DATA UNDERSTANDING

INDICE

1. Objetivo del Análisis

Comprender los datos de coches para predecir precios, analizar tendencias de consumo, o identificar características clave relacionadas con el rendimiento y la eficiencia.

---

2. Descripción de las distintas fuentes de datos.

Las páginas utilizadas para el desarrollo del proyecto son:

1- tesla.com
2- milanuncios.com
3- coches.net

2.1 Tesla.com

En el sitio web de Tesla, se ofrecen variedades. Los vehículos han pasado por un proceso de inspección y reacondicionamiento. Los precios varían dependiendo del año, kilometraje y estado del coche, con opciones que van desde unos 22,169 euros hasta 160,600 euros. Además, Tesla proporciona garantías y asistencia en carretera para los autos certificados. En general, el mercado de autos usados está viendo una caída de precios debido a un aumento en el inventario global de Teslas.

2.2 Milanuncios.com

Es una página es un tablón de anuncios y clasificados de segunda mano, compra-venta, inmobiliarias, empleo, servicios... Para temas de coches, es muy genérico y no se centra solo en teslas.

Durante la búsqueda de coches tesla se aparecen anuncios no relevantes a la búsqueda original(otra marca de coche), hay en cada oferta variedad de imágenes para apaciguar las dudas del comprador, la marca, un precio que cabe destacar que es mucho más bajo que los de la página tesla, ubicación del vehículo , un año de publicación, tipo de combustible(eléctrico en todos los casos) y una breve descripción para poner en contexto que es lo que se está comprando.

Dentro de cada oferta hay incluso más detalles como puertas, garantía, cv, cambio y la opción de comunicarse con el vendedor en caso de tener interés en el coche tesla

2.3 Coches.net

Es una página centrada única y exclusivamente en la venta de coches de segunda mano en España.

En el buscador de la propia página se puede buscar por diferentes marcas e incluso modelos, cada uno junta a un numero de ofertas, e incluso por provincia y precio, facilitando al usuario lo que busca.

Cada oferta tiene un numero de imágenes al igual que en milanuncios , con el modelo, el precio al contado y también si es financiado, la localización del coche, si el envió es posible, garantía, combustible(también eléctrico solo), el año en que se publicó y los kilómetros que ha hecho hasta ahora. Además de todo eso tiene un criterio del 1 al 5 de que tan justo es el precio.

Dentro de cada coche están los mismos datos, pero con comentarios del anunciante, cv, color, etiqueta, numero de puertas y la posibilidad de enviar tu contacto.

---

3. Descripción del Dataset

3.1 Tesla.com

- **Ubicación del dataset**:
  - data/Tesla.com.csv
- **Fuente**:
  - tesla.com
- **Fecha de creación/actualización**:
  - Actualizado el 05/01/2025.
- **Tamaño del dataset**:
  - 938 filas y 7 columnas.

3.2 milanuncios

- **Ubicación del dataset**:
  - data/milanuncios.com.csv
- **Fuente**:
  - milanuncios.com
- **Fecha de creación/actualización**:
  - Actualizado el 05/01/2025.
- **Tamaño del dataset**:

- 407 filas y 4 columnas.

3.3 cohes.net

- **Ubicación del dataset**:
  - data/cochesNet.csv
- **Fuente**:
  - coches.net
- **Fecha de creación/actualización**:
  - Actualizado el 05/01/2025.
- **Tamaño del dataset**:
  - 516 filas y 7 columnas.

3.4 Datos totales

- **Ubicación del dataset**:
  - data/full_data.csv
- **Fuente**:
  - tesla.com, coches.net, milanuncios.com
- **Fecha de creación/actualización**:
  - Actualizado el 05/01/2025.
- **Tamaño del dataset**:
  - 1845 filas y 5 columnas.

---

4. Características del Dataset

4.1 Dataset Tesla

Nombre de la columna	Descripción	Tipo de dato	Ejemplo
Sell_type	Estado del coche	Cadena de texto	"Nuevo"

Nombre de la columna	Descripción	Tipo de dato	Ejemplo
Car_type	Marca del coche	Cadena de texto	"Model S"
Year	Año de publicación	Cadena de texto	2024
Color	Color de coche	Cadena de texto	"WHITE"
Country	Ubicación geográfica	Cadena de texto	"spain"
Traction	Tracción del vehículo	Cadena de texto	"Plaid"
Price	Precio del coche	Cadena de texto	100000€
Km	Kilómetros que realizó	Cadena de texto	"Cuentakilómetros con 503 km"
Range	Rango que comprende el vehículo	Entero	634
Max_vel	Máxima velocidad	Entero	250
Zero_hundred	Tiempo que tarda en llegar a 100km/hora	flotante	3,2

4.2 Dataset Milanuncios

Nombre de la columna	Descripción	Tipo de dato	Ejemplo
Car_type	Modelo de coche	Cadena de texto	"TESLA - MODEL 3"
Year	Año que se publicó la oferta	Integer	2020
Price	Precio del coche	Cadena de texto	"34.000 €"
Km	Kilómetros realizados previamente	Cadena de texto	"42.100 kms"

4.3 Dataset Coches.net

Nombre de la columna	Descripción	Tipo de dato	Ejemplo
Car_type	Marca del coche	Cadena de texto	"Toyota"
Year	Modelo del coche	Cadena de texto	"Corolla"
Ubicación	Ubicación geográfica	Cadena de texto	"spain"
Link	Enlace de la oferta en concreto	Cadena de texto	"coches.net/covo.aspx"
Price	Precio del coche	Cadena de texto	"100.000 €"
Km	Tipo de transmisión	Cadena de texto	"Automática"
Tipo de vehículo	Tipo de motor usado	Cadena de texto	"Eléctrico"

4.4 Dataset Full_data

Nombre de la columna	Descripción	Tipo de dato	Ejemplo
Sell_type	Estado del coche	Cadena de texto	"Nuevo"
Car_type	Modelo del coche	Cadena de texto	"Model S"
Year	Año de publicación	Integer	2023
Price	Precio del coche	Flotante	100000.0
Km	Kilómetros realizados por el coche	Flotante	699.0

---

5. Descripción de las técnicas utilizadas para la obtención de los datos

-Tesla.com:

- Automatización del Navegador (Selenium):

  Utiliza un controlador web (c_driver) para navegar dinámicamente por diferentes URLs que se generan a partir de combinaciones de país, tipo de venta y modelo de coche.

- Construcción Dinámica de URLs:

  Se crean URLs específicas para cada combinación de filtros (tipo de coche, tipo de venta, color, país) para acceder a páginas de inventario.

- Uso de XPATH para Localizar Elementos:

  Se emplean expresiones XPATH para encontrar elementos específicos en la página, como opciones de color y detalles de los coches (precio, tracción, kilometraje, etc.).

- Extracción de Datos:

  Los datos extraídos de cada coche (precio, año, tracción, kilometraje, rendimiento, velocidad máxima, aceleración) se almacenan en un diccionario que se agrega a la lista data.

- Manejo de Errores:

  Se utilizan bloques try-except para capturar errores durante la navegación o extracción de datos, evitando que el programa se detenga.

- Control de Tiempos (Delay):

  Se incluye time.sleep(5) para esperar que las páginas carguen completamente antes de interactuar con ellas, evitando errores por carga incompleta.

-Milanuncios.com:

- Automatización del Navegador (Selenium):

  Se usa un navegador automatizado (c_driver) para acceder a las páginas de resultados de búsqueda de Tesla.

- Gestión de Cookies:

  Detecta y acepta el aviso de cookies haciendo clic en el botón correspondiente.

- Scroll Automático:

  Implementa desplazamiento suave hacia abajo (scroll) para cargar más resultados dinámicamente, simulando el comportamiento de un usuario.

- Extracción de Datos:

  De cada anuncio, se extraen detalles como modelo de coche, precio, kilometraje y año.

  Los datos recopilados se almacenan en la lista data.

- Manejo de Errores:

  Se utiliza try-except para evitar fallos si algún dato no está disponible (ej. anuncios incompletos).

- Navegación Paginada:

  Recorre las páginas del 1 al 11, actualizando la URL y repitiendo el proceso de extracción en cada página.

- Filtrado de Publicidad:

  Ignora anuncios tipo 'ma-AdCardCarousel' que no contienen información relevante.

-Coches.net:

- Automatización con SeleniumBase:

  Modo incógnito para boots o vpn (uc=True).

  Bloqueador de anuncios (ad_block=True).

  Navegación sin interfaz gráfica (headless=True).

- Extracción de Datos:

  Accede a cada página de resultados y busca los anuncios mediante selectores CSS.

  Extrae los datos de cada anuncio y los almacena en la lista arry_coches.

  Recorre hasta la página 18

- Eliminación de Duplicados:

Convierte la lista en un conjunto (set) usando los métodos de comparación personalizados para evitar registros repetidos.

- Función html_to_av:

Convierte el HTML de un anuncio en un objeto AnuncioVehiculo utilizando BeautifulSoup para extraer datos relevantes.

---

6. Resumen de los resultados del EDA + Data cleaning por cada fuente de datos.

- **Distribución de Precios:**

  Los precios de los vehículos presentan una distribución sesgada a la derecha, con la mayoría concentrándose en rangos medios y algunos valores extremos que podrían ser outliers.

- **Relación Kilometraje vs. Precio:**

  Existe una tendencia inversa: los vehículos con menor kilometraje suelen tener precios más altos. Sin embargo, se identificaron algunos vehículos de alto kilometraje con precios inusualmente elevados.

- **Precio según Tipo de Venta:**

  Los vehículos nuevos tienen precios consistentemente más altos, mientras que los de segunda mano muestran una mayor dispersión, reflejando variedad en modelos, años y condiciones.

- **Distribución de Colores:**

  Los colores más comunes son tonos neutros (blanco, negro y gris), lo que indica preferencias de mercado.

- **Aceleración (0-100 km/h) vs. Precio:**

  Los modelos con menor tiempo de aceleración (mejor desempeño) tienden a ser más caros, evidenciando la relación entre potencia y costo.

- **Correlación entre Variables:**

Existe una correlación moderada entre el precio y la aceleración, y una débil relación entre el precio y el kilometraje. La autonomía (range) también se relaciona positivamente con el precio

---