# D209 Data Mining 2

Professor Keiona Middleton

Mackenzie Simon

# Part 1 Research Question:

A1: As an analyst, our goal when looking at the Telecommunications Churn Data set is to figure out how to predict customer churn. We will figure this out by focusing on the Principal Component Analysis. Through using the PCA we can reduce the dimensionality of our data sets by transforming a large set of variables into a smaller set. We can then use our new reduced model from our PCA to predict for customer churn.

A2: The goal of our analysis is to use the Principal Component Analysis to create a new model for predicting customer churn. Using PCA, we can then attempt to cut down the number of variables for predicting churn.

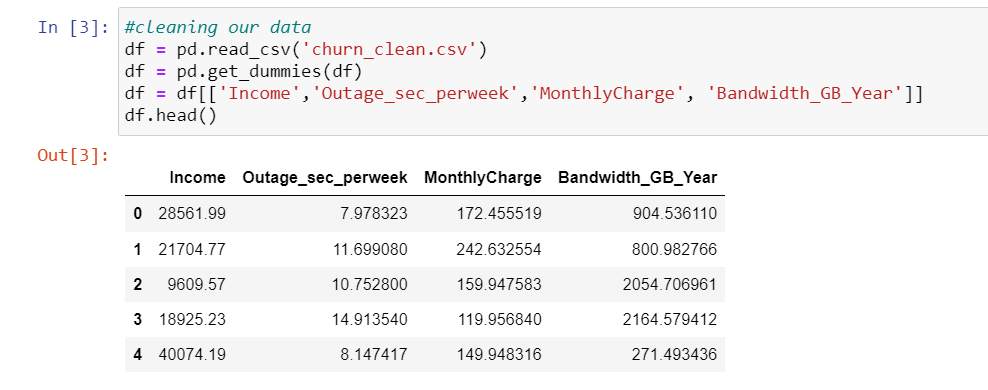
# Part 2 Method Justification

B1: The principal component analysis is a dimensionality-reduction method and allows us to cut down our number of variables in our dataset. Principal components are new variables that are constructed as linear combinations of the initial variables. PCA captures the variance of each component using eigenvectors and eigenvalues. Alto (2019) describes how these combinations are done in such a way that these new variables are uncorrelated and most of the information is stored within the initial first components. For our data set, we expect our first PCA to account for most of the variance, and the following components to account for less variance.

B2: One assumption for using PCA is there needs to be a linear relationship between all variables. Another assumption for PCA is that we are using multiple continuous variables. Jaadi (2021) explains the Principal Component Analysis as constructing linear combinations in order to put maximum possible information in the first component.

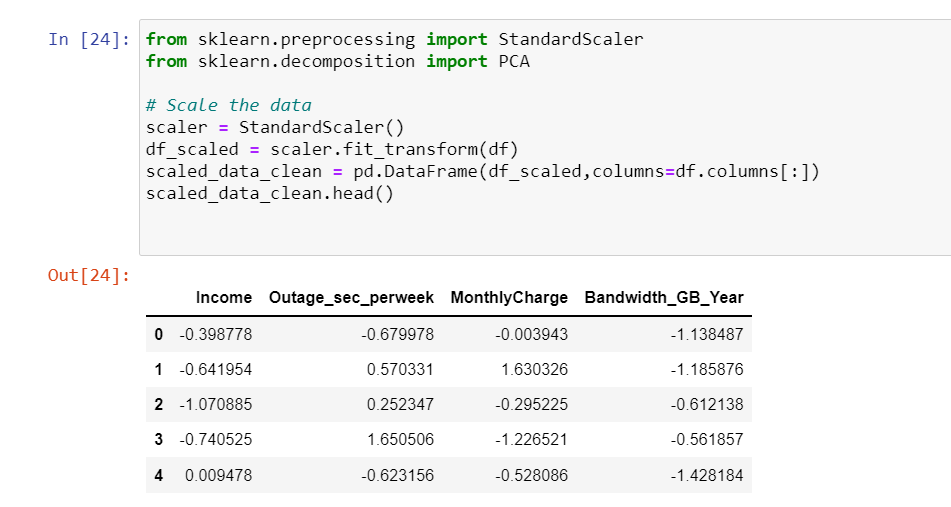
# Part 3 Data Preparation:

C1: Our variables for our PCA model are listed below. We will use the following continuous variables: income, outage seconds per week, monthly Charge, and bandwidth of gigabyte per year.



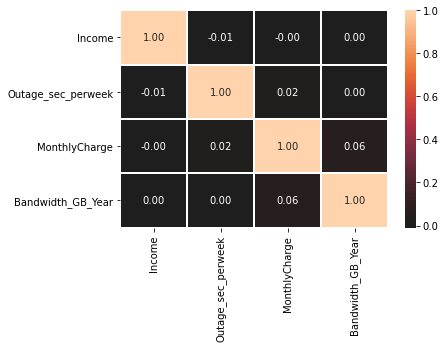
C2:

One important pre-classification step for our Principal Component Analysis is we must scale our data. Data scaling can be achieved by normalizing or standardizing variables.

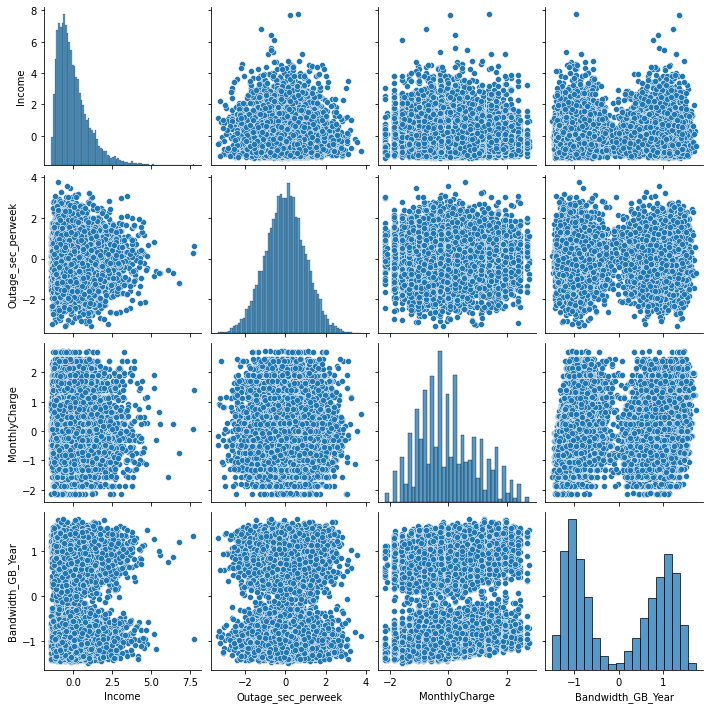


# Part 4 Analysis:

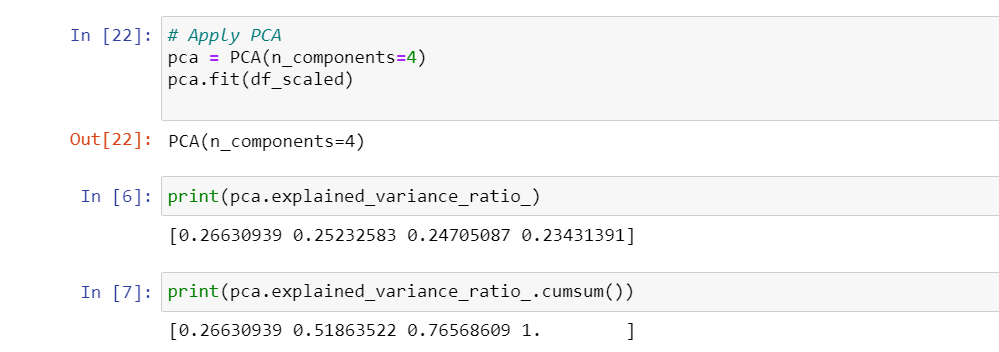
D1-5: We will start by analyzing our data with a heatmap for correlation between initial variables. Our data standardization was a success because we don’t see any one for one correlation.



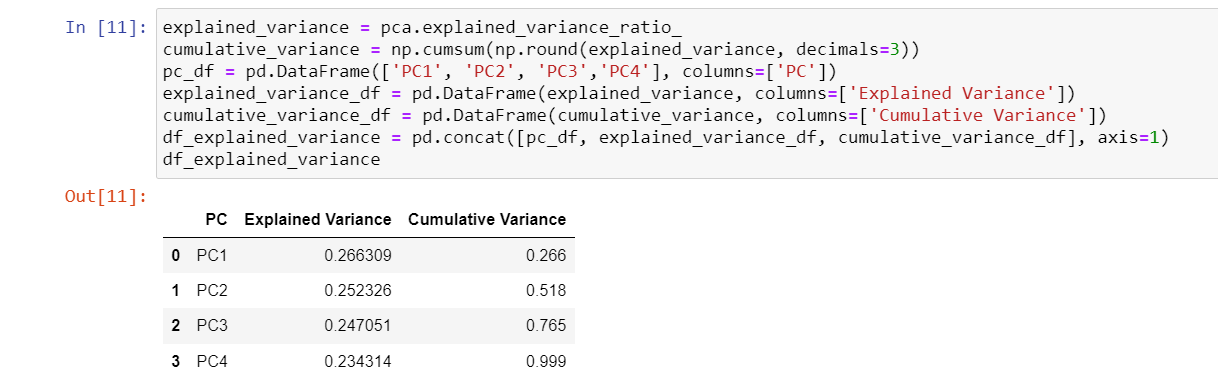
Next, I used seaborn pair plot feature to create a scatter plot matrix between all the initial variables. Like the heatmap, we don’t see any one for one correlation.

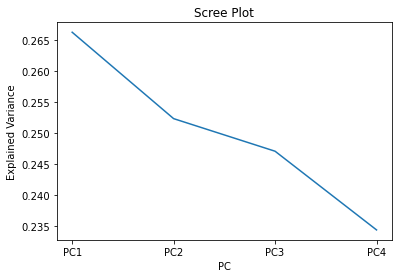


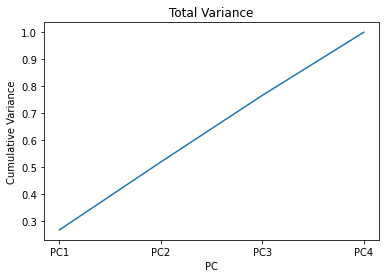
For our PCA I decided to use four components because our initial model was small consisting of only four variables.

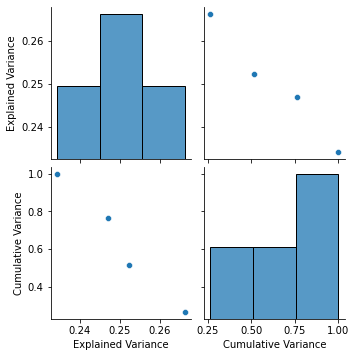


Through our PCA we see there isn’t a huge drop off in explained variance ration for each principal component. Each component contributes roughly the same amount of the cumulative explained variance. For explained variance we get the results PC1 = .266, PC2 =.252, PC3=.247, and PC4 =.234. We see that our first component is the largest component of explained variance .266, which is our expected outcome. For cumulative variance sum, we get the outcome PC1 = .266, PC2=.518, PC3=.765, and PC4 = 1. I then created a PCA matrix table to visualize our explained and cumulative variance. Based on our results, I would choose to keep all the components. There isn’t a huge bend in the knee for our scree plot because each component explains around the same proportion variance. I would choose to keep all components.









From our results, our components all contribute a similar amount of explained variance. There isn’t a huge dip in the elbow. We see our expected PCA results with the first principal component being the largest. The following principal components then sequentially reduce in size. From this result, I would choose to keep all the components for our PCA model. Running the model with a reduced number of components of n=2, we get similar results. I would suggest adding additional variables to our initial model to bring down the variance for each PC. If forced to cut down on PCA components, I would suggest using the first two components.

# References

Alto, V. (2019, July 13). *PCA: Eigenvectors and Eigenvalues*. Medium. https://towardsdatascience.com/pca-eigenvectors-and-eigenvalues-1f968bc6777a.

Jaadi, Z. (2021). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

**Resources for Python Libraries:**

https://matplotlib.org/

https://numpy.org/

<https://pandas.pydata.org/>

https://scikit-learn.org/stable/

https://seaborn.pydata.org/