



# Final Report

## Top Video Games Reviews

12.17.2021

---

Fay Dennis

## Problem Statement

A data science project to review how well a game does based on its summary, scores, and user reviews. I want to see how all of these top performing games line up together. The different platforms that the games are played on, and the summary for the game may present problems because every game is different and therefore hard to quantify a review of the game. So the reviews of the games can differ very much from game to game and are very personal to user experiences but what interests me is if there are any similarities between game summaries, and how well they perform based on user/metacritic reviews. So this project is also going to test if I can create a best-selling game, based on reviews/performances of best-selling games in the past.

For this project a hypothetical client could be a company that publishes and creates video games. Once we take in all of these factors and develop a model, it could be used to produce a game concept that would be a hit and thus profitable for the company.

## Data Wrangling

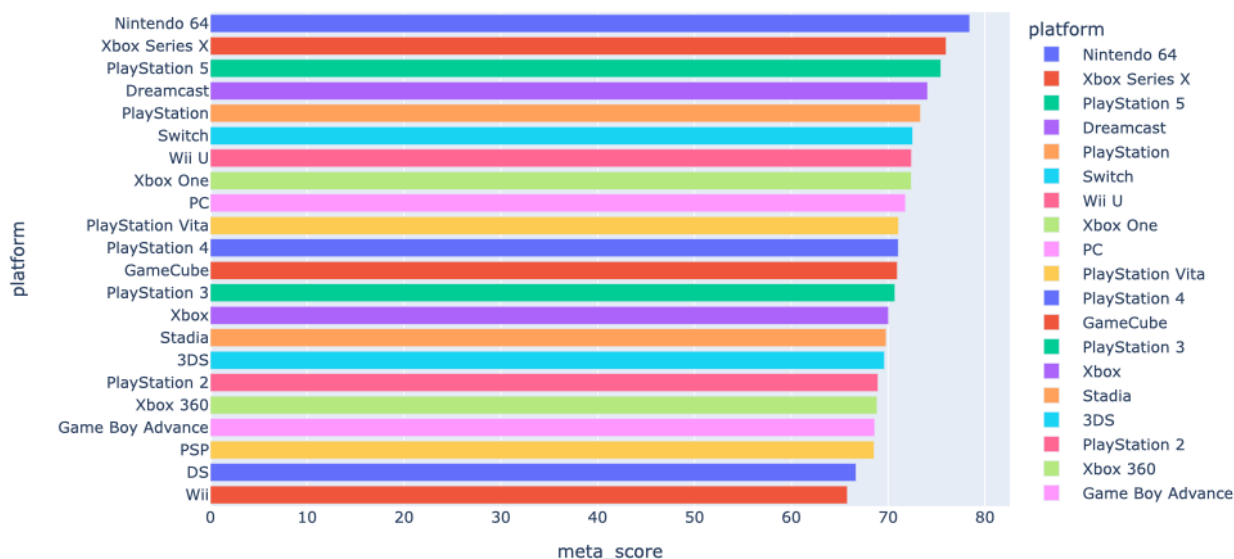
This dataset is scraped from metacritic.com, which user Deep Contractor put on Kaggle in CSV format. The dataset shows six columns, name of game, platform played on, release date, summary, meta\_score, and user\_review. 18,800 observations(rows) and 6 attributes(columns). It is quite big, however doesn't have that many columns but enough to work with. Most of the dataset are objects, specifically string objects or date time objects. The only numerical columns were the reviews columns.

There were some missing values in both the 'summary' and 'user\_review' columns. I decided not to fill in the 'user\_review' missing values with anything like an average and this was because it would have made the analysis of the comparison of the 'meta\_critic' reviews versus the 'user\_review' different further down the line. Because there were so few columns I didn't drop any of them. I also did not drop the rows with the missing 'summary' section. Since the dataset was pretty clean I then started to do some EDA.

## Exploratory Data Analysis

In the EDA section of my project I knew I wanted to do several things such as; distribution, platform performance, correlation and plotting them all to see what outcomes to draw an insight. First thing in my EDA I plotted the distribution of the reviews, so separately I plotted 'user\_reviews' and 'meta\_score' reviews. From this I found out that they were based on different rating scales from 0 to 10 and 0 to 100. I found their distributions to be pretty much the same from about a score of 50% through 90%. After distributing these reviews columns, I wanted to look at the correlation heat map and found that the two review columns have a solid 0.53 correlation.

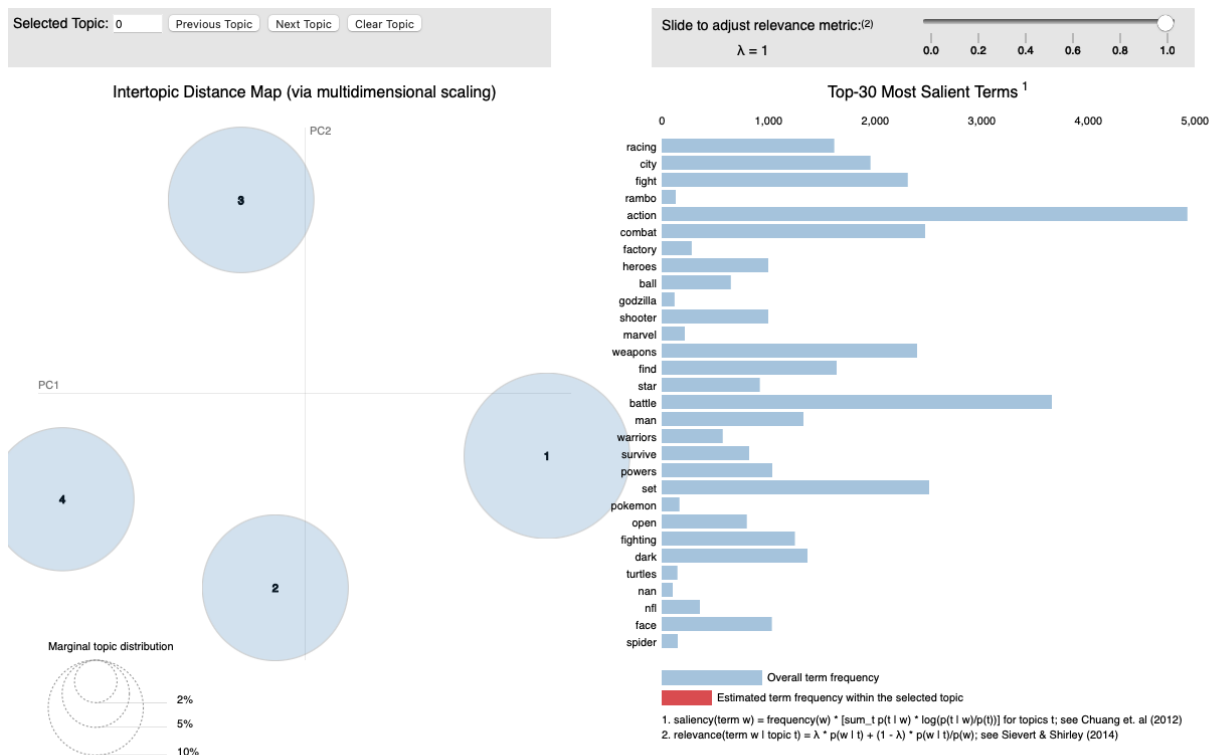
Even further into my EDA I wanted to see how a game performed based on the platform it was played on, since I am a Playstation fan myself I wanted to see how it did amongst cult-classics such as the Nintendo 64. First I did a simple groupby and later I plotted on a beautiful plotly.express plot. While the 'user\_reviews's top 3 performing platforms were: Nintendo 64, Dreamcast, and Playstation, the 'meta\_critic's top 3 performing platforms were: Nintendo64, Xbox Series X, PlayStation 5. It was interesting to see that Nintendo 64 was the top scoring in both of the 'user\_reviews' and 'meta\_score' reviews.



I also wanted to see what a scatter plot would look like in plotly so I thought I would plot the correlation between the reviews again. It was at this point in my EDA that I decided to stop and also realized that only doing modeling on the reviews was going to be boring and I instead wanted to focus on the summary section of the dataset.

## Modeling

In order to see what the top video games 'summary' section to develop a model to see the best performing game could be, I wanted to use text analysis of the 'summary' section. Luckily my mentor again was there to show me his code as to have a guideline of what to do in text analysis. He suggested the Gensim library, and pyLDavis library which I installed both of them. We also set up the stopwords and ran the model and visualized it. Then I found out that words normally used in games such as 'new', 'games', 'world' were way too frequent in the modeling. So I set up an extension for the stop words to filter out these words. And the visualization showed categories that are popular in games are as follows: action, experience, adventure, characters, and battle.

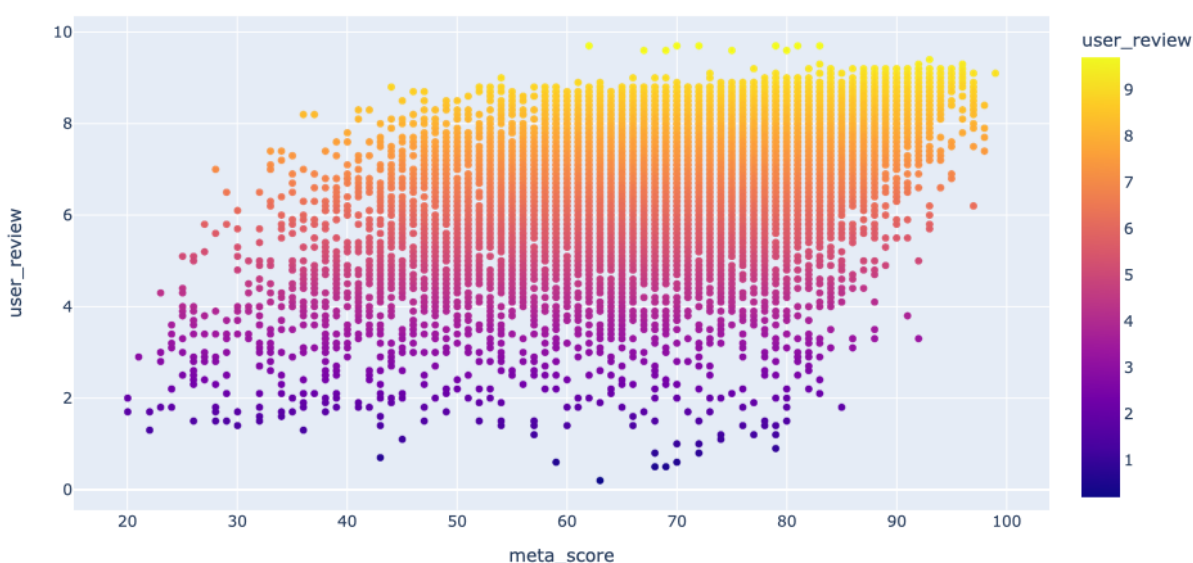


So if a company wants to make a best selling game based off the Top Video Games of 1995-2021 they should have a game that is an experience in action and adventure, but also have a sort of epic battle with lots of characters to interact with.

## Conclusion

In conclusion it was really fun to see all the different things I did in the EDA and the modeling, specifically seeing the reviews based on the platform. It was amazing to see cult classics like 'Zelda', 'Soul Calibur' and other beloved titles. It was very interesting to see the genres in the text analysis that we got through an unsupervised machine learning model. All in all it's really fun to see metrics on a dataset that interests you, specifically a hobby such as video games.


So again if a company wants to make a best selling game based off the Top Video Games of 1995-2021 they should have a game that is an experience in action and adventure, but also have a sort of epic battle with lots of characters to interact with.



## My Process

For this project I started with a dataset that I got from doing a search on Kaggle. This dataset interested me because I myself play a lot of video games, but I wasn't allowed to as a child through high school, and it was very interesting to see what video games were popular during my childhood.

The dataset specifically looks at the most popular video games from 1995 through this year 2021, so I feel like this is a perfect dataset to see what I was missing!




There are 18,800 observations(rows) and 6 attributes(columns). I didn't end up dropping any of the columns because it was already very limited. I did find that there were 2 of the 6 columns that had missing values, the `summary` column and the `user\_review` column did have missing values. The missing values in the `user\_review` column I ended up just leaving as Null because I didn't want to fill them with the average and mess up anything down the line.

Most of the dataset were categorical features(objects), but that's ok because a video game title would only be a string object. There was no need to convert those objects into numerical features.

In the EDA portion of this project I wanted to see the distribution of the different reviews. I was surprised to find out that `user\_reviews` only ranged from scores 6-9 (based out of 10). This means that almost all of the games did way better than getting a failing review of 5/10. Then I looked at the distribution of the `meta\_score` reviews, but before that I did some research and found that these reviews were based on player's personal websites reviewing the game and then imported as the `meta\_score` or "professional" reviews. So the distribution of the `meta\_score` reviews was mostly between 50-90 (based out of 100). This is interesting because the so-called professional reviews seem to be a bit more critical of the top video games of this time period. I then went and did a correlation map and then a correlation heatmap. The heatmap shows the correlation between the `meta\_score` and the `user\_review`, which was a solid 0.53 correlation. Yes it is a positive correlation but still not a perfect 1.

Even further into the EDA I wanted to see what the reviews were like based on the platform or console they were played on. My wonderful mentor Jeff suggested the Plotly library for python! It is very useful and also very beautiful! You can select, zoom in or out, and so many more things!! Anyway I wanted to plot the `meta\_score` and the `user\_review` based on the platform, so they are two separate plots. But at first the plot was not showing in descending order which platform had gotten the highest reviews. I then found some code to sort the platforms in a descending categorical order. While the `user\_reviews`'s top 3 performing platforms were: Nintendo 64, Dreamcast, and Playstation, the `meta\_critic`'s top 3 performing platforms were: Nintendo64, Xbox Series X, PlayStation 5. It was interesting to see that Nintendo 64 was the top scoring in both of the `user\_reviews` and `meta\_score` reviews.

I also wanted to see what a scatter plot would look like in plotly so I thought I would plot the correlation between the reviews again. It was at this point in my EDA that I decided to stop and also realized that only doing modeling on the reviews was going to be boring and I instead wanted to focus on the summary section of the dataset.



In the preprocessing, training, and modeling section of the project I wanted to continue my analysis of the summary section to see if we could do some text analysis. Lucky my mentor again was there to show me his code as to have a guideline of what to do in text analysis. He suggested the Gensim library, and pyLDAvis library which I installed both of them. We also set up the stopwords and ran the model and visualized it. Then I found out that words normally used in games such as 'new', 'games', 'world' were waaaaay too frequent in the modeling. So I set up an extension for the stop words to filter out these words. And the visualization showed categories that are popular in games are as follows: action, experience, adventure, characters, and battle.

In conclusion it was really fun to see all the different things I did in the EDA and the modeling, specifically seeing the reviews based on the platform. It was amazing to see cult classics like 'Zelda', 'Soul Calibur' and other beloved titles. It was very interesting to see the genres in the text analysis that we got through an unsupervised machine learning model.

All in all its really fun to see metrics on a dataset that interests you, specifically a hobby such as video games, I still believe people who love video games and grew up with certain classics have probably played one of theses games at some point, and in video-games should be accessible to everyone without gate-keeping.