**Introduction**

By now, you have gone through several modules that gave you knowledge and experience with Python fundamentals and two popular Python libraries for data science, NumPy and Pandas. In this assignment, you will put everything you have learned together to create your own Jupyter Notebook that performs data analysis on a real data set and incorporates best practices, such as code commenting and Markdown cell creation.

Scenario

You are working for **SmithWorks Studios**, a new movie studio company with the goal of producing the world's best movies. The company wants to analyze data on previous movies from **IMDb**, a popular source for movie ratings, to get ideas of the types of movies it should produce to maximize its profit. To help them, you will use a data set that contains the 100 top-rated movies from the past decade along with various pieces of information about the movie, its actors, and the voters who have rated these movies online.

About the Assignment

**Where do I get the data?**  
You can download the movie data set provided as an attachment at the bottom of this segment.

**Where do I write the code?**  
You will create your own Jupyter Notebook for submission. The Jupyter Notebook you create should have an appropriate quantity and quality of commented code and Markdown cells.

**How do I submit the assignment?**  
You will need to upload your Jupyter Notebook (a .ipynb solution file) under “Submission” on the last segment of this session.

Some Assignment Tips

Here are some tips for the assignment:

* Go through the data dictionary thoroughly before starting the assignment. That will give you a good idea of what each column represents before you begin the analysis.
* Read each instruction carefully, identify the task to be performed, and only then write the required code. The assignment is meant to be straightforward. You do not need to perform additional analyses that are not requested explicitly.
* Some of the tasks might require using functions you may not have used previously. In such cases, you should rely on the library documentation you referred to in the modules. Please understand that completing this assignment is a learning process, and research is part of it.
* Always run the cells of the Notebook sequentially, restart the kernel, and run all the cells to avoid runtime errors.
* Many of the questions will require you to view them from multiple angles. In other words, there will be no fixed answers. You are expected to apply your problem-solving skills to come up with solutions and also document your work appropriately; both of these are part of the assignment grading.

**Problem Statement**

Assignment Tasks

Here are the tasks you need to perform:

1. Import the data set into a Pandas DataFrame.
2. Generate descriptive statistics for the budget of all the movies.
3. Find out how many of the top-rated movies produced in the United States have a PG-13 rating.
4. Find out whether any of the top-rated movies produced in 2014 were not produced in the United States.
5. Find the percentage of the top-rated movies that are in:
   * 1 genre only
   * 2 genres only
   * 3 genres only
6. Convert the **budget** and **gross**values from “dollars” to “dollars in millions” for all top-rated movies. Round the converted values down to 3 decimal places. For example, a value of 192,345,273 should be converted to 192.345.
7. List all details for the top 10 movies with the highest profit, sorted from highest to lowest. **Hint**: Profit is not a column in the DataFrame. You will need to calculate it.
8. Generate a list of all the actors, in alphabetical order by the first name, that have starred in a top-rated movie. If an actor has starred in multiple movies, their name should appear only once on the list. Assume that all actors’ names are in the format *<first\_name> <last\_name>*.
9. The movie studio wants to determine who it should approach to act in its next movie production. Find the top 3 actors who appeared in the most top-rated movies.
10. Create a data visualization that shows each country and the number of top-rated movies produced in it. Find the country that produced the most top-rated movies.

**Evaluation Rubric**

The weights for the tasks in the assignment are given below.

| **Criteria** | **Meets Expectations** | **Does Not Meet Expectations** |
| --- | --- | --- |
| **Tasks #1–#2 (5% each)** | * Commands are syntactically correct. * Code output correctly answers the question. * Code is efficient; it uses built-in functions and libraries, when possible, rather than basic Python constructs (e.g., loops). | * Code has minor or major syntactical errors. * The functions used are only partially correct. * Code is inefficient, relying on basic Python constructs when built-in functions or libraries could have been used. * Code output does not directly or indirectly answer the question. * Results are unclear on how they address the tasks relevant to them. |
| **Tasks #3–#10 (10% each)** | * Commands are syntactically correct. * Code output correctly answers the question. * Code is efficient; it uses built-in functions and libraries, when possible, rather than basic Python constructs (e.g., loops). | * Code has minor or major syntactical errors. * The functions used are only partially correct. * Code is inefficient, relying on basic Python constructs when built-in functions or libraries could have been used. * Code output does not directly or indirectly answer the question. * Results are unclear on how they address the tasks relevant to them. |
| **Appropriate Markdown cells (5%)** | * Markdown cells are created to provide a high-level understanding of each analysis performed. * Markdown cells have no spelling errors and are written for the reader’s ease of understanding. * Markdown cells are well-formatted. | * Markdown cells are not used or are written poorly, and provide an ambiguous explanation of the analyses. * Markdown cells have spelling errors and/or the language is difficult to understand. * Markdown cells have not incorporated formatting elements to make reading easier. |
| **Appropriate code commenting and code style (5%)** | * Code comments cover all code written. * Code comments describe why code has been written. * Variable names are descriptive and unambiguous. * Code is structured for ease of reading. | * Code comments are either not used or only reiterate what the code is doing. * Variable names are vague. * Code is poorly structured, thus affecting readability. |