DS 2500 - Data Wrangling - Spring 2023

# MyAnimeList data analysis project

Hung Phan Quoc Viet

School of Computing, University of Utah

April 26, 2023

## 1 Introduction & Argument

Anime, or Japanese animation films has been so far one of the most popular animation at the meantime, with its fan base over 120 million users on Crunchyroll [1], and some interesting facts such that the US is the top 2 country where anime is the most popular [2], just behind Japan, where anime is originally from. Anime has such a big impact on popular culture. On the one hand, good anime can bring deep sorrowness, healing, and different kinds of emotions that can affect an individual to change themselves. On the other hand, anime is a way to introduce the nation's culture to the world, which helped to foster cultural exchange between Japan and other countries. One example is how many anime have appeared on Netflix nowadays, and got subtitled or dubbed in other languages for international distribution. Of course with a wide popularity and an enormous anime database like that, the quality of anime also varies. There are remarkable works such as Studio Ghibli's movies including My Neighbor Totoro, Spirited Away, Shinkai Makoto's works including Your Name, and other good, medium, bad anime. There are many factors that can affect anime quality. With this in mind, how do some basic metrics relate to the quality of an anime? Or to better rephrase this, how do some criterias include the rank, popularity, number of episodes, where does the anime originally from (e.g. original plot, from manga, light novel, etc.), what is the type of the anime (e.g. TV series, movie, etc.), affect the score of an anime? Looking in a statistical perspective, what does the correlation and causation relationship look like between these criterias and the score?

## 2 Dataset & Preprocessing

The dataset being used here is the MyAnimeList animelist.csv dataset, which was downloaded from Kaggle [3]. MyAnimeList is an anime database, so it contains all the basic description of an anime such as airing date, title, score, popularity, ranking. Therefore, all necessary variables are also contained in this dataset. The variables we are looking at in this analysis are continuous variables: *rank, episodes, popularity*, categorical variables: *source* and *type*, and the target variable: score. For convenience, I will call the target variable as the output, and others are input.

The data comes in fairly messy with a lot of missing data, at random places, with 31 columns in total. The first thing to do is to filter the most important variables described above, to not only make the dataset easier to read, but also easier to clean the data and minimize the accidentally missing rows because of missing data in other columns.

The most optimized way to clean data in this case is to drop all the missing values, this is because most missing values in this dataset is *rank*, and every rank is unique. As a result, if we impute the rank by using methods such as taking the mean, regression, it would create an inaccurate analysis. Another thing to notice is that in the *source* variable, apart from values like "Original", "Manga", or "Game", there is a variable "Unknown", that represents the missing values. With the same explanation with *rank*, these "Unknown" values also be dropped.

## 3 Methods

There are two main parts in the analysis process, the first one is looking at the correlation relationship, specifically looking at how the input relates to the output, which uses Exploratory Data Analysis (EDA) as the main tool. The second one is looking at the causation relationship, where we are examining how the input causes output, using OLS regression.

## 3.1 Exploratory Data Analysis

The general outline of the EDA process is first, examine the distributions, the count of variables, next do statistical tests for input against output, and finally visualize the correlation result.

### 3.1.1 Univariate visualizations

When discussing the first part, there are a total of 6 variables. However, only the score, type, and source are relevant for exploration, as the other variables, rank and popularity, consist of unique values. The plots being used in this part are histogram, and count plots. Although there may be cases where some anime share the same rank and popularity, these are only rare occurrences that would cause the x-range for the plot, such as a histogram, to become excessively large and difficult to read. Thus, there is no need to plot such complicated graphs, and this also applies to episodes.

### 3.1.2 Statistical tests

Talking about the statistical testing, the goal here is to test the correlation between each input with regard to the output. With that in mind, the tests being used are Pearson's correlation coefficient for testing the continuous inputs against the output, and the ANOVA test for categorical inputs against the output. To perform these tests, I used the SciPy library for the Pearson's correlation coefficient and statsmodels `anova_lm()` for ANOVA. SciPy also provided a function for ANOVA test, however it is not that interpretable when put into use, and `anova_lm()` performance showed that it is a good substitution.

### 3.1.3 Bivariate visualizations

Regarding the final visualization part, there are three aims here. First of all, is to explore the correlation between each variable, specifically the numerical, continuous variables, and heatmap was used to achieve this aim. Second of all, is to visualize the correlation between the continuous inputs with the output, which is done by using scatterplots. Finally, is the visualization of the categorical inputs with regards to the output, for which the categorical plot is an effective tool.

## 3.2 OLS Regression models

The aim for this part is looking at how the causation relationship looks like between the inputs and the output. After the EDA process, it is found that the rank and popularity are autocorrelated to each other, which violates the assumption of regression models. However, there is only one autocorrelation relationship, and that autocorrelation may not have a significant impact on the model in general.

To examine this, I built 2 groups of OLS models, the first group is without dealing with the auto-correlation and the second one is dealing with that using *Principal Component Analysis (PCA)*. Each group contains 2 models, with the first one not including the categorical variables, and the second one including all inputs. The reason I did not build a regression model that only includes categorical inputs, is that although they might be statistically significant, when put into a model, they are not as straightforward as the continuous variables. To investigate the performance, 3 metrics will be used here, including the *R-squared score*, *Akaike Information Criterion (AIC)* and *Bayesian Information Criterion (BIC)*, as they are fairly comprehensive to use.

# 4 Results

I am dividing the result into three subsections, with two showing the result for correlation and causation relationships, and one discussing any additional findings that are interesting to see.

## 4.1 Correlation relationships

### 4.1.1 Statistical tests

Initially, let us take a concise glance at the outcomes of the statistical tests:

```
rank against score:
Correlation: -0.8408879546529529
P value: 0.0

popularity against score:
Correlation: -0.7464144769017501
P value: 0.0

episodes against score:
Correlation: 0.11546623707857588
P value: 4.923739841775676e-28
```

*Figure 1: Pearson's correlation coefficient test results*

From the Pearson's correlation coefficient, we can see that there is a high correlation between *rank* and *popularity* with *score*. Specifically, at a value of approximately -0.84 and -0.75 respectively, *rank* and *popularity* implies a strong negative correlation with *score*, which can be explained that as the value of rank and popular-

ity decrease, the score will increase. In a more comprehensive saying, as the anime rating and popularity is higher, the score is higher. And the p-value at 0.0 for both of the variables also implies that these correlations are statistically significant.

Another interesting result to see is that while *episodes* showed a weaker correlation compared to two discussed variables, its significantly low p-value compared to 0.01 or 0.05 implies that this correlation is statistically significant, which means that this correlation relationship cannot be ignored.

score in relation to type

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(type) | 2591.565969 | 5.0 | 405.766742 | 0.0 |
| Residual | 11463.094718 | 8974.0 | NaN | NaN |

score in relation to source

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(source) | 3041.528843 | 14.0 | 176.849306 | 0.0 |
| Residual | 11013.131844 | 8965.0 | NaN | NaN |

*Figure 2: ANOVA test results*

The F score and the p-value interpreted pretty much the correlation relationships between *type* and *source*, which means that there is a strong correlation between type and source to the score. Therefore, whether the anime is a TV series or a movie, comes from a light novel or an original plot, all relate to how the score behaves.

### 4.1.2 Visualizations

Now let us have a look at how those statistical tests would look visually.

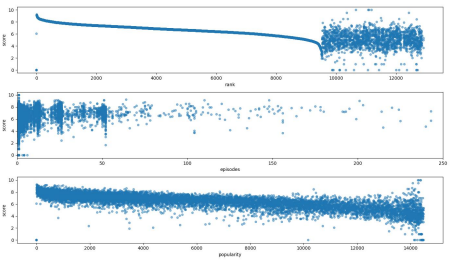

*Figure 3: Each continuous variable with regards to score*

In Figure 3, it is showing the relationship between each continuous variable (x-coordinate) against the score (y-coordinate), with the order *rank*, *episodes*, *popularity* consecutively. Looking at the figure, we can see that the result is same like what we discuss in statistical testing, where the *rank* and *popularity* show an obvious linear trend, despite that as the rank gets larger, the trend seems to be ambiguous, this might be due to the some noises in the data.

Figure 4 showing the relationship between type, and source (x-coordinate) respectively to

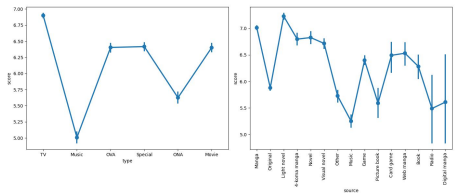

*Figure 4: Each categorical variable with regards to score*

the score (y-coordinate). While the figure is small and hard to see, here is some notes to make from that.

Regarding the type, TV series anime seems to provide the best quality to the audience, with the movie successively, and music videos anime would have the worst quality. This can be due to the lack of storytelling as a music video would only lasts for a few minutes.

Regarding the sources, the quality anime usually comes from manga or light novel, rather than the original plot, which has a far less preference compared to manga or light novel. Anime with its plot comes from game, or web manga, has a mid preference, not as quality as the light novel or manga, but better than the original plot.

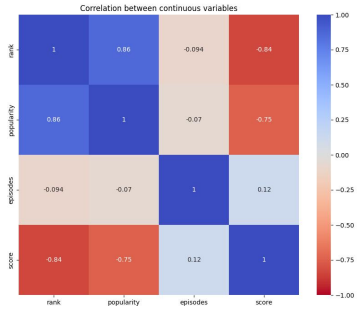


*Figure 5: Correlation between continuous variables*

Apart from the discussed correlations, there is another correlation between *rank* and *popularity*, which is considered as an auto-correlation that might be a problem when making an OLS model.

## 4.2 Causation relationships

Like discussed in the methodology, we are examining 2 groups of OLS models, one with and one without the solution to auto-correlation. However, after building the model, the result came out to be the same between the two. This means that the auto-correlation relationship does not

have a significant impact to the model in general. And so, I will not include the result for both groups here, but rather just the result for the group without the solution to the auto-correlation, as this group is simpler and does not require to deal with auto-correlation but still have a fair performance, therefore it is chosen over the other group.

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | score | **R-squared:** | 0.710 |
| **Model:** | OLS | **Adj. R-squared:** | 0.710 |
| **Method:** | Least Squares | **F-statistic:** | 7341. |
| **Date:** | Fri, 21 Apr 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 03:08:31 | **Log-Likelihood:** | -9188.5 |
| **No. Observations:** | 8980 | **AIC:** | 1.839e+04 |
| **Df Residuals:** | 8976 | **BIC:** | 1.841e+04 |
| **Df Model:** | 3 | | |
| **Covariance Type:** | nonrobust | | |

*Figure 6: OLS Regression with continuous inputs only*

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | score | **R-squared:** | 0.725 |
| **Model:** | OLS | **Adj. R-squared:** | 0.725 |
| **Method:** | Least Squares | **F-statistic:** | 1075. |
| **Date:** | Fri, 21 Apr 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 03:08:42 | **Log-Likelihood:** | -8951.2 |
| **No. Observations:** | 8980 | **AIC:** | 1.795e+04 |
| **Df Residuals:** | 8957 | **BIC:** | 1.811e+04 |
| **Df Model:** | 22 | | |
| **Covariance Type:** | nonrobust | | |

*Figure 7: OLS Regression with all inputs*

From Figure 6 and Figure 7, we can see that the R-squared score increases from 0.71 to 0.725 as we include categorical variables, which means that the model fits the data more with these inputs, and the decrements in AIC and BIC values also strengthen that argument. Therefore, an OLS Regression model, with all 5 inputs, will create an output that fits 72.5% to the data, and that there is a 72.5% accuracy in terms of causation effect.

### 4.3 Additional findings

Now we have discussed about the correlation and causation results. Other than that, there are also some more interesting findings, including the auto-correlation relationship between *rank* and *popularity* and details relating to the source and score.

Let us have a look first at the count of *type* and *source*:

Again it is hard to look clearly at the plot (figure 8), but the general thing is that, anime that have the original plot accounts for nearly half the data. However, the quality contradicts with the amount of original anime, where the score is significantly low compared to plots that

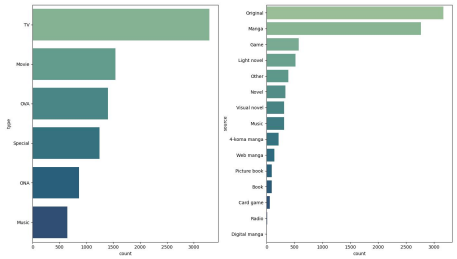

*Figure 8: Type and Source count*

comes from manga or light novel. Additionally, in contrast to the original plots, although there are only a few anime that come from light novel, the quality of those anime are far higher than original anime.

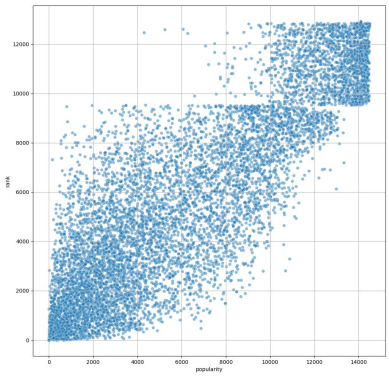

*Figure 9: rank and popularity relationship*

The scatter plot in figure 9 shows the relationship between rank and popularity. It is true that there is an underlying linear relationship between these two variables. However, it is not necessarily the case, since the graph shows a defined sparseness, which is caused by a relatively high volume of random values and noises. Therefore, it might be true that the rank and popularity are strongly correlated to each other, however, the figure is showing the sparseness of the values, which can explain why the auto-correlation is not significantly impactful to the OLS model that we built before.

## 5 Analysis limitations

So far, we have found out the correlation and causation relationship of the discussed variables. Throughout the analysis, there are some limitations that can be foreseen.

The first thing to talk about is the data itself. The data has a large amount of missing values, from actual missing values, to missing

values that have been replaced by some sort of replacements, such as the *source* missing data, which is replaced by "Unknown", and since the approach to handle missing data is to drop all of those, in total the number of rows left is 8980, which means that the original data reduced down nearly a half (raw data is 14477 rows). Therefore, doing this causes the analysis to be less accurate. However, with a situation like this data, it is inevitable to impute missing data without making it feel biased.

The second thing is regarding the dealing with auto-correlation. Since the models before and after PCA have exactly the same performance, it can be that the PCA has been done wrongly. Yet, with the existing knowledge, I tried to deal with it but did not have a strong foundation about PCA or any other auto-correlation solution technique before.

The third thing is about the OLS Regression. There is a limitation to this approach is that this type of model fits a straight line into the data, which would ignore many important information that seems random but they are not. Therefore, assuming that I have a better knowledge about different types of machine learning models, my approach to this would be putting different models into training, and then choose the best performance model.

## Conclusions

As of now, we have explored various aspects of how the quality of anime is determined. Our findings suggest that anime types, sources, rank, popularity, and episode count are correlated and causally linked to the score at a certain level. As previously mentioned, anime represents Japanese culture and has been remarkably successful over the years. While it is merely a type of animated film, with its beautiful artwork and stories, it has successfully promoted Japanese culture worldwide. Moreover, anime has a significant impact on individuals, influencing their behavior, mindset, and more. With the rise of social media platforms such as Tiktok, anime's popularity is growing and will continue to do so, creating a larger and stronger community in the future.

## References


[1] https://www.crunchyroll.com/anime-news/2021/08/03/crunchyroll-reaches-5-million-subscribers-announces-new-project

[2] https://www.epicdope.com/top-10-countries-where-anime-is-most-popular-and-why/

[3] https://www.kaggle.com/datasets/azathoth42/myanimelist?select=AnimeList.csv