Documentation of Development

**Finding a Dataset**

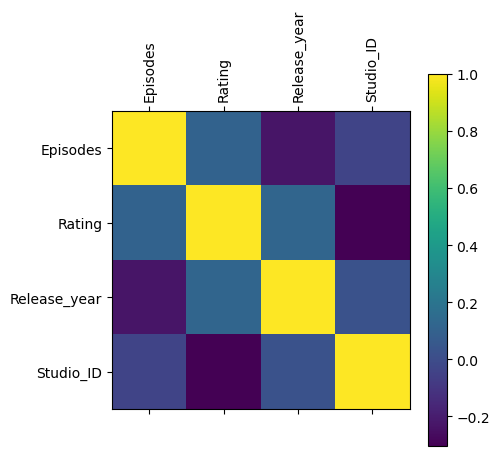
The first step of this project was to find a dataset I could use and would be interested in finding out some information form it, so I went to Kaggle and started looking for datasets that fall align with my interest, and so I started with a dataset about games, upon discussing with my lecturer I found that the dataset would not work, so I started looking for a different data set and, so I started looking at a dataset for anime and I found one I liked and had a good amount of data so I downloaded it and started working.

**Data Manipulation and Cleaning**

The next step was to import it into my code and start doing a deeper inspection into the data and clean it, so the first step was to import the data and print out a sample to ensure that the data was being imported properly, and when it was, I moved on to cleaning, at this point I already knew what information I wanted to find out and that was which production studio on average products the anime with the highest ratings, so with this in mind I started with the data cleaning and preprocessing, to so this I printed out the names of all the columns and decided what I thought would be need for what I wanted to do, but first I needed to add one data column of my own which was a studio ID as all the studios were named and for the analysis type I am going with I can not use strings, then once I had the ID for all the studios, I created a csv file to store the data so I could get the name of the studio later. Following this I started to remove the columns of data that I didn’t need those being End\_year, staff, Voice\_actors, Description, Name, Japanese\_name, Type, Tags, Related\_Mange, Related\_anime, Rank, Release\_season, Studio, and Content\_Warning and then dropped all the remaining rows that had missing data. At one point I did remove the Release\_Year but I was told I could maybe find out some interesting information from it so I added it back in.

**Correlation and Multivariate Analysis**

Following this I moved on to my correlation matrix to see how the parts of the data correlated with each other. Which can be seen on the correlation matrix this was done so I could have a rough idea of how much each of the variables relate to each other either positively or negatively and from what I was able to gather from this is that none of the variables correlate all that closely with each other, but is must be remembered that correlation does not imply causation, therefore other factors could be affecting the correlation. Following this I created scatter plots that would give me a visual idea of how the variables are interacting with each other and the correlations are about the same as in matrix but with this we can draw more information such as in the first plot we can see that longer anime typically were less likely they were to have ratings over 4, then in the second plot there is also a negative correlation which may be due to the people who made the list putting the production studios of the anime that on average did worse later in the list so their ID ends up being higher, then for the third plot, as years went on rating got higher either due to more people getting into anime or due to the fact that production quality has gotten higher and higher making people enjoy the anime allowing for higher ratings, in plot 4 we see a negative correlation which also could be due to how the anime industry has changed over time as more and more studios are producing 12-24 episode shows otherwise known as seasonal anime therefore in more recent times the episode counts that studios tend to have is lower, and in plot 5 we see something similar to was said in the last point about how in more recent times anime have become a lot shorter given a few outliers but this does still cause a negative correlation, and finally in the last plot we can see a positive correlation which makes sense as studios that have been around longer would have had more time to make anime then the new studios.



A graph with a red line and blue dots

Description automatically generatedA graph with a red line

Description automatically generated

A blue and red dotted chart

Description automatically generated A graph of blue dots

Description automatically generated

A graph with blue dots and red line

Description automatically generatedA graph with blue dots and a red line

Description automatically generated

Following this I got the PearsonRResult and the PValues relating to episodes and ratings, studio\_id and ratings, and the release year and rating and what can be drawn form them is with PearsonRResult being so low we at around 0.106 there is a weak correlation but with the PValue being so low at 3.88675987312317e-22 means that is statistically significance to it indicating that as the number of episodes increase there is a slight increase to ratings but as the graphs show us only to a point, then we can see a similar thing in in both the studio and the release year data with a low correlation but with significant PValues with the studio and ratings as it is 2.0459150762392743e-176 which is significant we can also observe that with a PearsonRResult of- 0.3044476447633157 we can now place a number to the negative correlation in the graph and finally as stated before there is also a low PearsonRResult once again showing a weak correlation but is also statistically significant as shows that newer anime tend to have higher ratings.

After this I looked at OLS regression results and as we can see in the first OLS that for each new episode of a show the ratings tend to increase by 0.0012 units, but the low R-squared value implies that while the relationship is statistically significant, the number of episodes is only explaining a very small portion of the information in the overall variation of rating being around 1.1%, then when looking the studio OLS we can see that the higher the studio\_id the lower rated the show will be which can likely be explained as the way the dataset creators made the dataset and then when looking at the R-Squared we can see that around 9.3% of the variation in the rating of anime can be based on the studio that makes it and then I checked the release year OLS was found that the as the years have gone on they contribute about 0.0042 units per year and that the R-Squared suggest that about 1.4% of variation can be explained by the release year lastly I did an OLS with all the data to see what I would get and I ended up getting very low coefficients for all the variables with them all being near zero, and there is likely an error somewhere with this OLS as the R-Squared is 1 implying that this data perfectly describes all the variation in the rating which from the other OLS regressions I know is not true. I then looked at getting the mean squared error, but I was getting nearly the same values as the last OLS meaning whatever the errors are, they present in both.

**Normalization, Weighting and Aggregation**

Following the main part of the multivariate analysis I moved on to normalization, weighting and aggregation I started by doing the normalization, which takes all the rating values and scales the values to be a range of 0 to 1 to ensure uniformity and comparability across the data after this I weight the data giving rating the highest weighting of 0.6 and gave episodes and release year both a rating of 0.2 each then for Aggregation I simply found the mean or average of all the data newly weighted data for each studio and printed that data.

**Getting composite score**

Finally, I need to get the composite scores and final index of my analysis, So first I added together all the weight values to together and printed out all the results with the composite scores, Then I found the top 10 studios and in doing this I also found the number 1 studio which I printed out under the scatter plot and bar chart I then printed out the scores of the top studio, and was able to find out that the studio with the best rating on average rating is Creative Power Entertaining, which was a studio I had never heard of so I looked into them, and found out that they are a Chinese animation company that have only ever made one series called “Pleasant Goat and Big Big Wolf” and is one of if not the biggest children’s TV show in China with 2789 episodes all being around 13 minutes long, but I do this that the amount of episodes in this show really played a role into why it was ranked first as even at a 0.2 weighting the amount of episodes made the weighted episodes value very high which likely throws off the data.

A graph with numbers and lines

Description automatically generatedA graph with numbers and dots

Description automatically generated

**Extras**

When discussing my project with my lecturer he pointed out that there could be some interesting information and so I decided to investigate it and that was the trend of how there seemed to be a lot of shows that did not have that many episodes in more recent years so to find out information on this I made a K-means cluster scatter plot with 3 clusters in it dividing up the data based on the amount of episodes they have, and then found the amount of anime in the first cluster that were released after 2015 and found out that out of the 7041 anime in that first cluster 2762 of them or a bit over 1/3 of all of them were released in or after 2015 meaning that in the 65 years or so before then that my data has were less then double the amount of the anime released in the past 9 or so years. Which as I stated before makes sense as in more recent years anime have become something known as seasonal mean they are typically only about 12 episodes where as in the past anime such as Naruto was able to have 700 episodes released almost weekly as well as One Piece which has over 1100 episodes which is something that just does not happen today.