

# Tipologia i cicle de vida de les dades

## Pràctica 2

---

Mireia Calzada i Noemi Lorente

## Pràctica 2

### Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de fins a 3 persones, o si preferiu, també podeu fer-ho de manera individual. Haureu de lliurar un sol fitxer amb l'enllaç Github (<https://github.com>) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu utilitzar aquests exemples com guia:

- Exemple: <https://github.com/Bengis/nba-gap-cleaning>
- Exemple complex (fitxer adjunt).

### Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

### Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

### Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (<https://www.kaggle.com>).

Alguns exemples de dataset amb els que podeu treballar són:

- Red Wine Quality (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al2009> ).
- Titanic: Machine Learning from Disaster (<https://www.kaggle.com/c/titanic> ).
- Predict Future Sales (<https://www.kaggle.com/c/competitive-data-sciencepredict-future-sales/>).

Els últims dos exemples corresponen a competicions actives a Kaggle de manera que, opcionalment, podrieu aprofitar el treball realitzat durant la pràctica per entrar en alguna d'aquestes competicions.

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

1. **Descripció del dataset.** Perquè és important i quina pregunta/problema pretén respondre?

El dataset forma part de la competició '[Predict Future Sales](#)' de la plataforma Kaggle. Aquesta competició serveix de projecte final del curs online "[How to win a data science competition](#)", gestionat per Coursera, que permet aplicar i millorar les habilitats i competències d'un científic de dades.

El dataset conté l'històric de les dades de les vendes diàries de l'empresa [1C Company](#), una de les companyies de programari russes més grans.

Abans de decidir quins camps utilitzarem i com els disposarem per a l'anàlisi, comencem fent un estudi de les dades de què disposem.

### Descripció dels fitxers

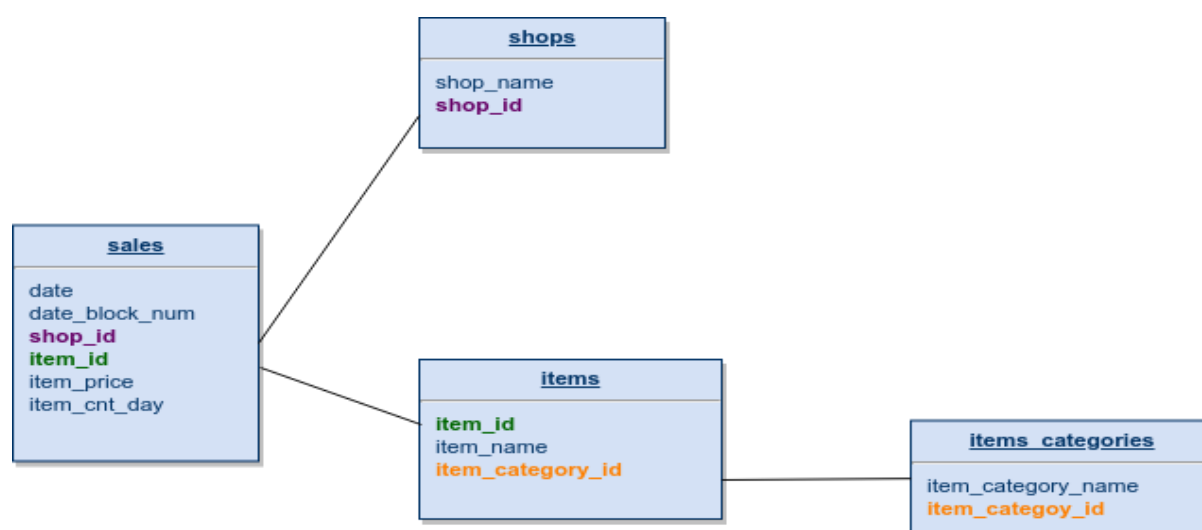
- **items.csv:** informació descriptiva sobre els articles / productes.
- **item\_categories.csv:** informació addicional sobre les categories dels productes.
- **shops.csv:** informació descriptiva sobre les botigues.
- **sales\_train.csv:** És el dataset d'entrenament i conté les dades històriques diàries des de gener de 2013 a octubre de 2015.
- **test.csv:** És el dataset de prova. Cal predir les vendes d'aquestes botigues i productes per a novembre de 2015.
- **sample\_submission.csv:** un fitxer de mostra de la presentació de la predicció per producte.

### Descripció dels camps de dades

Camp	Descripció
ID	Identificador que representa una tupla (un producte concret d'una botiga determinada) dins del conjunt de proves
shop_id	Identificador únic d'una botiga
shop_name	Nom de la botiga
item_id	Identificador únic d'un producte
item_name	Nom del producte
item_price	Preu de venda d'un producte
item_category_id	Identificador únic de la categoria d'un producte

item_category_name	Nom de la categoria del producte
item_cnt_day	Quantitat de productes venuts. L'objectiu és predir una quantitat mensual d'aquesta mesura
data	Data de la venda, en format dd / mm / aaaa
data_block_num	Número consecutiu que identifica els mesos. No s'inicialitza amb el canvi d'any. Per exemple, gener de 2013 és 0, febrer de 2013 és 1, ..., l'octubre de 2015 és 33.

### Model E-R de les vendes per botiga



### shops.csv

Comprovem que es tracta d'un dataset format per 60 registres amb 2 columnes:

- shop\_id: identificador enter per la botiga
- shop\_name: cadena de text amb el nom de la botiga.

```

>
> str(shops)
'data.frame': 60 obs. of 2 variables:
 $ shop_name: Factor w/ 60 levels "Адигея ТЦ \"Мера\"",...: 57 59 1 2 3 4 5 6 7 8 ...
 $ shop_id : int 0 1 2 3 4 5 6 7 8 9 ...
>
  
```

Es tracta d'un conjunt de dades referents a vendes en botigues russes, de manera que tant els noms de les botigues com dels diferents productes estan en aquest idioma, la qual cosa ens dificulta la cerca de possibles registres duplicats.

De manera que traduïm els noms de les botigues a l'anglès i comprovem si hi ha dades repetides:

```
>
> head(shops)
      shop_name shop_id
1 !Якутск Орджоникидзе, 56 фран 0
2 !Якутск ТЦ "Центральный" фран 1
3      Адыгея ТЦ "Мега"        2
4 Балашиха ТРК "Октябрь-Киномир" 3
5      Волжский ТЦ "Волга Молл"  4
6 Вологда ТРЦ "Мармелад"        5
>
```

	shop_name	shop_id
9	Voronezh SEC-City Park "Grad"	8
10	Comerç sortint	9
11	Zhukovsky Str. Chkalov 39m?	10
12	Zhukovsky Str. Chkalov 39 m²	11
13	Botiga en línia Emergència	12
14	Centre comercial de Kazan "Behetle"	13
15	Centre comercial de Kazan "ParkHouse" II	14
16	Kaluga SEC "Segle XXI"	15

Observem que possiblement les botigues amb identificadors 10 i 11 facin referència al mateix establiment, doncs la descripció és pràcticament igual.

Ho tindrem en compte a l'hora de netejar les dades per tal d'agrupar les vendes diferents d'aquestes dos botigues.

### items\_category.csv

Abans d'estudiar aquest fitxer de dades, traduïm els noms de les categories.

```
>
> str(item_categories_translated)
'data.frame':  84 obs. of  2 variables:
 $ PC...Auriculars...auriculars: Factor w/ 84 levels "Accessoris - PS2",...: 1 2 3 4 5 6 7 19 47 11 ...
 $ item_category_id              : int  0 1 2 3 4 5 6 7 8 9 ...
>
```

```
>
> head(item_categories_translated)
  item_category_name item_category_id
1 PC - Auriculars / auriculars      0
2      Accessoris - PS2             1
3      Accessoris - PS3             2
4      Accessoris - PS4             3
5      Accessoris - PSP             4
6      Accessoris - PSVita          5
>
```

Comprovem que es tracta d'un fitxer amb 84 registres, cadascun dels quals està format per 2 atributs:

- item\_category\_id: enter identificador de la categoria del producte
- item\_category\_name: cadena de text la descripció de la categoria.

### items.csv

```
>
> str(items)
'data.frame': 22170 obs. of 3 variables:
 $ item_name      : Factor w/ 22170 levels "007 Legends [PS3, русская версия]",...: 9929 1122 9909 10558 132
53 16234 20048 20047 21590 21984 ...
 $ item_id        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ item_category_id: int  40 76 40 40 40 40 40 40 40 40 ...
>
```

Comprovem en aquest cas que es tracta d'un fitxer amb 22170 registres formats per 3 variables que fan referència a:

- item\_category\_id: enter identificador de la categoria del producte
- item\_id: enter identificador del propi producte
- item\_name: cadena de text amb la descripció del producte

### sales\_train\_v2.csv

Analitzem el fitxer de les vendes que farem servir per a la predicció.

```
>
> str(sales_train_v2)
'data.frame': 2935849 obs. of 6 variables:
 $ date          : Factor w/ 1034 levels "01.01.2013", "01.01.2014",...: 35 69 137 171 477 307 35 103 341 69 .
..
 $ date_block_num: int  0 0 0 0 0 0 0 0 0 0 ...
 $ shop_id       : int  59 25 25 25 25 25 25 25 25 25 ...
 $ item_id       : int  22154 2552 2552 2554 2555 2564 2565 2572 2572 2573 ...
 $ item_price    : num  999 899 899 1709 1099 ...
 $ item_cnt_day  : num  1 1 -1 1 1 1 1 1 1 3 ...
>
```

En aquest cas tenim un conjunt de gairebé 3 milions de registres de 6 variables cadascun. Observem que conté dades de vendes diàries en les diferents botigues (`shop_id`) i per als diferents productes (`item_id`). Inclou a més, la data de la venda (`date`), el preu (`item_price`) i el recompte d'unitats venudes (`item_cnt_day`) des de l'1 de gener del 2013 al 31 d'octubre de 2015.

El camp "`data_block_num`" és una seqüència, començant pel zero fins al 33, que agrupa les vendes d'un mateix mes i any. 34 identificadors per als 12 mesos de l'any 2013 + 12 mesos de l'any 2014 + 10 primers mesos de l'any 2015.

Una observació interessant és que tenim quantitats negatives en la columna '`item_cnt_day`'. Considerem que aquestes quantitats negatives són degudes a devolucions, de manera que les tindrem en compte a l'hora de predir les vendes.

Cal tenir en compte que la llista de botigues i productes varia lleugerament cada mes i gestionar aquestes situacions forma part del repte.

Un cop analitzats aquests fitxers, ens plantegem els següents **objectius**:

- Volem analitzar quines variables influeixen més en el preu dels productes.
- Analitzarem si hi ha diferències estadístiques entre els diferents mesos, botigues,...
- Crearem models de regressió que permetin predir el preu dels productes

Aquests tipus d'anàlisis poden ser emprats com a estratègia empresarial d'una cadena de supermercats a l'hora de:

- Millorar la gestió del capital humà per donar resposta a l'increment de vendes i donar millor servei al client.
- Detectar en quin moment de l'any es produeixen menys vendes per incentivar-les, per exemple creant promocions.
- Calcular la demanda dels productes, quina estacionalitat tenen, per a poder anticipar-nos en les comandes per proveir les botigues i no trencar estocs, és a dir, predir quan hem de fer una comanda.

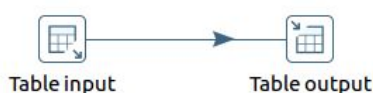
## 2. Integració i selecció de les dades d'interès a analitzar.

El conjunt de dades que analitzarem és sales\_train\_v2.csv i les variables que tindrem en compte són:

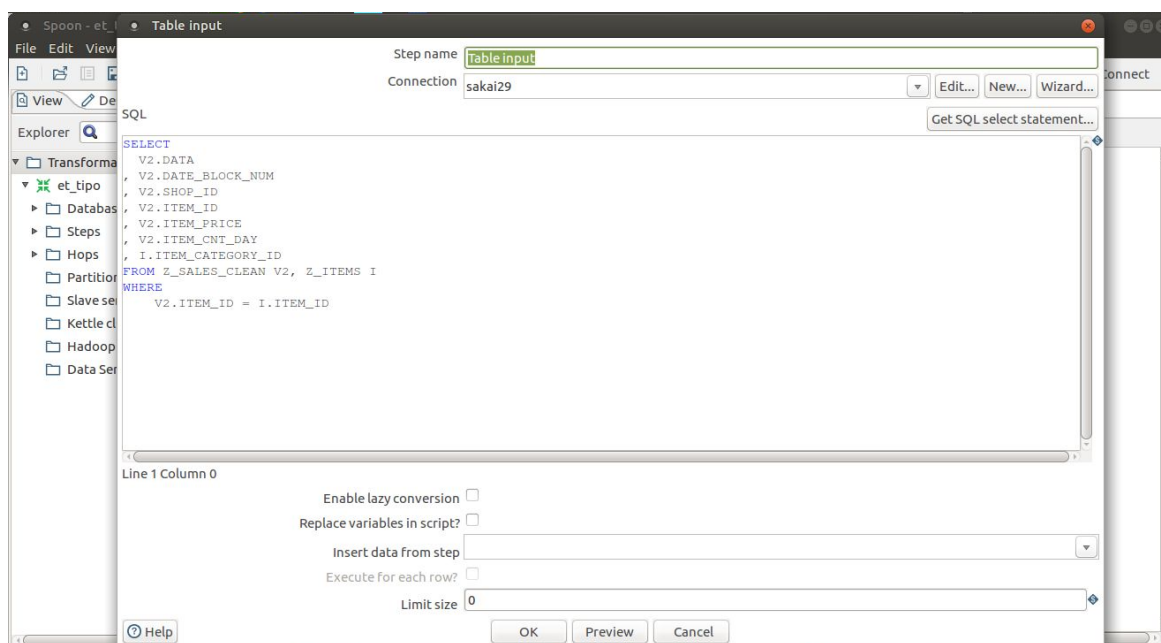
- date
- date\_block\_num
- shop\_id
- item\_id
- item\_price
- item\_cnt\_day

També agafarem la categoria del producte, que extraurem del fitxer ITEMS.csv

- Afegim el item\_category\_id a la taula sales mitjançant una senzilla transformació amb el Pentaho Spoon:



On fem un select de les vendes cercant el item\_category\_id per cada ítem en la taula de ítems i inserim els resultats en la taula Z\_SALES\_DEF\_V2, que serà la taula a partir de la qual generarem els datasets.



En el moment de generar els datasets per a l'anàlisi, afegirem les sumes agregades de preus i quantitats i el total de vendes calculat.



### 3. Neteja de les dades.

Un cop integrades les dades, llegim el nou fitxer generat `dataset_total.csv` amb la funció `read.csv`. El resultat és un objecte `data.frame` i el guardem en la variable `vendes`.

```
> vendes <- read.csv("dataset_total.csv", header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE)
```

També comprovem amb la funció `class()` que el tipus de dades assignat per R es correspon al domini correcte de dades.

```
> sapply(vendes, function(x) class(x))
```

```
DATE_BLOCK_NUM      SHOP_ID      ITEM_ID      ITEM_PRICE      ITEM_CNT_DAY
ITEM_CATEGORY_ID
"integer"    "integer"    "integer"    "numeric"    "numeric"    "integer"
```

Amb la funció `describe` del package `Hmisc` podem analitzar el nombre de variables del `data.frame`, quants registres té, el nombre de valors diferents de cada una de les variables, la mitjana, alguns dels percentils més interessants i la mesura de dispersió Gmd, diferència mitjana de Gini, entre altres.

```
> describe(vendes)
```

vendes

6 Variables 2935849 Observations

DATE\_BLOCK\_NUM

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2935849	0	34	0.999	14.57	10.84	1	2	7	14	23	28	31

lowest : 0 1 2 3 4, highest: 29 30 31 32 33

SHOP\_ID

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2935849	0	59	0.999	33	18.59	6	10	22	31	47	56	57

lowest : 0 1 2 3 4, highest: 55 56 57 58 59

ITEM\_ID

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2935849	0	21807	1	10197	7257	1540	2416	4476	9343	15684	19436	20949

lowest : 0 1 2 3 4, highest: 22165 22166 22167 22168 22169

ITEM\_PRICE

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2935849	0	19993	0.998	890.9	1025	99	149	249	399	999	1999	2690

Value	0	5000	10000	15000	20000	25000	30000	35000	40000	45000	50000	60000	310000
Frequency	2734703	175337	9436	7437	3879	3970	972	99	8	4	2	1	1
Proportion	0.931	0.060	0.003	0.003	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000

## ITEM\_CNT\_DAY

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2935849	0	198	0.281	1.243	0.4806	1	1	1	1	1	2	2

lowest : -22 -16 -9 -6 -5, highest: 624 637 669 1000 2169

## ITEM\_CATEGORY\_ID

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
2935849	0	84	0.989	40	19.19	19	19	28	40	55	65	71

lowest : 0 1 2 3 4, highest: 79 80 81 82 83

### a. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Revisant la informació *missing* de la funció *describe* podem comprobar que el dataset no conté elements buits. Encara que ho tornem a verificar amb la funció *is.na()* en cada una de les variables de *vendes*.

```
> sapply(vendes, function(x) sum(is.na(x)))
```

DATE_BLOCK_NUM	SHOP_ID	ITEM_ID	ITEM_PRICE	ITEM_CNT_DAY	ITEM_CATEGORY_ID
0	0	0	0	0	0

Les dades contenen zeros en les variables següents:

- *date\_block\_num* és igual a zero per identificar el més de gener de 2013
- *shop\_id* és igual zero per identificar una botiga concreta, 'Yakutsk Ordzhonikidze, 56 fr'
- *item\_id* és igual zero per identificar un article en concret, 'EN EL PODER DE LA FELICITAT (PLAST)

Aquests zeros els considerem valors correctes en les 3 variables.

Fem un subset de les dades per comprovar que les variables *item\_price* i *item\_cnt\_day* no contenen zeros.

```
> subset(vendes, ITEM_PRICE==0)
```

DATE_BLOCK_NUM	SHOP_ID	ITEM_ID	ITEM_PRICE	ITEM_CNT_DAY
[1]				

<0 rows> (or 0-length row.names)

```
> subset(vendes, ITEM_CNT_DAY==0)
```

DATE_BLOCK_NUM	SHOP_ID	ITEM_ID	ITEM_PRICE	ITEM_CNT_DAY
[1]				

<0 rows> (or 0-length row.names)

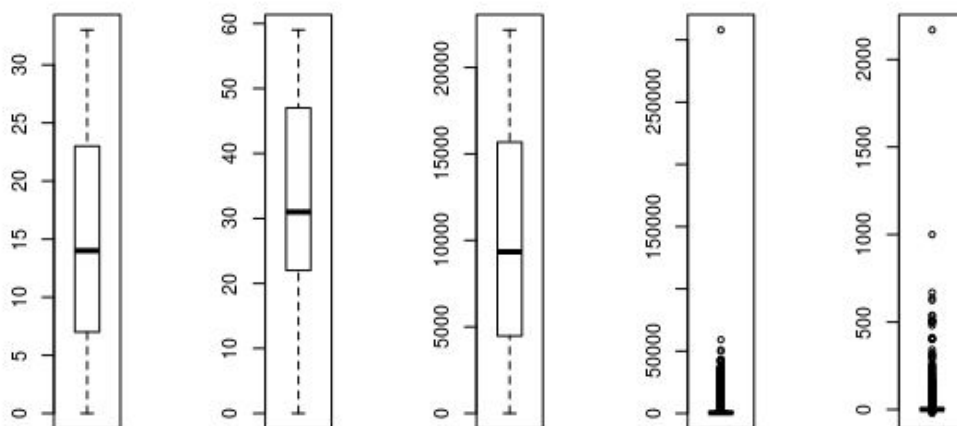
### b. Identificació i tractament de valors extrems.

Els valors extrems o outliers són valors numèricament distants de la resta de dades, fora de la distribució normal de les dades d'una variable. Per tant, per identificar els outliers podem emprar les alternatives següents:

- representar en un diagrama de caixa cada variable per separat i veure quines variables sobresurten del rang interquartílic
- visualitzar la distribució de les dades en intervals amb la funció *binnedCounts*
- calcular els outliers amb la funció *boxplot.stats*

Amb les gràfiques boxplot, observem que les dades presenten outliers.

```
# comprovar outliers
par(mfrow=c(1,5))
boxplot(vendes$DATE_BLOCK_NUM)
boxplot(vendes$SHOP_ID)
boxplot(vendes$ITEM_ID)
boxplot(vendes$ITEM_PRICE)
boxplot(vendes$ITEM_CNT_DAY)
```



La variable *item\_price* té 19993 valors diferents, la mitjana són 890,9 i amb la funció *binnedCounts* veiem la distribució de les dades per intervals:

```
> binnedCounts(vendes[, "item_price", drop=FALSE])
distribution of item_price
(-10000, 0]      (0, 10000]  (10000, 20000]  (20000, 30000]  (30000, 40000]  (40000, 50000]
1              2916103      12964      6541              226              11
(50000, 60000]  ..... (290000, 300000] (300000, 310000]
2              .....      0              1
```

A continuació fem la comprovació dels valors extrems amb la funció `boxplot.stats`, i amb el valor *out* de la funció podem comprovar que hi ha 258942 registres amb valors extrems en *item\_price*. Per tant, dels 2935849 registres inicials ens quedaríem amb 2676907.

```
> outliersItemPrice <-boxplot.stats(item_price)$out
> outliersItemPrice
> table(outliersItemPrice)
> indexItemPrice <- which( item_price %in% outliersItemPrice)
> length(indexItemPrice)
[1] 258942

> vendes<-vendes[-indexItemPrice,]

> dim(vendes)
[1] 2676907      6
```

Després d'eliminar els outliers detectats amb la funció `boxplot.stats`, veiem que només tenim productes amb *item\_price*  $\leq 2124$

```
> subset(vendes, item_price>2124)
[1] date          date_block_num shop_id          item_id          item_price
item_cnt_day
<0 rows> (or 0-length row.names)
```

La variable *item\_cnt\_day* té 198 valors diferents, la mitjana són 1,243 i amb la funció *binnedCounts* veiem la distribució de les dades per intervals:

```
> binnedCounts(vendes[, "item_cnt_day", drop=FALSE])
distribution of item_cnt_day
(-100, 0]      (0, 100] (100, 200] (200, 300] (300, 400] (400, 500] (500, 600] (600, 700]
  7356 2928355   100         15      4             7             7             3
(700, 800] (800, 900] (900, 1000] .... (2100, 2200]
          0          0           1           1
```

De la mateixa manera que hem fet amb la variable *item\_price*, comprovem ara els outliers de *item\_cnt\_day* amb el valor *out* de la funció `boxplot.stats`, i comprovem que hi ha 306477 registres amb valors extrems en *item\_cnt\_day*.

```
> outliersItemCnt <-boxplot.stats(item_cnt_day)$out
> outliersItemCnt
> indexItemCnt <- which( item_cnt_day %in% outliersItemCnt)
> length(indexItemCnt)
[1] 306477

> vendes<-vendes[-indexItemCnt,]

> dim(vendes)
[1] 2629372      6
```

Després d'eliminar els outliers detectats amb la funció `boxplot.stats`, veiem que només tenim productes amb `item_cnt_day <= 1`

```
> subset(vendes, item_cnt_day>1)
[1] date    date_block_num  shop_id  item_id  item_price  item_cnt_day
<0 rows> (or 0-length row.names)
```

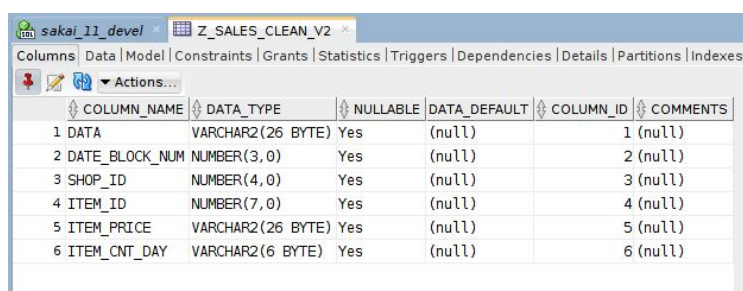
Per tant, després d'extreure els outliers tenim 2629372 registres amb `item_price <= 2124` i `item_cnt_day <= 1`

### c. Preprocessat de les dades.

Tal i com hem conclòs en l'apartat 2, processarem les dades per tal de:

- Fusionar les vendes de les botigues 10 i 11
- Convertim el `date_block_num` en una data referent al dia 1 del mes i any que identifica.

Importem el fitxer `sales_train_v2.csv` a una taula Oracle, ja que per tractar la quantitat de registres que conté ens resulta més pràctic.



COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1 DATA	VARCHAR2(26 BYTE)	Yes	(null)	1 (null)	
2 DATE_BLOCK_NUM	NUMBER(3, 0)	Yes	(null)	2 (null)	
3 SHOP_ID	NUMBER(4, 0)	Yes	(null)	3 (null)	
4 ITEM_ID	NUMBER(7, 0)	Yes	(null)	4 (null)	
5 ITEM_PRICE	VARCHAR2(26 BYTE)	Yes	(null)	5 (null)	
6 ITEM_CNT_DAY	VARCHAR2(6 BYTE)	Yes	(null)	6 (null)	

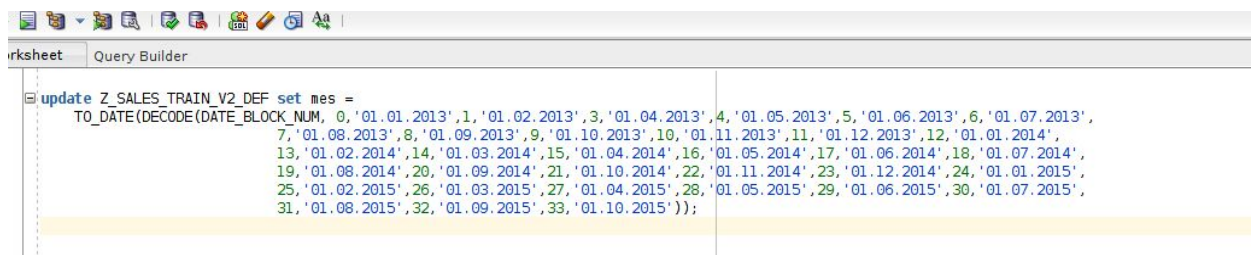
Un cop carregades les dades en una taula Oracle, apliquem les següents mesures per tractar-les:

- Fusionem les vendes de les botigues 10 i 11 mitjançant una sentència SQL:  

```
UPDATE Z_SALES_CLEAN_V2 SET SHOP_ID = 10 WHERE SHOP_ID = 11;
```

```
---
```

```
499 rows updated
```
- Convertim el `date_block_num` en una data referent al dia 1 del mes i any que identifica, mitjançant una consulta SQL que fa un decode de cada `date_block_num`:



Hem provat de formatar la data amb R però es penja. Per això decidim fer-ho amb una transformació ETL de Pentaho.

Finalment exportem la taula resultant en un fitxer csv -> DADES\_TOTALS.csv

#### 4. Anàlisi de les dades.

- a. Selecció dels grups de dades que es volen analitzar/comparar (planificació de les anàlisis a aplicar).

Creem uns datasets específics per als diferents objectius d'anàlisi.

#### **Objectiu 1: Quines variables influeixen més en el preu dels productes?**

Per a l'anàlisi d'aquest fem servir tot el conjunt de dades:

```
> vendes <- read.csv("dades_totals.csv", header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE)
```

```
> attach(vendes)
```

```
> str(vendes)
```

```
'data.frame': 2935849 obs. of 6 variables:
```

```
$ DATE_BLOCK_NUM : int 0 0 0 0 0 0 0 0 0 ...
```

```
$ SHOP_ID : int 25 25 25 25 25 25 25 25 25 ...
```

```
$ ITEM_ID : int 785 785 791 791 791 791 791 791 804 810 ...
```

```
$ ITEM_PRICE : num 349 349 600 600 600 600 600 600 240 199 ...
```

```
$ ITEM_CNT_DAY : num 1 1 2 1 2 1 1 1 1 1 ...
```

```
$ ITEM_CATEGORY_ID: int 49 49 73 73 73 73 73 73 49 77 ...
```

```
> dim(vendes)
```

```
[1] 2935849 6
```

## **Objectiu 2: La facturació és superior durant el segon semestre de l'any?**

Seleccionem d'una banda les files del fitxer DADES\_TOTALS.csv que contenen les vendes en els mesos del primer semestre, i d'altra banda les dades que contenen les vendes corresponents al segon semestre.

Un cop seleccionades retallem item\_price a 2 decimals i hi afegim columnes agregades per al sumatori de preus i quantitats així com una columna que indica les vendes totals (quantitat \* preu).

```
> vendes_1semestre <- read.csv("dataset_primer_semestre.csv", header=TRUE, sep=";", na.strings="NA",  
dec=".", strip.white=TRUE)
```

```
> vendes_2semestre <- read.csv("dataset_segond_semestre.csv", header=TRUE, sep=";", na.strings="NA",  
dec=".", strip.white=TRUE)
```

```
> str(vendes_1semestre)
```

```
'data.frame': 125763 obs. of 6 variables:  
 $ DATE_BLOCK_NUM: int 0 0 0 0 0 0 0 0 0 ...  
 $ MES           : Factor w/ 18 levels "01/01/13","01/01/14",...: 1 1 1 1 1 1 1 1 1 ...  
 $ ITEM_ID       : int 3140 2252 2222 3175 3438 1119 2058 1904 2522 481 ...  
 $ SUM_PRICE     : num 62604 93758 198 2990 43167 ...  
 $ SUM_COUNT     : int 65 177 1 10 12 6 2 56 2 26 ...  
 $ TOTAL        : num 4069260 16595193 198 29900 518004 ...
```

```
> str(vendes_2semestre)
```

```
'data.frame': 108149 obs. of 6 variables:  
 $ DATE_BLOCK_NUM: int 6 6 6 6 6 6 6 6 6 ...  
 $ MES           : Factor w/ 16 levels "01/07/13","01/07/14",...: 1 1 1 1 1 1 1 1 1 ...  
 $ ITEM_ID       : int 21996 19408 20139 10469 10733 2417 13370 12128 4163 3705 ...  
 $ SUM_PRICE     : num 449 27033.05 0.14 4389 8224.18 ...  
 $ SUM_COUNT     : int 1 14 20 11 31 69 14 45 227 69 ...  
 $ TOTAL        : num 449 378462.7 2.8 48279 254949.6 ...
```

### **b. Comprovació de la normalitat i homogeneïtat de la variància.**

Per a determinar si una variable segueix una distribució normal, emprarem el test de *Kolmogorov-Smirnov*. Si  $p \geq 0.05$  podem deduir que es tracta d'una distribució normal.

La comprovació de la normalitat ens permetrà decidir si emprar tests paramètrics per a distribucions normal o bé tests no paramètrics per a distribucions no normals.

Observem que en aplicar el test Kolmogorov-Smirnov el valor p-value és menor que 0.05, tant en la variable *item\_cnt\_day* com *item\_price*, per tant, cap de les dues variables segueix una distribució normal.

```
# estimar els paràmetres de la distribució normal a partir de la funció fitdistr del paquet MASS, en la
variable item_price
> require(MASS)
> ajust <- fitdistr(item_price,"normal")
> ajust
      mean      sd
564.2630146 471.6667710
( 0.2882827) ( 0.2038467)
```

```
> #test Kolmogorov-Smirnov per comprovar la normalitat. Si p<0.05
> Ks<- ks.test(item_price, "pnorm", mean =ajust$estimate[1], sd= ajust$estimate[2])
```

```
> Ks
```

One-sample Kolmogorov-Smirnov test

```
data: ITEM_PRICE
D = 0.20041, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
> # estimar els paràmetres de la distribució normal a partir de la funció fitdistr del paquet MASS, en la
variable item_cnt_day
> require(MASS)
> ajust <- fitdistr(item_cnt_day,"normal")
> ajust
      mean      sd
1.225364198 2.596753427
(0.001587136) (0.001122274)
```

```
> #test Kolmogorov-Smirnov per comprovar la normalitat. Si p<0.05
> Ks<- ks.test(item_cnt_day, "pnorm", mean =ajust$estimate[1], sd= ajust$estimate[2])
```

```
> Ks
```

One-sample Kolmogorov-Smirnov test

```
data: ITEM_CNT_DAY
D = 0.46319, p-value < 2.2e-16
alternative hypothesis: two-sided
```

No és necessari estudiar l'homogeneïtat de la variància perquè les dades no segueixen una distribució normal.



- c. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

### **Objectiu 1: Quines variables influeixen més en el preu dels productes?**

Realitzarem una anàlisi **de correlació** entre les diferents variables quantitatives per a determinar quines variables tenen major influència sobre el preu dels productes.

Com les dades no segueixen una distribució normal, utilitzem el coeficient de correlació de Spearman.

El coeficient rho de Spearman va de -1 a +1, on els extrems indiquen correlació perfecta i 0 significa no correlació. El signe és negatiu quan els valors grans d'una variable estan associats amb valors petits de l'altra variable i positius si ambdues variables solen ser grans o petites simultàniament.

Es desaconsella aplicar el test de correlació de Spearman en variables qualitatives, com SHOP\_ID, ITEM\_CATEGORY\_ID, ITEM\_ID, ja que no tenen escala ordinal (els seus valors no es poden ordenar) ni són dicotòmiques (només dos valors, 0 i 1).

La variable DATE\_BLOCK\_NUM sí que la considerem ordinal, ja que mostra una evolució en el temps.

Per tant, apliquem el test entre les variables ITEM\_PRICE i ITEM\_CNT\_DAY i ITEM\_PRICE i DATE\_BLOCK\_NUM.

Podem concloure que, com la variable *DATE\_BLOCK\_NUM* és la que té el coeficient *rho* més alt (0.1371966), és el mes de la venda la variable que més influeix en el preu. Tot i que considerem que es tracta d'un grau de correlació baix.

```
> cor.test(x = vendes$ITEM_PRICE, y = vendes$ITEM_CNT_DAY,  
+         alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Spearman's rank correlation rho

```
data: vendes$ITEM_PRICE and vendes$ITEM_CNT_DAY  
S = 4.024e+18, p-value < 2.2e-16  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
rho  
0.04586005
```

```
> cor.test(x = vendes$ITEM_PRICE, y = vendes$DATE_BLOCK_NUM,  
+         alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Spearman's rank correlation rho

```
data: vendes$ITEM_PRICE and vendes$DATE_BLOCK_NUM  
S = 3.6388e+18, p-value < 2.2e-16  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
rho  
0.1371966
```

Com el nombre de variables quantitatives són poques hem volgut comprovar també la correlació entre *ITEM\_CNT\_DAY* i *DATE\_BLOCK\_NUM*, però observem que el coeficient rho de Spearman és inferior a l'observat entre *ITEM\_PRICE* i *DATE\_BLOCK\_NUM*, per tant, el grau de correlació encara és menor.

```
> cor.test(x = vendes$ITEM_CNT_DAY, y = vendes$DATE_BLOCK_NUM,  
+         alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Spearman's rank correlation rho

```
data: vendes$ITEM_CNT_DAY and vendes$DATE_BLOCK_NUM  
S = 4.203e+18, p-value = 0.000000004857  
alternative hypothesis: true rho is not equal to 0  
sample estimates:  
rho  
0.003415363
```

Provem de tornar a aplicar el test Spearman havent eliminat els outliers de *ITEM\_PRICE*. Observem que no millora l'índex *rho* de Spearman en cap de les correlacions provades.

```
> # utilitzem boxplot.stats per a veure els valors outliers de ITEM_PRICE  
> outliersCnt <- boxplot.stats(vendes$ITEM_PRICE)$out  
> indexCnt <- which( vendes$ITEM_PRICE %in% outliersCnt)  
> length(indexCnt)  
[1] 258942  
  
> # eliminem els registres que contenen outliers de ITEM_PRICE  
> vendes<-vendes[-indexCnt,]  
> dim(vendes)  
[1] 2676907      6
```

```
> cor.test(x = vendes$ITEM_PRICE, y = vendes$ITEM_CNT_DAY,  
+         alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Spearman's rank correlation rho

data: vendes\$ITEM\_PRICE and vendes\$ITEM\_CNT\_DAY

S = 3.121e+18, p-value < 2.2e-16

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

**0.02377715**

```
> cor.test(x = vendes$ITEM_PRICE, y = vendes$DATE_BLOCK_NUM,  
+         alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Spearman's rank correlation rho

data: vendes\$ITEM\_PRICE and vendes\$DATE\_BLOCK\_NUM

S = 2.8737e+18, p-value < 2.2e-16

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

**0.1011314**

```
> cor.test(x = vendes$ITEM_CNT_DAY, y = vendes$DATE_BLOCK_NUM,  
+         alternative = "two.sided", conf.level = 0.95, method = "spearman")
```

Spearman's rank correlation rho

data: vendes\$ITEM\_CNT\_DAY and vendes\$DATE\_BLOCK\_NUM

S = 3.1966e+18, p-value = 0.8337

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

**0.0001283099**

## **Objectiu 2: La facturació és superior durant el segon semestre de l'any?**

Aquesta prova estadística es basa en un **contrast d'hipòtesi sobre dues mostres** per a determinar si la facturació és superior depenent del semestre de l'any (primer o segon semestre). Utilitzarem les mostres següents per aplicar-ho:

- `dataset_primer_semestre.csv`: conté les vendes del primer semestre
- `dataset_segond_semestre.csv`: conté les vendes del segon semestre

Com les dades no segueixen una distribució normal, emprarem el test no paramètric *Mann–Whitney–Wilcoxon (WMW)*. Aquest test també es coneix com *Wilcoxon rank-sum test* o *u-test*, i contrasta si dues mostres procedeixen de poblacions equidistribuïdes.

Aquest test es basa en la idea que si les dues mostres comparades procedeixen de la mateixa població, a l'ajuntar totes les observacions i ordenar-les de menor a major, caldria esperar que les observacions d'una i l'altra mostra estiguessin intercalades aleatòriament. I en el cas contrari, si una de les mostres pertany a una població amb valors majors o menors que l'altra població, a l'ordenar les observacions, aquestes tendiran a agrupar-se de manera que les d'una mostra quedin per sobre de les de l'altra.

Les hipòtesis que contrasta el test de *Mann–Whitney–Wilcoxon* no es basen en la mitjana sinó en:

- $H_0$ : La probabilitat que una observació de la població  $X$  sigui major que una observació de la població  $Y$  és igual a la probabilitat que una observació de la població  $Y$  sigui major que una observació de la població  $X$ .

$$P(X > Y) = P(Y > X)$$

- $H_a$ : La probabilitat que una observació de la població  $X$  sigui major que una observació de la població  $Y$  no és igual a la probabilitat que una observació de la població  $Y$  sigui major que una observació de la població  $X$ .

$$P(X > Y) \neq P(Y > X)$$

R disposa de la funció `wilcox.test()` que realitza el test entre dues mostres quan s'indica el paràmetre `paired = False` i, a més, genera l'interval de confiança per a la diferència de localització.

```
> wilcox.test(x = vendes_1semestre$TOTAL, y = vendes_2semestre$TOTAL, alternative = "less", mu =
0, paired = FALSE, conf.int = 0.95)
Wilcoxon rank sum test with continuity correction
data: vendes_1semestre$TOTAL and vendes_2semestre$TOTAL
W = 6744100000, p-value = 0.0002609
alternative hypothesis: true location shift is less than 0
95 percent confidence interval:
      -Inf -4.00007
sample estimates:
difference in location
      -38.00001
```

Hem tornat a aplicar el test de *Mann–Whitney–Wilcoxon* eliminant els outliers de la variable TOTAL en els dos datasets, però no observem massa canvis en els resultats llevat que l'interval de confiança és menor.

```
> #eliminem els outliers de vendes_1semestre
> # utilitzem boxplot.stats per a veure els valors outliers de vendes_1semestre$TOTAL
> outliers1sem <- boxplot.stats(vendes_1semestre$TOTAL)$out
> index1sem <- which( vendes_1semestre$TOTAL %in% outliers1sem)
> length(index1sem)
[1] 21673
> vendes_1semestre<-vendes_1semestre[-index1sem,]

> #eliminem els outliers de vendes_2semestre
> # utilitzem boxplot.stats per a veure els valors outliers de vendes_2semestre$TOTAL
> outliers2sem <- boxplot.stats(vendes_2semestre$TOTAL)$out
> index2sem <- which( vendes_2semestre$TOTAL %in% outliers2sem)
> length(index2sem)
[1] 18868

> vendes_2semestre<-vendes_2semestre[-index2sem,]

> wilcox.test(x = vendes_1semestre$TOTAL, y = vendes_2semestre$TOTAL, alternative = "less", mu =
0, paired = FALSE, conf.int = 0.95)
Wilcoxon rank sum test with continuity correction
data: vendes_1semestre$TOTAL and vendes_2semestre$TOTAL
W = 4612800000, p-value = 0.002861
alternative hypothesis: true location shift is less than 0
95 percent confidence interval:
      -Inf -0.000002762126
sample estimates:
difference in location
      -8.99995
```

Donat que p-value és menor que el valor de confiança fixat ( $p\text{-value} < 0.05$ ), podem rebutjar la hipòtesi nul·la i concloure que, efectivament, la facturació durant el segon semestre és major que durant el primer.

Normalment, es recomana utilitzar la prova de *Mann-Whitney-Wilcoxon* en lloc de *t-test* quan les mides de les mostres són petites i no hi ha evidències que les poblacions d'origen segueixin una distribució normal. Si bé aquesta pràctica està bastant fonamentada, cal no confondre-la amb l'ús de la prova de Mann-Whitney-Wilcoxon com a alternativa a *t-test* sempre que no es compleixi la normalitat i sense tenir en compte la mida de la mostra. A mesura que el nombre d'observacions augmenta, també ho fa la robustesa de *t-test* davant els desviaments de la normalitat.

Per tant, tot i que no segueixen distribució normal, com el dataset és gran provem a aplicar el test paramètric *t-test* i també es confirma que la mitjana del segon semestre (13797.85) és més gran que la mitjana del primer semestre (12866.84).

```
> t.test(vendes_1semestre$TOTAL,vendes_2semestre$TOTAL,alternative = "less")
```

```
Welch Two Sample t-test
data: vendes_1semestre$TOTAL and vendes_2semestre$TOTAL
t = -9.0774, df = 182480, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -762.3097
sample estimates:
mean of x mean of y
12866.84 13797.85
```

```
> t.test(vendes_1semestre$TOTAL,vendes_2semestre$TOTAL,alternative = "great")
```

```
Welch Two Sample t-test

data: vendes_1semestre$TOTAL and vendes_2semestre$TOTAL
t = -9.0774, df = 182480, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1099.715      Inf
sample estimates:
mean of x mean of y
12866.84 13797.85
```

**Objectiu 3: Crearem models de regressió que permetin predir el preu dels productes**

Un dels objectius principals de la pràctica és realitzar prediccions sobre el preu dels productes en funció de les característiques relacionades amb la venda. Amb el càlcul d'un **model de regressió** podrem obtenir un model predictiu que utilitzi les variables quantitatives i qualitatives per a poder realitzar les prediccions dels preus dels productes.

Com les dades no segueixen una distribució normal, cal emprar models de regressió no paramètrics. La regressió de *Kendall-Theil* s'adapta a un model lineal entre una variable  $X$  i una variable  $Y$  usant un enfocament totalment no paramètric.

Per a una regressió múltiple per trobar una relació lineal entre una variable dependent i una o més variables independents podem emprar la regressió *Quantile*.

Per aconseguir el model més eficient, hem provat a executar diversos models. La regressió *Kendall-Theil* no disposa del valor *r-squared* per a determinar quin és el millor model. Creiem que interpretant la pendent, l'intercept, el p-value i la variable MAD (mediana de la desviació) és suficient per a determinar el millor model.

Com R no pot executar el model de regressió amb el dataset inicial perquè té molts registres, hem creat una mostra del 0.1% (2936 registres), ja que R es queda penjat amb un percentatge superior de registres.

```
> #funcio sample_frac del package dplyr per crear una mostra  
> require(dplyr)  
> mostra <- sample_frac(vendes, 0.001, replace = FALSE)  
> dim(mostra)  
[1] 2936 6
```

Els models de regressió que hem provat amb Kendall-Theil són els següents:

- `model1 = mb1m(ITEM_PRICE ~ DATE_BLOCK_NUM, data=mostra)`

```
#Kendall–Theil Sen Siegel nonparametric linear regression
> set.seed(1234)
> library(mb1m)
> model1.k <- mb1m(ITEM_PRICE ~ DATE_BLOCK_NUM, data=mostra)
> summary(model1.k)
Call:
mb1m(formula = ITEM_PRICE ~ DATE_BLOCK_NUM, dataframe = mostra)
Residuals:
      Min       1Q   Median       3Q      Max
-538.8  -159.1   33.8  619.5 27603.3
Coefficients:
              Estimate      MAD    V value      Pr(>|V|)
(Intercept)    308.091 225.195 4021128    <2e-16 ***
DATE_BLOCK_NUM    7.143    17.262 2668666    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1877 on 2934 degrees of freedom
```

- `model2 = mb1m(ITEM_PRICE ~ SHOP_ID, data=mostra)`

```
> model2.k <- mb1m(ITEM_PRICE ~ SHOP_ID, data=mostra)
> summary(model2.k)
Call:
mb1m(formula = ITEM_PRICE ~ SHOP_ID, dataframe = mostra)
Residuals:
      Min       1Q   Median       3Q      Max
-454.4  -205.5  -55.5   544.5 27535.5
Coefficients:
              Estimate      MAD    V value      Pr(>|V|)
(Intercept)    454.47    366.38 4021236    < 2e-16 ***
SHOP_ID         0.00      9.47  1430684 0.000000546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1506 on 2934 degrees of freedom
```



- `model3 = mbIm(ITEM_PRICE ~ ITEM_CATEGORY_ID, data=mostra)`

```
> model3.k <- mbIm(ITEM_PRICE ~ ITEM_CATEGORY_ID, data=mostra)
```

```
> summary(model3.k)
```

Call:

```
mbIm(formula = ITEM_PRICE ~ ITEM_CATEGORY_ID, dataframe = mostra)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-470.9	-39.7	129.1	639.4	27362.7

Coefficients:

	Estimate	MAD	V value	Pr(> V )
(Intercept)	750.80	491.93	4079325	<2e-16 ***
ITEM_CATEGORY_ID	-10.29	12.41	591500	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1509 on 2934 degrees of freedom

- `model4 = mbIm(ITEM_PRICE ~ ITEM_CNT_ID, data=mostra)`

```
> model4.k <- mbIm(ITEM_PRICE ~ ITEM_CNT_DAY, data=mostra)
```

```
> summary(model4.k)
```

Call:

```
mbIm(formula = ITEM_PRICE ~ ITEM_CNT_DAY, dataframe = mostra)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1594	-150	0	600	27141

Coefficients:

	Estimate	MAD	V value	Pr(> V )
(Intercept)	349.0	518.9	3955381	<2e-16 ***
ITEM_CNT_DAY	50.0	222.4	19415950.102	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1516 on 2934 degrees of freedom

Revisant la pendent, intercept, MAD i p-value creiem que el millor model és el model 1. Tot i que també creiem que no té cap sentit aquesta fórmula per a predir el preu.

$$\text{ITEM\_PRICE} = 308.091 + 7.143 \text{ DATE\_BLOCK\_NUM}$$

Provem a fer la predicció de vendes d'un producte del mes '30', juliol 2015

```
> newdata <- data.frame(
+ DATE_BLOCK_NUM = 30
+ )
> predict(model1.k, newdata)
522.381
```

També hem executat el model de regressió *Quantile* amb valor  $\tau=0.5$ , tot i que creiem que no millora el model de predicció de *Kendall-Theil* ja generat.

```
model_q50 <- rq(ITEM_PRICE ~ DATE_BLOCK_NUM, tau = 0.5, data = vendes)
summary(modelo_q50)
```

```
Call: rq(formula = ITEM_PRICE ~ DATE_BLOCK_NUM, tau = 0.5, data = mostra)
tau: [1] 0.5
```

Coefficients:

	Value	Std. Error	t value	Pr(> t )
(Intercept)	367.73684	14.78255	24.87641	0.00000
DATE_BLOCK_NUM	5.21053	1.43332	3.63529	0.00028

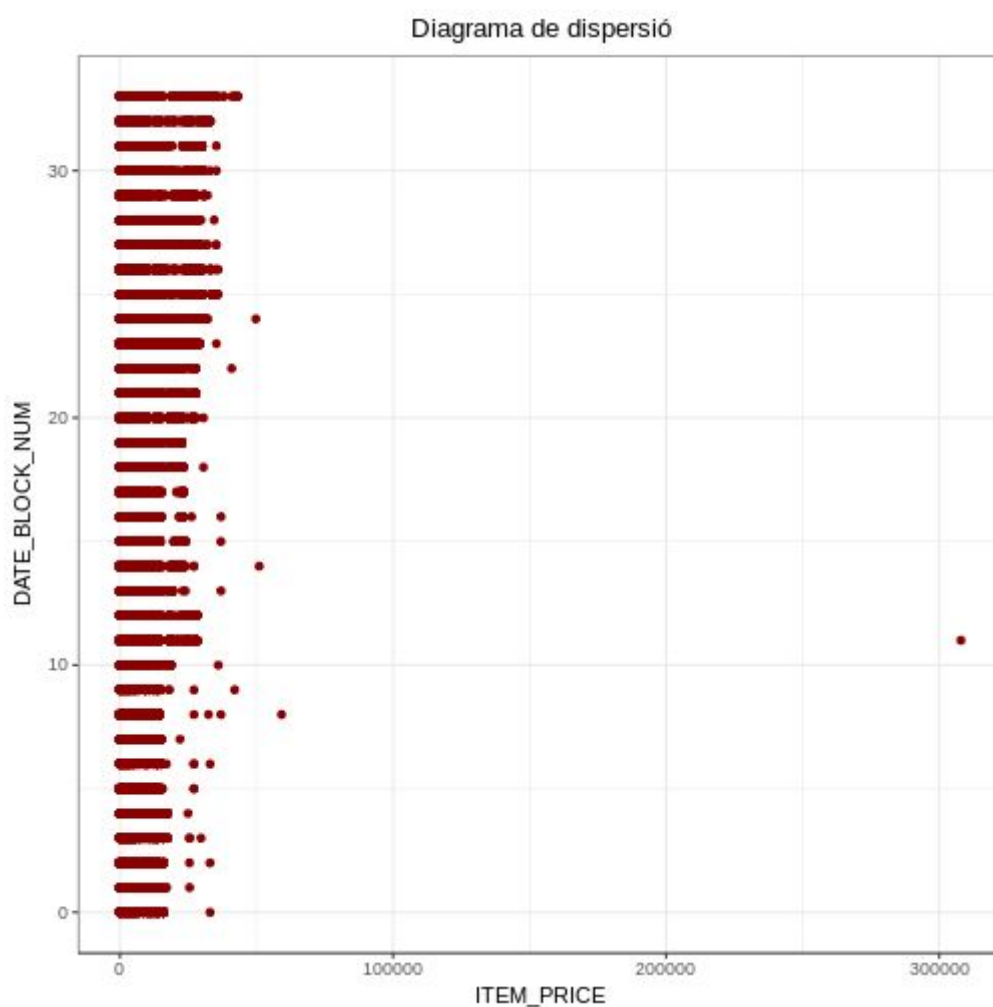
## 5. Representació dels resultats a partir de taules i gràfiques.

### **Objectiu 1: Quines variables influeixen més en el preu dels productes?**

Tot i que durant l'anàlisi hem observat que la variable *DATE\_BLOCK\_NUM* és la que més influeix en el preu dels productes, mitjançant el **diagrama de dispersió** podem comprovar que no existeix relació lineal entre *ITEM\_PRICE* i *DATE\_BLOCK\_NUM*.

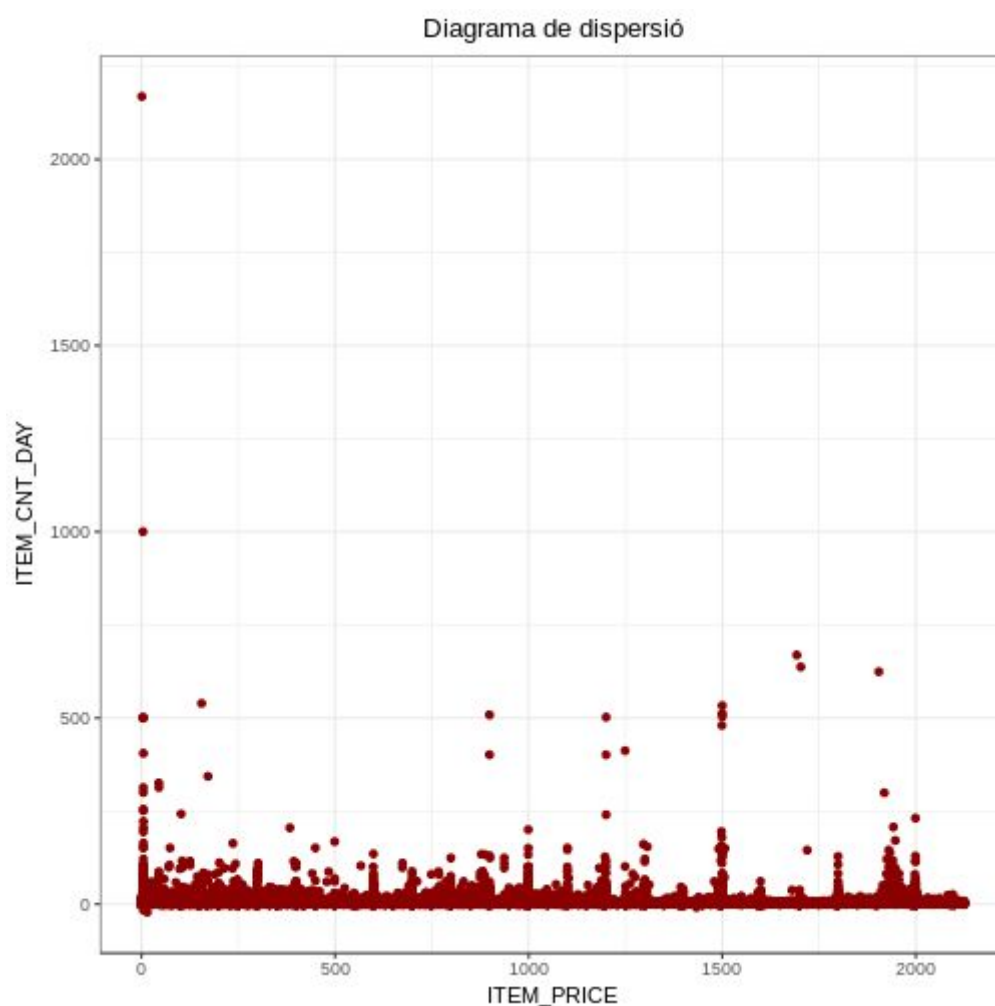
```
require(MASS)
require(ggplot2)

ggplot(data = vendes, aes(x = ITEM_PRICE, y = DATE_BLOCK_NUM)) +
  geom_point(colour = "red4") +
  ggtitle("Diagrama de dispersió") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



El diagrama de dispersió entre *ITEM\_PRICE* i *ITEM\_CNT\_DAY* es penja en R amb el dataset inicial però traient els outliers de *ITEM\_PRICE* sí que es genera el gràfic. Tampoc s'observa relació entre *ITEM\_PRICE* i *ITEM\_CNT\_DAY*.

```
ggplot(data = vendes, aes(x = ITEM_PRICE, y = ITEM_CNT_DAY)) +  
  geom_point(colour = "red4") +  
  ggtitle("Diagrama de dispersió") +  
  theme_bw() +  
  theme(plot.title = element_text(hjust = 0.5))
```



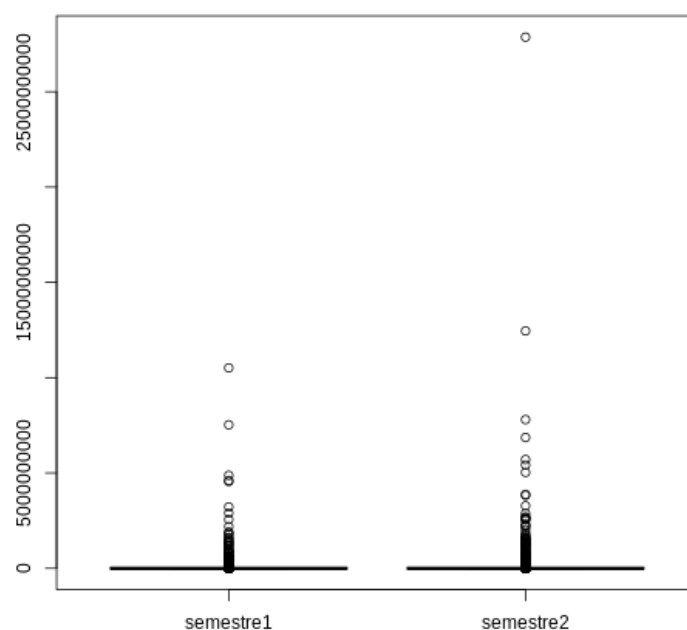
## **Objectiu 2: La facturació és superior durant el segon semestre de l'any?**

Amb el **diagrama de caixa** visualitzem el contrast d'hipòtesis que confirma que la facturació és superior durant el segon semestre.

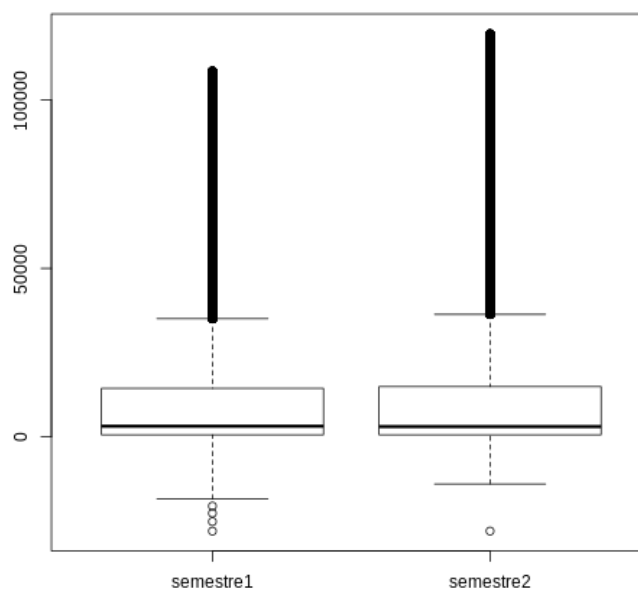
Es visualitza millor havent eliminat els outliers de la variable TOTAL en els datasets.

Diagrama de caixa sense eliminar outliers:

```
>boxplot(vendes_1semestre$TOTAL,vendes_2semestre$TOTAL, names=c("semestre1","semestre2"))
```



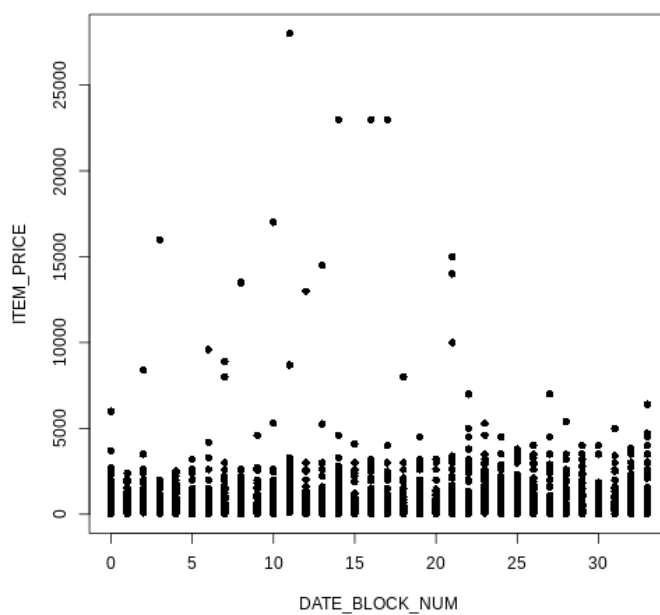
Amb el diagrama de caixa eliminant outliers, observem que la mitjana de vendes del segon semestre (13797.85) és lleugerament superior a la mitjana de vendes del primer semestre (12866.84).



**Objectiu 3: Crearem models de regressió que permetin predir el preu dels productes**

Amb el diagrama següent veiem que no tenen relació ITEM\_PRICE i DATA\_BLOCK\_NUM com per a crear un model predictiu fiable.

```
> plot(ITEM_PRICE ~ DATE_BLOCK_NUM, data = mostra, pch = 16)
```



6. **Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions?**  
Els resultats permeten respondre al problema?

**Objectiu 1: Quines variables influeixen més en el preu dels productes?**

En aquest objectiu hem buscat la correlació de les diferents variables per a determinar quines tenen major influència sobre el preu dels productes.

Havent comprovat que les dades no segueixen una distribució normal, hem utilitzat el coeficient de correlació de Spearman per cercar la relació entre les variables.

El test de correlació de Spearman es desaconsella en variables qualitatives, per tant, hem aplicat el test en les variables ITEM\_PRICE, ITEM\_CNT\_DAY i DATE\_BLOCK\_NUM. Aquesta última variable sí que la considerem ordinal, ja que mostra una evolució en el temps.

Amb els resultats obtinguts, concloem que la variable que més influència té sobre el preu del producte és el temps (DATE\_BLOCK\_NUM).

Tot i això ens hauria agradat poder estudiar més correlacions entre variables usant tant les quantitatives com les qualitatives, però no hem pogut trobar l'algorisme adequat per a fer-ho.

**Objectiu 2: La facturació és superior durant el segon semestre de l'any?**

Per a determinar si en el segon semestre de l'any hi ha més vendes que en el primer, hem aplicat una prova de contrast d'hipòtesi sobre dues mostres. Les mostres són:

- dataset\_primer\_semestre.csv: conté les vendes del primer semestre
- dataset\_segond\_semestre.csv: conté les vendes del segon semestre

Com les dades no segueixen una distribució normal, no podem emprar tests paramètrics, i per tant hem optat per utilitzar el test no paramètric *Mann–Whitney–Wilcoxon (WMW)*.

Després d'executar aquest test, hem obtingut un valor de p-value menor que el valor de confiança fixat i, per tant, podem rebutgem la hipòtesi nul·la i concloem que la facturació durant el segon semestre de l'any és superior a la facturació del primer.

**Objectiu 3: Crearem models de regressió que permetin predir el preu dels productes**

Tot i que és molt interessant poder realitzar prediccions sobre el preu dels productes en funció de les característiques relacionades amb la venda, no hem arribat a generar cap model de regressió que ens sembli correcte.

Hem tingut dificultats amb el dataset escollit, ja que no segueix una distribució normal, hi ha poques variables amb les que poder realitzar la inferència i la majoria de variables són qualitatives, en escala no ordinal per a poder aplicar els tests no paramètrics de models de regressió que podríem haver aplicat.

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

[https://github.com/mcalzada/predict\\_future\\_sales\\_cleaning/blob/master/code/predict\\_future\\_sales\\_cleaning.R](https://github.com/mcalzada/predict_future_sales_cleaning/blob/master/code/predict_future_sales_cleaning.R)



## Recursos web:

<https://github.com/kazimanil/predict-future-sales>

<https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet>

<https://www.kaggle.com/c/competitive-data-science-predict-future-sales/discussion/54949>

<https://techtravelo.wordpress.com/2014/03/21/installing-package-forecast-in-r/>

<https://estadisticaorquestainstrumento.wordpress.com/2013/04/30/tema-25-analisis-de-series-temporales/>

<http://www.doctormetrics.com/2017/04/27/introduccion-al-forecasting-con-r-statistics/#.WvibDRyxU5k>

<https://www.kdnuggets.com/2018/03/time-series-dummies-3-step-process.html>

<https://www.kaggle.com/vyordanov/simple-prediction-approach-to-get-you-in-top-70/code>

<https://www.otexts.org/fpp/9/4>

<https://www.kaggle.com/jagangupta/time-series-basics-exploring-traditional-ts>

[https://rstudio-pubs-static.s3.amazonaws.com/289564\\_7557e57a8aac42b1a8ca434689ee3cff.html](https://rstudio-pubs-static.s3.amazonaws.com/289564_7557e57a8aac42b1a8ca434689ee3cff.html)

<http://rfunction.com/archives/1692>

<https://www.marblestation.com/?p=794#00115>

[https://rpubs.com/Joaquin\\_AR/223351](https://rpubs.com/Joaquin_AR/223351)

[https://rpubs.com/Joaquin\\_AR/220579](https://rpubs.com/Joaquin_AR/220579)

[https://rpubs.com/Joaquin\\_AR/218456](https://rpubs.com/Joaquin_AR/218456)

[http://rcompanion.org/handbook/F\\_12.html](http://rcompanion.org/handbook/F_12.html)

[http://ri.uaemex.mx/bitstream/handle/20.500.11799/68226/CAPITULO%20DE%20LIBRO\\_estadistica.pdf?sequence=1&isAllowed=y](http://ri.uaemex.mx/bitstream/handle/20.500.11799/68226/CAPITULO%20DE%20LIBRO_estadistica.pdf?sequence=1&isAllowed=y)

Criteris de valoració

Tots els apartats són obligatoris. La ponderació dels exercicis és la següent:

- Els apartats 1, 2 i 6 valen 0,5 punts.
- Els apartats 3,5 i 7 valen 2 punts.
- L'apartat 4 val 2,5 punts.

Es valorarà la idoneïtat de les respostes, que han de ser clares i completes. Les diferents etapes han d'estar ben justificades i acompanyades del codi corresponent. També es valorarà la síntesi i claredat, a través de l'ús de comentaris, del codi resultant, així com la qualitat de les dades finals analitzades.