**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 3
2. Goal 3

**Part II: Method Justification**

1. Method 3
2. Assumption 4

**Part III: Data Preparation**

1. Variables 4
2. Standardize 4

**Part IV: Analysis**

1. Matrix 4
2. Principal components 5
3. Variance 5
4. Total variance 5
5. Summary 6

**Part V: References 7**

# Principal Component Analysis

**Part I: Research Question**

1. The goal of this principal component analysis is to reduce the dimension to find the principal component with the most variance. The research question to be answered is, “Which features contribute the most to the variance observed in our data?” Principal component analysis will answer this research question by simplifying complex data and preserving the essential patterns. First, in PCA, we standardize the range of continuous variables, next we compute covariance to identify correlation. Eigenvalues and eigenvectors are calculated which determine which PCs explain the most variance in the data. We choose how many PCs to retain based on that explained variance, and lastly, we recast the data along the principal component axes.
2. PCA is a tool used in machine learning and data science to answer questions about dimensionality reduction and feature extraction. PCA represents linear combinations of original features, the first PC explains the most variance, the second explains the second most variance, etc. To interpret the PCs you must understand which features contribute to each PC. The expected outcome of this analysis is to fulfill the goal which is to reduce the dimensionality and find the principal component with the most variance as described above.

**Part II: Technique Justification**

1. Principal component analysis is the appropriate technique to use when analyzing continuous data as we are in this dimensionality reduction analysis of the WGU Telecom company. We are looking at our customers' continuous characteristics to identify which characteristics have the most impact. PCA begins with standardization of the continuous data to ensure each variable contributes equally to the analysis. Next, the covariance matrix is computed which helps us understand which variables correlate. Then, we calculate the eigenvalues and eigenvectors of the covariance matrix, explaining the directions that the data vary. Then, we create a vector that captures the most significant variations. Lastly, we recast the data along the principal component axes. The expected outcome of this analysis is to reduce the number of features in the dataset while retaining as much information as possible. PCA reduces the impact of irrelevant features.
2. Principal Component Analysis assumes a linear correlation between original features.

**Part III: Data Preparation**

1. The continuous variables used in this principal component analysis include population, children, age, income, tenure, monthly charge, bandwidth\_gb\_year, outage\_sec\_week, email, contacts, and yearly\_equip\_failure.
2. To prepare the data set for PCA the data was standardized. See the attached cleaned data set.

A screenshot of a computer

Description automatically generated

**Part IV: Analysis**

1. See the matrix of all principal components, there are 11 in this analysis. PC1 is most significantly impacted by population with a higher weight of .99, PC2 holds its weight in income with a score of .99, PC3 is mostly comprised from tenure with .05 and yearly equipment failure at .04, PC4’s weight is from outage seconds per week at .9, PC5s weight is from children at .9, PC6s weight is from monthly charge at .73, PC7s weight is from monthly charge at .6, PC8s weight is from yearly equipment failure at .9, PC9s weight is from age at .9, PC10s weight is from contacts at .9, and PC11s weight is from bandwidth at .7.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

1. There is a total of 11 principal components in this analysis, we use the elbow method scree plot to determine the number of most significant PCs. The elbow method in the scree plot indicates one elbow and the most significant PC is PC2. A screen shot of a computer

   Description automatically generatedA screen shot of a computer

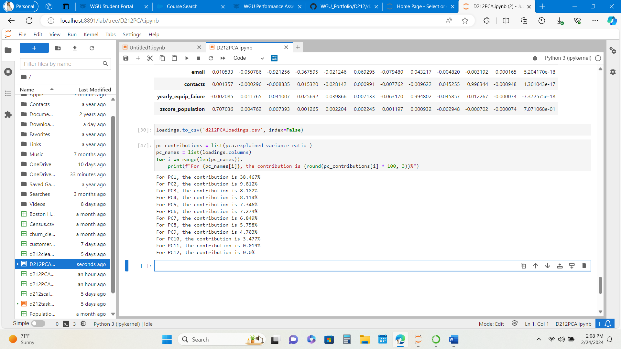
   Description automatically generated
2. The variance of each principal component is listed below in the screenshot. PC1 variance number percent is 23.821% and PC2 variance number percent is 12.148%.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

1. The sum of this PC variance is 35.969%. In the attached screenshot you can see PC1 variance percent is 23.821% and the PC2 variance percent is 12.148%. Combined they are 35.969.



1. To summarize the results, there are 11 principal components in this analysis. According to the elbow method scree plot there are two significant PCs which are PC1 and PC2. The variable that is most significant to PC1 is population. The variable that is most significant to PC2 is income. The variance of PC1 Is 23.821% and the variance of PC2 is 12.148. The total variance is thus 35.969%.

**References:**

I utilized Dr. Kesselly Kamaras' WGU instructor videos to reference segments of code to perform this analysis and D206 data cleaning course materials.

[Data Mining II D212 (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a88bed1d-12c6-45b7-9cde-aefc00d98bed&start=463.18608)

[PCA With Data Mining II - D212 (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=a7592d75-cc72-47fd-8a9a-b07a00efff97)

[Data Mining II - D212 PCA Part 2 (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8c618a96-fdb8-4757-abe2-b023018520ac)

[Data Cleaning | WGU-CGP-OEX](https://apps.cgp-oex.wgu.edu/wgulearning/course/course-v1:WGUx+OEX0026+v02/block-v1:WGUx+OEX0026+v02+type@sequential+block@f8572a9b3754417b85680e23bb7de7e6/block-v1:WGUx+OEX0026+v02+type@vertical+block@b768dbdb41dd41f9b3f865bf80ad3b1a)