EDA: Long Form!

Anna Wagman

Data Science (STAT 301-1)

Contents

Introduction:

As a subscriber to Hulu, AmazonPrime, and Disney+, I very excited to explore the similarities and differences between the three streaming services. In my EDA, I will explore the variables contained inside the individual data sets, and then compare specific variables across all three. Which service offers the most options? Does it differ depending on Movies vs TV Shows? Do Hulu, AmazonPrime, and Disney+ offer any of the same titles? How many? Which of these these services offer the most international streaming? In which countries? These are some of the questions I will be exploring and depicting throughout my Six Visualizations. The three data sets provide information on the titles available on Hulu, AmazonPrime, and Disney+ respectively. All three data sets include the same variables such as title, type, genre, rating, director, cast. . . etc

Data Sources:

https://www.kaggle.com/shivamb/hulu-movies-and-tv-shows/version/1. https://www.kaggle.com/shivamb/amazon-prime-movies-and-tv-shows https://www.kaggle.com/shivamb/disney-movies-and-tv-shows

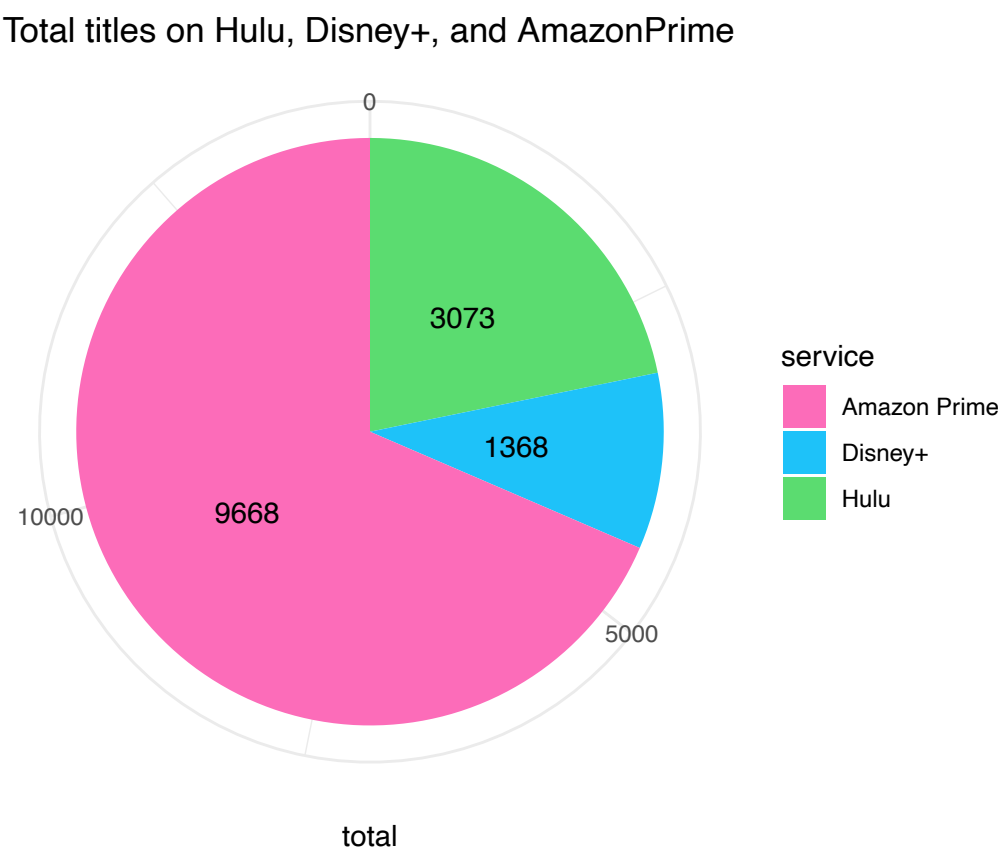
Data Cleaning:

In my initial data cleaning, I renamed the column var = "listed_in" -> to "genre" for all three data sets and I removed the last column var = "description" that gave a lengthy description of the plot for each. Throughout my EDA, I use functions such as as.numeric, extract_numeric, round(). . . and many more in order to clean my data. For example, to work with var = "duration", I had to parse the numerical value from "duration" for each separate type = "Movie" and type = "TV Show". For type = "Movie", duration is given in minutes and so I used the "as.numeric()" function. For type = "TV Show", I had to extract the numerical value from duration given by "# Seasons" where '#' is an integer. I used the function "extract_numeric(duration)". I cleaned and recoded each variable in many different ways for each visualization, which is specifically outlined in my "data_cleaning.R" & "data_cleaning.Rmd" files which can be found in the main foler EDA_Wagman.

Visualization 1: Total TV Shows and Movies

Q1. Which service has the most content? Is there a large disparity?

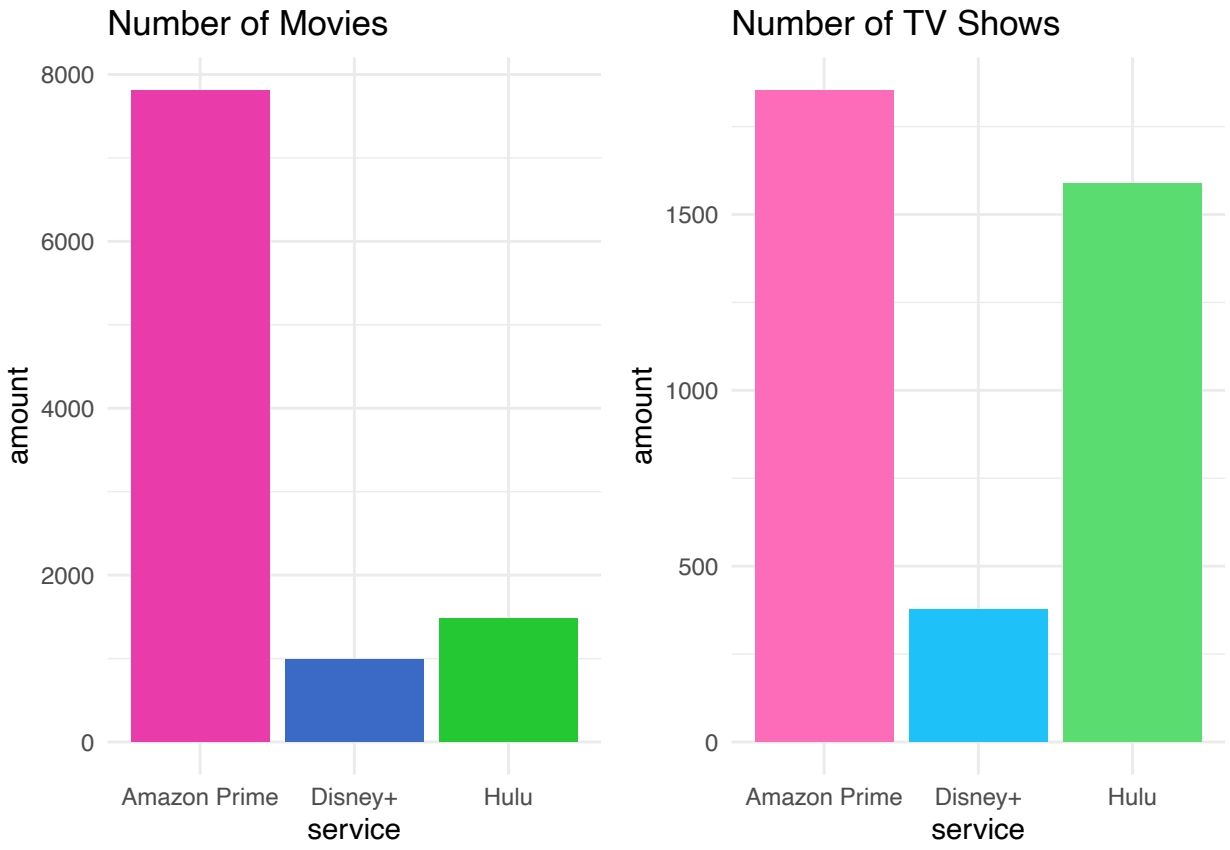
```
#plot 1
total_titles_plot
```



Visualization 2: Movies vs TV Shows on each service

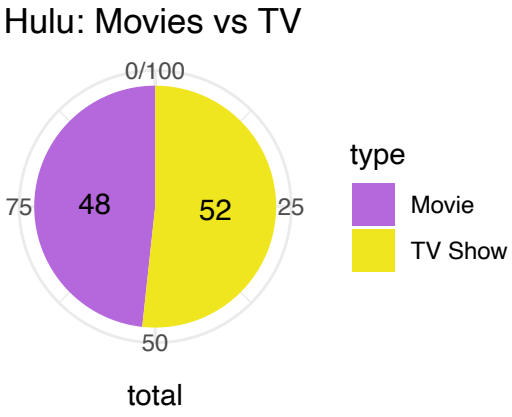
Q2. What is the distribution of Movies vs TV shows per service?

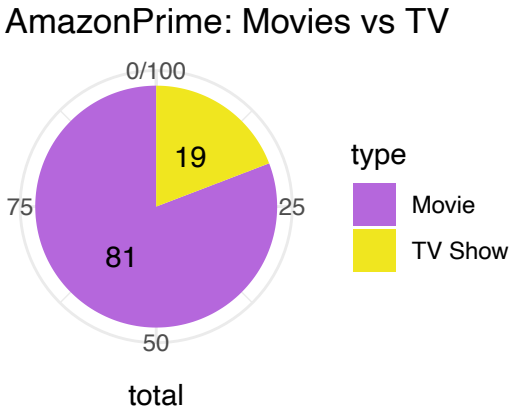
```
#plot side by side: Movies vs TV
plot_grid(total_movies_plot, total_TVShows_plot)
```

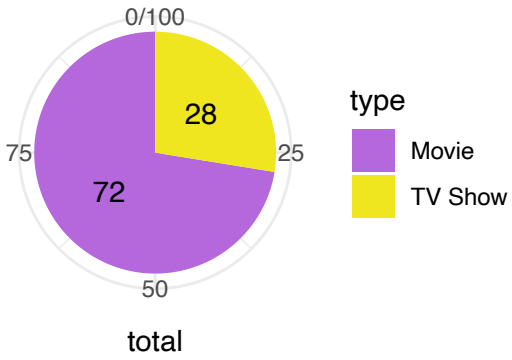


2b: Percentage of Movies vs TV Show on each

```
#plot percent Movie v TV of all 3 on one plot
plot_grid(Hulu_percent_plot, Amazon_percent_plot, Disney_percent_plot)
```



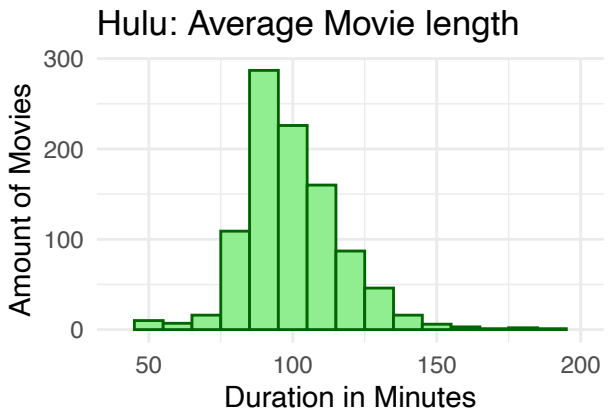


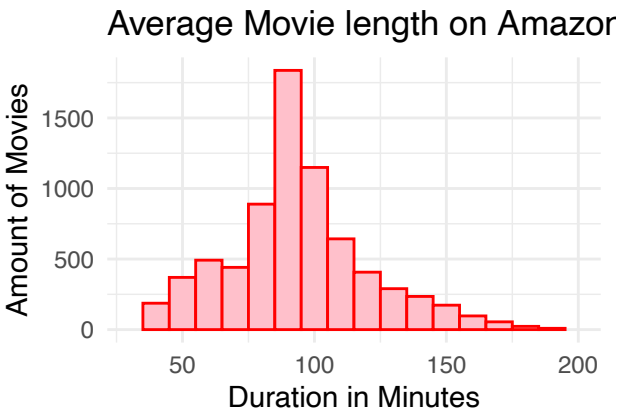


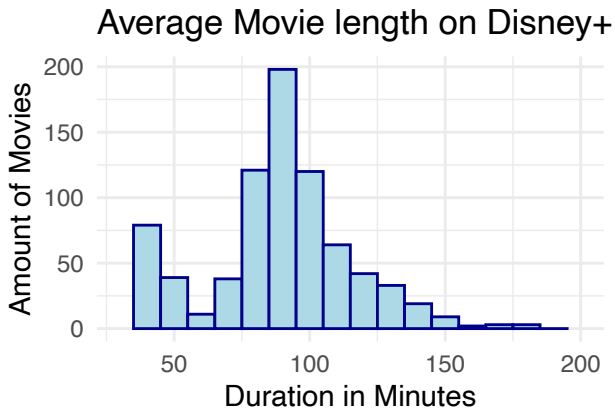
Visualization 3: Movie Duration distribution

Q3. Are movies on average the same length per service?

```
#plot all three service_duration plots created above on one plot!
plot_grid(hulu_duration_plot, amazon_duration_plot, disney_duration_plot)
```

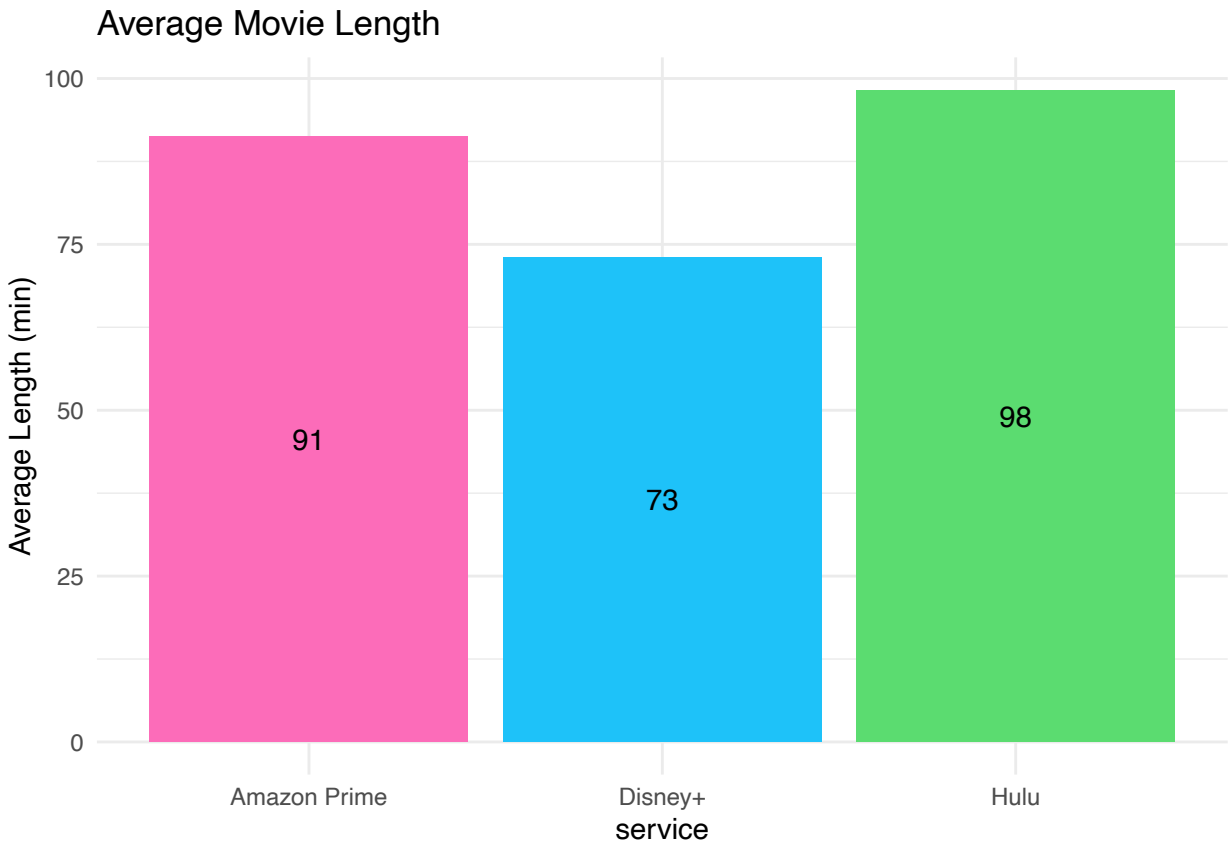






3b: Average (mean) of Movies Compared

```
Movie_av_plot
```



Visualization 4: Titles Available on Multiple services

Q4. Which services have overlap? How many titles do they have in common?

Total titles availale on multiple services:

```
Total_Overlap
```

```
## # A tibble: 3 x 2
##   services          titles_overlap
##   <chr>                      <int>
## 1 Hulu and Amazon              221
## 2 Disney and Hulu               31
## 3 Amazon and Disney             14
```

Titles availale on all 3 services:

```
total_total_overlap
```

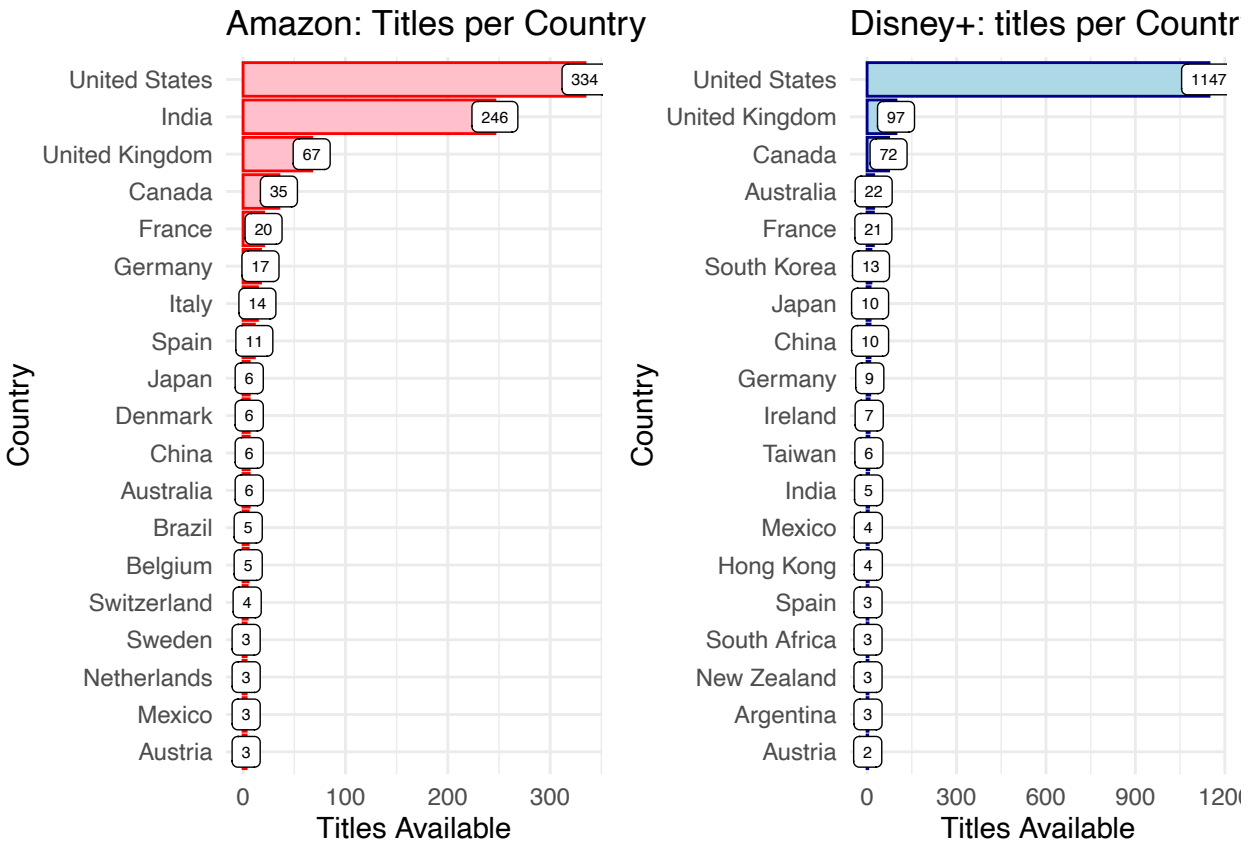
```
## # A tibble: 1 x 1
##   title
##   <chr>
## 1 10 Things I Hate About You
```

As you can see, Hulu and AmazonPrime have the most overlap by a significant margin with 221 titles in common. The only title that all 3 services have in common is the movie "10 Things I Hate About You".

Visualization 5: Worldwide Streaming

Q5. Which services offer international streaming? In which countries? How many titles are available in each?

```
plot_grid(amazon_country_plot, disney_country_plot)
```

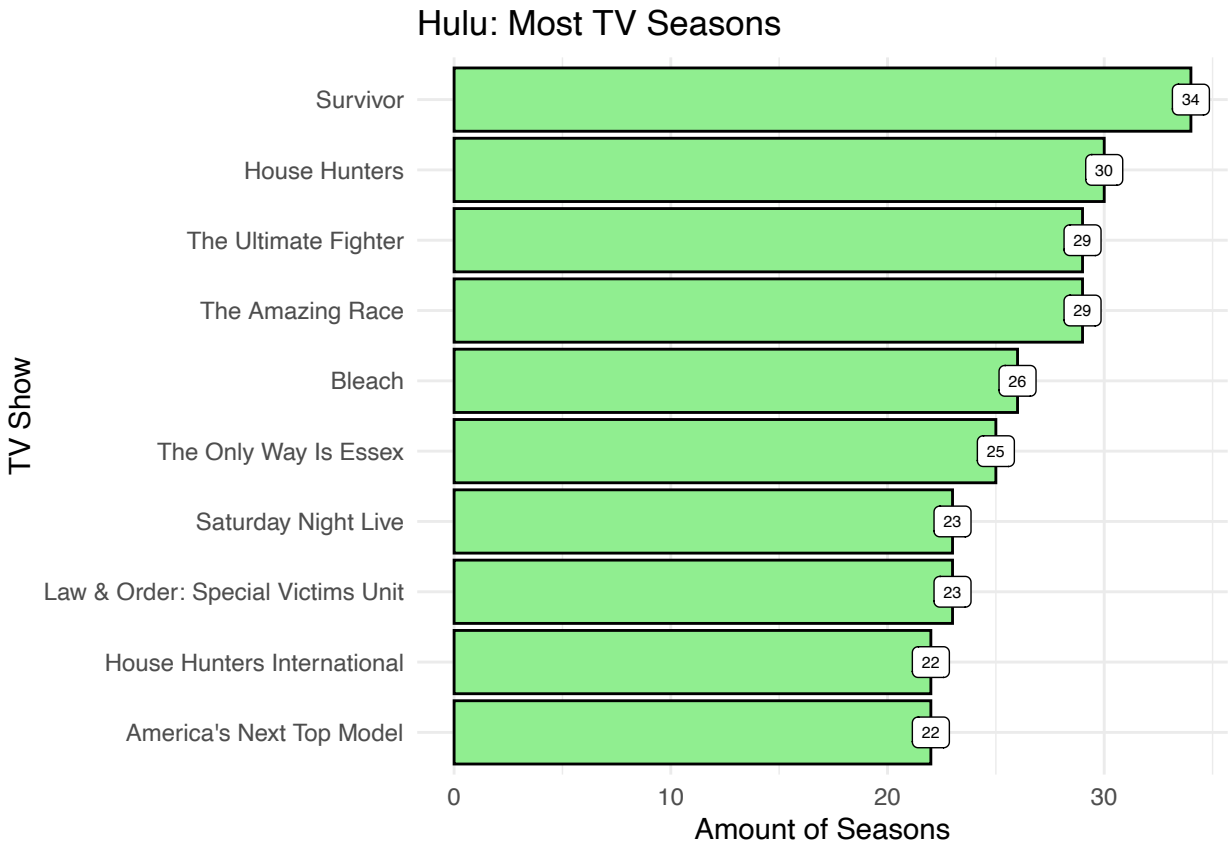


Hulu is only Available in the US and UK making for a very uninteresting graph, so I did not include Hulu. As shown above, the US has the most availability by a lot. The US has more titles than all of the other countries combined...

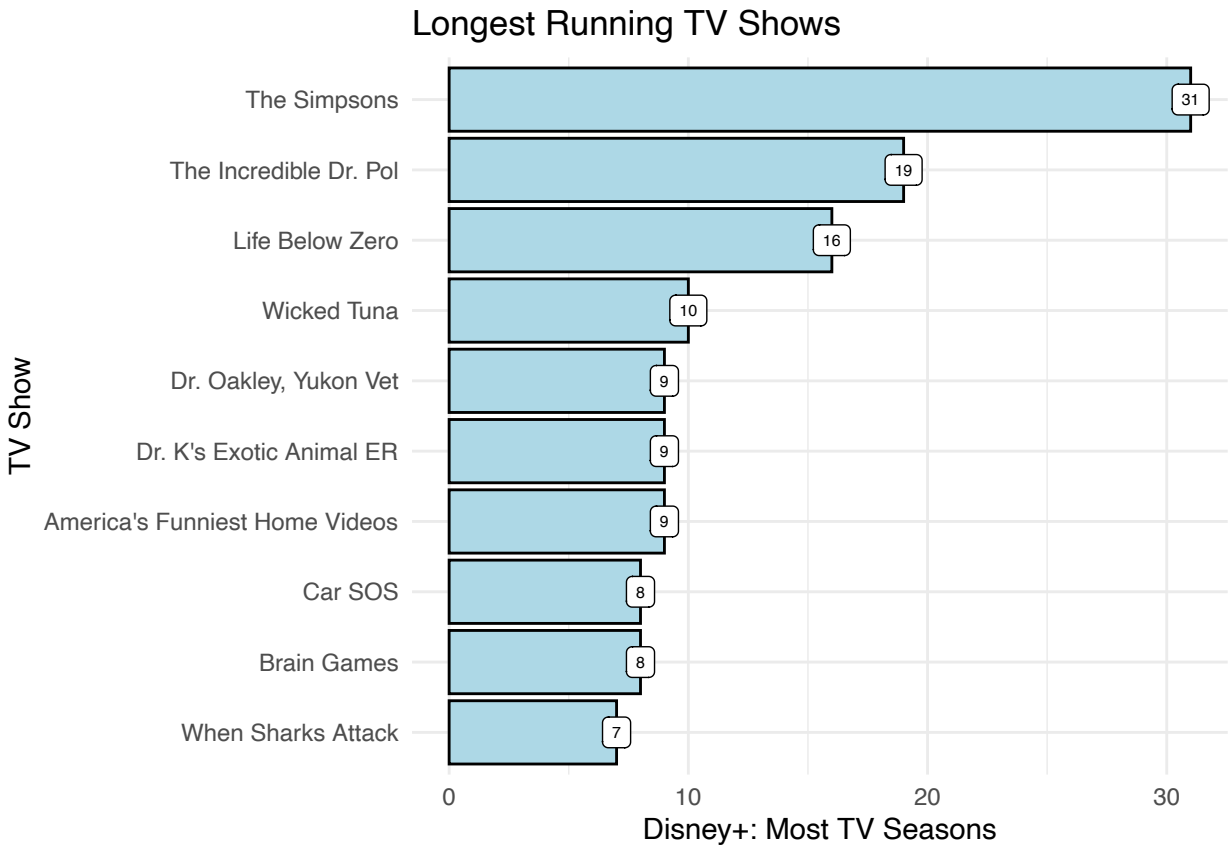
Visualization 6: Longest runnning TV Shows on each:

Q6. What are the longest running TV Shows on each? How many seasons are there?

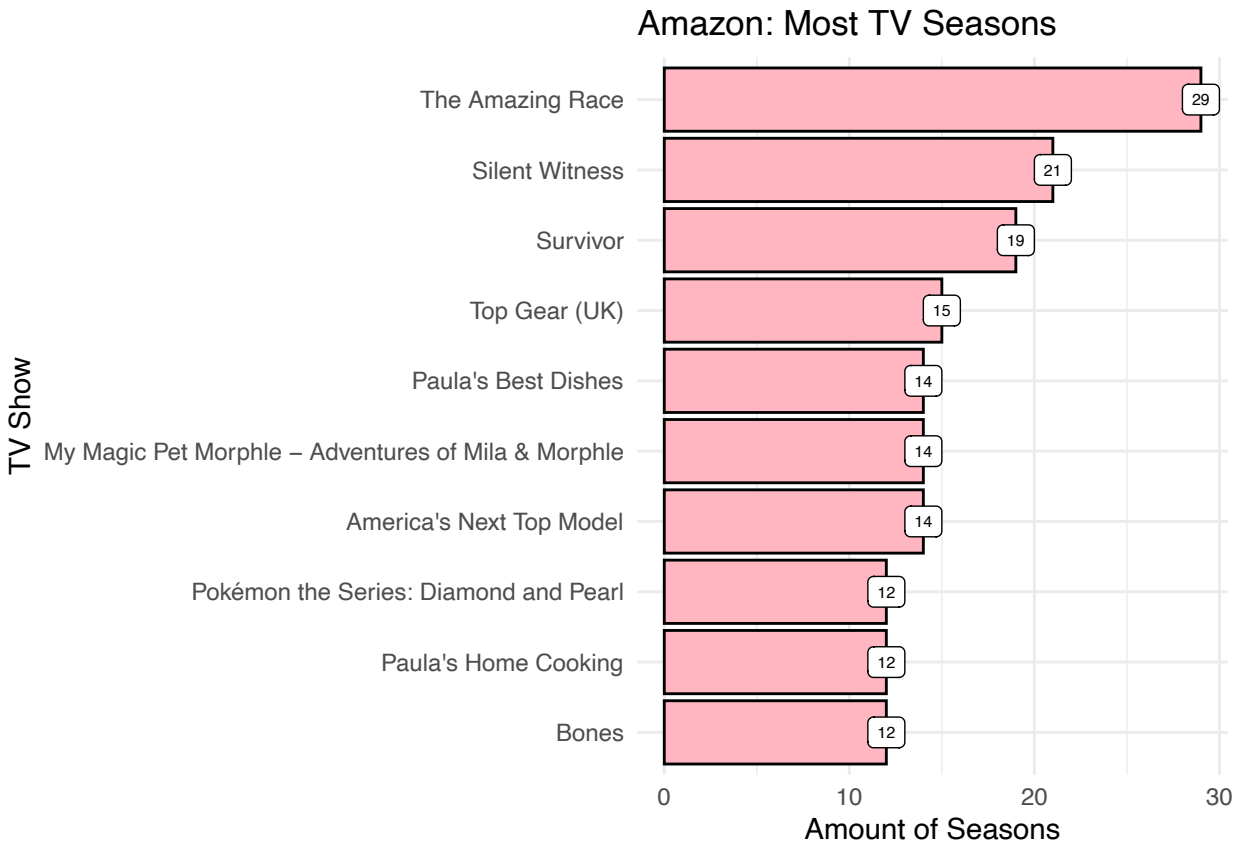
```
hulu_seasons
```




```
disney_seasons
```




```
amazon_seasons_plot
```



Hulu and AmazonPrime both have the TV Shows "Survivor" and "America's Top Model", both of which appear in their top 10 longest running shows. Fun fact: Survivor is my favorite TV Show ever and I have seen all 41 Seasons!