Análisis de ventas BigMart

Carlos Pérez Manzano

Tabla de contenidos

Resumen ejecutivo

Introducción

En este proyecto vamos a realizar un análisis gráfico de un problema de ventas. Contamos con un dataset de 14204 observaciones y 12 variables que recogen las ventas de 1559 productos en 10 establecimientos de la empresa BigMart en el año 2013. Se puede encontrar el conjunto de datos en https://zenodo.org/records/6509955.

Descripción de las variables:

- **Item_Identifier**: variable cualitativa nominal indicando el código del artículo.
- **Item_Weight**: variable numérica indicando el peso del artículo. No se nos informa la unidad de medida en la que se han recogido los datos.
- **Item_Fat_Content**: variable cualitativa ordinal indicando el nivel de grasa del artículo. Los posibles valores son `Low Fat`, `Regular`, `low fat`, `LF`, `reg`.
- **Item_Visibility**: valor numérico que indica cómo de visible es un artículo.
- **Item_Type**: variable categórica indicando el tipo de producto: `Dairy`, `Soft Drinks`, `Meat`, `Fruits and Vegetables`, `Household`, `Baking Goods`, `Snack Foods`, `Frozen Foods`, `Breakfast`, `Health and Hygiene`, `Hard Drinks`, `Canned`, `Breads`, `Starchy Foods`, `Others`, `Seafood`.

- **Item_MRP**: variable numérica indicando el MRP (Maximum Retail Price) del producto.
- **Outlet_Identifier**: variable categórica que contiene el identificador del establecimiento.
- **Outlet_Establishment_Year**: variable cuantitativa discreta indicando el año de inauguración del establecimiento.
- **Outlet_Size**: variable cualtitativa ordinal que muestra el tamaño del establecimiento. Toma los valores `Medium` , `High` y `Small`.
- **Outlet_Location_Type**: variable categórica para indicar la localización del establecimiento: `Tier 1`, `Tier 2`, `Tier 3`.
- **Outlet_Type**: variable categórica que indica el tipo de establecimiento: `Supermarket Type1`, `Supermarket Type2`, `Supermarket Type3`, `Grocery Store`.
- **Item_Outlet_Sales**: variable cuantitativa discreta que indica el número de productos vendidos.

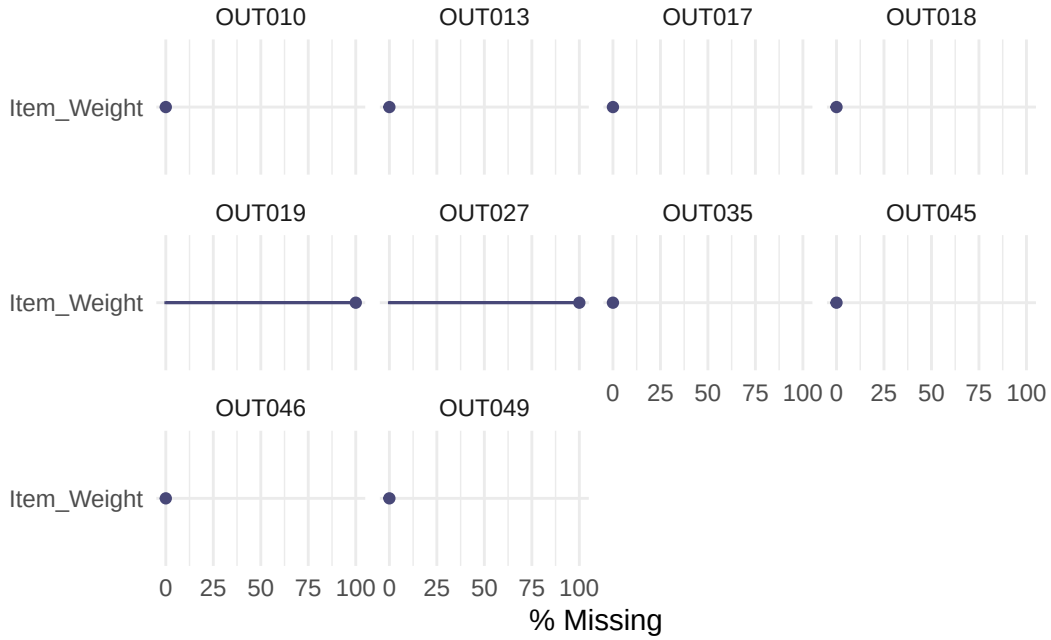
El objetivo es realizar un análisis exploratorio gráfico para comprender las propiedades de los productos y los establecimientos que pueden desempeñar un papel clave en el aumento de las ventas.

Preprocesamiento de datos

El preprocesamiento de los datos de origen ha consistido en hacer frente a los tres problemas siguientes:

1. Inconsistencia de las categorías de la variable `Item_Fat_Content`, pues tanto `Low Fat`, `low fat` y `LF` como `Regular` y `reg` representan la misma categoría respectivamente. Se han renombrado los niveles tomando solo dos: `Low Fat` y `Regular`.
2. Valores nulos en las variables `Item_Weight` y `Outlet_Size`.
3. Datos erróneos en la variable `Item_Visibility`, pues existen artículos con visibilidad igual a 0 con ventas.

Para el problema de los valores nulos, observemos la siguiente gráfica para los valores nulos de `Item_Weight`.



Vemos como los establecimientos OUT019 y OUT027 no han informado del peso del producto en ningún caso, mientras que los demás sí lo han hecho en todos ellos. Sin embargo, los productos cuyo peso no han sido informados son también vendidos por otros establecimientos. Por tanto, se imputan los valores nulos extrayendo el peso real de los productos del resto de establecimientos.

Para la variable `Outlet_Size` ocurre algo similar, pues en este caso OUT010, OUT017 y OUT045 no han proporcionado este dato. Se han imputado los valores nulos mediante mediante una predicción de manera exploratoria del tamaño de los establecimientos en función del número de ventas totales de estos.

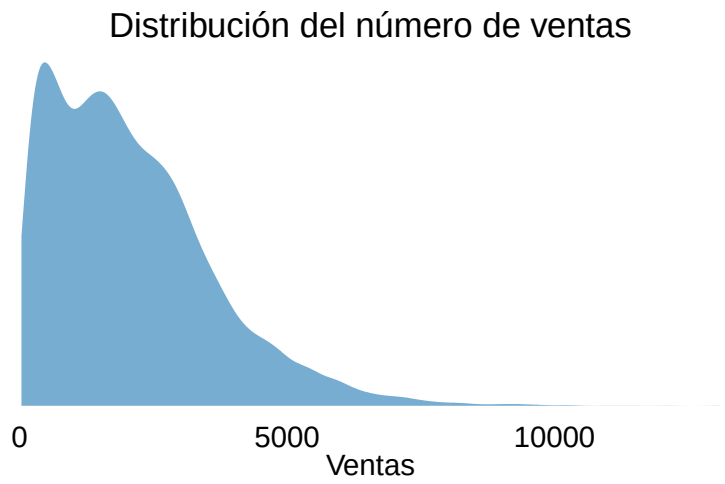
En cuanto al tercer problema, se han imputado los valores de `Item_Visibility` que son 0 por la media de la visibilidad de los productos del mismo tipo.

Análisis gráfico

Análisis univariante

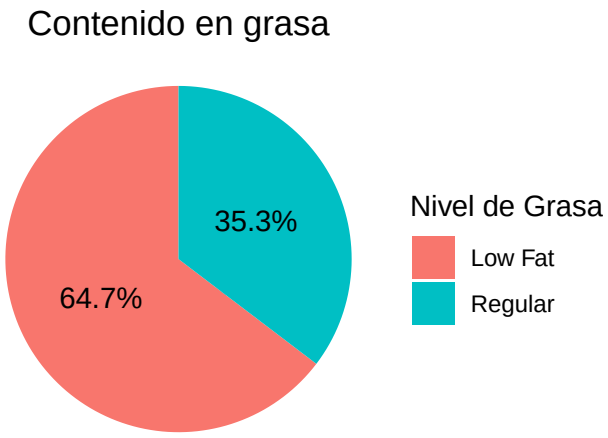
En primer lugar se ha procedido a el estudio univariante de los datos. Destaquemos algunos de estos análisis.

Mostramos la estimación de la densidad de la variable `Outlet_Item_Sales`, en la que podemos ver la distribución del número de ventas de cada producto en los distintos establecimientos.



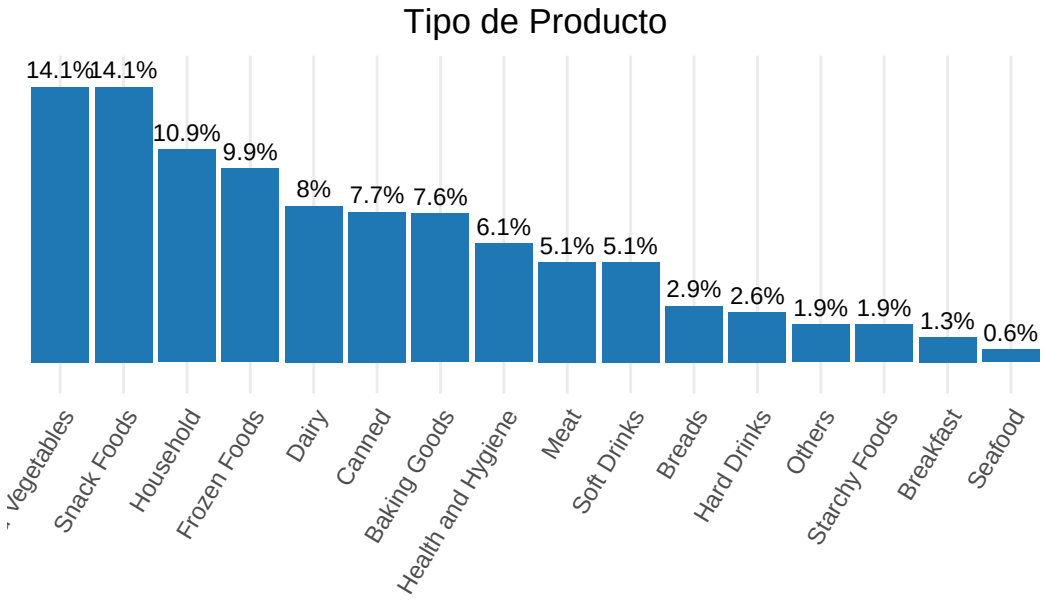
Observamos una distribución similar a una de grado en ley de potencia. El número de ventas de los productos en los establecimientos suele ser parecido, mientras que hay ciertos productos que tienen un número de ventas mucho mayor, o más bien productos que logran un mayor número de ventas en ciertos establecimientos.

Las proporciones del contenido en grasa de los productos es la siguiente:



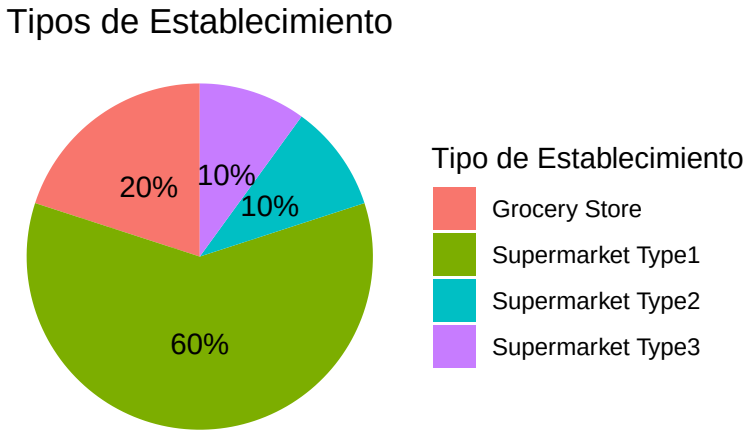
Vemos como hay un mayor número de productos con un bajo nivel de grasa.

En cuanto a los tipos de producto:



Los vegetales y los snacks son los tipos de comida con un mayor porcentaje de presencia entre los productos. En contraposición, los productos de desayuno y marisco están menos presentes.

Para las variables referentes a los establecimientos destacamos `Outlet_Type`.

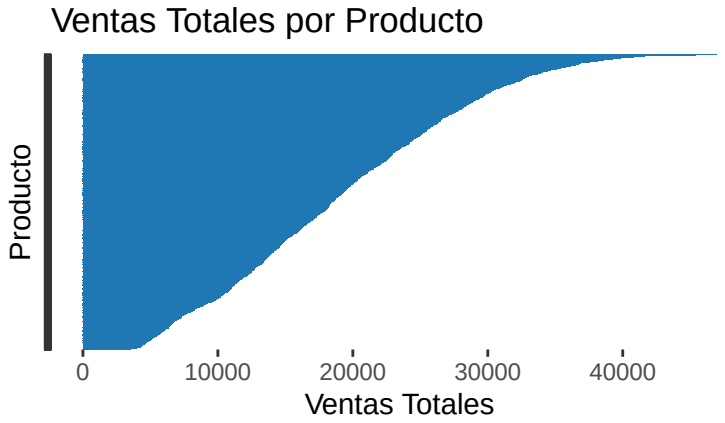


El tipo más común es el supermercado tipo 1. Por otro lado, los tipos `Grocery Store` (tienda de comestibles) y `SuperMarket Type3` representan cada uno un 10% del total. Teniendo en cuenta que contamos con un total de 10 tiendas, quiere decir que solo tenemos un establecimiento de cada uno de esos tipos.

Relación entre variables

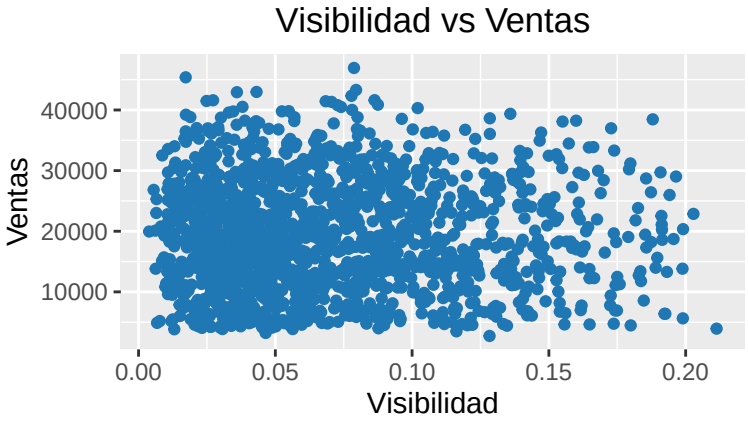
Puesto que lo que se pretende es detectar ciertos factores que influyen en el número de ventas, vamos a realizar una serie de gráficas que permitan notar las relaciones que esconden los datos.

Antes hemos visto las diferencias entre las ventas de los productos en los establecimientos. Un posible motivo sería que un mismo producto tiene un gran número de ventas en un establecimiento y en otras pocas debido al renombre del establecimiento y no es muy dependiente al producto. Para descartar esa posibilidad, veamos cómo se distribuyen las ventas de los productos en la totalidad de los establecimientos.



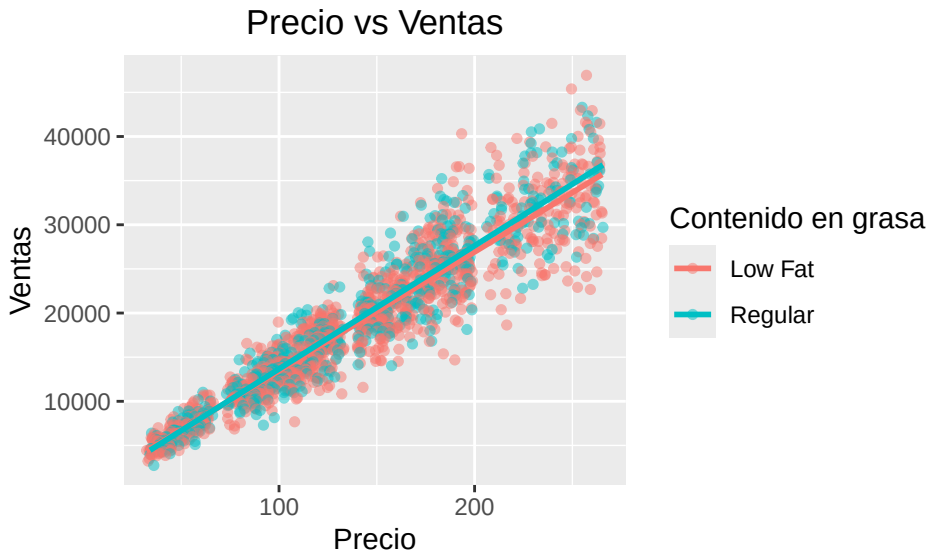
Queda claro que el motivo de la diferencia de las ventas de los distintos productos no depende únicamente del establecimiento, sino también del producto en sí. Esto es algo a tener en cuenta para el stock, pues es razonable tener una oferta de productos según la demanda de ellos.

Siguiendo con las características de los productos, veamos cómo se relacionan la visibilidad y las ventas.



Observamos como los productos con un mayor número de ventas no tienen una gran visibilidad, mientras que hay artículos con una mayor visibilidad que no necesariamente tienen un gran número de ventas. Esto se puede deber a que algunos de los productos que consiguen un mayor número de ventas se corresponden con productos con una gran demanda o necesidad, tanta que provoca que no sea necesario darles gran visibilidad. Por otro lado, se intenta potenciar a los productos que tienen un menor número de ventas (quizás por una baja demanda) dándoles mayor visibilidad.

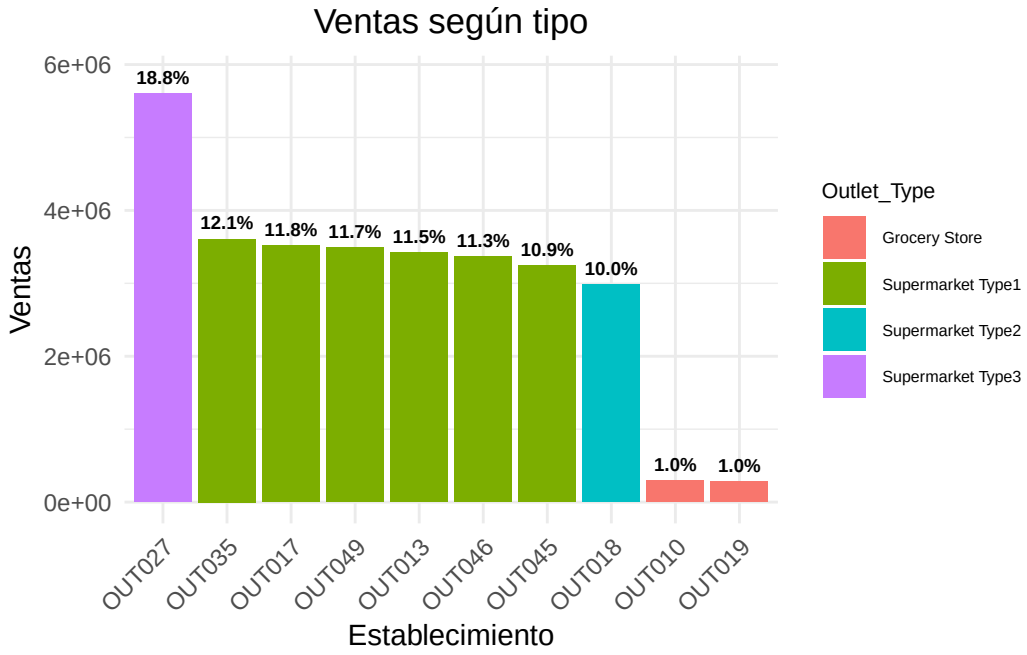
Veamos ahora cómo afecta el MRP (Precio máximo de venta al público) del producto en las ventas, en función también del contenido en grasa de estos.



De este gráfico podemos sacar aparentemente las siguientes conclusiones:

- Contamos con una clara relación lineal entre las ventas y el MRP Las ventas son sustancialmente mayores conforme el MRP de los productos aumenta. Notamos también como la fluctuación de las ventas es mayor a medida que se va aumentado el MRP del producto. El primero de los hechos se puede deber a muchos factores: una mayor calidad de los productos más caros, marcas más relevantes, más atractivos, productos que inicalmente eran más baratos pero tienen una clientela fiel que no renuncia a la compra pese a las subidas de precio, etc. En cuanto a lo segundo, puede deberse a que una subida repentina de los precios desenvoque en una baja considerable de potenciales clientes.
- No hay una relación clara aparente de los contenidos en grasa de los productos con los precios o las ventas de los productos.

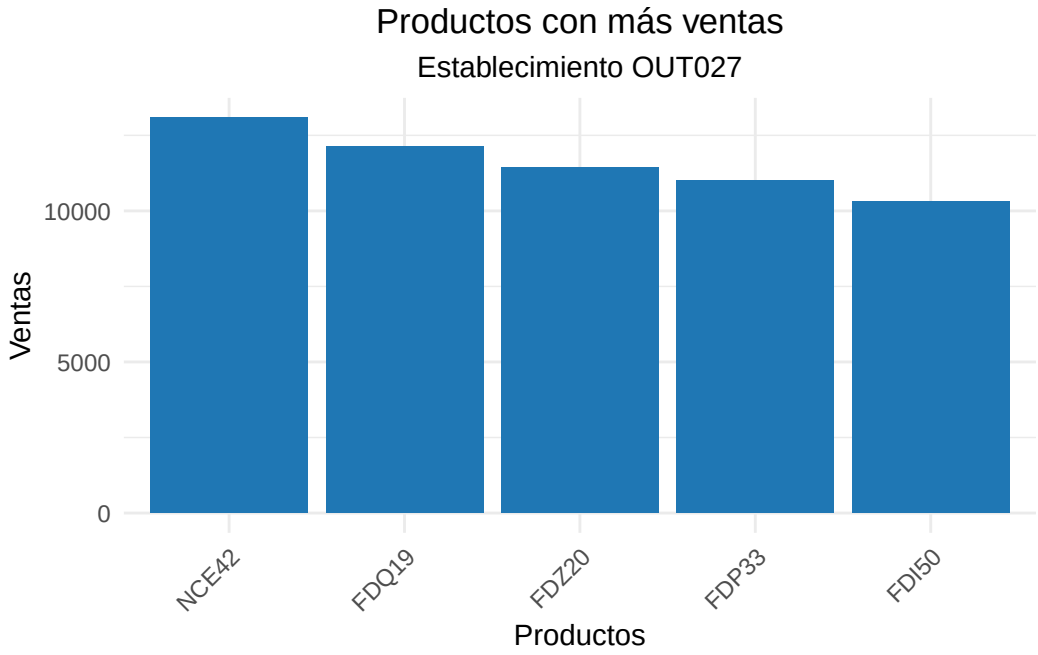
Pasemos ahora a considerar alguna gráfica que muestre la relación de las ventas con los establecimientos. De la que podemos sacar un mayor número de conclusiones es la siguiente.



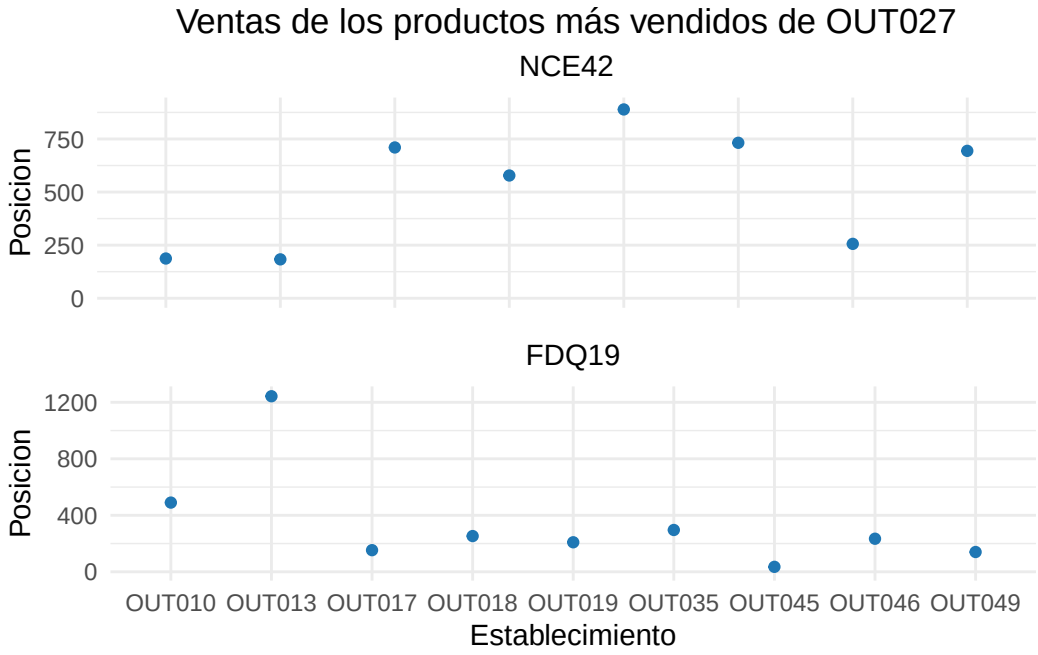
Destacamos en primer lugar el establecimiento OUT027 pues es con una gran diferencia el que realiza un mayor número de ventas, siendo estas un 18.8% del total. Los establecimientos OUT010 y OUT19 son los que menos realizan, siendo un 1% las ventas del total para cada una de ellas.

Observamos también una clara relación entre el tipo de supermercado y las ventas. Los establecimientos con menos ventas son tiendas de comestibles, lo que hace lógico ese número de ventas. El OUT027 es el único `SuperMarket Type3`. Obviando los costes de la constitución de este tipo de establecimientos pues no la conocemos, parece bastante razonable considerar la instauración de nuevos establecimientos de este tipo o la remodelación de los existentes.

Por último, sabiendo que OUT027 es que tiene mejores resultados, podemos tomar este como ejemplo a seguir para el resto de establecimientos. Una posible estrategia es potenciar en el resto de tiendas los artículos que más éxito tienen en OUT027 si estos no obtienen muchas ventas. Para ello, vamos a obtener los 2 ítems con un mayor número de ventas en OUT027 y vamos a ver qué posición ocupan en el ránking de ventas del resto de establecimientos.



Vemos como los ítems NCE42 y FDQ19 son los que más vendidos en el establecimiento OUT027.



Hay establecimientos como el OUT019, OUT045 o OUT017 que deberían potenciar más la venta del producto NCE42, mientras que el OUT013 es el que más destaca en los problemas

para vender el producto FDQ19.

En estos establecimientos podríamos tomar decisiones de una mayor inversión en marketing, dar una mayor visibilidad o realizar promociones para crear clientes con necesidad de estos productos, pues es quizás una de las mayores razones de éxito del establecimiento OUT027.

Anexo

El código completo y gráficas adicionales se puede encontrar en el siguiente enlace de Github