**Project II**

**Group 3**

**Members:**

Avick Haque (501037491)

Wenling Jiang (501003021)

Jyles Macdonell (500812711)

Tavish Ramasawmy (501137316)

Janahan Thanabalasinkam (500975108)

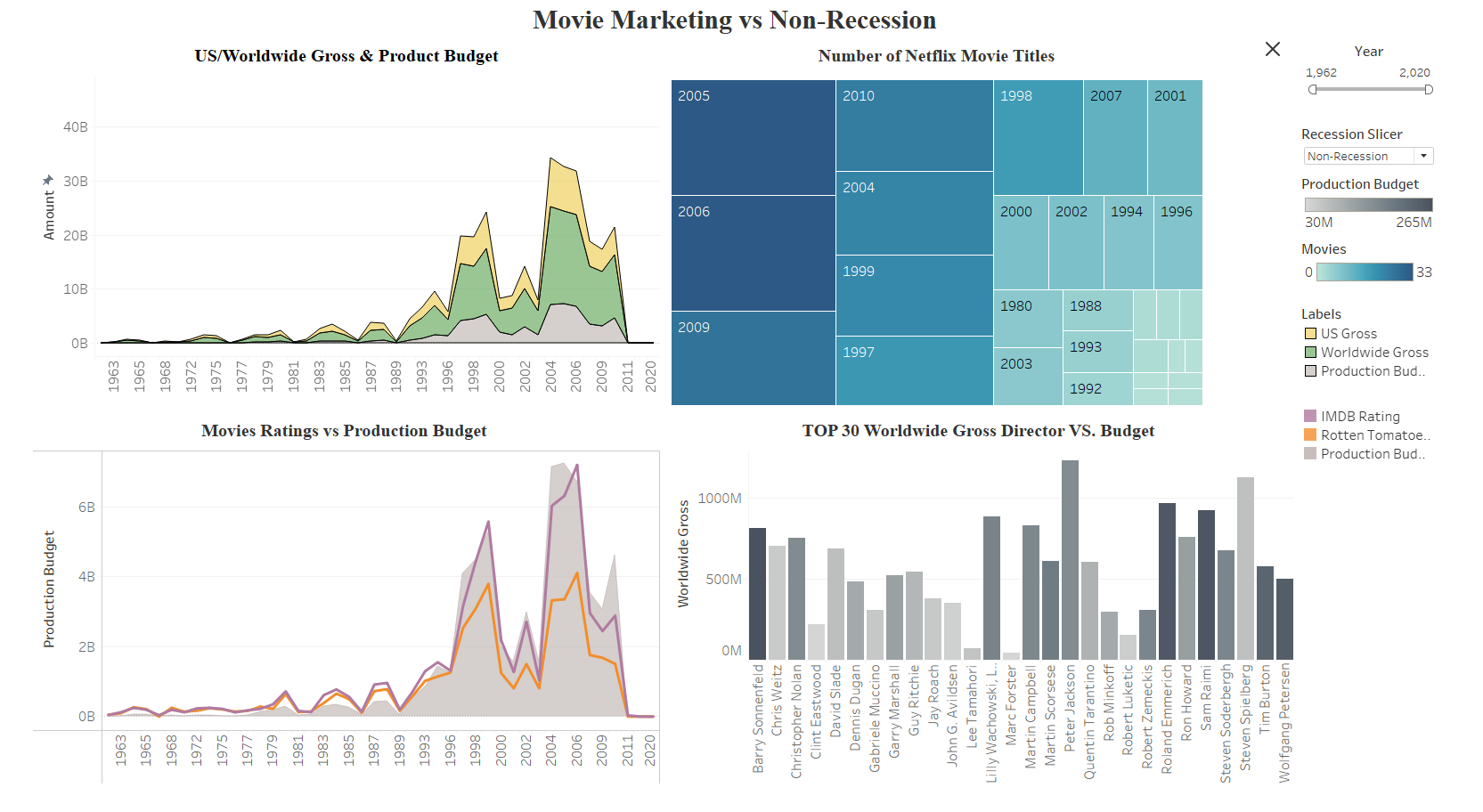
Thinesh Thevarajan (500962342)

Professor: Bachir Chehab

ITM 752 - Data Visualization

December 4th, 2023

**Movie Marketing vs. Recession/Non-Recession**



**Brief explanation of your visualization and what it shows.**

Our interactive dashboard illustrates the correlation between movie marketing and economic conditions (recession or non-recession) spanning the years from the 1960s to the 2020s. Users can easily filter between recession and non-recession periods and select specific years on the right side of the dashboard. Upon choosing the economic conditions, the first area graph displays the trends in US Gross, Worldwide Gross, and Product Budget over the specified time frame. A heatmap visualizes the total number of movies in the Netflix record for each year. Another area graph explores the relationship between movie ratings and production budget. The final bar chart is dedicated to the Top 30 Worldwide Gross Directors, revealing the connections between worldwide gross and production budget.

**Briefly explain the data transformation and cleaning process.**

In the Data transformation and cleaning process, the way data was handled was we identified the columns that we needed were necessary in order to convey our story that we were going to present within the dashboard. Therefore, we deleted the unnecessary columns within the dataset such as “distributor”, “source” and “creative type” as they were columns that we did not need to use within the dataset. In regard to the missing data within the dataset, such as the “add only year” and “add recession data'' column, first, we used the =Year() function in Excel to extract only the year from the “Release Year” column. Next, in order to add the recession data, we first formatted the “release year” column to follow the recession date column format, which was YYYY-MM-DD. Then, a =VLOOKUP function was used to find the corresponding date with the recession data file to get the recession data. Next, we cleaned up the data by removing the rows that gave “#n/a” in the recession data column. The reason why we did this is that these were the rows in the dataset that were not included in the form of the recession data, which were the years before 1960. Moving on, we also removed inconsistent rows in the dataset that did not apply with the data. For example, for the “US Gross” and “Worldwide Gross” columns, data that was not 0 or a string such as “Unknown” were removed from the dataset to make the data consistent. In addition, in the “release year” column, we removed any release years that were past 2022 as the movie was not released yet. Finally, as we were using the movie ratings column within the dataset, for any missing columns that had null values, we transformed them to be 0. The reason why we did this is that if we were to delete those columns, then a considerable amount of data would be deleted within the dataset.

**Brief description of the Tableau functionalities utilized and how/why you use them.**

The graph and dashboard incorporate various features, encompassing filters, parameters, and the creation of calculated fields. Filters are an essential part of analyzing data, and it has been used for years and recession. They serve the purpose of excluding irrelevant elements, while parameters and calculated fields are employed to establish a slicer, enabling users to choose between Recession and Non-Recession on the dashboard. LOD expressions have also been used to create the graphs. This feature has greatly helped to create flexible, granular calculations with ease. In the graph creation process, Measure Values are utilized to depict a dynamic grouping of multiple attributes. This functionality proves beneficial when dealing with multiple numerical elements in the dataset. For instance, US Gross, Worldwide Gross, and production budget are collectively "grouped" using Measure Values for trend analysis.

**Describe your data exploration process and what your final visualization shows. Describe the steps you have gone through, including the flow of your exploration, questions posed, and challenges.**

In our data exploration process, we posed initial questions regarding how the US Gross, Worldwide Gross and Production budget were affected by recession or not. In our initial graphs that we created for each variable, we discovered that US Gross, Worldwide Gross and Product budget all had similar trends in regard to recession versus non-recession. Consequently, this made us come to the conclusion that it would be best to include all three variables into 1 graph, as the trends regarding these variables within the dataset were all similar during recession and non-recession. Moving forward, it was evident that the 2008 recession had the greatest impact in regard to all the datasets. For example, US Gross, Worldwide Gross, Production budget, movie ratings and Netflix titles all had the greatest impact in 2008 from the graphs that we created. Adding on, we identified those who were outliers in our data exploration process for the movie ratings dataset. For example, we identified that year 2025 was included in the dataset regarding a movie. Consequently, this was misleading data as the year 2025 has not happened yet and the data regarding this movie was removed.

Lastly, challenges we faced during the data exploration process were regarding the correlation of the datasets and how we can connect them all together. In this situation, we had to connect 3 datasets. Netflix titles, Movie ratings and Recession data. Due to the volume of the dataset, it was extremely difficult to find out how to connect it, what data to remove and how to identify anomalies/outliers. Through Excel functions and Tableau features, we were able to connect the data into one file which worked coherently. While working on connecting the data, we were able to identify the approach we wanted to take towards creating a visualization. We found that the recession data was the most important in connecting the 3 datasets, so we used that as a basis during our visualization process. This helped us further examine ways we can connect Netflix data to recession.