Part I: Research Question

A. Describe the purpose of this data mining report by doing the following:

1. **Propose one question relevant to a real-world organizational situation that you will answer by using principal component analysis (PCA).**
   What is the smallest number of customer components we can analyze while still maintaining data integrity?

2. **Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.**
   One goal of the data analysis is to reduce the dataset to the optimal number of components for easier customer segmentation.

Part II: Method Justification

B. Explain the reasons for using PCA by doing the following:

1. **Explain how PCA analyzes the selected data set. Include expected outcomes.**
   PCA takes many dimensions and compresses them into fewer dimensions, attempting to lose the least amount of data possible in the process. Expected outcomes are the compressed variables comprised of mixtures of the initial variables. (Pramoditha, 2021)

2. **Summarize one assumption of PCA.**
   The variables used in the analysis must be continuous and standardized.

Part III: Data Preparation

C. Perform data preparation for the chosen dataset by doing the following:

1. **Identify the continuous dataset variables that you will need in order to answer the PCA question proposed in part A1.**
   Variables include Children, Population, Income, Tenure and Yearly_equip_failure.

2. **Standardize the continuous dataset variables identified in part C1. Include a copy of the cleaned dataset.**
   A copy of the cleaned dataset is provided in the Keim D212 Task Two Clean Data document.

Part IV: Analysis

D. Perform PCA by doing the following:

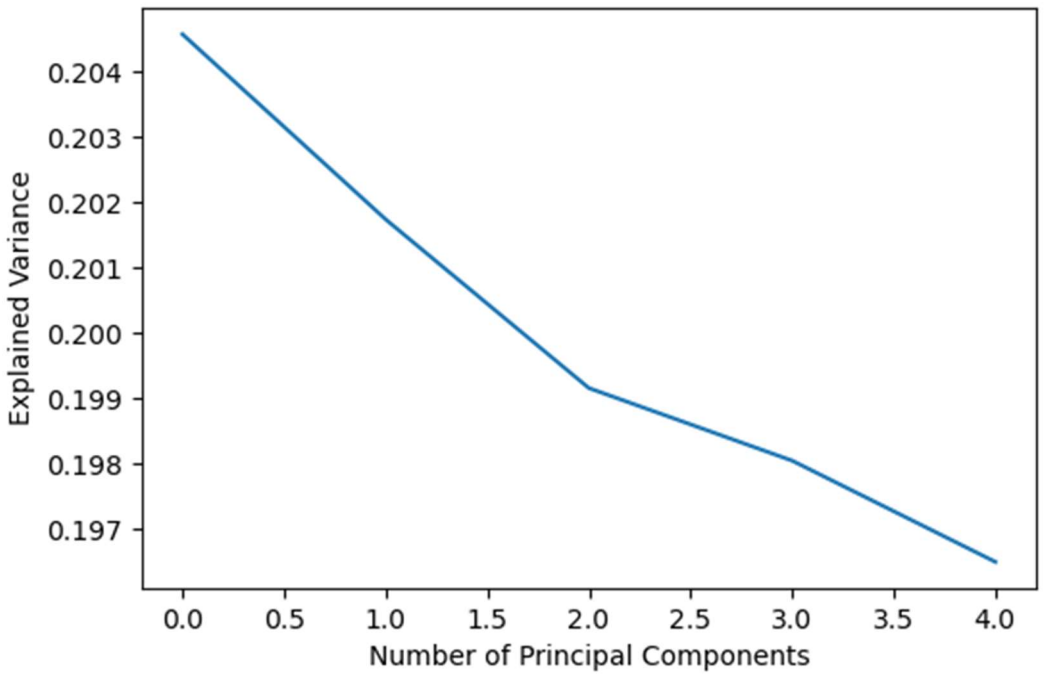
1. **Determine the matrix of *all* the principal components.**

```
[15]: loadings = pd.DataFrame(pca_5.components_.T,
      columns=['PC1','PC2','PC3','PC4','PC5'],
      index=reduced_variables.columns)
      loadings
```

[15]:

	PC1	PC2	PC3	PC4	PC5
Population	-0.450657	0.118529	0.727113	0.498767	-0.073474
Children	0.434291	-0.497231	0.469973	-0.252458	-0.528718
Income	0.508293	-0.274749	-0.162289	0.784294	0.157150
Yearly_equip_failure	0.510735	0.379836	0.462487	-0.217950	0.577475
Tenure	0.298476	0.720382	-0.100998	0.157724	-0.597398

2. **Identify the *total* number of principal components using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.**
   Using the elbow rule in conjunction with the total variance percentage, I've identified four principal components.



3. **Identify the variance of *each* of the principal components identified in part D2.**
   Component one has a variance of 20.46%, component two has a variance of 20.17%, component three has a variance of 19.92% and component four has a variance of 19.80%.

4. **Identify the *total* variance captured by the principal components identified in part D2.**
   The total variance captured by the four principal components is 80.35%.

5. **Summarize the results of your data analysis.**
   Reducing the data to four principal components takes the dimensionality from fifty to four, and still explains 80.35% of the data's variance, making the data much easier to segment and visualize.

Third Party Code Citation: (Pramoditha, 2021)

Works Cited

Pramoditha, R. (2021, December 21). *Principal Component Analysis (PCA) with Scikit-learn*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/principal-component-analysis-pca-with-scikit-learn-1e84a0c731b0