Data Mining II — D212

AUTHOR
Tyson Biegler

Part I: Research Question

A1: Can we identify distinct groups of customers based on their tenure and email interactions, using K-means clustering, to better understand churn behavior?

A2: The goal of this analysis is to group customers into distinct clusters based on their tenure and email frequency to identify patterns that could be indicative of the likelihood of churn.

Part II: Technique Justification

B1: I will be using k-means clustering with two continuous variables to group customers into clusters based on their similarity in `Tenure` and `Email`. The k-means algorithm assigns each customer to a cluster based on euclidean distance, ensuring that the customers within the cluster are more similar to each other than to customers in another cluster. The expected outcome is to identify customers with distinct characteristics, such as long time customers with low email engagement or newer customers with high email engagement. These types of clusters could help to identify patterns linked to churn.

B2: One assumption to k-means clustering is that the data is appropriately scaled. K-means clustering is based on euclidean distances and without the proper scale, the contributions from each variable would be less meaningful and accurate. Because the `Tenure` variable has a wider range than `Email`, it is important, in this case, to scale this data.

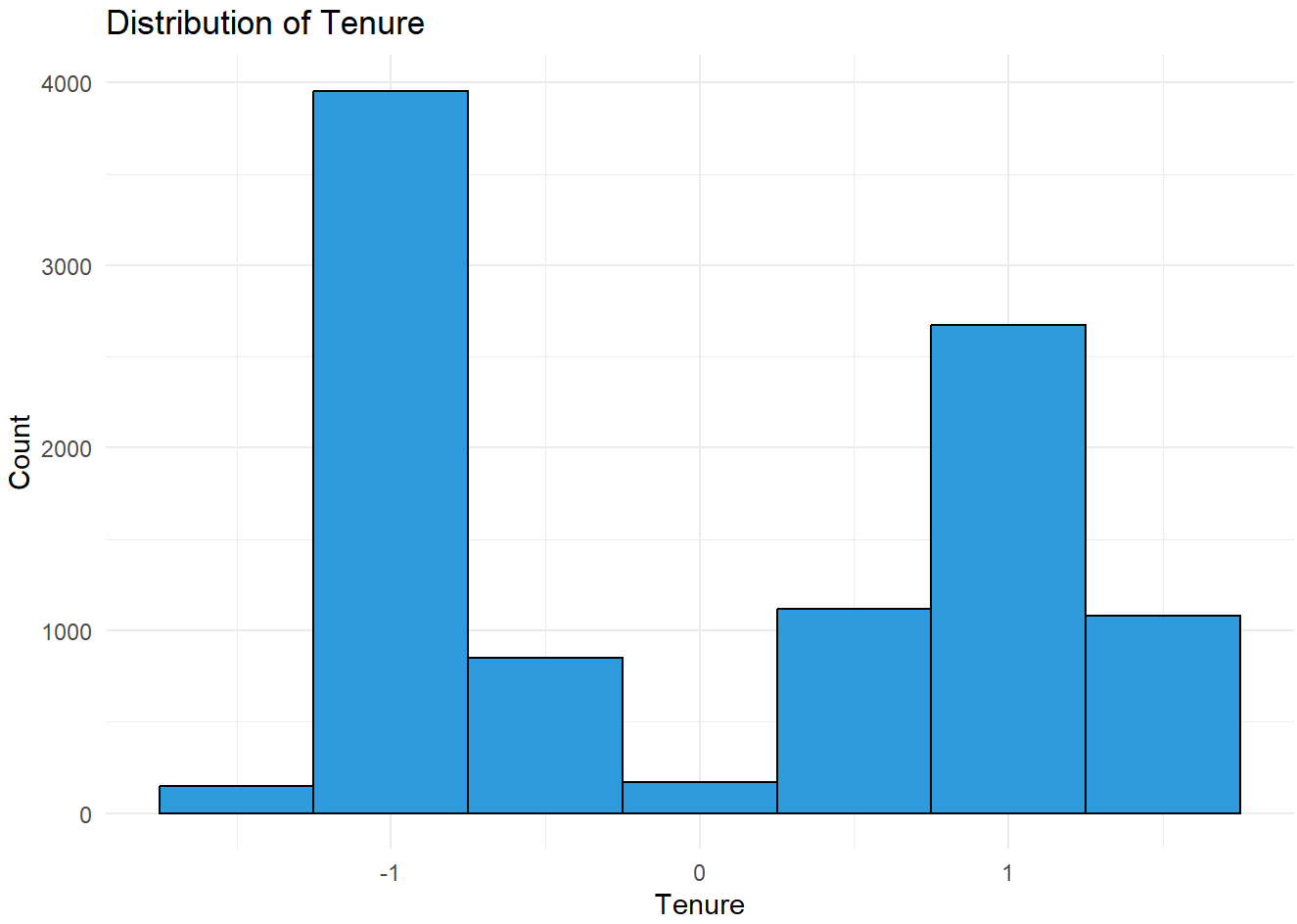
B3: In this analysis, I used `Tidyverse` for basic data manipulation and visualizations. `Cluster` was used for cluster analysis. Specifically I used `silhouette()` to calculate the silhouette score that measures the quality of the k-means clustering. Lastly I used `factoextra()` to visualize the scree plot for finding the optimal k value, plotting the clusters themselves, and for plotting the silhouette scores.

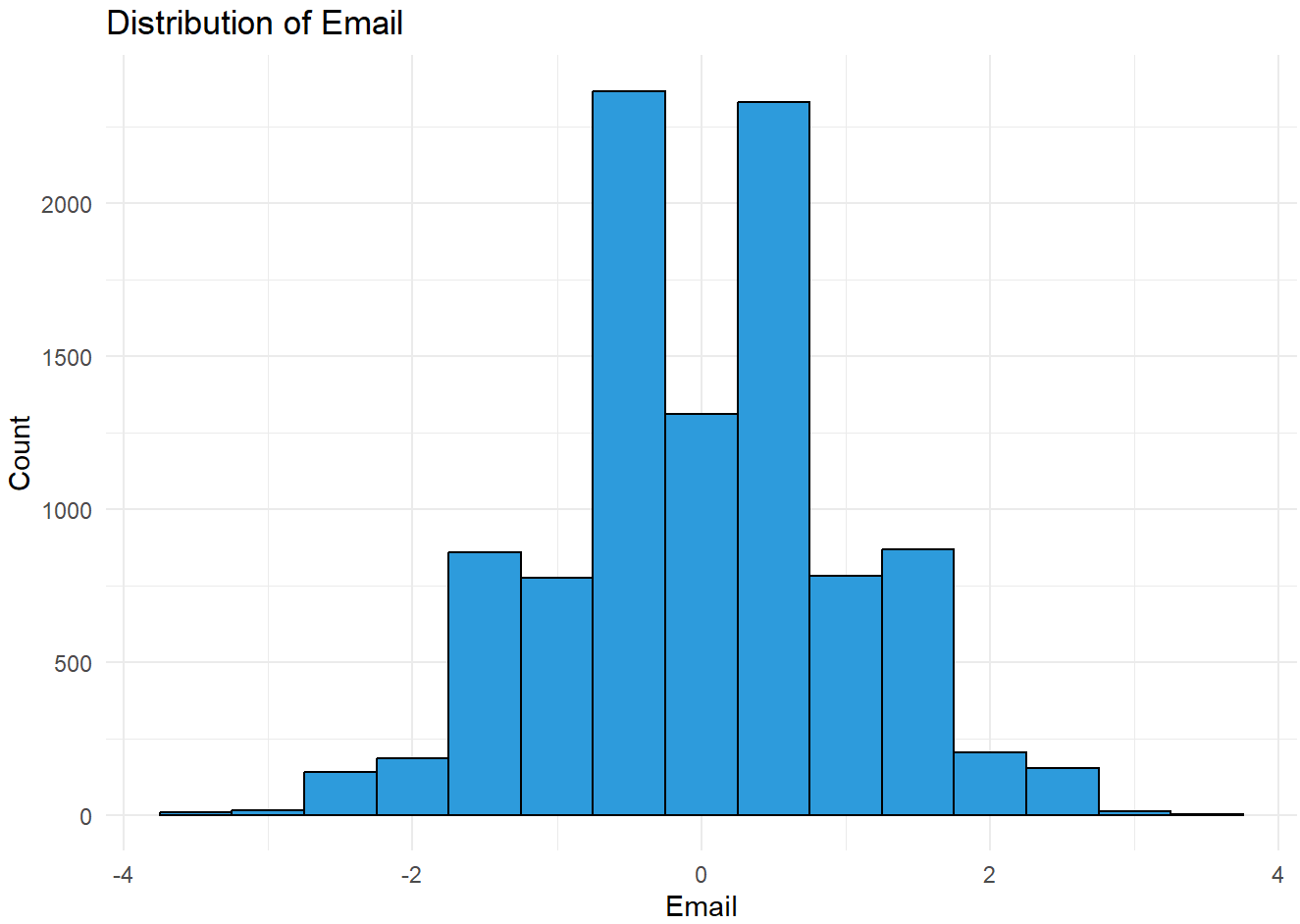
Part III: Data Preparation

C1: One preprocessing goal is to scale the variables so that they have equal weight in the k-means clustering algorithim becuase the range on the vales is vastly different. So it is essential that these variables are properly scaled.

C2: In his D212 webinar, Dr Kamara suggests that two variables is enough for for this assessment (Kamara, 2023). Therefore, I have chosen to investigate `Email` and `Tenure`. Email and Tenure can both help to determine likelihood of churn. Tenure indicates customer stability while email frequency could indicate retention efforts. When analyzing these together with k-means, I can identify the distinct groups and analyze these customers' churn behavior.

C3: I picked two numeric variables, `Tenure` and `Email`, and then I scaled them using `scale()`.


Distribution of Tenure
4000
3000
2000
1000
0
Count
-1
0
1
Tenure




```
# Preparing the data ----------------------------------------------------

churn <- churn[, c("Tenure", "Email")]  #picked only 2 variables (Kamara, 2023)

#scaling the data
churn <- as.data.frame(scale(churn))
```

C4: A copy of the cleaned dataset will be provided in the submission files and is named "***churn_cleaned_data.csv***". Below is a sample of the cleaned dataset. The output shows the standardized values where each variable has a mean of 0 and a standard deviation of 1.

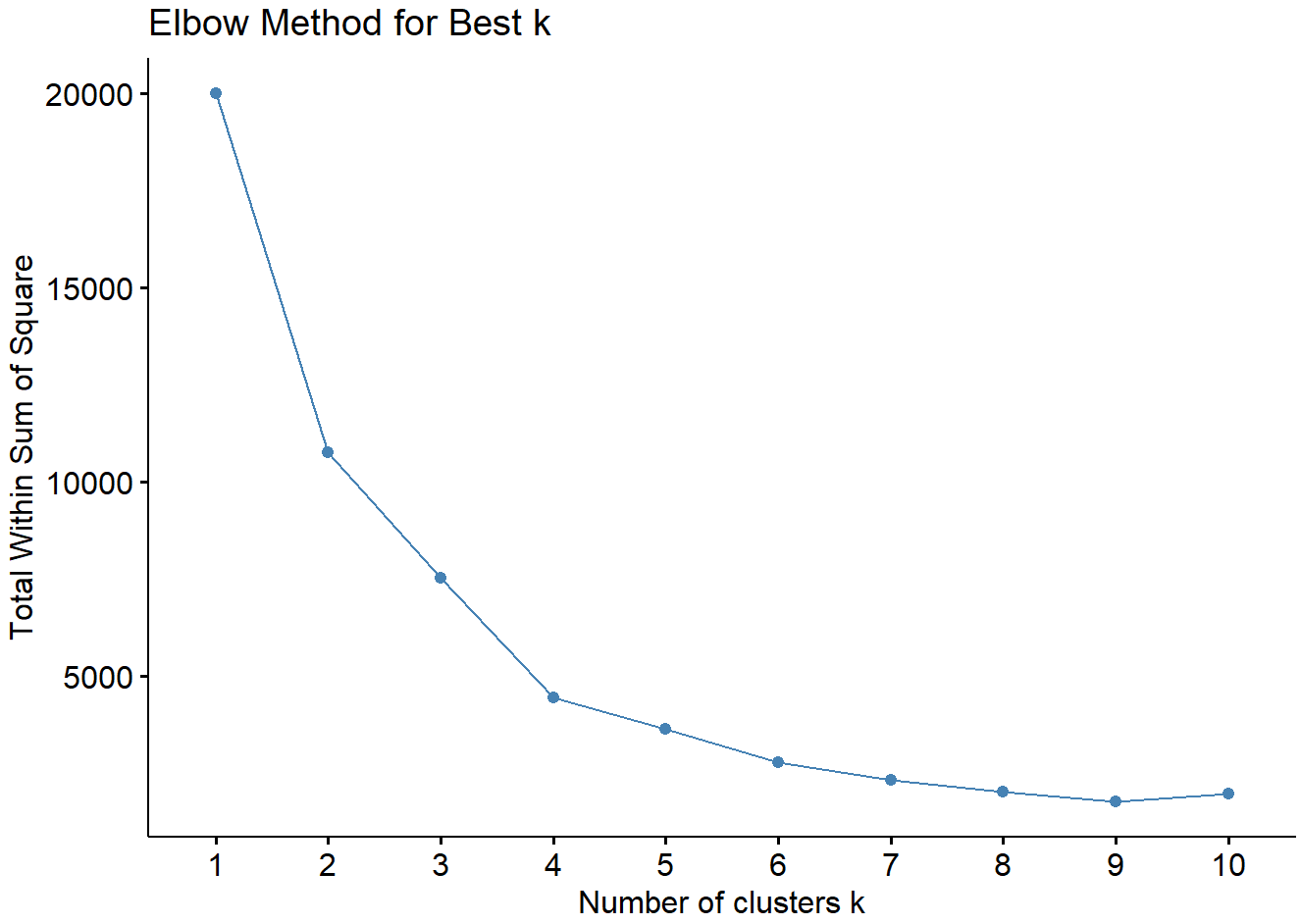
```
      Tenure       Email
1 -1.0486938 -0.666248466
2 -1.2619381 -0.005287686
3 -0.7099043 -0.996728856
4 -0.6594910  0.986153483
5 -1.2424891  1.316633873
6 -1.0409231  0.986153483
```

Part IV: Analysis

D1: To determine the optimal number of clusters, I used the elbow method. The scree plot below plots the total within sum of squares (WSS) and the number of clusters. From the plot it appears that the WSS change

slows significantly after 4 clusters. There is large changes from 1 - 4 but after 4 clusters the WSS appears to essentially level off. Therefore 4 clusters seems to be the optimal k value.

```
#scree plot to find the elbow and  best K value
fviz_nbclust(churn, kmeans, method = "wss") + # SOURCE: (Bobbitt,2022)
  labs(title = "Elbow Method for Best k")
```



D2: The following code performes the k-means clustering with 4 clusters *(centers = 4)* as mentioned in D1 and an nstart of 20. According to an article from Cmith College in 2016, "It is generally recommended to always run K-means clustering with a large value of nstart, such as 20 or 50..." (Smith College, 2016, under 'K-Means Clustering' section). So I decided to use 20 as my nstart value.

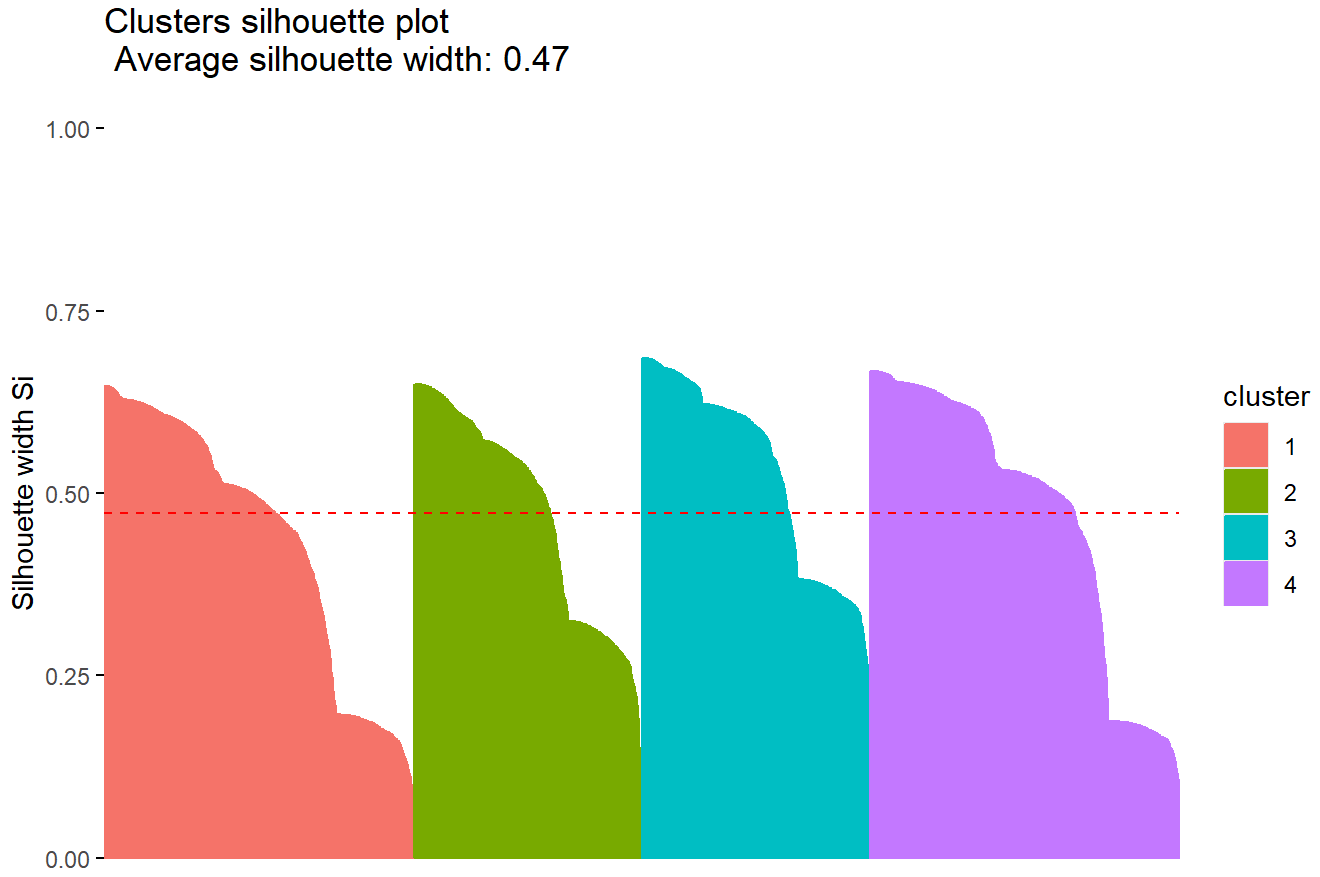
```
km <- kmeans(churn, centers = 4, nstart = 20)
```

Part V: Data Summary and Implications

E1: The quality of the clusters is evaluated by using a silhouette plot, generated withe `fviz_silhouette()`. The average silhouette width is 0.47. The silhouette scores range from -1 (bad) to +1 (good). So a width of 0.47 is slightly to the positive side of the range indicating a somewhat good quality.

```
  cluster size ave.sil.width
1       1 2886          0.44
2       2 2114          0.47
```

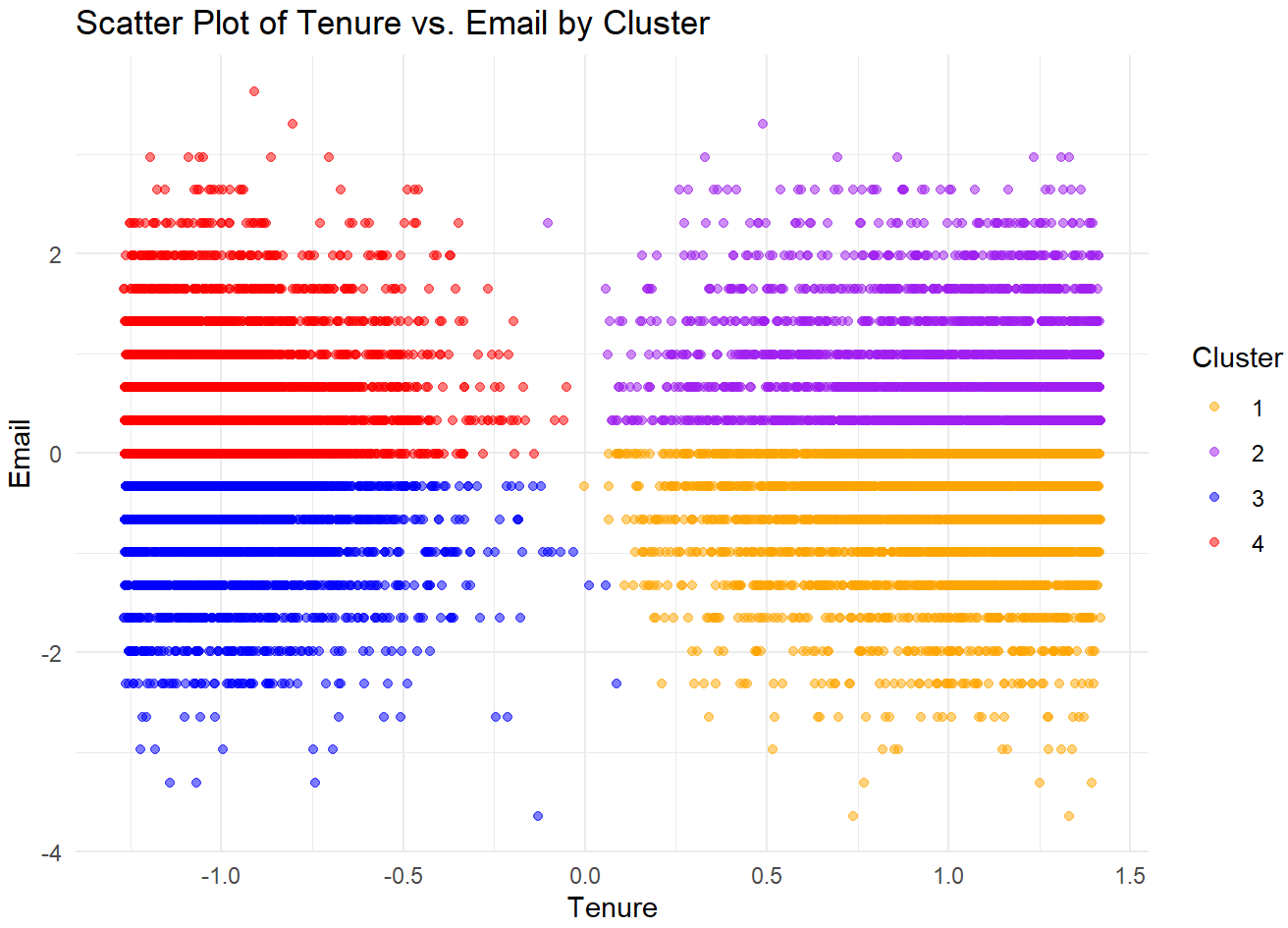
```
3       3 2127          0.53
4       4 2873          0.47
```



E2: This k-means algorithm was able to identify 4 distinct clusters in the data. When I print the cluster centroids it becomes apparent that the clusters are identified as follows:

- Cluster 1: High tenure and low email interactions. These customers have been with the company for a long time and either are happy with their service and require little to no communication, or they are not aware of new offerings which could make them vulnerable to switching to another provider if presented with a better offer.
- Cluster 2: High tenure and high email interactions. These customers have also been customers for a long time but they have high email interactions suggesting that this customer is well aware of products being offered or they just like to stay informed.
- Cluster 3: Low tenure and low email interactions. This customer group profile would suggest that maybe this customer is more prone to churn. They have low tenure and low email communications meaning that maybe they are not fully on-boarded or they haven't yet experienced the value of their services.
- Cluster 4: Low tenure and high email interactions. These customers are might be new customers who are being bombarded with email marketing or maybe on-boarding emails.

```
      Tenure      Email
1  0.9627003 -0.6971666
2  0.9573430  0.9283116
3 -0.9564788 -0.9081657
4 -0.9633643  0.6896069
```



E3: The main limitation to this analysis is that it only takes into account 2 variables and because of this, it might not capture the full complexity of a customer's behavior. Adding more features would help the algorithm cluster the customer into more accurate groups and allow me to make more accurate recommendations.

E4: Based on the customer profile information gleaned from the clusters, I recommend that the company pay close attention to the customers in cluster 1 and cluster 3. Both of these customers have low email communications. Cluster 3 has low tenure and low email communications, so perhaps the company could implement on boarding incentives to boost engagement. In contrast, cluster 1 has high tenure and low email communication. For this cluster the company might look into other forms of communication to keep these customers up to date on what is being offered.

Part VI: Demonstration

F: My panopto video link will be included in my submission files.

G-H: Code Sources:

- Bobbitt. (2022, September 8). *How to use the elbow method in R to find optimal clusters*. https://www.statology.org/elbow-method-in-r/
- Kamara, K. (2023, March 19). *Data mining II - D212 Webinar* [Video]. Western Governors University. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=afbc9be3-7f3a-48ef-a862-afcb0118b043&query=D212
- Smith College. (2016). *10.5.1 K-Means Clustering*. Retrieved from https://www.science.smith.edu/~jcrouser/SDS293/labs/lab16-r.html