A screenshot of a chart

Description automatically generated

**HEATMAP using CORRELATION MATRIX**

**Purpose**:

The purpose of using a heatmap correlation matrix is to identify relationships, patterns between several variables in our dataset. The colour gradients make it easy to spot the areas of high correlation. This visualization helps us to identify which variables are most significant for our analysis. The values range from -1 to 1; 1 being the strongest positive correlation and -1 being the strongest negative correlation and 0 being no correlation between the variables at all.

**Justification**:

**Position & Length**: Data points are arranged in a grid, and it allows for a clear, structured comparison between different categories. Unlike other plots, the length in a heatmap refers to the size of the grid cells and they are all uniform. This makes it consistent for an unbiased visual.

**Colour**: The use of colour gradient represents the intensity of a value. The changes in the colour gradient make it easier to spot the stronger correlations, lighter colours indicate a stronger correlation between the variables and darker colour indicate a weaker correlation.

**Size:** As explained in ‘Length’, the size for each grid is fixed, making each cell has the same dimensions. The absence of varying sizes makes it easier to focus purely on colour differences, making it clean and simple.

**Design trade-offs**: The limitation of this heatmap matrix is that it could only be used for numerical values therefore could lead to an incomplete understanding of the data. If there are datatypes other than float or integers, they could not be included, or these must be converted into numerical data if possible and then could be included in this plot. The other limitation is that correlations do not consider how data is distributed, so if there are outliers, the accuracy of the correlation will be affected.

**Interpretation**:

When two variables have a strong positive correlation, an increase in one variable is accompanied by an increase in the other, and a decrease in one lead to a decrease in the other. This means the relationship between the variables always moves in the same direction. If we set a minimum of %60 threshold between variables, the following variables have significant positive correlations.

1. US\_Gross & Worldwide\_Gross – 0.94 = As US Gross sale increases, naturally the Worldwide Gross sale increases too. They have a very strong positive correlation.
2. US\_Gross & US\_DVD\_Sales – 0.74 = Movies with higher gross sales tend to have a higher US DVD sale. (There are many missing values in US\_DVD\_Sales so it’s important to approach this analysis with caution.)
3. US\_Gross & Production\_Budget – 0.62 = As the Production Budget of the movie increases, US Gross sales tend to increase as well.
4. Worldwide\_Gross & US\_DVD\_Sales – 0.70 = Movies with higher worldwide gross sales tend to have a higher US DVD sale. There are many missing values in US\_DVD\_Sales so it’s important to approach this analysis with caution.)
5. Worldwide\_Gross & Production\_Budget – 0.67 = Movies with higher production budgets typically achieve higher Worldwide Gross sales.
6. IMDB\_Rating & Rotten\_Tomatoes\_Rating – 0.74 = There is a strong positive relationship, movies with higher IMDB Rating tend to have a higher Rotten Tomatoes Rating. (There are many missing values in Rotten\_Tomatoes\_Rating so it’s important to approach this analysis with caution.)

When two variables have a strong negative correlation, an increase in one variable corresponds to a decrease in the other, and a decrease in one lead to an increase in the other. This means the relationship between the variables always moves in opposite directions. There are no strong negative correlations on this matrix.