Data Science Society

PANDAS SANDBOX

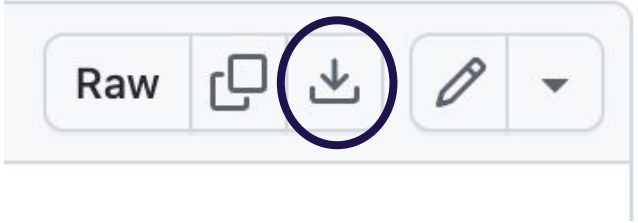
Wednesday, October 23rd

The Diamond, Workroom 2

16:30 - 18:30

- Google “ **Sheffield University Data Science Society”**
- **> 2024-2025 > Sandbox Sessions > Pandas > Pandas' SandBox.ipynb**

1- Download the file



2- Do one of the following:

- **open it using VS code**

  Or

- **open Google drive and upload the file to it**

Open Pandas Guide And VS code Guide from our website

Pandas Guide

VS Code Guide

https://uosdss.wordpress.com

Step1: Setup

Install , import Library , Import Data

challenge 1

- Display the first five rows of the data frame
. - Get a summary of the data.
- How many shows are there in the data set?
- Get the year that most shows were released on

challenge 2:

- How often does each country appear in the dataset?
- sort them descendingly
- Which country produces the most?

challenge 3:

- Sort the data by year in ascending order and inplace it.
- Filter the dataset to show only shows released in 2020.
- which year recorded the highest number of shows? What type of shows was mostly released in that year

challenge 4

- Create new 2 data frames, one for the data of movies only. One is for data on TV shows only.
- delete the type column from both
- add a column 'has_director' to indicate if the show has a director or not

challenge 5

Group the data by "type" (Movie/TV Show) And Count how many movies and TV shows were added each year.

Challenge 6

Create a new column, is_recent, that flags as "Recent" if it contains releases from the last five years (relative to the current year).

challenge 7

Task: Use a pivot table to summarise and analyse the distribution of content ratings by year and type (Movie/TV Show).
Fill empty spaces with 0

challenge 8

Modify the "rating" column to group similar ratings together. For example, group all "PG-13", "TV-14", and "TV-PG" ratings under a new label "Teen". and do a pivot table again