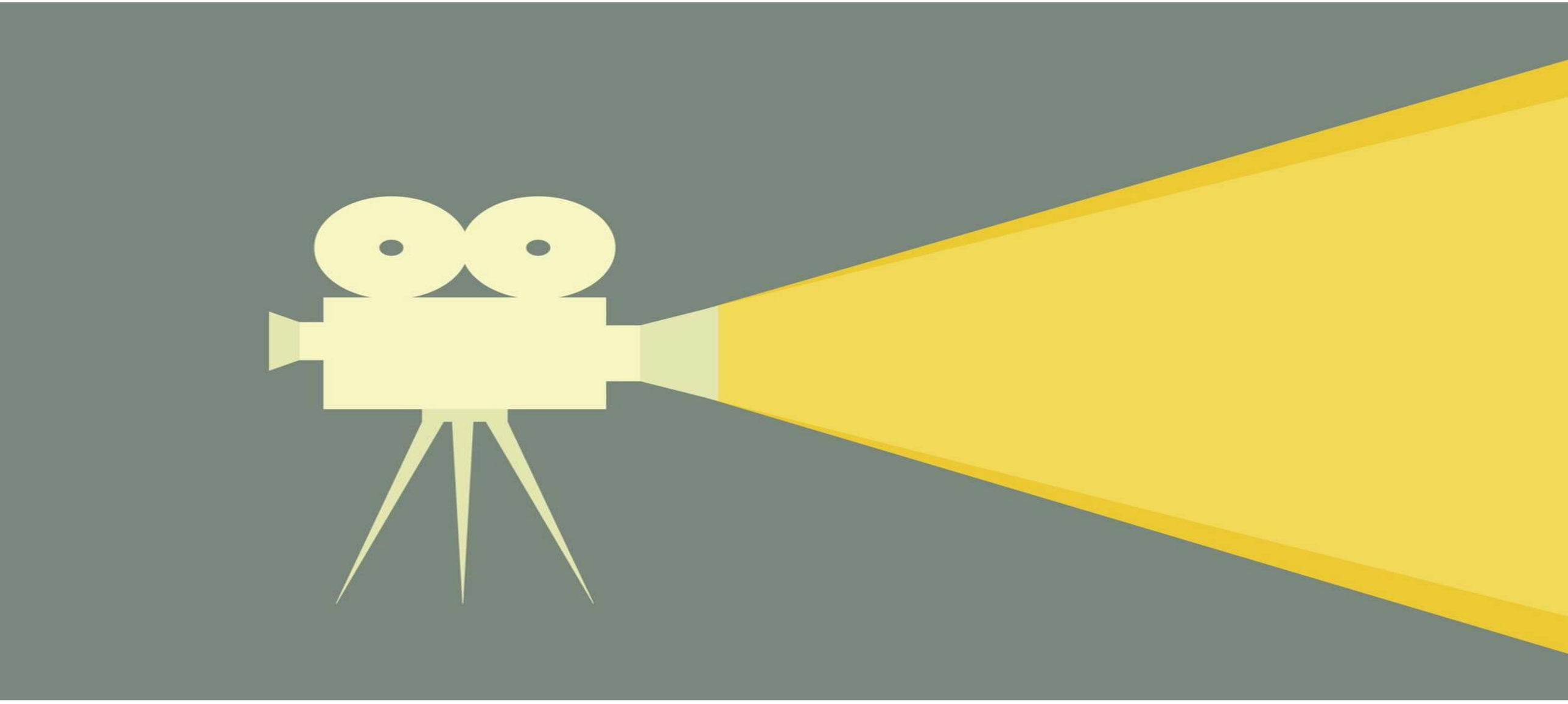


IMDB Movie Data Analysis

(2006 – 2016)

Project Presentation: Term-1

by **Sharad Goel**

Training Partner – **INSAID School**

Key takeaways from this dataset

1. Basic use of Pandas and Exploratory Data Analysis (EDA) which includes cleaning, combining, reshaping, slicing, dicing, and transforming data for analysis purpose.
2. Plotting graphs using pandas/seaborn like countplot, factorplot, swarmplot, barplot, violinplot, hexaplot, piechart, kdeplot, distplot, pairplot etc.
3. Using packages like matplotlib and seaborn to develop better insights about the data.
4. Create new features which will help in better prediction on hidden aspects of data.
5. Pandas Profiling to get an overall statistical knowledge of the data like any missing values and irregularities present in the data so as to normalize the data for better analysis.
6. Get to know correlation b/w different variables present in the data which might have an impact on overall finding.
7. Drawing final conclusion for the problem in hand.

Table of Content

Problem Statement

Analysis of last 10 yrs. (i.e. from 2006-2016) movies data on IMDB and come up with the “key success factors” for any movie based on past patterns and correlation b/w them

Data Description

The dataset contains 1000 observations of movies data hosted on IMDB. IMDB (Internet Movie Database) is an online database of information related to films, television programs, home videos and video games, and internet streams, including cast, production crew and personnel biographies, plot summaries, trivia, and fan reviews and ratings.

Users registered on this site are invited to rate any film on a scale of 1 to 10, and the totals are converted into a weighted mean-rating that is displayed beside each title.

It also displays the Metascore of each title. Metascore is the rating given by another movie rating company called Metacritic. However, unlike IMDB, they get ratings from registered well known rating agencies and calculates a weighted average of those ratings.

Data Dictionary

- •Rank - Movie rank order
- •Title - The title of the film
- •Genre - A comma-separated list of genres used to classify the film
- •Description - Brief one-sentence movie summary
- •Director - The name of the film's director
- •Actors - A comma-separated list of the main stars of the film
- •Year - The year that the film released as an integer.
- •Runtime - The duration of the film in minutes.
- •Rating - User rating for the movie 0-10
- •Votes - Number of votes
- •Revenue - Movie revenue in millions
- •Metascore - An aggregated average of critic scores. Values are between 0 and 100.

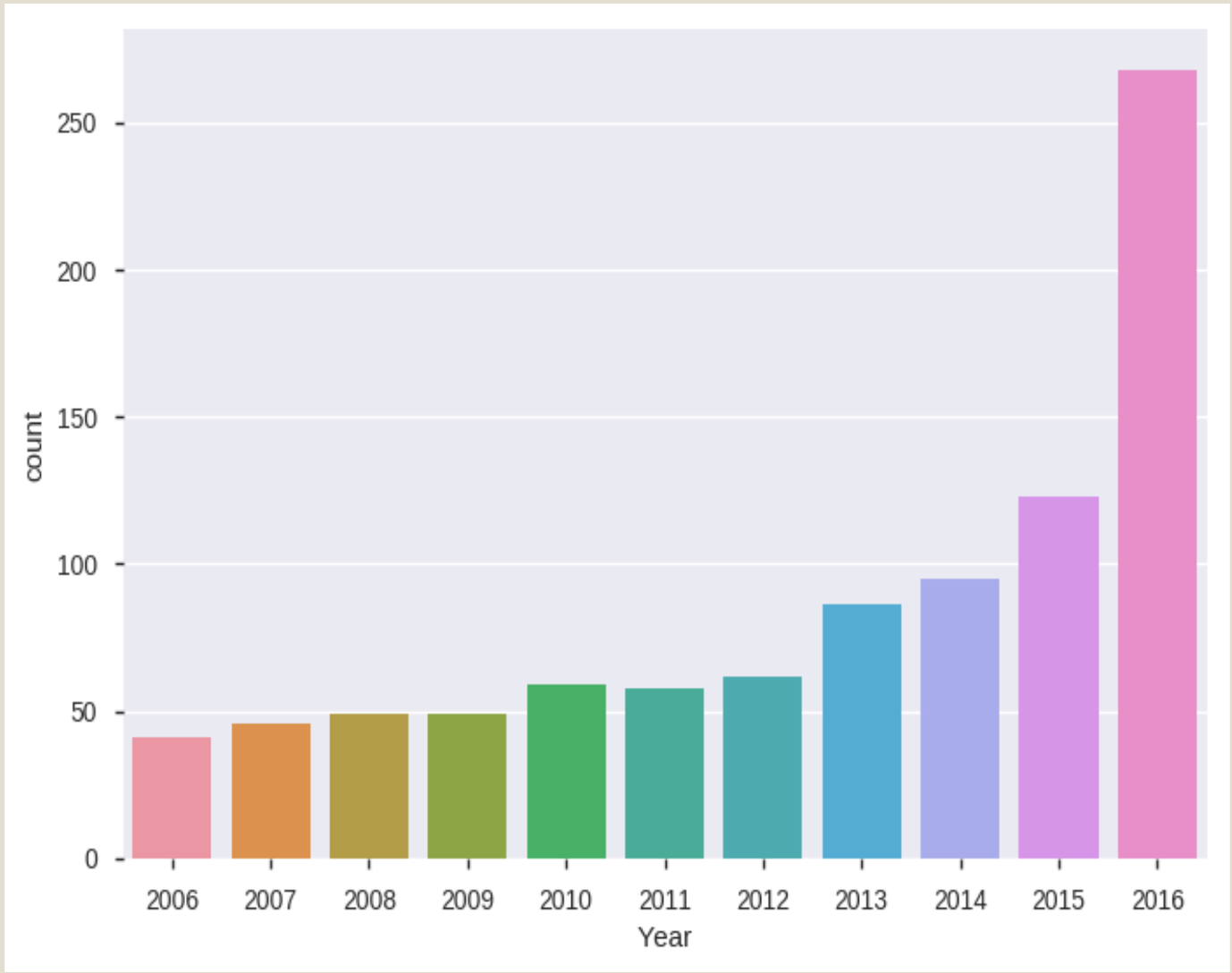
Higher scores represent positive reviews.

‘Year’ wise Trend and Findings..!!

How many movies got released Yr.-on-Yr.?

- As visible from the barplot b/w count of movies produced Yr.-on-Yr., there is an upward trend, i.e. more no. of movies are getting released in the showbiz

- The count has increased by more than 5 times from that of 2006, i.e. from 41 to 268

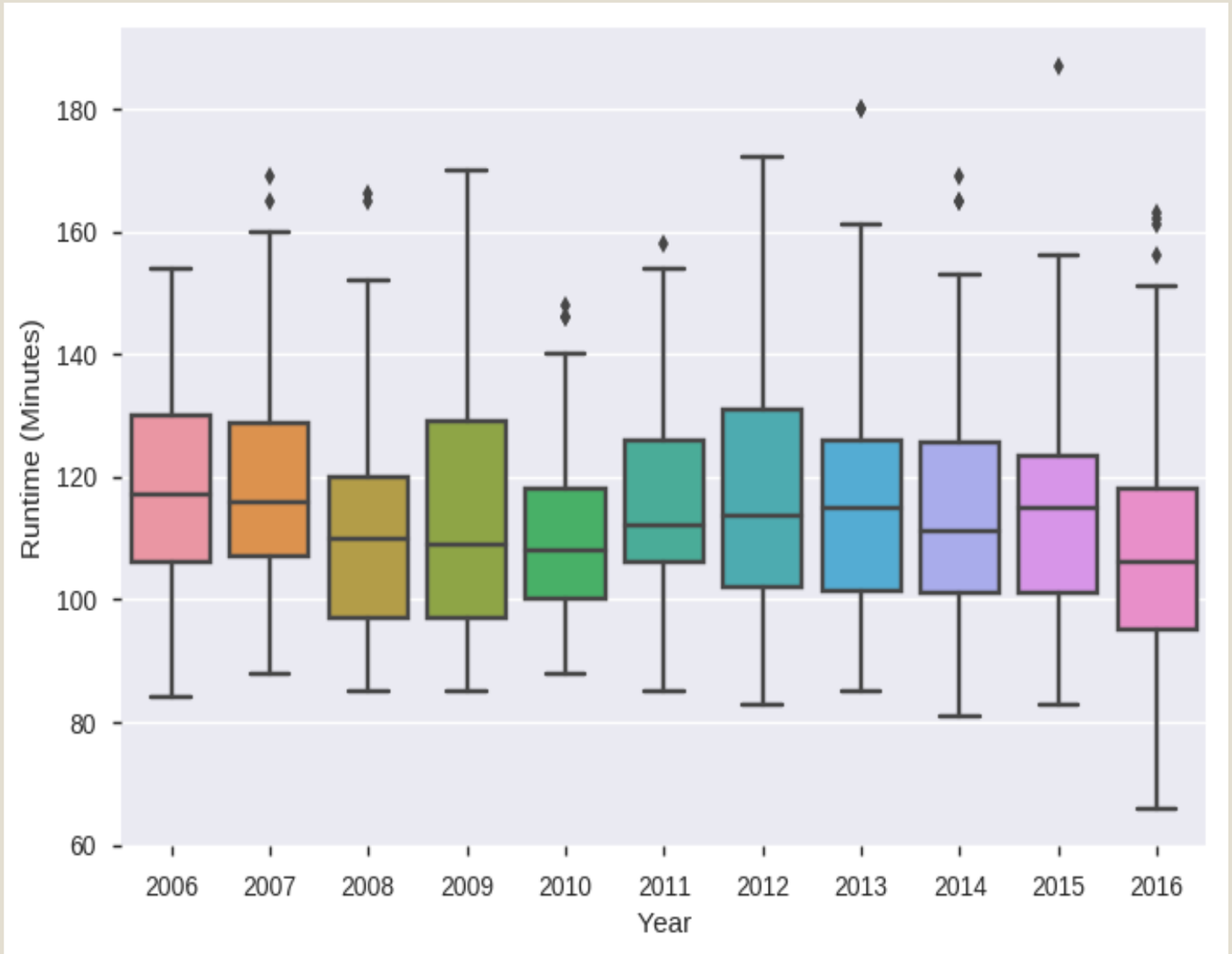


How does the duration of movies vary during 10 yrs.?

• From the boxplot, it is seen that the Runtime (in Minutes) of the movie has come down over a period of years

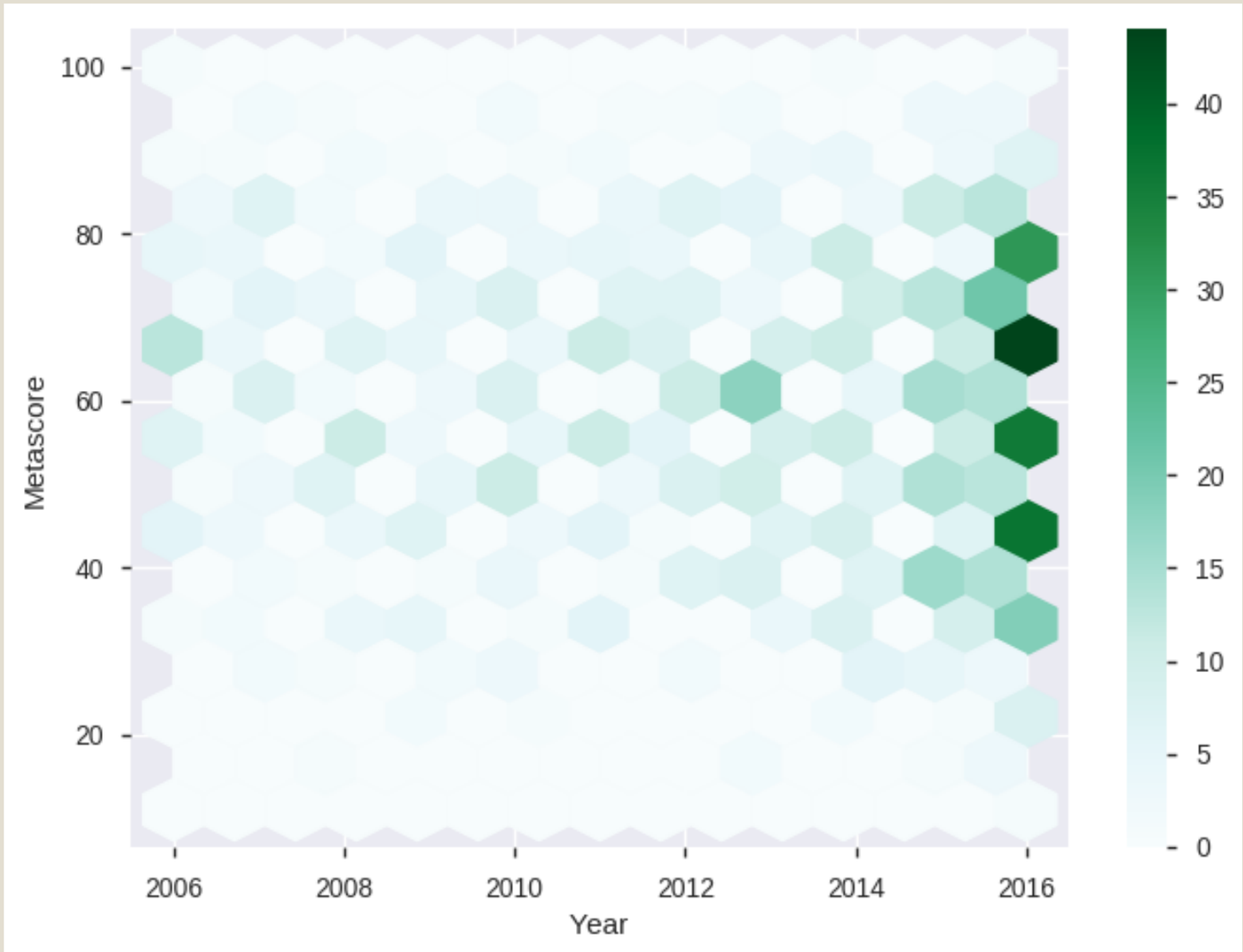
• Mean Runtime in 2006 was 119 mins. whereas it was 107 mins. In 2016, with most of the movies being produced in the UQR (upper quartile range)

• There were some outliers as well, but their count is almost negligible



How does Metascore vary at yearly intervals?

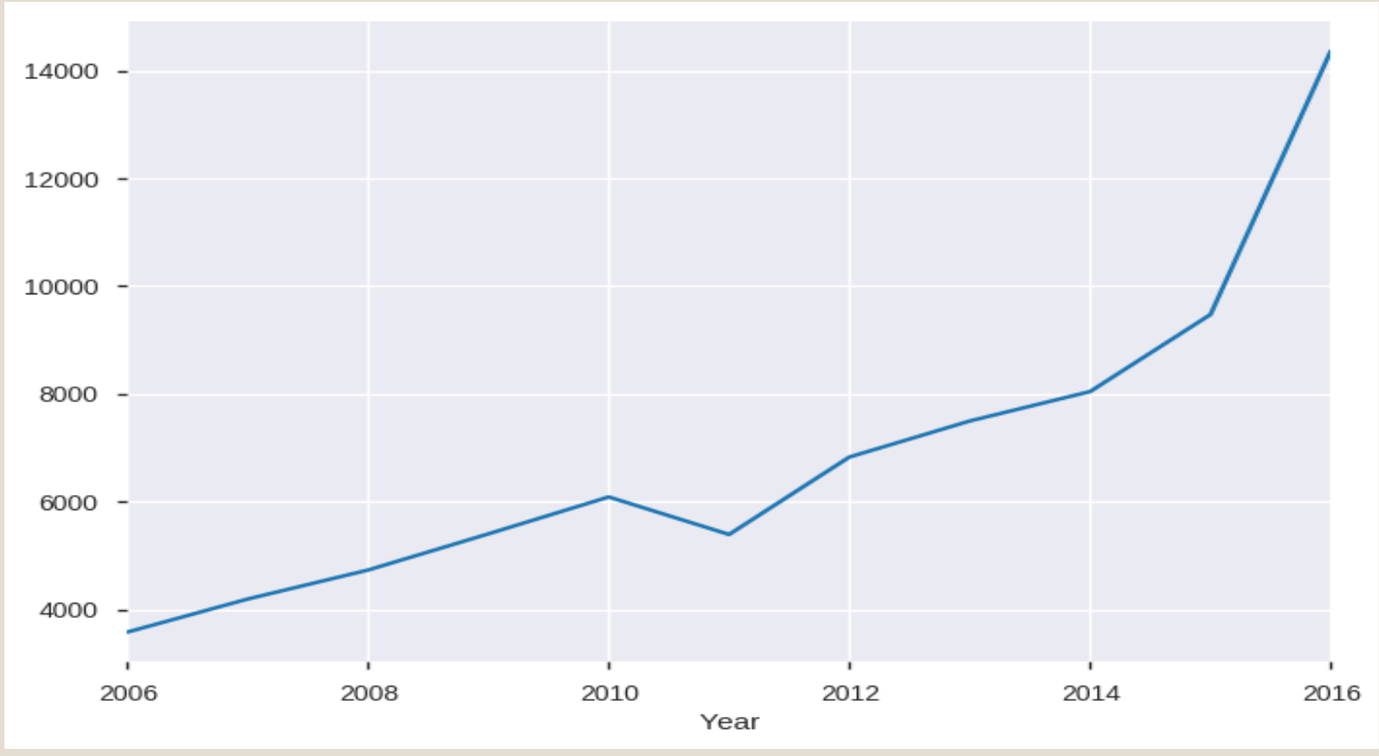
• From the hexaplot, it is clear that during initial years Metascore was not good, but it was very high for movies released in the later years i.e. 2016

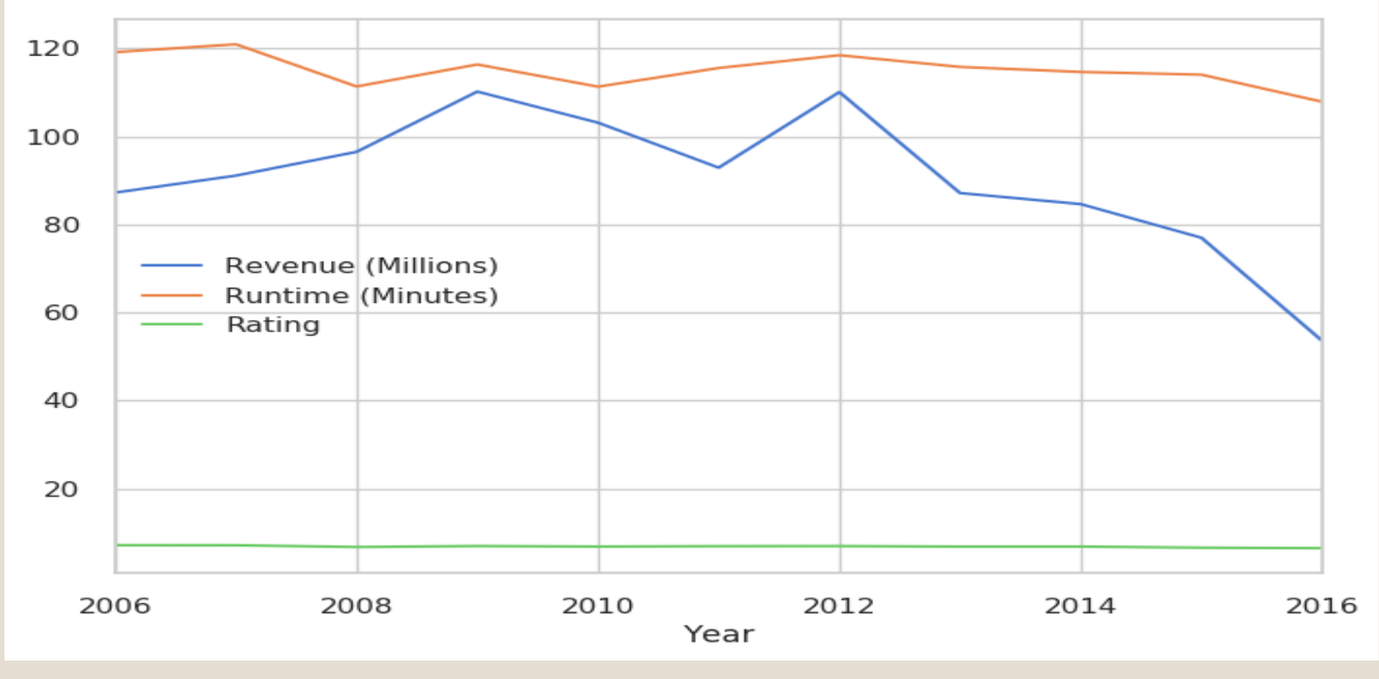


What is the performance trend on: Revenue, Runtime, Rating?

• From the first lineplot, the overall Revenue has been increasing due to high no. of movies getting released

• From the below chart, it is clearly visible that Revenue, Runtime and Rating have seen a downfall over a period of time

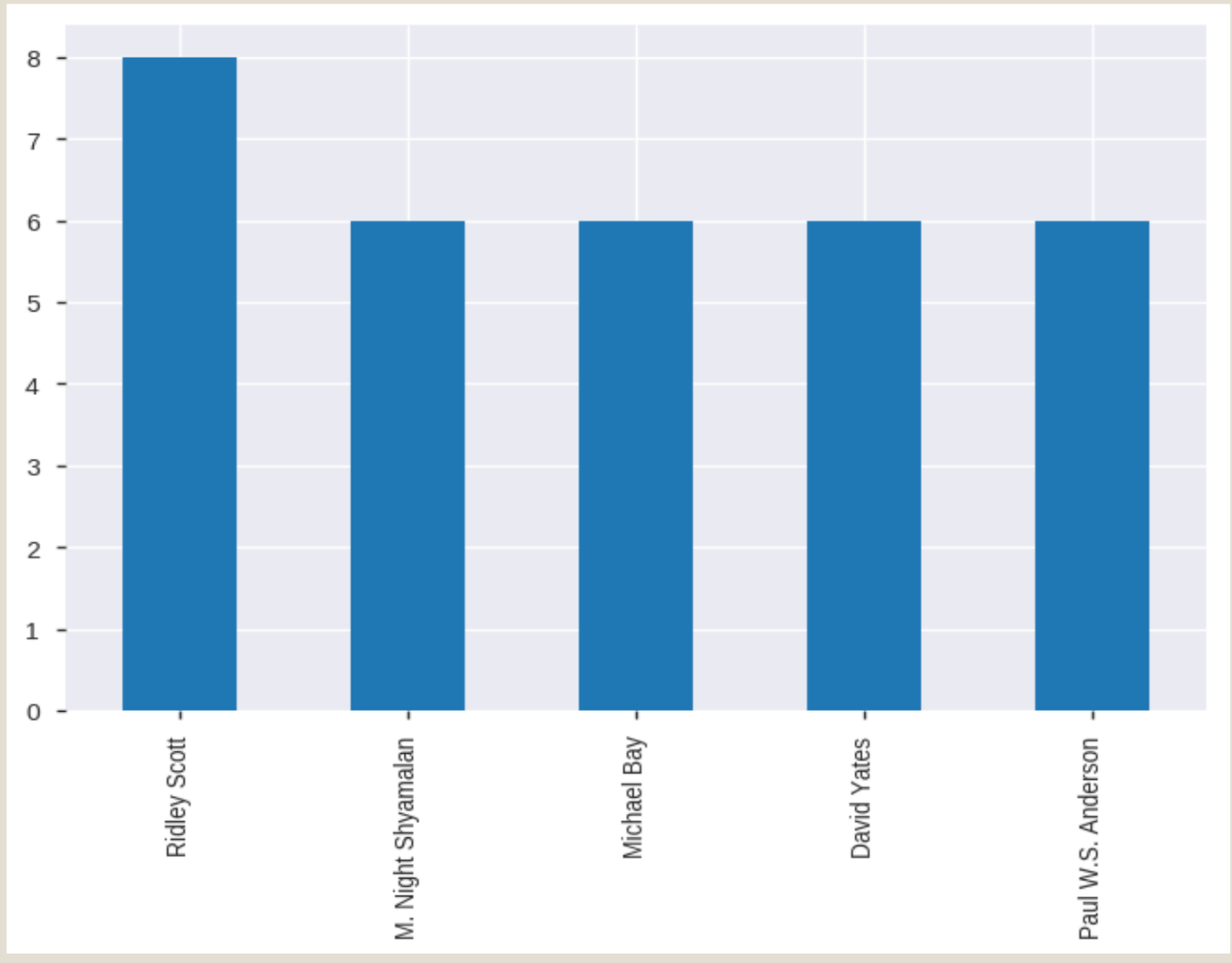




‘Director’ wise Trend and Findings..!!

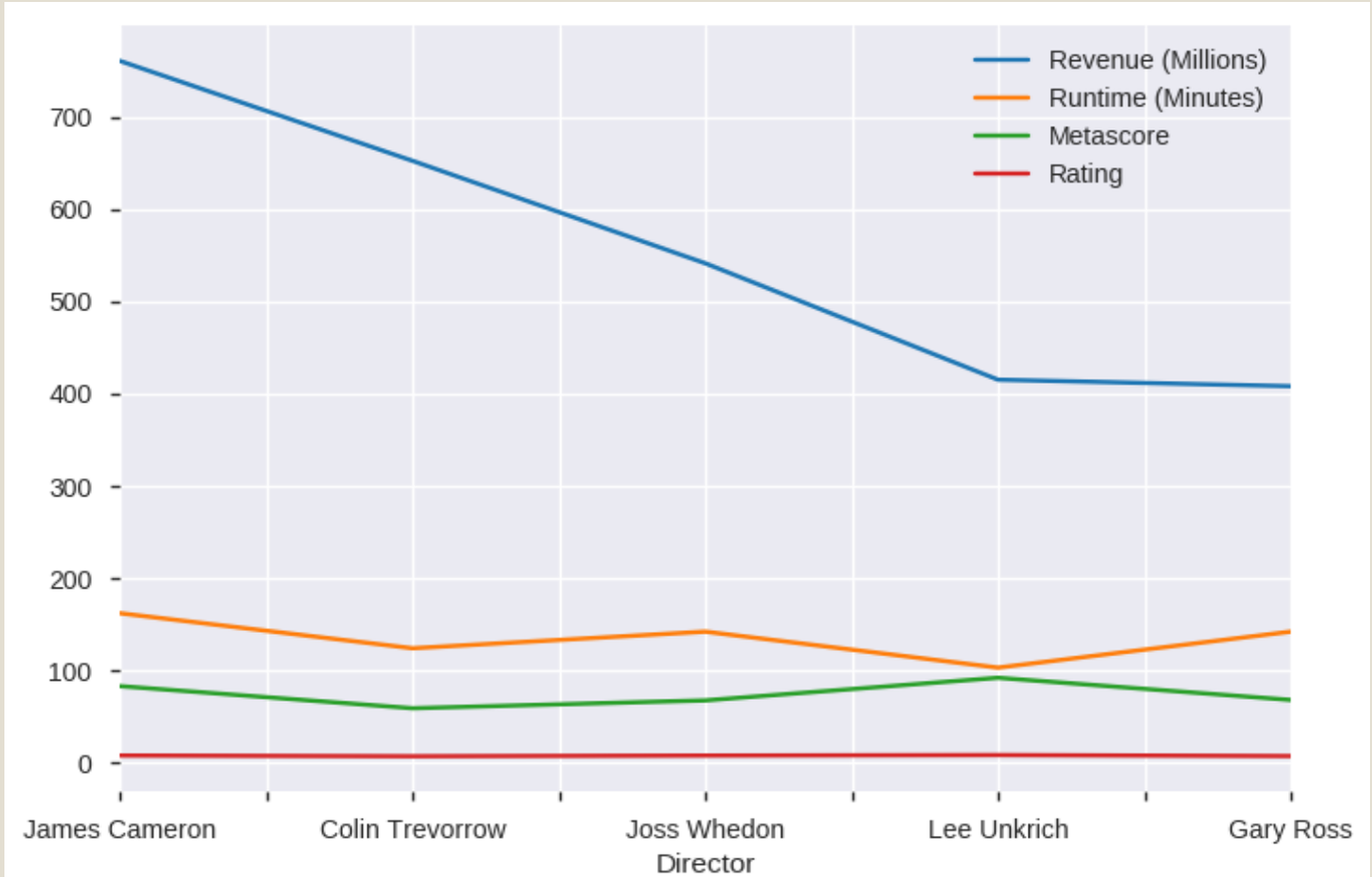
Which directors have produced highest no. of movies?

• The barplot depicts Directors (Top 5) producing highest no. of movies in the 10 yr. timeframe



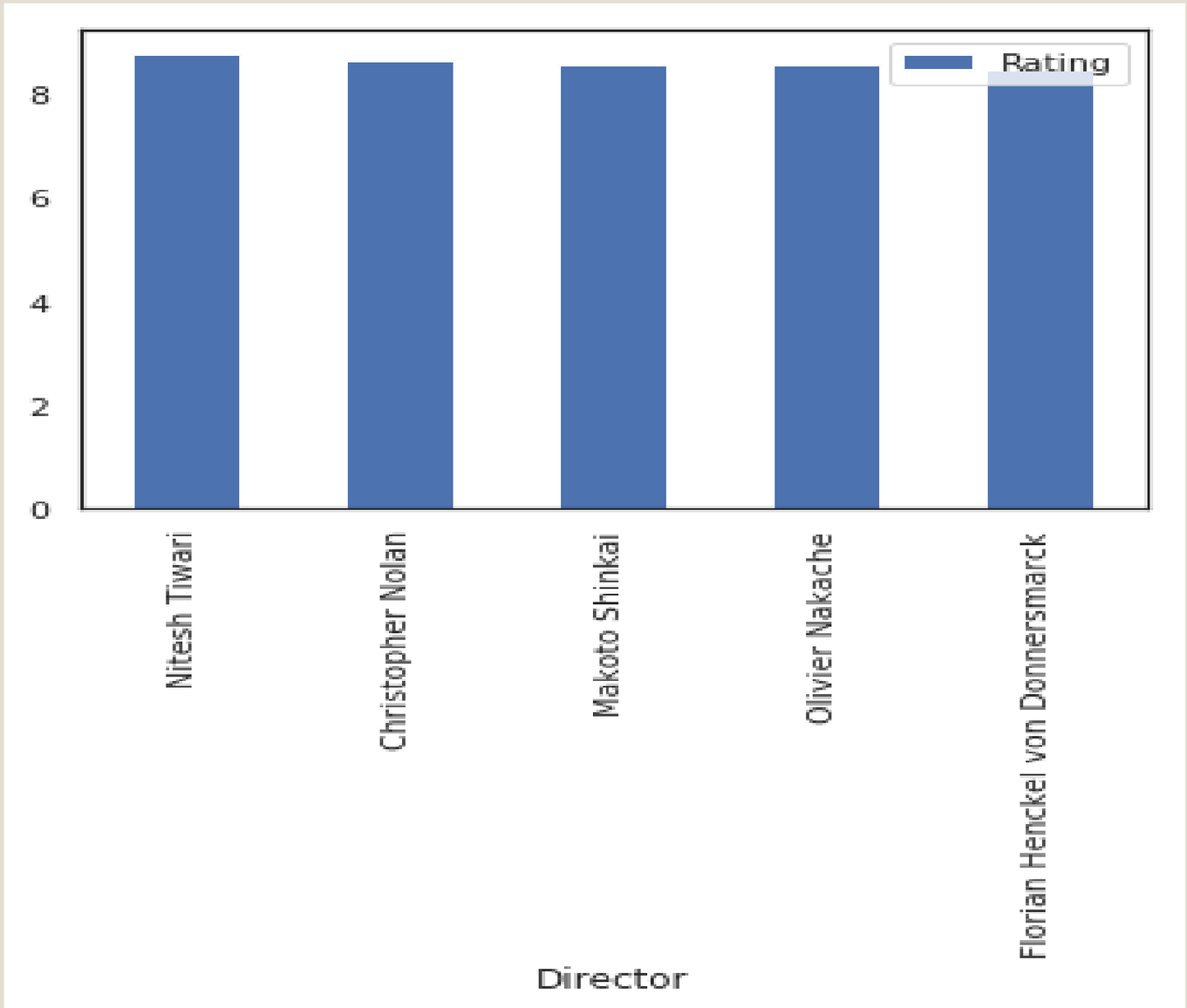
Who are the highest revenue generating directors and their performance on other parameters?

• The lineplot depicts Directors (Top 5) earning highest gross revenue and their measuring their performance on other important parameters



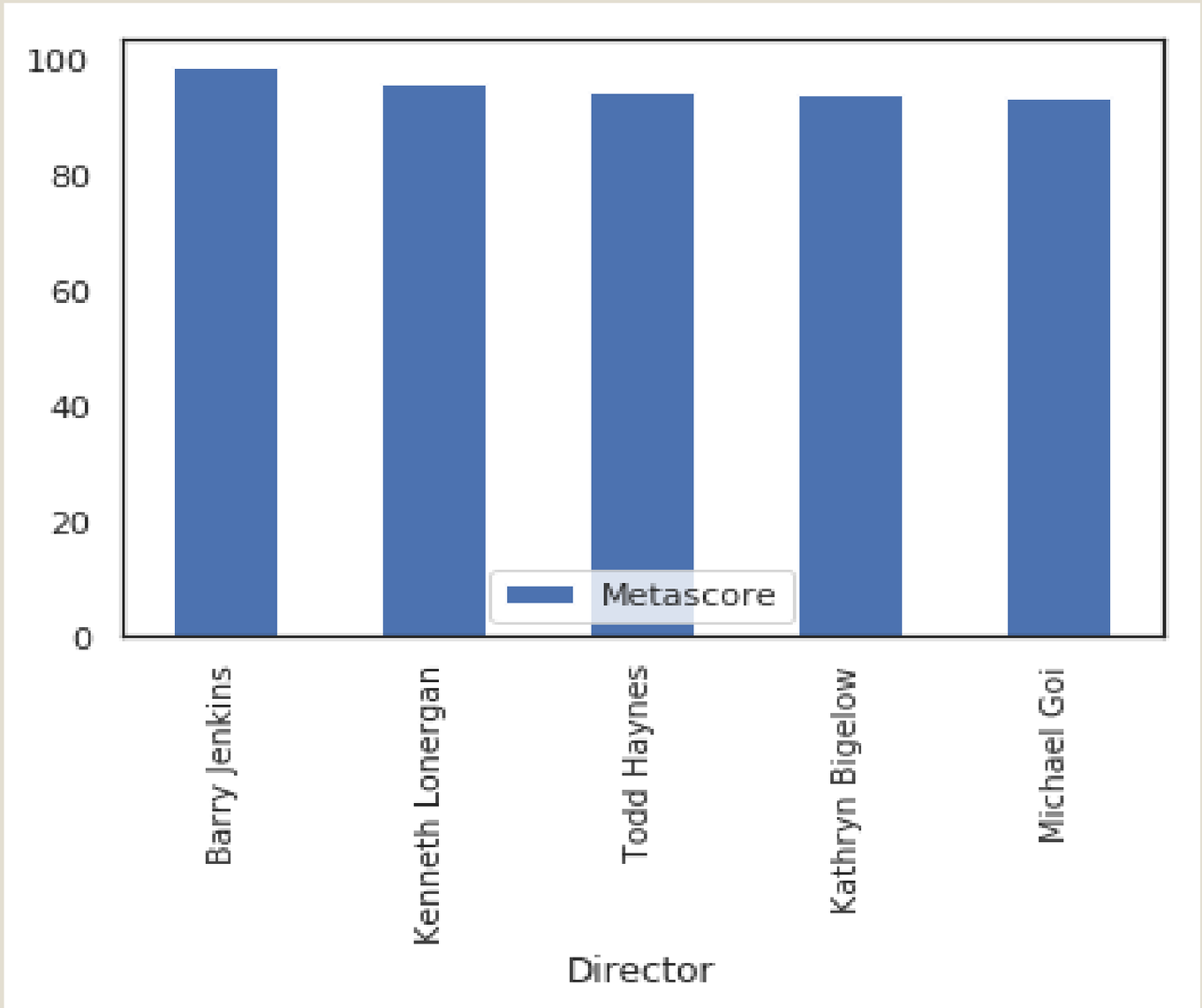
Who are the top directors with highest rating?

• The graph depicts directors list being rated highest on movie ratings



Who are the top directors with highest Metascore?

• The graph depicts directors list being rated highest on movie's Metascore



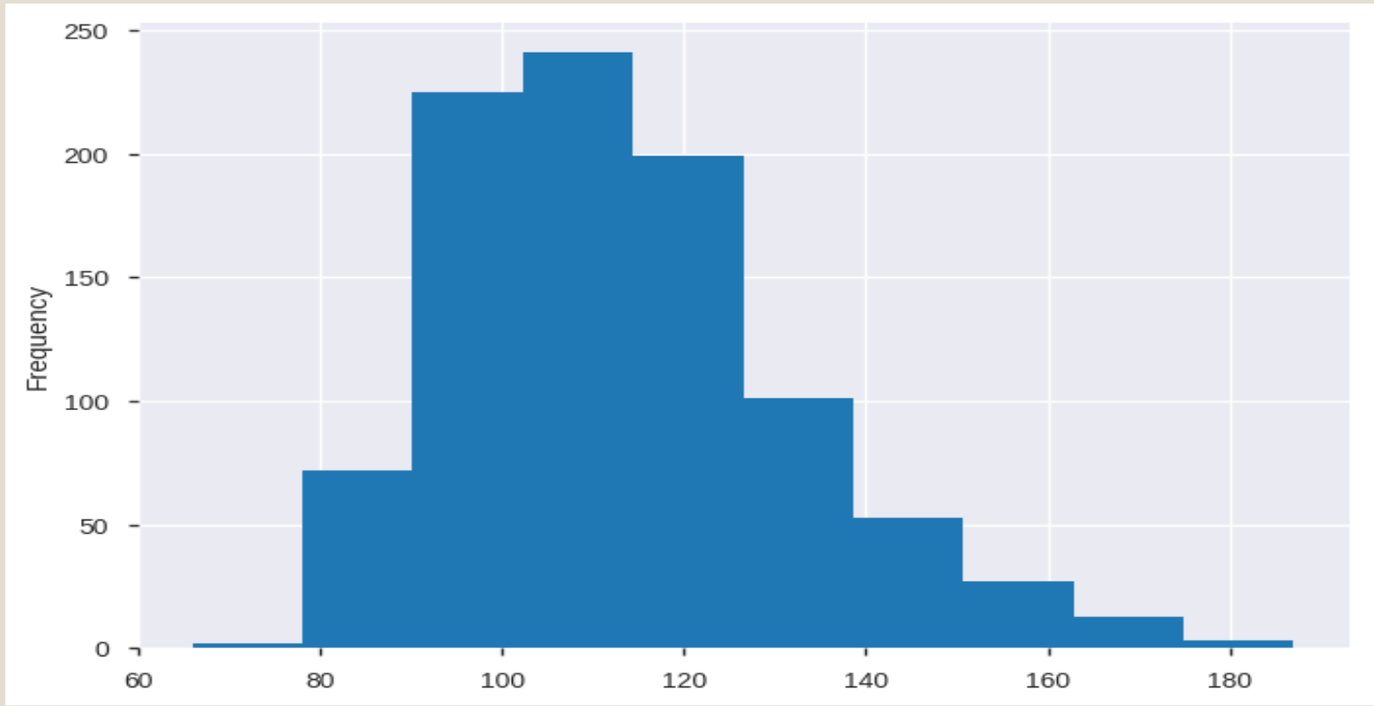
‘Runtime’ wise Trend and Findings..!!

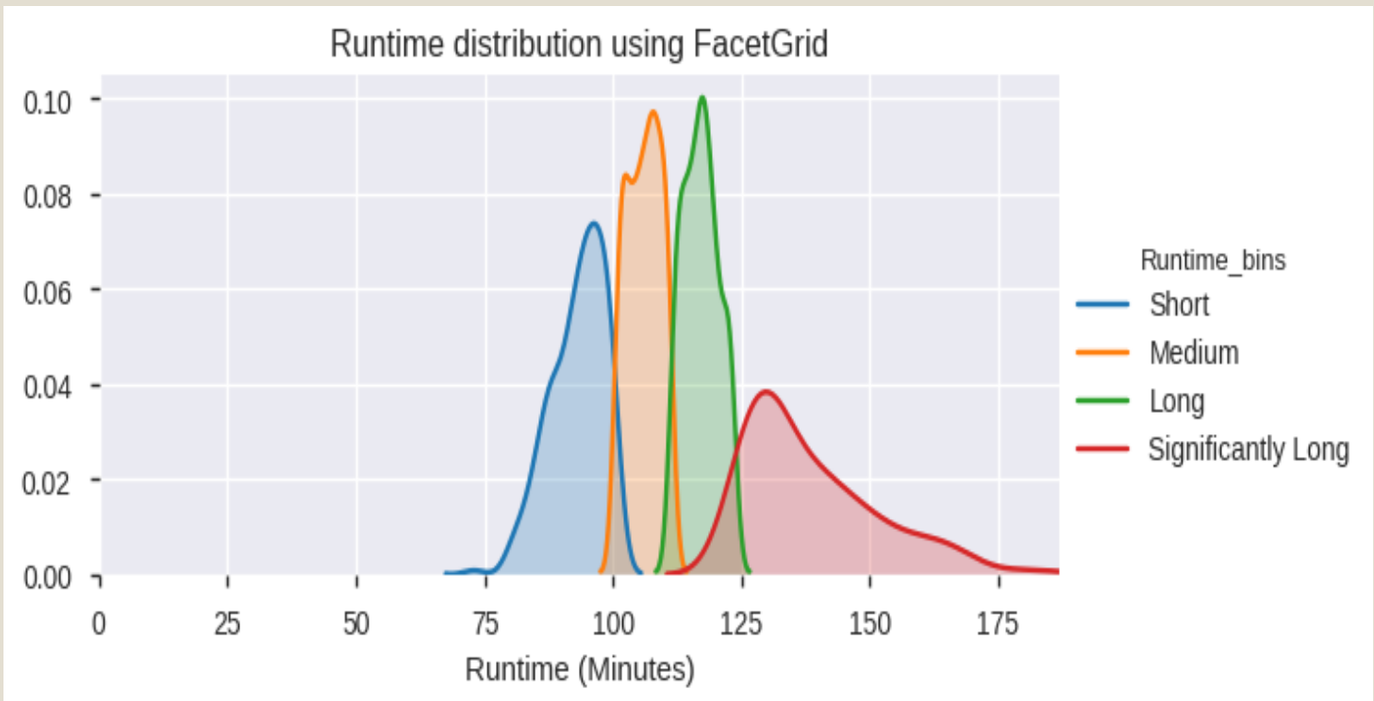
What is the Runtime distribution trend?

• We can see majority of the movies produced have Runtime b/w 100 to 115 mins.

• Both the charts depict the same but presented differently for ease of connect

• In the below Facetplot, grouping/bins have been created for better segregation

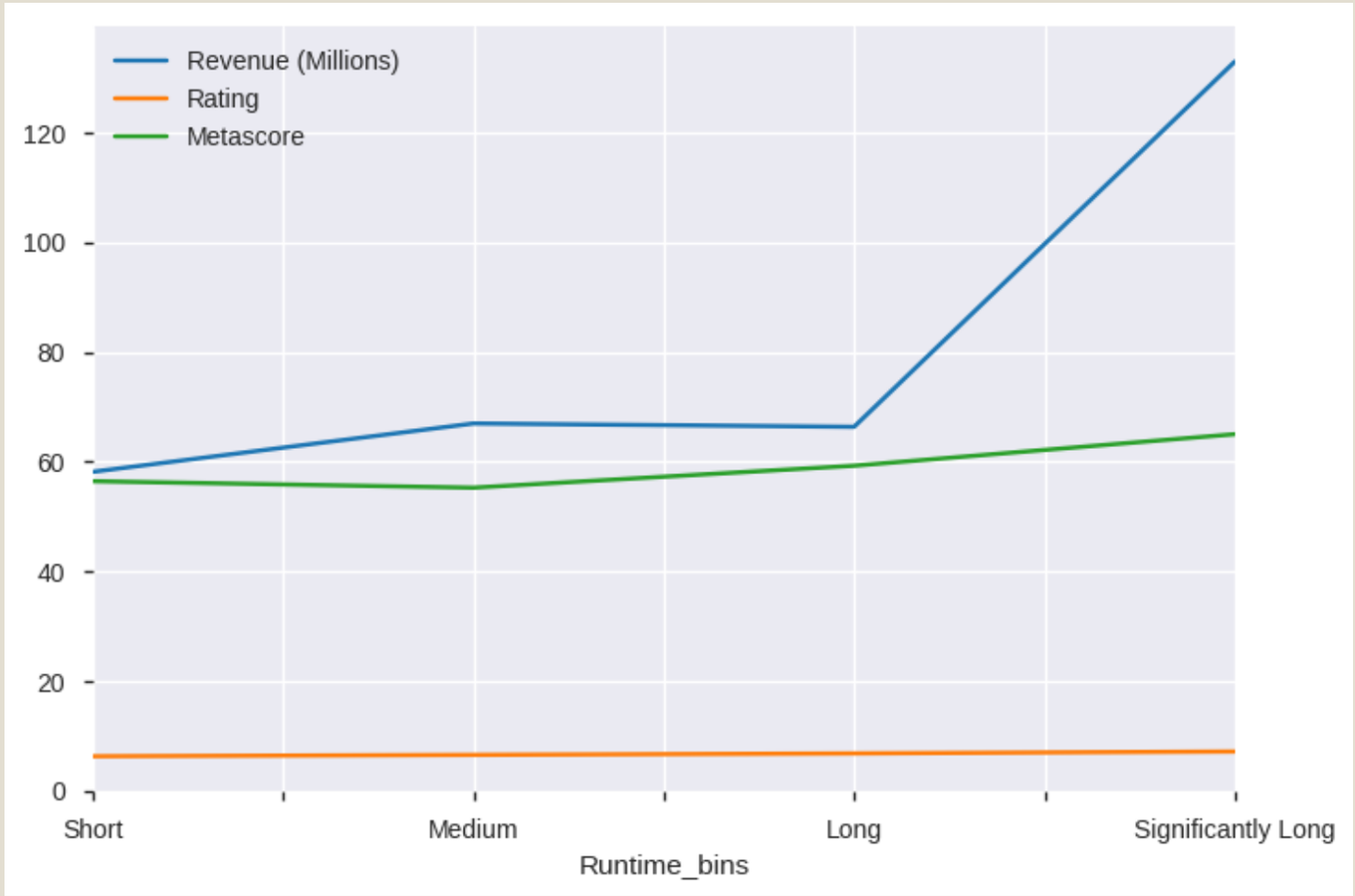




How does Runtime affect Revenue, Rating & Metascore?

- As visible, movies having higher Runtime tend to perform better on Revenue, Rating and Metascore

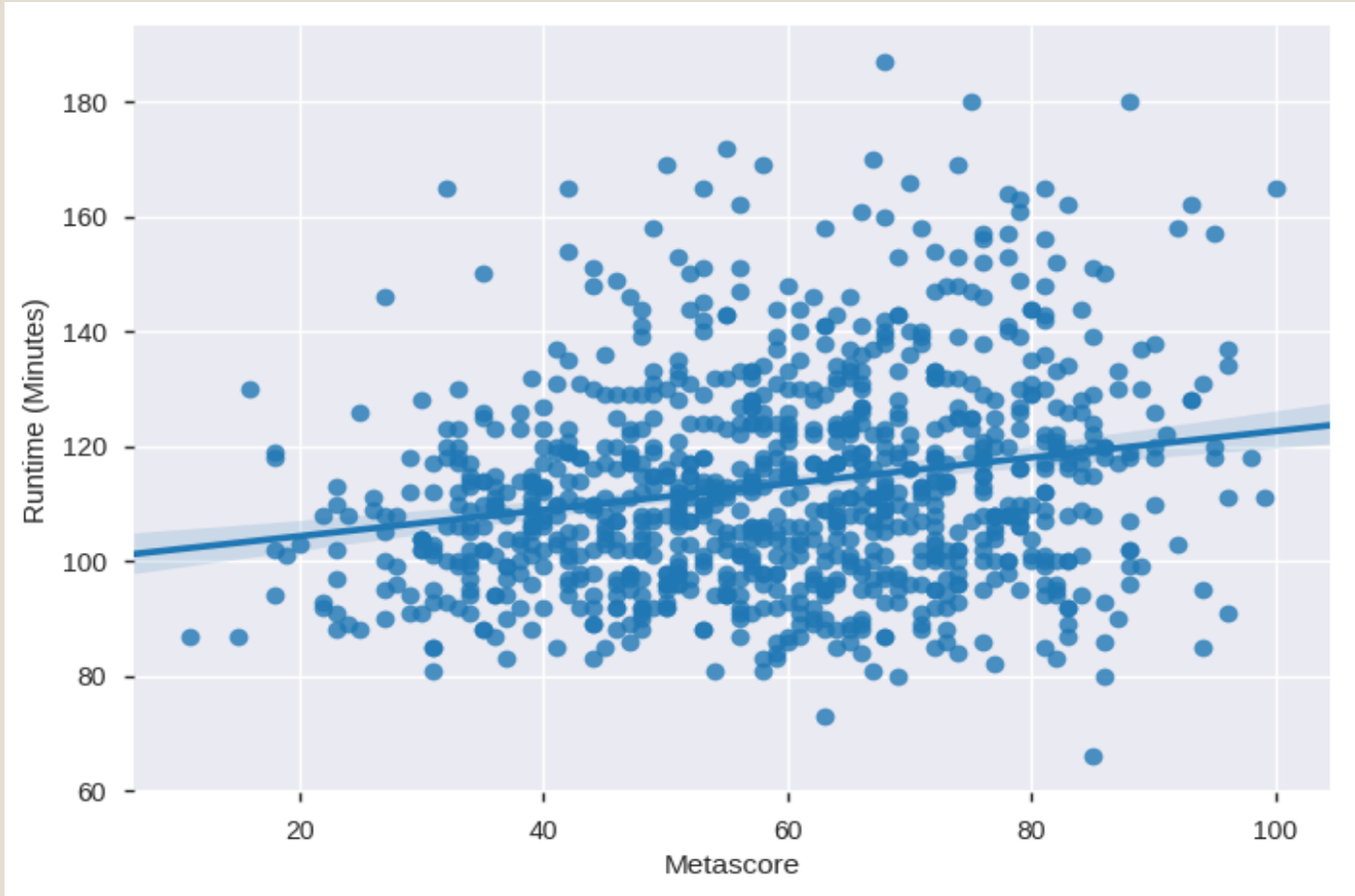
- Audience is more likely to enjoy movies with more than 2 hrs. duration



How are Runtime and Metascore related to each other?

- As we can see most of the concentration is b/w 40-80 (Metascore) and 80-140 (Runtime)

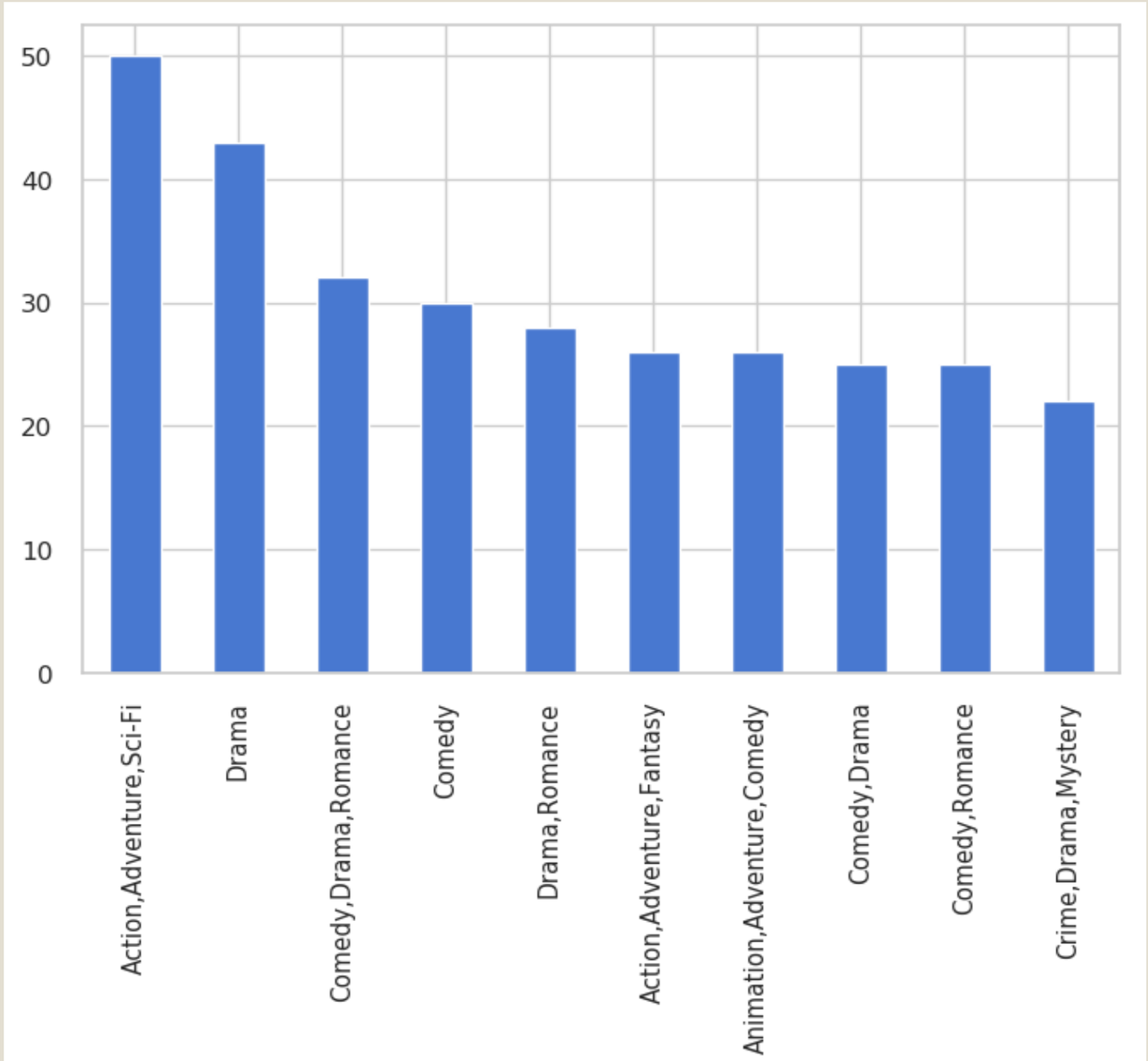
- But also noticed that better the Runtime (>140) higher is the Metascore



‘Genre’ wise Trend and Findings..!!

Which are the Top Genres count?

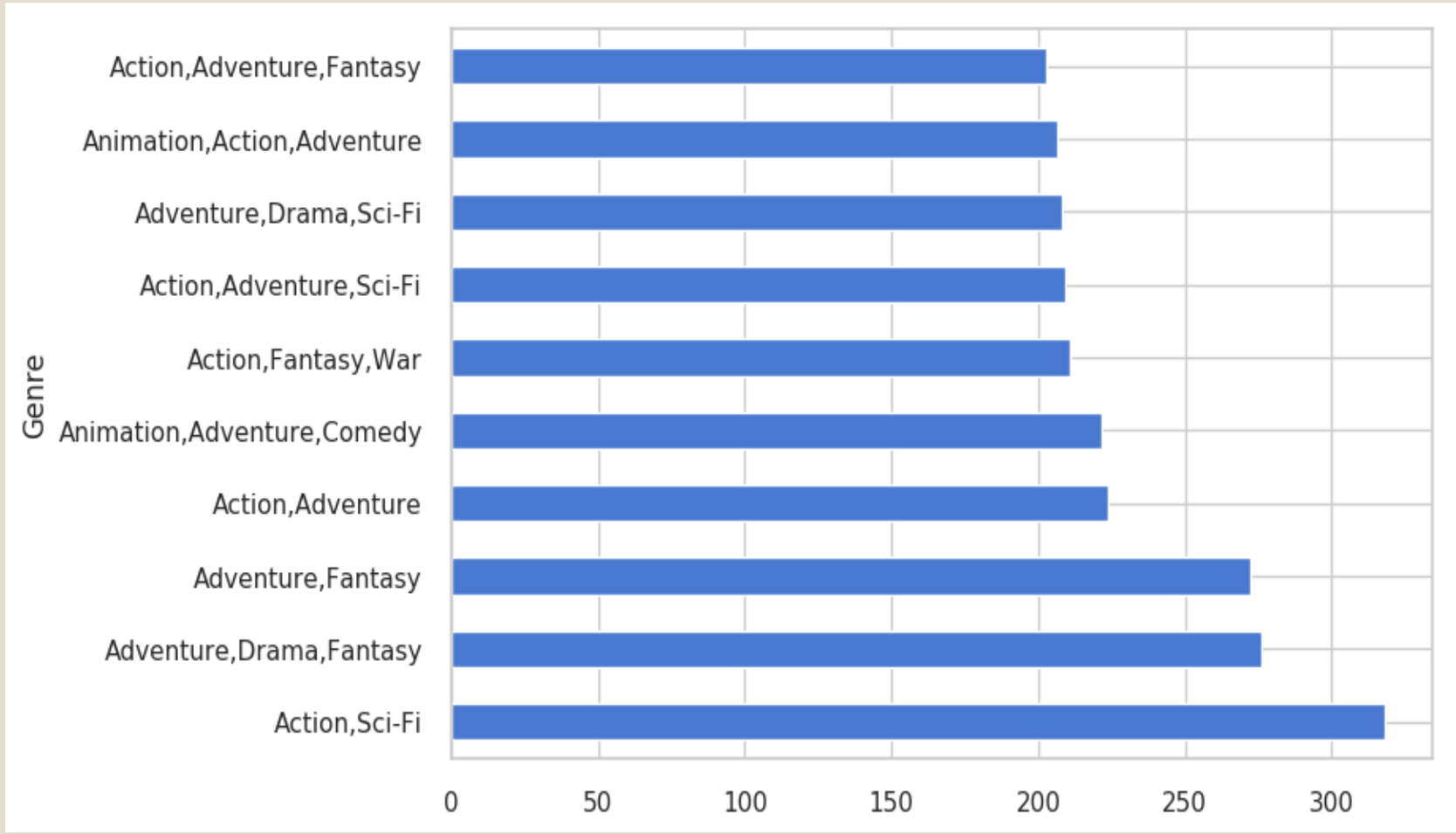
- As we can see, Action, Adventure and Sci-Fi tops the list in terms of production to be followed by Drama and other combinations



Which Genre movie has generated the highest Revenue?

- Action, Sci-Fi movies have generated the highest gross earnings followed by Adventure, Drama, Fantasy

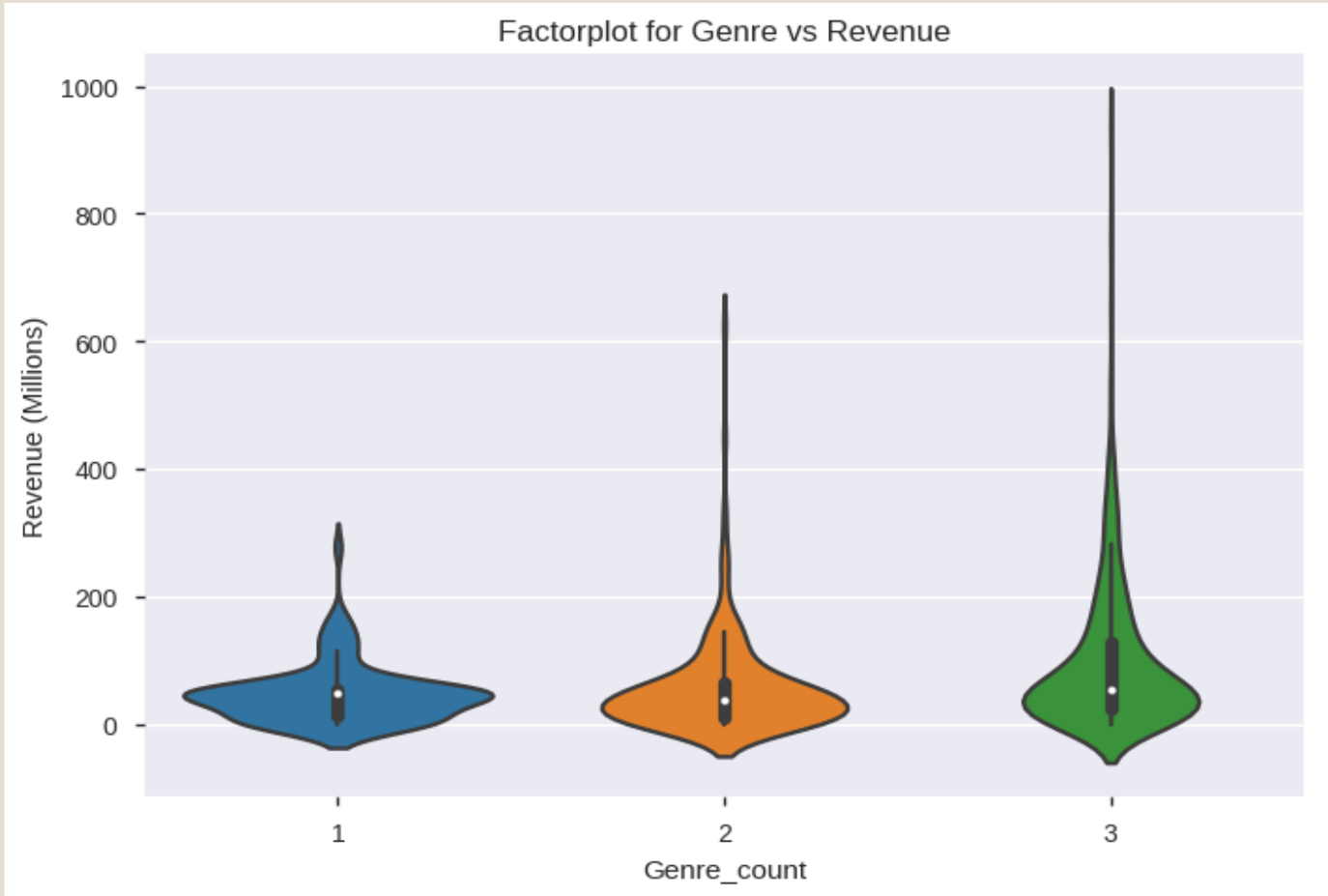
- Adventure and Action seems to be most preferred genres by audiences



How are Genre and Revenue linked to each other?

• As visible from the violinplot, movies having 3 genres have earned higher revenues than movies having 1 or 2 genres

• People are more likely to watch movies with variations combinations rather than any fixed theme



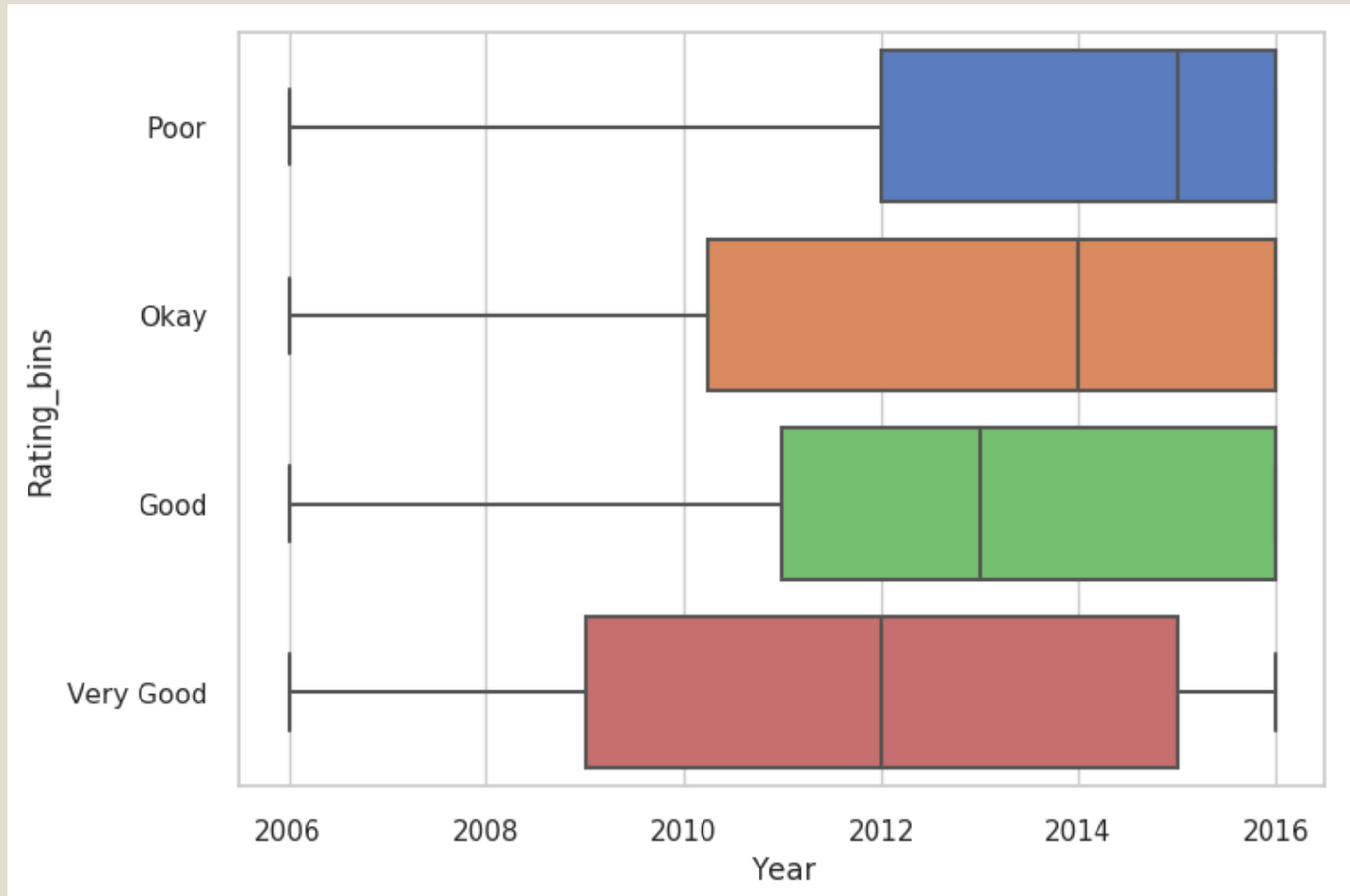
‘Rating’ wise Trend and Findings..!!

How Rating traverses through Year?

• Movies released during later 2016 have been rated lower than the one's produced during initial years

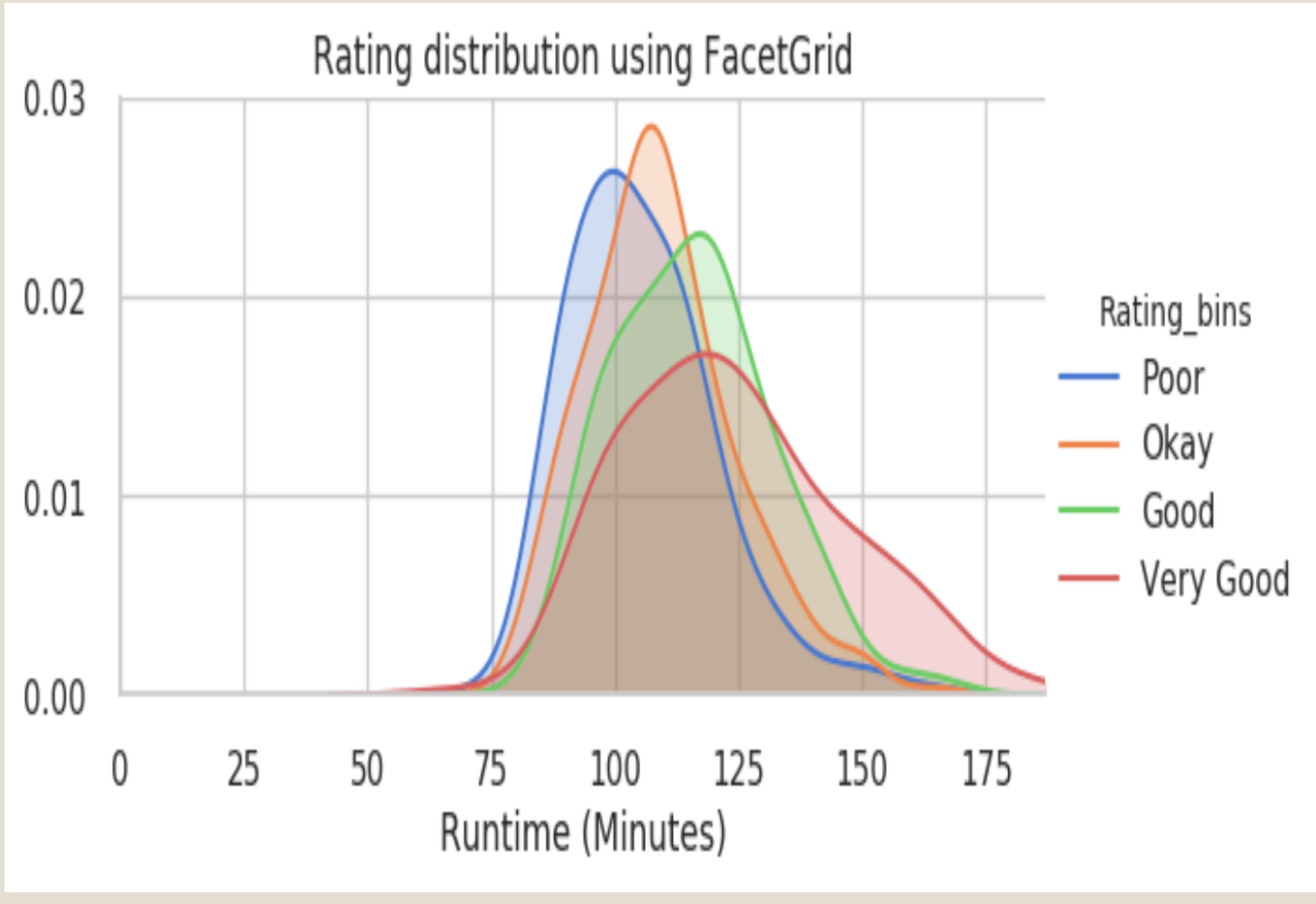
• From 2009 – 2014, better movies used to come in the market and liked by the audience

• From 2015 – 2016, most of the movies were able to secure either Okay or Poor customer ratings



How are Rating and Runtime related to each other?

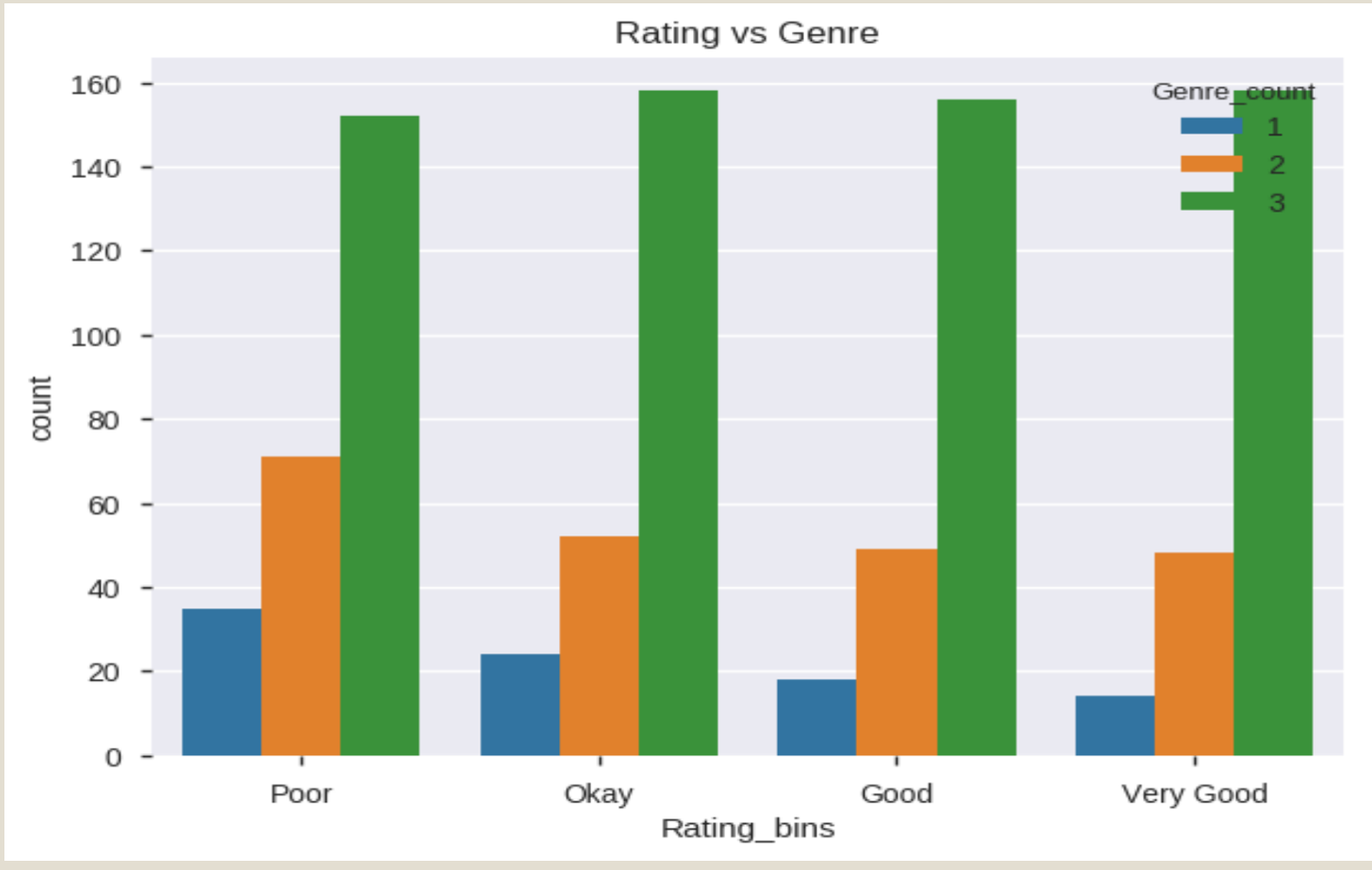
• As we can see movies which have higher Runtime have been rated either Good or Very Good



Does Rating and Genre affect each other?

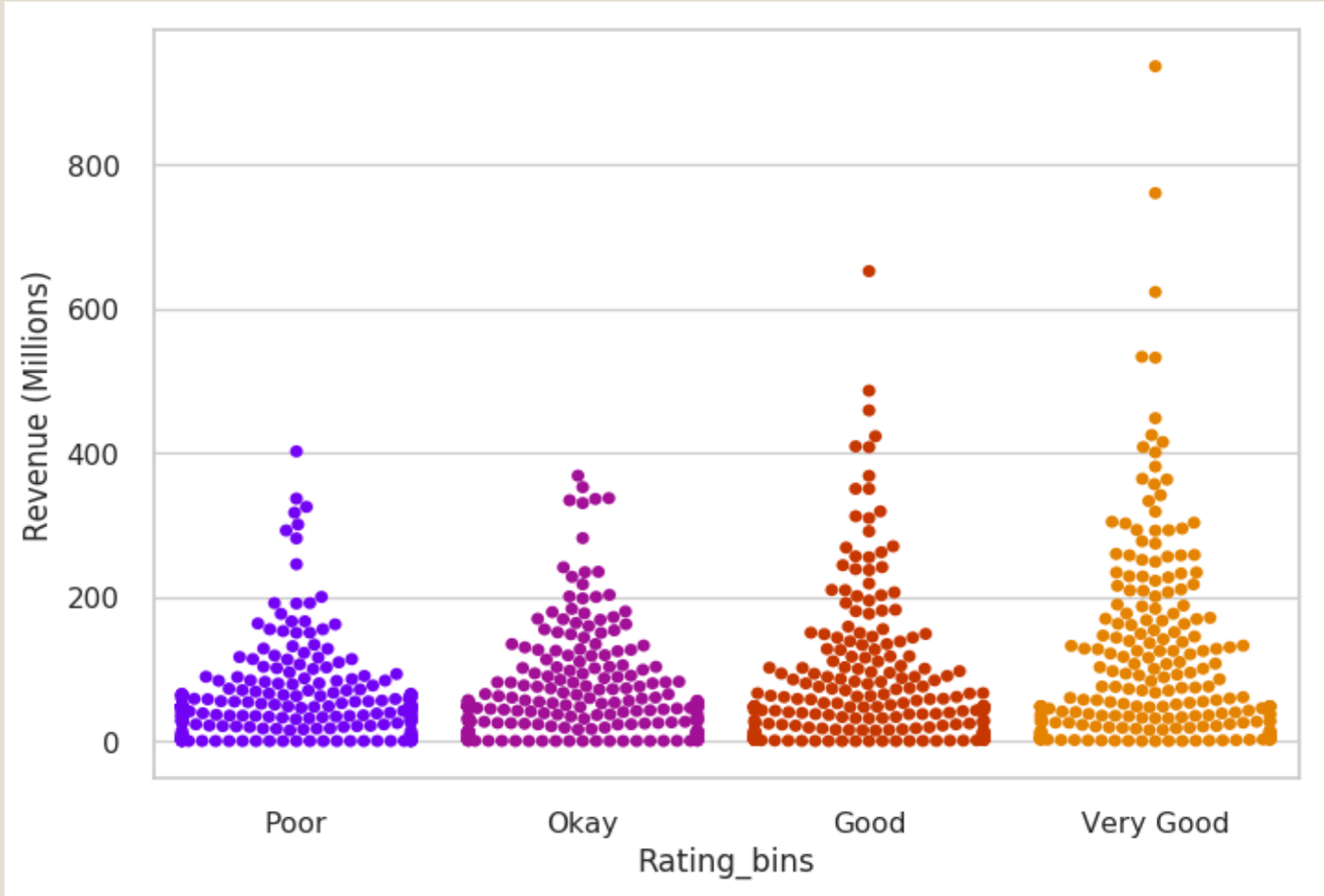
- As seen audience aren't able to judge much b/w different genres

- Ratings given by them is not affected by either type of genre or count of genre



Is there relation b/w Rating and Revenue?

• As clearly visible from the swarmplot, audience ratings have a direct influence on the Revenue and movies with better ratings have been watched more and by more people

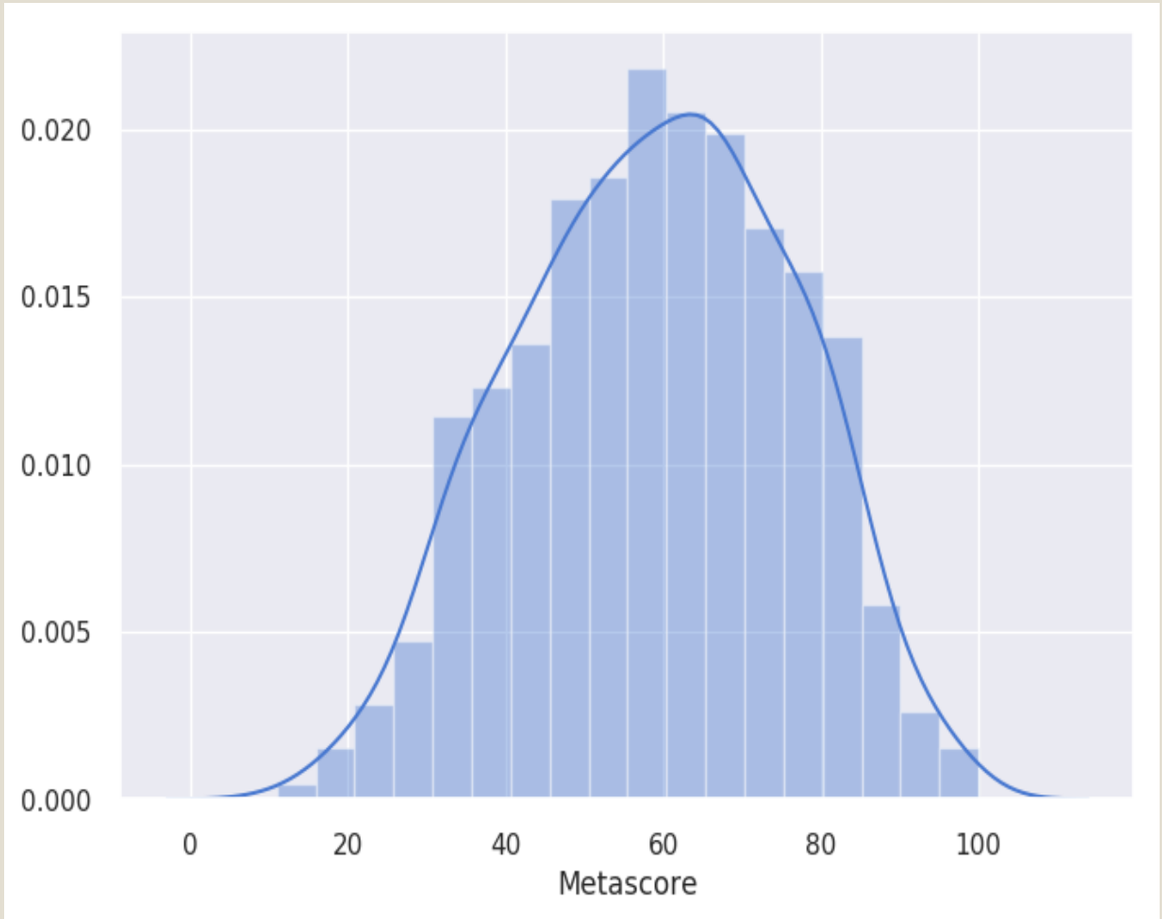


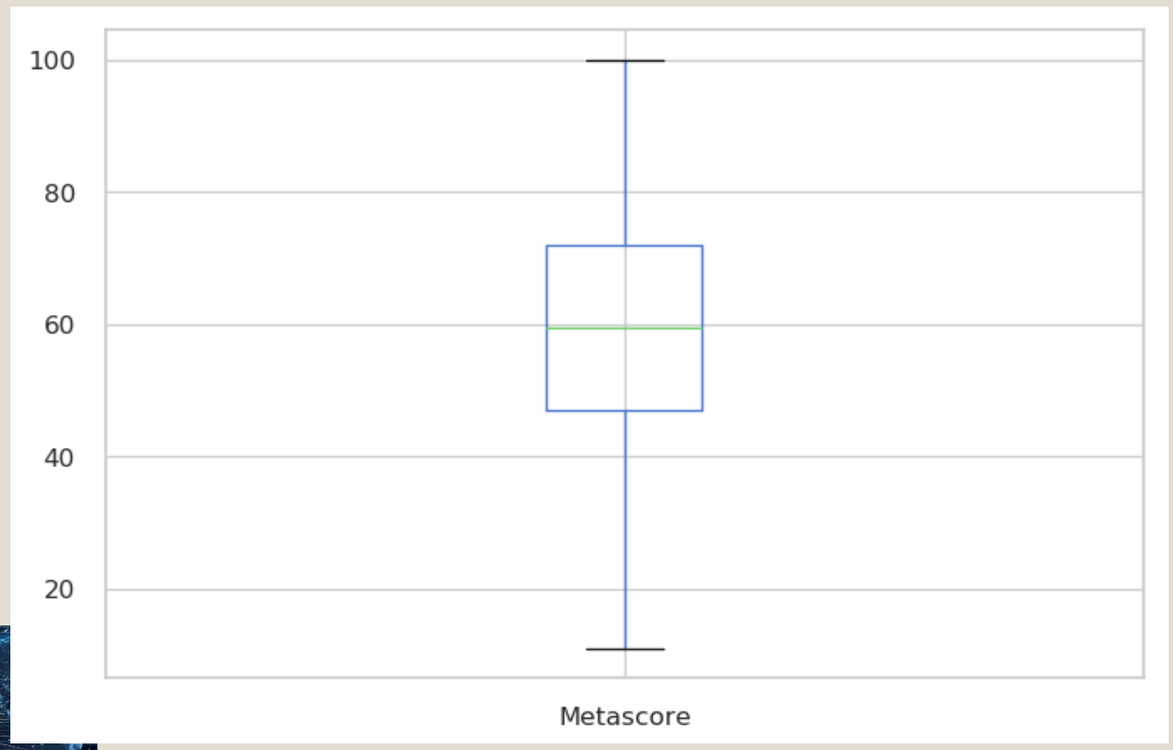
‘Metascore’ wise Trend and Findings..!!

How is the Metascore distribution?

•From the barplot, it is seen that 50% of the critics score stands at around 60

•It is more or less a uniform distribution curve with mean and median almost in same range





•Critics aren’t very much biased while scoring the movies and prefer to give a balanced judgement across different movie class and features which is a fair stance

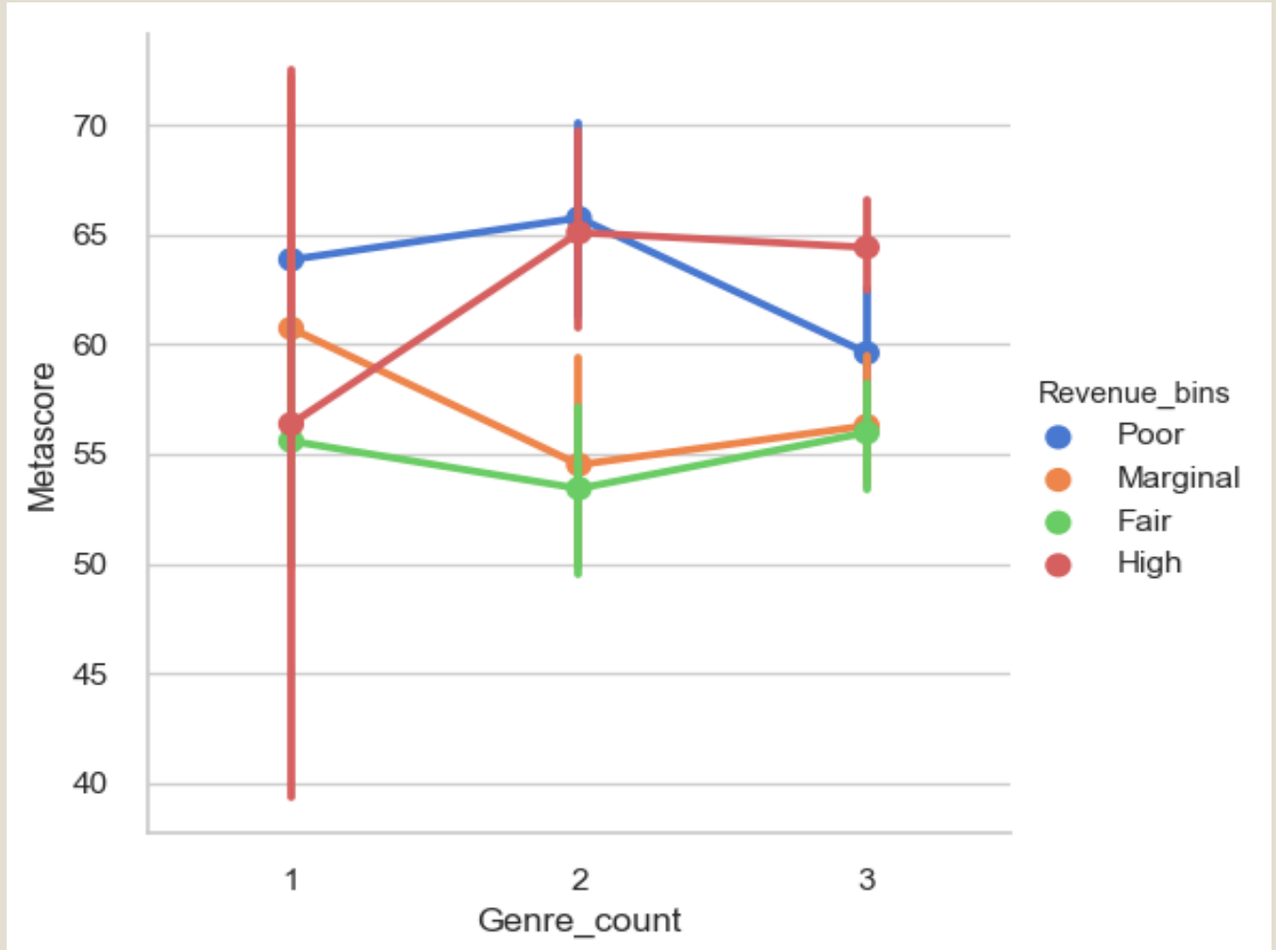
How does Genre and Metascore affect Revenue?

From Factorplot, it is seen that:

• Metascore and Revenue earning is high for movies with 3 genres.

• Metascore with 55 score and genre 1 has earned poor revenues while compared with genre 2&3 which have Metascore > 60

• Movies with more genres and good critics have high earning potential

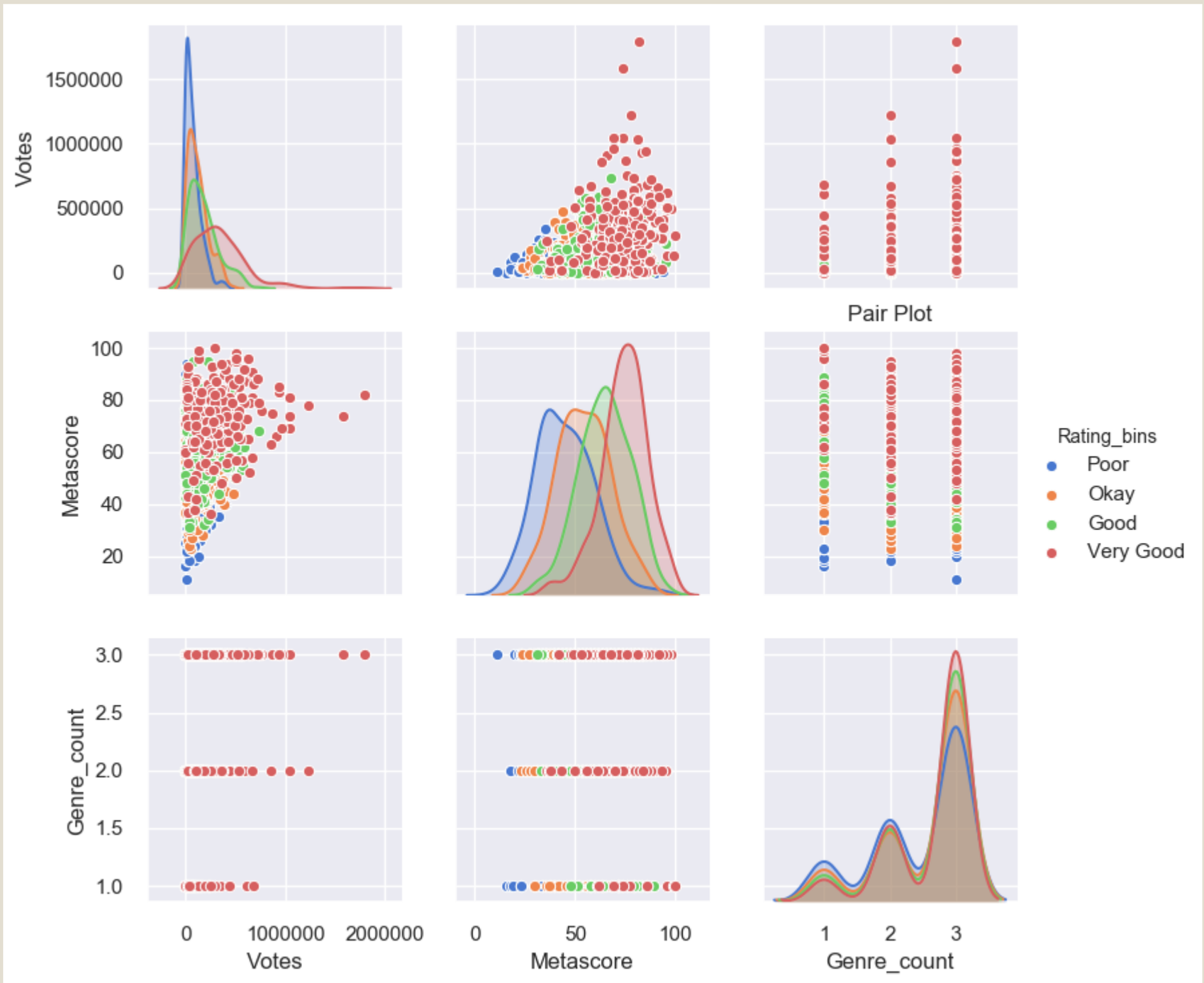


Overall Correlation..

How does Votes, Metascore & Genre affect the audience choice for any movie?

From the pairplot, it is observed that:

- Movies having higher ratings have received large no. of votes
- Higher Metascore ensures better ratings and higher votes
- Movies having 3 genres have performed better with good Metascore and larger count of votes as well



Correlation Heatmap (1)

•There seems to be good degree of correlation b/w Rating & Metascore

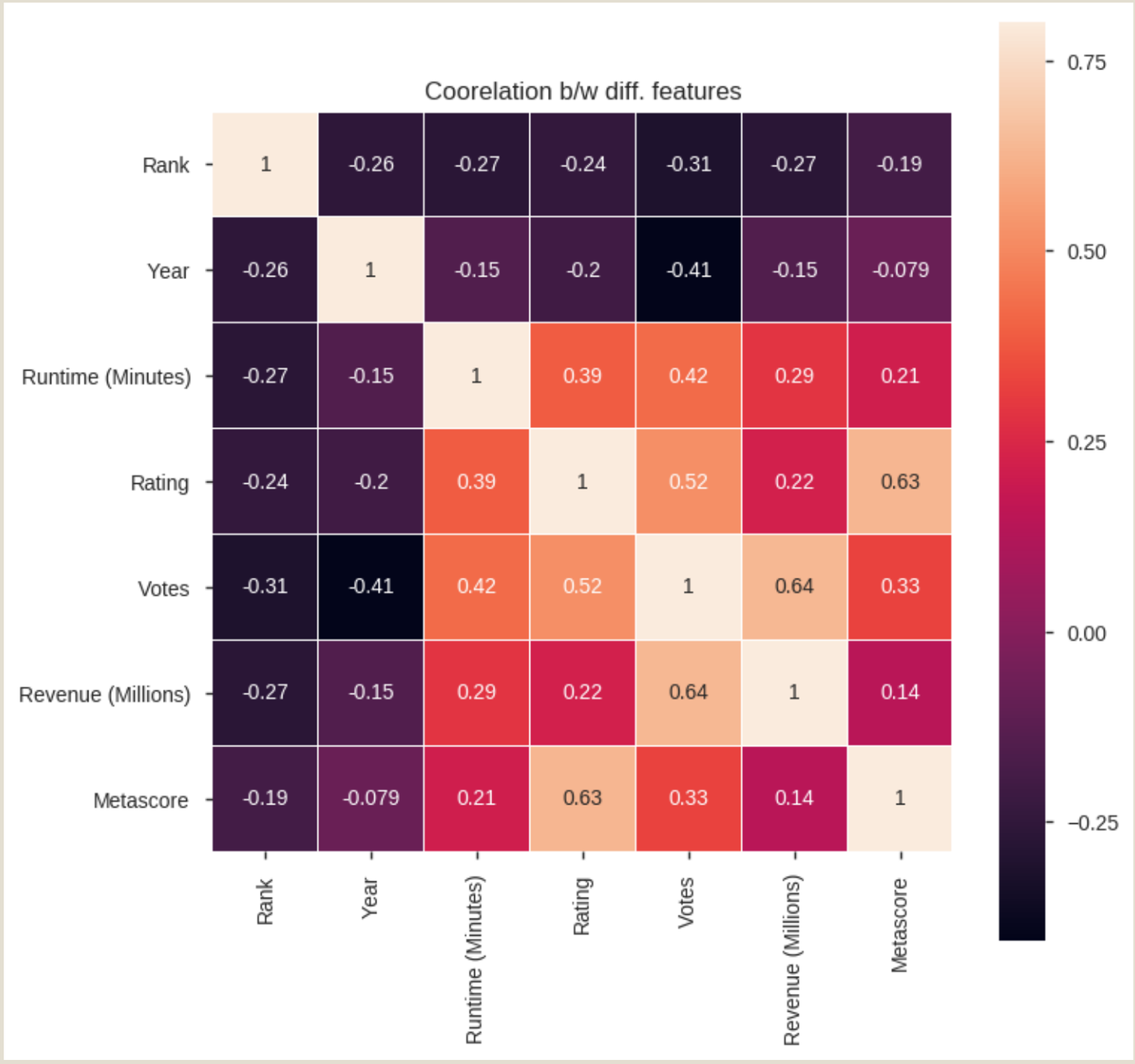
•Movies rated higher have earned more revenues

•People have voted more for movies with high Runtime

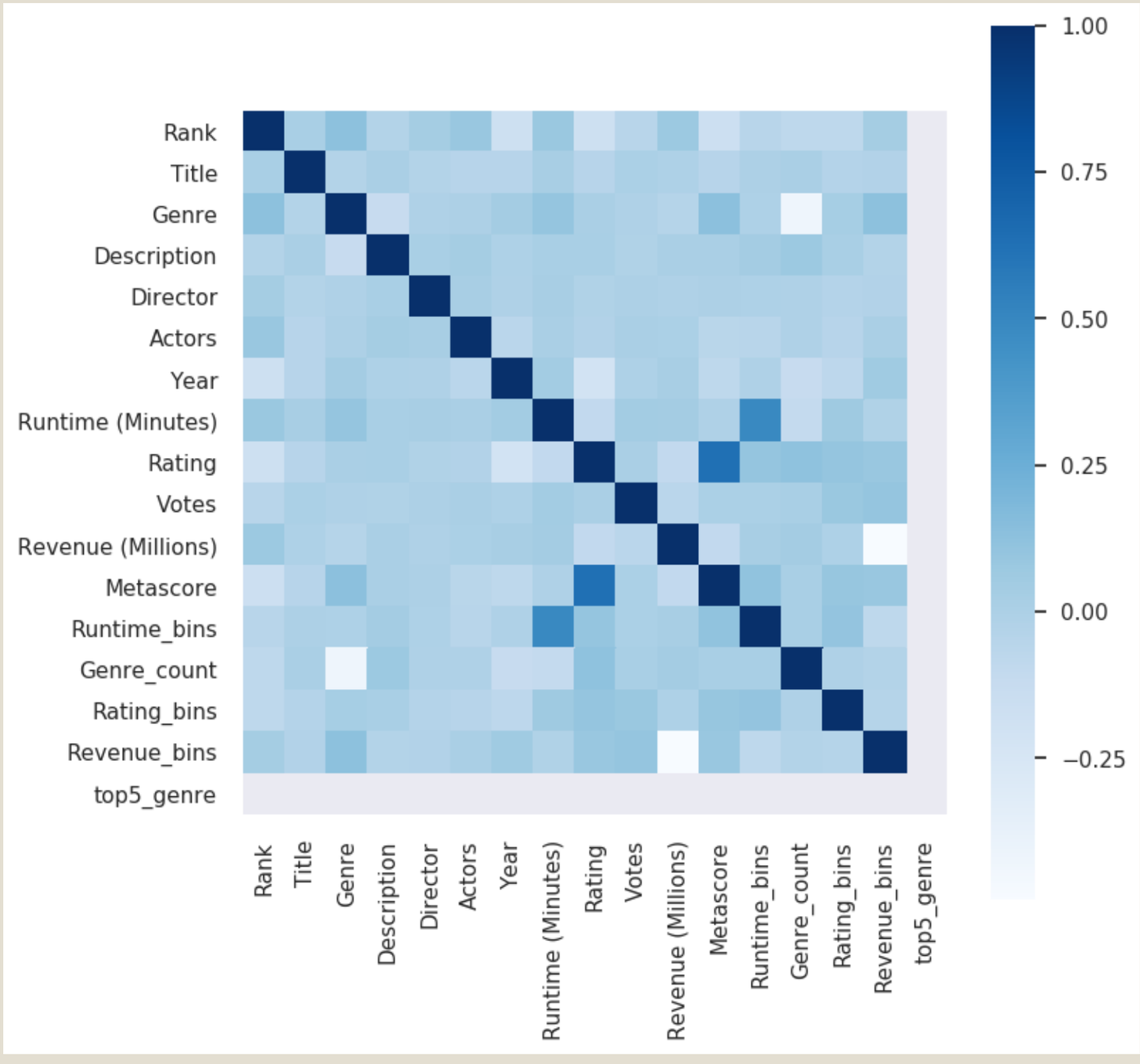
•Higher Runtime means better rating as well and higher earnings

•Votes are directly proportional to movie rating

•Votes, Rating, Revenue, Runtime, Metascore have direct correlation with each other though in different proportions



Correlation Heatmap (2)



Conclusion

Based on the data analysis done using Python as the programming language, below inferences can be summarized :

- Though the movies production has increased significantly, but the revenue, duration and user's rating have seen a downfall
- There is no direct relation b/w directors producing more number of movies vs their revenue earning
- User rating and critic's score stand in equilibrium for top rated directors
- Though there is similar ratio of movies by duration count, still longer duration movies have much better chances of generating positive ratings/score and higher revenue
- 4 genres vis-à-vis Adventure, Action, Sci-Fi & Drama have been liked by audience more than any other genre
- A combination of 3 genres (consisting above categories) attract larger people to the cinema halls thereby create more profitable business for producers, actors and directors
- Movies with more genres and good critics have high earning potential
- During 2010 to 2014, quality of movies were far better than any other years
- Critics have maintained a steady judgement over given timeframe and on an average 60% is their verdict score for most of the movies

Recommendations

Based on the findings from EDA would like to recommend below:

- Top directors with proven track record should be pitched in while going for any new movie production
- Movie duration should be maintained >2 hrs. for better results and audience appeal
- Irrespective of hovering around multiple genres, it's best to stick to combo of selected few one's having more customer liking
- User's ratings/votes are influenced by critic's review during any new movie release and therefore feedback shared by them should never be ignored while writing/ directing/producing new movies
- Focus should be more on quality of movies rather than the count in order to make market credibility and audience confidence

Suggestions welcome.. ☺

Thanks!

You can find me in LKG Medium of DS:

https://www.linkedin.com/in/goelsharad/

https://www.kaggle.com/sharad1501

https://github.com/shargoel

https://medium.com/@goelsharad15