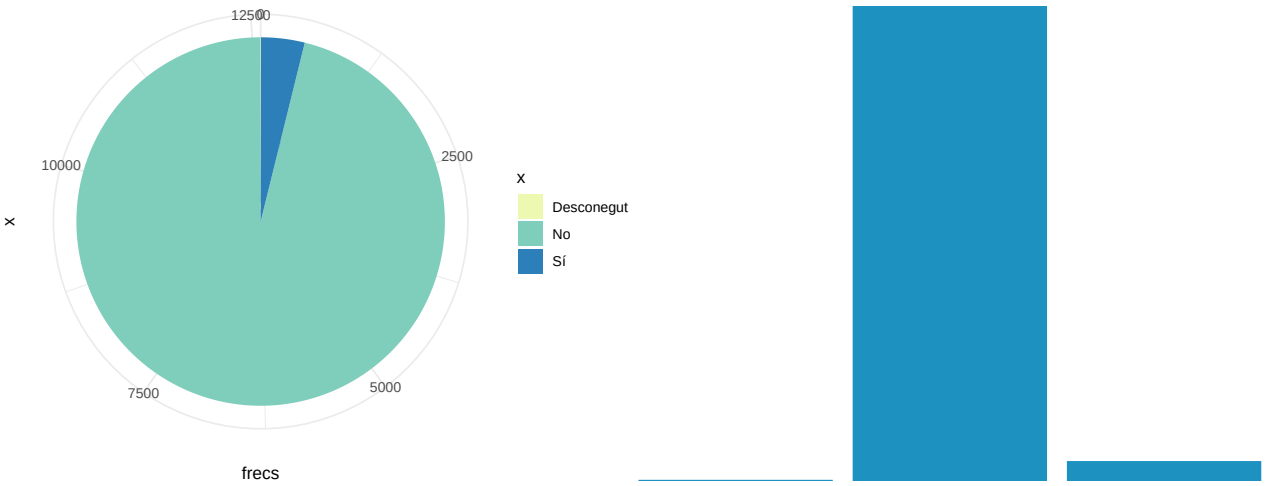


(a) Pie plot **(b)** Bar plot

Figura 14. Anàlisi descriptiu de la variable Tipus de víctima

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

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

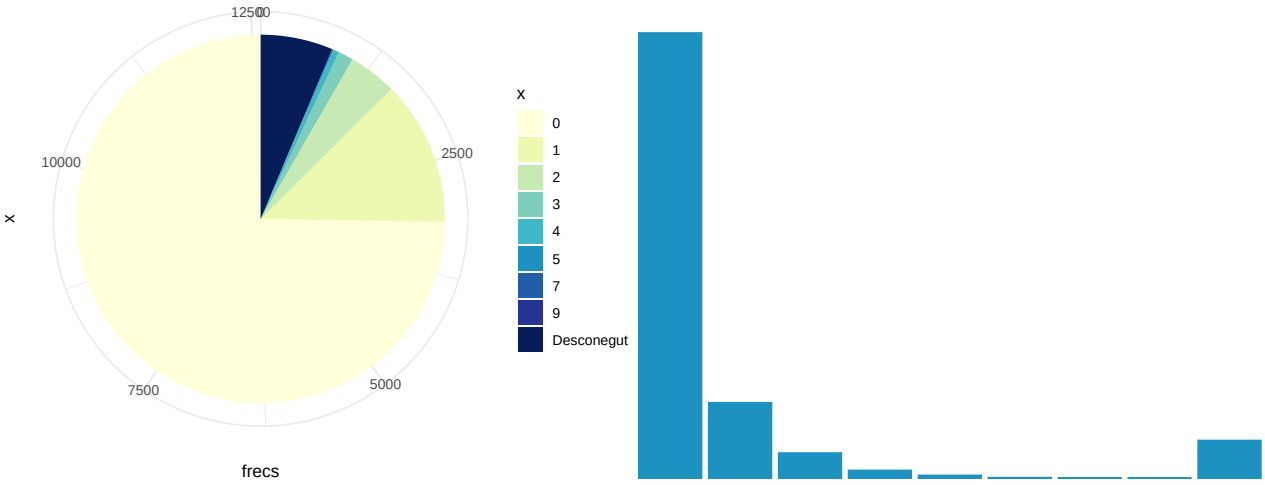


(a) Pie plot **(b)** Bar plot

Figura 15. Anàlisi descriptiu de la variable Identificació del vehicle

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)	Missing
Indicador d'existència de límit de velocitat permesa just abans de l'accident [factor]	1. 0	9415 (74.8%)	0 (0.0%)
	2. 1	1598 (12.7%)	
	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%)	



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

Figura 16. Anàlisi descriptiu de la variable Indicador d'existència de límit de velocitat