Anàlisi comparativa de Preus de Supermercat: Un Conjunt de Dades Extret a Partir de Web Scraping

Pràctica 2

Enric Sintes Arguimbau i Carlos Romero Matarin

Tipologia i cicle de vida de les dades -UOC

Memòria del Projecte: Anàlisi Comparativa de Preus de Supermercat

Integrants del Grup:

- Carlos Romero Matarin
- Enric Sintes Arguimbau

Enllaç al Lloc Web Elegit:

- Dia: https://www.dia.es/
- Consum: https://www.consum.es/
- Mas: https://www.supermercadosmas.com/

Enllaç al repositori amb el codi de la pràctica:
https://github.com/cromeroUOC/Analysis-Web-Scraping-Supermarket

Arxiu comprimit del repositori:
https://drive.google.com/file/d/1uDqjNv4UttjopvlR3GdPh2yKikkazA4d/view?usp=drive_link

Enllaç al vídeo de presentació de la pràctica:

- Carpeta de la pràctica:

https://drive.google.com/drive/folders/1KMSxFahQQ_2Rrl2fB1GJYMs2SHtmiO9-?usp=sharing

- Vídeo de la pràctica:

https://drive.google.com/file/d/1v97rjlEaDY6o1ZpCFpQ5LBja6RcDM1BC/view?usp=drive_link

- Vídeo comprimit de la pràctica:

https://drive.google.com/file/d/1VChiB5Hq-Uny2OgeZEFauaJQA0RaJMen/view?usp=drive_link

Practica 2

Enric Sintes Arguimbau i Carlos Romero Matarin

2024-06-04

Contents

1 Descripció del dataset

El conjunt de dades amb el qual s'ha treballat en aquest document està compost per una col · lecció d'articles disponbiles en diferents supermercats, incloent-hi un conjunt de característiques associades a cada article. Per tant, és un dataset que pot cobrar una considerable importància en el sector del retail, ja que permet comparar les característiques i preus d'articles de diferents cadenes.

La importància d'aquest conjunt de dades rau en la seva capacitat per respondre a diverses preguntes que poden optimitzar les estratègies de negoci en aquest sector. Per altra banda, pot servir per millorar l'experiències del client a l'hora de fer la seva compra a, triant la cadèna de supermercats que més s'adeqüi a les necessitats de cada un, tant sigui pel preu de certs productes a cada una de les cadenes, pel surtit d'articles que hi ha o per les marques que es troben a cada una.
Així doncs, els principals análisis que es poden fer de les dades son, classificació de productes, anàlisis de preus o comparació de marques.

El conjunt de dades inclou diverses variables que descriuen les característiques de cada producte:

- Nombre: descriptiu del producte.
- Marca: marca comercial del producte.
- Precio: preu de venta al públic.
- Supermercat: cadèna de supermercats al que pertany l'article.
- URL: direcció web on es troba l'article.
- Fecha: dia d'extracció.
- Hora: hora d'extracció.
- unidad: indica el tipus d'unitats amb el que es serveix producte.
- precio_unidad: preu per unitat en €.
- Categoria: secció de supermercat a la que pertany el producte. Tipus d'article.
- Subcategoria: subsecció de supermercat a la que pertany.
- Estado: disponibilitat de l'article al supermercat en questió.

Tot plegat comporta un conjunt de dades de 12 columnes i 25625 files.

Aquest volum de dades permet un anàlisi detallada i comparativa de productes, cadenes de supermercats i marques.

2 Integració i selecció

S'ha optat per modificar i ampliar el dataset original utilitzat a la Pràctica 1 per a l'activitat d'integració. El desig principal d'aquesta ampliació es proporcionar una base de dades més completa i detallada és allò que justifica aquesta decisió. Podem veure que aquest conjunt de dades es capaç de suportar anàlisis més profundes i variades sobre el comportament dels preus entre diferents cadenes de supermercats.

Dir que la limitació de la varietat i la profunditat de les dades inicials fa que calgui ampliar la les dades original. Com un exemple ràpid dir que hem millorat significativament la qualitat de l'anàlisi en afegir característiques com "Categoria", "Subcategoria" i "Estat" als productes, cosa que permet segmentacions i comparacions més precises entre els productes. Podem dir que és especialment útil als estudis de mercat on la categorització dels productes pot afectar la percepció del preu i les decisions de compra dels consumidors.

Per fer l'ampliació del conjunt de dades i incloure noves variables com "Categoria", "Subcategoria" i "Estado", s'ha modificat el script original de Python utilitzat per a l'extracció de dades via web scraping. El que hem desenvolupat és que les modificacions recullin informació addicional de les pàgines de productes. El script actualitzat està disponible en el repositori GitHub del projecte.

Fem aquí la càrrega de les dades en el rmd i inicien les llibreries necessàries:

Descripció del dataset:

Com ja hem indicat el conjunt de dades de productos.csv té 25.625 entrades, cadascuna de les quals representa un sol producte disponible per a la compra. Aquestes dades ofereixen una visió general del mercat actual en

incloure cinc característiques clau: nom, marca, preu, supermercat, URL, data d'obtenció, hora d'obtenció, unitat de mesura, preu unitari de mesura, categoria, subcategoria i estat.

Primeres Cinc Files:

El quadre de dades comença amb una varietat d'articles, com detergents, lleixiu, maquinetes d'un sol ús, cera per a terra, etc. Aquesta primera mostra la varietat de productes disponibles al supermercat Consum, amb informació detallada sobre la marca, el preu i enllaços directes als productes, també podem veure com hi ha registres sense categoria.

```
## # A tibble: 6 x 12
##   Nombre  Marca Precio Supermercado URL   Fecha      Hora   unidad precio_unidad
##   <chr>   <chr> <chr>  <chr>        <chr> <date>     <time> <chr>  <chr>
## 1 Deterg~ PERL~ 3,15 € Consum       http~ 2024-06-04 45'53" 1 Lv   0,13 €
## 2 Lejía ~ CONE~ 2,69 € Consum       http~ 2024-06-04 45'56" 1 L    1,35 €
## 3 Maquin~ GILL~ 2,89 € Consum       http~ 2024-06-04 45'57" 1 U    0,58 €
## 4 Cera S~ ASEVI 2,19 € Consum       http~ 2024-06-04 45'58" 1 L    2,19 €
## 5 Antica~ CALG~ 8,99 € Consum       http~ 2024-06-04 45'59" 1 U    0,60 €
## 6 Filtro~ MELI~ 2,45 € Consum       http~ 2024-06-04 46'00" 1 U    0,06 €
## # i 3 more variables: Categoria <chr>, Subcategoria <chr>, Estado <chr>
```

Descripció Estadística:

Podem veure que hi ha 24.399 noms de producte diferents, destacant la diversitat i amplada d'assortiment disponible.

Dir que encara que la majoria de marques són identificables, un total de 2.711 productes són etiquetats com a "Marca no disponible", reflectint possibles deficiències o inconsistències en la recopilació de dades. A més, 651 preus estan etiquetats com "no disponibles", indicant que aquestes entrades poden necessitar una verificació o actualització de dades.

Amb més cal dir que hi ha 7.031 preus per unitat únics i preus que varien àmpliament, per això podem dir que el conjunt de dades ofereix dades per a l'anàlisi de tendències de preus i comportaments de compra dins del context dels tres supermercats inclosos, destacant especialment Consum, on s'han registrat la majoria de les entrades.

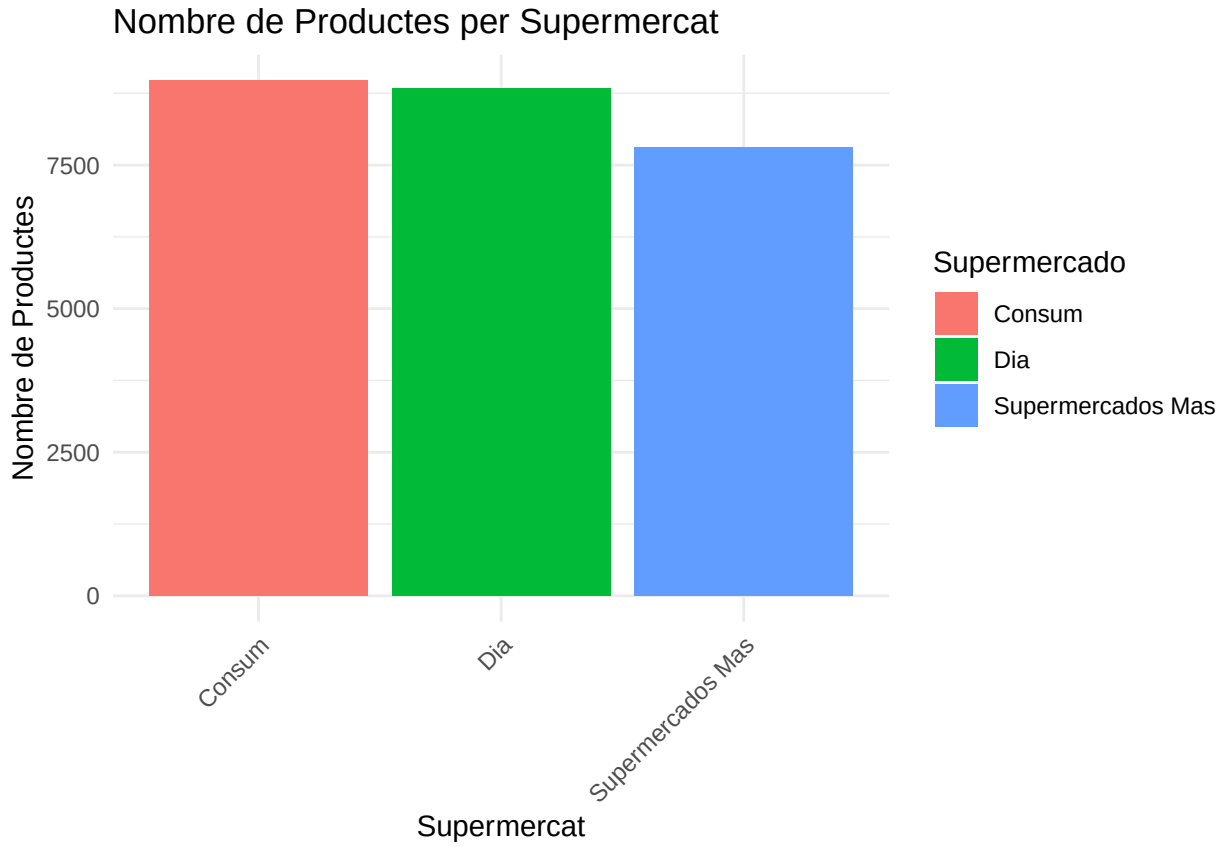
```
## productos
##
##  12  Variables      25625  Observations
## ---------------------------------------------------------------------------
## Nombre
##        n  missing distinct
##    25625        0    24399
##
## lowest : 100 Mejores Suplemento Alex Yañez             100% Real Whey Protein Cookies and Cream
## highest: Zumo Uva,Melocotón y Manzana Brik             Zumo Veggie Calabaza Zanahoria-Mango-Manza
## ---------------------------------------------------------------------------
## Marca
##        n  missing distinct
##    25159      466     3881
##
## lowest : 18/70       18/70 RUBIA 1880        1881        1906
## highest: ZERO        ZESPRI      ZIPPER      ZOCO        ZUM
## ---------------------------------------------------------------------------
## Precio
##        n  missing distinct
##    25625        0     2053
##
```

```
## lowest : 0,14 €               0,22 €               0,24 €               0,25 €               0,26 €
## highest: 9.99 €               95.75                99                   99,00 €              Precio r
## ---------------------------------------------------------------------------
## Supermercado
##        n  missing distinct
##    25625        0        3
##
## Value                 Consum              Dia Supermercados Mas
## Frequency               8972              8840              7813
## Proportion             0.350             0.345             0.305
## ---------------------------------------------------------------------------
## URL
##        n  missing distinct
##    25625        0    25625
##
## lowest : https://tienda.consum.es/es/p/-cafe-extra-intenso-20-capsulas/7407695              https://
## highest: https://www.supermercadosmas.com/vino-rioja-tinto-resera-azpilicueta-reserva-75cl/ https://w
## ---------------------------------------------------------------------------
## Fecha
##          n    missing   distinct       Info       Mean        Gmd
##      25625          0          1          0 2024-06-04          0
##
## Value      2024-06-04
## Frequency       25625
## Proportion          1
## ---------------------------------------------------------------------------
## Hora [secs]
##        n  missing distinct
##    25625        0    16599
##
## lowest : 00:45:53 00:45:56 00:45:57 00:45:58 00:45:59
## highest: 06:54:00 06:54:01 06:54:02 06:54:03 06:54:04
## ---------------------------------------------------------------------------
## unidad
##        n  missing distinct
##    25625        0     2337
##
## lowest : -30%AZUCAR 400GR 1-10 2C 1RX6     1 Dc              1 Do              1 Kg
## highest: Venta al peso    VF R070 P3+1X52G WHITE 400G       XXL BOLSA 260GR   ZERO LATA 33CL
## ---------------------------------------------------------------------------
## precio_unidad
##        n  missing distinct
##    25625        0     7031
##
## lowest : 0,00 € / Und.                  0,01 €                         0,01 €
## highest: 99,75 € / Kgr.                 99,80 €                        99,88 € / Litro
## ---------------------------------------------------------------------------
## Categoria
##        n  missing distinct
##    16653     8972      494
##
## lowest : 2ª al 50%                         Accesorios para el pelo            Accesorios para mani
## highest: Vodka                             Whisky                             Yogures y postres
## ---------------------------------------------------------------------------
```

```
## Subcategoria
##        n  missing distinct
##    16653     8972     7963
##
## lowest : Abono universal masso 1l                                      Abrillantador lavavajillas finish 5
## highest: Zumo refrigerado platano/coco/avena via nature bot 750ml Zumo refrigerado sin pulpa de naran
## ----------------------------------------------------------------------------------------------------
## Estado
##        n  missing distinct
##    25625        0        3
##
## Value                   Agotado              Añadir Estado no disponible
## Frequency                  2191                22939                  495
## Proportion                0.086                0.895                0.019
## ----------------------------------------------------------------------------------------------------
```

Nombre de productes per supermercat:

El gràfic mostra que el supermercat amb més productes registrats és "Consum" (8.972), seguit de prop per "Dia" (8.840) i "Supermercados Mas" (7.813). Segons aquesta distribució, "consum" i "dia" són els principals proveïdors de productes de la base de dades.



3 Neteja de dades

3.1 Valors faltants

En primer lloc s'ha identificat aquells valors de les variables que identifiquen valors faltants i s'ha trobat que a cada variable es registre un cadena de caràters que descriu que hi ha un valor faltant. Per exemple, a la variable `Nombre` s'han registrat els valors faltants com `Nombre no disponible`, a `Marca` `Marca no disponible`, i així successivament. Tots aquests valors s'han substituït per `NA` per poder tractar les variables correctament.

3.2 Tipus de dades

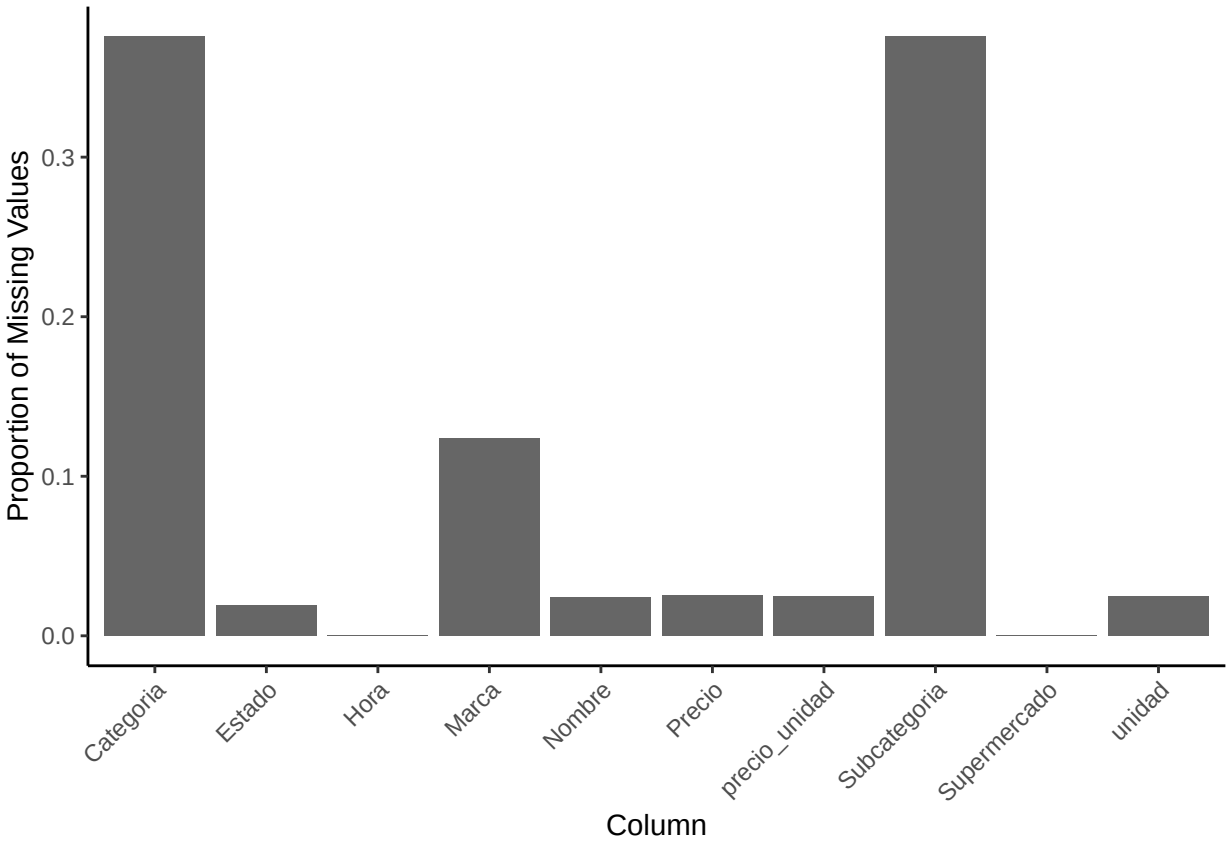
A continuació, s'han corregit totes les variables que contenien caràcters especials i que no permeten tractar les variables així com és degut. Per exemple, la variable `Precio` conté el signe € i per tant no permet tractar la variable com un valor numèric. Llavors:

- `Nombre`: cadena de caràcters.
- `Precio`: s'ha extret el signe €, els espais que hi podria haver i s'ha canviat les , dels decials per punts `.`. D'aquesta manera s'ha pogut transformar la variable a numèric.
- `unidad`: s'ha convertit a factor i s'han unificat les categories que significaven el mateix. Per exemple: `1 U`, `Und.` i `UNIDAD` s'ha unificat a un sol nivell `UNIDAD`.
- `precio_unidad`: s'ha fet el mateix tractament que la variable `Precio` extraient signes i canviar les , dels decimals per punts. També s'ha trobat algún cas que s'utilitzava el `.` com a separadors del milers i també s'ha eliminat. Transformant posterior a valor numèric.
- `Marca`: variable categòrica, per tant s'ha trasnformat a factor.
- `Supermercado`: transformat a factor.
- `Hora`: variable de tipus temps.
- `Fecha`: variable de tipus data.
- `Caregoría`: vairable categòrica, per tant, factor.
- `Subcategoria`: conversió a factor.
- `Estado`: conversió a factor.

Un cop transformades les dades correctament, s'estudien altres tipus de valors que poden significar pèrdua de dades. Com per exemple, valors de la variable `Precio` o `precio_unitario` que tinguin valors 0 o negatius. Entre aquests s'han trobat dos casos que tenen un `precio_unidad` igual a 0, degut al valor tant reduït que té cada unitat d'aquests productes. Per evitar que aparegui el valor 0, es calcularà manualment el valor real dividint el preu per les unitats de cada article.

3.3 Tractament de valors faltants

Un cop corregits els errors, hi ha que revistar i evaluar els valors faltants `NA` de cada una de les variables.



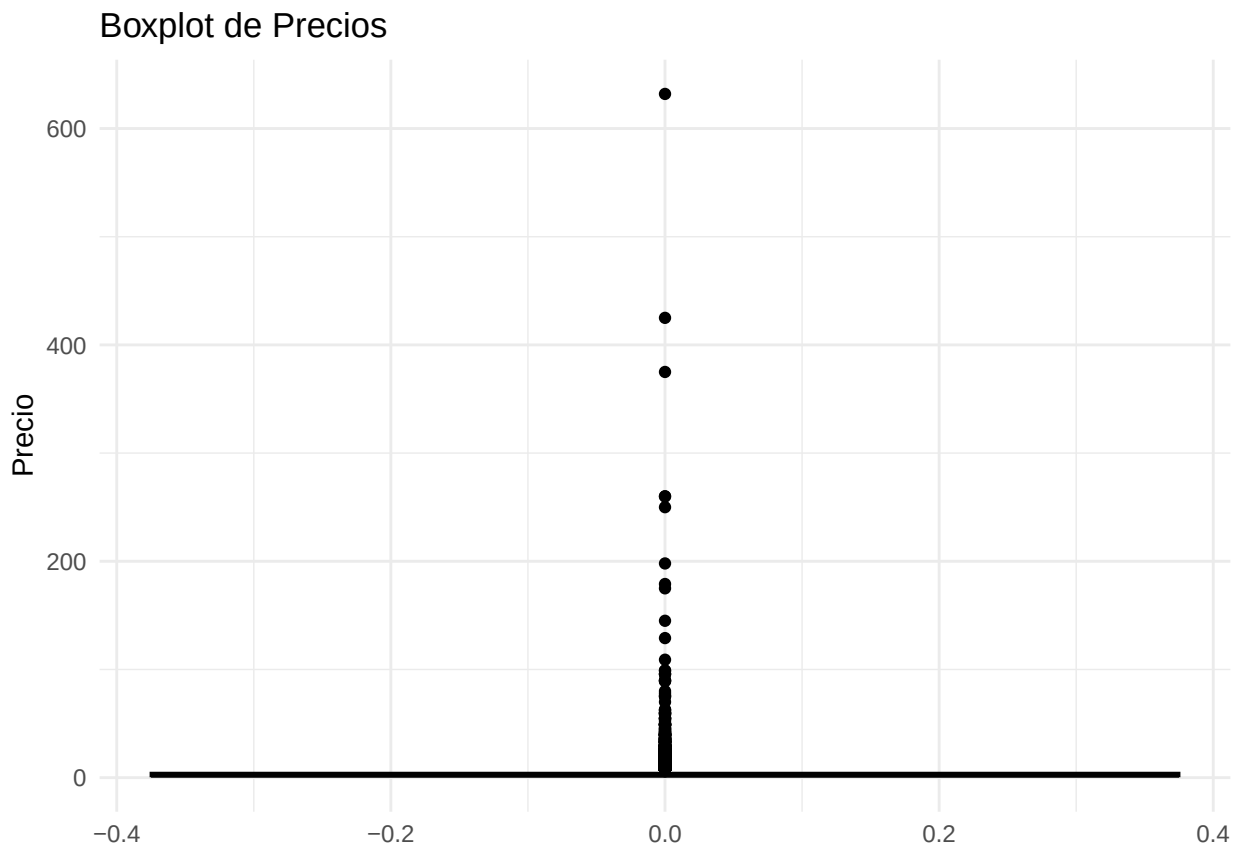
Les variables `Marca`, `Categoría` y `Subcategoría`, tenen una proporció de valors faltants elevats. És evident que si s'esborressin tots els casos faltants d'aquestes variables es perdrien molts casos i molta informació de valor per l'anàlisi posterior. Per tant, s'ha cercat una alternativa per evitar l'imputació de tots aquests valors. Aquesta alternativa consisteix en utilitzar un algoritme per completar les seccions faltants a partir del nom de cada producte. És a dir, a partir de les categories existents i dels noms dels productes d'aquestes categories, es fa una estimació de quina categoria pot ser la més adecuada per cada article li falta aquest valor. Exactament, el que es fa és, per un producte sense categoria i subcategoria, cercar els productes amb categoria i subcategoria conegudes amb més paraules coincidents i assigna la categoria i subcategoria més freqüents. Així i tot, hi ha productes que no s'ha pogut classificar correctament, però tot i així s'ha reduït considerablement els valors faltants.

Respecte a les altres variables es ronda un 2% de valors faltants i possiblament siguin els mateixos casos. Per tant, imputarem tots els valors que no tenen un registre ni a la varaibles `Nombre`, `Precio`, `unidad` ni `precio_unidad`, ja que no aporten cap valor analític. Aquests casos signifiquen exactament 0.0244 de les files.

3.4 Valors extrems

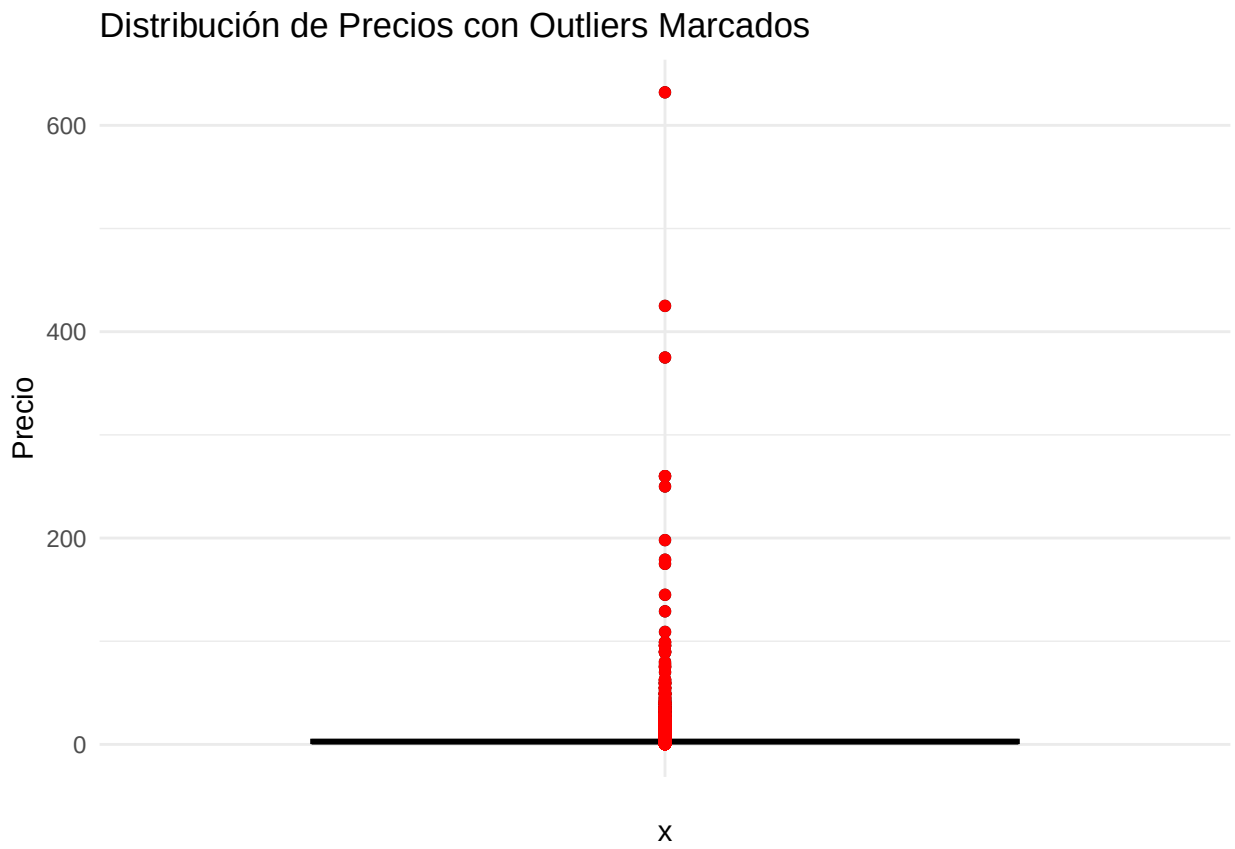
Per estudiar els valors extrems s'estudien els valors numèrics i estudiarem els valors que poden distorsionar els estudis analítics. Aquesta variable és: `Precio`.

El primer pas és estudiar la seva distribució, que es pot valorar tant amb un diagrama de barres com amb un boxplot, que és on es veuen millor els outliers.

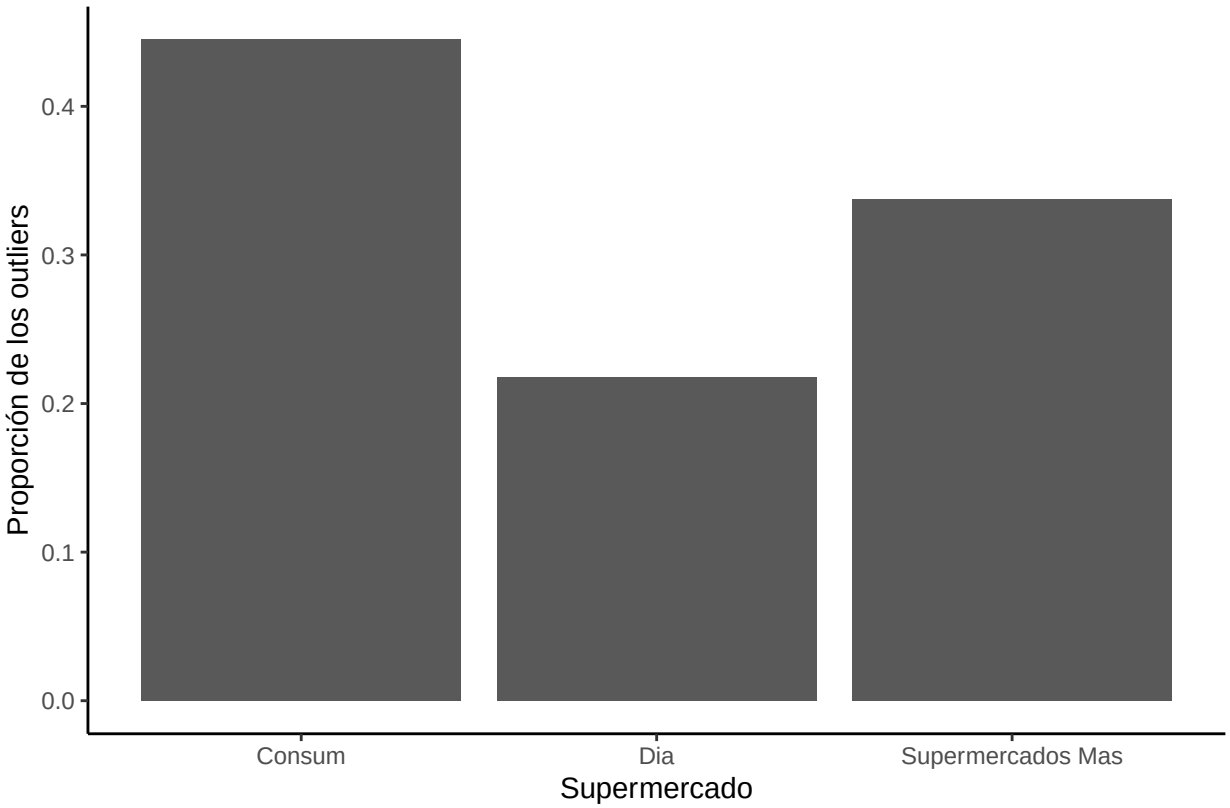


Es pot veure com la variable `Precios` sí té valors extrems que poden distorsionar les dades.

Un cop visualitzats els valors extrems, els podem detectar considerant que un outlier son tots els punts que es troben a una distància del primer o tercer quartil, més gran que 1.5 cops el rang interquantílic. D'aquesta manera, al tenir identificats els valors extrems, podem eliminar-los.



En total s'han trobat 2603 del quals es reparteixen entre els supermercats de la següent manera:

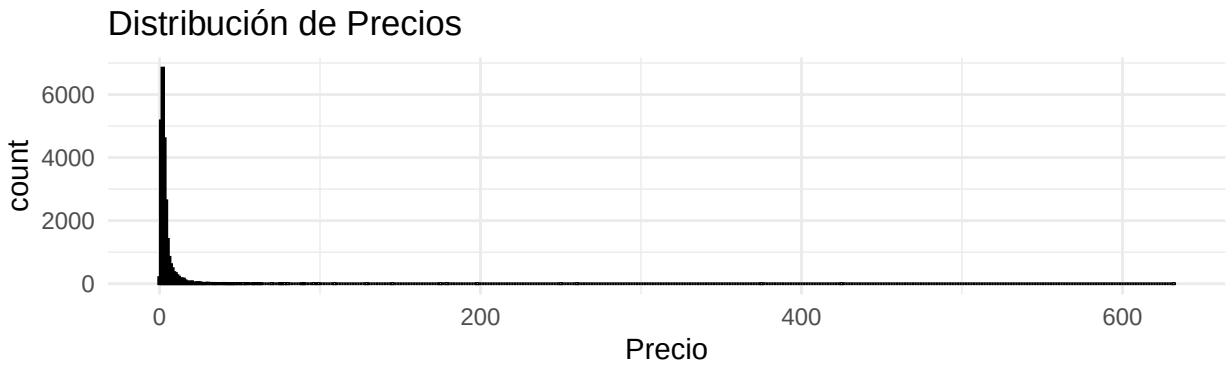


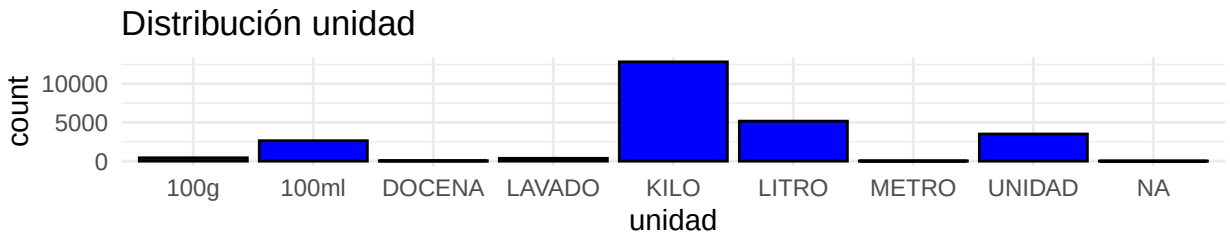
Distribución de los outliers entre los supermercados

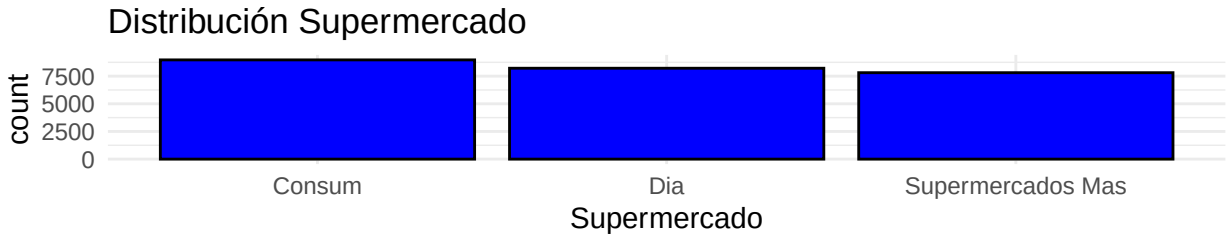
Per tant, es por veure que quasi la meitat dels outliers trobats es troben en el supermercat Consum, mentre que a Dia, és on hi ha menys freqüència de outliers. Això no vol dir que Consum sigui el supermercat més car, sinó que té més articles de preu elevat.

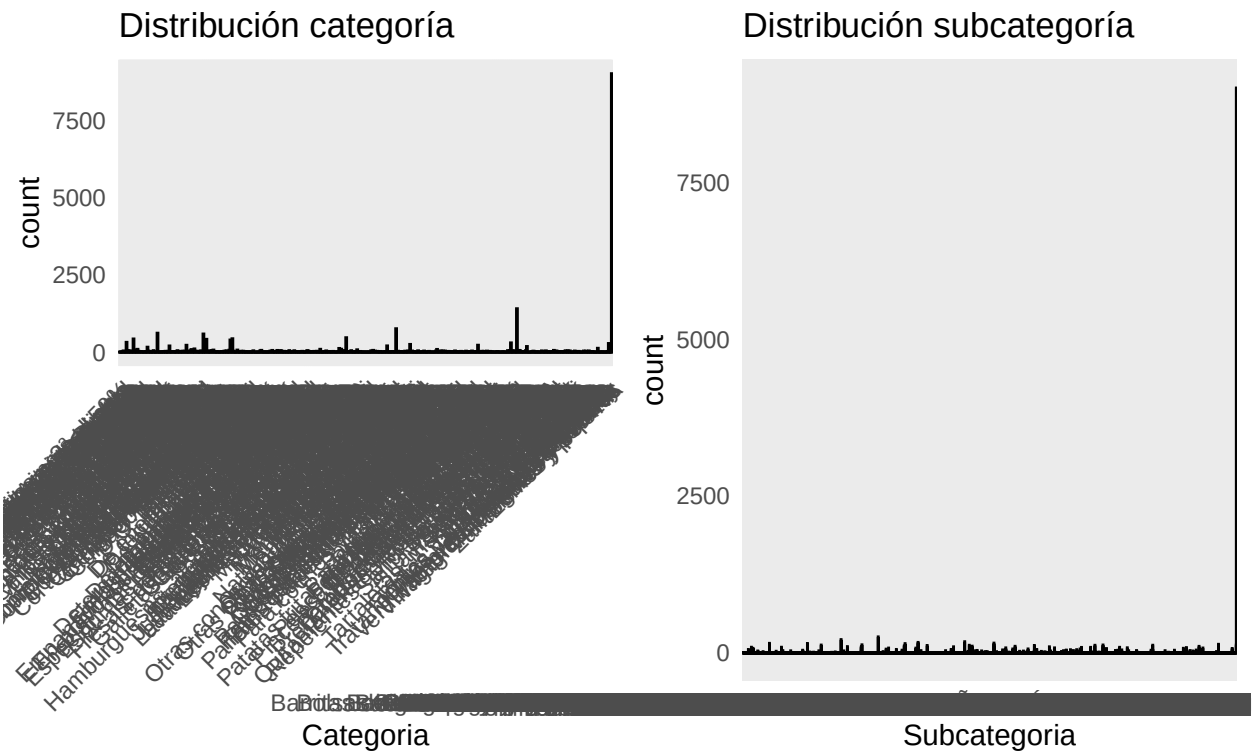
Tot i així, resvisant els valors extrems, considerem que no és necessari corregir ni eliminar-los, ja que no es tracten d'errors d'entrada, sinó que es tracta d'articles de més alta qualitat que tenen un preu alt, però formen part de l'assortiment normal d'un supermercat. Així doncs, formaràn part de les dades a analitzar.

3.5 Representació variables









4 Anàlisi de les dades

4.1 Models

4.1.1 Model supervisat

Basant-nos en les variables de unitat i supermercat, hem utilitzat un model de regressió lineal per predir el preu del producte a la secció de modelització supervisada. El model que hem definit cerca comprendre com aquests elements afecten el preu final del producte. Per si de cas s'ha eliminat la variable Preu els caràcters especials (€) i s'ha convertit de text a numèric.

```
# Asegurando que las variables están en formato adecuado
productos$Precio <- as.numeric(gsub(" €", "", gsub(",", ".", productos$Precio)))

# Modelo de regresión lineal donde predecimos 'Precio' basado en otras características:
modelo_lineal <- lm(Precio ~ unidad + Supermercado, data = productos_actualizados)
summary(modelo_lineal)
```

```
##
## Call:
## lm(formula = Precio ~ unidad + Supermercado, data = productos_actualizados)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.08  -2.30  -1.16   0.28 628.07
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.2493     0.3941   5.707 1.16e-08 ***
## unidad100ml                    3.3384     0.4216   7.919 2.50e-15 ***
```

```
## unidadDOCENA                    0.3366     1.3330   0.253  0.80064
## unidadLAVADO                    4.3098     0.5772   7.466 8.53e-14 ***
## unidadKILO                      1.3068     0.3974   3.288  0.00101 **
## unidadLITRO                     2.5454     0.4066   6.260 3.92e-10 ***
## unidadMETRO                     0.4068     1.6586   0.245  0.80627
## unidadUNIDAD                    2.5403     0.4137   6.140 8.38e-10 ***
## SupermercadoDia                -0.8466     0.1236  -6.848 7.66e-12 ***
## SupermercadoSupermercados Mas   0.3721     0.1255   2.964  0.00304 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.059 on 24957 degrees of freedom
##   (34 observations deleted due to missingness)
## Multiple R-squared:  0.01406,    Adjusted R-squared:  0.01371
## F-statistic: 39.56 on 9 and 24957 DF,  p-value: < 2.2e-16
```

Els resultats obtinguts ens indiquen que diferents unitats i els supermercats tenen un impacte estadísticament significatiu sobre el preu. Per exemple, els coeficients per a les unitats com '100ml' i 'LAVADO' mostren increments notables en els preus, cosa que suggeriria que aquestes unitats tendeixen a estar associades amb preus més alts.

Cal dir que , els supermercats també juguen un paper crucial en la determinació dels preus. En particular, els productes del supermercat Dia tendeixen a ser més barats en comparació amb la base. Aquest fet es reflecteix en el coeficient negatiu per a Dia. En contrast, els Supermercats Mas mostren un petit increment en els preus comparat amb el supermercat de referència.

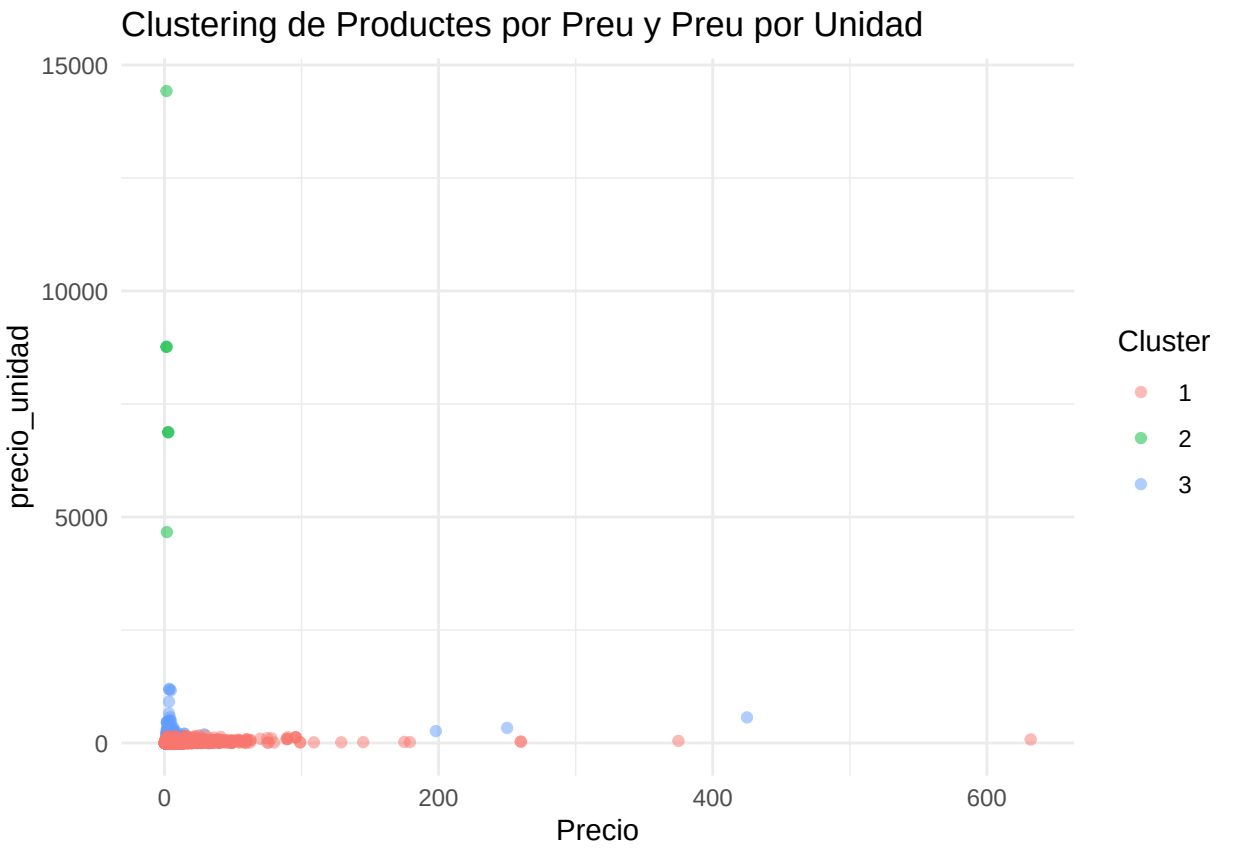
Podem veure que rl model té un R-quadrat ajustat relativament baix, cosa que indica que les variables incloses expliquen només una petita part de la variabilitat dels preus dels productes.

4.1.2 Model no supervisat

Hem desarrollat l'algorisme k-means per agrupar els productes en funció dels seus preus i preus per unitat per a l'anàlisi no supervisada. Primer per si de cas els preus s'han netejat i convertit en un format numèric. Despues l'agrupació esta definida en tres grups.

```r
productos$Precio <- as.numeric(gsub(" €", "", gsub(",", ".", productos$Precio)))
productos$precio_unidad <- as.numeric(gsub("€", "", gsub(",", ".", productos$precio_unidad)))

productos <- productos %>%
  filter(!is.na(Precio) & !is.na(precio_unidad) &
         !is.nan(Precio) & !is.nan(precio_unidad) &
         Precio != Inf & Precio != -Inf &
         precio_unidad != Inf & precio_unidad != -Inf)

set.seed(123)

# Ejecutar k-means con 3 centros
kmeans_resultado <- kmeans(productos[, c("Precio", "precio_unidad")], centers = 3)

# Agregar los clusters al dataframe para análisis posterior
productos$Cluster <- as.factor(kmeans_resultado$cluster)

# Visualizar los resultados de k-means

ggplot(productos, aes(x = Precio, y = precio_unidad, color = Cluster)) +
  geom_point(alpha = 0.5) +
```

```
  labs(title = "Clustering de Productes por Preu y Preu por Unidad") +
  theme_minimal()
```



Com odem veure en es el gràfic els productes han estat dividits en tres clústers diferents, cada un representat per un color diferent:

- Clúster 1 (Color vermell): Aquest grup conté productes amb preus més baixos i preus per unitat relativament baixos. Podem dir que son els productes bàsics o de consum diari més assequibles.
- Clúster 2 (Color blau): Els productes d'aquest grup semblen tenir preus més elevats, però encara amb preus per unitat baixos, podem dir que podria incloure articles venuts en major quantitat o en format a granel.
- Clúster 3 (Color verd): Aquí veiem pocs productes que tenen preus per unitat significativament alts. Podrien ser productes especialitzats o de luxe.

4.2 Prueba de Contraste de Hipótesis

Hem dut a terme una prova de contrast d'hipòtesis, enfocant-te específicament en els que contenien "Jamon" en nom seu. El procés inclou la neteja de dades, la conversió del preu de text a números i la creació d'una variable per determinar si el producte conté "Jamon" en el nom.

```
productos$Precio <- as.numeric(gsub(" €", "", gsub(",", ".", productos$Precio)))  # Limpieza de datos

# Agregamos una nueva columna para identificar si el nombre del producto contiene 'Jamon'
productos <- productos %>%
  mutate(contiene_Jamon = ifelse(grepl("Jamon", Nombre, ignore.case = TRUE), "Con Jamon", "Sin Jamon"))
```

```r
# Calculamos el precio medio general excluyendo NA
precio_medio_general <- mean(productos$Precio, na.rm = TRUE)

# Subconjunto de productos que contienen 'Jamon' en el nombre
productos_con_Jamon <- productos %>%
  filter(contiene_Jamon == "Con Jamon") %>%
  pull(Precio)

# Prueba de normalidad
shapiro_test_Jamon <- shapiro.test(productos_con_Jamon)
print(shapiro_test_Jamon)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  productos_con_Jamon
## W = 0.35624, p-value < 2.2e-16
```

```r
# Dependiendo de la normalidad, realizar t-test o Wilcoxon test
if (shapiro_test_Jamon$p.value > 0.05) {
  t_test_result <- t.test(productos_con_Jamon, mu = precio_medio_general)
  print(t_test_result)
} else {
  print("Distribución no normal, aplicando prueba no paramétrica.")
  wilcox_test_result <- wilcox.test(productos_con_Jamon, mu = precio_medio_general, alternative = "grea
  print(wilcox_test_result)
}
```

```
## [1] "Distribución no normal, aplicando prueba no paramétrica."
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  productos_con_Jamon
## V = 844, p-value = 0.9989
## alternative hypothesis: true location is greater than 4.055195
```

Podem veure a partir dels resultats de la prova de Shapiro-Wilk, s'ha determinat que la distribució dels preus dels productes que contenen "Jamon" no és normal, ja que el p-valor és significativament menor que 0.05. Per això ens portat a l'aplicació d'una prova no paramètrica, específicament el test de Wilcoxon, en comptes d'un t-test que requereix normalitat en les dades on el p-valor de 0.9989, suggerint que no hi ha evidència estadísticament significativa per rebutjar la hipòtesi nul · la que el preu mitjà dels productes que contenen "Jamon" és major que el preu mitjà general dels productes.

5 Resolució del problema:

Podem dir que hem arribat a diverses conclusions importants a partir de l'anàlisi i els resultats de la investigació del conjunt de dades dels preus dels supermercats que ens permeten abordar els primers problemes:

Conclusions:

- Variabilitat de preus entre cadenes: Podem veure que segons la cadena de supermercats, hi ha diferències significatives en el preu d'un producte. En particular dir que, els preus del supermercat Dia han estat generalment més baixos que els de Consum i Supermercats Mas.
- Impacte de les unitats de mesura en el preu: Podem veure que segons el model de regressió lineal, les unitats de mesura com a litres i quilograms tenen un impacte significatiu en el preu d'un producte. I

podem dir que els preus dels productes que es mesuren per unitats grans solen ser més alts, cosa que indica una estratègia de preus basada en el volum de venda.

- Segmentació de productes per preu: Podem veure ham la utilització de l'algorisme k-means per agrupar productes, que s'han identificat grups clars basats en el preu i el preu per unitat. Aquí es mostra la diversitat de l'oferta de les cadenes de supermercats i pot ajudar els clients a identificar quins supermercats ofereixen els millors preus per a productes específics o en grans quantitats.
- Normalitat i distribució de preus: I com punt final podem véreu que les proves de contrast d'hipòtesis han demostrat que la distribució de preus dels articles amb Jamon no segueix una distribució normal, cosa que ha requerit l'ús de proves no paramètriques. Aquestes proves han demostrat que no hi ha diferències significatives en el preu mitjà dels productes amb "Jamon" en comparació del preu mitjà general.
- Per concloure dir que els resultats de l'anàlisi demostren la utilitat dels models estadístics i les tècniques de mineria de dades en la comparació dels preus dels supermercats, proporcionant solucions clares i útils al problema plantejat.

6 Codi

El codi font desenvolupat per a la neteja, anàlisi i representació de dades ha estat principalment desenvolupat en R. Ja escollit per la seva flexibilitat, poder estadístic i la disponibilitat d'una àmplia biblioteca de paquets específics per a la manipulació de dades i visualització gràfica.

En elcodi en R hem utilitzat diverses llibreries crucials per al processament i anàlisi de dades com dplyr per a la manipulació de dades, ggplot2 per a la visualització, i readr per a la càrrega eficient de dades. Dir que per la neteja de dades, s'han implementat rutines específiques que tracten amb valors faltants, correcció de formats i la normalització de les cadenes de text. Aquestes tasques són necessàries per assegurar la qualitat i la consistència dels conjunts de dades amb els quals s'analitzen.

Podem veure també que en el codi per a l'anàlisi de les dades hem desenvolupat models estadístics com regressions lineals i algoritmes de clustering (k-means). Méstard hem fet les visualitzacions dels resultats d'anàlisis en difuntes tipus de gràfics.

Dir que el codi font està disponible de manera pública en el següent enllaç del repositori GitHub, que conté tots els scripts utilitzats així com documentació addicional sobre l'estructura i l'ús dels codis:

https://github.com/cromeroUOC/Analysis-Web-Scraping-Supermarket

7 Referències

- RPubs - Clase 6 - Limpieza de datos. (n.d.). https://rpubs.com/camilamila/limpieza
- kmeans function - RDocumentation. (n.d.). https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans
- RPubs - Introducción a los Modelos de Agrupamiento en R. (n.d.). https://rpubs.com/rdelgado/399475
- RPubs - Regresión lineal simple. (n.d.). https://rpubs.com/joser/RegresionSimple
- RPubs - Market Basket Analysis on Groceries Data. (n.d.). https://rpubs.com/Handedemirci/1011420

8 Taula de Contribucions

Table 1: Resum de Contribucions

Contribucions	Signatura
Investigació prèvia	CRM, ESA
Redacció de les respostes	CRM, ESA
Desenvolupament del codi	CRM, ESA
Participació al vídeo	CRM, ESA

CRM: Carlos Romero Matarin ESA: Enric Sintes Arguimbau