**Student: Seif Kungulio**

**Date: 04/27/2025**

**Subject: Project 7**

**Class: DSCI 512**

**Section: 01W**

**Instructor: Dr. Nengbing Tao**

**File Name: Project7\_Kungulio\_Seif.docx**

1. Load the dataset CreditCards.csv into memory.

A computer screen shot of a computer program

AI-generated content may be incorrect.

1. Perform the k-means cluster analysis
   1. Remove the first column: CUST\_ID since it doesn’t provide any info for cluster.

A computer screen with white text

AI-generated content may be incorrect.

* 1. Determine the optimal number of clusters. Justify your answer. It may take longer running time since it uses a large dataset.

A graph showing the growth of a number of statistic

AI-generated content may be incorrect.

The Gap Statistic Method determines the ideal number of clusters (k) by selecting the value that maximizes the gap statistic, a measure of clustering effectiveness. In this analysis, the gap statistic consistently rises and peaks at k = 7, signaling the best trade-off between cluster compactness and separation. Any increase in k beyond 7 offers little benefit, indicating diminishing returns. The error bars support the reliability of k = 7, confirming it as the most meaningful and appropriate segmentation for the dataset.

* 1. Perform k-means clustering using the optimal number of clusters.

A screenshot of a computer program

AI-generated content may be incorrect.

The k-means clustering analysis grouped 8,950 data points into 7 distinct clusters, each represented by a central point (centroid). The clustering result shows strong quality, with tightly grouped data (low within-cluster sum of squares) and clear separation between clusters (high between-cluster sum of squares). The algorithm completed successfully in just one iteration, with no errors, indicating the data was highly suitable for this clustering approach.

* 1. Visualize the clusters in different colors.

A colorful dots and circles

AI-generated content may be incorrect.

The visualization displays 7 clusters formed through dimensionality reduction (Dim1 and Dim2), which together account for 47.6% of the data’s variability. Certain clusters, such as Cluster 5, are clearly distinct and well-separated, while others, like Clusters 6 and 7, show slight overlap, suggesting some shared features. Tightly packed clusters like Cluster 2 indicate strong similarity among points, whereas broader clusters like Cluster 6 reflect more variation. Overall, the plot supports the clustering structure but also reveals areas—particularly where clusters overlap—that may benefit from further refinement.