Part I: Research Question

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

1. **Propose one question relevant to a real-world organizational situation that you will answer.**
   Using Tenure and MonthlyCharge as indicators, can we predict if a customer will churn?

2. **Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.**
   One goal of the data analysis is to use k-means clustering to determine if customers can be segmented by Tenure and MonthlyCharge to predict Churn. (Western Governor's University, 2021)

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 dataset. Include expected outcomes.**
   K-means clustering segments the dataset based on datapoints' mean distances from the centroids. When new datapoints are added, the distance between their means and other means in the dataset is used to allocate it to a cluster. The expected outcome are 'k' clusters whose points are as similar as possible, and as different from the other clusters as possible. (Dabbura, 2021)

2. **Summarize one assumption of the clustering technique.**
   Because k-means clustering depends on means to determine cluster allocations, the data passed to the algorithm must be numerical.

3. **List the packages or libraries you have chosen for Python or R and justify how *each* item on the list supports the analysis.**
   Pandas: to import the dataset, create a dataframe and create a crosstab
   Numpy: to create and use numpy arrays
   Matplotlib.pyplot: to visualize data
   Sklearn.preprocessing: to scale data with StandardScaler
   Sklearn.cluster: to initiate and use KMeans

Part III: Data Preparation

C. Perform data preparation for the chosen dataset by doing the following

1. **Describe one data preprocessing goal relevant to the clustering technique from part A1.**
   One data preprocessing goal was to ensure that my data was numeric and scaled to use the same units of measure.

2. **Identify the initial dataset variables that you will use to perform the analysis for the clustering question from part A1, and label *each* as continuous or categorical.**
   Churn is categorical, Tenure is continuous and MonthlyCharge is continuous.

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

First, I imported the dataset:

```
# Import dataset
churn_clean=pd.read_csv("C:/Users/hkeim/OneDrive/Documents/School/D212/churn_clean.csv")churn_clean.info()
```

Then I filtered the dataset to only include the dimensions I needed to answer my question from section 1A:

```
# Select dimensions
dimensions = churn_clean.filter(items = ['Churn','Tenure', 'MonthlyCharge'])
```

Next, I converted Churn to category codes, so it was numerical data:

```
# Convert Churn to category codes
dimensions['Churn'] = dimensions['Churn'].astype('category')
dimensions['Churn'] = dimensions['Churn'].cat.codes
```

I then used StandardScaler to scale my data to have the same units of measure:

```
# Import StandardScaler
from sklearn.preprocessing import StandardScaler

# Scale data
scaler = StandardScaler()
scaler.fit(dimensions)
StandardScaler(copy=True, with_mean=True, with_std=True)
dim_scaled=scaler.transform(dimensions)
```

4. **Provide a copy of the cleaned dataset.**

   The cleaned dataset is provided in the Keim D212 Task One Cleaned Data document.

Part IV: Analysis

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

1. **Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.**

   I first created an elbow graph to decide the number of clusters I would use for my k-means algorithm. I decided based on the graph to use four clusters and ran the k-means algorithm. Then I assigned Tenure to xs and MonthlyCharge to ys and created a scatterplot using the cluster labels as the color of the points. I assigned the cluster centers and added them to my scatterplot as well to enhance the visualization. I created a crosstab using the cluster labels and Churn to answer my question from A1.

2. **Provide the code used to perform the clustering analysis technique from part 2.**

   All code, intermediate calculations, and visualizations are included in the Keim D212 Task One Code document.

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

1. **Explain the accuracy of your clustering technique.**
   The elbow graph I created shows that k = four clusters has the lowest sum of squared errors that we can get while still having centroids as far away from each other as possible on the scatterplot.

2. **Discuss the results and implications of your clustering analysis.**
   While the k-means analysis did show labels and centroids for four clusters, two pairs of clusters can be seen very close together on the scatter plot, meaning it is not as successful at customer segmentation as an analysis showing completely separated clusters. However, the crosstab shows that Churn is somewhat defined or predicted by these clusters. As such, the answer to the original question: using Tenure and MonthlyCharge as indicators, can we predict if a customer will churn, is yes, with caution.

3. **Discuss one limitation of your data analysis.**
   One limitation of this data analysis is that the elbow plot does not have an extremely clear point on which to decide the 'k' for the k-means algorithm.

4. **Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.**
   Based on my discussion in part E2, I recommend using the k-means labels to further identify customers who are likely to Churn and targeting them with special deals and offers.

Works Cited

Dabbura, I. (2021, December 20). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Retrieved from towards data science: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Western Governor's University. (2021, December 20). *D212 Data Mining II Data Files and Associated Dictionary Files.* Retrieved from my.wgu.edu: https://access.wgu.edu/ASP3/aap/content/jf8rcds032ldktfces9r.html

No third party code was used.