# Tipología y ciclo de vida de los datos

Práctica 2: Limpieza y Análisis de Datos


Alicia Rodríguez Gómez

4 de enero de 2022

# Pregunta 1

**Descripción de Dataset. ¿ Por qué es importante y qué pregunta/ problema presente responder?**

El dataset elegido contiene información sobre productos vendidos en Agosto de 2020 en la plataforma digital Wish.com. Este está compuesto por los siguientes campos:

- Title: Título localizado para los países europeos. Puede ser el mismo que title_orig si el vendedor no ofrece una traducción.
- Title_orig: Título original en inglés del producto.
- Price: Precio que pagaría para obtener el producto.
- Retail_price: Precio de referencia de artículos similares en el mercado, o en otras tiendas o lugares. Utilizado por el vendedor para indicar un valor regular o el precio antes del descuento.
- Currency_buyer: Moneda de la venta.
- units_sold: Número de unidades vendidas.
- uses_ad_boots: Si el vendedor pagó para potenciar su producto dentro de la plataforma.
- Rating: Valoración media del producto
- Rating_count: Número total de valoraciones del producto
- Rating_five_count: Número de valoraciones de 5 estrellas.
- Rating_four_count: Número de valoraciones de 4 estrellas.
- Rating_three_count: Número de valoraciones de 3 estrellas.
- Rating_two_count: Número de valoraciones de 2 estrellas.
- Rating_one_count: Número de valoraciones de 1 estrella.
- badges_count: Número de insignias que tiene el producto o el vendedor.
- badge_local_product: Un distintivo que denota que el producto es un producto local. Las condiciones pueden variar (ser producido localmente, o

algo más). Algunas personas pueden preferir comprar productos locales en lugar de. 1 significa Sí, tiene el distintivo.

- badge_product_quality: Insignia concedida cuando muchos compradores dieron sistemáticamente buenas evaluaciones 1 significa Sí, tiene la insignia.
- badge_fast_shipping: Insignia concedida cuando el pedido de este producto se envía rápidamente de forma sistemática
- Tags: etiquetas establecidas por el vendedor
- product_color: Color del producto.
- product_variation_size_id: Una de las variaciones de tamaño disponibles para este producto.
- product_variation_inventory: Inventario que tiene el vendedor. La cantidad máxima permitida es de 50
- shipping_option_name
- shipping_option_price: Precio de envío.
- shipping_is_express: Booleano que indica si existe la posibilidad de envío express.
- countries_shipped_to: Número de países a los que se envía este producto.
- inventory_total: Inventario total de todas las variaciones del producto.
- has_urgency_banner: Booleano que indica si posee banner de texto.
- urgency_text: Un banner de texto que aparece sobre algunos productos en los resultados de la búsqueda.
- origin_country: País de origen del producto.
- merchant_title: Nick del vendedor.
- merchant_name: Nombre del vendedor.
- merchant_info_subtitle: Texto que muestra una descripción del vendedor.
- merchant_rating_count: Número de calificaciones del vendedor.

- merchant_rating: Calificación del vendedor.
- merchant_id: Identificador del vendedor.
- Merchant_has_profile_picture: Booleano que indica si el vendedor tiene imagen en su perfil.
- merchant_profile_picture: Imagen perfil del vendedor
- product_url: Enlace al producto.
- product_picture: Enlace a imagen del producto.
- product_id: Identificador del producto.
- Theme: Temporada del producto. Único valor verano.
- crawl_month: Mes de venta del producto. Único valor agosto 2020.

Los datos han sido obtenido de la siguiente página de Kaggle: https://www.kaggle.com/jmmvutu/summer-products-and-sales-in-ecommerce-wish?select=summer-products-with-rating-and-performance_2020-08.csv

El objetivo de este análisis es analizar cómo influyen tanto el precio como la puntuación que tienen plataformas online de venta de ropa a la hora de aumentar sus ventas. Para ello, utilizaremos una selección de atributos del conjunto anterior que nos den información sobre cómo es la venta en función de precios y de puntuaciones que los usuarios le han dado a los diferentes productos. En particular nos centraremos en datos de la plataforma de venta Wish, pero que perfectamente podrían ser extrapolables a cualquier otra plataforma de venta online.

# Pregunta 2

**Integración y selección de los datos de interés a analizar.**

En primer lugar, se procederá a la carga de datos del fichero csv.

```
#Lectura de datos
summerClothes <- read.csv(file = '~/R-projects/data/summer-products-with-rating-and-performance_2020-08.csv')
head(summerClothes)
```

No se muestra el resultado de la función head porque al tener tal cantidad de columnas la salida es demasiado grande. Se repetirá esta operación cuando se haga una selección de un subconjunto de columnas.

Además, en el siguiente paso vamos a ver los tipos de datos que se les han asignado a cada una de las columnas.

```
#Tipo de dato asignado a cada campo
sapply(summerClothes,  function(x) class(x))
```

```
                    title                  title_orig                       price                 retail_price
                "character"                 "character"                 "numeric"                   "integer"
            currency_buyer                  units_sold             uses_ad_boosts                      rating
                "character"                   "integer"                   "integer"                   "numeric"
             rating_count             rating_five_count          rating_four_count          rating_three_count
                "integer"                   "integer"                   "integer"                   "integer"
          rating_two_count              rating_one_count             badges_count         badge_local_product
                "integer"                   "integer"                   "integer"                   "integer"
    badge_product_quality             badge_fast_shipping                        tags               product_color
                "integer"                   "integer"                 "character"                 "character"
product_variation_size_id    product_variation_inventory        shipping_option_name        shipping_option_price
                "character"                   "integer"                 "character"                   "integer"
       shipping_is_express           countries_shipped_to           inventory_total             has_urgency_banner
                "integer"                   "integer"                   "integer"                   "integer"
              urgency_text              origin_country              merchant_title               merchant_name
                "character"                 "character"                 "character"                 "character"
     merchant_info_subtitle          merchant_rating_count          merchant_rating                 merchant_id
                "character"                   "integer"                   "numeric"                 "character"
merchant_has_profile_picture       merchant_profile_picture                 product_url             product_picture
                "integer"                 "character"                 "character"                 "character"
               product_id                       theme                 crawl_month
                "character"                 "character"                 "character"
```

Previo a hacer los próximos análisis se va a hacer una selección de los campos más relevantes para nuestro análisis. Rechazaremos los campos que no aportan información muy relevante.

```
#Seleccion de columnas
summerClothes <- summerClothes %>% select(product_id, price, retail_price, merchant_id,
                        merchant_rating, merchant_rating_count, origin_country,
                        inventory_total, countries_shipped_to, shipping_option_name,
                        shipping_option_price, product_color, product_variation_size_id,
                        product_variation_inventory, rating, rating_count, rating_one_count,
                        rating_two_count, rating_three_count, rating_four_count, rating_five_count,
                        units_sold, tags)

summerClothes
```

# Pregunta 3

**Limpieza de los datos.**

**3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?**

A continuación se va a calcular cuál de los campos tienen valores perdidos. En este caso no vamos a analizar los datos que contienen ceros puesto que en este conjunto de datos, el cero es un valor válido.

```
# Números de valores desconocidos por campo
sapply(summerClothes, function(x) sum(is.na(x)))
```

```
              product_id                    price             retail_price              merchant_id
                       0                        0                        0                        0
         merchant_rating     merchant_rating_count          origin_country          inventory_total
                       0                        0                        0                        0
     countries_shipped_to      shipping_option_name shipping_option_price           product_color
                       0                        0                        0                        0
product_variation_size_id product_variation_inventory                   rating             rating_count
                       0                        0                        0                        0
       rating_one_count         rating_two_count       rating_three_count        rating_four_count
                      45                       45                       45                       45
       rating_five_count              units_sold                     tags
                      45                        0                        0
```

Se observa que los campos que tienen alguno de valores nulos o perdidos son aquellos que cuentan el número de calificaciones con estrellas, desde la uno a la cinco. Haciendo un análisis de los datos, esto ocurre cuando el producto no ha tenido ninguna valoración, o lo que es lo mismo, el rating_count tiene una valor igual a cero. Es por ello que la decisión que se ha tomando sobre estos valores nulos, es rellenarla con un valor igual a cero.

```
# Completar con ceros los valores desconocidos
```

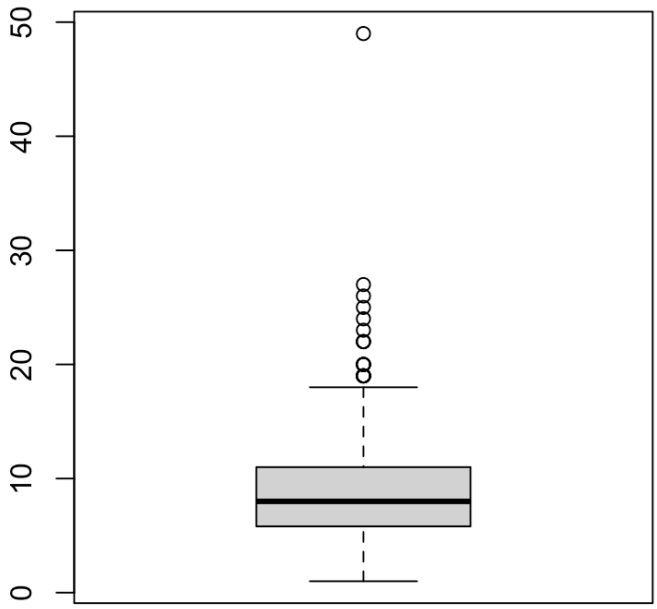
```
              product_id                    price             retail_price              merchant_id
                       0                        0                        0                        0
         merchant_rating     merchant_rating_count          origin_country          inventory_total
                       0                        0                        0                        0
     countries_shipped_to      shipping_option_name shipping_option_price           product_color
                       0                        0                        0                        0
product_variation_size_id product_variation_inventory                   rating             rating_count
                       0                        0                        0                        0
       rating_one_count         rating_two_count       rating_three_count        rating_four_count
                       0                        0                        0                        0
       rating_five_count              units_sold                     tags
                       0                        0                        0
```

### 3.2. Identificación y tratamiento de valores extremos.

Los valores extremos son valores anómalos con respecto al conjunto de datos. En este sentido, el análisis de valores extremos y su posterior tratamiento solo tiene sentido en el caso de los campos que son de tipo numérico. Es por ello que dicho análisis se va a realizar sobre los siguientes campos: **price, retail_price, merchant_rating, merchant_rating_count, inventory_total, countries_shipped_to, shipping_option_price, product_variation_inventory, rating, rating_count, rating_one_count, rating_two_count, rating_three_count, rating_four_count, rating_five_count, units_sold**.

Estos valores perdidos se podrían identificar visualmente con el gráfico boxplot como se ve un ejemplo a continuación.

```
#Valores perdidos
boxplot(summerClothes$price)
```




En el gráfico anterior se ve claramente el valor perdido de la variable *Price*. Señalar que en caso particular que vemos en el gráfico anterior, no podemos

descartar ningún valor como valor extremo, puesto que claramente es válido que un producto sobrepase los 10 euros y pueda llegar a los 50 euros de valor. Es verdad, que la mayoría de los productos están entre 5 y 12 euros, pero precios más altos no deben descartarse tampoco.

Para ser prácticos y no acumular gráficos, utilizaremos la función boxplot.stats() para el análisis del resto de variables.

```
boxplot.stats(summerClothes$price)$out
boxplot.stats(summerClothes$retail_price)$out
boxplot.stats(summerClothes$merchant_rating)$out
boxplot.stats(summerClothes$merchant_rating_count)$out
boxplot.stats(summerClothes$inventory_total)$out
boxplot.stats(summerClothes$countries_shipped_to)$out
boxplot.stats(summerClothes$shipping_option_price)$out
boxplot.stats(summerClothes$product_variation_inventory)$out
boxplot.stats(summerClothes$rating)$out
boxplot.stats(summerClothes$rating_count)$out
boxplot.stats(summerClothes$rating_one_count)$out
boxplot.stats(summerClothes$rating_two_count)$out
boxplot.stats(summerClothes$rating_three_count)$out
boxplot.stats(summerClothes$rating_four_count)$out
boxplot.stats(summerClothes$rating_five_count)$out
boxplot.stats(summerClothes$units_sold)$out
```

Los valores obtenidos como perdidos, no debemos considerarlos como tal. De todos los valores obtenidos, no se muestran por pantalla por su volumen, ninguno de ellos es un valor no posible. Es decir, no encontramos valores negativos para puntuaciones ni valores que no sean compatibles con posible soluciones. Es por ello, que se mantienen dichos valores y no se utilizará ningún método para cambiarlo.

# Pregunta 4

**Análisis de los datos.**

**4.1. Selección de los grupos de datos que se quieren analizar/ comparar (planificación de los análisis a aplicar)**

A continuación se han definido los grupos o conjuntos de datos que pueden ser interesantes de cara al posterior análisis. En particular, se agruparan los productos vendidos por valorización de los usuarios, por precio y por disponibilidad del artículo a la hora de comprarlo. El siguiente fragmento de código permite realizar los grupos mencionados.

```
# Agrupación por valoracion
summerClothes.bien_valorados    <- summerClothes[summerClothes$rating >= 2.5,]
summerClothes.mal_valorados  <- summerClothes[summerClothes$rating < 2.5,]

# Agrupación por precio
summerClothes.precio_bajo    <- summerClothes[summerClothes$price <10,]
summerClothes.precio_medio  <- summerClothes[summerClothes$price >= 10 && summerClothes$price <20,]
summerClothes.precio_alto  <- summerClothes[summerClothes$price >= 20,]

# Agrupación por disponibilidad del producto
summerClothes.baja_disponibilidad    <- summerClothes[summerClothes$inventory_total <20,]
summerClothes.media_disponibilidad  <- summerClothes[summerClothes$inventory_total >= 20 && summerClothes$inventory_total <40,]
summerClothes.alta_disponibilidad  <- summerClothes[summerClothes$inventory_total >= 40,]
```

### 4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para la comprobación de la normalidad, señalar que todas las variables con las que se trabaja en este conjunto de datos son de tipo cuantitativo por lo que se utilizará un único tipo de proceso para comprobar la normalidad de estas. En particular se utilizará la prueba de *Anderson-Darling* ya implementada en el ejemplo de la práctica. Es decir, se quiere calcular el p-value de cada una de las variables y comprobar si es superior o inferior al nivel de significancia establecido en 0.05. En caso de ser superior, podemos afirmar que la variable sigue una distribución normal.

```
for (i in 1:ncol(summerClothes)) {
  if (i == 1)
    cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(summerClothes[,i]) | is.numeric(summerClothes[,i])) {
    p_val = ad.test(summerClothes[,i])$p.value
    if (p_val < 0.05) {
      cat(colnames(summerClothes)[i])
      # Format output
      if (i < ncol(summerClothes) - 1)
        cat(", ")
      if (i %% 3 == 0)
        cat("\n") }
  }
}
```

Lo que se ha realizado es una adaptación de dicho código al caso de particular que nos ocupa.

```
Variables que no siguen una distribución normal:
price, retail_price,
merchant_rating, merchant_rating_count,
inventory_total, countries_shipped_to,
shipping_option_price, product_variation_inventory, rating,
rating_count, rating_one_count, rating_two_count,
rating_three_count, rating_four_count, rating_five_count,
units_sold
```

En los resultados obtenidos, se observa que ninguna de las variables sigue una distribución normal.

A continuación se estudiará la homogeneidad de varianzas con ayuda de la funcionalidad ya implementada en R del test de Fligner-Killeen. La homogeneidad se estudiará entre las diferentes puntuaciones de los productos. La hipótesis nula es que todas las varianzas son iguales.

```
# Test de homogeneidad de las varianzas
fligner.test(price ~ rating, data = summerClothes)
```

```
        Fligner-Killeen test of homogeneity of variances

data:  price by rating
Fligner-Killeen:med chi-squared = 219, df = 191, p-value = 0.0805
```

El resultado del test obtiene un p-value igual a 0.0805, es decir, mayor a 0.05. Por lo que se puede concluir que aceptamos la hipótesis nula y lo por tanto afirmar que las varianzas de todos los *ratings* son homogéneas.

**4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y objetivos del estudios, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.**

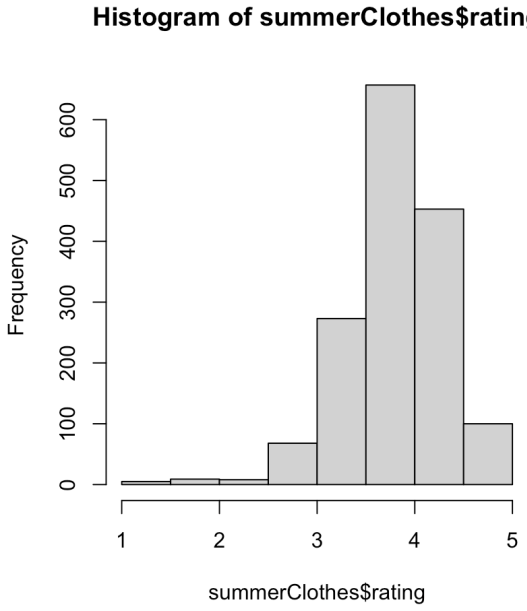
# Pregunta 5

**Representación de los resultados a partir de tablas y gráficas.**

A continuación, se van a mostrar una serie de gráficos que nos ayudaran a entender el comportamiento de los datos y serán útiles para obtener las conclusiones finales de estos.

En primer lugar, se hará uso de un histograma para mostrar la popularidad o puntuación con la que los diferentes usuario valoran la ropa vendida en la plataforma digital.

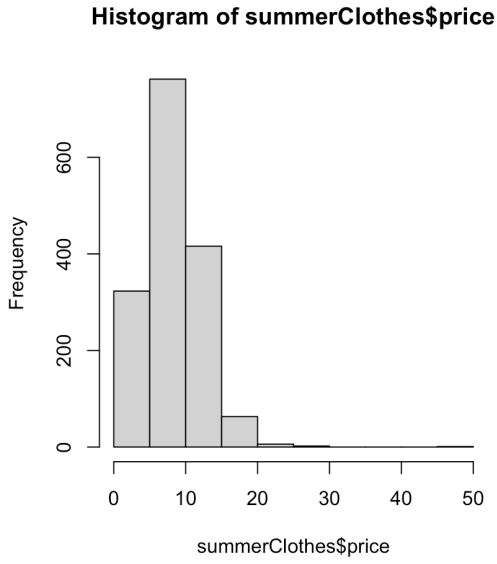
```
hist(summerClothes$rating)
```



Podemos observar que la mayoría de las puntuaciones se encuentran entre los valores 3 y 5. Teniendo en cuenta que la puntuación máxima es 5, podemos ver que la mayoría de usuarios están satisfechos con los productos adquiridos.

Seguidamente, se hará uso de un histograma para mostrar los rangos de precios de los diferentes productos vendidos en la plataforma digital.

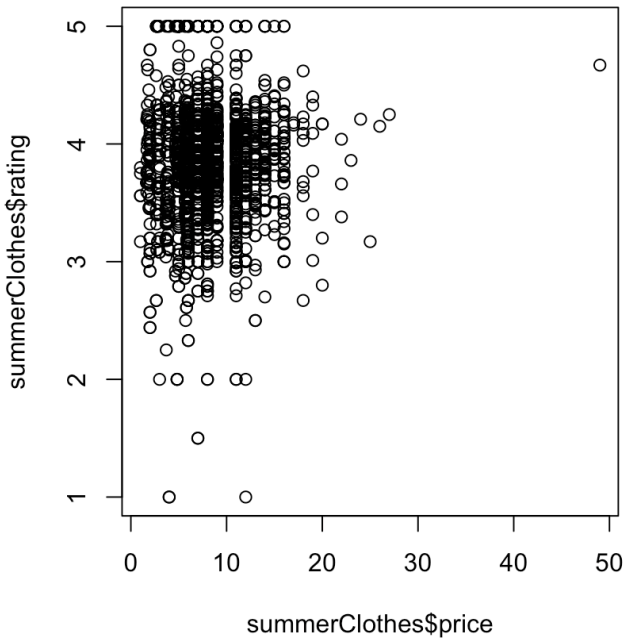
```
hist(summerClothes$price)
```




Se observa que, tal y como habíamos observado en el gráfico del apartado 3, los precios de la mayoría de los productos se encuentran en el intervalo de cero a veinte euros. Esto nos indica que es una plataforma con productos de bajo coste, que junto a la popularidad de estos que hemos observado en el gráfico anterior, nos hacer pensar que sus ventas deben ser altas.

Lo mencionado anteriormente, viene confirmado por la siguiente gráfica donde se observa la relación entre el precio y la valoración de los productos, donde los productos mejor valorados son aquellos que tienen menor precio.

```
plot(summerClothes$price, summerClothes$rating)
```




# Pregunta 6

**Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?**

Tras los resultados obtenidos tanto en el proceso de análisis como en el proceso de visualización estamos en condiciones de responder a la pregunta que nos hacíamos al principio de esta práctica.

Hemos observado que existe relación entre la ventas de productos con su precio y puntuación. Partíamos de la premisa que los datos se obtienen de una plataforma de bajo coste por lo que vemos que las ventas que estos tienen están apoyadas en mayor medida por este bajo coste de los productos. Pero además se ha detectado una nueva variable que influye en la venta y es la calificación que los diferente compradores hacen sobre un producto tras recibirlo y hacer uso de él. Vemos que cuando mayor es la puntuación mayor es la venta y esta variable es realmente importante. Los futuros vendedores de estas plataformas deben poner especial cuidado en las valoraciones que los usuarios hacen sobre sus ventas porque influyen directamente en sus futuras ventas.

## Pregunta 7

**Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.**

El código asociado a esta práctica está en el siguiente repositorio de GitHub: https://github.com/aliciarg6/Limpieza_Analisis_Datos