TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS
PRÁCTICA 2
ÁNGEL BALTASAR SÁNCHEZ

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Para tener coherencia con el esfuerzo realizado, utilizaremos el mismo dataset que obtuvimos con la primera práctica:

https://github.com/abaltasars/PRODUCTOS_BARRABES_ONLINE_10_2017

se trata del catálogo de productos de la tienda online www.barrabes.com ofertados a día 31/10/2017. Siendo una de las tiendas más completas que podemos encontrar intentaremos, a partir de los productos ofertados, establecer cual seria en precio medio y el precio mediano de adquirir un equipamiento de trekking para una travesía de un día para un montañero hombre.

En primer lugar estableceremos cuales son los productos que debemos llevar a una sesión de trekking, esto nos servirá para acotar dentro de la tienda sólo aquellos productos que necesitamos. Definiremos un conjunto de elementos necesarios , en el que seguro nos olvidamos de muchos elementos básicos que seguramente necesitaremos en la ruta. Comenzando de los pies a la cabeza necesitaremos:

- Botas de trekking
- Calcetines de trekking
- Pantalón de trekking
- Capa interior
- Capa intermedia
- Capa exterior
- Impermeable
- Gorro

2. Limpieza de los datos.

2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

De los datos que disponemos utilizaremos tres campos uno será el campo **categoría** que nos servirá para escoger los conjuntos de datos y una vez realizada esta selección tendremos en cuenta el campo **precio.** El tercer campo será **URL** y nos servirá para eliminar campos repetidos ya que tras estudiar los datos, hemos observado que un producto se repite si tiene diferentes tallas y tienen en común la URL.

En el dataset hemos detectado valores vacíos, por lo que procederemos a eliminar aquellos registros que tengan valor vacío en **precio o categoría**.

Tal como comentábamos anteriormente tenemos elementos repetidos y lo que haremos es eliminar los elementos repetidos dejando sólo la primera aparición de los registros repetidos.

Otras transformaciones que haremos son, la de convertir a UTF-8 las columnas **URL y categoría,** ya que la extracción de datos nos dejó los datos en ASCII y los caracteres especiales quedan como caracteres extraños.

En principio no deberían existir outliers ya que se tratan de los valores de los productos, pero consideraremos todos aquellos valores que estén fuera del rango media +/- 3 veces la desviación estándard como outliers, considerando que se trata de productos muy exclusivos y que no estrarán en las intenciones de compra de un montañero “medio”.

3. Análisis de los datos.
3.1. Selección de los grupos de datos que se quieren analizar/comparar.
3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.
3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Para cada uno de los productos el equipamento realizaremos una selección de datos, en contcreto crearemos las siguientes selecciones:

CalzadoHombre -> Calzado de Montaña > Hombre > Botas Trekking
CalcetinesHombre -> Calzado de Montaña > Hombre > Calcetines > Trekking
PantalonHombre -> Ropa Montaña Hombre > Pantalones > Trekking >
CamisetaHombre -> Ropa Montaña Hombre > Camisetas
CapaIntermedia -> Ropa Montaña Hombre > Chaquetas > Forros Polares
Chaqueta -> Ropa Montaña Hombre > Chaquetas
Impermeable -> Ropa Montaña Hombre > Chaquetas > Impermeables > Media Montaña
Gorros -> Ropa Montaña Hombre > Gorros y Tubulares

Estos subsets serán los que deberemos tomar en cuenta para nuestra respuesta final.

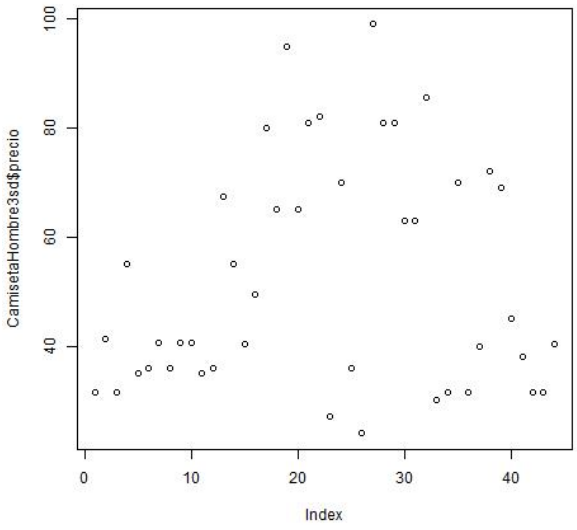
Para cada uno de estos grupos calculamos media, mediana, desviación estándar y varianza. A partir de estos cálculos calculamos los subgrupos donde eliminamos aquellos precios que están más allá de la media +/- 3 desviaciones estándares.

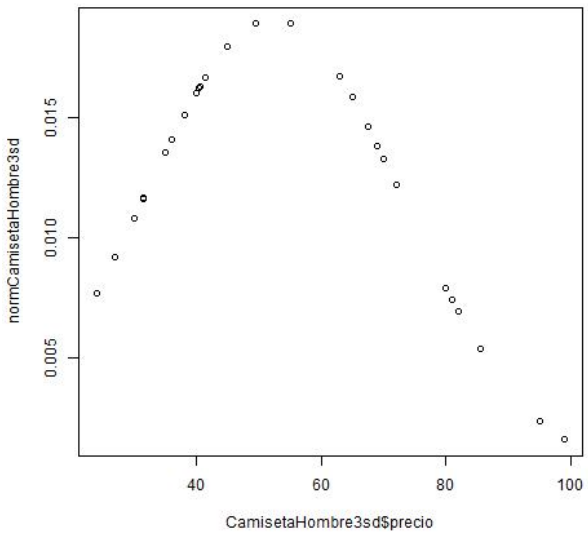
Los sumatorio de las medias y las medianas, nos darán los dos posibles precios de equipar a un montañero.

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

Para cada subset calculamos el histograma para ver cual es la distribución de los precios y también calculamos la distribución normal.

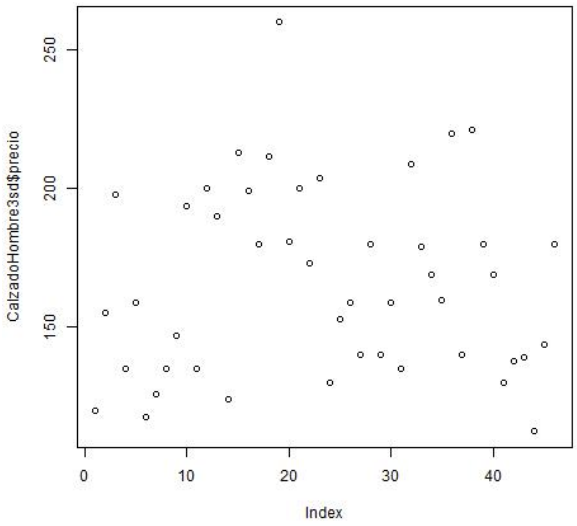
Subset: CalzadoHombre

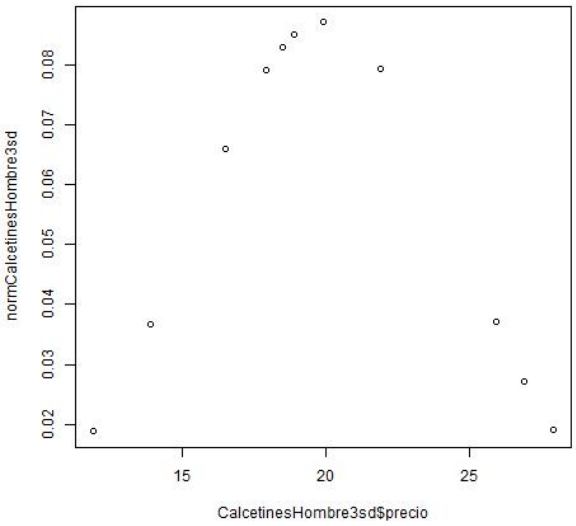




Figuras 1 y 2. Pantalones de hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

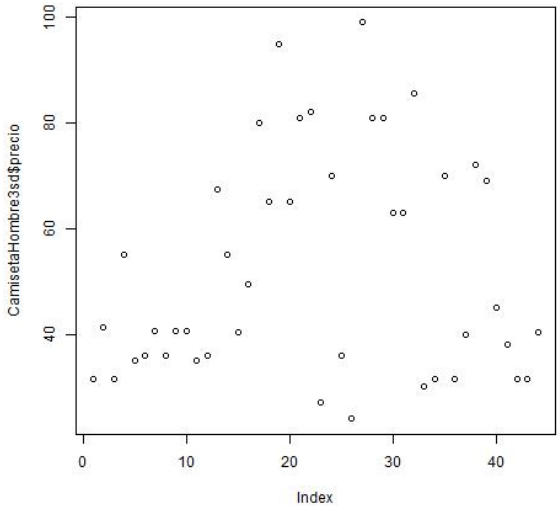
Subset: CalcetinesHombre

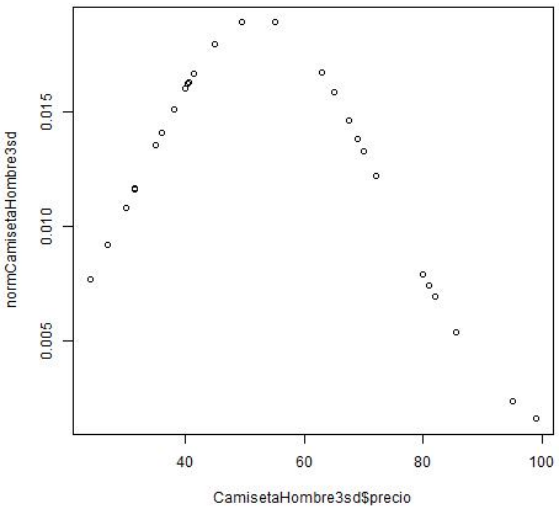




Figuras 3 y 4. Calcetines de hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

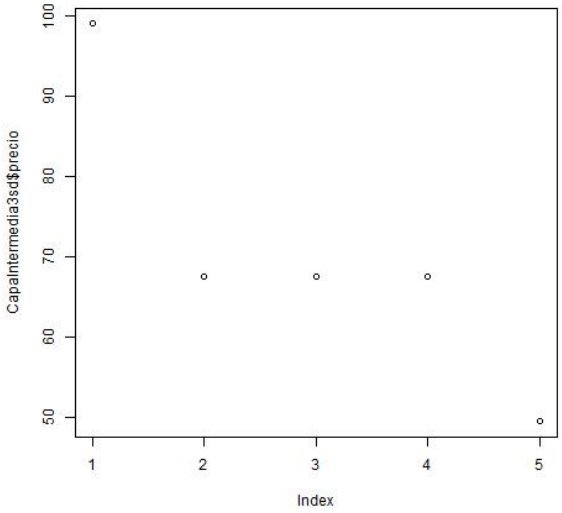
Subset: CamisetaHombre

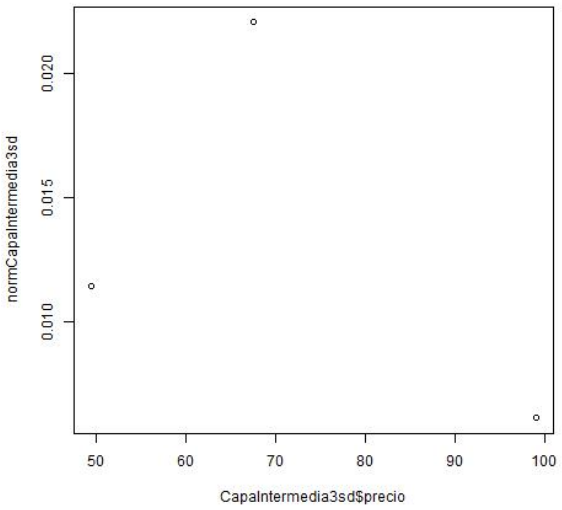




Figuras 5 y 6. Camisetas de hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

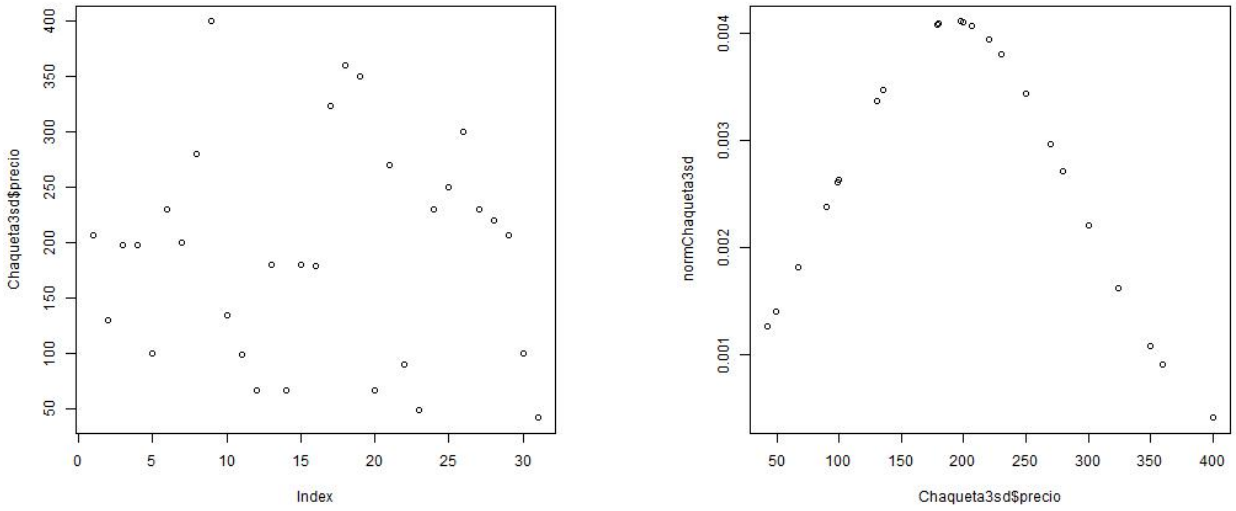
Subset: CapaIntermedia





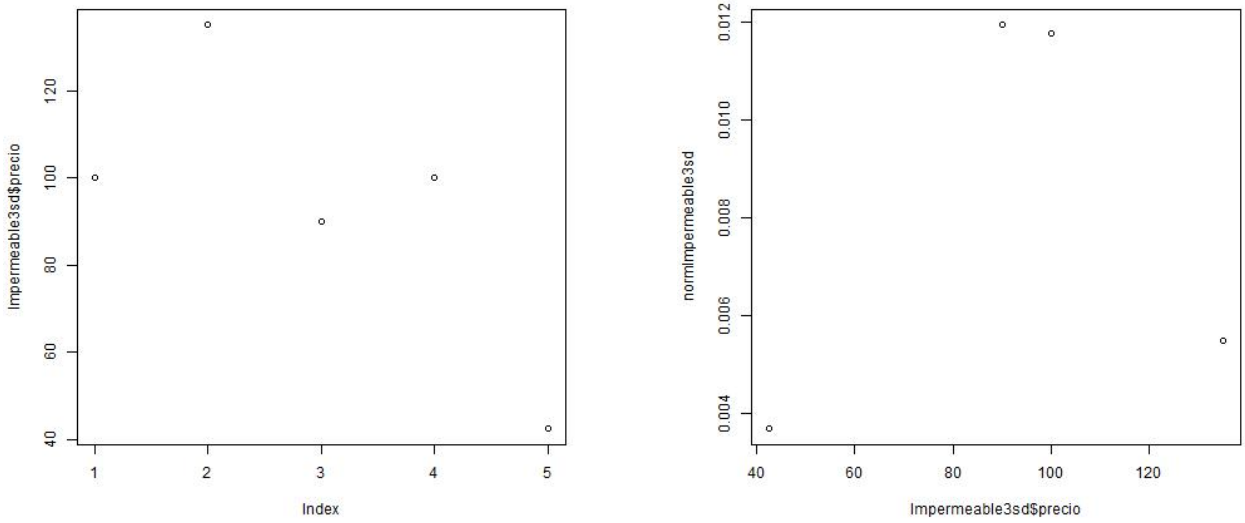
Figuras 7 y 8. Capa intermedia de hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

Subset: Chaqueta



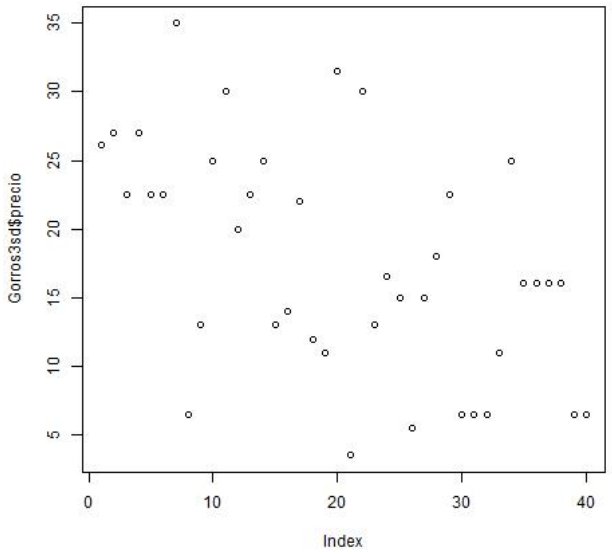
Figuras 9 y 10. Chaqueta hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

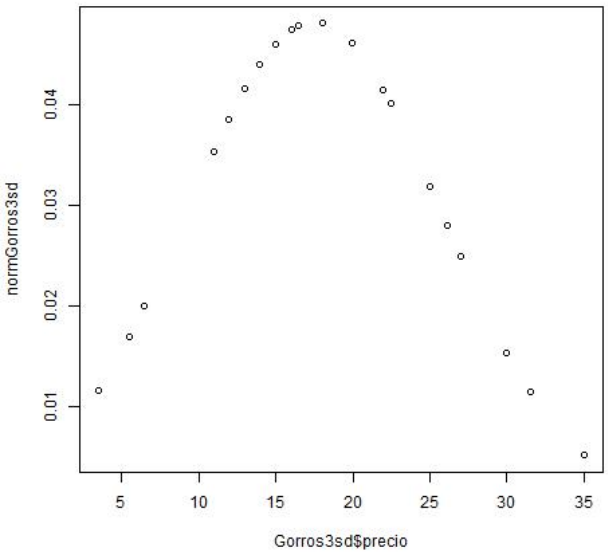
Subset: ChaquetasImpermeable



Figuras 10 y 11. Chaqueta impermeable hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

Subset: Gorros





Figuras 11 y 12. Gorros hombre. Histograma con distribución de precios, Curva normal de distribución de precios.

La tabla de datos calculados es la siguiente:

	N	Media	Mediana	Varianza	N (3sd)	Media (3sd)	Mediana (3sd)	Varianza (3sd)
CalzadoHombre	47	168,965	159,9	1497,558	46	166,117	159,45	1140,984
CalcetinesHombre	14	19,914	19,4	20,963	14	19,914	19,4	20,963
PantalonHombre	14	97,454	97,135	325,810	14	97,454	97,135	325,810
CamisetaHombre	44	52,253	40,94	436,038	44	52,253	40,94	436,038
CapaIntermedia	5	70,200	67,5	319,950	5	70,200	67,5	319,950
Chaqueta	32	202,852	199	13080,289	31	191,653	198	9369,648
Impermeable	5	93,500	100	1105,000	5	93,500	100	1105,000
Gorros	40	17,484	16	68,405	40	17,484	16	68,405

En algunos subsets, la N es muy pequeña como para que estadísticamente tengan valor, y sólo en dos de los subsets (calzadoHombre y chaqueta) hemos eliminado registros en los que hemos encontrado valores más allá del límite media +/- 3 desviaciones estándar.

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

Tras los datos obtenidos obtenemos los siguientes resultados:

	Precio segun media	Precio según Mediana
Total	708.5745	698.425

Según esto podríamos comprar el equipamento "medio" de un amante del trekking por 708,6€, aunque si tenemos en cuenta que el equipamiento "mediano" de un amante del trekking es la suma de valores reales de productos (en algún caso de la media de los valores centrales) podríamos afirmar que el precio más aproximado es el que nos da la mediana lo que implica 698,4€.

6. 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.

He desarrollado en R esta práctica. El código está adjunto en los ficheros subidos al repositorio de Github.

Nombre del fichero de código R: CodigoPRACTICA2.R

Nombre del fichero de fuente de datos:

- productos csv: lista de productos con la que realizaremos el estudio. Obtenida a partir de la práctica anterior
- categorias.csv: lista de categorias, extraída del fichero anterior, que nos servirán para escoger las diferentes prendas a estudiar.

Nombre de ficheros de resultados:

- estadisticas.csv: resultados del análisis estadísticos sobre los diferentes subsets.
- resultado.csv: resultado con el sumatorio de la media y la mediana obtenidos tras el análisis.

Link Github: https://github.com/abaltasars/Ropa_trekking_barrabes_online_31_10_2017