**Web Purchases Clustering**

In order to get a better grasp on the nature of the data file, I decided to do a fundamental EDA and get a look at the headers. The headers read as follows: "No.of.web.sites", “No.of.webpages", "No.of.prod.searches", "No.of.items.looked.at", "No.Items.in.cart", ”Total.Spend" and "Express.ship". Next, I though it would be worthwhile to get a statistical summary of the data complete with each column’s minimum, maximums, means, quartiles and medians. Here is a snapshot of those results:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

These results are quite informative for many reasons. For one, we can see that there is a considerable number of customers who bought nothing (1st quartile for ‘Items in Cart’ and ‘Total Spend’ are 0). Also, one can quickly notice that the popularity for express shipping is very low with a mean of .18 and a maximum of 1. Additionally, we can see that there is substantial variation in the number of webpages a customer will visit with a 1st quartile of 6 and a 3rd quartile of 15.

At this point, it was time to make a determination on how many clusters would be best for my analysis. First, I opted to use the elbow method to get an idea of the optimal number of groups.

A graph of a number of clusters

Description automatically generated

In my view, it seems as though the elbow method suggests that either 3 or 4 is the optimal number of groups. However, we have a few other methods we can try just to get a more holistic picture of what the data is trying to tell us.

Next, I decided to use the silhouette method to further investigate what number of groups is most appropriate.

A graph with a line

Description automatically generated

Here, we are looking for the point in the graph where the average silhouette width is the highest. Interestingly, this method offers a different perspective than the elbow method because this graph seems to be strongly suggesting that the optimal number of groups is 2, as we can clearly see that the average silhouette width drops off significantly after 2 clusters.

Yet, we still have a third option when it comes to choosing our number of clusters: the gap method. Unfortunately, the gap method did not necessarily corroborate the information from either of the previous methods.

A graph of a number of clusters

Description automatically generated

Here, the graph clearly seems to imply that there is *actually a loss* in optimization with two groups, whereas, the silhouette method seemed to suggest that two was optimal and the elbow method seemed to recommend 3-4. In this instance, I think it is ultimately necessary to use a principal component analysis (PCA) to graph out each of these options (2, 3 & 4 clusters) to see what the overlap is like when we graph the clusters visually.

A screenshot of a graph

Description automatically generated

To my surprise, the PCA I ran actually had a considerable amount of overlap across all three groups (2, 3, & 4 clusters). Thus, I though it might be a good idea to see what the same method would look like if we used 5, 6, 7, & 8 clusters.

A screenshot of a graph

Description automatically generated

Clearly, we did not fair any better in our efforts to avoid overlap in adding more groups. There is still a substantial amount of overlap when attempting to group these individuals into groups of more than 5.

After running some more ‘trial and error’ testing with ridiculously high cluster numbers (15, 18 and 25), I have concluded that the groups seem to get even more convoluted if you continue increasing the number of clusters. Thus, though it may sound strange and even lazy, I think the best number of groups is undoubtedly 2 because this is what was suggested by the silhouette method *and* the k2 graph appeared to have the least amount of overlap. I think it could be possible for one to use 3 or 4 clusters, but I don’t think there would be enough homogeneity within groups or difference between groups to draw meaningful conclusions.

Furthermore, I noticed that dimension 1 only accounted for 49% of the variance between the groups, which is not even half. I suppose this means that this group of shoppers just don’t seem to have that many characteristics that differentiates them into groups.

Now, using 2 clusters as my grouping moving forward, I immediately noticed that there were a few categories in which the groups differed tremendously, and a few categories in which they barely differed at all. Here is the output from my console file when I ran the model with only 2 clusters.

A close-up of numbers

Description automatically generated

Notice that there is barely any difference between categories like ‘Express Ship’ and ‘Number of Items in Cart’, but there is demonstrable difference in categories like ‘Number of Webpages’, ‘Number of Items Looked at’, ‘Total Spend’ and ‘Total Time in Seconds’. Given this information and the discussion our class had on clickstream data…I would say that the groups can best be characterized as men and women! In class we discussed that men and women have different shopping habits when it comes to their search methodology. We discussed how women seem to enjoy the experience more, whereas men are more product-focused; seeking out a specific product and not spending as much time in a store or going to as many stores as women overall. Hence, I would call group 1 men (because I recall that we discussed how these preferences are inverted when it comes to online shopping; that is, men prefer browsing more whereas women are more product-focused) and group 2 women because men spend more time browsing online (Total Time in Seconds= 1267.12) and ultimately spend less money (Total Spend=27.55) whereas women spend less time surfing (Total Time in Seconds=337.16) and spend more money (Total Spend=54.24).

When it comes to choosing ads to display to each of the groups, I will start by saying that it seems as though it would be a waste of time, money and effort to try and advertise express shipping to either of these groups because neither group had even a remotely high rate of choosing express shipping. When it comes to ‘value products’, I would say that these advertisements would be most appropriately applied to the males because men tend to spend less overall which means that they likely place high importance on ‘getting the best deal’. I think that the 10% off promotion would reap desirable results with both groups because women are likely to spend higher amounts which means that they know they can *also save* a considerable amount by buying in large quantities. However, since we already established that the men tend to be bargain hunters, it stands to reason that they would also respond well to the ‘10%’ off ads.