Anime Recommender

Thi Q, Saatvi Rajgopal, Alfred Custodio, Bethzaida Guerra

Introduction

The popularity growth of anime has been interesting to watch. *Cowboy Bepop (Live Action),* was the top watched show in the US last weekend, and at the time of writing this (11/30/2021) it is still one the top 10 list. Another show, *Attack on Titan,* beat out hit shows like *WandaVision* being the most demanded show in the week of January 31 (per Parrot Analytics). Who would've thought that this niche would become, according to Netflix "a common global currency"; that is set to be worth $48.3 Billion by 2030? But what is anime exactly?

Anime simply put is animation that comes from Japan (although to the Japanese the word refers to any type of animation). It started in 1917 and the first anime was less than five minutes long and was made with chalk. It began to grow in Japan and made its way here to the western world in 1961 with Astro boy and it proved that anime could be made for TV and it could be profitable. Its popularity began to grow till the anime "golden age" in the 1980's-1990's as more genres such as the anime sports genre with shows like *Captain Tsubasa,* shonen anime such as *Dragon Ball,* shoujo anime *Sailor Moon.* However anime's growth didn't stop there and with the help on the internet it has now become a source of r

Inspiration

- **Idea-** Our inspiration stems from the fact we all love anime, a lot of us, watching it from the time we were young. Our inspiration also stems from the fact that as much as we love anime, there is so much anime that we never know what to watch; there isn't time to watch all the anime that is out there. It's also worth noting that not all anime is worth watching, which is why we decided to build an anime recommendation page.
- **Color scheme-** A lot of anime follow a similar color scheme using complementary and split complementary colors. Take several of the popular anime: Naruto, Pokemon, Sailor Moon, Doraemon, Fullmetal Alchemist, DBZ ect. They all have a lot of hues of yellow, blue, red and oranges mixed with some achromatic colors.
  -
- **Tableau Inspiration-** We looked up a couple of Dashboards that were on Tableau public, used them as inspiration and as a guide for our own Tableau dashboard.
  - Introducing Naruto! | Tableau Public

- Anime Analytics through Studios | Tableau Public
- anime | Tableau Public

After looking at the previous data sources for inspiration, we made our own dashboards using the color scheme that was previously chosen.

Data Source:

Once the group decided on a topic, we began to search for a data source that could be used for this project. Ideally we wanted to have something that was updated for the current year. We thought about using the MyAnimeList API in order to get anime data; time was not on our side so we took that idea off the table.

Instead we used a database from Kaggle and it provided from MyAnimeList Database 2020. It provided us with the information that was needed, including but not limited to; the anime, its sub-genre, the score, source, etc. It was also used to create a machine learning model that recommends anime based on the information provided by the dataset.

Data Cleaning:

The original data had several unknown for all categories. Since “Score” is an important component to our data, unknown was dropped from the Score column. Hentai was also dropped from Genre as we wanted to keep the material class-appropriate. As mentioned previously, we had to break down the Genre section as each anime could fall under several types of genre. This created an obstacle when it comes to creating visualizations for Tableau and generating recommendations using KNN. Genres for each anime were then separated into its individual columns in reference to the anime. This suggests that the count of genres will be higher than the number of anime.

Website Design:

The idea for the website design was to make it user-friendly, be able to successfully show our data visuals, be able to successfully make recommendations, have it look “hella dope”,and be appreciated by both anime fans and people who did not know much about anime. The website design went through many changes to get to a point that we felt was satisfactory.

At first the idea was to make a MyAnimeList knockoff, however it was discarded as the site does not really appeal to the eyes or grab one's attention. It was also discarded because not many people are die-hard anime fans so the reference would go over people's heads.

The second idea was to make it like a stream site, having different anime playing, and other features. However building a site like this takes more than what was given, and sadly the idea was given up on.

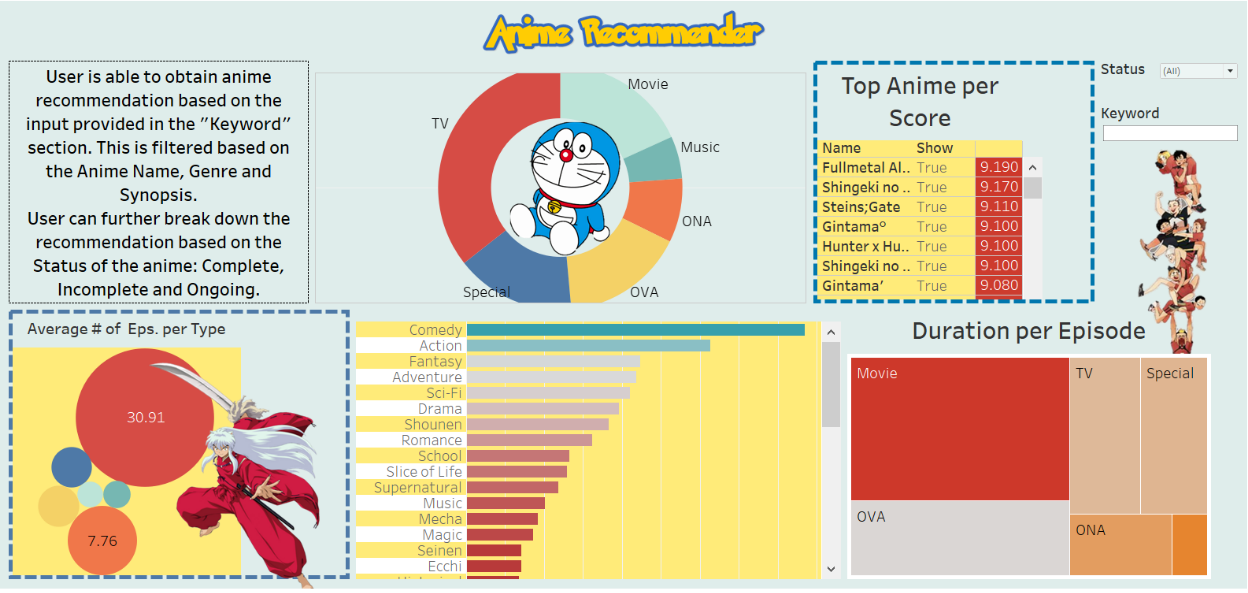
The third idea, and the one that stuck was making using videos and gifs to make the site appealing. There are either videos or gifs in every page to entertain the user, anime fan or not. More importantly this idea allowed for the site to be simple but eye catching, user-friendly and not overwhelm the user who was either looking at the data, using the recommender or reading the write up.

Tableau:

Our Tableau dashboards followed an overall color scheme of blues, yellows, oranges, and reds which was derived from some of the popular animes today. As you can see, the user viewing our dashboards can use the "Keyword" function to input their own suggestion. The data for each graph will be adjusted based on the Keyword. Output from keywords is generated from creating a new dimension using Calculated Fields of Genre, Name and Synopsis. The Parameters function was then used in parallel to create a search field for the input of strings. The user can also break down the data even further by filtering using the "Status" of the show- whether it's incomplete, complete, or still on-going. One finicky detail with the Tableau data is that the "Popularity" column is actually a ranked list, not an accumulation of scores from the viewers. This means that higher the "Popularity" number is for an anime, the lower the popularity is of an anime. With that given, there are a lot of elements from our data visualization that we could point out and confirm that the data makes sense. For example, our "Top 20 Most Popular Animes (meaning the ones ranked from #1-#20), is understandable, because the names listed would be familiar even to people who are not avid watchers of anime.

One other detail of the dataset that had to be adjusted prior to visualizing, was the "Genre" column. When we first looked at the dataset, each anime was given multiple different types of genres all at once, leaving us with thousands of different genre values because none of them matched. For example, one anime's genre could be "Comedy, Slice of Life, Dementia, Romance, Sci-Fi", and the one after could be "Comedy, Slice of Life, Dementia, Romance, Adventure". Since they are not exactly the same, they would not match up, requiring us to filter that column down to the most popular genres listed, and only apply the first genre of each list to the anime. This left us with "Comedy" and "Action" as our most popular anime genres, with "Fantasy" following closely behind.

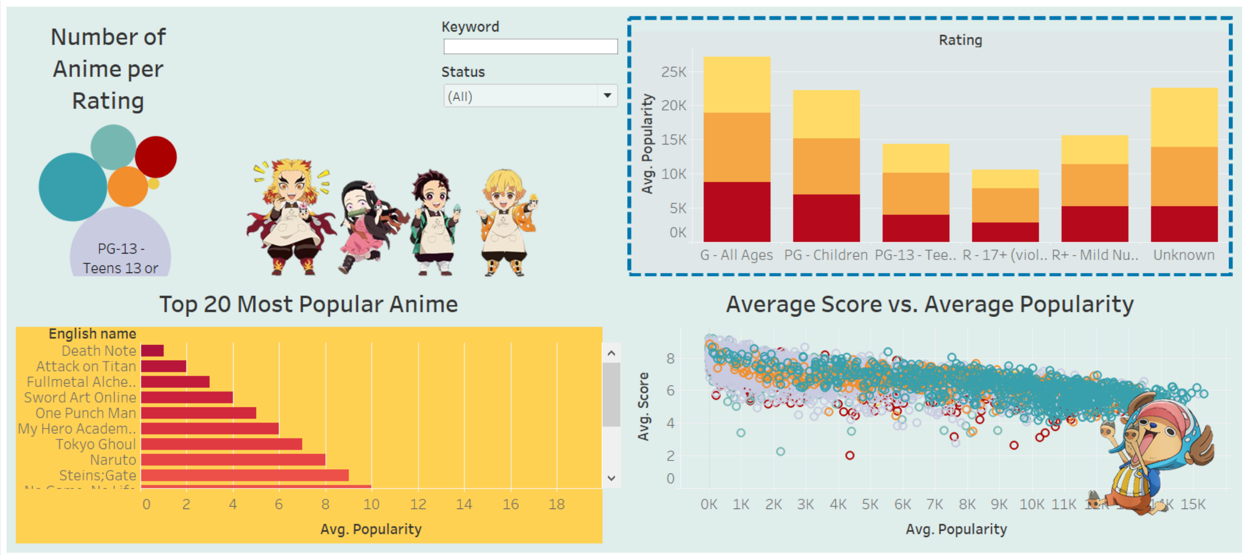
Overall, the visualizations confirmed a multitude of hypotheses for the dataset. TV is clearly the most popular form of anime, as all the other forms (ONA, OVA, Music, Movie, Special) aren't generally ranked as high, or are even produced for as long. Moving on, while the most amount of anime fall into the “PG-13” rating, the most popular anime go into the “R - 17+” category. (keeping in mind- the lower the “Popularity” number, the higher on the list it is. In the future, we can dive even deeper into the dataset and find other areas where we could clean the data further. Our data is derived from actual anime watchers, who gave their personal input online, so that can always be double checked in the future to confirm that all the information given is accurate.



User is able to obtain anime recommendation based on the input provided in the "Keyword" section. This is filtered based on the Anime Name, Genre and Synopsis.
User can further break down the recommendation based on the Status of the anime: Complete, Incomplete and Ongoing.

Top Anime per Score

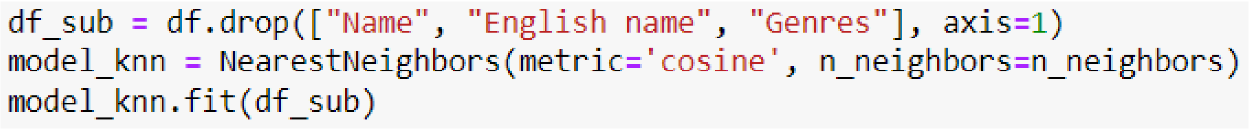
Name	Show	
Fullmetal Al..	True	9.190
Shingeki no ..	True	9.170
Steins;Gate	True	9.110
Gintama°	True	9.100
Hunter x Hu..	True	9.100
Shingeki no ..	True	9.100
Gintama'	True	9.080



Machine Learning

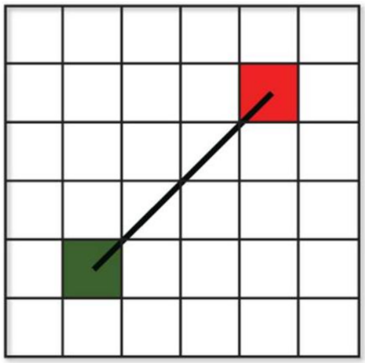
At the data preprocessing step, it was important to consider which features were needed to build the KNN model. Review score, maturity rating, and genres were the target features, and we were allowed to do normal one hot encoding on maturity rating. However, manual one hot encoding needed to be done for genres since multiple genres existed for each anime.

Afterwards, we built the KNN model using ‘NearestNeighbors’ instead of the ‘KNeighborsClassifier’. Instead of having a train/test split, the entire dataset post-processing was used to fit the model. The objective was to find the group of anime closest to the one selected in the function. Determining that distance is best performed by using the metric hyperparameter, cosine.



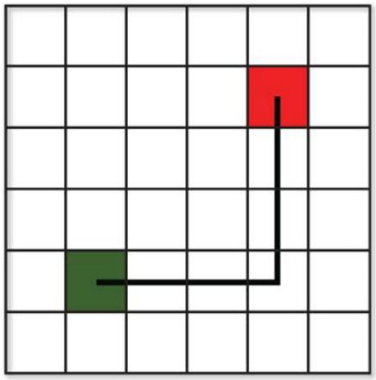
```
df_sub = df.drop(["Name", "English name", "Genres"], axis=1)
model_knn = NearestNeighbors(metric='cosine', n_neighbors=n_neighbors)
model_knn.fit(df_sub)
```

The default metric used for distance is euclidean which is simply the straight-line distance between two points.



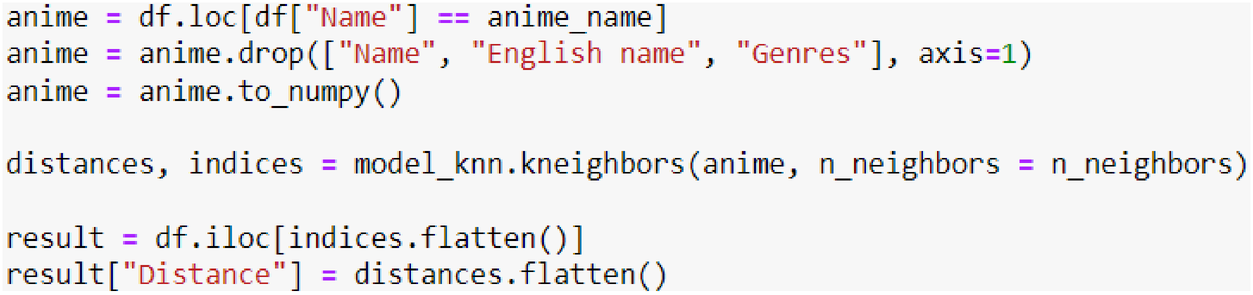
Euclidean Distance

The manhattan distance metric could be considered due to our dataset’s high dimension count. This method measures the sum of the absolute differences of their Cartesian coordinates.



Manhattan Distance

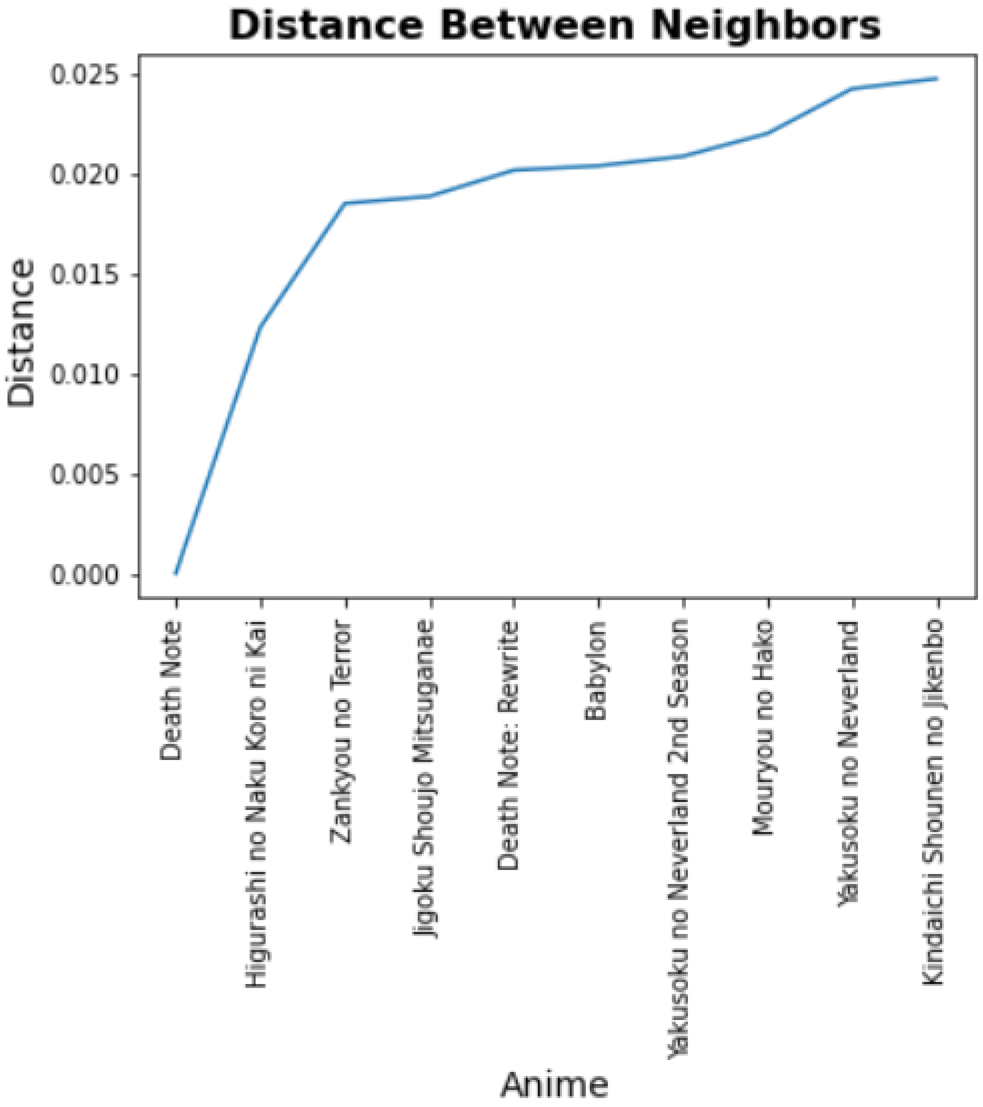
For the metric cosine distance, the cosine of the angle is measured between two vectors. Now, we’re observing if there’s a relationship between the lines’ distance and their direction. Since visualizing multi upon multi dimensional charting was not possible for our group at the moment, the KNN model provided a distance value using the selected anime as the point of origin and its nearest neighbors as recommended anime.



```
anime = df.loc[df["Name"] == anime_name]
anime = anime.drop(["Name", "English name", "Genres"], axis=1)
anime = anime.to_numpy()

distances, indices = model_knn.kneighbors(anime, n_neighbors = n_neighbors)

result = df.iloc[indices.flatten()]
result["Distance"] = distances.flatten()
```



Limitations and Future Work

There are always things we wish we could have done better when it comes to these projects. There never fails to be a list of things we wished we could have included in our final project. Sadly, it just was not possible due to time limitations. Given unlimited time, and unlimited knowledge here are some things we would have added to our project.

API- If given the opportunity then we would have loved to have updated data using an API from my anime list.

Website Design – Originally the web design concept in mind was to make a website in which the homepage had a stream like theme that would have certain anime playing their highlights using youtube embedded code. The site would also have a shared tab set of anime genre; having the top 15 animes of said genres and include their description, and links on where to watch the anime. However time was limited and a site like this was taking more time than expected.

Machine Learning - Add more inputs to the Recommender page by increasing feature selection and outputting a cleaner table. Experimented more with the hyperparameters and grid search to find possible optimizations with KNN. Instead of unsupervised KNN, use neural networks to process the huge dimension count and optimize.

Resources

Dataset -
https://www.kaggle.com/hernan4444/anime-recommendation-database-2020?select=anime.csv

KNN Model - https://www.kaggle.com/hernan4444/anime-content-collaborative-knn

What is KNN - https://realpython.com/knn-python/

Doing more with KNN - https://scikit-learn.org/stable/modules/neighbors.html

Distance Metrics -
https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html