Tucker Atwood
WGU MSDA
D212 Task 2
5/24/24

A1. Research Question

Using the provided telecommunications churn data set, I will answer the question, **"Can a Principal Component Analysis be used to reduce dimensionality and create a more efficient set of variables for analyzing telecom customers?"**

A2. Analysis Goal

The **goal of this analysis** is to use a Principal Component Analysis (PCA) to reduce dimensionality within the churn data set, allowing for a **simpler and more efficient** analysis of the quantitative continuous variables while **preserving much of the variance** explained by the original variables. This analysis will allow the telecom company to be better equipped in **understanding customer characteristics**, which will help inform future decisions and business strategies.

The variables to be reduced in the PCA will be:

- **Latitude** (represented as "Lat"), the latitude coordinate of each customer's location.
- **Longitude** (represented as "Lng"), the longitude coordinate of each customer's location.
- **Income**, the reported income of each customer in US dollars.
- **Outage seconds per week** (represented as "Outage_sec_perweek"), the estimated number of seconds each customer has experienced of system outage per week.
- **Tenure**, the number of months each customer has stayed with the service.
- **Monthly charge** (represented as "MonthlyCharge"), the average amount in US dollars charged to each customer's account per month.
- **Bandwidth gigabytes per year** (represented as "Bandwidth_GB_Year"), the estimated number of gigabytes used in data by each customer per year.

B1. PCA Technique

The Principal Component Analysis technique will be applied to the churn data set and used for analysis in the following manner:

- The continuous variables to be included in PCA are **scaled** by subtracting the mean value from each observation and dividing by the standard deviation. Section C2 includes further discussion of this step, as well as the resulting prepared data set.
- The PCA method uses the covariances between the scaled continuous variables (their correlations to each other) to create **principal components**, which are combinations of the original variables that aim to explain as much of the variation in the original variables as possible. The components are produced one at a time, with the first (PC1) capturing as much variation as possible from combinations of the original variables, and the second

(PC2) capturing as much of the remaining variance as possible, and so on. PC1 will inherently explain more variation than PC2, which will have more than PC3, and so on. The same number of components are created as original variables included; in this analysis, there are seven original variables, so **seven principal components are produced**.

- Statistics will be calculated and compared for each of the seven principal components, including:
    - The **variance of each individual component**, expressed as a percentage of the original variance in all original variables.
    - The **cumulative variance**, which represents the sum of individual variance and all variances of the preceding principal components.
    - The **eigenvalue**, which represents the variance explained relative to one of the original variables. For example, an eigenvalue of 1 means the component can explain just as much variance as one of the original variables.
- These statistics are used to reduce the number of principal components used for analysis, thereby reducing the dimensionality of the data set. The goal is to retain components such that the **data may be analyzed in a more efficient manner**, but with a **similar amount of information** to what the initial data could provide.
- The method that will be used to identify which principal components to retain is the **Kaiser Method**, which states that any component with an eigenvalue above 1 will be retained while components with eigenvalues below 1 will be removed from analysis. This method ensures that all remaining principal components provide more information than each of the original variables.
- The **expected outcome** of the **Principal Component Analysis** is to reduce dimensionality in the original data set and create a more efficient grouping of variables that allows for easier computational power in future analysis while maintaining a similar amount of information to the original data.

B2. PCA Assumption

To perform the Principal Component Analysis, the technique requires the **assumption** that the **variables of interest are all continuous**. For this analysis, all variables used are expressed as quantitative continuous variables, which ensures the **assumption has been met**.

This assumption was clarified at the following source:
https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php

C1. Data Set Variables

The variables to be used for this Principal Component Analysis are described as follows:

- **Latitude** (represented as "Lat"), a continuous quantitative variable that represents the latitude coordinate of each customer's location.

- **Longitude** (represented as "Lng"), a continuous quantitative variable that represents the longitude coordinate of each customer's location.
- **Income**, a continuous quantitative variable that represents the reported income of each customer in US dollars.
- **Outage seconds per week** (represented as "Outage_sec_perweek"), a continuous quantitative variable that represents the estimated number of seconds each customer has experienced of system outage per week.
- **Tenure**, a continuous quantitative variable that represents the number of months each customer has stayed with the service.
- **Monthly charge** (represented as "MonthlyCharge"), a continuous quantitative variable that represents the average amount in US dollars charged to each customer's account per month.
- **Bandwidth gigabytes per year** (represented as "Bandwidth_GB_Year"), a continuous quantitative variable that represents the estimated number of gigabytes used in data by each customer per year.

C2. Variable Standardization

To ensure all variables hold similar weight in the Principal Component Analysis method, the **features were scaled** using the R function "scale," which subtracts the mean value of the variable from each observation and divides the result by the standard deviation of the given variable. This ensures that each scaled variable is expressed such that its new mean is 0 and its new standard deviation is 1.

A CSV file containing the prepared data set, which includes only the relevant scaled variables, has been provided with the file name **"churn_cln_212T2.csv."**

D1. Principal Component Matrix

The following screenshot represents the loading **matrix** for all principal components as determined by the PCA method. This identifies the **contributions of each of the original variables to the principal components**, labeled PC1 as the first component, PC2 as the second, and so on.

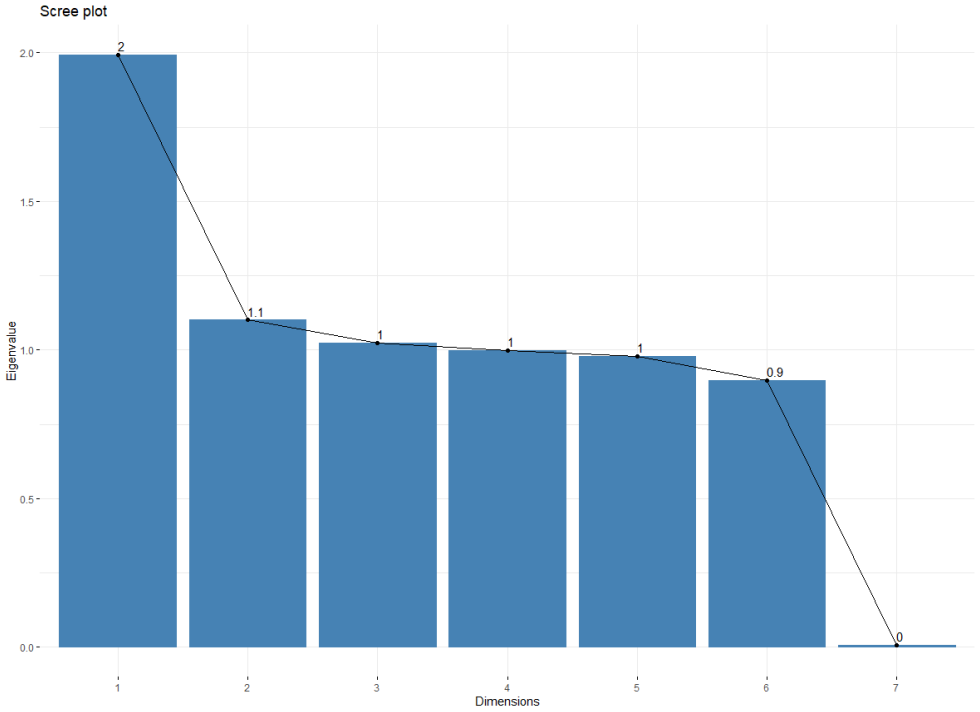
Values closer to 1 or -1 indicate stronger contributions than values close to 0. The sign of the contribution represents the direction. For example, Latitude and Longitude are both significant contributors to PC2, despite one being positive and the other negative.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Lat	-0.023929234	0.69947341	0.122082676	0.0246408043	-0.034349302	0.702475667	-1.028052e-03
Lng	0.007947700	-0.70647613	-0.003628747	0.0720628902	0.030740784	0.703332433	-7.552571e-04
Income	0.003750581	0.07202935	-0.330376255	0.9046198335	0.257293535	-0.033326517	1.254276e-03
Outage_sec_perweek	0.005784197	-0.02642439	0.687545305	0.0457085153	0.721765294	-0.059290290	-2.510506e-05
Tenure	0.705611996	0.02018042	-0.038143908	-0.0287092151	0.034436435	0.014294581	7.057160e-01
MonthlyCharge	0.040770002	-0.07139106	0.633843334	0.4158713883	-0.639925897	-0.083492313	4.537201e-02
Bandwidth_GB_Year	0.706941455	0.01542770	0.001818249	-0.0004778984	-0.006245335	0.007080253	-7.070383e-01

D2. Retained Principal Components

The **Kaiser Rule** was used to identify the principal component to be retained in this analysis. This rule states that components with eigenvalues above 1 will be retained, and components with eigenvalues below 1 will be removed from analysis. This achieves dimensionality reduction.

The following **scree plot** visualizes the eigenvalues for each number of dimensions within the principal components:



The labeled values have been rounded, so even though it appears that the first five dimensions all have eigenvalues of at least 1, a closer look at more precise eigenvalues is necessary. The following screenshot summarizes **important statistics** for each dimension of the PCA:

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.993780405	28.48257721	28.48258
Dim.2	1.101168955	15.73098507	44.21356
Dim.3	1.024381087	14.63401554	58.84758
Dim.4	0.997979593	14.25685132	73.10443
Dim.5	0.978373910	13.97677015	87.08120
Dim.6	0.897860751	12.82658216	99.90778
Dim.7	0.006455298	0.09221854	100.00000

Though dimensions 4 and 5 appeared on the scree plot to be at or above 1, their eigenvalues are slightly below 1; therefore, only **principal components PC1, PC2, and PC3 will be retained** in the analysis, per the Kaiser Rule.

D3. Individual Variance

The **variance explained by each** of the retained principal components as identified in section D2 is summarized as follows:

Principal Component	Variance
PC1	28.48%
PC2	15.73%
PC3	14.63%

D4. Total Variance

The **total variance captured** by all of the retained principal components as identified in section D2 is approximately **58.85%.**

D5. Analysis Results

The results of the Principal Component Analysis are as follows:

- **Seven quantitative continuous variables were selected** from the provided churn data set: Latitude, Longitude, Income, Outage seconds per week, Tenure, Monthly charge, and Bandwidth gigabytes per year.
- A Principal Component Analysis was conducted on the selected variables. **Seven new components were produced** using combinations of the original variables.
- The **Kaiser Rule** was used to identify the principal components to be retained: **PC1, PC2, and PC3**.
- **Principal Component 1** was primarily contributed to by **Tenure** and **Bandwidth gigabytes per year**, as can be seen in the PCA loading matrix. The **eigenvalue** of this component was calculated to be approximately **1.99**, which indicates that this variable explains almost as much variance in the original data as two of the original variables. The **percentage of variance** explained by PC1 was calculated to be approximately **28.48%.**
- **Principal Component 2** was primarily contributed to by **Latitude** and **Longitude**, as can be seen in the PCA loading matrix. The **eigenvalue** of this component was calculated to be approximately **1.10**, which indicates that this variable explains approximately 10% more variance than one of the original variables. The **percentage of variance** explained by PC2 was calculated to be approximately **15.73%.**
- **Principal Component 3** was primarily contributed to by **Outage seconds per week and Monthly charge**, with a notable but smaller contribution from **Income**, as can be seen in the PCA loading matrix. The **eigenvalue** of this component was calculated to be approximately **1.02**, which indicates that this variable explains approximately 2% more variance than one of the original variables. The **percentage of variance** explained by PC2 was calculated to be approximately **14.63%.**
- The **total variance** explained by the retained principal components was calculated to be approximately **58.85%.** Although this is adequate when considering the number of

variables was reduced from 7 to 3, this **may be an inadequate amount of information** depending on the type of research being conducted.

- As described in section D2, the **next two principal components** (PC4 and PC5) were very close to the boundary set by the Kaiser Rule. If these two components were retained as well, the explained variance would be increased to **approximately 87.08%**, which may be enough extra information to outweigh the addition of two more components to the final analysis. However, the Kaiser Rule was chosen as the method used to determine component retention, so these two components were not retained.
- **Overall, the Principal Component Analysis succeeded in reducing dimensionality and creating a more efficient set of variables to use for customer analysis, though a notable amount of information was lost in the process**. For future usage of the relevant data, it may be advisable to revise the selection method of the principal components.

E. Third-Party Code References

WGU Courseware was used as a resource to learn the methods, concepts, and functions used to create the codes in this project, including DataCamp course tracks (datacamp.com) and Dr. Kesselly Kamara's D212 Panopto videos. There are no codes that have been taken directly from any other resources.

F. Content References

The assumption of the inclusion of only continuous variables was clarified at the following source: https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php