analyzing and visualizing data

University of the Cumberlands - SPRING 2018

Final Research Paper

**Team:**

1. **Archi (SID?)**
2. **Nagendar (SID?)**
3. **Naveen (002836193)**
4. **Nrupen (SID?)**
5. **Ravi (SID?)**

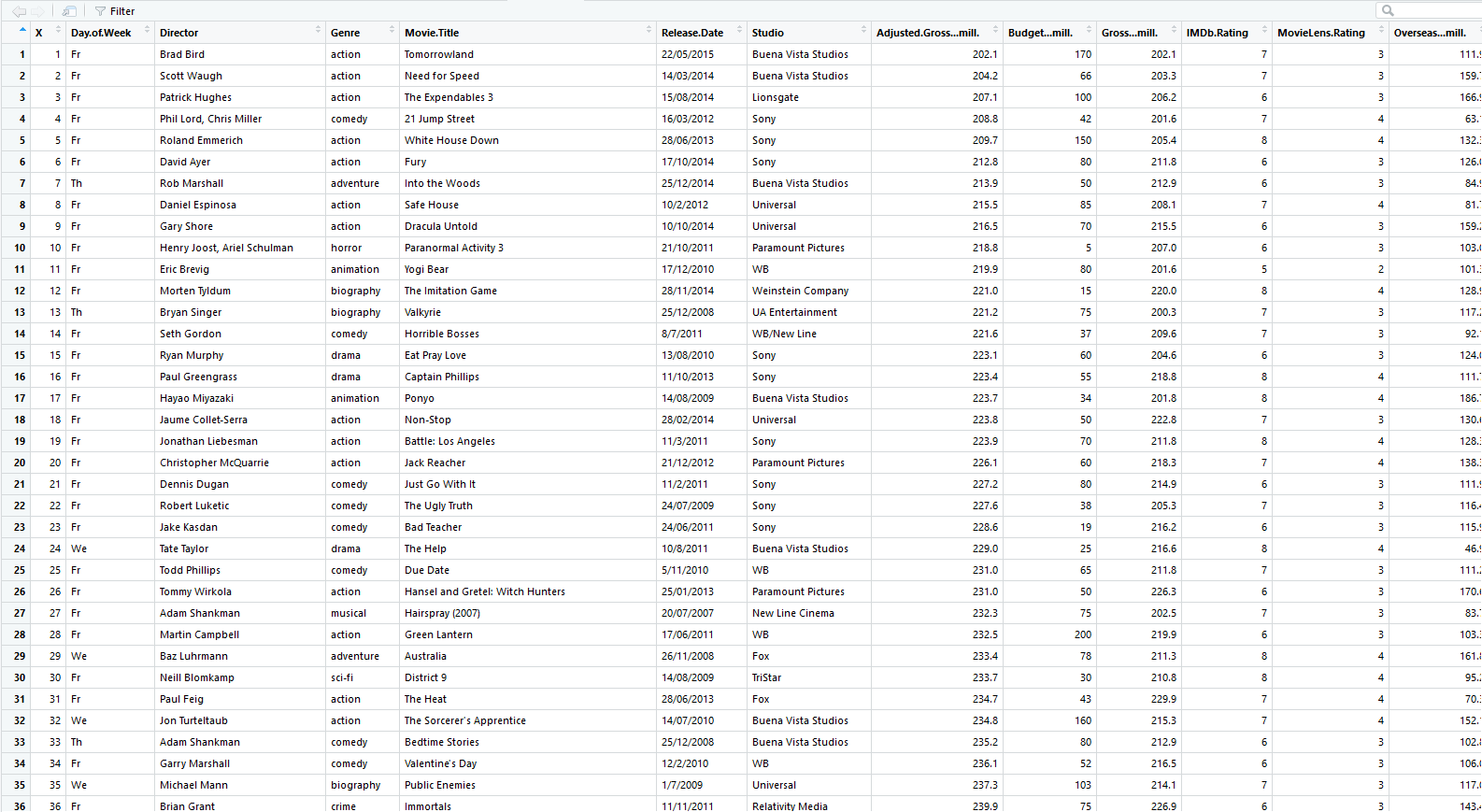
**Class #: ITS530 – Analyzing and Visualizing Data – 30**

**Professor: Dexter Francis**

**Date: 04/15/2018**

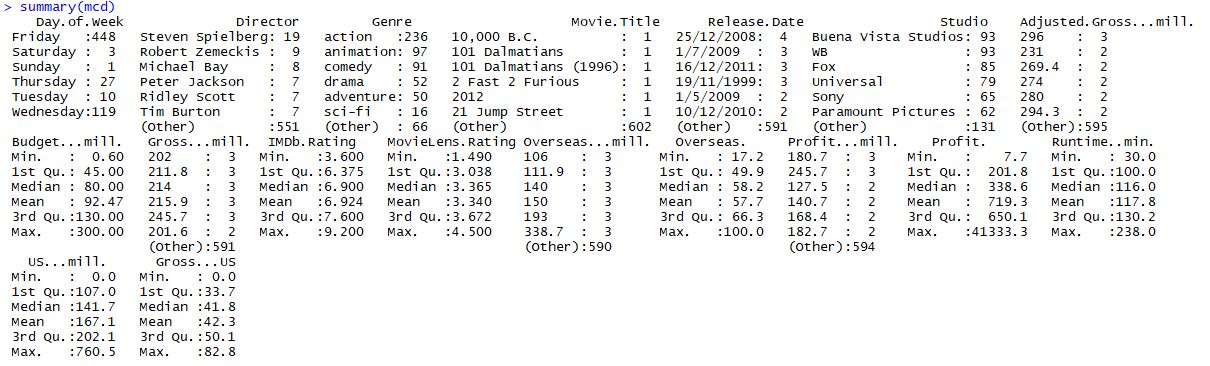
Section 2: Data Analysis (R Studio - R language - Library: mosaic)

* Describe the fields of the dataset.



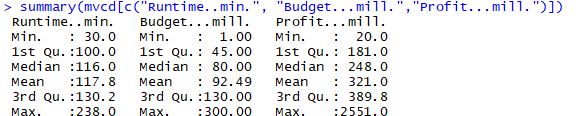
The dataset utilized in our analysis is Movies collection based on the studios and their release dates. It has many fields, where most of them talk about the numbers on the budget, gross collection, overseas collection and profit percentage. It also describes the data fields like IMDB rating and day in the week of movie released. So, based on these data fields we’ve considered the days of the week, genre and IMDB rating as categorical variables and the movie gross collection, profit percentage, runtime, overseas collection etc., as continuous variables.

* Create a summary of stats for the dataset.

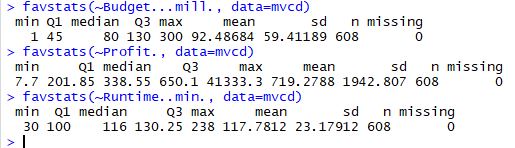


Given, is the summarized data of the movie collection data set we’ve considered.

* Discuss the Min, Max, Median, and Mean of the continuous fields.

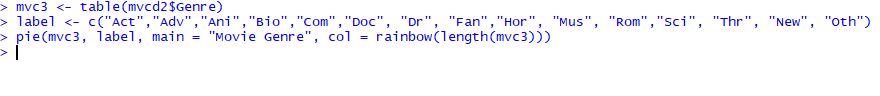
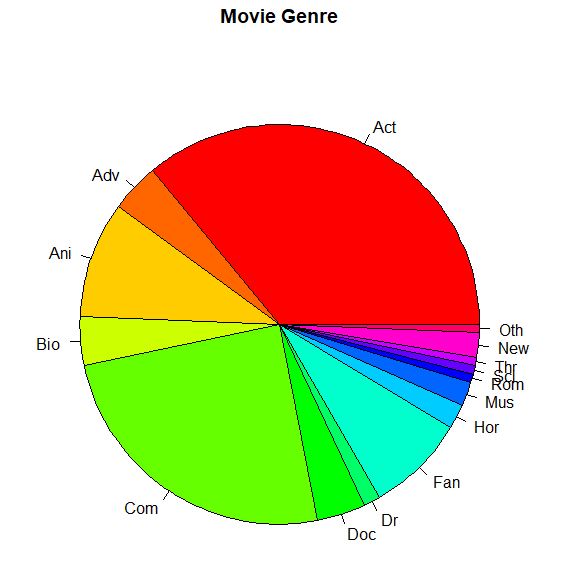


This data shows the min, median and max using the summary query for the continuous data fields considered.



This is another way to find the min, median and max using the favstats function in R studio

* Discuss the Counts and Percentages of the categorical fields. (ex. Pie Chart)



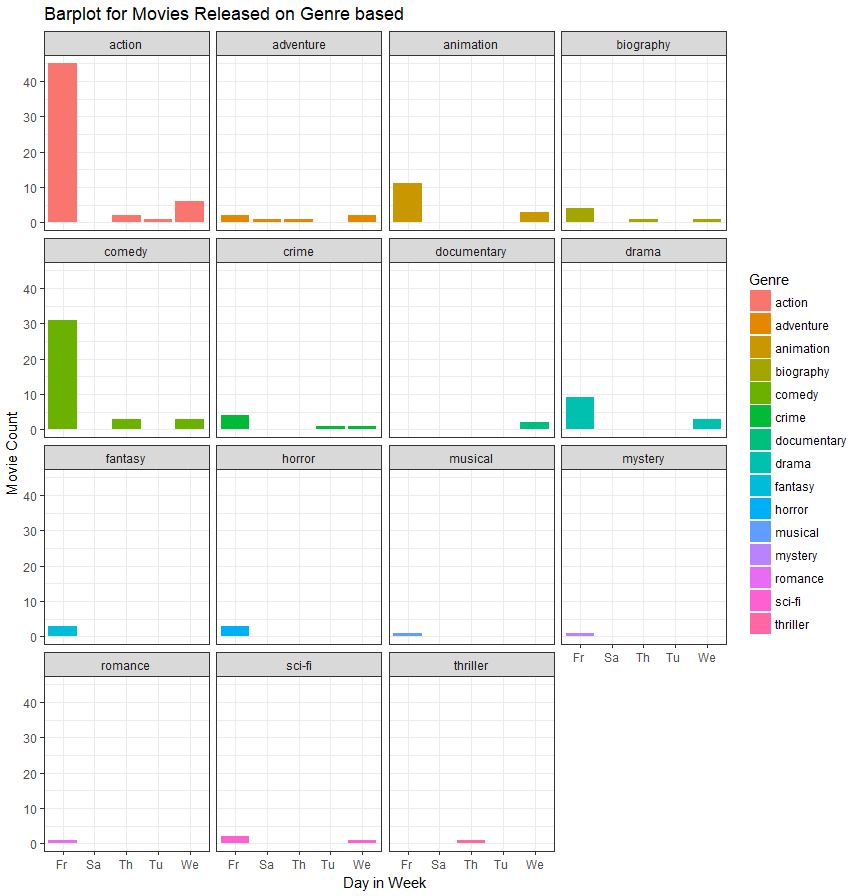
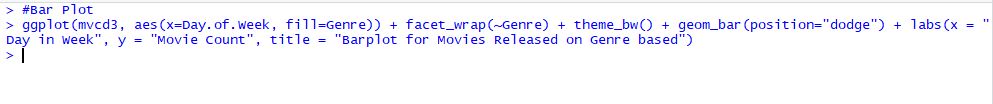
This shows the pie chart representation of the movie release day and its percentage, also the movie genre percentage. There is also given pie chart functions given.

* Discuss any missing data elements.

It would be nice if number of awards received for the movie, number of people watched in theater and actor information are included. I feel that the above mentioned are missing data elements, which would’ve helped to dig much deeper in the data analysis.

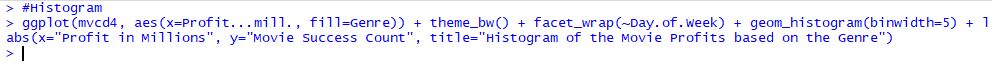
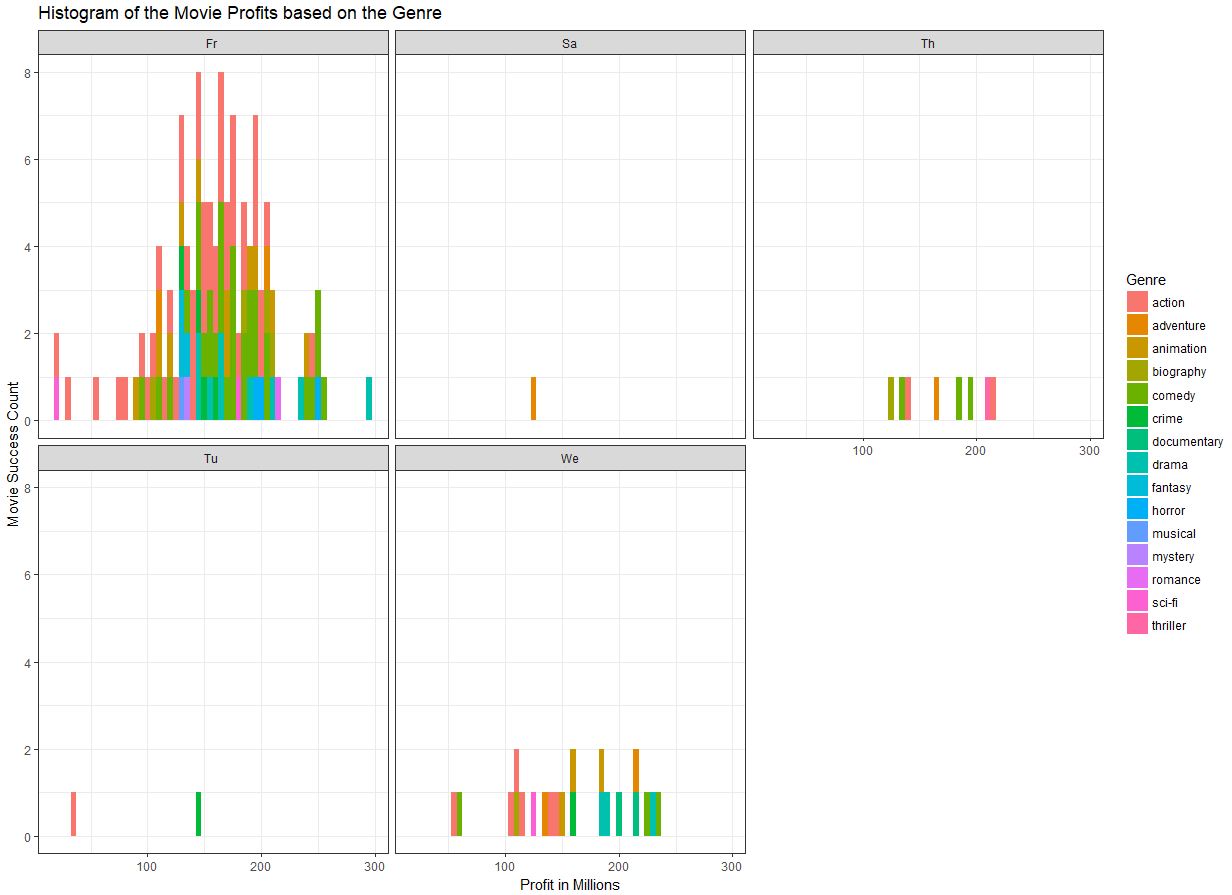
Section 3: Data Visualizations (R Studio - R language - Library: ggplot2)

* Bar plot:



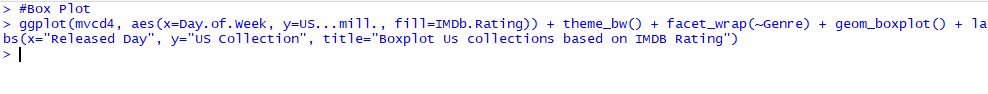
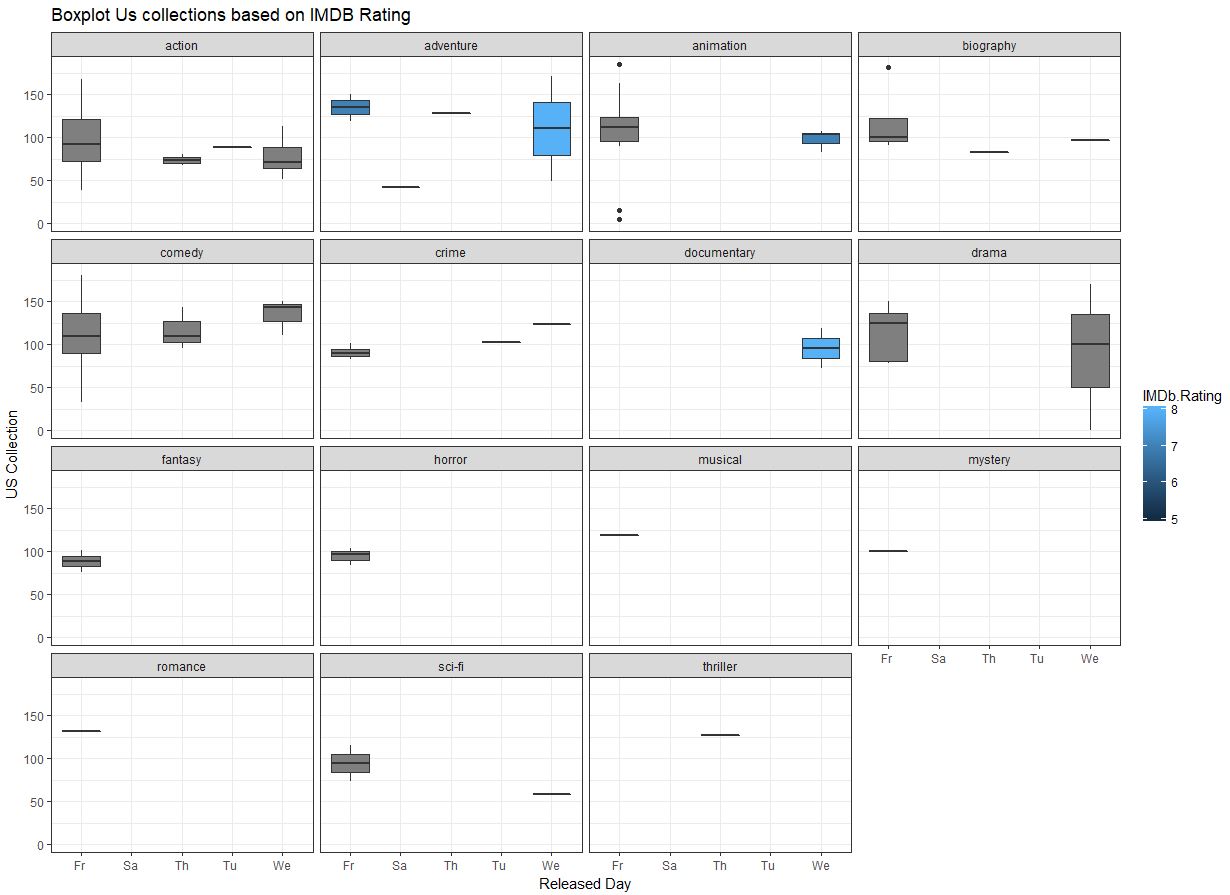
This above bar plot has been created using the ggplot function. Using the categorical data day of the week and genre it has been based on the genre of the movies. So, it gives an idea of how many movies have been released in a day of the week within the genre. Looking into this function there are large number of action movies got released on Friday which takes up by comedy to the next position.

* Histogram



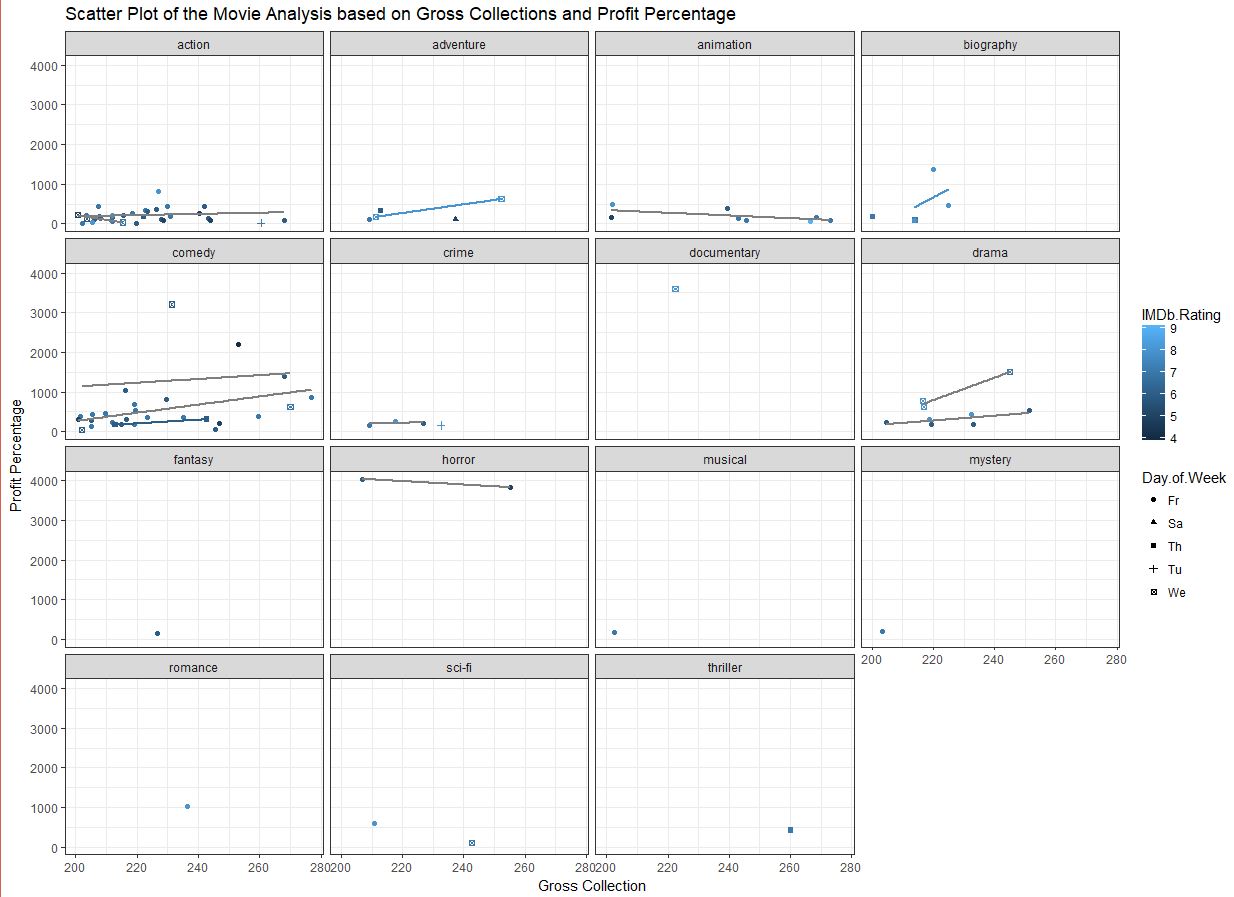
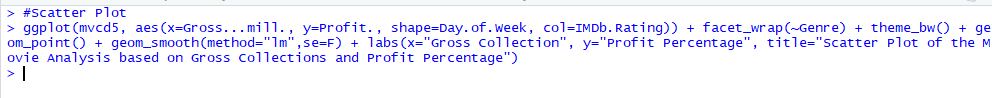
This histogram has been analyzed with the categorical genre data and continuous data profit of the movie. The facet wrap has been given by the day of the week another categorical data. So, this histogram clearly shows that there is very large number of movies released on Friday. When compared to remaining days in a week, also earned more profit in action genre.

* Box Plot



This box plot has been analyzed with the categorical day of the week data and continuous data US total collections of the movie and categorized IMDB rating. The facet wrap has been given by the genre another categorical data. So, this box plot clearly shows that there is very large number of movies released on Wednesday have got very good IMDB rating and in drama and adventure genre.

* Scatter plot



This scatter plot has been analyzed with the categorical day of the week data and continuous data gross collections of the movie and categorized IMDB rating and also the profit percentage. The facet wrap has been given by the genre another categorical data. So, this scatter plot clearly shows that there is very large number of movies in the horror genre has earned very IMDB rating and good profit percentage and so on.