Homework 3

Introduction

The goal of this assignment is to learn information about customers of a digital magazine company. Using clustering techniques and analyzing results, we can find out better ways to target groups of audiences. To learn about these customers two data sets were analyzed. The first looked at demographic and behavioral characteristics of customers, which included the fields: gender, age, income, browsing time, proportion of ads clicked, longest session, monthly visits, and number of days subscribed. The second set looks at the number of different article topics users read. Interpreting the data can help the company target customers with better ads and suggested articles, as well as understanding what topics and audience they should be focusing on the most.

Methods

The second model, for the article data, was a Hierarchical Cluster Model. No preprocessing was needed, as all data fields were on the same scale, being they were all total counts. For the two hyperparameters, the distance metric used was cosine similarity, and the linkage criterion was average linkage.

Behavioral Clustering Model

Pros and Cons

K-Means:
Pros:

- Lightweight
- Simply identifies spherical cluster patterns
- Good for dense separated circular clusters

Cons:

- Sensitive to outliers
- Not good for non normal distributions or weird shapes

GMMs:

Pros:

- Probabilities instead of hard assignments
- More flexible cluster shapes - ability to create elliptical clusters

Cons:

- Expensive
- Assumption of Gaussian distribution
- Harder to interpret (bic)

DBScan:

Pros:

- No need to specify cluster number
- Can identify complex cluster shapes
- Can completely ignore noise

Cons:

- Fine tuning of hyperparameters (epsilon and min points)

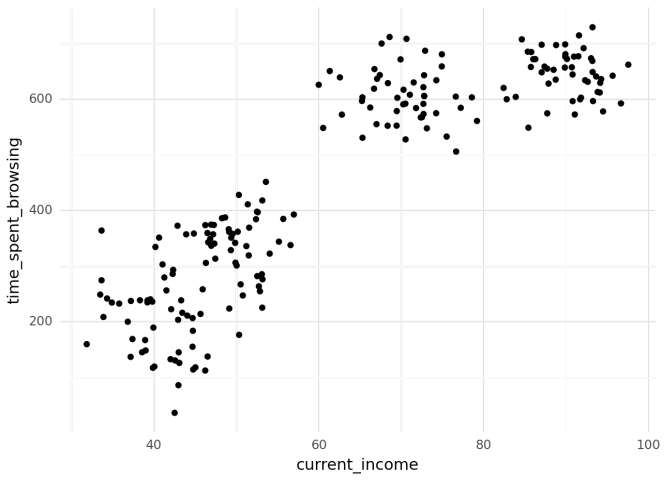
Hierarchical:

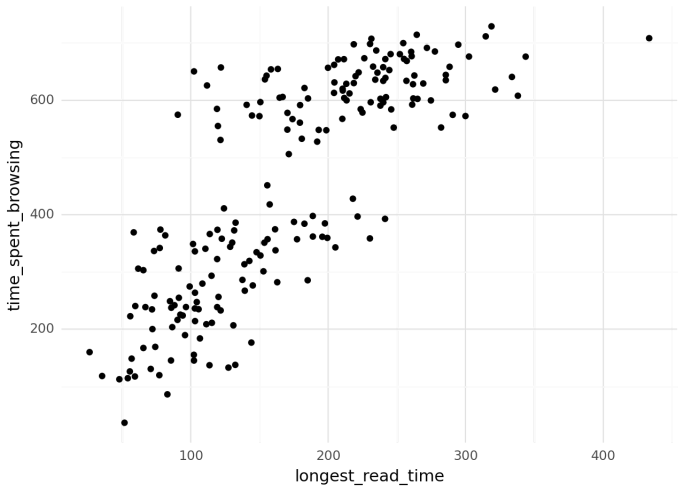
Pros:

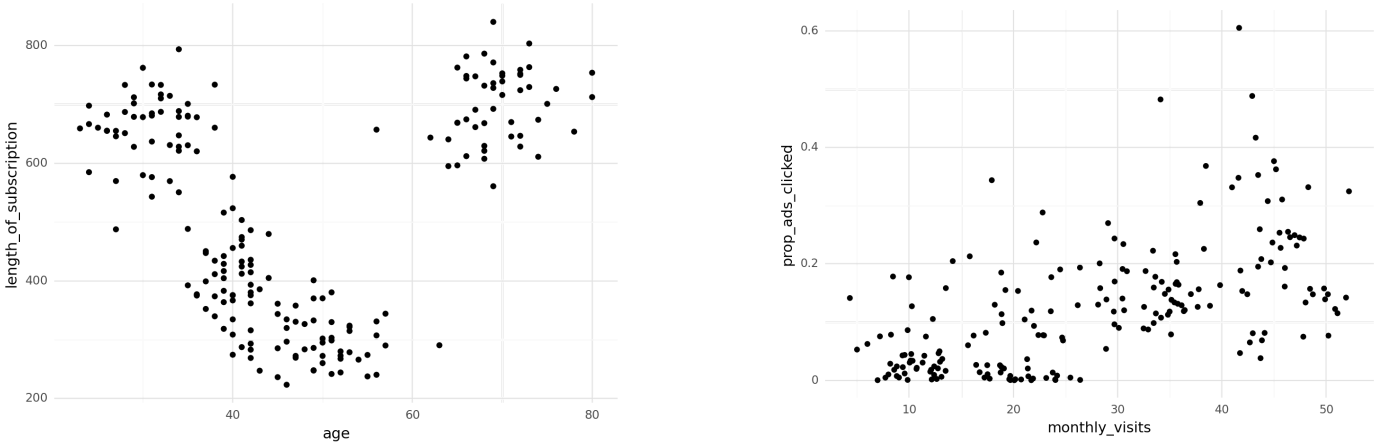
- Visualization - dendrogram
- Don’t need to specify number of clusters in advance
- Hierarchical relationship between points
- Shape is not important

Cons:

- Difficult with large datasets
- Sensitive to noise
- Useless if data has no hierarchical relationship







Scatter plots for various combinations of behavioral variables

Chosen Model Details

The first model is a Gaussian Mixture Model on the behavioral data. Looking at the scatterplots between pairs of variables, clusters were visually depictable. Both K-means and GMM were considered after looking at the scatter plot data, as well as the distinction of using 3 or 4 clusters. GMM was chosen due to the non-spherical shape of some of the clusters, and because the silhouette score was higher than the K-means model. I decided to go with 4 clusters because even though the BIC score was slightly higher, the silhouette score was significantly higher. Behavioral data values were continuous and needed to be z scored.

Article Clustering Model

The Article data was processed using a Hierarchical Clustering Model. This was chosen by the assignment. There was no processing needed, as all fields were just a total count of articles read by a user, sorted by topic. After plotting the dendrogram, I needed to select the number of clusters, the only hyperparameter needed for hierarchical clustering. Setting the dendrogram threshold to 0.55 created three distinct groups, which was confirmed by the fact that using 3 clusters in my model created the highest silhouette score of around 0.28.

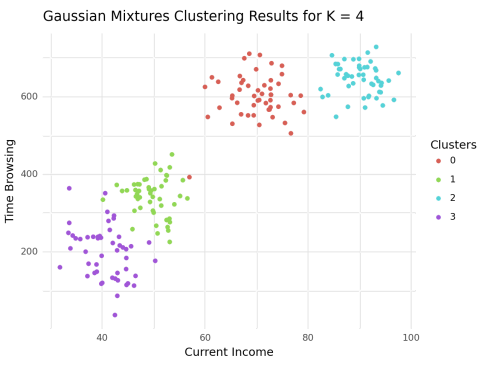
Results

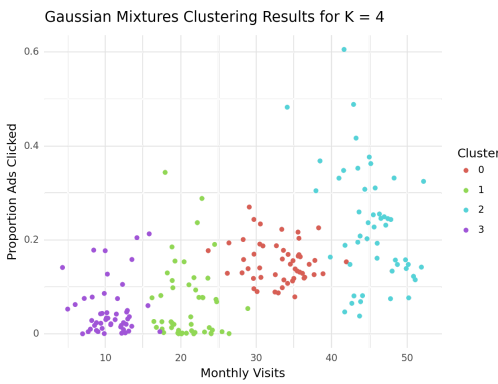
Behavioral Clustering Model

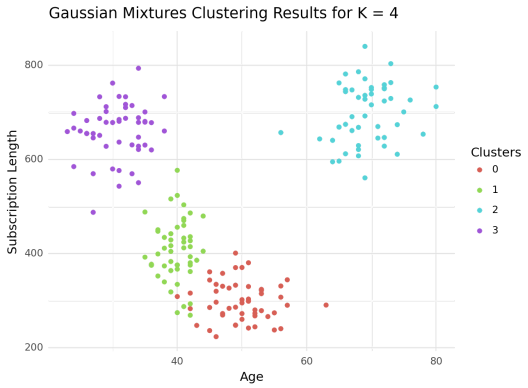
How did your model perform? What kind of customers are in each cluster? How might this information help the company? What does your PCA plot tell you about your clusters?

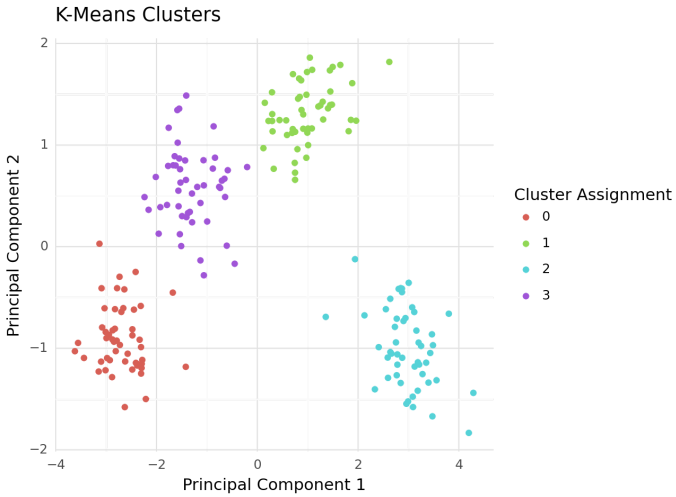
My model performed pretty well, and had a silhouette score of around 0.61. While I was not able to analyze all 21 combinations of predictors, only one that I looked at had any significant overlapping of clusters (and would have worked better with 3 clusters in any case, however all the others required 4). The PCA graph clearly shows 4 distinct clusters of customers. Next I created a table which showed the mean value of all of the variables for each of the 4 clusters, which gave a lot of insight to what type of people are in each cluster.

One cluster depicts a young, early career group. They make the lowest wages, spend the least time browsing, have the lowest number of monthly visits, have the shortest longest-read-time, and click on the least number of ads. However, this group has been subscribed to the magazine for almost the longest duration. This group seems to have little time to spend relaxing and reading this magazine due to the business that the early-thirties brings, and probably would value an efficiency-focused approach of the media company. Following an upward and to the right trend, the next group is about 40 years old, makes around 10k per year more than the last group, and spends a little less than half the time on the magazine. This group has actually been subscribed for a significantly shorter period of time. Next, there is a cluster of late 40s to early 50 year olds, mid to late career and making a lot more money (70k) . They spend almost double the time browsing as the last group, have 30 monthly visits (compared to 20 and 10), click on a lot more ads, and seem to have longer reading sessions. They have been subscribed for the shortest time, but at the age of kids going of to college, and being comfortable with their positions in their careers, probably realized they had more time on their hands for leisure. Finally, there is a group of around 70 year olds. This seems to be retirement age, yet this group has the highest income at 90k per year. They have been subscribed the longest, spend the most time browsing, have the highest monthly visits, longest reading sessions, and click on the most ads. It is very important for the company to consider this group of customers when making decisions.





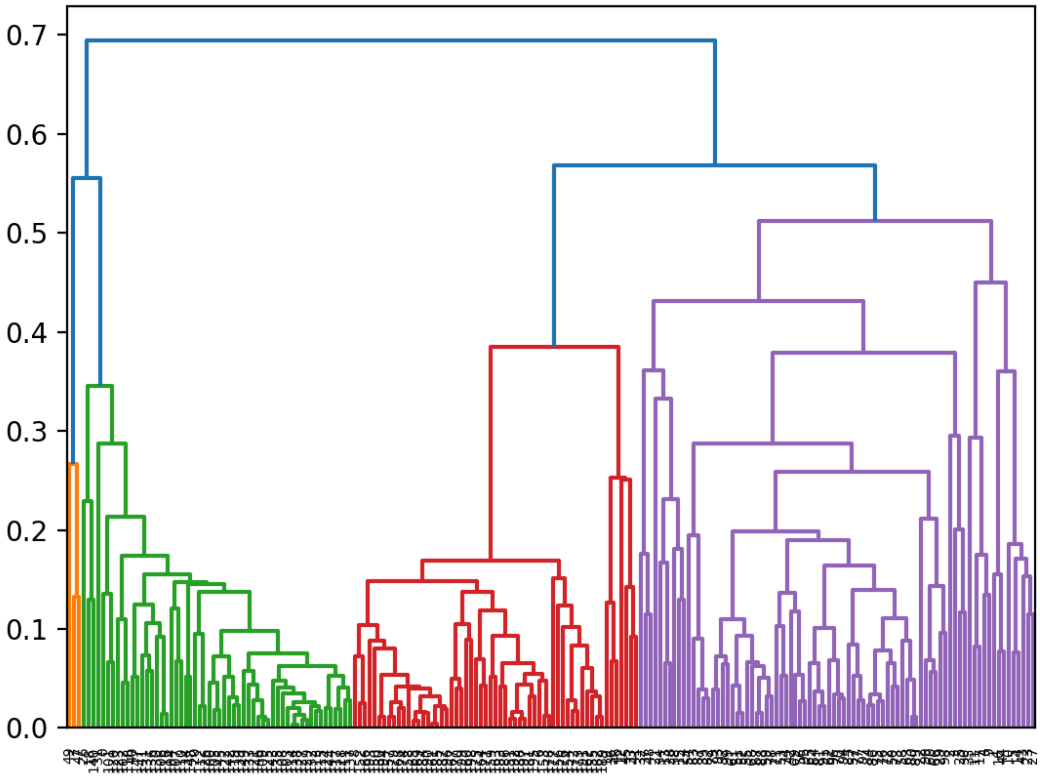




clusters	age	current_income	time_spent_browsing	length_of_subscription	monthly_visits	longest_read_time	prop_ads_clicked
0	30.70	40.71	199.09	660.92	10.66	89.71	0.05
1	49.86	69.98	600.80	298.50	33.10	206.72	0.15
2	69.38	90.15	645.68	698.61	45.22	244.24	0.23
3	39.86	49.74	342.06	407.62	21.03	140.82	0.07

Article Clustering Model

The hierarchical model seemed to perform ok. A 0.27 silhouette score is not great, but it does show that distinct clusters were made. The boxplot of the 3 clusters created showed which type of article each group (green, red, and purple) read the most. The green cluster had the most interactivity with Stocks, Productivity, Self Help, and Fitness. This group seems to be the “self improvement” audience. Next, the red group seemed to almost exclusively read from the Celebrity and Fashion topic sections, at a much higher rate than the other clusters. That seems to be the “pop culture” space. Finally, the purple group read about Cryptocurrency, Science, Technology, and AI, and it's clear to see that all of those topics are closely related as well. After looking at the box plot results, it seems that the model actually did a pretty good job in identifying groups of customers with distinct interests.



Dendrogram for article data

Discussion/Reflection

I really enjoyed this activity, and being a kinesthetic learner, the ability to apply what we learned about unsupervised models to an applicable model that people in industry would actually use, was really helpful for me to grasp the concepts beyond theory. I got to see how different clustering techniques grouped data, and why certain ones are better for certain occasions. If I was to perform this analysis again, I would start it earlier, because it took me a lot longer than the other homeworks.