MEMORIAS EDA DE JUAN IGNACIO MONTALBAN SARDI

# 1-. Primer paso

Lo primero ha sido elegir el tema del que quería hablar, sabía que quería realizar un análisis relacionado con el mundo del motor, aunque no estaba seguro de como enfocarlo. Primero pensé en hablar sobre la F1, pero al ser un tema que puede requerir mucha información no estaba seguro de poder conseguir los suficientes datos. Finalmente decidí realizarlo con el mercado de segunda mano ya que es un tema con mucho potencial:

* Desde el punto de vista de negocio tiene todo el sentido.
* Existe mucha información, aunque no siempre esté completa. La parte positiva es que, a pesar de que no sea fácil encontrar información a nivel europeo, a nivel estadounidense fue fácil encontrar un dataset con mucha información y completo.
* Se trata de un tema que considero que domino, ya que paso gran parte de mi tiempo libre buscando vehículos de segunda mano e información para poder elegir un buen vehículo, desde un coche para transporte/ para ocio a una moto o un kart para poder disfrutar los fines de semana.

# 2-. Segundo paso

La búsqueda de datasets la realice especialmente en la web [kaggle](https://www.kaggle.com/). Al buscar información del parque de vehículos de segunda mano solo encontré los datasets de estados unidos que estuviesen completos y con suficientes filas de datos.

Estuve pensando en realizar web scraping a una de las paginas de segunda mano en España como puede ser coches.net pero al pasar mucho tiempo en estas webs no quería que me bloqueasen la IP como le ha pasado a Andrés con rateyourmusic. Descarté esa opción y me decanté finalmente por realizar mi EDA con el dataset de estados unidos.

Una vez tuve los datos de los vehículos, tocaba tener datos sobre el sueldo medio por grupo de edad en estados unidos para poder obtener un presupuesto con el que limitar el precio de los vehículos.

Encontré unos datasets en [HESA](https://www.hesa.ac.uk/news/16-06-2022/sb263-higher-education-graduate-outcomes-statistics) pero la información era de salario en reino unido. Intenté hacerlos funcionar ya que no podía encontrar información de estados unidos agrupado por grupos de edad, pero durante la tutoría con Jhon encontramos dos webs que me daban la información necesaria.

La primera fue una que hablaba de la media de [salario por edad concreta](https://dqydj.com/average-median-top-income-by-age-percentiles/). La segunda, mucho más completa, permitía seleccionar la media de edad por estado y rangos de edad. Indicando el rango de sueldo más común, del que podía utilizar el [punto medio](https://www.averagesalarysurvey.com/united-states).

El mayor problema es que tenía que realizar web scraping y no tenía suficiente tiempo para poder realizarlo ya que tenia que dedicar el tiempo a limpiar y filtrar el dataset de vehículos. Por tanto, tuve que tomar la decisión de usar la información de la primera pagina a pesar de que los datos que podía obtener de la segunda eran mejores.

Por último, busqué documentación sobre fiabilidad según modelos o marcas de vehículos, pero al no encontrar datasets, solo artículos con una pequeña lista, tuve que utilizar los datos limitados que tenia para poder tener información sobre la fiabilidad de los vehículos y poder ordenar desde ahí.

# 3-. Tercer paso

La limpieza del dataset la comencé comprobando las columnas que tuviesen todos sus valores nulos, ya que así se cuales no me van a dar ningún valor. Después comprobé los valores nulos de la columna ‘imagen’ ya que en ningún caso elegiríamos un anuncio sin imagen para comprar un vehículo.

Lo siguiente fue retirar los valores nulos en las columnas ‘condition’ y ‘cylinders’ ya que son aspectos importantes por los que filtrar, por lo tanto, no puedo tener valores nulos. Retire los valores nulos del resto de columnas en los que no tener información fuese relevante y deje los valores nulos solo en las columnas que no eran relevantes como por ejemplo ‘paint\_color’.

# 4-. Cuarto paso

Una vez retirados los valores nulos, ya pude dedicarme a realizar filtraciones para poder quedarme con los valores necesarios. Comencé por retirar todas las filas en las que el numero de cilindros es mayor a 8 ya que son motores realmente grandes y nunca serian elegidos para un primer coche, además de ser vehículos mucho más costosos los que tienen estos motores.

Después lo que hice fue utilizar la información sobre el salario que recopilamos antes y termine obteniendo un limite por arriba de 20.000$, por lo que aplique el filtro para así retirar todos los valores superiores.

Lo siguiente que realicé fue, usando el dataset de los 35 fabricantes más fiables, retirar todas las filas que no estuviesen en el top 15 de fabricantes mas fiables, o lo que es lo mismo los fabricantes con una puntuación de 80 o mas en el dataset de fiabilidad.

Al terminar con la fiabilidad, convertí los años a valores integers y para trabajar mejor con ellos mas adelante.

Los coches tienen una vida útil de unos 12/15 años, por lo que no tendría sentido mantener vehículos con mas edad ya que son mas propensos a sufrir averías. Cuando retiré esos vehículos comencé a retirar todos los que tuviesen una valoración de ‘good’ o inferior ya que tenemos un buen presupuesto y podemos ajustar aún más los vehículos.

Para poder filtrar los vehículos según su ‘type’ lo que hice fue retirar los vehículos que su volumen fuese mas grande, como son los buses, camiones, furgonetas o minivanes; los motores de estos vehículos también suelen ser mas grandes y con mayor consumo. Por último, retiramos los convertibles o descapotables ya que no serían útiles debido a su escasa capacidad de maletero y necesitamos vehículos polivalentes. Los SUVs quedan fuera también debido a que se tratan de vehículos con mismas dimensiones que un sedán, pero con mayor consumo y menos seguros debido a su peor aerodinámica y un centro de gravedad más alto, siendo también mayor la distancia de frenado debido a su peso extra.

Comprobamos la mediana precio de los vehículos por numero de cilindros y vemos que solo los vehículos de 3 y 4 cilindros se encuentran en menos de 10000$, por lo que nos quedamos con los vehículos de 4 cilindros ya que los de 3 cilindros no son idóneos para cubrir las distancias que nos encontramos en estados unidos.

Por último, comenzamos a agrupar por precio, por kilometraje y por año para obtener un dataset ordenado especialmente por esos valores, retiramos los outliers como son los vehículos con 0 o 1 millas que no tienen la condición de nuevo o que su precio esta por debajo de los 5.000$ y ordenamos nuestro dataset según el numero de millas para conocer las mejores opciones, quedándonos con los 5 mejores vehículos.

# 5-. Quinto paso

Mediante tableau, utilizamos la información de los datasets que hemos obtenido filtrando para crear diferentes gráficas, usamos el dataset en diferentes puntos para comprobar la información en diferentes momentos.

El mayor problema para mi ha sido el poder realizar las graficas para poder dar información relevante sobre el dataset ya que la filtración ha sido prácticamente en su totalidad mediante información externa a las gráficas.

Mediante Python pude ver la densidad de precios, apreciando que la mayoría de los anuncios se sitúan entre los 5 y 10 mil dólares y mediante un histograma se puede ver que la cantidad de anuncios que utilizan números redondos (6000, 7000, 8000…) es mucho mayor que los valores intermedios como podría ser 7500 dólares.

Por último, traté de sacar información del dataset mediante un pairplot, pero era realmente difícil sacar algo en claro por lo que me decidí a utilizar un heatmap para comprobar la correlación de los diferentes valores, descubriendo que una vez filtrados los fabricantes y seleccionando solo los 15 mejores, el factor de la fiabilidad no influía en el resto de los valores, por lo que ya no eran relevantes.

# 6-. Sexto paso

La conclusión con la que puedo terminar es que a pesar de que ha resultado un tema que a mi parecer es muy interesante, considero que no podía mostrar su valor mediante las graficas por lo que no estoy seguro de que se pueda apreciar como es debido, espero que la presentación ayude en este aspecto.