INFO 250 Project 2A

Jerry Li - jl4533@drexel.edu

Lixiao Yang - ly364@drexel.edu

Mengyang Xu - mx73@drexel.edu

Tasks:

- Identify a dataset for your final project.
    a. We are using IMDB Top 250 Movies Dataset for our project 2.
    b. **https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset**

- Describe the dataset:
    a. (1) what it is about:
        - IMDB (Internet Movie Database) is one of the largest online databases for movies and television shows, providing comprehensive information about movies, including ratings and reviews from its vast user base. The IMDB ratings are widely used as a benchmark for the popularity and success of movies.
        - This dataset contains the top 250 rated movies on IMDB as of 2021, providing a snapshot of the most popular and highly rated movies of recent times. By analyzing this dataset, one can gain insights into the movie industry, such as trends in movie ratings and popular genres.
    b. (2) Other information about the dataset a designer should know (such as its history and conditions under which the data is collected)
        - The data for this dataset was scraped from the IMDB website. The top 250 movies were selected based on their IMDB ratings, and information such as movie title, director, cast, rating, votes, and year of release was collected.
    c. (3) Statistical descriptions of the variables that you are going to use.
        - rank - Rank of the movie (Ordinal)
        - name - Name of the movie (String)
        - year - Release year (Time)
        - rating - Rating of the movie (Ordinal)
        - genre - Genre of the movie (Nominal)

        - certificate - Certificate of the movie (Nominal)
        - run_time - Total movie run time (Time)
        - tagline - Tagline of the movie (String)
        - budget - Budget of the movie (Ordinal)
        - box_office - Total box office collection across the world (Ordinal)
        - casts - All casts of the movie (String)
        - directors - Director of the movie (String)
        - writers - Writer of the movie (String)

- Identify the target audience for your work (at least one key audience group) and discuss what they may know about your topic and what aspect of your data may be interested by them.
    a. Movie makers and producers
        - Movie producers can know from our dataset that this dataset does not just evaluate the quality of a movie from the perspective of the audience. It also includes data that producers and manufacturers are more concerned about, such as the theme of the movie, the production cost of the movie, and the box office of the movie, which allows them to better analyze the direction and investment of their own movie production. For example, for budget and box_office in this data set, movie manufacturers can count the production costs and box office of movies of various themes to determine their own production direction and effect evaluation.
    - Moviegoers
        - Moviegoers can know from the topic of our dataset that the movies in our dataset are all top-ranked movies in IMDB. Therefore, movie viewers should know that all the movies in our dataset are either with good reputation or high quality. Movie viewers may be most concerned about the rating and subject matter data in the data set. For example, viewers can find some movies with higher ratings among the subjects they are interested in from these movies.

- Identify at least two questions that the audience might be curious about your data/visualization.
    a. Is there a correlation between the budget and the rating/box office collection of a movie?
    b. How film budgets/length have trended in the last 100 years ?
    c. Which genre has the highest score on average?
    d. Do longer movies have better box office collections?
    e. What is the distribution of certificates for different movie genres?
    f. Who directed most movies in IMDB top movies?

- Design (but not implement) how your visualization will look like. (**Include the following subquestions**.)
    - What types of visualization are likely to be feasible?
        - Line chart, scatter plot
        - Line chart, scatter plot
        - Box plot
        - Line chart, scatter plot
        - Treemap, bar chart
        - Bar chart
    - What tools are you planning to use?
        - We are planning to use tableau for our project.
    - What data variables/observations would be necessary to answer your questions?
        - year, box_office, budget, genre, certificate, rating, run_time
    - What are the major mappings from data variables to visual representations?
        - year → coordinates
        - genre → color / label
        - budget → size / coordinates
        - box_office → coordinates
        - certificate → color
        - director → label