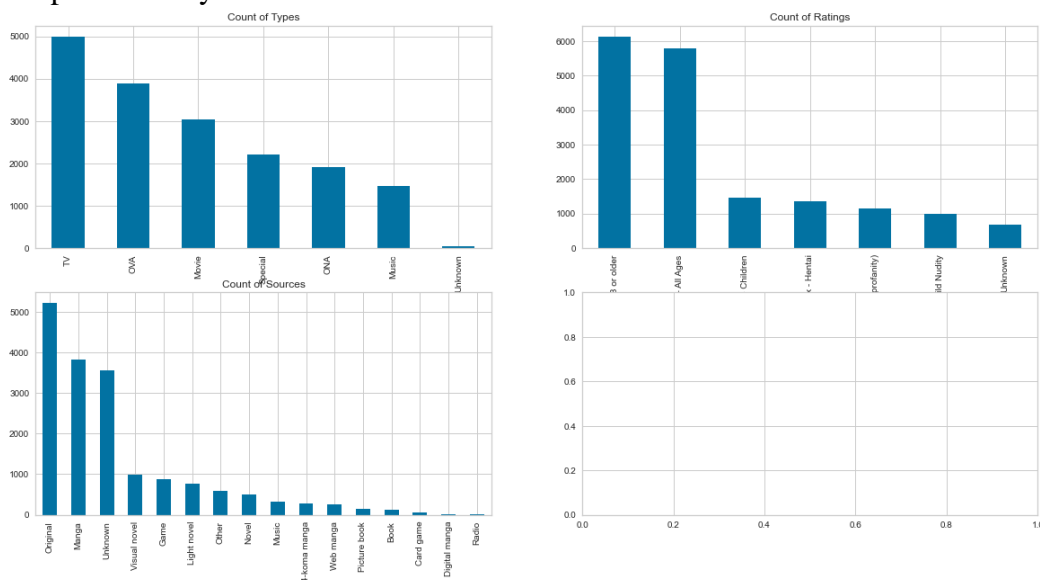


- Business Problem/Data

If you watch anime or read manga you might be able to resonate with this however, even if you do not, I am going to explain everything for you. When you look at the grand scheme of things a show can make money off being on a streaming site, television, publications, etc., these are some ways an anime can make revenue. But today we are looking past that we want to use that as a tool to gain popularity and that will ultimately allow the studio to sell merchandise and make a lot more profits. When you look at top anime like Naruto, One Piece, Dragon Ball are selling tons of merchandise, but they had to get to this spot somehow to gain this popularity.

The problem I am focusing on is how I could take data and provide the recipe for an anime studio to either pick up a manga for animation or create an original anime. The data set (<https://www.kaggle.com/hernan4444/anime-recommendation-database-2020?select=anime.csv>) I have used has great values that I have used to tackle this problem with important columns that can link to the popularity and in turn the revenue or ability to create revenue from a studio, such columns like Score, Episodes, Type, Watching. Completed, Scores 1-10. These are just some of the features that are in the data set that I am using to build my models off that can help giving the studio a reason to pick which type of anime to either create or pick up from a manga.

- Graphical Analysis

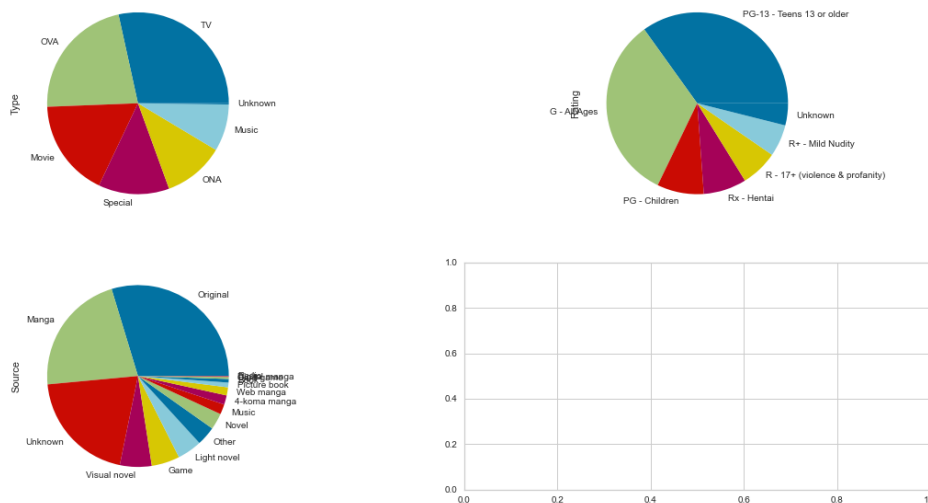


Looking at these bar charts I would start out by explaining what they represent, so starting with the “Count of Types” chart, we can see that the most popular form of anime is TV, now you cannot just say because you have an anime TV show that it will be successful, but we can say that there is a community there for anime on TV more than an OVA, movie, or special. From a real-world perspective explaining this to an anime studio should not need to happen but will give some credibility because they should understand

this about TV shows compared to a movie if they have ever experienced anime before hand, if not the information is available.

Next I want to take a look at the bottom left chart “Count of Sources”, here we can see that the best option by far is to make something original however, again you still cannot conclude that this will make your anime successful just that you should most likely make an original if you have the resources which means a story, characters, voice actors, animation, music, sound effects, all the things needed to create an anime from an art stand point. Now based on the resources of the company, choosing a manga could benefit the studio as they would have to do far less work, as well as the anime starting out with a fan based already from fans who already read the manga.

I wanted to wait for this visual for last because I wanted to demonstrate how the same features and values can pop in different visuals. For the “Count of Ratings” chart we can just see the ratings that are most common, but I do want to show a different type of visual that would better show this feature.



As you can see, I turned the bars into pie charts, but if you look top right again you can see more that the ratings “G-All Ages” and “PG-13” make up more than half of the total ratings of shows. This is something that we can for sure consider because before we said that we can not assume that TV will be successful but TV has different ratings you can put on it than a movie so this is showing us that if we wanted to go down that route then the studio might not want to pick up a higher rated anime than PG-13. The goal is to figure out the correct anime for a studio to decide on because their goals could be different than just revenue so, we need to look deeper than just what anime has done before, we need to see if we can find out why successful anime are so successful in which ever way their studio sees success.

- Dimensionality & Feature Reduction and Feature Engineering

I did have the issue of similar features in this data set if you look at the raw data, I found these features to be useless:

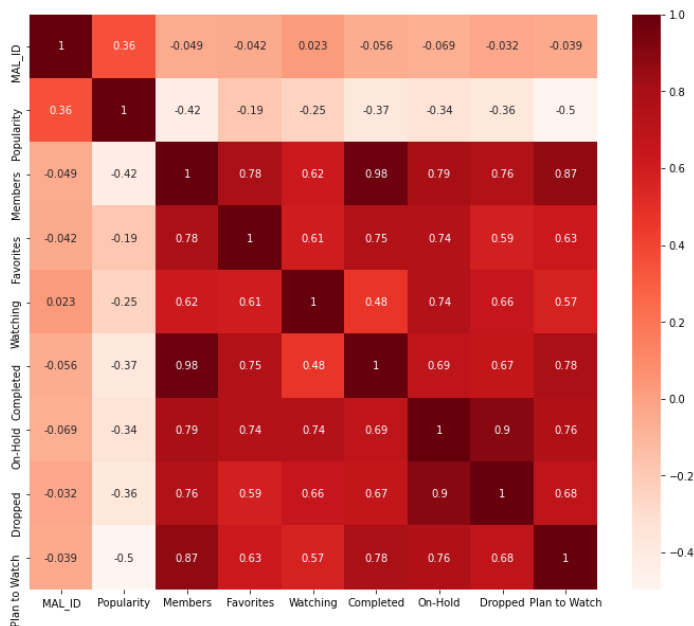
```
#Correlation with output variable
cor_target = abs(cor["Favorites"])

#Selecting highly correlated features
relevant_features = cor_target[cor_target>0.5]
relevant_features

Members      0.778583
Favorites    1.000000
Watching     0.607484
Completed    0.749235
On-Hold      0.739112
Dropped      0.588051
Plan to Watch 0.625559

#dropping the features
df2 = df.drop(['Completed', 'On-Hold', 'Plan to Watch', 'Dropped', 'Watching' ],1)
```

We can see from this all these features are all highly correlated, so we are going to choose Members and drop the rest because in linear regression, one of the assumptions are that the independent variables need to be uncorrelated to each other. We can create heat maps for the correlations between features.

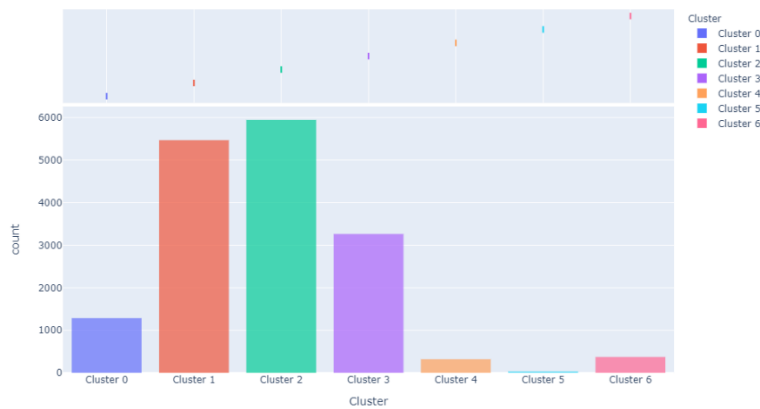
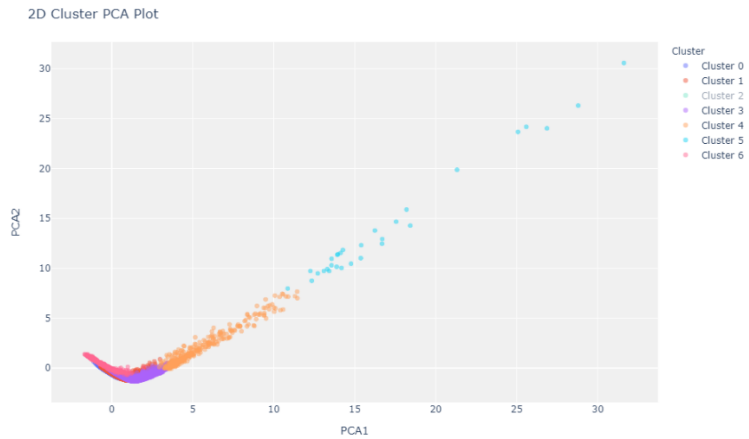


- Model Selection & Evaluation

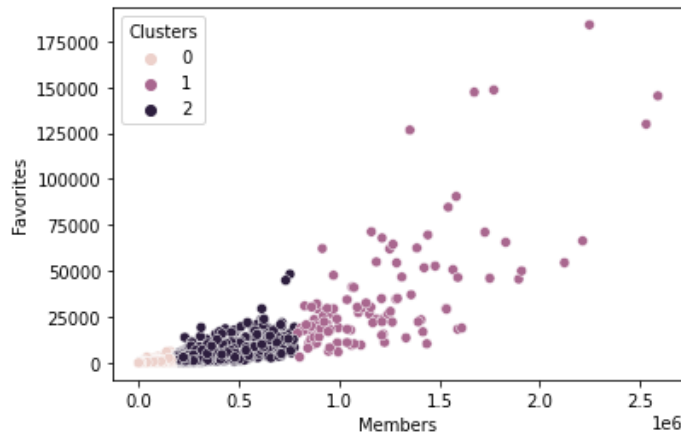
So of course, the problem is to help the studio figure out the right anime type, rating, genre, anything that can be a factor on why previous and current anime are successful and how I can use this data to help a team choose an anime. I then was thinking about what type of model would best help group anime together, which brought me to k-means clustering, this was the perfect model for this problem because we can just look at the

clusters and see what makes the best cluster for them or what cluster they want to be in with their anime.

I chose to first use the Pycaret library to create my first model and it was a very easy set up with some minor issues when it came to the library itself, eventually finding some very helpful forums. With this model I will be able to see the clusters and find the right fit for the studio and this can become modular being able to find the right scenario for different studios.



I thought it would be a great idea to make another k-means model using the Sklearn library because I want to see the differences between clusters, and it has on the data.



I am getting similar results for my clusters between the two models, and they even have a different number of clusters in them, yet the trend stays the same. The accuracy for these models are a bit above 95% but we can just round down to be safe, the more you run and train a model the more accurate it should become but for these it was about 300 iterations.

- Conclusion

With the Pycaret model there was an option to hover through the visuals so having that during the meeting with the studio would be helpful because you can hover and click on each cluster and point. But after seeing these models and going through all the processes with this data I can conclude that you can pick apart an anime that best fits the business' needs.

From the predictions loading the models that you see visuals for above we can constantly get the clusters and information for each as we go deeper into the predictions on why anime are being placed into these clusters. For example, a show that has a high rating of 6-7 on the Scores are being placed into the second cluster, although these are not similar types of anime, they have similar scoring across the board. And there are a lot of unknown information about original anime so that has been cluster 6. So if the studio had the problem of not knowing which direction they would want to go with their newest or first anime, this is a great place to start to really make a list and have some safe or even risky decisions. I wanted to really be able to go to someone and say here look at what I have let me match up your needs.

After this research and data wrangling I can say at least from the data I looked at, which included majority of the anime I could find with enough data and scoring, you cannot just decide on a type, or genre, maybe a rating if you know you wanted more people to be watching we see PG-13 is the cut off because anything more than that has less of a community to pull from excluding some outliers. We just need to match up the studio with these clusters and the one that works for their needs is going to have these features to help determine the route they should take on the anime that will be the best release.