AMDG

Faadil Ahmed, Tess Foral, Kyle Gwon, Frank Nieto, and Andrew Stranahan

IS 407

12/17/21

1. **Abstract**

As anime’s popularity spreads outside of Japan, we focus our research on the following question: *Can we determine the factors that influence an anime’s popularity*?Indeed, anime creators are surely asking themselves how they can make the next big anime. Given the sheer saturation of the anime market, many producers and distributors are eager to know which series they should put time and effort into and which ones they should avoid. In this study, we used multiple statistical analyses to see if we could determine if there are any variables that had an influence on the popularity (defined in this study as a high user-generated score) of an anime. We found that certain genres tended to be rated higher than the general anime population, but we could not determine if the relationship between a genre and score was causal or correlational. Similarly, we found that our selected variables related to viewership neither meaningfully contributed nor explained any given anime’s relative popularity. Finally, we found that the genre that we predicted to be the most popular was only the second most popular genre under our prediction analysis. This means that there is likely some other variable that cannot be quantified that influences a high rating, and that there are likely a multitude of factors that impact an anime’s popularity. Ultimately, we concluded that additional research with more metadata is needed to answer our research question adequately.

1. **Introduction and Literature Review**

Recent studies in the entertainment industry show that there is current and growing attention on how to predict the success of future media releases, including anime, using methods such as machine learning, statistical analyses, and algorithms. In this research paper, we will be using data from the MyAnimeList dataset from Kaggle, where users can give a score (numbered 1-10) for or “favorite” a series, to determine any factors that can influence an anime’s popularity. Commonly, an anime’s popularity is reflected in its income from DVD and merchandise sales, as well as its marketing campaigns, such as participation in anime conventions. Within the dataset, the column “Popularity” is a ranked list based on the number of users who have added the anime to their list of series to watch, i.e., the more users add an anime to their “list,” the more popular it is and *vice versa.* However, our research will define “popularity” as the user-generated score for an anime. The higher the score, the more popular it is among consumers and *vice versa*. Although this definition of popularity is more narrow than other definitions, we wanted to determine if there was any correlation between an anime that is perceived as popular and our data.

Other scientists have attempted similar feats. Research conducted by Abidi et al. (2020), Dhir and Raj (2018), and Kanitkar (2018) show that scholars frequently turn to online sources, such as IMDb, to gather data used for running predictive analyses on a movie’s success, prior to running machine learning processes on their collected data. Similarly, AlSulaim and Qamar (2021) applied deep learning methods to process sentiment analysis, and to predict an anime series’ success. Meanwhile, Cho et al. (2017) showed that anime consumers want recommendations for new series to watch, but encounter difficulty in finding them. Although each of these scholars used different techniques to uncover their findings, they all collected, normalized, and processed data on either films or anime to predict future popularity and consumers’ interest in their respective media. We modeled our own data analyses of anime based on this prior research.

Fortunately, we are able to combine our present study with the work conducted previously. Our own research relates to AlSulaim and Qamar (2021)’s research in that we are using the same dataset from MyAnimeList that they used, which allows us to check their results against our own. Meanwhile, the categorical variables used by Abidi et al. (2020) are fairly similar to the variables we will be using to conduct our analysis, such as genre. In addition, the information features found in Cho et al. (2017)’s study appear in our own dataset, such as genre, theme, and length, and we can potentially focus on those areas because they have the most impact in determining how popular an anime is. These examples show that our own research dovetails into the previous and current research that tries to predict the success of content released by the entertainment industry.

Having found our own place within recent scholarship, we can now construct our research question: *Can we determine the factors that influence an anime’s popularity?* This research is to determine if it is possible to use similar methods used by previous researchers to determine the popularity of an anime series among viewers based on a variety of variables and to study which variables in particular might be important when predicting an anime’s popularity.

Based on preliminary knowledge and experience, it seems likely that genre will have an effect on a given anime’s popularity. Thus, our hypothesis is that animes under the “*Shonen*”or “*Shounen*” (“teenage boy”) genre, and which is often characterized alongside action and adventure series, will be generally more popular with consumers. We plan to learn if there are predictable patterns that can identify which animes become popular and which ones fail to gain traction with the public. We hope to use our hypothesis, predictions, and analyses to guide future research in determining an anime’s popularity with viewers.

1. **Methods**

The following five subsections explain our data cleaning and statistical analysis methodologies. Because we did not run any hypothesis tests, defining statistical significance and alpha is not applicable to our research. All data cleaning and statistical tests were run through Austin, TX’s Anaconda3’s Jupyter Notebook 6.2.0. Also, we used Beaverton, OR’s Python 3 as our sole programming language.

1. **Preliminary Steps**

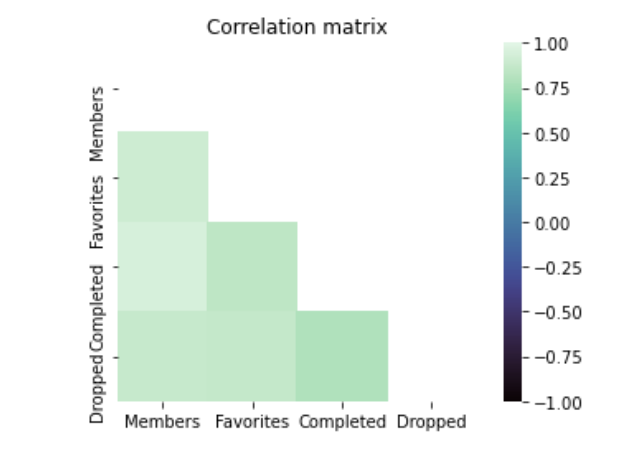
The first step was to read in the dataset as a CSV file using Python’s Pandas library package. After reading in the dataset, the next step was to use a listwise deletion to get rid of any rows that contained missing data, which could interfere with the rest of the data processing. In doing so, we removed about 5,000 rows that contained null entries. Next, after reviewing the 43 genre types in our dataset, we decided to focus on the 10 most prolific genres, which can be found in Table 1. Although this could bias our results, as more anime series are likely made within a more popular genre, we made this decision in order to narrow our attention on the most relevant data and genre types. To do so, Frank separated the data from the dataset into dataframes that contain each of the top 10 genres. To do so, he used a Boolean “true/false” operator in the dataset to identify if an anime was a part of that particular genre. If an anime series was indeed a part of that particular genre, then the Boolean operator returned “true,” and, if not, then the Boolean operator returned “false.” Frank converted the true outputs as “1” and the false outputs as “0” before dividing the anime series into dataframes. If an anime listed a particular genre, it was included within that dataframe.

1. **Multicollinearity**

Before we ran any statistical analysis tools, Tess used the Spearman method to determine if four quantitative independent variables related to viewership are correlated, which would answer the sub-research questions of: *Are these independent variables correlated*? *Will our variables be useful in answering our overarching research question*? Using the Spearman method, she found that the independent variables we had planned to use in the multivariable linear regression were strongly correlated, as seen in Figure 1. We did not include genre for this test, because its data is categorical.

1. **Measuring Central Tendency**

After Frank divided the genres into dataframes, Faadil then used the Pandas library package to determine the median and quartiles of each genre. He did this to compare them each to the general population of all animes to answer our sub-research question of: *Are more popular genres rated higher or are they similar to the general population*? The results can be seen in Table 1. Once he had found the measures of central tendency, he used Python’s Seaborn library package and created box plots that compared the genre-specific scores against the scores from the entire population. These results can be seen in Figures 2-11.

**

***Figure 1****. Spearman correlation matrix.*

1. **Multivariable Linear Regression**

Even though Tess had determined that our independent variables related to viewership were strongly correlated, Andrew decided to see what sort of results could be obtained from them still, especially if there were any variables that had an impact on the score of the anime. Moreover, he wanted to see if he could also find answers to the same sub-research questions that Tess addressed earlier. He used multivariable linear regression to study this and focused on r2 as his performance metric. To conduct multivariable linear regression, Andrew used Python’s Statsmodel module to analyze independent variables related to viewership. Meanwhile, he used “Score” as the dependent variable to determine how the chosen independent variables influenced the score. This would allow us to determine if there are relationships among these variables and score, which would help in the overarching goal of determining anime popularity. This can be seen in Table 2.

1. **Logistic Regression and Odds Ratio**

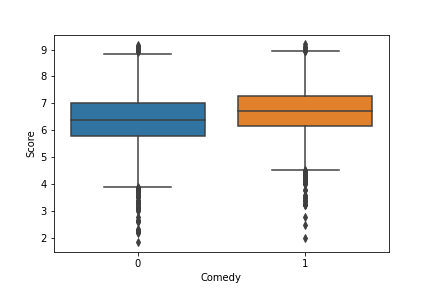
Finally, Kyle wanted to answer our sub-research question of: *Are there other indicators of whether or not an anime will be highly rated*? To do this, he ran a logistic regression model to determine if genre was a suitable method for determining what makes a “popular” (i.e., highly scored) anime. This meant he had to set a number that we would consider especially “popular.” Any scores above that number would be designated as a success and scores at or below that number would be designated as a failure. For the purposes of this study, he chose to limit the “popular” scores to above “7,” and “unpopular” scores to “7” or below. Then, Kyle compared each of the top ten genres against the high and low scores using the Sklearn library’s logistic regression function. This can be seen in Table 3. Similarly, Kyle used Numpy to calculate the odds ratio for all of the genres in order to predict which genre was likely to have the most popularity. This can be seen in Figure 12.

1. **Results and Discussion**
2. **Measures of Central Tendency Results**

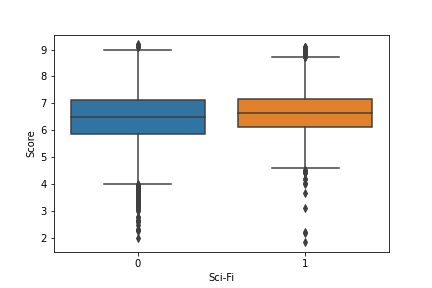
|  |  |  |  |
| --- | --- | --- | --- |
| **Genre** | **25th quartile** | **Median** | **75th quartile** |
| GENERAL POPULATION | 5.93 | 6.5114 | 7.14 |
| Comedy | 6.18 | 6.7041 | 7.29 |
| Sci-Fi | 6.13 | 6.75 | 7.27 |
| Adventure | 6.26 | 6.75 | 7.27 |
| Action | 6.22 | 6.7508 | 7.3325 |
| Fantasy | 6.14 | 6.752 | 7.22 |
| Slice of Life | 6.27 | 6.8388 | 7.46 |
| Romance | 6.4 | 6.8838 | 7.3925 |
| School | 6.41 | 6.9009 | 7.43 |
| Drama | 6.42 | 6.977 | 7.53 |
| *Shounen* | 6.5 | 7.0187 | 7.51 |

|  |
| --- |
| ***Table 1****. Measures of Central Tendency, sorted from the lowest to highest scores of median ratings*. |

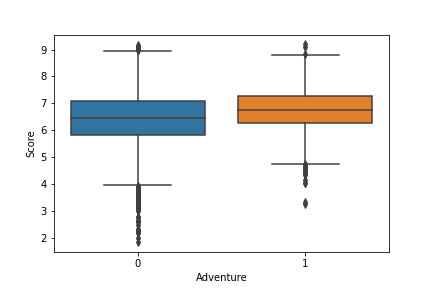
All of the genres we studied have a median score that is higher than that of the general population. This makes sense, as we studied the most prolific genres. The range of the scores of each individual genre tended to be higher as well. As predicted, the “*Shounen*” genre is the most highly rated. However, genres more closely aligned to “*Shojo*” or “*Shoujo*” (“teenage girl”) series, such as “School,” “Romance,” and “Slice of Life,” are rated higher than more typical *Shounen*-affiliated genres such as “Action” and “Adventure.” Thus, it is more accurate to say that the *Shounen* genre is most often rated highly, though genres under that umbrella, such as “Action” and “Adventure,” are not necessarily as highly rated.



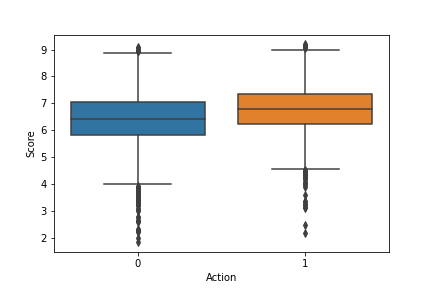
|  |
| --- |
| ***Figure 2****. Boxplot of the scores of the general population (0), and the anime under the “Comedy” genre (1).* |



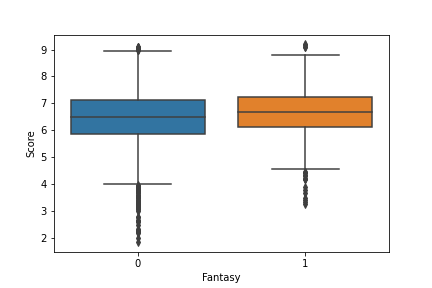
|  |
| --- |
| ***Figure 3****. Boxplot of the scores of the general population (0), and the anime under the “Sci-Fi” genre (1).* |



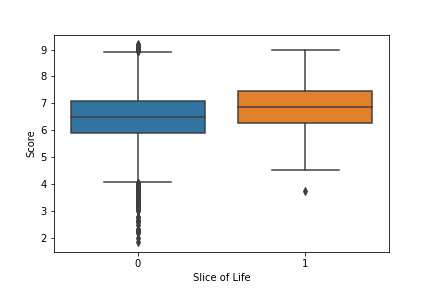
|  |
| --- |
| ***Figure 4****. Boxplot of the scores of the general population (0), and the anime under the “Adventure” genre (1).* |



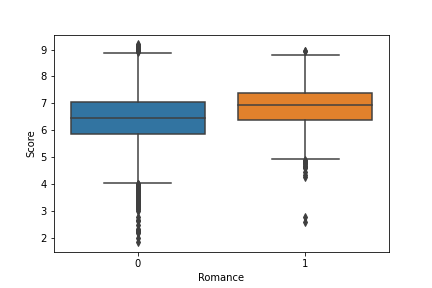
|  |
| --- |
| ***Figure 5****. Boxplot of the scores of the general population (0), and the anime under the “Action” genre (1).* |



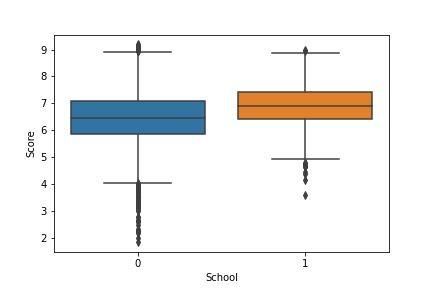
|  |
| --- |
| ***Figure 6****. Boxplot of the scores of the general population (0), and the anime under the “Fantasy” genre (1).* |



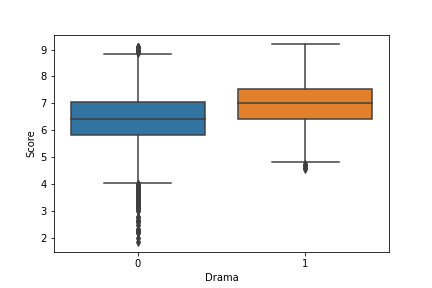
|  |
| --- |
| ***Figure 7****. Boxplot of the scores of the general population (0), and the anime under the “Slice of Life” genre (1).* |



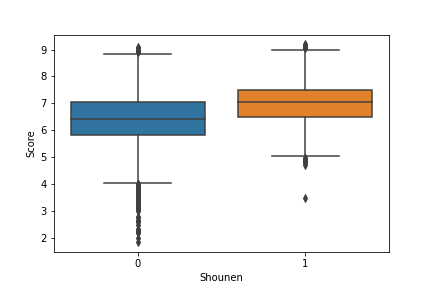
|  |
| --- |
| ***Figure 8****. Boxplot of the scores of the general population (0), and the anime under the “Romance”*  *genre (1).* |



|  |
| --- |
| ***Figure 9****. Boxplot of the scores of the general population (0), and the anime under the “School” genre (1).* |

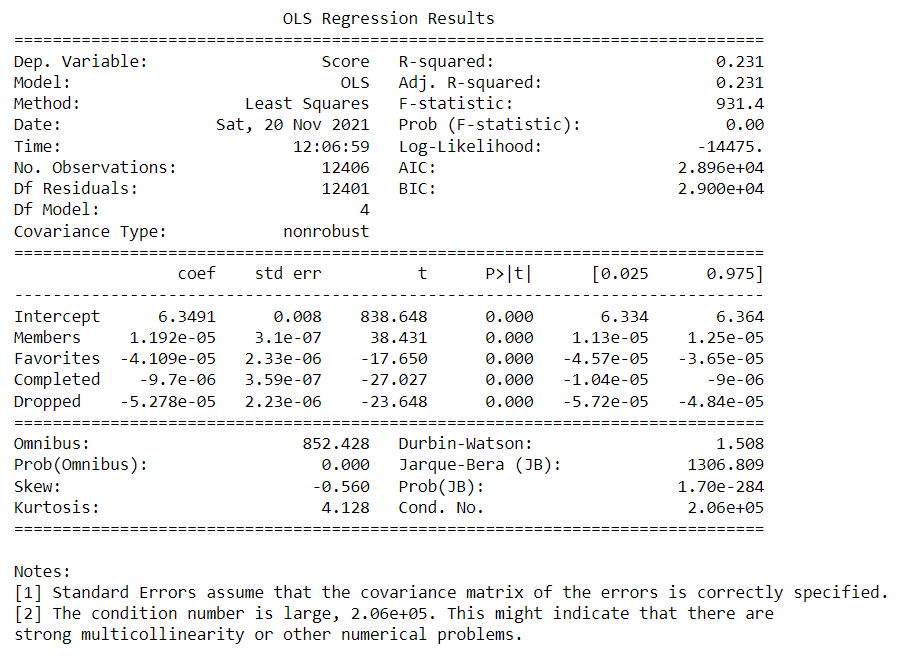


|  |
| --- |
| ***Figure 10****. Boxplot of the scores of the general population (0), and the anime under the “Drama” genre (1).* |



|  |
| --- |
| ***Figure 11****. Boxplot of the scores of the general population (0), and the anime under the “*Shounen*” genre (1).* |

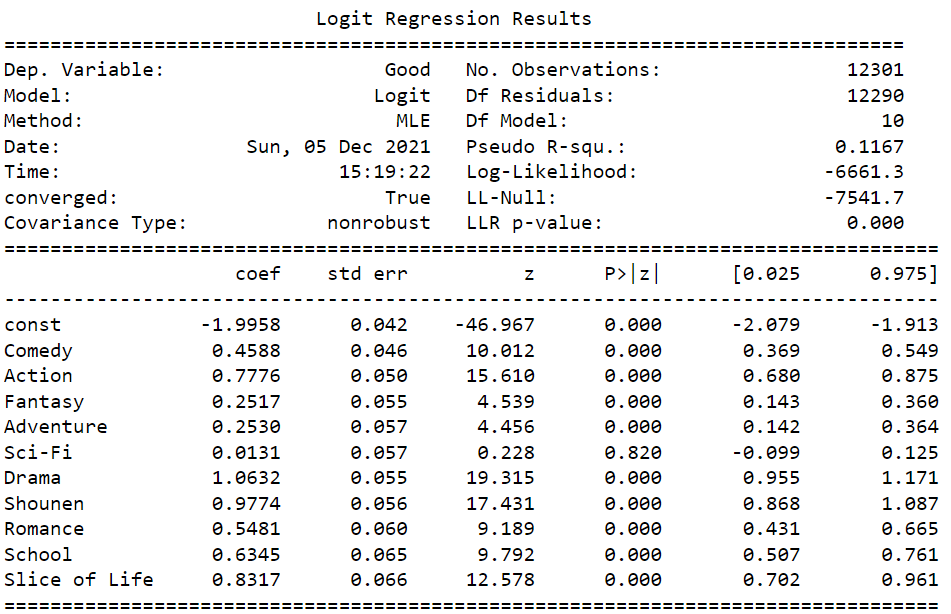
1. **Multivariable Linear Regression Results**



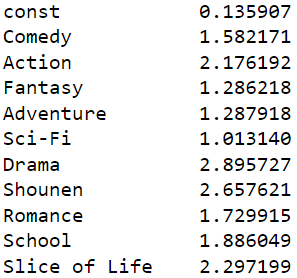
|  |
| --- |
| ***Table 2****. Output of the Statsmodel multivariable linear regression model.* |

We ran a multivariable linear regression model to determine if there is any relationship among the selected independent variables of “Members,” “Favorites,” “Completed” and “Dropped” to our dependent variable of “Score.” After analyzing the results, we discovered that r2 is equal to 0.231. Statistically speaking, this means that the relationship among our independent variables related to viewership and “Score” is weak. However, because we are basing our definition of popularity on user-generated scores, this means that our values are based upon human behavior. Thus, our r2 will not be very high. This makes it hard to gauge the true strength of the relationship among our independent variables to “Score.” In addition, multivariable linear regression assumes that there is no multicollinearity among the specified independent variables. However, as seen in Figure 1 and in Table 2, there is strong multicollinearity among the selected independent variables. Thus, our current multivariable linear regression model is insufficient to use for explaining the relationship among our independent variables to “Score.”

1. **Logistic Regression and Odds Ratio Results**



|  |
| --- |
| ***Table 3****. Output of the Logistic Regression Model.* |



***Figure 12.*** *Output of the Odds Ratio Test.*

Our logistic regression model analysis shows that genre alone is not the best or most relevant factor in determining the popularity of an anime, which we can tell from the low pseudo R2 value of 0.1167. From our hypothesis, we assumed that the “*Shounen*” genre would be the most popular. However, while the odds ratio of the “*Shounen*” genre is about 2.658, the “Drama” genre is actually the best indicator out of all the genres of whether or not an anime will be popular, as it has a higher score of about 2.896. Thus, it would be plausible to assume that there are other factors that influence an anime’s popularity beyond genre.

1. **Limitations**

In addition to our results, we identified limitations within the MyAnimeList dataset. For example, there are only a limited number of variables possible, some of which are difficult to use in determining an anime’s popularity, e.g., premiere season, an anime’s Japanese title, etc. In addition, each anime in our dataset had multiple genres associated with it, which made it impossible to put an anime into one, distinct category. Thus, anime genres should be viewed as part of a whole, rather than as distinct categories. Factors that might play into an anime’s popularity, e.g., marketing costs and social media content, are not present within this dataset. Meanwhile, although our chosen independent variables related to viewership exhibited strong multicollinearity, this does not mean that they are unusable in the context of analyzing anime popularity. This simply means that there are other independent variables or combinations of independent variables related to viewership that can better show the relationship between them and an anime’s popularity. Therefore, although we are unable to say definitively that there are no major factors that influence an anime’s popularity, we can state that none of the factors we examined had any major effects on influencing an anime’s popularity. In the future, it might be possible to gather more metadata, such as production budget, marketing budget, etc., and use them in the search to find any factors that influence the overall score of an anime. However, that is beyond the scope of this research.

1. **Conclusion**

None of the factors studied here had a meaningful effect on the score of any given anime. While series under the “*Shounen*” genre in our dataset tended to be rated more highly than other genres, we cannot claim that there is any statistical insight to this. Thus, while we can confidently say that *shounen* animes tended to be more popular than the general population of anime in our dataset, it is not accurate to say that they are more popular because they are a part of the *shounen* genre. For example, genres associated with *shoujo* animes, e.g., “Slice of Life” and “Romance,” are rated higher than genres associated with *shounen* animes, e.g., “Action” and “Adventure.” It is hard to determine why this is the case without more data and it is more likely that there are a variety of factors that play into this situation.

One factor that could play into an anime’s popularity is that many anime series are adapted from Japanese graphic novels or comic books called “*manga*.” In particular, Hodgkins (2019) found that *Weekly Shonen Jump*, a *shounen*-focused *manga* magazine, sold more copies than *Ciao*, a *shoujo*-focused *manga* magazine. As such, many people are likely to be invested in an anime because of the popularity of its manga, and, because *Weekly Shonen Jump* has a larger fanbase, their *manga* titles and subsequent anime adaptations are likely to be more popular. Thus, future research can study the correlation between an anime’s popularity and its connection to a preexisting *manga* series, if applicable.

Although our results were inconclusive, we hope that our research can contribute to future studies in determining an anime series’ potential popularity with viewers. As the anime industry continues to grow and influence consumers beyond Japanese borders, such research will become more necessary as time progresses. Thus, we look forward to the time when there is enough resources and expertise to begin analyzing the specific variables that make an anime so popular among viewers.

References

Abidi, S. M., Xu, Y., Ni, J., Wang, X., & Zhang, W. (2020). Popularity prediction of movies: From Statistical Modeling to Machine Learning Techniques. *Multimedia Tools and Applications*, *79*(47-48), 35583–35617.

AlSulaim, S. M., & Qamar, A. M. (2021). Prediction of anime series' success using sentiment analysis and deep learning. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif )*.

Anaconda, Version 3.9, Austin, TX

Cho, H., Schmalz, M. L., Keating, S. A., & Lee, J. H. (2017). Information needs for anime recommendation: Analyzing anime users' online forum queries. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Dhir, R., & Raj, A. (2018). Movie success prediction using machine learning algorithms and their comparison. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*.

Hodgkins, C. (2019, June 9). *JMPA reveals manga magazine circulation numbers for January to March 2019*. Anime News Network. https://www.animenewsnetwork.com/staff

Kanitkar, A. (2018). Bollywood movie success prediction using Machine Learning Algorithms. *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*.

Python, Version 3.10.01, Beaverton, OR