Diana Luz Hernandez Torres

Materia: Introduccion a la ciencia de datos

Jaime Alejandro Romero Sierra

21 de Octubre del 2025

Link: https://github.com/dianahdeztorres13-pixel/Base-limpia

Descripción

La base de datos proviene de keaggle, y fue recolectada en el contexto de hacer un análisis para reducir la brecha que existe en la formula 1. El objetivo principal es analizar el rendimiento entre escuderías a lo largo de las últimas temporadas mediante datasets, con el proposito de identificar patrones, tendencias y factores que influyen en la competitividad entre equipos, con el fin de detectar patrones y técnicos que influyen en las diferencias de rendimiento entre escuderías, considerando variables como los circuitos, condiciones meteorológicas, degradación de neumáticos.
El conjunto de datos contiene información sobre Sesion, Recorrido, Circuito, Clima Calido, frio, seco, lluvioso, nublado, pilotos, podios, puntos,
victorias, clasificación, etc, abarcando un total de 7517 filas y 22 columnas.

Tabla con las columnas:

Columna	Descripción
`Unnamed: 0`	Índice original del dataset (puede eliminarse si no aporta valor)
`Sesion`	Tipo de sesión (ej. clasificación, práctica, carrera)
`Recorrido`	Nombre del recorrido o evento específico
`Circuito`	Circuito donde se llevó a cabo la sesión
`Clima Calido`	Indicador binario de clima cálido
`Clima frio`	Indicador binario de clima frío
`Clima seco`	Indicador binario de clima seco
`Clima lluvioso`	Indicador binario de clima con lluvia
`Clima nublado`	Indicador binario de clima nublado
`Piloto`	Nombre del piloto participante
`Nacionalidad`	Nacionalidad del piloto
`Constructores`	Nombre del equipo constructor del vehículo
`Grilla`	Posición de salida en la grilla
`Podio`	Indicador binario si el piloto terminó en el podio
`Puntos_piloto`	Puntos obtenidos por el piloto en esa sesión
`Victorias_piloto`	Número de victorias acumuladas por el piloto
`Posicion_piloto`	Posición final del piloto en la sesión
`Puntos_constructores`	Puntos obtenidos por el equipo constructor
`Victorias_constructores`	Número de victorias acumuladas por el constructor
`Constructores_posicion`	Posición final del equipo constructor en la sesión
`tiempo_clasificación`	Tiempo registrado en la sesión de clasificación
`Edad_piloto`	Edad del piloto en el momento de la sesión

Proceso de limpeza

```
#Renombramos las columnas a español
df2 = df2.rename(columns={"season":"Sesion","round":"Recorrido","circuit_id":"Circuito", "weather_warm":"
                          "weather_cloudy":"Clima nublado","driver": "Piloto", "nationality": "Nacionalid
       'driver_wins':"Victorias_piloto", 'driver_standings_pos':"Posicion_piloto", 'constructor_points':"
       'constructor_wins':"Victorias_constructores", 'constructor_standings_pos': "Constructores_posicion
       'driver_age': "Edad_piloto"})
df2
```

```
#Se verifican las columnas con las que se cuenta
df.columns
```

```
Index(['Unnamed: 0', 'season', 'round', 'circuit_id', 'weather_warm',
       'weather_cold', 'weather_dry', 'weather_wet', 'weather_cloudy',
       'driver', 'nationality', 'constructor', 'grid', 'podium',
       'driver_points', 'driver_wins', 'driver_standings_pos',
       'constructor_points', 'constructor_wins', 'constructor_standings_pos',
       'qualifying_time', 'driver_age'],
      dtype='object')
```

```
#Importar librerias y cargar base de datos sucia
import pandas as pd
df = pd.read_csv("https://raw.githubusercontent.com/dianahdeztorres13-pixel/fh/main/j.csv")
df
```

	Unnamed: 0	season	round	circuit_id	weather_warm	weather_cold	weather_dry	weather_wet
0	14.0	1983.0	1.0	Auto%#	False	0.0	True	0.0
1	5.0	1983.0	1.0	jacarepagua	False	0.0	True	0.0
2	3.0	1983.0	NaN	jacarepagua	False	0.0	True	0.0
3	0.0	1983.0	1.0	jacarepagua	False	0.0	True	0.0
4	6.0	1983.0	1.0	jacarepagua	False	0.0	True	0.0

```
#Tamaño del dataframe original
df.shape
```

```
(16641, 22)
```

Proceso de limpeza

```
df2["Sesion"].unique()
```

```
array([1983.,   nan, 1984., 1985., 1986., 1987., 1988., 1989., 1990.,
       1991., 1992., 1993., 1994., 1995., 1996., 1997., 1998., 1999.,
       2000., 2001., 2002., 2003., 2004., 2005., 2006., 2007., 2008.,
       2009., 2010., 2011., 2012., 2013., 2014., 2015., 2016., 2017.,
       2018., 2019., 2020., 2021.])
```

```
df2["Sesion"].isnull().sum()
```

```
np.int64(481)
```

```
df2 = df2[df2["Sesion"].notna()]
```

```
df2["Sesion"].unique()
```

```
array([1983., 1984., 1985., 1986., 1987., 1988., 1989., 1990., 1991.,
       1992., 1993., 1994., 1995., 1996., 1997., 1998., 1999., 2000.,
       2001., 2002., 2003., 2004., 2005., 2006., 2007., 2008., 2009.,
       2010., 2011., 2012., 2013., 2014., 2015., 2016., 2017., 2018.,
       2019., 2020., 2021.])
```

```
df2["Clima Calido"].unique()
```

```
array([False, True, nan], dtype=object)
```

```
df2["Clima Calido"].isnull().sum()
```

```
np.int64(443)
```

```
df2 = df2[df2["Clima Calido"].notna()]
```

```
df2["Clima Calido"].unique()
```

```
array([False, True], dtype=object)
```

```
df2["Clima frio"].unique()
```

```
array([ 0., nan,  1.])
```

```
df2["Clima frio"].isnull().sum()
```

```
np.int64(720)
```

```
df2.to_csv("df2_limpio.csv", index=False)
```

```
!ls -l df2_limpio.csv
```

```
-rw-r--r-- 1 root root 952760 Oct 21 05:07 df2_limpio.csv
```

```
from google.colab import files
```

```
files.download("df2_limpio.csv")
```

Conclusion

Problemas principales detectados:
La base tuvo varios problemas iniciales:
Valores nulos en columnas clave como “Edad_piloto” y “tiempo_clasificación”.
Registros duplicados que podían distorsionar el análisis.
Inconsistencias en nombres de países, equipos y sesiones (variantes idiomáticas y errores tipográficos).
Columnas con nombres poco claros o formatos no estandarizados.
Técnicas aplicadas:
Para resolver estos problemas se aplicaron:
Se implemento estructural con “info()”, “isnull()”, para detectar irregularidad.
Renombrar de columnas.
Conversión de tipos de datos para facilitar cálculos y visualizaciones.
Aprendizaje:
Este proceso de limpieza de datos me permitió consolidar habilidades clave en la preparación de bases para análisis rigurosos. Más allá de aplicar funciones específicas, comprendí la importancia de adoptar una mentalidad mas crítica frente a los datos no solo con que estén completos, deben ser coherentes, legibles y representativos del fenómeno que se desea estudiar.
Uno de los aprendizajes más valiosos fue reconocer cómo pequeñas inconsistencias como nombres mal escritos, formatos desalineados o valores duplicados— pueden tener un impacto significativo en los resultados del análisis. Aprendí a detectar estos problemas no solo con funciones como isnull() o duplicated(), sino también de manera visual.