Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº 4

Esta vez nos toca trabajar con datos de `football_data` que van a encontrar en el campus.

Tema: *Relational Data*

Análisis del conjunto de datasets

Vamos a trabajar con 7 *datasets* de estadísticas de fútbol y apuestas de las 5 mejores ligas europeas entre los años 2014 y 2020. Los datos fueron tomados de acá.

1. Exploren todos los *datasets* para comprender los atributos de cada tabla. La fuente de los datos les puede resultar muy útil para ello (tengan en cuenta que algunos datasets fueron modificados y se les quitaron algunos atributos). En este link y este otro van a encontrar explicado qué son las Xstats. Pueden utilizar comandos como `glimpse()` o `colnames()` para este ejercicio.
2. Hagan un esquema a mano o utilizando una Hoja de Cálculo de Google con los atributos de cada *dataset* (algo así como hicimos para el dataset `vuelos` en la diapositiva 09 de la presentación de esta clase).
3. Utilicen el esquema armado para identificar las claves primarias y las claves foráneas en los diferentes *datasets* (o conjuntos de claves si no fuera suficiente con una sola). ¿Es posible detectar claves primarias en todos las tablas? Verifiquen que la clave o el conjunto de claves elegidas para cada *dataset* identifiquen de forma única cada observación. Para eso pueden utilizar el comando `count()` con las claves primarias y buscar las entradas con n mayor a uno (ver Cap. 13.3 de R para Ciencia de Datos).
4. Piensen entonces cómo podrían conectarse los diferentes *datasets*.

Trabajando con comandos de *Relational Data*

A partir de este momento es **fundamental** que inviertan tiempo en pensar cuál de todos los datasets es el más adecuado para responder la pregunta que les estamos o se están haciendo. Se pueden encontrar las mismas respuestas por diferentes caminos, pero si no lo piensan con detenimiento pueden tomar rumbos mucho más empantanados de lo necesario.

1. Encuentren los 10 equipos que más goles metieron en todas las Ligas de Europa. Para ello pueden usar los comandos `group_by()`, `summarise()`, `order()` o `arrange()` y `head()`. Si no saben cómo funcionan, busquen en la documentación o GOOGLE IT!.
2. Hagan un gráfico de barras que muestre la información del ejercicio anterior, donde cada barra tenga el nombre del equipo correspondiente.
3. Repitan los ejercicios 1. y 2. pero en vez de analizar la cantidad de goles totales, analizar los tiros al arco (*shots*).

4. Realicen un gráfico de dispersión (*scatter plot*) entre las variables goles totales y tiros al arco totales para cada uno de estos equipos. Analicen qué variable pondrían en cada eje. ¿Se observa alguna relación entre la cantidad de tiros al arco y la cantidad de goles convertidos?
5. Realicen el mismo gráfico de dispersión pero teniendo en cuenta todos los equipos (o sea, no tengan en cuenta el filtro de los ejercicios anteriores). ¿Ahora se observa alguna relación?
6. Elijan una de las 5 ligas europeas (*y por qué la Premier League*).
7. Quédense con los 5 jugadores que más goles han metido en esa Liga.
8. Analicen la distribución de tiempos (en minutos) en los cuales estos jugadores realizaron los goles **sin importar en qué club lo hicieron**. Para ello:

    → Consideren que todos los goles realizados son el resultado de tiros al arco (shot). Van a poder entonces encontrar todos los goles realizados a partir del atributo `shotResult` del dataset `shots`.

    → Puede serles útil el comando `semi_join()` como se explica en el Cap. 13.5 de R para Ciencia de Datos.

    → Pueden usar `ggridges` para mostrar las 5 distribuciones en un mismo gráfico.

Messismo explícito: para jugar después de clase

La idea de este ejercicio es comparar las estadísticas de Messi con otros jugadores que hayan tenido un gran desempeño en los aspectos a estudiar. Para ello:

1. Elijan la cualidad de los jugadores que quieren evaluar para su análisis. Por ejemplo, pero sin limitarnos a :
    - Rendimiento Físico
    - Compañerismo
    - Generación de situaciones de gol
    - Jugador clave
    - Fairplay
2. Fijen 2 variables de los datasets que representen la cualidad que quieren estudiar y expliquen por qué fueron estas.
3. Seleccionen algún intervalo de tiempo para su análisis. Pueden ser varias temporadas (*seasons*) o una sola. Por si les interesa:
    - Tengan presente que Messi ganó el balón de oro en los años 2016, 2019, 2020 y 2021. Cada premio tiene en cuenta el desempeño del jugador en la temporada anterior (por ejemplo, el premio del 2016 se otorgó por la temporada agosto 2015 - junio 2016). Sin embargo, consideren también que no sólo se evalúa el campeonato local del jugador,

sino su actuación en torneos internacionales, copas nacionales y su selección nacional.

4. Encontrar los 5 jugadores que tengan el mejor desempeño en cada variable analizada. Esto debería resultar en una lista de a lo sumo 10 jugadores (sin considerar a Messi).
5. Realicen un gráfico de dispersión que localice a cada uno de estos jugadores según las dos variables elegidas.
6. ¿Qué se puede decir de los jugadores a partir de esta comparación? ¿Dónde queda Messi en este plano de evaluación?
7. Repitan los ejercicios anteriores pero con los 5 jugadores con mejor desempeño **dentro de la misma Liga** (entre 2014 y 2020 Messi jugó en el Barcelona, en La Liga).

SÚPER EXTRA: Analizar específicamente las estadísticas de los jugadores del ejercicio 6. y las de Messi cuando se enfrentaron en partidos de La Liga.