Homework 3

Introduction

```
We're examining two sets of data—how people behave and what topics they
like on our online magazine. By looking at demographics and user behavior,
and also checking which topics are popular, we plan to group customers
based on these things. This will help their media team adjust what they
offer, how they promote it, and how they engage with users. Instead of
just glancing at basic info about people, we want to truely understand
what each person likes and does not like. This way, we can give them
content that matches their interests, making them happier and hopefully a
long term or life long customer.
```

Methods

I used K-means model on the Behavioral data, and Heiarchal clustering for the article data. I chose K-means for the simplicity and the fact that my data looked spherical and had distinct clusters.

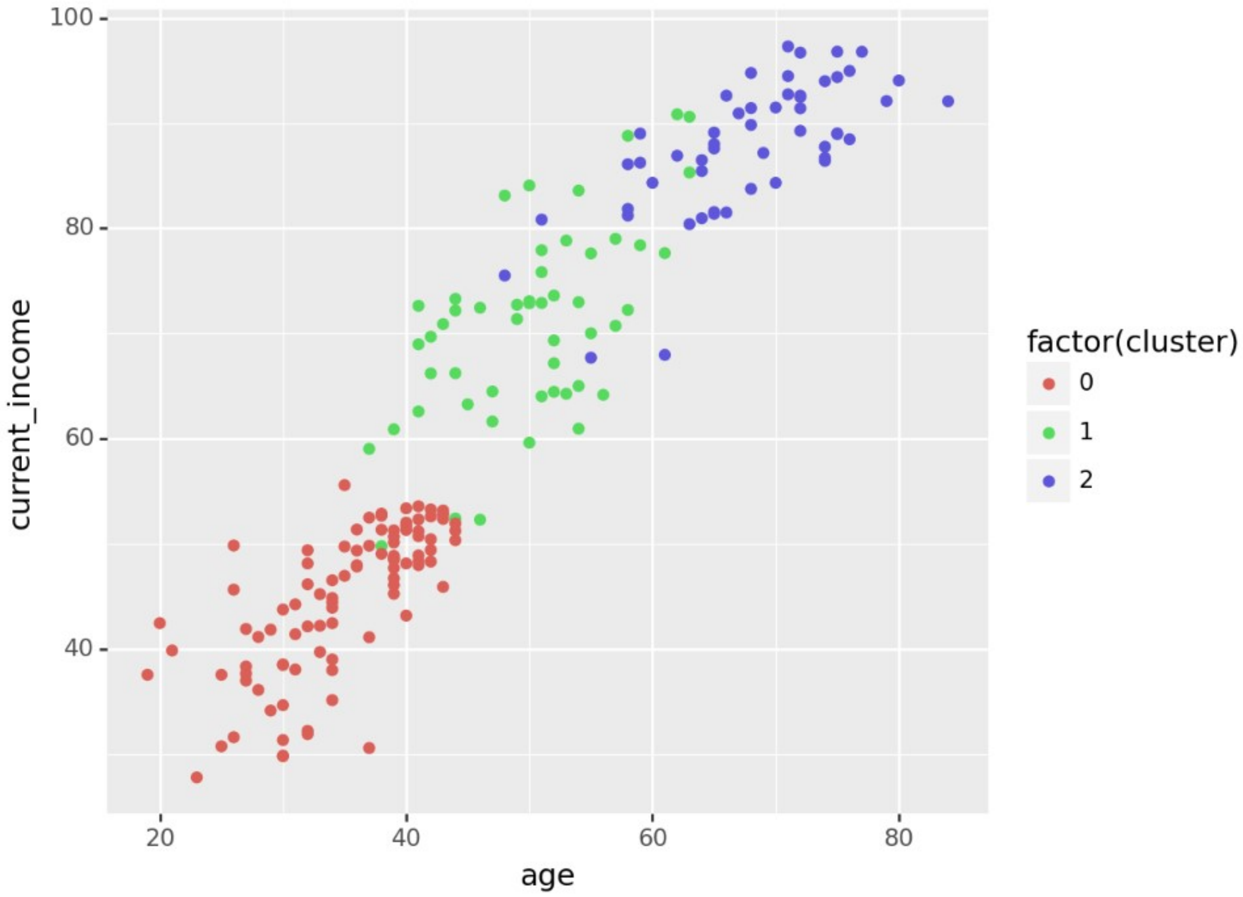
Behavioral Clustering Model

Pros and Cons from quarto doc

```
K-Means:
Pros:
1. Efficient and scalable for computers.
2. Works well with shperical clusters.
3. Simple and easy to understand
Cons:
1. Assumes clusters are spherical and equally sized
2. Sensitive to the initial choice of centroid.

Data Type:
```

```
Works well with data that has clear and seperated clusters.
DownSides: Sensitive to outliers and noise.

Gaussian Mixture Models (GMM):
Pros:
1. Accommodates clusters with different shapes and sizes.
2. Provides probabilistic cluster assignments.
3. More flexible than K-Means.
Cons:
1. More demanding on the computer than K-means.
2. Sensitive to the initial choice of parameters.

Data Type:
Works good with complex data and overlapping clusters.

Downsides:
Requires careful initialization, and number of components needs to be
specified.

DBSCAN:
Pros:
1. Does not assume a fixed number of clusters.
2. Can find clusters of arbitrary shapes.
3. Robust to outliers.
Cons:
1. Sensitive to choice of hyperparameters.
2. Not suitable for data with varying cluster densities.
3. May struggle with high-dimensional data.

Data Type:
Works well with data containing clusters of similar density.

Downsides:
Difficulty in finding clusters with varying densities.

Hierarchical Clustering:
Pros:
1. Captures hierarchical relationships.
2. No need to specify the number of clusters beforehand.
3. Can be visually represented as a dendrogram.
```

```
Cons:
1. Computationally more intensive, especially for large datasets.
2. Sensitive to the choice of distance metric.
3. Difficult to undo previous merges.

Data Type:
Works well with data containing nested clusters.

Downsides:
Not as scalable as other methods, especially for large datasets.
```



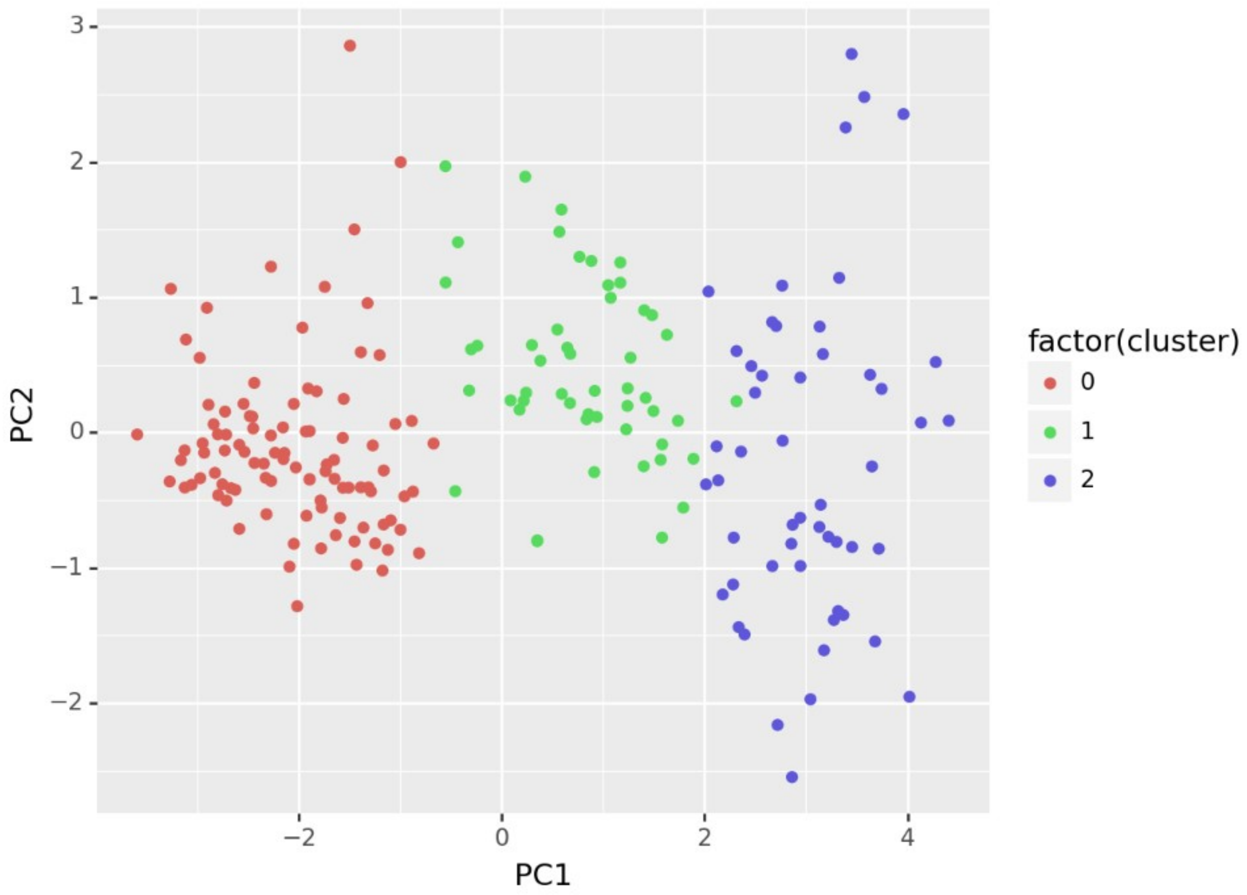
Age vs Income

Chosen Model Details

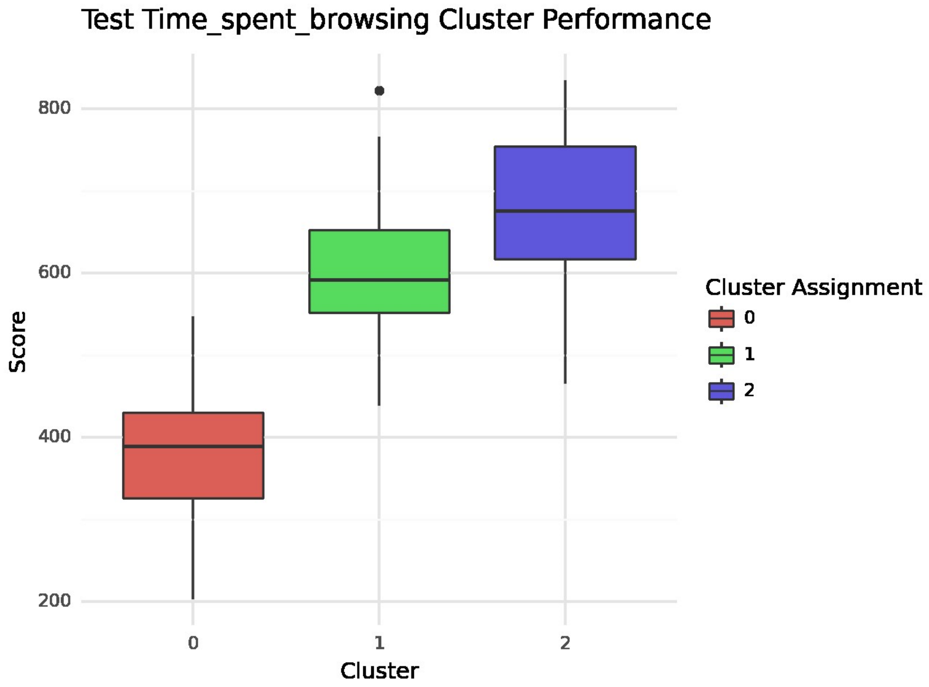
```
I chose K-means as the data appeared to have 3 distinct groups that were
all spherical. Now that I have finished the project though, I can kinda
see how GMM was probably the best method to use, as there are overlapping
clusters in many of my graphs. All I had to set was the k clusters being
3, I messed around with 2 as well as 4 and got 3 to be the best answer.
Preprocessing included one hot encoding gender variable.
```

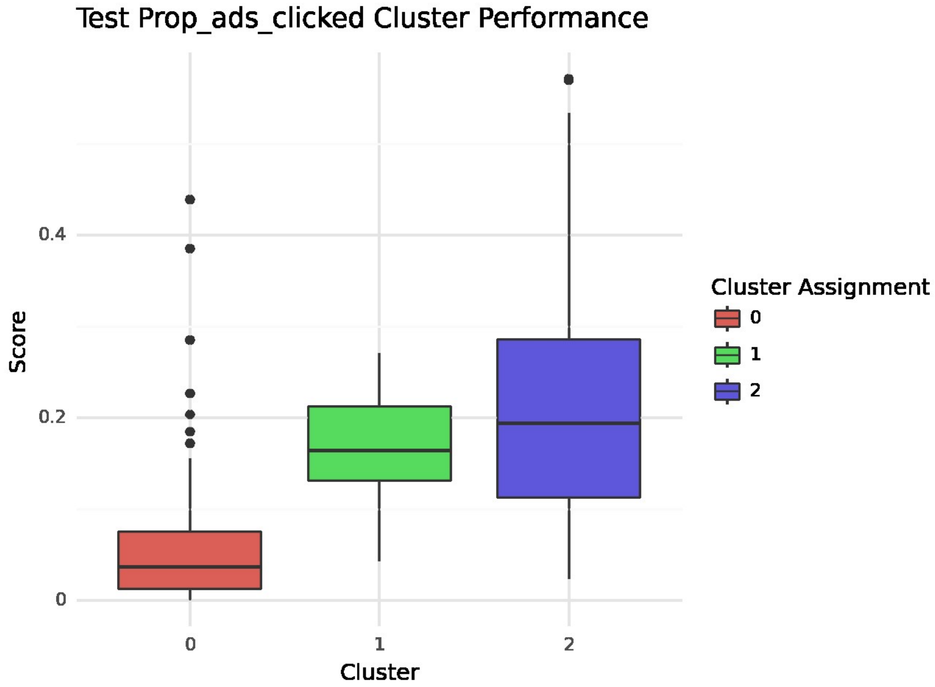
Results

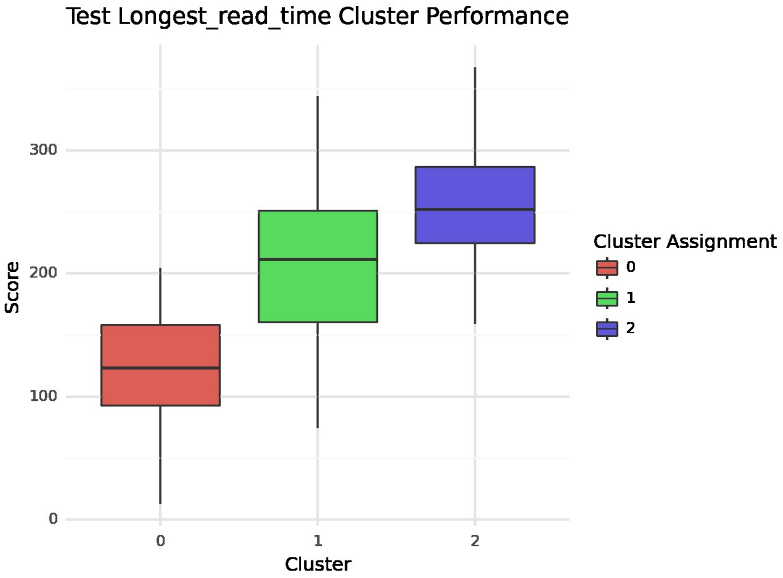
Behavioral Clustering Model

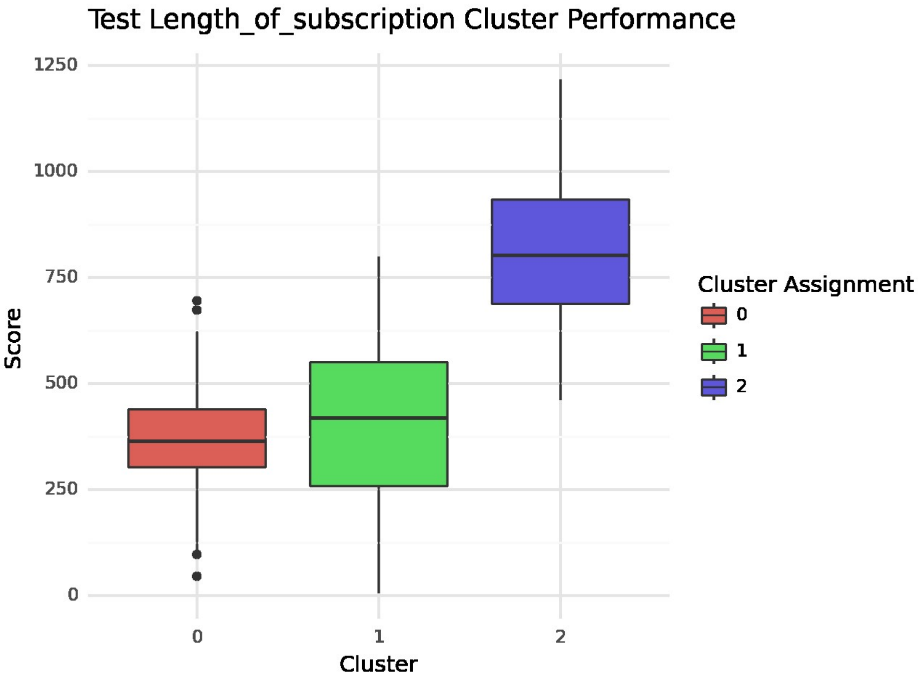


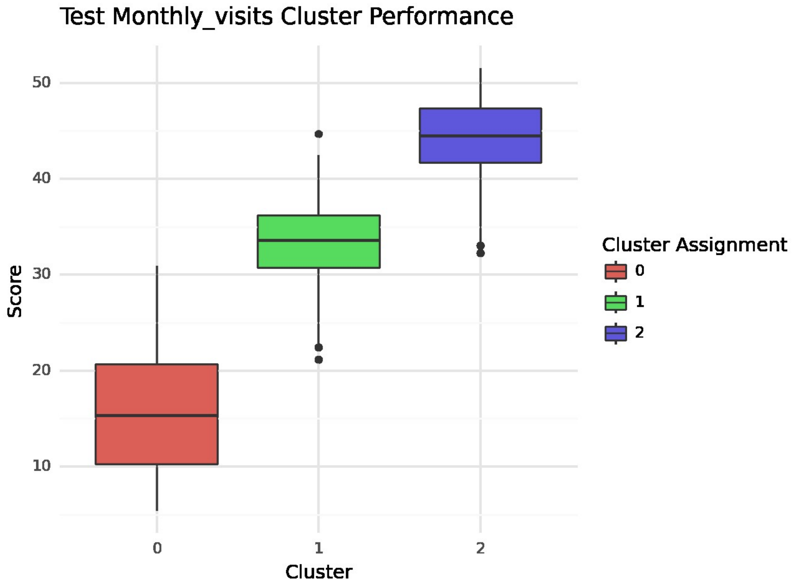
As we can see here, cluster 2 is heavily affected by PC1, as we see the large increase in x. While PC2 primarily affects cluster 0 and 1. And negatively affects cluster 2. From looking at the three clusters we can tell they are fairly separated as well. With barely any overlap.

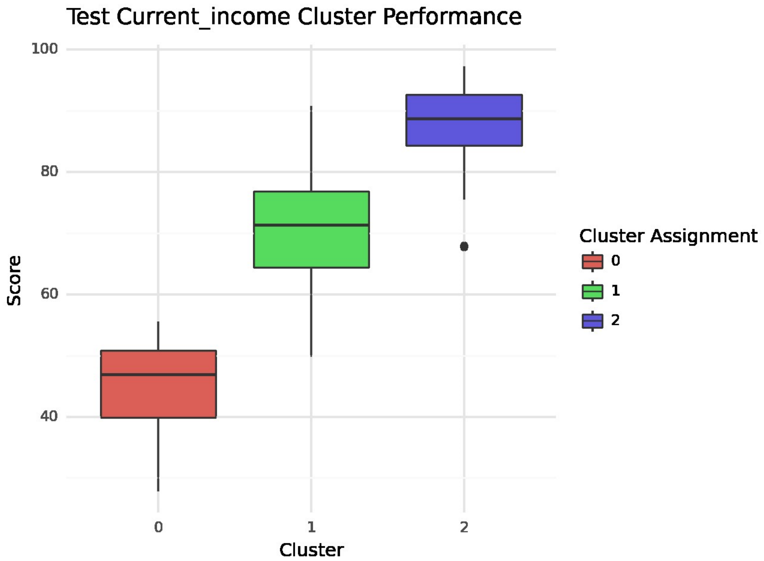

Test Time_spent_browsing Cluster Performance

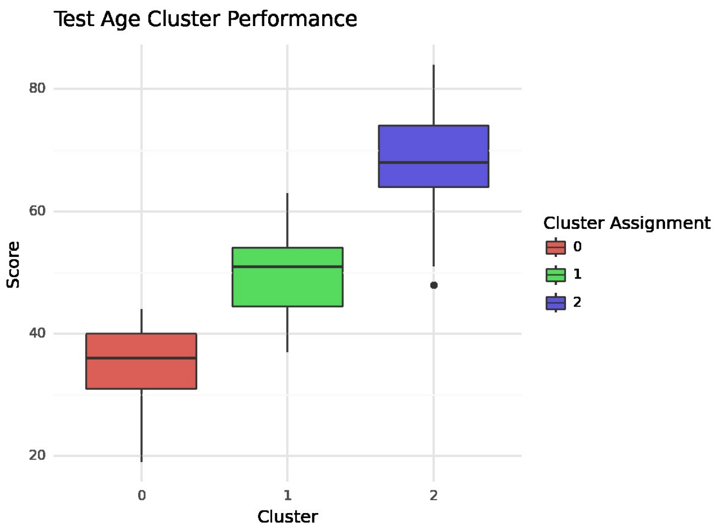

Test Prop_ads_clicked Cluster Performance


Test Longest_read_time Cluster Performance
Score
300
200
100
0
0
1
2
Cluster
Cluster Assignment
0
1
2


Test Length_of_subscription Cluster Performance
Score
1250
1000
750
500
250
0
0
1
2
Cluster
Cluster Assignment
0
1
2


Test Monthly_visits Cluster Performance
Score
50
40
30
20
10
0
1
2
Cluster
Cluster Assignment
🟥 0
🟩 1
🟦 2


Test Current_income Cluster Performance
Score
100
80
60
40
0
1
2
Cluster
Cluster Assignment
🟥 0
🟩 1
🟦 2



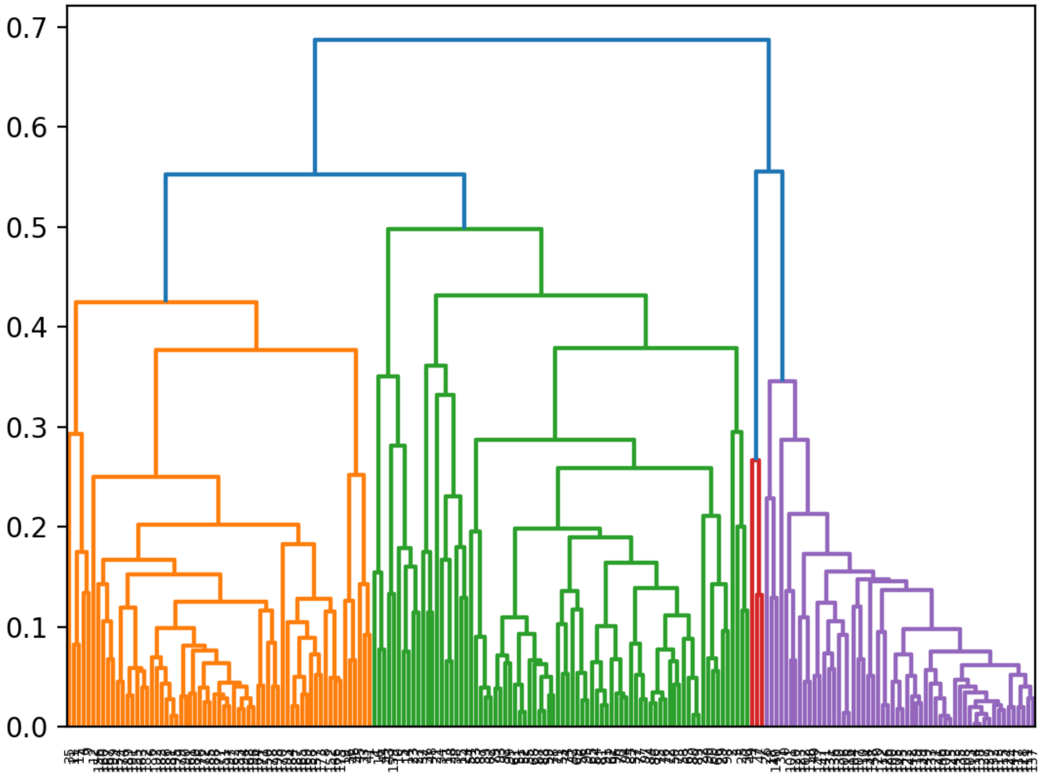
As we can see from the plots above, Cluster 2 always wins! There are some pretty common factors here, as we saw in the first figure. As age increases so does income. When someone has more money and is older, they probably have a lot of free time. This is explained by the longest read time, and time spent reading plots. Cluster 2 is the upper class, cluster one appears to be middle class, and cluster 0 appears to be lower class. What is interesting is the length of subscription time is nearly the same between lower and middle class. You would think lower class would ya know, not pay for subscriptions, but maybe that is why they are in the lower class anways, they pay for too much and do not make enough.

Article Clustering Model

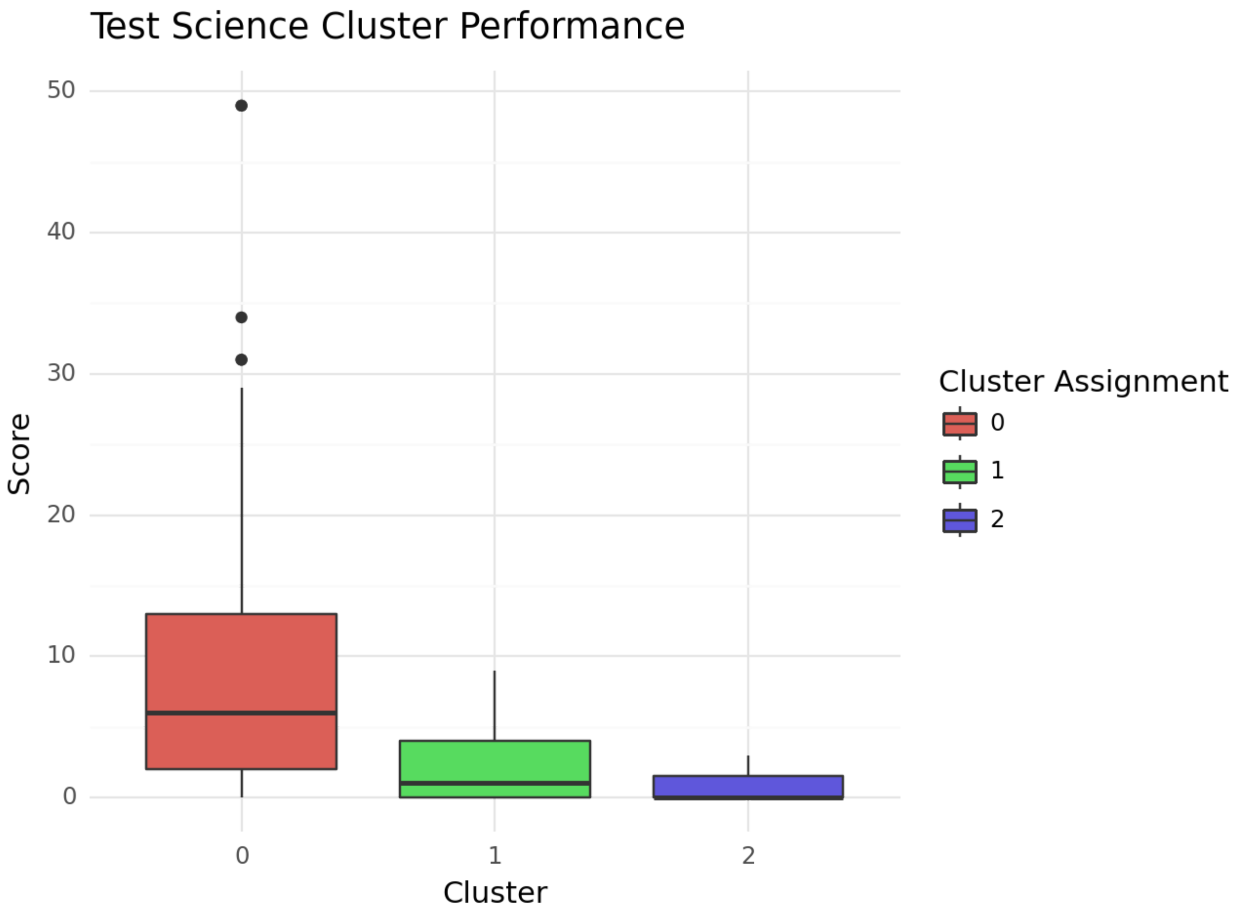
Cluster Summary:

```
          Stocks  Productivity    Fashion  Celebrity  Cryptocurrency  \
cluster
0       6.624113      7.312057   3.212766   2.035461        3.560284
1       5.000000      5.000000   1.000000   5.333333       12.666667
2       1.910714      2.767857  12.607143  17.250000        1.517857

         Science  Technology  SelfHelp   Fitness         AI          id
cluster
0       9.489362   10.262411  8.794326  5.638298  14.836879  101.049645
1       1.000000    3.333333  1.666667  1.000000   1.333333   94.000000
2       2.000000    2.517857  2.892857  1.607143   1.928571   99.464286
```

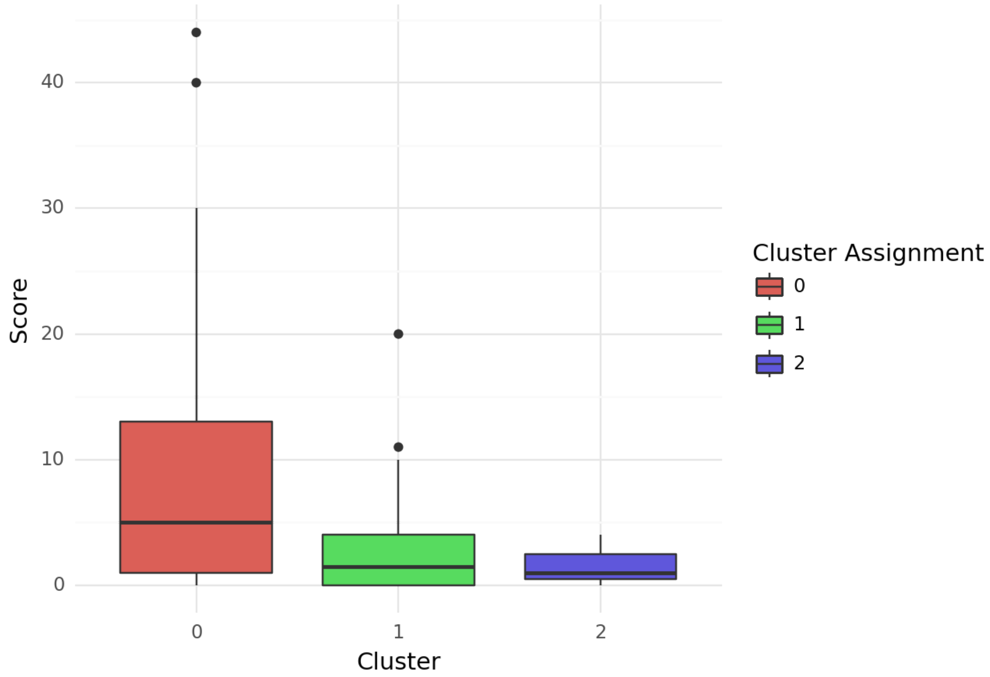
By focusing on this summary table above we see cluster zero likes Science and Technology as well as self help books and FItness. But this cluster really loves AI. Cluster 1 is more focused on crypto currency, stocks, productivity, and celebrities. They seem to be going for the get rich quick method. They are not interesting in science or fitness. Cluster 2 is really into fashion and celebrity magazines. Not so much anything else. All of this info is useful for a company that is trying to segment the market and advertise more directly and efficiently.



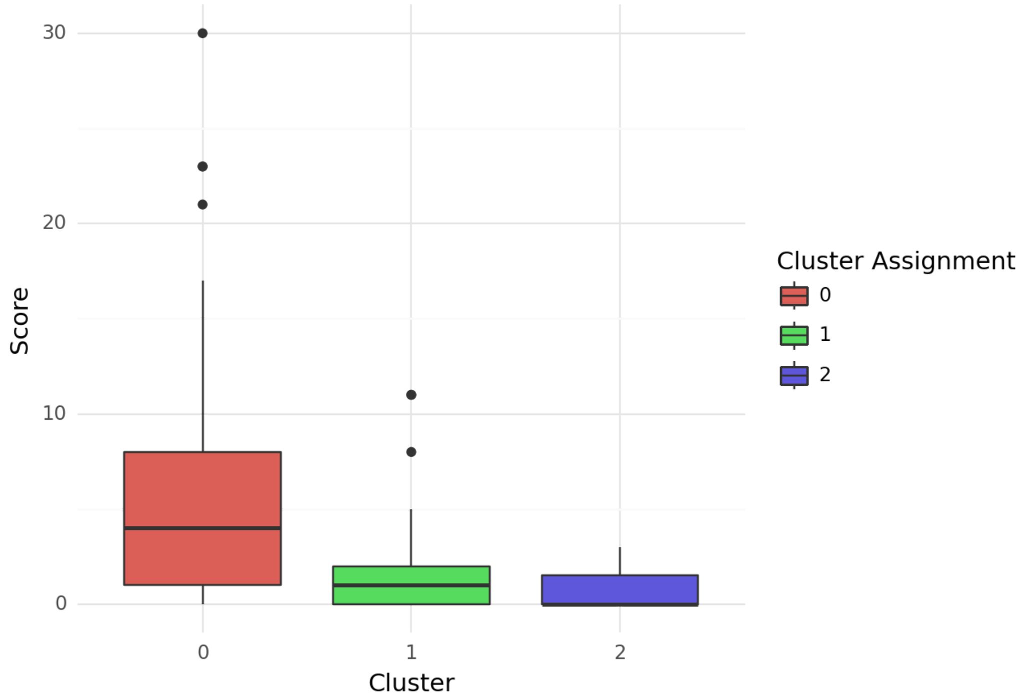
Dendrogram of Article Clustering data, with 3 main clusters.
The dendrogram performed good as they was a good amount of separation from Purple to green. As well as good separation from orange to green. We can see this by the height at which the clusters meet at. Longer vertical lengths means better separation. I chose 3 clusters but due to code error I could not get it to work. I chose 3 because the red cluster is tiny and does not affect much at all, we do not need 4 clusters.

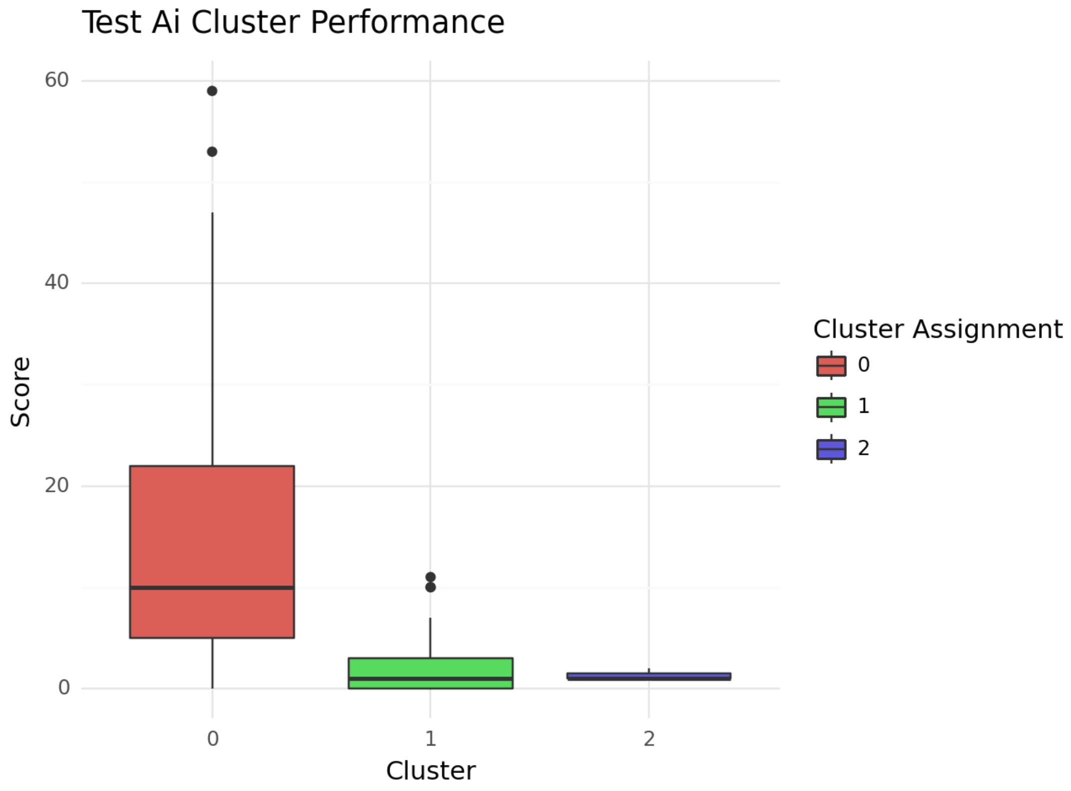

Test Science Cluster Performance
Score
50
40
30
20
10
0
0
1
2
Cluster
Cluster Assignment
0
1
2

Test Selfhelp Cluster Performance

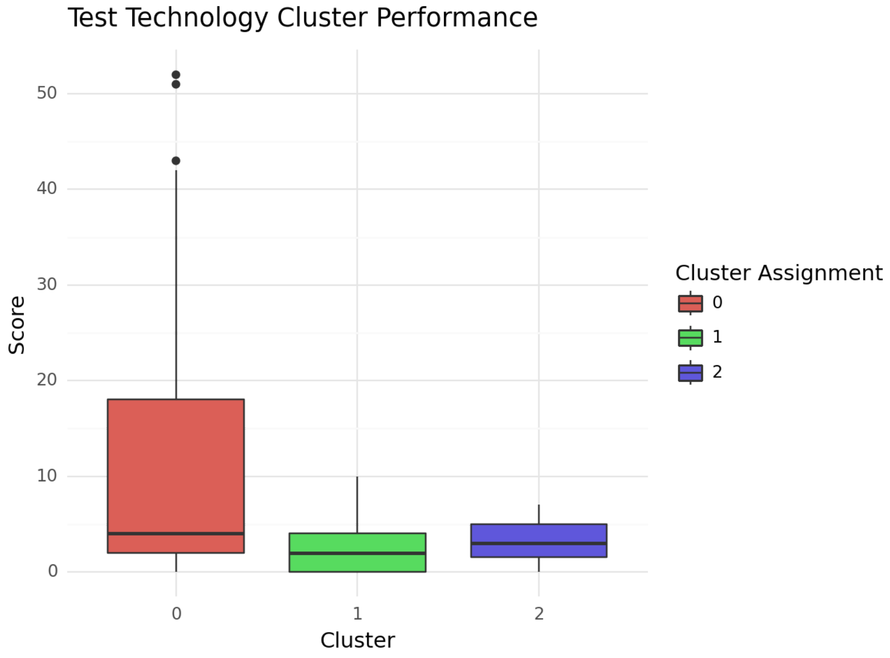


Test Fitness Cluster Performance

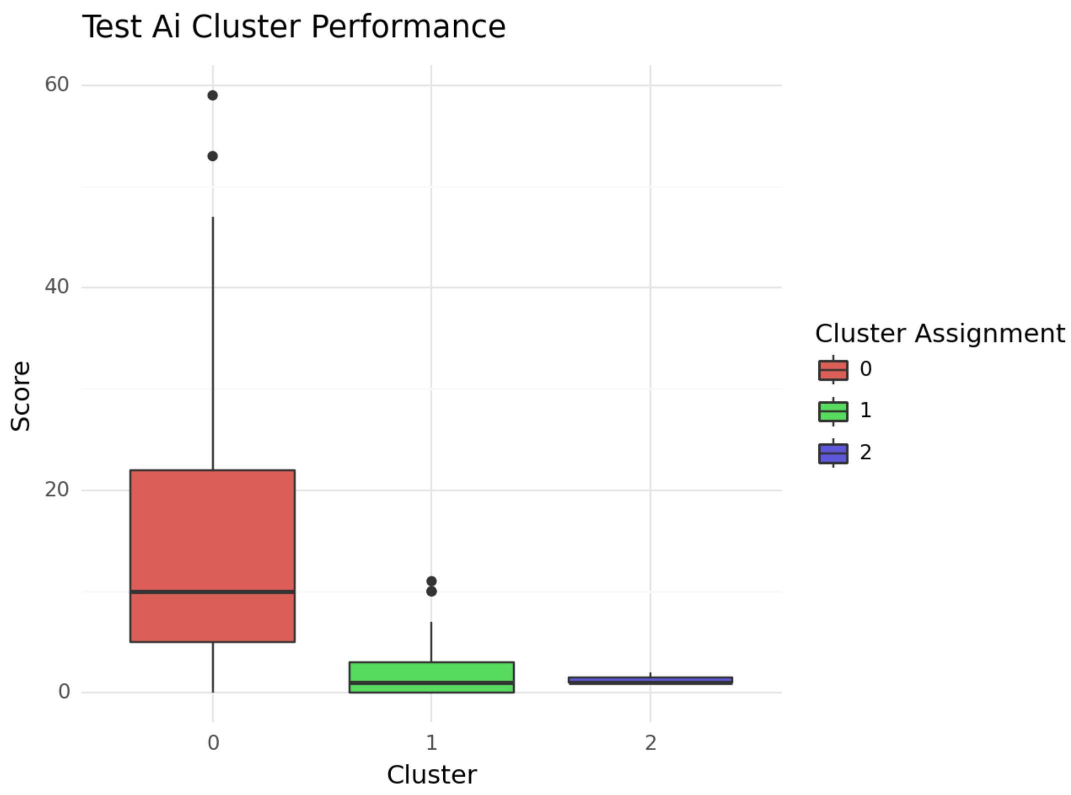




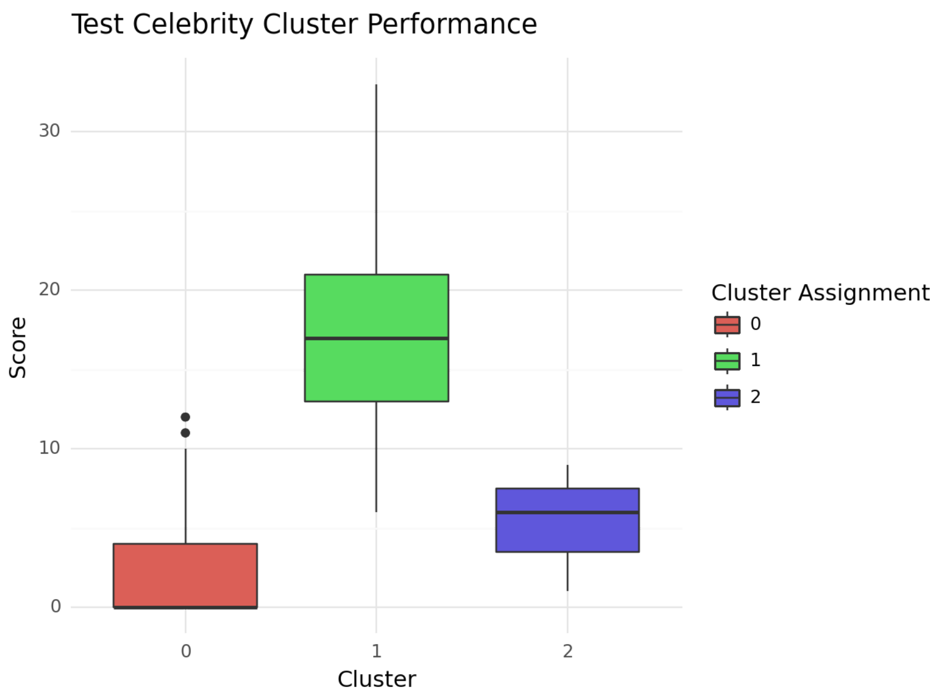
Test Ai Cluster Performance



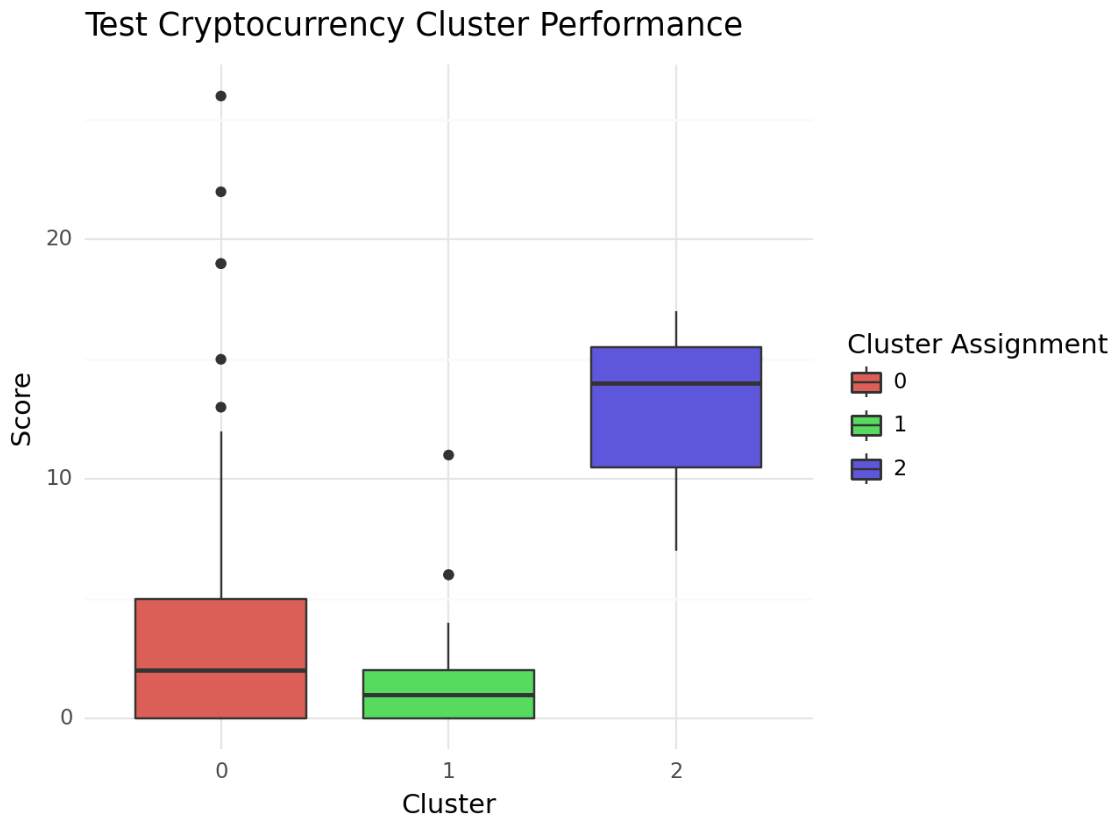
Test Technology Cluster Performance



It is clear that cluster 0 reads content that will make you smarter and better as a human. They read very knowledgeable articles. They also must be into bettering themselves as they read lots of self help and fitness articles.



Cluster 1 is probably mostly girls. They are into fashion and celeberties. Good to know, do not want to push sports magazines or etc onto them.



Cluster 2 gets the crypto nerd award. They are most likely into fast money, and we see that they do not read as extensively as some of the cluster zero members. We know that cluster zero is the most well off of the clusters as they read whats best for the brain. Cluster 2 most likely is looking for short term capital in my opinion.

Discussion/Reflection

I would use GMM next time as some of the clusters were overlapping in the behavioral model. I learned a lot of important things about feature selection. I think it was cool to visualize the data, and see how clusters were different from each other, really different in some cases.