Jeffery Bills

D212Task1

Churn Clustering Techniques

Scenario 1

One of the most critical factors in customer relationship management that directly affects a company’s long-term profitability is understanding the customers. When a company understands its customers’ characteristics, it is better able to target products and marketing campaigns for customers, resulting in better profits for the company in the long term.

You are an analyst for a telecommunications company that wants to better understand the characteristics of its customers. You have been asked to use clustering techniques to analyze customer data to identify groups of customers with similar characteristics, ultimately enabling better business and strategic decision-making.

Part I: Research Question

A. Describe the purpose of this data mining report by doing the following:

1. Propose one question relevant to a real-world organizational situation that you will answer using one of the following clustering techniques:

• k-means

• hierarchical

The question I am investigating is does k means clustering offer any significant insights to this data? This can be investigated by charting the within-cluster sum of squares against the number of clusters and locating the point when the rate of reduction changes drastically, the elbow approach assists in determining the ideal number of clusters. This point shows the ideal number of clusters to employ. Second, by computing the silhouette coefficient for every data point, the silhouette approach assesses the quality of the clusters. In comparison to other clusters, this coefficient indicates how similar a point is to its own cluster. We can determine the degree of cluster definition by averaging these scores. Better-defined clusters are indicated by an increased average silhouette score.

2. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

The goal of this analysis is to identify the optimal number of clusters using inertia and evaluate their meaningfulness with the silhouette score. First, we determine the optimal number of clusters by plotting inertia, which measures the within-cluster sum of squares, and identifying the "elbow" point where adding more clusters no longer significantly reduces inertia. This point indicates the best number of clusters. Then, we calculate the silhouette score to assess cluster quality. The silhouette score measures how similar each data point is to its own cluster compared to others, with higher scores indicating well-defined clusters. This approach ensures the clusters are both optimal in number and meaningful.

Part II: Technique Justification

B. Explain the reasons for your chosen clustering technique from part A1 by doing the following:

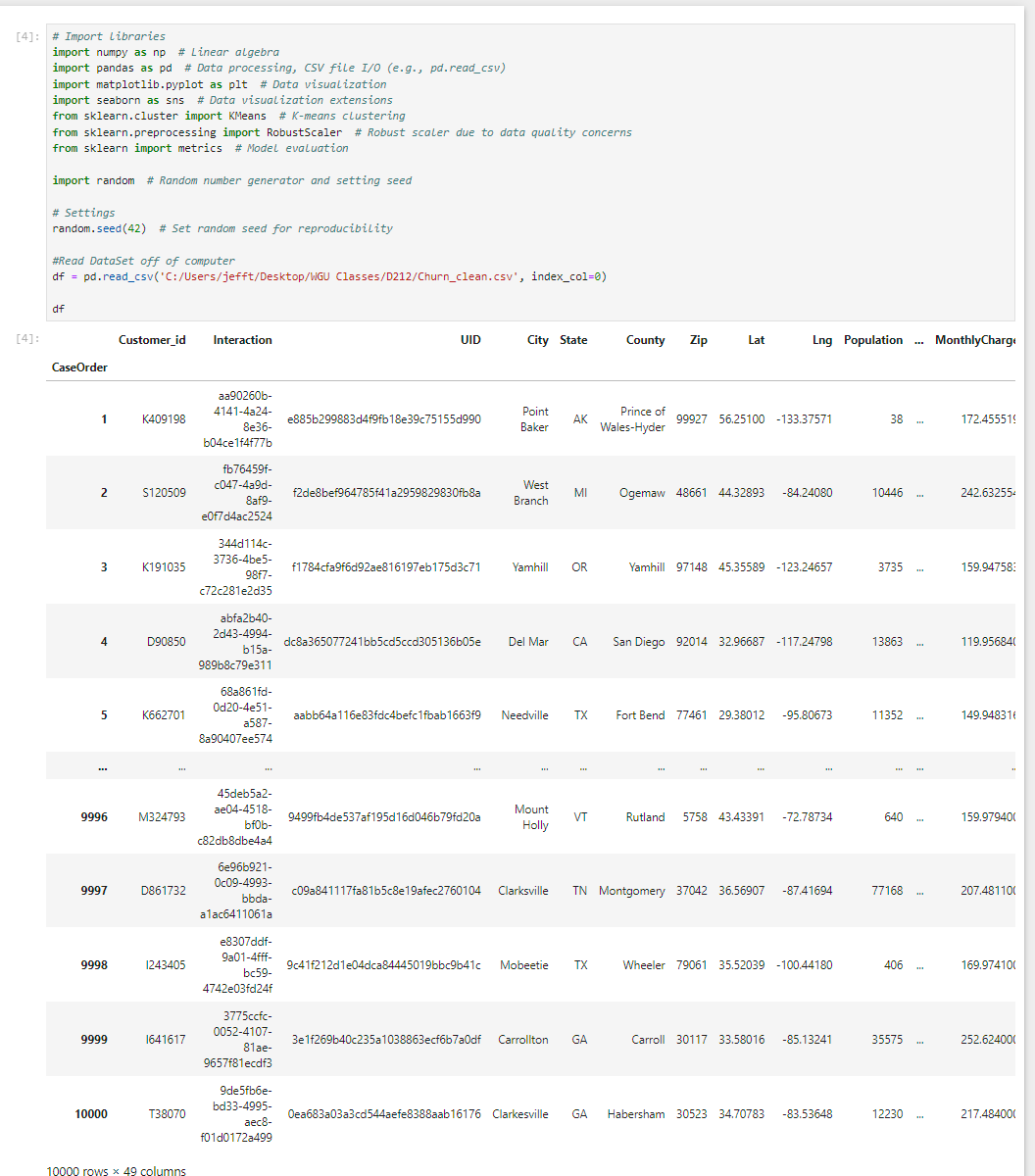
1. Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes.

K-means clustering is an unsupervised machine learning method used to group data points that are related. The procedure begins with randomly assigning data points to clusters and calculating each cluster's centroid. Each data point is then reassigned to the cluster with the closest centroid. This process is repeated until the centroids stop changing, indicating convergence. The final number of clusters is determined by the implementer's assessment.

1. Summarize one assumption of the clustering technique.

K-means clustering assumes that clusters are spherical and about the same size. This means it expects data points within each cluster to be evenly spread around the center and for all clusters to have similar spreads. If the actual clusters have different shapes or densities, this assumption can affect how well the method works.

1. List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

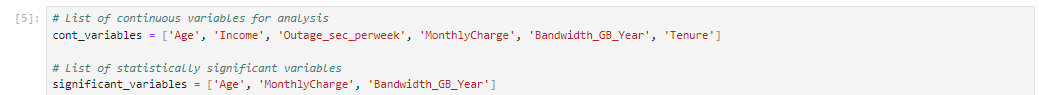


C. Perform data preparation for the chosen dataset by doing the following

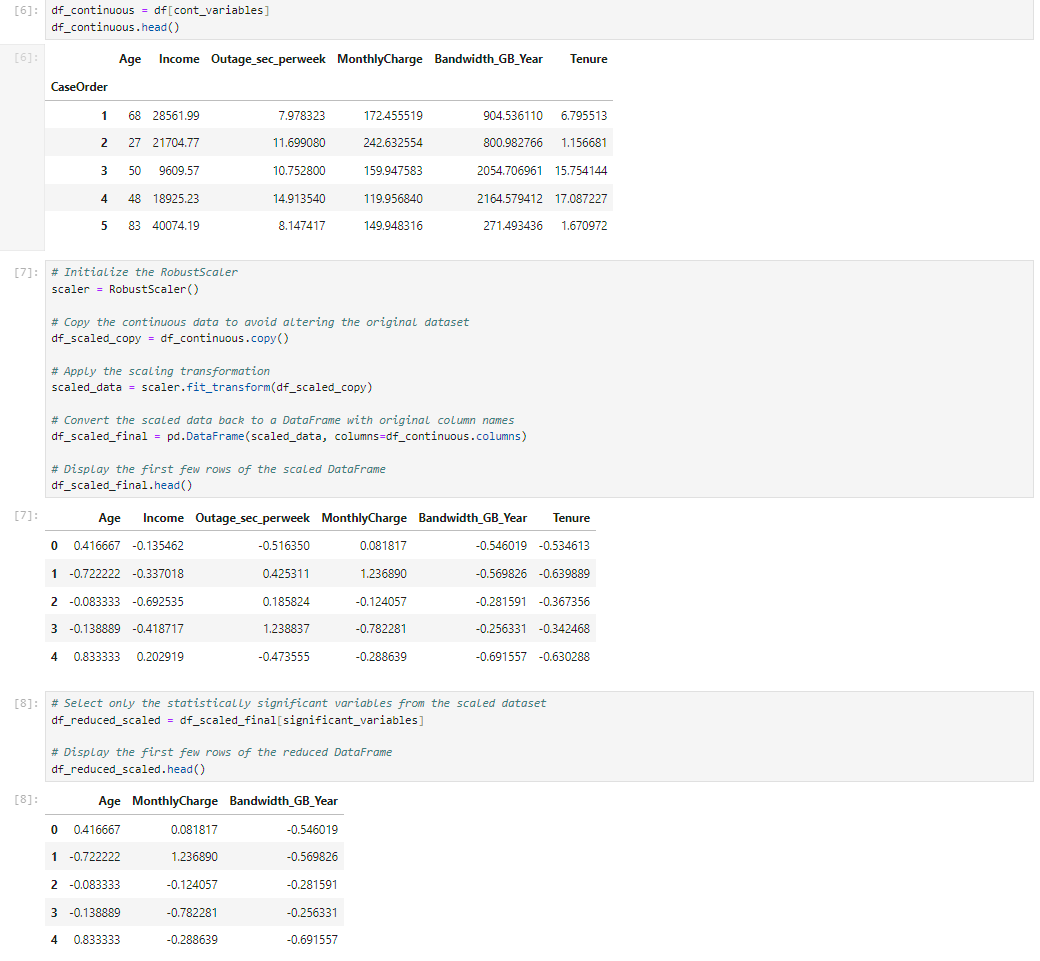
1. Describe one data preprocessing goal relevant to the clustering technique from part A1.

This project will involve two types of models: one using all continuous variables and another using only statistically significant variables from a previous linear regression analysis done in my spare time on the churn dataset.To use K-means clustering and PCA effectively, we need to focus on continuous data. This means we'll select specific parts of the datasets. After that, we'll scale the data using RobustScaler to handle outliers. RobustScaler is better than MinMaxScaler for this purpose because it isn't as affected by outliers, which are common in our dataset.

1. Identify the initial dataset variables that you will use to perform the analysis for the clustering question from Part A1, and label each as continuous or categorical.



1. Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step.



1. Provide a copy of the cleaned dataset.



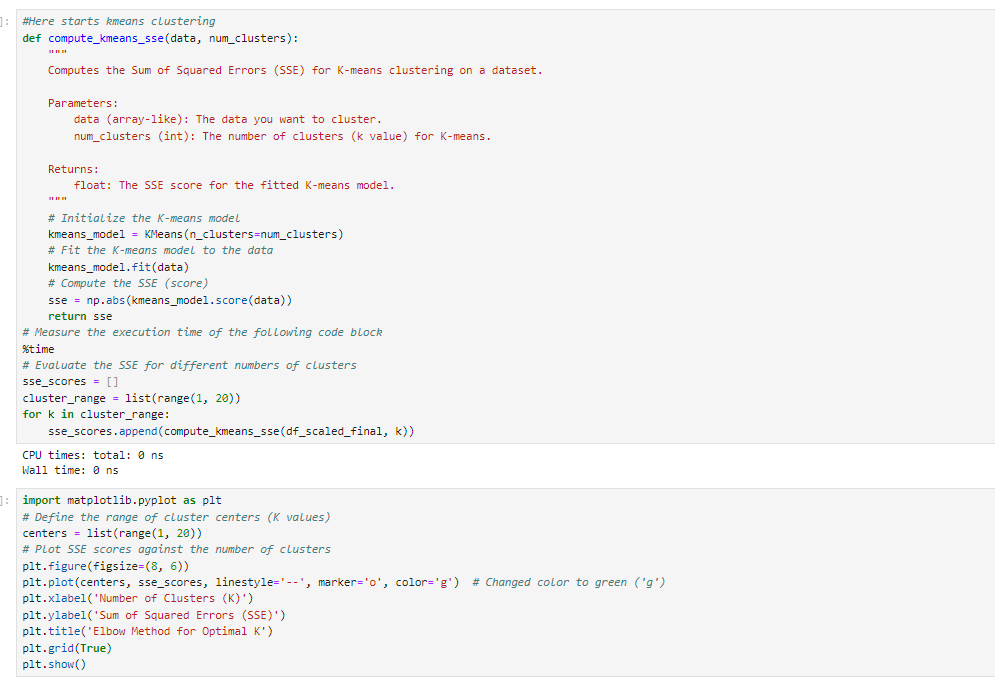
Part IV: Analysis

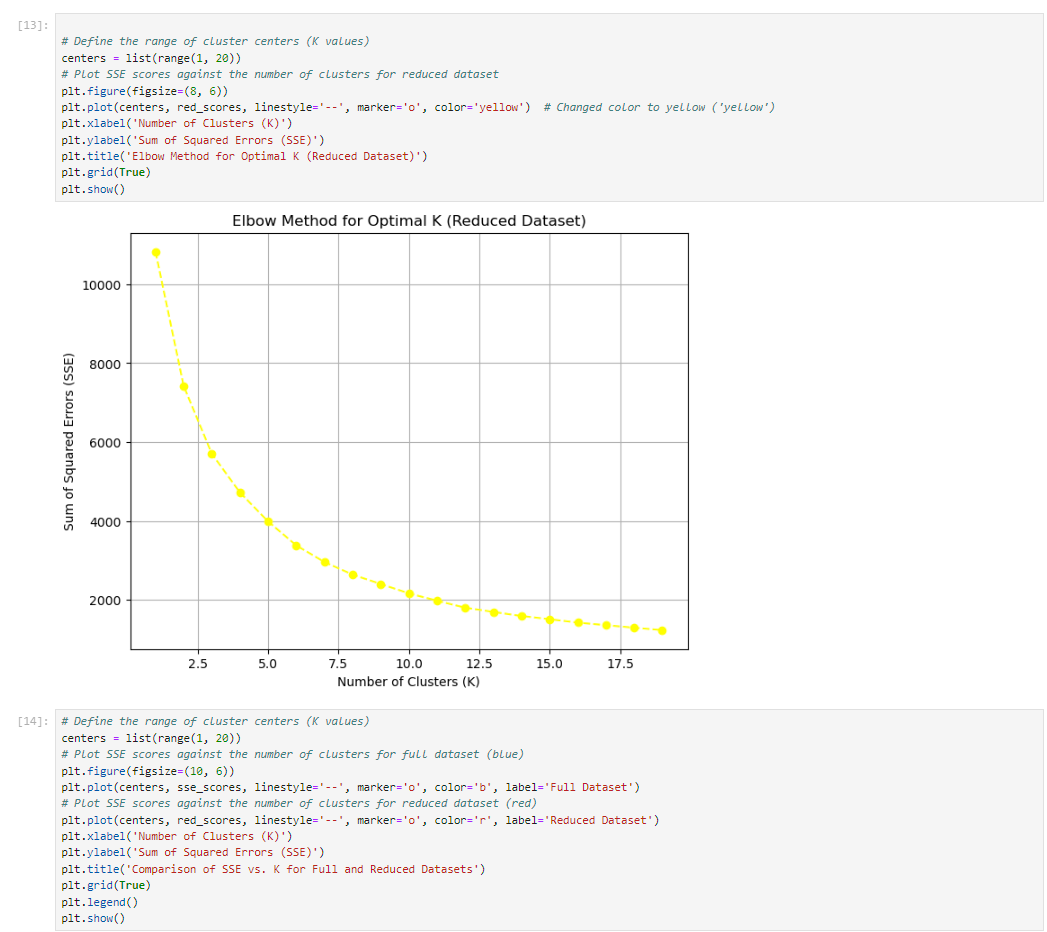
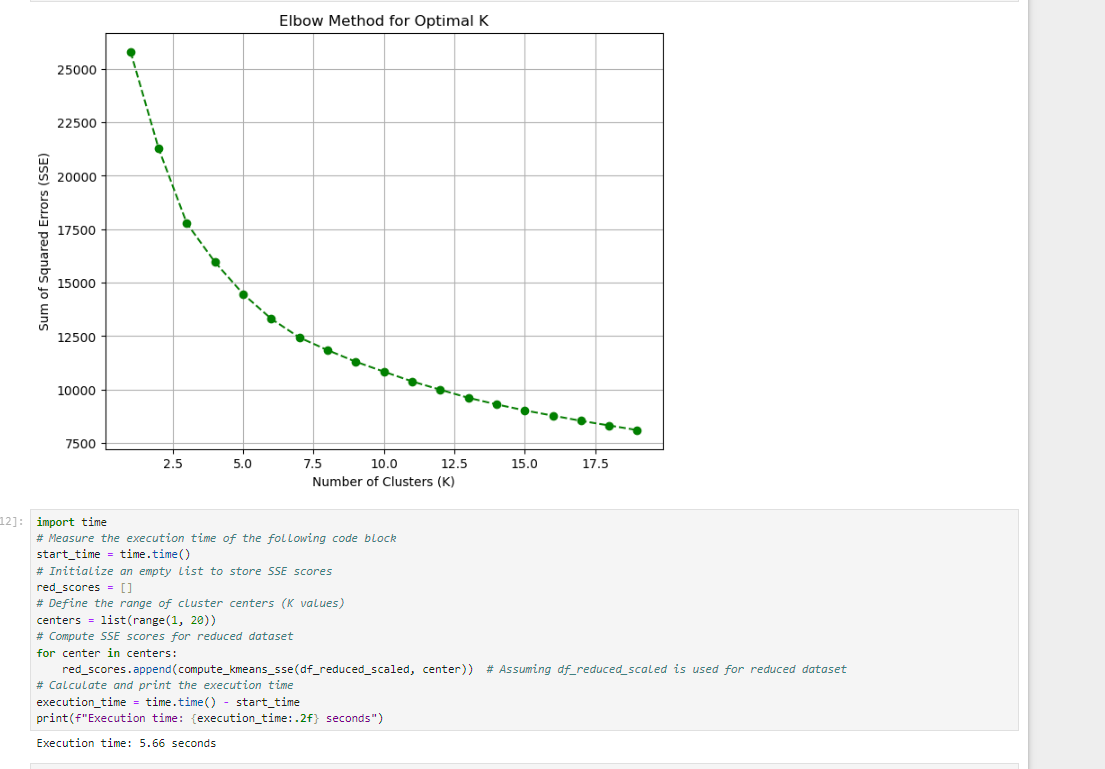
D. Perform the data analysis and report on the results by doing the following:

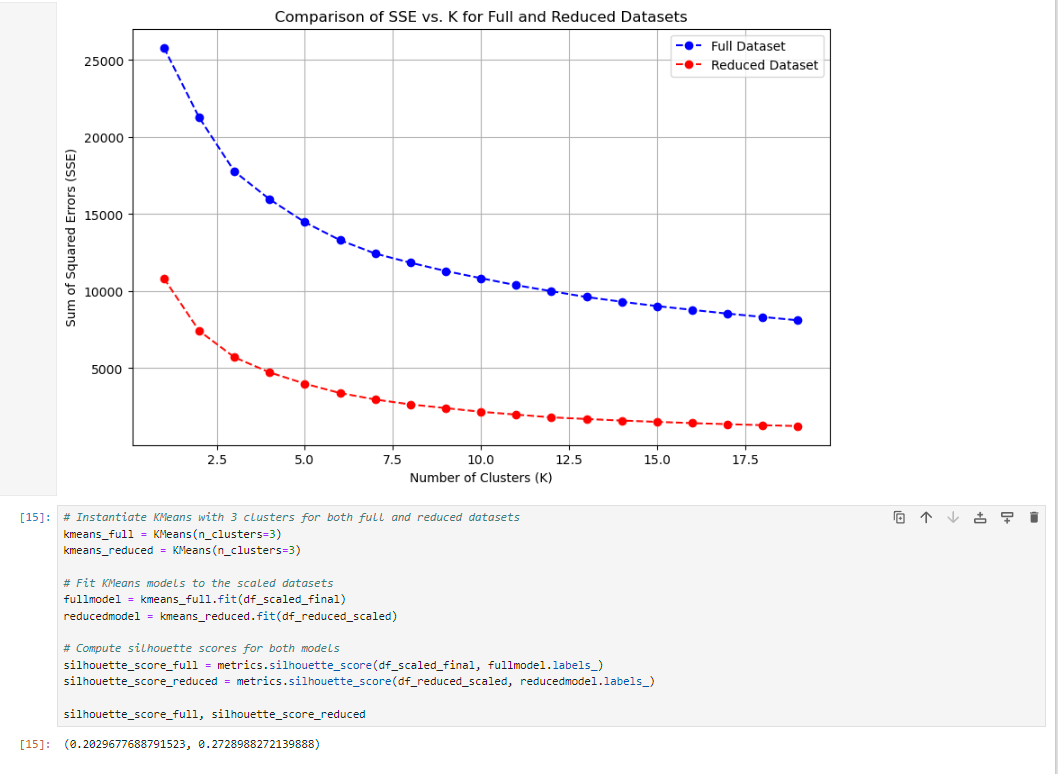
1. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

I utilized a function sourced from github that includes code for defining k-means clustering to test various k values on both the full and reduced datasets. We plotted the inertia (Sum of Squared Errors, SSE) to determine the optimal number of clusters. After identifying this optimal number from the elbow of the inertia plot, we evaluated cluster quality using the silhouette score.

Source: The function for defining k-means clustering was adapted from Smcgb. (n.d.). D212 Clustering Techniques. GitHub.<https://github.com/Smcgb/WGU_MSDA/blob/main/D212%20-%20Data%20Mining%20II/CLUSTERING%20TECHNIQUES/D212%20Clustering%20Techniques.ipynb>.







I have chosen to stop evaluating this clustering strategy further due to the continuously low silhouette scores seen in both the complete and reduced models. An essential tool for evaluating the quality of clusters is the silhouette score, which evaluates the cohesiveness and spacing of the clusters. A low silhouette score implies that the clusters are poorly defined and lack clear boundaries, which reduces the usefulness of the insights and visualizations obtained from displaying data.(W3 Schools) As such, using this model would not produce meaningful insights or useful results for our research goals.

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

1. Explain the accuracy of your clustering technique.

There is no target variable for comparing model accuracy in unsupervised learning. Instead, to evaluate performance, use measurements, such as the silhouette score. This score, which ranges from -1 (worst) to 1 (best), indicates how closely each object fits into its own cluster relative to others. Higher values denote better-defined clusters. The silhouette scores for the entire and reduced models of this model are 0.07 and 0.27, respectively. These scores imply that values from different groups overlap. Notably, with three clusters, the simplified model exhibits less overlap and a lower sum of squared error at the elbow point.

1. Discuss the results and implications of your clustering analysis.

The entire and reduced models' poor silhouette scores strongly imply that the clusters are not distinct and lack well-defined borders. This calls into question the usefulness of using K-means clustering with this dataset to obtain insightful knowledge about consumer segmentation. A key factor in this is the silhouette score, which compares intra-cluster cohesiveness to inter-cluster separation to determine how well-defined a cluster is. When scores show values that overlap between groups, it is clear that the data's natural structure might not be entirely consistent with the presumptions of K-means grouping. In this case, further in-depth feature engineering or the investigation of alternate clustering strategies may produce more insightful and useful segmentation results.

1. Discuss one limitation of your data analysis.

Limitation in this dataset concerns a number of continuous variables, chief among them income, which are dependent upon self-reported information provided by end-user consumers. This raises the possibility of errors because of different interpretations or reporting procedures' biases. Previous analyses have thoroughly addressed the consequences of these difficulties, raising worries about the dependability of the data and its potential impact on analytical conclusions. Taking steps to reduce reporting biases or looking into different data sources are two ways to address these limitations and improve the validity and robustness of further studies and conclusions drawn from the dataset.

1. Recommend a course of action for the real-world organizational situation from Part A1 based on your results and implications discussed in Part E2.

This model could be run with a different seed or with a different clustering method to investigate different strategies and evaluate possible differences in the outcomes. If such results continue, it may indicate that this dataset needs more attributes in order to identify relevant client categories. As an alternative, pre-assigning clients to clusters according to demographic data and then using clustering algorithms to validate these groupings could provide insights into the accuracy of the clusters and improve segmentation efforts. With regard to this dataset, these tactics seek to improve the efficiency and dependability of client segmentation techniques.

Sources:

W3Schools. (n.d.). *K-means clustering with Python*. Retrieved July 14, 2024, from<https://www.w3schools.com/python/python_ml_k-means.asp>