

Certificate in Introductory Data Analytics

Project Report

Disney+ Data Analysis

Written By: Peter Brady

Date of Submission: April 19 2023

Table of Contents

1. GitHub URL Page 3
2. Abstract Page 3
3. Introduction Page 3
4. Dataset Page 4
5. Implementation Process & Results Page 4
6. Insights Page 11
7. Benefit of Machine Learning Page 11
8. References Page 12

List of Figures

Figure 1: Importing Python Libraries

Figure 2: Importing Dataset into DataFrame

Figure 3: Total Number of Movies & TV Shows

Figure 4: NaN Values Present in DataFrame

Figure 5: Percentage of Missing Data

Figure 6: Removing NaN Values

Figure 7: Percentage of Movies vs TV Shows

Figure 8: Distribution of Ratings on Disney+

Figure 9: Distribution of Genres on Disney+

Figure 10: Create Function for Collecting Actor Data

# **GitHub URL**

<https://github.com/UCDPAPeter/UCDPA_PETERBRADY>

# **Abstract**

In this project, we will take a look at some important Disney+ data in order to understand what is best for their business. The data that we will analyse from Disney+ data are as follows:

1. understand what content is available
2. understand the similarities between the content in relation to ratings
3. understand the network of actors
4. understand what genres are being focused on
5. how Disney+ can improve its overall service

# **Introduction**

Data analytics can help companies harness their data and use it to identify new opportunities. That, in return, can lead to smarter business moves, more efficient operations, higher profits and happier customers. Businesses, like Disney, that analyse their data can benefit in many ways such as:

1. **Reduces Costs**

‘Big data technologies like cloud-based analytics can significantly reduce costs when it comes to storing large amounts of data and data analytics help organizations to find more efficient ways of doing business’ (Ahmad, 2018).

1. **Makes Faster & Better Decisions**

‘The speed of in memory analytics combined with the ability to analyse new sources of data, such as streaming data from smart devices helps businesses analyse information immediately and make fast, informed decisions that optimise the overall customer user experience’ (Ahmad, 2018).

1. **Develops New Products & Services**

‘Being able to gauge customer needs and customer satisfaction through analytics empowers businesses to give customers what they want and when they want it. With data analytics, more companies have an opportunity to develop innovative new products to meet their customers’ changing needs’ (Ahmad, 2018).

This project will focus on Disney’s streaming service as a use case. Disney+ is one of the largest providers of online streaming services. It collects large amounts of data because it has a large subscriber base.

We can analyse a lot of data and models from Disney+ because this platform has consistently focused on innovation based on Disney’s business needs and by adapting its classic business model to include focusing on growing their on-demand movie and TV show streaming service.

# **Dataset**

The dataset I used for this project is data consisting of TV shows and movies that can be streamed on Disney+ when the online streaming service was launched in 2019. The Disney+ dataset is provided by and stored on the Kaggle.com website (*Disney+ Movies and TV Shows*, 2021).

# **Implementation Process & Results**

**Note:** Where possible, comments have been added to the python script to outline some of my thought processes.

1. **Importing Python Libraries Needed for this Project**

As seen in Figure 1, firstly we need to import the python libraries needed to for this project.

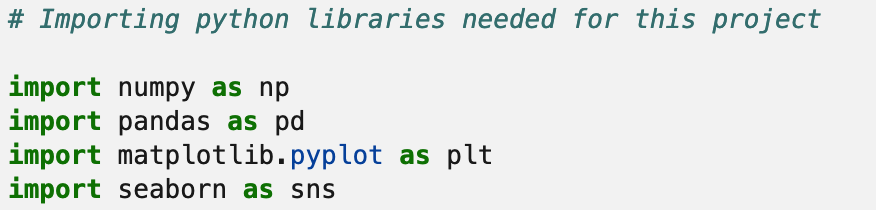


Figure 1: Importing Python Libraries

1. **Importing Dataset into DataFrame**

As seen in Figure 2, next we need to import the dataset into the DataFrame called ‘dataset1’.

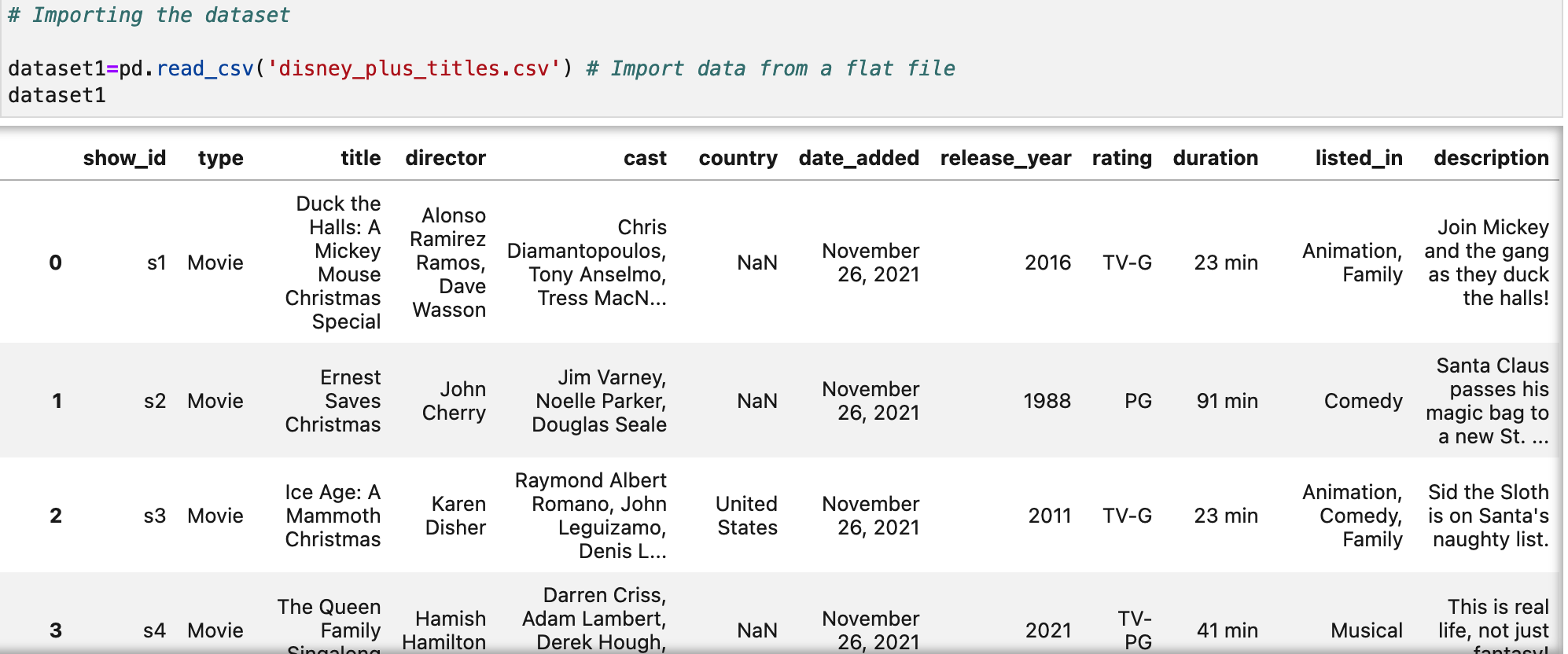


Figure 2: Importing Dataset into DataFrame

# **Handling Missing Column Data**

As seen in Figure 2, the DataFrame contains missing data or NaN values. First, we can check the total number of movies and TV shows on Disney+ as seen in Figure 3. There is a total of 1450.

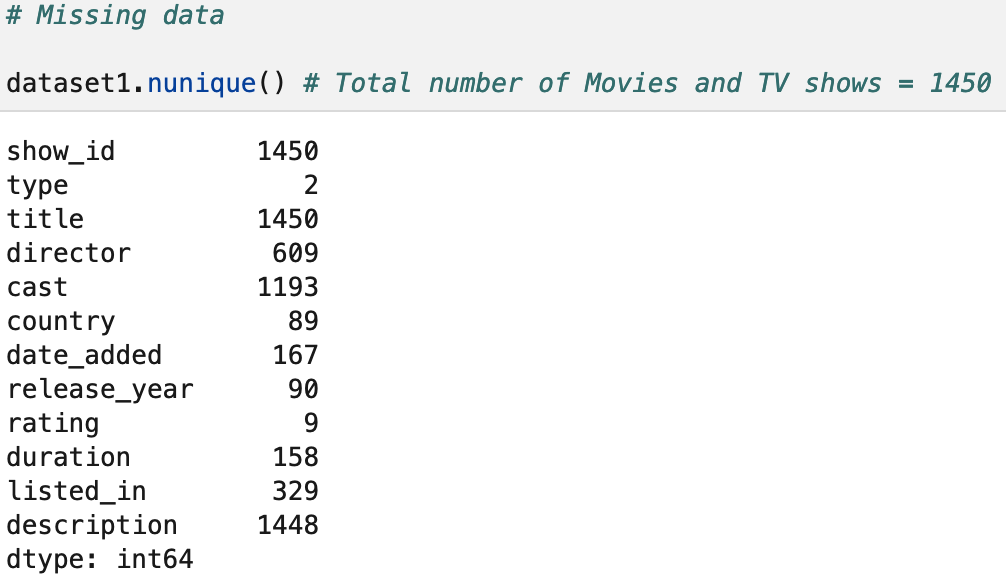


Figure 3: Total Number of Movies & TV Shows

# Next, we can check how of these NaN values are present in the DataFrame as seen in Figure 4. There is a total of 632 NaN values.

# 

Figure 4: NaN Values Present in DataFrame

# Next, we can check the percentage of NaN values in the Dataframe as seen in Figure 5. NaN values make up 43.5% of the data in the DataFrame.

# 

Figure 5: Percentage of Missing Data

# Therefore, with a large proportion of missing data, removing the NaN values may give a more accurate analysis as seen in Figure 6.

# 

Figure 6: Removing NaN Values

# **Comparing Movie & TV Show Content on Disney+**

# We can compare the distribution of movies and TV show content on Disney+ as seen in Figure 7. 73% of the content is movies and 27% of the content is TV shows.

# 

Figure 7: Percentage of Movies vs TV Shows

# **Distribution of Ratings on Disney+**

# We can also compare the ratings categories used in Disney+ as seen in Figure 8. In this case, ‘TV-PG’ is the most popular rating for both movies and TV shows on Disney+ and ‘TV-Y7-FV’ is the least popular rating for both movies and TV shows.

# 

Figure 8: Distribution of Ratings on Disney+

# **Most Popular Genre on Disney+**

# We can also find the most popular genre of movies and TV shows on Disney+ as seen in Figure 9. In this case, ‘Animation, Comedy, Family’ is the most popular genre on Disney+ and ‘Action-Adventure, Animation, Fantasy’ is the least popular genre.

# 

Figure 9: Distribution of Genres on Disney+

# **Create Function for Collecting Actor Data**

# Next, we need to create a function to collect the Top 10 actors appearing in the most Disney+ movies and TV shows as seen in Figure 10. Jim Cummings has appeared in most movies and TV shows on Disney+.

# 

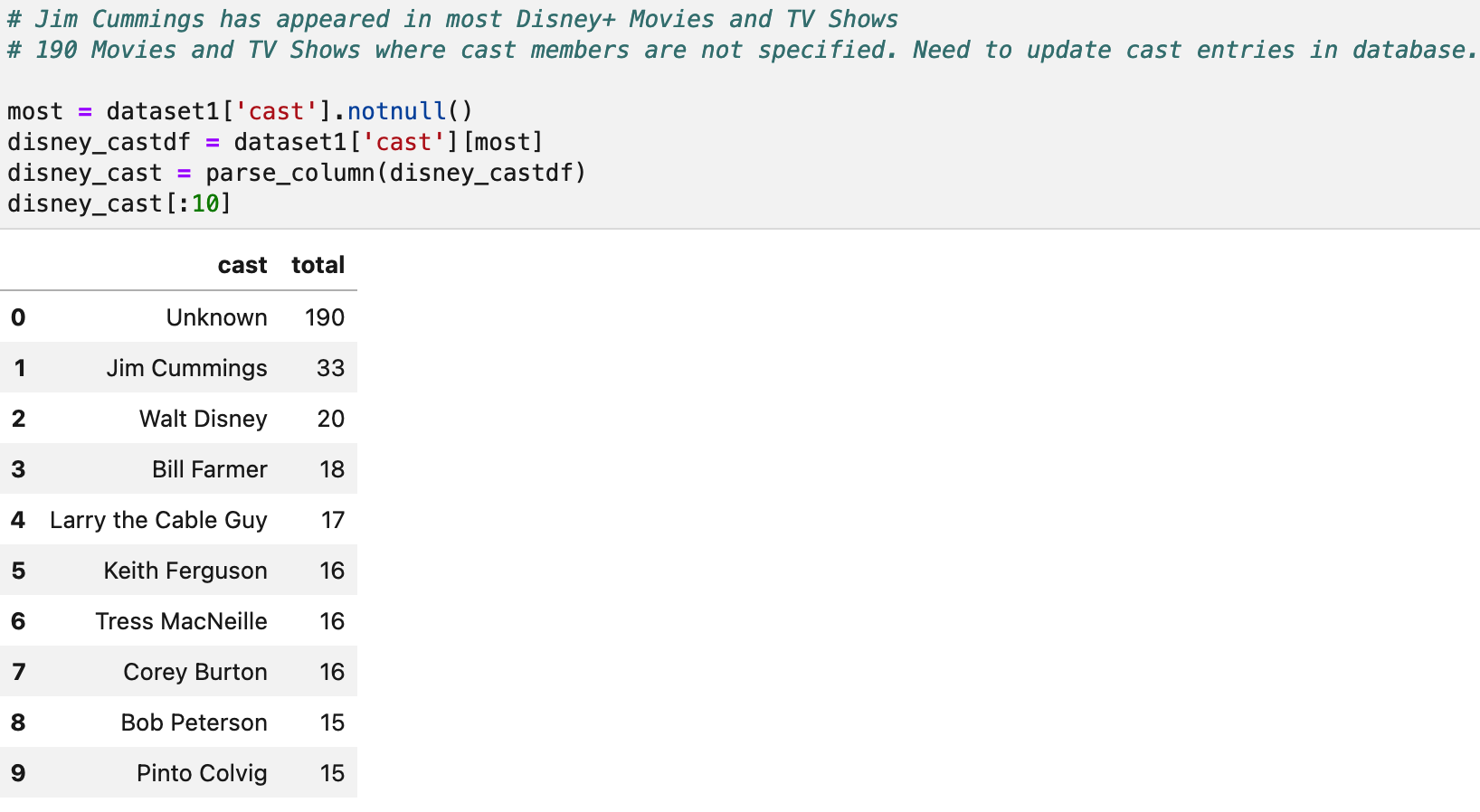


Figure 10: Create Function for Collecting Actor Data

# **Insights**

The insights from the results are as follows:

1. There is a total number of 1450 movies and/or TV shows on Disney+.

# 73% of the content is movies and 27% of the content is TV shows.

1. ‘TV-PG’ is the most popular rating for both movies and TV shows on Disney+ and ‘TV-Y7-FV’ is the least popular rating for both movies and TV shows. Therefore, Disney should consider producing more movies and TV shows with the ‘TV-PG’ rating.
2. ‘Animation, Comedy, Family’ is the most popular genre on Disney+ and ‘Action-Adventure, Animation, Fantasy’ is the least popular genre. Therefore, Disney should consider producing more ‘Animation, Comedy, Family’ genre movies and TV shows.

# Jim Cummings has appeared in most movies and TV shows on Disney+.

# 190 movies and/or TV shows have cast members that are not specified. Therefore, cast entries need to be updated on the database.

1. **Benefit of Machine Learning to Disney+**

The following examples are how machine learning can be used to improve Disney+:

### **Thumbnails**

‘The user places great importance on the thumbnail, which is becoming an extremely prevalent trend in modern times. The thumbnail alone is enough for many viewers to determine whether or not they should watch the video in question’ (Hinkle, 2021) and over time, Disney+ AI could provide more visually appealing thumbnails to entice viewers to watch their content by using machine learning to rank hundreds of frames taken from pre-existing movie or TV shows to determine which thumbnails are most likely to prompt a click from users.

### **Optimal Streaming Quality**

‘About 116 million subscribers actively use Disney+ each month. It can becomes very difficult to provide high-quality video to everyone at once under these conditions’. Video quality improvements for viewers even during busy viewing times can be achieved by placing video assets near subscribers locations in advance’ (Hinkle, 2021).

### **Tailored Movies Recommendation**

Disney+ customises its data recommendations for each customer. A single Disney+ account may be used in two distinct locations but you will be shown different recommendations in each. Disney+ AI is responsible for this function. The algorithm learns on its own and continues to gather information. Simply logging more hours on Disney+ increases the quality of the their recommendations sent to each subscriber’ (Hinkle, 2021).

1. **Classification or Regression Methods**

As Disney+ is a streaming service, I would use Classification algorithms for machine learning as the content on Disney+ is ‘arranged into discrete data categories’ (Brownlee, 2019) such as a Movie or TV Show.

‘A Classification algorithm can use either a fixed or categorical target variable. We can then use the algorithm to identify the most likely ‘class’ a target variable will probably fall into’ (Brownlee, 2019). Therefore, this type of algorithm can be best used to answer questions such as ‘Who is most likely to sign up for Disney+ ?’

# **References**

*Disney+ Movies and TV Shows* (2021). Available at: <https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows>.

Ahmad, Y. (2018) *5 Big Benefits of Data and Analytics for Positive Business Outcomes*. Available at: <https://www.teradata.com/Blogs/5-Big-Benefits-of-Data-and-Analytics-for-Positive-Business-Outcomes>.

Hinkle, D. (2021) *How Streaming Services Use Algorithms — AMT Lab @ CMU*. Available at: <https://amt-lab.org/blog/2021/8/algorithms-in-streaming-services>.

Brownlee, J. (2019) *Difference Between Classification and Regression in Machine Learning*. Available at: <https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>.