**D212 Data Mining II**

Nicole Reiswig

College of Information Technology, Western Governors University

Master of Science in Data Analytics

Kesselly Kamara

February 21, 2024

**Table of Contents**

**Part I: Research Question**

1. Question
2. Goal

**Part II: Method Justification**

1. Method
2. Assumption

**Part III: Data Preparation**

1. Variables
2. Standardize

**Part IV: Analysis**

1. Matrix
2. Principal components
3. Variance
4. Total variance
5. Summary

**Part V: Attachments**

# Principal Component Analysis

**Part I: Research Question**

1. The research question that will be answered is, “What are the characteristics of our customers?” We will identify characteristics of our customer base to understand better who our customers are and how to best provide services to them. We will identify groups of customers with similar characteristics ultimately enabling better business and strategic decision-making. The clustering technique that will be deployed is the KMeans clustering technique.
2. The KMeans clustering technique analysis aims to answer our research question, “What are the characteristics of our customers?” We will utilize this technique to identify characteristics and groups of our customer base, ultimately enabling better business and strategic decision-making for the Telecom company.

**Part II: Technique Justification**

1. KMeans clustering is the appropriate technique to use when analyzing continuous data as we are in this clustering analysis of the WGU Telecom company. We are looking at our customers' income and length of tenure to identify if there are patterns and groups of customers. The clusters are determined by the distance between data points. Data points that are closer together are grouped into a cluster together. The measurement of distance between data points is calculated using Euclidean's distance. This approach can tell us if groups of customers' income affect the length of tenure. For example, is there a pattern of high-income earners with longer lengths of tenure or shorter lengths of tenure?
2. To perform the KMeans clustering technique, we must know the distance between observations. KMeans clustering assumes that the variance of the distribution of each attribute is spherical and all variables have the same variance. This is why the data is scaled in preparation for the clustering analysis.
3. We will utilize Jupyter Lab, Anaconda, and Python programming language to complete this analysis. I chose Python for this project as I’ve previously utilized R and want to be fluent in both programming languages to expand employment opportunities after graduate school. The libraries and packages utilized for this analysis are: Scikit learn is used for supervised and unsupervised machine learning algorithms, Pandas is used for data manipulation and analysis, Warnings are used to disable warnings, Matplotlib is a plotting library for Python and its mathematics extension NumPy, and Seaborn is a powerful tool used in Python to create statistical graphics. Scikit learn supports the analysis by providing the capability to perform classification, regression, clustering, and dimensionality reduction. In this project we leverage the capability to perform clustering using KMeans. Pandas supports the analysis by giving us the capability to create data frames, data cleansing, data normalization, joins, visualizations, statistical analysis, data inspection and loading and saving data. In this project we utilize pandas to view, create and save data frames as well as explore the data. Matplotlib was utilized to create plots in our analysis. Seaborn was utilized to create statistical graphs for the analysis. Each of these libraries supported the analysis and was a necessary part of the results.

**Part III: Data Preparation**

1. A data preprocessing goal relevant to this analysis was to scale the data utilizing the Sklearn StandardScaler function. This is a necessary preprocessing goal for the KMeans clustering technique to accurately measure the distance between data points without biases in the data due to formatting or differences in scales. It normalizes the data for a more accurate view of the data. This is necessary because KMeans clustering assumes that the variance of the distribution of each attribute is spherical and all variables have the same variance. This is why the data is scaled in preparation for the clustering analysis.
2. The initial data set variables used to perform the analysis for clustering to answer the research question listed in part A1 is ‘income’ which is continuous and ‘tenure’ which is also a continuous variable. KMeans requires continuous variables to be used in the clustering analysis.
3. The steps taken to prepare the data for analysis were to explore the data set, check for nulls in the data set, rows, and columns, create a scatterplot to view the relationship of the features to be used in the analysis, using the describe function to explore the features, scale the data to prepare for analysis, explore the scaled data, visualize the scaled data with a scatterplot, create the clusters and centroids, create an elbow plot, and view the silhouette score. There were no null values in the variables selected.

**Part IV: Analysis**

1. Based on the elbow plot we’ve evaluated using standard opinion that we could have 3 clusters from our chosen data ‘income’ and ‘tenure’ from the customers.csv file as that is where the elbow bends. In addition to the elbow plot, we also calculated the silhouette score which returned .505. The silhouette score can range from -1 to 1, where -1 means data points are not similar and 1 means there’s similarity. Our score of .505 means we have some similarities. In the initial centeroid we had 2 clusters 0 and 1. 0 had less income and less tenure and cluster 2 had more income and more tenure. In the second centeroid, we have 3 clusters 0,1, and 2. In cluster 0 we have less income and less tenure, in cluster 1 we have high income and less tenure, and in cluster 2 we have less income and more tenure. For the customers in cluster 0 we would want to keep an eye on them, cluster 1 we would want to improve their happiness to improve their tenure length, and cluster 2 we would want to retain and maintain this customers' happiness as they have longer tenures.

**Part V: Data Summary and Implications**

1. The quality of the clusters is based on the silhouette score. First, we created the elbow plot to find the appropriate number of clusters using the standard opinion and found that we could have 3 clusters from our chosen data ‘income’ and ‘tenure’ from the customers.csv file as that is where the elbow bends. In addition to the elbow plot, we also calculated the silhouette score which returned .505. The silhouette score can range from -1 to 1, where -1 means data points are not similar and 1 means there’s similarity. Our score of .505 means we have some similarities. The closer to 1 the silhouette score is the more well-defined the clusters are, 0 means there's overlapping and -1 means the clustering is not good. This is a relatively accurate cluster as it is closer to 1.
2. The results of the KMeans clustering analysis we have 3 clusters 0,1, and 2. In cluster 0 we have less income at -.33 and less tenure at -.96, in cluster 1 we have high income at 1.79 and less tenure at -.01, and in cluster 2 we have less income at -.33 and more tenure at .97. For the customers in cluster 0 we would want to keep an eye on them, we could assume that due to their income, they found a lower rate or could no longer afford services. In cluster 1 we would want to improve their happiness to improve their tenure length since they have higher income they have higher spending potential, if we can find a way to retain them it means positive cash flow. In cluster 2 we would want to retain and maintain these customers' happiness as they have longer tenures even with lower incomes. We could explore the cause of their retention further to help retain other customers or use word of mouth and positive reviews to reach new audiences with similar characteristics.
3. The limitation of this analysis is that it only looks at continuous features. There are many other features we could also analyze to find similar characteristics of our customers to improve their satisfaction and length of tenure.
4. The recommendation would be to improve the happiness of cluster 1 to improve their tenure length since they have higher income they have higher spending potential, if we can find a way to retain them it means positive cash flow. In cluster 2 we would want to retain and maintain these customers' happiness as they have longer tenures even with lower incomes. We could explore the cause of their retention further to help retain other customers or use word of mouth and positive reviews to reach new audiences with similar characteristics. Another recommendation could be to perform a hierarchical cluster analysis of the survey response variables and explore if there are similarities there.

**References:**

I utilized Dr. Kesselly Kamaras' WGU instructor videos to reference segments of code to perform this analysis.