Homework 3

Introduction

Our goal is to better understand our readers' habits and preferences by grouping them into clusters based on their activity on our site and the types of articles they read. We're analyzing data from two samples of 200 customers each to uncover these patterns. This isn't just about numbers, but rather it's about enhancing how we engage with our readers and personalizing their experience. By examining factors like how often they visit our site, which articles they read, and their overall engagement, we'll be able to deliver more of the content they enjoy and find new ways to connect with them. The insights we gain will help us tailor our content more effectively, target our marketing efforts more precisely, and strengthen our position in the digital magazine market.

Methods

Pros and Cons of Each Clustering Algorithm

K-Means

- Pros:
  - Simple and easy to implement
  - Fast and efficient
  - Works well with spherical clusters
- Cons:
  - Assumes clusters are spherical
  - Assumes clusters are of roughly similar size
  - Assumes clusters are of similar density
  - Assumes clusters are well separated
- **Good for data that is well separated, spherical, and of similar size and density**

In Simple Words: It's a good choice if you're dealing with things that are neatly separated, round in shape, and about the same size and crowdedness.

Gaussian Mixture Model (GMM)

- Pros:
  - Can accommodate clusters of different shapes and sizes, particularly ellipsoidal clusters
  - Can accommodate clusters with different densities
  - Can accommodate clusters that overlap
  - Soft Assignment - each data point is assigned a probability of belonging to each cluster
- Cons:
  - Can be slow to converge
  - Can be sensitive to initialization values
- **Good for data that is not well separated, ellipsoidal, of different densities, or that overlaps**

In Simple Words: It's a great choice if you have things to group that aren't neatly separated, are more oval in shape, have different levels of crowdedness, or tend to overlap with each other.

Density Based Spatial Clustering of Applications with Noise (DBSCAN)

- Pros:
  - Can find clusters of any shape
  - Account for noise or outliers
- Cons:
  - Less Effective with high dimensional data
  - Not great with overlapping/touching clusters
  - Suboptimal results with clusters of varying densities
- **Good for data that is not well separated and is of shapes that are not spherical or ellipsoidal, but have roughly the same density and do not overlap**

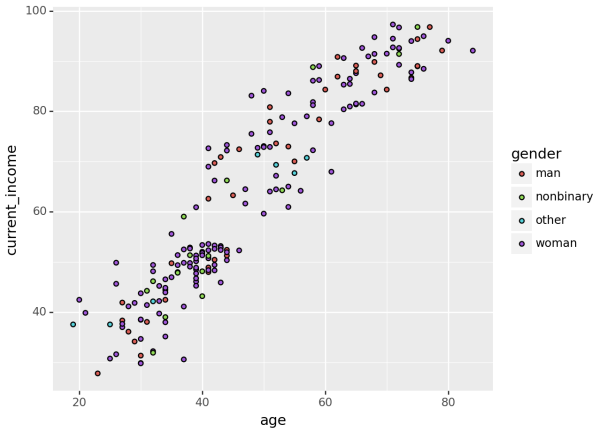
In Simple Words: It's suitable for situations where you have things to group that are not neatly separated and have complex shapes, but these groups should be more or less equally crowded and shouldn't be overlapping with each other.

Hierarchical Agglomerative Clustering (HAC)

- Pros:
  - Flexible Number of Clusters
  - Use hierarchical structure to model relationship between clusters
  - Flexibility with Linkage Methods
- Cons:
  - Computationally Expensive
  - Cannot un-merge clusters
- **Good for data in where there is a hierarchical relationship between clusters**

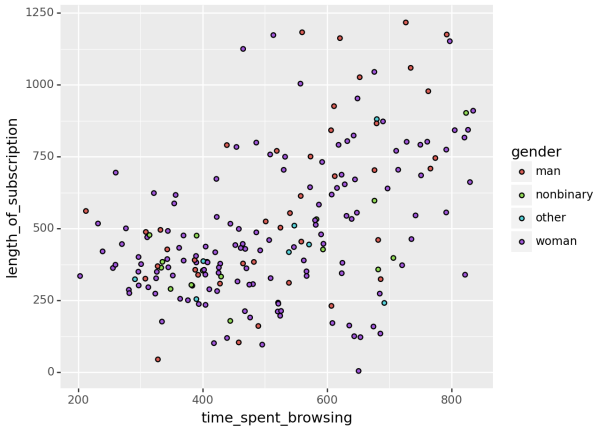
In Simple Words: It works well for situations where you need to group things based on a step-by-step relationship, kind of like a family tree, but it can be resource-intensive and doesn't easily allow for changes once the groups are formed.

Choosing Clustering Algorithm



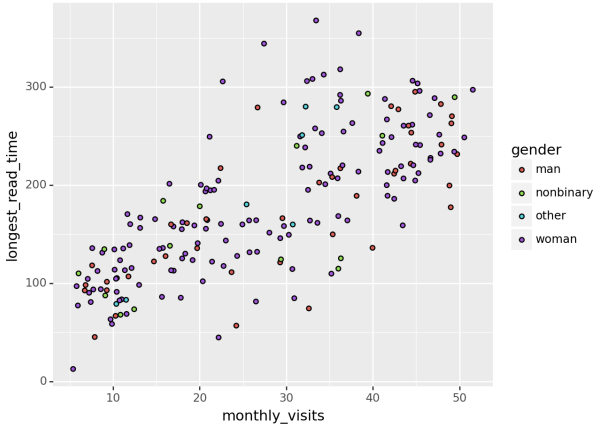
This plot shows a somewhat scattered distribution, with a slight indication that income may increase with age, but not in a strictly linear manner. The data points are not clearly grouped into distinct clusters, and aren't really in a spherical shape but rather are more spread out in a stretched ellipsoid. In addition, the data points are more dense in the lower left corner and less dense in the upper right corner.

In simple terms, the plot suggests that there might be a relationship between age and income, but it's not straightforward, and the way the data is arranged is more like a loosely stretched-out oval, with more data points gathered in one corner than in the other.



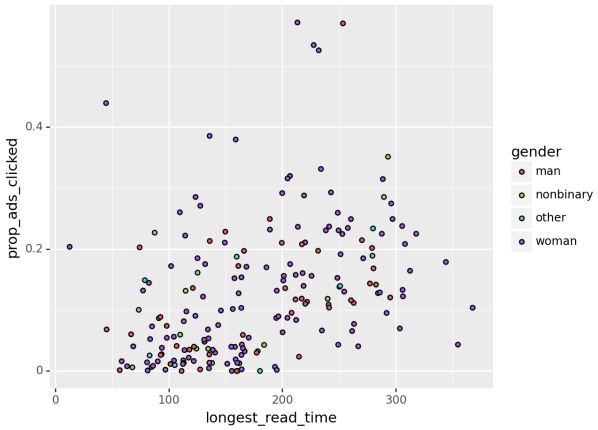
The data appears quite spread out. There are no clear, distinct groups that would indicate a good fit for K-Means, which looks for spherical clusters. It also seems to have a bit higher density in the bottom left corner, which would indicate that the clusters are not of similar size. There is also potential overlapping.

In simple terms, the data is too mixed up and uneven for a method like K-Means, which needs well-defined, evenly-sized, and round-shaped groups to work best.



Similarly, this plot does not show clear clusters. Instead, there is a wide dispersion of data points. Density does seem to be relatively uniform, though potential overlapping or at least bordering is a concern.

In simpler terms, the data on this plot is scattered broadly without forming any obvious groups. The points are evenly spaced out, but there's a chance that some of them might be overlapping or very close to each other, which could be an issue for using DBSCAN.



This plot shows some potential groupings or patterns but not clear spherical clusters. In addition density seems to be much higher in the bottom left corner, as well as potentially overlapping.

In simple terms, the plot suggests some loose groupings but not in neat, round shapes. The bottom left corner is particularly crowded, and there might be some areas where the data points are overlapping or nearly touching.

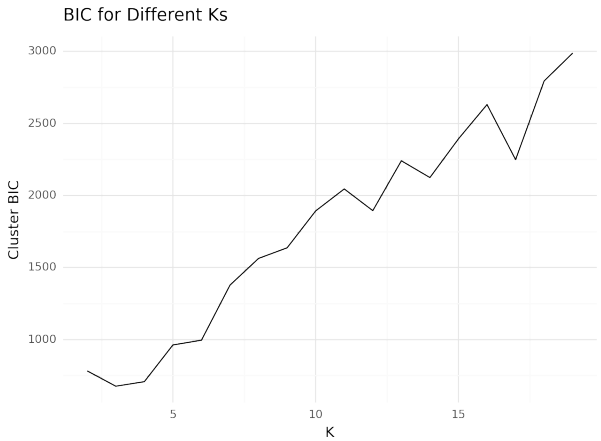
Behavioral Clustering Model: Gaussian Mixture Model (GMM)

Given that none of the scatter plots suggest clear spherical clusters and that the data points are quite dispersed without clear boundaries in addition to the fact that there isn't really any sort of natural hiearchical relationship between the data points, Gaussian Mixture Models (GMM) seems like an appropriate choice. GMM can accommodate the overlap and doesn't assume the clusters are of any specific shape or that there is a consistent density, which seems to match the pattern in the data.

In simpler terms, GMM is a flexible method that can adapt to the way the data is spread out and mixed together. It doesn't need neat, round groups or equal crowding, which seems to align well with the patterns in the data.

Chosen Model Details

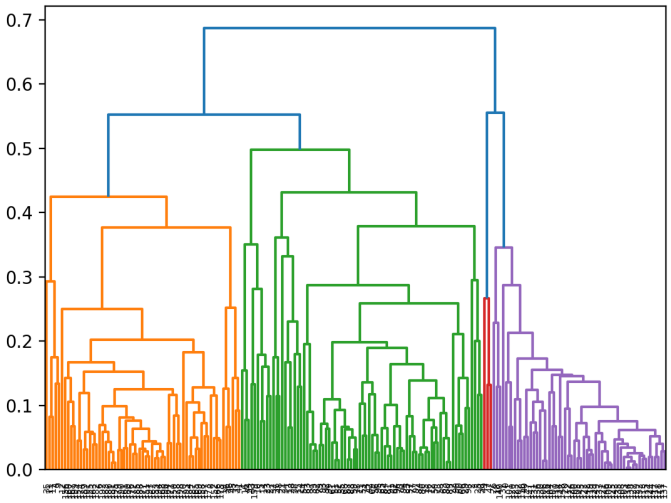
In terms of the pre-processing, we dropped rows with missing values, z score or standardize the values so they can be compared, and then chose the number of clusters via finding the BIC and graphing the BIC.



The best model, according to BIC, is the one with the lowest score. It's like finding the best set of instructions for your puzzle: detailed enough to be accurate, but not so complex that it's hard to follow. We use the elbow method to find the lowest BIC score, which corresponds to the best number of clusters. The elbow method is like finding the best number of groups for your data by looking for the point where adding more groups doesn't make your model much better. It's about finding the sweet spot where the model is accurate but not too complicated. The cluster size we chose was 4, which is the point where the BIC score starts to level off. We use a special method called PCA to simplify the data into just two main parts, which we then plot on a graph. This makes it easier to see and understand the patterns and groups in the data.

Article Clustering Model

In the Article Clustering Model, we perform pre-processing by dropping any rows with missing values, similar to what we did for the Behavior Clustering Model. We do not need to z score or transform the values to be compared since they are all counts.

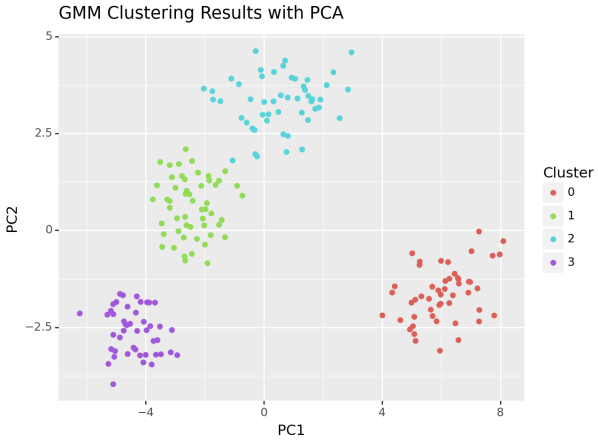


We then create a dendrogram, which is a tree-like diagram that shows the arrangement of the elements (like data points) that are grouped together in a hierarchy. Imagine it like a family tree, but instead of showing family members, it shows how different data points are connected or similar to each other.

In a dendrogram, you look at the heights of the branches, which represent distances or differences between data points. The taller the branch, the bigger the difference, so by finding where these tall branches are, you can determine the best number of clusters by seeing how the data naturally groups together at different levels of similarity. By looking for where big gaps occur in this 'tree', we can decide the best number of groups to split the data into, making it easier to understand and work with. Based on this, the number of clusters chosen was 4. We then use a HAC model to create the clusters, which is a method that groups data points together based on how similar they are to each other.

Results

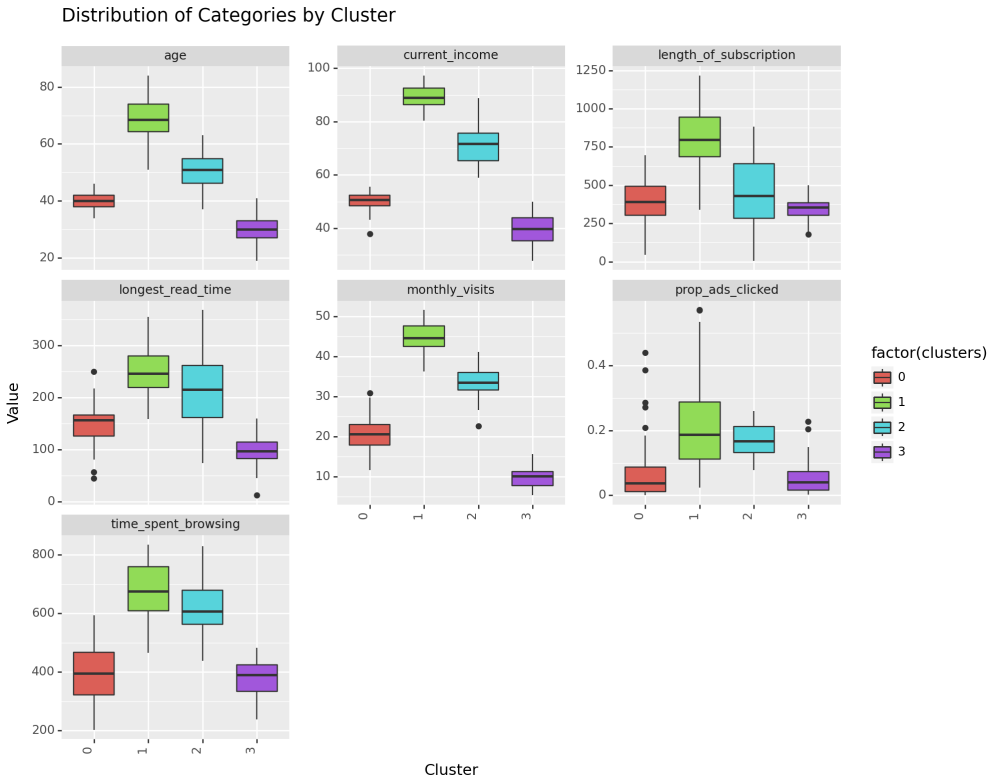
Behavioral Clustering Model



The PCA plot shows that the data is divided into four groups, and we can see them as separate from each other, which means the GMM method worked to some extent. There doesn't seem to be a lot of overlap or even bordering, so they are fairly distinct.

Dataset	Model	Silhouette Score
Behavior	GMM with PCA	0.1809
Articles	HAC	0.1488

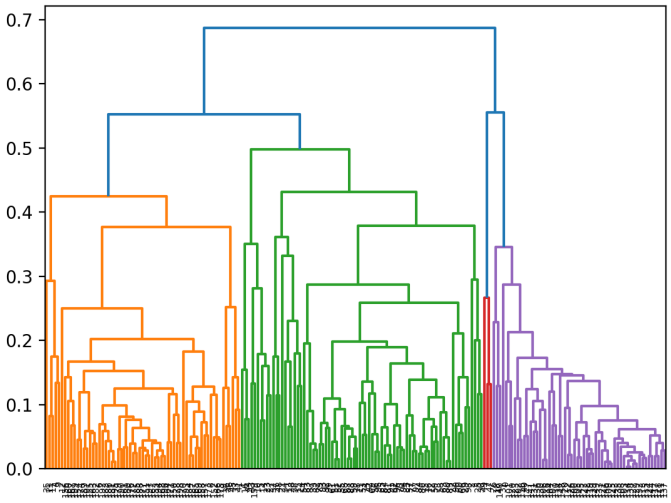
However, the silhouette score, a measure of how well-separated and tight these groups are, is only 0.18 out of a perfect 1.0. This tells us that the groups aren't very tightly packed or clearly distinct from each other; there could be some mixing or the groups might not be very neatly formed.



In terms of the 4 clusters for behavior, cluster 1 seems to be the oldest, averaging at 70, while cluster 3 seems to be the youngest, averaging at 30. There seems to be a general ocrrelation with age, income, and time spent browsing, with the older clusters having higher values for these variables. However, the proportion of ads clicked seems to be more evenly distributed, with cluster 1 having the highest proportion of ads clicked. The longest read time seems to be more evenly distributed as well, with cluster 3 having the highest longest read time. In general, it seems to as you increase age and/or income, you increase the length of subscription, time spent browsing, ads clicked, visit frequency, and longest read time.

In summary, these clusters represent distinct groups of customers with different behavioral patterns on the media site, from browsing habits to engagement with content and ads. This information could be instrumental in tailoring content and marketing strategies to each unique customer segment.

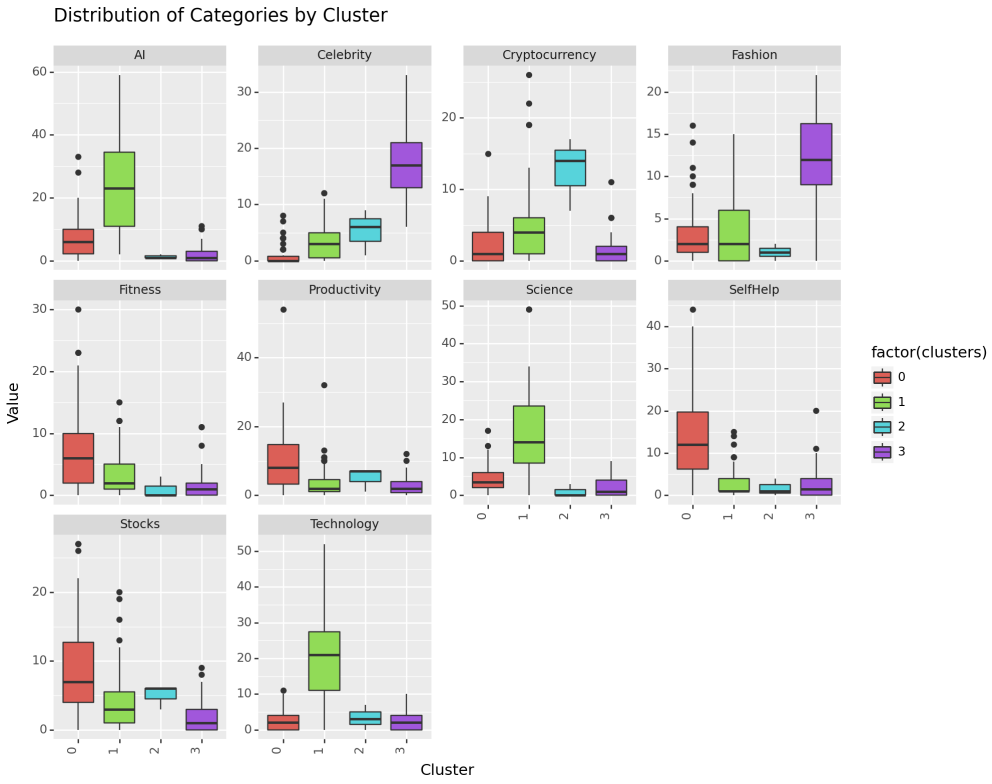
Article Clustering Model



Looking at the dendrogram we created for the articles dataset again, we can try to evaluate some of the model performance. The dendrogram presents a visual representation of the data points being merged into clusters based on their similarity. The height of the branches indicates the distance between clusters, with taller branches suggesting less similarity. For example, the red cluster does not seem to have a lot of datapoints, while the other three clusters seem to have a lot more. In addition, the red and purple clusters are closer to each other vs the green and orange clusters. In addiiton, near the "top" of the orange and green clusters, we can see a significant height increase, which suggests that the datapoints in these clusters are not very similar to each other.

Dataset	Model	Silhouette Score
Behavior	GMM with PCA	0.1809
Articles	HAC	0.1488

The silhouette score for the article clustering model is 0.1488, which is lower than the silhouette score for the behavior clustering model. This suggests that the clusters are not as well separated or distinct from each other, and there could be some mixing or overlap between them. This makes sense when we consider at the top of the orange and green clusters in the dendrogram, where there is a significant height increase, suggesting that the datapoints in these clusters are not very similar to each other.



In terms of the 4 different clusters, cluster 1 seems to be interested in STEM related articles, sporting high counts in AI, Science, and Technology, though surprisingly not cryptocurrency. There are however some outliers in cluster 1 that have high counts in that category. Cluster 0 seems to be most interested in fitness, productivity, self help, and stocks. Cluster 3 seems to enjoy reading about cryptocurrency the most, and cluster 4 seems to enjoy mostly reading about celebrity and fashion articles.

In summary, these clusters represent different categories that customers frequent. This information could be useful in seeing what types of articles are most popular and what types of articles are not as popular, which could be useful in tailoring content to each unique customer segment, such as bundling together articles that are popular with the same cluster in a package deal, such as an AI and Technology bundle for cluster 1.

Discussion/Reflection

Through this analysis, I've learned that different data types demand distinct clustering strategies, with visualization tools being invaluable for guiding these decisions. It was helpful to weigh the different pros and cons of utilizing the different clustering methods, and then analyzing the scatterplots of the behavioral dataset to determine which method would be most

appropriate. If I had to change on thing I would potentially use a different clustering algorithm for the article dataset, as I'm not completely convinced HAC is the most appropriate method to use. Looking at the dendrogram, you could see that the clusters were not very well separated, and there was a lot of overlap between them. I would potentially try to use a different clustering algorithm that would be better suited for this type of data, such as GMM or DBSCAN.