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

Víctor Blanes Martín  
Carlos Allo Latorre

06/06/2021

Tabla de contenido

[1. Descripción del dataset. 2](#_Toc73993420)

[2. Integración y selección 2](#_Toc73993421)

[3. Limpieza de los datos. 4](#_Toc73993422)

[3.1 Elementos vacíos 4](#_Toc73993423)

[3.2 Preparación de datos. 5](#_Toc73993424)

[3.3 Valores extremos. 9](#_Toc73993425)

[4. Análisis de los datos. 11](#_Toc73993426)

[4.0 Resumen del dataSet 11](#_Toc73993427)

[4.1 Selección de los grupos de datos 13](#_Toc73993428)

[4.2 Normalidad y homogeneidad de la varianza. 14](#_Toc73993429)

[4.3 Pruebas estadísticas 14](#_Toc73993430)

[Contraste de hipótesis 14](#_Toc73993431)

[Estudio de variables numéricas 16](#_Toc73993432)

[Estudio de variables categóricas 17](#_Toc73993433)

[Creación y evaluación del modelo 21](#_Toc73993434)

[5. Representación de los resultados 23](#_Toc73993435)

[6. Resolución del problema 23](#_Toc73993436)

[7. Código 23](#_Toc73993437)

[8. Contribuciones 24](#_Toc73993438)

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

Hoy en día, la compra de dispositivos electrónicos y 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:

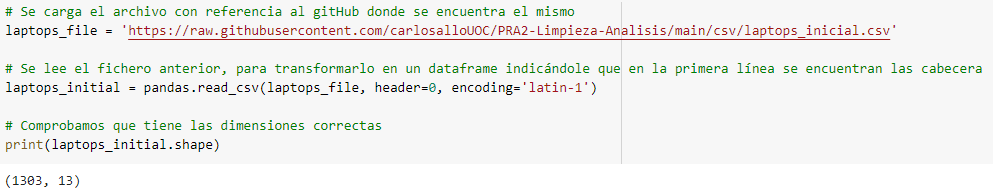
* Pueda hacerse una idea del presupuesto aproximado que tendrá el ordenador que desea en base a las características técnicas deseadas.
* Pueda conocer si es verdad que algunas marcas, como Apple, tienen un precio algo superior al resto de las marcas.
* Pueda conocer qué características son las más influyentes 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.

En el apartado 4, una vez los datos han sido preparados y limpiados, se mostrará visualmente y de una manera gráfica el dataset para entender de qué variables se parte para poder hacer un buen análisis posterior.

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

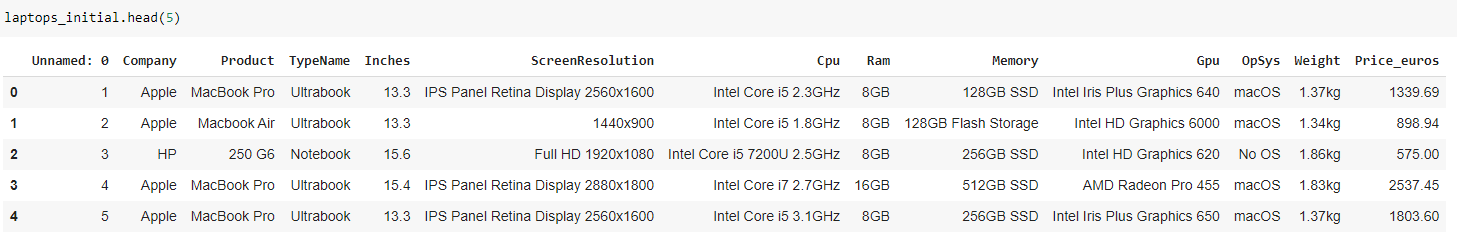
La integración de los datos será mínima ya que todos los datos a usar provienen de un mismo dataset. Como se usará el lenguaje de programación Python, en términos de programación la única importación que se realizará será la de carga del dataset al entorno, que se realizará mediante pandas apuntando al archivo comentado, que ha sido importado al GitHub del proyecto.



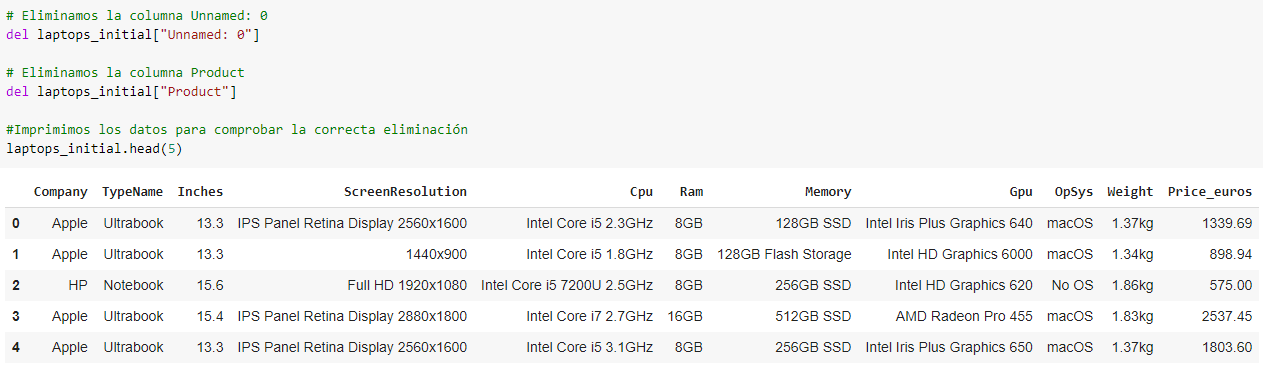
Cabe resaltar, que si se deseara usar dos datasets diferentes con las mismas características, se debería de realizar una integración de estos, donde habría que tener en cuenta aspectos como posibles repeticiones de objetos, que las propiedades se presenten en las mismas unidades... Este proceso se realizó 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. Además, nos encontramos ante un número de ordenadores (aproximadamente 1300) no tan grande como para tener que hacer reducción de la cantidad. Sin embargo, para el caso de tener que aplicar dicha técnica, 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, donde cada clúster podría estar correspondido por la marca o por intervalos de precio para asegurar que tenemos muestras de todos los precios.

Sobre las columnas podemos obtener un vistazo rápido de las mismas imprimiendo las 5 primeras columnas:



Aquí encontramos que 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 modelo del 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.

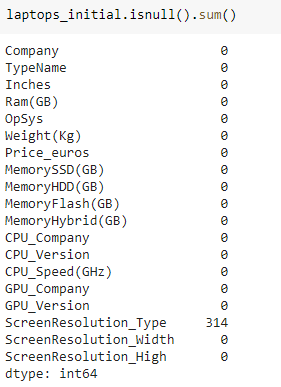


Por tanto, tras esta limpieza, las columnas que resultarán del dataset junto con su significado según se proporciona en el repositorio original serán:

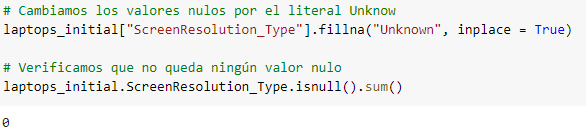
* Company --> Company Name
* TypeName --> Laptop Type
* Inches --> Screen Inches
* ScreenResolution --> Screen Resolution
* Cpu --> CPU Model
* Ram --> RAM Characteristics
* Memory --> Memory
* Gpu --> GPU Characteristics
* OpSys --> Operating System
* Weight --> Laptop's Weight
* Price\_euros --> Laptop's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

# 3. Limpieza de los datos.

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

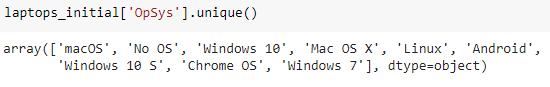
Tras la carga inicial, se aprecia como no se encuentra ningún valor vacío o cero. Sin embargo, tras el proceso de limpieza que se comentará posteriormente, se encuentran valores nulos en la variable “ScreenResolution\_Type”, que será creada a partir de la columna ScreenResolution, al no presentarse el tipo explícitamente en la variable inicial.

Para tratar estos nulos, se sustituirán los valores perdidos por una misma constante o etiqueta, en este caso, "Unknown". Se realiza de tal forma ya que falta un gran número de registros (un 33%), y el uso de otras técnicas como la sustitución por la moda o mediana harían que tuviéramos muchísimos datos 'no reales'.



Asimismo, tampoco tiene sentido aplicar técnicas de sustitución basadas en modelos de predicción, ya que la resolución de la pantalla no es una característica que dependa del resto de variables de la muestra.

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.

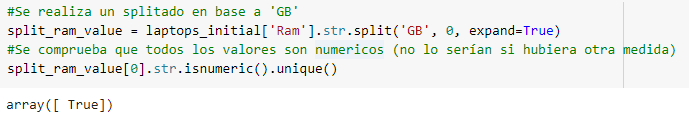


Sin embargo, este valor es válido, ya que algunos ordenadores pueden no tener sistema operativo preinstalado y, posiblemente, esto sea algo que abarate el coste del mismo.

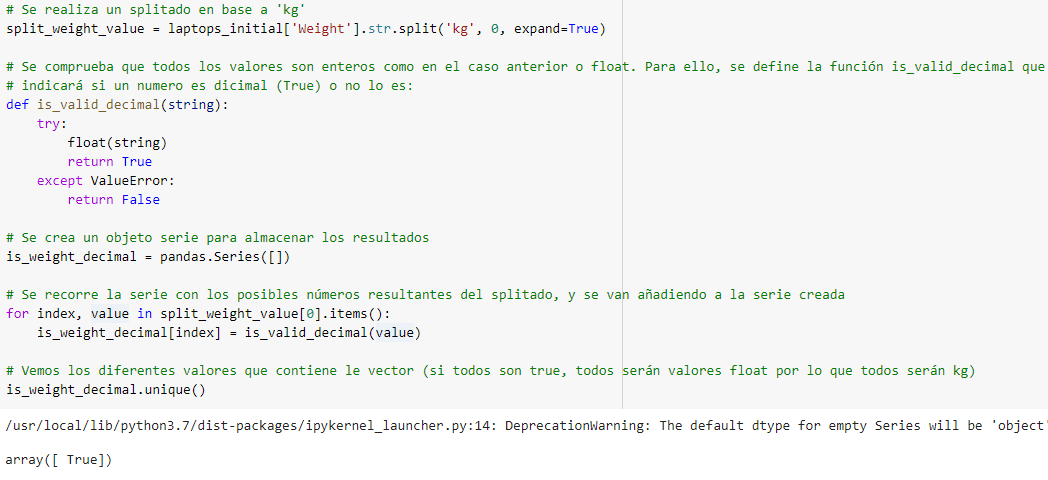
## 3.2 Preparación de datos.

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 primer lugar 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 comprobado en primer lugar que todos los valores son numéricos y que estaban indicados en las mismas unidades, GB y Kg respectivamente. Para el caso de la RAM, tendríamos el siguiente fragmento de código:



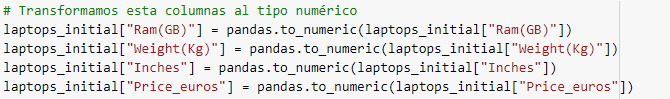
Mientras que, para el caso del peso, es algo más complejo al tener decimales, pero se persigue la misma idea:



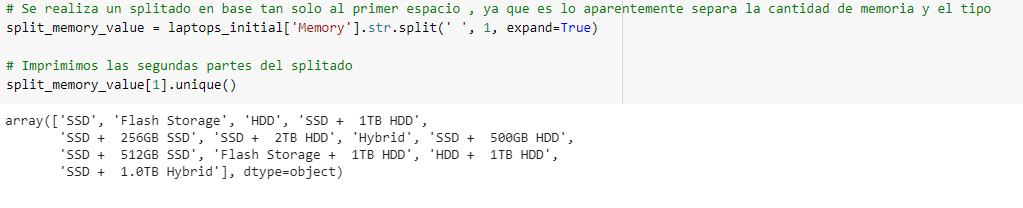
Una vez comprobados que todos los valores son numéricos y contienen la misma medida, se ha dejado únicamente el valor numérico como contenido del campo. Además, se ha modificado el nombre de las columnas para indicar las unidades Ram(GB) y Weight(Kg).



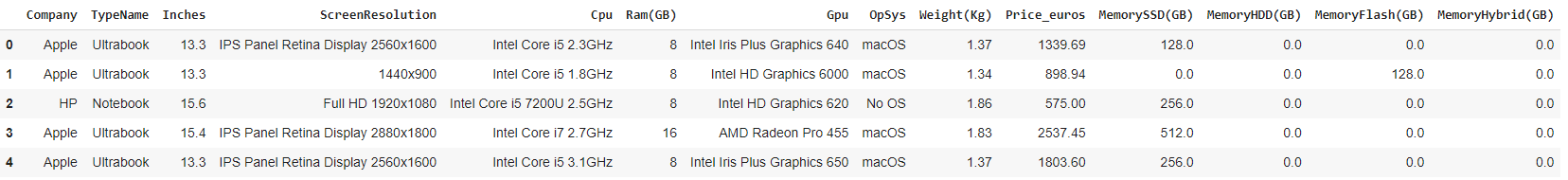
Finalmente en este proceso, se han dejado las columnas en formato numérico para que de esta manera, puedan ser tratados correctamente en los posteriores análisis que se harán:



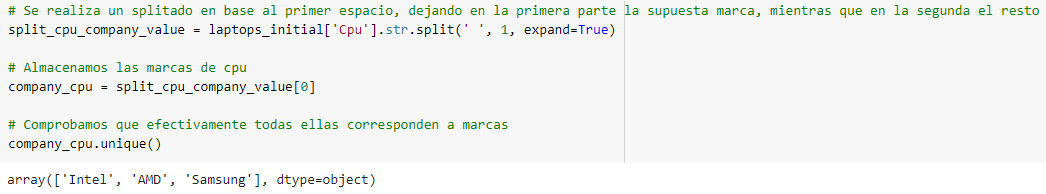
Siguiendo en la misma línea, se ha realizado un análisis del contenido de la variable memoria RAM. Se aprecian memorias individuales y también compuestas (híbridas) por varios tamaños (GB y TB) y tipos de tecnologías (SSD, Flash Storage, HDD e Hybrid):



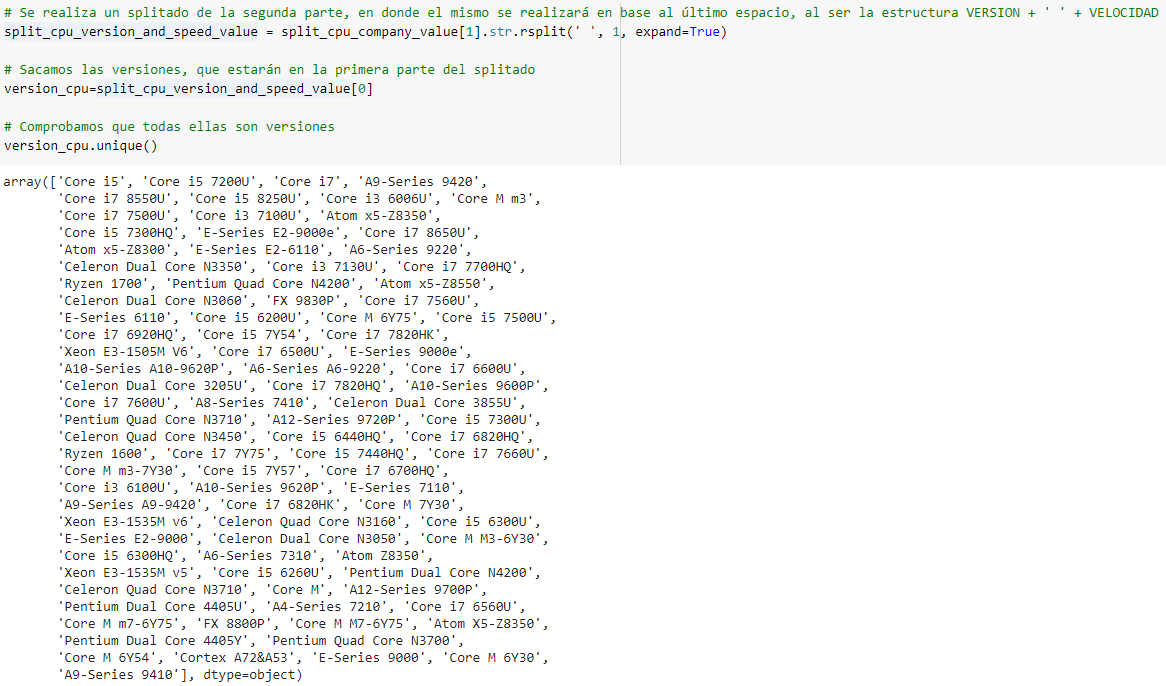
Para tratar este hecho, se ha hecho una función bastante larga (que no se adjunta imagen al ser demasiado larga pero se puede consultar en el apartado código si se desea), transformado todas las medidas a GB y se han creado cuatro nuevas columnas cuyo contenido será numérico: MemorySSD(GB), MemoryHDD(GB), MemoryFlash(GB), MemoryHybrid(GB). Se consigue por tanto, que el dataSet en este punto presente la siguiente estructura:



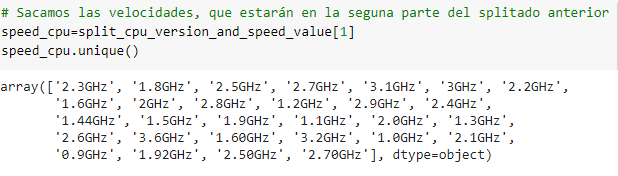
Para el caso de la CPU, se ha identificado que el patrón que sigue esta columna es el siguiente: en primer lugar se da el fabricante, en segundo lugar la versión que se proporciona y por último, la velocidad de reloj de la misma. El proceso seguido para conseguir su separación, ha sido en primer lugar, separar el string en base al primer espacio para conseguir la marca:



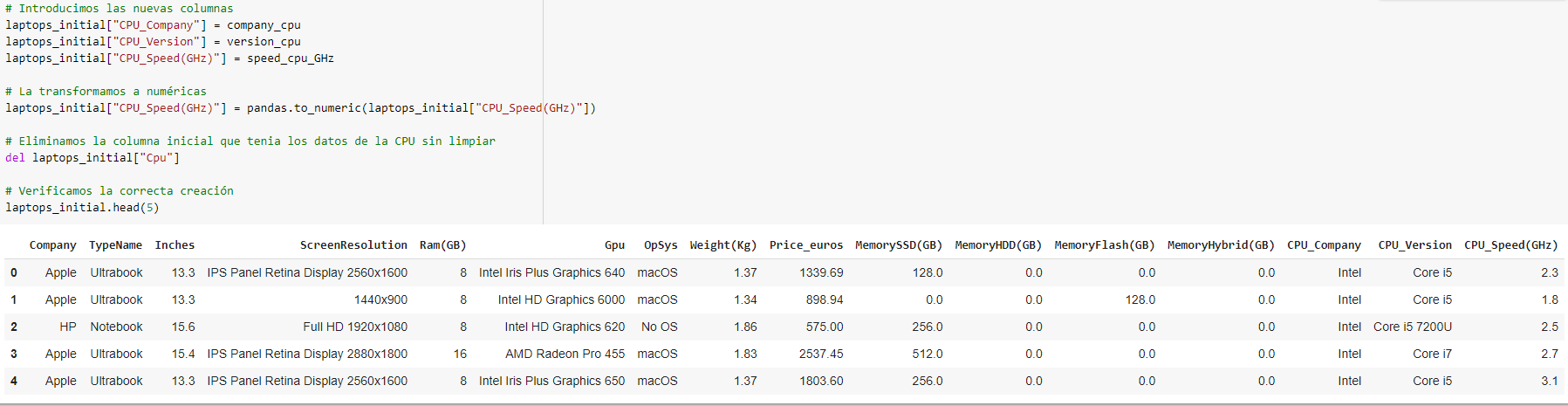
En segundo lugar, con la segunda parte de la separación anterior, se separará la versión y la velocidad realizando en este caso el splitado en el último espacio. Así pues, el proceso para obtener la versión será:

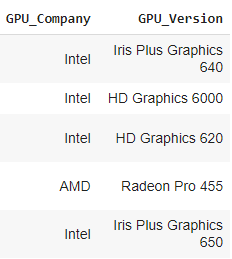


Y de la parte resultante corresponderá a la velocidad:

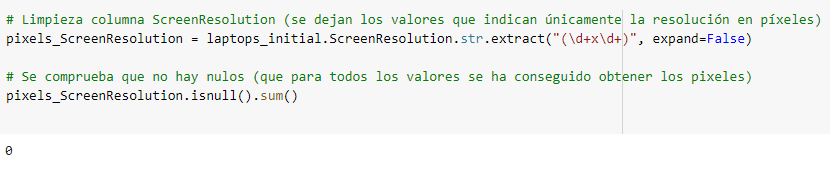


Que como se aprecia, presenta todas las unidades de medida en GHz, por lo que se realizará la separación de la medida como en el apartado anterior. Por tanto, realizando este una separación de estos 3 campos, resultan las columnas CPU\_Company, CPU\_Version, CPU\_Speed(GHz).

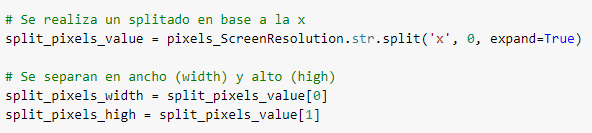


En la misma línea con la GPU, resulta claro ver cómo esta presenta la estructura de fabricante de la GPU seguido por la versión, por lo que da lugar a las nuevas columnas GPU\_Company, GPU\_Version. Extrapolando el método anterior, y para no sobrecargar el informe de código parecido, se muestran las columnas añadidas.

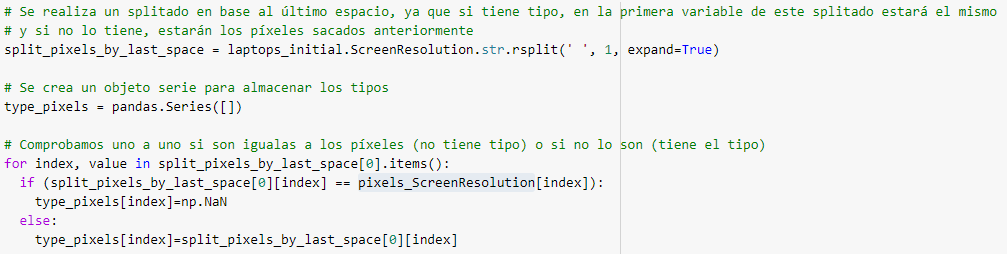
Finalmente, con la resolución de pantalla se observa el siguiente patrón: en primer lugar, en algunas ocasiones, se describe cualitativamente la resolución de la pantalla (p.ej.: Full-HD), seguido por el tamaño de pantalla en píxeles con el patrón alto x ancho. Por tanto, se procederá a fraccionar esta variable en estos 3 atributos nuevos. Para ello, en primer lugar se extraerá el tamaño de la pantalla mediante el uso de una expresión regular:



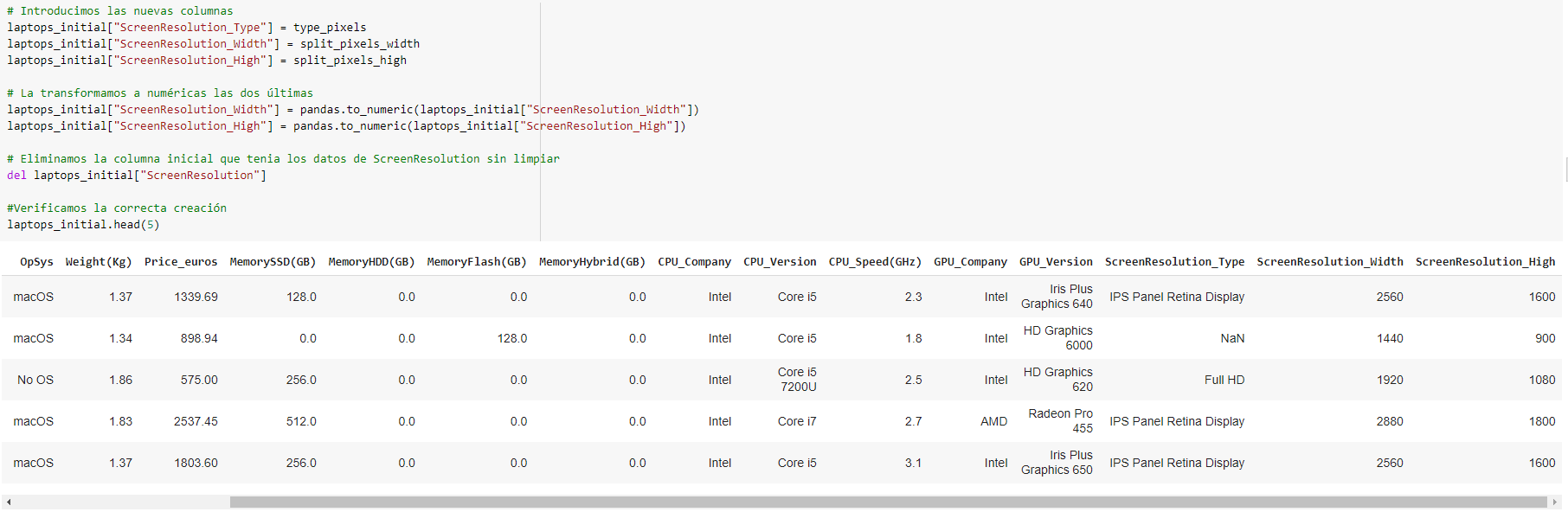
Para, seguidamente, mediante la función de splitado, dividir el valor en alto y ancho:



En segundo lugar, se procede a extraer el tipo de resolución. Para ello, se hará un splitado en base al primer espacio. En este caso, como no todos los registros poseen un valor, si en la posición 0 de este splitado contamos con la dimensión de pantalla, significará que no tiene tipo indicado, por lo que en este caso dejaremos el tipo a nulo. Esto, traducido a código será:



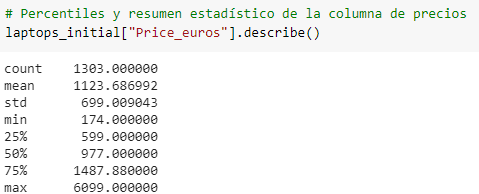
Será en esta variable donde aparezcan, como era de esperar, los valores nulos que se han comentado en el apartado anterior. Una vez tenemos los datos preparados, se crearán las 3 nuevas columnas en el dataset ScreenResolution\_Type, ScreenResolution\_Width, ScreenResolution\_High, y se eliminarán las columnas de las que se han partido.



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. De forma previa a realizar este procesamiento, muchas de las variables estaban en formato string ya que incluían la unidad de medida o información adicional en el propio valor. Esto nos impedía su uso para hacer buenos análisis o modelos al ser muchas de ellas variables categóricas y no numéricas.

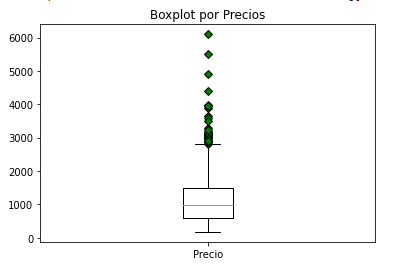
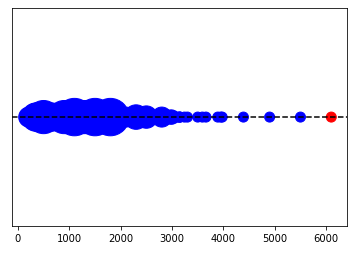
## 3.3 Identificación y tratamiento de valores extremos.

Se analiza en primer lugar la variable precio, variable numérica más importante de nuestro estudio. Se procede en primer lugar a realizar un pequeño resumen estadístico de esta variable:

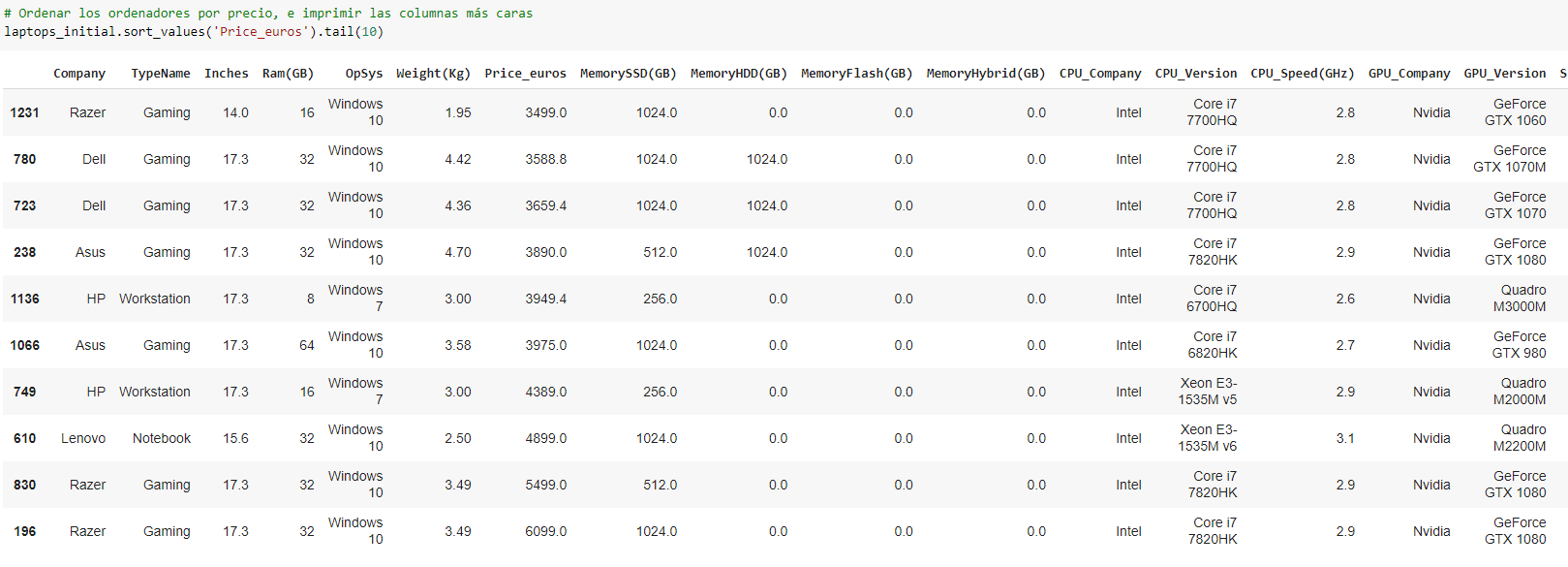


Se aprecia como aparentemente, aunque el percentil 50 está en 977 y su 75 en 1487, la media es de 1123, lo que es indicio de que, tras el percentil 75 encontraremos algún valor más elevado, hasta llegar al máximo de 6099.

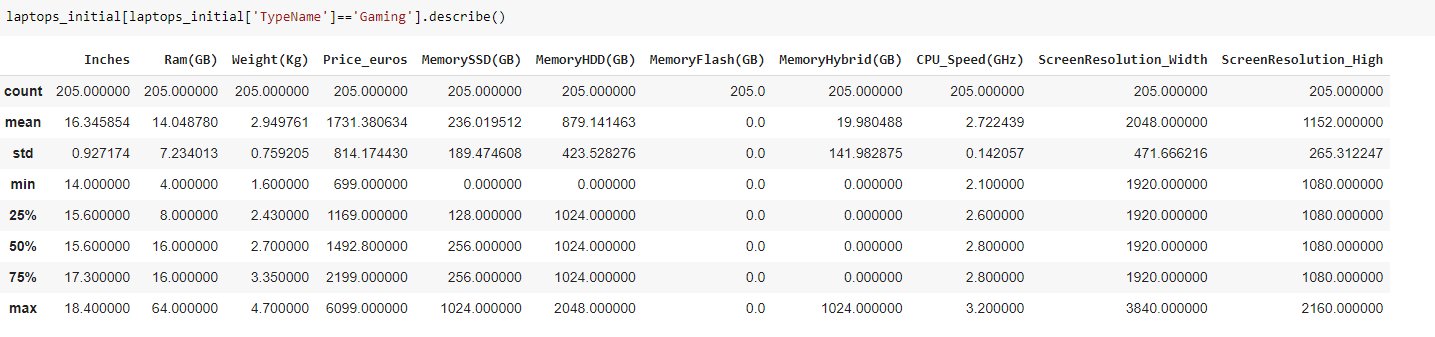
Gracias a las representaciones realizadas en el Notebook, se aprecia como hay una gran concentración de datos en la zona en torno a 1000 euros, y que va disminuyendo hasta llegar hasta los 3000, donde los precios empiezan a encontrarse más distantes entre sí hasta llegar a los 6000, dato que podríamos plantearnos como punto alejado o outlier. Las dos gráficas más representativas de este hecho, serán las que se presentan a continuación:



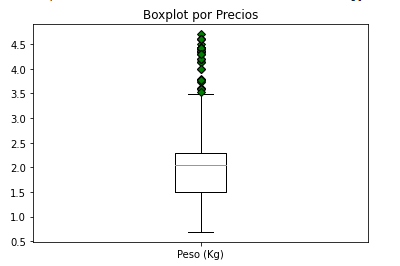
Sin embargo, al analizar estos puntos ‘alejados’ para estudiar si realmente son outliers que hay que eliminar o si son datos posibles, se aprecia que la gran mayoría de portátiles de precios elevados están catalogados como ordenadores Gaming.



Este tipo de ordenadores destacan por necesitar de una potencia gráfica y de procesamiento muy superior a la media, hecho que también hace incrementar su precio. Este hecho es fácilmente visible si realizamos un estudio estadístico de los ordenadores del tipo Gaming:



Tras una rápida consulta de precios en el mercado de ordenadores Gaming en las webs de los principales vendedores de elementos electrónicos, confirmamos que no es descabellado que se den dichos precios en este sector específico, con lo que hemos considerado que han de mantenerse en el dataset.

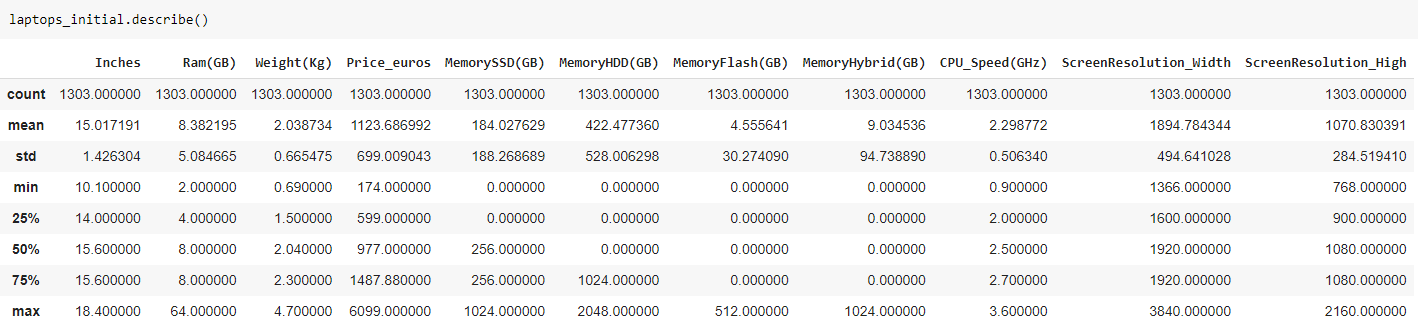
Además, realizando el mismo estudio para el caso del peso se aprecia que ocurre algo similar, donde algunos ordenadores se alejan bastante de la mayoría por la parte superior.

Si analizamos la tipología de son estos ordenadores, corresponden también a ordenadores Gaming, conocidos por unas pantallas más grandes por lo general y que requieren de un cuerpo mayor para disponer de una mayor capacidad de disipación del calor, hecho que también hace aumentar su peso. En base a estos análisis, no hemos considerado necesario considerar ningún valor del peso como outlier.

# 4. Análisis de los datos.

## 4.0 Resumen del dataSet

En un primer lugar, se realizará un estudio de las variables numéricas. Para ello, mostraremos sus valores medios, desviaciones, medias al igual que sus percentiles.

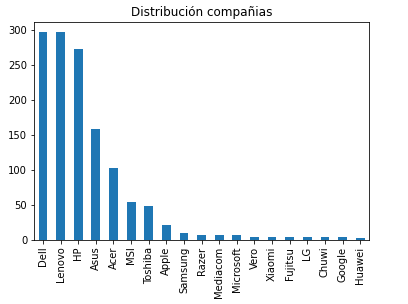


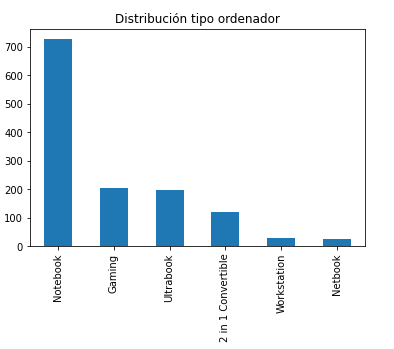
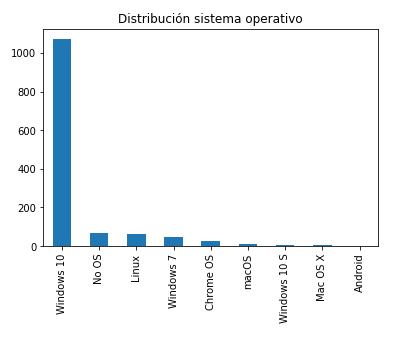
Con lo que se ha visto hasta el momento, de esta tabla resumen lo primero que llama la atención es que se puede confirmar la afirmación mostrada anteriormente. Todos los valores presentan medias y percentiles menores que los que se han dado en los ordenadores gaming, menos en la memoria Flash, ya que los ordenadores de estas características priorizan los otros tipos de memoria, en especial la SSD. Por tanto, se aprecia como lo descrito anteriormente es cierto.

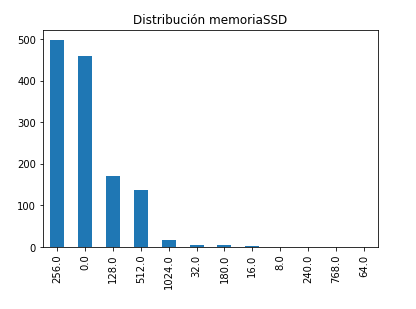
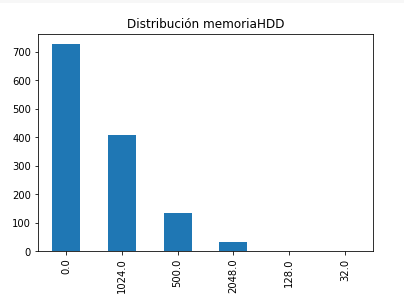
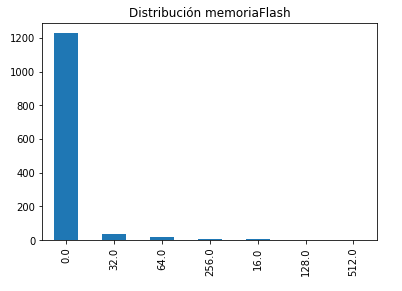
Por otro lado, si ignoramos las variables de memoria, no se producen desviaciones estándar muy elevadas, excepto en el precio (que tiene sentido al tener ordenadores de diversas marcas, características) y en la resolución de pantalla, que tiene sentido también ya que son números más elevados y es lógico que de un notebook a un ordenador gamer se encuentre esta diferencia en la pantalla.

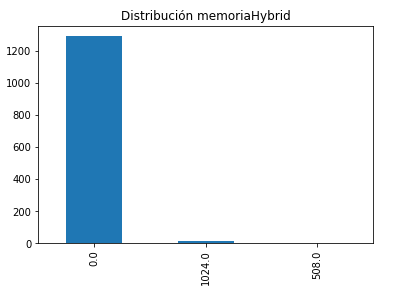
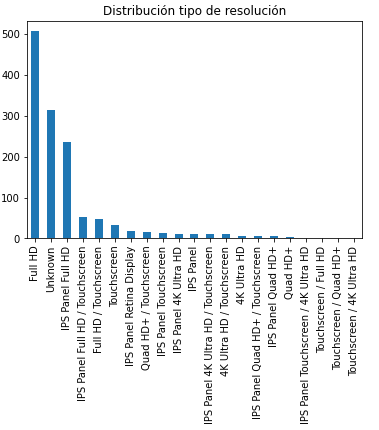
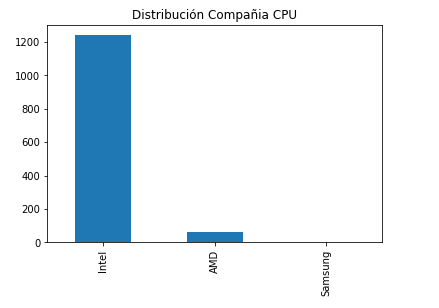
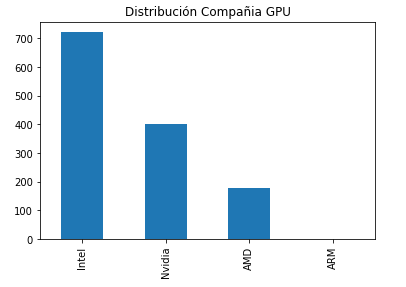
En las memorias vemos como las desviaciones son mayores (teniendo en cuenta su unidad de medida, medias y límites), y esto es así por el hecho de que algunos ordenadores poseen un tipo de memoria y otros dos. Este hecho, por la decisión que se ha tomado antes, hace que si no se dispone de un tipo de memoria, se ponga el valor de 0 para ese caso, por lo que en una columna se pueden apreciar datos algo diversos por el hecho de que un ordenador tenga una memoria de un tipo pero de otra no. Debido a este hecho, y a que estas columnas se podrían llegar a considerar discretas (las memorias pueden ser de unos valores en concreto), se harán una representación gráfica de las mismas como con las categóricas.

Para las variables categóricas, siempre que no nos encontremos con una cantidad de posibles valores desorbitada, se realizarán gráficos de barra que permiten ver de una forma más visual el tipo de datos con los que se trabaja. En caso de que se encuentren muchas más categorías, se realizará una tabla de frecuencia relativa a esa variable, hecho que ocurre tanto para las versiones de GPU como de CPU como para los tipo de resolución .

Por tanto, el conjunto de gráficas resultantes serán:

De estas se pueden extraer de primera mano varias conclusiones, como son:

- El dataset cuenta con gran cantidad de marcas de ordenadores diferentes, pero se observa una predominancia de Dell, Lenovo y HP. Posiblemente sea porque son los que más modelos tengan.

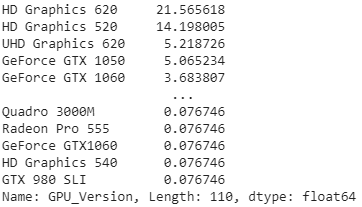
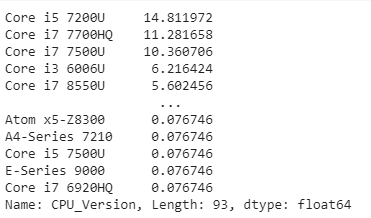
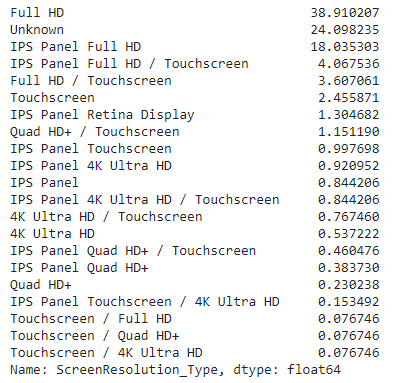
- Los ordenadores de tipo Notebok (al ser los más comunes en el día a día), son los que más porción ocupan del dataSet.

- El sistema operativo por excelencia es Windows 10, seguido muy de lejos por ordenadores libres y a continuación Linux.

- En el caso de contar memoria SSD (que más de la mitad cuentan con este tipo de memoria), la misma es de 256 GB. En el caso de tener HDD, es 1024GB, aunque predominan los ordenadores sin este tipo de memoria (habitualmente es porque los ordenadores, como se ha podido ver en la limpieza, es más habitualmente que posean memoria SSD que HDD al ser esta más rápida). Para los casos de flash e Hybrid, no la gran mayoría no disponen de esta memoria.

- Se puede afirmar, que Intel es la marca líder en el mercado en “darle vida al ordanador”. Esta marca es la líder en tanto en GPU como en CPU. En GPU, es seguida por Nvidia, sistema operativo presente en la gran mayoría de ordenadores Gaming como se puede apreciar en la tabla del apartado anterior.

Para las variables que no se han mostrado gráficamente, se muestran las tablas de frecuencia para las mismas (en algunas al tener tantos valores solo se mostrarán los máximos y mínimos).

Respecto a estas tablas de frecuencias, se aprecia como ha gran cantidad de versiones tanto de CPU como de GPU. Estos son principalmente el core i5 7200U, que se lleva casi el 15% de ordenadores (este hecho puede ser debido a que es una versión de CPU de calidad baja válida para trabajos diarios), seguidos por CPU de calidad algo mayores al hablar de core i7. Para el cado de las GPU, predominan los HD Graphics 620 y 520, llevándose estas el 35% de los ordenadores. Finalmente, con respecto al tipo de resolución predomina la HD en su mayor parte, aunque en un 25% de las ocasiones, no se ha indicado el tipo de la misma como se ha analizado en el apartado de limpieza.

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

A lo largo de este apartado y los siguientes, se han realizado análisis tanto cuantitativos como cualitativos que permiten ver qué variables son las mejores candidatas para formar parte de un modelo o estudio que permita responder a las dudas/objetivos que habíamos planteado al inicio, es decir, que el comprador:

* 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.
* Pueda hacerse una idea del presupuesto aproximado que tendrá el ordenador que desea en base a las características técnicas deseadas.

Para responder a la primera de las cuestiones, se ha planteado un contraste de medias que nos indicará si los productos de Apple son significativamente más caros, de media, que el resto de los productos de otras marcas.

Para el segundo objetivo, se ha evaluado la correlación de las variables numéricas respecto a la variable dependiente del precio, para identificar cuáles de ellas tienen más influencia en la variabilidad del precio.

Para el tercero, se han generado modelos de regresión lineal que predicen el precio esperado a partir de un conjunto de características seleccionadas a lo largo del objetivo anterior.

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

El estudio de la normalidad y homogeneidad de la varianza se incluye en el apartado “Contraste de muestras”, correspondiente a la pregunta 4.3.

## 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.

### Contraste de hipótesis

Se plantea si los precios de los portátiles de Apple tienen un precio superior, de media, al resto de productos de otras compañías. Para poder verificar estadísticamente dicha afirmación, se plantea un contraste de hipótesis como el siguiente:

Para llevar a cabo el contraste de muestras, separaremos en dos el dataset de tal forma que contengan respectivamente los precios de los productos de Apple y del resto de marcas.



Una vez separadas las muestras, estudiamos la normalidad de cada una de ellas:

Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente



En base a los resultados, dado que p-value es menor que el valor de significancia 0.05, podemos descartar la hipótesis nula de normalidad, por lo que no podemos considerar que la distribución que siguen los precios de apple se ajuste a una normal. Asimismo, en el caso del subset de Apple, no sería posible tampoco aplicar el teorema del límite central al no superar las 30 muestras.

Texto

Descripción generada automáticamente con confianza media

Histograma

Descripción generada automáticamenteImagen que contiene Forma

Descripción generada automáticamente

Imagen que contiene Patrón de fondo

Descripción generada automáticamente

Nombre de la empresa

Descripción generada automáticamente con confianza media

En base a los resultados, dado que p-value es menor que el valor de significancia 0.05, podemos descartar la hipótesis nula de normalidad, por lo que no podemos considerar que la distribución que siguen los precios de productos distintos a apple se ajuste a una normal.

En este sentido, para poder confirmar si podemos asumir homocedasticidad (igualdad de varianzas entre muestras) debemos aplicar el test de Fligner-Killeen (no paramétrico) al no poder suponer normalidad:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

A raíz de los resultados (p value >> 0.05) no podemos descartar la hipótesis nula, por lo que se confirma la homocedasticidad. No obstante, debido a que no hemos podido afirmar que sigan distribuciones normales, no podremos aplicar un contraste de muestras paramétrico (t-Student), si no que tendremos que aplicar uno no paramétrico (Mann-Whitney):

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

A partir de los resultados obtenidos del test de Mann-Whitney (p-value << 0.05), podemos rechazar la hipótesis nula en favor de la hipótesis alternativa que, en este caso, correspondía con que el precio de los productos de Apple es superior de media que para el resto de marcas.

### Estudio de variables numéricas

En este apartado se pretende estudiar qué variables numéricas son más influyentes a la hora de determinar el valor que toma la variable dependiente Precio\_euros.

Para ello, realizamos una matriz de correlación, que nos podrá indicar con qué variables numéricas tiene más relación el valor del precio:

Texto

Descripción generada automáticamente con confianza media

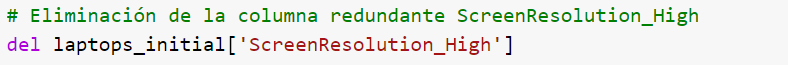
Gráfico de rectángulos

Descripción generada automáticamente

Tal y como se puede apreciar en el mapa de calor, la variable numérica con la que tiene más correlación el precio es con la cantidad de RAM que tenga el dispositivo, seguido de la cantidad de memoria SSD, la resolución de pantalla y la CPU que monte el dispositivo. Todas ellas, serán candidatas para formar parte del modelo de predicción del precio de un dispositivo.

Al contrario, las variables que menos influyen en la variabilidad de precio son el resto de tipos de almacenamiento (HDD, Flash e Híbrido), así como las pulgadas del portátil. El peso, pese a que influye en la valoración del portátil, lo hace de forma muy débil.

Dado que se identifica que las resoluciones de pantalla a lo ancho y alto están completamente correlacionadas entre sí, se elimina la columna ScreenResolution\_High para evitar redundancias.



### Estudio de variables categóricas

A continuación, vamos a estudiar con más detalle algunas de las variables categóricas que pensamos pueden afectar más al rendimiento del modelo, para identificar visualmente si vemos conveniente que también formen parte del mismo:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Respecto a las marcas, si bien hay algunas con rangos claramente superiores o inferiores, la mayoría de ellas mantienen una media de precios estable entorno a los 1000-2000€, rango que por otra parte se ha comprobado anteriormente como la media del dataset. En este sentido, no haremos uso de esta variable para el modelo.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En este caso, se ve una clara diferencia de medias entre las CPU de Intel y las de AMD, que son las dos que abarcan la mayoría de dispositivos. Vemos conveniente por tanto disponer de dicha información en el modelo.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En este caso, aunque no de forma tan clara como en el de las CPU, se pueden apreciar lo que serían 3 gamas de GPU. De más barata a más cara estarían: AMD, Intel y Nvidia. Por lo tanto, consideramos relevante añadirlo al modelo.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En tipos de dispositivos, si bien hay alguna categoría que destaca (como Gaming), el resto de los dispositivos parece que abarcan una franja amplia de precios, con lo que no se ve recomendable incorporarla al modelo.

A continuación procedemos a codificar las variables categóricas como numéricas mediante el método de one-hot para su inclusión en el modelo que generaremos a posteriori. Dado que desgranamos una variable categórica en tantas variables como categorias tenga (k), siempre podremos eliminar una de estas variables resultantes para mejorar el rendimiento del modelo y evitar redundancias.

Imagen que contiene Texto

Descripción generada automáticamente

Imagen que contiene interior, taza, computadora, estante

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza media

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Para tratar de ver el efecto real de estas variables, ejecutaremos de nuevo la matriz de correlación y veremos si la relación entre las variables que hemos pasado a numéricas y el precio es buena para incluirlas en el modelo.

Logotipo

Descripción generada automáticamente con confianza baja

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

A partir de los resultados, vemos que la única variable que podría ser relevante para el modelo es la de GPU\_Nvidia, al tener una correlación con el precio de 0.35. También podría llegar a ser candidata CPU\_Intel, al haber obtenido un 0.18 de correlación.

### Creación y evaluación del modelo

A partir de las variables que se ha identificado tienen más correlación e influencia sobre la variable que queremos estimar, Price\_euros, generamos un subset de datos que nos servirá para generar el modelo de predicción:



Procedemos a normalizar los valores de todas las variables para evitar que las diferentes magnitudes de cada una de ellas afecten al rendimiento del modelo final:

Imagen que contiene Diagrama

Descripción generada automáticamente

Tabla

Descripción generada automáticamente con confianza baja

A continuación, procederemos a dividir el dataset en un subset de entrenamiento y otro de test para evaluar el rendimiento del modelo.



A continuación generamos el modelo de regresión lineal a partir de los subsets de entrenamiento y test:

Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Los resultados de esta regresión no dan un resultado excesivamente alto respecto a la precisión del modelo, ya que esta se queda en un coeficiente de determinación de 0.7 (siendo 0 el peor valor y 1 el mejor), lo que indica que el 70% de la variabilidad de la variable dependiente está explicada por las variables independientes. No obstante, puede ser considerado suficiente para identificar si un precio de mercado está por encima o por debajo de lo esperado por sus especificaciones.

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

A lo largo del informe se han ido mostrando las tablas y gráficas necesarias para realizar los análisis propuestos.

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

Si nos basamos en los objetivos que nos hemos planteado, para el podemos llegar a las siguientes conclusiones:

* Se ha obtenido un modelo que es capaz de aproximar un precio justo para el tipo de producto que se está tratando. Aún así, hay desviaciones en varias ocasiones respecto al precio esperado, dado que el modelo no está arrojando unos valores de precisión demasiado elevados. Este último hecho tendría fácil solución añadiendo más datos de ordenadores de marcas diferentes para enriquecer el mismo y trabajar con más de 1300 muestras.
* Se ha podido confirmar, en base al contraste de muestras realizado, que los dispositivos de Apple son de media más caros que el resto de los dispositivos de otros fabricantes. Por lo tanto, la afirmación que la gente suele suponer de que la marca “Apple” se paga, según el análisis, es una afirmación correcta.
* Se ha podido verificar cuales de las características de los portátiles eran más influyentes a la hora de subir o bajar el precio del producto. La característica más correlacionada con el precio es la de la memoria RAM.

Asimismo, se han podido extraer otras conclusiones no perseguidas en base al estudio y preparación del dataset. Destacarán, entre otras:

* La categoría de portátiles más caros es la de Gaming, así como la que tiene componentes de mejor prestación pero que, a su vez, hace que los portátiles sean más pesados.
* La compañía que sin lugar a dudas posee el liderazgo en GPU y CPU es Intel.
* Windows es el líder en sistemas operativos, específicamente en la fecha de toma de la muestra, con la versión Windows 10.

# 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 mostrado en este informe con todos los pasos detallados con comentarios explicativos, al igual que código adicional de partes que no se han adjuntado por no hacer más largo el informe o evitar repeticiones, en el siguiente enlace de GitHub (<https://github.com/carlosalloUOC/PRA2-Limpieza-Analisis>), 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 visualización y un extracto del mismo en formato .py. Adicionalmente, por si se tuviera algún problema de visualización, se ha añadido en la misma carpeta un archivo .pdf con el código ejecutado (aunque es preferible por estética ver el .html al ser interactivo y tener estilos que hacen más bonito el estudio y ser este formato interactivo).

# 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 |