Project 1: Netflix User’s Information Analysis Report

Pei-Yu Jheng

ALY6040 – Data Mining Applications, Northeastern University

Professor Justin Grosz

February 22, 2024

Introduction

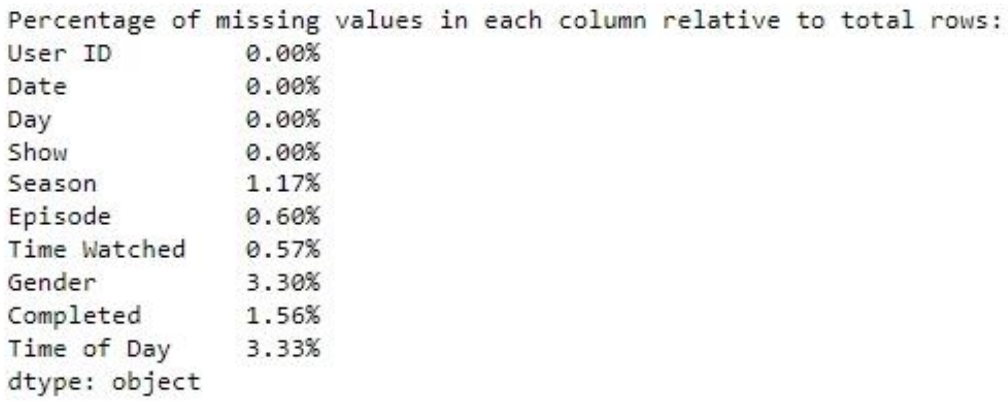
The goal of this project was to present the data mining decisions made on the Netflix User’s Information Dataset through the analysis processes. The data was collected between February 9th and 12th, 2019. Various features of a Netflix user base are presented in this dataset, including user ID, day of watching, name of the show, season of the show, episode of the show, total time watched, gender, completed status, and time of day. To achieve this goal, we explored the data to gain important insights from users' experiences and tried to determine the appropriate decisions for dealing with missing values. Analyzing and cleaning a Netflix user base can provide insight into user trends, preferences, and revenue generation.

Data Cleanup

During the data cleaning process, we carefully examined the dataset to ensure its accuracy and reliability for the subsequent analysis. The initial step involved displaying summary statistics and data types for the dataset. This examination revealed six numerical values and four categorical values. Among the numerical variables, we identified two binary variables, namely “Completed” and “Time of Day”. Additionally, we confirmed that the data types were correct.

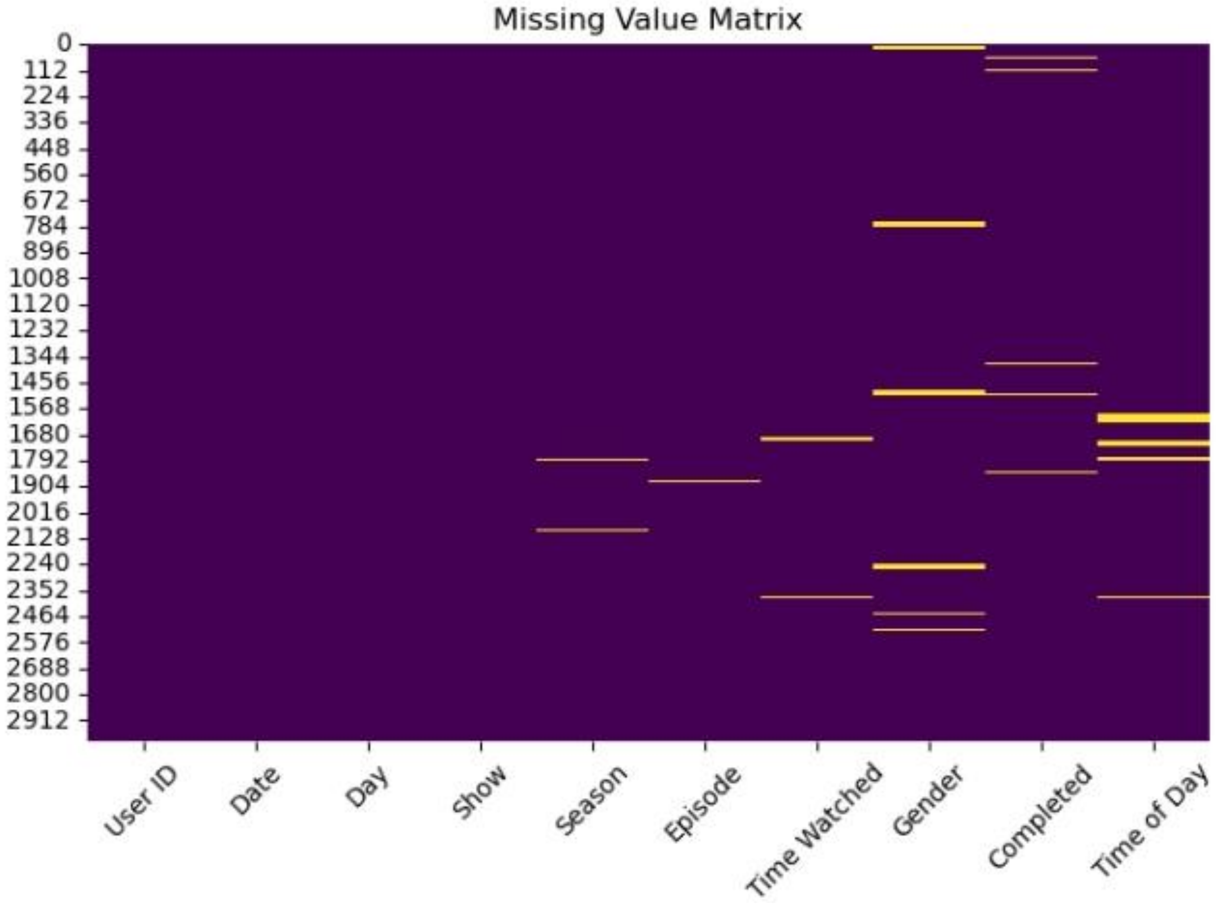
The next step in the cleaning process was to address missing values. We found that six variables contained missing values. To quantify the extent of missing data, we calculated the missing value matrix and the percentage of missing values in each column. We found that the missing values account for a small proportion of the dataset, with each column having less than 10% missing values individually. The percentages of missing values in each column relative to the total number of rows are as follows:

Table 1



```
Percentage of missing values in each column relative to total rows:
User ID         0.00%
Date            0.00%
Day             0.00%
Show            0.00%
Season          1.17%
Episode         0.60%
Time Watched    0.57%
Gender          3.30%
Completed       1.56%
Time of Day     3.33%
dtype: object
```

Figure 1



The third step involves decision-making based on our investigation of missing values.

- For the variable with the highest percentage of missing values, “Time of Day,” we have decided not to analyze this column due to its high rate of missing values. Therefore, we will drop this column to preserve the integrity of the remaining data.

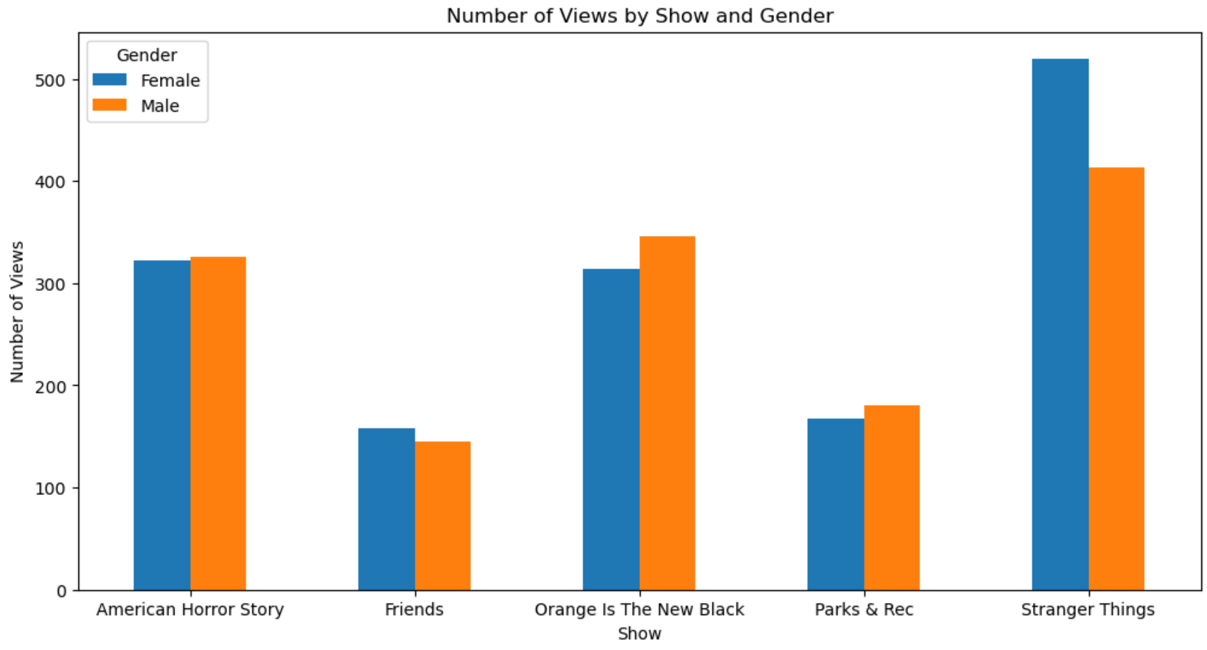
- Regarding the variable with the second-highest percentage of missing values, "Gender," we will replace the missing values with the most frequent gender within each show.
- For the remaining variables with missing values (Season, Episode, Time Watched, and Completed), since the proportion of missing values is less than 3%, which is considered a tiny proportion for this dataset, we have decided to drop these missing values. After ensuring that no missing values are remaining in the dataset, we will proceed to the next phase of our analysis.

Data Hypothesis Exploration

Q1: Which gender, male or female, tends to enjoy watching each show more, and what is the difference in their viewing preferences?

To answer this question, we utilized the pivot table function to display the number of views for each show and gender combination. Subsequently, we created a bar chart based on the pivot table data.

Figure 2



The top three most popular shows from February 9th to 12th, 2019, were “Stranger Things,” “Orange is The New Black,” and “American Horror Story,” respectively. The show with the fewest views was “Friends.”

In analyzing the different preferences by gender, we observed that “Stranger Things’ had more female viewers than males. Conversely, “Orange is the New Black” had more male viewers than females, while other shows did not show significant differences in viewership by gender.

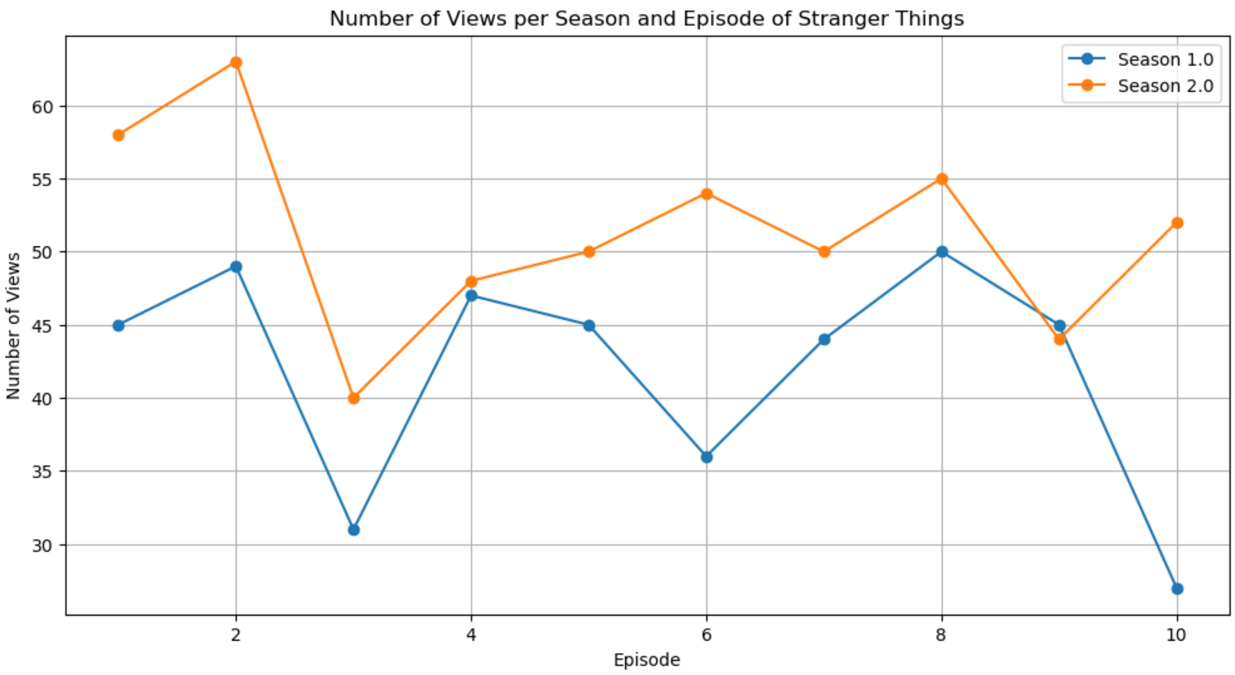
For Netflix, understanding the viewing preferences of its audience and the popularity of different shows is crucial for several reasons:

1. **Targeted Marketing:** Understanding the popularity of different shows helps Netflix target its marketing and promotional efforts more effectively. They can tailor their campaigns to appeal to specific audience segments based on these viewing preferences.
2. **Customer Retention and Growth:** Offering a diverse range of popular shows that cater to different audience segments can improve customer retention and attract new subscribers, ultimately leading to business growth.

Q2: What is the number of views for each season and episode of 'Stranger Things'

To answer this question, we filtered the data frame for only “Stranger Things,” calculated the number of views for each season and episode, and plotted the result using a line chart.

Figure 3



Comparison of the Two Seasons:

- Season 2 generally had higher views compared to Season 1, with more episodes having views above 45. This indicates that Season 2 was more popular overall, as more episodes in Season 2 received higher viewership compared to Season 1.
- Season 1 had more fluctuation in views, with some episodes having significantly lower views compared to Season 2. Each episode likely featured key moments or events that captured audience interest, leading to fluctuations in viewership, indicating that Episodes 3, 6, and 10 might not have been as attractive to the audience.
- Despite these fluctuations, both seasons were able to maintain audience engagement, as evidenced by the consistent pattern of views increasing and decreasing throughout each season.

Q3: How does the total time watched vary by day for the dataset?

To answer this question, we sorted the "Day" column in descending order, starting from Monday to Sunday. We then calculated the total time watched for each day and plotted two dot plots with lines based on the result.

Figure 4

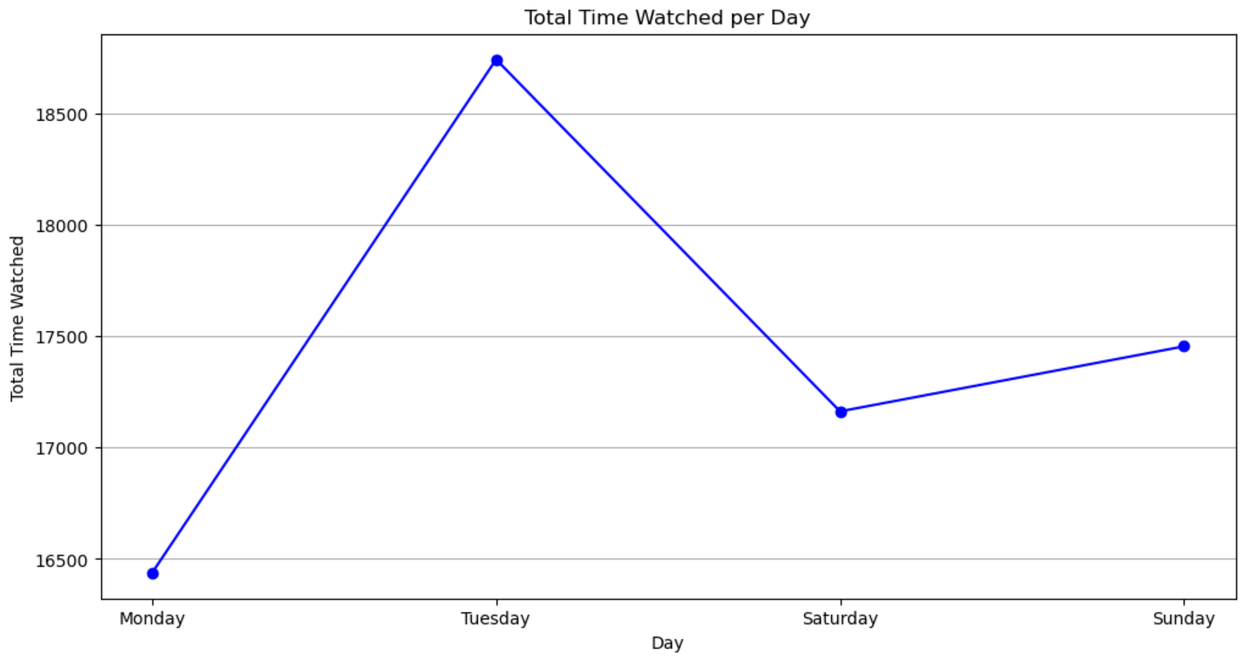
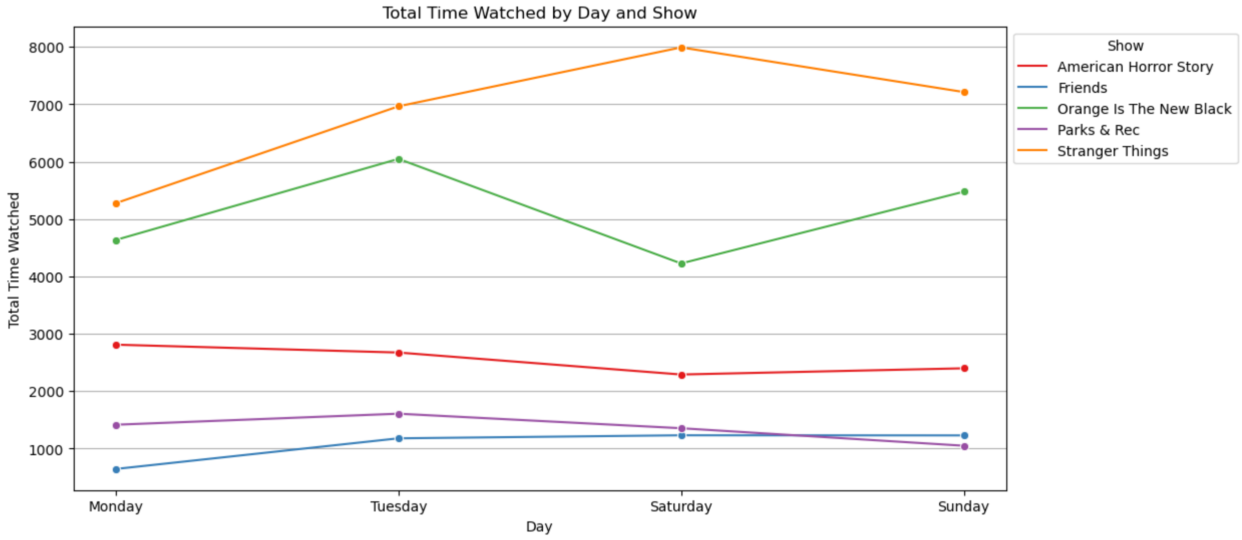


Figure 5



From the results of Q3, we can gain several insights:

1. **Peak Viewing Day:** we identified Tuesday had the relatively high total time watched, indicating peak viewing day. In addition, even though there is some rise on Saturday for “Stranger Things;” however, Tuesday is the overall high views in Figure 5. This information can be valuable for scheduling content releases or promotions to maximize viewership and engagement.
2. **Audience Behavior:** by understanding how total time watched varies by day, we can gain insights into audience behavior patterns. We observed that viewers didn’t watch more content during weekends than on weekdays, which can inform content scheduling decisions.
3. **Show Performance:** Figure 5, showing the total time watched by day and show, helps us understand the popularity of different shows over time. We can identify trends in viewership for each show and use this information to evaluate the performance of existing shows and plan for future content.

Finally Learnings

The analysis of viewing patterns revealed that Tuesday had the highest total time watched, indicating it as a peak viewing day and suggesting that scheduling content releases or promotions on Tuesdays could maximize viewership and engagement. Surprisingly, viewers did not watch more content during weekends than on weekdays, challenging the conventional belief that weekends are prime time for viewership. The analysis also provided insights into the popularity of different shows over time, with “Stranger Things” showing a noticeable rise in views on Saturdays, indicating its weekend popularity. However, Tuesday stood out as the overall high views day, suggesting a consistent trend across all shows. These findings provide valuable insights into viewing patterns, audience preferences, and show performance, enabling Netflix to make informed decisions to enhance its content strategy and drive viewer engagement.

Reference

Data Source:

ALY6040 Data Mining Application Course of Northeastern University, https://northeastern.instructure.com/courses/165136/assignments/2071167