Saemi Ramirez

011926418

D212 PA 1 Clustering Techniques

9/28/2024

WGU

1. Research Question
   1. Research question and clustering technique  
      My research question is “How many groups can k-means cluster form based on the numerical features such as “Initial\_days” and “TotalCharge”?”
   2. Define a goal  
      The goal of using k-means clustering for this research question is to identify patterns or groups among patients based on the numeric variables selected
2. Reasons of chosen clustering technique
   1. Explain how the chosen clustering technique will analyze the selected dataset & Expected outcome  
      K-means clustering technique is to partition a set of n observations into k-clusters, where each observation is assigned to the cluster with the closest mean or centroid, which serves as the representation of that cluster.  
      K-means clustering begins by randomly selecting K points from the dataset. Each data point is then assigned to the cluster with the nearest centroid, based on the calculated distance between the points and centroids. Once all the points are assigned, the centroids are recalculated. This process is repeated until the algorithm converges. (PulkitS)   
      The expected outcome will be the dataset grouped by the optimal number of clusters that is determined by the elbow method and silhouette score.
   2. One assumption of the clustering technique   
      The assumption that the anisotropic data points indicate clusters that are elongated and non-spherical, with varying spreads across different dimensions. This breaks the algorithm’s assumption of spherical clusters, leading to reduced accuracy. (Demonstration)
   3. List the packages or libraries used in Python and justify

|  |  |
| --- | --- |
| **Packages/Libraries** | **Purpose** |
| Pandas | To handle the dataset/dataframe |
| Numpy | To perform mathematical operations or values |
| Matplotlib.pyplot | To visualize the data via graphics |
| Seaborn | To create the statistical graphics |
| StandardScaler from sklearn.preprocessing | To normalize the numeric values by reducing the size of the data to match the original form |
| KMeans from sklearn.cluster | To cluster similar data points and uncover hidden patterns or structures within the dataset |
| Silhouette\_score from sklearn.metrics | To calculate the average silhouette coefficient for each sample across different clusters |

1. Perform data preparation
   1. Describe one data preprocessing goal relevant to the clustering technique
   2. Identify the initial data set variables (continuous or categorical)

|  |  |  |  |
| --- | --- | --- | --- |
| **Initial\_days** | continuous | **TotalCharge** | continuous |

* 1. Explain each of the steps used to prepare the data for the analysis
     1. The medical\_clean.csv dataset was imported, checked the general information about the dataset using info()  
        A screenshot of a computer

        Description automatically generated
     2. Medical\_clean.csv was filtered where HighBlood was “Yes”  
        A screenshot of a computer code

        Description automatically generated
     3. Created 2 new data frames
        1. New\_df with Gender, Age, Income, Initial\_days, and TotalCharge for the final result
        2. New\_df2 with Age, Income, Initial\_days, and TotalCharge for KMeans calculation and graphs  
           
     4. Generated a heatmap using new\_df2 to find what 2 columns are the most correlated. Initial\_days and TotalCharge were found that their correlation score is 0.99  
        A screenshot of a graph

        Description automatically generated
     5. Created another data frame new\_df3 with Initial\_days and TotalCharge only. Graphed the scatterplot between Initial\_days and TotalCharge with new\_df3  
        A screen shot of a graph

        Description automatically generated
     6. Normalized new\_df3 using StandardScaler()  
        A screenshot of a computer

        Description automatically generated
     7. Check if the standard deviation is 1.00 for both Initial\_days and TotalCharge using describe() function  
        A screenshot of a calculator

        Description automatically generated
  2. Provide a copy of the cleaned data set  
     cleaned\_dataset\_copy.csv is submitted

1. Perform analysis and report on the results
   1. Determine the optimal number of clusters in the data set and describe the method used to determine this number
      1. The elbow method was used to determine the optimal number of clusters. It seems to be the 2 is the best number for the clusters  
         A screenshot of a computer code

         Description automatically generated  
         A graph with a line

         Description automatically generated
      2. Then used the silhouette method to confirm if 2 is the best fit  
         A screenshot of a computer program

         Description automatically generated  
         A graph with blue dots

         Description automatically generated
      3. The final\_model was created with n\_clusters of 2, n\_init of 25, and random\_state of 30  
         A screenshot of a computer

         Description automatically generated
      4. Centroid was also collected using the cluster\_centers\_  
         A computer screen shot of a computer code

         Description automatically generated
      5. Final graph was plotted with colors and clusters  
         A screen shot of a computer code

         Description automatically generated

A blue and orange dotted line

Description automatically generated

* + 1. Added a ‘Cluster’ column to new\_df dataframe then used the get\_dummies for the deeper analysis without dropping the first column  
       A screenshot of a computer

       Description automatically generated  
       A screenshot of a computer

       Description automatically generated
    2. Statistically analyzed each column using either mean (Gender\_female, Gender\_Male, and Gender\_Nonbinary) or median (Age, Income, Initial\_days, and TotalCharge) before the last analyze with clusters  
       A screenshot of a computer code

       Description automatically generated
    3. Final statistical analyze chart which is grouped each column by cluster  
       A screenshot of a computer

       Description automatically generated
  1. Provide the code used to perform the clustering analysis technique  
     “Saemi Ramirez D212 PA1 – Clustering Techniques.ipynb” is submitted

1. Summarize the data analysis
   1. Explain the quality of the clusters created  
      As the screenshot below, the accuracy is quite high as little over 80% when the clusters are set to 2. However, the accuracy drops drastically to approximately 67.5% and 54% when the number of the clusters are set to 3 and 4.   
      A graph with blue dots

      Description automatically generated
   2. Discuss the results and implications of the clustering analysis  
      Before proceeding with the clustering analysis, a general statistical overview is provided using median for “Initial\_days” and “TotalCharge.”  
      Patients were hospitalized for an average of 31 days during their initial visit, with an average daily charge of approximately $5,155.

A screenshot of a computer

Description automatically generated  
  
The screenshot below is the statistics of the each group. The first cluster, which is the cluster 0, patients were hospitalized 7.6 days and approximately charged $3,222 per day in average. The second cluster, which is the cluster 1, patients were hospitalized 61 days and approximately charged $7,549 per day in average.  
  
A screenshot of a computer

Description automatically generated

* 1. Discuss one limitation of your data analysis  
     One of the limitations was that the categorical features were excluded for k-means clustering method which could group the patients better with more detailed information
  2. Recommend a course of action for the real-world organizational situation

I would recommend the WGU medical team to add more numerical variables on their dataset because “Initial\_days” and “TotalCharge” were correlated to each other with 0.99 but the next highest correlated score was 0.0037 with “Age” and “Initial\_days”

1. Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=601a55c0-1dfd-4e35-b9a0-b1f900fbb30f>
2. Third-party code to support the analysis
   1. *Machine Learning – K-means*. W3schools. <https://www.w3schools.com/python/python_ml_k-means.asp>.
   2. Ankita. *K-Mean: Getting the Optimal Number of Clusters*. Analytics Vidhya. (July 02, 2024). <https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters>.
3. Third-party in-text citations and references
   1. *Demonstration of K-Means Assumptions*. GeeksforGeeks. (December 09, 2023). <https://www.geeksforgeeks.org/demonstration-of-k-means-assumptions>.
   2. PulkitS. *Introduction to K-Means Clustering*. Analytics Vidhya. (August 22, 2024). <https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering>.