**Reporte: Proceso de Limpieza de Datos**

Tema: Análisis y predicción del tiempo de venta de autos Usados

Giovanni Leonel Chillopa Martínez

Materia: Introducción a la Ciencia de Datos

Jaime Alejandro Romero Sierra

20/10/2025

Link al repositorio de Git-Hub: <https://github.com/GiovanniChTO/Proyecto_Base_Car_Prices>

**Descripción inicial de la base de datos**

Fuente o contexto:

La base de datos fue obtenida de la página web de Kaggle, es una recopilación de venta de autos en casas de subastas de estados unidos, recopilando le auto, modelo, color, precio, millaje y fecha a lo largo de un periodo de años determinado.

Descripción general del contenido:

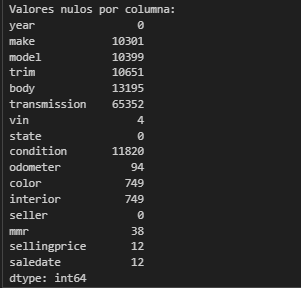
El dataset contiene 558837 registros de autos usados con 16 columnas. El objetivo es limpiar los datos para poder predecir el tiempo que un auto permanece en venta antes de ser vendido y de es amanera poder saber que autos es mejor comprar para vender rapidamente.

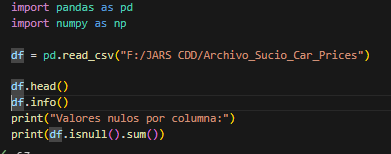
|  |
| --- |
| |Columna Descripción |
| | `year` | Año de fabricación del vehículo |
| | `make` | Marca del auto |
| | `model` | Modelo del auto |
| | `trim` | Versión o configuración del modelo |
| | `body` | Tipo de carrocería (sedan, coupe, etc.) |
| | `transmission` | Tipo de transmisión (automática o manual) |
| | `vin` | Número de identificación del vehículo |
| | `state` | Estado donde se vendió el auto |
| | `condition` | Condición del auto |
| | `odometer` | Kilometraje del auto |
| | `color` | Color exterior del auto |
| | `interior` | Color interior del auto |
| | `seller` | Nombre del vendedor |
| | `mmr` | Precio estimado de mercado |
| | `sellingprice` | Precio final de venta |
| | `saledate` | Fecha en que se vendió el auto |

**Proceso de Limpieza (con evidencias)**

Revisión de datos faltantes

Se utilizó isnull() e info() para identificar valores faltantes por columna.

df.isnull().sum()



Detección y manejo de duplicados

Se detectaron filas duplicadas con duplicated() y se eliminaron conservando la primera aparición.

df = df.drop\_duplicates()

Corrección de valores atípicos o inconsistentes

Se corrigieron precios fuera de rango en las columnas de tipo numérico.

df = df[(df['sellingprice'] >= 500) & (df['sellingprice'] <= 100000)]

df = df[(df['year'] >= 1980) & (df['year'] <= 2024)]



Detección y corrección de palabras mal escritas

df['make'] = df['make'].str.lower().str.title()

df['model'] = df['model'].str.lower().str.title()

Traducción de textos al español

Se tradujeron valores de columnas categóricas como condition, transmission, body, etc.

transmission\_map = {

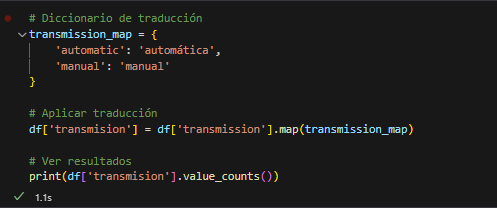
'automatic': 'automática',

'manual': 'manual',

'automanual': 'automática secuencial'}

df['transmision'] = df['transmission'].map(transmission\_map)

print(df['transmision'].value\_counts())



Cambio de nombres de columnas para mayor claridad

Se renombraron columnas al español.

df.rename(columns={

'year': 'año',

'make': 'marca',

'model': 'modelo',

'transmission': 'transmision',

'sellingprice': 'precio\_venta',

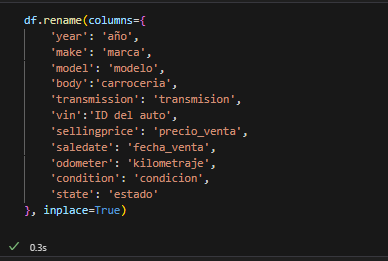
'saledate': 'fecha\_venta',

'odometer': 'kilometraje',

'condition': 'condicion',

'state': 'estado'

}, inplace=True)

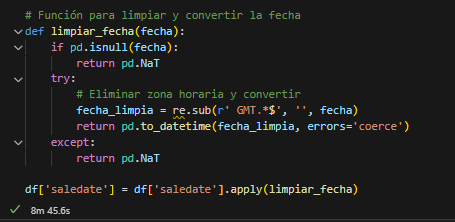


Conversión de tipos de datos

Se convirtieron columnas como fecha\_venta a tipo datetime, y precio\_venta a entero.

df['fecha\_venta'] = pd.to\_datetime(df['fecha\_venta'], errors='coerce')

df['precio\_venta'] = df['precio\_venta'].astype(int)

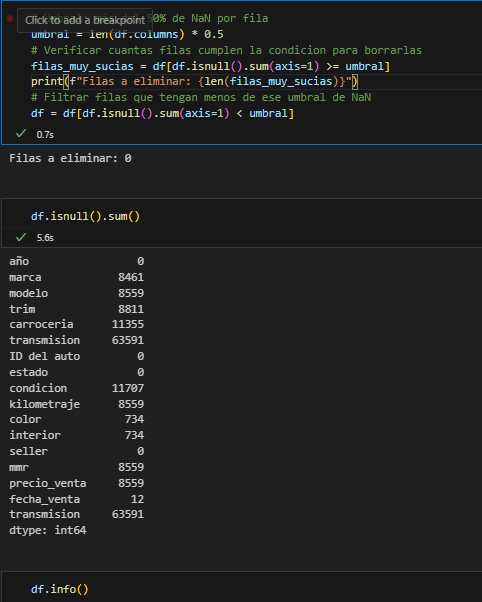


Validación final

Se verificó que no haya valores inconsistentes, que los tipos de datos sean correctos y que no haya filas con más del 50% de valores faltantes.

df.isnull().sum()

df.info()



**Conclusiones**

La base de datos sucia presentaba varios problemas, algunas columnas mezclaron sus datos, estaba todo en inglés, habían varios datos tipo Nan y nulos, además de que tenía que eliminar los datos no correspondientes en columnas numéricas o en la columna de estado.

Utilice varios códigos vistos en las clases de Ciencia de datos, como los comandos de: sum, map, isnull, info, len, duplicates y dos eliminaciones de columnas duplicadas.

Con este proyecto he aprendido lo laborioso que puede llegar a ser analizar la base por completo y lograr eliminar los datos que no corresponden, están vacíos y no están dentro de un rango común, las bases se pueden limpiar sin utilizar un dropna entero y es muy necesario limpiarla para poder hacer nuestros modelos predictivos o de machine learning.