Performance Assessment

WGU | 211

D212 PA1

Narcisse, Laurie

2023

Contents

[**Part I: Research Question** 2](#_Toc156301992)

[A.  Describe the purpose of your data mining report by doing the following: 2](#_Toc156301993)

[1.  Propose one question relevant to a real-world organizational situation that you will answer using one of the following clustering techniques: 2](#_Toc156301994)

[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. 2](#_Toc156301995)

[**Part II: Technique Justification** 2](#_Toc156301996)

[B.  Explain the reasons for your chosen clustering technique from part A1 by doing the following: 2](#_Toc156301997)

[1.  Explain how the clustering technique you chose analyzes the selected data set. Include expected outcomes. 2](#_Toc156301998)

[2.  Summarize one assumption of the clustering technique. 3](#_Toc156301999)

[3.  List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis. 3](#_Toc156302000)

[**Part III: Data Preparation** 5](#_Toc156302001)

[C.  Perform data preparation for the chosen data set by doing the following: 5](#_Toc156302002)

[1.  Describe one data preprocessing goal relevant to the clustering technique from part A1. 5](#_Toc156302003)

[2.  Identify the initial data set variables you will use to perform the analysis for the clustering question from part A1, and label each as continuous or categorical. 5](#_Toc156302004)

[3.  Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step. 5](#_Toc156302005)

[4.  Provide a copy of the cleaned data set. 7](#_Toc156302006)

[**Part IV: Analysis** 7](#_Toc156302007)

[D.  Perform the data analysis, and report on the results by doing the following: 7](#_Toc156302008)

[1.  Determine the optimal number of clusters in the data set, and describe the method used to determine this number. 7](#_Toc156302009)

[2.  Provide the code used to perform the clustering analysis technique. 13](#_Toc156302010)

[**Part V: Data Summary and Implications** 16](#_Toc156302011)

[E.  Summarize your data analysis by doing the following: 16](#_Toc156302012)

[1.  Explain the quality of the clusters created. 16](#_Toc156302013)

[2.  Discuss the results and implications of your clustering analysis. 17](#_Toc156302014)

[3.  Discuss one limitation of your data analysis. 18](#_Toc156302015)

[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. 18](#_Toc156302016)

[G.  Record the web sources you used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable. 19](#_Toc156302017)

# **Part I: Research Question**

## A.  Describe the purpose of your 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, using only continuous variables

#### •  hierarchical

How can distinct customer segments be identified based on their responses to an eight-question survey, where they rate the importance of factors such as timely response, timely fixes, timely replacements, reliability, options, respectful response, courteous exchange, and evidence of active listening?

The aim is to utilize hierarchical clustering to enhance the telecommunication company's comprehension of customer segments and their preferences regarding these specific survey items.

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 goal is to employ hierarchical clustering to segment customers into meaningful clusters based on their responses to an eight-question survey. By identifying distinct groups with shared characteristics related to factors like timely response, timely fixes, reliability, options, respectful response, courteous exchange, and evidence of active listening, the objective is to gain insights that can enhance customer relationship management and effectively reduce churn for the telecommunication company.

# **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 data set. Include expected outcomes.

I embarked on this research journey with the aim of identifying distinct customer segments through the application of hierarchical clustering. The dataset I chose comprises responses from customers who participated in an eight-question survey, wherein they rated the importance of factors such as timely response, timely fixes, reliability, options, respectful response, courteous exchange, and evidence of active listening. This research question seeks to unravel meaningful clusters among customers based on shared characteristics inferred from their survey responses.

In my pursuit of meaningful segmentation, hierarchical clustering stands out as a powerful technique. This approach constructs a hierarchical tree-like structure, known as a dendrogram, by iteratively merging clusters of customers with similar survey item responses. The methodology involves measuring the similarity or distance between customers, selecting a linkage method to guide the merging process, and ultimately identifying clusters at a specific dendrogram height. The resulting clusters represent distinct groups of customers with shared characteristics, providing a comprehensive understanding of the diverse preferences within the customer base.

In conclusion, my exploration of hierarchical clustering within the context of customer segmentation aims to unravel nuanced patterns within the survey data. By employing this analytical technique, I anticipate uncovering clusters that encapsulate distinct customer characteristics, thereby providing actionable insights for the telecommunication company to tailor their approaches and better serve the diverse needs of their customer base.

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

One assumption of hierarchical clustering is that the data points within a cluster are more similar to each other than to data points in other clusters. This assumption is fundamental to the concept of clustering, as it implies that clusters represent groups of observations that share common characteristics or patterns. The algorithm relies on the notion that the chosen similarity or distance metric accurately captures the underlying structure of the data, allowing for meaningful grouping of similar data points. If the assumption of intra-cluster similarity is violated, the effectiveness of the clustering results may be compromised, leading to less meaningful and less interpretable groupings. Therefore, careful consideration of the appropriateness of the similarity metric and the nature of the data is crucial in ensuring the reliability of the clustering outcomes.

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

A screenshot of a computer program

Description automatically generated

1. pandas (pd): Data manipulation and analysis library for structured data.
2. CategoricalDtype (from pandas.api.types): Handles categorical data types in pandas.
3. numpy (np): Fundamental package for scientific computing with support for arrays.
4. matplotlib.pyplot (plt): Plotting library for creating visualizations in Python.
5. seaborn (sns): Statistical data visualization library based on Matplotlib.
6. scipy.cluster.hierarchy: SciPy module for hierarchical clustering and dendrogram plotting.
7. linkage, fcluster, dendrogram, cophenet: Functions for hierarchical clustering in SciPy.
8. silhouette\_score (from sklearn.metrics): Metric for clustering evaluation.
9. pdist (from scipy.spatial.distance): Computes pairwise distances for clustering algorithms.

A screenshot of a computer

Description automatically generated  
 A screenshot of a computer

Description automatically generated

# **Part III: Data Preparation**

## C.  Perform data preparation for the chosen data set by doing the following:

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

The main challenge with this dataset is the format of the survey results. The data dictionary indicates that a survey score of 1 represents the highest importance, while 8 is the lowest. This counterintuitive format complicates comparisons. To address this, the data should be remapped so that 1 becomes 8 and vice versa. Additionally, the variables need to be cast as floats instead of ints for compatibility with the linkage matrix.

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

1. Item1: Timely response - Ordinal Categorical
2. Item2: Timely fixes - Ordinal Categorical
3. Item3: Timely replacements - Ordinal Categorical
4. Item4: Reliability - Ordinal Categorical
5. Item5: Options - Ordinal Categorical
6. Item6: Respectful response - Ordinal Categorical
7. Item7: Courteous exchange - Ordinal Categorical
8. Item8: Evidence of active listening - Ordinal Categorical

I chose to classify each column as ordinal categorical based on the nature of the data and the context of the survey responses. In the provided dataset, customers were asked to rate the importance of various factors on a scale of 1 to 8, where each numerical value represents a category indicating the perceived importance of the corresponding factor. The key reason for selecting ordinal categorical is the inherent order in these ratings – a higher numerical value implies a lower perceived importance.

For instance, in Item1 ("Timely response"), a rating of 1 signifies the highest importance, while a rating of 8 indicates the least importance. This ordinal relationship is consistent across all survey items. The ratings represent ordered categories rather than continuous measurements, and the intervals between the categories are not assumed to be equal.

Recognizing these characteristics, I opted for the ordinal categorical classification to accurately reflect the nature of the data. This choice is essential for appropriate statistical analysis, especially in methods that account for the ordinal nature of the variables, such as hierarchical clustering with appropriate distance metrics for ordinal data. It ensures that the analysis respects the meaningful order of the categories and provides more accurate insights into the relationships between the survey items.

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

The data will be cleaned using much of the same code that was utilized in my previous assessments.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

I am creating a mapping to reverse the values of survey questions (1 < 8). Then, I loop through each survey item (Item1 to Item8) in the DataFrame (df), remap their values using the established mapping, and convert the columns to float data type. This is done to ensure that the data reflects the correct order for analysis, and the float conversion is necessary for certain operations.

A screenshot of a computer program

Description automatically generated

I'm using a loop to iterate over the questions in the survey. The loop runs from 1 to 8 (inclusive), representing the eight questions.

Inside the loop, I calculate the mean overall score and standard deviation for each question by accessing the corresponding columns in the DataFrame (df). The round function is used to round the calculated values to three decimal places.

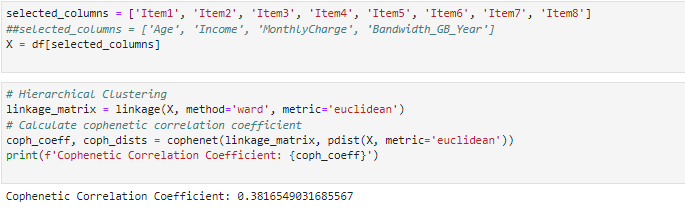
Finally, I use the print statement to display the results in a formatted string. Each iteration of the loop prints information about a specific question, including the question number, mean score, and standard deviation. This makes the output clear and concise, providing insights into the survey results for each question.

4.  Provide a copy of the cleaned data set.  
 

# **Part IV: Analysis**

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

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



I am performing hierarchical clustering on a dataset represented by matrix X. The linkage\_matrix is calculated using the 'ward' method and the 'euclidean' distance metric. Afterward, I compute the cophenetic correlation coefficient, a measure of how well the hierarchical clustering preserves the pairwise distances between original data points. The obtained coefficient is printed, and in this specific run, it is 0.3816549031685567.

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer code

Description automatically generated

I am performing hierarchical clustering on a DataFrame (df) using the Ward linkage method. The result is a linkage matrix, which captures the hierarchical structure of the clusters based on the input data. Subsequently, I am assigning cluster labels to each data point in the DataFrame.

To achieve this, I use the fcluster function, specifying 3 as the number of clusters to form. The criterion='maxclust' parameter indicates that the number of clusters is determined by the previously specified value of 3. The resulting cluster labels are stored in a new column named 'ward\_cluster\_labels' within the DataFrame.

This output signifies that there are 3085 data points assigned to cluster 1, 3742 data points to cluster 2, and 3173 data points to cluster 3 based on the hierarchical clustering process using the Ward linkage method.

A screenshot of a computer code

Description automatically generated

I begin by setting the figure size to [20, 10] using plt.figure(figsize=[20, 10]). Subsequently, I use a loop to create subplots for each of the eight survey items. These subplots are arranged in a 2x4 grid, and for each subplot, I set the title to indicate the distribution of scores for a specific survey item based on cluster labels. The countplot function from the Seaborn library is employed to visualize this distribution, with distinct colors representing different clusters. Legends, x-axis labels (indicating survey item scores), and y-axis labels (representing the number of customers) are appropriately set.

Following the visualization, I adjust the layout using plt.tight\_layout() to ensure a clean and organized appearance of the subplots. The subsequent part of the code involves calculating the mean scores for each survey item within each cluster. A nested loop iterates through each survey item and each cluster label, computing the average score for that specific combination. The output is a series of print statements, providing insights such as, "For Item1, respondents from Cluster 1 scored this at 6.272, on average," revealing the average scores for each survey item within each identified cluster. These average scores serve as a valuable summary, offering a nuanced understanding of how different clusters of respondents perceive and score various survey items.

A screenshot of a document

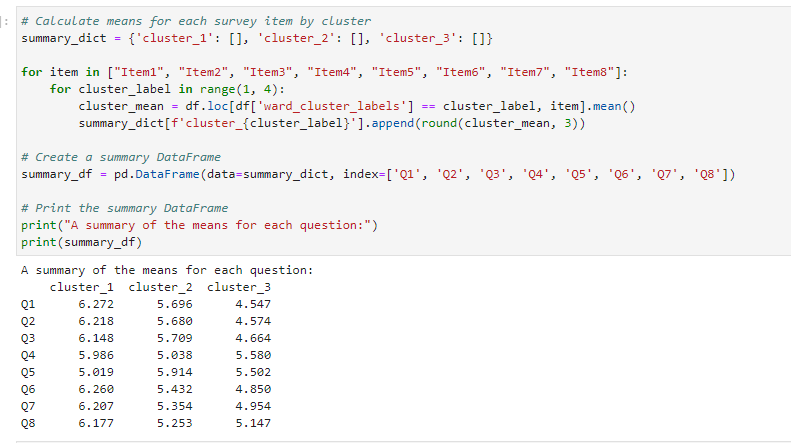
Description automatically generated

A close-up of a graph

Description automatically generated

A close-up of a graph

Description automatically generated



I am calculating the means for each survey item within three participant clusters and summarizing the results in a DataFrame. Firstly, I initialize an empty dictionary summary\_dict with keys representing each cluster ('cluster\_1', 'cluster\_2', 'cluster\_3') and empty lists as their corresponding values.

Next, I use nested loops to iterate through each survey item and each cluster label (ranging from 1 to 3). Within these loops, I calculate the mean score for the current survey item within the specified cluster using the Pandas DataFrame df. The calculated means are rounded to three decimal places and appended to the corresponding cluster's list in summary\_dict.

After gathering the means, I create a summary DataFrame named summary\_df using the collected data in summary\_dict. The DataFrame is structured with survey items ('Q1' to 'Q8') as rows and clusters ('cluster\_1', 'cluster\_2', 'cluster\_3') as columns.

Finally, I print the summary DataFrame to the console using print(summary\_df). The output presents a clear summary of the means for each question across the participant clusters, providing a concise overview of how different clusters perceive and score each survey item.

A graph with lines and numbers

Description automatically generated with medium confidence

I'm using Matplotlib and Seaborn to create a line plot visualizing the distribution of average survey scores per question across participant clusters. I set the figure size, specify the data (summary\_df), and add markers to the line plot for better visibility. The title, x-axis label ("Survey Question Number"), and y-axis label ("Mean Score") provide context to the visualization. This line plot offers a concise overview of how average scores vary across different survey questions and participant clusters.

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

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a document

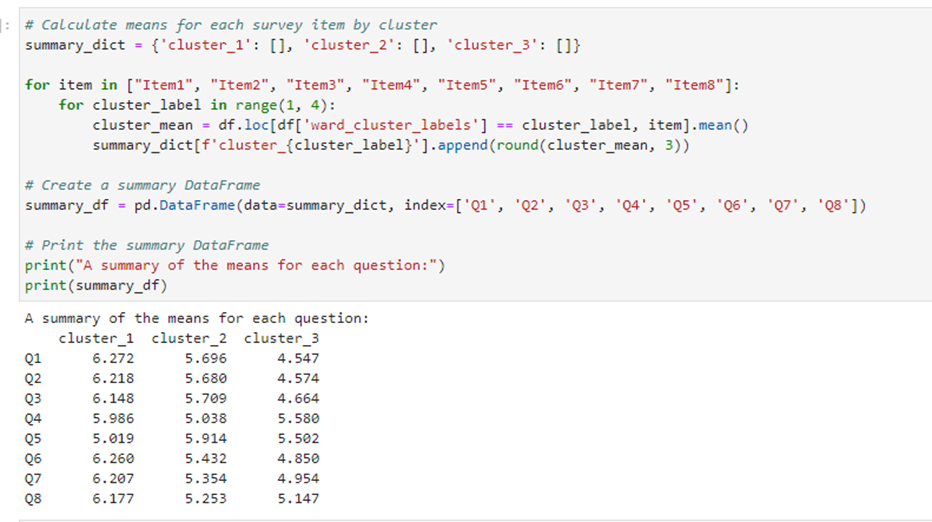
Description automatically generated

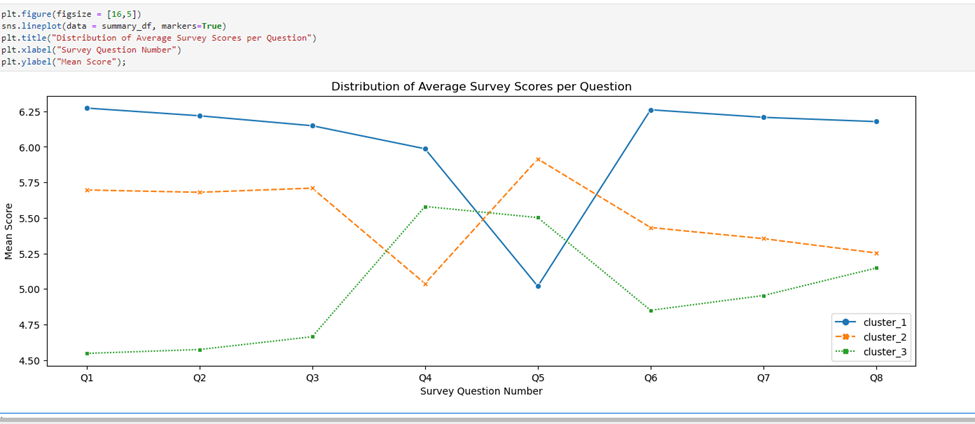
A close-up of a graph

Description automatically generated

A close-up of a graph

Description automatically generated





# **Part V: Data Summary and Implications**

## E.  Summarize your data analysis by doing the following:

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

A white background with text

Description automatically generated

The clusters created in this hierarchical clustering analysis exhibit a silhouette score of 0.108. Interpreting this score suggests that the clusters are somewhat separated, but the value is relatively low, indicating potential overlap or less distinct boundaries between clusters. A higher silhouette score, closer to 1, would imply more well-defined and clearly separated clusters. The assessment of cluster quality is context-dependent, and factors like the dataset's nature and the chosen clustering method play a role. In this case, further exploration and refinement may be necessary to enhance the quality of the clusters, such as experimenting with different clustering algorithms or adjusting the features used in the analysis.

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

A screenshot of a computer code

Description automatically generated

A screenshot of a graph

Description automatically generated

Analyzing the results of the clustering analysis provides valuable insights into the distinct patterns of responses across different clusters. Looking at the mean scores for each survey question in clusters labeled cluster\_1, cluster\_2, and cluster\_3, notable trends emerge. Questions Q1 to Q3 exhibit a descending trend in mean scores, with cluster\_1 consistently having the highest scores, followed by cluster\_2 and then cluster\_3. This suggests varying preferences or perspectives among the clusters for these initial questions.

Moving on to question Q4, there is a divergence in mean scores, with cluster\_1 and cluster\_3 scoring higher than cluster\_2. This indicates differences in sentiments or opinions on this particular question. Question Q5 stands out, with cluster\_2 showing the highest mean score, suggesting a distinctive perspective or preference compared to the other clusters.

Examining questions Q6 to Q8, a similar descending trend in mean scores across clusters is observed. Generally, respondents in cluster\_1 tend to provide higher scores, reflecting a more positive outlook or response pattern. Cluster\_2 exhibits mixed responses, indicating nuanced sentiments on specific aspects. Conversely, respondents in cluster\_3 consistently provide lower mean scores, suggesting a potentially less favorable response pattern.

Understanding the characteristics of each cluster is crucial for drawing meaningful implications. Respondents in cluster\_1 consistently express positive sentiments, while those in cluster\_2 demonstrate a mixed response pattern. Cluster\_3, on the other hand, tends to provide lower scores across most questions, indicating a potentially less favorable viewpoint.

These findings can be interpreted to inform targeted strategies or interventions based on the preferences and sentiments expressed by each cluster. For a more nuanced understanding, further investigation into the specific features contributing to the observed patterns would be valuable. In conclusion, the clustering analysis provides a structured approach to understanding respondent preferences, offering insights that can guide tailored strategies and decision-making based on the distinct characteristics of each cluster.

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

One limitation of the analysis is the sensitivity to the choice of clustering algorithm and parameters. In hierarchical clustering, the selection of the linkage method, distance metric, and the number of clusters can significantly impact the resulting clusters and, consequently, the interpretation of the analysis. If alternative clustering methods or parameter configurations were considered, the outcomes might differ, leading to potentially different cluster structures and interpretations.

Additionally, the analysis focuses on survey item responses as features for clustering. While this provides insights into respondent preferences, it may overlook other relevant factors or features that could contribute to more nuanced interpretations.

Furthermore, the interpretation of clusters relies on the assumption that similar mean scores within a cluster represent similar attitudes or preferences. However, the analysis does not account for potential variations or heterogeneity within clusters, and individual differences within each cluster may exist. This limitation highlights the need for careful consideration and validation of the results to ensure they align with the underlying characteristics of the data.

In conclusion, while hierarchical clustering provides valuable insights, its sensitivity to algorithmic choices, assumptions about data structure, and potential oversight of important features represent limitations that should be acknowledged and considered in the interpretation of the analysis results.

### 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.

I suggest starting by refining our survey questions, ensuring they align with our organizational goals and truly capture crucial aspects of the customer experience. After cleaning and preprocessing our survey data, it's vital to normalize or standardize the data for comparability. Additionally, considering feature selection or dimensionality reduction techniques will help us focus our clustering analysis on the key factors influencing customer segmentation.

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

Bock, T. (2020, December 9). *What are the Strengths and Weaknesses of Hierarchical Clustering?* Displayr. <https://www.displayr.com/strengths-weaknesses-hierarchical-clustering/>

*Clustering using AgglomerativeClustering and silhouette scoring*. (n.d.). Gist. <https://gist.github.com/bitsnaps/12415200fc62539fff852a1b46168d0a>

*Dendrograms*. (n.d.). <https://plotly.com/python/dendrogram/>

Ellis, C. (2022, June 8). *When to use hierarchical clustering - Crunching the Data*. Crunching the Data. <https://crunchingthedata.com/when-to-use-hierarchical-clustering/>

Firdose, T. (2023, December 8). Hierarchical clustering in Python: unraveling data structures. *Medium*. <https://tahera-firdose.medium.com/hierarchical-clustering-in-python-unraveling-data-structures-b646b5abb8df>

Girgin, S. (2021, December 9). Hierarchical Clustering Model in 5 Steps with Python. *Medium*. <https://medium.com/@sametgirgin/hierarchical-clustering-model-in-5-steps-with-python-6c45087d4318>

*Hierarchical Clustering in Python: A Comprehensive Implementation Guide*. (2024, January 8). Quantitative Finance & Algo Trading Blog by QuantInsti. <https://blog.quantinsti.com/hierarchical-clustering-python/>

Keita, Z. (2023, January 19). *An introduction to hierarchical clustering in Python*. <https://www.datacamp.com/tutorial/introduction-hierarchical-clustering-python>

Koli, S. (2023, March 2). How to evaluate the performance of clustering algorithms using Silhouette Coefficient. *Medium*. <https://medium.com/@MrBam44/how-to-evaluate-the-performance-of-clustering-algorithms-3ba29cad8c03>

Pai, P. (2022, January 6). Hierarchical clustering explained - Towards Data Science. *Medium*. <https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8>

*Plot Hierarchical Clustering Dendrogram*. (n.d.). Scikit-learn. <http://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html>

*Python Machine Learning - Hierarchical Clustering*. (n.d.). <https://www.w3schools.com/python/python_ml_hierarchial_clustering.asp>

Sharma, P. (2023, October 23). *What is Hierarchical Clustering in Python?* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/>

Shukla, P. (2023, November 23). *Hierarchical clustering in machine learning*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/11/hierarchical-clustering-in-machine-learning/>