**INFO 6151 Data Visualization**

**Project Requirements:**

**Data validation and Preprocessing:** Clean andpreprocess the data. This step includes handling any missing values, outliers and ensuring that the data is ready for analysis.

**Chart 1: Distribution of TV Shows and Movies by Type**

* Use ML algorithm to analyse the distribution of TV shows and movies.
* Visualize the proportion using a pie chart, where each slice represents the percentage of TV shows and movies.

**Chart 2: Relationship between Release Year and Duration**

* Perform predictive analysis to understand the relationship between release year and duration of TV shows or movies using machine learning regression techniques.
* Develop a predictive model to estimate the duration based on the release year.
* Create a scatter plot visualization to depict the relationship between release year and duration.

**Chart 3: Distribution of Content by Country**

* Aggregate TV shows and movies by country using ML technique.
* Visualize the distribution using a bar chart, where each bar represents a country and the height represents the number of TV shows or movies produced.

**Chart 4: Distribution of Content by Rating**

* Analyse the distribution of TV shows and movies by rating using ML clustering techniques.
* Visualize the distribution using a bar chart, where each bar represents a rating category (e.g., G, PG, PG-13, R, etc.) and the height represents the number of TV shows or movies in each category.

**Chart 5: Release Year Trends in Content Addition (Line Chart)**

* Track the trends in content addition over the years using machine learning time series analysis techniques.
* Use a line chart to display the number of TV shows and movies added annually.

**Chart 6: Content Duration by Type (Box Plot)**

* Analyse the distribution of content duration (in minutes) by type (TV show or movie)
* Visualize the distribution using a box plot, where each box represents the distribution of duration for TV shows and movies separately.

**Chart 7: Distribution of Content by Genre**

* Aggregate TV shows and movies by genre
* Visualize the distribution using a stacked bar chart, where each bar represents a genre and the stacked segments represent the proportion of TV shows and movies within each genre.

**Chart 8: Content Addition by Date**

* Analyse the pattern of content addition by date (month and year) using ML technique
* Visualize the pattern using a heatmap, where each cell represents the number of TV shows and movies added on a specific date.

**Note:** *Please use ML models wherever necessary even if not specified – apply you critical thinking and analytical skills*

**TECHNICAL EXPECTATIONS:**

* Use appropriate tools and libraries when necessary
* Data Analysis – Pandas, NumPy, etc.
* Visualization – Matplotlib, Seaborn, Plotly
* Interactive Dashboard – Dash/Streamlit

**CAPSTONE PRESENTATION:**

* Include the preprocessing methods employed in your project
* Provide the insights and interpretations of the visualizations
* Explain in detail the analysis and the visualization techniques used and why?
* Provide explanation for the choice of the ML algorithm

**Meeting notes:**

Meetings on Mondays 11am -12pm

We will use 1 google colab document.

Make a copy and make sure to eliminate most errors before updating main file.

Make sure to comment what code should be changed and date the code.

Each person just does the next chart the last person left off at.

The next person to work on the project reviews the last persons work and does the next chart.