# Descripción del conjunto de datos

## 1. Introducción

La Fórmula 1 (F1) es una de las competencias automovilísticas más prestigiosas y tecnológicamente avanzadas del mundo. En ella, equipos compiten durante una temporada por puntos que se otorgan según el rendimiento en diferentes carreras alrededor del mundo.

Cada evento es una combinación de estrategia, talento humano y desarrollo técnico. Por eso, el análisis de datos en este contexto no solo permite entender mejor el rendimiento de pilotos y escuderías, sino también predecir resultados y optimizar decisiones deportivas.

---

## 2. Descripción del dominio: Fórmula 1

La Fórmula 1 es una competición automovilística organizada por la Federación Internacional del Automóvil (FIA). Cada temporada consta de múltiples Grandes Premios (GPs), que se celebran en circuitos de todo el mundo.

Cada equipo o “constructor” inscribe a dos autos (pilotos), que compiten por puntos individuales y de equipo. Los puntos son otorgados a los primeros 10 puestos en cada carrera. Existen múltiples variables que influyen en los resultados: el tipo de circuito, las condiciones climáticas, el rendimiento del auto, la estrategia de boxes y la habilidad del piloto.

---

## 3. Cantidad de ejemplos y atributos

El conjunto de datos contiene:

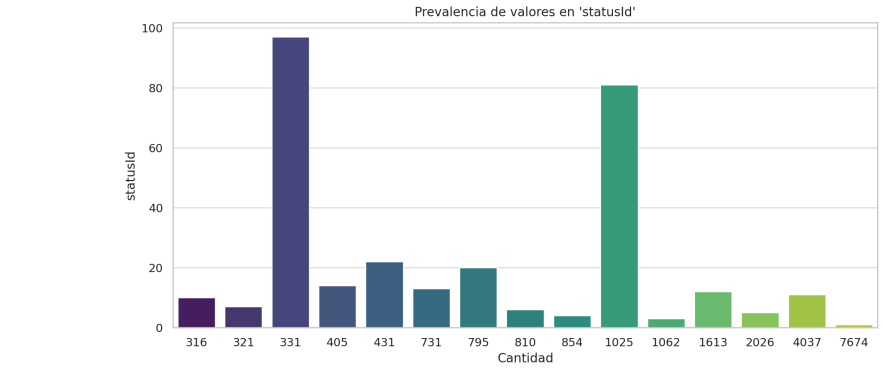
- **Cantidad de ejemplos (filas):** 25.155
- **Cantidad de atributos (columnas):** 15

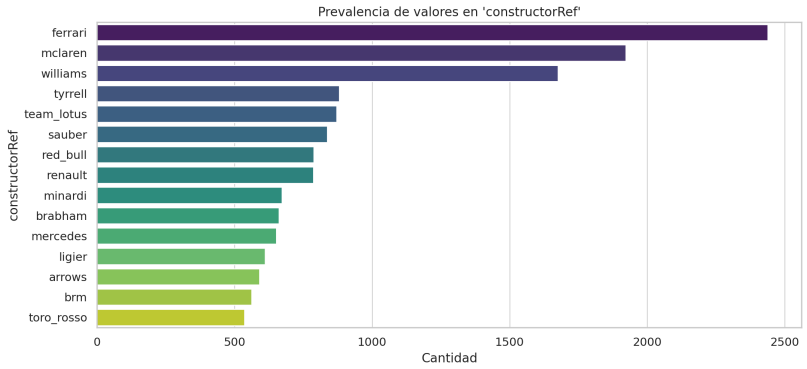
### Atributos:

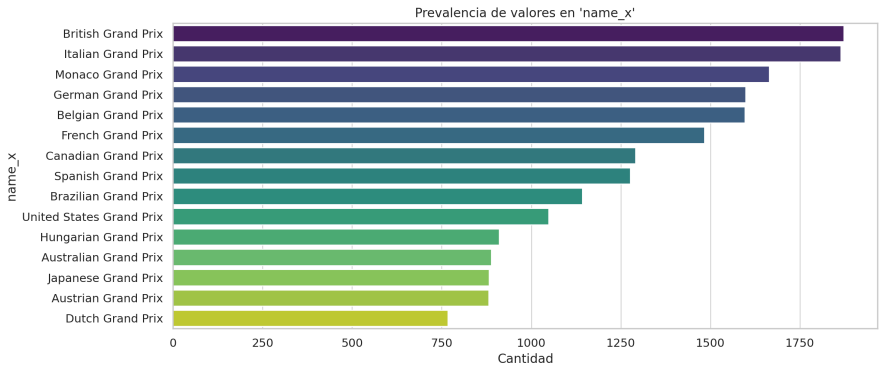
| Atributo | Tipo | Descripción |
|---|---|---|
| `raceId` | Numérico | Identificador único de cada carrera |

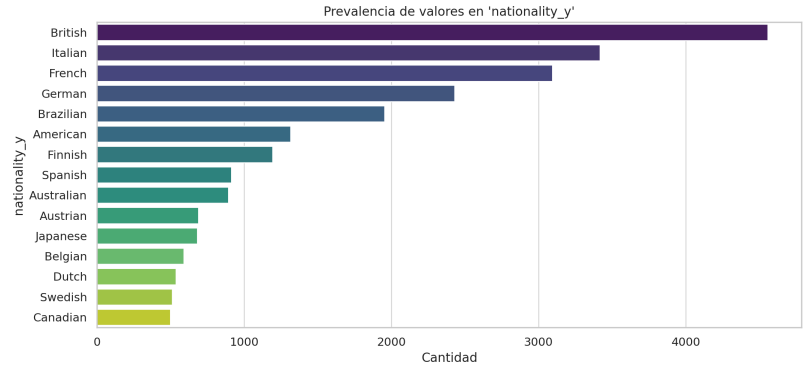
| | | |
|---|---|---|
| year | Numérico | Año en el que se corrió la carrera |
| round | Numérico | Número de la carrera dentro del calendario del año |
| circuit_name | String | Nombre del circuito donde se corrió la carrera |
| circuit_location | String | Ubicación (ciudad o región) del circuito |
| driver_name | String | Nombre del piloto que corrió esa carrera |
| constructor_name | String | Nombre de la escudería del piloto |
| grid | Numérico | Posición de largada del piloto |
| position | Numérico | Posición final del piloto (puede haber valores nulos si no terminó la carrera) |
| status | String | Estado final del piloto (ej. Finished, Accident, Disqualified) |
| points | Numérico | Puntos obtenidos en esa carrera |
| fastestLap | Numérico | Número de vuelta más rápida del piloto (puede ser nulo) |
| fastestLapTime | String | Tiempo de la vuelta más rápida del piloto (puede ser nulo) |
| fastestLapSpeed | Numérico | Velocidad promedio de la vuelta más rápida (puede ser nulo) |
| rank | Numérico | Ranking de velocidad de vuelta entre todos los pilotos en esa carrera |

## El siguiente es un grafico de la prevalencia de los distintos atributos categoricos:


Prevalencia de valores en 'statusId'
statusId
Cantidad
100
80
60
40
20
0
316
321
331
405
431
731
795
810
854
1025
1062
1613
2026
4037
7674


Prevalencia de valores en 'constructorRef'
constructorRef
Cantidad
ferrari
mclaren
williams
tyrrell
team_lotus
sauber
red_bull
renault
minardi
brabham
mercedes
ligier
arrows
brm
toro_rosso
0
500
1000
1500
2000
2500


Prevalencia de valores en 'name_x'
name_x
Cantidad
British Grand Prix
Italian Grand Prix
Monaco Grand Prix
German Grand Prix
Belgian Grand Prix
French Grand Prix
Canadian Grand Prix
Spanish Grand Prix
Brazilian Grand Prix
United States Grand Prix
Hungarian Grand Prix
Australian Grand Prix
Japanese Grand Prix
Austrian Grand Prix
Dutch Grand Prix
0
250
500
750
1000
1250
1500
1750


Prevalencia de valores en 'nationality_y'
nationality_y
Cantidad
British
Italian
French
German
Brazilian
American
Finnish
Spanish
Australian
Austrian
Japanese
Belgian
Dutch
Swedish
Canadian
0
1000
2000
3000
4000


Prevalencia de valores en 'positionText'
positionText
Cantidad
R
F
3
4
2
5
1
6
7
8
9
10
11
12
13
0
2000
4000
6000
8000

## 4. Forma de recolección de los datos

El dataset fue descargado de la página de “kaggle.com”, en la misma página mencionan el origen de los datos:

> The data is downloaded from http://ergast.com/mrd/ and refreshed after each Grand Prix weekend. The data was originally gathered and published to the public domain by Chris Newell.

El dataset está compuesto por distientas tablas que se relacionan entre si. Para obtener una tabla completa lo que hicimos fue agrupar las anteriormente mencionadas con un script de python:

```
import pandas as pd
import os

datos_Path = "datosFormula1/"
datos = ["results.csv","constructors.csv","races.csv","drivers.csv","circuits.csv

datosdf = []
for i in datos:
    if os.path.exists(datos_Path + i):
        print(f"El archivo {i} existe")
        datosdf.append(pd.read_csv(datos_Path + i,na_values=['\\N']))
        print(f"El archivo {i} se ha cargado correctamente")
    else:
        print(f"El archivo {i} no existe")

races_circuits = pd.merge(datosdf[datos.index("races.csv")],datosdf[datos.in
results_constructor = pd.merge(datosdf[datos.index("results.csv")],datosdf[da
results_constructor_drivers = pd.merge(results_constructor,datosdf[datos.inde
races_circuits_results_constructor_drivers = pd.merge(races_circuits,results_c

races_circuits_results_constructor_drivers.to_csv("datosEntrega/races_circuits
```

## 5. Información adicional

- **Datos faltantes:** Algunos atributos (como `position`, `fastestLap`, `fastestLapTime`) tienen valores faltantes. Esto es esperable en carreras donde el piloto no termina o no realiza una vuelta rápida válida.
- **Unificación de valores:** Algunos atributos como `status` contienen múltiples valores para una misma categoría conceptual (por ejemplo, "Accident" y "Collision"). Puede ser útil agruparlos para ciertos análisis.
- **Redundancia y relaciones:** Las variables `grid` y `position` están altamente correlacionadas en pilotos que completan la carrera sin incidentes, pero no siempre reflejan el rendimiento real debido a abandonos u otras situaciones.
- **Valor predictivo:** Atributos como `constructor_name`, `grid` y `fastestLapSpeed` podrían ser útiles para modelos de predicción del resultado (`position`) o puntos obtenidos.