Tutorial

1. Join the discord: <https://discord.gg/PxDbvZu>
2. Download archive.zip (taken from Kaggle)
3. Extract the zipfile, and launch Jupyter Notebook
4. Load the data into pandas

Activities

1. Summarise the data
   1. What does .head() do?
   2. What does .describe() do?
   3. How can you select a specific column?
   4. How many movies were produced per year? Can you show this in a bar chart?
2. Slicing data
   1. If you want a specific row, how do you get it?
   2. What’s the difference between iloc and loc?
   3. Can you show all the rows with more than the average number of votes?
3. Copy
   1. When creating a new dataframe from your previous one (df2 = df), what happens if you edit one?
   2. What does .copy() do? Why might it be useful?
4. Create new columns
   1. Create a new column called ‘centered rating’. This column will contain the rating minus the average (if the average was 5, and an original rating was 8, the new rating will be 3)
5. Apply
   1. What does .apply() do?
   2. Can you use it to change all the runtimes that are greater than 100, to just 100?
6. Final challenge
   1. Create a new column for each genre. Fill these with 1s or 0s depending on if the movie has that genre or not