Machine Learning

Task 2

Nathan Hefner

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

1. Propose one question relevant to a real-world organizational situation

One question relevant that we will be solving with k-means is what is the largest factor and contributor to causing customers to churn, and having the largest impact on customer retention.

2. Define one goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

One goal of this analysis will be to find the largest contributor of churned customers and to find the largest impact of keeping customer retention. With both of the observations, we can then adjust the business model to reduce the impact it has that causes customers to churn and boost the impact of keeping customer retention.

C. Explain the reasons for your chosen clustering technique from part B1 by doing the following:

1. Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes.

K-means clustering is a machine learning algorithm that partitions the data into distinct, non-overlapping clusters based on feature similarity. After preprocessing the data the k-means algorithm initializes by randomly selecting k points as the initial centroids. It will then assign each customer to the nearest centroid using Euclidian distance. After it will recalculate the centroids as the mean of all points in each cluster. Following that it will repeat all steps until cluster assignments no longer change.

The expected outcome will be; clustering labels, each customer gets a cluster label, cluster centroids, each cluster has a centroid showing the average customer in that group, and group characteristics, every cluster can be analyzed to identify common behavioral patterns and distinct churn profiles.

2. Summarize one assumption of the clustering technique.

An important assumption of k-means is that the clusters are spherical and equally sized in feature space. This means that data points in each cluster are grouped around a central point, the centroid, all clusters have similar variance, and the distance between points accurately reflects their similarity.

Without these k-means may misclassify points or produce inaccurate groupings.

3. List the packages or libraries you have chosen for Python

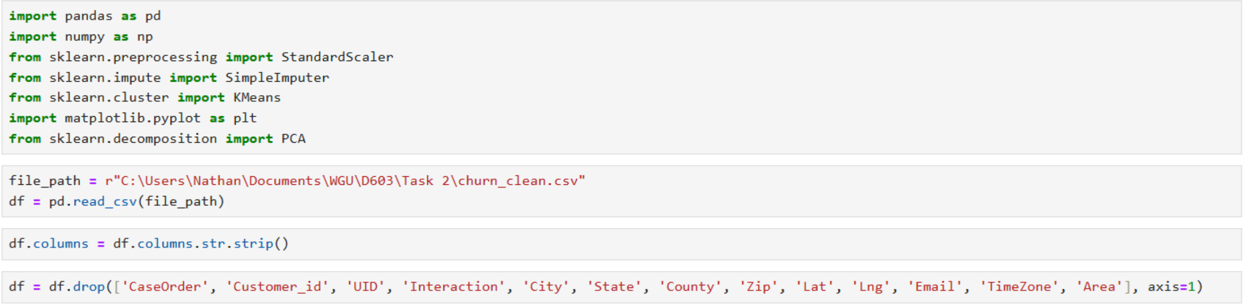
Here are all the packages and libraries we will need to complete this task:

1. Pandas - Loads the dataset, allows Python to read the dataset, and helps clean the dataset if necessary.
2. NumPy - Used for efficient mathematical computations
3. Scikit-learn - Used for the K-Means algorithm and preparing for machine learning models by imputing, scaling, and clustering the data
4. MatPlotLib & Seaborn - Visualizes plots or graphs created from a dataset

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

1. Describe one data preprocessing goal relevant to the clustering technique from part B1.

The first thing for preprocessing the data we need to do is ensure there are no empty or null cells, make sure there are no leading or trailing spaces, and drop all columns that are not relevant to us. The code is as seen below.



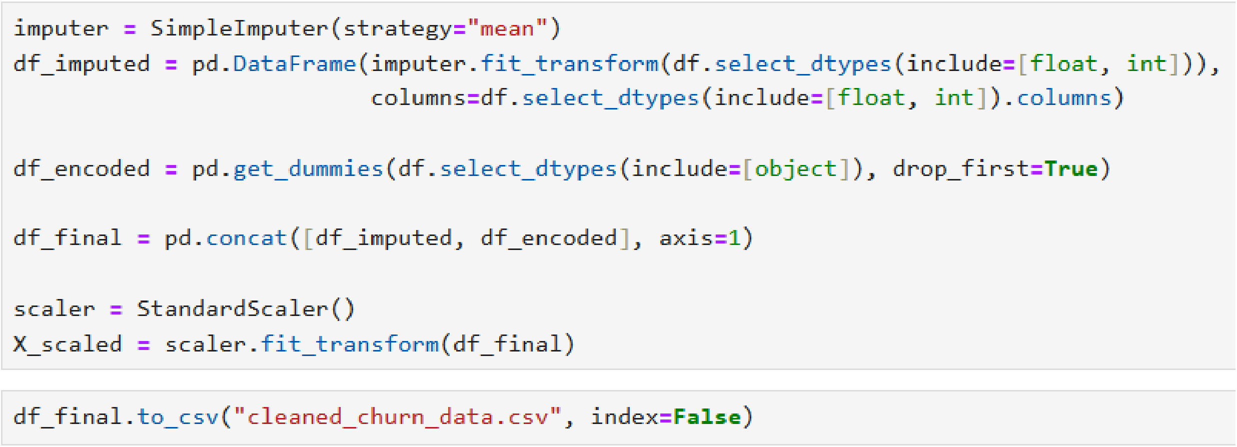
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
```

```
file_path = r"C:\Users\Nathan\Documents\WGU\D603\Task 2\churn_clean.csv"
df = pd.read_csv(file_path)
```

```
df.columns = df.columns.str.strip()
```

```
df = df.drop(['CaseOrder', 'Customer_id', 'UID', 'Interaction', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng', 'Email', 'TimeZone', 'Area'], axis=1)
```

After this, we will now impute any missing data cells, encode the categorical data, because k-means clustering can only handle numeric data, and finally scale and fit the data to help shape and ensure all features contribute equally.



```
imputer = SimpleImputer(strategy="mean")
df_imputed = pd.DataFrame(imputer.fit_transform(df.select_dtypes(include=[float, int])),
                          columns=df.select_dtypes(include=[float, int]).columns)

df_encoded = pd.get_dummies(df.select_dtypes(include=[object]), drop_first=True)

df_final = pd.concat([df_imputed, df_encoded], axis=1)

scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_final)
```

```
df_final.to_csv("cleaned_churn_data.csv", index=False)
```

2. Identify the initial dataset variables that you will use to perform the analysis for the classification question from part B1, and classify each variable as continuous or categorical.

After dropping the irrelevant data columns I separated the remaining ones into continuous or categorical below.

Categorical:

- Job
- Marital
- Gender
- Churn
- Techie
- Contract
- Port_modem
- Tablet
- InternetService
- Phone
- Multiple
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- PaperlessBilling
- PaymentMethod

Continuous:

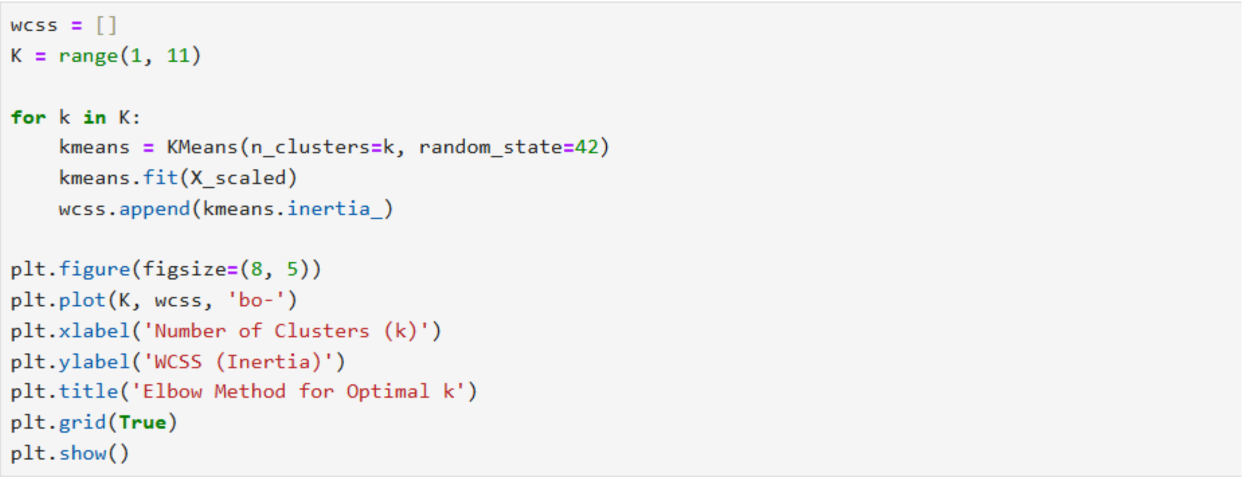
- Population
- Children
- Age
- Income
- Outage_sec_perweek
- Contacts
- Yearly_equip_failure
- Tenure
- MonthlyCharge
- Bandwidth_GB_Year
- Item1
- Item2
- Item3
- Item4

- Item5
- Item6
- Item7
- Item8

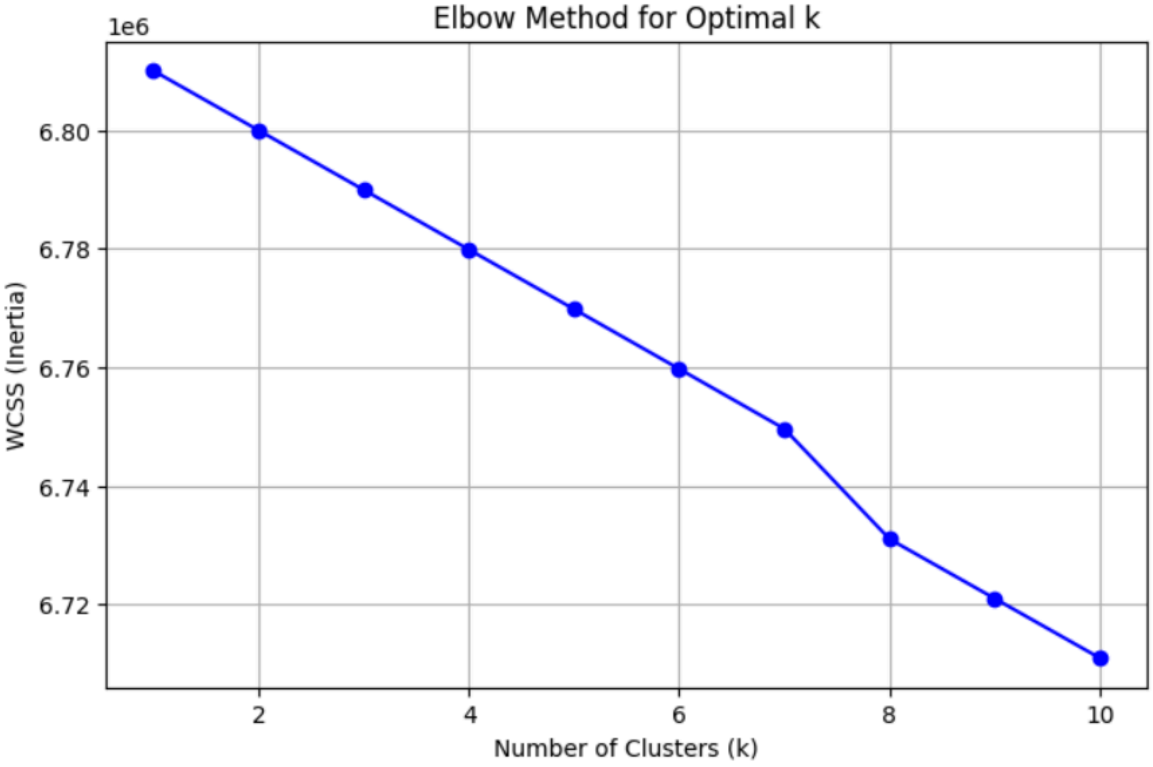
E. Perform the data analysis and report on the results by doing the following:

1. Determine the optimal number of clusters in the dataset, and describe the method used to determine this number.

The method I chose was the elbow method, what it does is simply plot the WCSS (Within-Cluster Sum of Squares) versus the number of clusters (k). We then look for the “elbow” the point where adding more clusters yields diminishing returns. Our graph ended up being mostly linear, resulting in no clear bend in the graph, which is common for complex or overlapping data, but k = 8 appears to be our best candidate because of the clear change of trajectory in the graph.



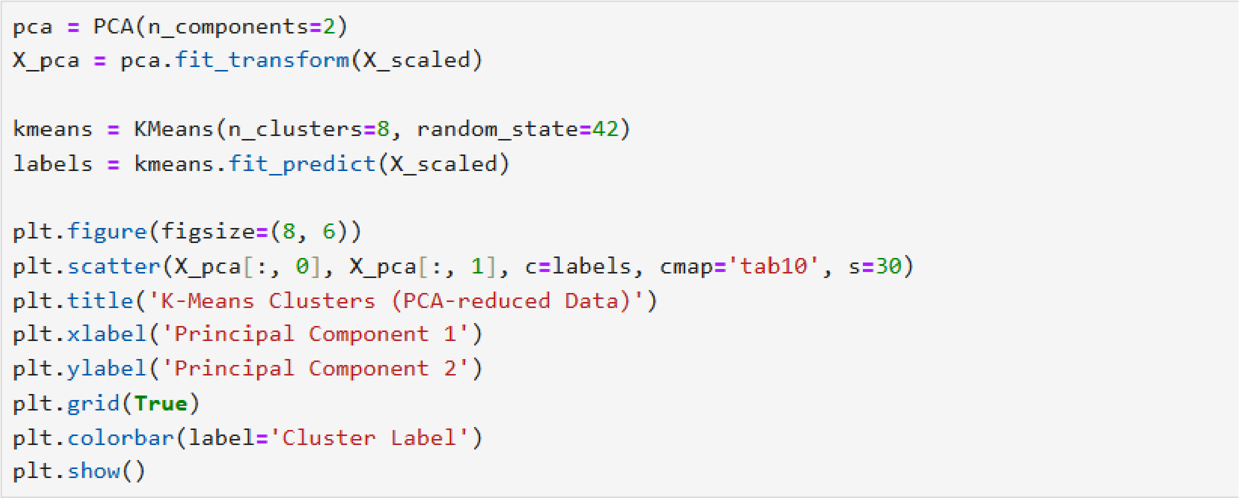
```
wcss = []
K = range(1, 11)

for k in K:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(8, 5))
plt.plot(K, wcss, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('WCSS (Inertia)')
plt.title('Elbow Method for Optimal k')
plt.grid(True)
plt.show()
```



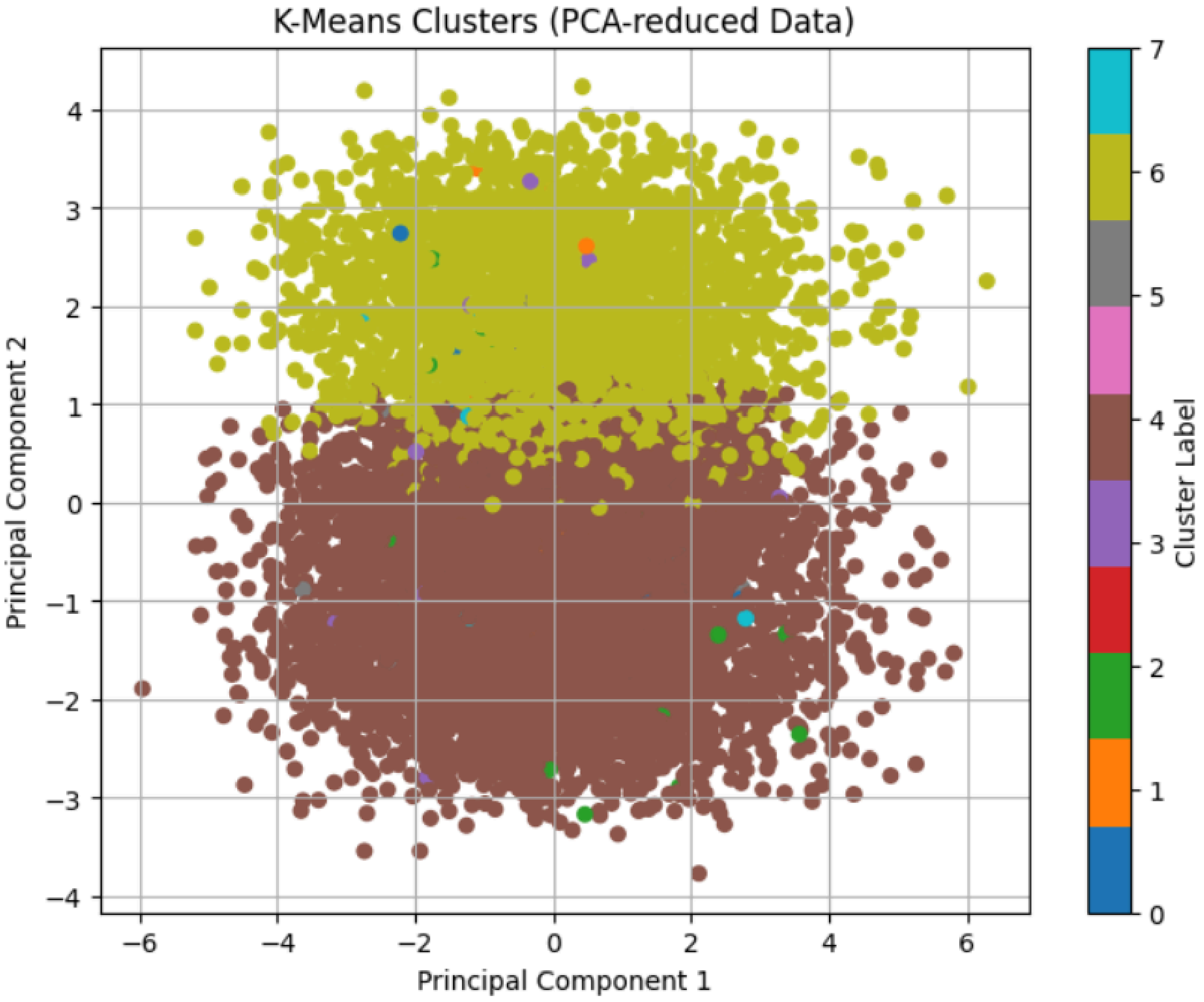
F. Summarize your data analysis by doing the following:

1. Visualize the clusters and explain the quality of the clusters created.

To visualize the clusters we use Principal Component Analysis to reduce the features from 3d to a 2d plot. Below is the code and the graph, using 8 as the number of clusters.



```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

kmeans = KMeans(n_clusters=8, random_state=42)
labels = kmeans.fit_predict(X_scaled)

plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=labels, cmap='tab10', s=30)
plt.title('K-Means Clusters (PCA-reduced Data)')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.grid(True)
plt.colorbar(label='Cluster Label')
plt.show()
```



As we can see, there are two main clusters, labels 4 and 6. This suggests there are two main factors contributing to why customers churn. As for the other smaller clusters they appear to be minimal and too scattered and lack a strong presence on the graph, suggesting that these clusters are simply outliers. There is also a clear split horizontally at about 0.5; however, there is some overlap, indicating that some features are too similar and k-means can't separate them well. K-means also assumes clusters are roughly spherical and evenly sized, if not k-means will struggle to properly visualize.

2. **Discuss the results and implications of your clustering analysis.**

The results show an overwhelming amount of presence for two features specifically causing customers to churn. This implies that these two features are the biggest causes for customers to churn; therefore, once we discover why these features cause customers to churn, we can then reduce this by targeting the specific problem.

3. **Discuss one limitation of your data analysis.**

A limitation of k-means is the clustering is not labeled so there is no output to show the correct groups. We find the patterns without knowing how accurate they are to the real-world scenarios, thus may lead to misguided business decisions.

4. **Recommend a course of action for the real-world organizational situation**

My recommended course of action would be to identify these two features and then see if they are reasonable values to cause customers to churn. Once we can do that we would then try to reduce this amount by implementing new changes or policies into our business.

Sources:
No Sources were used except for WGU Material