TERM PROJECT

Netflix Movies and TV Shows 2019

https://www.kaggle.com/datasets/PromptCloudHQ/imdb-data/data

Create a dashboard: Power BI and Jupyter

MILESTONE 1

BACKGROUND

Provide a general background on the scenario; why are you interested in this specific scenario?

- I was searching for a set of databases from different sources on the internet and came across Kaggle, which has a ton of datasets available. After navigating the different types of datasets, I chose the following dataset regarding tv shows and movies available on Netflix as of 2019.

RESEARCH QUESTION/PROBLEM STATEMENT

State the main question you are trying to answer or problem you are trying to solve. Explain why this question is relevant.

- Performance: Analyze the overall performance of the movie vs. shows added to Netflix. Are there any significant changes in metrics like type and rating within the countries available?
- The dataset contains the following data points: show_id title, director, cast country,date_added,release_year,rating,duration, listed in, and description type. This information is relevant to the question, and I believe I can get a lot of information to complete the milestones and generate diagrams.

PROPOSED GOALS

Provide some suggested goals for your analysis; these goals can change later in the next milestone and be adjusted along the way. Even after exploring the data, you may come up with new ideas.

Regional Analysis: I want to investigate differences in the types (movies vs. TV shows) and ratings of content added to Netflix across different countries. Also, explore how the popularity of content varies by country.

Rating Distribution: Being able to visualize the distribution of ratings for both movies and TV shows. Also try to identify any significant differences in rating distributions between the two types of content.

Monthly Distribution: Add visualization of the number of shows released each month throughout the year.

PREVIOUS RESEARCH

Identify at least one previous study in the literature that is related to your scenario.

- With the dataset publication there are different dataset notebooks regarding the findings with the dataset. Here is a list of a few published notebooks:
- https://www.kaggle.com/code/obayprogrammer/tmdb-movies-full-eda
- https://www.kaggle.com/code/yunasheng/netflix-movies-and-tv-shows-analysis
- https://www.kaggle.com/code/abdelrahmanaboelnaga/eda-for-netflix-movies-and-tv-shows

DATA COLLECTION

Explain how you are planning to obtain or collect the dataset. Identify the data set(s) you want to use.

Indicate the citation for your data and a link to the source. Is the size appropriate? Is it too small, too large? What about the data's quality? Does the data set have many missing values, extreme observations, or outliers?

Also, make sure that you comply with any regulations. Some websites do not allow the harvesting of their data. Other sites have APIs to obtain data, but they have limitations, or you may need to pay. Other datasets are public but could be unreliable.

- The dataset that I am obtaining has already been constructed by the user who published the dataset within Kaggle. The dataset is downloadable via csv, which can be imported into a compatible tool to further analyze the data. This dataset is composed of the 5,000 most popular movies and shows that were added to Netflix in 2019. The dataset contains

the following data points: show_id title, director, cast country,date_added,release_year,rating,duration, listed in, and description type. I believe that the amount of data is enough to reach this milestone, and I plan on doing some data cleaning regarding some data fields that aren't relevant. Regarding complying with regulations, all I had to do was create an account on Kaggle to have access to the dataset.

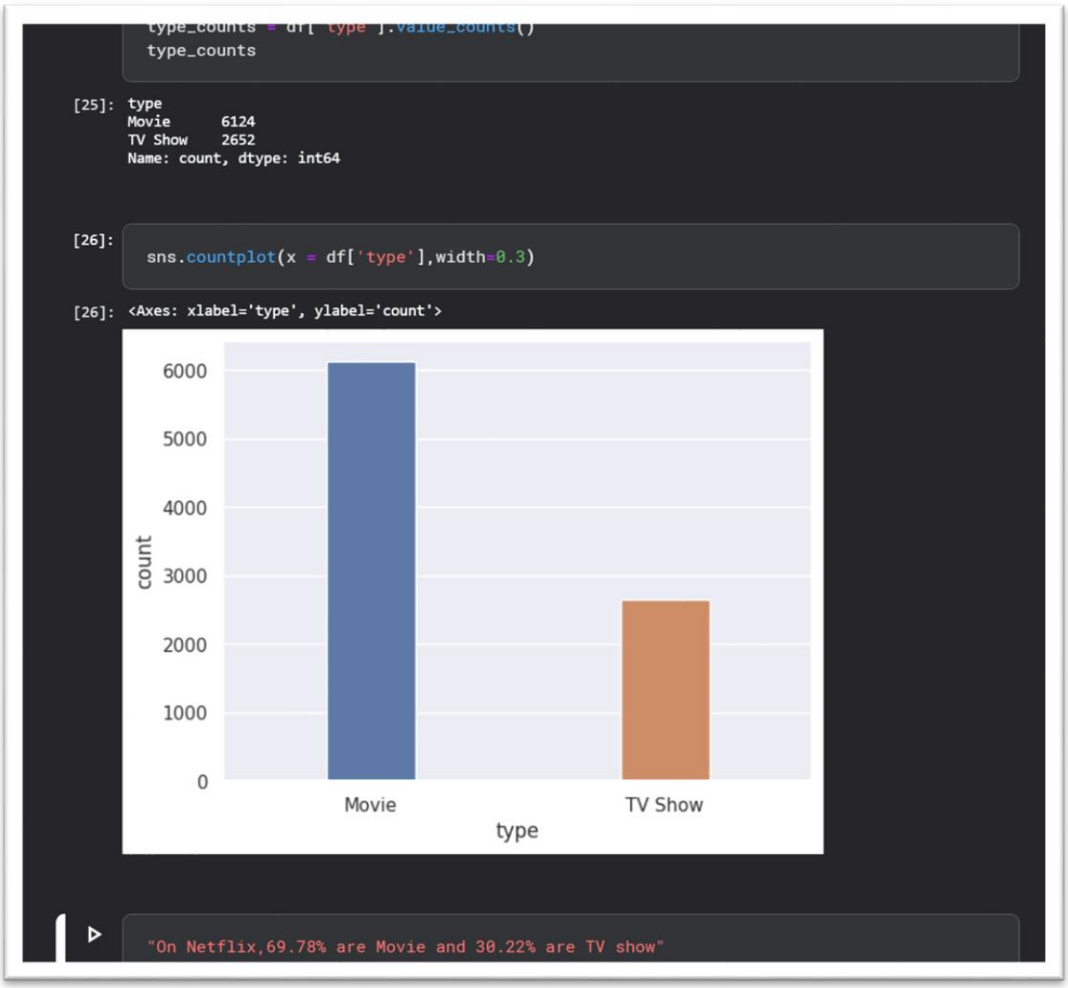
DATA CLEANING

Explain any data cleaning: transformation, filtering, imputation. Give a summary of the cleaning/joining of data that you expect to do upfront.

- The data is formatted correctly and has enough information to answer my research questions and data collection. There are some fields within the dataset that do not seem to be relevant to the types of diagrams that I will be outputting from the software. I plan on going through the dataset and deleting the cast and description fields, as the results from that data will be very minimal for comparison. Additionally, within different other projects I analyzed, there are suggestions on data cleaning, which consists of fixing values and deleting irrelevant data, which I took upon and adjusting.

EXPLORATORY DESCRIPTIVE ANALYTICS

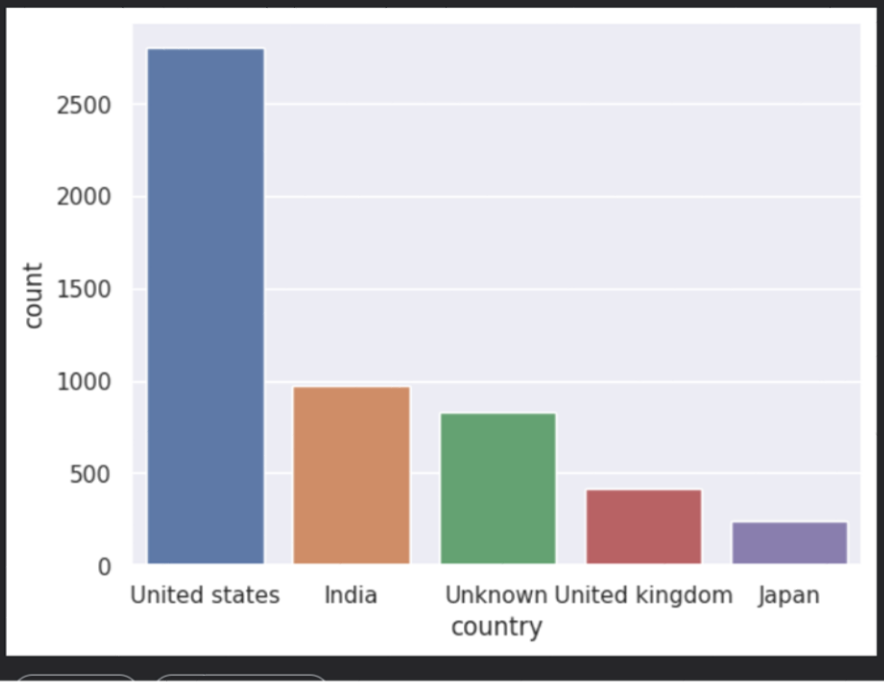
Become familiar with the data set. Explore the data by running descriptive analytics using the software of your choice. Also, create visualizations as appropriate, including histograms, box plots, correlation charts, and line charts suitable for each relevant variable. Provide a summary of the findings based on descriptive analytics. Based on the descriptive analysis, try to predict some of your results. Do the visualizations support your research question? For each variable, describe the type of variable (categorical, ratio….), range, and applicable statistical measures.



- Using Jupyter that's integrated within Kaggle, I configured the dataset to display a comparison between the number of movies and TV shows that were added to Netflix throughout the year. I also picked up on methods to filter out data that isn't relevant. As a result, the data shows that 6,124 movies were added, roughly 68.78%, and 2,652 shows were added, roughly 30.22% of the entire dataset.

```
top_countries = country_counts.nlargest(5)
top_countries
sns.countplot(x=df['country'], order=top_countries.index, width=0.8)
```

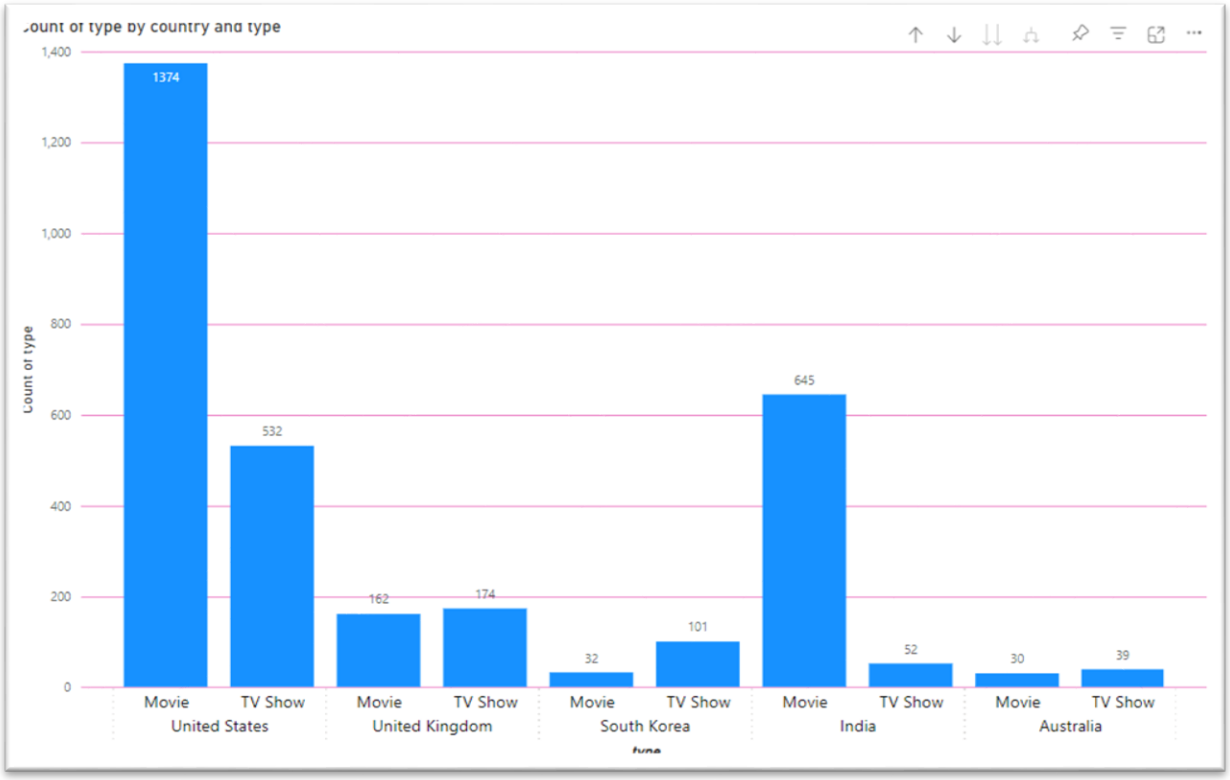
```
[37]: <Axes: xlabel='country', ylabel='count'>
```



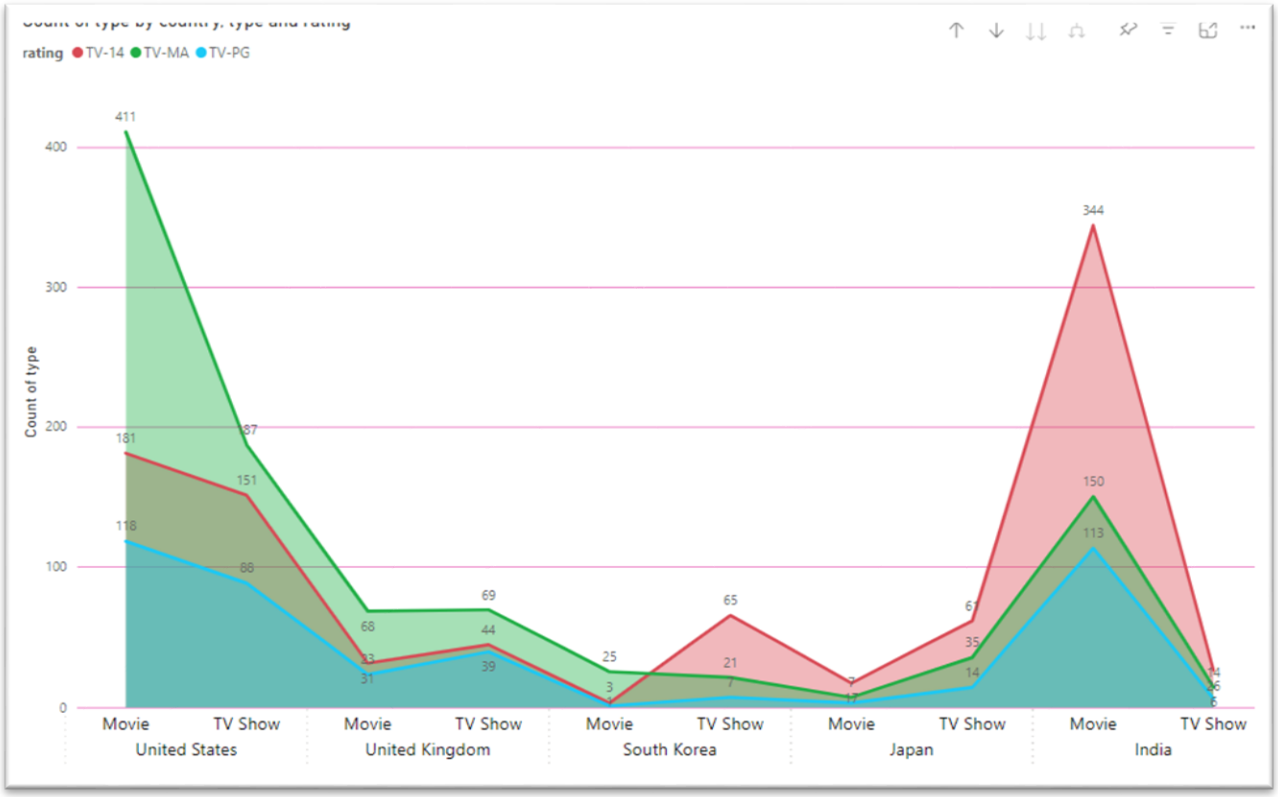

```
[30]: country_counts = df['country'].value_counts()
      country_counts

[30]: country
      United states     2803
      India              972
      Unknown            829
      United kingdom     418
      Japan              243
```

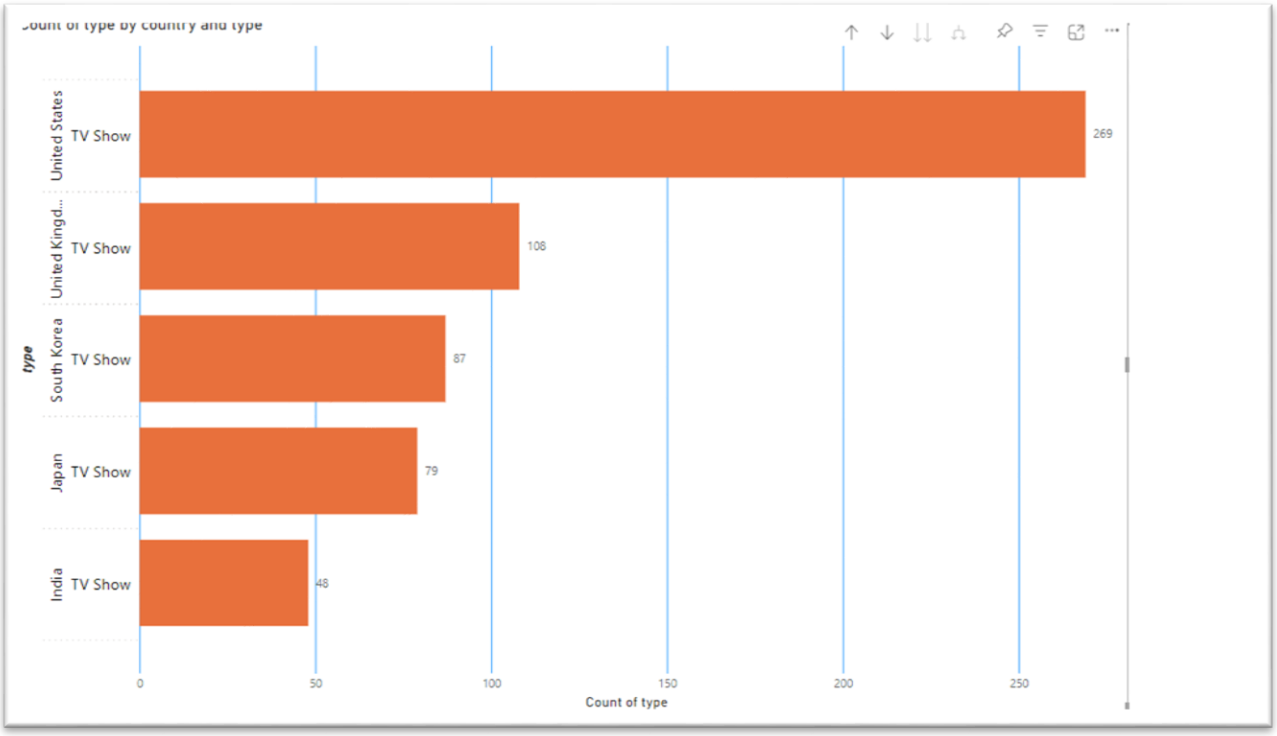
- With further experimenting with Jupyter, the following chart displays the top 5 countries that had the most shows produced within their country. Overall, the United States has the highest number of shows produced that are available on Netflix.



- The next three diagrams were created using PowerBI. In this scenario, I wanted to categorize the movies and shows within different countries. I went through and added different countries, as in the Jupyter, and I can see that the United States still has a significant amount compared to the rest of the world. India is the second most popular, followed by the United Kingdom, South Korea, and finally Australia.



- Next, I wanted to analyze the ratings of movies and shows within each following country and see what I could find. The following ratings were selected: TV-14, TV-MA, and TV-PG. In the United States, the movies and shows added throughout 2019 rated as TV-MA were higher than the other two. Next, the United Kingdom has the same outcome as the United States. Next, South Korean movies added were rated TV-MA more than the others, but regarding shows added, TV-14 was much more than the rest. A shift is in Japan; both movies and shows added were TV-14 compared to the rest. The same goes for India; although there isn't a significant difference in shows, movies are much more popular.



- With further analysis within the shows category, I went ahead and analyzed the shows with one season, as it was the most popular of all countries. As a result, the United States is the leading country with added shows to Netflix with one season compared to the rest of the countries. As a company created in the United States, it's expected to have a greater distribution of shows and movies compared to the rest, but it's still interested in seeing what origin everything else comes from.

PROPOSED SOFTWARE TOOLS

Identify the software tools you are planning to use for analyzing the data. Explain if you have installed them on your computer or if you are planning to access them online. Explain how you are planning to practice and gain skills with the software.

- Throughout this course, we have been able to use different software to analyze data and meet milestones. I plan on using Power Bi and a Jupyter notebook that is integrated with Kaggle to create diagrams for explanation. I will trace back to assignments and notes that I have done throughout this course to use as a reference to create diagrams with this dataset.

PROPOSED METHODOLOGIES

Identify the methods you are planning to use to analyze the data. Is the methodology appropriate for the type of data?

- In this dataset, dimensions I could include are attributes like genre, month, or even descriptive attributes like language or country of origin. In terms of measures, I plan on including metrics like movies vs. shows, and ratings. Overall, the type of analyzes on this milestone is more towards a categorical and comparative analysis.