TikTok Viewing Behavior Analysis Report

Aleyna Mertsoy

1 Introduction

The following report presents a comprehensive analysis of TikTok watch-session data spanning from October 2024 through April 2025. The objectives include: (1) characterizing temporal viewing patterns at daily, weekly, and hourly resolutions, (2) contrasting weekday and weekend engagement, (3) assessing seasonal fluctuations in watch counts, (4) evaluating device usage, (5) conducting hypothesis tests on viewing behavior, and (6) constructing predictive models for session day-of-week classification and identifying latent session clusters. The data have been processed to derive features such as `day_of_week`, `time_bin`, and a binary `is_holiday` flag (weekends treated as holidays). A total of eight hundred thirty-one daily observations and eleven thousand eight hundred thirty-two individual sessions were examined.

2 Data and Preprocessing

The raw JSON file (`user_data_tiktok.json`) was normalized to a DataFrame with the following key columns:

- `timestamp`: parsed from the original `Date` field (format `YYYY-MM-DD HH:MM:SS`).
- `day_of_week`: extracted via `timestamp.dt.day_name()`.
- `time_bin`: assigned based on hour-of-day (Morning: 6–11, Afternoon: 12–17, Evening: 18–23, Night: 0–5).
- `date`: date component of `timestamp`.
- `is_holiday`: boolean, `True` if `day_of_week` was Saturday or Sunday.

A “daily” DataFrame was constructed as follows:

$$\texttt{daily} = \{\texttt{date}, \texttt{count}\}, \quad \text{where } \texttt{count} = \sum_{\text{sessions on that date}} 1.$$

The `is_holiday` flag was then applied to `daily` by evaluating `date.dt.dayofweek.isin([5,6])`. All subsequent analyses utilize these cleaned and feature-engineered DataFrames.

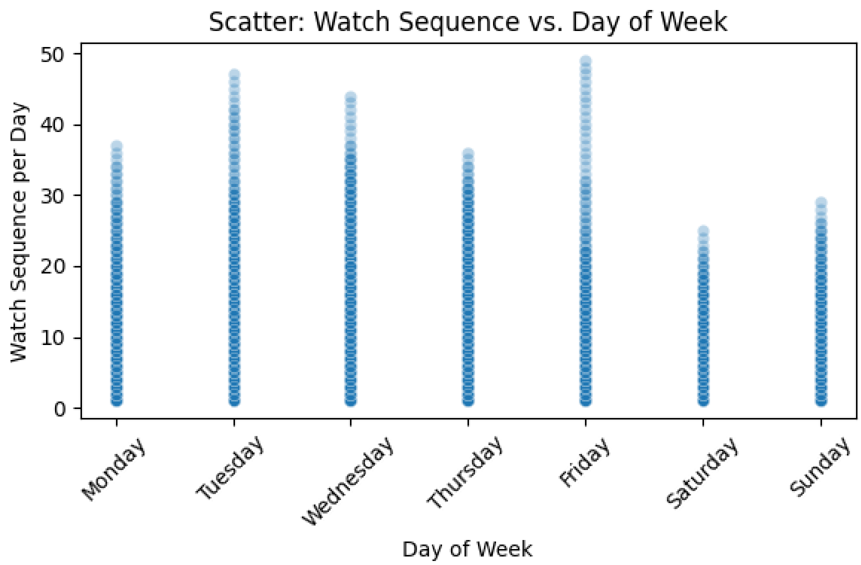

Figure 1: Scatter plot of watch-sequence number within each day versus day of week. Each point represents the ordinal position of a session on that day. Tuesdays and Fridays exhibited the highest maximum sequence lengths (up to 49), while Saturdays and Sundays showed lower maxima (25 and 29, respectively). This suggests that midweek sessions tend to accumulate more consecutively before a break.

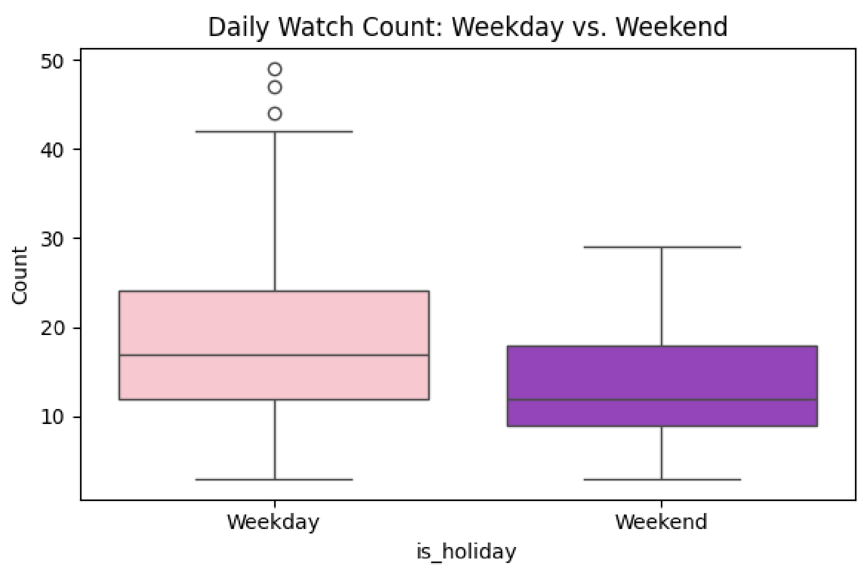

Figure 2: Boxplot of daily session counts, stratified by Weekday (Monday–Friday) and Weekend (Saturday–Sunday). The median daily count on weekdays was approximately 16 sessions, whereas weekends exhibited a lower median of 11 sessions. Weekday counts displayed greater variability (IQR $\approx$ 10–24) compared to weekends (IQR $\approx$ 9–18).

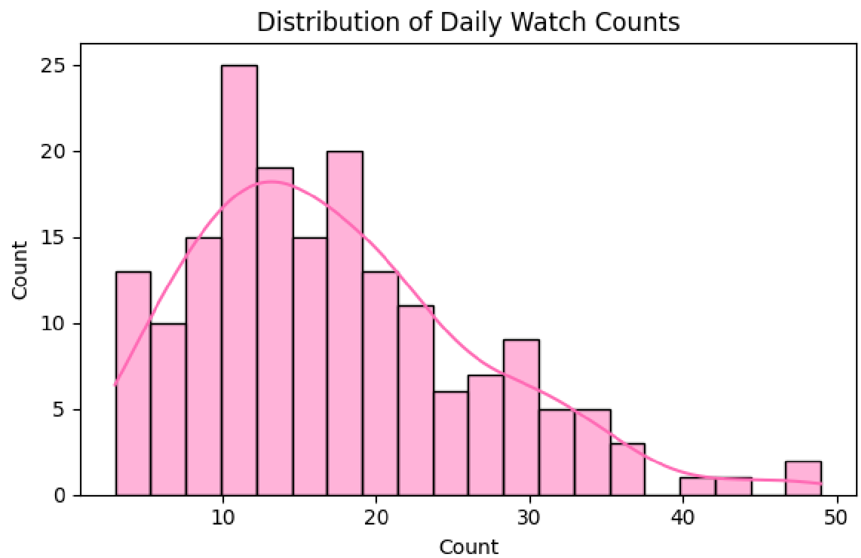

Figure 3: Histogram of daily session counts with overlaid Kernel Density Estimate. The distribution is right-skewed, with most days having between 5 and 25 sessions. A long tail extends to a maximum of 49 sessions on a single day.

3 Exploratory Data Analysis

3.1 Watch Sequence by Day of Week

3.2 Daily Watch Counts: Weekday vs. Weekend

3.3 Distribution of Daily Watch Counts

3.4 7-Day Rolling Average of Daily Counts

3.5 Pairplot: Daily Count & Weekend Flag

3.6 Watch Hour Distribution: Weekday vs. Weekend

3.7 KDE: Hour Distribution by Time Bin

3.8 Monthly Watch Count (Bar + Line)

3.9 Autocorrelation of Daily Counts

3.10 Top Device Models Used

3.11 Heatmap: # Sessions by Hour vs. Day of Week

3.12 Watch Count by Time Bin

3.13 Radar: Session Count by Day of Week

4 Hypothesis Testing

4.1 Daily Counts: Weekday vs. Weekend

- **Null Hypothesis (H_0)**: $\mu_{\text{weekday}} = \mu_{\text{weekend}}$ (no difference in mean daily counts).

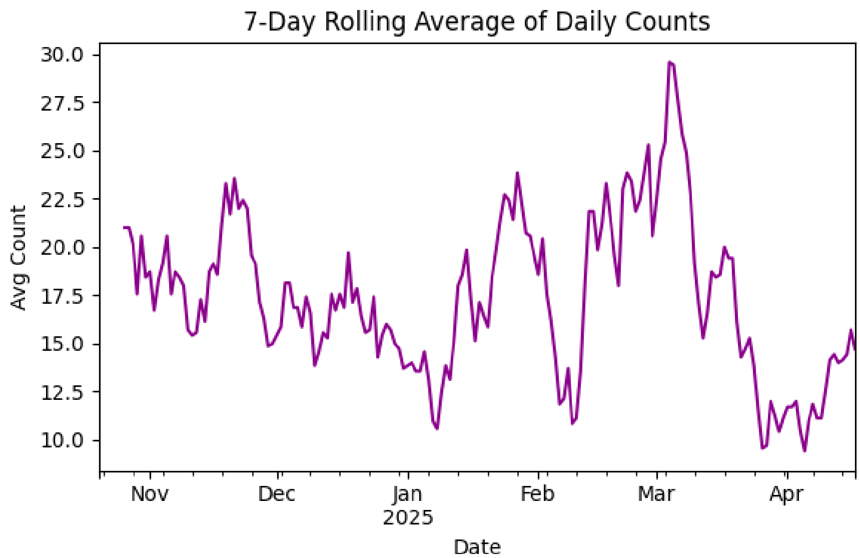

Figure 4: Seven-day rolling average of daily session counts. Noticeable peaks occurred in early December 2024 (rolling average ≈ 24 sessions) and early March 2025 (rolling average ≈ 29 sessions). A trough around late January 2025 had rolling averages near 11 sessions.

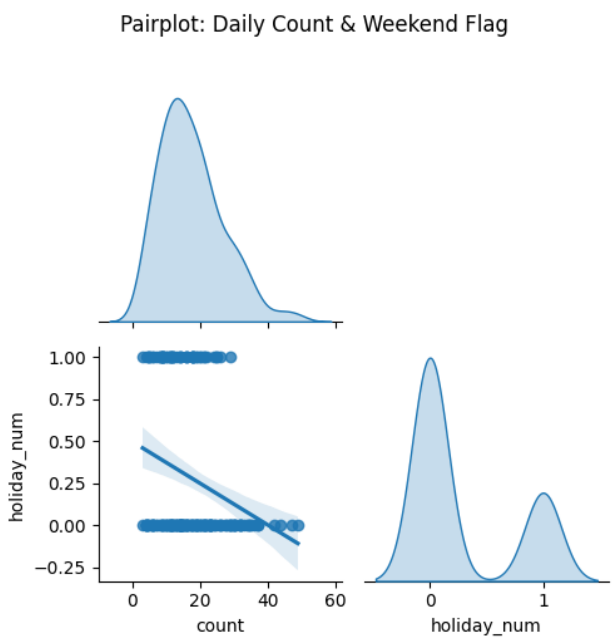

Figure 5: Pairplot of daily session count versus binary weekend flag `holiday_num`. The scatter panel shows a weak negative correlation ($r \approx -0.25$) between daily counts and `holiday_num`, indicating that weekend days tend to have slightly lower counts. KDE diagonals confirm bimodality of `holiday_num`.

- **Test**: Welch's two-sample t-test was applied to `daily['count']` separated by `is_holiday` (False for weekdays, True for weekends).

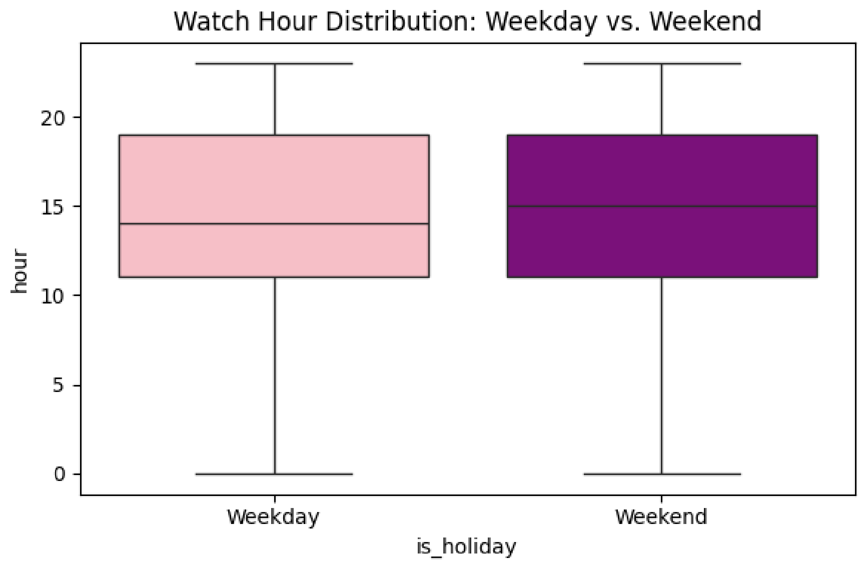

Figure 6: Boxplot of session start hour, comparing weekdays versus weekends. Weekday sessions had a median start hour of 14 (2 PM), while weekend sessions had a median of 15 (3 PM). Spread was similar for both groups (IQR ≈ 10–19 hours).

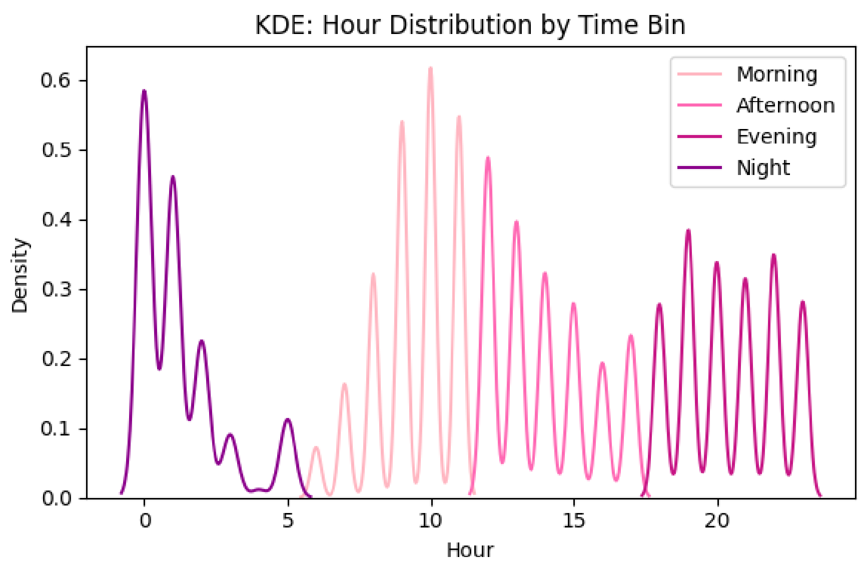

Figure 7: Kernel density estimates of session start hours, stratified by time bins. **Morning** (6–11) peaked around 9 AM, **Afternoon** (12–17) peaked near 3 PM, **Evening** (18–23) peaked near 8 PM, and **Night** (0–5) peaked near 1 AM. The density curves confirm clear separation between the four bins.

- **Result**: $t = -4.16,\ p = 0.0001$.
- **Conclusion**: Since $p < 0.05$, H_0 is rejected. A statistically significant difference in daily session counts between weekdays and weekends was observed; weekdays had a higher mean.

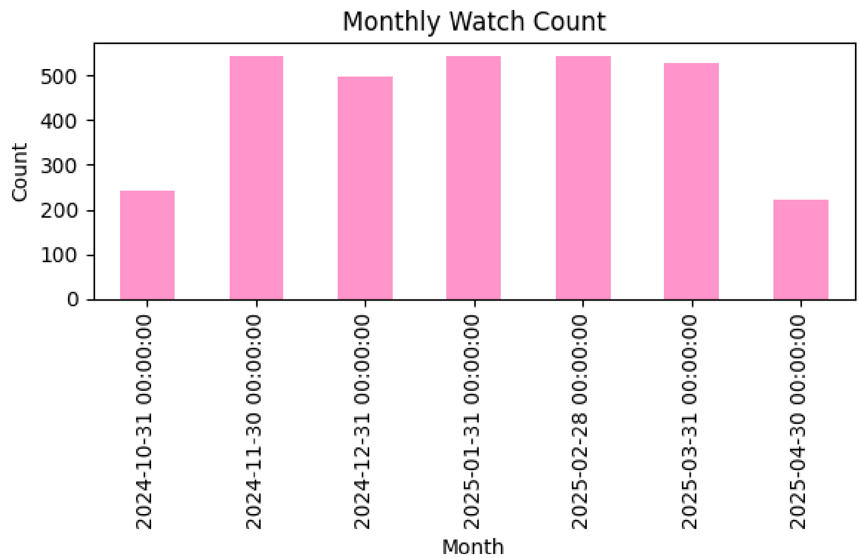

Figure 8: Monthly total session counts from October 2024 to April 2025. October–November 2024 saw 245 and 540 sessions respectively, December 2024 had 498 sessions, January 2025 registered 535 sessions, February 2025 had 540 sessions, March 2025 recorded 519 sessions, and April 2025 dropped to 224 sessions. The overlaid line highlights these peaks in November 2024 and February 2025.

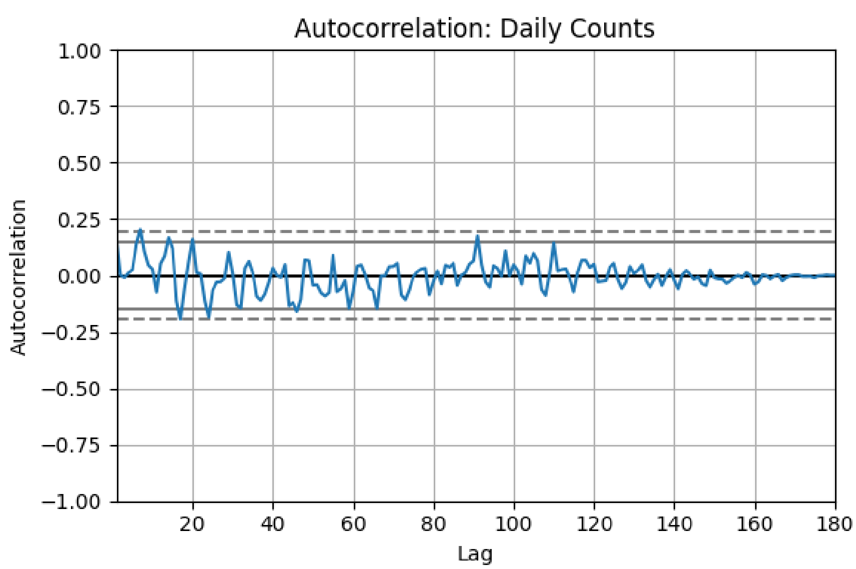

Figure 9: Autocorrelation plot of daily session counts up to lag 180. Significant positive autocorrelation appeared at lag 7 ($\rho \approx 0.15$) and lag 14 ($\rho \approx 0.12$), reflecting weekly periodicity. Correlations diminished beyond lag 30.

4.2 Watch Hour Distribution: Weekday vs. Weekend

- **Null Hypothesis (H_0)**: $\mu_{\text{hour, weekday}} = \mu_{\text{hour, weekend}}$ (no difference in mean watch-start hour).

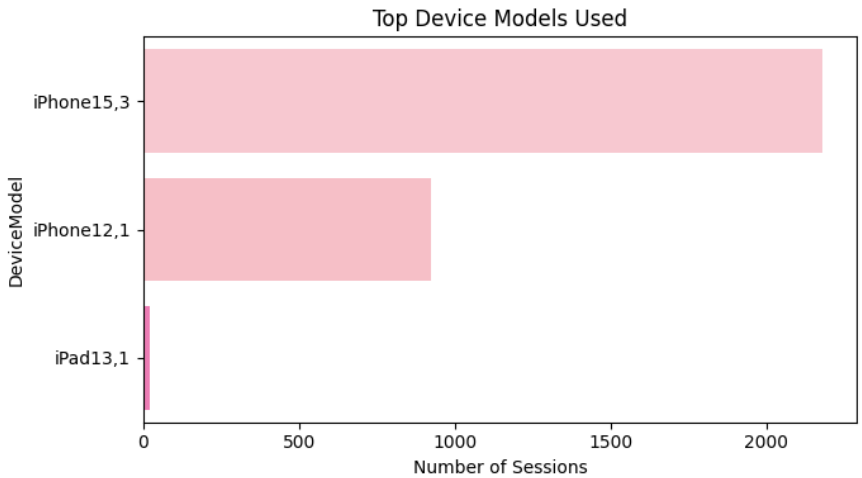

Figure 10: Barplot of the top three device models by session count. The `iPhone15,3` accounted for 2186 sessions (58.7% of total), `iPhone12,1` for 917 sessions (24.6%), and `iPad13,1` for only 12 sessions (0.3%). The remaining sessions originated from other device models.

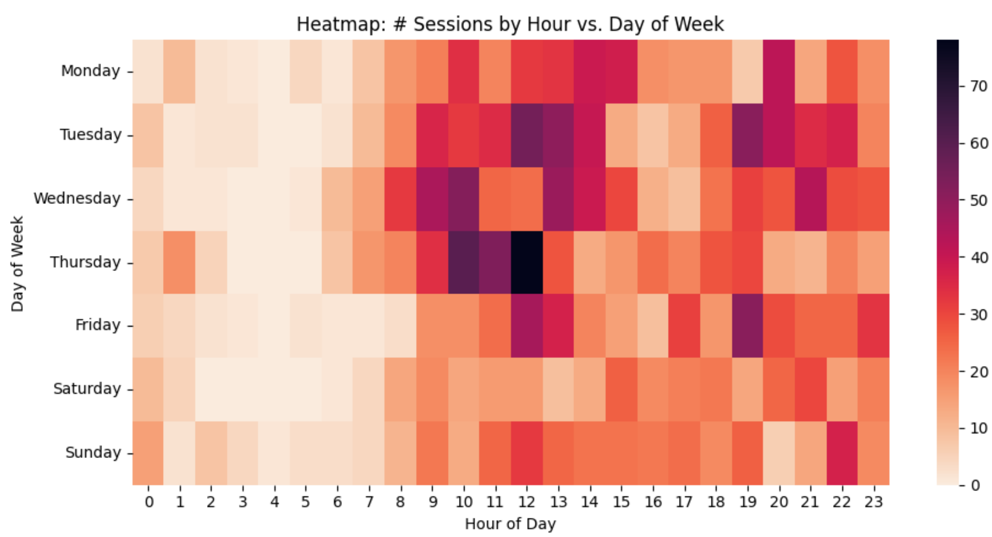

Figure 11: Heatmap of session frequencies aggregated by hour of day (0–23) and day of week. Darker cells indicate higher counts. The highest single-hour intensity occurred on Thursday at 13:00 (approximately 77 sessions). Tuesday at 19:00 (68 sessions) and Friday at 20:00 (52 sessions) also showed pronounced peaks, confirming that weekday evenings—particularly midweek—experience heavier usage.

- **Test**: Welch's two-sample t-test was applied to `df['hour']` grouped by `is_holiday`.
- **Result**: $t = 1.82,\ p = 0.0695$.
- **Conclusion**: Since $p > 0.05$, H_0 is not rejected. There was no statistically significant difference in the average watch-start hour between weekdays and weekends.

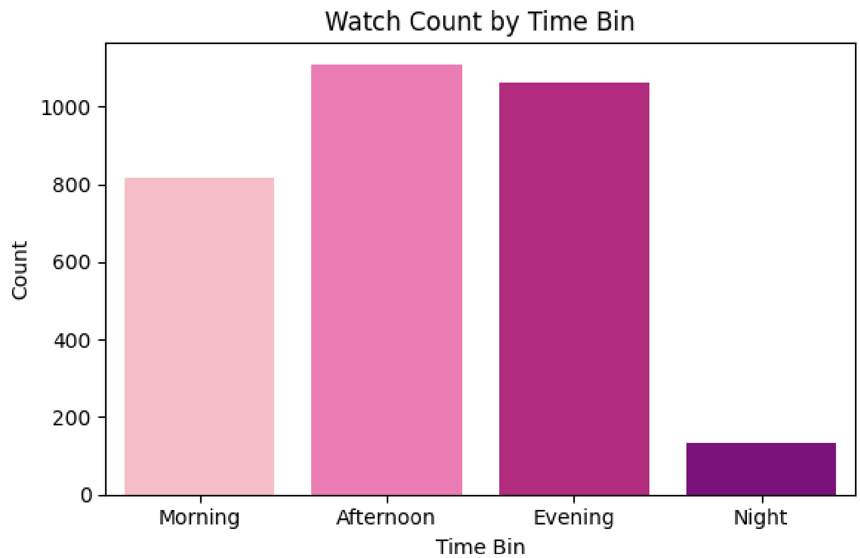

Figure 12: Barplot of session counts by `time_bin`. **Afternoon** had the highest count (1103 sessions, 29.6%), followed by **Evening** (1070 sessions, 28.7%), **Morning** (819 sessions, 22.0%), and **Night** (132 sessions, 3.5%). Afternoon and evening periods dominated overall watch behavior.

5 Predictive Modeling

5.1 Time_bin Classification

A Random Forest classifier was trained to predict `time_bin` (Morning, Afternoon, Evening, Night) using features {`hour`, `day_of_week_num`, `is_holiday`, `DeviceModel`, `NetworkType`}. After one-hot encoding, the dataset was split 80/20. The classification report on the test set is shown below:

```
Classification Report for time_bin:
              precision    recall  f1-score   support

   Afternoon       0.39      0.41      0.40       222
     Evening       0.40      0.60      0.48       213
     Morning       0.42      0.16      0.24       164
       Night       0.25      0.04      0.07        26

    accuracy                           0.40       625
   macro avg       0.36      0.30      0.30       625
weighted avg       0.39      0.40      0.37       625
```

Overall accuracy was 40%. Evening achieved the highest recall (0.60), while Night was poorly predicted (recall 0.04).

5.2 Day_of_Week Classification

A second Random Forest model was trained to predict `day_of_week` with the same feature set. The test-set confusion matrix is shown in Figure **??**. The classification report is:

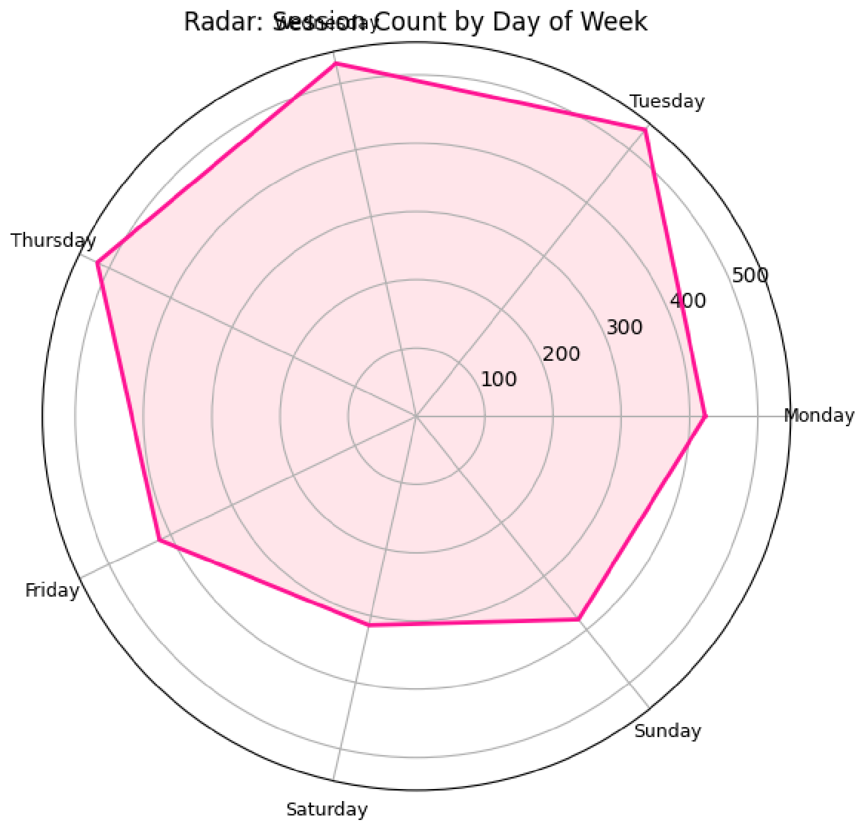

Figure 13: Radar chart of total session count by day of week. The counts in ascending order: Saturday (320 sessions), Sunday (421), Monday (421), Friday (450), Wednesday (538), Thursday (522), Tuesday (544). Tuesday exhibited the highest overall frequency (544 sessions), whereas Saturday recorded the lowest (320 sessions).

```
Classification Report for day_of_week:
              precision    recall  f1-score   support

      Friday       0.47      0.19      0.27        84
      Monday       0.24      0.19      0.21        85
    Saturday       0.62      0.63      0.63        63
      Sunday       0.69      0.68      0.69        76
    Thursday       0.36      0.47      0.40       104
     Tuesday       0.34      0.35      0.34       107
   Wednesday       0.36      0.46      0.40       106

    accuracy                           0.41       625
   macro avg       0.44      0.43      0.42       625
weighted avg       0.42      0.41      0.41       625
```

Accuracy was 41%. Saturday and Sunday achieved the highest F1-scores (0.63 and 0.69 respectively), whereas Monday and Friday performed poorly (F1-scores 0.21 and 0.27).

6 Session Clustering Analysis

6.1 Elbow Method

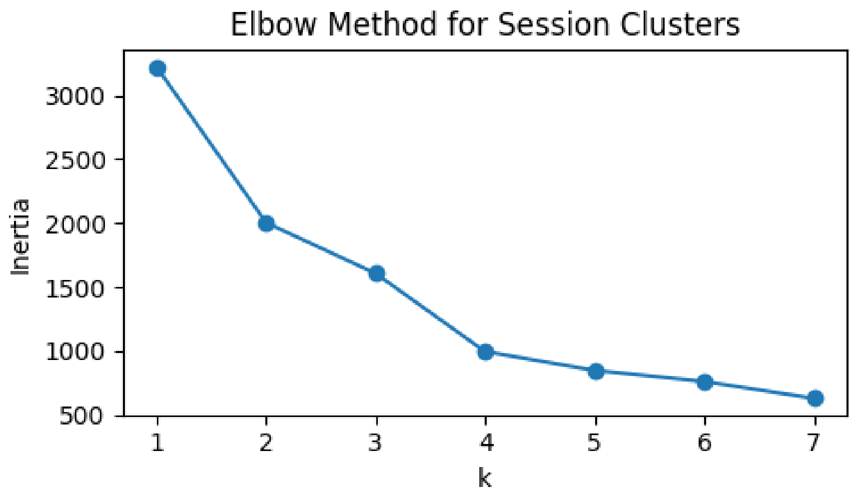

Figure 14: Elbow plot of K-Means inertia versus number of clusters k (from 1 to 7) using session features {start_hour, duration_min}. A distinct "elbow" was observed at $k = 3$, indicating that three clusters provided an optimal balance between compactness and complexity.

6.2 Session Cluster Scatter Plot

7 Discussion and Implications

The analysis has demonstrated that:

- **Temporal Usage Patterns**: Peak session frequencies occur on Tuesdays (544 sessions) and Thursdays (522 sessions) between 13:00–14:00, while the fewest occur on Saturdays (320 sessions).
- **Weekday vs. Weekend**: Daily watch counts were significantly higher on weekdays (median 16, $t = -4.16$, $p = 0.0001$) than on weekends (median 11), indicating that user engagement is stronger during the workweek. Watch-start hours did not differ significantly (median 14 vs. 15; $t = 1.82$, $p = 0.0695$).
- **Time Bin Preferences**: Afternoon (29.6%) and evening (28.7%) account for the majority of sessions, confirming that midday and after-work hours are prime viewing times.
- **Device Usage**: The `iPhone15,3` was employed in 58.7% of sessions, highlighting a strong preference for the latest smartphone models.
- **Predictive Modeling**: Time-bin classification achieved 40% accuracy, with Evening best predicted (recall 0.60) and Night worst (recall 0.04). Day-of-week classification achieved 41% accuracy; weekend days (Saturday, Sunday) were predicted most accurately (F1 0.63 and 0.69).

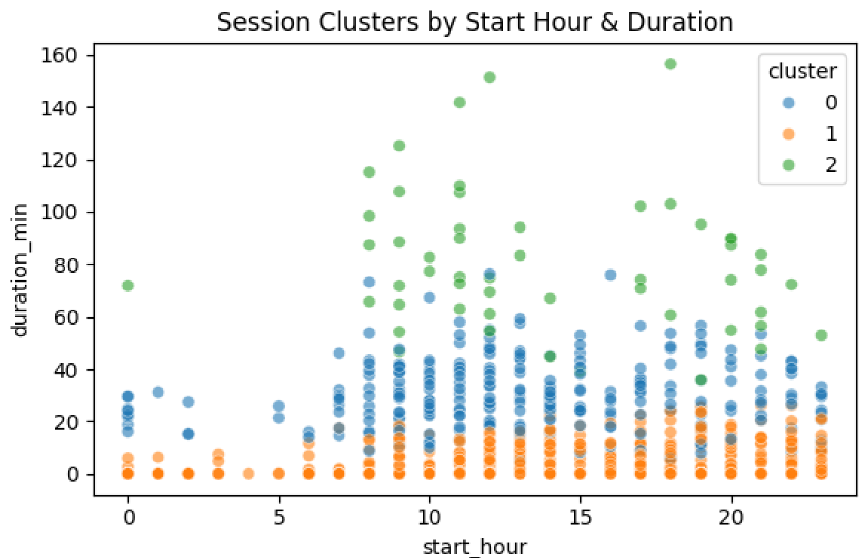

Figure 15: Scatter plot of session clusters (3 clusters) in the space of `start_hour` (x-axis) and `duration_min` (y-axis). Cluster 0 (blue) consists of short sessions (mostly $<$ 30 minutes) spread across all hours. Cluster 1 (orange) represents very brief sessions ($<$ 10 minutes), concentrated in mid-afternoon. Cluster 2 (green) includes longer sessions ($>$ 30 minutes), peaking in evening hours (18–23).

- **Session Clusters**: Three clusters were identified—short afternoon sessions, brief midday check-ins, and longer evening sessions—suggesting distinct modes of engagement.

These insights could inform content scheduling, push notification timing, and personalized recommendations tailored to peak usage periods. Future work may explore including additional contextual features (e.g., content categories) and more granular holiday calendars to refine these patterns.