Clustering Techniques

Matthew E. Heino

Data Mining II

Introduction

In this assessment, we are asked to create a cluster of data based on a given dataset. The options for clustering are hierarchical and K-means clustering. The chosen clustering method will discussed in a subsequent section of the paper. There will be a research question that will be used to help develop an appropriate, clustering method, extract the appropriate data that can be used to answer the question, and be the right data to be used in developing the clustering model. Each section of this paper will map to a section within the rubric and the requirements of the assessment.

Background

This assessment revolves around one of the scenarios that were given in the requirements section of the assessment. This document will focus on the characteristics of the patients and what they want from the hospital. The dataset provides a set of data that can be used to answer this question. The **medical_clean.csv** file can be used to answer the question that will proposed in a future section of this document.

The medical dataset is composed of 10,000 rows and 50 columns. These columns are used to store data about the patient. Not all the columns of the CSV file will be used to create the clustering model that is the focus of this paper. The columns that will be used will be discussed in another section of the paper.

The language that will be used for this assessment will be Python and this will utilize a Jupyter Notebook. The use of the Jupyter Notebook will allow the reproducibility of the code and allow for stepping over the code for each section of the assessment during the video presentation. Please note that not all sections of the assessment required coding and only text-

based answers will be provided. If code is required there will be a reference to the appropriate section of the Jupyter Notebook so the reviewer can see and, if applicable, run the code that is referenced in that section.

Part I: The Research Question

In this section, there will be coverage of the research question. The type of clustering that will be used to meet the requirements of the assessment will be discussed in this section. There will be a discussion of the goal of the research question and how it pertains to the data that is found within the CSV file.

A1. The Research Question. In the dataset, there are a series of survey questions the patients were asked to answer. These questions asked the patient to rank the importance of various factors when it comes to the services or how they are treated at the hospital. The question that pertains to the scenario is the following "What are the most important factors that the patient needs to see in a hospital?" This question addresses the characteristics of what a patient deems most important when it comes to receiving service and using the hospital.

This question will seek to address what characteristics of a hospital and its staff are important to the patient. Through this research, we can see what the patient needs in a hospital and its staff. The survey questionnaire can be used to answer this question and meet the requirements of the assessment. Using the right clustering model can be used to answer questions. The right model will addressed in another section of the document.

A2. Goal of the Data Analysis. The goal of the data analysis is to find out which of the survey questions matter most to the patients. This is within the scope of the data and can be modeled successfully using the right model. The clustering model will be used to determine if

there are distinct clusters that are exhibited by the answers the patients had given. If there are distinct clusters how do they relate to the patients?

The overarching goal of the analysis is to find factors that matter most to the patients. With this analysis, the hospital can go about improving the areas that matter most to the patient. For example, if timely admission is found to be the most important to the patient. It might be beneficial to look into ways the hospital can speed up the admission process.

The actual results of the analysis will be discussed in subsequent sections of the paper.

Part II: Technique Justification

This section will cover the reasons for choosing the clustering technique that was used to model the data. This section of the paper will summarize one assumption of the clustering technique. In the last part of this section, there will be a listing of the packages that were used to create the assessment and the models.

B1. Explanation of How the Clustering Technique Analyzes. The chosen clustering technique for the data was the hierarchical clustering technique. The clustering technique works in the following manner. The first step is where each of the samples in the dataset begins within its cluster. The next step is to look at the next closest clusters and merge them. This merging will decrease the number of clusters. This clustering of data elements will continue until there is only one cluster left (Wilson, n.d.).

The type of clustering that was used in this assessment is referred to as agglomerative clustering. In this type of clustering the clustering begins at the bottom with each cluster being an element of the dataset. As described each of these elements is then grouped with a nearest neighbor. The large cluster will be the last one created and it will contain all the elements in the

dataset. This can be best illustrated using a dendrogram. The image below shows what the clustering model created for this assessment (Wilson, n.d.).

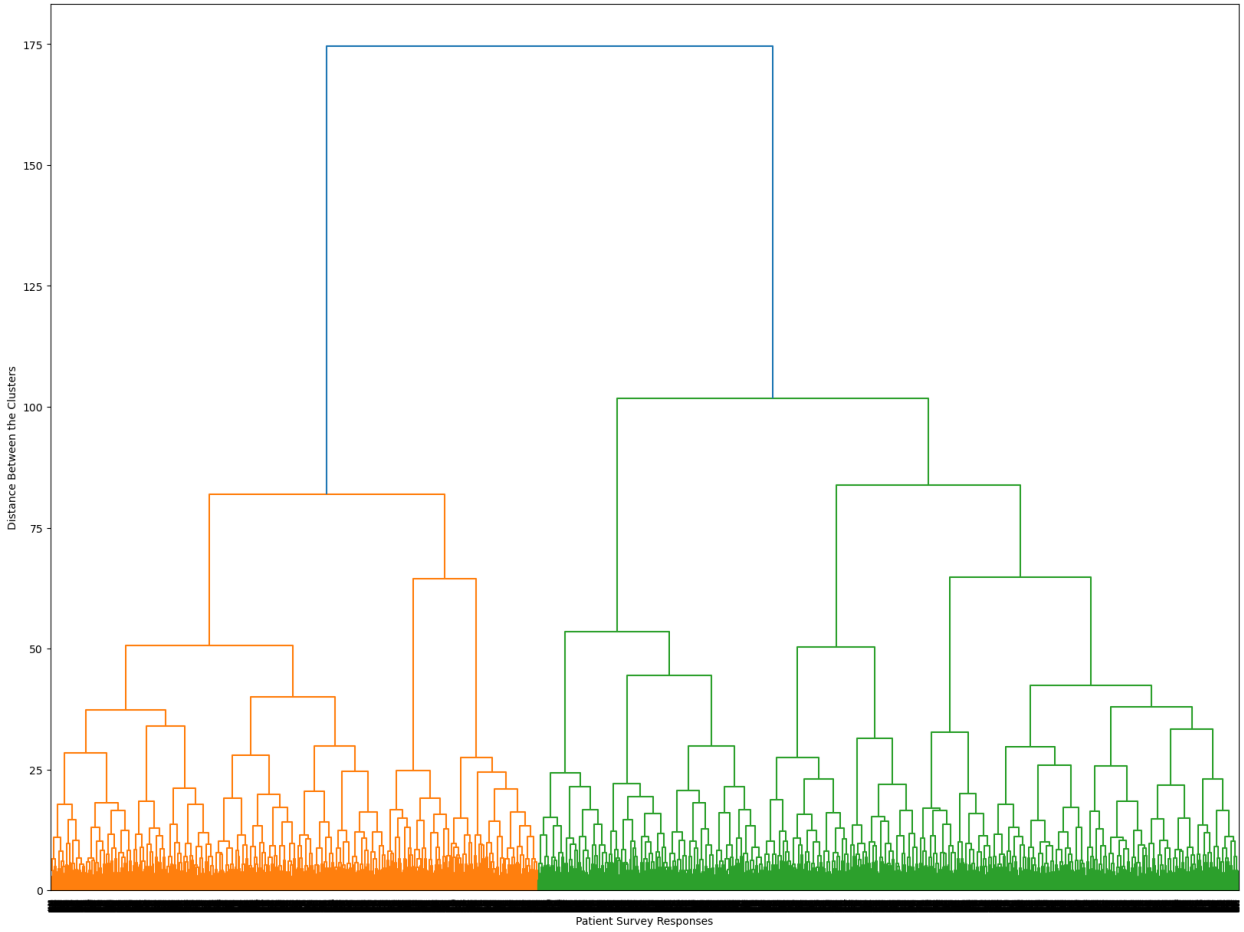

Image 1. Dendrogram of the clusters.

If you look at the bottom of the dendrogram you can see that each element of the dataset is grouped into its cluster. Subsequently, each of these clusters is grouped until you arrive at one cluster that contains all elements. This illustrated exactly what happens during the execution of the hierarchical algorithm.

The clustering was accomplished using SciPy's **linkage** method. This method will create the clusters that are visually depicted in the graph above. More about the process of creating this will be discussed in **Section D** of this document.

B2. Summary of One Assumption of Hierarchical Clustering. One assumption of the hierarchical clustering method is the method uses a measurement of distance between the elements in the dataset. The algorithm will try to group or cluster those elements that are near each other. The assumption is that the elements occupy the same scale. If the scale is not the same then the elements that will compose the cluster are on the same scale. This why methods like the **StandardScaler** or **Normalizer** will be employed depending on the data types that need to be scaled and the type of model that is to be created (Maniriho et al., 2022).

The data that was used to create this assessment did not need to undergo the process of standardization. All the data elements used the same scale. In the case of the survey data, it was on a scale of 1 to 8. There were no other values used that fell outside this range. None of the other features used a scale that varied greatly, therefore, implementing standardization or normalization did not seem to be required. This will be further discussed in the data preparation section of this document.

B3. Libraries Used in the Modeling. This assessment used the Python programming language version 3.7. This made it possible to use all the libraries that will be discussed in this section. Using older or newer versions of the language may not support all the tools that this assessment utilizes. The table below lists the libraries that were used to create the model and the libraries that were used to support other areas of data analysis.

Python Library	Description
matplotlib.pyplot	For plotting the graphs that were used in the assessment.
pandas	For reading the data from the CSV file and manipulating the data that is contained in the data frame.

seaborn	For creating graphs for the assessment. Used to help visualize the distribution of the answers that each cluster contained.
scipy.cluster.hierarchy	• **dendrogram** – used to create the dendrogram graph (see Image 1) • **fcluster** – this is used to flatten clusters to obtain the clusters and any possible values (see section D for more information) • **linkage** – a method that performs the hierarchical/agglomerative clustering. [1]
sklearn.metrics	• Provided the method for calculating the silhouette score for cluster quality.

These were the libraries that were used to create the model and other components of the assessment. Please refer to the Jupyter Notebook for their use and other documentation.

Part III Data Preparation

This section will cover what was required to get the chosen data into a state that can be used to create the model that was discussed in the previous section of this document. This section will list the variables that were required to create the model. There will be an explanation of the steps used to create the cleaned dataset. There will be a listing of the cleaned data file.

[1] The citation is scipy.cluster.hierarchy.linkage — SciPy v1.11.4 Manual. (n.d.). Retrieved January 3, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

C1. Goal of Data Preparation. One of the goals of data preparation is to get the data into a form that is suitable for use within the model. The data that is the focus of the analysis is the survey data. The data is currently stored in a counterintuitive way. It currently stores data in an order that assumes 1 is more than 8. To make the data more intuitive the data will be "flipped" in its order. There is a need to change the data type of the data. The reason for this will be covered in a subsequent section.

C2. Dataset Variables. The dataset variables will come from the **medical_clean.csv** file. The variables are listed in the table below along with a brief description of the information that is included in the variable. Please note that some of these variables will be renamed to make them easier to work with.

Variable Name	Type	Renamed To	Description
Item1	Qualitative (categorical)	timely_admis_surv	Response of whether timely admission is a factor.
Item2	Qualitative (categorical)	timely_treatment_surv	Response of whether timely treatment is a factor.
Item3	Qualitative (categorical)	timely_visits_surv	Response of whether timely visits are a factor.
Item4	Qualitative (categorical)	reliability_surv	Response of whether reliability is a factor.
Item5	Qualitative	options_surv	Response of whether availability of options

	(categorical)		is a factor.
Item6	Qualitative (categorical)	hours_of_treatment_surv	Response of whether hours of treatment are a factor.
Item7	Qualitative (categorical)	courteous_staff_surv	Response of whether courteousness of the staff is a factor.
Item8	Qualitative (categorical)	active_listening_surv	Response of whether the staff is actively listening is a factor.

Table 1. Variables used in the model.

The variables that were used in this model are ordinal in nature. So there is an implied hierarchy that can be used to create a hierarchical model that has been discussed previously.

C3. Steps Used to Prepare the Data. The steps used to prepare the data only involved data that was used to create the model. Any of the columns that were not used were not prepared and are not the subject of discussion in this section.

The data that was provided did not need much in the way of cleaning. There were no null values in the data provided. Each of the features that were used contained a value in the expected range. The range of values ran from one to eight.

Since the data all used the same "scale" the use of a standardization method or function was not required. This step was skipped but would have been required if the data varied

drastically in the size of the feature values. The boxplot below shows the distribution of the features. The data exhibits very similar statistics.

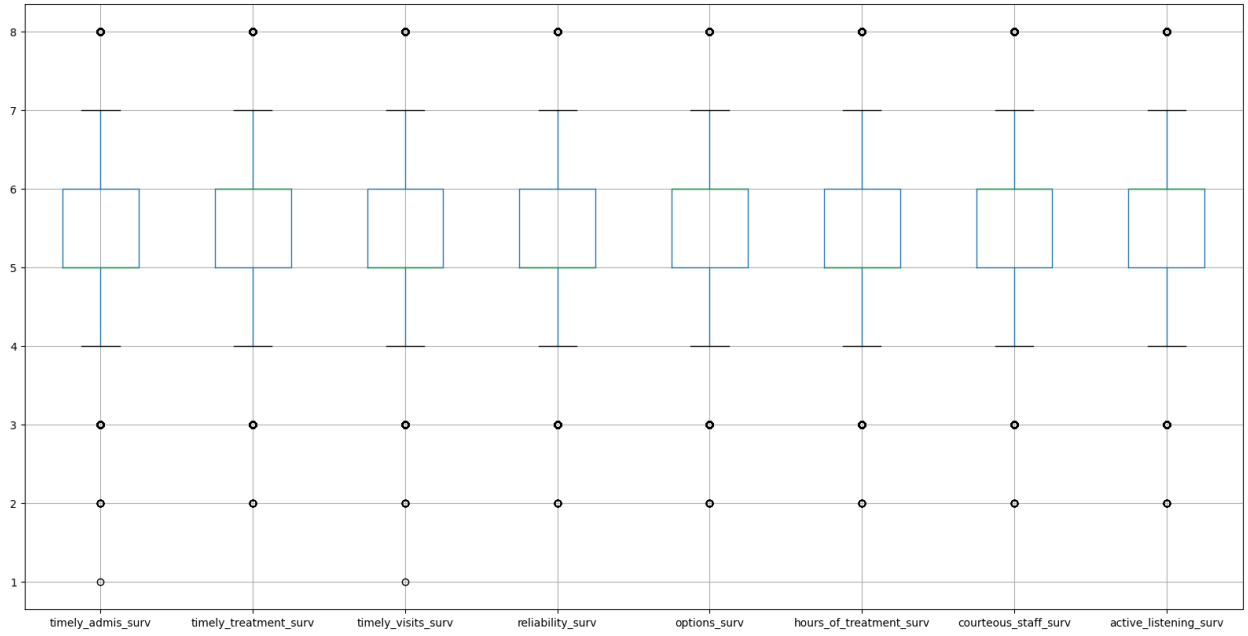

Image 2. Boxplot of the Features.

This visualization was further supported by creating some summary statistics about the features. The results of the execution of the **describe** method are shown in the following image.

	count	mean	std	min	25%	50%	75%	max
timely_admis_surv	10000.0	5.4812	1.031966	1.0	5.0	5.0	6.0	8.0
timely_treatment_surv	10000.0	5.4933	1.034825	2.0	5.0	6.0	6.0	8.0
timely_visits_surv	10000.0	5.4889	1.032755	1.0	5.0	5.0	6.0	8.0
reliability_surv	10000.0	5.4849	1.036282	2.0	5.0	5.0	6.0	8.0
options_surv	10000.0	5.5031	1.030192	2.0	5.0	6.0	6.0	8.0
hours_of_treatment_surv	10000.0	5.4775	1.032376	2.0	5.0	5.0	6.0	8.0
courteous_staff_surv	10000.0	5.5060	1.021405	2.0	5.0	6.0	6.0	8.0
active_listening_surv	10000.0	5.4903	1.042312	2.0	5.0	6.0	6.0	8.0

Image 3. Summary Statistics of the Features.

The only data preparation steps that were carried out were the following:

1. Change the column names from the uninformative Item format to something easier to remember and something more descriptive in terms of the content the column contains.
2. To make it more intuitive there was a change in how the surveys were ranked. This may not have been needed, but it made sense to do. There is a further explanation for this in the Data Analysis section of this paper.

The results of this were then output to a CSV file whose file name is given in the next section.

C4. Copy of the Cleaned Data

The file that contains the cleaned data is the following. It will only contain the columns that were used to create the model. It does not contain the columns that make up the rest of the dataset. It will contain 8 columns and 10,000 rows. This row count is the same as the original dataset.

- **Heino_cleaned_medical_task1.csv**

Part IV. Analysis

In this section, there will be the determination of the number of clusters in the data set. There will be a description of the method used to determine the number of clusters for the dataset.

D1. Number of Clusters. The number of clusters to be used was derived by using the following method. This assessment makes use of the hierarchical clustering algorithm. It will use the methods that were **linkage** method that was supplied by the Scipy library.

The **linkage** method was used in coordination with the dendrogram graph to illustrate how the data was clustered. Reviewing the graph it seems that the data was clustered into two

groups as indicated by the orange and green groupings. The dendrogram below illustrates this clustering.

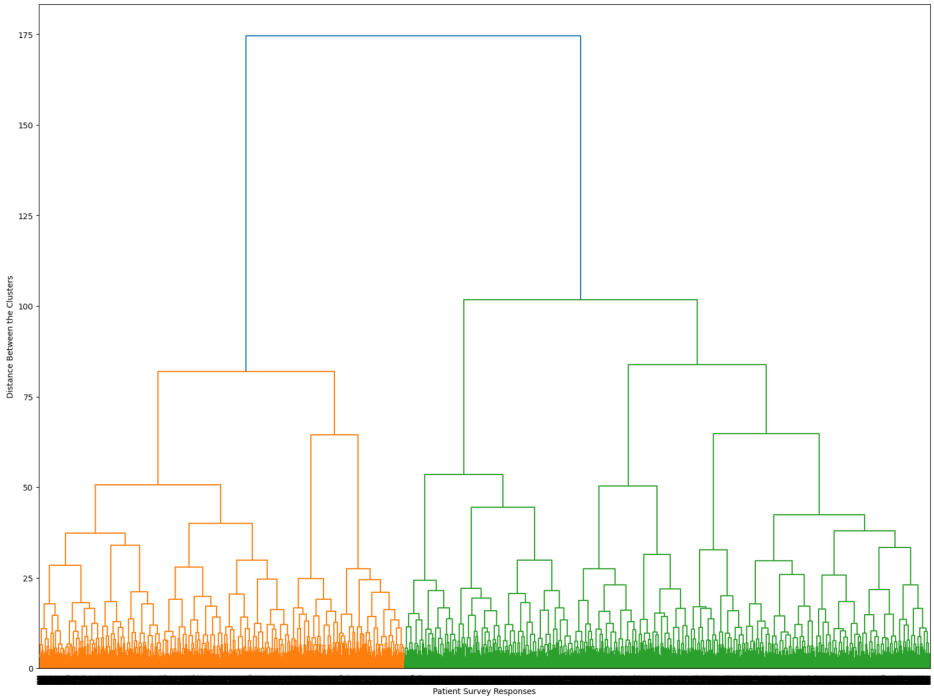

Image 4. Dendrogram of the clusters.

The **linkage** method used the **Ward** method to create the clusters. The Ward method is sometimes referred to as the Minimum Variance Clustering Method. During the clustering process, the Ward method tries to keep variance between the clusters to a minimum. Where variance is the amount of difference between the members of the cluster (Tim, n.d.).

The other method that was used to look at the model was the **fcluster** method. This was to retrieve information about how elements composed the clusters. The number of members of Cluster 1 was 4105 and for Cluster 2 5895. This distribution is loosely illustrated by the dendrogram graph in Item 2 shown above.

Other combinations of parameters were tried with **flcuster**. These proved to be no better.

There was the output of the sizes of the clusters for each of the new clusters. See a sample of the output after trying different sizes.

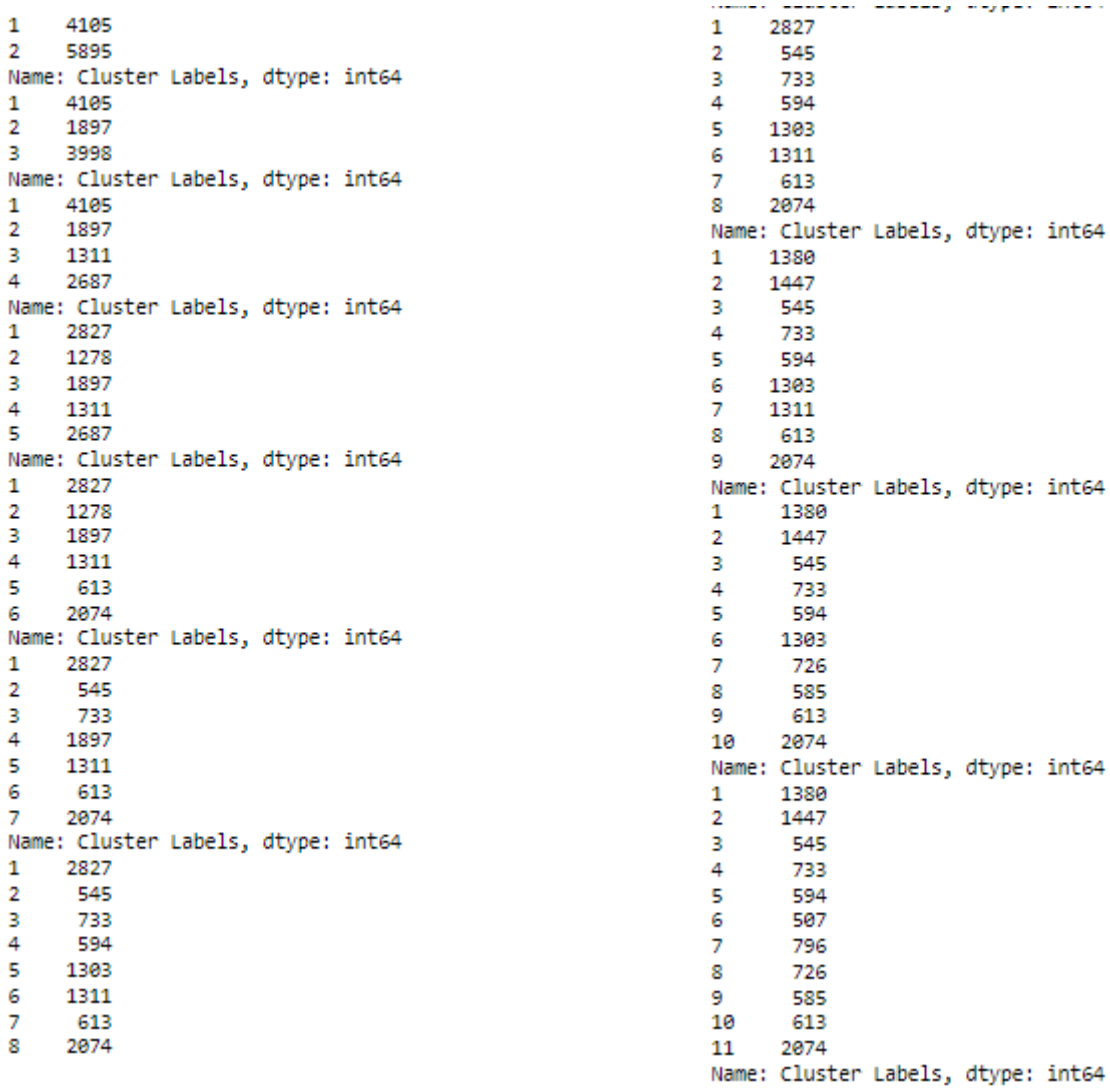

Image 5. Different Clusters and their sizes.

Some of the other parameters yielded either a large number of clusters with minimal. An example of other silhouette scores can be found in the graph below.

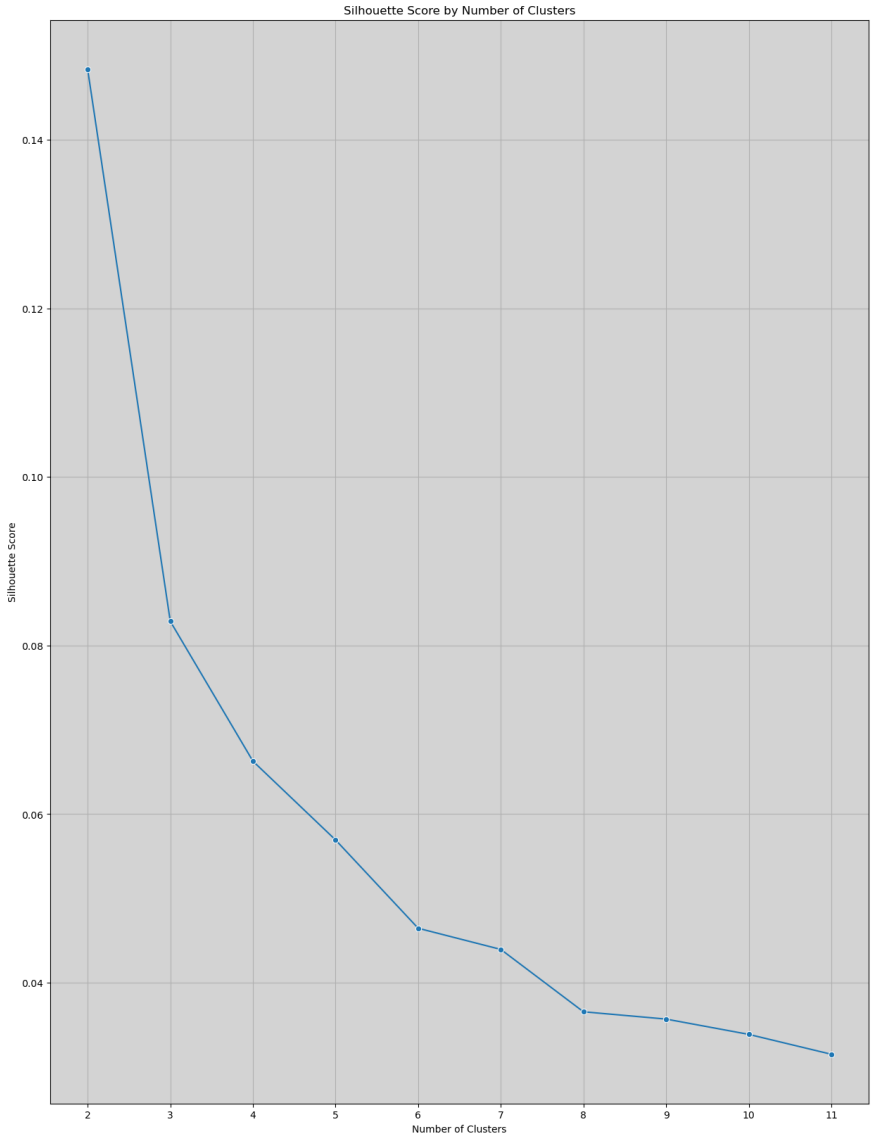

Image 6. Different Clusters and their sizes.

There will be further discussion about the results in section E1 – Quality of the Clusters.

D2. Code for the Analysis.

In sections D1 and D2 of the Jupyter Notebook. Please refer to that section of the notebook for the exact code along with additional comments about how the analysis was created

and implemented. The code will not be included in this section but can be best viewed in the notebook. The Notebook is included and it does a better job of keeping the formatting of the code and it provides a way to show the code in context with the other code that was used to create this assessment.

Part V. Data Summary and Implications.

In this section, there will be a discussion of the quality of the clusters that were created. A short discussion of the implementations of the clustering analysis. A discussion of the limitations of the analysis that was performed. After this section, there will be a recommended course of action based on the analysis of the dataset.

E1. Quality of the Clusters. The quality of the clusters will be measured using a silhouette score or silhouette coefficient. The silhouette coefficient or score will give a value that determines how well the clusters are separated from each other. The range of values for the silhouette score range from -1 to 1. A value that is close to -1 indicates that a sample has been assigned to the wrong cluster. A value near zero indicates that there is some degree of overlapping of the clusters. This means that the sample will be very close to the decision boundary. This could lead to a data element being assigned to two possible clusters. A value that approaches 1 is where the clusters are far away from any neighboring clusters (*Sklearn.Metrics.Silhouette_Score*, n.d.).

The score that was received for this model was approximately .14837. This indicates that there is some degree of overlap between the two created clusters. A screenshot of the output can be found below.

```
# Print the silhouette score.
print("The silhouette score for {} clusters is {}".format(int(scores_df.iloc[0]["Number of Clusters"])
                                                          , round(scores_df.iloc[0]["Silhouette Score"]
                                                                  , 5)))
```

```
The silhouette score for 2 clusters is 0.14837
```

The code for this can be found in section **E1. Quality of the clusters → The Silhouette Score**. As stated earlier other cluster sizes were tried and yielded the following results.

Number of Clusters	Silhouette Score
2	0.148369
3	0.082875
4	0.066284
5	0.056939
6	0.046459
7	0.043949
8	0.036552
9	0.035686
10	0.033872
11	0.031512

Image 7. Different Clusters and Their Silhouette Scores.

E2. Implications of the Cluster Analysis. The implications of the cluster analysis can be interpreted by looking at a few graphs that were created previous section of the assessment. The graphs that were created in Section D [2] show that the patients did not have a real preference when it came to the importance of the different factors that were presented to them. All the question factors were ranked in the middle with no clear winner when it came to a factor. Each factor was ranked equally when looked at in aggregate. The plots below help to illustrate this observation.

[2] The code that was used to create these graphs can be found in this section of the document. It will not be inserted into this section of the paper.

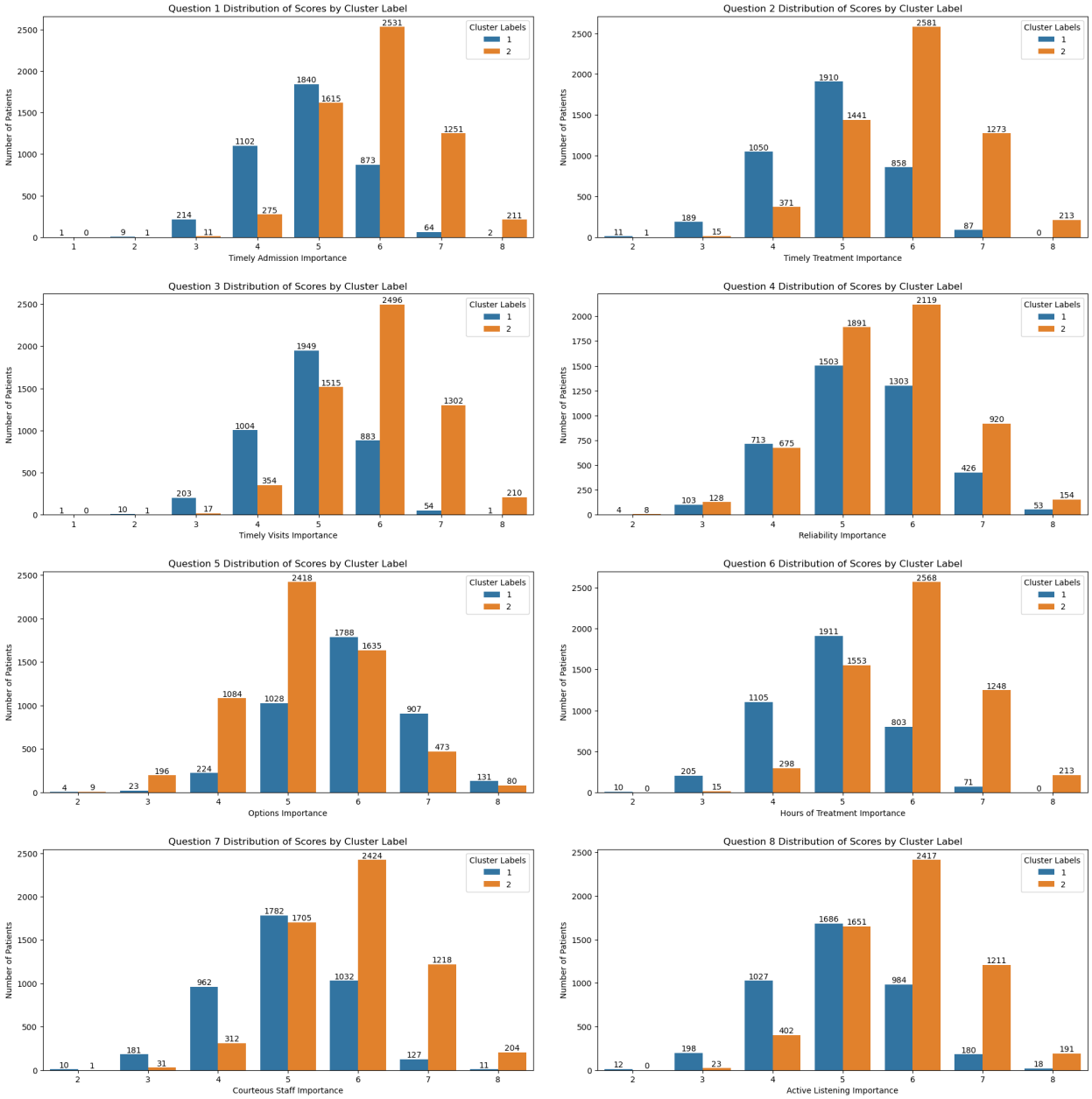

Image 8. Cluster distribution with counts per score.

Reviewing these graphs the data is mostly located around two values 5 and 6. This means that there is no strong feeling as to what factors are important to the patient. This is a problem when we want to answer the question that was proposed in section A of the paper. Without a clear inclination in preference, it will be difficult to propose a plan for the organization.

By looking at the results by cluster we can derive what questions are more important to each of the clusters. For Cluster 1 there was an equal score for all the questions except for question 5, which seemed to be the only question that scored slightly higher in this cluster. This was the Options questionnaire. For Cluster 2, the cluster ranked most questions about a 6. The only exception is that Question 5 (options survey) was ranked the lowest overall.

By observing the graphs you can deduce that Cluster 1 tends to rank some questions on the lower end of the scale. While Cluster 2 tends to rank questions a little higher. Please keep in mind that the survey ranking was remapped during the data preparation stage to make this type of comparison easier and more intuitive. Another thing to consider is the size. The first cluster has fewer elements in it.

E3. Limitations of the Data Analysis. The limitations of the analysis are that using hierarchical clustering there is a chance that arbitrary decisions are being made when it comes to grouping the data. This is an inherent in the algorithm. The other limitation is the choice of the distance metric. The metric that was chosen for clustering was the Euclidean metric. This model could have made use of a different metric for the model but that was not attempted.

The most limiting factor is that as the size of the dataset increases there is an exponential increase in the computational time to create the clusters within the hierarchy (Daityari, n.d.). The graph below shows this for a sample range of up to 10,000 data points.

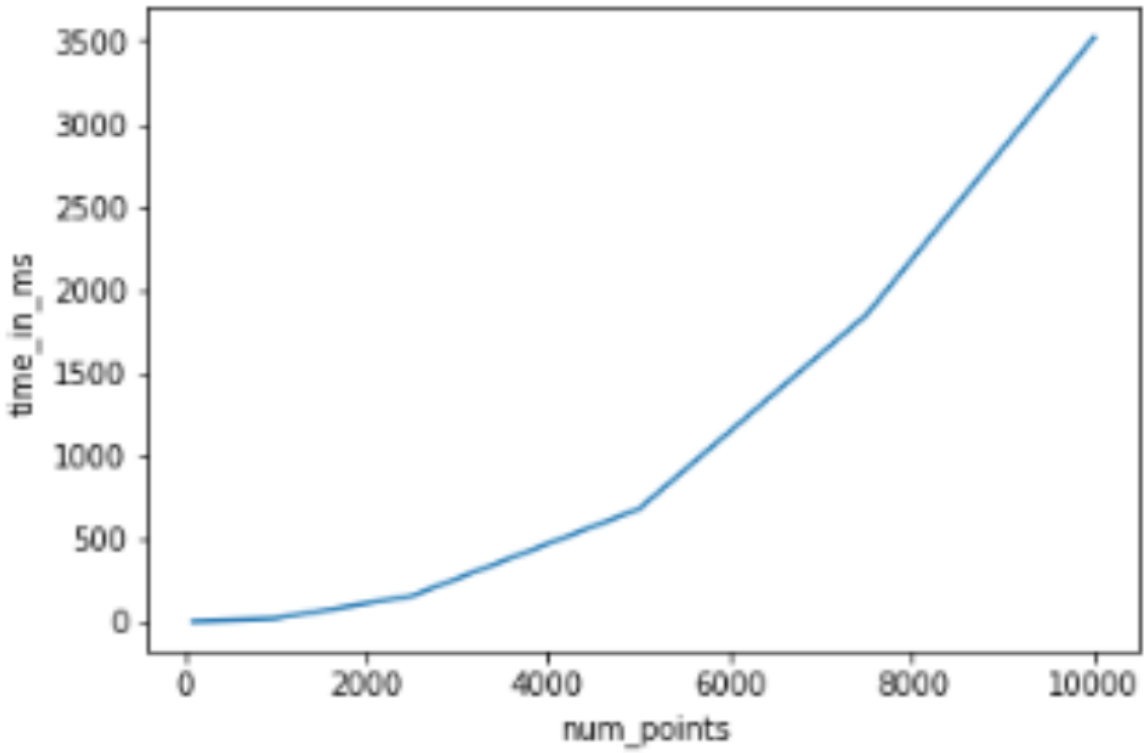

Image 9. Cluster distribution with counts per score. Image Credit: Daityari, n.d.

Reviewing the image above you can see that as you increase the number of samples within the dataset you will get a corresponding increase in computational time. While this is adequate for the small dataset that was used in this assessment. It will not scale well with a larger dataset that may number into the millions of data points.

For example, it could not be utilized in a large dataset composed of Amazon or Walmart customers. The sheer number of transactions could reach easily into the millions for a given day, let alone a month or year. If this algorithm were to be employed in datasets of that size and caliber there would be a need to partition the data into smaller chunks either by day or in the case of Walmart by the store. This may bring the size of the dataset down to something this algorithm can handle in a reasonable amount of computational time.

E4. Course of Action. A recommended course of action is the following. Looking at how the data was distributed and the clusters that were formed. It is difficult to propose a course of action. The first course of action should be to create questions that are engaging to the people. Looking at the distributions and the CSV file it is easy to notice that most of the values that were exhibited in the data are in the middle. There are very few values that would help to indicate what factors are most important to the patient. All the factors were "ranked" more or less the same.

There were eight questions along with eight possible answers. It may be beneficial to change the questionnaire in a manner that more likely invokes more time and consideration when it comes to ranking the factors want most. Looking at the answers that were given it is clear that the patients did not like the questions, they did not take the time to read them and just put any suitable answer, or the questions were not worded in a way that made the patient want to engage with the questionnaire. Writing engaging questions will make it easier for the patient to answer and for subsequent analysis to occur. The answers that were given were not thought through. The questionnaire was useless in gaining information that could be used to make the patients' lives easier when they visit the hospital and make it easier for the hospital to provide services that are desired and/or expected by the patient.

Part VI. Demonstration

F1. Demonstration. The link below is to the Panopto presentation of the code being executed within a Jupyter Notebook. The Jupyter Notebook can also be found as an included file with this assessment.

References

G. Web Sources Use

This section includes citations that were used that were not included in the resources that were provided by the university and not included in the DataCamp Videos.

Awan, A. A. (2022, November 7). *4 Ways to rename pandas Columns – KDNuggets*. KDnuggets. Retrieved January 2, 2024, from https://www.kdnuggets.com/2022/11/4-ways-rename-pandas-columns.html

pandas.Series.map — pandas 2.1.4 documentation. (n.d.). Retrieved January 3, 2024, from https://pandas.pydata.org/docs/reference/api/pandas.Series.map.html

scipy.cluster.hierarchy.linkage — SciPy v1.11.4 Manual. (n.d.). Retrieved January 3, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

H. In-text Citations

This section includes the citations that were used in the creation of the written document. These are citations that were not used to create code that can be found in various sections of the Jupyter Notebook.

Daityari, S. (n.d.). *Hierarchical Clustering - Limitations of hierarchical clustering*. DataCamp. Retrieved January 8, 2024, from

https://campus.datacamp.com/courses/cluster-analysis-in-python/hierarchical-clustering-c5cbdf0e-e510-4e0a-8437-4df11123fd58?ex=11

Maniriho, P., Mahmood, A. N., & Chowdhury, M. J. M. (2022). A study on malicious software behavior analysis and detection techniques: Taxonomy, current trends and challenges. *Future Generation Computer Systems*, *130*, 1–18. https://doi.org/10.1016/j.future.2021.11.030

sklearn.metrics.silhouette_score. (n.d.). Scikit-learn. Retrieved January 4, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Tim. (n.d.). *Ward's Method (Minimum variance method) - Statistics How To*. Statistics How To. Retrieved January 4, 2024, from https://www.statisticshowto.com/wards-method/

Wilson, B. (n.d.). *Visualization with Hierarchical Clustering and t-SNE – Visualizing hierarchies*. DataCamp. Retrieved January 1, 2024, from https://campus.datacamp.com/courses/unsupervised-learning-in-python/visualization-with-hierarchical-clustering-and-t-sne?ex=1