Programación para *Data Science*

Unidad 8: Visualización de datos en Python

Introducción

En este Notebook encontraréis un conjunto de **actividades evaluables** como PEC de la asignatura.

Veréis que cada una de ellas tiene asociada una puntuación que indica el peso que tiene la actividad sobre la nota de la PEC. Adicionalmente, hay un ejercicio opcional, que no tiene puntuación dentro de la PEC, pero que se valora al final del semestre de cara a conceder las matrículas de honor y redondear las notas finales. Podéis sacar la máxima nota de la PEC sin necesidad de hacer este ejercicio! El objetivo de este ejercicio es que sirva como pequeño reto para los estudiantes que quieran profundizar en el contenido de la asignatura.

Además, veréis que todas las actividades tienen una etiqueta que indica los recursos necesarios para llevarla a cabo. Hay tres posibles etiquetas:

- `NM` **Sólo materiales**: las herramientas necesarias para realizar la actividad se pueden encontrar en los materiales de la asignatura.
- `EG` **Consulta externa guiada**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, pero el enunciado contiene indicaciones de dónde o cómo encontrar la información adicional necesaria para resolver la actividad.
- `EI` **Consulta externa independente**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, y el enunciado puede no incluir la descripción de dónde o cómo encontrar esta información adicional. Será necesario que el estudiante busque esta información utilizando los recursos que se han explicado en la asignatura.

Es importante notar que estas etiquetas no indican el nivel de dificultad del ejercicio, sino únicamente la necesidad de consulta de documentación externa para su resolución. Además, recordad que las **etiquetas son informativas**, pero podréis consultar referencias externas en cualquier momento (aunque no se indique explícitamente) o puede ser que podáis hacer una actividad sin consultar ningún tipo de documentación. Por ejemplo, para resolver una actividad que sólo requiera los materiales de la asignatura, puedéis consultar referencias externas si queréis, ya sea tanto para ayudaros en la resolución como para ampliar el conocimiento!

En cuanto a la consulta de documentación externa en la resolución de los ejercicios, recordad **citar siempre la bibliografía utilizada** para resolver cada actividad.

Ejercicios para la PEC

A continuación, encontraréis los **ejercicios que debéis completar en esta PEC** y que forman parte de la evaluación de esta unidad.

Ejercicio 1

Goodreads es una comunidad web donde los lectores y usuarios pueden mantener un registro de sus libros, sus hábitos de lectura y revisar y evaluar los libros.

En este ejercicio exploraremos los datos de **Goodreads** con el dataset `books.csv`. Este dataset se ha extraído y adaptado de Kaggle. `NM`

1. Cuántos libros hay según el idioma (*language_code*)? Cuál es el idioma más frecuente? Haz una gráfica del número de libros por idioma sin considerar el más frecuente. (1 punto)

1. Muestra el top5 de escritores con más libros publicados. (1 punto)

1. Qué escritor del top5 tiene la puntuación media (*average_rating*) más alta? Responde gráficamente esta pregunta utilizando *violinplots*. (0.5 puntos)

Nota 1: Podéis consultar información referente a *violinplots* en este link.

Nota 2: Los *violinplots* son similares a los *boxplots*, pero adicionalmente también muestran la densidad de los datos en los diferentes valores. Normalmente un *violin plot* incluye un marcador para la media de los datos y un marcador que indica el intervalo intercuartil.

Representa las respuestas gráficamente.

(2.5 puntos)

In [1]:

```
# Respuesta
```

Ejercicio 2

Un equipo de investigación de los Estados Unidos está haciendo un estudio del impacto del COVID-19 en los diferentes estados del país. Es por eso que requiere de nuestro servicio para estudiar la tasa de mortalidad del país. En este ejercicio trabajaremos con el *dataset* `covid19_US.csv`. EG

Se nos han asignado dos tareas:

1. Estudiar la relación entre la tasa de mortalidad (*Mortality_Rate*) y las tasas de hospitalización (*Hospitalization_Rate*) y de tests (*Testing_Rate*). Para esta pregunta podéis utilizar el tipo de gráfico que os parezca más adecuado para contestar la pregunta gráficamente. (1 punto)
2. Construir un mapa **interactivo** que muestre los puntos donde hay más mortalidad del país. Para esta tarea nos piden crear un mapa utilizando la librería *geoplotlib* que nos debe permitir ver la distribución del Covid-19. Es importante que en el mapa se distingan las zonas con la tasa de mortalidad inferior y superior a 4% en diferentes colores (verde y rojo respectivamente). (1 punto)

(2 puntos)

In [2]:

```
# Respuesta
```

Ejercicio 3

Teniendo en cuenta los datos obtenidos en el ejercicio anterior, el equipo americano también nos pide llevar a cabo un análisis descriptivo. En este ejercicio utilizaremos el *dataset* `covid19_US_cases.csv`, que muestra la evolución de los casos de Covid-19 en los diferentes estados desde finales de enero de 2020. NM

1. Cuál es la distribución de casos en Estados Unidos el último mes (abril)? **Muestra gráficamente los resultados para cada grupo (activos, recuperados y muertos), pero sin separar por estados.** (1 punto)
2. Adicionalmente, se nos pide generar una gráfica que muestre la evolución de los casos confirmados en el estado de Nueva York, Massachusetts, Washington, California y Texas desde el inicio de la pandemia. Qué región es la más afectada? (1.5 puntos)

Nota 1: En el apartado 1, para sumar todos los casos de US se puede utilizar la función groupby de pandas que hemos visto en lecciones anteriores.

Nota 2: La visualización de los datos debería ser atractiva y clara a fin de que se puedan sacar conclusiones de ellas.

(2.5 puntos)

In [3]:

```
# Respuesta
```

Ejercicio 4

A partir de los datos de twitter se pueden crear redes de relaciones entre sus usuarios. En este ejercicio trabajaremos con el conjunto de datos `politics_twitter.csv`, que constituye una red de las relaciones entre diferentes políticos de España generada a partir de extracciones de twitter. EG

Cada fila representa un link entre 2 políticos (P1 y P2) y el valor (*Weight*) indica la fuerza de la conexión, en este caso 1 ya que sólo hace referencia a la presencia de esta relación.

1. Cuántas comunidades se pueden detectar en el grafo? (0.5 puntos)
2. Representa gráficamente el grafo y pinta de diferente color cada comunidad detectada. (1.5 puntos)
3. Cuál es el político más importante de esta red? (1 punto)

3. Cuál es el político más importante de esta red? (1 punto)

Nota: La importancia de un nodo dentro de una red se puede medir con el número de nodos a los que está conectado. Esta característica se conoce con el nombre de **degree_centrality**. Puede ver este link para más información:

(3 puntos)

In [4]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import networkx as nx

# Cargamos los datos del fichero book1.csv en un dataframe
twitter = pd.read_csv('data/politics_twitter.csv')

# Visualizamos las primeras filas del dataset
twitter.head()
```

Out[4]:

	P1	P2	Weight
0	KRLS	IreneMontero	1
1	LauraBorras	IreneMontero	1
2	gabrielrufian	IreneMontero	1
3	AdaColau	IreneMontero	1
4	Jaumeasens	IreneMontero	1

In [5]:

```
# Datos disponibles para el ejercicio

# Creamos un objeto grafo vacío
G_twitter= nx.Graph()

# Iteramos para todas las filas del dataset
for _, edge in twitter.iterrows():
    G_twitter.add_edge(edge['P1'], edge['P2'], weight=edge['Weight'])
```

In [6]:

```
# Respuesta
```

Ejercicio Opcional

La saga Marvel es conocida por las complicadas interacciones entre sus personajes. Al estar formada por diferentes películas, tiene sentido que la importancia de los personajes cambie durante la saga. Así pues, estudiaremos la evolución de los personajes mediante la red de relaciones de las películas de *Avengers* (Avengers, Avengers infinity war, Avengers age of Ultron, Avengers End game).

Trabajaremos con la base de datos `marvel_movies.csv`, que constituye una red de las relaciones entre los personajes de toda la saga de *Marvel*. `EI`

- Cuál es el personaje más importante de cada película de Avengers?
- Muestra gráficamente cómo varía la importancia de Tony Stark, Thor, Steve Rogers, Bruce Banner y Natasha Romanov en cada película.

In [7]:

```
# Respuesta
```