A screenshot of a chart

Description automatically generated

**HEATMAP using CORRELATION MATRIX**

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.

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
2. US\_Gross & US\_DVD\_Sales – 0.74
3. US\_Gross & Production\_Budget – 0.62
4. Worldwide\_Gross & US\_DVD\_Sales – 0.70
5. Worldwide\_Gross & Production\_Budget – 0.67
6. IMDB\_Rating & Rotten\_Tomatoes\_Rating – 0.74

What this information tells me that I need to dive into these variables for more analysis.

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.

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.