PECL 6 - Fundamentos de la Ciencia de Datos

Visualización de datos

Marcos Barranquero Eduardo Graván
Adrián Montesinos

10 de diciembre de 2019

1. Introducción

En esta práctica no contamos con ejercicios guiados por el profesor, por lo que se ha dividido en dos partes propias. La primera estudia varios métodos y librerías de R que permiten visualizar los datos mediante diferentes gráficos y modelos. La segunda es un estudio sobre la correctitud en la representación de los datos, la manipulación que podemos encontrar tras estos, y cómo corregirla.

2. Apartado 1 - Visualización de datos en R

En esta sección vamos a utilizar varias formas de visualizar datos en R. El conjunto de datos elegido para los ejemplos del estudio está conformado por información sobre productos de supermercado, y se encuentra en el archivo *Dataset.xlsx*.

2.1. Diagrama de dispersión - ScatterPlot

El diagrama de dispersión nos permite visualizar la relación entre dos variables. La variable independiente es aquella que se emplea para medir. Normalmente incrementa de forma sistemática, y va en el eje X. La variable dependiente es aquella variable medida en función de la independiente, y se representa en el eje Y.

Para este ejemplo, pondremos como variable independiente el índice de MRP de los productos en relación a la visibilidad del producto.

En primer lugar, cargaremos las librerías necesarias para leer los datos, y para realizar el diagrama de dispersión. También necesitamos cargar los datos:

```
> library("readxl") # Libreria para leer archivos de excel
> library("ggplot2") # Libreria para hacer plots avanzados
> # Leemos los datos desde el excel
> datos <- read_excel("Dataset.xlsx")
```

Después, podemos dibujar el gráfico con la función *geompoint*

```
# Representamos los datos con un gráfico de dispersión
print(ggplot(datos, aes(Item_Visibility, Item_MRP))
    + geom_point(aes(color = Item_Type))
    + scale_x_continuous("Visibilidad de los articulos",
    breaks = seq(0,0.35,0.05))
    + scale_y_continuous("MRP de los articulos",
    breaks = seq(0,270,by = 30))
    + theme_bw() + labs(title="Estudio de la relacion
    entre variables con Scatterplot"))
```

Obteniendo el siguiente gráfico:

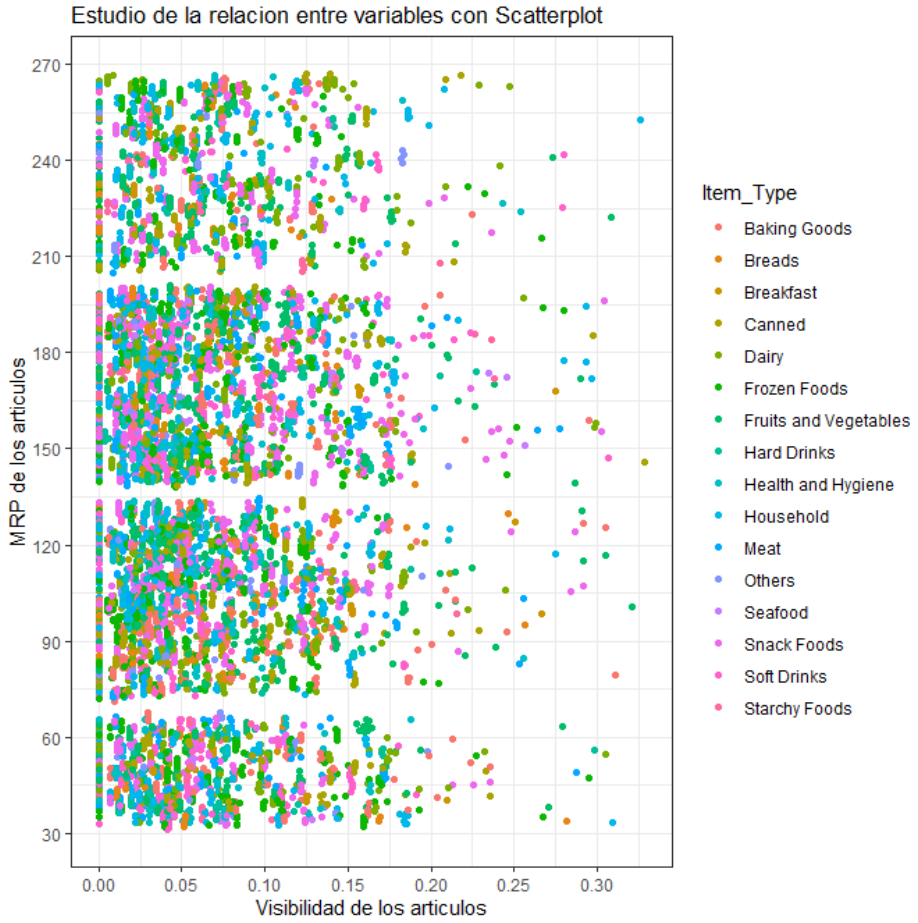

Figura 1: Diagrama de dispersión

Ahora, queremos ampliar el diagrama, realizando varios diagramas de dispersión en función de la categoría del artículo. Para ello, realizamos la misma llamada añadiendo una línea adicional indicando que queremos que agrupe por la variable *ItemType*:

```
 # Representamos los datos separando en categorias con
 # un gráfico de dispersión
print(ggplot(datos, aes(Item_Visibility, Item_MRP))
        + geom_point(aes(color = Item_Type))
        + scale_x_continuous("Visibilidad de los articulos",
        breaks = seq(0,0.35,0.05))
        + scale_y_continuous("MRP de los articulos",
        breaks = seq(0,270,by = 30))
        + theme_bw() + labs(title="Estudio de la relacion
        entre variables separando en categorias con Scatterplot")
        + facet_wrap( ~ Item_Type))
```

Y obtenemos el siguiente gráfico:

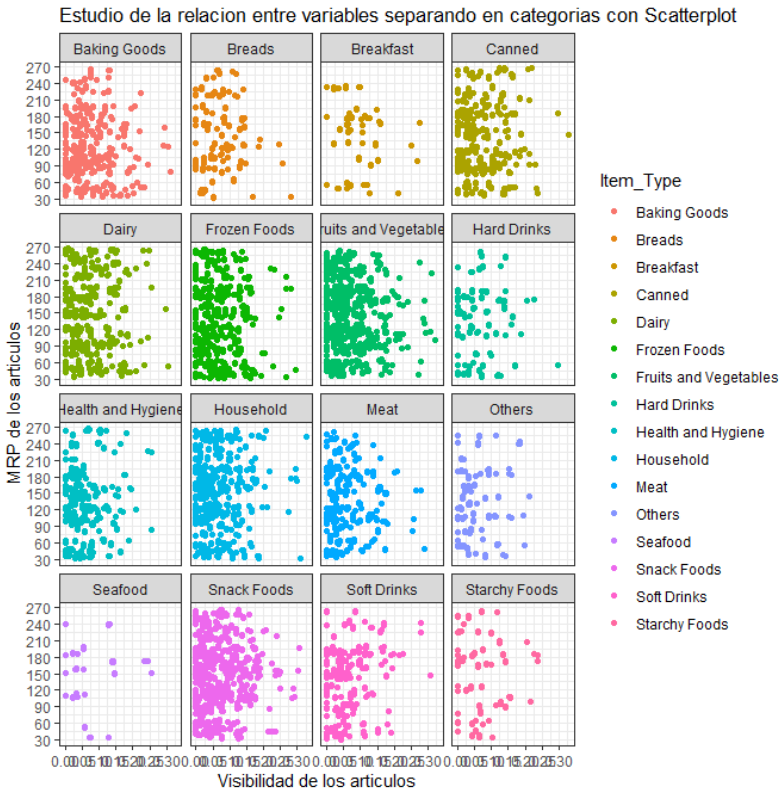

Figura 2: Diagrama de dispersión por categorías

2.2. Diagrama de temperatura - Heat Map

Un mapa de temperatura o *HeatMap* muestra la relacion entre dos o tres variables usando la intensidad de colores en una imagen bidimensional. La intensidad del color actúa como una tercera dimensión.

Para el ejemplo, queremos relacionar categorías de productos con su *outlet*, en función del MRP de los productos. Como ya se tienen cargados los datos y las librerías, no hace falta volver a cargarlos.

Volvemos a llamar a ggplot con la función *geomRaster* para realizar el mapa de temperatura:

```
# Representamos los datos con un Heat Map
print(ggplot(datos, aes(Outlet_Identifier, Item_Type))
        + geom_raster(aes(fill = Item_MRP))
        + labs(title ="Heat Map",
        x = "Identificador de outlet", y = "Tipo de articulo")
        + scale_fill_continuous(name = "MRP"))
```

Y obtenemos el siguiente diagrama:

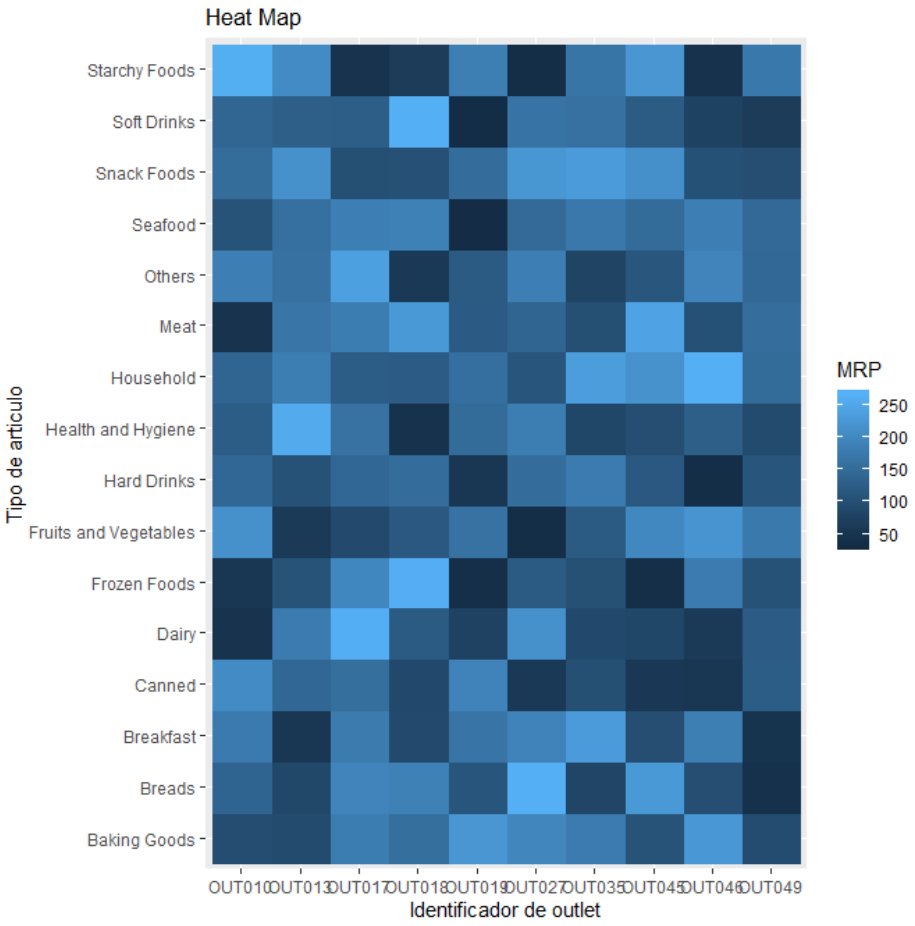

Figura 3: Diagrama de temperatura

Observamos que la intensidad del color define el MRP.

2.3. Diagrama de datos anómalos - Box Plot

Este diagrama es útil para visualizar la dispersión de los datos y detectar datos anómalos. Muestra 5 datos significativos:

- El valor mínimo
- El percentil 25
- La mediana
- El percentil 75
- El valor máximo

Para el ejemplo, vamos a identificar datos anómalos en la cantidad de ventas de items, en función de su temporada o *outlet*.

Para ello, ejecutamos el siguiente script haciendo uso de *geomBoxplot*.

```
# Representamos los datos usando caja y bigotes, pudiendo verse los outliers
print(ggplot(datos, aes(Outlet_Identifier, Item_Outlet_Sales))
        + geom_boxplot(fill = "red")
        + scale_y_continuous("Ventas de articulo", breaks= seq(0,15000, by=500))
        + labs(title = "Datos anomalos con Box Plot",
        x = "Identificador de articulo"))
```

Al ejecutarlo, obtenemos el siguiente gráfico:

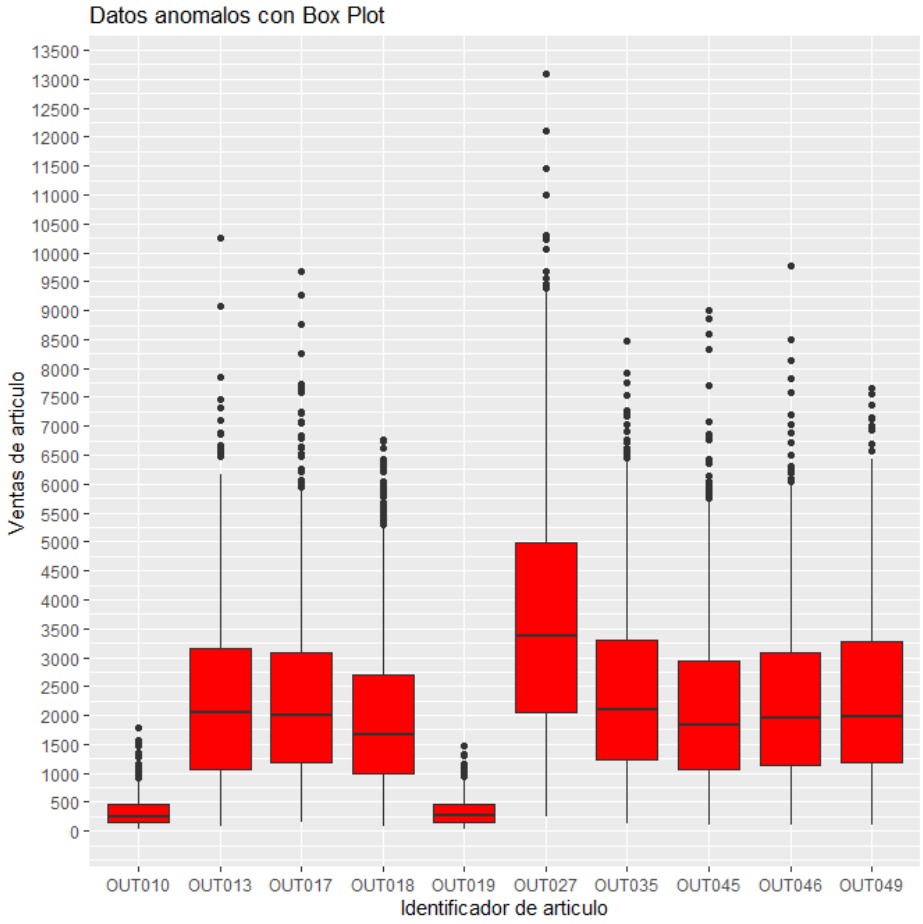

Figura 4: Diagrama de datos anómalos

Los puntos negros son *outliers* o datos anómalos. Las barras negras parten del mínimo y llegan al máximo, mientras que los bloques rojos indican el percentil en el que se encuentran:

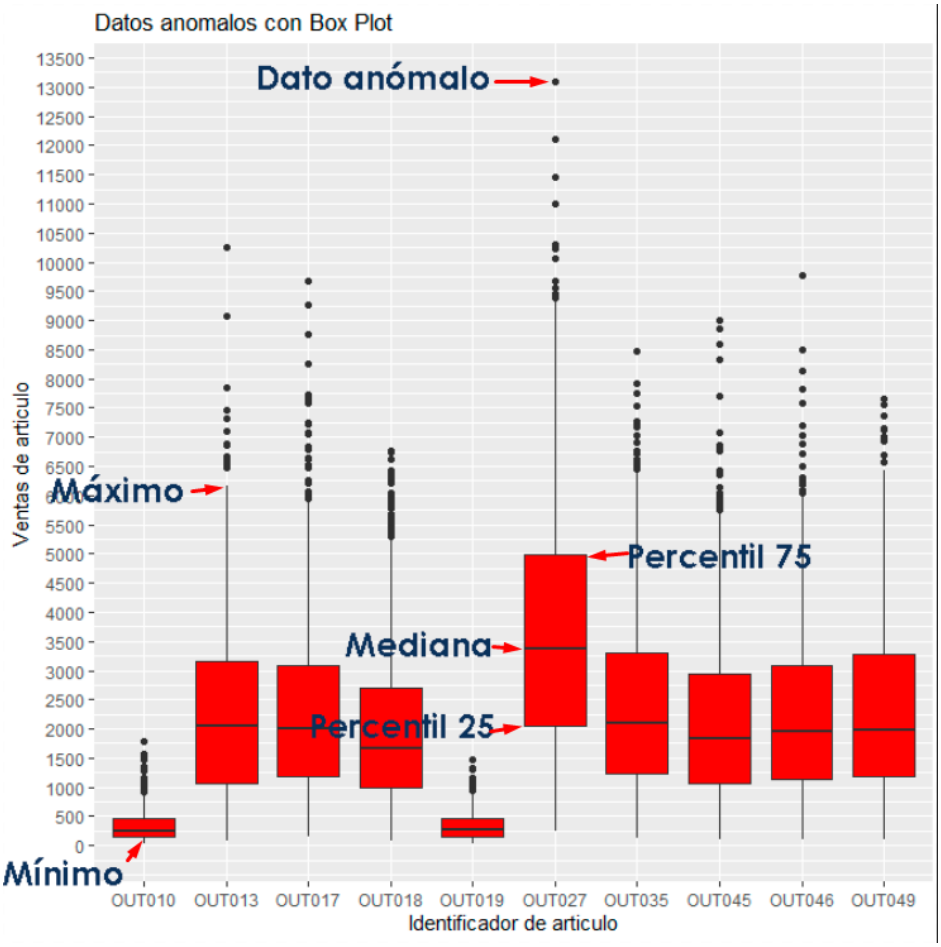

Figura 5: Diagrama de datos anómalos

3. Apartado 2 - Estudio de correctitud y representación de datos

Según la manera que elijamos de representar y visualizar los datos, podemos influenciar la opinión y tergiversar lo que los datos representan y la información que transmiten.

Nuestro objetivo debe ser transmitir la información representada de manera neutra y ética, sin modificar el mensaje que los datos transmiten.

Para llevar a cabo esto, debemos tener los siguientes criterios en cuenta:

- Nuestro gráfico debe contener el mínimo número de elementos gráficos para representar la información que queremos transmitir.
- Que la línea base comience en 0 y los ejes tengan nombres y valores.
- Buscaremos que tenga una correspondencia exacta con la información textual, es decir, que no tergiverse la información.
- En matera de diseño, buscaremos evitar colores llamativos y estilos innecesarios, que entorpezcan el envío de información.
- En relación a los outliers, si deforman mucho el gráfico, podemos recortar la manera de representarlos para que no sesguen el resto de elementos.
- Los valores negativos tienen preferencia para representarse en vertical.
- Para múltiples datos, queremos evitar contrastes llamativos y no incluir más de 4 o 5 categorías, para no crear un gráfico demasiado confuso.
- Los gráficos con ejes deben de ser limpios y comprensibles, no deben mostrarse como gráficos *spaggueti*.

A continuación, veremos varios ejemplos de gráficos que no se adecuan a la realidad y se saltan uno o varios de estos criterios, y como se podrían corregir.

3.1. Datos negativos e inversión del gráfico

En la figura 6 podemos observar una representación de la variación del precio de la luz. Encontramos varios errores:

- En primer lugar, la variación puede ser menor o mayor, pero nunca puede ser negativa.
- En segundo lugar, observamos que las barras del eje x no son congruentes: hasta cierto punto muestran la varaición por años, y luego pasan a mostrarla por meses.

Si corregimos estos fallos, obtenemos un gráfico bastante diferente, mostrado en la figura 7.

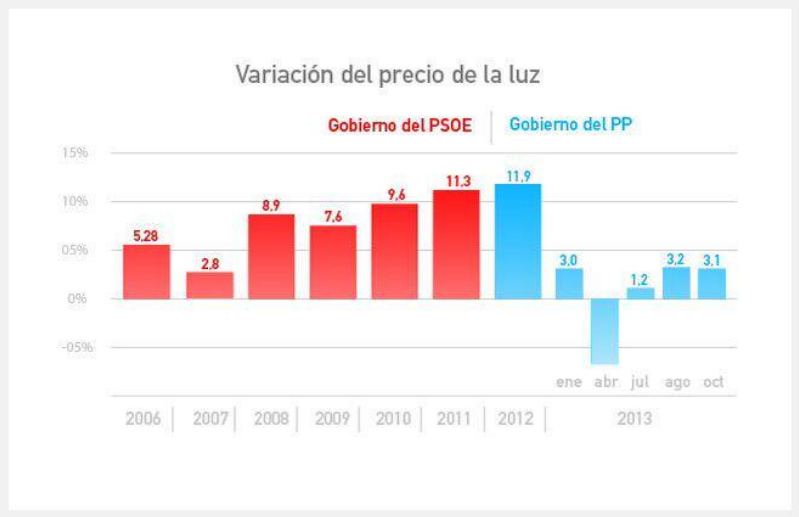

Figura 6: Gráfico erróneo de variación del precio de la luz

Otro gráfico errático que hace uso de invertir los ejes es el mostrado a la izquierda en la figura 8. Este intento de manipulación muestra el gráfico al revés para causar la impresión de que el número de víctimas ha subido, en lugar de disminuir. A la derecha, podemos ver el gráfico corregido. Cabría resaltar también que años contabilizados en el eje x no están del todo claros.

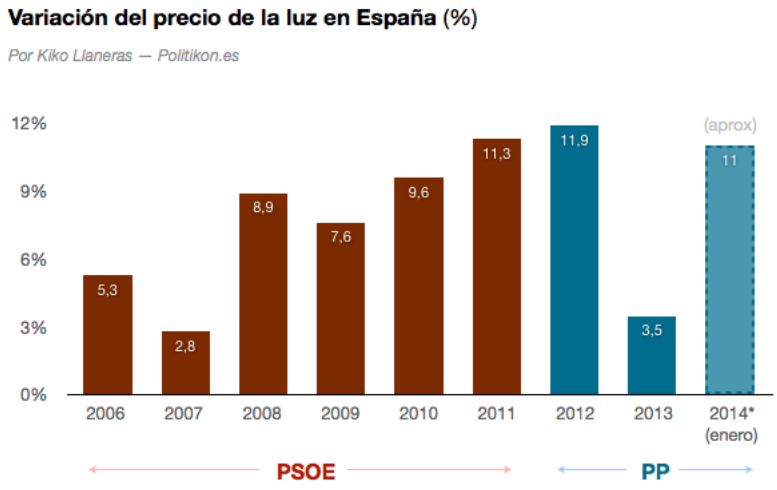

Figura 7: Gráfico correcto de variación del precio de la luz

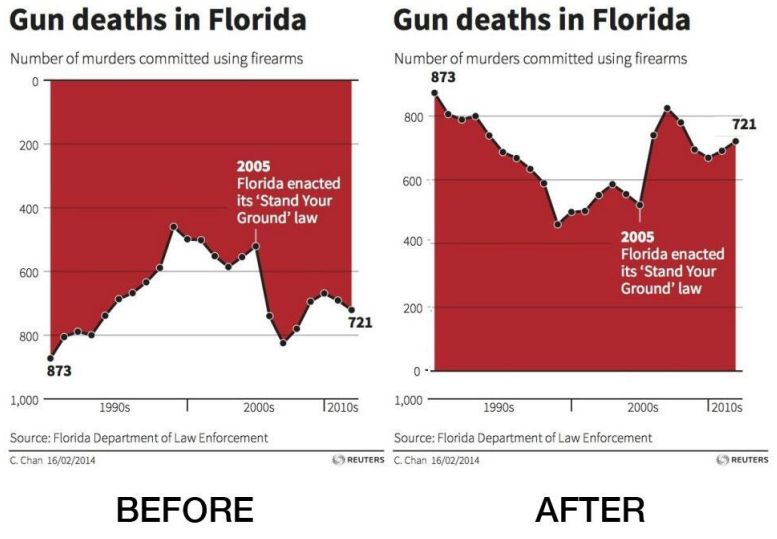

Figura 8: Gráfico incorrecto y corregido del número de muertes por armas en EEUU.

3.2. Ejes cortados

En la figura 9, mostrada por RTVE en un noticiero, podemos hallar los siguientes errores:

- El eje y no comienza en 0 y no contiene ninguna linea de escala que permita discernir los valores de dicho eje.
- El eje x tan solo muestra dos elementos, por lo que se encuentra limitado a un dominio muy bajo que resulta insuficiente para poder extraer conclusiones.

Podemos ver una versión corregida en la figura 10, donde ampliamos el dominio del eje x, establecemos en 0 el comienzo del eje y, y podemos ver lineas de escala.

Figura 9: Gráfico incorrecto del paro de 2013 en España

3.3. Gráficos 3D

Los gráficos circulares son una pésima herramienta de comunicación de información. En la figura 11 observamos los siguientes errores:

- El gráfico se encuentra rotado y girado hacia el público, de modo que produce la sensación de que la porción de Apple es mucho más grande de lo que en realidad es.
- El sector de .ºtros"se encuentra en posición opuestl de Apple, haciendo que, con la distorsión por la rotación, parezca más pequeño de lo que en realidad es.

Si empleamos un gráfico de barras, podemos observar que a Apple no le iba tan bien como quisieron hacer ver allá por el 2008:

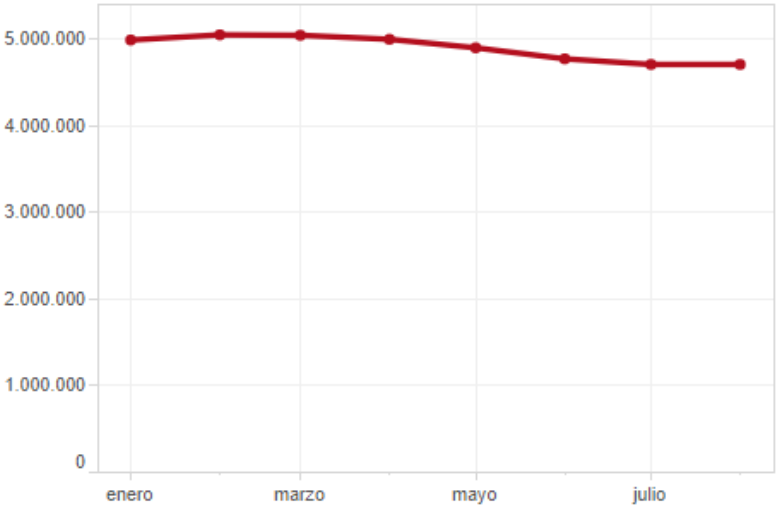

Figura 10: Gráfico correcto del paro de 2013 en España

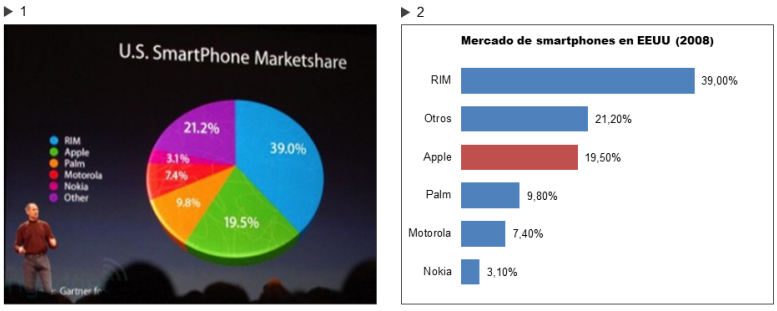

Figura 11: Gráficos de ventas de Apple

3.4. Gráficos de área

Un gráfico de área es un gráfico de líneas en el que el área entre la línea y el eje aparece sombreada con un color. Estos gráficos normalmente se usan para representar los totales acumulados a lo largo del tiempo y son la forma convencional de visualizar líneas apiladas. Sin embargo, también se pueden utilizar para comunicar un mensaje que no se corresponda con la realidad.

En el gráfico de la izquierda de la figura 12 vemos un gráfico de líneas apilado. En este, da la impresión de que el crecimiento económico de España es acorde al de otros países europeos. Sin embargo, la realidad es que no ha cambiado: a lo largo del eje x, el área ha sido la misma. Si lo representamos con un gráfico de líneas como el que se muestra a la derecha, podemos visualizarlo.

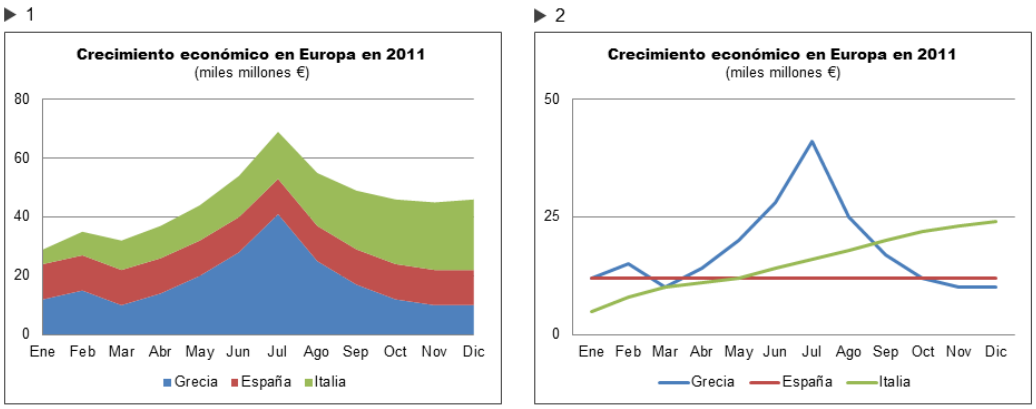

Figura 12: Gráficos de crecimiento económico

3.5. Conclusiones

La visualización de datos es una parte fundamental en todos los estudios de análisis de datos. Apoyándose en la visualización, el analista va descubriendo los secretos enterrados en los datos. Al desarrollar gráficos y formas de visualizar dichos datos, debemos aproximarnos con cautela: es fácil cometer estos errores y tergiversar los resultados en busca de una opinión o idea que no está reflejada en los propios datos originales.