**Data Mining II Performance Assessment Task #1**

**D212**

**By Josue Gonzalez**

Contents

[Part I: Research Question 3](#_Toc180522048)

[Part II: Technique Justification 3](#_Toc180522049)

[Part III: Data Preparation 5](#_Toc180522050)

[Part V: Data Summary and Implications 9](#_Toc180522051)

[1. Explain the Quality of the Clusters: 9](#_Toc180522052)

[2. Discuss the Results and Implications: 9](#_Toc180522053)

[3. Discuss One Limitation of the Analysis: 10](#_Toc180522054)

[4. Recommend a Course of Action: 10](#_Toc180522055)

# Part I: Research Question

**A. Purpose of the Data Mining Report**

**A1.Proposed Research Question:**

* + *How can customer characteristics be grouped to reveal distinct customer segments where we can base it off their service usage, tenure, and time they spend without service?*

**A2.Goal of the Analysis:**

* + The goal is to identify customer segments based on similarities in service usage patterns, tenure, and satisfaction levels. doing it this way will help the company unearth distinct customer groups, allowing the company to design better marketing strategies that would result in better outcomes.

# Part II: Technique Justification

**B. Chosen Clustering Technique:**

**B1.Technique chosen**

The chosen technique is **K-means clustering**, a method well-suited for identifying distinct customer segments based on continuous numerical data. K-means works by assigning each data point to one of k clusters in a way that minimizes the variance within each cluster, thereby ensuring that customers within the same cluster exhibit similar characteristics. This is achieved by iteratively adjusting cluster centroids to find the most compact groupings.

The continuous variables selected for this analysis—**Tenure**, **MonthlyCharge**, **Bandwidth\_GB\_Year**, and **Outage\_sec\_perweek**—are essential indicators of customer behaviors and experiences, which makes K-means an ideal choice. By clustering customers based on these metrics, we aim to uncover patterns in service usage, spending, and service reliability. This segmentation will allow the company to tailor marketing and retention strategies for different customer groups, leading to better-targeted interventions.

The expected outcome is the creation of clusters that group customers with similar behaviors and service usage patterns. These clusters are anticipated to provide actionable insights, such as identifying high-value customers for upselling opportunities and segments at risk of churn for retention efforts.

**B2. Assumptions:**

K-means clustering assumes that clusters are spherical in nature, meaning that each cluster is rotating around the center similar to a solar system in this this case it reflects the average characteristics of the data points within this study. Not only has Standardization become crucial for clustering techniques like k-means to ensure that each variable contributes equally to the analysis (Jain, 2010).

**B3.Libraries and packages used:**

* + We will use the following packages in Python:
    - **Pandas** for data manipulation and cleaning.
    - NumPy is used for number manipulation and other processing duties.
    - Scikit-learn is employed to implement the k-means clustering algorithm, facilitating model training and evaluation.
    - **Matplotlib/Seaborn** for visualizing the clusters and interpreting the results.
  + Each of these packages supports different steps of the analysis, from data preparation and clustering to visualization.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Part III: Data Preparation

**C. Data Preparation Steps:**

The data preprocessing stage ensures that the selected variables are prepared for effective clustering with the K-means algorithm. Given that K-means clustering is sensitive to the scale of data, we first standardized the continuous variables to have a mean of 0 and a standard deviation of 1. This scaling process is essential to prevent any single variable from disproportionately influencing the formation of clusters due to differences in magnitude.

The selected continuous variables for this analysis include **Tenure**, **MonthlyCharge**, **Bandwidth\_GB\_Year**, and **Outage\_sec\_perweek**. These variables were chosen because they represent critical aspects of customer behavior and service usage. Standardization of these variables ensures that each contributes equally to the clustering process, allowing K-means to group customers based on meaningful patterns across multiple dimensions. We avoided categorical and ordinal variables (such as survey response data) because they are less compatible with K-means clustering, which is optimized for continuous numerical data.

In this step, we also handled any missing values in the selected variables by either removing or imputing them, ensuring a complete dataset for clustering.

1. **Variables:**

This data analysis incorporates both service usage patterns and survey responses to provide a comprehensive view of customer characteristics. By including survey results alongside service metrics, we aim to capture both behavioral and attitudinal factors, enhancing the effectiveness of the clustering model. This approach helps ensure that segmentation reflects not only usage patterns but also customer perceptions of service quality.

The selected variables for this analysis include both the original continuous variables and the survey response variables:

**Outage\_sec\_perweek (continuous):** Indicates the average number of seconds per week that the customer experiences service outages. This variable can help identify customers who might be dissatisfied due to reliability issues.

**Tenure** (continuous): Indicates the length of time the customer has been with the service.

**MonthlyCharge** (continuous): Reflects the monthly charge incurred by the customer.

**Bandwidth\_GB\_Year** (continuous): Measures the customer's yearly bandwidth consumption.

These variables were selected because they provide a balanced view of both customer behavior and service experience. By focusing on these continuous metrics, the analysis is better aligned with K-means clustering requirements, leading to clearer and more actionable customer segments.

This combined set of variables is expected to provide a more robust clustering model that identifies distinct customer segments based on both their service behaviors and perceptions.

A screenshot of a computer

Description automatically generated

1. **Data Preparation Steps**

In our data preparation process, we began by renaming variables **item1** through **item8** to more descriptive names reflecting the survey responses. We then addressed missing values by either removing or imputing them as needed in the selected variables. Next, we applied standard scaling to ensure that each variable had a mean of 0 and a standard deviation of 1, making the data more suitable for clustering. Additionally, to enhance the clustering accuracy, we performed a feature selection step. This involved generating a correlation heatmap and analyzing feature importance, which helped us identify and remove less relevant variables that could clutter the clustering process. As a result, we updated the list of selected variables to include only the most significant ones, leading to a more focused and interpretable model.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph with numbers and lines

Description automatically generated

A screenshot of a computer code

Description automatically generated

We will then apply K-means and Fit the algorithm to the data we standardized to find those distinct custome segments if there are any.

* + - Cleaned code

A screenshot of a computer code

Description automatically generated

**D. Clustering Analysis Steps**

**Determine the Optimal Number of Clusters:**

To determine the optimal number of clusters, we applied the **Elbow Method**, which involves plotting the Within-Cluster Sum of Squares (WCSS) against various cluster counts. The goal is to identify the point were adding more clusters results in minimal reductions in WCSS, indicating a natural 'elbow' in the graph. In this analysis, the third point on the graph represents the most suitable number of clusters, as increasing the clusters beyond this point yields diminishing returns in terms of improved grouping.

A graph with numbers and lines

Description automatically generated

A screenshot of a computer code

Description automatically generated

After determining that three clusters would be optimal using the Elbow Method, the K-means clustering algorithm was applied to the standardized dataset, focusing on four key continuous variables: **Tenure**, **MonthlyCharge**, **Bandwidth\_GB\_Year**, and **Outage\_sec\_perweek**. This model configuration allowed us to capture distinct patterns in customer behavior, with each cluster representing a unique customer segment. The resulting cluster distribution shows that Cluster 0 contains 4,956 customers, Cluster 1 has 3,170 customers, and Cluster 2 includes 1,874 customers. This relatively balanced distribution indicates that the clustering effectively captures different types of customer behaviors and service experiences.

**Perform K-means Clustering:**

A screen shot of a graph

Description automatically generated

To further interpret the clusters, a scatter plot was generated using two of the most influential features, **Tenure** and **Bandwidth\_GB\_Year**. The visualization reveals clear separations between the clusters, with each group occupying a distinct region in the scatter plot. Customers in Cluster 0, for example, tend to have moderate tenure and bandwidth usage, whereas those in Cluster 1 exhibit longer tenure and higher bandwidth consumption. Meanwhile, Cluster 2 represents customers with shorter tenure and lower service usage. This visual distinction aligns with the quantitative results from the silhouette score and provides deeper insights into each segment’s characteristics, supporting targeted strategies based on customer behavior and service needs.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer code

Description automatically generated

To assess the quality of the clusters, we calculated the silhouette score, which measures how well each data point fits within its assigned cluster relative to other clusters. The model achieved a silhouette score of **0.69**, which indicates that the clusters are fairly well-separated and cohesive. This score suggests that customers within the same cluster share similar characteristics, while the distinctions between clusters are well-defined. The silhouette score provides a quantitative validation that K-means has successfully identified meaningful groupings within the data.

D2. Excution of code

A graph with lines and dots

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer code

Description automatically generated

# Part V: Data Summary and Implications

E.

## 1. Explain the Quality of the Clusters:

The K-means clustering analysis resulted in well-defined clusters, supported by a silhouette score of **0.69**, which suggests relatively strong separation and compactness of clusters. The clusters were created based on key continuous variables such as **Tenure**, **MonthlyCharge**, **Bandwidth\_GB\_Year**, and **Outage\_sec\_perweek**, which capture important dimensions of customer behavior and service experience. The scatter plot further confirms the quality of these clusters, with each group appearing more compact and distinct when plotted along the two most influential features, **Tenure** and **Bandwidth\_GB\_Year**. These well-separated clusters provide a solid foundation for interpreting customer segments and designing targeted strategies.

## 2. Discuss the Results and Implications:

The analysis identified three customer segments with distinct characteristics:

Cluster 0: Customers with moderate tenure and bandwidth usage.

Cluster 1: Customers with longer tenure and higher bandwidth usage.

Cluster 2: Customers with shorter tenure and lower bandwidth usage.

Implications for Business Strategy:

These clusters present clear opportunities for targeted marketing and retention strategies. For example, **Cluster 1** consists of customers with high tenure and bandwidth usage, making them ideal candidates for loyalty rewards or upsell promotions. Meanwhile, **Cluster 2** includes customers with shorter tenure and lower service usage, suggesting a higher churn risk. Retention efforts could be prioritized for this group, focusing on addressing service concerns to reduce churn and improve customer satisfaction.

## 3. Discuss One Limitation of the Analysis:

## Despite a strong silhouette score, there remains some overlap between clusters, particularly between **Clusters 0 and 1**. This suggests that additional variables—especially categorical ones, such as **contract type** or **payment method**—might help refine cluster distinctions and reduce overlap. Additionally, K-means clustering assumes spherical clusters, which may not fully capture the natural shape and distribution of customer groups. This limitation may restrict the algorithm's ability to recognize clusters with irregular shapes, which is a known drawback of K-means clustering (Xu & Wunsch, 2005).

## 4. Recommend a Course of Action:

**Refine Marketing Strategies:** Tailor campaigns based on the specific traits of each cluster. For example, focus upsell strategies on Cluster 1, while prioritizing retention for Cluster 2 customers.

**Further Data Exploration:** Consider incorporating additional categorical variables to improve cluster differentiation. Future analyses could also experiment with hybrid clustering techniques that handle both continuous and categorical data to enhance model performance.

**Enhance Retention Efforts:** Develop targeted retention plans for at-risk customers in **Cluster 2**. This might include addressing their specific service concerns, based on insights gathered from customer feedback and survey responses, to improve satisfaction and loyalty.

F. Panopto Recording

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f3f25379-b958-4d8c-a7db-b21b0060e070>

G Code sources

*Silhouette\_score*. scikit. (n.d.-e). https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.silhouette\_score.html

<https://www.geeksforgeeks.org/k-means-clustering-introduction/>

H.References

Jain, A. K. (2010). "Data Clustering: 50 Years Beyond K-Means." *Pattern Recognition Letters*, 31(8), 651-666.

*K means: Unraveling patterns in data with K means clustering*. FasterCapital. (n.d.). https://fastercapital.com/content/K-means--Unraveling-Patterns-in-Data-with-K-means-Clustering.html

*Linkage#*. linkage - SciPy v1.14.1 Manual. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

Xu, R., & Wunsch, D. (2005). "Clustering Algorithms in Biomedical Research: A Review." *IEEE Reviews in Biomedical Engineering*.