Part 1: Research Question

A1) Research Question

The question for this analysis is if we can use the K-means clustering technique to achieve our goal of segmenting the data into clusters of related variables. I am mainly looking at the variables of tenure and bandwidth usage to see how they are related in the clusters.

A2) Define Goal

Our goal of the analysis is to segment the data to find distinct groups of related variables within the data set.

Part 2: Technique Justification

B1) Technique Explanation

For this analysis I have chosen to use the K-means clustering technique. The K-means algorithm analyzes data by starting with an initial group of randomly selected centroids, which are then used as starting points for every cluster. (1) Each point of data is assigned to a cluster based on distance. After all data points have been assigned, the center of each cluster is recalculated by taking the mean of all the data points assigned to that cluster. The algorithm then performs iterative calculations to optimize the starting positions of the centroids. This will continue until the defined number of iterations has been achieved. Once the iteration stops, you have the final clusters. (1) My expected outcome of this analysis is to hopefully gain some insight into commonalities between groups of customers for targeting strategies.

B2) Summary of Assumption

One assumption of k-means clustering is that the variance of the distributions of the attributes is spherical. (2)

B3) Packages and Libraries

|  |  |
| --- | --- |
| Pandas | Used to import my data into a DataFrame for manipulations |
| Numpy | Used to provide arrays for any required calculations |
| Matplotlib.pyplot  Seaborn | Used for making graphs and visualizations |
| Sklearn.cluster import KMeans | Enables use of the K-means clustering technique |
| Sklearn.metrics import silhouete\_score | Used to get an accurate measure of how good the clusters are |
| Sklearn.preprocessing import robustscaler | Used to scale the data for better clustering accuracy |
| import warnings  warnings.filterwarnings('ignore') | Used to hide warnings given in the python enviroment |

Part 3: Data Preparation

C1) Preprocessing Goal

The preprocessing step I will be taking for this analysis is scaling the variables using the robustscaler() method.

C2) Variable Identification

The variables I will be using for this analysis are:

* Tenure (Continuous)
* Bandwidth\_GB\_Year (Continuous)

C3) Steps for Analysis and Code

Step 1 – Load dataset

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

*import pandas as pd*

*import numpy as np*

*import matplotlib.pyplot as plt*

*import seaborn as sns*

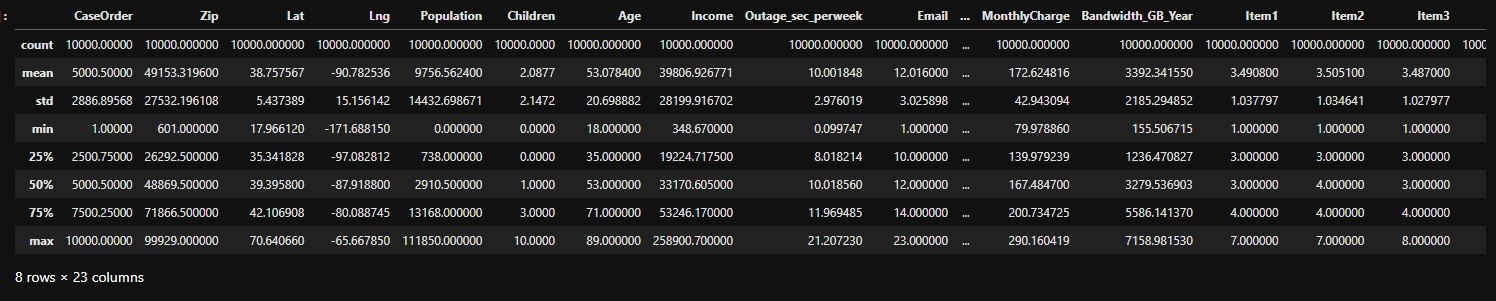
*from sklearn.cluster import KMeans*

*from sklearn.metrics import silhouette\_score*

*from sklearn.preprocessing import RobustScaler*

*df = pd.read\_csv('churn\_clean.csv')*

*df.describe()*

**

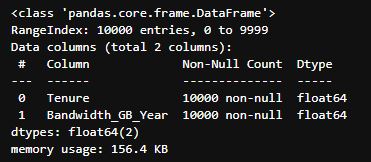
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Step 2 – Select the two variables I’ll be looking at, check for null values and describe the data

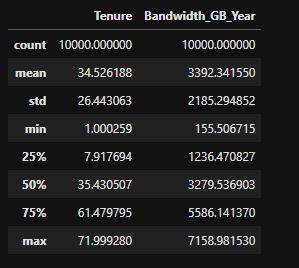
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

*X = df[['Tenure', 'Bandwidth\_GB\_Year']].copy()*

*X.info()*



*X.describe()*



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

*Step 3 – Scale the data then save as prepared csv file*

*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\**

*robust = RobustScaler()*

*scaled\_var = robust.fit\_transform(X)*

*scaled\_df = pd.DataFrame(scaled\_var, columns = ['Tenure', 'Bandwidth\_GB\_Year'])*

*scaled\_df.to\_csv('churn\_prepared.csv', index = False)*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Part 4: Analysis

D1) Optimal Clusters

The optimal number of clusters for this analysis was 2 clusters. In order to determine this number, I created an elbow plot (which is shown below) and saw that there is a very sharp decline at 2, then it goes fairly smoothly from there. This shows that the best number to use for clusters was 2.

D2) Code

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

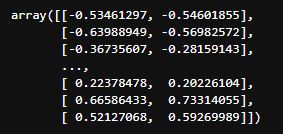
*df = pd.read\_csv('churn\_prepared.csv')*

*df.head()*

*df\_cluster = df[['Tenure', 'Bandwidth\_GB\_Year']]*

*df = df\_cluster.iloc[:,:].values*

*df*



*inertia = []*

*for k in range(1,11):*

*kmeans = KMeans(n\_clusters = k, init = "k-means++")*

*kmeans.fit(df)*

*inertia.append(kmeans.inertia\_)*

*plt.figure(figsize = (12, 6))*

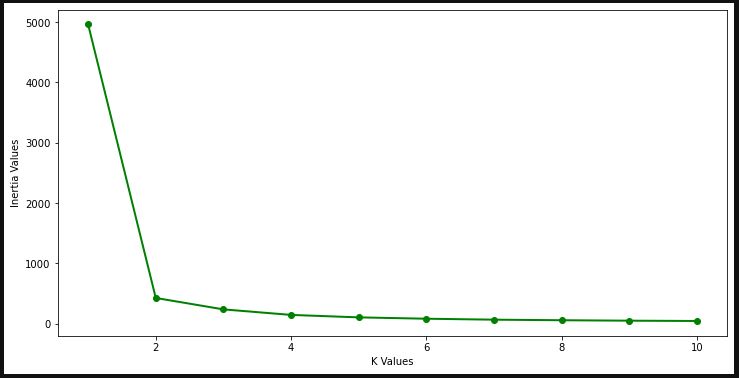
*plt.plot(range(1,11), inertia, linewidth = 2, color = "green", marker = "o")*

*plt.xlabel("K Values")*

*plt.ylabel("Inertia Values")*

*plt.show()*

*#plt.savefig('elbow.jpg')*



*km = KMeans(n\_clusters = 2)*

*kmeans = km.fit\_predict(df)*

*plt.figure(figsize = (12,12))*

*plt.scatter(df[:, 0], df[:,1], c = kmeans, cmap = 'Spectral')*

*plt.scatter(km.cluster\_centers\_[:, 0], km.cluster\_centers\_[:, 1], marker = 'X', color = 'black')*

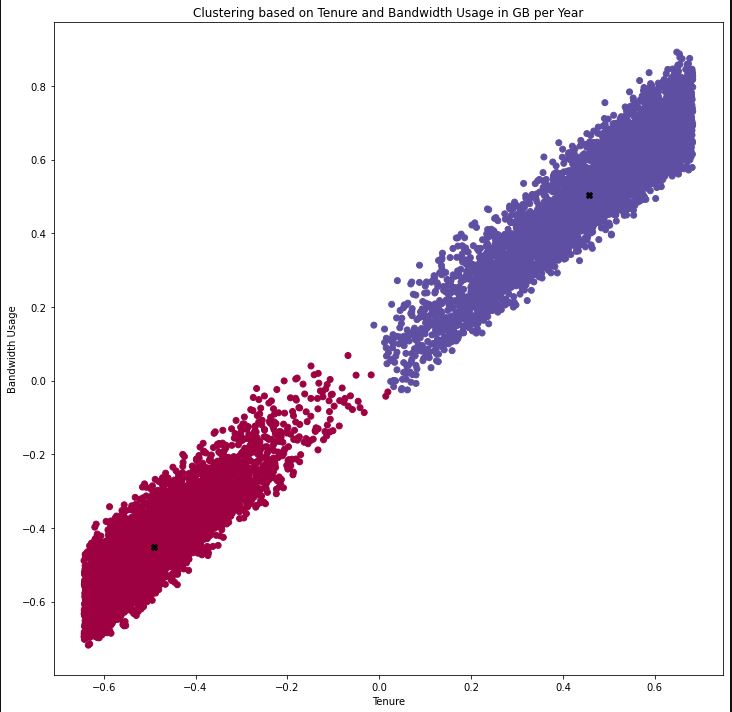
*plt.xlabel('Tenure')*

*plt.ylabel('Bandwidth Usage')*

*plt.title('Clustering based on Tenure and Bandwidth Usage in GB per Year')*

*plt.show()*

*#plt.savefig('cluster.jpg'*)



*score = silhouette\_score(df, kmeans)*

*print(f"Silhouette Score: {score}")*



Part 5: Data Summary and Implications

E1) Quality of Clusters

After running K-means clustering, I ran the silhouette score and got 0.81, which seems to suggest that the quality of the clusters is good. Looking at the scatterplot itself, they seem evenly dispersed between the two clusters, going up like a line graph.

E2) Results and Implications

Given the silhouette score I think we can determine that the analysis seems to be decently strong. Looking at the scatter plot we see that the two clusters are separate but the tail ends seemingly moving towards each other. Visually, this suggests there could be a more linear relationship between them, with higher tenure related to higher usage. The centroids also appear to be more towards the back of each cluster with the data points in a tight group, getting more stretched out towards the other side. I feel this also gives strong implications of a linear relationship. While the top cluster is larger, it does appear to be a mirror image of the bottom cluster, yet they are oval shaped rather than the spherical shape that K-means should produce.

E3) Limitation of Analysis

One assumption of K-means is that the clusters are spherical and equally sized, and the implications of my analysis listed above show a more linear relationship. So already one limitation is the fact my clusters are more of an oblong shape.

Given the above, another limitation of my analysis is that if the data does have a linear relationship, then K-means clustering might not be the most efficient algorithm to use for this specific instance.

E4) Recommendation

Since tenure and bandwidth usage are clustered in the manner they are, suggesting a linear relationship between them, a good course of action for the company would be to do further analysis into the reason why customers with higher tenure are using more data. Knowing how the company operates would also provide more insight into how the data came out the way it did. If the company has a cap on usage and longer tenured customers have a higher cap could explain it. We could also do further analysis using the Monthly Charge variable as well to determine if longer tenured customers get discounts.

F) Panopto Video

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=cca109ce-24ab-4a75-8465-b06e0141808b>

G) Web sources

1) Ecosystem, E. (2022, May 17). Understanding K-Means clustering in machine learning. *Medium*. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

2) *K-means clustering is not a free lunch*. (2015, January 16). Variance Explained. http://varianceexplained.org/r/kmeans-free-lunch/