Homework 3

Introduction

There are two models that we are creating on two different datasets. The first takes continuous variables age, current income, time spent browsing, proportion of ads clicked, longest read time, length of subscription, and monthly visits. The second takes data on the number of articles a given customer reads within a range of categoeries: stocks, productivity, fashion, celebrity, cryptocurrency, science, technology, selfhelp, fitness, and AI. This customer data will be used to cluster different customers into groups. These clusters are meant to identify different types of customers, which can be used to help target and market towards their traits and tendencies.

Methods

Our first model took data from the bevhaiour data set. The model chosen was a KNN model, after we observed some trends between age and current income. Afterwards we tested the Sillhouette Scores for each K to decide our K and then plugged that into our model.

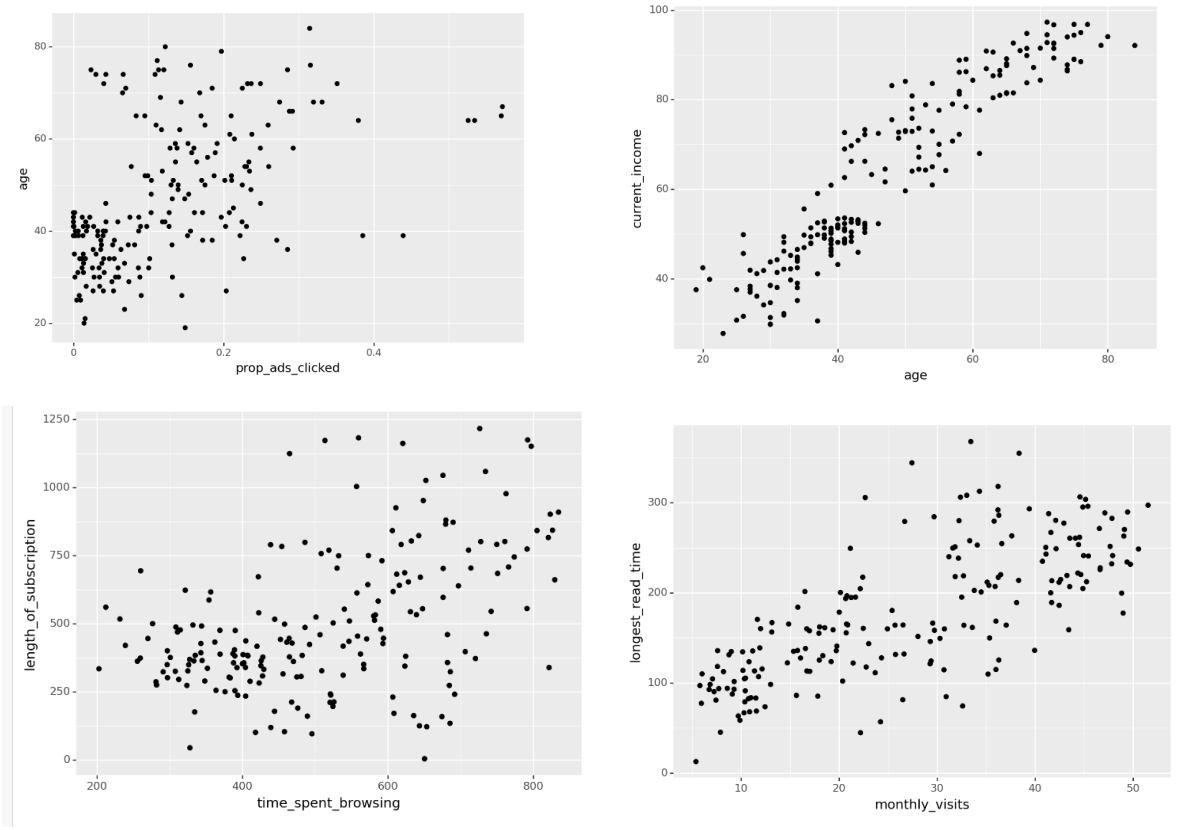
Our second model was a hierarchical clustering model, which used cosine distance and average linkage. After running a Scree Plot and Cumulative Variance Plot, we divided the dendrogram into 7 clusters.

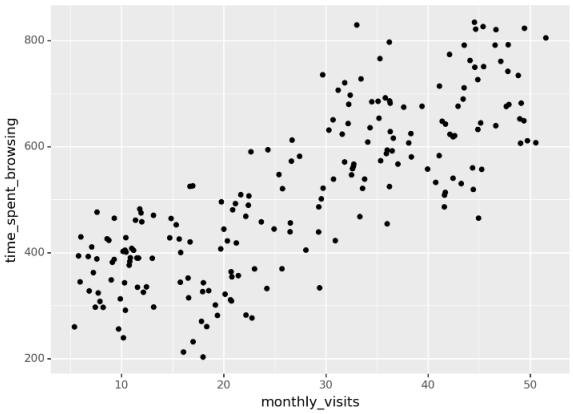
Behavioral Clustering Model

Pros and Cons

KMeans is a simple agorithm that is great with continuous data but can have problems with higher dimensions and the values such as the mean of a cluster can be affected by outer noise. Gaussian Mixture Models also suffers from that problem, but differs in two major ways: every data point has some probability of being a part of every cluster, and there are differenes in variances between each cluster. This is great for clusters with differing sizes, but it will struggle to incorporate categorical features. In order to avoid variances and means from being affected

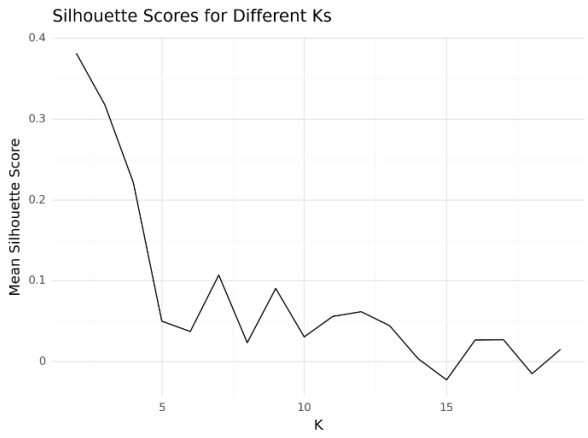
by noise points, you can use DBSCAN, which will group points based on the hyperparameters epsilon and minpoints. This allows DBSCAN to work well with clusters of different shapes, and also the number of clusters don't need to be specified. One major issue with DBSCAN however is when two clusters come into contact it will likely group the two together into one. The last clustering algorithm is hierarchical clustering, which allows for clusters within clusters. This allows ordering of objects, and also doesn't require a specified number of clusters. A disadvantage of hierarchical clustering however is that you cannot undo any clustering and it is very slow.





Chosen Model Details

The model chosen was K-Means, due to the many continuous variables and its easy implmentation. The only hyperparameter that needed to be chosen was the number of clusters, which we used the Sillhouette Scores to find the optimal K value, which was 6.



Article Clustering Model

Hierarchical Clustering was used for our article clustering model. It is displayed using a dendrogram and we used cosine distance with average linkage. We chose these because we are looking to find the coorelation between the different variables. The only hyperparameter that needed to be chosen was the threshold. The threshold was changed according to the vertical heights of the groups within the dendrogram. The final result for the threshold was 0.5, which lead to four clusters.

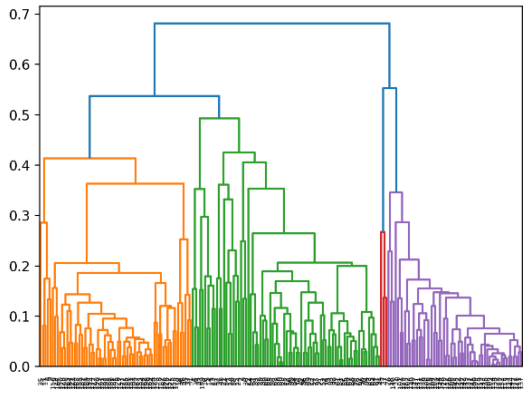


Figure 1: article dendrogram

Results

We can see through these metrics that there is coorelation between three different groups. These I classify as the technology group, the stocks group, and the fashion group. That being said, AI was a commonly coorelated with almost every genre.

```
article.groupby(['Technology']).mean().sum()
```

```
Stocks            166.496951
Productivity      154.974010
Fashion           164.208198
Celebrity         138.610248
Cryptocurrency    152.077113
Science           527.068203
SelfHelp          138.332574
Fitness           173.137695
AI                775.159127
id               3616.781435
dtype: float64
```

```
article.groupby(['Stocks']).mean().sum()
```

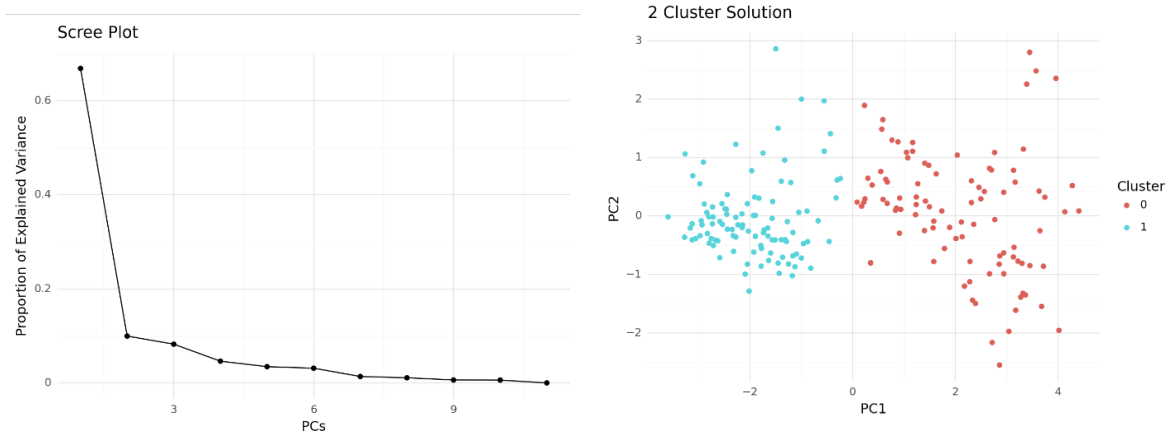
```
Productivity      193.572805
Fashion            88.558352
Celebrity          79.550103
Cryptocurrency     63.305506
Science           180.625741
Technology        198.884337
SelfHelp          239.876121
Fitness           149.031921
AI                333.445490
id               2660.009734
dtype: float64
```

```
article.groupby(['Fashion']).mean().sum()
```

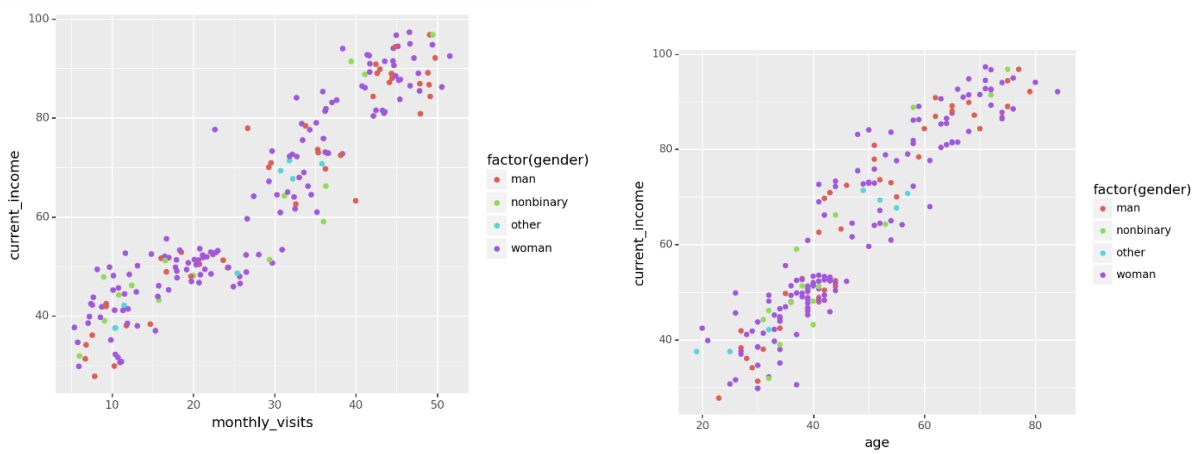
```
Stocks              86.297955
Productivity        97.900305
Celebrity          250.026129
Cryptocurrency      59.882173
Science            135.133364
Technology         149.678999
SelfHelp           105.246337
Fitness             81.321856
AI                 189.294292
id                2267.248107
dtype: float64
```

Behavioral Clustering Model

The clusters of our model didn't group them very well, so reduced dimensionality using Principal Component Analysis. We found the number of PC using the Scree Plot and ended with 2 clusters:



Lastly, we would like to illustrate the relationship between income, monthly visits, and age. A user that is older in age also has more income, and people with higher income typically visit more often. This helps us market towards older audiences as they can typically afford the subscription and use it more often. Also as shown in the plots, gender does not play a significant role in any of the above parameters.



Article Clustering Model

Our dendrogram split the groups well, however there was one outlier in the form of a fourth cluster. This likely meant that in a DBSCAN context it would've been a few noise points.

We can also see that the AI, Science, and Technology group saw the most articles read amongst all users.

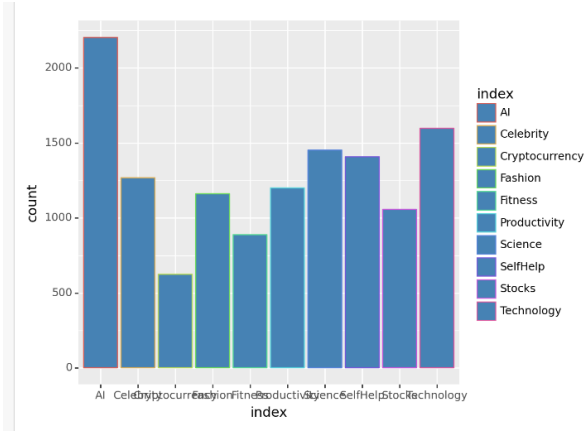


Figure 2: Genre Popularity

Discussion/Reflection

A few sentences about what you learned from performing these analyses, and at least one suggestion for what you'd add or do differently if you were to perform this analysis again in

the future.

These clustering algorithms will group customers to help target and adjust a businesses marketing strategy. Using our behaviour model we can see a coorelation between income, age, and monthly visits. This can be applied by marketing and catering more towards older users, who have more income and will visit the site more often. Using the article model, we can see the popularity of the technology, science, and AI group, and might want to create more articles involving those topics. If I were to change anything about my models, I think I might try to search for more relationships between the different behaviour predictors.