Data Investigation

Cleaning data:

1 - I have dropped the following columns (homepage, tagline, overview, id, imdb_id)

2- I have filled nuke values with “UNKNOWN” value.

3- I changed the datatype of release_year to string

4- rewrote the year in release date then changed the data type to date time.

5- I have done 3 copies of the data frame to separate the actors of each movie, genres and the production companies.

6-Did a new column of the final revenue and budget that I should calculate on.

Extracted data:

1- The oldest 2 movies were produced in 1960 were : The Unforgiven
and The Brides of Dracula.

2- The newest two movies produced in 2015 were: 1- Open Season: Scared Silly
2- Martyrs.

3- The most popular movie is 'Jurassic World’

4- The least popular movie is 'North and South, Book I’

5- The highest vote average movie is "The Story of Film: An Odyssey”

6- In this data the most actress acted in movies is "Robert De Niro" who acted in 72 movie

7- In this data the most production company produced movies is "Universal Pictures" who produced 522 movie

8- Drama is the most produced genre

Sources:

1- Stack over flow
2- Pandas documentation
3- geek for geeks