**Data Analysis Case Study #2 Streaming Content**To practice more Data Analysis skills, I will make a second case study. I will take an existing dataset and make a new realistic scenario I will put myself in to try and solve a data problem. In this situation I would work for a fictional streaming company Watchalot trying to look at trends in the current competition and how they can use it to their advantage, taking a look at Netflix data to compare. I will follow the steps to come to a conclusion.

**ASK:**

Context: Watchalot is an up and coming streaming company who wants to enter the field but wants to look at competition beforehand to understand their competition and see how they excel and to use it to benefit their endeavors and to potentially make up for blind areas in the market. I am tasked to look at the biggest competitor, Netflix, to judge their success in movie listing and licensing to make an informed decision to shareholders.

**Business Task:**How can Watchalot decide on what is currently in demand in the streaming field and look for titles and properties to add to their service? And find genres in demand to fill a niche that isn’t actively held.

* What movies had the highest rating and views
* What genres and style are most popular
* Are there any genres that are popular yet underused

**Stakeholders:**Watchalot Media Department – The people who are responsible in the company to try and acquiesce media from all sources to stream in their service.

Watchalot Executives – The higher up members of the company responsible for overseeing all decisions and making approvals for any company actions. As the main stakeholders they are the focus of the presentation.

**Report Requirements –**

1. A clear summary of the business task

2. A description of all data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

6. Your top high-level content recommendations based on your analysis

**Relevant Questions:**

* What movies and shows does Netflix have that keep people watching and subscribing to them?

* How can Watchalot use this information to get into the streaming market?
* Are there any niches in the industry that Watchalot can fill

**PREPARE:**  
This dataset was provided by Kaggle user **OctopusTeam**. They got the data from IMDB or the Internet Movie Data Base which provided much of the ratings and votes in it. IMBD is a very valuable and reliable internet archive of movie information so the data is validated. The dataset is also licensed with a **CC0 1.0 Universal** Public Domain license. This means that anyone can use the dataset provided without any rights or residuals, so its free to use for this Case Study.   
  
Regardless the Case Study is for a fictional company, so it does not violate any laws or regulations and is only used for analysis purposes

The data is contained in one large .CSV file. It has a large list of Netflix shows and movies. It contains roughly 20584 rows containing the name of the media listed and the rating and information from it. For Watchalot’s case what we’re really looking for is the genre, rating, and number of votes to see which type of media people like and are watching to see what we can include in our service. Then we can observe the trends and see how we can emulate that success in our service.

The columns are divided into

* Title – The name of the media
* Type – If the media is a movie or tv series
* Genres – The Genre or multiple Genres the media is categorized in
* Release Year – The year the media came out and released to the public
* IMDBID – The Id of the media on IMDB
* IMBD Average – The average score of the media on IMBD
* IMBD Number Of Votes – How many votes from users the media had received
* Available Countries – Where the media is currently available to watch on the service

The release year, the IMBD ID, and the Available countries are largely irrelevant to the case study. So we need to focus on the number of votes and the ratings to see the highest and most valuable media the service has and the correlating Genre to pick up on any trends. One of the goals is to compare the most popular genres with the highest ratings to see any correlation. Available countries may be important to see what’s trending in the biggest and most valuable markets.

**PROCESS:**

Since both files are relatively small, Spreadsheets will be the best way to organize and manipulate the datasets.   
  
But the datasets must be processed before the Analysis stage.

The IMBD dataset has 20584 rows and a few cells spots are left blank. We use COUNTA to see how many empty cells exist in the important columns.

A screenshot of a computer

Description automatically generated  
We see that there are quite a few rows without information. For the best data we need to remove these from the dataset. We subtract the total by the lowest value, 20584 – 18958 which equals 1626 so we need to remove that many rows from the dataset if they contain no values.

Do to so we select each column, Title, Genre, Rating, and NumVots, and use the Search and Find button and go to Special and select Blank. This selects all the blank values in the column. With them selected we go to the delete button and select the Delete Sheet Rows option to get rid of the rows with blank values. We be thorough and do the same for four of the important columns.

A screenshot of a spreadsheet

Description automatically generated

Now all of them have the same value of 18957 ensuring that they have no blank values.

We check for the same blank values in the Netflix Engagement sheet using the same method. But it returns no null values outside of the release date which doesn’t matter for this study.  
  
We also need to quantify how many values are in each genre. So we compile every genre listed in the dataset and try to use a **COUNTIF** formula to count every time in the genre column a particular genre was listed. But since the column is organized with multiple genres, it can’t work with just a simple COUNTIF as it will only mark it if the cell only contains the value. So we use a specific formula **COUNTIF(range, "\**text*\*")** with the asterix being added so that Excel will mark if the selected text appears in any position. Doing so to every genre listed we get.

A screenshot of a computer

Description automatically generated

With this we have a better idea of the quantity of each genre has represented in Netflix. We can already make assumptions which is better left to the Analysis Section.

**ANALYSIS:**

For the business question, we need to find what is the most popular media Netflix has by IMBD Average and Vote Number. And then find which Genre has the highest ratings and votes to see the most on demand and rated to invest in making. We can also potentially compare it to the amount of media in that genre compared to how high the rating is to see if there’s any niche market that Netflix hasn’t taken advantage of.

To find the highest rating media is easy. We just select the Sort By Largest button in the Sort and Filter tab over the Rating column.

A screenshot of a computer

Description automatically generated  
Here one who is knowledgeable enough about movies and tv can see some obvious selections that make sense for the highest ratings such as Breaking Bad, Avatar the Last Airbender TV, and The Dark Knight. Also, usually the most notable and recognizable properties also have the higher number of votes.

We also should see the highest number of votes to the media. We do the same thing but instead sort the NumVotes column from largest to smallest.

A screenshot of a computer

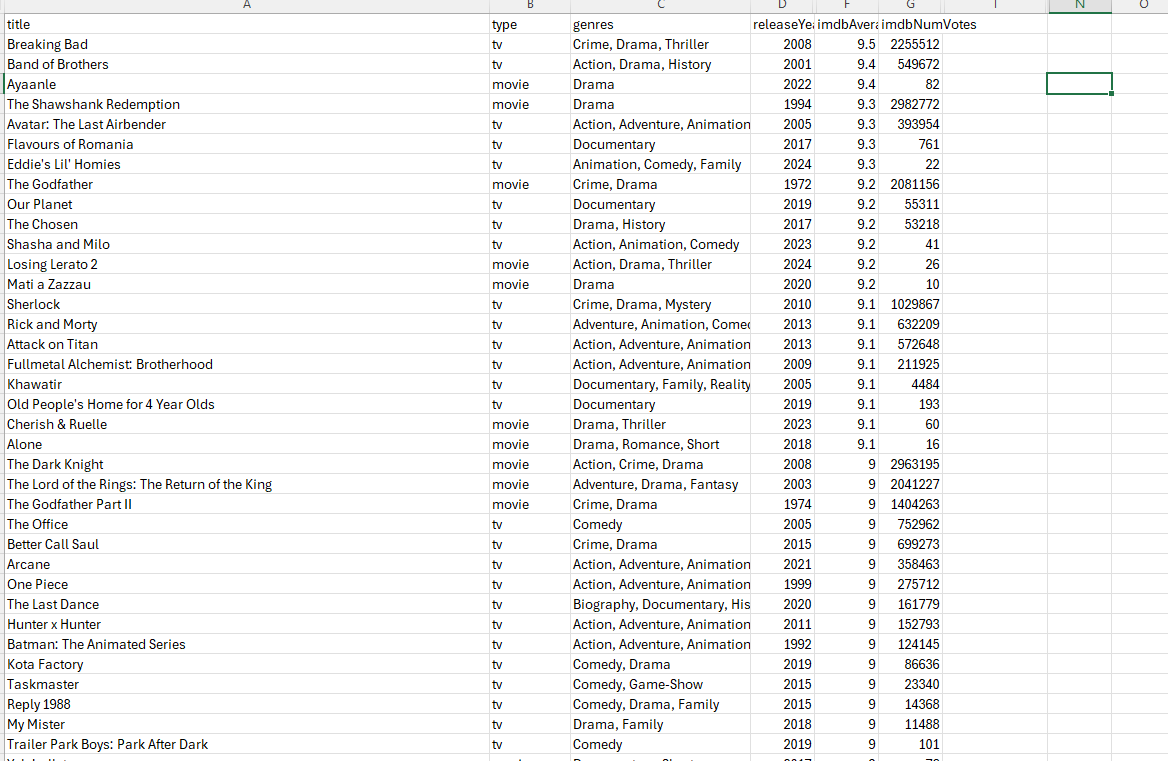
Description automatically generated  
With the table arranged by largest number of Votes we also see that most of the larger number of votes are very popular and noteworthy movies and tv shows. Quite a few also are near the top of both lists like Breaking Bad.

But we want to try and find the highest rated and the highest voted media. So we make a new custom sort. First we sort by the highest ratings and then by the highest votes in their respective columns.

A screenshot of a computer

Description automatically generated

Thus this sorting gives us the highest rated and highest voted media at once



This may be important for later analysis so let’s copy and save the data about 7 into a separate excel sheet.

But we also need to find out how much media has high ratings and votes and the corresponding genres. That way we can get a better idea on what types of media we should make and acquire. In order to do this we do another COUNTIF function with COUNTIFS to have multiple critera, but we also add extra parameters to check Column F if the rating is greater than or equal to 7.5 and if the NumVotes is greater than or equal to 10,000. To do this we make a new formula = **=COUNTIFS(C:C, "\*genre\*", F:F, ">=7", G:G, ">=10000")** and fill in for every Genre.

A screenshot of a spreadsheet

AI-generated content may be incorrect.  
So with this we now have the amount of media that is 7 or over and above 10000 votes. We can see a stark drop compared to the amount of media to the media that fit the criteria. We can already make judgements, but to play it safe its best that we also use the same formula but instead to check the individual requirements to see how the data skews by votes and ratings.

So to check by only media with a rating 7 and over:

A screenshot of a computer

AI-generated content may be incorrect.

And to make a second column for votes 10,000 and over:

A screen shot of a grid

AI-generated content may be incorrect.

We have a much more detailed idea of what movies are the most popular and highly rated but in order to get the best understanding of it compared to the total we convert them into percentages. To do this we make a formula to divide the countif by the total amount and then convert the form to percentage from decimal.

A screenshot of a spreadsheet

AI-generated content may be incorrect.  
Now with the percentages we have a much better idea of how successful each genre is in the ratings and votes. This allows us to make new judgements

* **Action, Adventure, Sci-Fi, Horror, Mystery, have very high number of votes to the total amount on the dataset**
* **Animation, Music, and especially Documentaries have low number of votes to the total on the dataset.**
* **Animation, Biography, and Documentary have high ratings compared to the total.**
* **Action, Adventure, Animation, Biography, Crime, War, and Western have a good amount of both high ratings and votes.**
* **Despite Comedy having decent ratings and votes its combined statistics are low**
* **Documentaries have the highest ratio of high rating to low votes**
* **Horror has the highest ratio of low ratings to high votes.**

With this data we have enough to start to make visualizations and graphs to better showcase the data found.

**SHARE:**

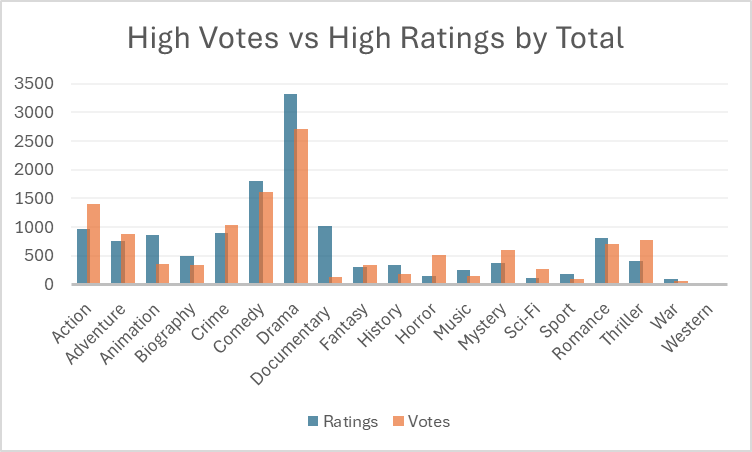
Our shareholders need visualization aide in our report to better understand the points gathered from the dataset we’re given. As such we need to make graphs from our data. Luckily Spreadsheets already gives us a great option for making graphs from the existing data so we don’t need to export our data into another program.

First we just need to make a bar graph to show off the total number of media documented by the dataset in the IMBD record. So that the shareholders have a good idea of the media that’s documented.

Here we get a better idea of the amount of media listed in the dataset. Now we also have to make similar graphs showing the amount of votes and ratings compared to the total. We’ll use a stacked Column chart to show the proportions of the data compared to the total. These bars are stacked on top of each other and not inside of each other.

Now with those we need to make a chart for the Countif column of high votes and high ratings. Here we get a better proportion of both.

So now we’ve compared the different statistics between the total. We can see a stark difference in some and a minimal difference in others. But more importantly we need to show the data between the votes and the ratings in a visual graph. This is one of the most important elements we need to. address. We’ll use the 7 or Higher Rating for more broad selections.



Here the genres with higher votes and ratings are directly showcased. This leads a better visualization of which genres excel in which categories which is important for analysis. Such as Action, Thriller, Crime, Horror, and Mystery having more votes than ratings and Drama, Comedy, Romance, and Animation having more ratings than votes. But its best that we also visualize the percentage of each genre.

Here it is much easier to compare the statistics from votes and ratings to the total. But we’ll also need to compare all the percentages at once. So we make one more graph with High Votes, Ratings, and Both.

With the charts made and presentable we can move on to the next step and finalize our findings and data into a visual presentation for our stakeholders to give the appropriate call to action.

**ACT:**So with our data compiled we will present our findings in the form of a Power Point. We’ll state our business question, give the appropriate data context, and then present our findings with help of the graphs we made to visualize the data and then at the end conclude with business strategies and suggestions. The slides are shown in the order they will be presented in.

To see the report please open the Watchalot Data Report Powerpoint Presentation to view the information inside.