1. **Calculating statistics for rankings:**

Based on rankings for actors and directors in each movie descriptive statistics is calculated as a new column . Ex : For “Normalised\_Actor\_Rank “ , statistics like mean, min , maximum, mode,standard deviation

1. **Filtering variables by observations:**

Few variables are removed just by visualising at the column , such as standard deviation , mode for each ranking as they cannot be useful to build a model.

1. **Correlation Matrix**:

Correlation matrix explained correlation between independent variables and variables having high correlation with others removed.

1. **Scatter Plots:**

Scatter plots used to check is there any particular pattern between dependent and independent variables.

1. **Useful variables:**

Variables with “max” value for each ranking are kept to predict hit/flop. Maximum rank gives ideas about actors or directors having famous or critically high value in movies.

# **Adding new variables derived from existing variables:**

Variables added which are derived from old variables to get useful information in building model.

**Final dataset :**

Finally a dataframe with only useful 18 variables is kept .

Dependent variable hit\flop movies are divided into “0” and “1”.