|  |
| --- |
|  |
| PRA2: Limpieza y análisis de datos |
| M2.851 - Tipología y ciclo de vida de los datos |

Víctor Blanes Martín  
Carlos Allo Latorre

28/05/2021

Tabla de contenido

[1. Descripción del dataset 2](#_Toc73106472)

[2. Integración y selección de los datos 2](#_Toc73106473)

[3. Limpieza de los datos. 3](#_Toc73106474)

[3.1 Ceros y elementos vacíos 3](#_Toc73106475)

[3.2 Preparación de datos. 3](#_Toc73106476)

[3.3 Identificación y tratamiento de valores extremos. 4](#_Toc73106477)

[4. Análisis de los datos. 5](#_Toc73106478)

[4.1 Selección de los grupos de datos 5](#_Toc73106479)

[4.2 Normalidad y homogeneidad de la varianza. 5](#_Toc73106480)

[4.3 Pruebas estadísticas para comparar los grupos de datos. 5](#_Toc73106481)

[5. Representación de los resultados.. 5](#_Toc73106482)

[6. Conclusiones 5](#_Toc73106483)

[7. Código. 5](#_Toc73106484)

[8. Contribuciones 5](#_Toc73106485)

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

Hoy en día, la compra de dispositivos electrónicos, en especial de portátiles, está a la orden del día y es más frecuente que nunca. A raíz del desarrollo de nuevas tecnologías, cada vez es más frecuente disponer de una oferta de productos más amplia, excesiva en ocasiones, que dificulta la toma de decisiones en cuanto a la compra de dichos productos.

En función del perfil del comprador, el desconocimiento de la gama de productos y sus características puede que los lleve a tomar una decisión de compra poco adecuada o ajustada en precio. Por este motivo, sería interesante conocer qué variables o características son más influyentes en el precio a la hora de compra de un ordenador, para que de esta manera, el comprador:

* Pueda hacerse una idea del presupuesto aproximado que tendrá el ordenador que desea en base a las características técnicas deseadas.
* Pueda conocer si es verdad que algunas marcas, como Apple, tienen un precio algo superior al resto de las marcas.
* Pueda conocer qué características son las más influyentes en el precio para en base a estas, poder centrarse en lo que realmente necesita para incrementar o decrementar su presupuesto.

Para ello, se podría usar el dataSet de ordenadores portátiles construido en la práctica 1, en donde, entre otros objetivos a responder, el recién presentado era uno de ellos. Sin embargo, tras un comienzo con este dataSet, surgió el problema de que no poseían suficientes datos ni características como para realizar buenos análisis ni modelos, por lo que se decidió buscar bajo el mismo objetivo a tratar, otro dataSet. Tras la búsqueda se seleccionó el conjunto de datos “Laptop Prices” de Kaggle (<https://www.kaggle.com/ionaskel/laptop-prices>), que posee 1303 registros de ordenadores con sus correspondientes características y precio.

# 2. Integración y selección de los datos de interés a analizar.

La integración de los datos será mínima ya que todos los datos a usar provienen de un mismo dataset. Como se usará el lenguaje de programación Python, en términos de programación la única importación que se realizará será la de carga del dataSet al entorno, que se realizará mediante pandas apuntando al archivo comentado, que ha sido importado al GitHub del proyecto. Cabe resaltar, que si se deseara usar dos dataSets diferentes con las mismas características, se debería de realizar una integración de los mismos en donde habría que tener en cuenta posibles repeticiones de objetos, que las propiedades se presenten en las mismas unidades... Este proceso se realizó en la práctica 1, en el momento en el que se integraban dos dataSets diferentes (uno de cada web en donde se realizó WebScraping), en uno sólo.

Respecto a la selección de los datos, nos quedaremos con todas las filas, ya que cada una de ellas corresponde a un ordenador diferente y aporta información al estudio que se está realizando. Ademas, nos encontramos ante un número de ordenadores (1003) no tan grande como para tener que hacer reducción de la cantidad. Sin embargo, para el caso de tener que realizar el mismo, consideramos que las dos mejores formas de hacerlo serían el método de muestra aleatoria simple sin sustitución (para no tener repeticiones de ordenadores en el dataSet resultante), o muestra de clústeres, en donde cada cluster podría estar correspondido por la marca o por intervalos de precio para asegurar que tenemos muestras de todos los precios.

Sobre las columnas, encontramos la primera de ellas que no nos da ninguna información útil para el estudio y ser un simple id ascendente, con lo cual la eliminaremos. Haremos lo mismo con la columna que se refiere al producto, pues el objetivo en todo momento es realizar una comparación en base a características o marcas, pero no en base al modelo del portátil directamente.

Por tanto, tras esta limpieza, las columnas que resultarán del dataSet junto con su significado según se proporciona en el repositorio original serán:

* Company --> Company Name
* TypeName --> Laptop Type
* Inches --> Screen Inches
* ScreenResolution --> Screen Resolution
* Cpu --> CPU Model
* Ram --> RAM Characteristics
* Memory --> Memory
* Gpu --> GPU Characteristics
* OpSys --> Operating System
* Weight --> Laptop's Weight
* Price\_euros --> Laptop's Price

# 3. Limpieza de los datos.

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

Tras la carga inicial, se aprecia cómo no se encuentra ningún valor vacío o cero. Sin embargo, tras el proceso de limpieza que se comentará posteriormente, se encuentra en la variable “ScreenResolution\_Type” que será una creada a partir de la columna ScreenResolution, se encuentran algunos valores nulos al no presentarse el tipo explícitamente en la variable inicial. Para su tratamiento, se realizará el método de indiciar los valores perdidos como la sustitución por una misma constante o etiqueta, en este caso, "Unknown". Esto se hace ya que falta un gran número de registro (un 33%), y el uso de otras técnicas como la moda harían que tuvieramos muchísimos datos 'no reales'. Otras técnicas basadas en modelos que lo predicen, tampoco las vemos oportunadas, ya que siendo un atributo como es el tipo de resolución, podría darse el caso de que dos ordenadores tuvieran las mismas características, y sin embargo, un tipo de resolución diferencia.

Con respecto al resto de variables, tras realizar un pequeño estudio, se aprecia que no hay datos perdidos o que indiquen la pérdida de valor. Si bien es verdad que para sistema operativo encontramos 'No OS', que puede dar lugar a confusión. Sin embargo, este valor es válido, ya que algunos ordenadores pueden no tener sistema operativo integradado, y por tanto, posiblemente, abarate el coste del mismo.

## 3.2 Preparación de datos.

Nos hemos sentido libres de añadir este apartado, por el hecho de que antes de pasar al apartado 4 en donde se analizan más profundamente los datos y se elaboran modelos o incluso del apartado 3.3 consistente en la búsqueda de valores extremos, se ha considera que es necesario un proceso de preprocesado o data cleaning de los mismos.

Para ello, se han centrado los esfuerzos en variables como Ram o Weight, que son variables que se han de tratar como numéricas pero que al tener la unidad de medida en cada uno de sus valores, dificulta su tratamiento. Para ello, se ha realizado una comprobación de cada una de las medidas, y se ha verificado que tan solo aparecían las medidas de Gb y Kg y se ha trasladado este valor a las columnas, quedando estas en Ram(GB) y Weight(Kg). Siguiendo en la misma línea, el tipo de memoria se aprecia que incluye tanto medidas individuales (256GB SSD) como combinadas (128GB SSD + 1TB HDD) de las medidas de memoria que se presentan en 4 tipos: SSD, Flash Storage, HDD e Hybrid . Para tratar este hecho, se transformarán todas las medidas a GB y se crearán cuatro nuevas columnas que podrán tratarse como numéricas, siendo estas MemorySSD(GB), MemoryHDD(GB), MemoryFlash(GB), MemoryHybrid(GB).

Para en el caso de la CPU, para poder análisis más al detalle, se ha encontrado el patrón de esta columna en que en primer lugar se da la empresa que la proporciona, en segundo la versión que se proporciona y por último, la velocidad de la misma. Por tanto, realizando este splitado, resultan las columnas CPU\_Company, CPU\_Version, CPU\_Speed(GHz).

En la misma línea con la GPU, resulta claro ver como esta presenta la estructura de vendedor de la GPU seguido por la versión que proporciona, dando lugar a las nuevas columnas GPU\_Company, GPU\_Version.

Finalmente, con la resolución de pantalla se observa el patrón de en un primer lugar como parte opcional el tipo de resolución, seguido por el tamaño de pantalla con el patrón alto x ancho, que se separarán en dos columnas. De esta forma, resultarán 3 columnas más, siendo estas: ScreenResolution\_Type, ScreenResolution\_Width, ScreenResolution\_High.

Una vez se ha llegado a este punto, ya se poseen datos más preparados para poder realizar análisis posteriores, y muchas variables numéricas listas para trabajar con ellas. Antes, aunque estaban presentes también, el incluir la unidad de medida de las mismas o el estar en un string con otra información, no se permitía su uso para hacer buenos análisis o modelos al tratar muchas de ellas como variables categóricas y no numéricas.

## 3.3 Identificación y tratamiento de valores extremos.

Se analizará en primer lugar la variable precio, variable numérica más importante de nuestro estudio. Se aprecia como aparentemente, aunque el percentil 50 está en 977 y su 75 en 1487, la media es de 1123, lo que es indicio de que, tras el percentil 75 encontraremos algún valor más elevado, hasta llegar al máximo de 6099.

Gracias a las representaciones realizadas en el Notebook, Se aprecia como hay una gran concentración de datos en la zona en torno a 1000 euros, y que va disminuyendo hasta llegar hasta los 3000, en donde los precios a partir de este punto empiezan a encontrarse más distantes hasta llegar a los 6000, que la podría catalogarse como punto alejado o outlier.

Sin embargo, al analizar estos puntos ‘alejados’ para estudiar si realmente son puntos outliers que hay que eliminar o si son datos posibles, se aprecia como sin duda, en la gran mayoría encontramos en esta sección ordenadores Gaming, los cuales necesitan una capacidad mayores que los ordenadores estándar para el día a día. Además, vemos por ejemplo que la diferencia entre los dos últimos reside en la diferencia de la memoria SSD, ya que las demás características son iguales. Por lo general, vemos que todos estos ordenadores, presentan una CPU y GPU muy potente, que consultando el precio de estas en el mercado, parece ser un elemento que aumentará bastante el precio. Por lo tanto, no se puede considerar como puntos a tratar ya que corresponden a una pequeña parte del dataset, que por su tipología, han de mantenerse en el dataSet.

Además, realizando el mismo estudio para el caso del peso, se aprecia que ocurre algo similar, en donde algunos ordenadores se alejan bastante de la mayoría por la parte superior. Si analizamos cuáles son estos ordenadores, corresponden a ordenadores gaming, por lo que se confirma de nuevo que estos ordenadores forman un subgrupo que, aunque sus precios, peso u otras características se alejan del resto, son datos completamente válidos y que se han de tratar tal y cómo son.

# 4. Análisis de los datos.

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

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

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

# 5. Representación de los resultados a partir de tablas y gráficas..

DUDA: Qué se espera exactamente de este apartado? En el Notebook que se ha desarrollado la práctica se han elaborado varias gráficas e incluso comentado y analizada las mismas… Consiste en introducir estas aquí?

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

Mencionar que ordenadores gaming aumentan precio al tener características mejores.

Posible modelo adicional que se ajuste mejor.

Hipótesis ordenadores Apple.

# 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.

Se puede encontrar el código desarrollado en el siguiente enlace de GitHub, dentro de la carpeta src. Aquí, se encuentra el Notebook desarrollado en formato .ipynb, al igual que un archivo con extensión .html para facilitar su corrección.

# 8. Contribuciones

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa teórica (lectura material UOC, tutoriales y documentación de limpieza de datos, revisión ejemplos anteriores, etc.) | VB, CA |
| Investigación y elección dataSet (estudio de sus características e idoneidad para llevar a cabo los objetivos). | VB, CA |
| Desarrollo limpieza de datos | VB, CA |
| Desarrollo análisis de los datos | VB, CA |
| Elaboración informe | VB, CA |