

# Final Project

Franchesca Johnson

07/30/2023

Franchesca Johnson 7/30/2023

Introduction It's January 2020 and the first confirmed case of Covid-19 is in the U.S. Even though China has been dealing with the spread of this unknown virus for the past month, it hasn't really hit the Western part of the world, yet.

Slowly, more positive cases are starting to emerge throughout the west coast of the U.S. As the virus starts to spread worldwide affecting every country in different ways, it still hasn't completely affected the Americas. Around springtime of 2020, more and more positive cases of covid are emerging in our country. Employers and state officials are taking drastic steps to ensure the safety of their employees and the general population. Schools are closing and going virtual, daycares are shutting down, curfews are being enforced, jobs are becoming virtual and you're being told to stay home at all times.

You're not supposed to leave your house unless necessary, so what are you to do? I guess it's time to Netflix and Netflix. There was some chilling going on too, especially since there were so many covid babies. Either way, I'm going to be researching the most watched show/movies between 2020 to March 2022.

Research Questions Other than the increase in births between that time and netflix's stock, what show or movie was most binged watched between that time? What are the top 10 shows that were watched? Who watched it (age group) and why?

Was Netflix watched more during certain times of the day? Did location, gender, age, class have any influence as to what was watched? Was the show new or an old series? Did the program(s) match the current culture of that time? Are the shows/movies still in the top line up since new series have started to broadcast? Will this data tell us anything about our culture? Approach First reviewing the data, will it be able to answer all my questions? Meaning is there enough information/variables will I need to eliminate certain questions immediately and potentially add new ones based on the review? Once I can do a process of elimination, if necessary, then I'll see if I notice any additional information that stands out that may need clarification. How your approach addresses (fully or partially) the problem. I think it will partially address the problem based on how much data is there.

Data Kanawattanachai, Prasert. "Netflix Daily Top 10." Kaggle, 12 Mar. 2022, [www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us](https://www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us). Netflix movies and TV shows dataset - dataset by crawlfeeds. data.world. (2022, August 8). [https://data.world/crawlfeeds/netflix-movies-and-tv-shows-dataset/workspace/file?filename=netflix\\_movies\\_and\\_tvShows\\_sample\\_dataset\\_sample.csv](https://data.world/crawlfeeds/netflix-movies-and-tv-shows-dataset/workspace/file?filename=netflix_movies_and_tvShows_sample_dataset_sample.csv) Mahmood, A. (2022, October 3). Netflix Movies Dataset. Kaggle. <https://www.kaggle.com/datasets/anasmahmood000/netflix-movies-dataset> Required Packages I'm using ggplot to plot a scatterplot, tidyverse to clean the data, and data explorer. The scatterplot will be a helpful tool, especially when doing the linear regression analysis. I started out with the summary function to get an overview of each variable and the min. max of each. From there I'll determine what questions I can answer based on the data and which one I can eliminate. I'm hoping new questions will emerge based on the information I see within the data.

```
library(plyr) library(tidyverse) library(DataExplorer) library(ggplot2) library(GGally) install.packages("janitor")
install.packages("data.validator") library(janitor) library(data.validator)
```

```
install.packages("Rcmdr", dependencies = TRUE) library(Rcmdr)
```

```
installed.packages(lib.loc = ) library(lubridate) library(magrittr) library(markdown) library(Matrix)
```

library(performance) library(pastecs) library(polycor) library(purrr) library(proxy) library(readxl) library(rlang) library(tibble) library(parallel) library(parameters) library(compiler) library(datasets)

### 9.3 Final Project Step 2

1. How to import and clean my data?
  - look for any missing data, cells.
  - duplicate throughout the dataset
  - Put a min and max parameter in to confirm only shows between 2020-2022 are part of the data.
2. What does the final data set look like?
  - I've sorted the data and i haven't been able to discover any type of immediate information for my questions. I think changing up one of the datasets to get a clearer picture should be my next step if the answers aren't clear enough.
3. Questions for future steps.
  - Is there a yearly trend with the same show.
  - Did the type of show/movie, etc change throughout the years. Better to worse or worse to better as lock down and death numbers increased and later decreased.
4. What information is not self-evident?
  - The age and geographic location.
5. What are different ways you could look at this data?
  - Possible look at it as the current mental health of the world at the time.
6. How do you plan to slice and dice the data?
  - Make three tables by year.
  - Sort each year by rank, year to date rank, released date, days in top 10, viewership score and Title.
7. How could you summarize your data to answer key questions?
  - By sorting the information in the format above, it will tell me the shows that are the most popular and for how long. Based on the content of the show, it could potentially explain the age group and well as the overall feeling of the culture at the time.
8. What types of plots and tables will help you to illustrate the findings to your questions?
  - Linear regression is a useful plot. It break up the data and potentially show me the genre that was the most popular.
9. Do you plan on incorporating any machine learning techniques to answer your research questions?

Explain.

- At this point, I think logistic regression could be interesting to include especially since Covid number are starting to rise again (Just today had 3 positive covid patients this week). I'm not sure if it could get as extreme as before since majority of the population is vaccinated thus creating a strong herd immunity. In addition, this is no longer new, our immune system has seen this virus now, so even without having any type of vaccination, the severity shouldn't be the same \

10. Questions for future steps.

- As of right now, I don't have any but I do would like to figure out what type of shows could be in the lead this year (flu/covid season).