**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

Which customer segments exist based on income and number of children? This question will be answered using hierarchical clustering.

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.

Group telecom customers by income and number of children. These groupings can be used to perform targeted upselling of telecom services.

**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.

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample (scikit learn).

2.  Summarize **one** assumption of the clustering technique.

It is typical to normalize (standardize) continuous variables by subtracting the mean and dividing by the standard deviation. Otherwise, variables with large scale will dominate the clustering process (B,B,G, 2019).

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

|  |  |  |
| --- | --- | --- |
| **Library** | **Module** | **Purpose** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Part III: Data Preparation**

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.

The data preprocessing goal is to standardize the data.

2.  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.

Income and number of children are both numeric variables. See univariate statistics in [CODE].

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

The data was prepared according to the following steps:

* LIST
* THEM
* HERE

See [CODE] for more.

**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.

Hierarchical clustering was used to analyze the data. Once the data was processed, it was loaded into [SKLEARN CLUSTERING MODEL]. Plots were produced to visualize these results. See [CODE] for more.

**Part V: Data Summary and Implications**

E.  Summarize your data analysis by doing the following:

1.  Explain the accuracy of your clustering technique.

The silhouette coefficient was used a proxy for clustering accuracy. According to the output, the hierarchical clustering model had a silhouette coefficient of approximately 0.41 at n=3 clusters.

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

The dendrogram shows three primary clusters. The model produced the highest silhouette coefficient when using n=3 clusters. According to the analysis, there are three customer segments that exist based on income and number of children.

3.  Discuss **one** limitation of your data analysis.

I did not evaluate cluster performance with different linkage methods. Performing hierarchical clustering with different linkage methods may have resulted in higher silhouette scores for the model.

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

While the model produced the highest silhouette score at n=3 clusters, the performance is still relatively low. I would not use the current iteration of the model for customer segmentation. I recommend re-performing this analysis to evaluate cluster performance using untested linkage methods (single, average, complete) to find a model with a better silhouette score.

**Part VI: Demonstration**

F.  To view a walkthrough demonstration of the code and programming environment, refer to the following link: PANOPTO LINK

G.  Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

dendrogram code: https://scikit-learn.org/stable/auto\_examples/cluster/plot\_agglomerative\_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Bruce, P., Bruce, A., & Gedeck, P. (2019). *Practical Statistics for Data Scientists : 50+ Essential Concepts Using R and Python* (2 ed.). O'Reilly Media, Incorporated. Retrieved March 2023

info: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/