Improving Customer Experience in Telecomunication:

Customer Churn Prediction and Market Segmentation

Let’s Get Started!

Team Profile

Dimas Surya Prasetyo
Bachelor of Informatics Engineering

Bagaskara Adhi Pradana
Bachelor of Informatics Engineering

Abigeil Febiola Sabatini
Bachelor of Informatics Engineering

Introduction

Background of the topics and issues raised

Understanding customer behavior is a key challenge for the telecommunications industry, which is characterized by ever-increasing customer needs and innovation. In this case, the main topic discussed is customer behavior analysis. These topics include understanding service usage patterns, components that influence customer satisfaction, and the formation of useful knowledge. To maintain competition and understand customer preferences, rapid changes are needed in the industry. On the other hand, there is a direct link between understanding customer preferences and the ability to provide an exceptional experience. Understanding customer behavior has become critical to maintaining enterprise success and ensuring telecommunications companies remain at the forefront in a dynamic and competitive landscape as the industry becomes increasingly connected.

Data & Tools

Dataset & Tools Used in Research

Data Source :

Data Challenge DSW 2023 - Student & Junior Pro
The dataset shows the usage of telecommunication services in Q3 of a particular year adapted from a Kaggle public dataset with several modifications.

Tools :

- **Google Colaboratory**
  For data processing, analysis, modeling, and visualization.
- **Excel / Spreadsheet**
  To see an overview of the data and prepare the data for further analysis.

Methodology

Method used in Research

Data Cleaning & Pre - processing

This phase involves cleaning, transforming, and organizing data to ensure it is suitable for analysis.

Modelling

In this step, statistical or machine learning models are applied to the prepared data to make predictions or draw conclusions.

Exploratory Data Analysis

EDA focuses on understanding the dataset through visualization and statistical tools to uncover insights and patterns.

Evaluating

The final step assesses the model's performance by using various metrics to determine its accuracy and effectiveness.

About Data

Understanding Data

```
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Customer ID        7043 non-null   int64
 1   Tenure Months      7043 non-null   int64
 2   Location           7043 non-null   object
 3   Device Class       7043 non-null   object
 4   Games Product      7043 non-null   object
 5   Music Product      7043 non-null   object
 6   Education Product  7043 non-null   object
 7   Call Center        7043 non-null   object
 8   Video Product      7043 non-null   object
 9   Use MyApp          7043 non-null   object
 10  Payment Method     7043 non-null   object
 11  Monthly Purchase   7043 non-null   object
 12  Churn Label        7043 non-null   object
 13  Longitude          7043 non-null   object
 14  Latitude           7043 non-null   object
 15  CLTV               7043 non-null   object
dtypes: int64(2), object(14)
```

Kaggle

Data Challenge DSW 2023
- Student & Junior Pro Dataset

7.403
Rows

16
Columns

Customer ID (A unique customer identifier)

Tenure Months (How long the customer has been with the company by the end of the quarter specified above)

Location (Customer's residence - City)

Device Class (Device classification)

Games Product (Whether the customer uses the internet service for games product)

Music Product (Whether the customer uses the internet service for music product)

Education Product (Whether the customer uses the internet service for education product)

Call Center (Whether the customer uses the call center service)

Video Product (Whether the customer uses video product service)

Use MyApp (Whether the customer uses MyApp service)

Payment Method (The method used for paying the bill)

Monthly Purchase (Total customer's monthly spent for all services with the unit of **thousands of IDR**)

Churn Label (Whether the customer left the company in this quarter)

Longitude (Customer's residence - Longitude)

Latitude (Customer's residence - Latitude)

CLTV (Customer Lifetime Value with the unit of **thousands of IDR** - Calculated using company's formulas)

Data Pre - processing

Cleaning and Transforming Data

Missing Value

Statistical Summary

Handling Outlier

Missing Value

Handling the missing value

```
# checking null value
df.isnull().sum()
```

```
Customer ID          0
Tenure Months        0
Location             0
Device Class         0
Games Product        0
Music Product        0
Education Product    0
Call Center          0
Video Product        0
Use MyApp            0
Payment Method       0
Monthly Purchase     0
Churn Label          0
Longitude            0
Latitude             0
CLTV                 0
dtype: int64
```

There's **no missing value**

```
#  read dataset info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Customer ID        7043 non-null   int64
 1   Tenure Months      7043 non-null   int64
 2   Location           7043 non-null   object
 3   Device Class       7043 non-null   object
 4   Games Product      7043 non-null   object
 5   Music Product      7043 non-null   object
 6   Education Product  7043 non-null   object
 7   Call Center        7043 non-null   object
 8   Video Product      7043 non-null   object
 9   Use MyApp          7043 non-null   object
 10  Payment Method     7043 non-null   object
 11  Monthly Purchase   7043 non-null   object
 12  Churn Label        7043 non-null   object
 13  Longitude          7043 non-null   object
 14  Latitude           7043 non-null   object
 15  CLTV               7043 non-null   object
dtypes: int64(2), object(14)
memory usage: 880.5+ KB
```

```
[73] df.duplicated().sum()
```

```
0
```

There's **no duplicated data**

Statistical Summary

Statistical Summary of Numeric Data

```
# statistical summary
df.describe()
```

	Tenure Months	Monthly Purchase	CLTV
count	7043.000000	7043.000000	7043.000000
mean	32.371149	84.190200	5720.384481
std	24.559481	39.117061	1537.974298
min	0.000000	23.725000	2603.900000
25%	9.000000	46.150000	4509.700000
50%	29.000000	91.455000	5885.100000
75%	55.000000	116.805000	6994.650000
max	72.000000	154.375000	8450.000000

Conclusion :

Tenure Months data shows a **wide variation in customer loyalty** or subscription duration, suggesting the need for strategies to retain customers over time.

Monthly Purchase information **highlights varying spending patterns** among customers, with a moderate level of spending diversity.

CLTV data **indicates differences** in customer lifetime value, highlighting the need of focusing marketing efforts on client segments.

Statistical Summary

Statistical Summary of Categorical Data

```
df.describe(include='object')
```

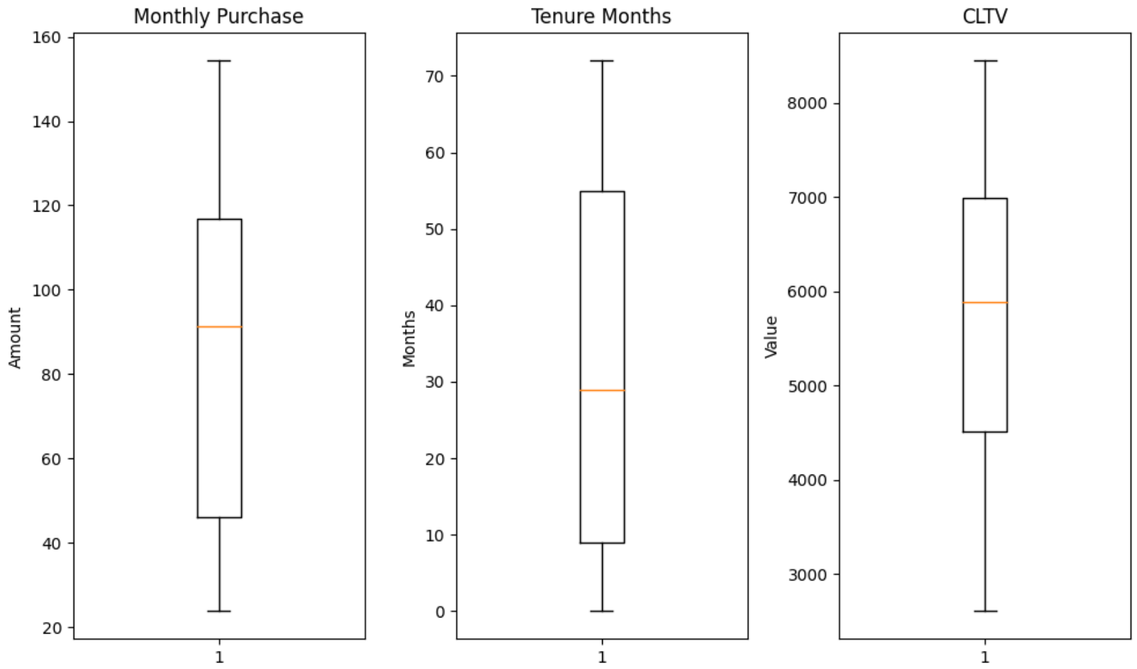
	Customer ID	Location	Device Class	Games Product	Music Product	Education Product	Call Center	Video Product	Use MyApp	Payment Method	Churn Label
count	7043	7043	7043	7043	7043	7043	7043	7043	7043	7043	7043
unique	7043	2	3	3	3	3	2	3	3	4	2
top	0	Jakarta	High End	No	No	No	No	No	No	Pulsa	No
freq	1	5031	3096	3498	3088	3095	4999	2810	2785	2365	5174

Conclusion :

From the top value and their frequency, we gain insight into the **most common** data points.

Outliers

Handling Outlier



Other outliers indicate data diversity. In the numeric data, **there are no outliers**, so further analysis can be carried out.

Exploratory Data Analysis

Data Distribution

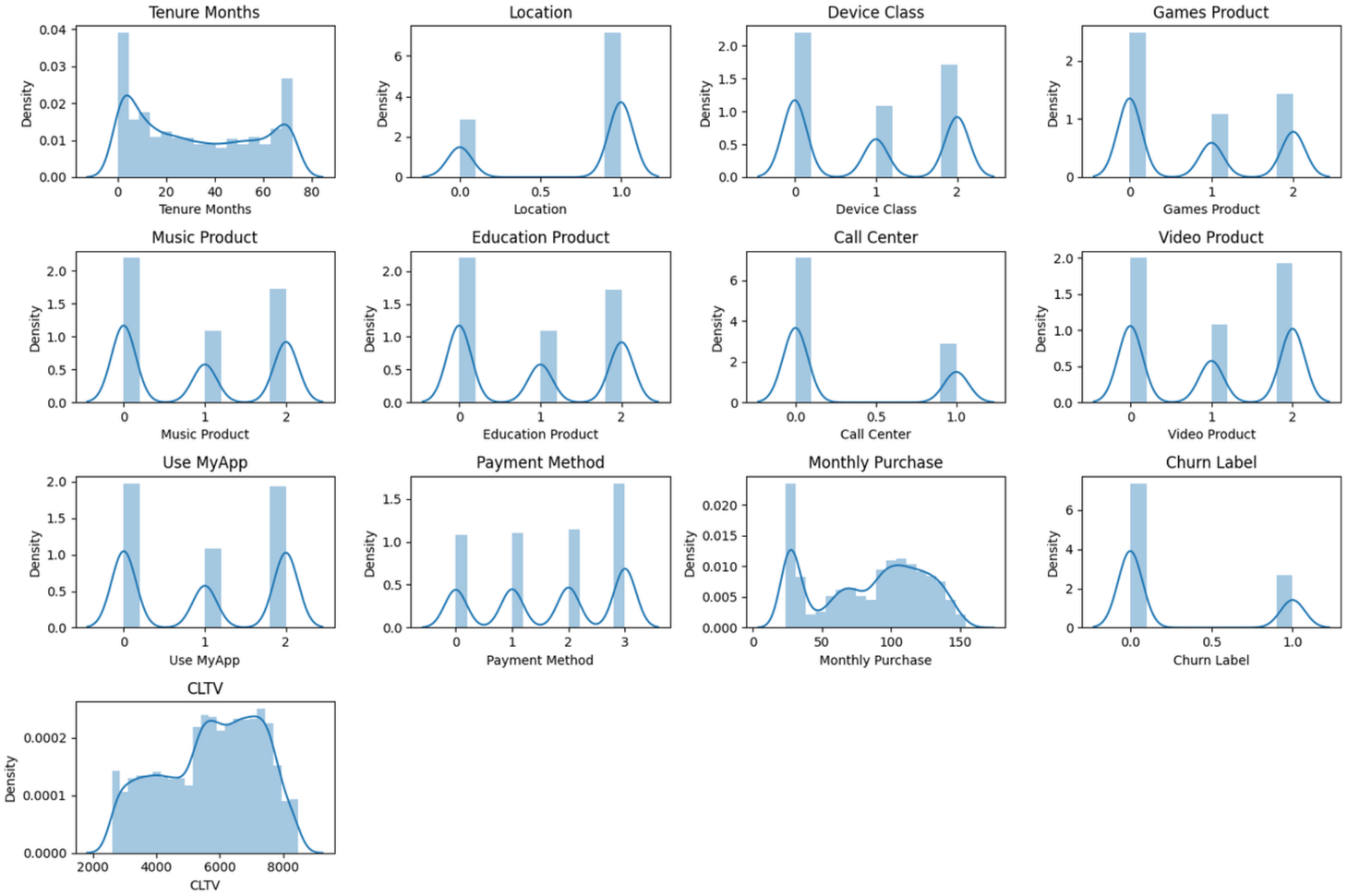
Data Visualization

Deep-dive Exploration

Conclusion

Data Distribution

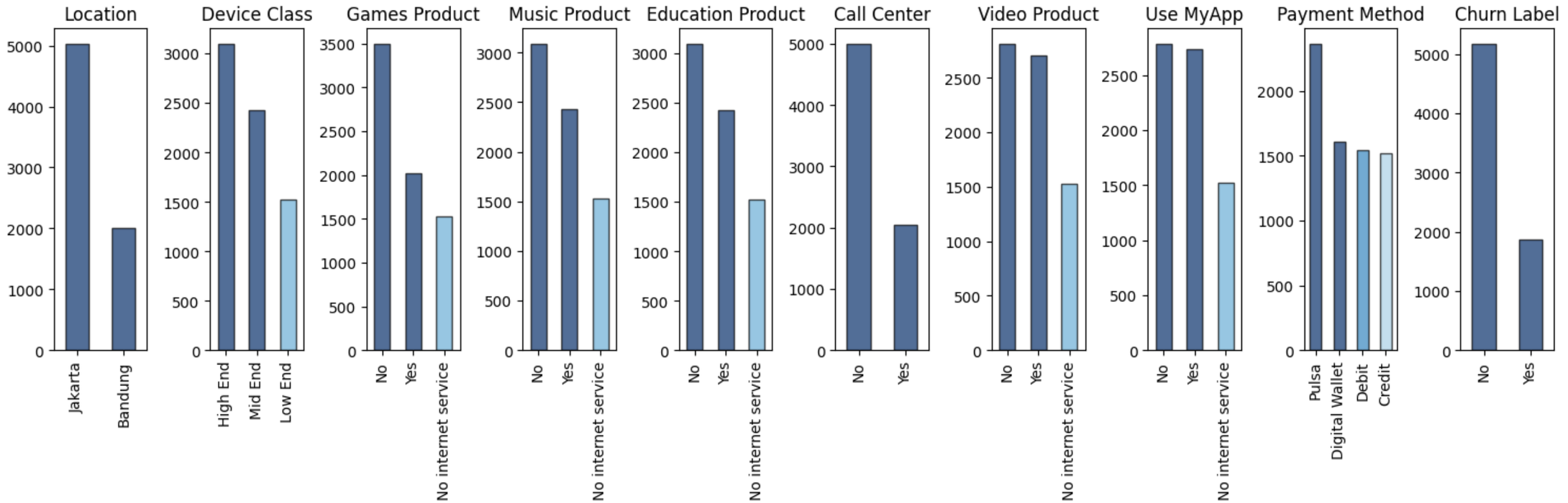
Data Distribution



Conclusion :

The spread is quite diverse

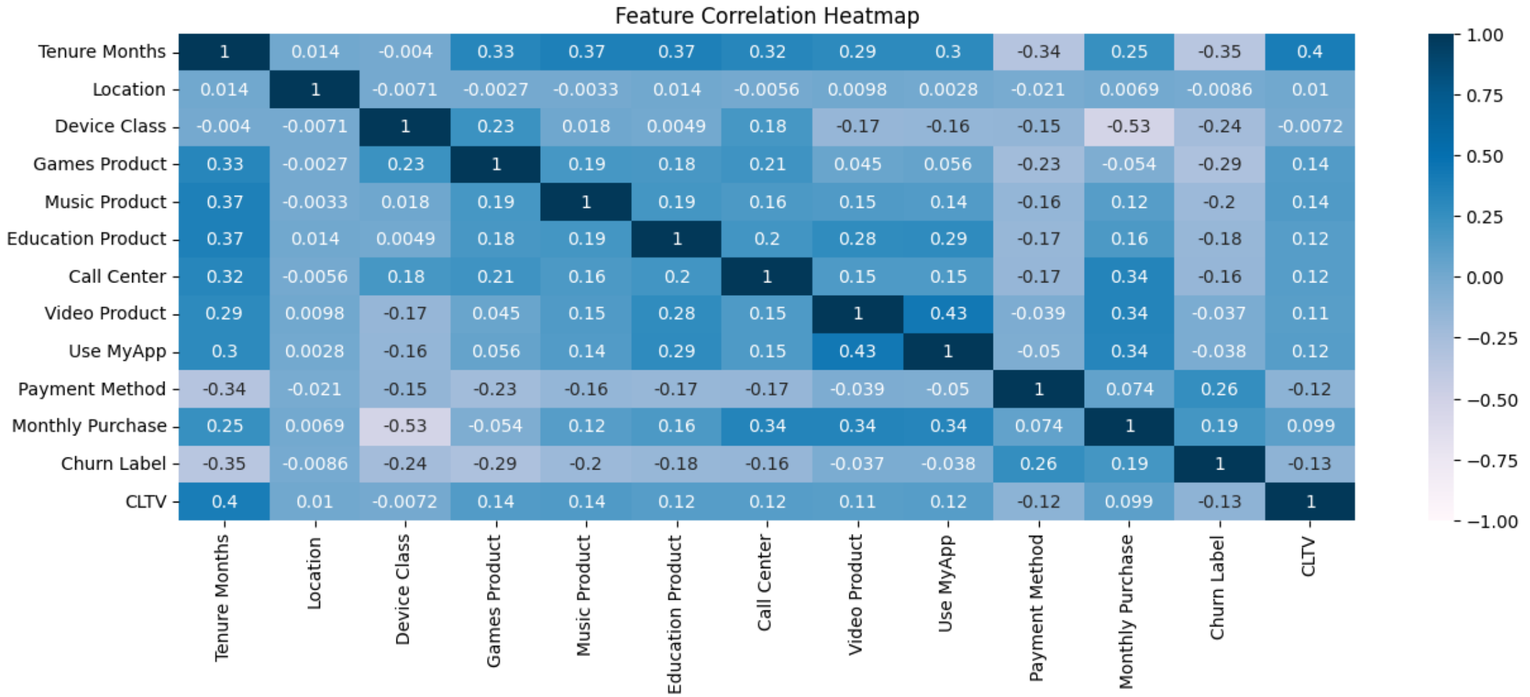
Univariate Analysis

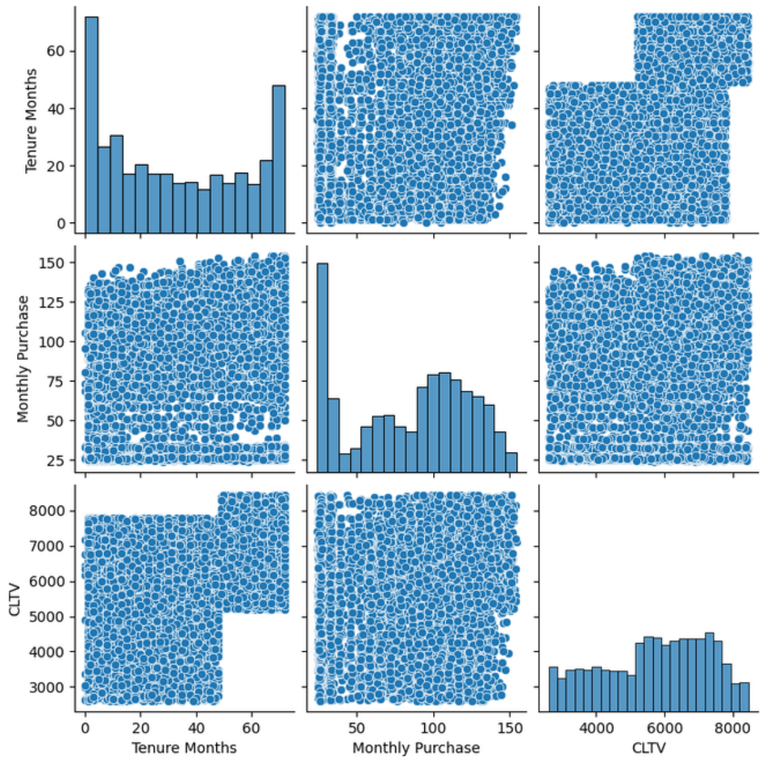


Conclusion :

the categorical columns tend to be **imbalance**. this will have an impact when segmentation is performed later.

Multivariate Analysis

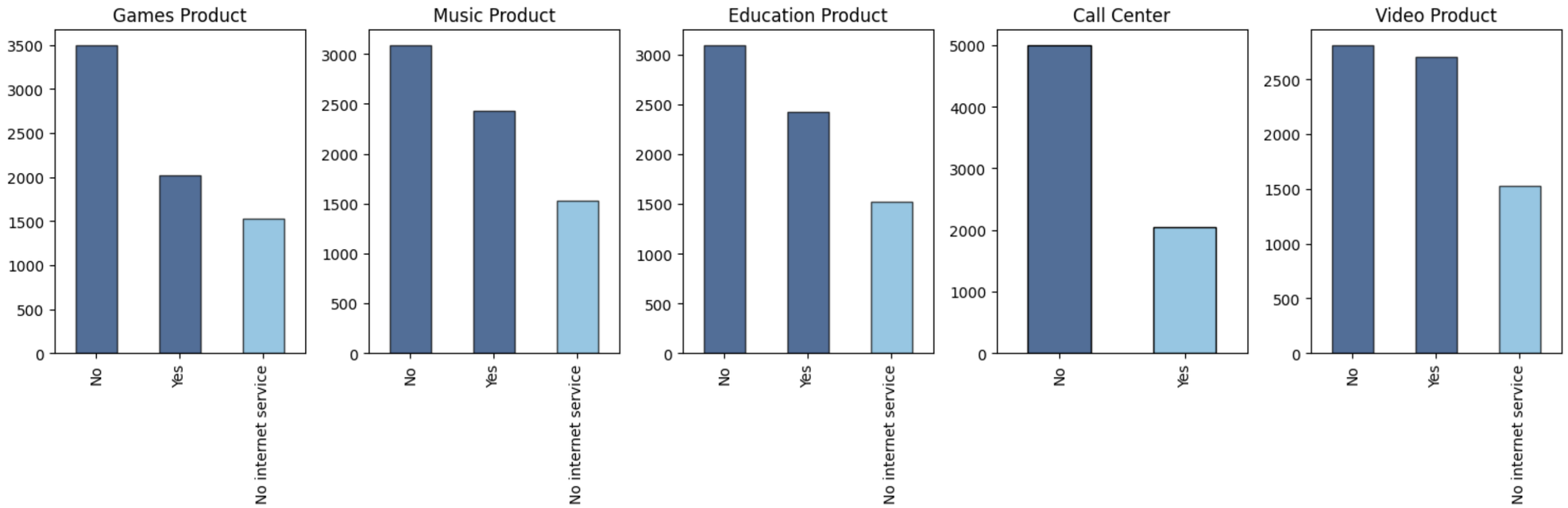




low correlation will have impact on segmentation later

Data Visualization

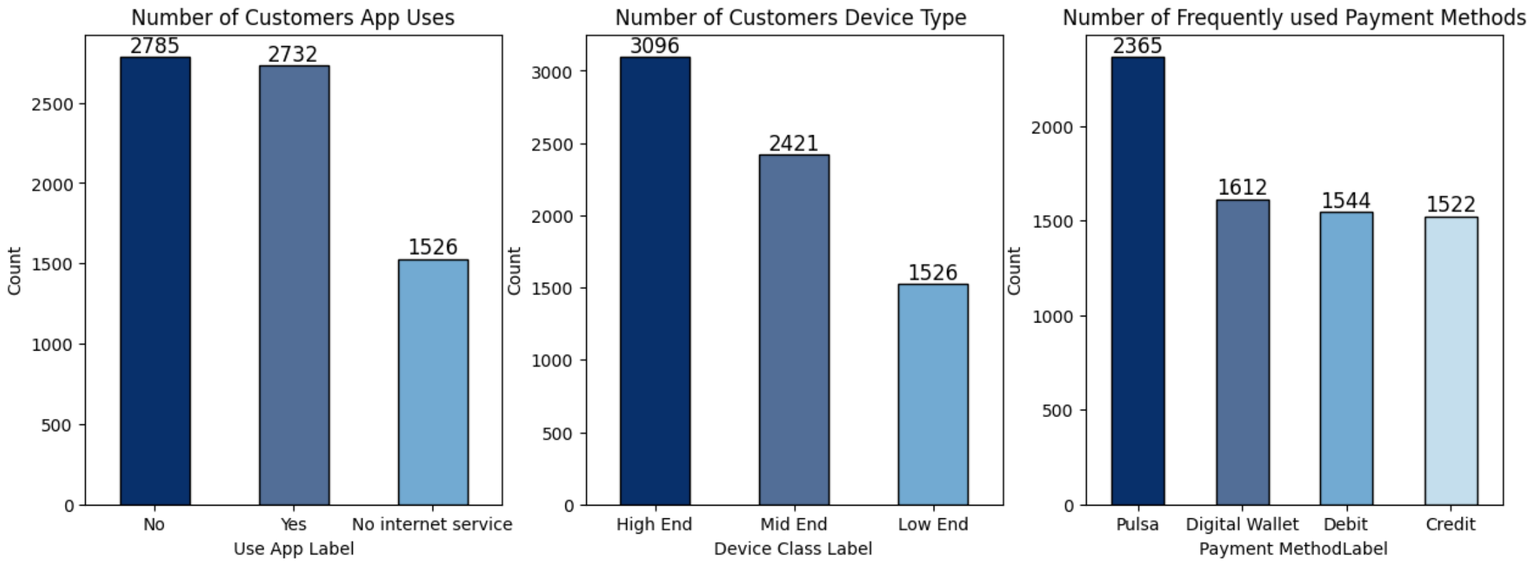
Number of Products Type Uses



Conclusion :

- Video services are the **most popular** among customers, with usage almost even between "Yes" and "No".
- Games, Music and Education services were **less popular**.
- 1526 customers **do not have internet service** in their packages.

Number of Payment Method Uses



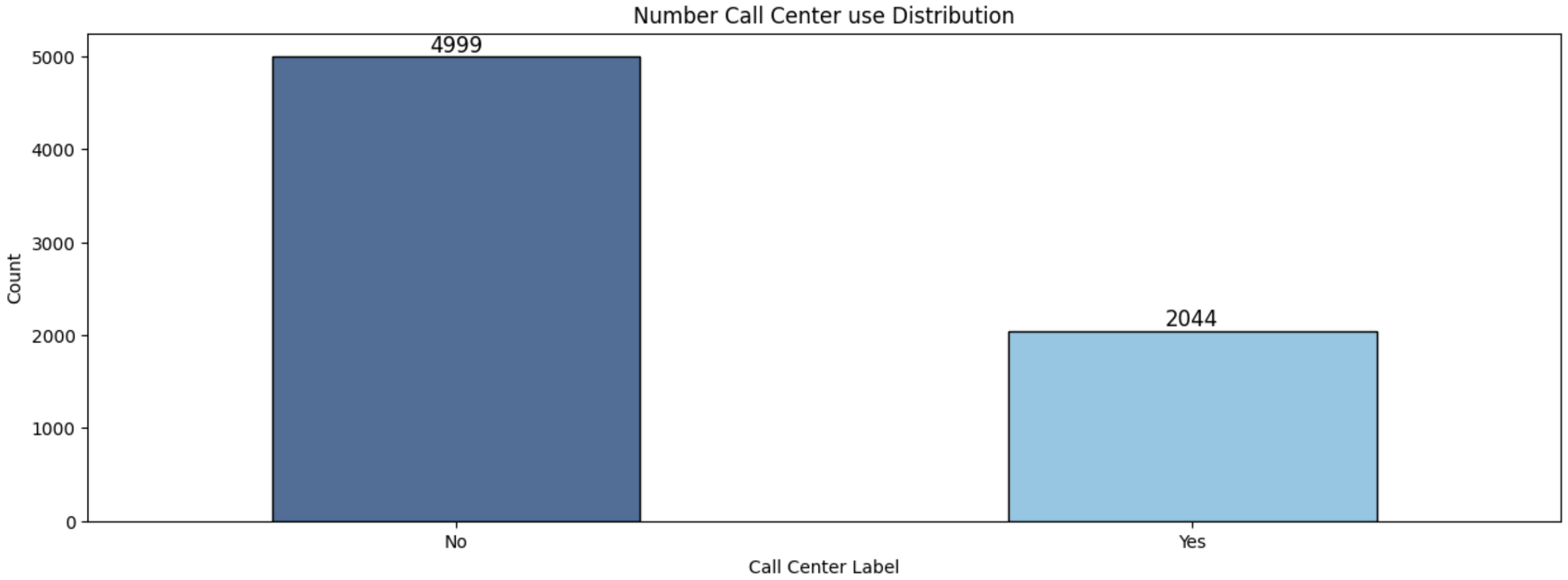
Conclusion :

- Pulsa has the highest share in the number of transactions and is also the most dominant payment method with 33.6% of total transactions. This shows that more customers choose to use Pulsa as a payment method.

- Digital Wallet, despite having a significant number of transactions, is still below Pulsa. This shows that although quite popular, the use of digital wallets has not yet reached the level of credit adoption.

- Debit and Credit are in almost equal proportions, with Debit being slightly higher. This indicates that most customers prefer to use their debit or credit card as a payment method.

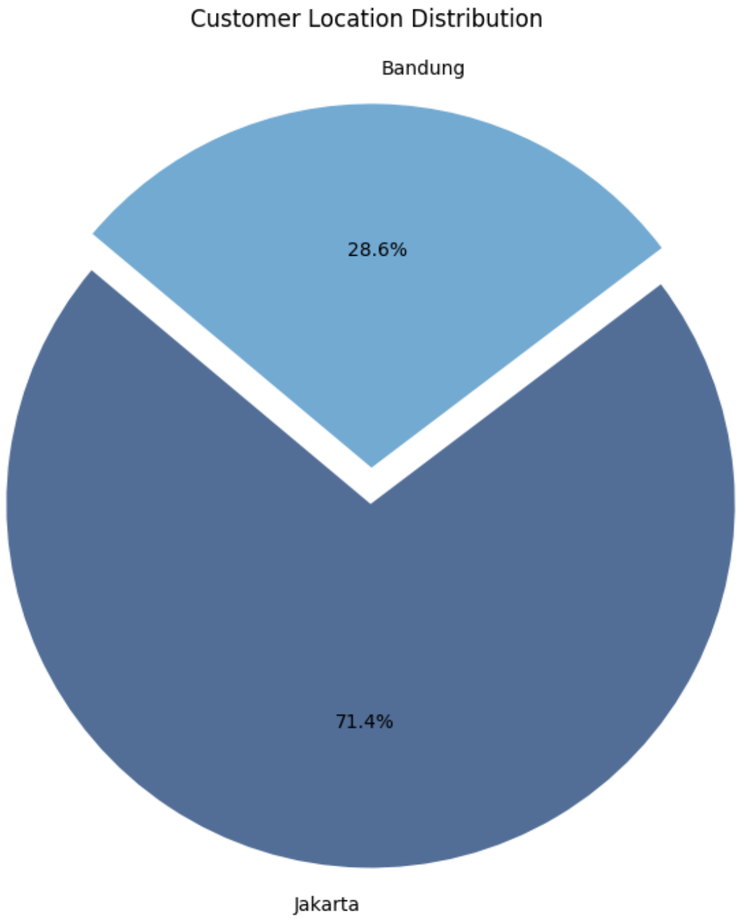
Number of Call Center Uses



Conclusion :

The total number of customers who use call center service is 2044, while those who do not use are 4999.

Customer Location Distribution



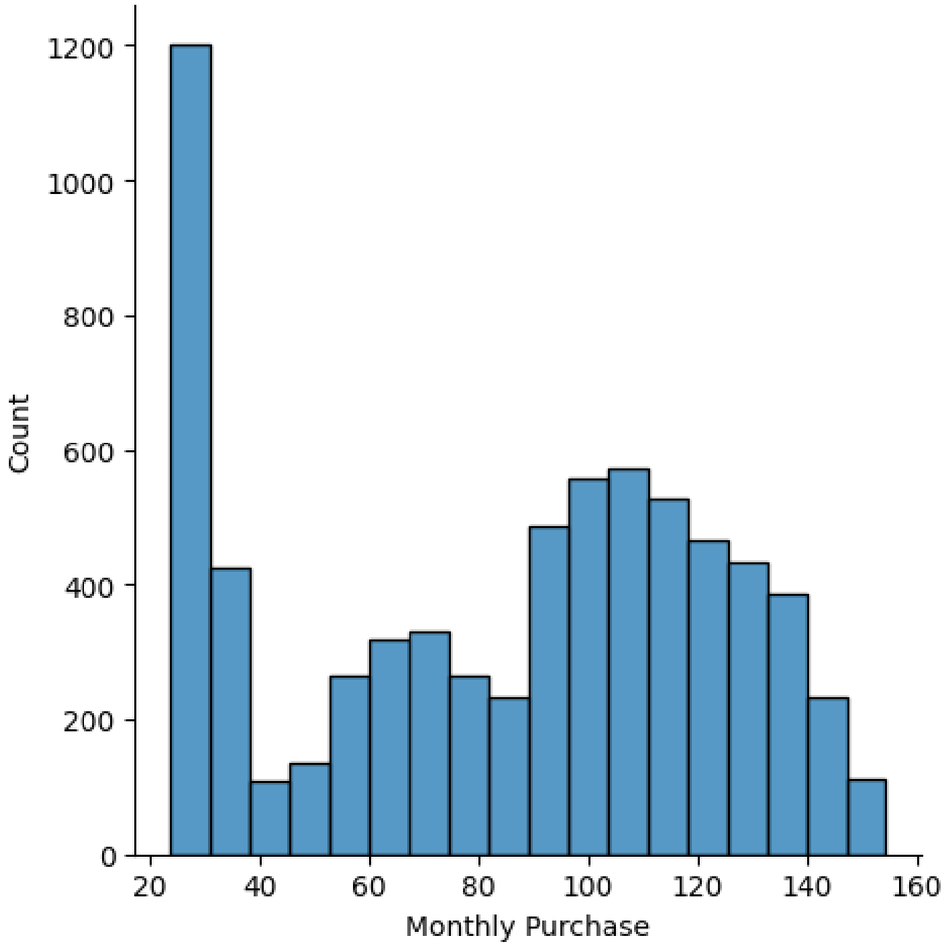

```
Jakarta    5031
Bandung    2012
Name: Location, dtype: int64
```

Conclusion :

- Jakarta has more customers than Bandung, with 5031 customers (around 71.4%) located in Jakarta and 2012 customers (around 28.6%) located in Bandung.

- This depicts the distribution of customers per location clearly, with the majority of customers being in Jakarta.

Monthly Purchase Distribution



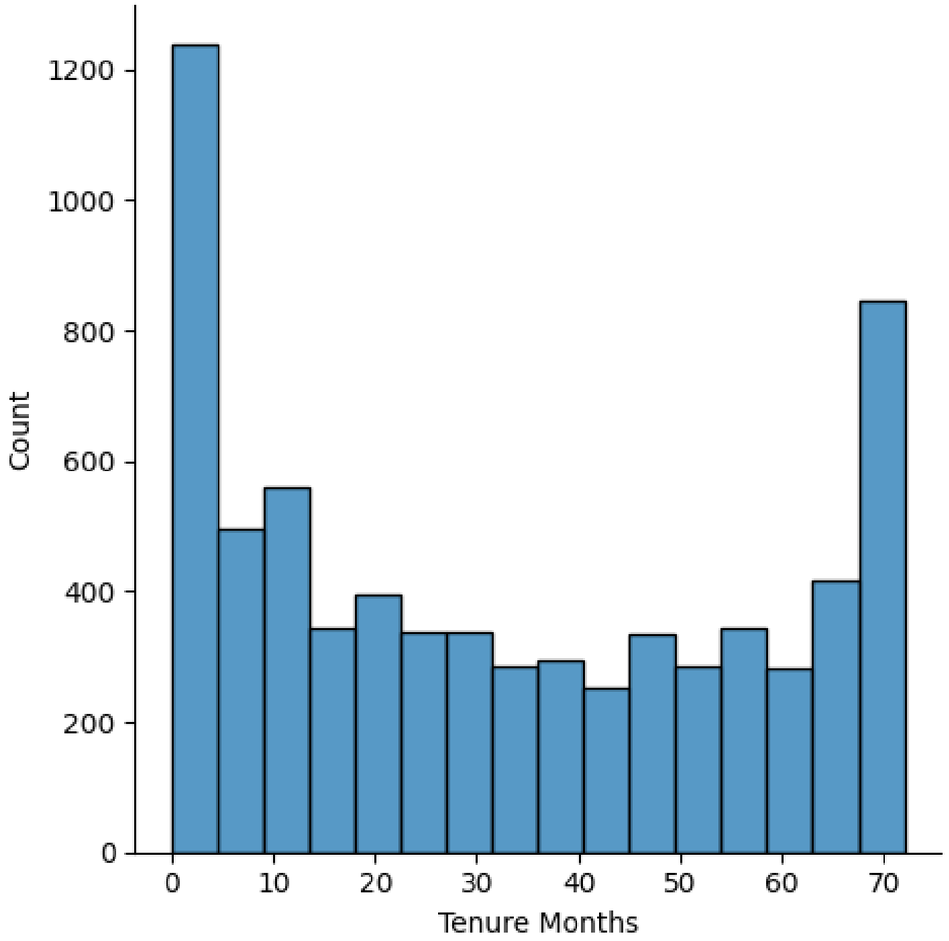
Conclusion :

- The highest purchase frequency is in the range of 20-40 thousand rupiah, with a total of 1600. This shows that most customers make purchases of **relatively small value**.

- Only a small percentage of customers make purchases of relatively large value.

- Purchase frequency patterns **tend to decrease** from the lower value range to the higher value range.

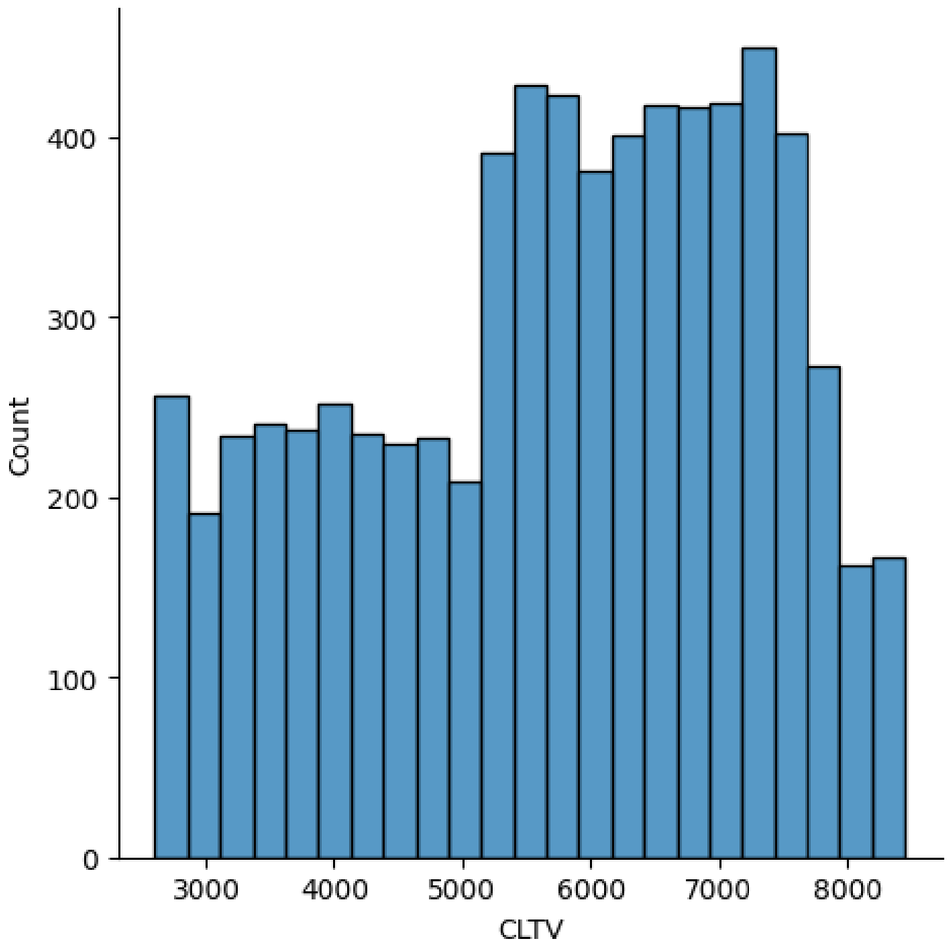
Tenure Months Distribution



Conclusion :

- The distribution of length of service usage is quite even, with the majority of customers (50%) having a usage period of less than 29 months.
- There was significant variation in length of use of the service, with some customers being new and some customers having been using the service for a long time.
- The existence of customers who have used the service for 6 years (72 months) shows that there is a very loyal group of customers.

Customer Lifetime Value Distribution

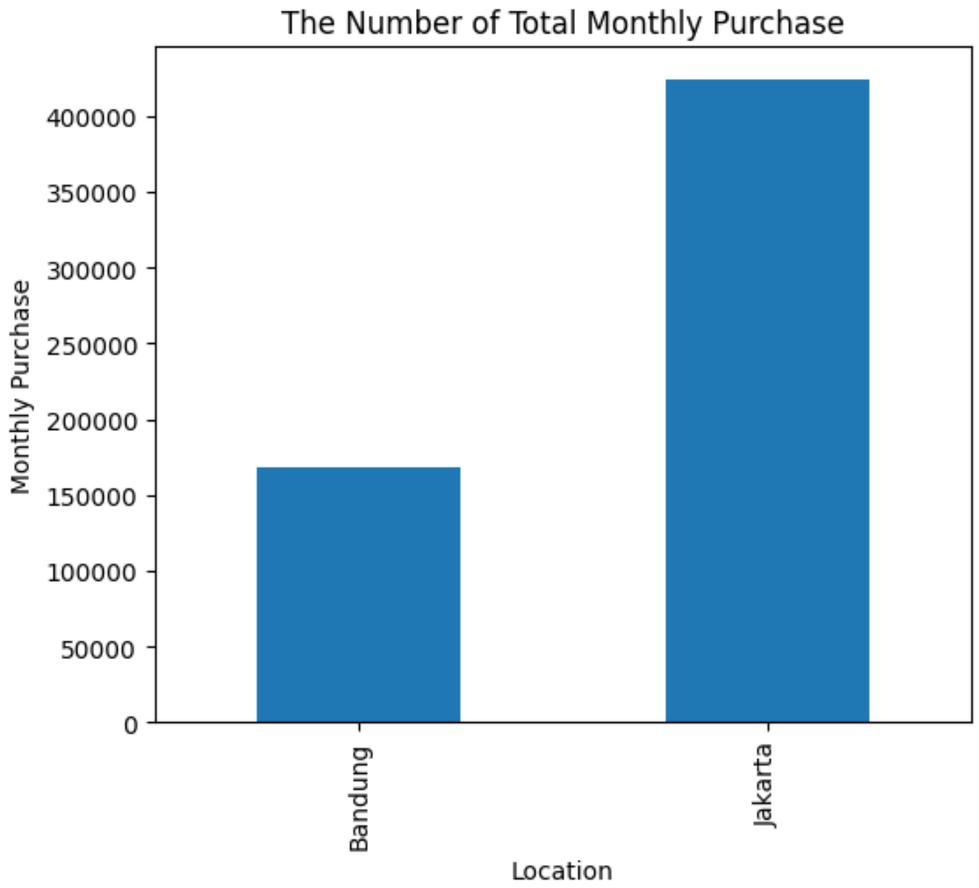


Conclusion :

- The distribution of CLTV values is quite varied, with the majority of customers (50%) having CLTV values below 5885.10.
- There is significant variation in CLTV values, with some customers having low CLTV values and some customers having high CLTV values.

Deep-dive Exploration

Monthly Purchase by Location



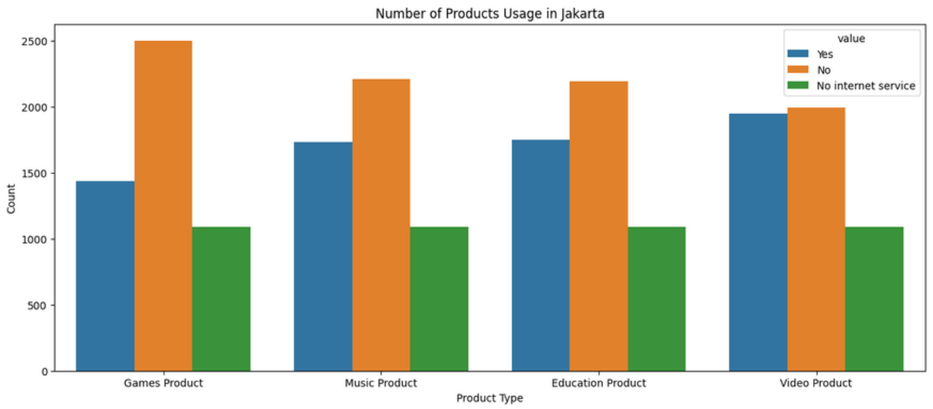

```
region_average = df.groupby('Location')['Monthly Purchase'].mean()
region_average
```

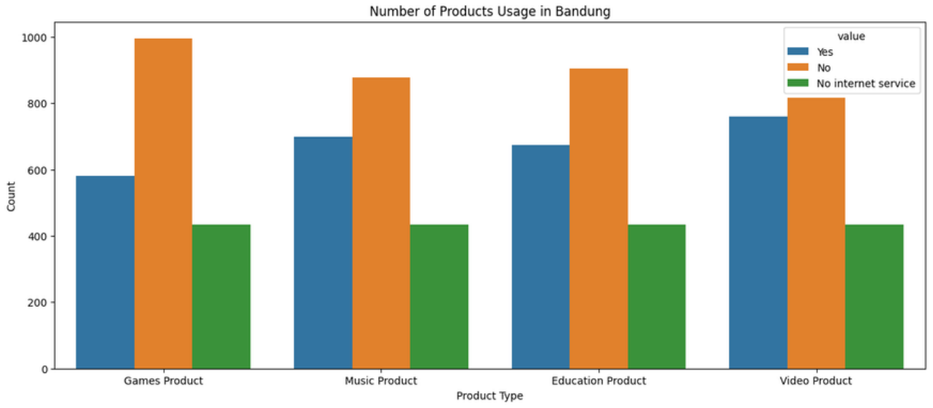
```
Location
Bandung    83.760447
Jakarta    84.362067
Name: Monthly Purchase, dtype: float64
```

Recommendatiion :

- Segment the customers based on their purchasing behavior and demographics.
- Explore the monthly purchase trends over time for both locations. Look for any seasonal patterns, trends, or anomalies that could provide insights into the purchasing behavior of customers.
- Evaluate the impact of marketing campaigns on customer engagement and purchasing behavior.

Products Usage by Location





Recommendation:

Product Optimization in Each Location:

Identify products that have a low frequency of use in each location.

Check whether the product is not in demand or whether additional marketing strategies are needed.

Marketing Strategy Adjustments:

Customize marketing strategies for each product based on customer preferences in each location.

There may need to be promotions or special offers tailored to the needs and preferences of customers in Bandung and Jakarta.

Joint Product Analysis:

Conduct further analysis to see if there are patterns or trends in the use of product combinations at each location.

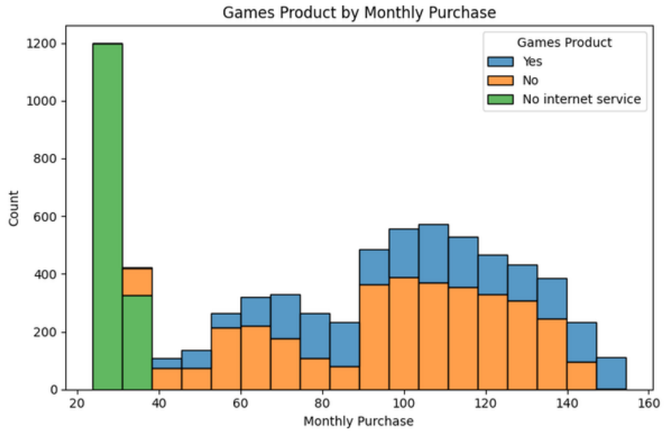
Consider putting together product packages or offering combination promotions based on customer preferences in each location.

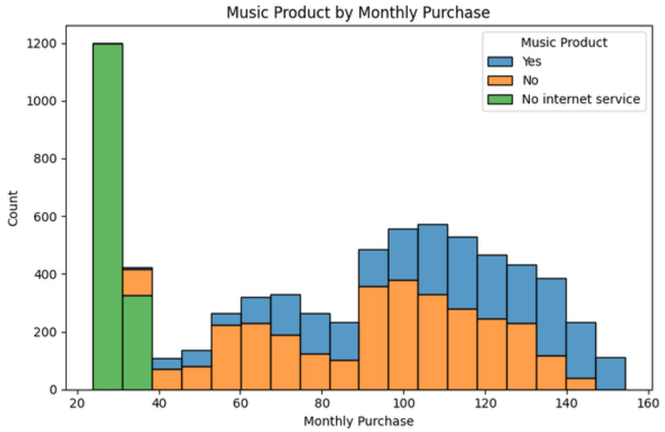
Internet Service Evaluation:

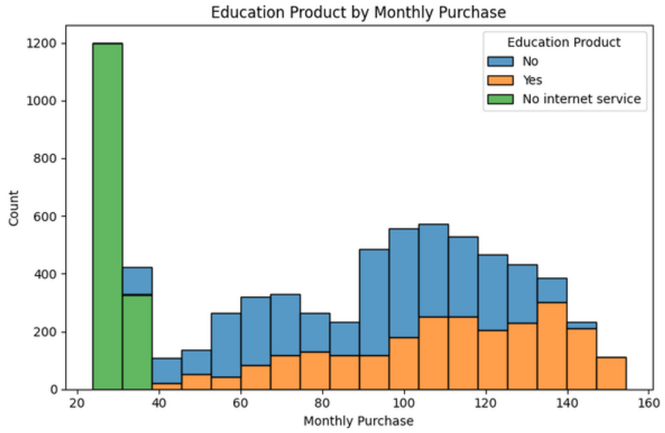
Check customers who have the label "No internet service" to see if they really don't use internet service.

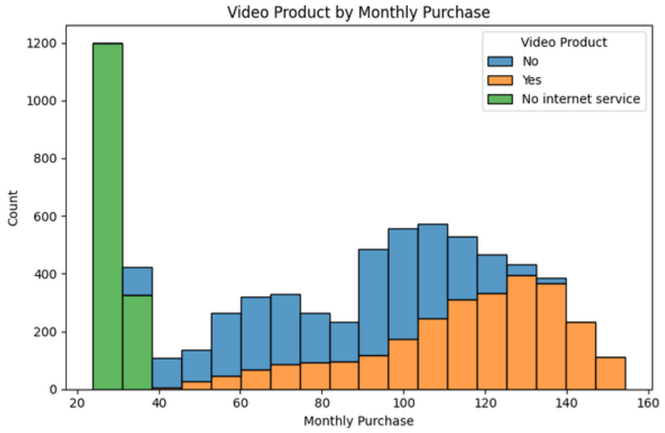
Consider offering appropriate internet packages or marketing strategies, especially if there are customers in both locations who do not use internet service.

Products Usage by Monthly Purchase









Recommendation :

- It is necessary to identify the cause of "No internet service" and ensure that customers in all locations have adequate access to the internet.
- Consider creating product bundle offers or subscription packages that can increase customer interest and motivate the purchase of more products.

Conclusion

In majority, Jakarta has the highest monthly purchase

Recommendation :
It is necessary to evaluate product marketing in Bandung, and adjust customer preferences and improve service in each location.

No internet service is Quiete High in Each Products

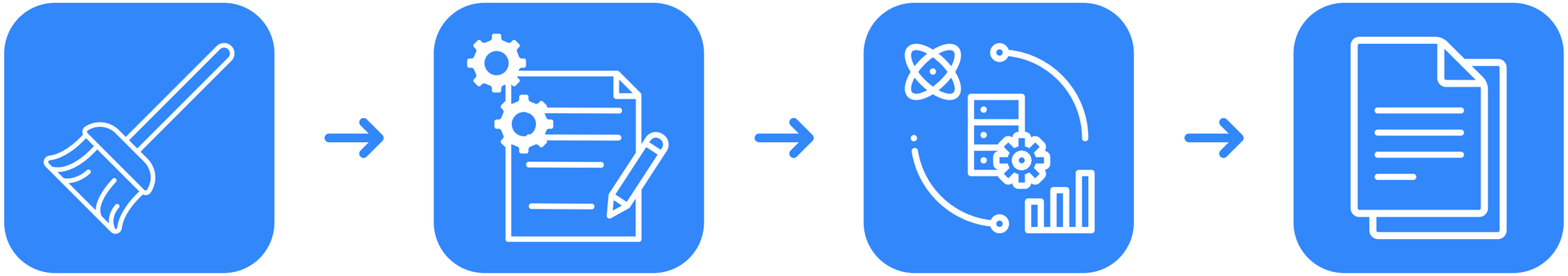
Recommendation :
because the number of customers who experience no internet service greatly influences customer churn. It is best to immediately identify the cause of no internet service in customers in each area to improve service performance.

Machine Learning

Churn Prediction

Customer Segmentation

Step



Data Cleaning

Feature Engineering

Modeling

Evaluating

Data Cleaning

Missing Value

There’s no missing value in the dataset

Duplicated Value

There’s no duplicate value

Outliers

No Outliers in dataset

Feature Engineering

Encoding

10 Variable

- Location
- Device Class
- Games Product
- Music Product
- Education Product
- Call Center
- Video Product
- Use MyApp
- Payment Method
- Churn Label

Scaling

Standard Scaler

Imbalance

Imbalance will be left to avoid overfitting

Modeling

Logistic Regression Model

Random Forest Model

Logistic Regression Model Evaluation

```
Confusion Matrix :
[[902 107]
 [198 202]]


Classification Report :
              precision    recall  f1-score   support

           0       0.82      0.89      0.86      1009
           1       0.65      0.51      0.57       400

    accuracy                           0.78      1409
   macro avg       0.74      0.70      0.71      1409
weighted avg       0.77      0.78      0.77      1409


All AUC Scores :
[0.81252452 0.85927983 0.83304772 0.84047208 0.82636353 0.79565366
 0.82028155 0.80926571 0.80893472 0.82431395]


Mean AUC Score :
Mean AUC Score - Logistic Regression :  0.8230137264348834
```

- The model successfully identified 902 negative cases and 202 positive cases correctly.
- There were 107 cases that should have been negative but were incorrectly identified as positive.
- A total of 198 positive cases were not detected by the model.
- The overall accuracy of the model was 0.78, which reflects the extent to which the model actually classified instances correctly
- The model tends to be better at classifying the negative class (0) than the positive class (1).
- This Logistic Regression model has good performance with an AUC of around 0.82, however, it is still possible to carry out further exploration to improve performance, especially for positive class recall.

Random Forest Model Evaluation

```
Confusion Matrix :
[[897 112]
 [210 190]]


Classification Report :
              precision    recall  f1-score   support

           0       0.81      0.89      0.85      1009
           1       0.63      0.47      0.54       400

    accuracy                           0.77      1409
   macro avg       0.72      0.68      0.69      1409
weighted avg       0.76      0.77      0.76      1409


All AUC Scores :
[0.78822291 0.82953771 0.82744203 0.82848395 0.80414568 0.78475677
 0.81038281 0.80172013 0.79245751 0.81771287]


Mean AUC Score :
Mean AUC Score - Random Forest:  0.8084862380305914
```

- The model successfully identified 895 negative cases and 184 positive cases correctly.
- There were 114 cases that should have been negative but were incorrectly identified as positive.
- A total of 216 positive cases were not detected by the model.
- The overall accuracy of the model was 0.77, reflecting the extent to which the model actually classified instances correctly.
- The mean AUC score of 0.81 indicates that the model has good performance in distinguishing between positive and negative classes. AUC above 0.8 is considered good.
- The model tends to be better at classifying the negative class (0) than the positive class (1).

Conclusion

VS

Logistic Regression Model

- Accuracy:
  Logistic Regression had slightly higher accuracy (78% vs. 77%).
- Precision:
  Both Logistic Regression and Random Forest have higher precision for class 0
  (negative) than for class 1 (positive).
- Recall:
  Both Logistic Regression and Random Forest have higher recall for class 0
  (negative) than for class 1 (positive).
- F1-Score:
  Both Logistic Regression and Random Forest have a higher F1-score for class 0
  (negative) than for class 1 (positive).
- AUC-ROC:
  Logistic Regression has a slightly higher Mean AUC Score (0.82 vs. 0.81).

Random Forest Model

Conlcusion :

- In general, both models show similar performance, with Logistic Regression tending to provide a slight edge in some evaluation metrics.

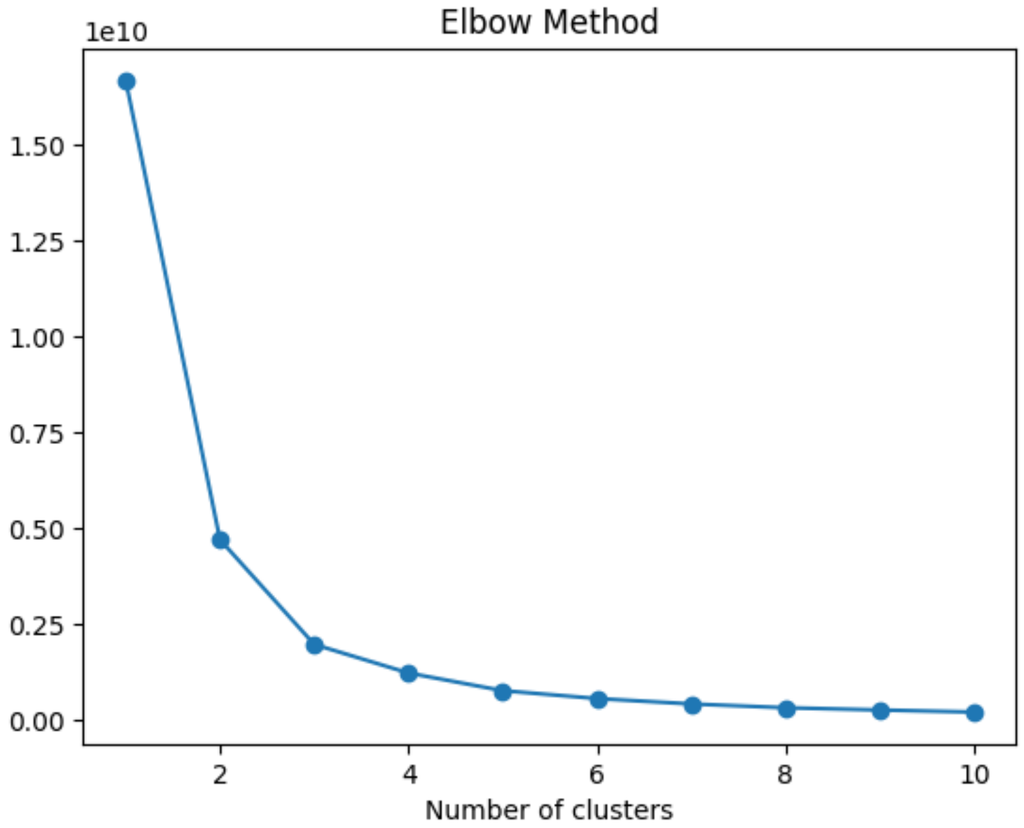
- The choice between Logistic Regression and Random Forest can depend on data characteristics, model interpretation, and specific project needs.

- If model interpretability and ease of configuring the model are important, Logistic Regression may be preferred.

- If model complexity and the ability to handle non-linear features are priorities, Random Forest can be a good choice.

Customer Segmentation

Modeling: Elbow Method



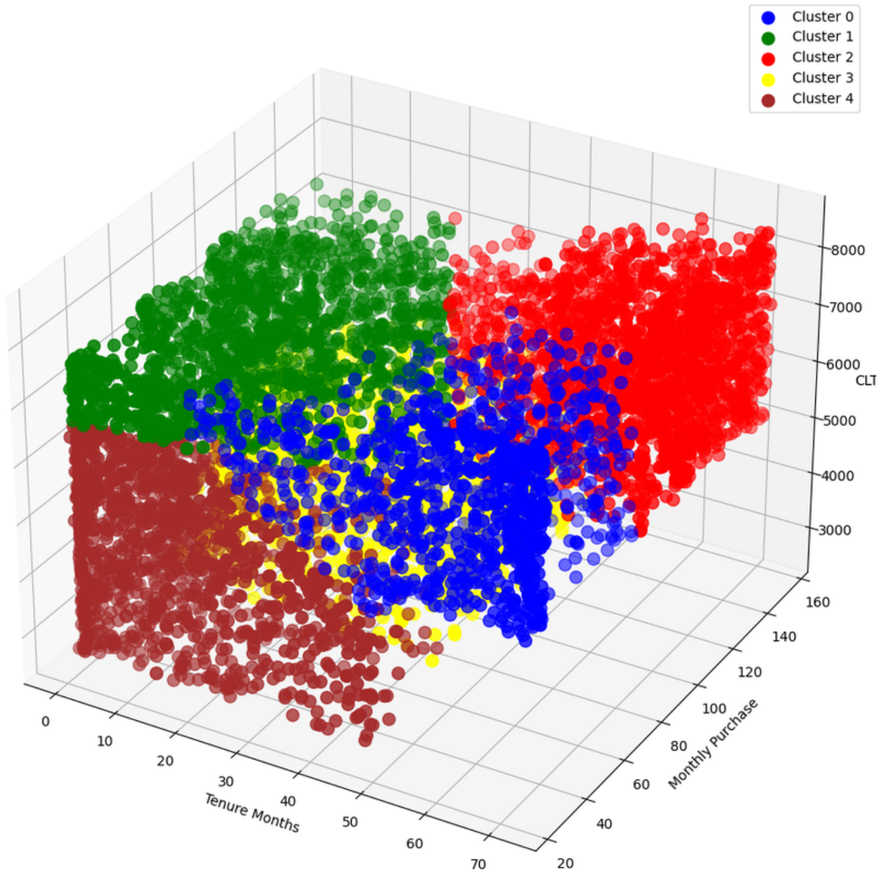
Modeling: Shillouette Score

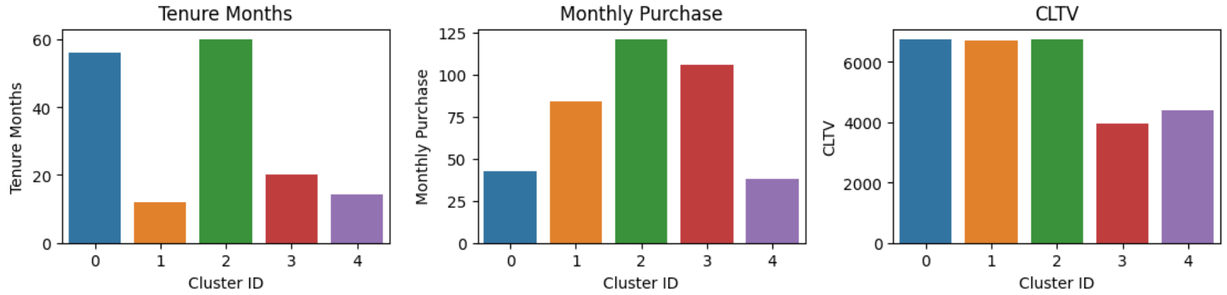
```
For n_clusters=2, the silhouette score is 0.3391790795928419
For n_clusters=3, the silhouette score is 0.3278001841544866
For n_clusters=4, the silhouette score is 0.3367210529285514
For n_clusters=5, the silhouette score is 0.3559082414346754
For n_clusters=6, the silhouette score is 0.3532375478748478
For n_clusters=7, the silhouette score is 0.3221663151871398
For n_clusters=8, the silhouette score is 0.3144665933691782 6
```

Conclusion :

- The best number of clusters (n_clusters) may be the value that gives the highest Silhouette score. In this case, n_clusters=5 has the highest Silhouette score.
- It is important to remember that Silhouette Score evaluation should be used in conjunction with an understanding of the data domain and external evaluation to ensure selection of the optimal number of clusters.
- The higher the Silhouette score, the better the cluster is considered to be within clearer boundaries.

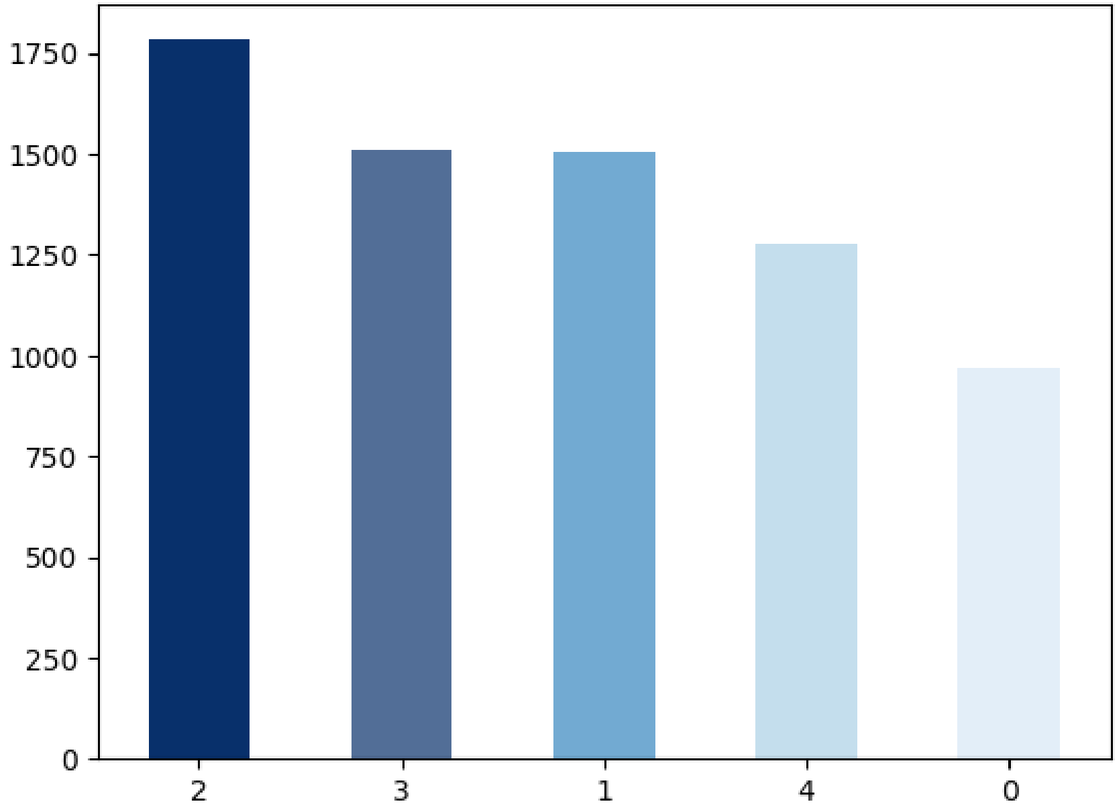
Modeling: K-means Clustering





- Cluster 0 and Cluster 2 have high CLTV values, indicating good customer value, while Cluster 3 and Cluster 4 have lower CLTV values.
- Cluster 0 and Cluster 2 also have longer tenure and high monthly purchase rates, indicating loyal customers with good financial contributions.
- Cluster 1 has a high CLTV value despite a low monthly purchase rate, indicating the growth potential of these customers with the right marketing strategy.
- Cluster 3 and Cluster 4, although having moderate to long tenure, have lower CLTV values, and may need special attention to increase their financial contribution.

Evaluating: Cluster



	Customer ID	Tenure Months	Monthly Purchase	CLTV	Cluster ID
0	0	2	70.005	4210.7	4
1	1	2	91.910	3511.3	3
2	2	8	129.545	6983.6	1
3	3	28	136.240	6503.9	2
4	4	49	134.810	6942.0	2

Conclusion :

Cluster 2 and Cluster 3 are the largest clusters with a significant number of customers.

Cluster 0 has the smallest number of customers.

Analysis: Market Segmentation

High - Value Loyal Customer

Customers with high monthly purchase rates, high lifetime value (CLTV), and long service life.

Potential Growth Customers

Customers with growth potential, good lifetime value (CLTV) despite low monthly purchase rates, and shorter service life.

Moderate-Value Loyal Customers

Customers with medium to low monthly purchase rates and lifetime value (CLTV), but a long service life.

Low - Value Customer

Customers with low monthly purchase rates and lifetime value (CLTV), as well as shorter service life.

Conclusion

Monthly Purchase:
There is significant variation in monthly purchasing levels within each cluster.
Cluster 2 has a high monthly purchase rate, while clusters 0, 1, 3, and 4 tend to have more stable rates.

Customer Lifetime Value (CLTV):
Clusters 0, 2, and 4 have high lifetime values, indicating the potential for significant financial contributions.
Cluster 3 has lower lifetime value, and can be the focus of marketing strategies to increase value.

Tenure Months:
Clusters 0, 2, and 4 have a long period of service use, showing the potential to become loyal customers.
Clusters 1, 3, and 4 have shorter tenures, requiring more intensive retention strategies.

Recommendation

Personalize Customer Experience:
Based on segmentation analysis, personalization of customer experience can be implemented according to the profile and preferences of each cluster.

The Right Retention Strategy:
Clusters with shorter tenures (Clusters 1, 3, and 4) require more intensive retention strategies. Special offers, discounts or loyalty programs can be used to extend their usage period.

Optimize Services for High CLTV Clusters:
Clusters with high lifetime value (Clusters 0, 2, and 4) can be targets for premium services or exclusive offers.

Product Combination Analysis:
Carry out further analysis of the product combinations used by customers in each cluster to optimize package offers or product promotions.

Continuous Monitoring and Evaluation:
Conduct continuous monitoring and evaluation of customer responses to implemented marketing and retention strategies. Adapt strategies according to changing trends and customer preferences.

Improved Communication:
Improve communication with customers through channels that suit the preferences of each cluster.
Use sentiment analysis tools to understand customer feedback and respond quickly.

Thank You!

Any Question? Reach me at:

diimprasetyos@gmail.com

https://www.linkedin.com/in/diimprasetyo/

Google Colaboratory: https://colab.research.google.com/drive/1S4Fym4g-GCBtgHoGxpnqMqF-Uljyr44w?usp=sharing