Python Correlation 🡪 Movies Dataset with about 6000 rows to analyze.

Importing necessary libraries and read the csv file

A screenshot of a computer

Description automatically generated

Cleaning the data

1.Check for null values 🡪 drop or fill based on different cases

A screenshot of a computer

Description automatically generated

2. Check the datatypes of each column change it depending on user needs

A screenshot of a computer

Description automatically generated

3. check for data mismatch

A screenshot of a computer

Description automatically generated

Example here the year is different

Going by the released column having the right year

A screenshot of a computer

Description automatically generated

Extract the year from the released column

company, genre, budget , star, rating and runtime according to me affect the gross of the movie

Checking if BUDGET VS GROSS are correlated

A screenshot of a computer

Description automatically generated

From the graph it seems as though they are correlated the extent of correlation is yet to be determined.

A graph with red dots and a blue line

Description automatically generated

A screenshot of a computer

Description automatically generated

Budget seems to be highly correlated with gross.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Now that the dataset is in a numerical format, it becomes suitable for performing correlation

A screenshot of a computer

Description automatically generated

Budget and votes are the highest correlated columns of all columns

A screenshot of a computer

Description automatically generated

Correlation of each column with the rest of the columns in a linear fashion

A screenshot of a computer

Description automatically generated

Using that to find the pairs with the highest correlation.

Analysis : I believed company, genre, budget , star, rating and runtime would be the columns that would have a high correlation with gross.

Although budget was votes is the second highest correlation.