### Final Report

This document serves as the **final report** for my individual data science project titled "Can Netflix Genre Preferences Predict Menstruation Periods?". The project combines personal behavioral data with statistical and machine learning techniques to uncover meaningful patterns in entertainment consumption and its potential links to menstrual cycles.

Prepared by:  
**Pırıl Deniz Zorlutuna**  
**Student ID: 32160**  
**Course: DSA210 – Introduction to Data Science**  
**Semester: Spring 2024-2025**

Thank you for reviewing my work.

**Introduction**

This project investigates whether hormonal fluctuations associated with the menstrual cycle influence personal media consumption habits—specifically genre preferences and viewing intensity on Netflix. The analysis is based on two personal datasets: menstruation records tracked via the Apple Health app since 2019, and Netflix viewing history collected from 2018 onward.

Throughout the study, I have combined and preprocessed these data sources to examine how menstruation status may correlate with changes in the type and frequency of content consumed. The primary focus is to identify any behavioral patterns that emerge during menstruation periods compared to non-menstruation periods, using both statistical and machine learning techniques.

The project follows a structured approach: data preparation, exploratory data analysis (EDA), hypothesis testing, and predictive modeling. In particular, it aims to answer the following research question:

**Hypotheses**

* **Null Hypothesis (H₀)**: Menstruation has no significant impact on viewing behavior (e.g., genre preference, view count).
* **Alternative Hypothesis (H₁)**: Menstruation significantly influences viewing behavior, leading to measurable differences in genre preference and/or view count.

At this final stage of the project, this report presents the complete analysis and findings. All results are critically examined, including visual exploration of genre trends, statistical tests to validate observed differences, and classification models developed to predict menstruation status based on Netflix viewing features.

Beyond evaluating the personal impact of hormonal cycles on entertainment habits, this project also reflects on the potential for integrating such behavioral insights into personalized content recommendation systems or targeted marketing strategies in the future.

**Exploratory Data Analysis (EDA)**

To better understand the structure and distribution of the data, exploratory data analysis was conducted on both temporal and content-based features. This phase aimed to uncover patterns in viewing behavior related to genre preferences, watch frequency, weekdays, seasons, and menstruation status.

The visualizations provided in this section highlight key trends in the dataset, such as dominant genres, viewing intensity by time variables, and differences observed during menstruation versus non-menstruation periods. These insights offer a foundation for subsequent hypothesis testing and machine learning tasks.

Plot 1 – Dominant Genre Frequency

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This bar chart displays the frequency of each dominant genre watched in the dataset. **Drama** stands out as the most frequently consumed genre, followed by **Comedy** and **Animation**. The clear dominance of emotionally engaging genres suggests a preference for narrative-driven or comforting content, which will be further explored in relation to menstruation status in later analyses.

Plot 2 – View Count Distribution

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This histogram shows the distribution of how many times each title was watched. The data is **right-skewed**, meaning most contents were watched only once or twice, with a few being viewed more frequently. This suggests that binge-watching or repeated viewing occurred only for selected shows, while casual one-time viewing was more common overall.

Plot 3 – View Count by Genrediyagram, çizgi, ekran görüntüsü, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This boxplot compares the distribution of view counts across different genres. While the median number of views remains fairly similar for most genres, **Comedy** and **Historical Drama** show higher variability and upper-range outliers, suggesting they were more frequently rewatched. In contrast, genres like **Sci-Fi**, **Reality**, and **Documentary** had consistently low view counts, indicating lower engagement or one-time viewing behavior.

Plot 4 – View Count by Weekday diyagram, çizgi, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This boxplot illustrates how viewing behavior varies by day of the week. The median view count remains relatively stable across all days; however, **Wednesday** shows a slightly higher median and a wider interquartile range, indicating more variability in viewing intensity. Several outliers are present throughout the week, especially on **Friday and Saturday**, which may reflect occasional binge-watching patterns tied to weekend routines.

Plot 5 – View Count by Season

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This boxplot shows the variation in view counts across different seasons. **Summer** stands out with a higher median and greater spread, suggesting that viewing activity tends to increase during this period—possibly due to more free time such as summer breaks. In contrast, **Winter, Spring,** and **Fall** show similar and more compact distributions, with fewer extreme values and lower maximum view counts.

Plot 6 – Genre by Menstruation Status

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This grouped bar chart compares the frequency of watched genres between menstruation and non-menstruation days. While **Drama** remains the dominant genre in both periods, it is relatively more prominent during menstruation. Other genres such as **Animation** and **Comedy** also show higher counts during menstruation, suggesting a possible shift toward emotionally comforting or light-hearted content during that time. These trends align with the project’s hypothesis that genre preference may vary based on hormonal cycles.

Plot 7 – Average View Count by Menstruation

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This bar plot compares the average number of contents watched per day during menstruation and non-menstruation periods. The average view count is slightly lower on menstruation days; however, the difference is minimal. This aligns with the t-test results, which showed no statistically significant difference between the two groups (p = 0.4924). Therefore, while genre preference may vary, overall viewing volume appears to remain stable across hormonal phases.

Plot 8 – Genre Trend Over Time (30-Day Rolling Average)diyagram, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This line plot illustrates the normalized frequency of each genre over time, smoothed with a 30-day rolling average. The results reveal clear **temporal patterns** in genre preferences. **Drama** consistently remains the most dominant genre, while others like **Comedy**, **Animation**, and **Crime** exhibit fluctuating popularity over the years. These patterns highlight evolving viewing interests and external factors (e.g., seasonal mood, availability of new content), and suggest long-term shifts in emotional or narrative content preferences.

Plot 9 – Monthly Average View Countmetin, ekran görüntüsü, yazı tipi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This bar chart displays the average number of contents watched per day for each month. The data reveals a noticeable **peak in June (month 6)**, where the average daily view count exceeds 4. This suggests a seasonal pattern, likely influenced by summer breaks or reduced obligations during that time. Other months show relatively stable viewing behavior, with slight increases in **November** and **January**, potentially related to holidays or exam recovery periods.

Plot 10 – Viewing by Weekdaymetin, ekran görüntüsü, yazı tipi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This bar chart shows the total number of contents watched for each day of the week. **Saturday** clearly stands out as the most active viewing day, followed by **Sunday** and **Friday**, indicating a strong weekend viewing trend. Weekdays such as **Tuesday** and **Thursday** have relatively lower totals, which may reflect academic or work-related constraints. This pattern supports the idea that leisure time availability significantly influences viewing behavior.

**Plot 11 – Viewing by Season**

**metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

This bar chart presents the total number of contents watched across seasons. **Summer** has the highest total view count, followed by **Spring**, while **Fall** and **Winter** show significantly lower activity. The increased viewing during summer may be attributed to academic breaks or holidays, offering more free time for entertainment. These seasonal differences suggest that environmental or lifestyle factors may shape media consumption behavior independently of menstruation status.

Plot 12 – View Count by Genre & Menstruation

diyagram, metin, ekran görüntüsü, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This grouped boxplot compares view count distributions for each genre, split by menstruation status. For most genres, the median view count remains similar between menstruation and non-menstruation periods. However, **Historical Drama**and **Comedy** show notably higher view counts during menstruation days, while **Drama** and **Animation** are relatively stable across both conditions. These observations suggest that while overall viewing frequency may not change drastically, **certain genres become more prominent during menstruation**, supporting the idea of mood-driven content preferences.

Plot 13 – Genre vs Weekday Heatmap

metin, ekran görüntüsü, dikdörtgen, paralel içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This heatmap visualizes the distribution of genre preferences across different days of the week. **Drama** is the most consistently watched genre every day, peaking on **Saturday**. Other genres like **Comedy**, **Animation**, and **Crime** are also regularly consumed, with noticeable activity midweek and over the weekend. Less common genres such as **Historical Drama**, **Horror**, and **Psychological** appear sporadically. The heatmap highlights **Saturday and Wednesday** as high-activity days across multiple genres, reinforcing patterns seen in earlier weekday analyses.

Plot 14 – Correlation Matrix

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This heatmap displays the Pearson correlation coefficients between numerical variables: view\_count, year, month, and day. As shown, **no strong correlations** exist among these variables. The highest observed relationship is between view\_count and month (0.047), which is still negligible. This indicates that **temporal variables alone do not linearly explain changes in view count**, suggesting that genre, mood, or hormonal cycles may play a more significant role in influencing viewing behavior.

Plot 15 – Pairplot of Viewing Features

metin, ekran görüntüsü, diyagram, paralel içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This pairplot visualizes the relationships between key numerical variables: view\_count, year, month, and day. The diagonal shows the distribution of each variable individually, while the scatterplots in the off-diagonal sections reveal how these features interact. Notably, **view\_count is right-skewed**, and no strong linear relationships are visible between view count and date components. However, we observe that **recent years (2023–2025)** include more high-count entries, which might reflect increased viewing frequency or better data logging. The plot also confirms the **temporal diversity and completeness** of the dataset.

Plot 16 – Monthly Comparison of Menstruation Days and Netflix Viewing Days

çizgi, metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This line chart compares the **number of menstruation days and Netflix viewing days** on a monthly basis from 2019 to 2025. While menstruation days fluctuate month to month (generally between 5–10 days), **Netflix viewing days appear consistently low and static**, forming a flat line. This visual contrast highlights a limitation in the viewing dataset: the **granularity of viewing records** does not reflect individual viewing days, likely due to aggregation during preprocessing. Still, the menstrual cycle trend is visibly regular and cyclical, forming a basis for potential behavioral prediction

**Plot 17 – 30-Day Rolling Average of Menstruation and Netflix Viewing**

metin, çizgi, yazı tipi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

This line graph displays the **30-day rolling average** of menstruation and Netflix viewing occurrences. The **menstruation trend line exhibits periodic fluctuations**, consistent with biological cycles. In contrast, the Netflix viewing line remains flat at the bottom, indicating that daily viewing activity was not granular or consistently recorded on a per-day basis in the dataset. This reinforces the limitation identified earlier: **Netflix viewing behavior appears underrepresented at the daily level**, making it difficult to correlate with natural rhythms like menstruation in a temporal trend analysis.

### **Machine Learning Analysis**

In the final phase of the project, machine learning techniques were employed to test whether menstruation periods could be predicted based on Netflix viewing habits. The classification models used were:

* K-Nearest Neighbors (KNN)
* Logistic Regression
* Random Forest

These models were trained using features such as genre, viewing time, weekday, season, and view count, with the target variable being whether the day was within a menstruation period. SMOTE was applied to address class imbalance.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.1. K-Nearest Neighbors (KNN)

KNN achieved:

* **High recall (0.92)** for menstruation days, meaning it correctly identified most true positive cases.
* However, **precision was low (0.69)**, leading to many false positives.
* This model is better for cases where missing a menstruation period is more critical than predicting a false one.

2. Logistic Regression

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Logistic Regression showed:

* **High precision (0.95)** and **overall balanced performance**.
* Recall (0.66) was lower compared to KNN, meaning more menstruation days were missed, but fewer false alarms were raised.
* Given its interpretability and solid F1-score (0.78), it is suitable for understanding which features are most influential.

3. Random Forest

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Random Forest performed the best overall:

* **Accuracy: 86%**, **Precision: 0.94**, **Recall: 0.76**, **F1-Score: 0.84**
* Balanced performance across all metrics.
* As an ensemble method, it reduces overfitting and captures complex patterns.

### Overall Comparison

| **Model** | **Accuracy** | **Precision (1)** | **Recall (1)** | **F1-Score (1)** |
| --- | --- | --- | --- | --- |
| K-Nearest Neighbors | 75% | 0.69 | 0.92 | 0.79 |
| Logistic Regression | 81% | 0.95 | 0.66 | 0.78 |
| Random Forest | 86% | 0.94 | 0.76 | 0.84 |

* **Random Forest** emerged as the best-performing model, striking a strong balance between precision and recall.
* These results demonstrate that personal behavioral data, even with limited features, can be predictive of sensitive conditions such as menstruation status.

This opens a path for further work on personalized content delivery, suggesting that future platforms could adapt recommendations based on inferred hormonal cycles—if such data is responsibly and ethically handled.

### Conclusion

This project explored the relationship between menstruation status and Netflix viewing behavior using personal data spanning several years. Through exploratory data analysis, statistical testing, and machine learning, several key insights were discovered:

* **Genre preferences** significantly differ during menstruation, with a higher tendency toward genres like drama and comedy.
* **Total view count** did not show a statistically significant change between menstruation and non-menstruation periods.
* **Seasonal and weekday patterns** influenced viewing habits—summer and weekends saw the highest activity.
* **Machine learning models**, especially Random Forest, were able to predict menstruation status from viewing behavior with up to **86% accuracy**, indicating that personal media habits carry behavioral signals related to hormonal changes.

These findings suggest that with appropriate privacy measures, platforms could one day use such behavioral patterns to inform **personalized content recommendations** or **wellness insights**, enhancing user experience and self-awareness.