**Performance Assessment: Task 1**

**A1. Proposal of Question**

My research question for this performance assessment is, “Can patients with similar characteristics be grouped together using the k-means clustering technique?”

**A2. Defined Goal**

The goal of this analysis is to determine whether k-means clustering is an effective method for identifying groups of patients with similar characteristics. This can help the hospital better understand their patients and lower their cost of care.

**B1. Explanation of the Clustering Technique**

K-means clustering is an unsupervised machine learning algorithm used to group a dataset into k clusters, with ‘k’ being the number of labeled data points, or centroids, around which the clusters are formed. The dataset points are assigned to the k centroid with the nearest Euclidean distance. The centroids are then recalculated, and the process repeats until there are no significant changes to the centroids.

**B2. Summary of the Technique Assumption**

One assumption of the k-means clustering technique is that all clusters have a similar spread, or variance. To meet this assumption, the data must be normalized prior to applying k-means.

**B3. Packages or Libraries List**

For this assignment, I chose to use Python via Jupyter Lab. I used the pandas package to load the data, evaluate my model, and create dummy variables. I used the NumPy package to impute the outliers with nulls and replace with the median value. The matplotlib and seaborn packages were both used to create visualizations. The scikit-learn package was used to normalize the data, create my k-means model, and calculate the silhouette score of my model.

**C1. Data Preprocessing**

One data preprocessing goal relevant to k-means clustering is standardization of the data. This is done to meet the assumption of homogeneity of variance. To accomplish this goal, the scikit-learn StandardScaler package was utilized.

**C2. Data Set Variables**

The initial data set variables I will use to perform the analysis are:

* Income (continuous)
* Total Charges (continuous)

**C3. Steps for Analysis**

The first step in preparing the data was to obtain a summary of all the variables using the .head() and .info() functions. I used the .duplicated().value\_counts() function to assess for duplicates. Next, I used the .isnull().sum() function to assess for nulls. The matplotlib and seaborn packages were utilized to examine the distribution of the variables and to check for outliers. I imputed the outliers found in ‘Income’ with the median value. I normalized the data using the .fit\_transform() function and the StandardScaler package. Once normalization was confirmed, I exported the cleaned data to a csv file using the .to\_csv() function.

**C4. Cleaned Data Set**

See ‘D212\_MV\_medical\_clean.csv’ file for cleaned data set.

**D1. Output and Intermediate Calculations**

I built my initial model using the KMeans package of sklearn using a k-value of 3. I evaluated the model by calculating and plotting the centroids and clusters of my initial model. I then utilized the within cluster sum of squares (WCSS) standard to create an ‘elbow’ plot of my model. According to that plot, the optimal k-value was somewhere between 3-5. Since there was not a clear optimal k-value, I also calculated the silhouette score to create a silhouette plot. The silhouette plot clearly demonstrated that the optimal k-value was 3. Based on the methods used determine the optimal k-value, I chose to keep my initial model with a k-value of 3.

**D2. Code Execution**

See below for code execution of cluster technique analysis:

A screenshot of a computer

Description automatically generated  
A screen shot of a graph

Description automatically generatedA black screen with a black rectangle

Description automatically generatedA screen shot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA black screen with colorful text

Description automatically generatedA graph with a line

Description automatically generatedA screenshot of a computer program

Description automatically generatedA graph with a line

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**E1. Quality of the Clustering Technique**

The quality of my k-means clustering model was validated by the silhouette score. The silhouette score of my model was calculated to be 0.52. The closer to 1 a silhouette score is, the higher the quality of the model and the closer to -1, the lower the quality. Since my score is above 0.5, my model is generally considered to be a high quality model.

**E2. Results and Implications**

My final model resulted in three clusters. In cluster ‘0’, we see patients with a lower income and lower total charges. In cluster ‘1’, we see patients lower income but higher total charges. In cluster ‘2’, we see higher income patients with both high and low total charges which average out to the mean of the total charges. The breakdown of each cluster by gender and age appeared to be similar to the total population. No notable differences for these measures were observed.

**E3. Limitation**

One limitation of my data analysis was that the model only examined two of the 50 variables in the medical data set. However, adding more than two variables to a k-means clustering analysis can make visualizing and interpreting the model more challenging.

**E4. Course of Action**

As mentioned above, this analysis only examined two of the 50 variables in this data set. I would recommend expanding the analysis to determine if there is another set of variables whose model may have a higher silhouette score than my initial model. I would also recommend expanding on the analysis of my initial model to compare the averages of more variables in addition to gender and age.

**F. Panopto Video of Code** See ‘D212 Task 1- Vicente Panopto Recording’ for link to Panopto video recording.

**G. Sources for Third-Party Code**

Kamara, Kesselly. *D212 Task 1 Overview & Tutorial.* Retrieved December 5, 2024.

**H. Sources**

Kamara, Kesselly. *D212 Task 1 Overview & Tutorial.* Retrieved December 5, 2024.