Práctica 1

Web Scraping

Gisele Guadalupe Almeida dos Santos Maia

Octubre 2019.

Universitat Oberta de Catalunya.

Máster de Ciencia de Datos (Data Science).

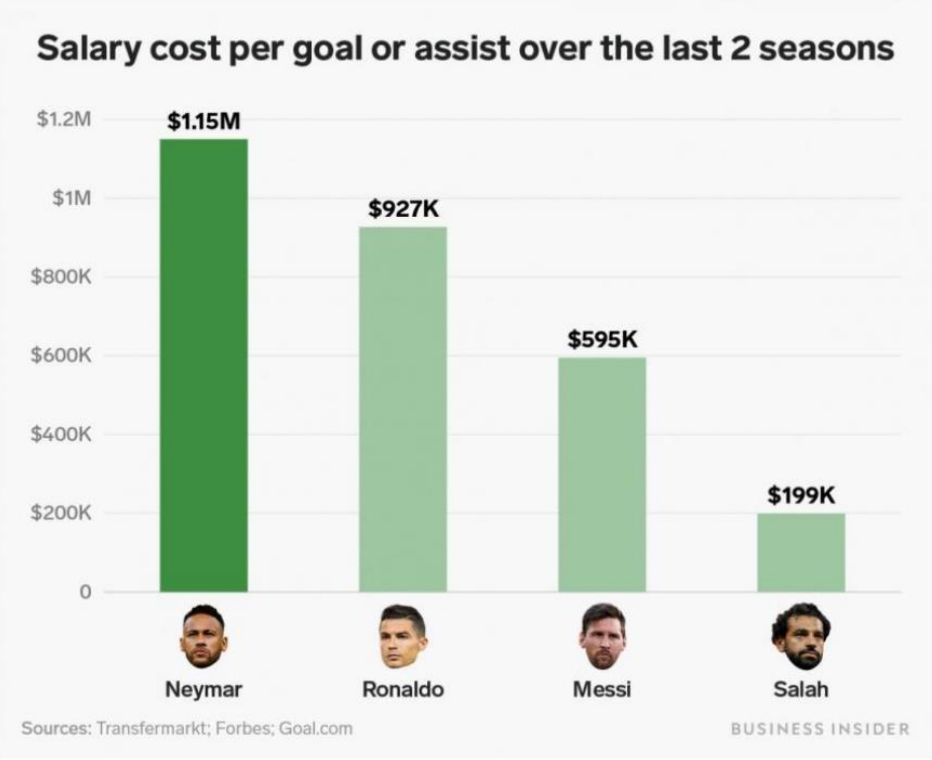
Tipología y ciclo de Vida de los datos

Práctica 1: Web Scraping

1. Contexto

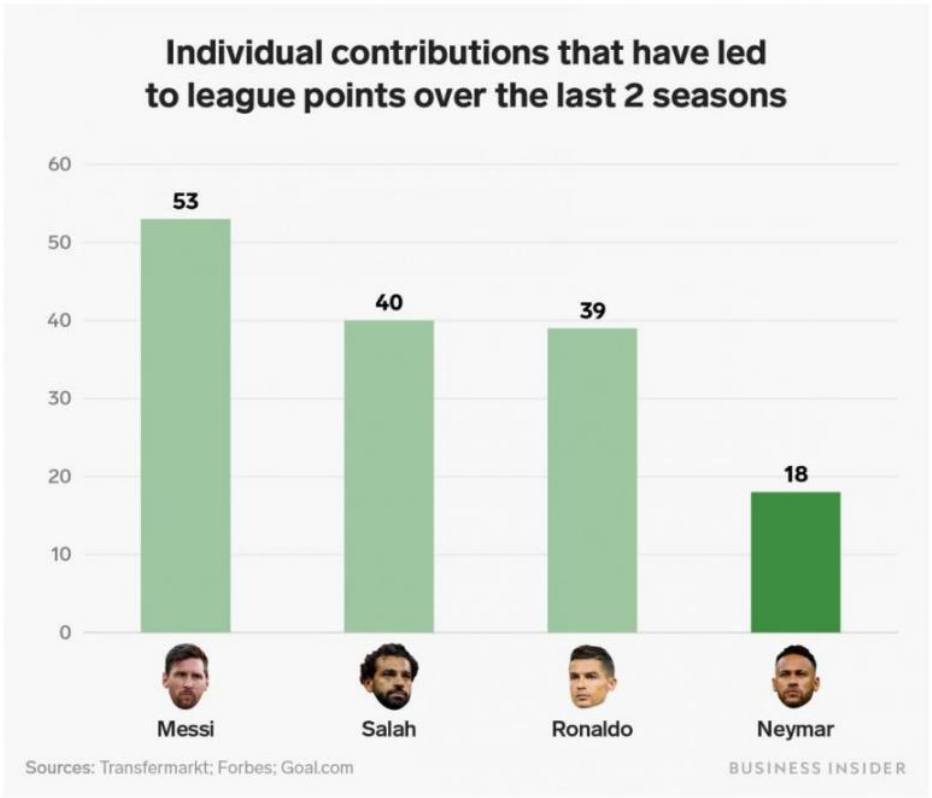
El futbol es el deporte más popular en el mundo ya que es uno de los más antiguos. Además, es sencillo, con reglas simples de entender, de fácil diseminación y da igual la clase social del jugador. Hoy en día, la publicidad ayuda en dejarlo todavía más popular y tiene un papel fundamental en la facturación de los clubes.

También es un deporte que emplea a muchas personas y cómo podemos mirar en la noticia del periódico de Expansión de 30 de enero de 2019: "El fútbol español supone ya el 1,37% del PIB" así, también sabemos que el futbol hay un gran peso en la economía del país y esto peso aumenta a cada año con la ayuda de la televisión y la contratación de jugadores de otros países que influencian en la globalización y conocimiento de los clubes. Por ejemplo, en el periódico Business Insider de 02 de septiembre de 2019 donde se habla de la contratación de Neymar por 220 millones de euros por el equipo Paris Saint-German y "se suponía que iba a ser la gran superestrella del equipo, pero en vez de eso ha resultado ser un enorme desperdicio de dinero para el club". En el noticiario se habla también de su salario comparado con la cuantidad de goles que se ha hecho y la comparación con otros jugadores:



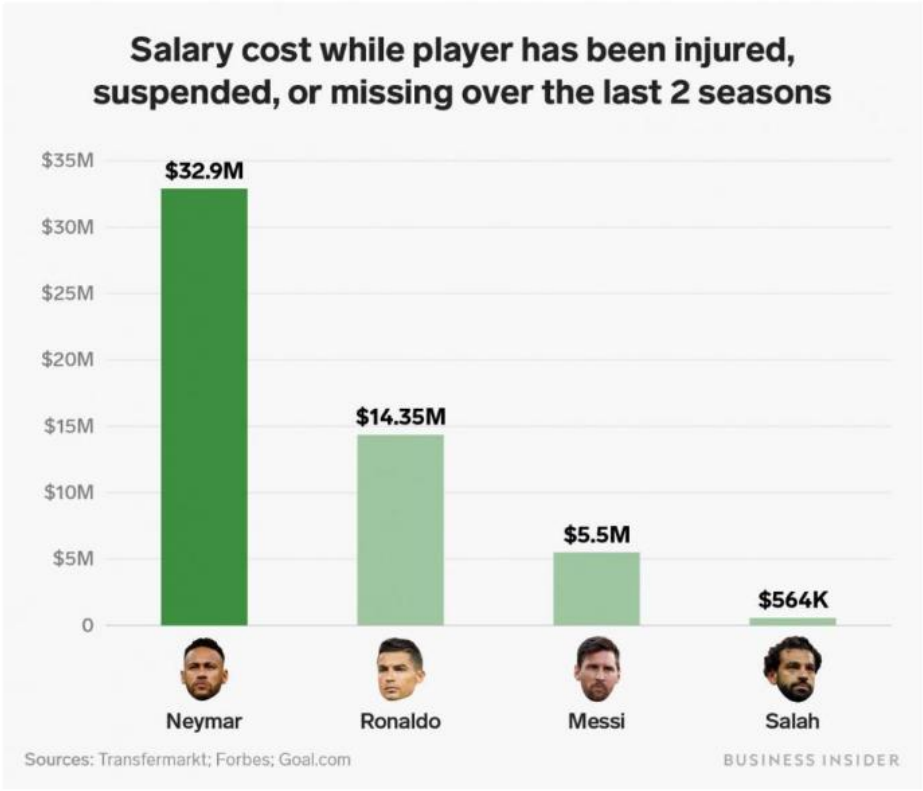
Shayanne Gal/Business Insider

Miramos aquí que Messi hay sido el más eficaz de Europa y Neymar es lo más caro para no mucho eficaz. Además, mirando la cantidad de puntos que cada uno has conseguido en la liga, Neymar es el peor:



Shayanne Gal/Business Insider

Otro problema es Neymar no jugar en 38 partidos (hasta la fecha de publicación del noticiario) por lesión



Shayanne Gal/Business Insider

En este contexto, y sabiendo que más de 3 millones de personas (según la FIFA sobre el mundial de 2018 en 21 de diciembre de 2018) miran en futbol y la importancia que el deporte tiene en la economía mundial, la contratación de un jugador no es hecha de una hora para otra, es necesario analizar los jugadores, sus habilidades, lo que hacen en campo y, cómo la imagen del jugador también es importante y está vinculada con el club, es necesario también analizarnos su vida personal para que no traiga problemas en el futuro.

Así, lo que vamos a buscar para analizar son:

1. Dados en campo de los jugadores brasileños en uno de los principales campeonatos brasileños: La Serie A que es unos de los campeonatos más importantes en Brasil;

El punto 1. vamos a buscar en la página web https://fbref.com/en/comps/24/1559/stats/2017-Serie-A-Stats con el histórico de los años de 2016 hasta 2019 (mismo que todavía no has finalizado el campeonato de la serie A). Esta página web fornece estadísticas de fácil acceso e informaciones muy completas de algunos deportes cómo: basquetbol, baseball, hockey etc., y también el futbol. La página web tiene constante actualización y algunos datos primarios son fornecidos por empresas cómo: Gracenote (Nielsen), Sportradar y DSG (Data Sports Group). Esto último es más específico del futbol.

2. Título para el dataset

En el script tenemos el dataset definido con el nombre: players_brasil_seriea_16_19.csv. Así, su título es:

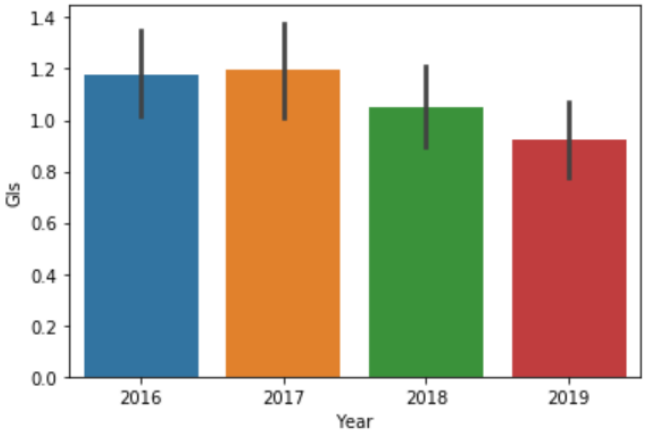
Brasil Soccer Players Performance Serie A 2016 – 2019 ya que tenemos algunos datos del jugador (nombre, edad, equipo y distrito del equipo) y estadísticas de los jugadores de 2016 hasta 2019 (con actualización hasta octubre de 2019).

3. Descripción del dataset

El dataset contiene 25 variables que incluye datos personales: nombre, edad, año del nascimiento, equipo y posición del jugador y su rendimiento en campo con la cantidad de goles, ayudas, minutos en juegos, intento a penaltis etc. Hay las informaciones de 2016 hasta octubre de 2019, el total por año de cada jugador.

4. Representación gráfica.

Uno de los puntos principales que vamos a analizar son los goles. Así, miramos la media de goles por año:



Y, de una manera general, la media de goles por posición del jugador. Además, vemos también la cantidad de jugadores que tenemos en cada posición.

Medias Cantidad

	Pos	Gls
0	DF	0.537649
1	DF,FW	0.000000
2	DF,MF	0.500000
3	DF,MF,FW	1.000000
4	FW	2.135338
5	FW,DF,MF	0.000000
6	FW,MF	2.307692
7	GK	0.014851
8	GK,DF	0.000000
9	MF	1.068287
10	MF,DF	0.797297
11	MF,FW	2.406593
12	MF,FW,DF	0.500000
13	MF,GK	0.000000

	Pos	Gls
0	DF	757
1	DF,FW	1
2	DF,MF	18
3	DF,MF,FW	1
4	FW	532
5	FW,DF,MF	1
6	FW,MF	26
7	GK	202
8	GK,DF	2
9	MF	864
10	MF,DF	74
11	MF,FW	91
12	MF,FW,DF	2
13	MF,GK	1

Algunos juegan en más de una posición.

5. **Contenido.**

El dataset contiene 2.614 líneas con jugadores y su rendimiento entre los años de 2016 hasta 2019. Las informaciones de los jugadores están en 25 variables que son:

Rk: contaje de las filas, no hay ordenación

Player: Nombre del jugador

Nation: La nacionalidad del jugador (en este datasete solamente tenemos brasileños

Pos: Posición que juega en el año de referencia: GK= portero; DF=defensor; MF=mediocampista; FW=delantero; FB=fullback; CD=defensa central; DM=mediocampista defensivo; CM=mediocampista central; WM=mediocampista ancho; AM=centrocampistas de ataque

Squad: nombre de equipo que juega en Brasil

Age: Edad en el año de referencia

Born: año de nascimiento del jugador

Apps: apariciones del jugador o equipo

Starts: juego(s) iniciado por el jugador

Min: minutos de juego

Mn/Ap: minutos de juego por apariciones

Gls: total de goles del jugador (actuación)

Ast: ayudas (actuación)

PK: penaltis realizados (actuación)

PKatt: intentos de penaltis (actuación)

SoT: disparos en el blanco (actuación)

Fls: faltas cometidas (actuación)

CrdY: tarjetas amarillas (actuación)

CrdR: tarjetas rojas (actuación)

Gls_90min: goles hechos por 90 minutos

G+A_90min: goles y ayudas por 90 minutos

G-PK_90min: goles menos penaltis realizados por 90 minutos

G+A-PK_90min: goles más ayudas menos penaltis hechos por 90 minutos

SoT_90min: disparos en blanco por 90 minutos

Year: año de referencia del campeonato de la Serie A - de 2016 hasta octubre de 2019.

Los datos fueran recogidos utilizando web scraping en python con BeautifulSoup. La conexión fue hecha más de una vez porque recogemos datos en 4 páginas web distintas. Es posible mirar el paso a paso del script de nombre: Web Scarping - Players Soccer – Practica1.

6. Agradecimientos.

Agradecer a la página web https://fbref.com y sus colaboradores: Data Sports Group y Stats Bomb por dejar las estadísticas de los jugadores actualizadas. Las informaciones fueran recogidas en octubre de 2019. Con la última conexión hecha en 09 de octubre de 2019.

7. Inspiración

Cómo he comentado en el contexto, el futbol es uno de los principales deportes, el más conocido mundialmente y tiene una gran importancia en la economía de los países. Así, tener en cuenta el jugador que se está contratando es muy importante y el primer paso es saber su rendimiento en campo, se ayuda el equipo, cuantos intentos hace para marcar un gol si es agresivo (tiene histórico de tarjetas amarillas o rojas) etc. Estas son algunas preguntas que se puede responder.

8. **Licencia.** Para el dataset la licencia utilizada es: Released Under CC BY-NC-SA 4.0 License o Creative Communs Non-Commercial y Share-Alike donde es una licencia Internacional donde se puede compartir, hacer copia, unir, transformar o crear utilizando el dataset. Pero, hay que disponer e indicar las modificaciones con el mismo tipo de licencia. Y no se puede utilizarlo para fines comerciales.

9. **Código**

Para la captura de los dados se ha utilizado web scraping con Python y en el Jupyter Notebook con utilización de librerías cómo: Pandas, Requests, BeautifulSoup. El nombre del código es: ***Web Scarping - Players Soccer - Practica1.ipynb.*** Es posible mirar el código paso a paso comentado.

10. **Dataset**: El nombre del dataset es: ***players_brasil_seriea_16_19.csv***. Se guarda en la extensión .csv.

11. Componentes del grupo

Contribuciones	***Firma***
Investigación previa	*Gisele Guadalupe Almeida dos Santos Maia*
Redacción de las respuestas	*Gisele Guadalupe Almeida dos Santos Maia*
Desarrollo código	*Gisele Guadalupe Almeida dos Santos Maia*

12. Referencias

- Expansión [2019] [En línea]

  España: Unidad Editorial Información Económica S.L [consulta en 26 de octubre de 2019]

  https://www.expansion.com/directivos/deportenegocio/2019/01/30/5c517ee9e2704e22598b45d6.html

- Business Insider España [2019] [En línea]

  España: Barnaby Lane [consulta en 26 de octubre de 2019]

  https://www.businessinsider.es/neymar-graficos-muestran-ha-sido-derroche-dinero-psg-484239

- Fifa [2019] [En línea]

  Zurich: Fedération Internationale de Football Association [consulta en 26 de octubre de 2019]

  https://es.fifa.com/worldcup/news/mas-de-la-mitad-del-planeta-disfruto-de-un-mundial-incomparable-en-2018