# Data Preparation and Tableau Implementation

## Data Cleaning and Preparation

The attributes highlighted in bold are those that have been added following data cleaning and preparation for the visualisations. Date, Month, and Year have been extracted from the 'released' attribute to produce the 'Date of Release', showing only the Dates, and 'Net Profit' is the result of Gross minus Budget variable (see file *"movies_modified.xlsx"* in the appendix). The variables highlighted in both bold and red are the attributes used in creating the data visualisation. This dataset initially consisted of 7,688 observations, however, after data cleaning and subsetting the dataset to extract the required information, the dataset now consists of 1,417 observations.

| Name | Description | Data type |
|---|---|---|
| Name | Name of the movie | Categorical (nominal) |
| Rating | Ratings of the movie | Categorical (nominal) |
| **Genre** | **Main genre of the movie** | **Categorical (nominal)** |
| Year | Year of Release | Integer (continuous) |
| Released | release date (YYYY-MM-DD, country)) | Integer (continuous) |
| **Scores** | **IMDb user rating** | **Categorical (continuous)** |
| Votes | number of user votes | Integer (continuous) |
| Director | the director | Categorical (nominal) |
| Writer | writer of the movie | Categorical (nominal) |
| Star | Main actor/actress of the movie | Categorical (nominal) |
| **Country** | **Country of origin** | **Categorical (nominal)** |
| Budget | the budget of a movie. | Integer (continuous) |
| Gross | revenue of the movie | Integer (continuous) |
| Company | the production company | Categorical (nominal) |
| **Company Numerical** | **The numerical of the production company (just the top 5)** | **Integer (continuous)** |
| Runtime | duration of the movie | Integer (continuous) |
| **Date Of Release** | **Release of Date, Month, Year** | **Integer (continuous)** |
| **Country Of Release** | **Country of release** | **Categorical (nominal)** |
| **Net Profit** | **Gross-Budget** | **Integer (continuous)** |

## Implementation

The first trend line plot was produced by first creating a calculated field called 'Net Profit,' where this attribute is a calculation of the gross income less the budget, as shown below.


***Figure 1:*** *Producing a new variable called 'Net Profit' within Tableau.*

As Tableau registered the 'Date of Release' attribute as 'Date' data type, the month part was easily moved into the column section. The net profit was then moved to the rows section. The years from the 'Date of Release' was moved to the colour 'Marks' section, for the three groups of years to be colour encoded. The filter for the top 5 companies is shown and is applied to the box plot and the TreeMap, to allow interactive brushing and linking (see *Figure 2*).


***Figure 2:*** *The worksheet of the line plot, showing where each attribute was moved to accordingly. The "Marks" section allows for the encoding of the objects such as colour and size.*

The box plot was produced by moving the 'Company' attribute to the column section, followed by moving the 'Net Profit' into the row section. On the y-axis, showing the 'Net Profit', a box plot reference was added, ensuring that "Hide underlying marks (expect outliers)" is clicked and the colour was changed accordingly. This transformed the plot that was initially a

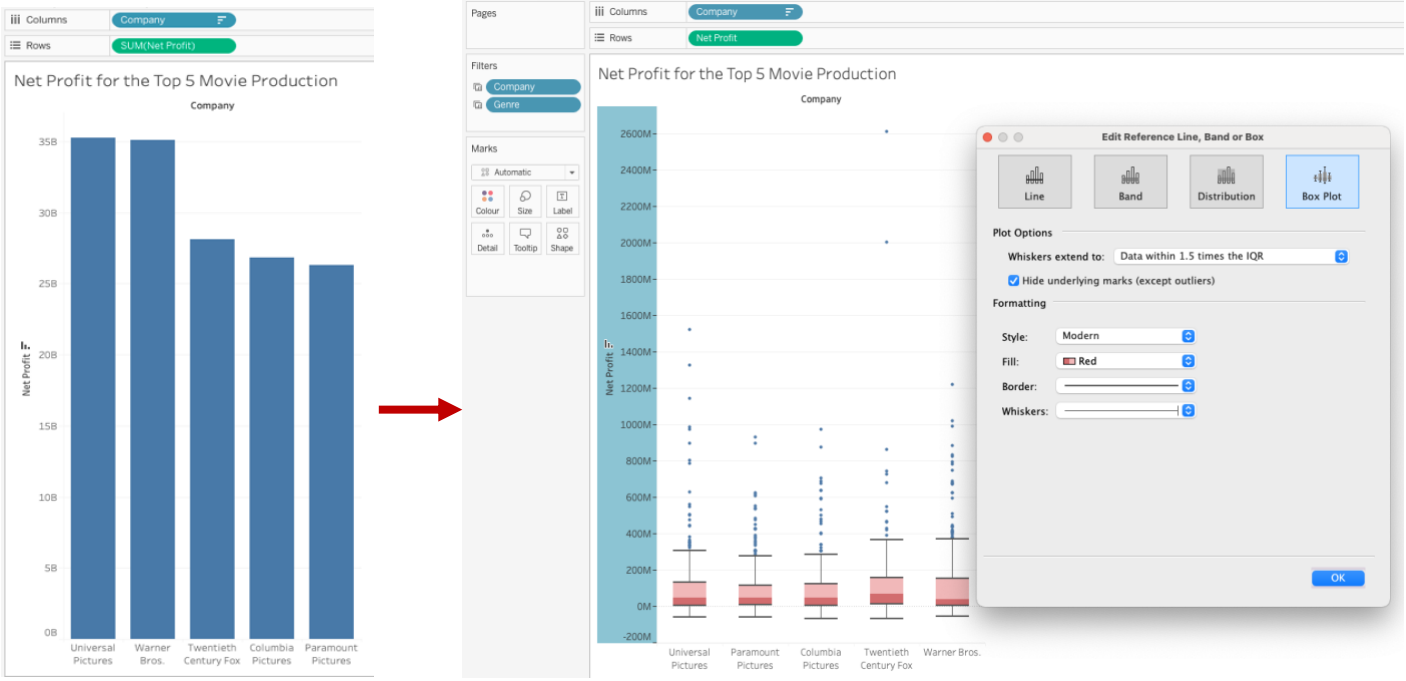


***Figure 3:*** *An illustration of producing and transforming the box plot on Tableau.*

bar chart to a box plot. However, the values were aggregated to the median line, therefore, the measurement was un-aggregated, and the size of the data point was adjusted using the size on the "Marks" (see *Figure 3*).

The TreeMap was created by moving the 'Company Numerical' attribute to the size and the 'Genre' attribute to the colour within the "Marks" section, respectively. The 'Genre' was applied as a filter section for brushing and linking for the TreeMap and the grouped bar plot. The 'Company' and sum of movie 'Score' was added to label within "Marks" (see *Figure 4*).

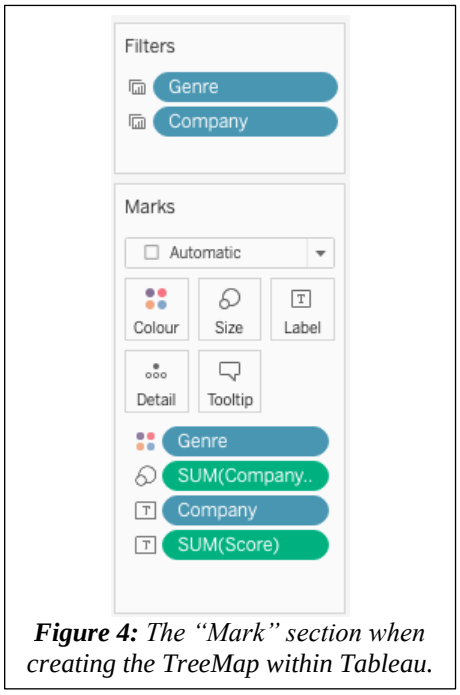


***Figure 4:*** *The "Mark" section when creating the TreeMap within Tableau.*

The final plot was created by adding 'Country' to columns, filtering only the UK, US and Europe data, and 'Scores' to rows, and 'Genre' was colour encoded, in line with the TreeMap colours. This plot was transformed into a grouped bar plot with the aid of the "Show me" section. 'Genre' and 'Country' were arranged in descending order. Logarithm was applied to the y-axis to improve the scales and enable better viewing of the values as shown in *Figure 5*.

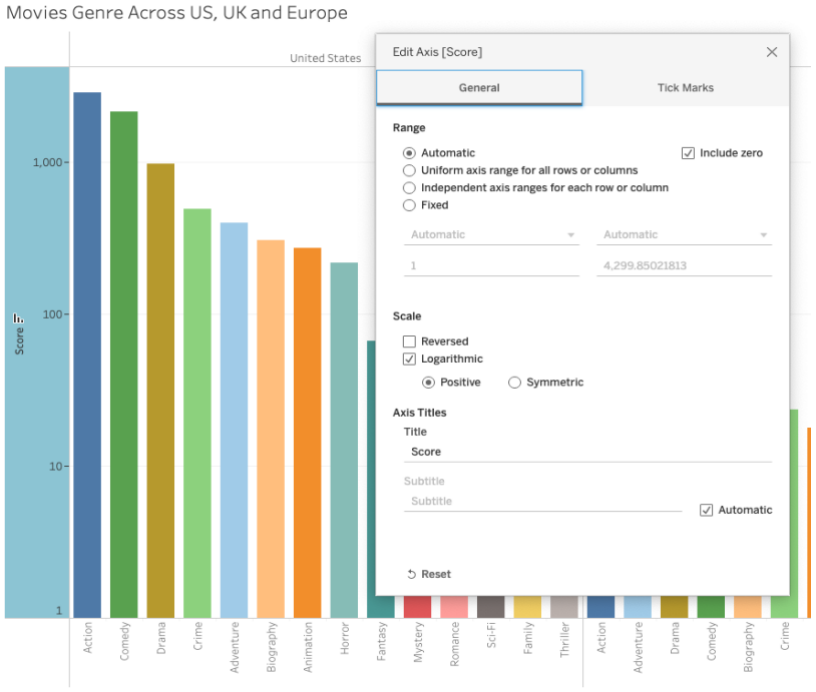


***Figure 5:*** *An abstract of the group bar plot showing how the logarithmic scale was applied on the y-axis.*