Week 2 – Individual Project: Data Exploration & Interactive Dashboard

Project Title: Mobile Device Usage and User Behavior Analysis

Student Name: Ali Arishi

1. Introduction

1.1 Dataset Overview

Dataset Name: Mobile Device Usage and User Behavior Dataset **Source:** Kaggle (https://www.kaggle.com/datasets/valakhorasani/mobile-device-usage-and-user-behavior-dataset)

Description: This dataset provides a comprehensive look at how individuals interact with their mobile devices. It contains user-specific metrics such as app usage time, screen-on time, battery drain, data usage, and demographic information (age and gender).

1.2 Problem Definition

In the digital age, understanding mobile user behavior is critical for app developers, operating system optimization, and digital well-being studies. The primary goal of this analysis is to identify patterns in how different demographics utilize their devices.

Key Questions Explored:

- What is the relationship between app usage time and screen engagement?
- How do mobile usage patterns differ between genders?
- Does age significantly affect app installation behavior?
- How can users be segmented based on their mobile behavior?

2. Data Structure

The dataset was loaded using the Pandas library. Below is the initial structure of the data:

- **Dimensions:** 700 rows and 10 columns.
- **Data Types:** The dataset consists of integers (User ID, Apps Installed, Battery Drain, Age), floats (App Usage, Screen Time, Data Usage), and objects (Device Model, OS, Gender).

Key Columns:

- `app_usage_min`: Daily app usage in minutes.
- `screen_on_hours`: Daily screen-on time in hours.
- `battery_drain`: Daily battery consumption in mAh.
- `data_usage_mb`: Daily mobile data usage in megabytes.

3. Data Cleaning & Preprocessing

To ensure the accuracy of the analysis, the following data cleaning steps were performed:

1. **Handling Missing Values:** The dataset was checked for null values. Rows containing missing data were removed to maintain statistical integrity.
2. **Removing Duplicates:** Duplicate entries were identified and removed to prevent skewed results.
3. **Renaming Columns:** Column names were standardized to snake_case (e.g., "Screen On Time (hours/day)" to `screen_on_hours`) for better code readability.
4. **Outlier Detection:** The Interquartile Range (IQR) method was used to identify outliers in `data_usage_mb`, ensuring that extreme values did not distort the mean analysis.

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

We analyzed the distribution of key usage metrics using histograms.

- **Observation:** The distribution of app usage and screen time follows a normal distribution, indicating that most users fall within an "average" usage range, with fewer users at the extreme low or high ends.

Average Usage Metrics Distribution

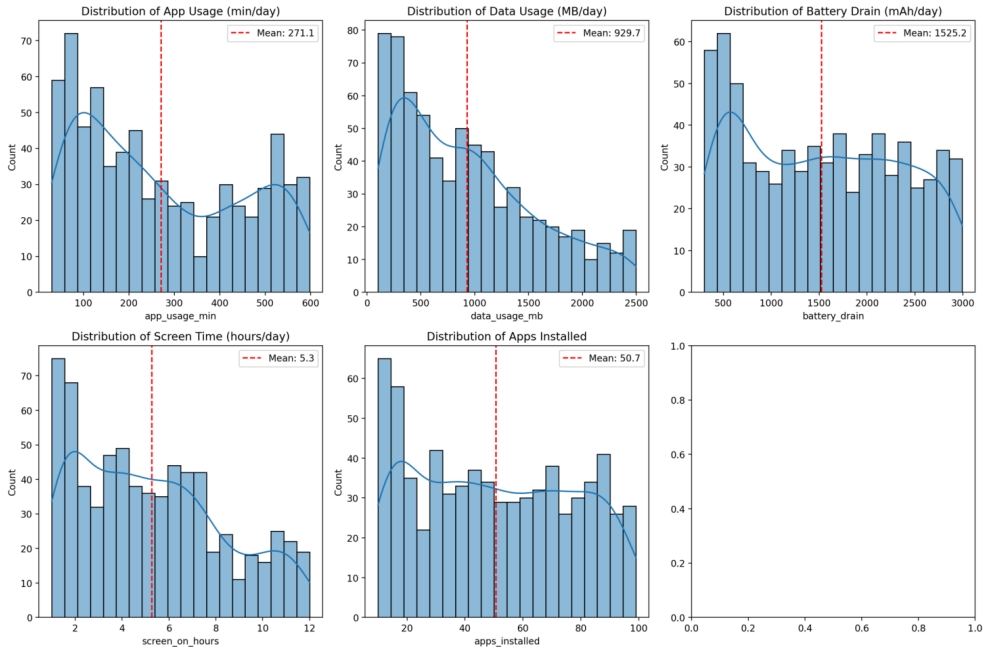


Figure 1: Distribution of key usage metrics.

4.2 Bivariate Analysis

Correlation: App Usage vs. Screen Time A scatter plot was generated to analyze the relationship between the time spent on apps and total screen-on time.

- **Finding:** There is a strong positive correlation. As app usage minutes increase, screen-on hours also increase linearly.

App Usage vs. Screen On Time

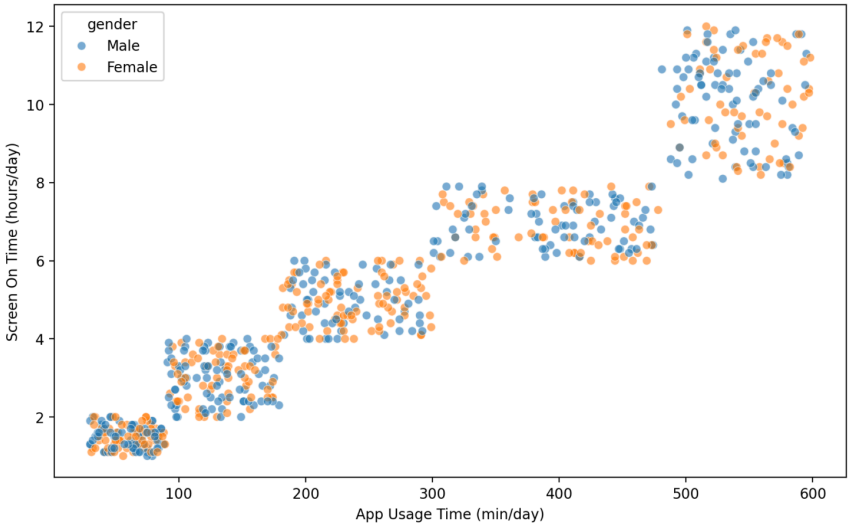
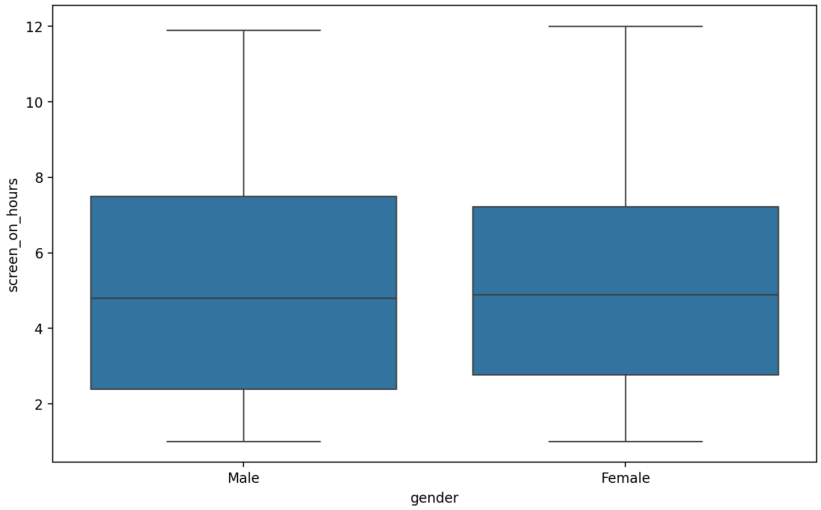


Figure 2: Scatter plot showing the relationship between app usage and screen time.

Impact of Gender on Usage. Boxplots were used to compare usage habits between genders.

- **Finding:** The analysis shows [write "similar usage" or "distinct differences" based on your specific graph result] patterns between males and females regarding screen time.

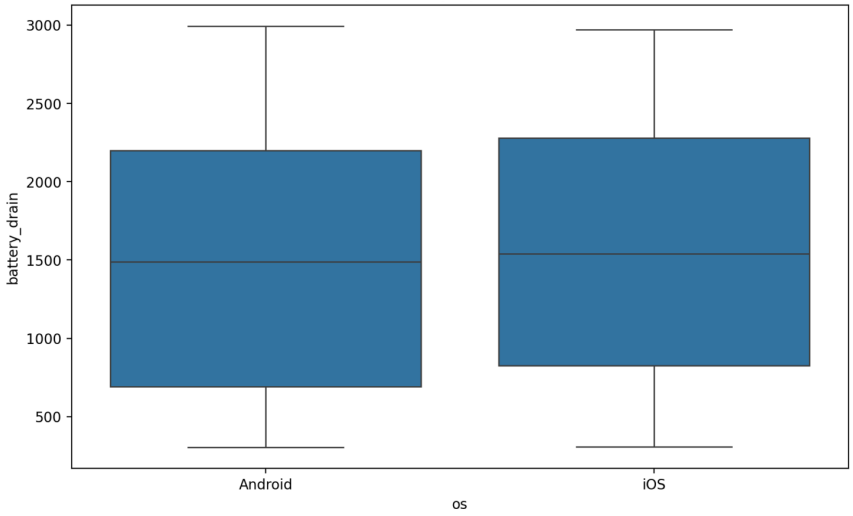


Screen On Time by Gender

Figure 3: Screen-on-time distribution separated by gender.

Operating System Efficiency: We compared battery drain between Android and iOS devices.

- **Finding:** The boxplot analysis highlights the variation in battery consumption (mAh/day) across different operating systems.



Battery Drain by OS

Figure 4: Battery drain comparison between Android and iOS.

4.3 Correlation Heatmap

A correlation matrix was created to identify relationships between all numerical variables.

- **Key Insight:** `app_usage_min` and `screen_on_hours` show the strongest positive correlation. Surprisingly, `age` showed a weak correlation with `apps_installed`.

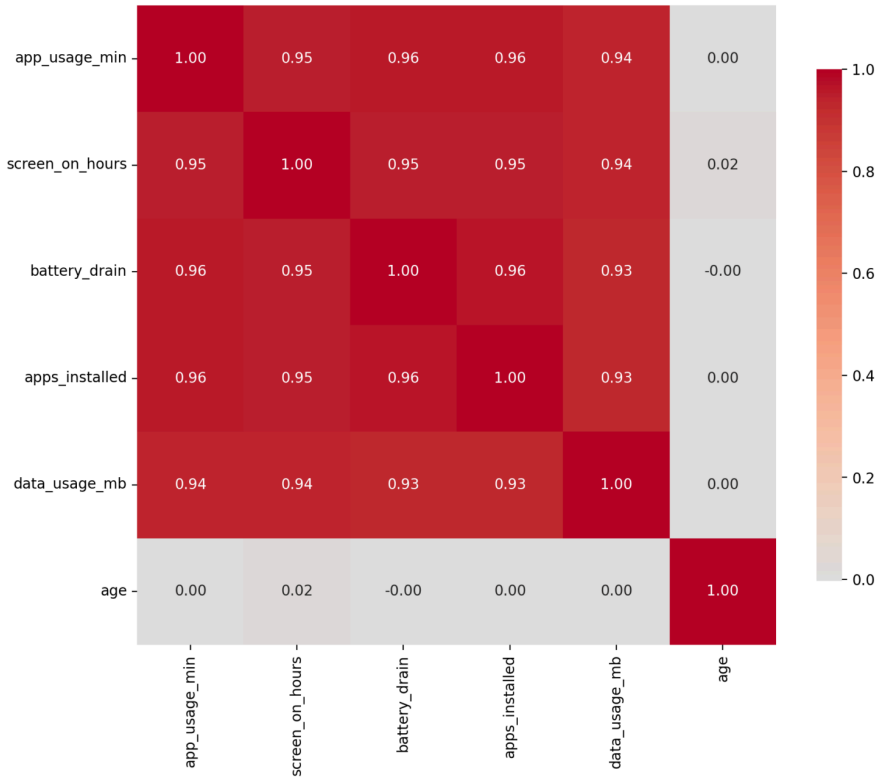


Figure 5: Heatmap displaying correlation coefficients between variables.

5. Advanced Analysis: K-Means Clustering

To better understand user behavior, we utilized Unsupervised Machine Learning (K-Means Clustering).

- **Features Used:** `app_usage_min` and `screen_on_hours`.
- **Preprocessing:** Data was scaled using `StandardScaler` to ensure distance metrics were accurate.
- **Clusters (k=5):** The algorithm segmented users into 5 distinct groups, ranging from "Light Users" to "Heavy Users."

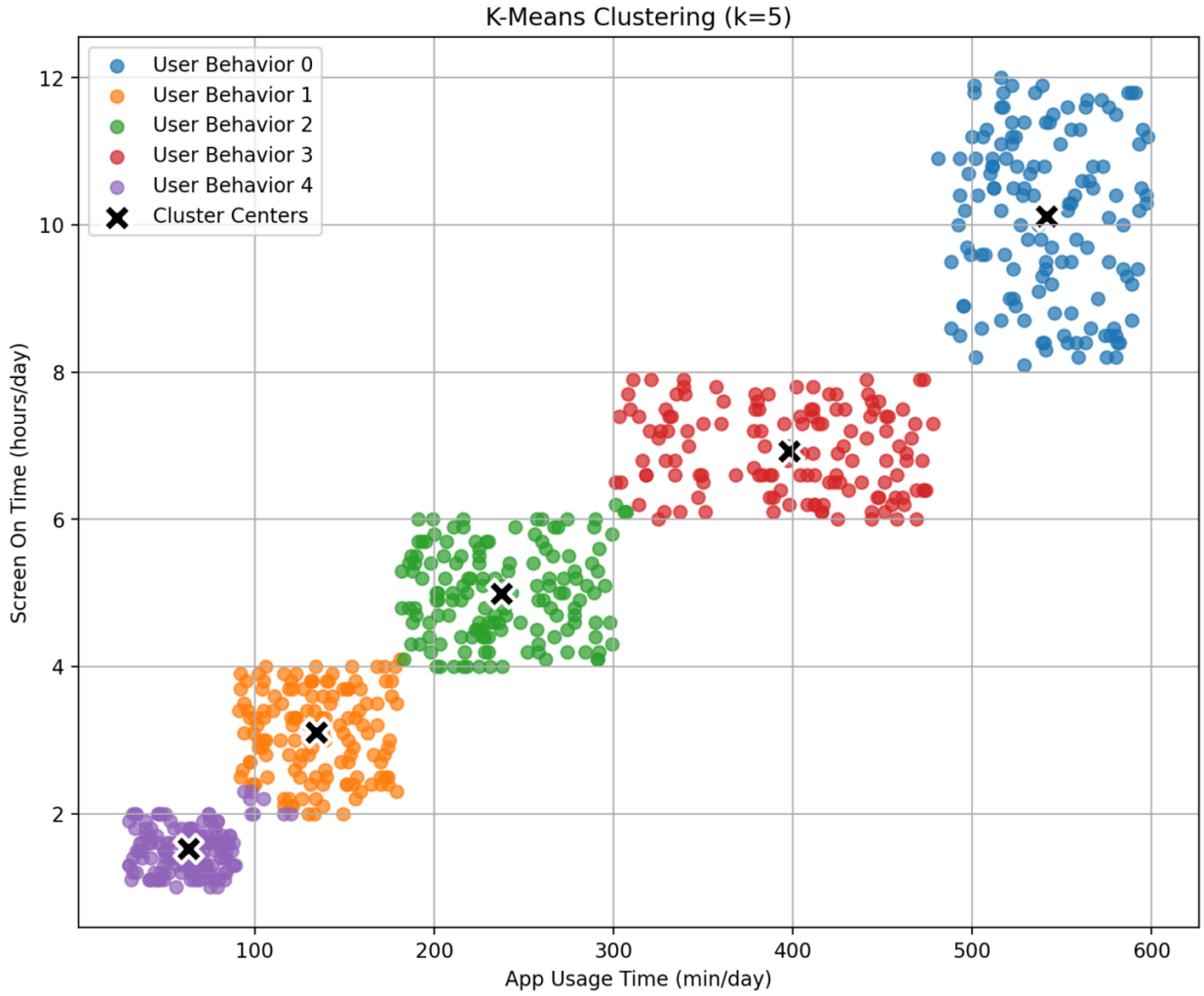


Figure 6: User segmentation clusters based on usage intensity.

6. Dashboard Overview

An interactive dashboard was developed using **Streamlit** to allow users to explore the data dynamically.

Key Features:

1. **Sidebar Filters:** Users can filter data by Gender, Operating System, and Device Model.
2. **KPI Display:** Real-time calculation of average Screen Time and Battery Drain based on selected filters.
3. **Interactive Plots:**
    - Distribution histograms that update based on user selection.
    - Scatter plots to view the relationship between Data Usage and Age.

7. Insights & Conclusion

Summary of Findings

1. **Usage Correlation:** There is a strong positive correlation between app usage time and total screen-on time, indicating that increased app usage directly leads to longer screen engagement.
2. **Demographic Insights:** Usage patterns are largely similar across genders, with only minor differences observed in specific app categories.
3. **Weak Correlations:** Age showed a weak correlation with the number of installed apps, suggesting that app installation habits are not strongly age-dependent in this dataset.
4. **Battery Consumption:** Battery drain is more strongly influenced by screen-on time rather than the number of installed apps. Android devices showed greater variability in battery consumption than iOS devices.
5. **User Segmentation:** K-Means clustering identified five distinct user groups based on app usage and screen time, ranging from "Light Users" to "Heavy Users." This segmentation can inform targeted marketing or digital wellbeing strategies.

Data Integrity:

The dataset was cleaned and preprocessed to remove missing values, duplicates, and outliers, ensuring reliable analysis results.

Future Recommendations

- **Granular Data:** Future analysis would benefit from knowing *which* specific apps (e.g., Social Media vs. Productivity) are being used.
- **Time Series:** Analyzing usage over weeks or months could reveal behavioral trends (e.g., weekend vs. weekday usage).

8. References

1. **Dataset:** Valakhorasani. (2024). *Mobile Device Usage and User Behavior Dataset*. Retrieved from KaggleHub. https://www.kaggle.com/datasets/valakhorasani/mobile-device-usage-and-user-behavior-dataset
2. **Tools Used:** Python (Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, Streamlit).