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.
- Hora: hora d'extracció de la dada.
- 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

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:

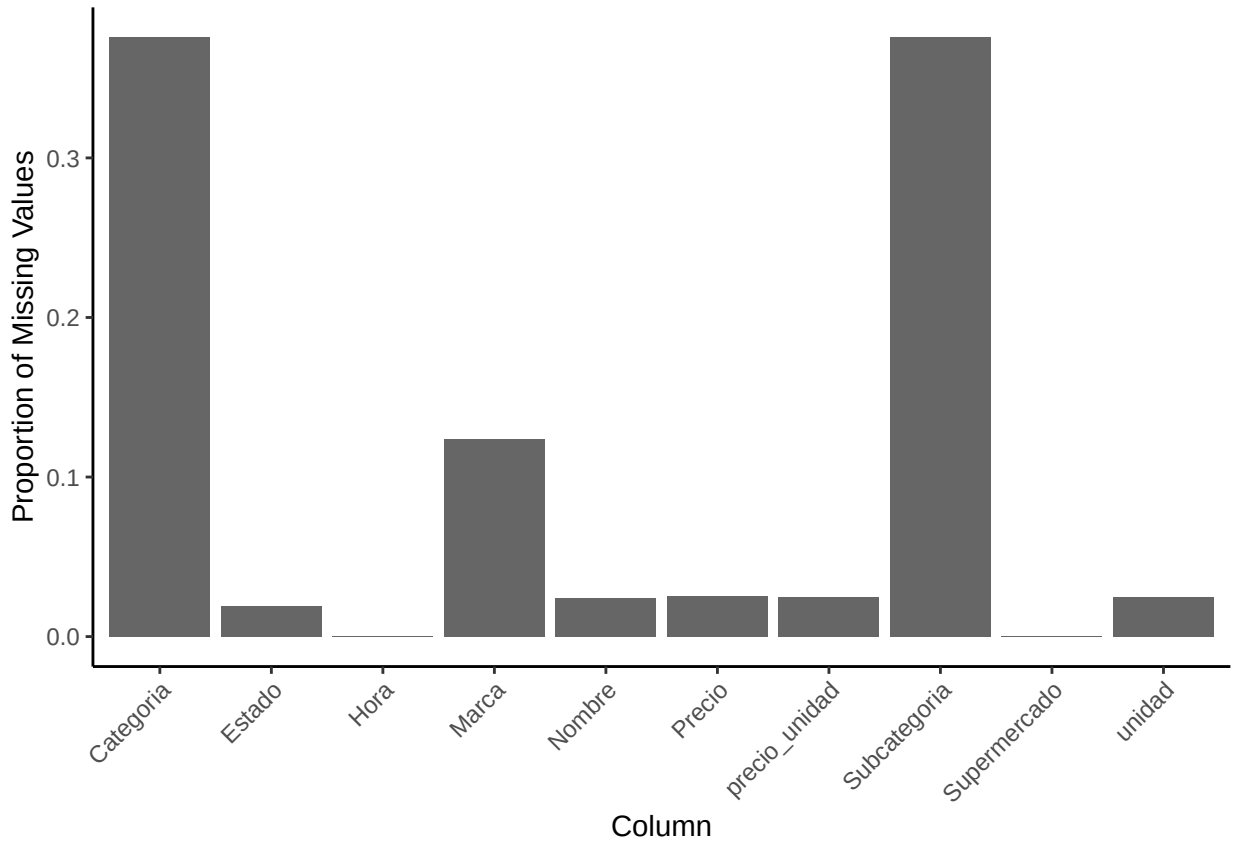
- `Nombre`: cadena de caràcters.
- `Precio`: s'ha extret el signe €, els espais que hi podria haver i s'ha canviat les , dels decials per punts .. D'aquesta manera s'ha pogut transformar la variable a numèric.

- `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.
- `Supermercat`: 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

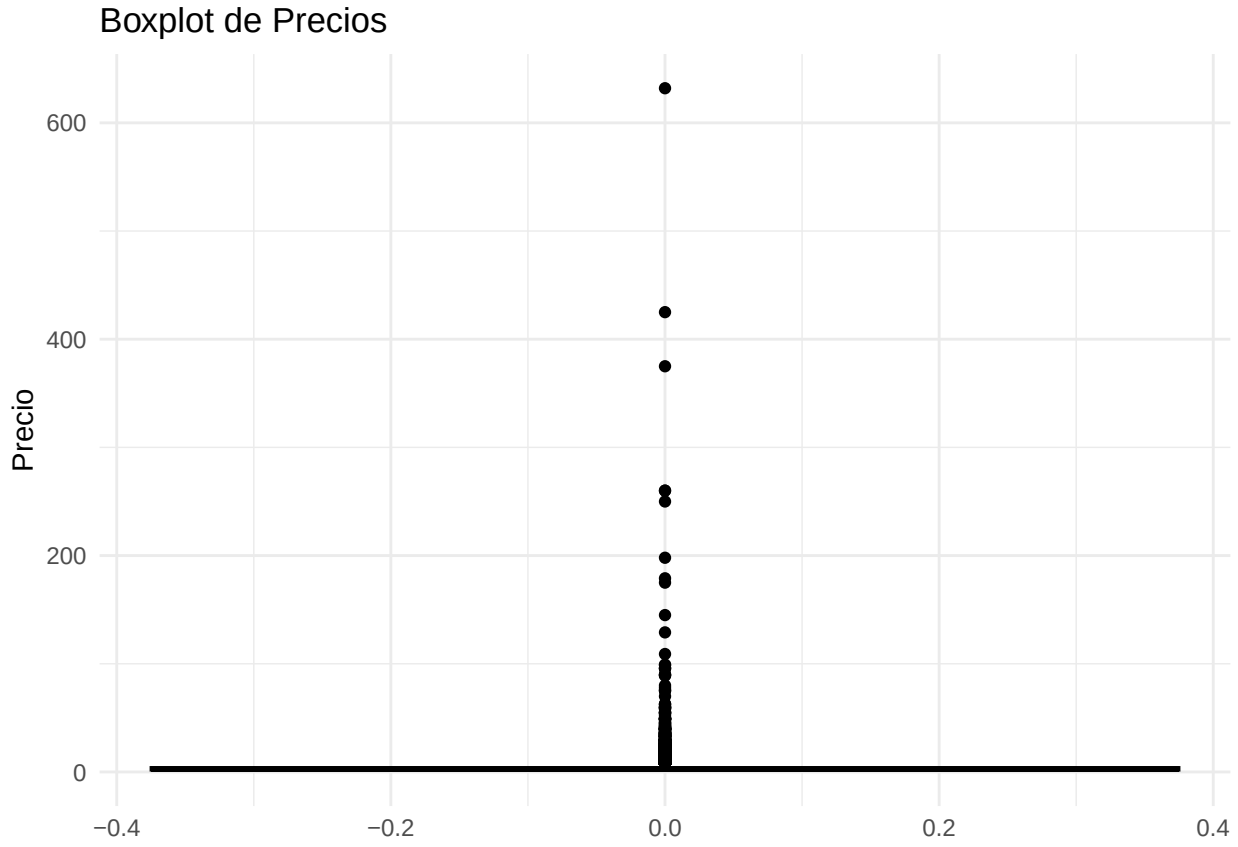
nom de cada producte. És a dir, a partir de les categories existents i dels noms dels productes d'aquestes categories, es fa una estimació de quina categoria pot ser la més adecuada per cada article li falta aquest valor. Exactament, el que es fa és, per un producte sense categoria i subcategoria, cercar 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

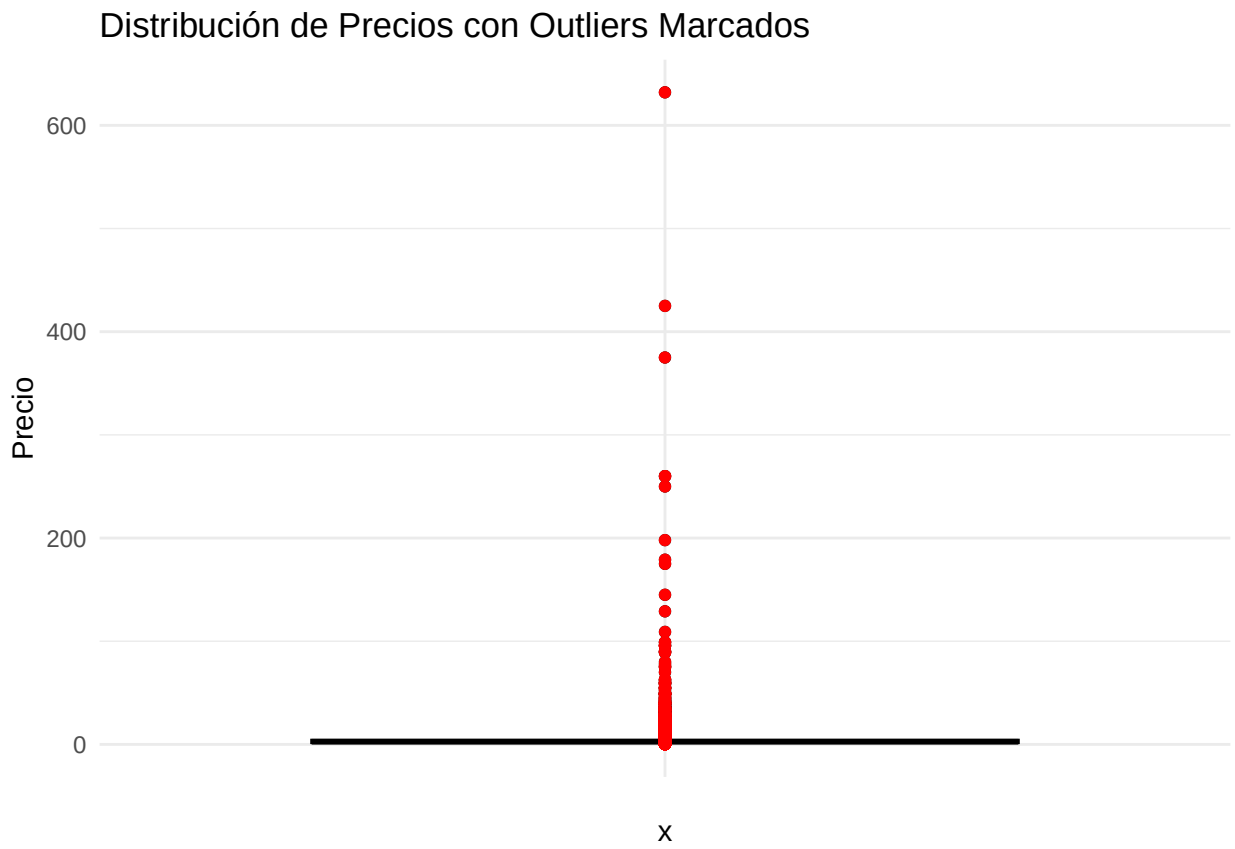
Per estudiar els valors extrems s'estudien els valors numèrics i estudiarem els valors que poden distorsionar els estudis analítics. Aquesta variable és: `Precio`.

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.

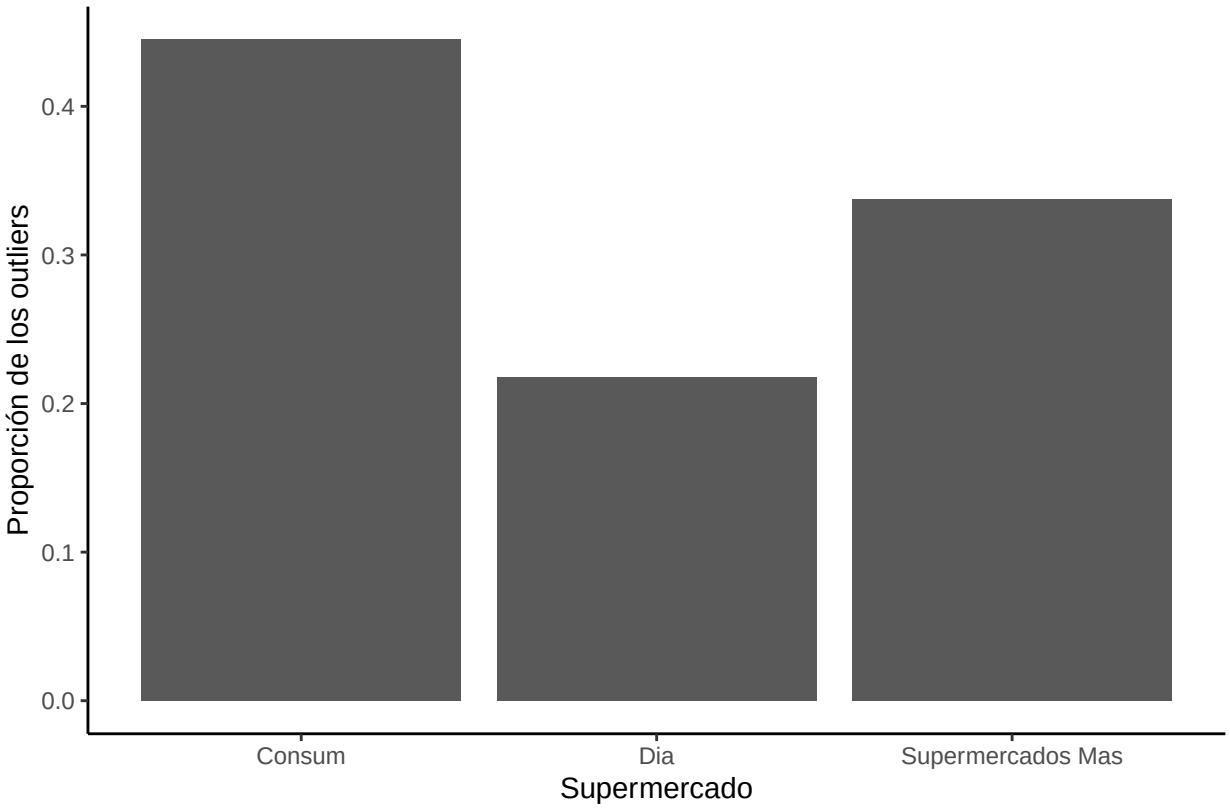


Es pot veure com la variable `Precios` sí té valors extrems que poden distorsionar les dades.

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:

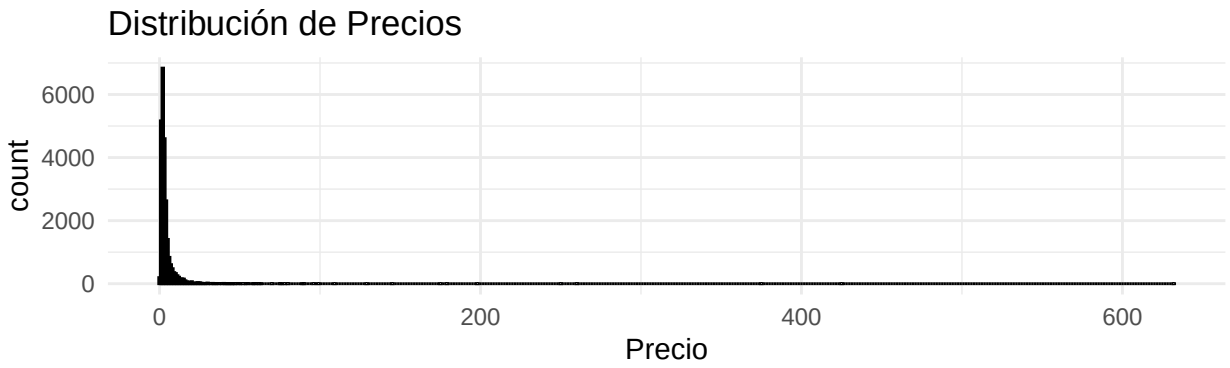


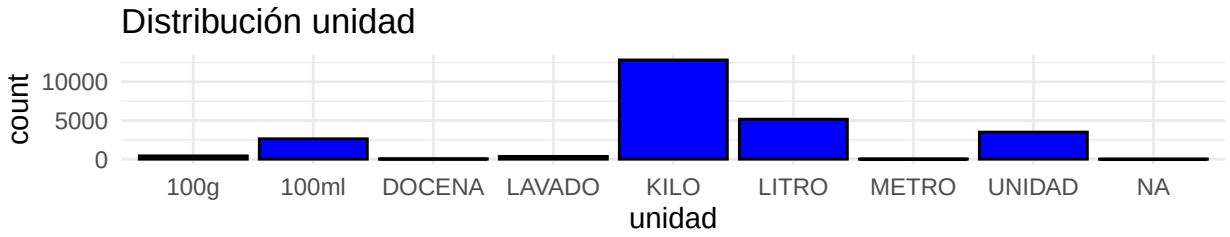
Distribución de los outliers entre los supermercados

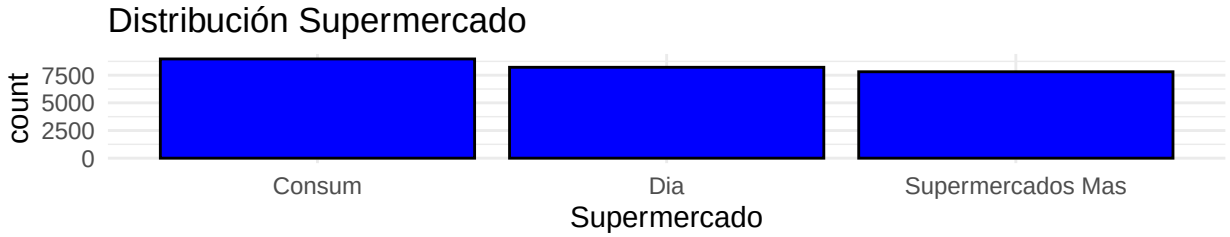
Per tant, es por veure que quasi la meitat dels outliers trobats es troben en el supermercat Consum, mentre que a Dia, és on hi ha menys freqüència de outliers. Això no vol dir que Consum sigui el supermercat més car, sinó que té més articles de preu elevat.

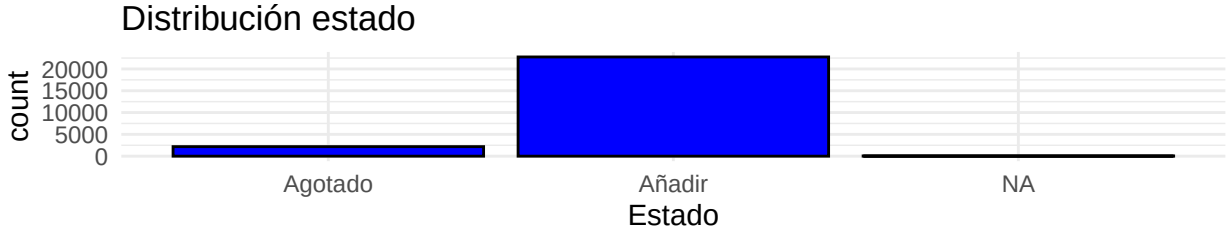
Tot i així, resvisant els valors extrems, considerem que no és necessari corregir ni eliminar-los, ja que no es tracten d'errors d'entrada, sinó que es tracta d'articles de més alta qualitat que tenen un preu alt, però formen part de l'assortiment normal d'un supermercat. Així doncs, formaràn part de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

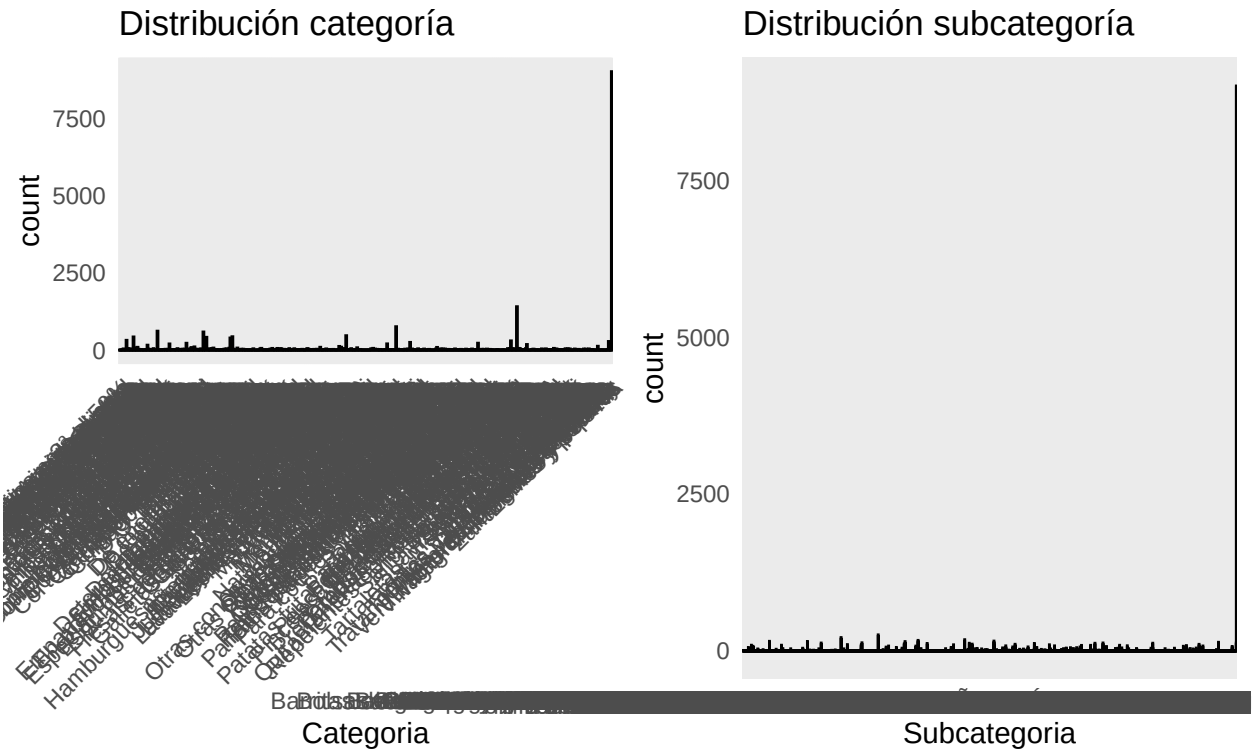
3.5 Representació variables











4 Anàlisi de les dades

4.1 Models

4.1.1 Model supervisat

4.1.2 Model no supervisat

4.2 Prueba de Contraste de Hipótesis