1. Title

Social Media User Behavior Clustering
(Using survey data — HybridDataset.csv)

---

2. Introduction

Unsupervised learning helps uncover hidden patterns in data without predefined labels. In this project, we use **K-Means clustering** to segment social media users based on their digital behavior.

Objective:
Identify distinct user groups from survey data and profile them (e.g., *Heavy Users*, *Casual Browsers*, *Premium Subscribers*, *Aware but Distracted*).

---

3. Dataset Description

- **File:** HybridDataset.csv
- **Rows:** 310 (survey responses)
- **Columns:** 19 (age group, occupation, hours online, device, number of platforms, platform lists, primary use, activities, subscription behavior, distraction details, awareness category, etc.)
- **Types of variables:** mixture of single-choice, yes/no, numeric-like text, and multi-select answers.

---

4. Preprocessing (implemented in code)

Steps performed in the script:

- **Column cleaning:** removed unnamed index columns and trimmed spaces in column names.
- **Binary encoding:** converted Yes/No responses into 0/1 (*_bin).
- **Numeric extraction:** extracted numbers from text (e.g., “5 hours” → 5 in *_num).
- **Multi-select handling:** for answers like *“Which platforms do you use?”*:
    - Created _count column = number of options selected.

    - Created binary flags for the top 8 most common platforms/activities (*_has_facebook, etc.).

- **One-hot encoding:** for small categorical fields, created indicator columns for the top categories.
- **Feature selection:** kept numeric + engineered features (*_num, *_bin, *_count, *_has_, *_is_).
- **Missing values:** imputed with median.
- **Scaling:** standardized all features using StandardScaler.

---

5. Exploratory Data Analysis (EDA)

- Checked value counts for age group, occupation, hours online, and number of platforms.
- **Distributions:** Histograms showed skewed distributions (e.g., most users spend 2–6 hours online).
- **Platforms:** Facebook, Instagram, WhatsApp, and YouTube were among the most frequent.
- **Outliers:** Some extreme “hours per day online” values were found but mitigated with median scaling.

---

6. Dimensionality Reduction (PCA)

- Fitted **PCA** for visualization.
- First **2 components** captured significant variance.
- Plotted **cumulative explained variance curve** to justify dimensionality reduction.
- Used PCA1 & PCA2 for cluster scatter plots.

---

7. Clustering & Model Tuning

- Applied **K-Means** with k = 2…8.
- Evaluated:
    - **Inertia (Elbow method)** → to check the “bend” in the curve.

  - **Silhouette score** → to evaluate separation quality.
- **Rule:** chose k with the highest silhouette score.
- Final model refit with best_k and labels assigned.

---

8. Results

- **Chosen k:** (from silhouette analysis — insert the number you got when you ran the code).
- **Cluster sizes:** printed in script (e.g., Cluster 0: 120, Cluster 1: 100, Cluster 2: 90).
- **Cluster profiles:** from cluster_means.csv, examples could be:
  - **Cluster 0 – Heavy Users:** higher hours online, more platforms, frequent distraction, high posting activity.
  - **Cluster 1 – Casual Browsers:** fewer hours, fewer platforms, lower distraction.
  - **Cluster 2 – Premium Subscribers:** medium hours but high subscription likelihood, awareness of monetization.
  - **Cluster 3 – Aware but Distracted:** moderate usage, low adoption of distraction management tools.
- **Top features per cluster:** listed in script output (those most different from the global mean).
- **Visuals produced by code:**
  - Elbow plot
  - Silhouette score plot
  - PCA 2D scatter with clusters & centroids

---

9. Insights & Recommendations

- **Marketing:** promote premium services to clusters that showed subscription behavior.
- **Digital wellbeing:** target heavy/distraction-prone users with focus tools and reminders.
- **Personalization:** recommend relevant content for platform-heavy clusters.

10. Conclusion

This project successfully grouped survey respondents into meaningful social-media usage clusters. These insights can support:

- **Marketing campaigns** (cluster-specific targeting).
- **Product development** (features for different user types).
- **Digital wellbeing programs.**

11. Appendix — Deliverables

- **kmeans_tuning.csv** → table with k, inertia, silhouette.
- **cluster_profile_means.csv** → average values per cluster.
- **hybrid_dataset_with_clusters.csv** → dataset with PCA coordinates + cluster labels.
- **Plots:** elbow, silhouette, PCA scatter.