**STATISTICS DIGITAL ASSIGNEMNT 2**

**IMDB MOVIE DATASET ANALYSIS.**

Data: The data was extracted by scraping the iMDB website using Python and the Beautiful Soup library. The data collected consists of information on both Movies and TV-Shows. The information was extracted from the page of “featured films / tv-series” from the iMDB website. It contains the following information for each show:

1. Name of the Movie
2. Movie Year: Denotes the release year of Movies. Left ‘-‘ for TV-Shows.
3. Type of Show: Movie / TV\_Show
4. Genre: A list specifying the genre these shows can be categorized into.

[‘Animation’, ‘Drama’, ‘Adventure’, ‘Fantasy’, ‘Horror’, ‘Mystery’, ‘Comedy’, ‘Sci-Fi’, ‘Crime’, ‘Action’, ‘Romance’, ‘Thriller’, ‘Biography’, ‘Documentary’]

1. Movie Score: As rated in the iMDB website.
2. Meta Score: It is the given Metascore. Available for movies only.
3. Duration: Denotes the total running time of the movie. Given ‘0’ for TV\_Shows.
4. Movie Color: Denotes if the show is Color or Black & White.
5. Movie Language: The original language of release of the movie.
6. Movie Worldwide Gross: Denote the worldwide gross in USD. Given ‘0’ for TV\_Shows.
7. Movie URL: Contains the link to the iMDB page containing the information on the respective show.
8. Total Votes: Denotes the total number of voters, voting for the show.

A total of 5036 rows of data has been extracted, containing information on both Movies and TV\_Shows.

Note: A few rows contain ‘NA’ values. This was done to avoid collecting incomplete data on un-released shows that were also listed on the iMDB website.

Probable Visualizations, Inferences and Predictions (Initial list):

1. Determining correlation between various attributes. Plotting scatter matrix.
2. Visualizing number of movies with respect to year, genre, language.
3. Visualization of Movie iMDb score.
4. Analysis of performance between ‘Black & White’ and ‘Color’ movies: No. of movies and Average Rating(both iMDB and Metascore)
5. Relationship between iMDB Scores and Genre, iMDB Scores and Worldwide Gross, iMDB Scores and Metascore, Metascore and Worldwide Gross, iMDB Score and Total Voters, iMDB Score and Duration, Metascore and Duration.
6. Plotting top 50 movie names with highest worldwide gross, iMDB Score, Metascore, Total Voters, Duration
7. Comparing performance
8. Developing a regression model for score prediction based on the available iMDB Score, Meta Score, totalVoters and Worldwide Gross.