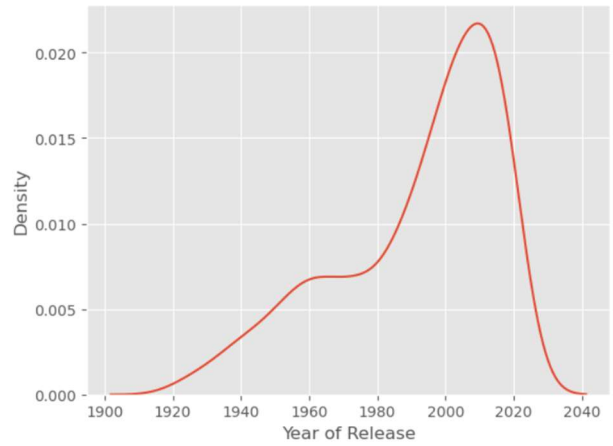
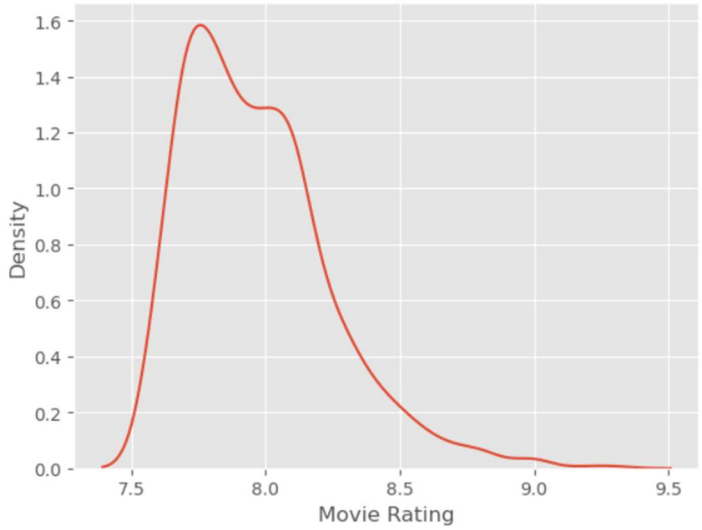
Problem 2 – probabilities (40%)

Next, you are to generate visualizations of various probability distributions (univariate, multivariate), and interpret them.
You can use any of the visualization methods presented in the lectures (and reading lists.)

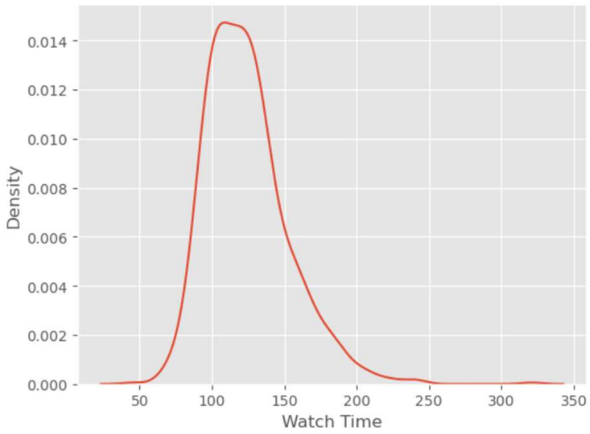
Univariate Visualization



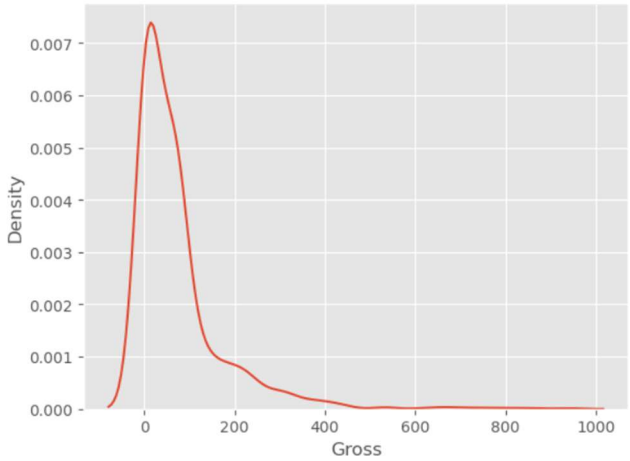
The above density plot for Year of Release appears to be left skewed, which tells me that the mean for Year of Release in terms of movies is less than the median. I find this interesting, as this suggests that in recent years, there was a spike of movies released that made it to IMDB's top 1000. More specifically, there was a lot of movies produced between the years 2000 to 2023 that made the charts. I wonder if this is because of movie release trends, as when I was reading over an article that argued that older movies receive higher ratings because they are "quality" compared to newer movies, which are comparable in quality. Medium Article for reference.



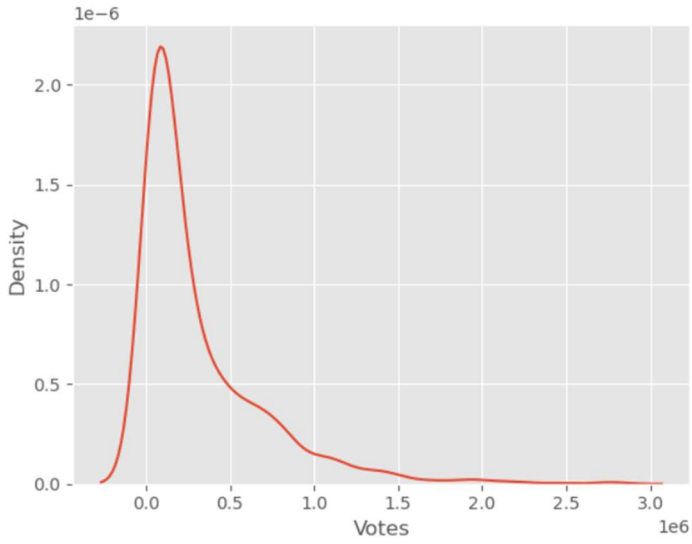
The above density plot for Movie rating appears to be right skewed and with two peaks. This tells me that a lot of movie ratings for the top 1000 IMDB movies tend to concentrate around 7.75 and 8.2. This also tells me that the mean is greater than the median for Movie Ratings. I find this interesting because since the dataset is IMDB's top 1000 movies, I expected higher ratings.



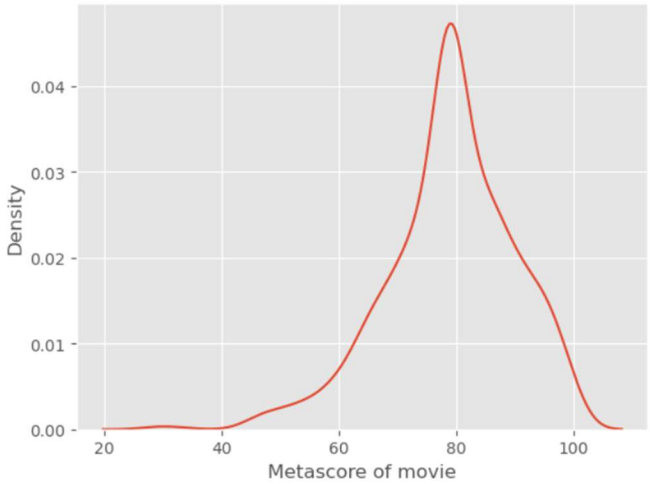
The density plot for Watch time appears to be right skewed with one peak. This tells me that most of the movies are concentrated with watch time of about 118 minutes (estimate) and that the mean is greater than the median for Watch Time univariate statistics. I find this interesting as I feel like this is generally average for most movies. I assume that this is because this is a dataset about movies, but even so, finding some movies with watch times of around 28 minutes to 5 hours makes me wonder about what defines a movie.



The density plot for Gross appears to be right skewed with one peak. This tells me that most movies have Gross concentrated around 60-70, and that the mean is greater than the median for Gross univariate statistics. This is interesting to me, as I feel like the plot shows that there is a big range for gross in terms of movies. Perhaps a successful movie is not determined by gross, as I feel like the range for the top 1000 movies is very wide.



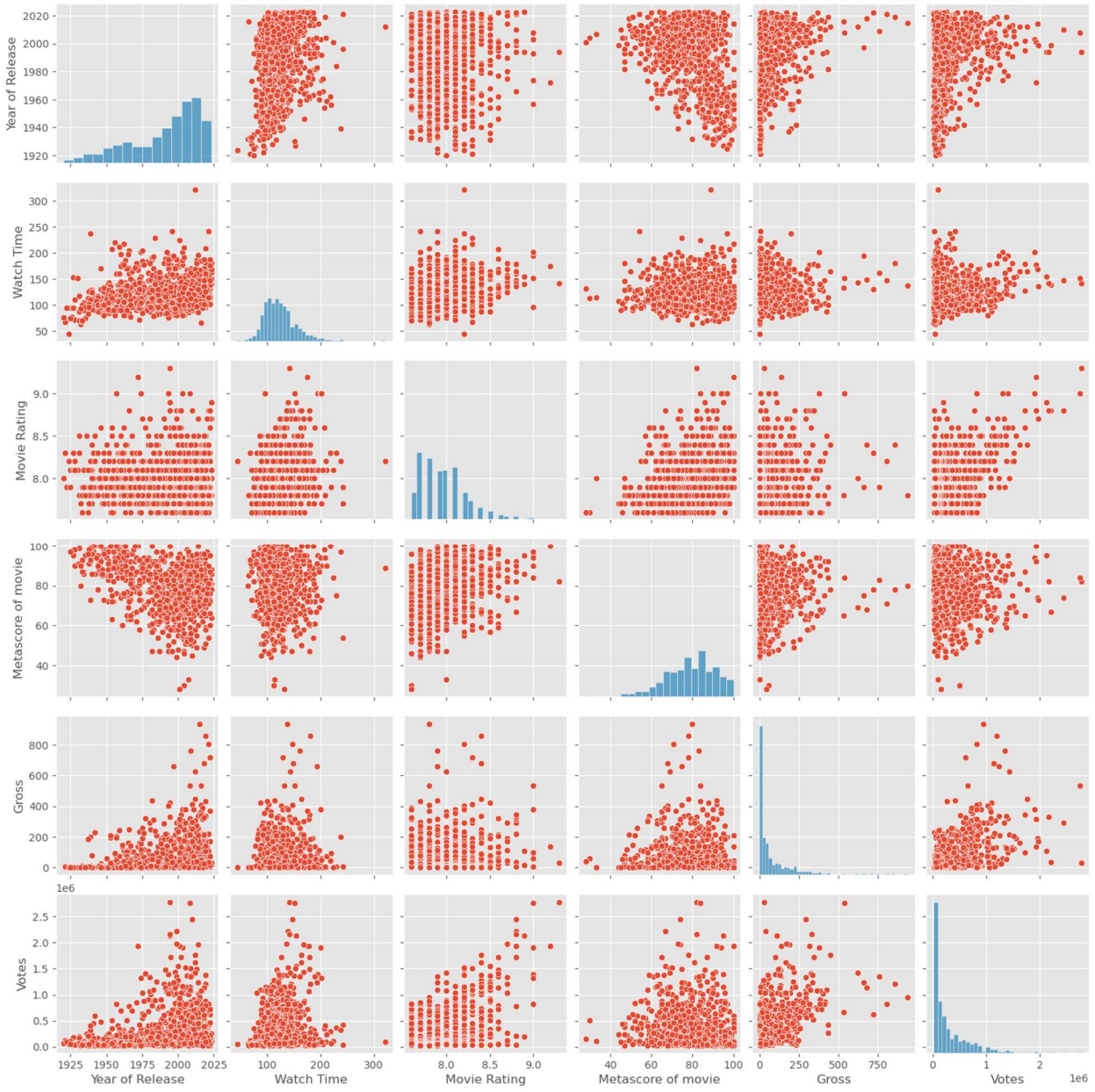
The density plot for Votes appears to be right skewed with vote concentration being around 250,000. This is interesting to me, as I feel like the range for votes is pretty wide, and that there are some extreme values on the long tail of the distribution. This also tells me that some movies may have a significantly larger amount of votes than others.



The density plot for Metascore of movie appears to be left skewed, with the Metascore concentration being around 80. This tells me that a large amount of movies in the dataset have Metascores of 80, which makes sense to me, as it means that most of the movies are positively received by audiences. This aligns with the fact that the dataset includes the top 1000 movies.

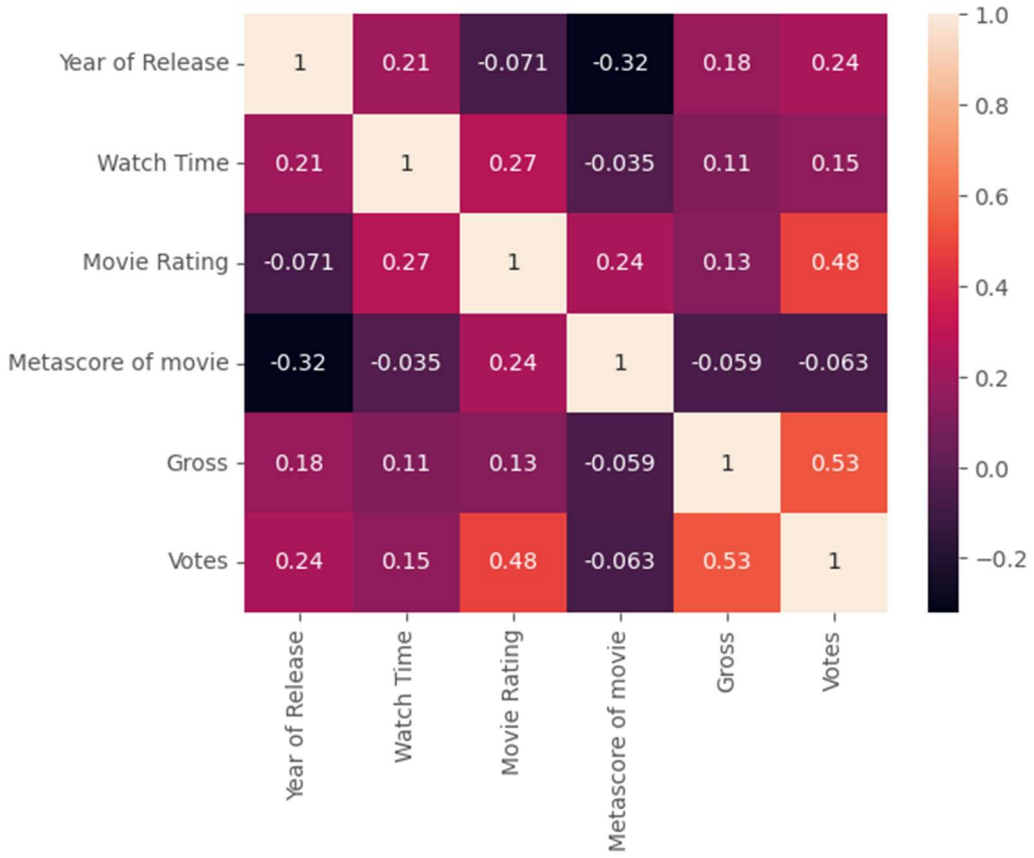
Multivariate Visualization

Multivariate examples are scatter plots, contour plots, scatter matrix

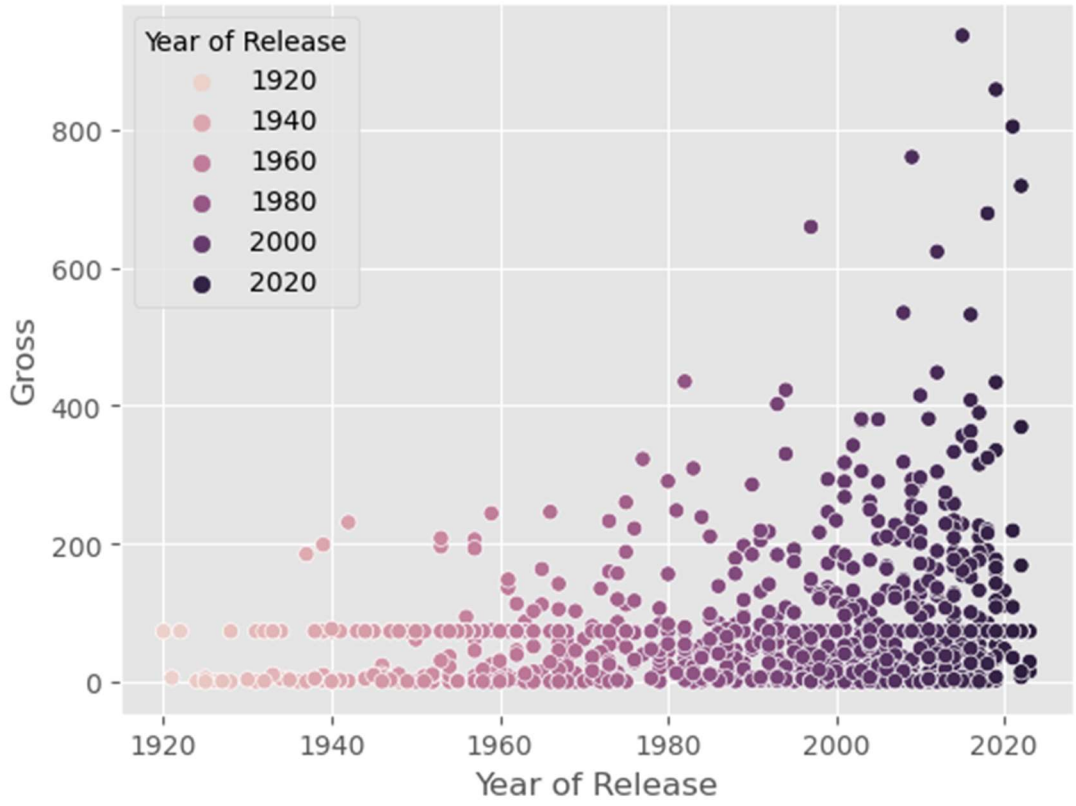


This is the overall pair plot for all the columns in the dataset, which I used to visualize correlation.

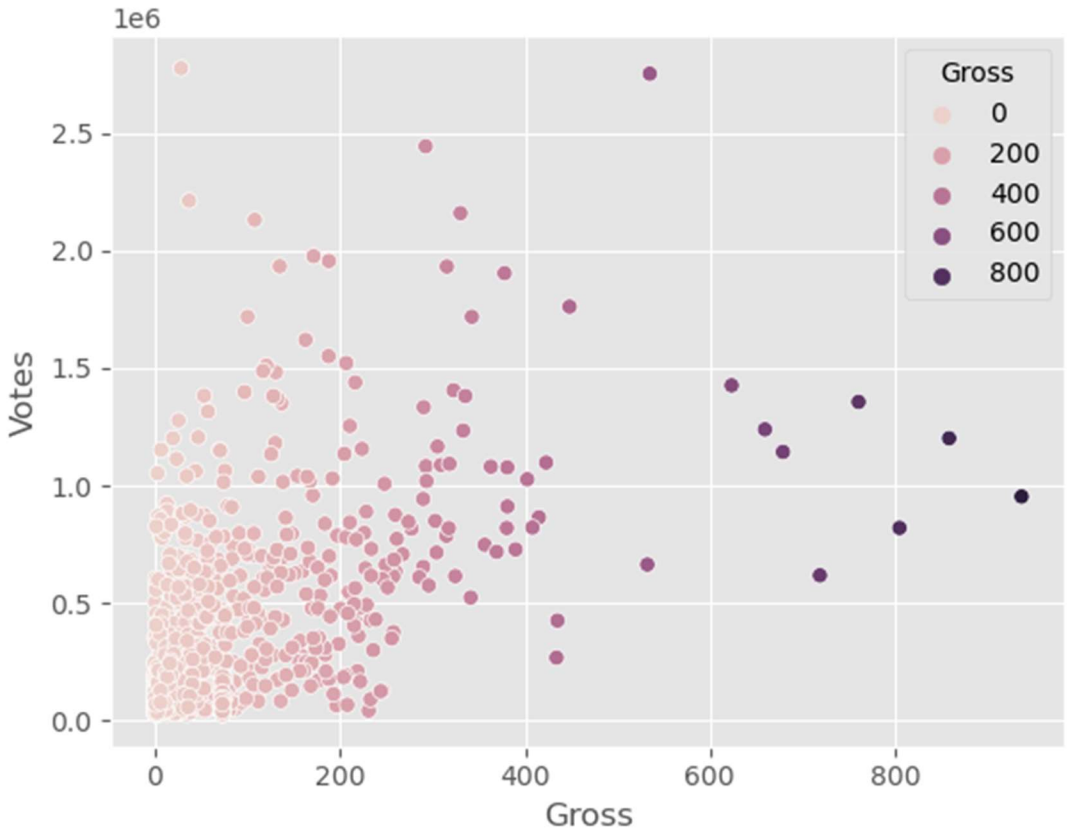
```
In [34]: sns.heatmap(df_corr, annot=True)
Out[34]: <AxesSubplot:>
```



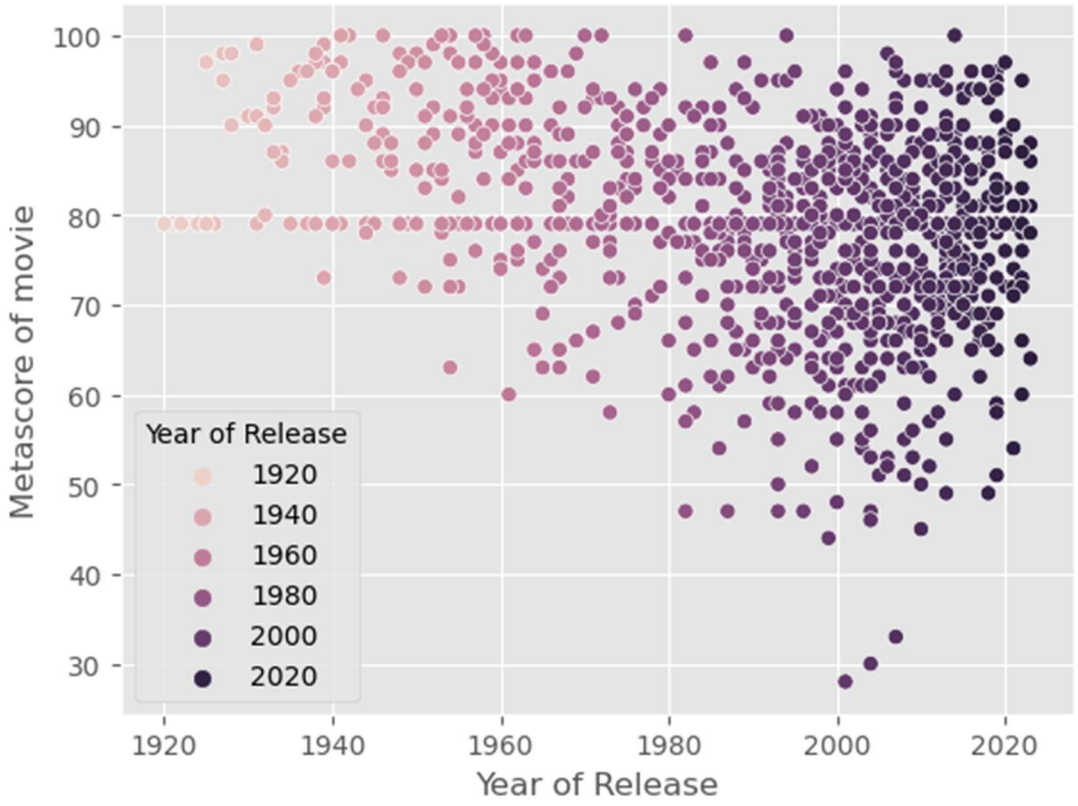
In addition to this, a visualization of correlation from P1 is provided in a heatmap. In the case of multivariate observations, I will observe the more notable two variable relations with moderate or small correlations, such as Movie Rating and Votes, Gross and Votes, and Year of Release and Metascore of movie. For the sake of visualization, all the plots are shown below, but the three mentioned in this paragraph will be the only ones that are gone into depth with explaining.



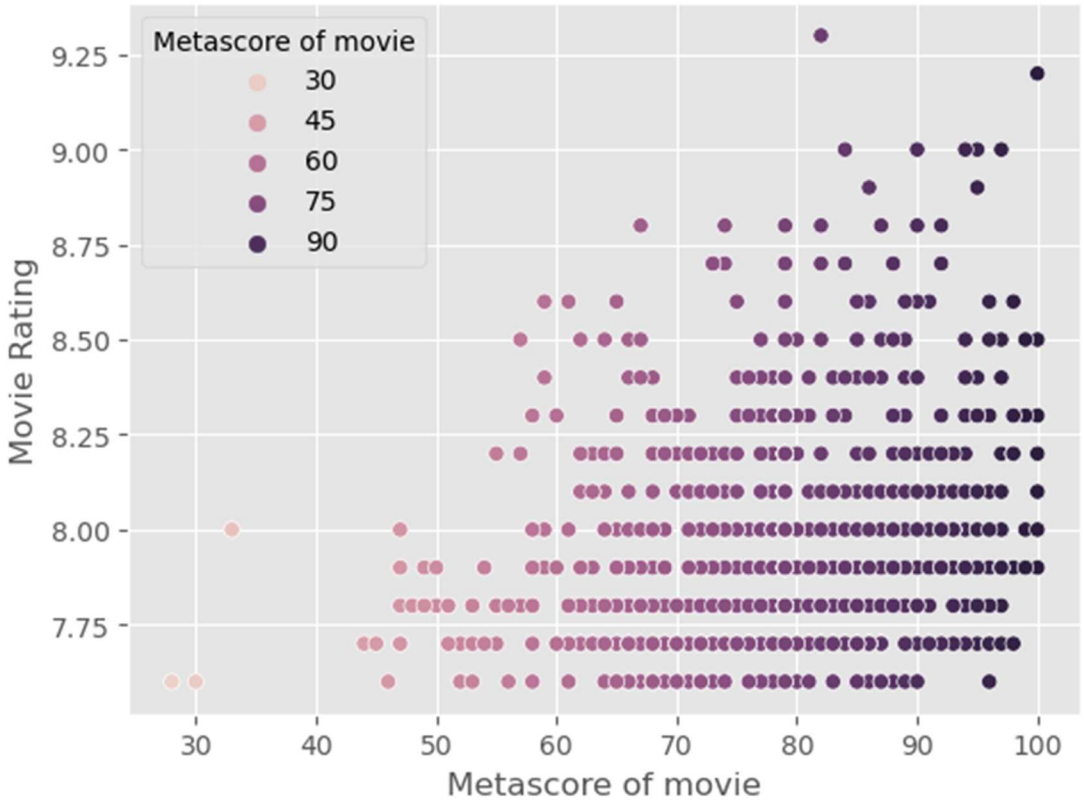
For the scatter plot multivariate visualization between Year of Release and Gross, it appears that there is a growing trend for movies to be increasing in Gross. Whether this be because of more people being able to watch movies or inflation, the maximum gross per each year appears to be on the increase.



This visualization between variables Gross and Votes interested me as it appears that there is a significant amount of clustering in the left most corner. I'm not sure if I can determine a positive correlation or not, but the correlation chart says that there is moderate positive correlation between the two variables. Most notably, the clustering in the bottom left corner appears to be the most notable, and I suppose that it makes sense, because higher Gross may encourage more audience members to watch a movie, consequently resulting in higher votes.



There appears to be negative correlation between the two variables Year of Release and Metascore of movie, as when we observe the Metascores of movies as the years progress, we notice that the minimum Metascore for each year gets lower and lower. I wonder if this is because of bias regarding older movies and the idea of a movie being a “classic” makes people remember movies to be better than they might have been initially received.



In this scatterplot between Movie Rating and Metascore of movie, there appears to be a positive correlation between the two variables. This is because as the Metascore of movie increases, it appears that the maximum movie rating for a given Metascore increases. Perhaps this is a sign that Metascore has an effect on Movie Rating, as it might be a determining value if an audience will receive a movie well.