DATA MINING II - D212
Task 1
Eric D. Otten
Student ID: 011399183

A1. Describe the purpose of your data mining report by doing the following: Propose one question relevant to a real-world organizational situation that you will answer using one of the following clustering techniques: k-means (using only continuous variables), hierarchical

My research question for D212 is “How can we segment our telecommunications customers based on their demographic characteristics, service usage, and account information for market segmentation purposes”

This question is relevant to a real world organization for market segmentation purposes. By segmenting customers by similar features, an organization can tailor its marketing strategy to suit each demographic separately, creating creatives and marketing messages that resonate with each group (cluster) and maximize the organization’s return on ad spend.

For my analysis I will be using the k-means clustering technique. My rationale for selecting this clustering technique is described in Part B1.

A2. 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.

The goal of this data analysis is to segment customers into distinct groups based on continuous variables related to their demographics, service usage, and account information. These distinct groups will provide insights to decision makers for market segmentation purposes.

B1. Explain how the clustering technique you chose analyzes the selected data set. Include expected outcomes.

K-means clustering will be used to partition the customers into K distinct clusters based on their continuous features. This technique minimizes the within-cluster variance, creating clusters where customers within each cluster have similar characteristics.

K-means clustering is designed to partition a data-set into K distinct clusters based on continuous features, aiming to minimize within-cluster variance. The process begins with the **initialization** step, where the number of clusters K is chosen, and K centroids are randomly initialized in the data space. These centroids represent the initial cluster centers.

In the **assignment** step, each data point in the dataset is assigned to the nearest centroid based on a distance metric, typically the Euclidean distance, effectively forming K clusters.

Following this, the **update** step involves recalculating the centroids by computing the mean of all data points assigned to each cluster. These updated centroids serve as the new cluster centers. The algorithm then iterates between the assignment and update steps; data points are reassigned to the nearest centroid, and centroids are recalculated based on the new cluster compositions.

This process continues until convergence is reached, which occurs when the centroids no longer change significantly between interactions. This can be identified using the elbow method.

The result is K clusters where customers within each cluster exhibit similar characteristics, ensuring that variance within each cluster is minimized. This can then be used for market segmentation purposes.

B2. Summarize one assumption of the clustering technique.

K-means assumes that the data is continuous and that the number of clusters (K) is predefined. It also assumes that the clusters are spherical and equally sized, which may not always reflect real-world data distributions.

B3. List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

I chose to use Python for my analysis due to its flexibility and extensive libraries, of which I will be using:

- **Pandas**: For data manipulation and cleaning.
- **NumPy**: For numerical operations.
- **Scikit-learn**: For implementing the K-means clustering algorithm and other preprocessing steps.
- **Matplotlib/Seaborn**: For data visualization and plotting the results of clustering.

C1. Describe one data preprocessing goal relevant to the clustering technique from A1.

The goal is to clean and normalize the data to ensure all continuous variables contribute equally to the distance calculations used in K-means clustering. Additionally, categorical columns will need to be encoded using one-hot encoding.

C2. Identify the initial data set variables you will use to perform the analysis for the clustering question from A1, and label each as continuous or categorical.

I chose the following variables for my initial data set:

- Age (continuous)
- Income (continuous)
- Outage_sec_perweek (continuous)
- MonthlyCharge (continuous)

C3. Explain each of the steps used to prepare the data for the analysis. Identify the code segment for each step.

The dataset has already been mostly cleaned. However, in order to apply k-means clustering to categorical variables, one hot encoding must be used. The code below selects the columns listed in Part C2 and applies one hot encoding to categorical columns. Additionally, I used StandardScaler to normalize the continuous variables.

Code
import pandas as pd from sklearn.preprocessing import StandardScaler from sklearn.cluster import KMeans import matplotlib.pyplot as plt from sklearn.metrics import silhouette_score

```
import numpy as np

# Load the data
df = pd.read_csv('churn_clean.csv')

# Select the relevant columns
selected_columns = ['Age', 'Income', 'Outage_sec_perweek', 'MonthlyCharge']
df = df[selected_columns]

# Normalize the continuous variables
scaler = StandardScaler()
df[selected_columns] = scaler.fit_transform(df[selected_columns])

# Display the first few rows of the preprocessed data
print(df.head())
```

Output

	Age	Income	Outage_sec_perweek	MonthlyCharge
0	0.720925	-0.398778	-0.679978	-0.003943
1	-1.259957	-0.641954	0.570331	1.630326
2	-0.148730	-1.070885	0.252347	-0.295225
3	-0.245359	-0.740525	1.650506	-1.226521
4	1.445638	0.009478	-0.623156	-0.528086

C4. Provide a copy of the cleaned data set.

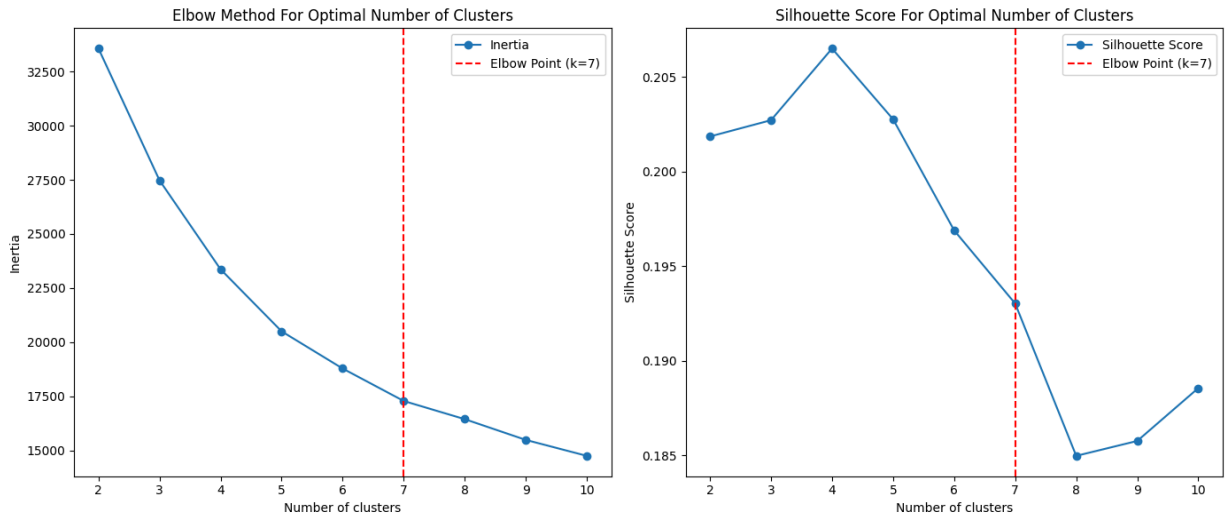
See data_cleaned.csv

D1. Determine the optimal number of clusters in the data set, and describe the method used to determine this number.

After the data has been prepared, I used python to determine the optimal number of clusters in the data set by doing the following:

- Calculate the inertia (within-cluster sum of squares) for cluster numbers between 2 and 10. Inertia is the measure of how tightly the data points in a cluster are grouped around the centroid of the cluster. It is calculated as the sum of squared distances between each data point and the centroid.
- Calculate the second derivative (y") of the inertia values and find the elbow point by selecting the inertia value with the minimum second derivative. This identifies the number of clusters after which, additional clusters lead to diminishing returns of inertia.

The results of my analysis have been plotted in figures below, where Figure 1 on the left plots Inertia as a function of number of clusters, and Figure 2 on the right plots the silhouette score as a function of the number of clusters.



The elbow curve at k=7 indicates that 7 is the optimal number of clusters.

D2. Provide the code used to perform the clustering analysis technique.

Code

```
# Calculate inertia and silhouette scores for a range of cluster numbers
inertia = []
silhouette_scores = []
K_range = range(2, 11)

for k in K_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(df[selected_columns])
    inertia.append(kmeans.inertia_)
    score = silhouette_score(df[selected_columns], kmeans.labels_)
    silhouette_scores.append(score)

# Calculate the second derivative of the inertia values to find the elbow point
diffs = np.diff(inertia)
second_diffs = np.diff(diffs)
elbow_point = np.argmin(second_diffs) + 2  # Adding 2 to account for the double difference

# Plot the Elbow Curve
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
plt.plot(K_range, inertia, marker='o', label='Inertia')
plt.axvline(x=elbow_point, linestyle='--', color='r', label=f'Elbow Point (k={elbow_point})')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method For Optimal Number of Clusters')
plt.legend()

# Plot the Silhouette Scores
plt.subplot(1, 2, 2)
```

```
plt.plot(K_range, silhouette_scores, marker='o', label='Silhouette Score')
plt.axvline(x=elbow_point, linestyle='--', color='r', label=f'Elbow Point (k={elbow_point})')
plt.xlabel('Number of clusters')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score For Optimal Number of Clusters')
plt.legend()

plt.tight_layout()
plt.show()

df.to_csv("data_cleaned.csv")
```

See main.py for the full executable.

E1. Explain the quality of the clusters created.

The quality of the clusters will be evaluated using the Silhouette Score. Figure 2 shows a Silhouette Score of roughly 0.225. A score around 0.225 is relatively low, indicating that the clusters are not very well defined. There is a significant overlap between the clusters, meaning that the data points are not clearly separated into distinct groups.

E2. Discuss the results and implications of your clustering analysis.

After performing the clustering analysis using K-means, we identified the optimal number of clusters based on the Elbow Method and Silhouette Score to be 5. This indicates that customer demographics and behavior can be categorized into 5 major groups, which simplifies decision making.

An organization can, rather than treating each customer individually, treat several groups who share similar demographic and behavioral characteristics as a whole. This is further described in Part E2.

E3. Discuss one limitation of your data analysis.

One significant limitation of the data analysis is the use of a limited number of variables. This might lead to oversimplified clustering, which may not capture the full complexity of customer behaviors and preferences.

E4. Recommend a course of action for the real-world organizational situation from A1 based on the results and implications discussed in E2.

As an organization with limited resources, customer segmentation via clustering techniques is useful in order to understand the behavior and demographics of the customer base as a whole.

In this example, a telecommunications company has 10,000 customers. It would be highly inefficient for an organization to create a detailed customer retention plan for each individual customer. Therefore, clustering customers into an optimal number saves the organization time and energy.

The result of my analysis indicates that the optimal number of clusters (ie. customer segments) is 5. Therefore, an organization can group their customer base using the variables described in Part C2 into 5 groups and create a customer retention plan for each group.

G. Web Sources

StandardScaler docs –
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

silhouette_score docs –
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

KMeans docs – https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

H. Works Consulted
None