

PRÀCTICA 2 TIPOLOGIA DE DADES - RTP Puntualitat

Eva Solernou

juny 2020

Table of Contents

1. INTRODUCCIÓ	2
2. DESCRIPCIÓ DEL DATASET	2
2.1. Hores de pas per parada	3
2.2. Precipitació diària	5
2.3. Dades de trànsit.....	5
3. INTEGRACIÓ I SELECCIÓ DE LES DADES D'INTERÈS.....	7
4. NETEJA DE LES DADES	11
4.1. Zeros i elements buits	11
4.2. Valors extrems de la variable retard	15
4.3. Valors extrems en sensors	20
5. ANÀLISI DE LES DADES.....	28
5.1. Selecció dels grups de dades que es volen analitzar	28
5.2. Comprovació de la normalitat i homogeneïtat de la variància.	43
5.3. Aplicació de proves estadístiques per comparar els grups de dades.....	45
5.3.1. Contrast d'hipòtesis: la L20 té una distribució de retards diferent a les altres línies?.....	45
5.3.2. PCA.....	48
5.3.3. FAMD - Factor Analysis of Mixed Data	49
5.3.4. Kernel Regression	53
5.3.5. Model no supervisat: clustering: hi ha retards diferents en les parades? o en la combinació parada+línia? o parada+línia+expedició?.....	57
5.3.6. Correlació: hi ha relació entre els retards i els dies de pluja?.....	61
6. Representació dels resultats a partir de taules i gràfiques.....	63
7. Resolució del problema. A partir dels resultats obtinguts	63

1. INTRODUCCIÓ

Els autobusos urbans de Reus estan equipats amb un SAE (Sistema d'Ajuda a l'Explotació) que, esquemàticament, consisteix en una balissa GPS muntada a cada autobús que envia cada pocs minuts dades a un servidor, les quals són processades per un software amb múltiples objectius, entre ells estimar els temps de pas reals per parada i oferir-los als usuaris.

El sistema guarda registre històric d'algunes de les dades, fet que fa possible realitzar a posteriori anàlisis més complets del servei i plantejar millores en base als resultats obtinguts.

S'ha observat en anàlisis anteriors que la línia 20 concentra el major número d'incidències de servei i de queixes dels usuaris. És la línia que es realitza amb més vehicles i la més llarga, i es planteja la pregunta de si aquest fet és degut a que té més retards que les altres línies, o a altres motius com que té més hores de servei i més usuaris (i per tant més volum de queixes).

Per respondre aquesta pregunta hem descarregat del sistema els registres del pas per parada dels autobusos durant els mesos de febrer, març i abril de 2020 (arxiu 'HoraPasoParada.csv').

Per complementar l'anàlisi, hem incorporat dos conjunts de dades complementaris: per un costat les dades de precipitació a Reus (descarregades de la web de dades obertes d'aemet en format json), i per altra banda les dades d'unes càmeres de visió artificial que hi ha instal·lades en un punt del recorregut de la línia 20 que registren diversos paràmetres relacionats amb la fluïdesa del trànsit ('sensorsRiera.csv').

Així doncs, aquesta pràctica pretén donar resposta a les següents preguntes:

- A. Quina distribució segueix la puntualitat de pas en cada parada?
 - B. La línia 20 té més retards que la resta?
 - C. Els retards es veuen afectats per la quantitat de pluja?
 - D. En quina mesura afecta la congestió del trànsit?
-

2. DESCRIPCIÓ DEL DATASET

El conjunt de dades amb el que treballarem està format per tres datasets que caldrà relacionar, i que es poden descarregar en aquests enllaços:

<https://drive.google.com/open?id=1cfC4JjEQVNFEN5gHinprNtCBHuyEmQQr>

<https://cloud.amersam.net/index.php/s/GVrs7W5sHql8G8b/download>

2.1. Hores de pas per parada

L'arxiu s'ha extret de la base de dades del SAE, en concret totes les dades dels mesos de febrer, març i abril d'aquest any. Consta d'un total de 300.556 registres.

No es disposa d'una explicació dels camps, el seu significat s'ha deduït del nom dels camps i del coneixement del sistema. Els que poden ser interessants per l'anàlisi són els següents:

- dtHora_BD: moment previst d'arribada de l'autobús a la parada, segons horaris teòrics (publicats a les guies)
- dtHora_LlegadaReal: moment real d'arribada a la parada
- iIdExpedicion: cada valor representa una 'expedició' concreta, una 'volta' de l'autobús al conjunt de parades que ha de recórrer
- iIdParada: codi de la parada objecte del registre (integrarem el nom de parada provinent d'una altra taula)
- iIdLinea: codi de la línia que està realitzant el vehicle
- iIdTrayecto: codi que defineix el recorregut, que pot variar segons l'hora del dia. El més habitual són el '1' (recorregut complet), el '201' (part de la volta a l'inici de servei) i el '301' (part de la volta al final del servei diari)
- iIdConductor: codi del conductor

Variables de les quals desconexim el significat, pot estar relacionat amb el codi de colors que es mostra en pantalla de l'operador, s'intentarà deduir de l'anàlisi descriptiu de dades:

- iTipoAccion: Els valors registrats són 0,1,2,3
- iEstadoHRLlegada: Els valors registrats són 0,1,2,3,4
- iEstadoHRSalida: Els valors registrats són 0,1,2,3,4,5
- iOrdenExpedicion: Indica l'ordre de pas de cada parada en una expedició concreta

```
dades <- as_tibble(read.csv(file = "HoraPasoParada.csv", header = TRUE, sep = ","))
```

```
dades <- dades[, -c(4,6,7,8,15,18,19,22,23,25,26,27)]
```

```
# revisem i corregim el tipus de dades
```

```
#str(dades)
```

```
dades$dtHora_BD <- as.POSIXct(dades$dtHora_BD)
```

```
dades$dtHora_LlegadaReal <- as.POSIXct(dades$dtHora_LlegadaReal)
```

```
dades$iIdExpedicion <- as.ordered(dades$iIdExpedicion)
```

```
dades$iIdParada <- as.factor(as.numeric(dades$iIdParada))
```

```
dades$iIdLinea <- as.factor(dades$iIdLinea)
```

```
dades$iIdTrayecto <- as.factor(dades$iIdTrayecto)
```

```
dades$iTipoAccion <- as.factor(dades$iTipoAccion)
```

```
dades$iIdAutobus <- as.factor(dades$iIdAutobus)
```

```
dades$iIdConductor <- as.factor(dades$iIdConductor)
```

```
dades$iEstadoHRLlegada <- as.factor(dades$iEstadoHRLlegada)
```

```
dades$iEstadoHRSalida <- as.factor(dades$iEstadoHRSalida)
```

```
dades$iOrdenExpedicion <- as.ordered(dades$iOrdenExpedicion)
```

```
# resum
```

```
summary(dades)
```

```
## iIdHoraPasoParada iTiempoEnParada dtHora_BD
## Min. :8662289 Min. : 0.00 Min. :2020-02-01 05:20:00
## 1st Qu.:8737428 1st Qu.: 8.00 1st Qu.:2020-02-11 12:53:00
## Median :8812566 Median : 16.00 Median :2020-02-20 17:57:00
## Mean :8861316 Mean : 36.19 Mean :2020-03-03 03:05:59
## 3rd Qu.:9067661 3rd Qu.: 33.00 3rd Qu.:2020-04-03 11:03:00
## Max. :9142800 Max. :9981.00 Max. :2020-04-30 22:46:00
##
## dtHora_LlegadaReal iIdExpedicion iIdParada
## Min. :2020-02-01 05:17:58 117785 : 2328 93 : 5955
## 1st Qu.:2020-02-11 12:55:04 117786 : 2325 198 : 5754
## Median :2020-02-20 18:02:17 117802 : 2319 130 : 4416
## Mean :2020-03-03 03:07:59 117520 : 2318 12 : 4120
## 3rd Qu.:2020-04-03 11:04:06 117800 : 2301 51 : 4116
## Max. :2020-04-30 22:43:06 117505 : 2293 88 : 4113
##
## (Other):286672 (Other):272082
## iIdLinea iIdTrayecto iTipoAccion iOrden iIdAutobus
## 20 :123035 1 :282184 0:269861 Min. : 0.00 28 : 28
638
## 10 : 63088 201: 4807 1: 16317 1st Qu.:10.00 23 : 27
843
## 21 : 29738 301: 4970 2: 32 Median :23.00 27 : 26
934
## 60 : 23200 401: 7471 3: 14346 Mean :26.27 26 : 26
792
## 11 : 20418 501: 933 3rd Qu.:38.00 22 : 26
776
## 50 : 15702 701: 191 Max. :74.00 19 : 26
249
## (Other): 25375 (Other):137
324
## iIdConductor iEstadoHRLlegada iEstadoHRSalida iOrdenExpedicion
## 88 : 13818 0: 323 0: 457 0 :144301
## 27 : 13369 1: 13857 1: 12480 1 : 34397
## 94 : 12920 2:265327 2:280544 2 : 22916
## 65 : 12306 3: 14722 3: 811 3 : 18530
## 49 : 11912 4: 6327 4: 5459 4 : 14649
## 91 : 11628 5: 805 5 : 12405
## (Other):224603 (Other): 53358
```

```
# afegim les dades de nom de parada, que es troben en una altra taula
# parades.JSON recull les parades en servei actualment
parades.JSON <- as_tibble(fromJSON("parades.json", simplifyMatrix = TRUE)
)
parades <- as_tibble(parades.JSON$stops)
parades$stop_id <- as.factor(as.numeric(parades$stop_id))

dades <- left_join(dades,parades[,c(1,2)], by=c("iIdParada"="stop_id"))
dades$stop_name <- as.factor(dades$stop_name)
```

```
# paradas.Sae recull Les parades registrades al mapa del SAE
paradas.Sae <- read.csv("ParadasSae.csv", encoding = "UTF-8")
paradas.Sae$iIdParada <- as.factor(paradas.Sae$iIdParada)
```

2.2. Precipitació diària

Les dades s'han obtingut del portal de dades obertes de l'Agència Estatal de Meteorologia (<https://opendata.aemet.es/centrodedescargas/inicio>).

S'ha descarregat un fitxer json amb les dades meteorològiques diàries de l'estació meteorològica anomenada Reus Aeroport, compreses entre el gener i l'abril de 2020, en total 121 registres.

D'aquest arxiu, ens interressarà el valor de la precipitació diària, que segons s'explica a l'arxiu de metadades, correspon a la variable 'prec': * prec: precipitació diària, de 07h a 07h, en mm. (lp = inferior a 0,1 mm) * fecha: data, en format AAAA-MM-DD

```
aemet <- as_tibble(fromJSON("aemet.json", simplifyMatrix = TRUE))

# revisem el tipus de dades (només ens interessa data i precipitació)
#str(aemet)
aemet$fecha <- as.Date(aemet$fecha)

# hem de canviar les comes per punts en la precipitació
aemet$prec <- sapply(aemet$prec, gsub, pattern = ",", replace=".")
aemet$prec <- as.numeric(aemet$prec)

# resum
summary(aemet$prec)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##  0.000   0.000   0.000   2.197   0.000  45.700         3

summary(aemet$fecha)

##      Min.      1st Qu.      Median      Mean      3rd Qu.
## "2020-01-01" "2020-01-31" "2020-03-01" "2020-03-01" "2020-03-31" "2020
-04-30"
```

Observem que dels 121 registres diaris, en 32 dies ha plogut.

2.3. Dades de trànsit

Les dades s'han obtingut dels registres de les 11 càmeres de visió artificial situades a diferents punts de la ciutat per regular l'accés al casc antic.

La càmera que pot resultar més interessant per estimar la fluïdesa del trànsit és l'anomenada S7, situada al nord de la Riera Miró, monitoritzant el trànsit en sentit Av. St. Jordi.

Segons la informació facilitada, cada sensor SX registra les següents dades, en intervals d'un minut:

Ocupació del tram indicat a la càmera, aquest es parteix en 6 zones e indica el % del temps ocupat durant lectura de dades:

SX_0_20 : Percentatge ocupació de la zona 0 a 20 % del tram.

SX_20_40 : Percentatge ocupació de la zona 20 a 40 % del tram

SX_40_60 : Percentatge ocupació de la zona 40 a 60 % del tram.

SX_60_80 : Percentatge ocupació de la zona 60 a 80 % del tram.

SX_80_100 : Percentatge ocupació de la zona 80 a 100 % del tram.

SX_100_-1 : Percentatge ocupació de la zona de més del 100% del tram.

Dades del tram indicat, a la càmera durant la lectura de dades :

SX_distancia_vehicle : Distància mitjana entre vehicles en el tram (m).

SX_velocitat : Velocitat mitjana dels vehicles en el tram (km/h).

SX_volum : Número de vehicles detectats en el tram.

Dels registres, ens interessen els següents camps: * sensor: identificació del sensor i paràmetre de lectura * value: valor de lectura, segons la descripció anterior * event_timestamp: moment de la lectura

Cal tenir present que les lectures dels paràmetres velocitat i distància entre vehicles no es fan servir actualment i no han estat calibrades.

Donada la gran quantitat de dades generades per 11 sensors amb lectures de 9 paràmetres cada minut, s'ha optat per exportar les dades que continguin una setmana de febrer, que considerarem com a model de comportament del trànsit a la ciutat. Per motius tècnics, s'han facilitat les dades en 4 arxius que cal agrupar.

```
sentilo1 <- as_tibble(read.csv(file="sensorsRiera1.csv", header = TRUE))
sentilo2 <- as_tibble(read.csv(file="sensorsRiera2.csv", header = TRUE))
sentilo3 <- as_tibble(read.csv(file="sensorsRiera3.csv", header = TRUE))
sentilo4 <- as_tibble(read.csv(file="sensorsRiera4.csv", header = TRUE))

# agrupament dades en una sola taula
sentilo <- rbind(sentilo1,sentilo2,sentilo3,sentilo4)
#str(sentilo)

# eliminem les columnes que no aporten informació i reviseu format
sentilo <- sentilo[,-c(4:7)]
sentilo$event_timestamp <- as.POSIXct(sentilo$event_timestamp)

veloc_S7 <- sentilo %>%
  filter(sensor=="S7_velocitat")

dist_veh_S7 <- sentilo %>%
```

```

filter(sensor=="S7_distancia_vehicle")

S7_ocup_60_80 <- sentilo %>%
  filter(sensor=="S7_60_80")

S7_ocup_80_100 <- sentilo %>%
  filter(sensor=="S7_80_100")

S7_100 <- sentilo %>%
  filter(sensor=="S7_100_-1")

```

Són un total de 800000 registres.

3. INTEGRACIÓ I SELECCIÓ DE LES DADES D'INTERÈS

Per relacionar puntualitat i pluja, incorporarem al dataset 'dades' la variable 'prec' del dataset 'aemet'. Com que la dada de precipitació és diària, caldrà repetir-la en tots els valors de pas per parada d'aquell dia.

Per fer la integració, necessitem abans haver creat a 'dades' un camp amb la data sola, sense hora i minut, per buscar la coincidència amb el camp 'fecha' de la taula 'aemet'.

Com que previsiblement voldrem analitzar la puntualitat segons l'hora del dia i el dia de la setmana, creem també aquestes variables.

```

dades$dataR <- as_date(dades$dtHora_LlegadaReal)
dades$horaR <- hour(dades$dtHora_LlegadaReal)
dades$minR <- minute(dades$dtHora_LlegadaReal)
dades$dataT <- as_date(dades$dtHora_BD)
dades$horaT <- hour(dades$dtHora_BD)
dades$minT <- minute(dades$dtHora_BD)
dades$diastna <- as.factor(lubridate::wday(dades$dtHora_LlegadaReal, label=TRUE, week_start=1))

# afegim el nom de parada a partir del codi
dades <- left_join(dades,paradas.Sae)

## Joining, by = "iIdParada"

# afegim la columna precipitació
dades <- left_join(dades,aemet[,c(1,7)], by=c("dataT"="fecha"))

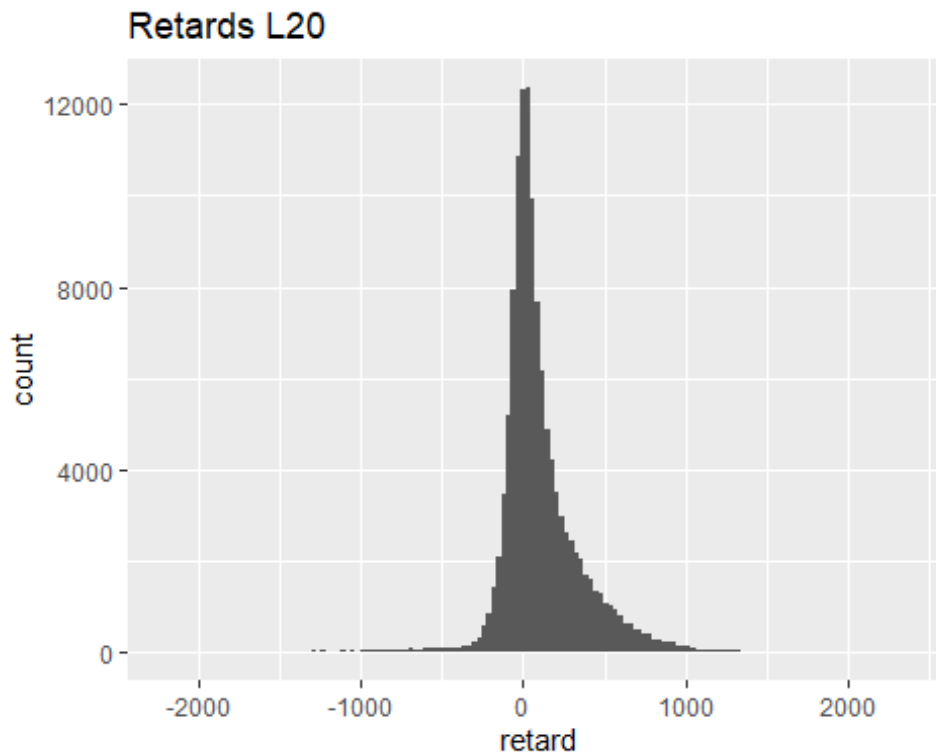
```

Per determinar la puntualitat de pas a cada parada generem una nova variable, anomenada 'retard', que indiqui quants segons passen entre l'hora programada i l'hora real de pas. Aquesta variable pot ser negativa, donat que en alguns casos els vehicles arriben abans d'hora a la parada. En aquests casos, cal que esperin fins l'hora prevista de sortida (valor reflectit en la variable iTiempoEnParada).

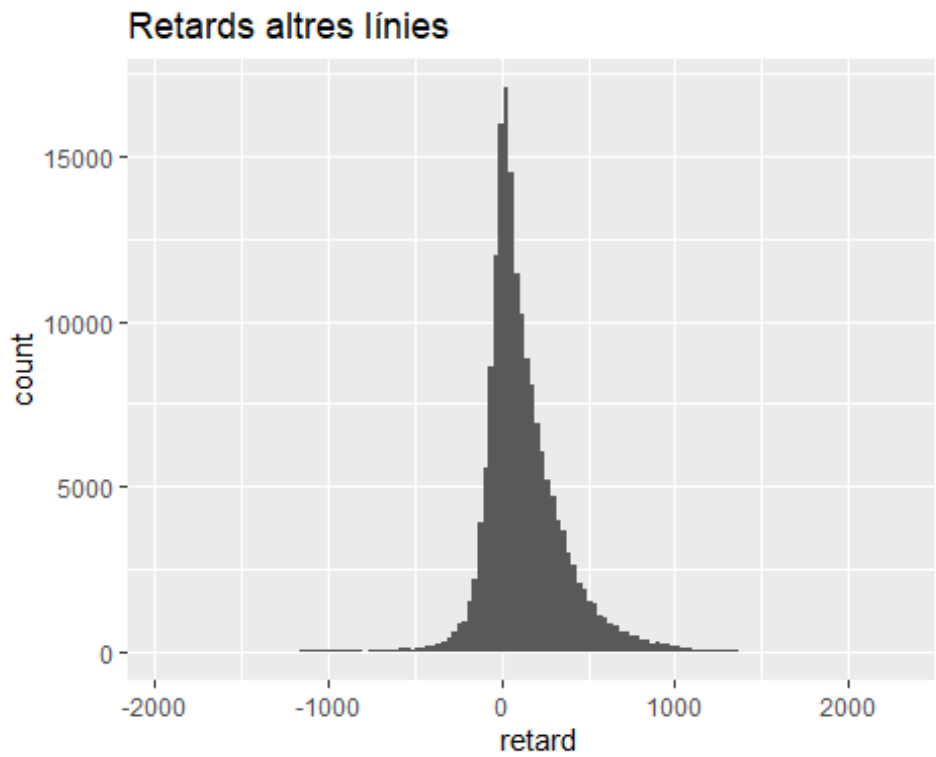
```
dades$retard <- as.numeric(dades$dtHora_LlegadaReal-dades$dtHora_BD)
summary(as.numeric(dades$retard))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -9911     -13      67     120     217    8880
```

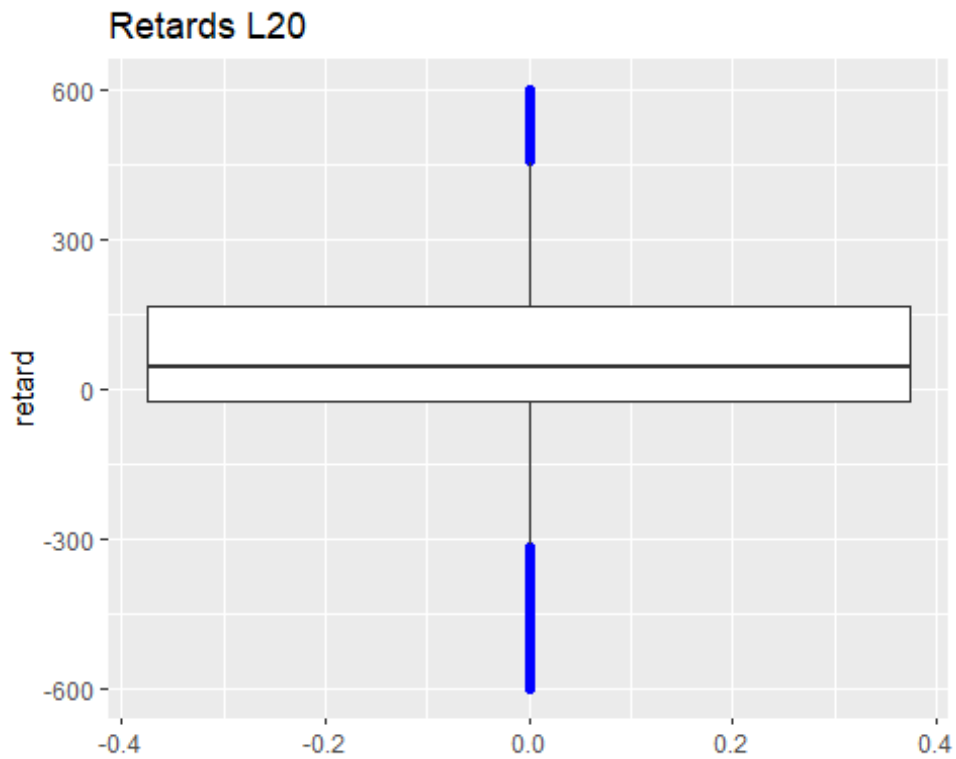
```
dades %>%
  filter(retard<2400 & retard>-2400) %>%
  filter(iIdLinea==20) %>%
  ggplot(aes(x=retard)) +
  geom_histogram(binwidth = 30)+
  labs(title = "Retards L20")
```



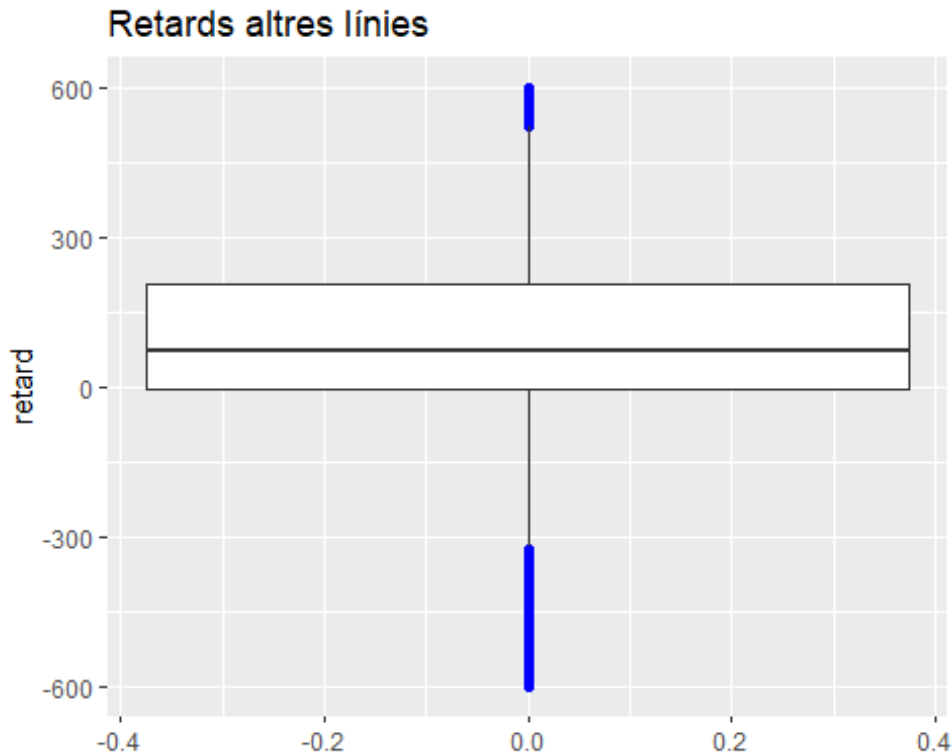
```
dades %>%
  filter(retard<2400 & retard>-2400) %>%
  filter(iIdLinea!=20) %>%
  ggplot(aes(x=retard)) +
  geom_histogram(binwidth = 30)+
  labs(title = "Retards altres línies")
```

```
dades %>%  
  filter(retard<600 & retard>-600) %>%  
  filter(iIdLinea==20) %>%  
  ggplot(aes(y=retard)) +  
  geom_boxplot(outlier.color = "blue")+  
  labs(title = "Retards L20")
```



```
dades %>%  
  filter(retard<600 & retard>-600) %>%  
  filter(iIdLinea!=20) %>%  
  ggplot(aes(y=retard)) +  
  geom_boxplot(outlier.color = "blue")+  
  labs(title = "Retards altres línies")
```



Aquesta primera exploració mostra una distribució de retards similar a una normal, esbiaixada cap als valors positius, i amb valors extrems molt allunyats dels valors centrals.

La comparació de la línia 20 respecte les altres línies no sembla indicar que la L20 tingui retards superiors a les altres línies, tot i que més endavant ho analitzarem en profunditat.

Les dades de fluïdesa de trànsit no s'integren en aquest moment donat que són una referència de volum de trànsit setmanal. Si pertoca, s'incorporaran més endavant, comparant els moments de congestió (dia de la setmana i hores punta) amb els de retards.

4. NETEJA DE LES DADES

4.1. Zeros i elements buits

En aquest punt analitzem la presència de valors que puguin indicar una manca de dades o lectures errònies.

```
# veiem quants NA tenim a Les dades
sapply(dades, function(x) sum(is.na(x)))

##   iIdHoraPasoParada   iTiempoEnParada   dtHora_BD dtHora_Llegad
aReal
##                0                0                0
0
```

```
##      iIdExpedicion      iIdParada      iIdLinea      iIdTra
yector
##              0              0              0
0
##      iTipoAccion      iOrden      iIdAutobus      iIdCond
uctor
##              0              0              0
0
##      iEstadoHRLlegada      iEstadoHRSalida      iOrdenExpedicion      stop
_name
##              0              0              0
0
##              dataR      horaR      minR
dataT
##              0              0              0
0
##              horaT      minT      diastna      sDenomin
acion
##              0              0              0
0
##      iCoordenadaUTMX      iCoordenadaUTMY      prec      r
etard
##              0              0              0
0

sapply(sentilo, function(x) sum(is.na(x)))

##      id      sensor      value      event_timestamp
##      0      0      0      0
```

No tenim valors perduts a cap de les variables.

Tenint en compte que recollim dades llegides d'instruments, és fàcil que hi hagi valors numèrics que indiquin pèrdua de dades, com valors registrats que no formin part del domini de la variable. Mirem d'identificar-los amb la funció summary, que ens indicarà els valors extrems de cada variable.

```
summary(dades)

##      iIdHoraPasoParada      iTiempoEnParada      dtHora_BD
##      Min.      :8662289      Min.      : 0.00      Min.      :2020-02-01 05:20:00
##      1st Qu.:8737428      1st Qu.: 8.00      1st Qu.:2020-02-11 12:53:00
##      Median :8812566      Median : 16.00      Median :2020-02-20 17:57:00
##      Mean   :8861316      Mean   : 36.19      Mean   :2020-03-03 03:05:59
##      3rd Qu.:9067661      3rd Qu.: 33.00      3rd Qu.:2020-04-03 11:03:00
##      Max.   :9142800      Max.   :9981.00      Max.   :2020-04-30 22:46:00
##
##      dtHora_LlegadaReal      iIdExpedicion      iIdParada
##      Min.      :2020-02-01 05:17:58      117785 : 2328      Length:300556
##      1st Qu.:2020-02-11 12:55:04      117786 : 2325      Class :character
##      Median :2020-02-20 18:02:17      117802 : 2319      Mode  :character
##      Mean   :2020-03-03 03:07:59      117520 : 2318
```

```

## 3rd Qu.:2020-04-03 11:04:06 117800 : 2301
## Max. :2020-04-30 22:43:06 117505 : 2293
## (Other):286672
## iIdLinea iIdTrayecto iTipoAccion iOrden iIdAutobu
s
## 20 :123035 1 :282184 0:269861 Min. : 0.00 28 : 28
638
## 10 : 63088 201: 4807 1: 16317 1st Qu.:10.00 23 : 27
843
## 21 : 29738 301: 4970 2: 32 Median :23.00 27 : 26
934
## 60 : 23200 401: 7471 3: 14346 Mean :26.27 26 : 26
792
## 11 : 20418 501: 933 3rd Qu.:38.00 22 : 26
776
## 50 : 15702 701: 191 Max. :74.00 19 : 26
249
## (Other): 25375 (Other):137
324
## iIdConductor iEstadoHRLlegada iEstadoHRSalida iOrdenExpedicion
## 88 : 13818 0: 323 0: 457 0 :144301
## 27 : 13369 1: 13857 1: 12480 1 : 34397
## 94 : 12920 2:265327 2:280544 2 : 22916
## 65 : 12306 3: 14722 3: 811 3 : 18530
## 49 : 11912 4: 6327 4: 5459 4 : 14649
## 91 : 11628 5: 805 5 : 12405
## (Other):224603 (Other): 53358
## stop_name dataR horaR
## Oques 3 : 5955 Min. :2020-02-01 Min. : 4.00
## Hospital 1 : 5754 1st Qu.:2020-02-11 1st Qu.:10.00
## RENFE 1 : 4416 Median :2020-02-20 Median :13.00
## Avinguda de Sant Jordi 1: 4120 Mean :2020-03-02 Mean :13.25
## Pompeu Fabra 1 : 4116 3rd Qu.:2020-04-03 3rd Qu.:17.00
## Niloga 1 : 4113 Max. :2020-04-30 Max. :22.00
## (Other) :272082
## minR dataT horaT minT
## Min. : 0.00 Min. :2020-02-01 Min. : 5.00 Min. : 0.00
## 1st Qu.:15.00 1st Qu.:2020-02-11 1st Qu.:10.00 1st Qu.:15.00
## Median :30.00 Median :2020-02-20 Median :13.00 Median :29.00
## Mean :29.32 Mean :2020-03-02 Mean :13.22 Mean :29.37
## 3rd Qu.:44.00 3rd Qu.:2020-04-03 3rd Qu.:17.00 3rd Qu.:44.00
## Max. :59.00 Max. :2020-04-30 Max. :22.00 Max. :59.00
##
## diastna sDenominacion iCoordenadaUTMX
## lu:46479 Oques 3 : 5955 Min. :841772
## ma:46804 Hospital 1 : 5754 1st Qu.:844228
## mi:50068 RENFE 1 : 4416 Median :844757
## ju:50019 Avinguda de Sant Jordi 1: 4120 Mean :844779
## vi:47182 Pompeu Fabra 1 : 4116 3rd Qu.:845346
## sá:36639 Niloga 1 : 4113 Max. :848610

```

```
## do:23365 (Other) :272082
## iCoordenadaUTMY prec retard
## Min. :4561415 Min. : 0.0000 Min. : -9911
## 1st Qu.:4563337 1st Qu.: 0.0000 1st Qu.: -13
## Median :4563867 Median : 0.0000 Median : 67
## Mean :4563947 Mean : 0.9462 Mean : 120
## 3rd Qu.:4564563 3rd Qu.: 0.0000 3rd Qu.: 217
## Max. :4566510 Max. :25.7000 Max. : 8880
##
```

```
summary(sentilo)
```

```
## id sensor value
## Min. :173000065 S1_0_20 : 12729 Min. : -1.00
## 1st Qu.:173508084 S1_100_-1: 12729 1st Qu.: 0.00
## Median :174003018 S1_20_40 : 12729 Median : 0.00
## Mean :174018963 S1_40_60 : 12729 Mean : 13.44
## 3rd Qu.:174525279 S1_60_80 : 12729 3rd Qu.: 3.00
## Max. :175064287 S1_80_100: 12729 Max. :1691.00
## (Other) :723626
## event_timestamp
## Min. :2020-01-31 18:08:00
## 1st Qu.:2020-02-02 23:07:00
## Median :2020-02-05 04:11:00
## Mean :2020-02-05 04:17:14
## 3rd Qu.:2020-02-07 09:28:00
## Max. :2020-02-09 14:33:01
##
```

*# veiem quants casos tenen value -1, que indica error de lectura, i els e
liminem de l'anàlisi*

```
sentilo %>%
  filter(value==-1) %>%
  select(sensor, event_timestamp) %>%
  summary()
```

```
## sensor event_timestamp
## S1_distancia_vehicle:9952 Min. :2020-01-31 18:08:00
## S4_distancia_vehicle:8436 1st Qu.:2020-02-02 23:43:00
## S7_distancia_vehicle:6267 Median :2020-02-05 03:05:00
## S2_distancia_vehicle:5857 Mean :2020-02-05 04:11:49
## S6_distancia_vehicle:4089 3rd Qu.:2020-02-07 10:00:00
## S3_distancia_vehicle:3989 Max. :2020-02-09 14:33:01
## (Other) :3743
```

```
sentilo <- sentilo[sentilo$value!=-1,]
```

```
dist_veh_S7 <- dist_veh_S7 %>%
  filter(value!=-1)
```

Amb el resum de valors de la taula 'dades' observem que els temps i volums de precipitacions registrats són positius, que les dates estan dins els intervals analitzats, i les dades que defineixen el servei també estan dins el domini.

En canvi, al resum de la taula 'sentilo' observem per un costat valors dels sensors de -1, que poden indicar mesura errònia, i per l'altre un valor màxim molt elevat que caldrà analitzar a l'apartat de valors extrems. Els valors -1 corresponen tots a mesures de distància entre vehicles. Donat que aquests registres no aporten cap informació, procedim a eliminar-los. Les dades de sensors passen de 800.000 registres a 757.667.

4.2. Valors extrems de la variable retard

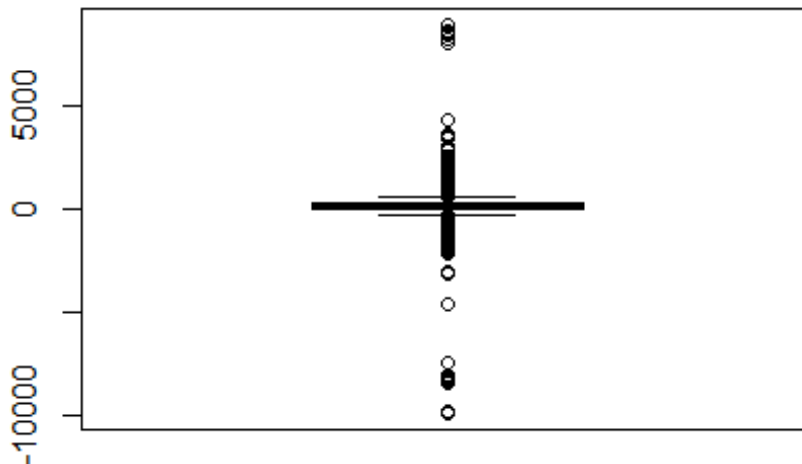
A l'exploració preliminar de la variable 'retard' ja apareixia una quantitat important de valors extrems, aprofundirem en el seu estudi per decidir si són o no valors legítims per l'anàlisi que volem realitzar.

De la funció summary observem una gran diferència entre la mitjana (119.954521) i la mediana (67), que assenyala un biaix de la distribució cap als valors positius.

```
summary(dades$retard)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	-9911	-13	67	120	217	8880

```
retard.bp <- boxplot(dades$retard)
```



```
# valors del diagrama de caixa:  
retard.bp$stats
```

```
##      [,1]
## [1,] -358
## [2,]  -13
## [3,]   67
## [4,]  217
## [5,]  562

# taula resum dels valors extrems en retards

parades.out <- dades %>%
  filter(retard %in% retard.bp$out) %>%
  group_by(iIdParada, sDenominacion, iIdLinea) %>%
  summarise(
    casos=n(),
    retard.avg=round(mean(retard),1),
    retard.median=median(retard)
  )
```

Conmbinacions parada línia amb més outliers

```
knitr::kable(head(arrange(parades.out, desc(parades.out$casos)), n=20))
```

ildParada	sDenominacion	ildLinea	casos	retard.avg	retard.median
198	Hospital 1	20	1425	-722.8	-687.0
13	Barri Montserrat	10	890	-355.3	-540.0
198	Hospital 1	21	461	-884.5	-918.0
1	Aeroport	50	326	-150.6	-451.5
134	Riera d'aAragó 2	20	311	745.7	719.0
23	Camí de Tarragona 2	20	308	747.1	728.0
197	Hospital Consultas Externas 2	20	278	695.0	717.0
202	Hospital Regulació	20	275	741.2	714.0
25	Camí de Valls 2	20	266	733.6	714.0
50	Pompeu Fabra 2	20	256	775.0	717.0
151	Universitat 2	20	246	658.7	709.5
75	Mas Abelló 2	20	246	694.1	720.0
199	Hospital 2	20	242	627.7	707.0
40	Dom Bosco 2	20	240	720.2	710.0
117	Sant Josep 2	20	239	669.1	714.0
11	Avinguda de Sant Jordi 2	20	237	771.4	724.0
13	Barri Montserrat	11	224	193.8	580.5
147	Tanatori 2	20	220	695.3	719.5
72	Llibertat	20	219	760.3	707.0
93	Oques 3	30	188	-401.2	-420.0

Conmbinacions parada línia amb arribades d'autobus abans d'hora

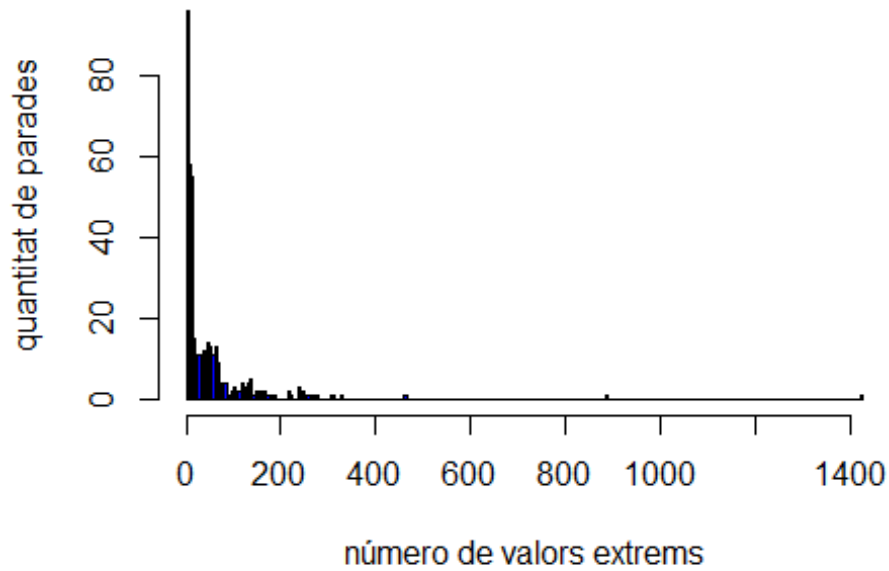
```
knitr::kable(head(arrange(parades.out, parades.out$retard.avg), n=20))
```


ildParada	sDenominacion	ildLinea	casos	retard.avg	retard.median
198	Hospital 1	21	461	-884.5	-918.0
198	Hospital 1	20	1425	-722.8	-687.0
93	Oques 3	33	73	-718.4	-467.0
156	Ventura Gassol 1	32	1	-649.0	-649.0
31	Cementiri	32	46	-641.3	-648.0
69	Jutjats 1	32	1	-614.0	-614.0
132	Reus Transport	31	58	-553.7	-554.5
129	RENFE 2	33	1	-497.0	-497.0
214	Aeroclub de Reus 1	50	32	-446.3	-659.0
88	Niloga 1	32	1	-444.0	-444.0
116	Plaça Almostrer 2	33	1	-429.0	-429.0
67	Joan Rebull 1	32	1	-426.0	-426.0
130	RENFE 1	32	1	-403.0	-403.0
93	Oques 3	30	188	-401.2	-420.0
52	Flix 1	10	69	-385.5	-556.0
51	Pompeu Fabra 1	32	2	-376.0	-376.0
198	Hospital 1	60	117	-375.4	-417.0
93	Oques 3	32	7	-368.0	-684.0
12	Avinguda de Sant Jordi 1	32	1	-363.0	-363.0
13	Barri Montserrat	10	890	-355.3	-540.0

Quantitat d'outliers per parada

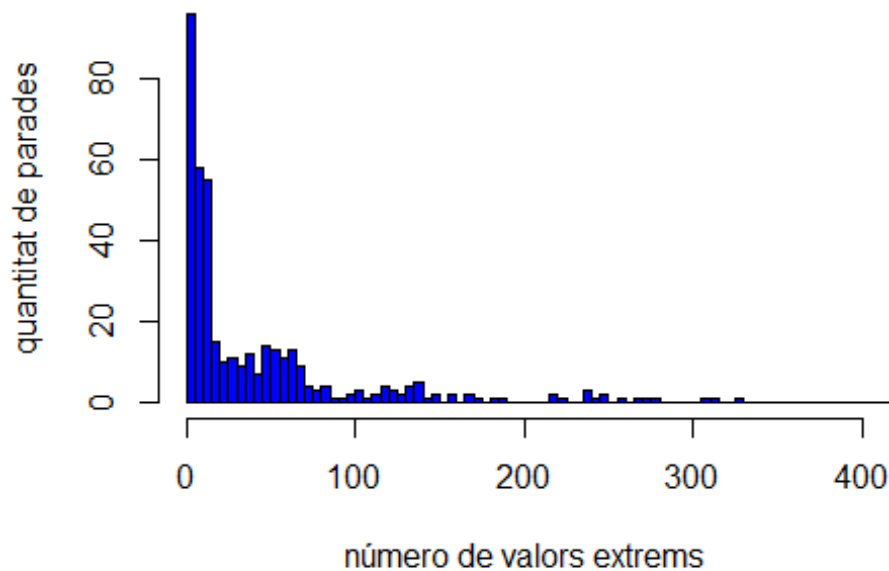
```
hist(parades.out$casos, breaks = 500, col="blue", main="Quantitat d'outliers per parada", xlab = "número de valors extrems", ylab = "quantitat de parades")
```

Quantitat d'outliers per parada

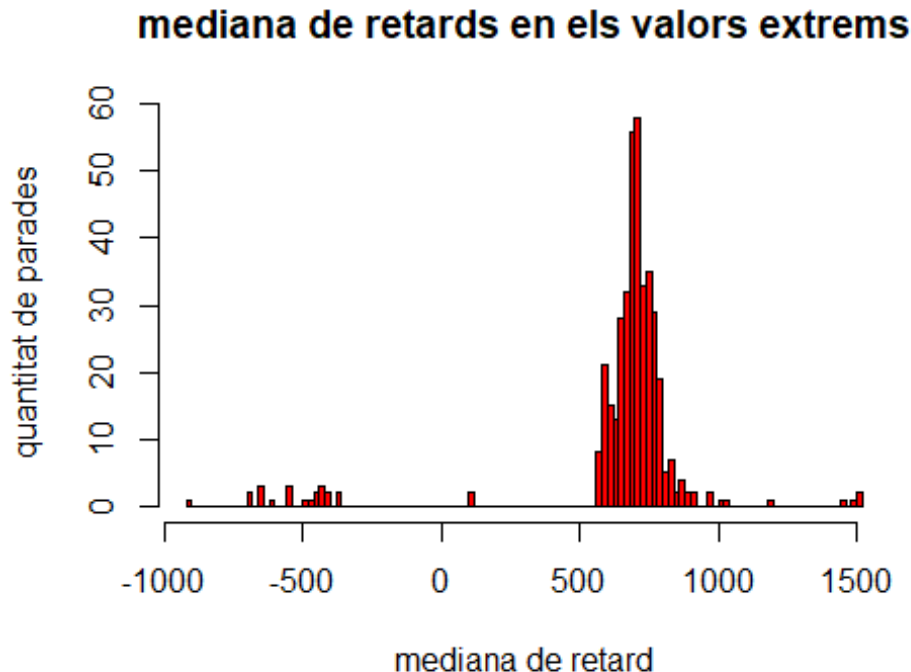


```
hist(parades.out$casos, breaks = 500, xlim = c(0,400), col="blue", main="Quantitat d'outliers per parada", xlab = "número de valors extrems", ylab = "quantitat de parades")
```

Quantitat d'outliers per parada



```
hist(parades.out$retard.median, breaks = 100, col="red", main="mediana de retards en els valors extrems", xlab = "mediana de retard", ylab = "quantitat de parades")
```



```
# eliminem els registres amb valors extrems de retard negatius
dades.clean <- dades %>% filter(!(retard %in% retard.bp$out))
```

Dels llistats i histogrames anteriors observem que:

- Les parades amb més quantitat de valors extrems són Mas Abelló 1 i Barri Montserrat. A més, els valors centrals dels retards en aquestes parades són negatius
- Altres parades amb retards extrems negatius són Misericòrdia, Cementiri, Tanatori, etc.

Del coneixement del servei sabem que aquestes parades són “de regulació”: parades on s’envia l’autobús saltant-se part de la ruta per a posar-se en hora, normalment quan porta més de deu minuts de retard respecte l’horari teòric, i si va molt a prop d’algun altre autobus de la línia.

En aquests casos és normal que el sistema registri un horari d’arribada molt diferent al previst, però és una situació forçada per reconduir horaris, de manera que cal eliminar aquests registres de l’anàlisi. A l’histograma veiem que es dona en poques ocasions, correspon als valors de mediana entre -400 i -600, aproximadament. En aquest pas hem eliminat aproximadament el 7% dels registres.

En canvi, si analitzem les parades amb valors extrems de retards positius, veiem que són parades diverses i que es dona una gran quantitat de vegades, especialment retards entre 600 i 900 segons, que corresponen a 10-15 minuts i pel coneixement del servei podem dir que són situacions que es donen en ocasions.

En aquest cas, optem per mantenir tots els registres positius, encara que siguin valors extrems, donat que l'objecte de l'anàlisi és justament estudiar-los.

4.3. Valors extrems en sensors

Per analitzar els valors dels sensors que ens interessen cal prèviament seleccionar el sensor i paràmetre que volem estudiar i filtrar-lo en una nova taula.

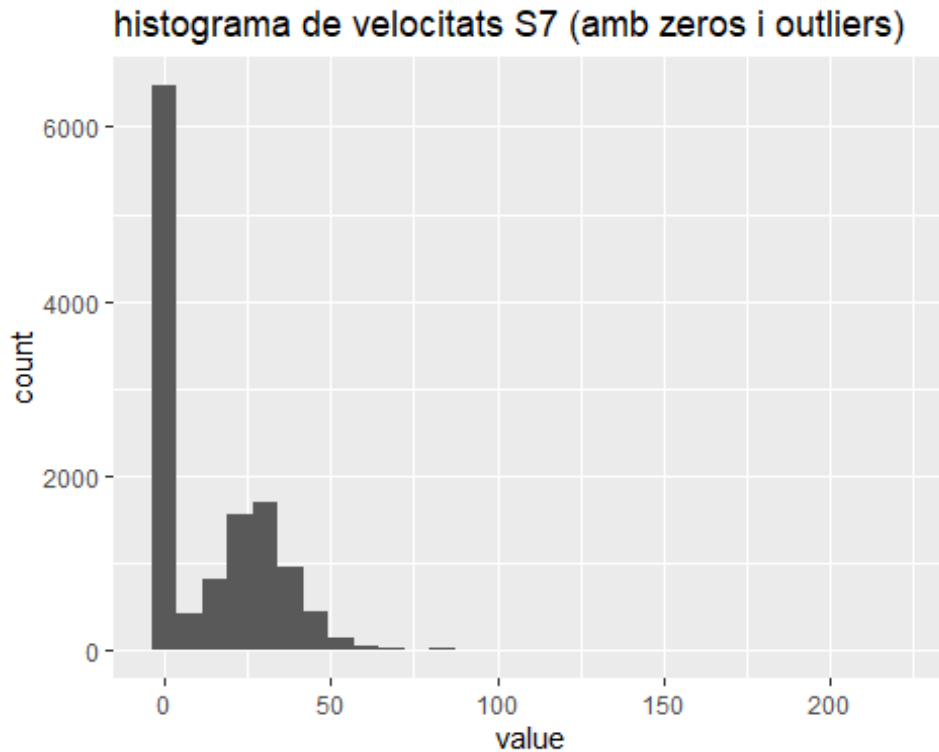
Per analitzar les afectacions al servei de bus urbà ens podrien ser útils la velocitat mitjana, la distància entre vehicles, el percentatge de temps d'ocupació del 60 al 80% del tram, del 80 al 100 % del tram i el del 100% del tram, tots ells del sensor S7, situat entre les parades Dom Bosco 1 i Pompeu Fabra 1. Ja disposem de les taules específiques, creades a l'apartat de càrrega de dades. No es poden unificar en una sola taula perquè les mesures estan preses en moments diferents, no són diferents paràmetres d'un mateix esdeveniment, sinó esdeveniments diferents.

```
# velocitat mitjana
summary(veloc_S7$value)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   0.00   0.00   13.96   27.00   220.00

veloc_S7 %>%
  filter(value>=0) %>%
  ggplot(aes(x=value)) +
  geom_histogram()+
  labs(title = "histograma de velocitats S7 (amb zeros i outliers)")

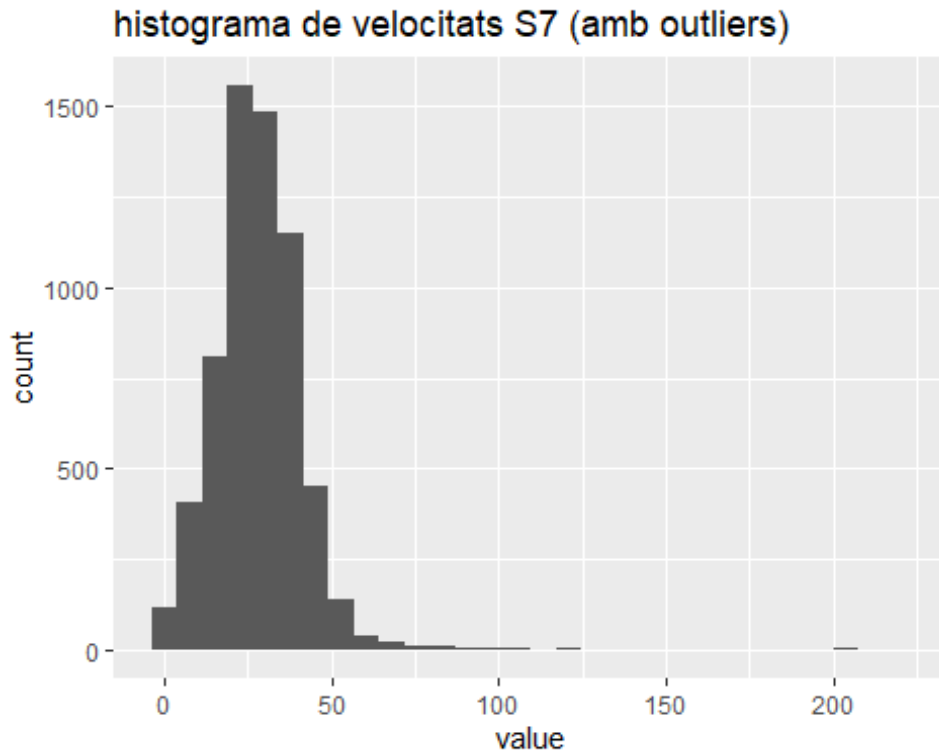
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# Les lectures =0 són eliminades dels registres, per interpretar que són lectures sense vehicles, no és creïble que els vehicles estiguin sempre a turats
veloc_S7 <- veloc_S7 %>%
  filter(value!=0)

veloc_S7 %>%
  ggplot(aes(x=value)) +
  geom_histogram()+
  labs(title = "histograma de velocitats S7 (amb outliers)")

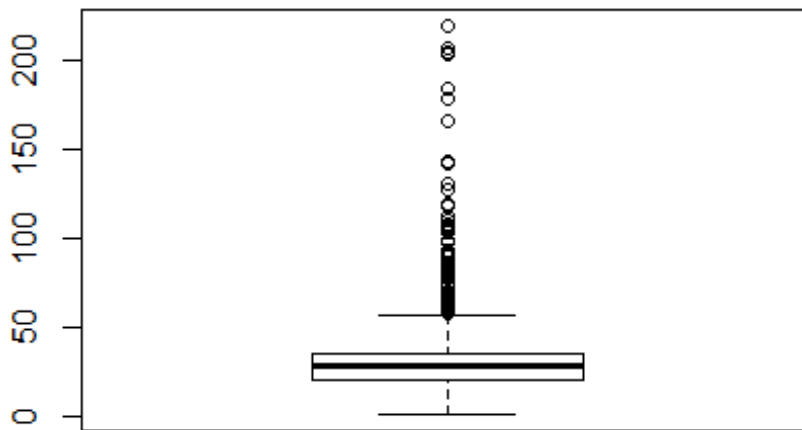
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



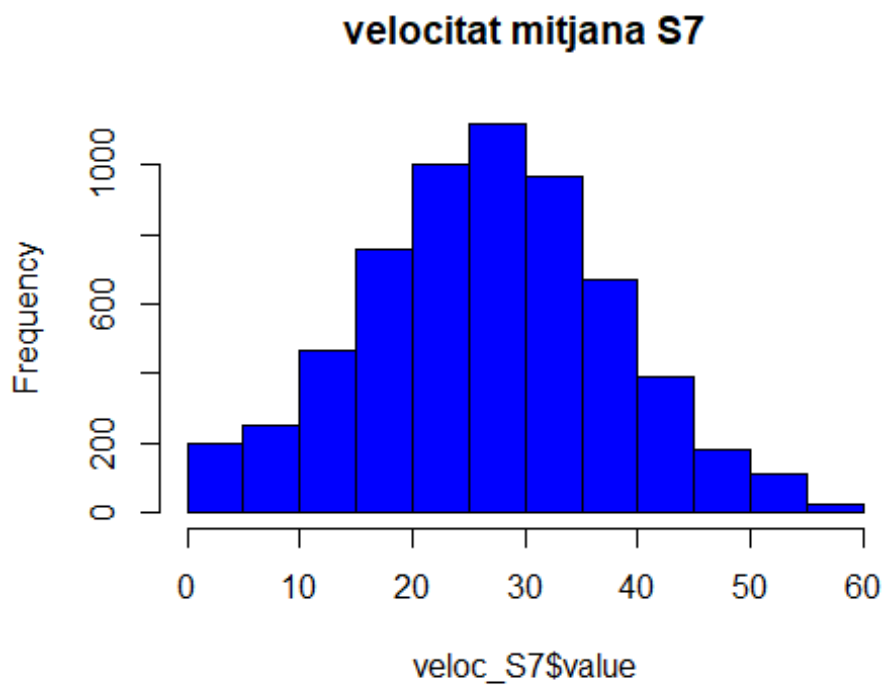
```
summary(boxplot(veloc_S7$value)$out)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   58.00  64.00   72.00   85.71  91.75  220.00

# eliminem els valors extrems segons el boxplot, tots a la part alta de v
# elocitats
veloc_S7 <- veloc_S7 %>%
  filter(!(value %in% boxplot(veloc_S7$value)$out))
```



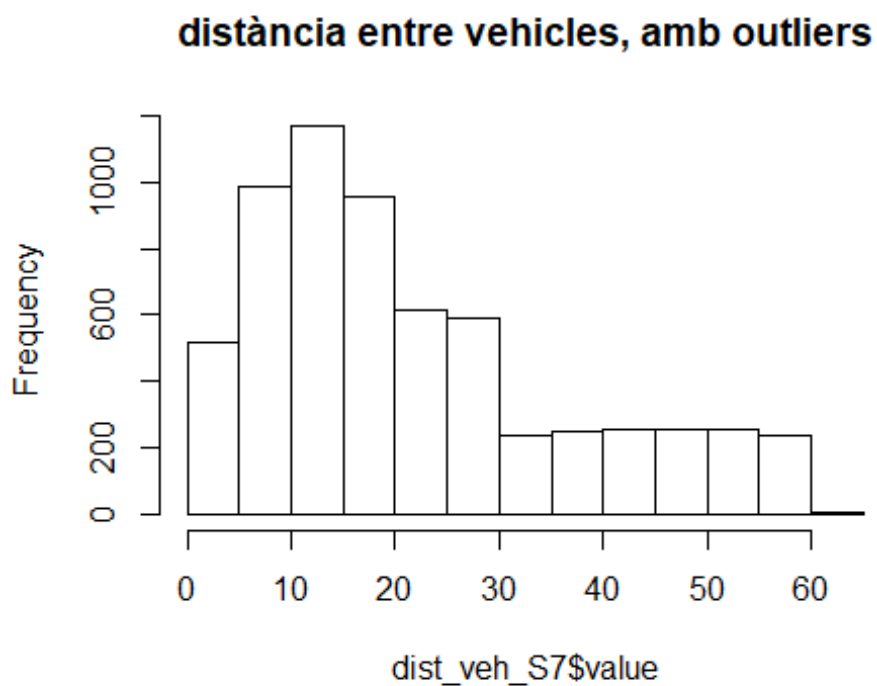
```
hist(veloc_S7$value, col="blue", main="velocitat mitjana S7")
```



```
# distància entre vehicles
summary(dist_veh_S7$value)
```

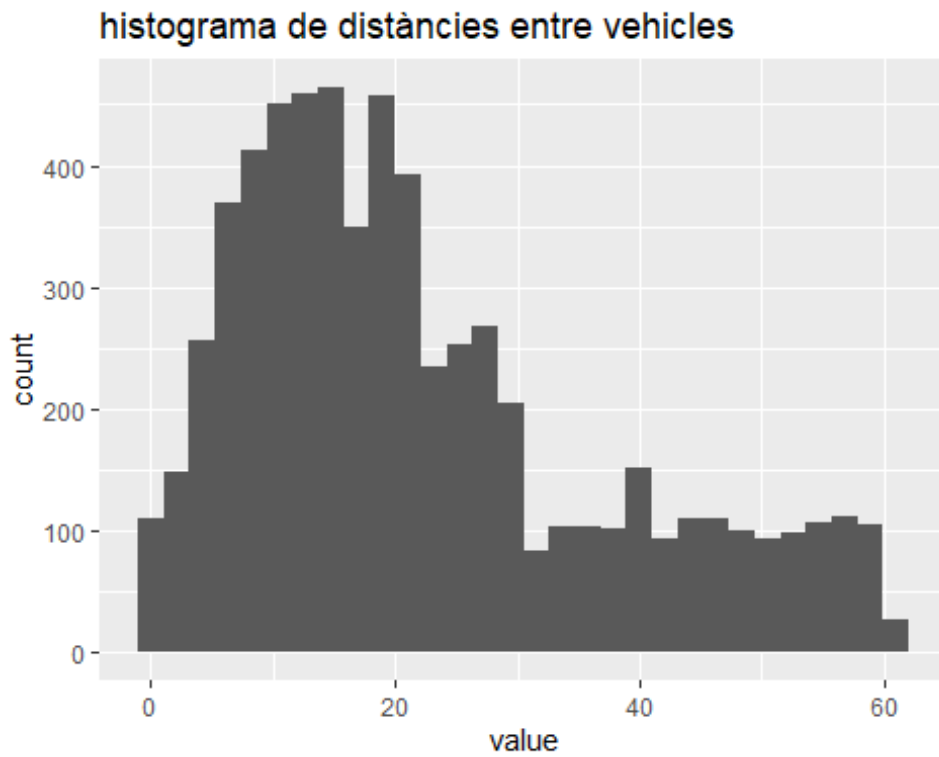
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00  11.00   18.00   22.24  29.00   61.00
```

```
hist(dist_veh_S7$value, main="distància entre vehicles, amb outliers")
```

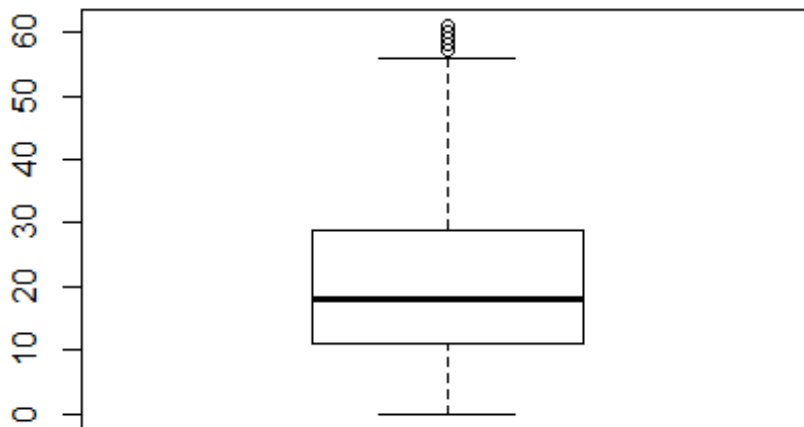


```
dist_veh_S7 %>%
  ggplot(aes(x=value)) +
  geom_histogram() +
  labs(title = "histograma de distàncies entre vehicles")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
# valors extrems, no els eliminem perquè estan en el domini de la variable  
summary(boxplot(dist_veh_S7$value)$out)
```



```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    57.00  57.00   58.00   58.29  59.00   61.00

# ocupacions de carril, no s'observen valors erronis o extrems
summary(S7_100$value)

##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    0.0000  0.0000  0.0000   0.0833  0.0000 100.0000

summary(S7_ocup_80_100$value)

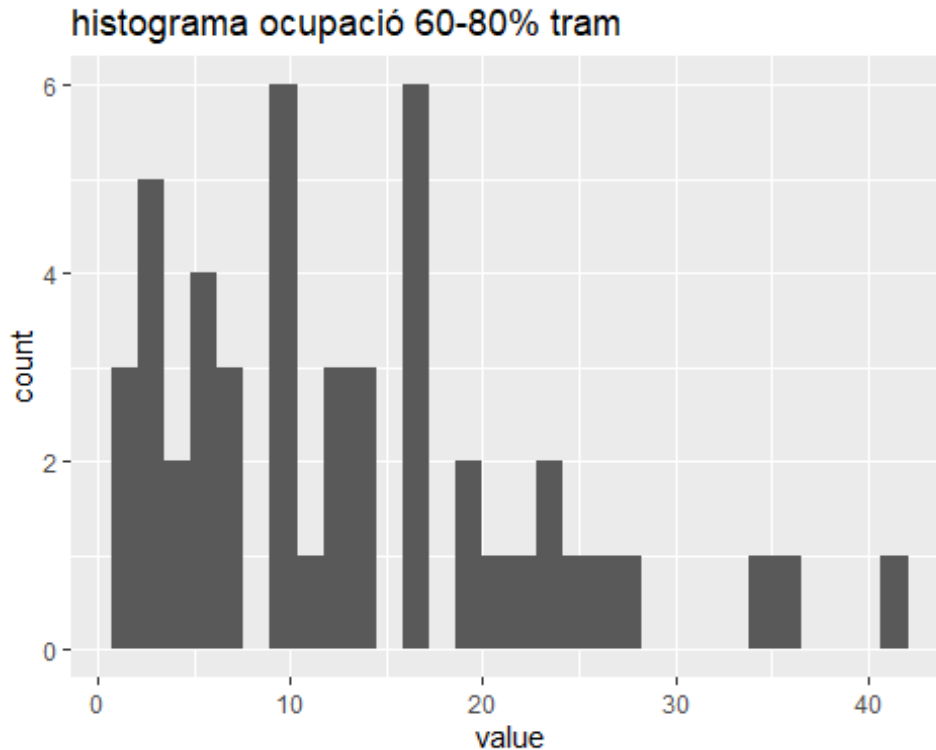
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    0.00000  0.00000  0.00000  0.03685  0.00000  51.00000

summary(S7_ocup_60_80$value)

##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    0.00000  0.00000  0.00000  0.05003  0.00000  42.00000

S7_ocup_60_80 %>%
  filter(value>0) %>%
  ggplot(aes(x=value)) +
  geom_histogram() +
  labs(title="histograma ocupació 60-80% tram")

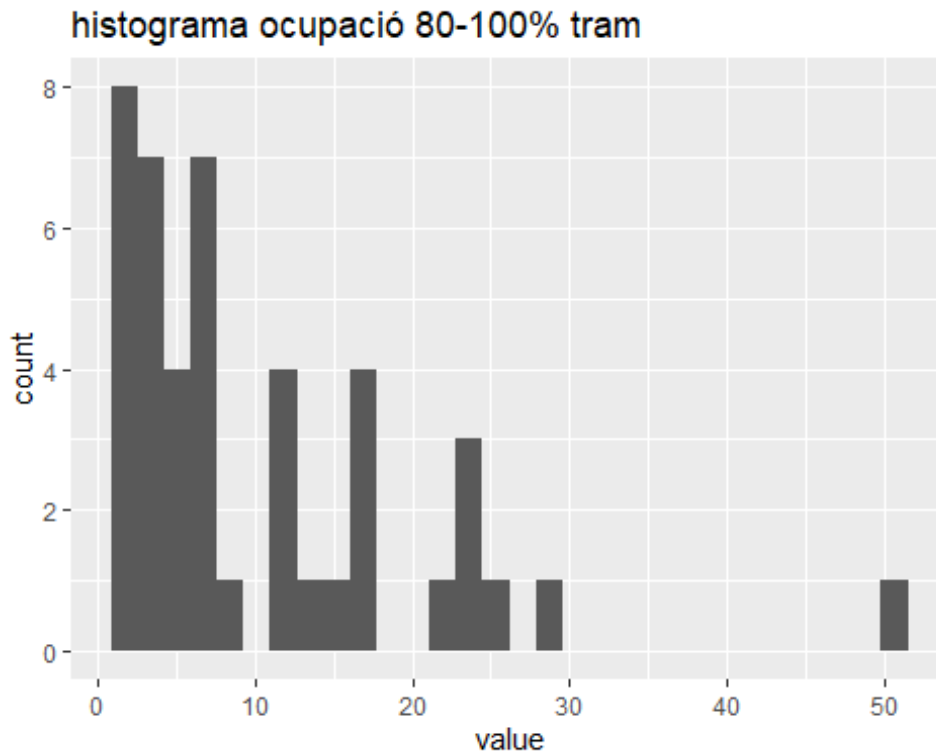
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
S7_ocup_80_100 %>%
  filter(value>0) %>%
```

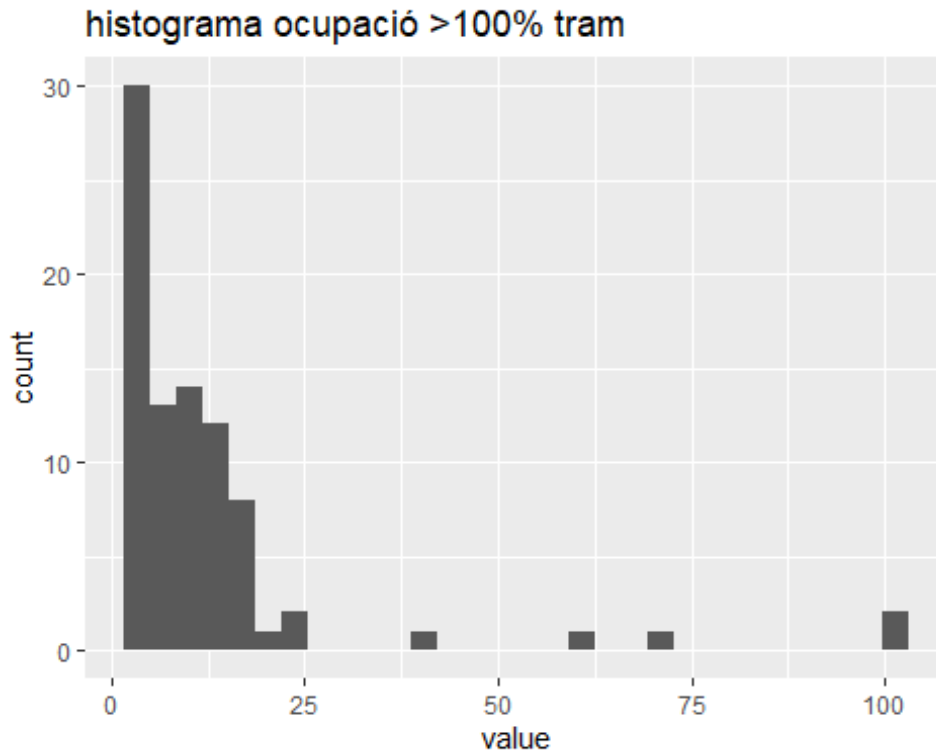
```
ggplot(aes(x=value)) +  
geom_histogram() +  
labs(title="histograma ocupació 80-100% tram")
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



```
S7_100 %>%  
filter(value>0) %>%  
ggplot(aes(x=value)) +  
geom_histogram() +  
labs(title="histograma ocupació >100% tram")
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



5. ANÀLISI DE LES DADES

5.1. Selecció dels grups de dades que es volen analitzar

La variable objecte de l'anàlisi és la variable retard (del dataset dades.clean).

En un dels anàlisis, volem determinar quines de les diferents variables que tenim disponibles influeixen en els retards, i en quina mesura, així que treballarem amb aquestes variables, que són totes les que d'entrada són susceptibles d'influir:

- retard
- dataT
- horaT
- minT
- diastna
- expedicio
- línia
- parada
- ordre.parada
- ordre.expedicio
- conductor

- prec

```
dades.w <- dades.clean[ , c("retard", "dataT", "horaT", "minT", "diastna",
"iIdExpedicion", "iIdLinea", "sDenominacion", "iOrden", "iOrdenExpedicion",
"iIdConductor", "prec")]
```

```
dades.w$iOrden <- ordered(dades.w$iOrden)
```

```
dades.w$horaT <- ordered(dades.w$horaT)
```

```
dades.w$minT <- ordered(dades.w$minT)
```

El conjunt de dades consta de 281174 registres. En aquest punt realitzarem algunes visualitzacions per intentar observar possibles relacions, però un volum de dades tan gran dificulta la comprensió d'algunes gràfiques, de manera que en ocasions treballarem amb una mostra de les dades, del tipus SRSWOR.

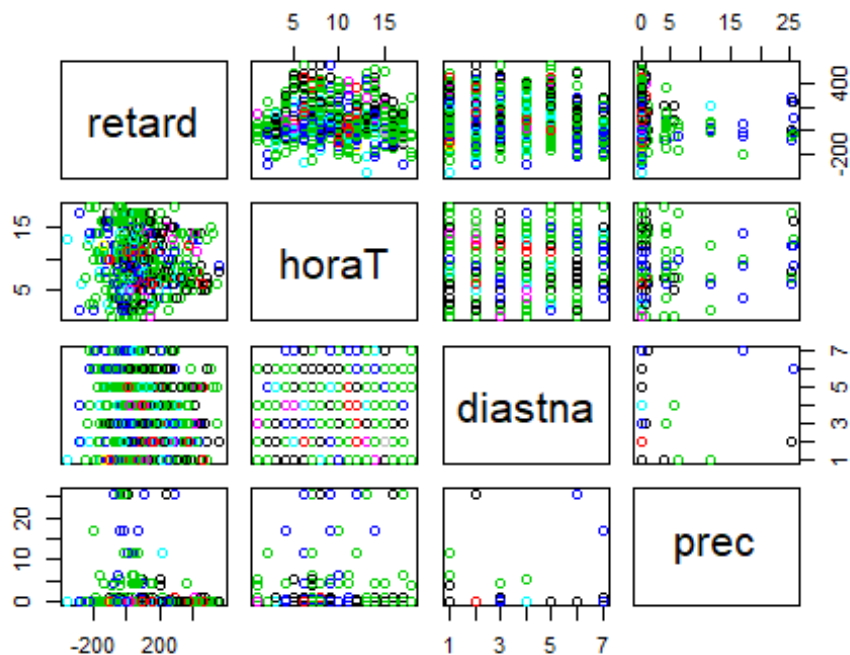
```
# mostra de 500 registres
```

```
dades.sample <- dades.w[sample(nrow(dades.w),500), ]
```

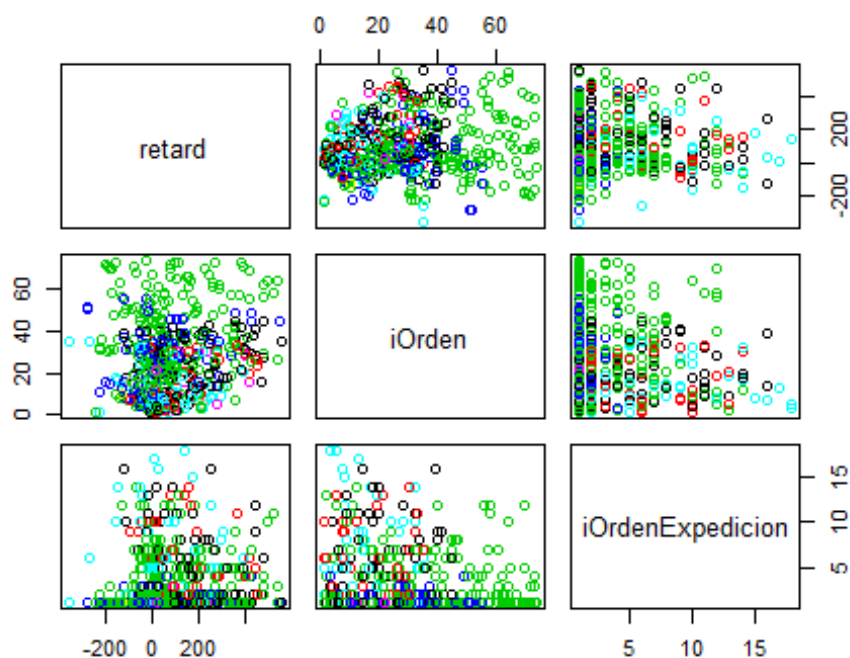
```
summary(dades.sample)
```

```
##      retard      dataT      horaT      minT      di
astna
## Min.   :-356.00   Min.   :2020-02-01   10    : 40   40    : 17   lu
:81
## 1st Qu.: -15.25   1st Qu.:2020-02-12   18    : 38   9     : 16   ma
:80
## Median :  59.50   Median :2020-02-21   14    : 37   19    : 15   mi
:93
## Mean   :  91.12   Mean   :2020-03-04   11    : 36   31    : 14   ju
:65
## 3rd Qu.: 189.50   3rd Qu.:2020-04-04   16    : 36   45    : 13   vi
:86
## Max.   : 558.00   Max.   :2020-04-30   12    : 34   0     : 12   sá
:60
##                                     (Other):279   (Other):413   do
:35
## iIdExpedicion   iIdLinea      sDenominacion   iOrden
## 117521 :  7    20      :223   Dom Bosco 1    : 11   13    : 16
## 117784 :  7    10      : 97   La Salle 1    :  9    3     : 15
## 117786 :  7    21      : 55   Oques 2      :  9    6     : 15
## 117801 :  7    11      : 35   Oques 3      :  9   31    : 14
## 117511 :  6   60      : 35   RENFE 1      :  9    9     : 13
## 117518 :  6   50      : 23   Camí de Valls 1:  8   17    : 13
## (Other):460   (Other): 32   (Other)      :445   (Other):414
## iOrdenExpedicion iIdConductor   prec
## 0      :245      88      : 30   Min.     : 0.00
## 1      : 68      94      : 27   1st Qu.: 0.00
## 2      : 27      12      : 25   Median : 0.00
## 3      : 25      85      : 24   Mean    : 1.06
## 4      : 25      37      : 20   3rd Qu.: 0.00
## 5      : 22      91      : 20   Max.    :25.70
## (Other): 88      (Other):354
```

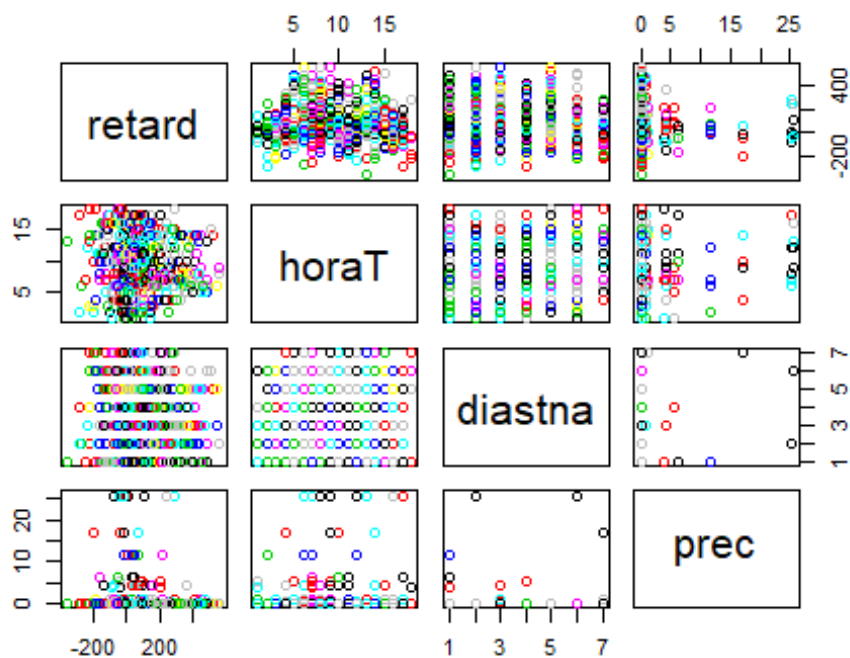
```
# gràfica de punts entre variables amb distinció de línia per color
plot(dades.sample[,c("retard", "horaT", "diastna", "prec")], col=dades.s
ample$iIdLinea)
```



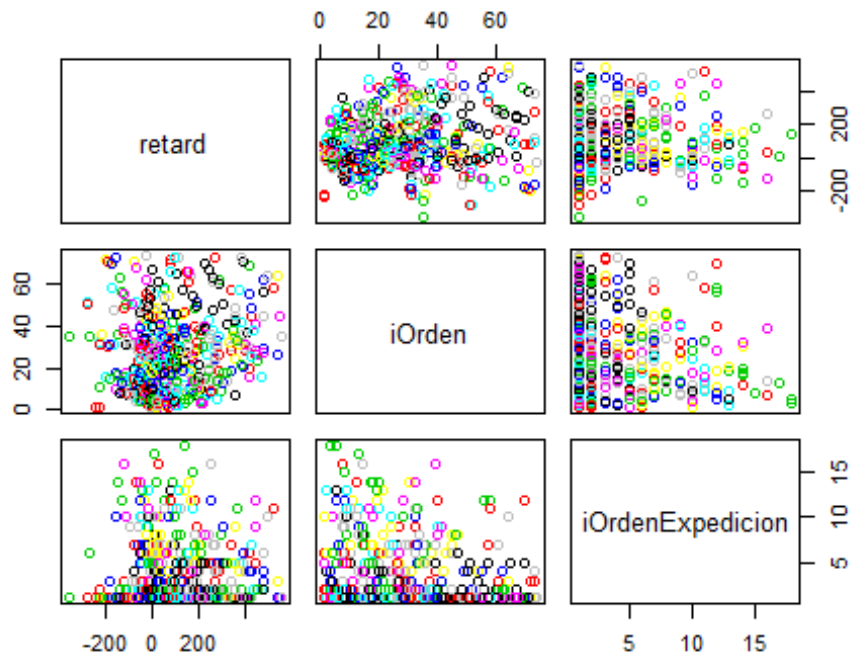
```
plot(dades.sample[,c("retard", "iOrden", "iOrdenExpedicion")], col=dades
.sample$iIdLinea)
```



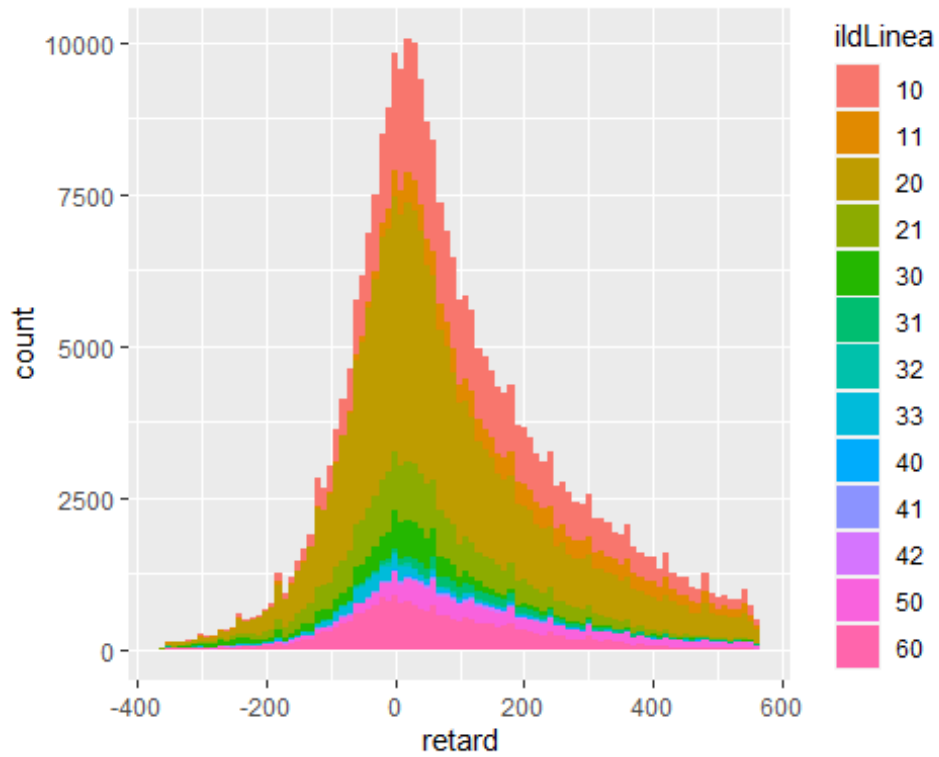
```
# gràfica de punts entre variables amb distinció de conductor per color
plot(dades.sample[,c("retard", "horaT", "diastna", "prec")], col=dades.sample$iIdConductor)
```



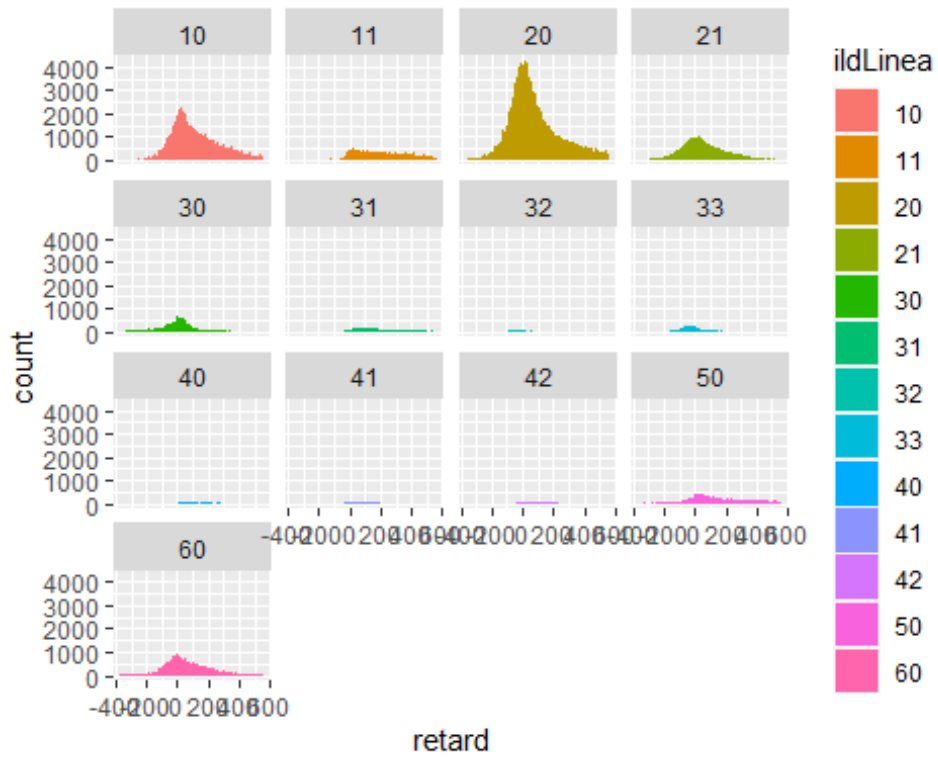
```
plot(dades.sample[,c("retard", "iOrden", "iOrdenExpedicion")], col=dades.sample$iIdConductor)
```



```
# histograma de retards en funció de La línia
dades.w %>%
  ggplot(aes(retard, fill=iIdLinea)) +
  geom_histogram(binwidth = 10)
```

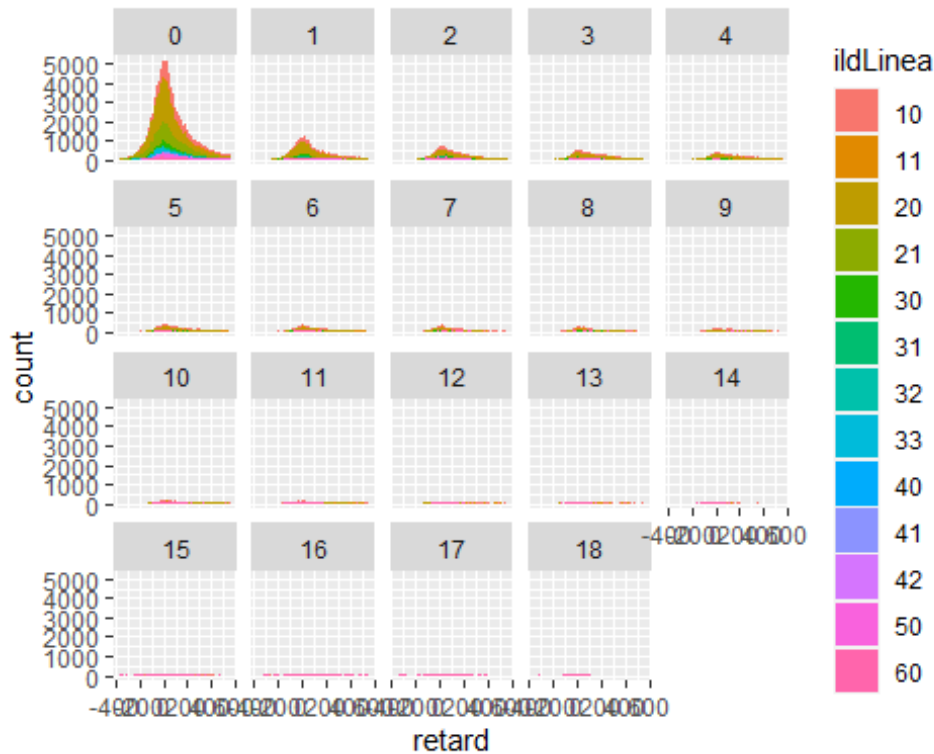
```
dades.w %>%  
  ggplot(aes(retard, fill=ildLinea)) +  
  geom_histogram(binwidth = 10) +  
  facet_wrap(dades.w$ildLinea)
```



```
# histograma de retards en funció del conductor
dades.w %>%
  ggplot(aes(retard, fill=iIdLinea)) +
  geom_histogram(binwidth = 10) +
  facet_wrap(dades.w$iIdConductor)
```



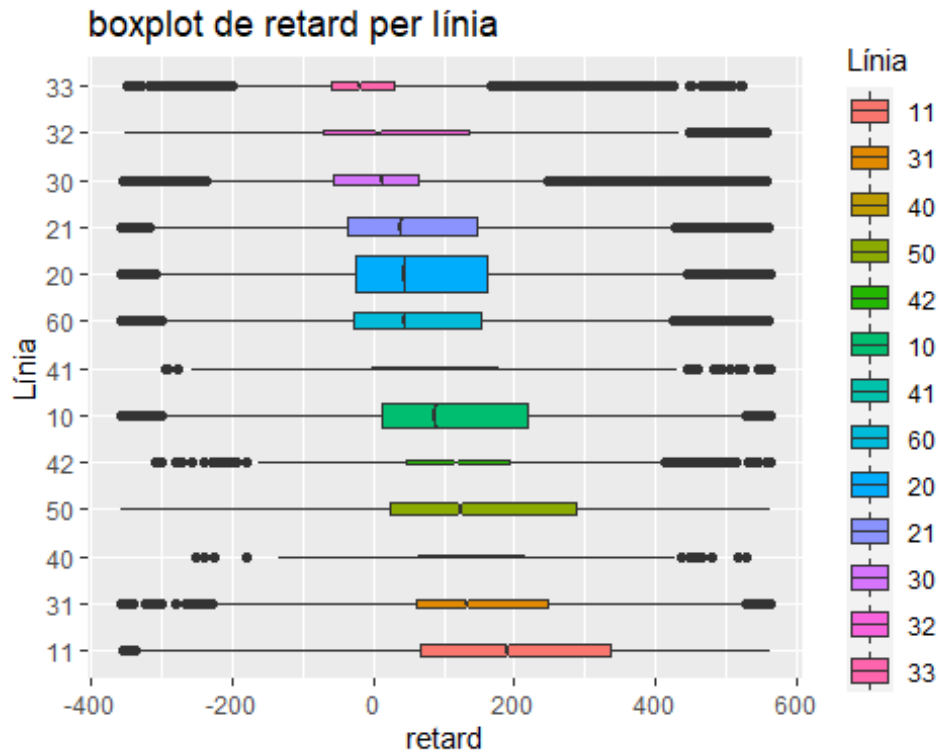
```
# histograma de retards en funció de l'ordre d'expedició
dades.w %>%
  ggplot(aes(retard, fill=iIdLinea)) +
  geom_histogram(binwidth = 10) +
  facet_wrap(dades.w$iOrdenExpedicion)
```



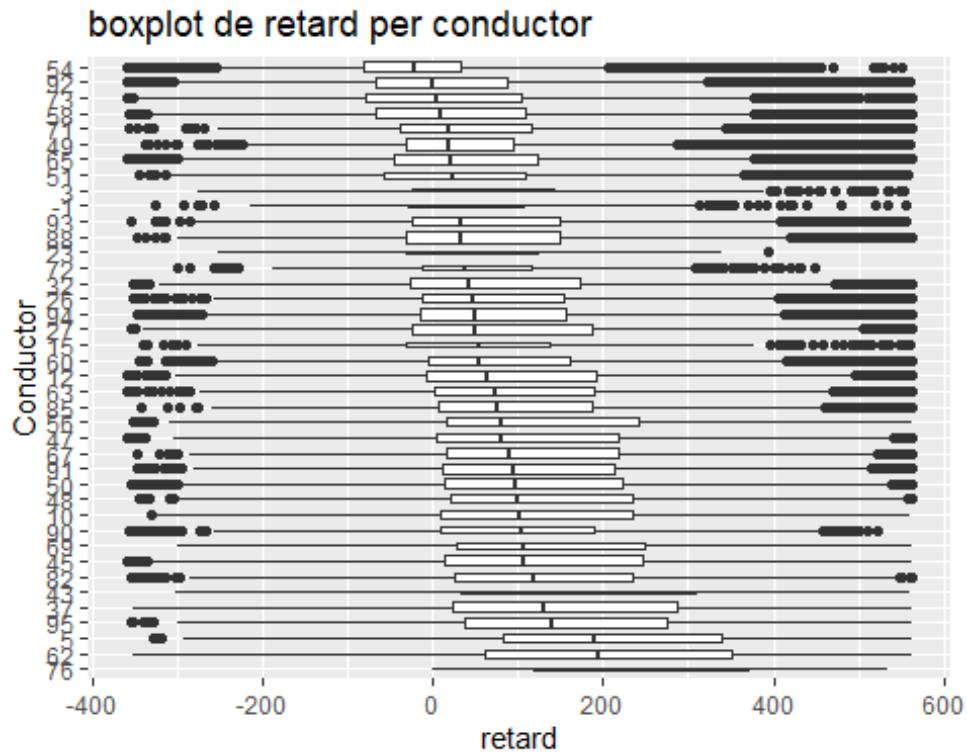
```
# histograma de retards en funció de l'ordre de parada
#dades.w %>%
# ggplot(aes(retard, fill=iIdLinea)) +
#   geom_histogram(binwidth = 10) +
#   facet_wrap(dades.w$iOrden)

# mitjana de retard per ordre de parada i línia
#dades.w %>%
#   ggplot(aes(x=iOrden, colour=iIdLinea)) +
#     geom_bar(dades.w$retard)

# boxplot de retards en funció de línia
dades.w %>%
  ggplot(aes(y=fct_reorder(iIdLinea, retard, FUN = median, na.rm=TRUE, .desc = TRUE), x=retard))+
  geom_boxplot(aes(fill=fct_reorder(iIdLinea, retard, FUN = median, na.rm=TRUE, .desc = TRUE)), notch = TRUE, varwidth = TRUE) +
  labs(title="boxplot de retard per línia", y="Línia" ) +
  scale_fill_discrete(guide=guide_legend(title="Línia"))
```

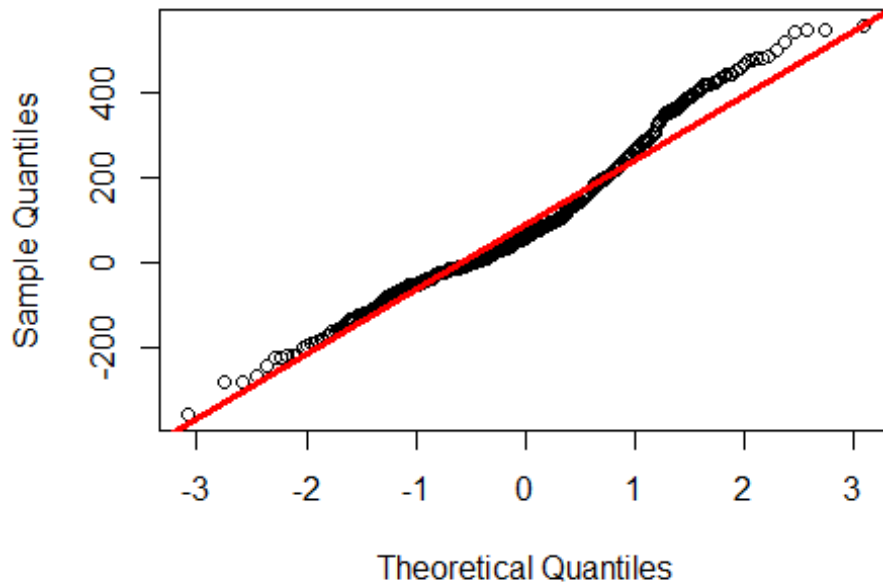


```
# boxplot de retards en funció del conductor
dades.w %>%
  ggplot(aes(y=fct_reorder(iIdConductor, retard, FUN = median, na.rm=TRUE),
    .desc = TRUE), x=retard))+
  geom_boxplot(varwidth = TRUE) +
  labs(title="boxplot de retard per conductor", y="Conductor" ) +
  scale_fill_discrete(guide=guide_legend(title="Línia"))
```



```
# visualitzem si la distribució dels retards s'acosta a una distribució normal
qqnorm(dades.sample$retard, main="Gràfica Q-Q per la variable retard")
qqline(dades.sample$retard, col="red", lwd=3)
```

Gràfica Q-Q per la variable retard



Sembla que la distribució de retards s'allunya de la distribució normal, especialment en els extrems. Això pot ser degut a que no hem eliminat els valors extrems.

Repetim la visualització aquest cop amb la distribució de retards sense valors extrems:

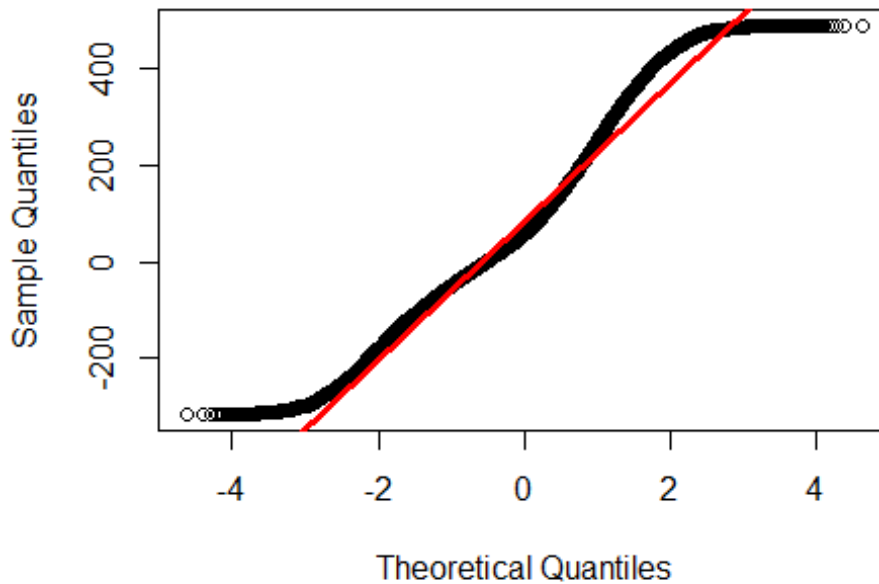
```
out.ret.w <- boxplot.stats(dades.w$retard)$out
summary(out.ret.w)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -358.0   501.0   520.0   451.9   540.0   562.0

dades.w.sense.out <- dades.w %>%
  filter(!(retard %in% out.ret.w))

qqnorm(dades.w.sense.out$retard, main="Gràfica Q-Q per la variable retard
, sense outliers")
qqline(dades.w.sense.out$retard, col="red", lwd=3)
```

Gràfica Q-Q per la variable retard, sense outliers



La distribució de retards sense valors extrems presenta un millor ajust a una distribució normal però no sembla que la segueixi.

Dels histogrames i diagrames de caixa anteriors observem que sembla que els retards es veuen influïts per:

- hora
- dia de la setmana
- línia
- ordre de parada
- conductor

En canvi, la precipitació no sembla influir.

La variable que representa l'ordre d'expedició també sembla influir clarament, però buscant explicació a la quantitat anormal de zeros s'observa que quan es desvia un autobús de la ruta per recuperar horari, s'assigna un zero com si fos una primera expedició. Això pot confondre els resultats.

Pel que fa a la variable ordre, que representa la posició ordenada de les parades a cada línia, veiem que es pot millorar l'exactitud de la seva importància si creem una nova variable que la normalitzi en relació amb la quantitat de parades de la línia objecte de la mesura (en una parada poden passar diverses línies).

Creem per tant una nova parada que contingui la posició relativa entre 0 i 1. Cal tenir present que les línies són circulars, de manera que sobre el 50% del recorregut hi ha un final de línia camuflat, on l'autobús disposa d'un temps de coixí per posar-se en hora.

```
# table(dades.w$iOrden,dades.w$iIdLinea)

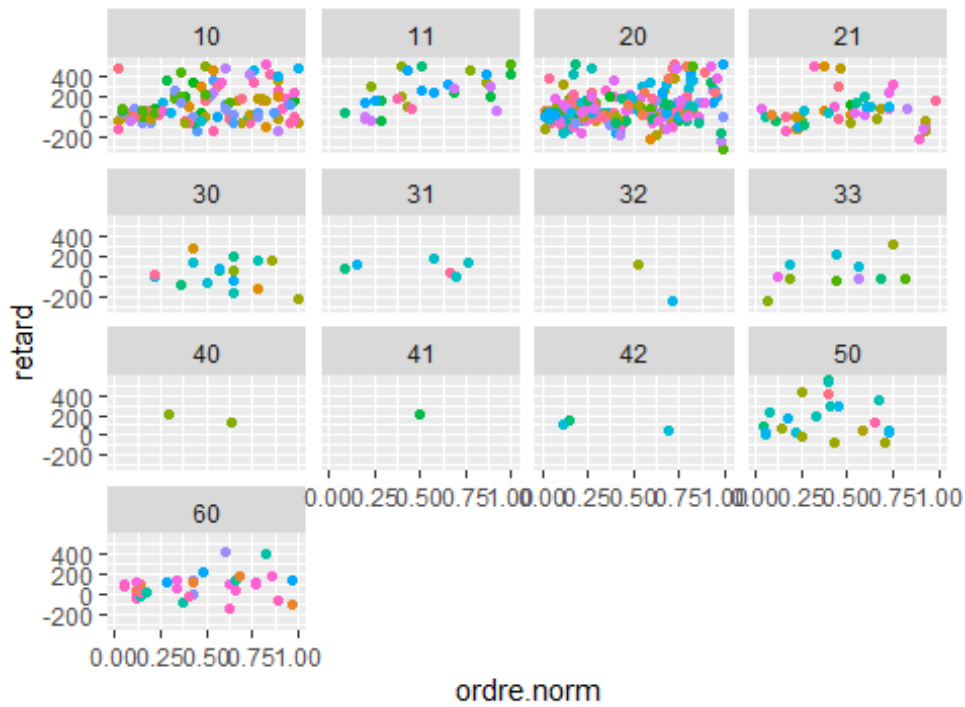
for (a in levels(dades.w$iIdLinea)) {
  # número de parades de cada línia
  assign(paste0("max.L",a),max(as.numeric(dades.w$iOrden[dades.w$iIdLinea
==a])))
}

# afegim una columna amb el número de parades total de la línia
for (i in seq(nrow(dades.w))) {
  a=get(paste0("max.L",dades.w$iIdLinea[i]))
  dades.w$ordre.max[i]=as.numeric(a)
  # dades.w$ordre.norm[i]=round(dades.w$iOrden[i]/as.numeric(a),2)
}

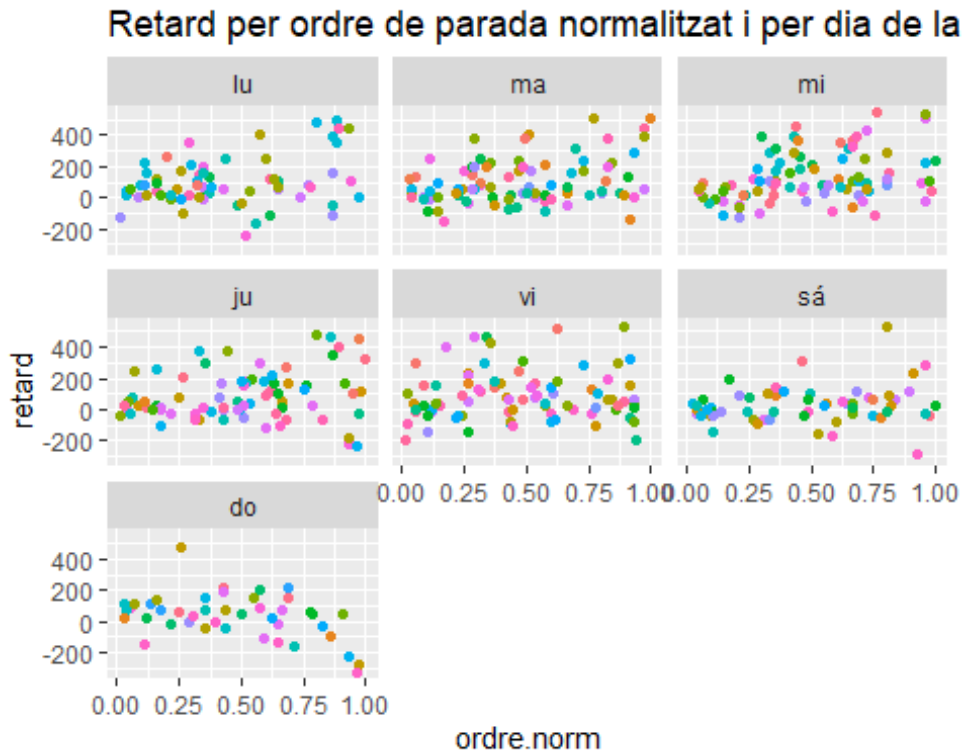
# afegim una variable amb la posició de la parada respecte el total de la
línia
for (i in seq(nrow(dades.w))) {
  dades.w$ordre.norm[i] <- as.numeric(dades.w$iOrden[i])/dades.w$ordre.ma
x[i]
}

dades.w[sample(nrow(dades.w),500), ] %>%
  ggplot(aes(x=ordre.norm, y=retard)) +
  geom_point(mapping = aes(color=iIdConductor), show.legend = FALSE) +
  facet_wrap(~iIdLinea) +
  labs(title="Retard per ordre de parada normalitzat i per conductor (mos
tra n=500)")
```

Retard per ordre de parada normalitzat i per conductor



```
dades.w[sample(nrow(dades.w), 500), ] %>%
  ggplot(aes(x=ordre.norm, y=retard)) +
  geom_point(mapping = aes(color=iIdConductor), show.legend = FALSE) +
  facet_wrap(~diastna) +
  labs(title="Retard per ordre de parada normalitzat i per dia de la setmana (mostra n=500)")
```



5.2. Comprovació de la normalitat i homogeneïtat de la variància.

Aplicuem a les dues variables numèriques (retard i precipitació) els test de Kolmogorov-Smirnov i de Shapiro-Wilk, que assumeixen com a hipòtesis nul·la que les dades segueixen una distribució normal.

```
ks.test(dades.w$retard, pnorm, mean(dades.w$retard, sd(dades.w$retard)))

##
## One-sample Kolmogorov-Smirnov test
##
## data: dades.w$retard
## D = 0.49085, p-value < 2.2e-16
## alternative hypothesis: two-sided

ks.test(dades.w$prec, pnorm, mean(dades.w$prec, sd(dades.w$prec)))

##
## One-sample Kolmogorov-Smirnov test
##
## data: dades.w$prec
## D = 0.5, p-value < 2.2e-16
## alternative hypothesis: two-sided

# el test de Shapiro-Wilk està limitat a una mostra de tamany 5000
dades.sample5000 <- dades.w[sample(nrow(dades.w), 5000), ]
shapiro.test(dades.sample5000$retard)
```

```
##
## Shapiro-Wilk normality test
##
## data:  dades.sample5000$retard
## W = 0.96398, p-value < 2.2e-16

shapiro.test(dades.sample5000$prec)

##
## Shapiro-Wilk normality test
##
## data:  dades.sample5000$prec
## W = 0.2501, p-value < 2.2e-16
```

El p-valor obtingut en tots els casos és inferior al nivell de significació 0,05, de manera que rebutgem la hipòtesis nul·la i concloem que les dades no segueixen una distribució normal.

No obstant, donat el tamany de la mostra i aplicant el teorema central del límit, podem considerar que la distribució de la mitjana de retards segueix una distribució normal de mitjana = mitjana(retards) i variança = sd(retards)^2/N.

Per comprovar l'homocedasticitat de la variable retard segons els diferents factors, apliquem el test de Fligner-Killeen, ja que les dades no compleixen amb la condició de normalitat.

```
fligner.test(retard~iIdLinea, data=dades.w)

##
## Fligner-Killeen test of homogeneity of variances
##
## data:  retard by iIdLinea
## Fligner-Killeen:med chi-squared = 4103.5, df = 12, p-value < 2.2e-16

fligner.test(retard~iIdConductor, data=dades.w)

##
## Fligner-Killeen test of homogeneity of variances
##
## data:  retard by iIdConductor
## Fligner-Killeen:med chi-squared = 5715, df = 39, p-value < 2.2e-16

fligner.test(retard~iOrden, data=dades.w)

##
## Fligner-Killeen test of homogeneity of variances
##
## data:  retard by iOrden
## Fligner-Killeen:med chi-squared = 23435, df = 74, p-value < 2.2e-16
```

Donat que en tots els casos el p-valor és inferior al nivell de significació, es rebutja la hipòtesis nul·la d'homocedasticitat, i es conclou que el retard presenta variacions estadísticament diferents segons la línia, conductor i ordre de parada.

5.3. Aplicació de proves estadístiques per comparar els grups de dades.

5.3.1. Contrast d'hipòtesis: la L20 té una distribució de retards diferent a les altres línies?

Pel teorema central del límit podem aplicar proves paramètriques per determinar si hi ha diferències significatives en el retard entre els grups definits per les altres variables, de manera que podem aplicar t-test.

Hipòtesis nul·la: mitjana de retards de la L20 = mitjana de retards de la resta de línies.

```
# creació dels dos grups diferenciant L20 de la resta
for (i in seq(nrow(dades.w))) {
  if (dades.w$IdLinea[i]==20) {
    dades.w$isL20[i] = 1
  } else {
    dades.w$isL20[i] = 0
  }
}

## Warning: Unknown or uninitialised column: `isL20`.

# contrast d'hipòtesis
t.test(retard~isL20, data=dades.w)

##
## Welch Two Sample t-test
##
## data:  retard by isL20
## t = 37.46, df = 246910, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  22.25758 24.71525
## sample estimates:
## mean in group 0 mean in group 1
##      104.31865      80.83224

wilcox.test(retard~isL20, data=dades.w)

##
## Wilcoxon rank sum test with continuity correction
##
## data:  retard by isL20
## W = 1.0515e+10, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Obtenim un p-valor més petit que el nivell de significació, fet que implica que hi ha diferències significatives entre els retards de la L20 i de la resta de línies, tot i que els retards són en sentit contrari de l'esperat, en resulta que la L20 té menys retards de mitjana (81

segons) que la resta de línies (104 segons). El resultat coincideix si apliquem les proves de Wilcoxon i Mann-Whitney per a dades que no segueixen una distribució normal.

Donat que tenim diverses línies, volem aprofundir en la variació de la mitjana de retards en elles, pel que apliquem un test ANOVA, tant pel que fa a les línies com a les altres variables categòriques.

```
res.aov.linea <- aov(retard~iIdLinea, data = dades.w)
summary(res.aov.linea)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## iIdLinea      12 5.455e+08 45454797   1824 <2e-16 ***
## Residuals    281161 7.008e+09    24925
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.cond <- aov(retard~iIdConductor, data = dades.w)
summary(res.aov.cond)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## iIdConductor   39 6.521e+08 16721223   681.2 <2e-16 ***
## Residuals     281134 6.901e+09    24548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.exp <- aov(retard~iOrdenExpedicion, data = dades.w)
summary(res.aov.exp)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## iOrdenExpedicion  18 2.541e+08 14119209   543.9 <2e-16 ***
## Residuals       281155 7.299e+09    25962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.hora <- aov(retard~horaT, data = dades.w)
summary(res.aov.hora)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## horaT          17 6.118e+08 35990887   1458 <2e-16 ***
## Residuals     281156 6.942e+09    24689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.min <- aov(retard~minT, data = dades.w)
summary(res.aov.min)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## minT           59 1.375e+08 2330289    88.33 <2e-16 ***
## Residuals     281114 7.416e+09    26380
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

res.aov.parada <- aov(retard~sDenominacion, data = dades.w)
summary(res.aov.parada)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## sDenominacion   204 9.671e+08 4740906   202.2 <2e-16 ***
## Residuals      280969 6.586e+09   23441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(aov(retard~iIdConductor+iIdLinea+ordre.norm+horaT+diastna+sDenomi
nacion, data=dades.w))

##              Df      Sum Sq   Mean Sq F value Pr(>F)
## iIdConductor   39 6.521e+08 16721223   975.3 <2e-16 ***
## iIdLinea       12 5.411e+08 45095118  2630.3 <2e-16 ***
## ordre.norm      1 3.469e+08 346850224 20231.2 <2e-16 ***
## horaT          17 5.272e+08 31011700  1808.9 <2e-16 ***
## diastna         6 9.307e+07 15511758   904.8 <2e-16 ***
## sDenominacion  204 5.772e+08  2829640   165.0 <2e-16 ***
## Residuals     280894 4.816e+09   17144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

En tots els casos s'obté que els retards són estadísticament diferent segons els grups (línia, conductor, expedició, hora, minut, parada).

Repetim l'anàlisi sense valors extrems (en l'apartat de preparació de dades havíem decidit mantenir tots els valors extrems positius), que podrien emascarar diferències no significatives en la majoria de casos.

```

res.aov.linea.o <- aov(retard~iIdLinea, data = dades.w.sense.out)
summary(res.aov.linea.o)

##              Df      Sum Sq   Mean Sq F value Pr(>F)
## iIdLinea      12 4.413e+08 36775855   1731 <2e-16 ***
## Residuals    274331 5.827e+09   21242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.cond.o <- aov(retard~iIdConductor, data = dades.w.sense.out)
summary(res.aov.cond.o)

##              Df      Sum Sq   Mean Sq F value Pr(>F)
## iIdConductor   39 5.112e+08 13108303   624.5 <2e-16 ***
## Residuals     274304 5.758e+09   20990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.exp.o <- aov(retard~iOrdenExpedicion, data = dades.w.sense.out)
summary(res.aov.exp.o)

##              Df      Sum Sq   Mean Sq F value Pr(>F)
## iOrdenExpedicion 18 2.057e+08 11430166   517.2 <2e-16 ***

```

```
## Residuals      274325 6.063e+09    22101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.hora.o <- aov(retard~horaT, data = dades.w.sense.out)
summary(res.aov.hora.o)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## horaT         17 4.793e+08 28191944   1336 <2e-16 ***
## Residuals    274326 5.789e+09    21104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.min.o <- aov(retard~minT, data = dades.w.sense.out)
summary(res.aov.min.o)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## minT          59 1.116e+08 1891422    84.26 <2e-16 ***
## Residuals    274284 6.157e+09    22448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

res.aov.parada.o <- aov(retard~sDenominacion, data = dades.w.sense.out)
summary(res.aov.parada.o)

##              Df      Sum Sq Mean Sq F value Pr(>F)
## sDenominacion 204 7.561e+08 3706149   184.3 <2e-16 ***
## Residuals    274139 5.513e+09    20109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(aov(retard~iIdConductor+iIdLinea+horaT+diastna+sDenominacion, data=dades.w.sense.out))

##              Df      Sum Sq Mean Sq F value Pr(>F)
## iIdConductor   39 5.112e+08 13108303   867.7 <2e-16 ***
## iIdLinea       12 4.406e+08 36719329  2430.7 <2e-16 ***
## horaT          17 4.287e+08 25215696  1669.2 <2e-16 ***
## diastna         6 7.230e+07 12049614   797.6 <2e-16 ***
## sDenominacion  204 6.757e+08 3312407   219.3 <2e-16 ***
## Residuals    274065 4.140e+09    15107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Es confirma que la mitjana dels retards és diferent en cadascun dels grups.

Els valors de F indiquen que les diferències més grans es donen en l'ordre norm seguides a molta diferència per la línia, l'hora i el conductor.

5.3.2. PCA

Aplicarem la tècnica de l'Anàlisi de Components Principals per veure quins factors són els que influeixen més en el retard mesurat a cada parada. Aquest mètode treballa amb

similituds i distàncies. Les dades haurien d'estar prèviament normalitzades per comparar entre els grups. Només es pot aplicar a les variables numèriques.

```
# normalitzem les variables retard, precipitació, ordre de parada relatiu a la línia
```

```
dades.w.norm <- as.tibble(scale(dades.w[,c(1,12,14)]))
```

```
# afegim les variables categòriques
```

```
dades.w.norm <- cbind(dades.w.norm,dades.w[,c(2:5,7,10,11)])
```

```
PCA.dades.norm <- prcomp((dades.w.norm[,c(1:3)]))
```

```
summary(PCA.dades.norm)
```

```
## Importance of components:
```

```
##                PC1      PC2      PC3
## Standard deviation    1.1058 1.0021 0.8793
## Proportion of Variance 0.4076 0.3347 0.2577
## Cumulative Proportion 0.4076 0.7423 1.0000
```

```
# matriu de canvi de coordenades
```

```
PCA.dades.norm$rotation
```

```
##                PC1      PC2      PC3
## retard      0.7102320 -0.01719037  0.7037577
## prec        -0.2684630  0.91753816  0.2933451
## ordre.norm  0.6507673  0.39727600 -0.6470500
```

El resultat d'aquest anàlisi indica que no hi ha reducció de dimensionalitat, tant les variables retard, com precipitació, com ordre de parada (normalitzat) tenen influència significativa en l'explicació de la variança de les dades originals.

5.3.3. FAMD - Factor Analysis of Mixed Data

Aquest mètode és una variació de l'anàlisi de components principals (PCA) per treballar amb datasets que tinguin variables categòriques i numèriques.

```
res.famd2 <- FAMD(dades.w[,c(1,3,7,11,12,14)], ncp=8, graph=FALSE)
get_eigenvalue(res.famd2)
```

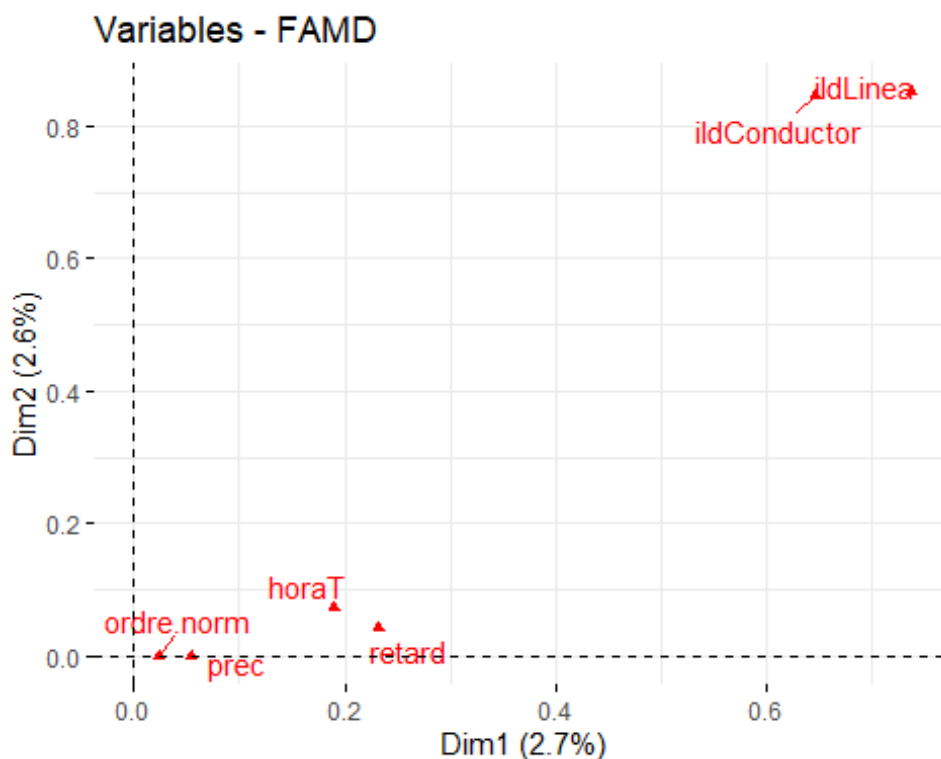
```
##      eigenvalue variance.percent cumulative.variance.percent
## Dim.1    1.883677         2.653067          2.653067
## Dim.2    1.813821         2.554678          5.207745
## Dim.3    1.607873         2.264610          7.472355
## Dim.4    1.583827         2.230743          9.703098
## Dim.5    1.458849         2.054716         11.757814
## Dim.6    1.430281         2.014481         13.772295
## Dim.7    1.306296         1.839854         15.612149
## Dim.8    1.292671         1.820664         17.432813
```

```
# contribució de les variables a les dimensions
```

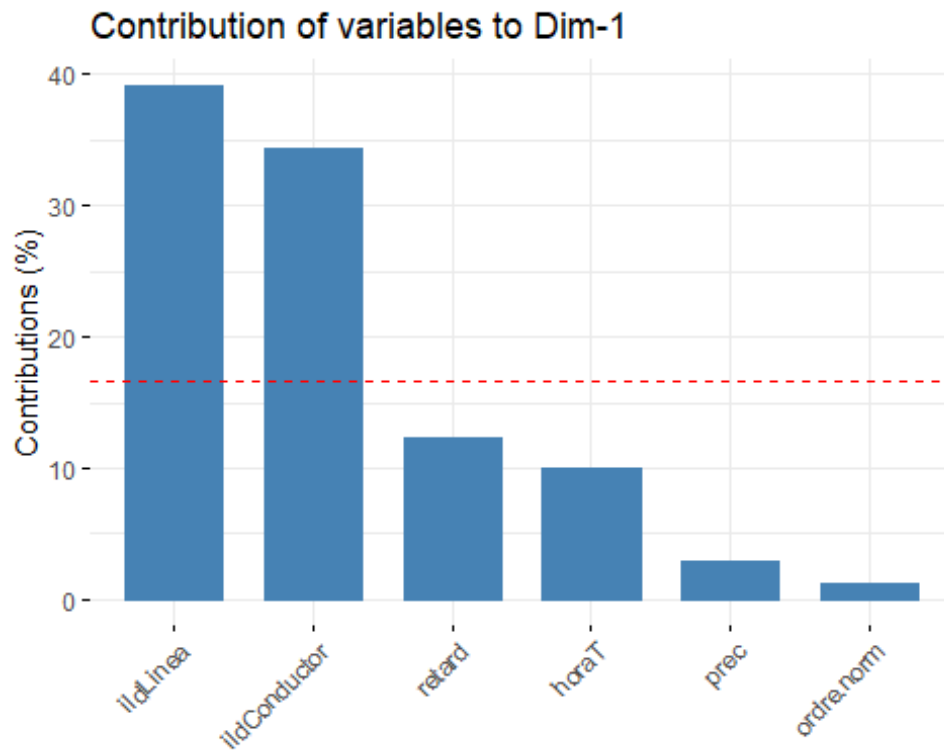
```
get_famd_var((res.famd2))$contrib
```

```
##          Dim.1      Dim.2      Dim.3      Dim.4      Dim.5
## retard    12.289671  2.366842e+00  8.770575e-04  0.06058012  5.700177
## prec      2.910457  1.988699e-04  2.033310e-01  0.24968476  1.969902
## ordre.norm 1.294262  4.042294e-02  6.407025e-01  0.04504588  1.731245
## horaT     10.074307  4.079584e+00  4.921643e+00  3.33953494  34.817488
## iIdLinea   39.147983  4.690756e+01  4.741257e+01  48.64816042  20.837483
## iIdConductor 34.283319  4.660539e+01  4.682087e+01  47.65699387  34.943704
##          Dim.6      Dim.7      Dim.8
## retard    15.533275  0.5807155  6.777235
## prec      10.461652  1.6334614  9.949773
## ordre.norm 1.023595  0.1525199  9.904781
## horaT     4.140632  18.1173896  14.952883
## iIdLinea   32.170608  42.7293182  30.714697
## iIdConductor 36.670238  36.7865955  27.700632
```

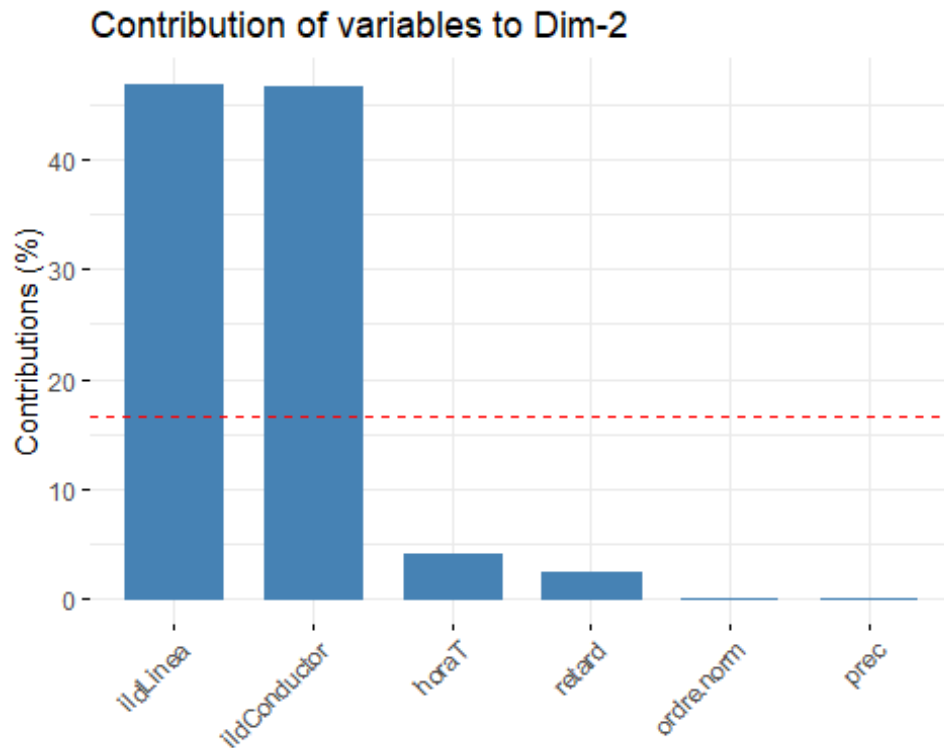
```
# gràfica amb la primera i segona dimensió
fviz_famd_var(res.famd2, repel=TRUE)
```



```
# contribucions de les variables a cada dimensió
fviz_contrib(res.famd2, "var", axes=1)
```

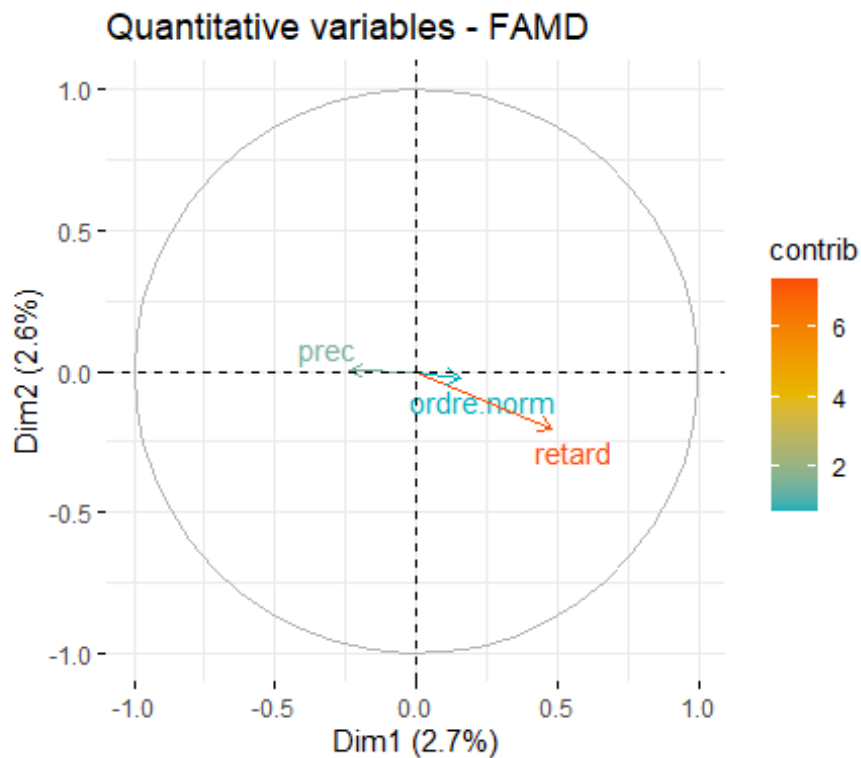


```
fviz_contrib(res.famd2, "var", axes=2)
```



```
# variables quantitatives  
quanti.var <- get_famd_var(res.famd2, "quanti.var")
```

```
fviz_famd_var(res.famd2, "quanti.var", repel=TRUE, col.var="contrib", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
```



```
quanti.var$cos2
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
retard	0.23149777	4.293028e-02	1.410197e-05	0.0009594844	0.0831569
prec	0.05482363	3.607145e-06	3.269305e-03	0.0039545756	0.0287378
ordre.norm	0.02437972	7.332000e-04	1.030168e-02	0.0007134490	0.0252562

	Dim.6	Dim.7	Dim.8
retard	0.22216951	0.007585864	0.08760738
prec	0.14963104	0.021337843	0.12861788
ordre.norm	0.01464029	0.001992361	0.12803628

Després de diverses combinacions de variables, la que més variances explica és retard + precipitació + ordre.norm + hora + línia + conductor. Tot i així, la contribució és molt pobre, entre un 2 i un 3% per cada vector propi.

Tant a la taula com a les gràfiques de contribucions, observem que les 4 primeres dimensions tenen una contribució molt forta de les variables línia i conductor (en igual magnitud) i una contribució menys important de l'hora i retard.

Finalment, els valors de \cos^2 per a les variables numèriques representen la qualitat de la representació de la variable en les dimensions 1 i 2. La variable amb \cos^2 més alt és retard, amb 0.23, que indica que com les altres no està ben representada en dos dimensions, (el \cos^2 seria proper a 1 si ho estigués).

5.3.4. Kernel Regression

Aquest algorisme es pot aplicar per a obtenir una regressió quan la variable resposta és contínua i les variables predictores són diverses i contenen tant variables contínues com discretes o factors.

```
```{r kernel regression, eval=FALSE, include=FALSE}

nprebw(formula = retard ~factor(ildLinea) + factor(ildConductor) + prec + ordre.norm +
ordered(horaT), data=dades.w, regtype = "lc", nmulti=2)

out <- capture.output(

 res.kernel <- nprebw(formula = retard ~ factor(ildLinea) + factor(ildConductor) +
ordre.norm , data=dades.sample5000, regtype="lc", nmulti=2)

)

res.kernel

Regressió

fit.kernel <- npreg(res.kernel)

summary(fit.kernel)
```

```
> summary(fit.kernel)

Regression Data: 5000 training points, in 3 variable(s)
 factor(iIdLinea) factor(iIdConductor) ordre.norm
Bandwidth(s): 0.0153565 0.3149655 0.1094529

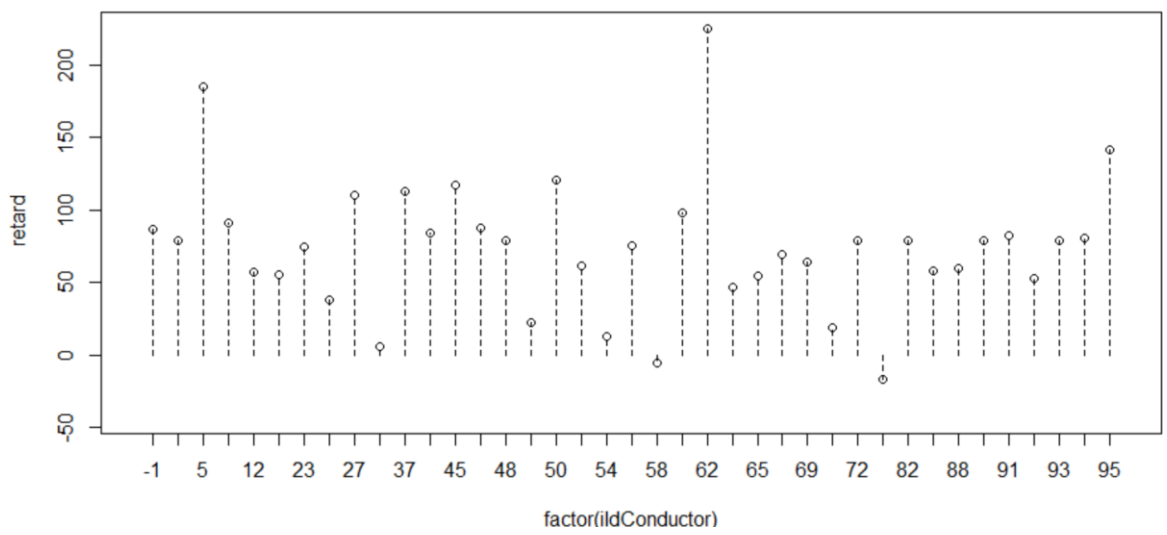
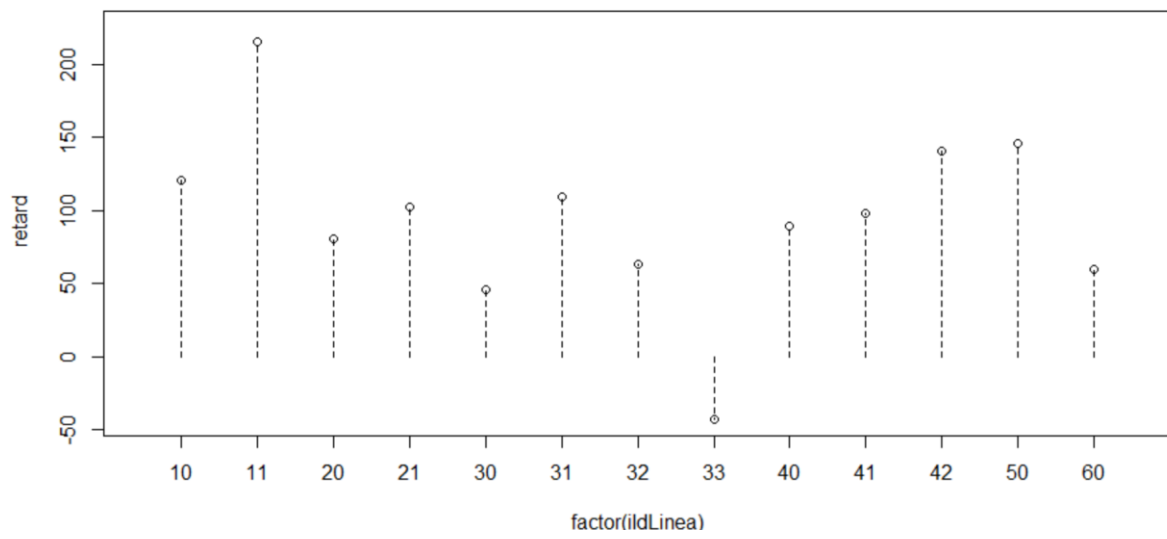
Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed
Residual standard error: 127.5077
R-squared: 0.3743655

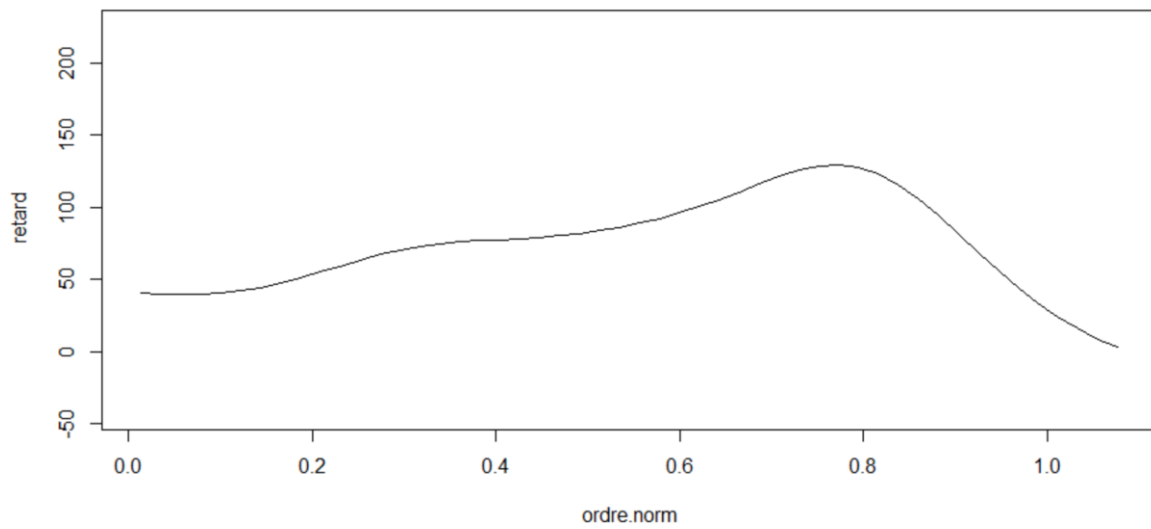
Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 2
```

```
gràfiques dels efectes marginals de cada variable sobre el retard
```

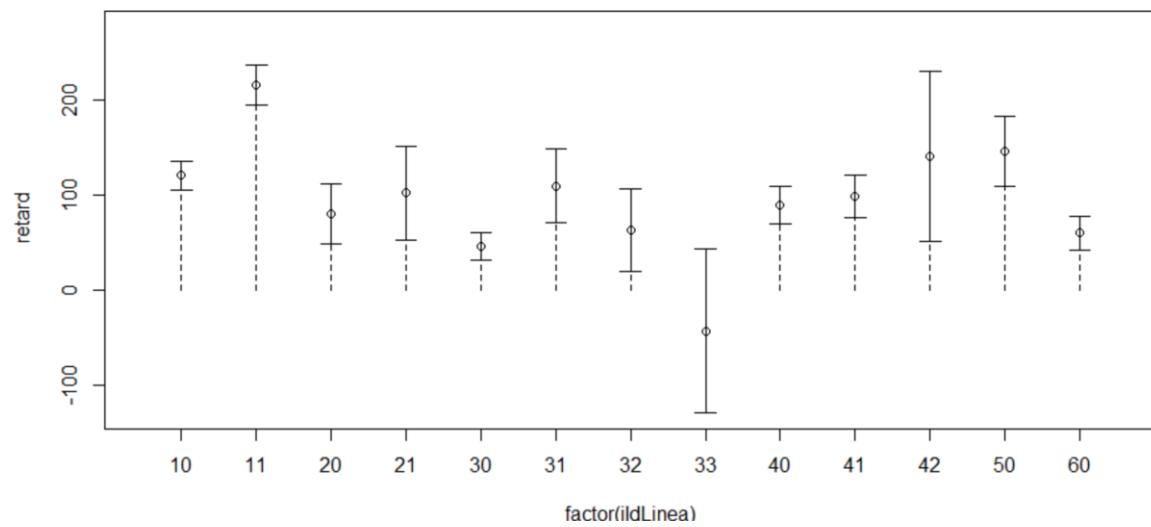
```
plot(fit.kernel, plot.par.mfrow=FALSE)
```

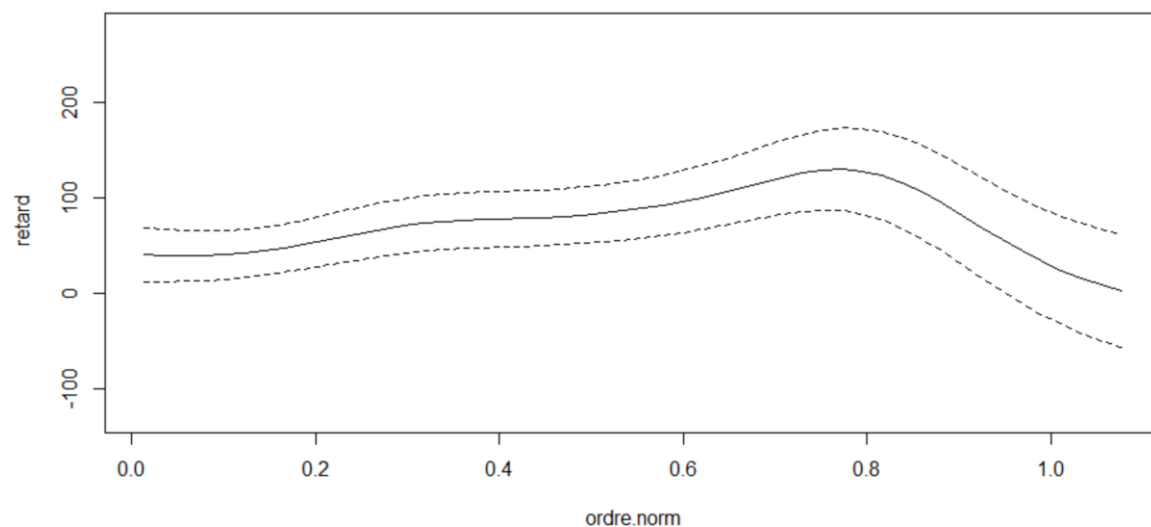
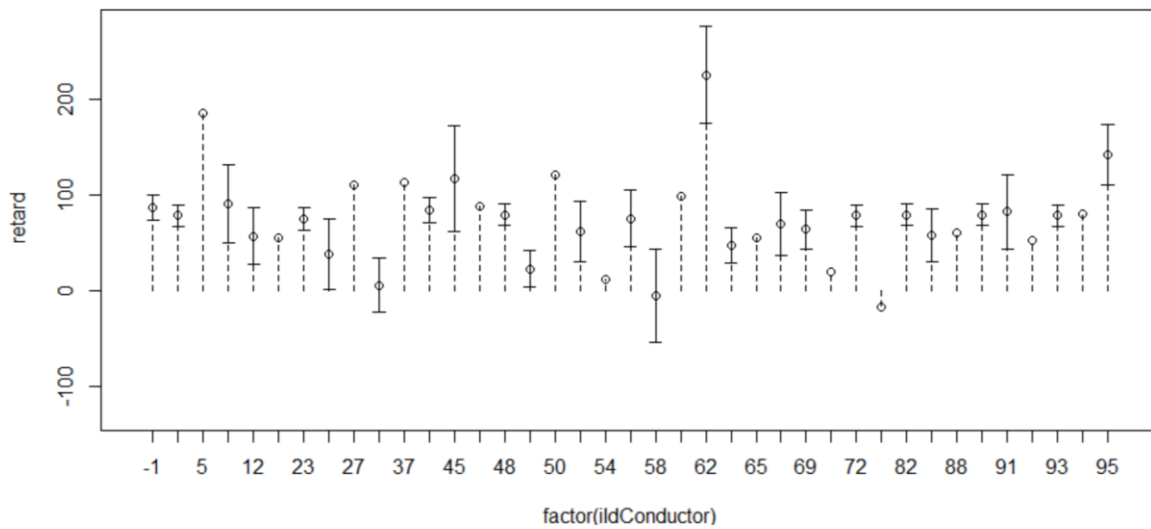




# intervals de confiança pels efectes marginals de cada variable sobre el retard

plot(fit.kernel, plot.errors.method="bootstrap", plot.par.mfrow=FALSE)





...

Com a resultat de la funció summary obtenim uns coeficients per cada variable que reforcen el protagonisme del conductor i les gràfiques amb els intervals de confiança coincideixen en bona mesura amb les gràfiques descriptives de les dades reals. Aquest mètode té una funció predict, que podria estimar els retards mitjans d'un determinat conductor, en una determinada línia i ordre de parada.

Un resultat molt interessant d'aquest anàlisi és l'efecte de l'ordre de la parada (normalitzat en cada línia) en el retard, com mostra que els retards s'acumulen fins cobrir el 75-80% de la línia, moment en que es redueixen fins a valors negatius, reflectint molt probablement la regulació del servei (que un autobús es salti part del trajecte per anar directament a una parada a recuperar horari previst).

No obstant, cal tenir en compte que l'ajust del model a les dades és molt pobre ( $R^2 = 0.374$ )



### 5.3.5. Model no supervisat: clustering: hi ha retards diferents en les parades? o en la combinació parada+línia? o parada+línia+expedició?

Per aplicar un mètode de clustering sobre variables mixtes (numèriques, categòriques i factors ordenats) contemplem una funció distància i un algorisme que ho permetin: la distància de Gower i l'algorisme PAM Partitioning around medoids. L'algorisme és robust en front a valors extrems, donat que treballa amb medianes i no amb mitjanes.

```
matriu amb les distàncies de Gower
gower_dist <- daisy(dades.sample5000[,c(1,3,5,7,14,12,11)], metric="gower")
gower_mat <- as.matrix(gower_dist)

els casos més similars són
dades.sample5000[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind = TRUE)[1,],]

A tibble: 2 x 14
retard dataT horaT minT diastna iIdExpedicion iIdLinea sDenominacion
<dbl> <date> <ord> <ord> <ord> <ord> <fct> <fct>
1 44 2020-02-19 6 33 mi 116892 20 Riudoms
2 38 2020-02-05 6 33 mi 116892 20 Riudoms

... with 6 more variables: iOrden <ord>, iOrdenExpedicion <ord>,
iIdConductor <fct>, prec <dbl>, ordre.max <dbl>, ordre.norm <dbl>

els casos més dissimilars són
dades.sample5000[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind = TRUE)[1,],]

A tibble: 2 x 14
retard dataT horaT minT diastna iIdExpedicion iIdLinea sDenominacion
<dbl> <date> <ord> <ord> <ord> <ord> <fct> <fct>
1 534 2020-02-29 7 52 sá 117282 10 Urbanit
2 -163 2020-04-21 18 15 ma 117784 20 Hospital

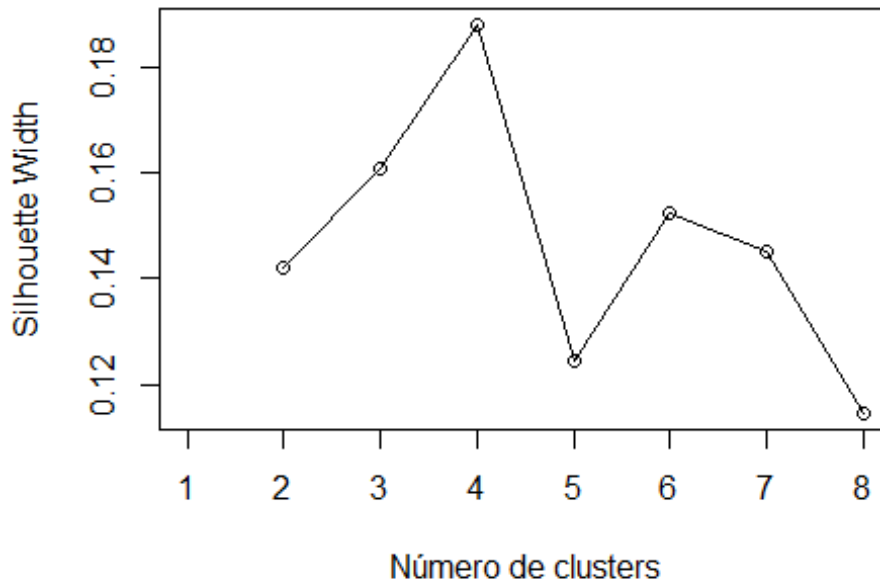
... with 6 more variables: iOrden <ord>, iOrdenExpedicion <ord>,
iIdConductor <fct>, prec <dbl>, ordre.max <dbl>, ordre.norm <dbl>

escollim el número de clústers més baix amb un bon coeficient silhouette
sil_width <- c(NA)
for(i in 2:8){
 pam_fit <- pam(gower_dist, diss = TRUE, k = i)
 sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:8, sil_width,
```

```

xlab = "Número de clusters",
ylab = "Silhouette Width")
lines(1:8, sil_width)

```



```

k=4 ens dona el màxim coeficient silhouette
descripció de cada cluster
k=4
fit.pam <- pam(gower_dist, diss=TRUE, k)

elements representatius del cluster
dades.sample5000[fit.pam$medoids,c(1,3,5,7,14,12,11)]

A tibble: 4 x 7
retard horaT diastna iIdLinea ordre.norm prec iIdConductor
<dbl> <ord> <ord> <fct> <dbl> <dbl> <fct>
1 33 13 ju 20 0.36 0 49
2 116 14 ma 60 0.514 0 54
3 111 12 ju 10 0.578 0 26
4 -31 15 vi 21 0.607 0 32

característiques dels clústers
fit.pam$clusinfo

size max_diss av_diss diameter separation
[1,] 2248 0.5008145 0.2754476 0.8013066 0.005319981
[2,] 739 0.5179819 0.2934381 0.7539769 0.008854626

```

```
[3,] 1252 0.4829787 0.2696645 0.8417871 0.018199761
[4,] 761 0.5460047 0.3039854 0.8075162 0.005319981

resum dels elements de cada clúster
res.pam <- dades.sample5000 [,c(1,3,5,7,14,12,11)] %>%
 mutate(cluster = fit.pam$clustering) %>%
 group_by(cluster) %>%
 do (the_summary = summary(.))
res.pam$the_summary

[[1]]
retard horaT diastna iIdLinea ordre.norm
Min. :-348.00 9 : 165 lu:274 20 :1977 Min. :0.01
333
1st Qu.: -27.00 7 : 160 ma:285 11 : 74 1st Qu.:0.21
212
Median : 41.00 11 : 153 mi:389 30 : 62 Median :0.40
000
Mean : 75.71 14 : 153 ju:418 50 : 62 Mean :0.44
248
3rd Qu.: 150.00 8 : 150 vi:367 31 : 24 3rd Qu.:0.66
667
Max. : 562.00 17 : 140 sá:337 21 : 12 Max. :1.00
000
##
(Other):1327 do:178 (Other): 37
prec iIdConductor cluster
Min. : 0.0000 49 : 200 Min. :1
1st Qu.: 0.0000 65 : 195 1st Qu.:1
Median : 0.0000 88 : 190 Median :1
Mean : 0.8361 27 : 178 Mean :1
3rd Qu.: 0.0000 94 : 169 3rd Qu.:1
Max. :25.7000 85 : 113 Max. :1
##
(Other):1203
##
[[2]]
retard horaT diastna iIdLinea ordre.norm
Min. :-329.0 14 : 70 lu:223 60 :388 Min. :0.01786
1st Qu.: -11.0 15 : 70 ma:250 11 :124 1st Qu.:0.25357
Median : 68.0 10 : 67 mi:121 50 : 68 Median :0.48571
Mean : 95.2 17 : 66 ju: 81 30 : 64 Mean :0.49220
3rd Qu.: 190.0 16 : 61 vi: 64 31 : 25 3rd Qu.:0.72638
Max. : 560.0 12 : 58 sá: 0 33 : 22 Max. :1.00000
##
(Other):347 do: 0 (Other): 48
prec iIdConductor cluster
Min. : 0.0000 54 :132 Min. :2
1st Qu.: 0.0000 90 : 78 1st Qu.:2
Median : 0.0000 67 : 60 Median :2
Mean : 0.6372 10 : 57 Mean :2
3rd Qu.: 0.0000 48 : 41 3rd Qu.:2
Max. :25.3000 50 : 38 Max. :2
```

```

(Other):333
##
[[3]]
retard horaT diastna iIdLinea ordre.norm
Min. :-336.0 13 :122 lu:151 10 :1061 Min. :0.0222
2
1st Qu.: 17.0 10 :109 ma:178 11 : 69 1st Qu.:0.3111
1
Median : 104.0 11 :108 mi:241 50 : 26 Median :0.5555
6
Mean : 132.6 9 :103 ju:240 30 : 23 Mean :0.5391
1
3rd Qu.: 234.2 16 : 94 vi:198 31 : 22 3rd Qu.:0.7647
1
Max. : 562.0 8 : 91 sá:176 21 : 20 Max. :1.0000
0
(Other):625 do: 68 (Other): 31
prec iIdConductor cluster
Min. : 0.0000 26 :153 Min. :3
1st Qu.: 0.0000 12 :111 1st Qu.:3
Median : 0.0000 93 :105 Median :3
Mean : 0.7276 95 : 78 Mean :3
3rd Qu.: 0.0000 91 : 77 3rd Qu.:3
Max. :25.7000 45 : 74 Max. :3
(Other):654
##
[[4]]
retard horaT diastna iIdLinea ordre.norm
Min. :-351.00 18 : 79 lu: 63 21 :461 Min. :0.0178
6
1st Qu.: -46.00 16 : 71 ma: 65 50 : 83 1st Qu.:0.3035
7
Median : 33.00 20 : 67 mi:110 30 : 77 Median :0.5714
3
Mean : 58.76 14 : 59 ju: 82 11 : 57 Mean :0.5412
1
3rd Qu.: 148.00 19 : 51 vi:146 33 : 28 3rd Qu.:0.7714
3
Max. : 552.00 8 : 50 sá:130 20 : 18 Max. :1.0000
0
(Other):384 do:165 (Other): 37
prec iIdConductor cluster
Min. : 0.000 32 :171 Min. :4
1st Qu.: 0.000 73 : 44 1st Qu.:4
Median : 0.000 50 : 40 Median :4
Mean : 1.918 88 : 33 Mean :4
3rd Qu.: 0.000 63 : 32 3rd Qu.:4
Max. :25.700 94 : 32 Max. :4
(Other):409

```

```
visualització en un espai 2D
tsne_obj <- Rtsne(gower_dist, is_distance=TRUE)

tsne_data <- tsne_obj$Y %>%
 data.frame() %>%
 setNames(c("X", "Y")) %>%
 mutate(cluster = factor(fit.pam$clustering))

ggplot(aes(x=X, y=Y), data=tsne_data) +
 geom_point(aes(color = cluster))
```



Hem trobat que el número òptim de clústers és  $k=4$ , i a la visualització en 2D veiem que estan molt ben diferenciats, excepte el clúster número 1, que està més repartit a l'espai.

### 5.3.6. Correlació: hi ha relació entre els retards i els dies de pluja?

Hem comprovat prèviament que no es compleix el criteri de normalitat i homocedasticitat, de manera que caldrà aplicar la correlació de Spearman, que té com a únic requeriment sobre les dades que es mesurin en una escala ordinal. Fem servir el conjunt de dades sense normalitzar i sense valors extrems, per evitar que interfereixin en els coeficients.

```
```{r correlació, eval=FALSE, include=FALSE}
```

```
cor.test(dades.w.sense.out$retard, dades.w.sense.out$prec, method = "spearman")
```

```
> cor.test(dades.w.sense.out$retard, dades.w.sense.out$prec, method = "spearman")
```

```
spearman's rank correlation rho
```

```
data: dades.w.sense.out$retard and dades.w.sense.out$prec
S = 3.7987e+15, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.1038329
```

```
cor.test(dades.w.sense.out$retard, dades.w.sense.out$ordre.norm, method = "spearman")
```

```
> cor.test(dades.w.sense.out$retard, dades.w.sense.out$ordre.norm, method = "spearman")
```

Spearman's rank correlation rho

```
data: dades.w.sense.out$retard and dades.w.sense.out$ordre.norm
S = 2.8271e+15, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1784957
```

```
corrplot::corrplot.mixed(cor(dades.w.sense.out[,c(1,12,14)], method="spearman"))
```



...

Obtenim com a resultat en els dos casos que es rebutja la hipòtesis nul·la i per tant el coeficient de correlació és significativament diferent de zero, en els dos casos és una correlació baixa.

Per la variable precipitació, s'estima en -0,1, fet que tot i que sigui el contrari a l'esperat, es pot interpretar com que els valors positius de pluja redueixen el retard en una proporció petita.

Per la variable ordre de parada, a mesura que augmenta el recorregut de la línia augmenta el retard, amb un coeficient de correlació de 0,18.

6. Representació dels resultats a partir de taules i gràfiques.

La representació gràfica s'ha anat introduïnt en cada apartat.

7. Resolució del problema. A partir dels resultats obtinguts

Si bé des de les primeres anàlisis descriptives s'ha intuït una relació del retard amb la línia i el conductor, no s'ha pogut formular aquesta dependència de manera prou senzilla i intuïtiva.

S'ha comprovat també que les distribucions dels retards no segueixen una distribució normal.

La recerca dels factors que més influeixen en els retards ha resultat molt complicada degut a la gran quantitat de variables predictores possibles, i degut a la varietat de tipus.

Per altra banda, es pot concloure que la L20 té en realitat menys retard que les altres línies, fet contrari al suposat prèviament. Reflexionant sobre aquest fet, ens podem adonar que el model plantejat en aquesta pràctica no reflecteix bé els casos en que un autobús és posat en hora saltant-se part del trajecte: per aquest model, a partir de la regulació, aquella expedició esdevé puntual, mentre que per l'usuari que està esperant a la parada, el servei té un retard important perquè un autobús va tant retrassat que ni passa. Aquest retard no es correspon a cap registre i degut a la quantitat de regulacions de la L20 és molt possible que la puntualitat millori artificialment en aquest anàlisi. En futurs anàlisis, es podria introduir aquest fet assignant un retard important a cada parada on no hagi passat el bus.

Els diferents mètodes d'anàlisi aplicats confirmen la importància de gran quantitat de variables, no essent possible trobar-ne un conjunt petit que expliqui en mesura suficient la variable retard. El mètode de clústering sí que ha separat les dades en 4 grups ben diferenciats, però degut a la tipologia de les dades resulta difícil interpretar el resultat i aplicar-lo a la pràctica.

Finalment, hem comprovat que la pluja afavoreix lleugerament la puntualitat. No hem introduït les dades dels sensors de trànsit, donat que no hem pogut determinar moments del dia o de la setmana amb pitjor comportament.

Com a conclusió podem dir que no s'ha resolt el problema plantejat, però en canvi s'ha mostrat el camí per a futurs anàlisis.

Com a material complementari, s'han consultat entre d'altres, els següents recursos:

<http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/#graph-of-individuals>

<https://cran.r-project.org/web/packages/PCAmixdata/PCAmixdata.pdf>

<https://bookdown.org/egarpor/PM-UC3M/npreg-multmix.html>

<https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3>