Tucker Atwood
WGU MSDA
D212 Task 1 PA
5/25/24

A1. Research Question

Using the provided telecommunications churn data set, I will answer the question, **"Can a k-means clustering method be used to identify groups of telecommunications customers with similar characteristics so that products and marketing campaigns may be better targeted at specific groups of customers?"**

A2. Analysis Goal

The goal of this analysis is to use reported customer characteristics to **identify groups** of customers for whom **different products and campaigns may be targeted at specific groups of customers.** A k-means clustering method using information on customer **income and bandwidth usage** will be applied to determine these groups.

B1. K-Means Clustering Technique

The data mining method referred to as **k-means clustering** will be used to perform this analysis. The process is summarized as follows:

- The number of **centers** is determined prior to clustering. In this analysis, **three centers** will be used; the process used to determine this is described in section D1.
- The variables of interest, Income and Bandwidth, are plotted on a coordinate graph. The k-means algorithm initializes three points with random coordinates, referred to as **cluster centroids**, among the churn data.
- **For each observation** in the churn data, the **Euclidean distance** is calculated from the observation **to each cluster centroid**. This is calculated using the following formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Each observation is **assigned to an initial group** based on its closest cluster centroid.
- The **centers of these initial groups** are calculated using the mean Income and Bandwidth of each group. The **cluster centroids are then moved** to these locations.
- Once again, for each observation, the Euclidean distance to each centroid is calculated. The points are **re-assigned to the new centroid they are closest to**, if necessary.
- The **process continues until no observations are re-assigned** and it is no longer necessary for the centroids to be moved.

The **expected outcome** for this technique is to be able to use the resulting clusters from this process to define groups of customers based on their Income and Bandwidth values. These groups can then be used to inform future business strategies and decisions such as product outreach and marketing.

B2. Clustering Assumption

To perform the k-means clustering method, the technique requires the **assumption** that the **variables of interest have equal variance**. For this analysis, the Income and Bandwidth variables will be **scaled** so that both variables will be expressed with similar centers and spread. This scaling is described in section C1 and ensures the variance assumption has been met.

B3. Programming Environment and Packages Used

This research question will be answered using the programming language R in the RStudio environment. Within R, the following packages will be utilized:

- **purrr**: This package will be used to utilize the "map_dbl" function which will assist in calculating the total within-cluster sum of squares for k-means clustering methods with a varying amount of centers.
- **ggplot2**: This package will be used to create visualizations such as a plot of the total within-cluster sum of squares for various centers, and a scatterplot of the Income and Bandwidth data colored by cluster.
- **plyr** and **dplyr**: These packages will be used to perform data frame manipulation tasks, such as the functions "revalue," "mutate," and "count" which will assist the analysis of clusters created by the k-means method.
- **cluster**: This package will be used to utilize the "silhouette" function which will assist in analyzing the effectiveness of the k-means clustering method.

C1. Preprocessing Goal

To prepare the churn data for analysis and meet the k-means clustering assumption of equal variance, a **data preprocessing goal** has been set to **scale the variables of interest**, Income and Bandwidth, such that they are expressed with similar centers and spread. The **"scale" function** will subtract each value from the variable's mean, then divide by the variable's standard deviation.

This is relevant to the clustering technique because Euclidean distance is used to determine the proximity of observations to the cluster centroids. Since this distance calculation treats x and y values with equal weight, the variables need to be expressed in such a way as to not give inherited extra weight to one variable over the other.

C2. Data Set Variables

The variables to be used for the k-means clustering method are as follows:

- **Income**, a continuous quantitative variable that represents the reported income for each customer in US dollars.

- **Bandwidth_GB_Year**, a continuous quantitative variable that represents the amount of data used by each customer, in gigabytes per year.

C3. Data Preparation

The following steps were taken to prepare the churn data for k-means analysis:

- The variables **Income and Bandwidth_GB_Year were scaled** for the reasons described in section C1.
- The **code** used in this portion of data preparation is provided:

```
churn$Income <- scale(churn$Income)
churn$Bandwidth_GB_Year <- scale(churn$Bandwidth_GB_Year)
# scales necessary variables
```

- The **relevant variables were selected** from the larger churn data set to represent the cleaned data set: Income and Bandwidth_GB_Year. This was done to simplify the process of conducting the k-means analysis.
- The **code** used in this portion of data preparation is provided:

```
churn_cln <- churn %>%
  select('Income','Bandwidth_GB_Year')
# selects variables relevant to k-means clustering
```

C4. Cleaned Data Set

A CSV file containing the cleaned data set, including scaled Income and Bandwidth variables and converted Churn values, has been provided with the file name "churn_cln_212T1.csv."

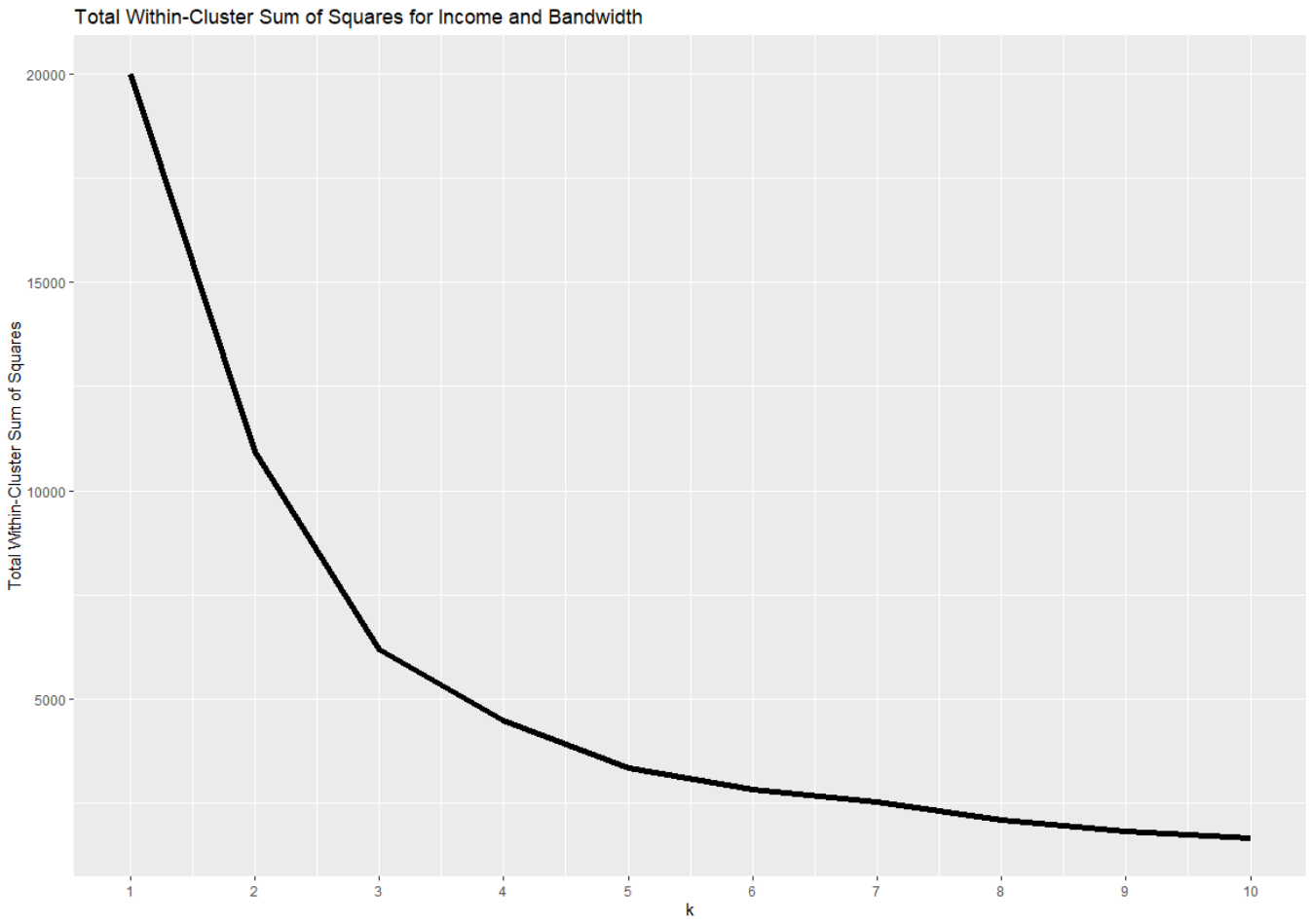
D1. Optimal Number of Clusters

To determine the optimal number of clusters in the k-means clustering method, the **total within-cluster sum of squares (WCSS)** metric was calculated and compared for values of k from 1 to 10. The WCSS is calculated by measuring the Euclidean distance from each observation to its cluster centroid, squaring this value, and calculating the sum of all these values for all observations.

The WCSS naturally decreases when k increases, so it is not enough to simply minimize the value. Rather, a plot of the WCSS values by number of centers is visualized and the **elbow method** is used to locate the optimal number of centers. This method identifies this optimal value based on where the WCSS's decreasing pattern begins to slow down. This point represents the "elbow" in

the plot and can be used to show where any further increases in k do not provide enough added cluster accuracy to counteract the associated lower precision.

The following visualization plots the number of centers, k, against the calculated WCSS for the scaled Income and Bandwidth variables in the churn data set. Using the elbow method, it was determined that the **optimal number of centers is three.**



D2. Clustering Analysis Code

The following code was used to **determine the optimal number of centers** as described in section D1:

```
set.seed(8)
# sets seed at specific value to ensure consistency in k-means analysis which uses random number generations

tot_withinss <- map_dbl(1:10, function(k){
```

```
  model <- kmeans(churn_cln[,1:2], centers = k)
  model$tot.withinss})
# calculate total within-cluster sum of squares for k-means analyses from 1 to 10 centers

elbow <- data.frame(
  k = 1:10,
  tot_withinss = tot_withinss)
# creates data frame to be used to create plot of total within-cluster sum of squares by number of centers

ggplot(elbow, aes(x = k, y = tot_withinss)) + geom_line(linewidth = 2) +
scale_x_continuous(breaks = 1:10) +
  labs(title = "Total Within-Cluster Sum of Squares for Income and Bandwidth") +
  ylab("Total Within-Cluster Sum of Squares")
# creates plot of total within-cluster sum of squares by number of centers
```

The following code was used to **conduct the k-means clustering method** as described in section B1:

```
set.seed(8)
# sets seed at specific value to ensure consistency in k-means analysis which uses random number generations

kmodel <- kmeans(churn_cln[,1:2], centers = 3)
# creates k-means model with number of centers determined by elbow method
```

The following code was used to **analyze the results** of the k-means clustering method, which is discussed further in sections E1 and E2:

```
churn_clust <- mutate(churn_cln, cluster = kmodel$cluster)
# creates new data frame with cluster designations appended to churn data used to generate k-means model

count(churn_clust, cluster)
# provides number of data points in each cluster

churn_clust %>%
  group_by(cluster) %>%
  summarize_all(list(mean))
# provides mean of relevant variables for k-means model and analysis
```

```
ggplot(churn_clust, aes(x=Income, y=Bandwidth_GB_Year, color = factor(cluster))) +
geom_point(size = 2) +
  labs(title = "Income and Bandwidth, Scaled and Clustered") +
  ylab("Bandwidth GB Per Year")
# creates scatterplot of relevant variables, colored by cluster

SIL <- silhouette(churn_clust$cluster, dist(churn_clust[,1:2]))
summary(SIL)
# calculates average silhouette width of data points in k-means model
```

E1. Cluster Quality Analysis

To analyze the quality of the clusters created by the k-means model, the **average silhouette width** of all observations was calculated. The silhouette width metric compares the distance from each observation to other observations within its cluster and the distance to observations within the nearest neighbor cluster. The calculation produces a value between -1 and 1 for each observation. A value close to 1 indicates the observation fits well within its cluster; a value close to 0 indicates the observation may fit equally well in its cluster or the neighboring cluster; and a value closer to -1 indicates the observation fits better in the neighboring cluster. These same interpretations of silhouette width values can be applied to aggregates such as means and medians, which can also be calculated for each individual cluster.

A **summary of the silhouette width statistics** for the observations of Income and Bandwidth as clustered using the k-means model described in this analysis is provided:

```
 Cluster sizes and average silhouette widths:
     1556      4213      4231
0.2349064 0.5689406 0.5971061
Individual silhouette widths:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1846  0.4239  0.6019  0.5289  0.6796  0.7296
```

The overall **average silhouette width for the observations in this analysis was 0.5289**. The **median silhouette width was 0.6019.** Since these values are positive and roughly halfway between 0 and 1, this indicates that the **created clusters are of adequate quality.** However, a noticeable difference in average silhouette width between each cluster is apparent: Cluster 1 has a significantly lower value (0.235) than Cluster 2 (0.569) and Cluster 3 (0.597). The minimum value also indicates at least one observation has a negative silhouette width. While the measures of center for each cluster are all positive and the quartiles imply that there are significantly more positive values than negatives, the differences between clusters and the presence of negative values indicate further analysis may be necessary.

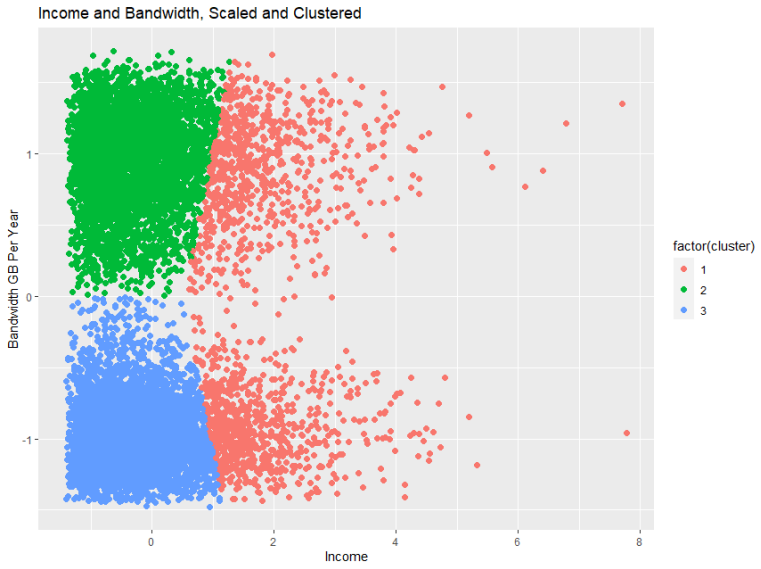
E2. Results and Implications

Using a k-means clustering method, the churn data set was split into three distinct groups based on Income and Bandwidth values. Relevant summary statistics for each cluster are provided as follows:

cluster	n
1	1556
2	4213
3	4231

	cluster	Income	Bandwidth_GB_Year
	<int>	<dbl>	<dbl>
1	1	1.80	-0.0159
2	2	-0.330	0.967
3	3	-0.333	-0.957

A scatterplot of the Income and Bandwidth data is provided as follows, with observations colored by their cluster value (1, 2, or 3):



These clusters, defined by the k-means clustering method with three cluster centroids, can be described as follows:

- **Cluster 1** represents a group of customers with relatively high income and any amount of bandwidth used. This group is represented by the pink cluster of observations in the scatterplot; all customers with an income past a certain threshold will fall into this cluster.
- **Cluster 2** represents a group of customers with relatively low income and a relatively high amount of bandwidth used. This group is represented by the green cluster of observations in the scatterplot.
- **Cluster 3** represents a group of customers with relatively low income and a relatively low amount of bandwidth used. This group is represented by the blue cluster of observations in the scatterplot.

The implications of these results are as follows:

- With customers split into groups based on their income and bandwidth characteristics, business strategies and decisions may be made to **target specific groups of customers** for product sales and marketing campaigns, potentially leading to better company performance and profits.

E3. Limitation

The following is a limitation to the k-means cluster process and analysis which must be kept in mind before using the analysis to draw conclusions or implement future plans:

- The average silhouette width for observations in Cluster 1 is noticeably lower than the corresponding values for Clusters 2 and 3, indicating a lower quality of that cluster than the others. The shape of Cluster 1 in the scatterplot is also potentially problematic: the cluster includes all observations with a relatively high income regardless of bandwidth, and the borders between Cluster 1 and each of the neighboring clusters blend into each other with no clearly defined separation from one cluster to the other. These issues may indicate that the number of clusters could be changed to achieve better results, or that k-means clustering was not the best method of analysis for these variables. Further research should be conducted to confirm or deny whether a different method of analysis provides better quality information.

E4. Recommendations

The following courses of action are recommended based on the results of this analysis:

- The clusters determined in this analysis can be used to better understand customers and their relevant characteristics. Further research should be conducted to determine products and services that align with the clusters of customers formed in this analysis.
    - If a product or marketing campaign aligns with customers who have **relatively high income, Cluster 1** should be used as a target for these strategies.
    - For customers with **relatively low income**, products and campaigns that have been determined to be successful for customers with **relatively high bandwidth usage should be targeted at Cluster 2**.
    - For customers with **relatively low income and relatively low bandwidth usage**, products and campaigns that align with these characteristics should be **targeted at Cluster 3**.
- Therefore, **separating customers into clusters** based on their similarities in income and bandwidth usage helps **inform future business strategies and decisions**. This will allow the telecommunications company to be **more successful and more profitable** in the long term as its products and services will be better aligned with its customers.

F. Panopto Video

A link to a Panopto video has been provided which shows the execution of the code used to prepare for and conduct the described k-means clustering process.

G. Third-Party Code References

WGU Courseware was used as a resource to learn the methods, concepts, and functions used to create the codes in this project, including DataCamp course tracks (datacamp.com) and Dr. Kesselly Kamara's D212 Panopto videos. There are no codes that have been taken directly from any other resources.

H. Content References

There is no content in this analysis that has been quoted, paraphrased, summarized, or otherwise requires direct citation.