Estudiante:	Anddy Paúl Aldave Valle, aaldave@uoc.edu Xus García de la Vega Matas, jgarcia_de_la_vegam@uoc.edu
Asignatura:	Tipología y ciclo de vida de los datos
Practica:	Práctica 1
Titulación	Máster en Ciencia de datos

Práctica 1

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique porque el sitio web elegido proporciona dicha información.

El máximo organismo europeo a nivel futbolístico es la UEFA (Unión de Asociaciones Europeas de Fútbol) y esta se encarga de organizar los distintos campeonatos que se puedan realizar en Europa, así como promover, desarrollar y controlar cualquier aspecto a nivel de dicho deporte.

Cualquier *site* a parte de tener una interfaz amigable, una navegación fácil e intuitiva, un buen diseño… debe tener un contenido de calidad. Esto último implica que la información que se facilite debe estar destinada al usuario, aportándoles valor y que puedan profundizar sobre el tema. Y lo más importante, que sean veraces.

La UEFA consigue trasladar estas características en su página oficial; donde podemos encontrar un apartado de noticias, las distintas competiciones que organiza con sus correspondientes resultados, las clasificaciones, videos resúmenes y un largo etcétera, todo ello relacionado con la actividad que desarrollan. De esta forma consiguen trasladar al usuario toda su actividad.

En la práctica que nos atañe, hemos trabajado en el apartado relacionado con los miembros de dicha asociación. Donde por un lado se puede ver el resumen de las principales ligas y en la parte derecha, el enlace a todas las ligas que forman parte del organismo.

Este último apartado es donde se ha focalizado la recolección de datos, de los que se han creado los distintos códigos para tener una clasificación por países y dentro de cada uno de ellos, los equipos que lo forman. Por cada uno de los equipos se ha recopilado la información relacionada a la liga que pertenece, obteniendo así la posición en la clasificación, los partidos jugados, los puntos…

2. Definir un titulo para el dataset. Elegir un titulo que sea descriptivo.

Clasificación actualizada de equipos de fútbol profesional en las ligas europeas

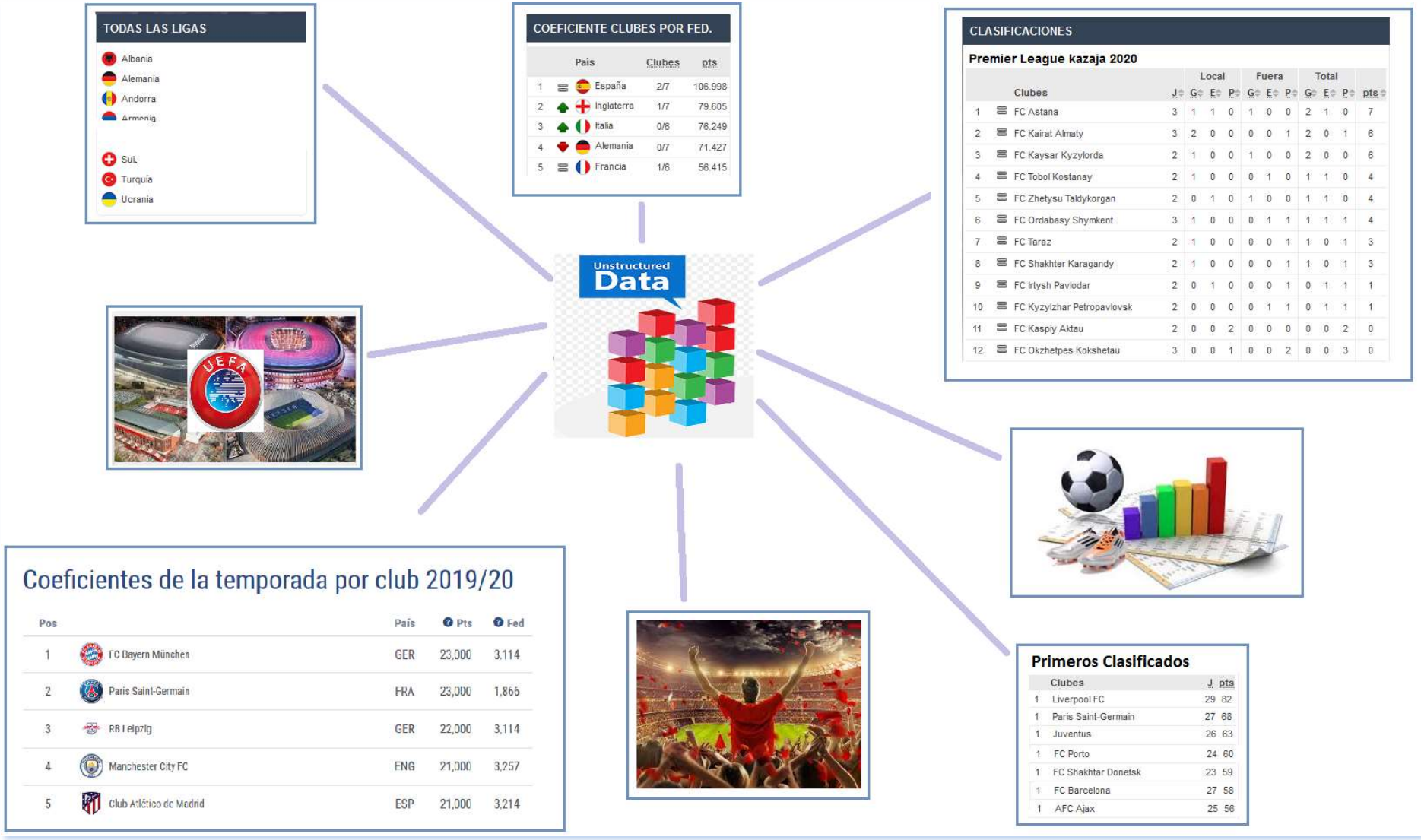
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Un dataset es un una colección de datos, normalmente tabulada. En este caso el conjunto de datos procede de la página oficial del mayor organismo futbolístico a nivel europeo.

La información que pública la UEFA corresponde, entre otra, a todas las ligas de fútbol profesional europeas con sus correspondientes clasificaciones de los equipos según los partidos que han jugado.

Los datos extraídos corresponden precisamente a todos los equipos de estas ligas europeas con los puntos, posiciones, partidos… de cada uno de ellos. De ahí que se haya escogido el nombre anteriormente mencionado: *Clasificación actualizada de equipos de fútbol profesional en las ligas europeas*.

4. Representación grafica. Presentar una imagen o esquema que identifique el dataset visualmente.



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y como se ha recogido.

Para la creación del dataset, se ha accedido a la web de trabajo y se ha descargado mediante la librería *BeautifulSoup*. Para cada link que encontramos, se guarda el link que será el que va a contener la información por países.

Por países vamos obteniendo de forma iterativa los equipos que componen dicho país con los correspondientes datos de la clasificación. Por cada equipo obtenido, se guarda la información en un vector que contendrá todos los atributos y que posteriormente, una vez recopilados en su totalidad [los equipos] los guardaremos en el *csv*.

Los datos han sido capturados desde la página web oficial de la UEFA y los datos sobre las clasificaciones son datos oficiales hasta la última jornada que se pudo practicar este deporte. Cabe decir que no existe histórico de ningún tipo, esto implica que si se quisiera guardar la información de las clasificaciones por fechas o algún otro criterio, se debería ejecutar el programa de forma semanal y guardarlos.

El dataset incluye los siguientes campos:

Variable	Descripción
countryName	País al cual pertenece el equipo
teamName	Nombre del equipo de futbol
gamesPlayed	partidos jugados
wHome	partidos ganados en casa
dHome	partidos empatados en casa
lHome	partidos perdidos en casa
wAway	partidos ganados fuera de casa
dAway	partidos empatados fuera de casa
lAway	partidos perdidos fuera de casa
wTotal	partidos totales ganados
dTotal	partidos totales empatados
lTotal	partidos totales perdidos
points	total de puntos obtenidos
position	posición en la clasificación

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

La UEFA (Unión de Federaciones Europeas de Fútbol), es la confederación europea de asociaciones nacionales de fútbol y máximo organismo de este deporte en el continente europeo. Agrupa a todas las federaciones nacionales a lo largo de toda Europa.

Fue fundada en 1954 y es la encargada de organizar los campeonatos nacionales de Europa, la Eurocopa masculina y femenina, la Liga de Campeones masculina y femenina y la Europa League.

Sus objetivos principales, además de organizar los principales campeonatos de Europa, son: promover el fútbol en espíritu de unidad, solidaridad, paz, comprensión y juego limpio sin ningún tipo de discriminación y apoyar y salvaguardar a las federaciones por el bienestar general de fútbol europeo.

Hemos realizado un análisis del fichero robots.txt ubicado en https://es.uefa.com/robots.txt ya que en este fichero se indican las restricciones que deberíamos tener en cuenta cuando realizamos un rastreado y hemos comprobado que nuestro directorio https://es.uefa.com/memberassociations no está restringido y por lo tanto, es apto para poder ser rastreado a través de una herramienta como la que hemos desarrollado.

Existen diversas webs que resumen las clasificaciones de las principales ligas europeas como son: Alemania, España, Italia, Inglaterra, Francia, Holanda, pero no hemos encontrado ninguna que informe de las clasificaciones de todos los países europeos que están asociados a la UEFA.
Las webs que hemos encontrado que muestran la clasificación de las principales ligas europeas son:

- https://www.mismarcadores.com
- https://www.superdeporte.es/deportes/futbol/
- https://www.scoreboard.com/es/futbol/

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El fútbol es el deporte más popular no solo del continente europeo sino del mundo, se estima que tiene más de 4 mil millones de seguidores, alcanzando cifras del 60% de la población total.
Por ello, consideramos que la extracción de cualquier tipo de información de fuentes oficiales asociadas al fútbol y generación de datos listos para ser analizados por distintas plataformas, webs, prensa, aplicaciones móviles, etc., puede llegar a tener una repercusión muy importante teniendo en cuenta la cantidad de usuarios finales a quienes les puede llegar la información a través de diferentes medios.

Las preguntas que se pretenden responder son:

- ¿Qué liga europea de fútbol es la más competida y complicada de ganar? ¿Y la más fácil?
- ¿Qué liga europea de fútbol tiene más goles de media?
- ¿Qué país tiene la liga de fútbol con más equipos en primera división? ¿Y el que menos?
- ¿Qué equipo / equipos de fútbol europeo son los más goleadores? ¿Y los menos goleados?
- ¿Qué equipos son los mejores en casa a nivel de cada país y también a nivel de toda Europa?
- ¿Qué equipos son los mejores fuera de casa a nivel de cada país y también a nivel de toda Europa?
- ¿Hay equipos invictos (que no han perdido aún) en Europa?
- A final de temporada, se puede realizar un comparativo entre todas las ligas y poder responder:
    - Qué equipo de Europa ha obtenido la mejor puntuación, el más goleador y el menos goleado? ¿Cuántos puntos hacen falta de media para ser campeón en cada país europeo?
- Debido a la situación actual por la pandemia del COVID-19:
    - ¿Qué ligas se pueden dar por finalizadas, según la diferencia de puntos entre los primeros clasificados?
    - ¿Cuántas jornadas faltan para dar por terminadas las ligas europeas de fútbol por países?

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

Consideramos que nuestro dataset debería tener una licencia del tipo Released Under CC BY-NC-SA 4.0 License. La elección de esta licencia se basa en:

- Cualquier persona o empresa que la utilice deberá reconocer la autoría de quienes construyeron el dataset y también indicar si realizaron cambios o no en el mismo. De esta forma se nos reconocerá nuestro trabajo.
- No permitimos el uso comercial, ya que, no nos gustaría que otras empresas obtengan beneficios económicos a costa de nuestro trabajo.
- Los nuevos desarrollos o adaptaciones realizadas sobre nuestro dataset, deberá difundir nuestras contribuciones de la misma forma que la licencia original, de tal forma, que se seguirá reconociendo nuestra autoría, después de cualquier actualización.

Teniendo en cuenta que de todos los tipos de licencias proporcionados en el enunciado, la única que se ajusta a nuestro modo de trabajo y reconocimiento seleccionamos Released Under CC BY-NC-SA 4.0 License.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código fue desarrollado en Python utilizando el Entorno de Desarrollo Integrado (IDE) Wing Personal 7.2, el código fuente se encuentra en el fichero comprimido código.zip y también en el repositorio de github: https://github.com/aaldaveva/web-scraping
El proyecto consta del siguiente código desarrollado en Python:

- country.py: Modela un país
- countryTeam.py: Modela un equipo de un país
- scraper.py: Clase inicial que será la encargada de desencadenar todo
- scraperCountries.py: Clase que contiene toda la lógica de negocio de la aplicación

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

El proyecto se encuentra alojado en el repositorio de Zenodo:
https://zenodo.org/record/3748300#.XpF4ac1S9hE

11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

El proyecto se encuentra alojado en el repositorio de Github:
https://github.com/aaldaveva/web-scraping

Contribuciones	Firma

Investigación previa	APAV, XGdlVM
Redacción de las respuestas	APAV, XGdlVM
Desarrollo código	APAV, XGdlVM