Mid Term - NED Academy
PGD - Data Visualization - Batch 9

Data Visualization (IMDb Dataset)

Total Marks: 30
Duration: 24 Hours
Tools Allowed: Python (Pandas, Matplotlib, Seaborn or any tool you can use easily)

Link for Dataset:

IMDB Movies Dataset | Kaggle

Section 1: Conceptual Questions (5 Marks)

1. **Why is data visualization important in analyzing movie trends?** Give two key reasons. (2 Marks)
   Hint: Think about how visual representation helps in identifying patterns, trends, and outliers.
2. **When analyzing a dataset with numerical columns like IMDb ratings, runtime, and gross revenue, which type of plots would you use for:**
   - Understanding distribution?
   - Comparing categories?
   - Identifying correlations? (3 Marks)
     Hint: Consider plots like histograms, boxplots, scatter plots, and bar charts.

Section 2: Python Coding Tasks (25 Marks)

Task 1: Dataset Exploration (5 Marks)

- Load the IMDb dataset using Pandas.
- Display the number of rows and columns, and show the first five rows.
- Find the top 5 most frequently occurring movie genres.

Hint: The genre column contains multiple values per movie—consider splitting it to count individual genres.

Task 2: IMDb Ratings Distribution (5 Marks)

- Create a **histogram** of IMDb ratings.
- Describe whether the ratings are normally distributed.

Hint: Use a histogram with an appropriate number of bins and check if there's a peak or skew in the distribution.

Mid Term - NED Academy
PGD - Data Visualization - Batch 9

Task 3: Genre Popularity (5 Marks)

- Create a **bar plot** of the top 10 movie genres.

Hint: Some movies belong to multiple genres—be sure to account for that when counting occurrences.

Task 4: Yearly Trends in Movie Production (5 Marks)

- Create a **line plot** showing the number of movies released each year.

Hint: Ensure the **Released_Year** column is numeric before plotting the trend.

Task 5: Director Analysis (5 Marks)

- Find the **top 10 directors** with the most movies.
- Create a **bar plot** to visualize this.

Hint: Some directors may appear more frequently—count their occurrences and plot accordingly.

Additional Questions

Task 6: IMDb Rating vs. Number of Votes (5 Marks)

- Create a **scatter plot** to analyze the relationship between IMDb ratings and the number of votes a movie has received.

Hint: Are highly-rated movies always the most popular? Consider adding a trendline to observe patterns.

Task 7: Highest-Grossing Movies (5 Marks)

- Identify the **top 10 highest-grossing movies** and visualize them using a bar chart.

Hint: The **Gross** column may contain missing values—handle them before sorting and plotting.

Mid Term - NED Academy
PGD - Data Visualization - Batch 9

Bonus Task (Extra Credit)

Task 8: IMDb Rating vs. Runtime (5 Marks, Bonus)

- Create a **scatter plot** of IMDb ratings vs. runtime.
- Add a trendline to analyze if longer movies tend to have higher ratings.

Hint: The **Runtime** column may contain non-numeric characters—clean it before plotting.