Homework 3

Tiffany Le

Introduction

In the rapidly evolving media (written and/or visual) industry, companies actively pursue strategies to boost revenue by offering products and services aligned with current media trends and their current and future consumers. This data analysis aims to assist a company in understanding its customer's behaviors and preferences through two distinct clusterings: behavioral and article clusterings.

"Behavioral clustering," in this analysis, involves grouping the company's customers into different clusters to identify patterns, similarities, and differences between the customers. To achieve this goal, a customer dataset provided by the media company on their customers provides key independent variables to build a clustering model. These key independent variables are the customers:

- **age** : age in years
- **current_income** : self-reported current annual income in thousands
- **time_spent_browsing** : average number of minutes spent browsing website per month
- **prop_ad_clicks** : proportion of website ads that they click on (between 0 and 1)
- **longest_read_time** : longest time spent consecutively on website in minutes
- **length_of_subscription** : number of days subscribed to the magazine
- **monthly_visits**: average number of visits to the site per month

"Article clustering," in this analysis, refers to identifying customers with similar reading patterns by analyzing the number of articles a customer has read in each specified topic over the past three (3) months and clustering or grouping them with others who share common interests in the given topics. To achieve this goal, an article dataset provided by the media company on the number of articles customers read in each topic provides key independent variables to build a hierarchical clustering model. These key independent variables are the following topics in customer data:

- **Stocks**
- **Productivity**
- **Fashion**

- **Celebrity**
- **Cryptocurrency**
- **Science**
- **Technology**
- **SelfHelp**
- **Fitness**
- **AI**

These clustering models can be incredibly valuable and helpful if successful. By accurately clustering and identifying the similarities and differences between their customer's preferences, the company can tailor their products and service to the customers' preferences. Furthermore, these clustering algorithms can lead to customers sharing the benefits of this company with their friends, family, and peers, leading to the company's growth, diversified consumer population, and higher customer retention.

Methods

The data analysis of clustering the company's customer and their article preferences were independently completed by two different clustering algorithms and datasets: the Behavioral Clustering Model (using K-Means Clustering on the Behavioral dataset) and the Article Clustering Model (using Hierarchical Clustering on the Article dataset).

Behavioral Clustering Model

Before fitting the independent, continuous variable data to the clustering model, it was prepared by dropping the "null" or empty values from the dataset and resetting its indices. Additionally, all of the variables except "gender" and "id" were added to a list containing all of the independent, continuous variables. Following this, scatterplots of pairs of the independent variables were created to give some information about the data and assist with deciding which clustering algorithm to choose. The scatterplots of the pairs are displayed below in Figures #1-4.

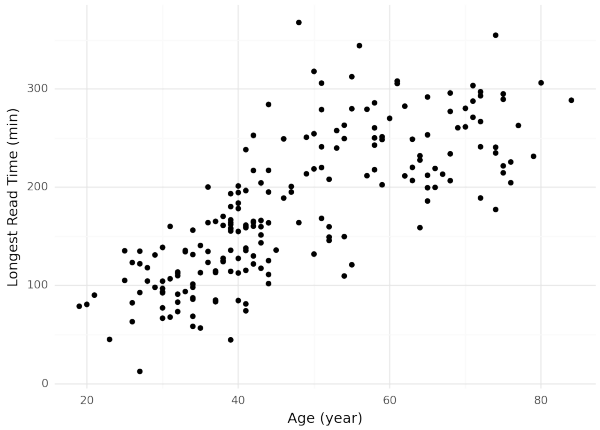

Figure 1: A scatterplot graph that shows the relationship between a customer age (years) and their longest read time (minutes)

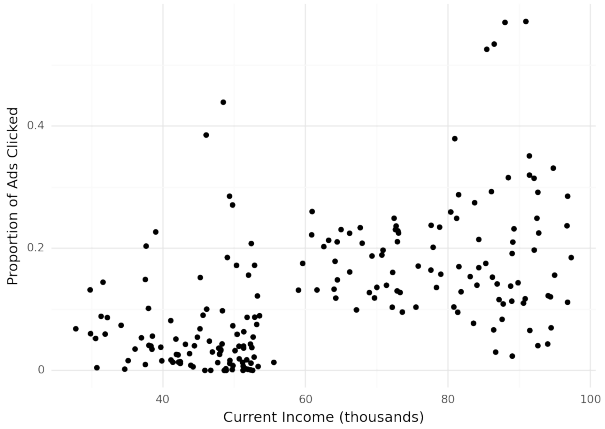

Figure 2: A scatterplot graph that shows the relationship between a customer income (thousands) and their proportion of ads clicked

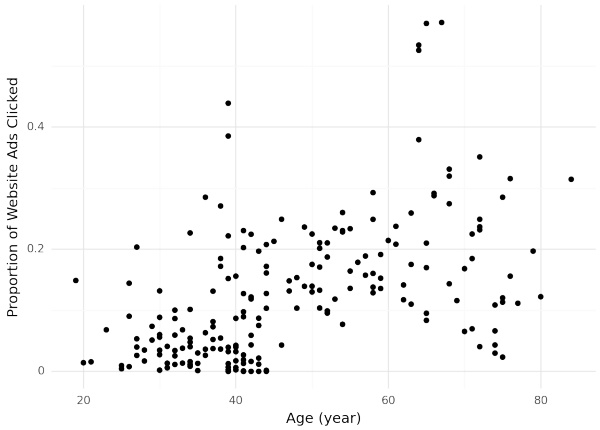

Figure 3: A scatterplot graph that shows the relationship between a customer age (years) and their proportion of ads clicked

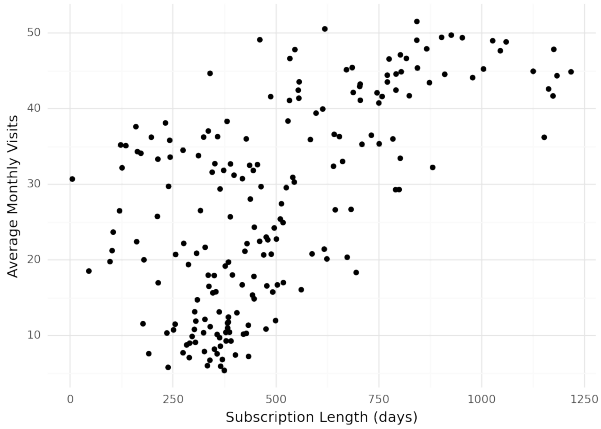

Figure 4: A scatterplot graph that shows the relationship between a customer subscription length (days) and their monthly visit count

Pros and Cons

In order to determine the algorithm that will be utilized for this dataset, an analysis on each algorithm was conducted to analyze their algorithm, benefits, drawbacks, and ideal data. The following algorithms were a part of the analysis: K-Means, Gaussian Mixture Model (GMM), Density-Based Spatial Clustering Applications with Noise (DBSCAN), and Hierarchical Agglomerative Clustering (HAC).

K-Means (KM)

K-means is a centroid-based clustering algorithm where data points are sectioned into k clusters based on their similarity to the centroid of a cluster. This algorithm chooses k random points to be the clusters' center. Following this, each data point is assigned to the cluster whose center is closest to it. After assigning the data points, the centers of the clusters are recalculated. Steps #2 and #3 will be repeated until convergence occurs (cluster membership doesn't change, and/or the centers only change by a small amount).

Ideal Data and Assumptions

This clustering algorithm works well with spherical, well-separated clusters. This algorithm assumes that clusters are spherical, equally sized, and the variance within each cluster is similar.

Pros and Cons

Compared to the other algorithms, K-means is fast, efficient, and simple to understand and implement. That said, K-means clustering requires the initial number of centroids to be specified. Additionally, K-means assumes that clusters are spherical and equally sized, so this clustering model is not as flexible as other models.

Gaussian Mixture Models (GMM)

GMM is a probabilistic model representing data as a mixture of multiple Gaussian (Normal) distributions. This clustering algorithm is similar to k-means, except it doesn't assume spherical variance within the clusters; somewhat, the clusters can be elliptically shaped. Additionally, this algorithm doesn't give a hard assignment for the data points similar to K-Means. Instead, this algorithm calculates the probability of data points in a cluster. This is done by estimating the clusters' means and the variance of each predictor. Additionally, the cluster means and variances are calculated using every data point weighted by the probability of a data point belonging to that cluster.

Ideal Data and Assumptions

This clustering algorithm is effective on data with different shapes and orientations. This algorithm assumes that data points are generated from a mixture of normal distributions and its underlying structure is probabilistic.

Pros and Cons

Compared to K-means, GMM is more flexible and can model complex cluster shapes. That being said, GMM is also sensitive to the number of components specified and can be more computationally expensive than K-Means.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

This clustering model works by grouping data points on density and considering regions with high density as clusters. The remaining data would be identified as noise and outliers.

Ideal Data and Assumptions

This clustering model is adequate for data with potentially arbitrary cluster shapes and density variability. This algorithm assumes that clusters have a higher point of density than the areas between them.

Pros and Cons

DBSCAN can find clusters of different shapes, which makes the model more flexible than GMM and K-Means. Additionally, this algorithm is immune to the effects of noise and outliers, and it can detect them. That being said, DBSCAN is sensitive to overlapping clusters. Similar to the following algorithms, it is sensitive to hyperparameters specified. Finally, DBSCAN may struggle with clusters of varying densities.

Hierarchical Agglomerative Clustering (HAC)

This clustering model builds a tree of clusters, either top-down or bottom-up, based on the pair distances between the data points.

Ideal Data and Assumptions

This clustering model is suitable for various cluster shapes and sizes. This model doesn't make any explicit assumptions like the previous algorithms.

Pros and Cons

Hierarchical clustering is about to provide a hierarchy of clusters and there is no requirement to specify the number of clusters beforehand. That being said, this algorithm can be computationally expensive and is sensitive to the choice of distance metric.

Chosen Model Details

The algorithm that was chosen for this data set is K-Means. While looking through the plotted data, the data showed signs of spherical variance clustering. While there were forms of overlap between the "clusters," there was a distinction between the clusters, so K-Means appeared to be the best option. Not only is K-Means a more efficient algorithm, the structure of the data was spherical-like. In regards to pre-processing, all of the independent variables were z-scored. Regarding K-Mean's hyperparameter `number of clusters`, two plots were created showing the increases and decreases of the clusters' Sum of Squared Errors and their silhouette scores by changing the number of clusters (k). Displayed below in Figure #5 and #6 are the "Within Cluster SSE for Different Ks" and "Silhouette Scores for Different Ks" plots.

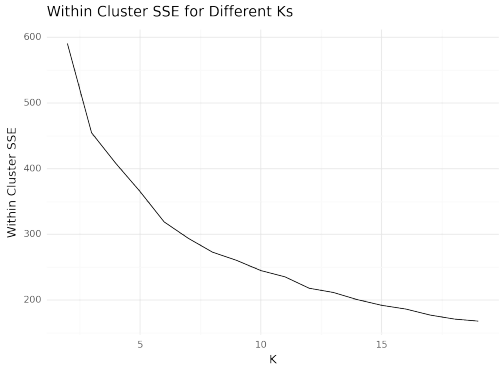

Figure 5: A graph that shows the Within Cluster Sum of Squared Errors (SSE) in relation to the number of clusters created by the K-Means

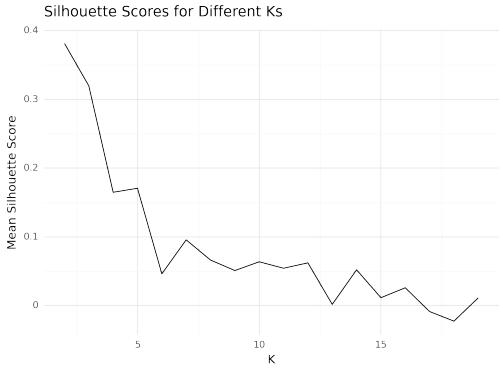

Figure 6: A graph that shows the Mean Silhouette Score in relation to the number of clusters created by the K-Means

Article Clustering Model

Data Preparation

Before fitting the independent, continuous variable data from the Article dataset to the clustering model, it was prepared by dropping the "null" or empty values from the dataset and resetting its indices. Additionally, all of the variables except "id" were added to a list ('X') containing all of the independent, continuous variables. Besides this step, the data was not put through pre-processing steps such as z-scoring or train test splits.

Hierarchical Clustering

After preparing the data, an empty Hierarchical Clustering (Agglomerative Clustering) model was created with the following hyperparameters set: "cosine" (affinity) distance metric, "average" linkage, 0 distance threshold, and 0 clusters. A distance metric in HAC determines how the distance between data points is measured. In the case of "cosine," we measure the degree of the angle between two documents, vectors, or clusters. A linkage criteria in HAC determines how the distance between two clusters is measured. In our case, an "average" linkage was chosen which defines the distance as the average distance between the data points of each cluster. The distance threshold and cluster parameters were set to 0 in order to create a dendrogram of the hierarchical clustering data, and another hierarchical clustering model will be created with the number of clusters set after looking at the dendrogram. Following this, an empty pipeline was fitted with the Hierarchical Clustering model, and the pipe was used to "fit predict" on the ('X') independent variable data.

Dendrogram

A dendrogram is a tree-like diagram that is useful for Hierarchical Clustering (HC) analysis, and it is a method utilized to build and visualize the hierarchy of clusters. Displayed below IN FIGURE #7 is the dendrogram of the Article dataset using the above Hierarchical Clustering parameters.

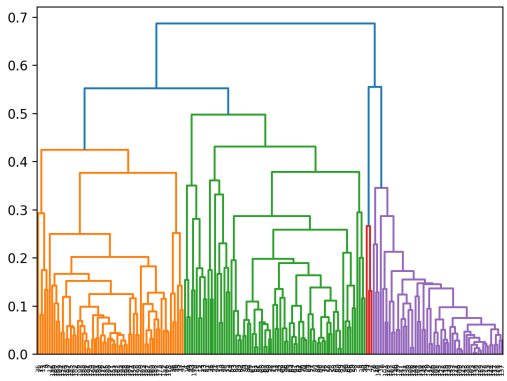

Figure 7: A dendrogram that shows the hierarchical relationship between clusters and datapoints in the article dataset.

Hierarchical Clustering cont.

After analyzing the hierarchical structure of the data and determining the presence of four clusters (distinct groupings), a new hierarchical clustering model with the specified presets. This involved implementing the agglomerative clustering algorithm with the chosen hyperparameter values: "cosine" (affinity) distance metric, "average" linkage, and four (4) clusters.

Results

Behavioral Clustering Model

After clustering the data from the "Behavioral" dataset into two clusters, the model's performance was measured by its silhouette score and observing the scatterplot of the clusters by plotting the first Principal Component (PC1) on the x-axis, and the second Principal Component (PC2) on the y-axis. By observing Figure #8 (the Principal Component Analysis Scatterplot), we see that the clustering model did a good job in clustering the data points into two distinct clusters. Furthermore, we are able to see in the PCA plot that the data points share similar characteristics with how cohesive and dense the centers of the clusters are, especially in Cluster 0.

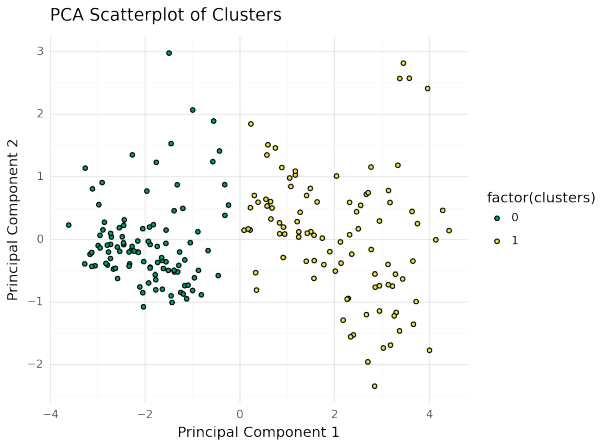

Figure 8: A graph that shows the relationship between Principal Component 1 and 2 and the clustering completed

To comprehend the characteristics of each cluster, the mean values of each variable were computed from each cluster. Through this, we are able to visualize the average consumer profile from each cluster. Displayed below in Figure #9 compares the mean values of the customers' age, current income, time spent browsing, length of subscription, monthly visits, and longest read time from each cluster. Displayed in Figure #10 compares the proportion of ads clicked by customers from each customer. Figure #11 and 12 is a table that displays the mean and standard deviation of each variable from Cluster 0 and 1, respectively.

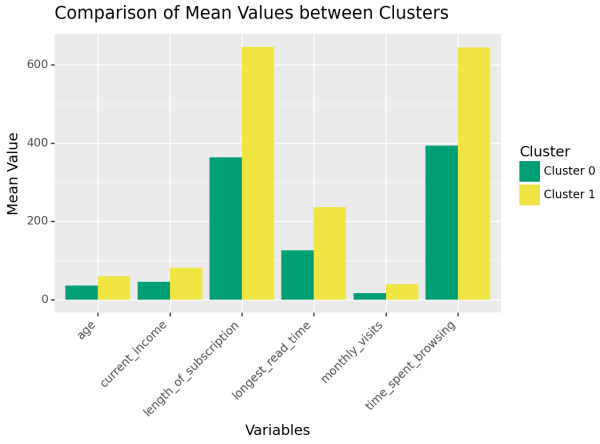

Figure 9: A bar graph that compares the mean values of each variable between Cluster 0 and 1

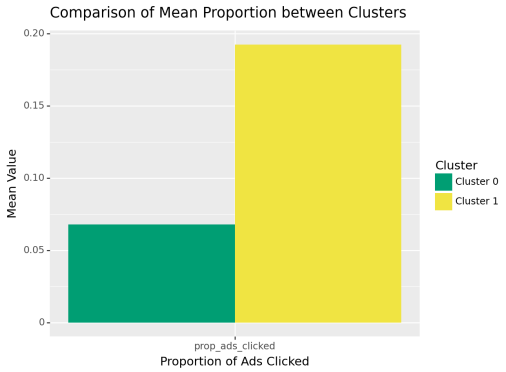

Figure 10: A bar graph that compares the mean proportion of ads clicked between Cluster 0 and 1

Figure #11 - Cluster 0

	Age	Current Income	Time Spent Browsing	Length of Subscrip-tion	Monthly Visits	Longest Read Time	Proportion of Ads Clicked
Mean	35.46	45.88	393.65	363.87	16.36	125.51	0.068
Standard Deviation	6.03	7.67	87.96	125.83	7.23		
42.50	0.08						

Figure #12 - Cluster 1

	Age	Current Income	Time Spent Browsing	Length of Subscrip-tion	Monthly Visits	Longest Read Time	Proportion of Ads Clicked
Mean	60.41	80.97	644.29	645.19	39.37	236.60	0.192
Standard Deviation	10.54	10.35	97.87	276.41	6.67	53.74	0.11

As we see in the figures, the average profile of a customer in Cluster 0 is an "Early Middle Age" (age 35) compared to Cluster 1 where the average profile is a "Late Middle Age" (age 60) customer who has a 15k higher annual salary. Additionally, customers in Cluster 1 spend more resources and time towards the media company–higher time spent browsing per month) (~250 minutes), subscription length (~281 days), monthly visits (~23), longest read time (~111), and proportion of ads clicked (0.12 or 12%).

Through this information, the media company is able to allocate resources towards understanding the average media trends and favorites from each cluster. Additionally, the company is able to understand their customer age demographic and loyalty to the program, and they are able to find trends on increasing customer retention and loyalty.

Article Clustering Model

Dendrogram Results

After creating the dendrogram (Figure #7), the hierarchical structure of the data revealed by the dendrogram shows that the density is somewhat high due to the tall heights of the first y-axis connecting two data points to one another. Additionally after 0.3 on the y-axis, we begin to observe that the clusters are dissimilar to one another, resulting in a tall y-axis line connecting the clusters to one another. As a result, there were four distinct clusters under the threshold.

Clustering Model

Based on the dendrogram results, we see that there was high separation between the clusters completed by the Hierarchical Agglomerative Clustering algorithm. This is proved by the y-axis heights of the stems connecting one cluster to another. While the density of the data was not as high, the model is able to do a good job in combining the similar data points into their respective clusters.

To comprehend the characteristics of each cluster, the mean values of each variable were computed from each cluster. Through this, we are able to visualize the average number of articles read in each topic from each cluster. Displayed below in Figure #13 compares the mean values of the number of articles read for the following topics: stocks, productivity, fashion, celebrity, cryptocurrency, science, technology, selfhelp, fitness, and AI from each cluster. Figures #14, 15, 16, and 17 is a table that displays the mean and standard deviation of each variable from Cluster 0, 1, 2, and 3 respectively.

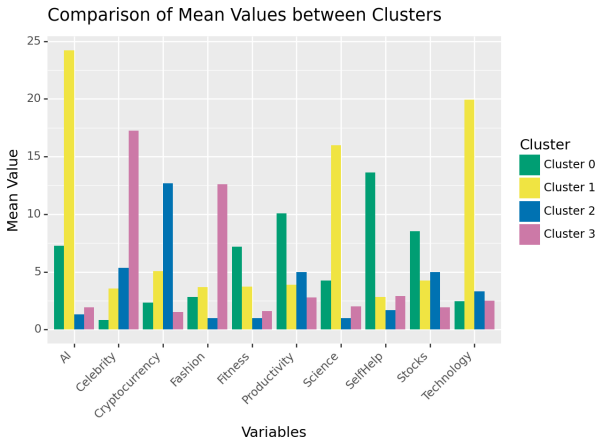

Figure 11: A bar graph that compares the mean values of each variable between Cluster 0, 1, 2, and 3

Please see the following abbreviations for the article topics

- **Stocks** : St
- **Productivity** : Product
- **Fashion** : Fash
- **Celebrity** : Celeb
- **Cryptocurrency** : Crypto
- **Science** : Sci
- **Technology** : Tech
- **SelfHelp** : SelfHelp
- **Fitness** :Fn
- **AI** : AI

Figure #14 - Cluster 0

	St	Product	Fash	Celeb	Crypto	Sci	Tech	SelfHelp	Fn	AI
Mean	8.54	10.08	2.83	0.82	2.35	4.24	2.46	13.63	7.18	7.27
Standard Deviation	6.68	8.88	3.30	1.71	2.90	3.58	2.53	9.87	6.27	6.55

Figure #15 - Cluster 1

	St	Product	Fash	Celeb	Crypto	Sci	Tech	SelfHelp	Fn	AI
Mean	4.25	3.89	3.68	3.54	5.06	15.98	19.92	2.81	3.73	24.21

	St	Product	Fash	Celeb	Crypto	Sci	Tech	SelfHelp	Fn	AI
Standard Deviation	4.63	5.03	4.40	3.50	5.48	10.87	12.52	3.37	3.60	14.21

Figure #16 - Cluster 2

	St	Product	Fash	Celeb	Crypto	Sci	Tech	SelfHelp	Fn	AI
Mean	5.00	5.00	1.0	5.33	12.67	1.00	3.33	1.6	1.00	1.33
Standard Deviation	1.73	3.46	1.0	4.04	5.13	1.73	3.51	2.08	1.73	0.58

Figure #17 - Cluster 3

	St	Product	Fash	Celeb	Crypto	Sci	Tech	SelfHelp	Fn	AI
Mean	1.91	2.77	12.61	17.25	1.52	2.00	2.52	2.89	1.61	1.93
Standard Deviation	2.40	2.74	4.93	5.83	1.96	2.18	2.59	3.78	2.46	2.70

As we see in the figures, customers in cluster 0 are consistent in the number of articles read across all topics, but their top topics are "Self Help" and "Productivity". We can surmise that customers in this cluster are looking for articles regarding life advice. Customers in Cluster 1 are the leading factors in the increase in article reads, especially in AI, Technology, and Science. We can surmise that customers in this cluster are interested in the sciences (specifically technology). Customers in cluster 2 average the second smallest number of articles read, but they focus their time on reading into cryptocurrency. Finally, customers in cluster 3 average the smallest number of articles read, but their focuses are on celebrity and fashion. We can surmise that they would prefer to read celebrity/lifestyle magazines.

Through this information, the customer is able to understand their customer demographics and which topics are the most appealing to them. Additionally, the company can combine the two datasets (given that the IDs match) to visualize if there is a latent structure between consumer profile and the topics they are interested in reading about.

Discussion/Reflection

Through the conducted analyses, latent structures and trends have emerged among customers in the media company based on their consumer profiles and preferred topics for reading. This underscores the importance of accurate clustering in revealing true patterns and trends among

customers. Utilizing the K-Means clustering or similar algorithms enable media companies to gain a deeper understanding of their target demographic, facilitating improved customer retention and loyalty through tailored product offerings.

If these analyses were to be replicated, a valuable addition would be the incorporation of columns capturing customers' preferences for literature consumption (e.g., paper, digital, audio) and their typical reading goals (e.g., leisure, education). This enhancement would provide additional insights into the types of readers who engage most with the company's products and services.

With the ever changing world of media consumption, understanding how and why customers choose to utilize a media company's services is essential for fostering growth and profitability within the industry.