# Customer Segmentation Analysis

# Introduction

The goal of this project is to segment a digital magazine’s customers into meaningful groups based on their behavior and article consumption. We have two datasets: Behavioral Data (demographics, browsing behavior, ad clicks, subscription length, and engagement metrics) and Article Data (counts of articles read across topics such as Stocks, Productivity, Fashion, Celebrity, Technology, AI, etc.).

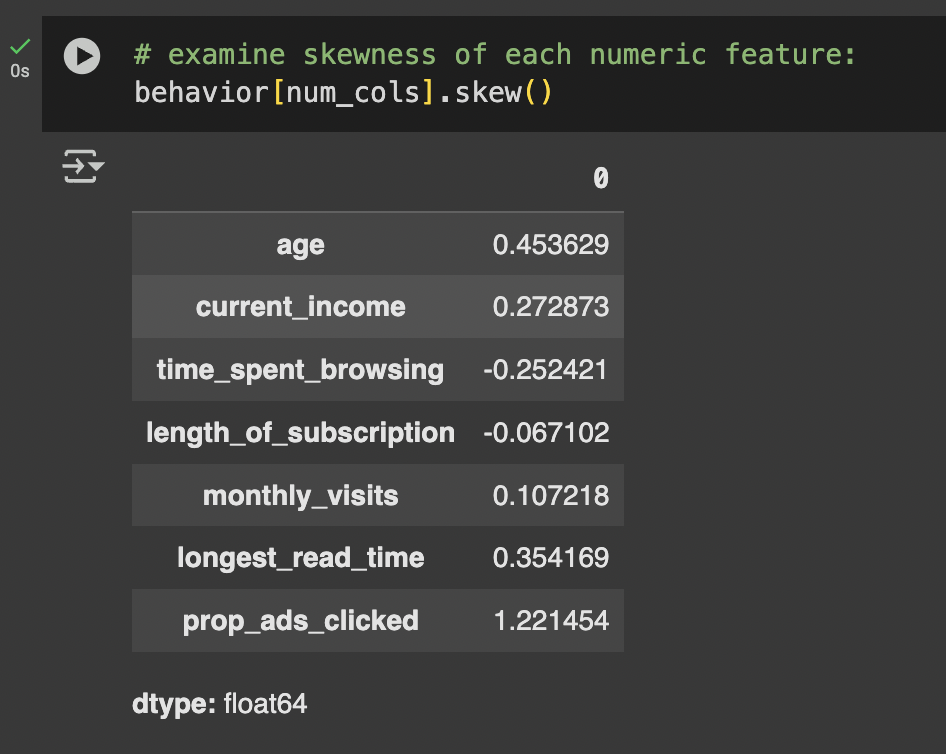
If successful, this segmentation can help the company tailor content, improve retention, and optimize ad targeting. Understanding these clusters will allow the company to personalize experiences, increase subscription revenue, and maximize ad engagement.

# Methods

## Behavioral Clustering Model

### Pros and Cons

* **K-Means:** Works well for spherical clusters and balanced data. Sensitive to skewness, outliers, and variable scaling. Requires standardization and log transformation for skewed features.
* **Gaussian Mixture Models (GMM):** Assumes elliptical Gaussian clusters. Can handle soft clustering but is sensitive to skewness and outliers. Requires preprocessing.
* **DBSCAN:** Density-based, robust to cluster shape. Works best with large datasets; struggles with small samples like ours (200 customers).
* **Hierarchical Agglomerative Clustering (HAC):** Flexible, interpretable, no assumptions about cluster shape. Sensitive to scaling/outliers, but computationally feasible with small datasets.

**Based on these pros and cons,** we decided to choose a model by looking at the data’s spread or shape, numbers of outliers, and skewness. Upon some exploration of each features via histograms and boxplots, the variables of interest we see are ‘time\_spent\_browsing’, ‘longest\_read\_time’, and ‘prop\_ads\_clicked’ because these graphs showed that these columns had outliers and were left/right skewed.

**Figure 1:** ‘longest\_read\_time’ histogram

Exact skewness shows that time\_spent\_browsing, longest\_read\_time, and prop\_ads\_clicked are moderately skewed. We can log transform 'prop\_ads\_clicked' since it is right skewed and left skewed variables can be transformed by reflecting them first so that they are right skewed, then log transformed.

**Figure 2a: Figure 2b: Figure 2c:**

### 

### 

### 

From the boxplots above, we see that the features 'prop\_ads\_clicked', 'longest\_read\_time', and 'length\_of\_subscription' have outliers.

**Justification:** We chose **HAC** for its interpretability, flexibility with small datasets, and ability to visualize clusters via dendrograms.

### Chosen Model Details

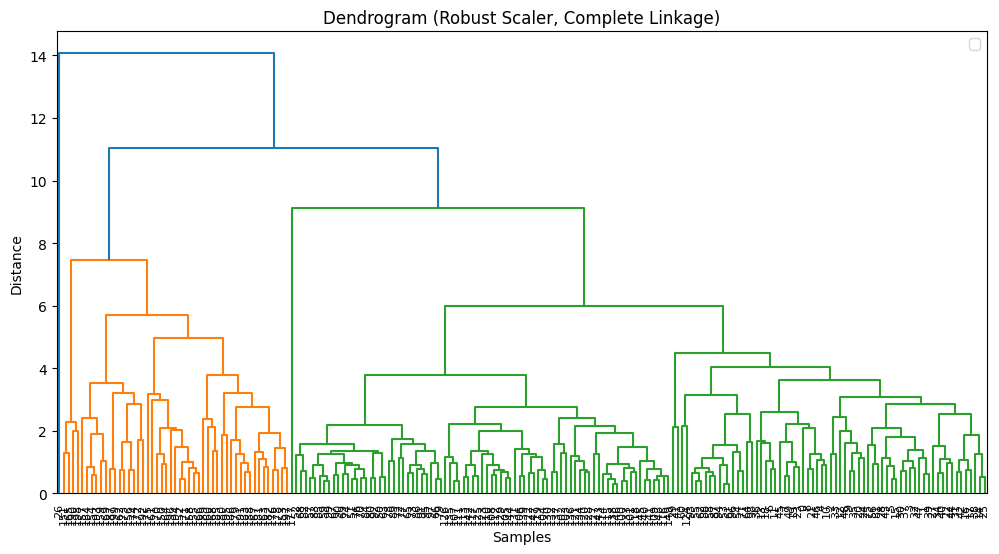
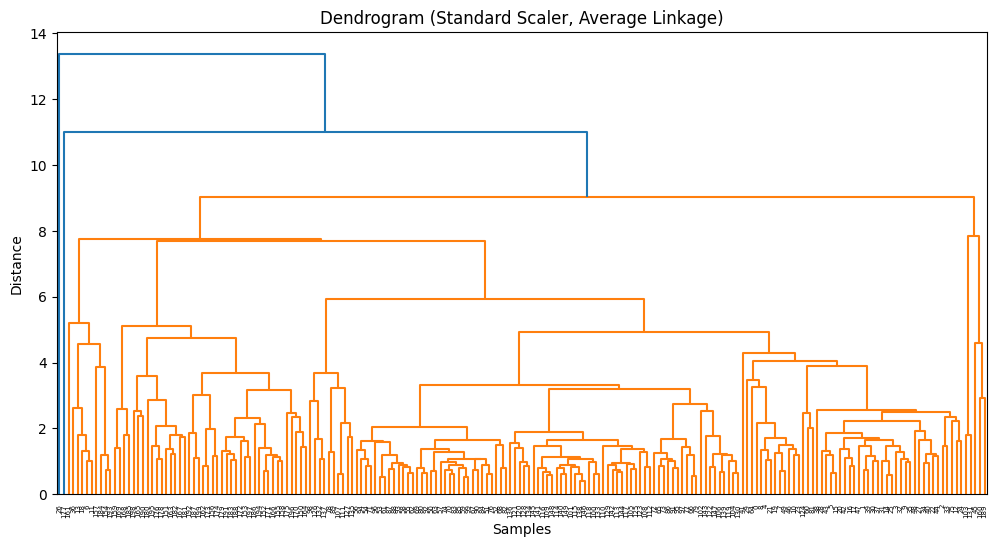
For behavioral clustering, we chose Hierarchical Agglomerative Clustering (HAC) due to its flexibility and interpretability for a small dataset. Preprocessing included dropping the id column, one-hot encoding gender, log-transforming skewed features (prop\_ads\_clicked, time\_spent\_browsing, longest\_read\_time), creating interaction features (e.g., income\_times\_ads, age\_times\_read), and standardizing all features using RobustScaler.

The key hyperparameters were found using a hyperparameter tuning/optimization technique called grid search, where linkage (complete) and number of clusters (2) were selected based on silhouette scores and dendrogram analysis to ensure well-separated and meaningful clusters.

After finding the most optimal hyperparameters, we scaled our data and fed it into the Hierarchical Agglomerative Clustering model with the hyperparameters found and experimented with two different scalers: Standard Scaler and Robust Scaler. We experimented with StandardScaler versus RobustScaler to handle the impact of outliers and skewed features in the behavioral data. StandardScaler standardizes features by subtracting the mean and dividing by the standard deviation. It works well for normally distributed data but is sensitive to extreme values, which can distort the clustering. RobustScaler scales features using the median and interquartile range, making it much less sensitive to outliers.

By comparing both, we found that RobustScaler produced better silhouette scores and more balanced clusters, because it reduced the influence of extreme browsing times, ad-click proportions, and other skewed variables. This ensured that the HAC model focused on the typical behavior of most customers rather than being dominated by outliers.

**Figure 3a: Figure 3b:**



## 

Figure 3a: Dendrogram shows poor distinguishing performance because clusters who are grouped together have high dissimilarity that is indicated by the "tallness" of the bars at the bottom. It only distinguishes two clusters (blue and orange) however the orange cluster contains almost all of our data supporting poor distinguishing performance.

Figure 3b: The dendrogram reveals two distinct clusters: a large mainstream segment (green) and a smaller, distinct segment (orange). The biggest jump in branch height indicates these groups are well-separated, consistent with the high silhouette score.

## Article Clustering Model

We used Hierarchical Agglomerative Clustering (HAC) with average linkage and cosine distance to group articles based on topic counts. No preprocessing such as scaling was applied because the features are counts, and cosine similarity is scale-invariant. The dendrogram was used to determine the number of clusters, and we observed five clear clusters based on the largest jumps in branch heights. These clusters represent distinct thematic groups of articles.

# Results

## Behavioral Clustering Model

The behavioral clustering model performed well, producing two clearly separated clusters with a high silhouette score of 0.712, indicating strong intra-cluster similarity and inter-cluster dissimilarity. Cluster 0 consists of younger, lower-income customers who are highly engaged, with long subscriptions, heavy browsing, and low ad-click activity—representing loyal readers driving subscription revenue. Cluster 1 includes older, higher-income customers with shorter subscriptions, less engagement, but higher ad-click rates, representing ad-responsive users who contribute to advertising revenue. The PCA scatterplot confirms the separation of these clusters, showing that the model effectively distinguishes between the two main behavioral segments. This information allows the company to tailor strategies: focusing on retention and premium content for Cluster 0, and targeted ads and promotions for Cluster 1.

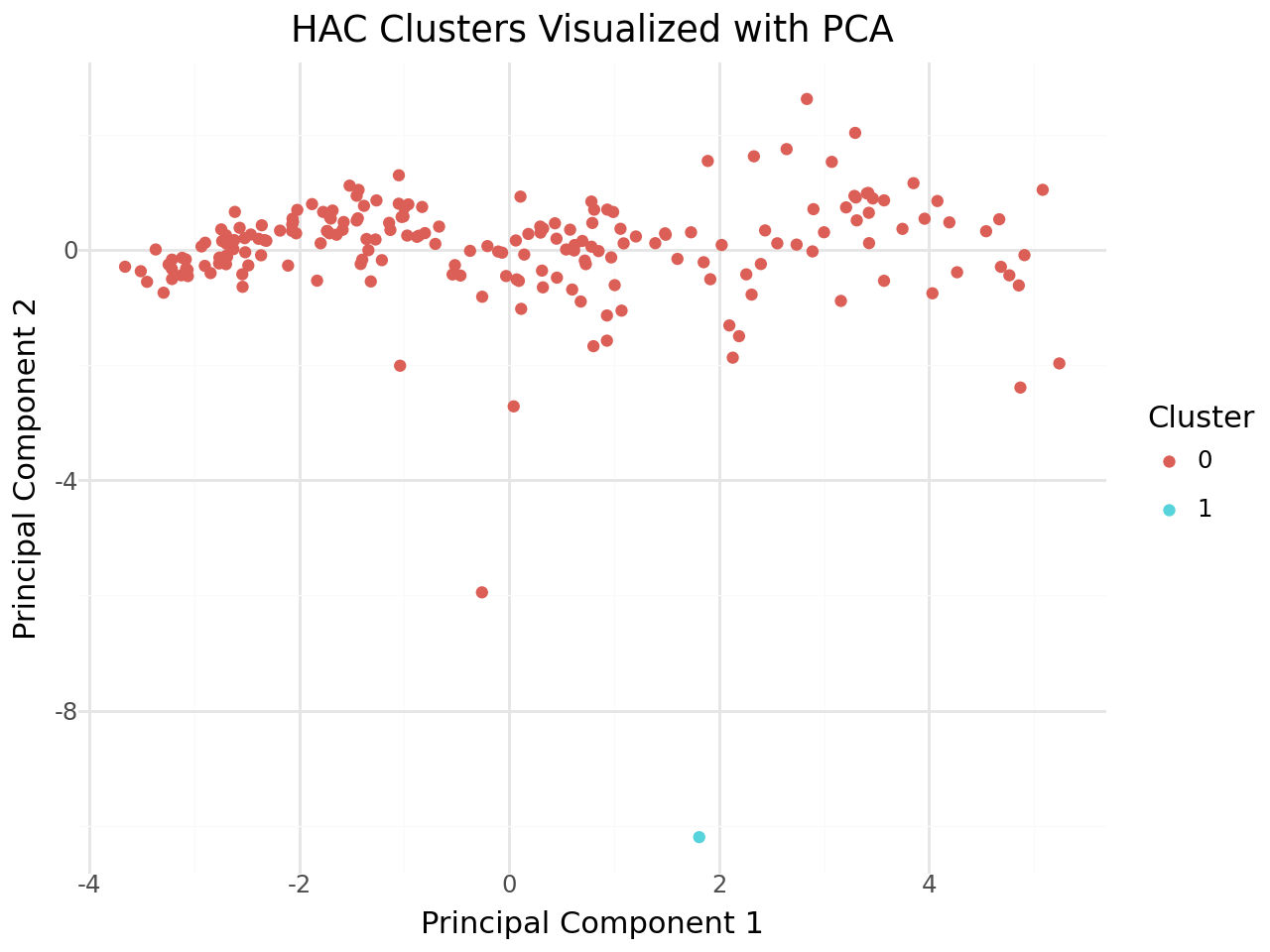


Figure 4: PCA Graph of Clusters

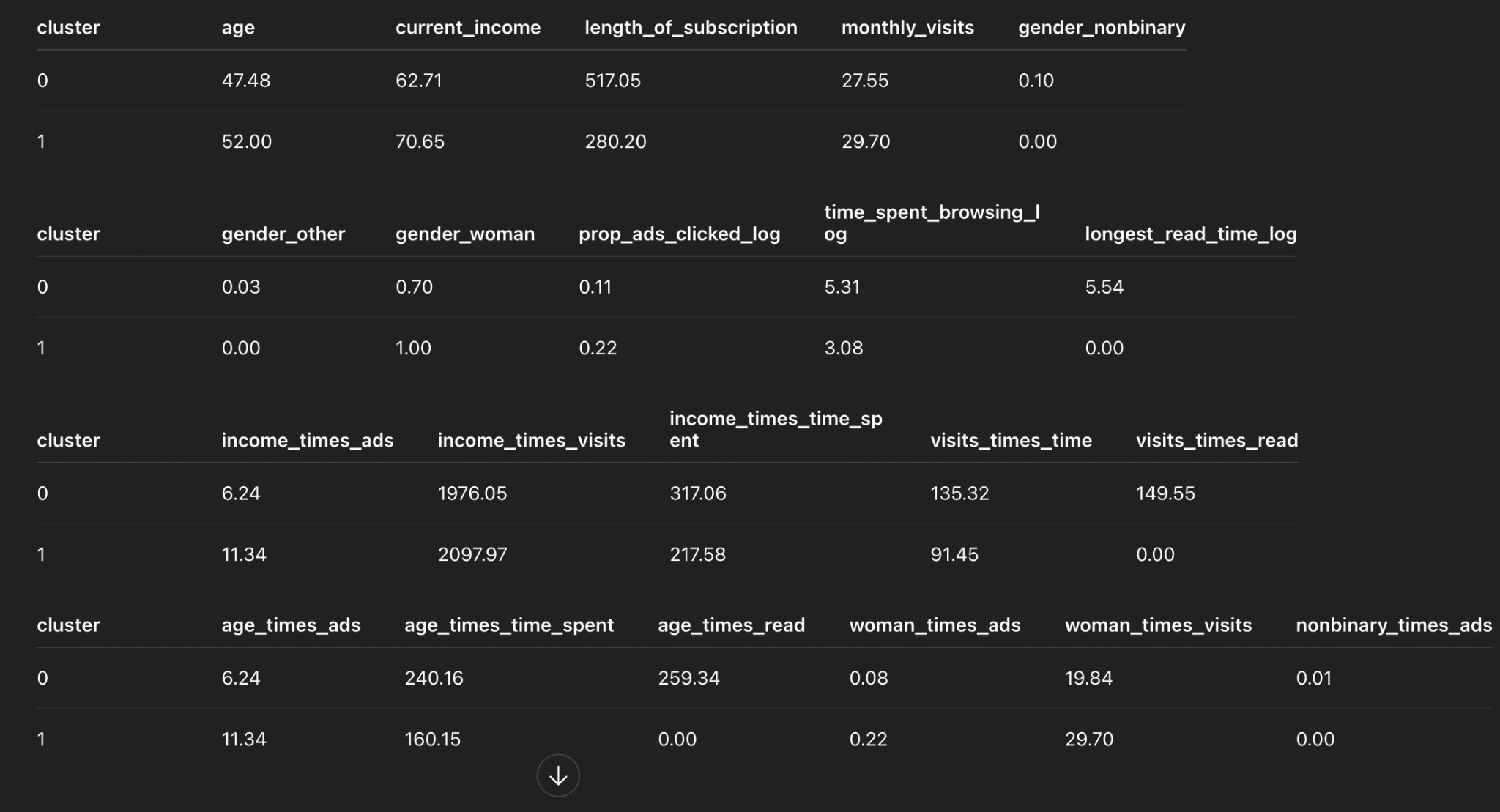


Figure 5: Cluster Averages Tables

Cluster Analysis Table:

| **Cluster** | **Characteristics** | **Business Insight / Strategy** |
| --- | --- | --- |
| 0 | Younger (47), lower income ($63k), long subscriptions (~517 days), heavy browsing/reading, low ad engagement | Loyal readers driving subscription revenue; focus on retention, premium content, community perks |
| 1 | Older (52), higher income ($71k), shorter subscriptions (~280 days), less engaged browsing, higher ad clicks | Ad-responsive users driving ad revenue; focus on targeted ads, cross-promotion, retention offers |

## Article Clustering Model

Upon using an sklearn Pipeline to build and fit a Hierarchical Clustering model (HAC) using all the variables except id. We used the parameters of cosine similarity as our distance metric (also called affinity) and average linkage. We created a dendrogram and used it to determine the number of clusters we used.

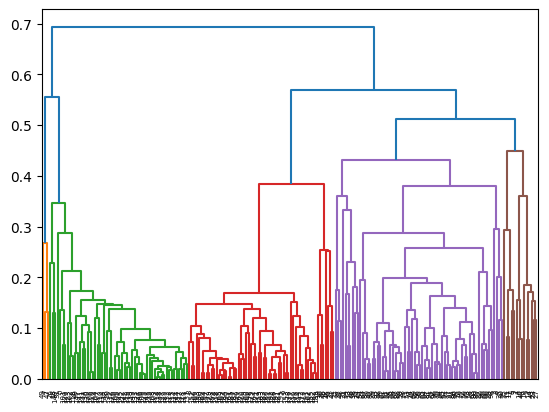


Figure 6: Article HCA Clustering Dendrogram

From the dendrogram above, we can see that the HAC model segments 5 clear clusters indicated by the colors.

The article cluster profiles were created by calculating the average count value of each feature for the articles in each cluster.



Figure 7: Summary table of average counts of articles by topic in each cluster

The model segmented the articles into five clusters, each with a clear thematic focus:

* Cluster 0 – Science & Fitness with some Lifestyle content: High counts in Science (8.14) and Fitness (8.14), notable coverage of Fashion (5.21) and Celebrity (4.57), moderate in SelfHelp (3.64), Cryptocurrency (2.71), Stocks (2.0), and Productivity (2.64), and low in AI (2.64) and Technology (2.86). This cluster combines scientific and wellness topics with some lifestyle content.
* Cluster 1 – SelfHelp, Productivity & Finance: High counts in SelfHelp (15.04), Productivity (11.19), Stocks (9.54), Fitness (6.43), and AI (7.79), low in Celebrity (0.53), Fashion (2.31), Technology (2.28), Science (3.87), and Cryptocurrency (2.5). Focused on career growth, personal development, and finance-related topics.
* Cluster 2 – Technology & AI: High counts in Technology (26.17), AI (26.10), Science (14.27), moderate in Cryptocurrency (4.03), Stocks (3.51), Productivity (3.78), and low in Fashion (3.22), Celebrity (2.75), SelfHelp (3.36), and Fitness (3.36). Strong tech and science emphasis, with minimal lifestyle content.
* Cluster 3 – Finance & Pop Culture: High counts in Cryptocurrency (12.67), Celebrity (5.33), Stocks (5.0), Productivity (5.0), moderate in Technology (3.33) and SelfHelp (1.66), low in Fashion (1.0), Science (1.0), Fitness (1.0), and AI (1.33). Appeals to users interested in digital finance and entertainment.
* Cluster 4 – Celebrity & Fashion: High counts in Celebrity (17.25) and Fashion (12.61), moderate in SelfHelp (2.89), Productivity (2.77), Technology (2.52), and low in Stocks (1.91), Fitness (1.61), Science (2.0), AI (1.93), and Cryptocurrency (1.52). Primarily lifestyle and entertainment-focused, with minimal technical or finance content.

**Business Insight**

These clusters allow the company to understand which types of articles appeal to different reader segments. For example, tech-savvy readers can be targeted with more Technology & AI content, while lifestyle readers may engage more with Celebrity & Fashion articles. This segmentation can inform content recommendations, newsletter targeting, and marketing campaigns.

# Discussion/Reflection

Through this project, I learned how clustering can reveal meaningful patterns in both customer behavior and article topics. Behavioral clustering highlighted distinct customer segments, loyal long-term readers versus ad-responsive short-term subscribers, which can guide personalized engagement strategies. Article clustering revealed thematic content groups, showing which topics naturally attract similar readership.

If I were to perform this analysis again, I would consider incorporating temporal patterns in customer behavior (for example, seasonal browsing trends) and engagement metrics like time spent per article or click-through rates. For article clustering, experimenting with topic modeling or TF-IDF weighting could provide deeper insight into content similarity beyond raw counts.