TENNIS-STADISTICS



Un EDA sobre las estadísticas en el mundo del tenis

# TEMÁTICA

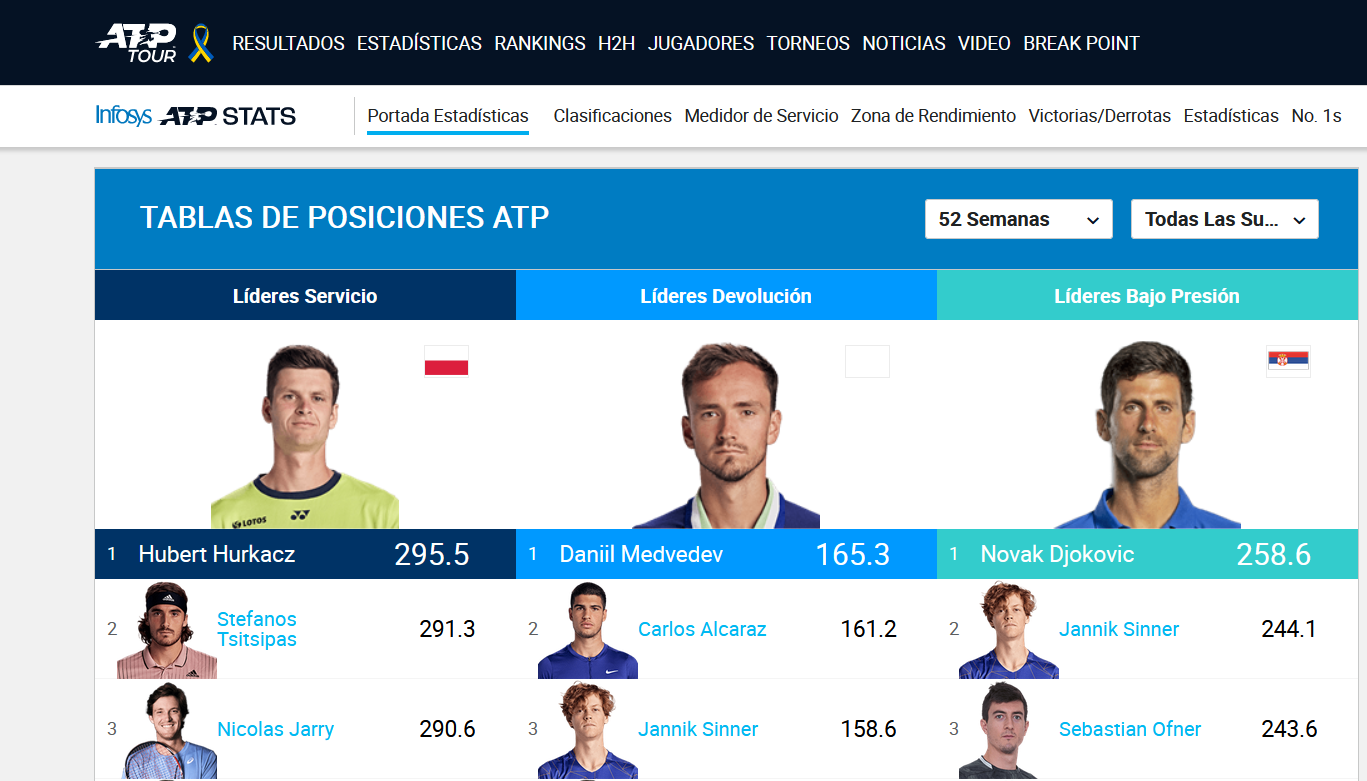
El tema elegido ha sido el tenis.

¿Por qué me he decidido por el tenis?

Principalmente, porque es uno de los deportes que más me gusta, e incluso lo he jugado, con lo que es un tema que no me es desconocido y en el que me puedo defender.

Además, es un campo en el que hay mucha información disponible, con lo que hacer un estudio es sencillo y ofrece muchas posibilidades, tanto para este EDA como de cara a futuro con Machine Learning (o al menos eso espero ☺).

De hecho, la propia ATP (Asociación de Tenistas Profesionales) dispone ya de un apartado de estadísticas en su propia web con información desde 1990 de jugadores en activo y retirados:



# DATASETS UTILIZADOS

He utilizado dos datasets para realizar el EDA.

## FICHERO CSV CON LOS RESULTADOS DE PARTIDOS

### Origen

Este dataset lo he sacado de Kaggle

(<https://www.kaggle.com/datasets/dissfya/atp-tennis-2000-2023daily-pull>) y contiene información de todos los partidos de la ATP desde el año 2000. Se trata de un formato CSV.

### Información contenida

Sobre el CSV inicial he aplicado una serie de transformaciones y adaptaciones para poder trabajar mejor con él:

* He transformado la columna “Date” a tipo Fecha para poder filtrar por fechas.
* He eliminado las columnas “Pts\_1”, “Pts\_2”, “Odd\_1” y “Odd\_2, ya que la información que contenían no la iba a utilizar.

Tras esto, las columnas que quedan son las siguientes:



Esta tabla se puede encontrar también en el notebook “2\_Preparacion\_Datos.ipynb”, presente en el zip con el EDA.

### Dificultades

No he encontrado demasiadas dificultades a la hora de trabajar con este fichero. El principal problema fue el campo de la fecha, ya que hasta el año 2001 siempre mostraba la misma fecha para todo el torneo, cuando lo que contiene es la fecha del partido que está en esa fila.

## API CON ESTADÍSTICAS DE JUGADORES

### Origen

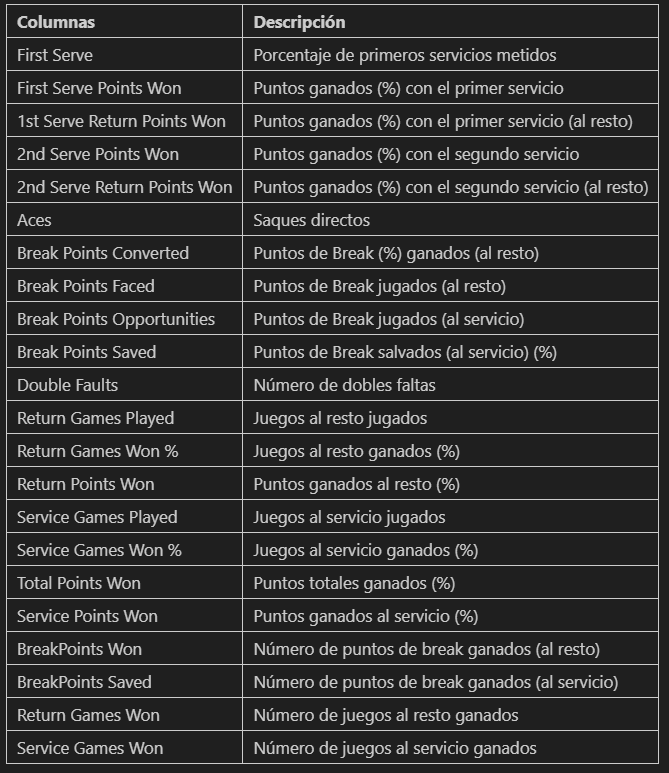
Esta API se encuentra en RapidAPI, que es una web que contiene muchas APIS de distintas temáticas. La API que he elegido ([Ultimate Tenis](https://rapidapi.com/cantagalloedoardo/api/ultimate-tennis1)) contiene información sobre los jugadores en el ranking ATP, WTA (femenino)(tanto personales como profesionales), así como de datos de partidos en directo y las apuestas de esos partidos. Por ello es una API muy versátil de cara a futuro.

### Información contenida

Para este EDA me he quedado con la información del top-50 de la ATP (datos de servicios, juegos ganados…). He aplicado también algunas transformaciones en estos datos.

* El formato del nombre de las columnas no era “normal”, ya que entre cada palabra había un carácter extraño que hacía imposible consultarlas poniendo el nombre, con lo que primero las renombré todas. El nombre, visualmente hablando, es el mismo que tenían antes, pero ahora las palabras están separadas por un espacio en blanco.
* El 90% de las columnas originales son numéricas, pero tenían formato “object”, con lo cual tuve que cambiarles el tipo a float (las que eran porcentajes, tipo XY%) o int64 (las que eran cifras a secas).
* Como a pesar de que venía mucha información, me faltaban algunos datos he añadido columnas que informan de:
  + Puntos de break ganados ("BreakPoints Won")
  + Puntos de break salvados ("BreakPoints Saved")
  + Juegos al servicio ganados ("Service Games Won")
  + Juegos al resto ganados ("Return Games Won")

Con todo esto, las columnas quedan así:



### Dificultades

Los únicos problemas han sido el formato del nombre de las columnas, y la conversión de los tipos de la mayoría de ellas

# HIPOTESIS

Para este EDA he planteado 6 hipótesis, 3 por cada dataset.

## Hipótesis 1

* H0 -> No hay evidencia significativa de la correlación entre número de aces y el número de servicios ganados
* H1 -> Existe una evidencia significativa de la correlación entre número de aces y el número de servicios ganados

## Hipótesis 2

* H0 -> No hay evidencia significativa de la relación entre número de dobles faltas y el número de Break Points salvados
* H1 -> Existe una evidencia significativa de la relación entre número de dobles faltas y el número de Break Points salvados

## Hipótesis 3

* H0 -> La distribución de las columnas numéricas sigue una distribución normal
* H1 -> La distribución de las columnas numéricas NO sigue una distribución normal.

## Hipótesis 4

* + H0 -> No hay evidencia significativa de la correlación entre número de TB jugados en hierba y el número de TB jugados en arcilla
  + H1 -> Existe una evidencia significativa de la correlación entre número de TB jugados en hierba y el número de TB jugados en arcilla

## Hipótesis 5

* + H0 -> No hay evidencia de diferencias significativas en el número de sets jugados en los Grand Slam, Masters 1000 y Copa de Maestros
  + H1 -> Existe una evidencia de diferencias significativas en el número de sets jugados en los Grand Slam, Masters 1000 y Copa de Maestros

## Hipótesis 6

* H0 -> No hay evidencia significativa de la correlación entre número de sets jugados y el número de Tie Breaks jugados
* H1 -> Existe una evidencia significativa de la correlación entre número de sets jugados y el número de Tie Breaks jugados