**Part I: Research Question**

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

1. Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following clustering techniques:

What are the main characteristics that govern what groups our customers fit into?

• ***k*-means, using only continuous variables**

• hierarchical

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.

The goal of the data analysis is to group customers with similar characteristics together into distinct groups.

**Part II: Technique Justification**

B. Explain the reasons for your chosen clustering technique from part A1 by doing the following:

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

K-means clustering partitions the dataset into k groups by minimizing the variance within each cluster. The iterative nature of clustering techniques like K-Means involves repeatedly assigning data points to clusters and recalculating cluster centroids until the assignments stabilize or a predefined stopping criterion is met.

**Expected Outcome**

What I expect will happen is that the customers will, at minimum, break off into two large groups. Since I am using Tenure as a variable it will likely break off into one long tenured group and a shorter tenured group.

2. Summarize **one** assumption of the clustering technique.

K-means assumes that clusters are spherical (or isotropic) and evenly distributed

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

* **Scikit-learn:** k-means and utilities to help determine optimal numbers, standard scaler is also included in this (it helps reduce bias by adjusting features to a similar range)
* **Numpy:** numerical computations, linear algebra
* **Matpotlib:** cluster visualization
* **Pandas:** data cleaning, manipulation, base library to generally download in python

**Part III: Data Preparation**

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

1. Describe **one** data preprocessing goal relevant to the clustering technique from part A1.

One data preprocessing goal that is relevant is the standardization of the data using the StandardScaler function. It ensures that features are on the same scale, making models more robust and faster to converge.

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

Tenure - continuous

Bandwidth\_GB\_year - continuous

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

* **As far as preparing the data for the analysis. What I did was select the columns that I needed from the data and put them into their own dataframe. The code for this was:**

columns = churn\_df[['Tenure', 'Bandwidth\_GB\_Year']]

* **The next thing I did was run the dropna() function on the dataframe I just made to drop any null values that may be in the data. The code for that is:**

columns = columns.dropna()

* **The last data preparation step I did was to standardize my data that so that each value holds equal weight. I did this by using the code here:**

scaler = StandardScaler()

scaled\_columns = scaler.fit\_transform(columns)

4. Provide a copy of the cleaned data set.

\*Attached with submission\*

**Part IV: Analysis**

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

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

The optimal number of clusters in the data set was 2. I determined this using the elbow method. It involves plotting the **Within-Cluster Sum of Squares (WCSS)**—a measure of how close data points are to their cluster centroid—against different values of k.

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

**The code used is below:**

wcss = []

for k in range(1, 11):

kmeans = KMeans(n\_clusters=k, random\_state=42)

kmeans.fit(scaled\_columns)

wcss.append(kmeans.inertia\_)

**Part V: Data Summary and Implications**

E. Summarize your data analysis by doing the following:

1. Explain the quality of the clusters created.

I calculated the Silhouette Score to assess the quality of my clusters. The value that I got was 0.8126288795519078. This value is close to 1 (perfect clustering) meaning that the clusters are of good quality.

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

**Results**

* The clustering analysis shows two groups. The lower, purple group is the short tenure/low GB group and the higher, yellow group is the long tenure/high GB group.
* The clustering also shows that as tenure increases, the GB per year also increases.

**Implications**

* The positive correlation between tenure and GB usage suggests that the service becomes more integral to users over time.
* This also implies that Long-term users might represent a core customer base, and understanding their needs could help sustain and grow your customer pool.

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

K-means works well only if the clusters in the data are roughly circular in shape and have similar densities. If the data contains elongated, irregularly shaped, or overlapping clusters, k-means may fail to correctly identify the cluster boundaries.

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

* For the low tenure/GB group, come up with an onboarding strategy/routine that can be used to encourage the lower tenured people to stay for longer.
* For the long tenure/high GB group, focus on keeping the people who have longer tenures happy, a survey could be sent out asking what things they like and dislike so improvements can be made to keep them and potentially help keep the shorter tenured people as well.

**Part VI: Demonstration**

F. Provide a Panopto video recording that includes the presenter and a vocalized demonstration showing all code used, the code being executed, and the results of all code used in the task.

1. Include the presenter and a vocalized demonstration describing the programs used to complete this task in the Panopto video recording.

*Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.*

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.  
  
To submit your recording, upload it to the Panopto drop box titled “Data Mining II – OFM4” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

G. Record the web sources you used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I. Demonstrate professional communication in the content and presentation of your submission.