Anime Dashboard

Maya Farmer
October 28, 2021

The data from this dashboard was retrieved from Anime Recommendation Database 2020 dataset on kaggle.com (link: https://www.kaggle.com/hernan4444/anime-recommendation-database-2020)
A pdf of this dashboard can be found in the ‘PowerBIDashboards’ repository on my Github.

There are 5 different csv files in this dataset, but I’m only working on the anime.csv file. This file contains a total of 17562 rows and 35 columns containing information on:

- Anime name
- score
- rank
- genre
- type
- number of episodes
- aired and premiere dates
- source (manga, original, video game, etc.)
- producers and studios
- popularity

This project in particular is a really fun one for me because I LOVE Japanese anime and manga. This analysis will focus on identifying some of the most popular anime to date. There was a single data table, containing the rows listed above, organized as a fact table in Excel (using Inset > Table).

Data Prep in Excel

The raw data in the anime.csv file are very messy. Most cells contain a lot of information and the data hasn’t been entered in a way that will make it useful for analysis. For example, each cell in the Genre column contains about 5 different genre instead of one. Also, in the duration column, instead of listing the duration in minutes (as a whole number) each cell contains additional text (e.g. 24 min. per episode).

I will go column by column to clean up the data.

Steps and Calculations

First, I’ll remove the Japanese name column.

Next, I want to add individual genre columns. I chose the following genre because these are typically my favorite anime: Action, Adventure, Comedy, Drama, and Mystery. I created 5 genre columns using the following formula:

> Action = IF(ISNUMBER(SEARCH("Action",D2)), 1, “”)
>
> This formula returns a 1 if it finds the word Action in the original Genre column. This was repeated for the remaining genre.

Then, I extracted Start Date and End Date Columns from the original Aired Column.
> I removed all “?” from the aired column using find/replace (~?)
>
> Start Date = IF(ISNUMBER(SEARCH("to",H2)), TRIM(LEFT(H2,FIND("to",H2)-1)), TEXT(H2,"MMM DD, YYYY"))
>
> This formula returns the first date in the cell if it finds the word “to” in the Aired column and trims/removes everything else after that
>
> End Date = IF(ISNUMBER(SEARCH("to",H2)), TRIM(RIGHT(H2,LEN(H2)-1-FIND("to",H2))), "")
>
> This formula returns the second date in the Aired column if it finds the word “to” in the Aired column and trims/removes everything before that. If it doesn’t find the word “to”, nothing is returned

Next, I created a Year Premiered column by extracting only the year from the Premiered column.

> Year Premiered = IFERROR(TEXTJOIN("",TRUE,IFERROR((MID(I2,ROW(INDIRECT("1:"&LEN(I2))),1)*1), ""))*1, "")
>
> This formula parses out the numbers in a string of text and joins them together. Multiplying the join by 1 converts the output from text to a number.

Then, I created a Duration in minutes column.
First, I removed all “.” from the duration column using find/replace

> Duration (min) = IF(ISNUMBER(SEARCH("hr",O2)), 1440*SUBSTITUTE(SUBSTITUTE(O2,"hr",":"),"min",":"), TEXTJOIN("",TRUE,IFERROR((MID(O2,ROW(INDIRECT("1:"&LEN(O2))),1)*1),""))*1)

> Similar to the previous formula, this one also extracts all the numbers in a column and joins them together. It also searches for the word “hr” and converts those cells with data in hours to minutes.

Finally, I transposed the Producer, Licensor, and Studio Columns.

> To get these, I use the text to columns in the Data tab and used the commas as delimiters. For the Licensor and Studio columns, I’m only interested in the first 2 entries in each cell (ex: if a studio cell contained TV Tokyo, Gallop, and Gonzo, I only want TV Tokyo and Gallop). For the Producer column I want the first 3 entries in each cell.

After cleaning the data in the main anime dataset, I used the data from the 5 genre columns I created to make a summary table. Here I wanted the total number of anime that were classified in each genre (=SUM(anime!E:E) and I wanted the average score of the anime in each genre (=AVERAGEIF(anime!E:E,"=1",anime!C:C)).

In Power BI

Because the anime and summary tables were both fact tables, no modelling was necessary.
I did however, reclassify the columns that contained dates, whole numbers, and decimal numbers.

Visualizations and Findings

As I stated earlier, I’m a huge fan of Japanese anime. One of the first plots I created was the Anime by Source donut chart. Source refers to whether the anime originated as a manga, video game, light novel or visual novel, etc. I thought that most anime would be sourced from manga, but actually, most anime are originals.
*I filtered the data in this plot to only show the top 4 sources.

Next I looked at the top 10 anime people are currently watching. The anime listed are a pretty good mix of older/classic series like Hunter x Hunter, Naruto, and One Piece (which is still airing) and popular new series like Shingeki no Kiogen (Attack on Titan), Black Clover, and My Promised Neverland.

I was also interested in finding out which producers and studios created some of the highest rated anime using a decision tree. First, I included anime score and filtered the data so that it only included anime with scores > 8. You can see that Aniplex and TV Tokyo are the top 2 producers creating highly ranked anime; as well as Bandai Visual. There also doesn’t seem to be a lot of overlap in the studios utilized by different producers. For example, A-1 Pictures works with Aniplex, but they don’t show up as a major contributor with any other producers.

The last 2 visualizations show how the popularity of anime has changed over time. The first area plot shows that the number of anime series or movies that release each year has increased significantly, particularly in the early 2000's. The popularity of anime has trended in a similar way over time as well, likely indicating why more anime content is being produced and released.

The second page of the dashboard compares the different genre I chose. I was really surprised that the average rating among the genres was basically the same, and that the Mystery genre had the highest score. The lists below each genre show the top rated series for each and actually gave me some great recommendations for the series I might want to binge next!