Performance Assessment 1 – Submission 2

D212 – Data Mining II

Jessica Hosey

MSDA, College of Information Technology

Western Governors University

October 12th, 2024

Part I: Research Question

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

While using hierarchical clustering, can we find any beneficial insights within patient groups/clustering?

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

The use of hierarchical clustering will help find similar clusters or groups of patients within the data. That information can help several areas within the company, such as patients, doctors, and billing, create the best quality care for our visitors. The ward's method will be specifically used to help limit noise outliers within the chosen variables (IBM 2024).

Part II: Technique Justification

B. 1. Explain how your chosen clustering technique analyzes the selected data set. Include expected outcomes.

Hierarchical clustering uses data to find patterns. It is an unsupervised machine-learning technique that measures the distance between these groups of patterns after creating a dendrogram diagram (IBM 2024). A dendrogram resembles a family tree or pedigree diagram used in introductory biology courses. Each branch connects the different groups or patterns within the data. The expected outcome is the creation of a dendrogram and the distance between the clusters (IBM 2024).

The dataset is a healthcare-related source of information about a day's operation and the patients who entered their care that day. Hierarchical clustering is excellent for the healthcare and bioinformatics field as it can show connections between similar groups or individuals. This technique will help us "categorize mixed populations into more homogeneous groups" (IBM 2024).

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

One assumption for HCA clustering technique is that all the variables being analyzed are scaled similar;y to one another. The use of the Standard Scaler package from sklearn will be used to appropriately scale the variables to ensure that the data is on the same scale prior to completing the HCA technique.

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

- Pandas: pandas framework and CategoricalDtype – These allow us to load the dataset and remap/recategorize the variables in the dataset.
- Numpy – This will allow us to complete basic calculations like mean and standard deviation.
- Matplotlib.pyplot - This allows us to create visualizations of the data.
- Seaborn – This allows us to create visualizations of the data.

- Scipy.cluster.hierarchy: linkage, fcluster, and dendrogram – This allows us to create, calculate, and refine the results of the clustering method chosen.
- Sklearn.metrics: silhouette_score – This allows us to score our dendrogram based on the means of the variables used.
- sklearn.preprocessing: StandardScaler – This allows us to standardize all the variables so there are no differences in the weight of a 1 (no response) versus a 53 (age) in the data.

Part III: Data Preparation

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

The biggest issue in the original dataset was that the categorical data was in string format (yes/no). I needed to convert these to a Boolean data type to complete hierarchical clustering. This will change the variables to responses such as '0' or '1' for 'no' and 'yes' replies (Hosey 2024).

Hierarchical clustering analysis (HCA) can handle many continuous and categorical data types and is used in this project. The more variables used, the longer it will take for HCA to create the dendrogram (IBM 2024).

C. 2. Identify the initial data set variables you will use to analyze the clustering question from part A1, and label each as continuous or categorical.

Variable Title	Data Type
Children	Continuous
Age	Continuous
Readmis	Categorical
Doc_visits	Continuous
Soft_drink	Categorical
HighBlood	Categorical
Stroke	Categorical
Overweight	Categorical
Arthritis	Categorical
Diabetes	Categorical
Hyperlipidemia	Categorical
BackPain	Categorical
Anxiety	Categorical
Allergic_rhinitis	Categorical
Reflux_esophagitis	Categorical
Asthma	Categorical

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

The exact steps for processing have been used for other WGU course task submissions. The following code segments were used and what they do to prepare the data for analysis:

Code Segment	Explanation
`.astype`	This is used to convert data types. A few variables were converted from categories to strings, strings to Booleans, etc. This was also used to remap the survey responses so the computer could weigh an '8' as least essential and a '1' as most important.
`.map`	This is used to aid in converting strings to Boolean data types above. This was also used to remap the survey responses so the computer could weigh an '8' as least essential and a '1' as most important.
`.copy()`	Used to gather the variables that we were interested in for HCA clustering.
`StandardScaler()`	Used to standardize all variables in the dataset so no two variables of the same input had different weights.
`.fit_transform()`	Used to fit and transform the data after using the standard scaler code.

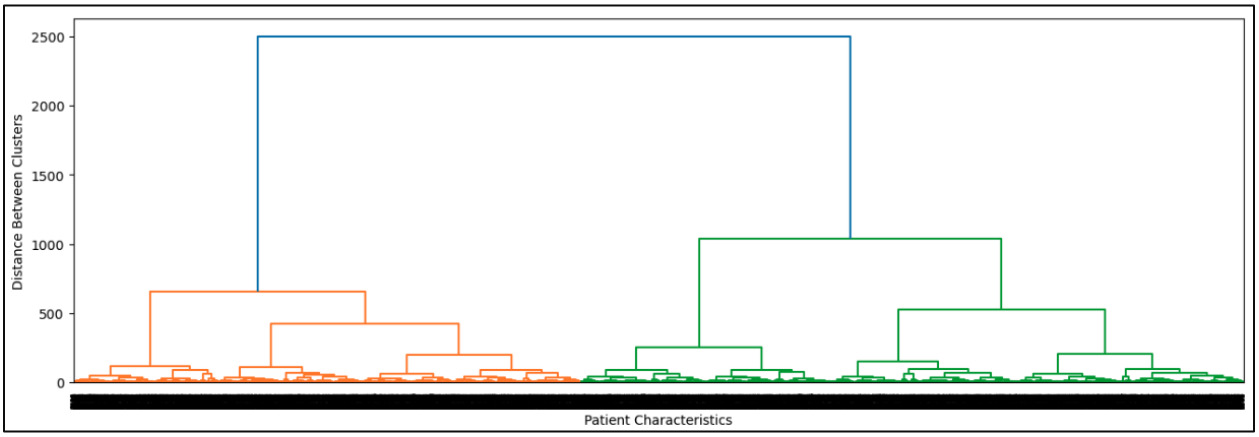
C. 4. Provide a copy of the cleaned data set.

See the attached files in my submission.

Part IV: Analysis

D. 1. Determine the optimal number of clusters in the data set and describe the method used to determine this number.

Based on the results of the created dendrogram, two significant clusters are found within the selected variables, as seen below.



In addition to the clusters, `fcluster` labels these clusters and can give us some information about the data points within the clusters. To figure out how many of the data points are within each cluster, the use of `value_counts` and `sort_index` gave us the number to see the difference in size. Cluster 1 had (as seen below) 43.35% of the dataset, and Cluster 2 had most of the data at 56.65%.

```
ward_cluster_labels
1    4335
2    5665
Name: count, dtype: int64
```

D. 2. Provide the code used to perform the clustering analysis technique.

See the attached files in my submission.

Part V: Data Summary and Implications

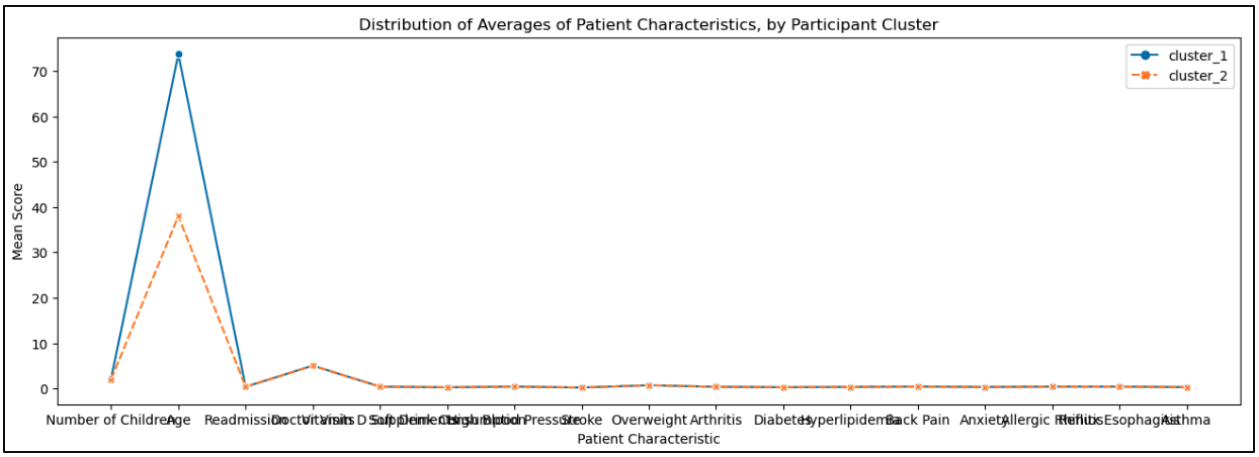
E. 1. Explain the quality of the clusters created.

Using the silhouette score to evaluate the cluster quality, the created hierarchical clustering technique received a score of 0.586 (as seen below). A silhouette score can range from –1 to 1, and the higher the score, the better the clustering will be created. Silhouette scores are a means of all silhouette scores of every data point in the dataset (Koli 2019). In our project, the score was high and almost near 1. Therefore, the quality of clusters created was relatively high.

```
The silhouette score of this hierarchical clustering is: 0.586
```

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

	cluster_1	cluster_2
Number of Children	2.192388	2.024360
Age	73.740484	38.032127
Readmission	0.373241	0.362048
Doctor Visits	5.031603	4.997352
Vitamin D Supplements	0.396309	0.400883
Soft Drink Consumption	0.257670	0.257370
High Blood Pressure	0.415456	0.404060
Stroke	0.203460	0.196117
Overweight	0.706574	0.711562
Arthritis	0.360554	0.354987
Diabetes	0.274971	0.272904
Hyperlipidemia	0.335409	0.338570
Back Pain	0.416609	0.407414
Anxiety	0.325029	0.318800
Allergic Rhinitis	0.402076	0.387996
Reflux Esophagitis	0.402999	0.421536
Asthma	0.293195	0.286320



Using the screenshots above, we can see the general information results found after using the Ward method for hierarchical clustering. The two clusters within the created dendrogram had a silhouette score of 0.586 (as seen in the previous question).

Cluster 1 represents 43.35% of the dataset, with all the variables being close to one another except for the variable Age. Age has a Mean Score above 70. The other increased mean score, which is different from the other variables, is doctor visits, with a score of 7-9.

Cluster 2 represents 56.65% of the dataset, with all the variables being close to one another except for the variable Age. Age has a Mean Score of 40 in this cluster. The other increased mean score, which is very different from the other variables, is doctor visits, with a score of 7-9.

The line graph above shows that both clusters have nearly identical mean scores for each variable. This means that no matter your Age (within this WGU-created dataset), being affected by these symptoms or diseases is not solely due to Age. Therefore, the dataset could be a limitation in finding specific insights about the variables for a company to use. If there was more data from more than one day of hospital operation, there might be more variation in the Mean Scores.

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

The most significant limitation of this project is that the data is from only one day of hospital operation. If there was more data from a range of dates, we could pull some real insights from the dataset that would be useful to the company. Without additional data, this created model is purely used as practice and not for pulling legitimate insights for a company looking to solve a problem.

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

In addition to gathering more data from various dates of hospital operation, the next step would be to create other dendrograms with the removed variables within the dataset to see if these variables would change or enhance our results from the original dataset. These two recommendations would be best for the company if it wishes to gather better insights from the data.

Part VI: Demonstration

F. Panopto Video

See the attached files in my submission.

G. Sources

Hosey, Jessica. (2024). Performance Assessment 2, D208 – Predictive Modeling. [Unpublished manuscript]. WGU.

Hosey, Jessica. (2024). Performance Assessment 1, D209 – Data Mining I. [Unpublished manuscript]. WGU.

IBM. (December 7, 2021). What is hierarchical clustering? IBM. https://www.ibm.com/think/topics/hierarchical-clustering

IBM. (August 5, 2024). Hierarchical cluster analysis. IBM Documentation. https://www.ibm.com/docs/en/spss-statistics/beta?topic=features-hierarchical-cluster-analysis

Koli, Shubham. (2019, March 31). How to evaluate the performance of clustering algorithms. Medium. https://medium.com/@MrBam44/how-to-evaluate-the-performance-of-clustering-algorithms-3ba29cad8c03