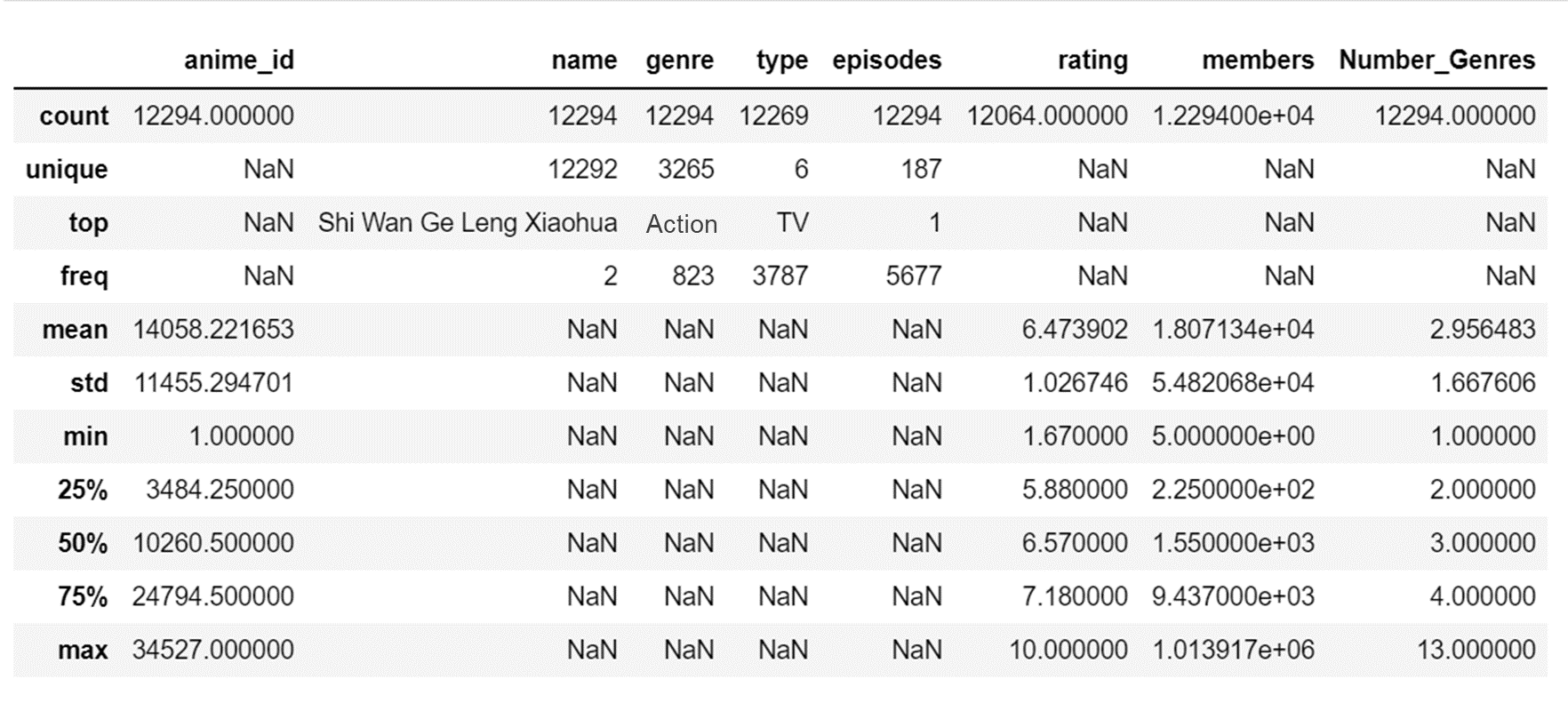


# Exploratory Analysis:

The second section of the assignment hopes to conduct extensive exploratory analysis to unearth statistically significant insights or relationships between the variables. The relationships will identify trends which lead to an anime’s success and identify similarity patterns of content consumption for several user groups.

## Basic Descriptors:

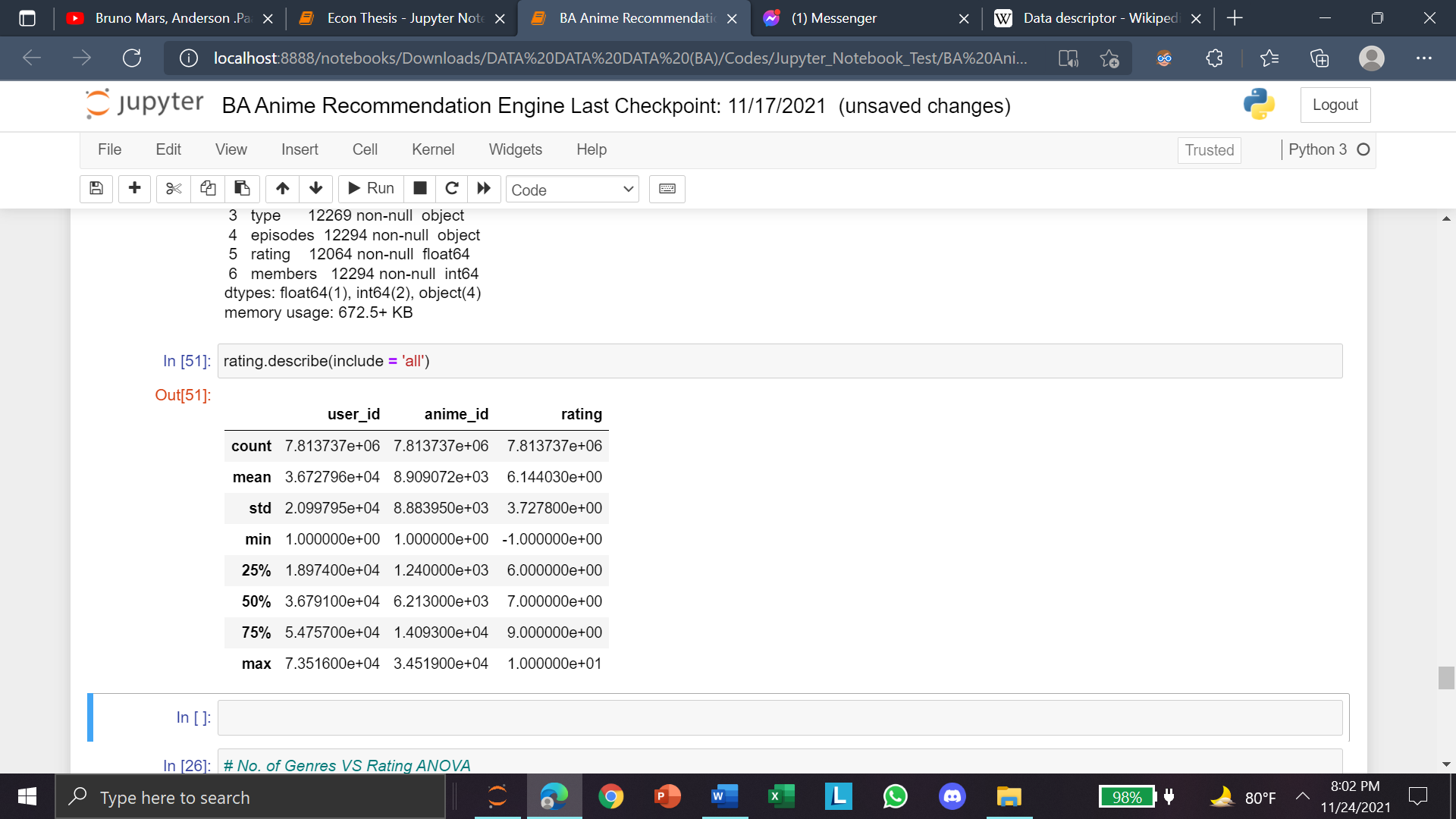
This section looks into the key basic data descriptors to get basic idea about the central tendency and distribution of the data. For the anime dataset containing the variables the descriptors for the variables are as follows:



For the output it can be interpreted that the anime\_id, rating, members and Number\_Genres variables are quantitative variables and the name, genre, type and episodes are the categorical variables. However, the anime\_id should be an object variable and episodes should be an integer variable, the change will be addressed in the preprocessing section of the study.

As per the basic descriptors it can be seen that the standard deviation for members section is extremely high compared to its mean. However, other than that no other variables have displayed out of the ordinary behavior.

For the ratings dataset containing the variables the descriptors for the variables are as follows:



From the output it can be identified that the rating variable has a minimum value of -1. However, practically a value of -1 is not possible and therefore, this value will be removed in the preprocessing section. The rows with ratings of -1 will be removed since even after the removal of the negative data rows, the dataset still has 6,337,241 data rows and removal of the negative rows in the ratings dataset does not impact the anime dataset in any way rather would reduce anomaly in the dataset.

## ANOVA:

### 2.1 Variation with Number of Genres:

In the visualization section it was seen that the ratings varied with the number of genres. To find out whether differences are statistically significant the following hypothesizes have been developed:

**H0:** There is no difference across mean ratings of anime with different number of genres.

**H1:** There is significant difference across mean ratings of anime with different number of genres.

As per the basic descriptors of number of genres in the earlier section, the minimum number of genres is 1 and the maximum is 13, creating 13 different groups.

|  |  |  |
| --- | --- | --- |
| Source Of Variance | Formula | Df Value |
| Among Groups | (k-1) | (13-1) = 12 |
| Within Group | (k-n) | (12294-13) = 12281 |

The sources of variance lead to a critical F value of 1.75 at a Alpha value of 5%. The following value have been calculated.

|  |  |  |  |
| --- | --- | --- | --- |
| **Calculated F Ratio** | **P-value** | **F-critical** | **Decision** |
| 2396.43 | 0.0 | 1.75 | Reject Null Hypothesis |

Since the p-value is less than alpha, the null hypothesis has to be rejected, suggesting that there is statistically significant difference across mean ratings of anime with different number of genres.

### 2.2 Variation with Type of Release:

In the visualization section it was seen that the ratings varied with the type of release. To find out whether differences are statistically significant the following hypothesizes have been developed:

**H0:** There is no difference across mean ratings of anime with different types of release.

**H1:** There is significant difference across mean ratings of anime with different types of release.

There are total of 6 unique types of release which results in the following values of sources of variance.

|  |  |  |
| --- | --- | --- |
| Source Of Variance | Formula | Df Value |
| Among Groups | (k-1) | (6-1) = 5 |
| Within Group | (k-n) | (12294-6) = 12288 |

The sources of variance lead to a critical F value of 2.21 at a Alpha value of 5%. The following value have been calculated.

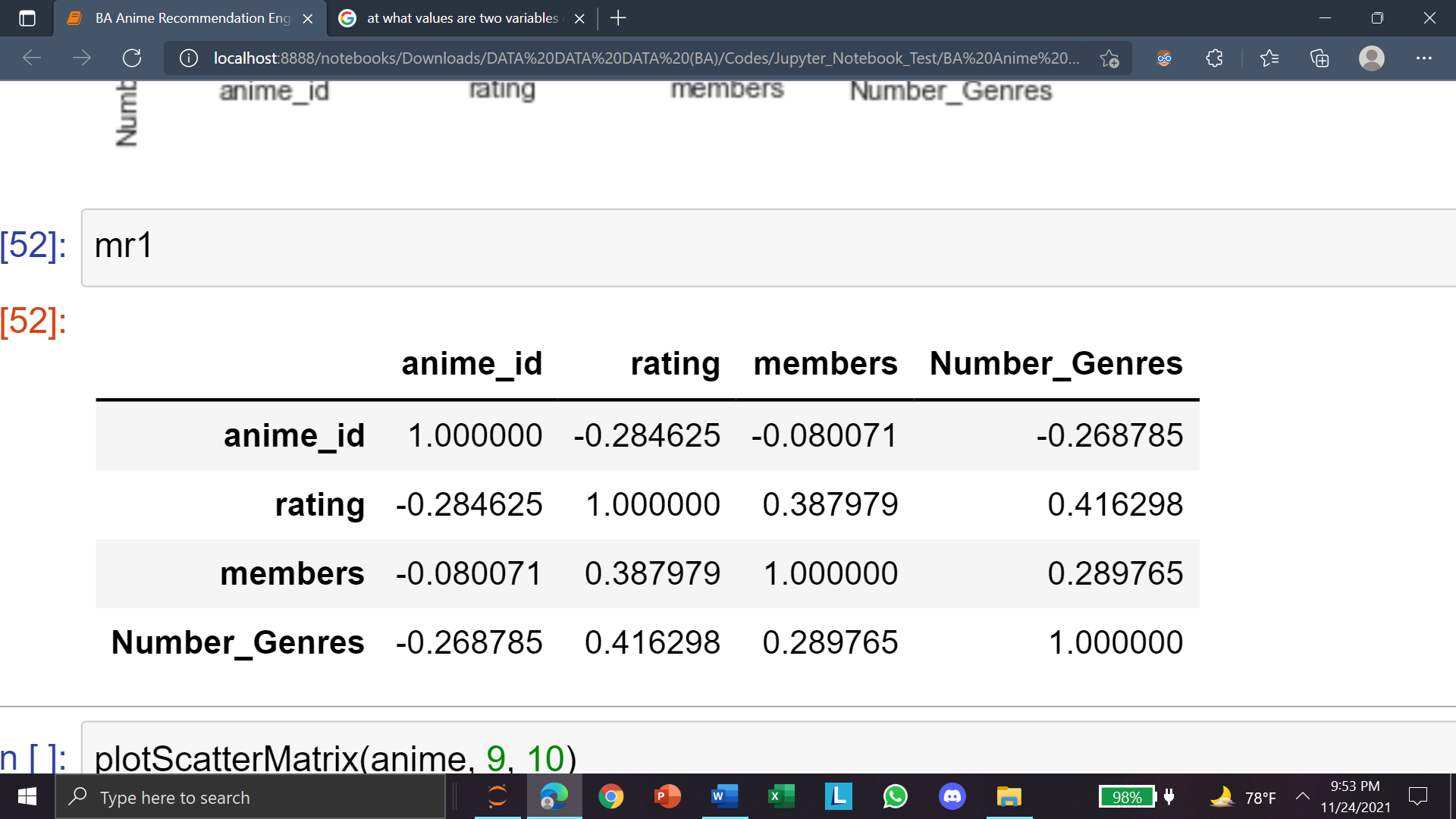
|  |  |  |  |
| --- | --- | --- | --- |
| **Calculated F Ratio** | **P-value** | **F-critical** | **Decision** |
| 1597.55 | 0.0 | 2.21 | Reject Null Hypothesis |

Since the p-value is less than alpha, the null hypothesis has to be rejected, suggesting that there is statistically significant difference across mean ratings of anime with different types of release.

## Correlation

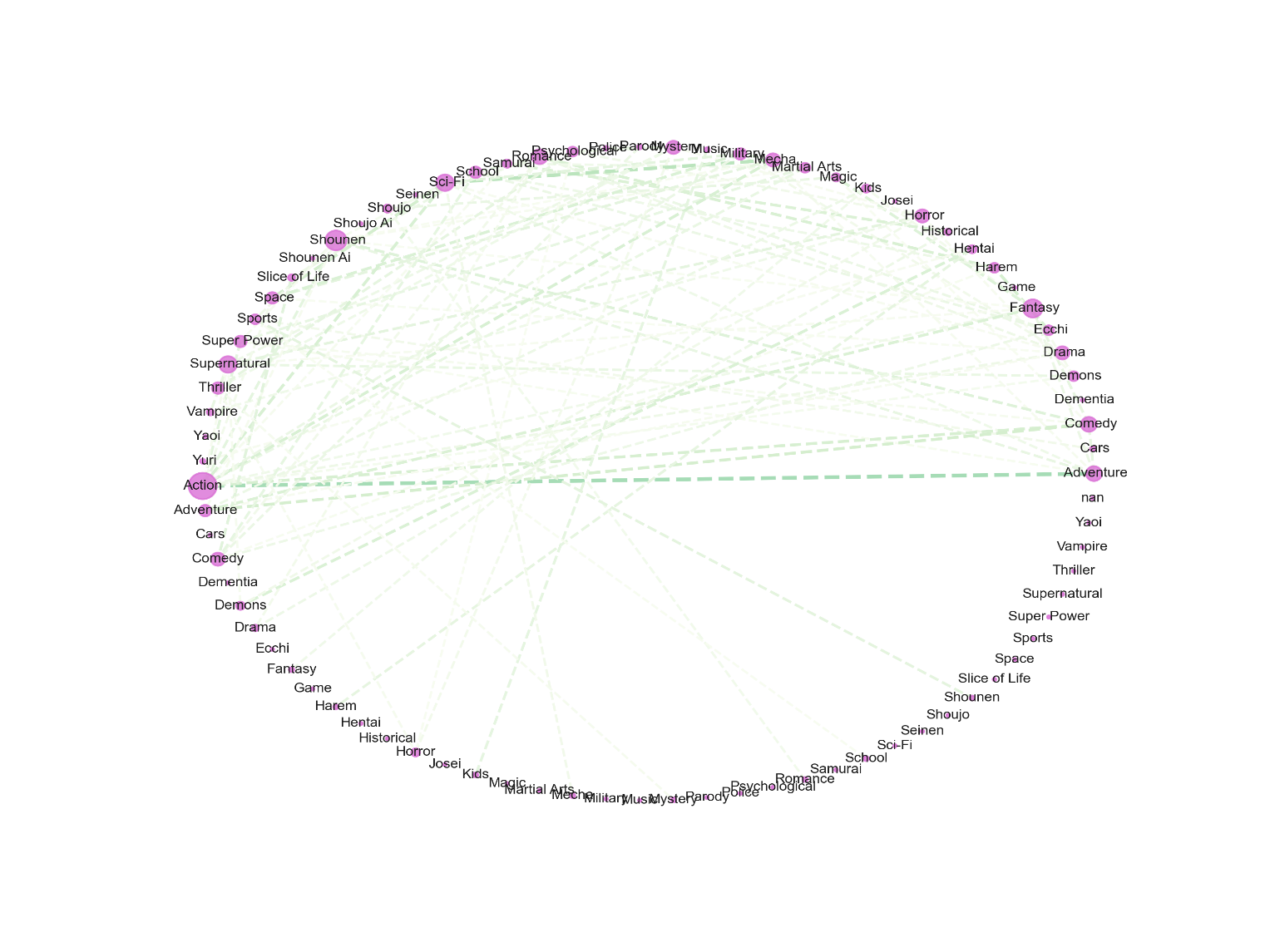
### 3.1 Correlation among Variables:

For the quantitative variables a correlation matrix was plotted. Rating and the number of genres is shown to have a strong correlation with a correlation value of 0.416. Furthermore, moderate levels of correlation can be seen between members and anime rating have a correlation value of 0.388 and the number of genres with members having a correlation value of 0.290. Other than that most variables have weak correlation or are uninterpretable (for comparison with anime\_id).



### 3.2 Correlation among Genres:

The network diagram displays the correlation of the between genres. The strongest associations are depicted by thicker green lines. From observation it can be seen that Action and Adventure, Sci-fi and Mecha are the most strongly correlated.



# Appendix

#### Data Link:

* <https://www.kaggle.com/CooperUnion/anime-recommendations-database?select=rating.csv>

#### Code Link:

* [Project-Otaku-An-Anime-Recommendation-Engine/BA Anime Recommendation Engine.ipynb at main · Ferdostto/Project-Otaku-An-Anime-Recommendation-Engine (github.com)](https://github.com/Ferdostto/Project-Otaku-An-Anime-Recommendation-Engine/blob/main/BA%20Anime%20Recommendation%20Engine.ipynb)