User Profiling and Segmentation

1. PROJECT OBJECTIVE

The primary objective of this project is to build a robust and scalable **User Profiling and Segmentation System** that categorizes users into meaningful clusters based on demographic, behavioral, and interest-based attributes. This enables businesses to:

- Improve personalized marketing
- Optimize advertising campaigns
- Enhance user engagement
- Increase conversion rates while minimizing ad spend

---

2. TOOLS AND TECHNOLOGIES USED

To develop this project, we used the following Python libraries and machine learning tools:

- **Python**: Core programming language
- **Pandas & NumPy**: Data manipulation and numerical operations
- **Matplotlib & Seaborn**: Visualization
- **Scikit-learn**: Machine learning algorithms (KMeans, Logistic Regression, SVM)
- **Joblib**: Model persistence
- **ydata-profiling (optional)**: HTML-based EDA report

---

3. DATA COLLECTION AND DATASET DESCRIPTION

The dataset comprises **1,000 user profiles** with **16 features** collected from an online platform. These include:

- **Demographics**: Age, Gender, Income Level, Education
- **Behavioral**: Likes, Reactions, Followed Accounts, Device Usage
- **Engagement**: Time spent online (weekdays/weekends), Click-through rates, Conversion rates
- **Ad Interaction**: Time spent engaging with ads
- **User Interests**: A multi-label categorical field indicating top interest categories

Source: Dataset Link

User Profiling and Segmentation

4. FEATURE ENGINEERING

To extract meaningful patterns from the data:

- Categorical variables were encoded using **One-Hot Encoding**
- The "Top Interests" column was transformed using **MultiLabelBinarizer**
- All numeric features were **scaled** using **StandardScaler**

5. SEGMENTATION TECHNIQUE

We applied **KMeans Clustering** with n_clusters=4 based on domain experimentation. This unsupervised method groups users by similarity in their scaled feature space.

- After clustering, a Segment label was assigned to each user
- **Principal Component Analysis (PCA)** was used for 2D visualization of clusters

6. USER PROFILE GENERATION

For each cluster, a segment profile was generated using:

- **Segment-wise means** of all features
- **Boxplots** to visualize feature distribution across segments
- **Heatmaps** for feature correlation

This provided clear insights into the key behavioral and demographic traits per cluster.

7. PREDICTIVE MODELING

To simulate real-time user classification into segments:

- **Logistic Regression** and **Support Vector Machine (SVM)** models were trained using the segmented dataset
- The models achieved strong classification performance (accuracy metrics provided in the output logs)

These classifiers enable fast, consistent assignment of new users to segments.

User Profiling and Segmentation

8. OUTPUT AND VISUALIZATION

Enhanced visualizations were generated and saved in the output/ folder:

- **PCA Cluster Scatter Plot**
- **Segment Distribution Bar Plot**
- **Box Plots** for Top Varying Features
- **Correlation Heatmap**
- **Segment Summary CSV**

9. Alignment with Problem Statement

We addressed all major components of the original task:

1. **Objective Definition**: Optimizing ads, improving personalization
2. **Data Source**: Cleaned and structured online user dataset
3. **Feature Creation**: Categorical encodings, binarization, scaling
4. **Segmentation Technique**: KMeans Clustering, PCA
5. **User Profiles**: Segment summaries, visualization, classification
6. **End Goal**: Enabling businesses to align marketing efforts to specific user clusters for better ROI

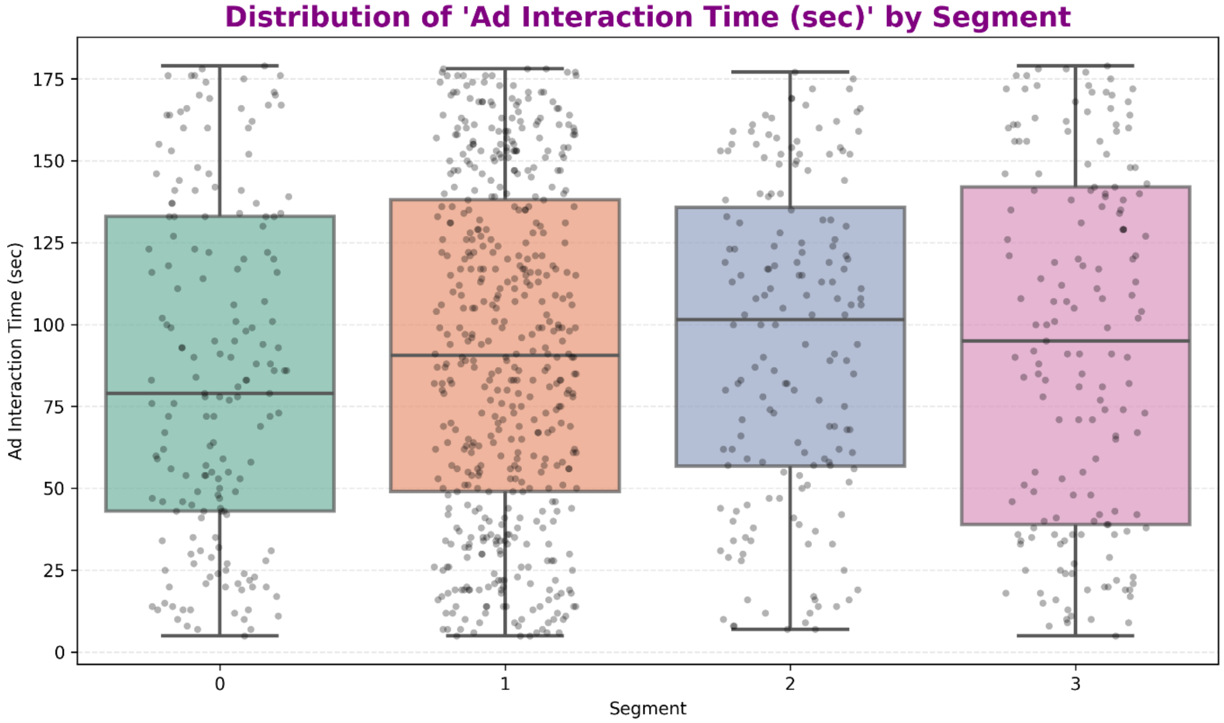
10. Conclusion

This project provides a full-stack solution for data-driven user segmentation, with modular components for profiling, clustering, visualization, and prediction. It is scalable and can be deployed in customer intelligence systems or recommender engines.

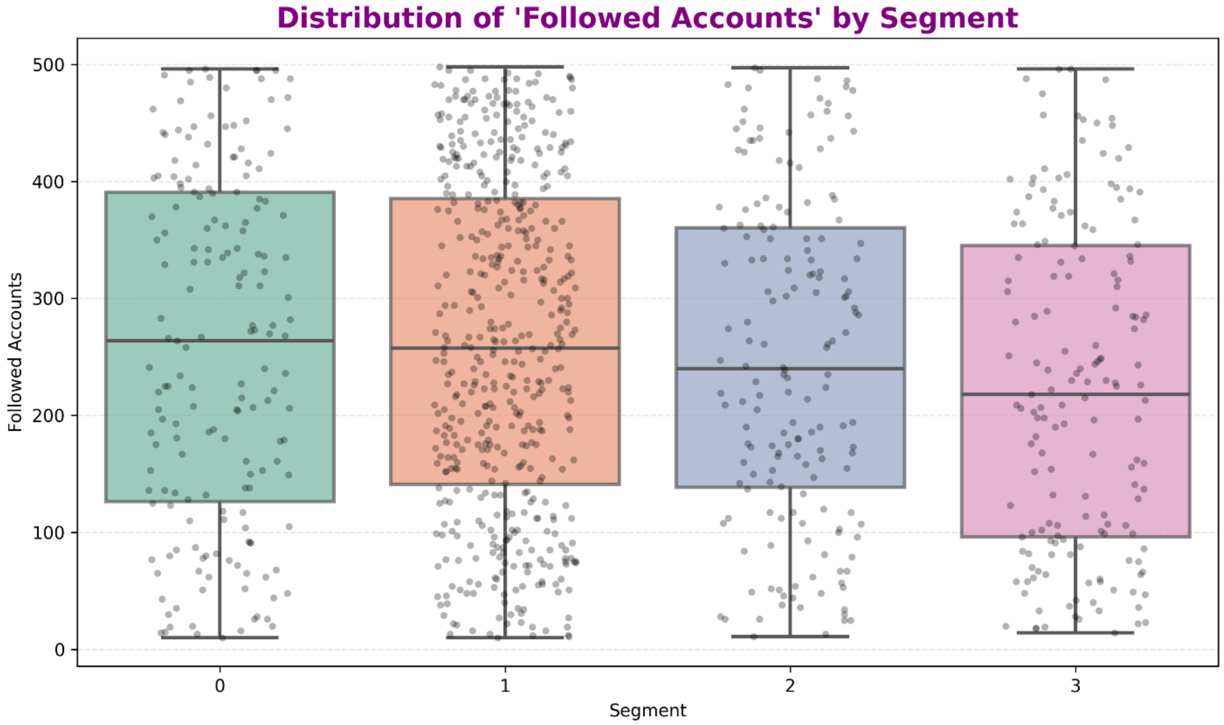
All models and reports are stored under the output/ directory for review and deployment.

11. VISUAL ANALYSIS OF SEGMENTS

1. Distribution of 'Ad Interaction Time (sec)' by Segment This boxplot shows how long users in different segments interact with ads. We observe that segments 1 and 3 have higher median ad interaction times, indicating strong engagement with advertisements in these groups.

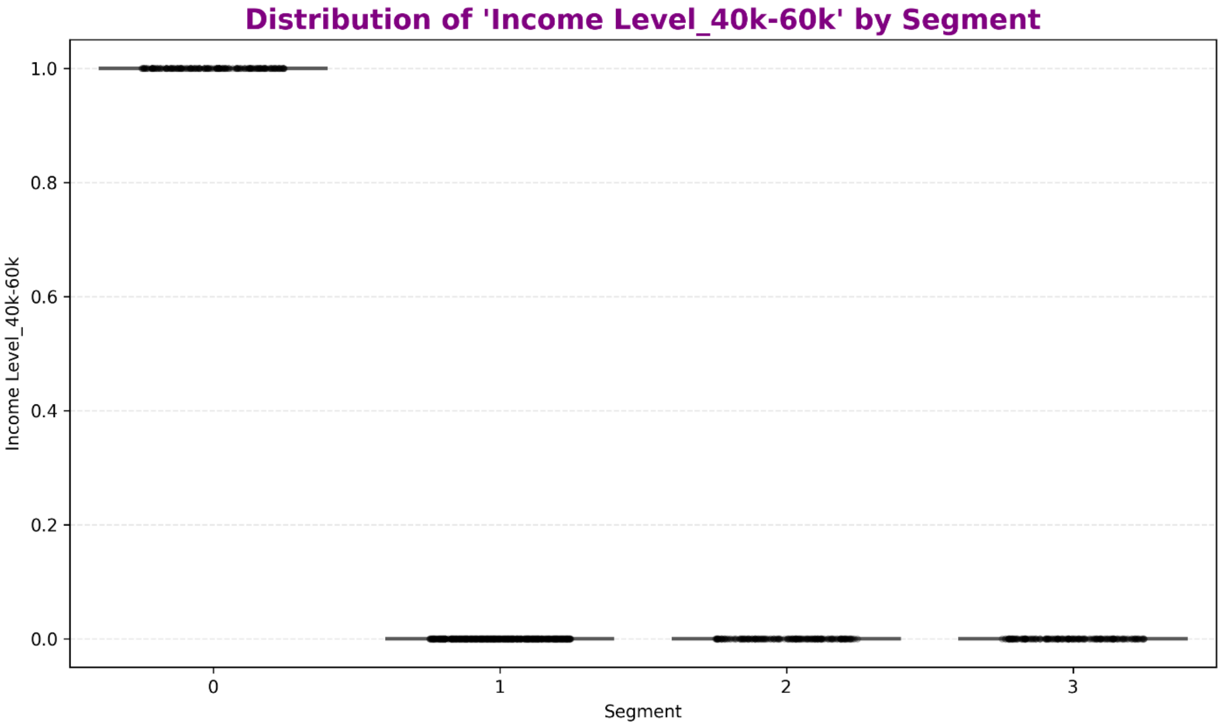


2. Distribution of 'Followed Accounts' by Segment This plot reveals the spread of the number of accounts followed by users in each segment. Segment 0 and 1 users tend to follow more accounts, suggesting higher social engagement and broader content exposure.

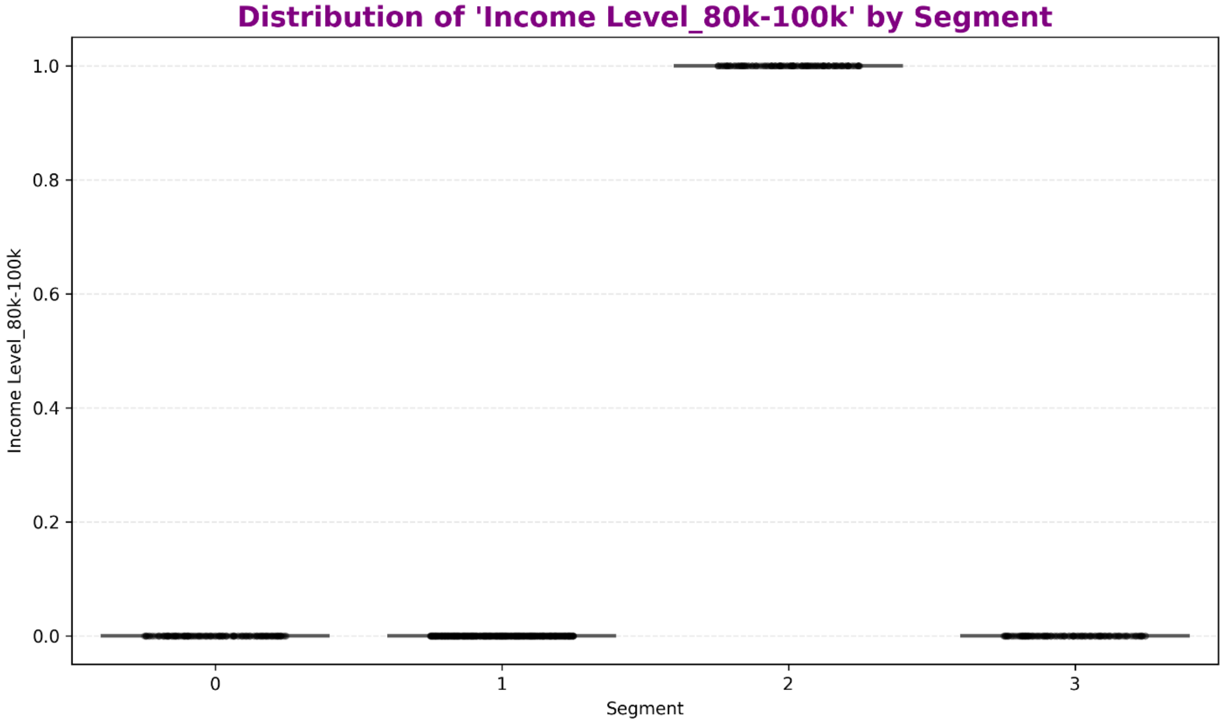


User Profiling and Segmentation

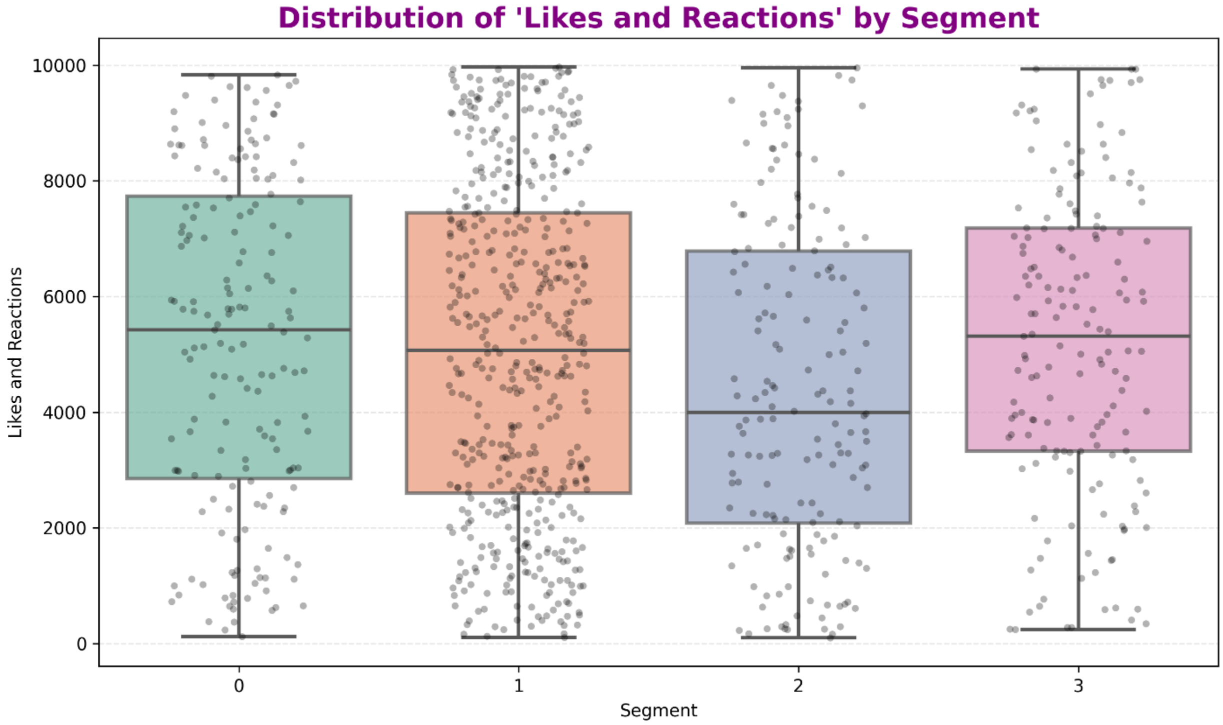
3. Distribution of 'Income Level 40k-60k' by Segment This binary plot shows that users in Segment 0 predominantly fall in the 40k-60k income bracket, helping marketers target this middle-income group effectively.



4. Distribution of 'Income Level 80k-100k' by Segment This visualization highlights that Segment 2 includes users mainly in the 80k-100k income range. These users may be more responsive to premium products and services.

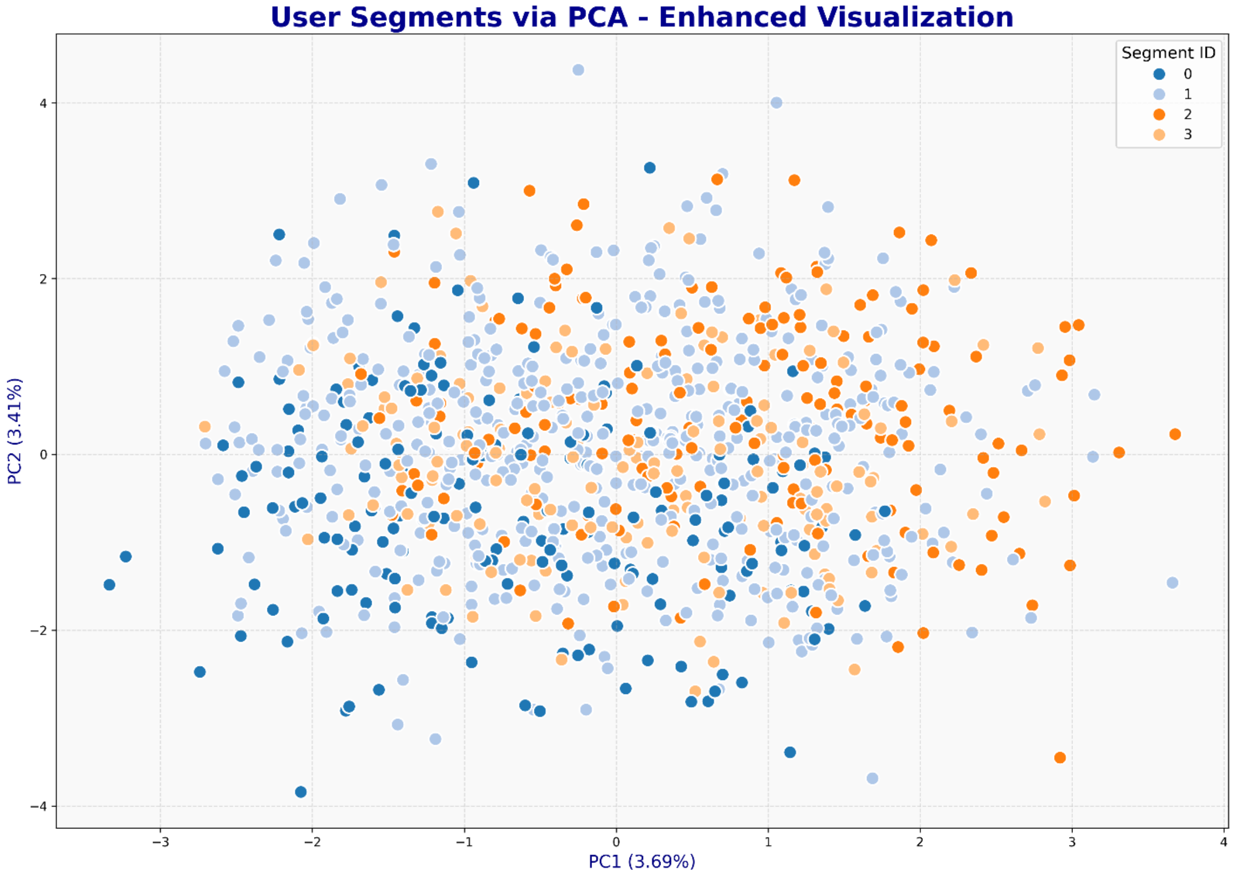


5. Distribution of 'Likes and Reactions' by Segment The likes and reactions metric reflects content engagement. Segment 0 shows the highest engagement level, making it ideal for viral content marketing.



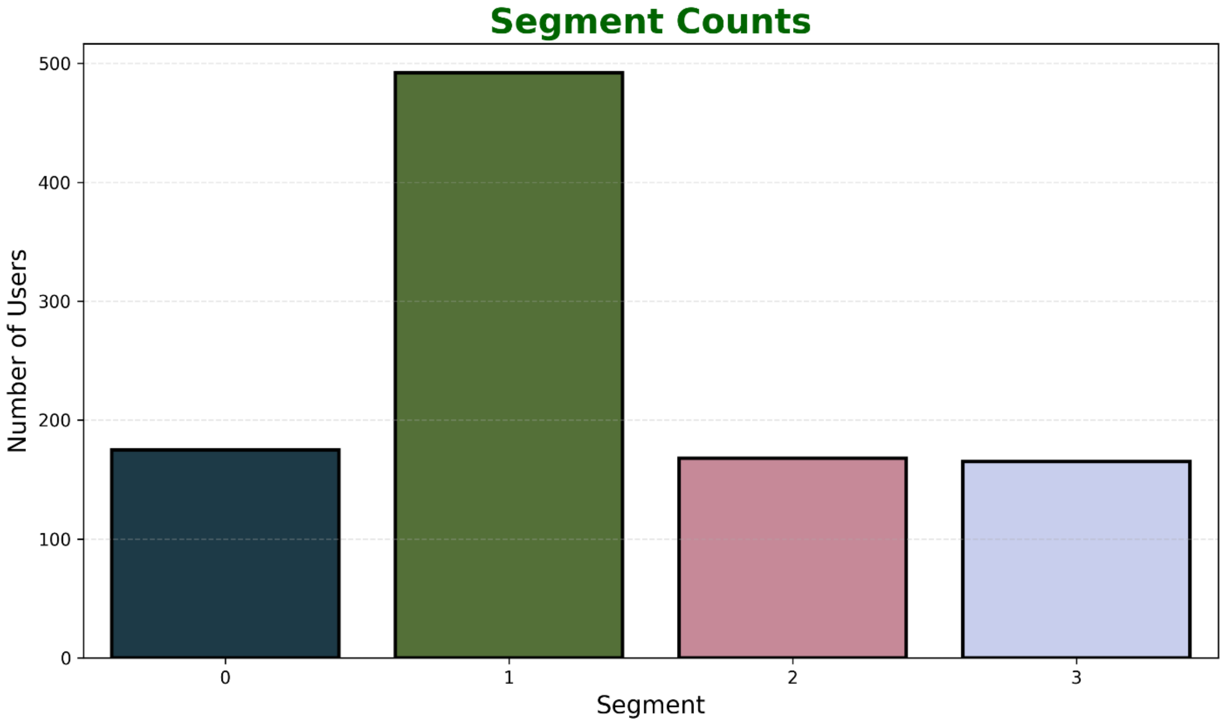
User Profiling and Segmentation

6. PCA-Based User Segment Visualization Using PCA, we projected high-dimensional data into 2D for visualization. This scatter plot clearly shows how the KMeans model formed distinct clusters of users based on underlying patterns.



User Profiling and Segmentation

7. Segment Count Bar Plot This bar chart shows the number of users in each segment. Segment 1 has the largest group, suggesting this segment's behavior is the most common in our dataset.



User Profiling and Segmentation

INTERNSHIP CONTEXT:

This project was undertaken as part of a **45-day internship** at **Bharathversity**, hosted on-site at **BITS Pilani, Hyderabad Campus, Secunderabad**. It was completed by **[Your Name]**, a student from **[Your University Name]**, under the guidance and mentorship provided during the internship program.

Prepared by: kuppili.sivamani