INFO 212: Data Science Programming I

Data Science Assignment 1: Movie Data Analysis

Objective: Analyze a dataset of movies to uncover insights related to genres, ratings, and revenue trends.

Dataset: Use the "The Movies Dataset" available on Kaggle (link below). This dataset contains metadata on over 45,000 movies released on or before July 2017. It includes information such as genres, ratings, revenue, and more.

1. Data Cleaning (10 marks):

- Load the dataset using pandas.
- Handle missing values and clean the data appropriately.
- Convert data types if necessary (e.g., converting strings to numerical values).

2. Exploratory Data Analysis (EDA) (10 marks):

- Generate summary statistics for the numerical columns.
- Explore the distribution of movie ratings and revenues.
- Identify the top 10 highest-grossing movies.

3. Genre Analysis (20 marks):

- Create a new column that lists the genre of each movie.
- Determine the most common movie genre.
- Analyze the average rating by genre.

4. Temporal Trends (20 marks):

- Examine how the number of movies released has changed over time.
- Investigate trends in movie revenues over the years.

5. Basic Calculations (20 marks):

- Calculate the average revenue of movies for each year.
- Determine the average movie rating for each genre.

INFO 212: Data Science Programming I

6. Insights and Conclusion (20 marks):

- Summarize your key findings from the EDA and Genre Analysis.

- Discuss any interesting trends observed in the Temporal Trends section.

Dataset - https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Requirements

- Provide full and detailed answer to each question.
- Work in group of 4 individuals. Less teams members can be accepted on a case-by-case basis. The compositions of the teams should appear as a comment in the solutions sent to me and in the report.
- University Academic Honesty Rules and Procedures will be adhered to strictly.
- There are 6 topics. The total marks are 100.

- Please name your assignment file as INFO212-assign1-group-number.zip, with the following **Deliverables:**
    - A Jupyter notebook containing the cleaned dataset, analysis code, and a summary of findings.
    - A brief pdf report (2-3 pages) outlining the approach taken and key observations.
        - List of members of the team
        - A general explanation of the solution idea for each task
        - A summary of the obtained results
        - Main obstacles faced during the execution of the project and how the team circumvented them
        - A quick summary of the responsibilities of each team member

- Submit your assignment file via moodle until the due **Due Date: 15-May-2024**
    - There will be a 10% (absolute value) deduction for each day of lateness, to a maximum of 3 days; assignments will not be accepted beyond that point. Missing work will earn a zero grade.

- Images must be clear and legible. Assignments will be judged on the basis of code, visual appearance, grammatical correctness, and quality of writing, as well as their contents. Please make sure that the text of your assignments is well-structured, using paragraphs, full sentences, and other features of well-written presentation. Text font size should be at least 11 points.