Individual Project Description Markdown

Ian Brown

3/11/2024

CSC-324, Professor Jimenez

Project purpose:

The purpose of this project is to examine Netflix's usage on a monthly and daily basis. This was done by collecting data from the United Kingdom's Netflix servers and individual users from 2017 to half-way through 2019. My target audience for this project is anyone interested in Netflix's user traffic. My visualizations can be used to understand the monthly user data, the amount of time users spend watching a film which shows user interest in films, which days users spend the most time on Netflix, and which films were the most popular and what films were the most popular given a certain day. The goal of my project was to first identify when users are more likely to watch Netflix, and then to determine if certain films being added or removed contributed to user traffic on the site.

Data description:

The two data sets that I will be working with will be merged into one dataset. They will provide me with information regarding when Netflix movie titles went live on the platform, as well as when users watched the titles. The individual UK user streaming data encompasses data collected from 1/1/17 to 6/30/19. The Netflix live titles data encompasses data collected between 1/1/08 to 9/25/21.

The data were obtained from the "Netflix Movies and TV Shows" data set created by Shivam Bansal on kaggle.com [6] and the "Netflix audience behaviour - UK movies" was created by Stephen Follows and Eliel CM on kaggle.com [7].

What questions are you trying to answer?

1. What is the monthly distribution of Netflix viewership?
2. Given a specified date range, are viewers watching the full length of a film? Are they watching longer or are they just viewing small amounts and moving to a different title?
3. What days during the analyzed time period is Netflix traffic the highest?
4. On a given day, what is the most watched film?

What insights did you get from your data?

1. Given the month viewership distribution, I was able to discern that January is the month with the most traffic, and it slowly tapers off. This potentially correlates to when Netflix adds the newest titles to their viewing list.
2. When provided with a scatter plot showing the user watch time of a film to the length of the film, the distribution shows that most viewings spend less than the length of the movie. This shows that users will view parts of a film, get disinterested, and either move onto a new title or move onto a different activity.

3. When analyzing what days have the most traffic on Netflix given a heat map of a calendar, it is surprising to see that these days seem to be random. This would indicate that Netflix is not just a weekend or Holiday endeavor, but something else entirely. This could be a possible indicator for viewership increases on days when new titles are added to the Netflix library.
4. The information presented by a plot of the total watch time of a film given a certain day interesting for two reasons. One, it shows that the most watch film, Layer Cake, was streamed for over 3.6 million minutes on June 5, 2018. Two, it ideally shows that when new titles are added they rapidly become the most watched films in the coming days. The best example is ANIMA, a film added on June 17th, 2019, which immediately becomes the most watched film withing the next couple weeks, as shown on June 28th when it was streamed for over 1.2 million minutes.

Describe the process to make your work reproducible: (for example, tidying data and pipeline process.)

To make my data tidy and clean for use I took multiple steps in the beginning of my app compilation. First, I loaded the individual streaming data, the Netflix titles available, and the time log I kept for the project. Second, I cleaned the data. In this process I filtered the Netflix titles by country (UK), date added to Netflix (January 1, 2017), and type (Movie). I also filtered the user streaming data by filtering watch time (>600 seconds, “10 mins”). I them merged this data together by movie title to be able to access all data quickly and simply in a single file. I then accessed this data in each of my four visualizations.

1. In the first visualization, the histogram of monthly viewership, I simply filtered the individual user data by month, and plotted the number of times each month appeared for every user. This, once tabulated, gave me the data desired.

2. In the second visualization, the scatter plot of watch time and movie length, I filtered the merged data by the date range specified by the slider. This then gave me the data needed to correlate the x and y axis in an accurate manner.

3. In the third visualization, my calendar heat map, I did a multitude of adjustments to the merged data. The first adjustment was to ensure that the “date” variable was encoded in the correct format. Then I filtered the data by the specified year from the drop-down menu. Finally, I created a new aggregated data set which aggregated all the data based on the day of the specified year. This was paired with a new variable, the mean watch time variable which calculated the mean watch time of all viewers on a given day. This allowed me to create an accurate heatmap displaying the most viewers on a given day.

4. In the fourth visualization, my line chart, I created a similar aggregated data set to the heatmap. However, this differed because instead of grouping the data based on the day of the year, I grouped the data by title and date, then summarized the watch time of all the users. This manipulation allowed for an integrated analysis of the most watched films on a given day.

Description of design decisions (encoding/mapping):

Visualization 1 Idiom: Histogram

1. What: In this histogram the user is seeing a combination of clustered data with a geometric distribution method. I use the histogram to user traffic presented in my data set by month. I then display this in a simple bar chart (histogram).
2. Why: The reason for creating this visualization was to make a simple consumption element to show the monthly user traffic. It does not feature an isolation or search method. However, it does allow for certain queries by comparison. This is done by comparing one month to another. For example, it is clear to see that viewership spikes highest in January compared to other months.
3. How: This visualization uses an encoding method which arranges data by aligning the monthly viewership information for comparison.

Visualization 2 Idiom: Scatter Plot

1. What: In this scatter plot I take a cluster of user data showing the time spent watching a film. This is displayed in a dot plot format to show the difference between how many people watch a film for the entirety of the film. This allows for a comparison between the amount of people who are perusing the site for something to watch, to the people who are watching a film for the entire duration.
2. Why: This visualization was created for the consumption of user data given a time scale. Users can compare the overall watch time of a user to the film's duration, giving them insight into how many users watch a film for the entirety to those who are just looking. Users are also able to search by selected dates, narrowing the time scale from 2.5 years down to a single day if they wish. This also allows for multiple query options, such as comparison between times as well as identifying different trends.
3. How: This visualization used an encoding method which arranges data by separating different plots of watch time vs. film duration. Users are also able to reduce the data by filtering date spreads to desired time spans.

Visualization 3 Idiom: Heatmap

1. What: In this heatmap I have used daily user data to describe the difference in Netflix traffic across different days. I have chosen to use this field version of data visualization because I believe it most accurately depicts the difference across days, weeks, months and even years.
2. Why: This visualization was created for the consumption of user data given a certain year. Users can discover trends in the heatmap by identifying which days represent the highest viewership counts. Users are also able to search/browse different years allowing for a query into comparison across time to take place.
3. How: This visualization used an encoding method used to map the viewership counts on specific days. The user can reduce by filtering the year which they wish to view.

Visualization 4 Idiom: Line Chart

1. What: In this line chart I have used daily user data to show the most popular film watched on each day. This correlated both user traffic as well as time spent watching the film. I chose the network data visualization method to best overlay the data all at once. This best shows the distinction of different film viewership over time.

2. Why: This visualization was created for the consumption of user data given a certain day spanning the 2.5 years of data. Users are also able to browse, locate, and explore different trends of films over this time. This allows for queries identifying which films are most popular on what days and compare viewership spikes to analyze the most popular films in the Netflix library.

3. How: This visualization used an encoding method by expressing the change in daily views of a film by connecting scatter plot points with lines. The user can manipulate the graph by navigating and selecting certain lines. The user can reduce the data by viewing embedded information, such as date, watch time, and movie title by selecting a distinct line.

What needs improvement? (Wish list)

If I could make improvements, I would first change the y-axis of the histogram of month viewership data. In this case, scientific notation of numbers is not nearly as clear as simple numbering. I would also like to fix the clarity of the heatmap’s legend. In a few cases, the legend appears out of order. I am not sure why this is occurring, and I spent a lengthy amount of time trying to figure it out and was unsuccessful. Finally, I would add a filter to the line chart where users could specify the films which they wanted plotted. Currently the visualization is quite messy. Although it is easily deciphered with the ‘plotly’ ability for cursor hovering, it would be much easier to read if a filter were added. In making these improvements, I believe I could present a complete and more useful product, however given the time constraint, I was not able to accomplish these “wishes.”

Acknowledgments:

- library(tidyverse)
- library(readr)
- library(dplyr)
- library(shiny)
- library(shinydashboard)
- library(lubridate)
- library(rsconnect)
- library(ggplot2)
- library(plotly)

1. [1]D. Royé, “A heatmap as calendar: R-bloggers,” R-Bloggers, 20-Dec-2020. [Online]. Available: https://www.r-bloggers.com/2020/12/a-heatmap-as-calendar/ (https://www.r-bloggers.com/2020/12/a-heatmap-as-calendar/). [Accessed: 10-Mar-2024].
2. [2]“Faithful,” Shiny, 11-Aug-2014. [Online]. Available: https://shiny.posit.co/r/gallery/start-simple/faithful/ (https://shiny.posit.co/r/gallery/start-simple/faithful/). [Accessed: 10-Mar-2024].
3. [3]J. Bryan, “Chapter 42 Building Shiny apps,” Stat 545. [Online]. Available: https://stat545.com/shiny-tutorial.html#shiny-tutorial-1 (https://stat545.com/shiny-tutorial.html#shiny-tutorial-1). [Accessed: 10-Mar-2024].
4. [4]“Shinydashboard,” RStudio. [Online]. Available: https://rstudio.github.io/shinydashboard/get_started.html (https://rstudio.github.io/shinydashboard/get_started.html). [Accessed: 10-Mar-2024].

5. [5]Y. Holtz, “Basic scatterplot with R and GGPLOT2,” the R Graph Gallery. [Online]. Available: https://r-graph-gallery.com/272-basic-scatterplot-with-ggplot2.html (https://r-graph-gallery.com/272-basic-scatterplot-with-ggplot2.html). [Accessed: 10-Mar-2024].
6. [6]S. Bansal, “Netflix movies and TV shows,” Kaggle, 27-Sep-2021. [Online]. Available: https://www.kaggle.com/datasets/shivamb/netflix-shows?rvi=1 (https://www.kaggle.com/datasets/shivamb/netflix-shows?rvi=1). [Accessed: 10-Mar-2024].
7. [7]E. CM and S. Follows, “Netflix audience behaviour - UK movies,” Kaggle, 05-Feb-2021. [Online]. Available: https://www.kaggle.com/datasets/vodclickstream/netflix-audience-behaviour-uk-movies?rvi=1 (https://www.kaggle.com/datasets/vodclickstream/netflix-audience-behaviour-uk-movies?rvi=1). [Accessed: 10-Mar-2024].

Appendix:

```
knitr::include_graphics("hist_time_log.png")
```

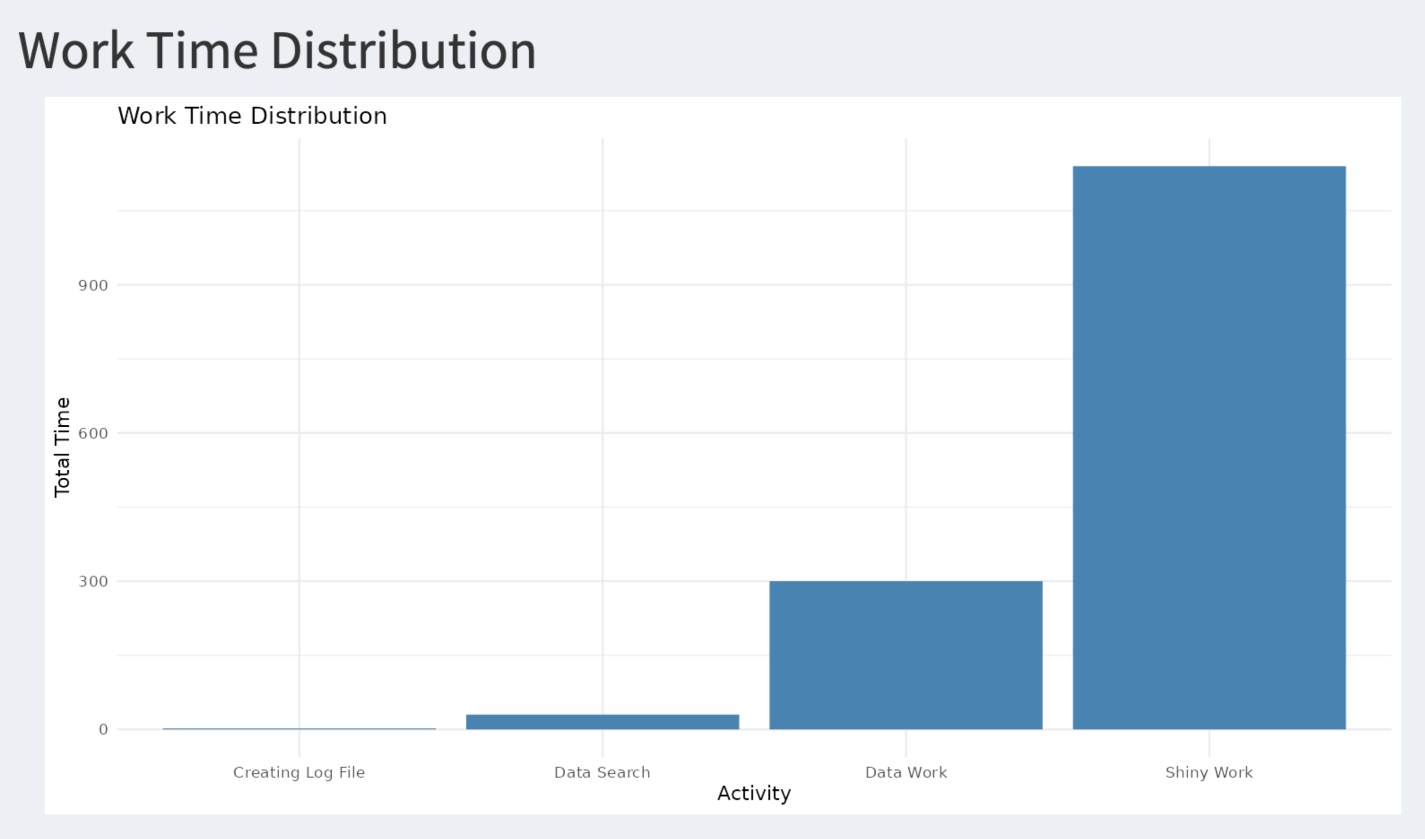
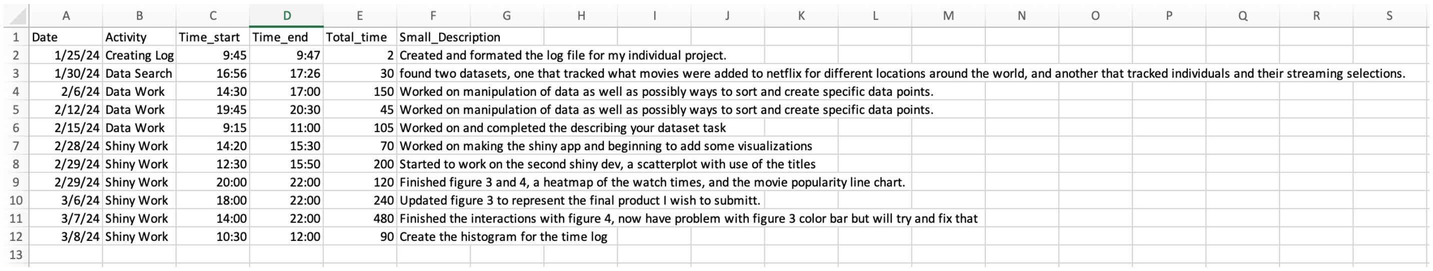


Figure 1: Time/Tasks Log Chart. This figure represents the time spent on each individual task during this project. As can be seen, the most time spend was on Data Work and Shiny Work. Data Work represents the cleaning of data while the Shiny Work represents the time spent on figures and visualizations for the Netflix data. Total Time is represented in minutes. (Going to need to knit this in as .png)

```
knitr::include_graphics("time_log_dataset.png")
```



Date	Activity	Time_start	Time_end	Total_time	Small_Description
1/25/24	Creating Log	9:45	9:47	2	Created and formated the log file for my individual project.
1/30/24	Data Search	16:56	17:26	30	found two datasets, one that tracked what movies were added to netflix for different locations around the world, and another that tracked individuals and their streaming selections.
2/6/24	Data Work	14:30	17:00	150	Worked on manipulation of data as well as possibly ways to sort and create specific data points.
2/12/24	Data Work	19:45	20:30	45	Worked on manipulation of data as well as possibly ways to sort and create specific data points.
2/15/24	Data Work	9:15	11:00	105	Worked on and completed the describing your dataset task
2/28/24	Shiny Work	14:20	15:30	70	Worked on making the shiny app and beginning to add some visualizations
2/29/24	Shiny Work	12:30	15:50	200	Started to work on the second shiny dev, a scatterplot with use of the titles
2/29/24	Shiny Work	20:00	22:00	120	Finished figure 3 and 4, a heatmap of the watch times, and the movie popularity line chart.
3/6/24	Shiny Work	18:00	22:00	240	Updated figure 3 to represent the final product I wish to submitt.
3/7/24	Shiny Work	14:00	22:00	480	Finished the interactions with figure 4, now have problem with figure 3 color bar but will try and fix that
3/8/24	Shiny Work	10:30	12:00	90	Create the histogram for the time log

Figure 2: Time Log Table. The data used to create the histogram in figure 1.