Introduction : What worked well and what worked badly with AI

Using AI during this project worked well in several areas. It helped to generate code, such as cleaning titles, converting data types and helping to create visualisations. The AI was particularly useful when I needed quick explanations of errors or unfamiliar code (like regex). It also made it easier to explore different types of charts. However, there were also limitations. Sometimes the AI provided code that didn't fully match the structure of my dataset, which caused errors until I adjusted it manually. It occasionally suggested overly complex solutions when simpler ones existed, and it could not see my data directly unless I explained it thoroughly or uploaded it. Because of this, I still had to understand the code myself rather than rely on AI completely. Overall, AI was very helpful for guidance and troubleshooting, but it still required careful checking, testing, and my own understanding to get everything working correctly.

Prompts used;

- *How do I remove the '$' and ',' in my pandas df column 'Gross' and store it as a float and fill empty cells?*

- *What was the need for r and the \ in the [] and what is regex?*

- *I have the following site https://www.the-numbers.com/box-office-records/domestic/all-movies/cumulative/all-time. In this site there is a table, and the site has many pages that follow a consistent URL pattern. The 2nd page is same URL /101, the 3rd is same URL /201, etc. I want to build a list of all URLs to scrape using a for loop, how do I create this?*

- *I am trying to extract the text from each of the table data from the table of a site using this code:*

```
table = soup.find("table")
rows = table.find_all("tr")

for row in rows:
    cols = row.find_all("td")
    rank = cols[0].text.strip()
    year = cols[1].text.strip()
    movie = cols[2].text.strip()
    distributor = cols[3].text.strip()
    domestic = cols[4].text.strip()
    international = cols[5].text.strip()
    worldwide = cols[6].text.strip()
    all_data.append([rank, year, movie, distributor,
    domestic, international, worldwide])
```

but I get this error:

```
IndexError                                Traceback (most recent
call last)
/tmp/ipython-input-1591964235.py in <cell line: 0>()
36     for row in rows:  # skip header row
37         cols = row.find_all("td")
```

```
---> 38          rank = cols[0].text.strip()
     39          year = cols[1].text.strip()
     40          movie = cols[2].text.strip()

IndexError: list index out of range
```

Explain this error.

- *I want to clean domestic, international and worldwide income columns the same as the gross column how do I do that?*

- *I have two movie datasets one from IMDb (Series Title) and one from The Numbers (Movie). The titles have slight differences or punctuation. Write a Python function to clean and standardise the titles (lowercase, punctuation, and extra spaces) so that I can merge the datasets by the movie name as the common column.*

- *I have a pandas df with the column movies how do I check the column for duplicated entries.*

- *Duplicates = df[df['movies'].duplicated(keep=False)] print(duplicates) , what does keep = false do ?*

- `sns.histplot(final_merged["Runtime"], bins=10, kde=True) plt.title("Distribution of movie runtimes") plt.xlabel("Runtime Mins") plt.ylabel("Number of Movies") plt.show()`
  How do I only take the first 10 rows ?

- *How do I check the type of values stored in a column*

- *I have the columns director, movies and world gross I want to group the movies by director and display the highest earning directors on a bar chart.*

- How do I take one of my directors :Anthony Russo and display his top three movies a Bar chart based off of gross income

- I have the column distributors and movies how do I display the number of movies made by distributors