Mobile Usage Behavior Clustering Based on Demographic and Technological Factors

Tomer Shulman (208997551)
Shaked Shabat (315112227)
Guy Dulberg (206562977)
Oren Raz (206390114)

May 20, 2025

Research Question

How do demographic attributes (such as age and gender) and technological factors (such as operating system) influence mobile usage patterns?

1. Methodology

This project aimed to identify distinct patterns in mobile phone usage by analyzing a structured dataset containing behavioral, demographic, and technological features. We developed a clean and reusable machine learning pipeline to group users into clusters based on their smartphone usage behavior.

Problem-Solving Pipeline

1. Data Preprocessing

- No missing values were detected in the dataset.
- All time-related features were converted to minutes.
- Numeric features were standardized using Z-score normalization via `StandardScaler`.
- Categorical features such as gender, operating system, and age group were one-hot encoded using `OneHotEncoder`.

2. Feature Engineering & Selection

Using domain knowledge, we engineered additional features:

- **Battery per Minute**: Battery drain normalized by screen-on time.
- **Data per Minute**: Data consumption per screen-on minute.
- **App Usage Ratio**: Ratio of app usage time to screen-on time.
- **Apps per Hour**: Installed apps per hour of screen time.

Only interpretable and relevant features were retained for modeling.

3. Model Building

We implemented a `KMeans` clustering model inside a `Pipeline` along with the preprocessing steps. The model was set to 4 clusters (`k=4`), which we justified based on evaluation.

4. Parameters Tuning

To tune `k`, we applied two strategies:

- **Elbow Method**: Plotting inertia for `k=2` to `9` to identify the optimal inflection point (see Figure 1).
- **Silhouette Score**: Used to measure cluster quality. The best result was obtained at `k=4` with a silhouette score of 0.221 (see Figure 2).

Figure 1: Elbow Method showing optimal number of clusters (k=4)

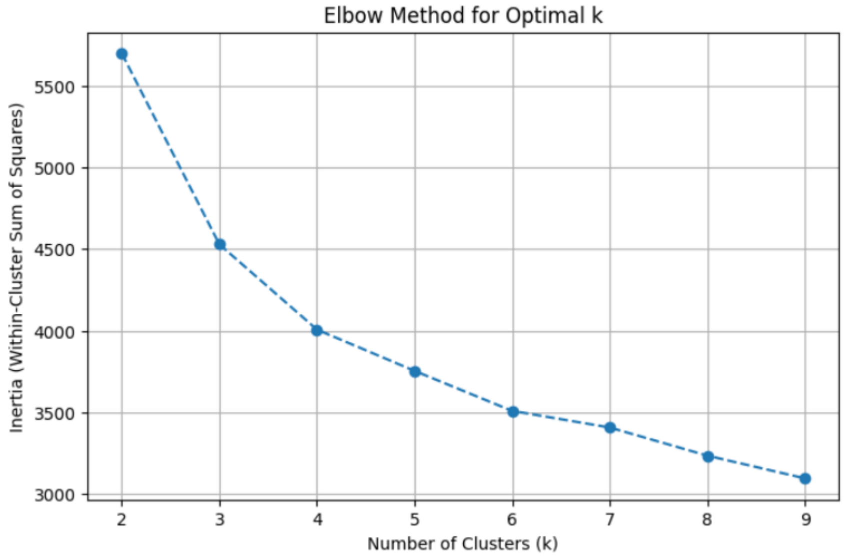
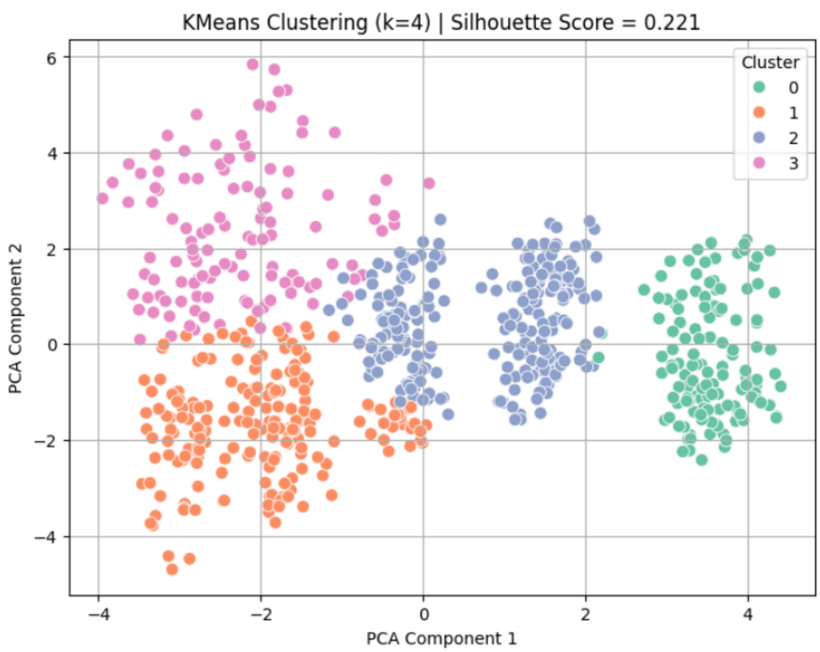


Figure 2: PCA visualization of KMeans clustering with k=4

Silhouette Score: 0.221



5. Pipeline Automation

We built a fully modular `Pipeline` using `scikit-learn` to automate preprocessing and clustering. This ensures reproducibility and scalability for other datasets.

Why This is the Best Solution

1. Our pipeline is modular, reproducible, and easily scalable.

2. It uses interpretable and engineered features for rich user profiling.

3. The use of silhouette score and PCA visualization helps validate and interpret cluster structures.

4. It outperforms static segmentation approaches by uncovering behavioral segments directly from data.

Software/System Implementation

- **Programming Language**: Python 3.11
- **Version Control**: Codebase managed via GitHub (Link)
- **Dependencies**: Listed in `requirements.txt` for reproducible setup
- **Reproducibility**: Random seeds were fixed in model initialization to ensure consistent results across runs
- **Modularity**: The code was structured into modular blocks within a single colab notebook, separating preprocessing, modeling, evaluation and visualization
- **Scalability & Efficiency**: Although the dataset was relatively small, the use of a pipeline structure ensures the solution can scale to larger datasets without changes

2. Evaluation

We evaluated our unsupervised model using both internal clustering quality metrics and system-level considerations.

Clustering Metrics

To assess the quality of the KMeans clustering, we used two key metrics:

- **Silhouette Score:** We achieved a silhouette score of 0.221 for $k = 4$, indicating moderate cluster cohesion and separation.

- **Elbow Method:** The plot of inertia from $k = 2$ to $k = 9$ showed a distinct elbow at $k = 4$, confirming our cluster choice.

We also visualized the clusters using PCA, which showed reasonably well-separated groupings in a 2D space.

Cluster Insights

To better understand each cluster, we computed the average of key behavioral and demographic features.

- **Cluster 0** (138 users): Heavy users with very high app usage (540 min/day), longest screen-on time (604 min/day) and highest app usage ratio (0.91), indicating most of the screen time is active use. These users also had the highest data consumption and battery drain.

- **Cluster 1** (198 users): Older users (average age 40.2) with low app usage (114 min/day), fewer apps (27), and lower battery/data consumption. They also had the lowest app usage ratio (0.63), suggesting passive phone use.

- **Cluster 2** (244 users): Balanced users. Moderate app use (327 min/day), high number of apps (61) and strong battery performance. App usage ratio (0.91) matches Cluster 0, but with less intensity overall.

- **Cluster 3** (120 users): Youngest group (age 35.6) with short screen time but extremely high battery per hour and apps per hour. This could indicate users with power-intensive apps or background processes.

Demographic distributions showed that Cluster 2 had the highest proportion of female users (52%), while Clusters 0 and 3 skewed slightly male. Cluster 3 also had the youngest average age (35.6), while Cluster 1 had the oldest. For visualizations of the insights, see Figures 3, 4, 5, and 6.

Visual Summary of Clustering Results

Figure 3: Heatmap of average feature values for each cluster

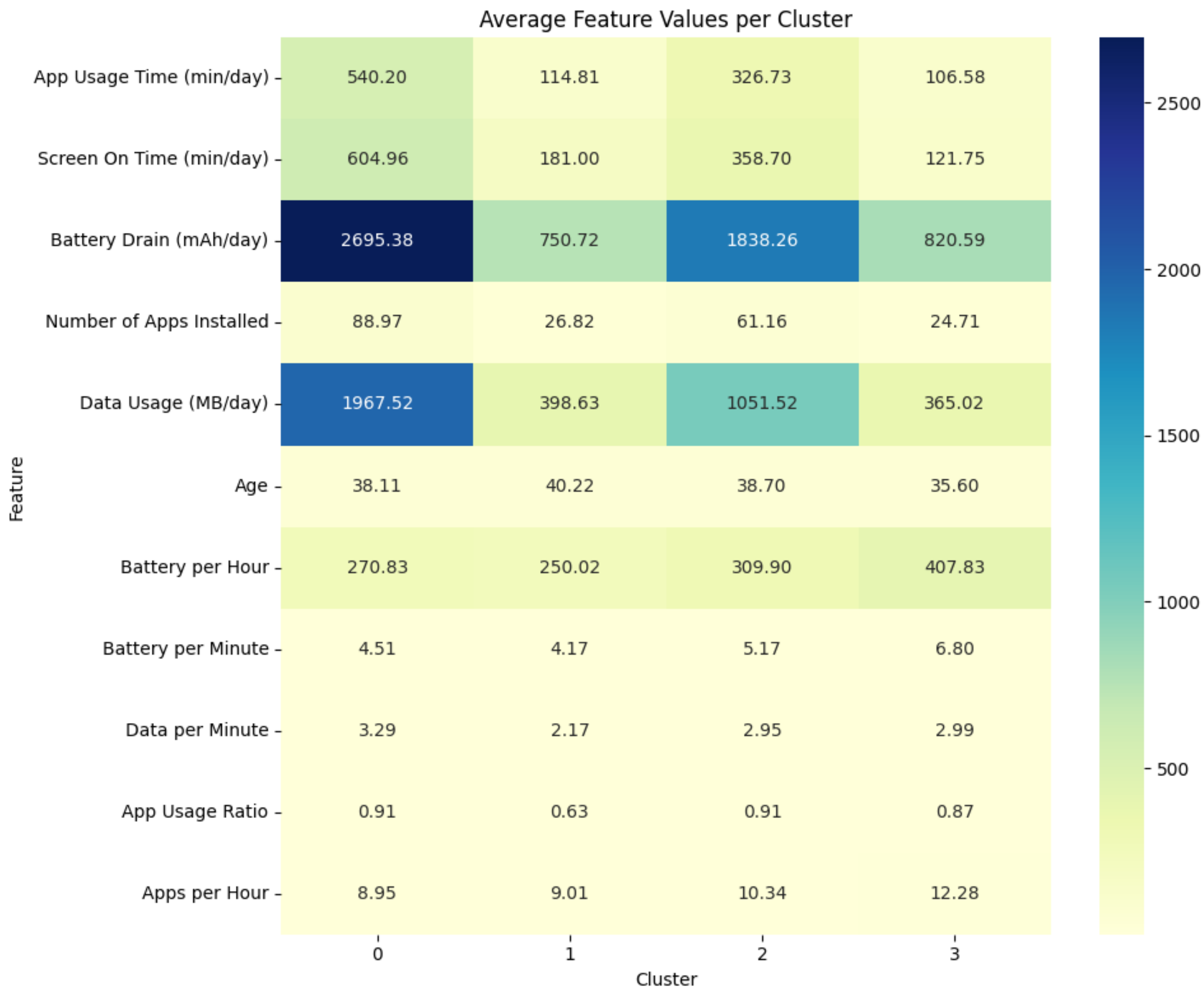


Figure 4: Number of users in each cluster

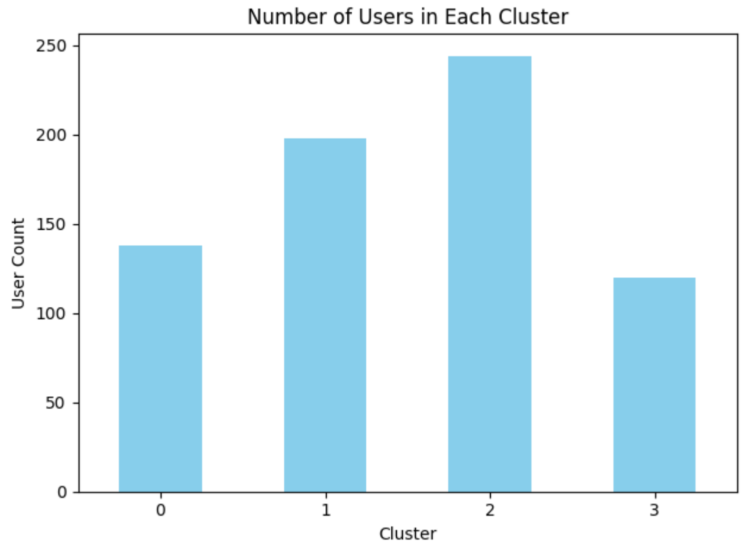


Figure 5: Proportion of female users and iOS users per cluster

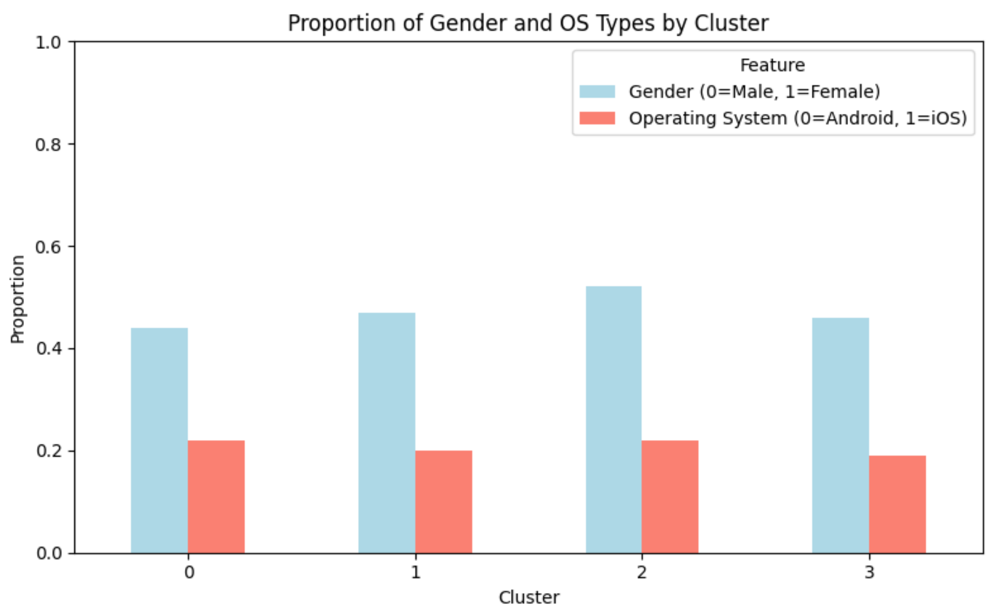


Figure 6: Average age of users per cluster