Pràctica 2: Neteja i validació de les dades

Nom: Carlos A. García Pérez

Professora col·laboradora: Laia Subirats

Data de lliurament: 03/01/2019

1. Descripció del Dataset

El dataset se correspon amb les dades ampliades de la primera pràctica d'obtenció de dades. Són **1.515 files i 18 columnes** (variables).
Les variables són:

- **SearchConcept.** Concepte de cerca. Tots estan relacionats amb la fotografia i són Nikon, Canon, Leica i Fujifilm. Evidentment, deixem fora altres marques com Olympus o Sony. El dataset és ampliable amb tots els conceptes que volguem estudiar.
- **Page.** Ordre de la pàgina de cerca. És d'esperar que l'importància del producte està relacionada amb l'ordre de pàgina en la que apareix.
- **ImgName.** Nom del fitxer de l'imatge descarregada. En un procés de Machine Learning es pot analitzar l'imatge per extreure conclusions: Qué és el producte? A quin tipus es correspon? Està realment connectat amb el search concept?
  Molts de fitxers es repeteixen (l'imatge 51Ir4alVqeL._AC_US218_.jpg apareix 42 vegades). Això és perque molts de productes tenen diferents versions però només una imatge.

- **Code.** Codi del producte per a Amazon. Tenim les mateixes repeticions que al punt anterior. Per exemple, l'imatge anterior es correspon amb el producte B01N36317N. Tenim fins a 317 productes sense codi.
- **Description.** Descripció del producte mostrada a Amazon. Les repeticions de registres pràcticament es corresponen amb els casos anteriors.
- **Price.** Preu tal i com apareix a la web d'Amazon. Pot aparèixer de dues formes:
  - **Moneda i valor.** Exemple: EUR 249,29
  - **Moneda i rang** (mínim i màxim): EUR 37,87 - EUR 41,21
- **PriceValue.** Quantitat del preu. Si és un rang, es correspon amb al mínim d'aquest rang.
- **PriceCurrency.** Moneda del preu. En el nostre cas, sempre és EUR.
- **PriceMax.** Quantitat del preu. Si és un rang, es correspon amb al màxim d'aquest rang.
- **Image.** URL on es pot trobar l'imatge original del producte. És on l'hem d'obtenir si la volem tornar a recuperar. Els valors a partir de l'url base es correspon amb imgName, així que no cal estudiar per separat aquesta variable.

- **Puntuation.** Puntuació del producte tal i com apareix a la web d'Amazon. Exemple: "4,4 de un máximo de 5 estrellas".
- **PuntuationValue.** Valor numèric de la puntuació. En el nostre exemple anterior: 4.4.
- **AltImage.** Texte que apareix a la web quan naveguem per l'imatge. Es correspon amb la descripció; per això, eliminarem la columna. Es pot veure que són iguals executant:

```
Rcmdr> countDistinct <- length(which(PhotoDataset$description != PhotoDataset$altImage))

Rcmdr> countDistinct
[1] 0
>
```

- **Stock.** Quan existeix un stock limitat, texte tal i com apareix a la web. Per exemple: "Sólo hay 3 en stock. Cómpralo cuanto antes".
- **StockValue.** Valor numèric del stock quan hi apareix. En l'exemple anterior: 3.
- **Date.** Data de creació del registre del dataset (dia, hora, minuts, segons i milisegons). Aquesta variable no l'inclourem a l'anàlisi; simplement serveix a nivell informatiu.
- **DateDay.** Data de creació del dataset (dia). Aquesta variable no l'inclourem a l'anàlisi; simplement serveix a nivell informatiu.
- **DateHour.** Data de creació del dataset (hora, minuts i segons). Aquesta variable no l'inclourem a l'anàlisi; simplement serveix a nivell informatiu.

2. Resum de dades

El resum de les dades és:

searchConcept	Canon	Fujifilm	Leica	Nikon
# registres	374	380	379	382

page	Pàgina	Registres
	1	79
	2	80
	3	77
	4	79
	5	79
	6	79
	7	78
	8	78
	9	77
	10	80
	11	79
	12	77
	13	79
	14	79
	15	77
	16	80
	17	78
	19	79
	20	24

imgName	Imatge	Registres
	51Ir4alVqeL._AC_US218_.jpg	42
	41yp0fRMGqL._AC_US218_.jpg	22
	51ruwrL+HvL._AC_US218_.jpg	22
	41bC3fBd2SL._AC_US218_.jpg	21
	41kvtnySSHL._AC_US218_.jpg	21
	41sc43hZMzL._AC_US218_.jpg	21
	Altres	1364

code	Imatge	Registres
	B01N36317N	42
	B01MZBY4WQ	22
	B071P55939	22
	B01N5J5F5E	21
	B06W9K61SB	21
	NA	317
	Altres	1.070

description	
Descripció	Registres
Fujifilm X100F - Cámara compacta de 24.3 MP (pantalla de 3, visor hibrido, video Full HD)	43
Fujifilm X-T20- Cámara EVIL de 24 MP (pantalla de 3, visor electrónico, resolución máxima 4K) negro - kit cuerpo con objetivo XC 16-50 mm F3.5-5.6 OIS II	22
Powerextra Batería Nikon EN-EL14 EN EL14a para Nikon D3100 D3200 D3300 D5100 D5200 D5300 D5500 Coolpix P7000 P7100 P7700 P7800 EN EL14a DSLR Cámara	22
Canon EOS M5 - Kit de Cámara Evil de 24.2 MP con Objetivo EF-M 15-45 S (Pantalla Táctil DE 3.2, DIGIC, NFC, CMOS, Bluetooth, ISO, EF Lenses, Remote Shooting, Full HD, WiFi) Negro	21
Canon EOS M50 - Kit de cámara EVIL de 24.1 MP y vídeo 4K con objetivo EF-M 15-45mm IS MM (pantalla táctil de 3, estabilizador óptico, Wifi), color negro	21
Canon EOS M6 - Cámara Evil de 24.2 MP (Pantalla táctil de 3.0, DIGIC 7, NFC, Dual Pixel CMOS AF, Bluetooth, 5 - Axis Digital IS, Full HD, WiFi) Plata - Kit Cuerpo con Objetivo EF-M 15-45	21
Altres	1.366

Price

El resum del preu tal qual apareix a la web d'Amazon és:

price	preu	Registres
	EUR 1.261,01	42
	EUR 19,99	38
	EUR 29,99	34
	EUR 49,99	25
	EUR 36,99	23
	EUR 999,99	23
	Altres	1.330

Això no ens dóna gaire informació. Si anem al detall del priceValue (valor numèric del preu), podem començar a veure un poc més de detall (tots els valors són en euros):

priceValue	Clau	Valor
	Mínim	1,05
	Primer quartil	22,79
	Mitja	340,20
	Tercer quartil	396,00
	Màxim	33.182,98

Com que pocs productes tenen un rang de preu, la columna priceMax és molt semblant a priceValue:

priceMax	Clau	Valor
	Mínim	1,05
	Primer quartil	22,99
	Mitja	340,25
	Tercer quartil	396,0
	Màxim	33.182,98

Puntuation

El resum de les puntuacions tal qual apareix a la web d'Amazon és:

puntuation	puntuation	Registres
	4,3 de un máximo de 5 estrellas	153
	5 de un máximo de 5 estrellas	150
	4,4 de un máximo de 5 estrellas	123
	4 de un máximo de 5 estrellas	101
	4,6 de un máximo de 5 estrellas	95
	NA	317
	Altres	576

Per lo vist, és d'esperar que les notes siguin molt altes. Si mirem el detall numèric:

puntuationValue	Clau	Valor
	Mínim	1,000
	Primer quartil	4,000
	Mitja	4,223
	Tercer quartil	4.600
	Màxim	5,000

Stock

El resum de l'etoc tal qual apareix a la web d'Amazon és:

puntuation	puntuation	Registres
	Sólo hay 1 en stock. Cómpralo cuanto antes	101
	Sólo hay 2 en stock. Cómpralo cuanto antes	49
	Sólo hay 3 en stock. Cómpralo cuanto antes	49
	Sólo hay 4 en stock. Cómpralo cuanto antes	30
	Sólo hay 5 en stock	35
	Altres	1.231

És a dir, només s'especifica quan queden 5 o menys unitats. Si mirem el detall numèric:

puntuationValue	Clau	Valor
	Mínim	1,000
	Primer quartil	1,000
	Mitja	2,428
	Tercer quartil	3,000
	Màxim	5,000

3. Tractament previ de dades

S'ha fet una neteja prèvia de dades a la pròpia captura del dataset. El detall es troba al fitxer dataCleaningUtils.py. Com a resum:

Neteja prèvia

L'entorn R no acepta tots els caràcters. Falla amb els caràcters (", ' i #). Com que no són rellevants per a l'anàlisi, s'eliminen i no s'escapen.

```
36      def deleteInvalidChar(value):
37          valueIn = value.replace('\"', "")
38          valueIn = valueIn.replace("\'", "")
39          valueIn = valueIn.replace('#', "")
40          return valueIn
```

D'igual forma, no ens interessen els productes gratis o que estan inclosos a una subscripció.

```
42      def isValidRow(value):
43          if (value in ('Escuchar con Unlimited', 'GRATIS')):
44              return 1
45          return 0
```

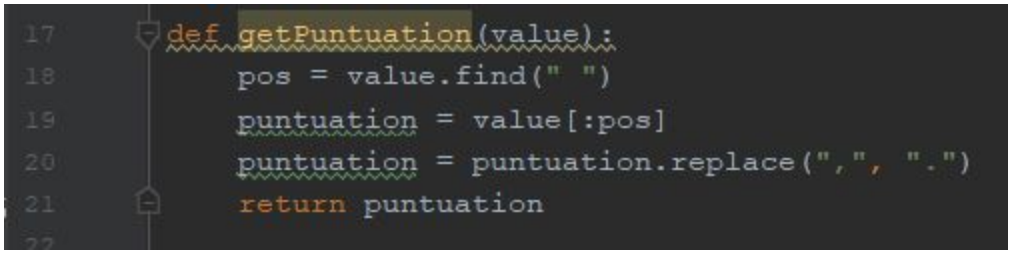
Neteja de preu

El codi distingueix dos casos: valor únic o rang. Es pot distinguir entre els dos casos perque els rangs s'especifiquen amb un guió. Una vegada fet això, es cerquen els espais en blanc entre preu i moneda. Així podem retornar els valors per separat (preu, moneda i preu màxim).

```
1   def getPrice(value):
2       posGui = value.find("-")
3       if (posGui == -1):
4           pos = value.find(" ")
5           price = value[pos+1:]
6           currency = value[:pos]
7           price = price.replace(".", "")
8           return price.replace(",", "."), currency, price.replace(",", ".")
9       else:
10          pos = value.find(" ")
11          currency = value[:pos]
12          price = value[pos+1:posGui-1]
13          posEnd = value.rfind(" ")
14          priceEnd = value[posEnd+1:]
15          return price.replace(",", "."), currency, priceEnd.replace(",",".")
```

Neteja de puntuació

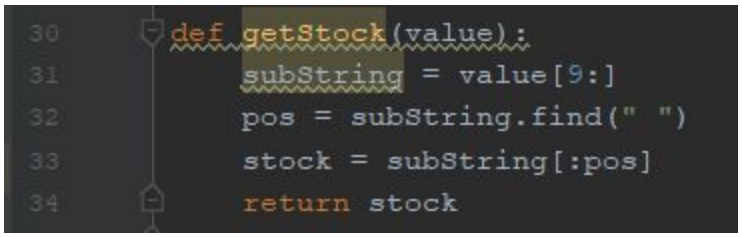
Com es pot veure, si hi han puntuacions, sempre tenen la estructura “4,4 de un máximo de 5 estrellas”. Obtenir la puntuació és molt senzill: basta agafar el valor fins al primer espai en blanc.



```
17  def getPuntuation(value):
18      pos = value.find(" ")
19      puntuation = value[:pos]
20      puntuation = puntuation.replace(",", ".")
21      return puntuation
```

Neteja d’estoc

L’estoc sempre s’especifica de la forma “Sólo hay 4 en stock.[...]”. Com que l’estoc sempre és una xifra, bastaria agafar el valor que ocupa la posició 10. Per a evitar errors, agafem des de la posició 9 fins al següent espai en blanc.



```
30  def getStock(value):
31      subString = value[9:]
32      pos = subString.find(" ")
33      stock = subString[:pos]
34      return stock
```

4. Objectiu i neteja de dades

La nostra hipòtesi és que els productes a Amazon apareixen a les primeres pàgines en funció de la seva rellevància. És a dir: com més relevant és un producte, abans sortirà com a resultat. Les variables amb les que contam per a obtenir la rellevància són:

- El preu. És de suposar que, com més elevat és el preu, més rellevant és el producte.
- La puntuació. Com més elevada és la puntuació, més rellevant és del producte.
- L'estoc. Els estocs limitats poden afectar al resultat. No tenc clar de quina forma, però hi ha dues hipòtesi raonables:
    - Mostrar primer els productes amb estoc limitat. Això suposa una pressió cap al client, que es veu obligat a prendre una decisió.
    - Mostrar els productes amb estoc limitat al final. Vendre primer això que tenim "il·limitat".

El tractament de neteja de dades se farà amb R. Les pases seguides són:

Càrrega de dades i tractament inicial

A aquest punt:

- Carreguem el dataset general al projecte. Se correspon amb el fitxer que es pot trobar al Git Hub.
- Validem que les dades són correctes i s'ha carregat de forma adequada.
- Comprovem que hi ha columnes redundants (description i altImage, per exemple)
- Eliminem les columnes redundants. Com tenim els valors numèrics del preu, puntuació, stock, … no necessitem els valors originals.
- Eliminem les columnes que no farem servir; per exemple, l'instant concret de captura no és rellevant en el nostre cas.
- Com que hi ha productes que es repeteixen a diferents pàgines, el que farem és eliminar els duplicats; ens quedem amb la mínima pàgina. Això té sentit, ja que l'ordre de la pàgina és un indicador de com d'important és el producte. Si apareix més d'una vegada, ens quedem amb el registre més rellevant.

```
#Carreguem el dataset
PhotoDataset <- read.table("D:/Users/cagarcia/uoc/M2951/data/productData.csv",
header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE)

#Visualitzem les dades
fix(PhotoDataset)

#Resum de dades
summary(PhotoDataset)
```

#Comprovem que les columnes description i altImage són iguals
countDistinct <- length(which(PhotoDataset$description != PhotoDataset$altImage))
countDistinct

#Eliminem les columnes redundants
PhotoDataset <- subset(PhotoDataset, select = -c(price, stock, puntuation, stock, altImage, image))

#Ens quedem amb el dia de la captura i eliminem dia i hora
PhotoDataset <- subset(PhotoDataset, select = -c(date, dateHour))

#Estudiem la correlació entre variables numèriques
cor(PhotoDataset[,c("page","priceMax","priceValue","puntuationValue","stockValue")], use="complete")

```
Rcmdr> cor(PhotoDataset[,c("page","priceMax","priceValue","puntuationValue","stockValue")], use="complete")
                    page    priceMax  priceValue puntuationValue stockValue
page             1.00000000 -0.07575320 -0.07575320     -0.01615406  0.1884974
priceMax        -0.07575320  1.00000000  1.00000000     -0.07652032  0.1672411
priceValue      -0.07575320  1.00000000  1.00000000     -0.07652032  0.1672411
puntuationValue -0.01615406 -0.07652032 -0.07652032      1.00000000 -0.1229245
stockValue       0.18849743  0.16724106  0.16724106     -0.12292451  1.0000000
> |
```

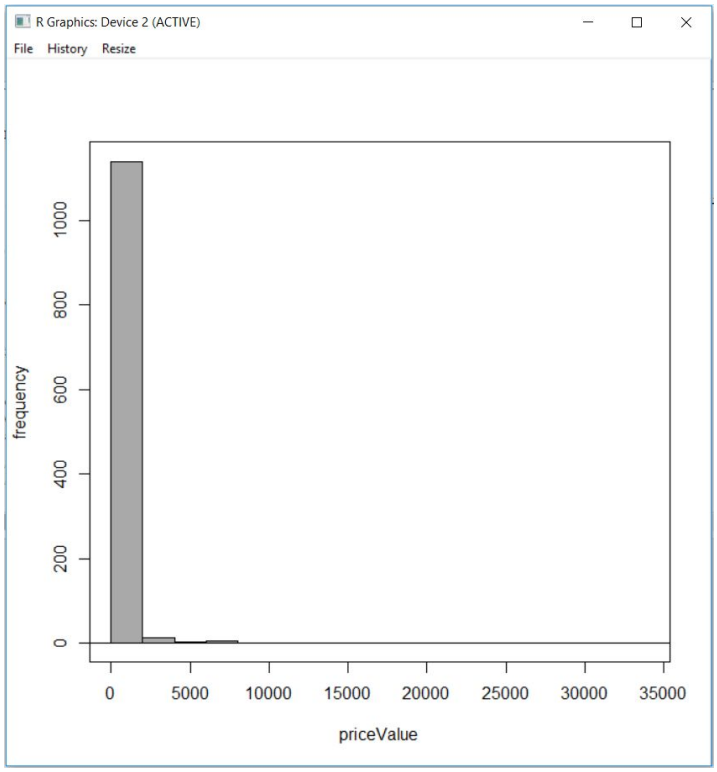
Si tenim en compte les tres variables (podem considerar priceValue i priceMax com a iguals), podem veure que les dues més correlacionades són page i stockValue.

#Eliminem duplicats, agafant la mínima pàgina
require(sqldf)
PhotoDataset <- sqldf("select distinct searchConcept,code,imgName,description,priceValue,priceCurrency,priceMax,puntuationValue, stockValue,dateDay,min(page) as page from PhotoDataset group by searchConcept,code,imgName,description,priceValue,priceCurrency,priceMax,puntuationValue, stockValue,dateDay", drv = 'SQLite')

#Obtenim les dades
sqldf("select page, count(*) from PhotoDataset group by page order by page asc", drv = 'SQLite')

Neteja de preu

La gran majoria dels preus es troben en el rang baix de preus, com es pot veure a la gràfica.



Es veu de forma exacta si agrupem en rangs de 500 euros

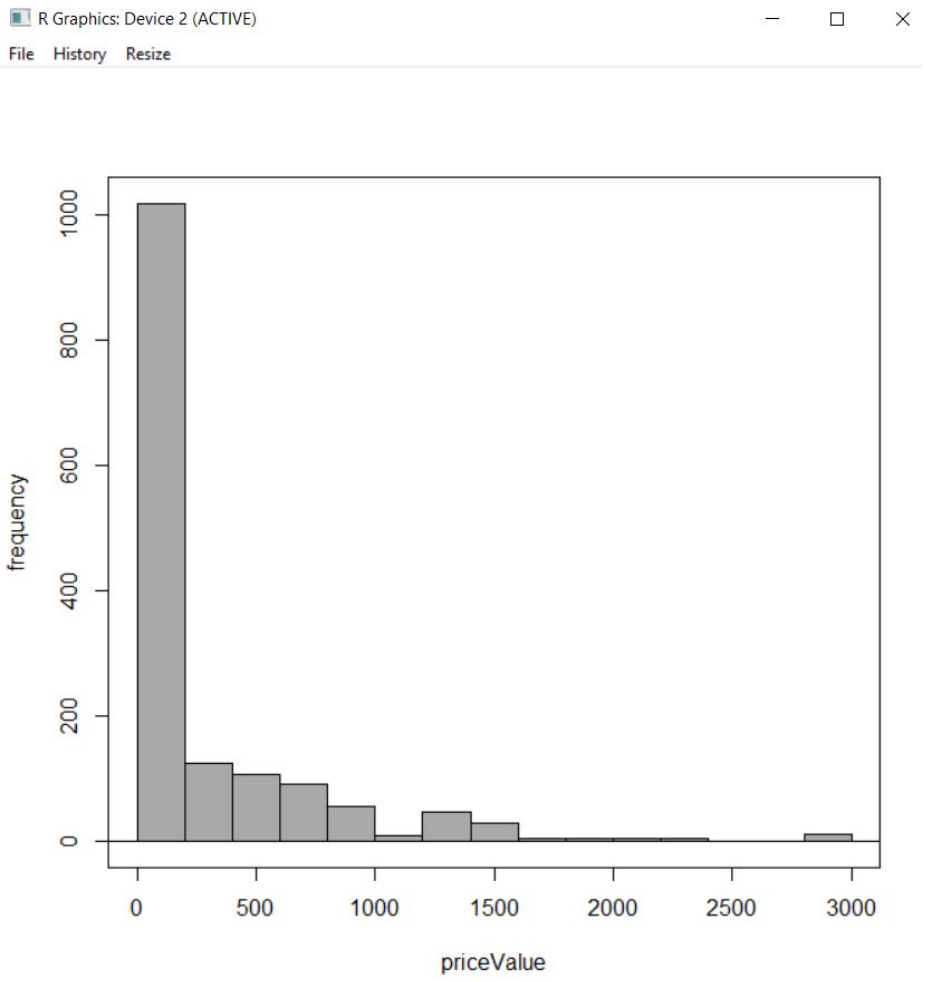
price	Rang	Registres
	0-500	1196
	501-100	199
	1001-1500	84
	1501-2000	14
	2001-2500	8
	2501-3000	3
	> 3000	11

Si calculem la desviació estàndard del preu

Rcmdr> sd(PhotoDataset$priceValue, na.rm = FALSE)
[1] 1080.242

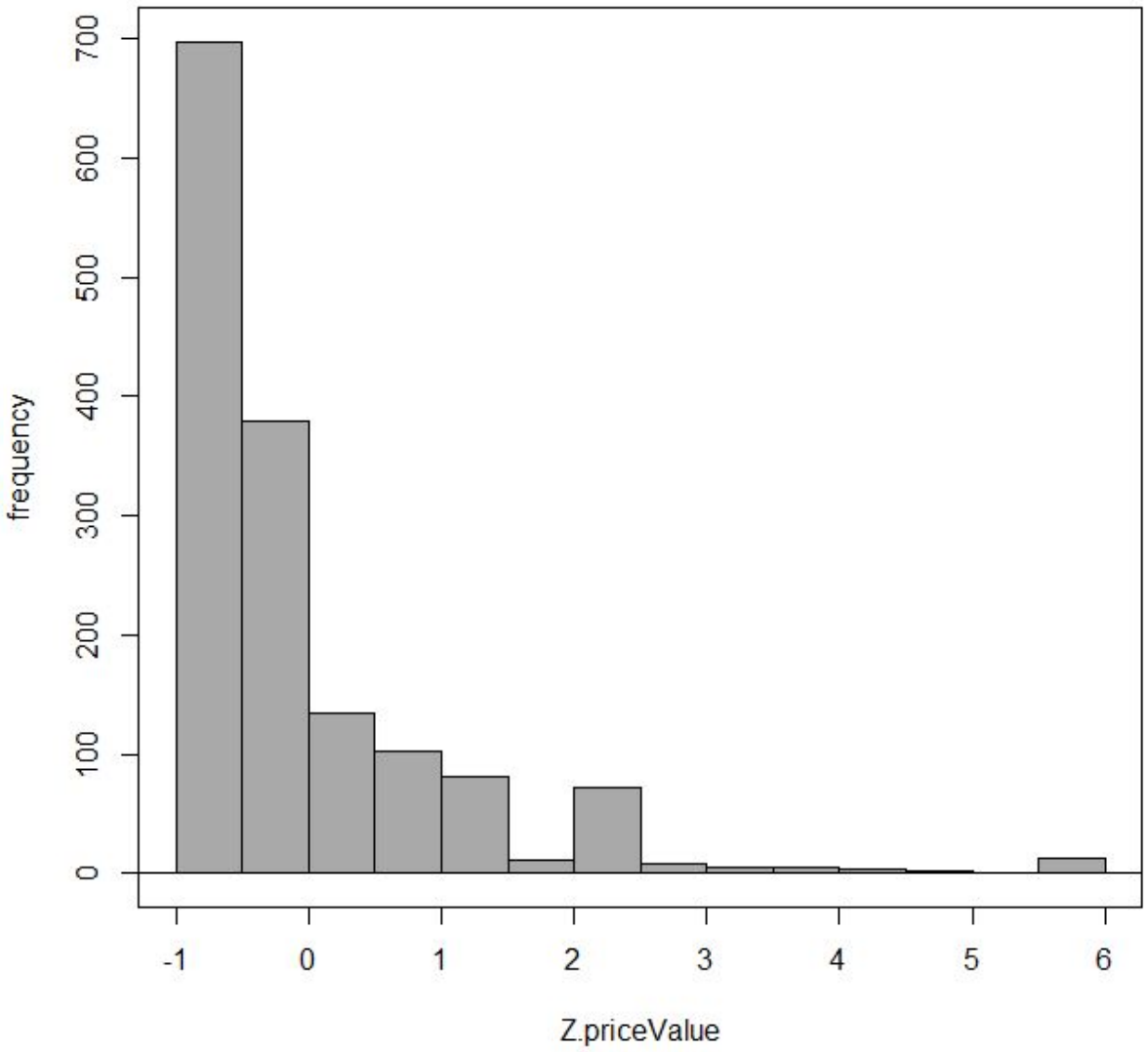
Està clar que els valor elevats (més de 3.000€) estan distorsionant la distribució, malgrat són només 11 registres. Assignem el valor 3.000 als registres amb un valor superior.

```
#Assignem el valor de 3.000 als preus superiors
PhotoDataset$priceValue <- ifelse(PhotoDataset$priceValue > 3000, 3000,
PhotoDataset$priceValue)
```



Si estandaritzem els valors per a que sigui més manejable, tenim una variable equivalent

```
#Standaritzem el preu
PhotoDataset <- local({
  .Z <- scale(PhotoDataset[,c("priceValue")])
  within(PhotoDataset, {
    Z.priceValue <- .Z[,1]
  })
})
```

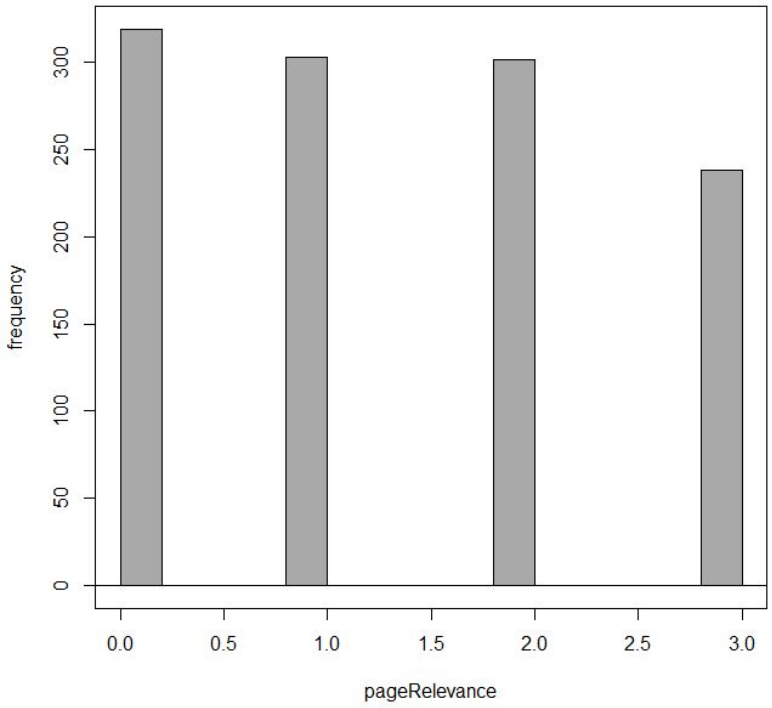

frequency
0
100
200
300
400
500
600
700
-1
0
1
2
3
4
5
6
Z.priceValue

Neteja de pàgina

La pàgina a la que apareix un resultat és molt important. Sabem que, com més relevant és un resultat, abans apareixerà a la web. Si volguessim crear un producte nou, el nostre principal objectiu comercial hauria de ser aparèixer a les primeres pàgines.
Després de l'acció ja definida d'agafar el mínim de pàgina i eliminar duplicats, tenim:

```
# Agrupem les pàgines
# Considerem que les pàgines 1-5 són més rellevants que les de 6-10, i aquestes més que les 11-15, ...
PhotoDataset$pageRelevance <- trunc(PhotoDataset($page-1) / 5)
sqldf("select page, count(*) from PhotoDataset group by page order by page asc", drv = 'SQLite')
sqldf("select pageRelevance, count(*) from PhotoDataset group by pageRelevance order by pageRelevance asc",
drv = 'SQLite')
```

page	Rang	Registres
	1-5	319
	6-10	303
	11-15	301
	16-20	238



Neteja d'estoc

Realment, podem considerar dos casos:

1. L'estoc és limitat i hem de gestionar un número limitat d'unitats
2. L'estoc és ilimitat

#Creem la columna que indica si hi ha stock limitat (1) o no (0)
PhotoDataset$stockIsLimited <- ifelse(is.na(PhotoDataset$stockValue), 0, 1)

#Validem que l'existència d'stock s'ha calculat ok
sqldf("select stockValue, stockIsLimited, count(*) from PhotoDataset group by stockValue, stockIsLimited", drv = 'SQLite')

```
Rcmdr> sqldf("select stockValue, stockIsLimited, count(*) from PhotoDataset group by stockValue, stockIsLimited", drv = 'SQLite')
  stockValue stockIsLimited count(*)
1         NA              0      963
2          1              1       78
3          2              1       46
4          3              1       29
5          4              1       29
6          5              1       16

Rcmdr> sqldf("select stockIsLimited, count(*) from PhotoDataset group by stockIsLimited", drv = 'SQLite')
  stockIsLimited count(*)
1              0      963
2              1      198
>
```

És a dir: tenim 963 registres sense estoc limitat i 198 amb estoc limitat

Neteja de puntuació

Si mirem el detall de les puntuacions:

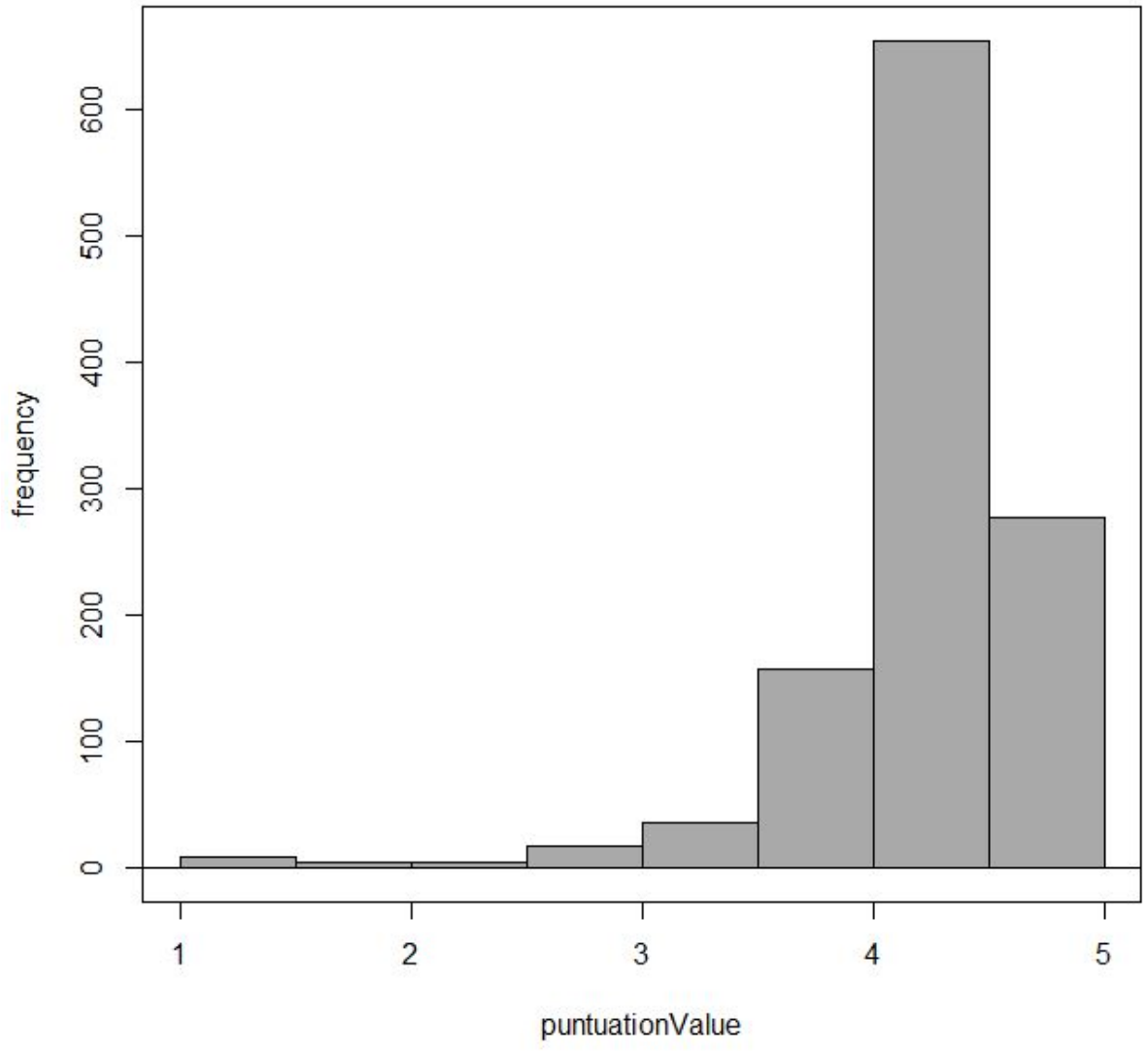
```
Rcmdr> summary(PhotoDataset$puntuationValue)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   4.000   4.300   4.276   4.600   5.000     257
>
```

Podem veure que hi ha poca dispersió. De fet, si no tenim en compte els NA, la desviació típica és de 0.5989409:

```
Rcmdr> sd(PhotoDataset$puntuationValue, na.rm = TRUE)
[1] 0.5989409
>
```

Per a poder tractar els NA com a valors, crec que en aquest cas el millor és assignar-lis la mitja:

```
#Assignam la mitja a les puntuacions NA
PhotoDataset$puntuationValue[is.na(PhotoDataset$puntuationValue)] <-
mean(PhotoDataset$puntuationValue, na.rm = TRUE)
```



5. Anàlisi de resultats

Matriu de correlacions

Si obtenim la matriu de correlacions de les variables numèriques, podem veure que inclús després del tractament, no estan forçament relacionades:

#Matriu de correlacions
cor(PhotoDataset[,c("page","priceMax","priceRang","priceValue","puntuationValue","stockValue","Z.priceValue", "pageRelevance", "stockIsLimited")], use="complete")

```
> cor(PhotoDataset[,c("page","priceMax","priceRang","priceValue","puntuationValue","stockValue","Z.priceValue", "pageRelevance", "stockIsLimited")]
)
Warning in cor(PhotoDataset[, c("page", "priceMax", "priceRang", "priceValue",  :
  the standard deviation is zero
                      page    priceMax   priceRang  priceValue puntuationValue   stockValue Z.priceValue  pageRelevance stockIsLimited
page            1.0000000000 -0.11665429 -0.05880271 -0.09620844   0.00042229030  0.118406947  -0.09620844  0.96608802680             NA
priceMax       -0.1166542876  1.00000000  0.89134628  0.87059844  -0.02571294006 -0.174557286   0.87059844 -0.11274934038             NA
priceRang      -0.0588027058  0.89134628  1.00000000  0.96733840  -0.04352311594 -0.188489146   0.96733840 -0.06473891707             NA
priceValue     -0.0962084400  0.87059844  0.96733840  1.00000000  -0.03433580855 -0.181418306   1.00000000 -0.09742795991             NA
puntuationValue  0.0004222903 -0.02571294 -0.04352312 -0.03433581   1.00000000000  0.006488645  -0.03433581 -0.00007664498             NA
stockValue      0.1184069469 -0.17455729 -0.18848915 -0.18141831   0.00648864524  1.000000000  -0.18141831  0.09866793283             NA
z.priceValue   -0.0962084400  0.87059844  0.96733840  1.00000000  -0.03433580855 -0.181418306   1.00000000 -0.09742795991             NA
pageRelevance   0.9660880268 -0.11274934 -0.06473892 -0.09742796  -0.00007664498  0.098667933  -0.09742796  1.00000000000             NA
stockIsLimited             NA          NA          NA          NA              NA           NA           NA             NA              1
```

Com es pot veure, la variable pageRelevance està correlacionada:

- Un -9.742796% amb la variable priceValue
- Un -18.141831% amb la variable stockValue
- Un -3.433581% amb la variable puntuationValue

La variable més correlacionada és l'estoc, després el preu i finalment la puntuació. Ho podem veure més al detall:

Relació de la pàgina amb el preu

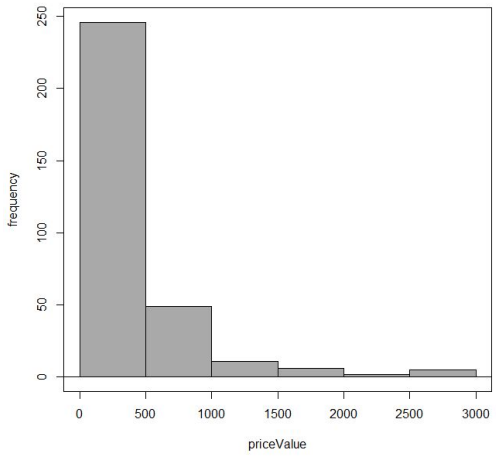
Si carreguem l'informació en funció de la rellevància de la pàgina:

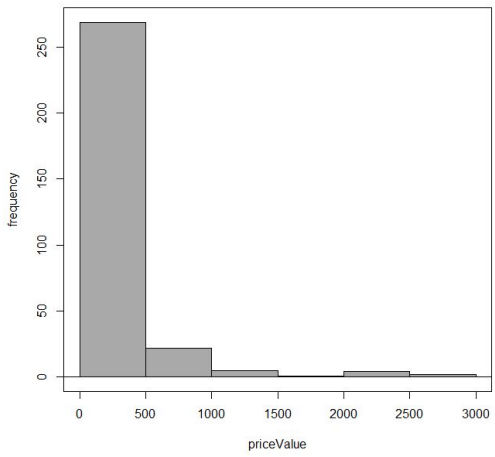
#Detall de les correlacions de preu
pageRelevance0 <- PhotoDataset[PhotoDataset$pageRelevance==0,]
pageRelevance1 <- PhotoDataset[PhotoDataset$pageRelevance==1,]
pageRelevance2 <- PhotoDataset[PhotoDataset$pageRelevance==2,]
pageRelevance3 <- PhotoDataset[PhotoDataset$pageRelevance==3,]
pageRelevance4 <- PhotoDataset[PhotoDataset$pageRelevance==4,

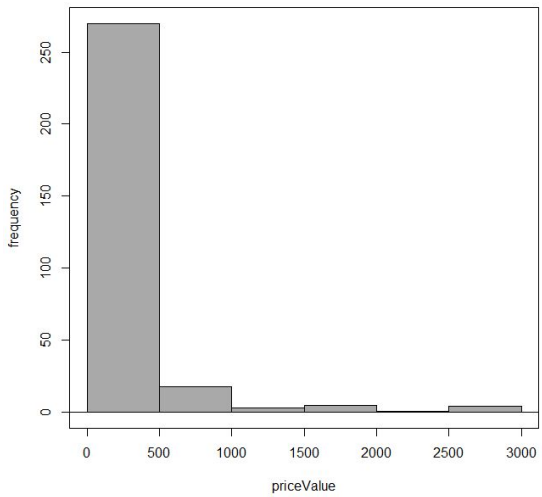
Podem veure clarament que el preu influeix a la posició seguint la nostra hipòtesi incial:

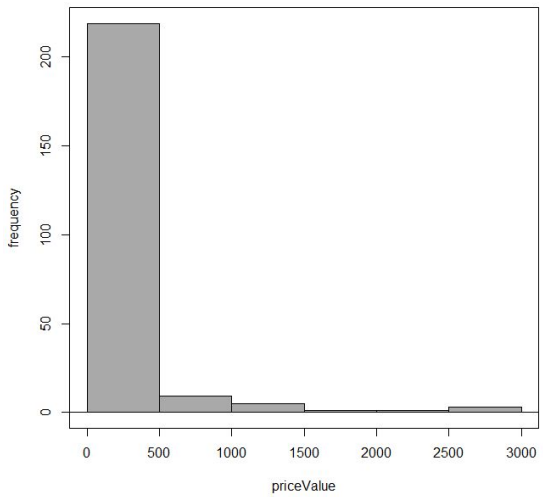
pageRelevance	0	1	2	3
Mitja (€)	360.8	204.02	202.63	184.17

A la següent gràfica veim la distribució de preus dels distints pageRelevance (ordenats d'esquerra a dreta i d'adalt abaix). Com es pot veure, els preus es mouen cap a l'esquerra a mida que s'incrementa la pàgina. Això vol dir que Amazon tendeix a considerar com a menys rellevants els productes més baratos.









Relació de la pàgina amb l'estoc

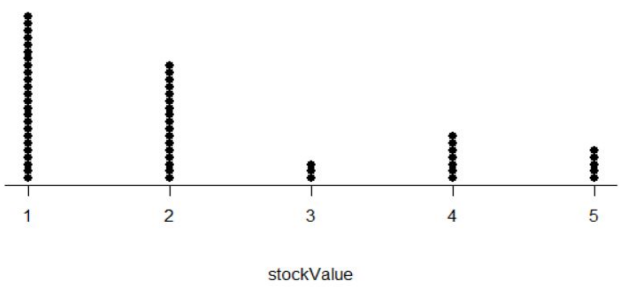
De les dues possibles hiòtesi:

1. Es mostren abans els productes amb estoc limitat, per a presionar els clients
2. Es dóna prioritat als productes amb estoc il·limitat, ja que hi ha més

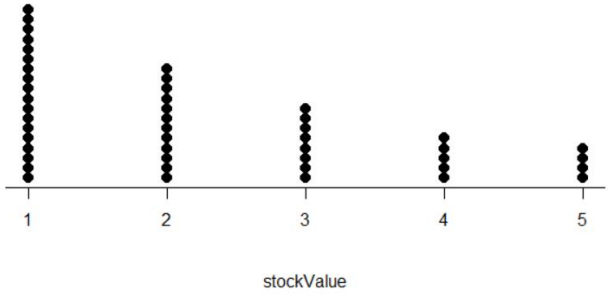
Podem comprovar que es dóna el primer cas si l'estoc està indica i el segon si no. A mida que creix el número de pàgina, augmenta l'estoc si està indicat, però també disminueix el nombre de NA.
En qualsevol cas, no sembla que pugui afectar; sembla que és més una conseqüència (els productes que es mostren primer, s'acaben abans)

pageRelevance	0	1	2	3
estoc	2.143	2.255	2.255	2.545
NA	263	256	250	194

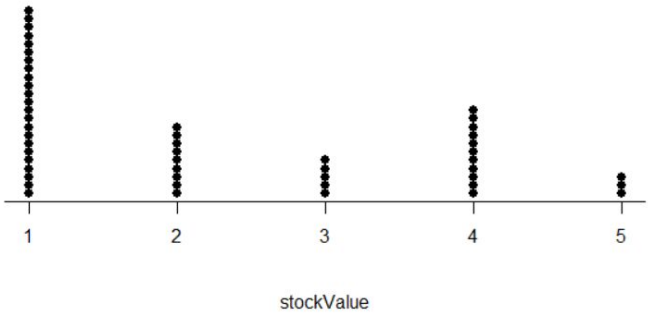
pageRelevance0



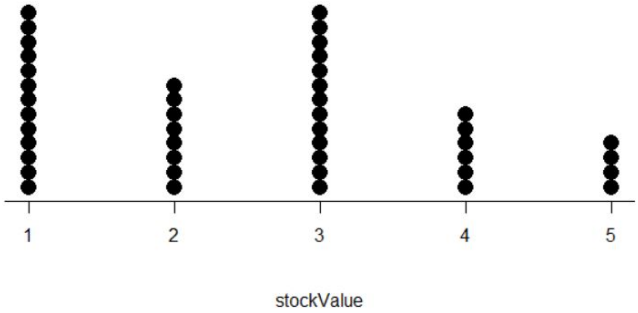
pageRelevance1



pageRelevance2



pageRelevance3



Relació de la pàgina amb la puntuació

No existeix cap tipus de relació entre la puntuació d'un producte i l'ordre en el que apareix. Els resultats per pageRelevance són pràcticament idèntics:

page	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
pageRelevance0	1.000	4.200	4.300	4.279	4.500	5.000
pageRelevance1	1.000	4.200	4.276	4.275	4.500	5.000
pageRelevance2	1.000	4.200	4.276	4.280	4.500	5.000
pageRelevance3	1.000	4.125	4.276	4.269	4.500	5.000

Podem afirmar que Amazon no té en compte la puntuació global d'un producte a l'hora de mostrar-ho. Això és una sorpresa relativa, ja que sembla que Amazon està més interessat en els perfils individuals (historial de navegació, productes comprats, etc.) que en els resultats globals.