**CLUSTERING TECHNIQUES (OFM4 TASK 2)– D212**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

***Part I: Research Question***

**A-1**

**What are the features that a principal component analysis would determine as the best while retaining most of the variance?**

**A-2**

One goal of this analysis is to achieve dimensionality reduction. This takes a large number of features and reduces it to simplify the analysis while keeping most of the variance.

***Part II: Method Justification***

**B-1**

**The PCA simplifies the continuous variables to find the principal components. It requires the data to be standardized to allow each variable to have an equal effect on the analysis. The covariance matrix shows the features covariance to each other. This indicates if two features have correlate or if they vary independently. Eigenvectors and eigenvalues of the covariance matrix are calculated to show the directions the data vary. The principal components are selected next based on their eigenvalues. The principal components are ordered where the most variance is the first one and the second most variance is the second one and this continues. The PCA then reduces the dimensionality to the top principal components that explains most of the variance in the data. Lastly the PCA transforms the data into a lower-dimensional space. This is what we can expect from the outcome of the PCA, a reduced number of features in a dataset that keeps most of its variance.**

**B-2**

In doing a PCA, it is assumed that the relationship between variables is linear. This is called linearity.

***Part III: Data Preparation***

**C-1**

**The continuous data set variables that are used to answer the PCA question from part A-1: Lat, Lng, Population, Age, Income, Outage\_sec\_perweek, Email, Contacts, Yearly\_equip\_failure, Tenure, MonthlyCharge, and Bandwidth\_GB\_Year.**

**C-2**

Please see attached d212scaledt2 csv file for cleaned data set.

***Part IV: Analysis***

**D-1**

**Please see below screenshot of the matrix for all the principal components.**

A table with numbers and symbols

Description automatically generated

**D-2**

Using the Kaiser criterion that drops components with eigenvalues that are less than 1, the total number of principal components is 6. I included the Eigenvalues below because it is difficult to see in the graph.

A graph with a line

Description automatically generated A screenshot of a computer screen

Description automatically generated

**D-3**

Below is the variance for each of the six principal components:

* PC1 – 16.62
* PC2 – 10.28
* PC3 – 8.75
* PC4 – 8.54
* PC5 – 8.46
* PC6 – 8.38

**D-4**

The total variance captured by the six principal components is 61.03.

**D-5**

The original PCA of 12 principal components had an explained variance of 100%. The principal component with the most explained variance, PC1, gives 16.62%. PC1 is heavily influenced by Tenure and Bandwidth\_GB\_Year with both having over 0.70 in the loading with no other original variable over 0.1. The Kaiser criterion did give a reduction from 12 principal components to six principal components. The results when running the PCA again with the 6 principal components, but it only had an explained variance of 61.03. This below what I wanted to see and would not be confident in recommending using it with almost 40% of the variance from the original data being lost.

***Part V: Attachments***

**E**

No web sources were used to acquire data or segments of third-party code.

**F**

No sources were used to cite in-text or quoted.

**G**

The content in this Performance Assessment is set up and presented with the highest professional standards.