Taylor Cina

D212 – Data Mining II

***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*

What relationships, if any, exist between customer tenure and monthly charge that can be classified using k-means clustering?

*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 of this analysis is to determine if clustering relationships exist between customer tenure and monthly charge.

***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.*

K-means clustering is an unsupervised machine learning method that splits data into clusters and classifies them without existing knowledge of data labels. The centroid is defined as the mean of all datapoints within the cluster and is decided iteratively using random initial conditions. Each iteration assigns points to their closest centroid and ends when the centroids converge, or the maximum specified iterations is reached. The result is a list of cluster centroids that have the smallest distance to all points within their cluster.

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

K-means clustering assumes each cluster has a similar shape and size. For datasets with significantly different cluster sizes, k-means may not be appropriate.

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

Pandas – dataframes.

Matplotlib – base implementation on which python graphs are built.

Seaborn – pretty graphs with simple implementation.

Sklearn – 3 libraries are used here.

-StandardScaler – used to scale the dataset to improve the performance of the K-means cluster analysis.

-KMeans – K-means cluster analysis model package.

-silhouette\_score – used to determine the optimal number of clusters in the k-means analysis.

***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.*

Data used for k-means clustering performs much better when standardized. The data will be scaled using a standard scaler from scikit-learn.

*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.*

The only two variables used are Tenure and MonthlyCharge, both of which are continuous. Using categorical variables for unsupervised cluster analysis is not effective.

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

1.Data is loaded into a dataframe using pandas. .info() and .describe() are used to give an overview of the data set.

2. The data is subset to only include the Tenure and MonthlyCharge columns.

3. The data is scaled using the StandardScaler() from sklearn preprocessing.

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

Cleaned dataset is attached as the file “TCina D212 T1 Scaled.csv”

***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.*

The optimal number of clusters in the data set is 4, which was determined using silhouette scores.

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

Code is attached as the file “TCina D212 T1.ipynb”

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

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

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

The quality of the clusters was determined using silhouette scores, the average score was near 0.50, which is acceptable. Silhouette scores are a measure of cluster distinction and overlap. Scores near 1 are good, scores near -1 indicate overlap and indistinct clusters. The clusters generated by this analysis have some overlap but are distinct enough to be useful.

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

The clustering analysis generated 4 clusters with centroids in a rectangular formation pictured below.

A screenshot of a graph

Description automatically generated

The clusters can be roughly categorized as follows:

-low tenure, low charge

-low tenure, high charge

-high tenure, low charge

-high tenure, high charge

This implies that most customers fall into four categories depending on their tenure and monthly payment, with very little customers being near the middle ranges of tenure, but normally distributed across their monthly charges. The distribution across monthly charge has overlap with other clusters but is decently distinct.

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

Cluster analysis works best using continuous variables. The vast majority of variables in the dataset are either discrete numerical variables or categorical variables. There is a lot of extra information in the set that is not utilized by using k-means cluster analysis.

*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.*

The cluster labels can be applied to the dataset and analysis can be run on members of each cluster as their own dataset in order to gain better understanding of the properties of each cluster. Of most importance are the high monthly payment customers, and understanding this demographic could help retain more of them.

***Part VI: Demonstration***

*F.  Provide a Panopto video recording that includes the presenter and a vocalized demonstration showing all code used, the code being executed, and the results of all code used in the task.*

Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=19b2edf6-3392-4156-a67e-b21f0041807f>

*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.*

No web sources used.

*H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.*

No external references used.