



Telco Company Churn Analysis

Data Science Weekend 2023





DIKER TEAM

IN DATA WE TRUST



DHIEMAS ADY K W
COMPUTER SCIENCE
UPN 'VETERAN'
YOGYAKARTA (2019–2023)



KELVIN ANDERSEN
BACHELOR OF COMPUTER SCIENCE
INSTITUT TEKNOLOGI SEPULUH
NOPEMBER (2019–2023)

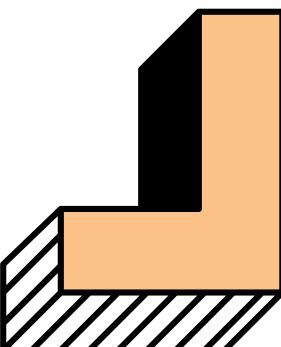


RIZAL MUJAHIDDAN
COMPUTER SCIENCE
IPB UNIVERSITY (2019–2023)



BACKGROUND STORY

- The company experienced a significant loss of potential customers amounting to 10 billion in Q3 (Based On Exploratory)
- Participants are tasked with conducting a comprehensive analysis to provide valuable insights for stakeholders
- Building strategies and measures to mitigate the likelihood of customers wanting to churn





METHODOLOGY AND TOOLS

- **Business Understanding**
 - Define problem, identify goals
- **Data Understanding**
 - Collect, explore, assess data quality.
- **Data Preparation**
 - Clean, preprocess, handle missing values.
- **Modelling**
 - Choose algorithms, split data, train models.
- **Deployment**
 - Deploy, monitor, ensure integration.

01

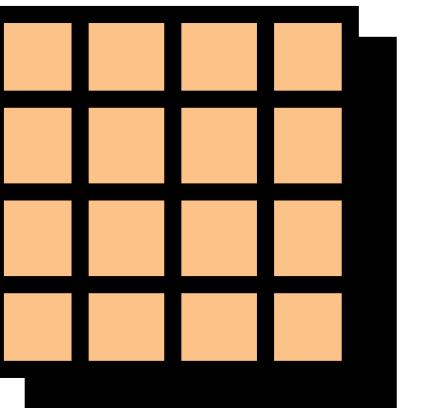
Python

02

Tableau

03

Google
Collab



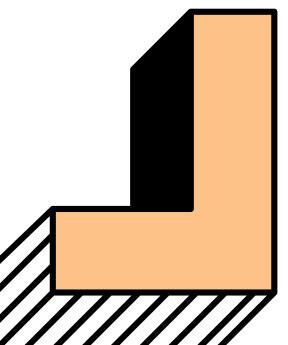
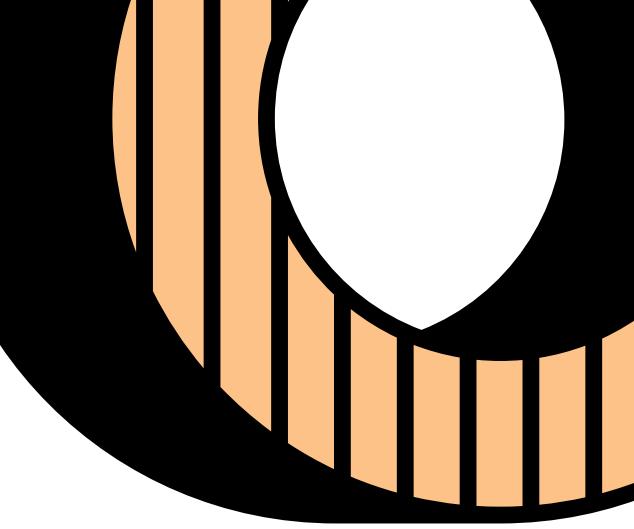
BUSINESS UNDERSTANDING

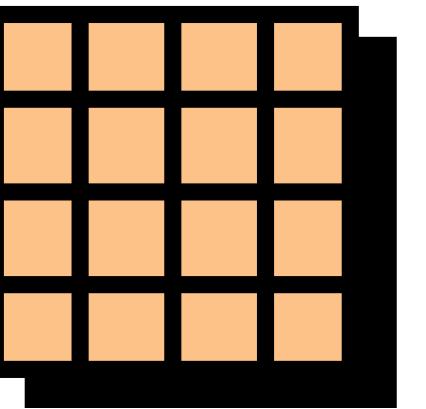




BUSINESS UNDERSTANDING

- What patterns define customer behavior in our telco services, revealing their unique interactions and preferences?
- Which factors hold the most influence on customer churn in our telco services, and how can we address them to strengthen customer loyalty





DATA UNDERSTANDING





THE STORY OF DATASET



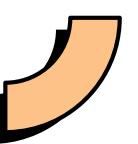
DESCRIPTION

The dataset presents telecommunication service utilization, such as games, music and education, during the third quarter of a specific year.

A screenshot of a dataset page. At the top, there's a search bar with a magnifying glass icon and the word "Search". Below it is a header with the "DATA SCIENCE INDONESIA" logo, the text "DATA SCIENCE INDONESIA · UPDATED A MONTH AGO · PRIVATE", and a download button labeled "Download (575 kB)". To the right of the download button is a three-dot menu icon. In the center, the title "Data Challenge DSW 2023 - Student & Junior Pro" is displayed in bold black text, followed by the subtitle "Data Challenge DSW 2023 Students & Junior Professional Category". To the right of the title is a logo for "DSW DATA SCIENCE WEEKEND" and "Powered by Telkomsel". Below the title, there are several sections: "Data Card" (underlined), "About Dataset", "Problem Statement", and "Dataset Description". The "About Dataset" section includes a detailed description of the dataset's purpose and participants' freedom to build AI models. The "Dataset Description" section provides a brief summary of the dataset's content. On the right side of the page, there are sidebar cards for "Usability" (4.71), "License" (Unknown), "Expected update frequency" (Never), and "Tags" (Business). A decorative orange ribbon graphic is located at the bottom right.

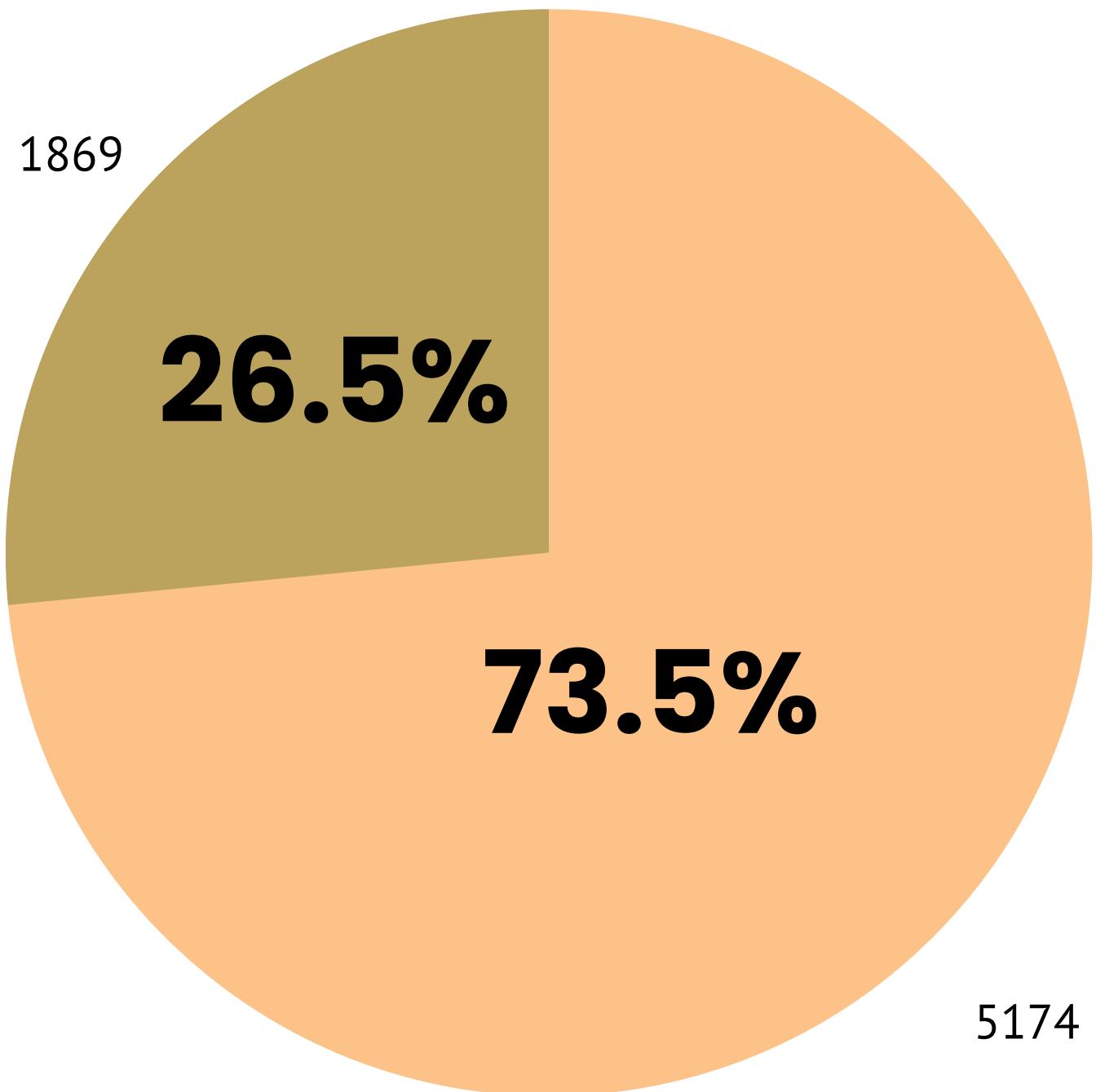
SIZE & QUALITY

Data has good quality with no missing or duplicate data, but too little features (for example: no gender or age data). The data set consist 7043 rows, 16 columns



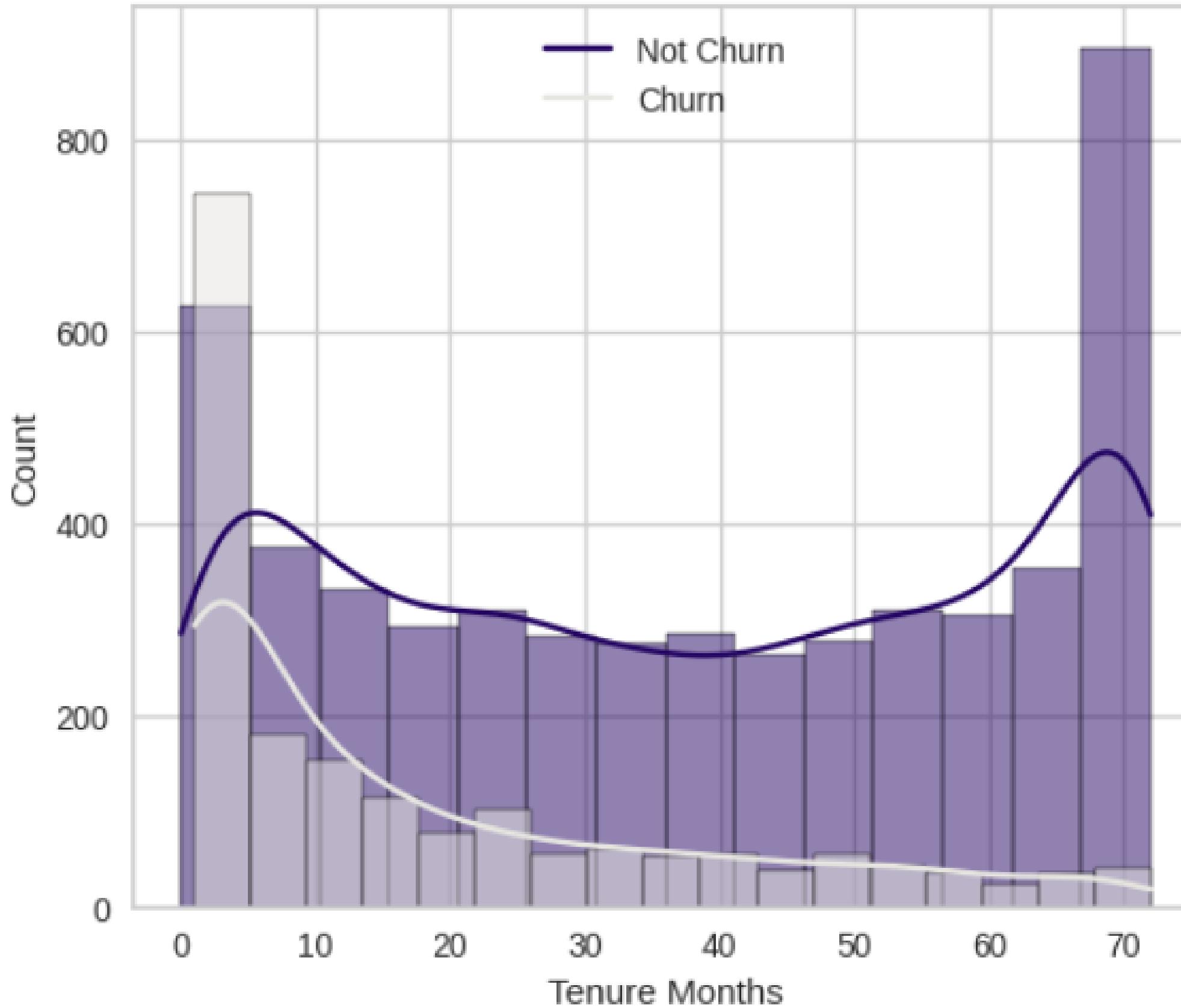


Churn percentage in Q3



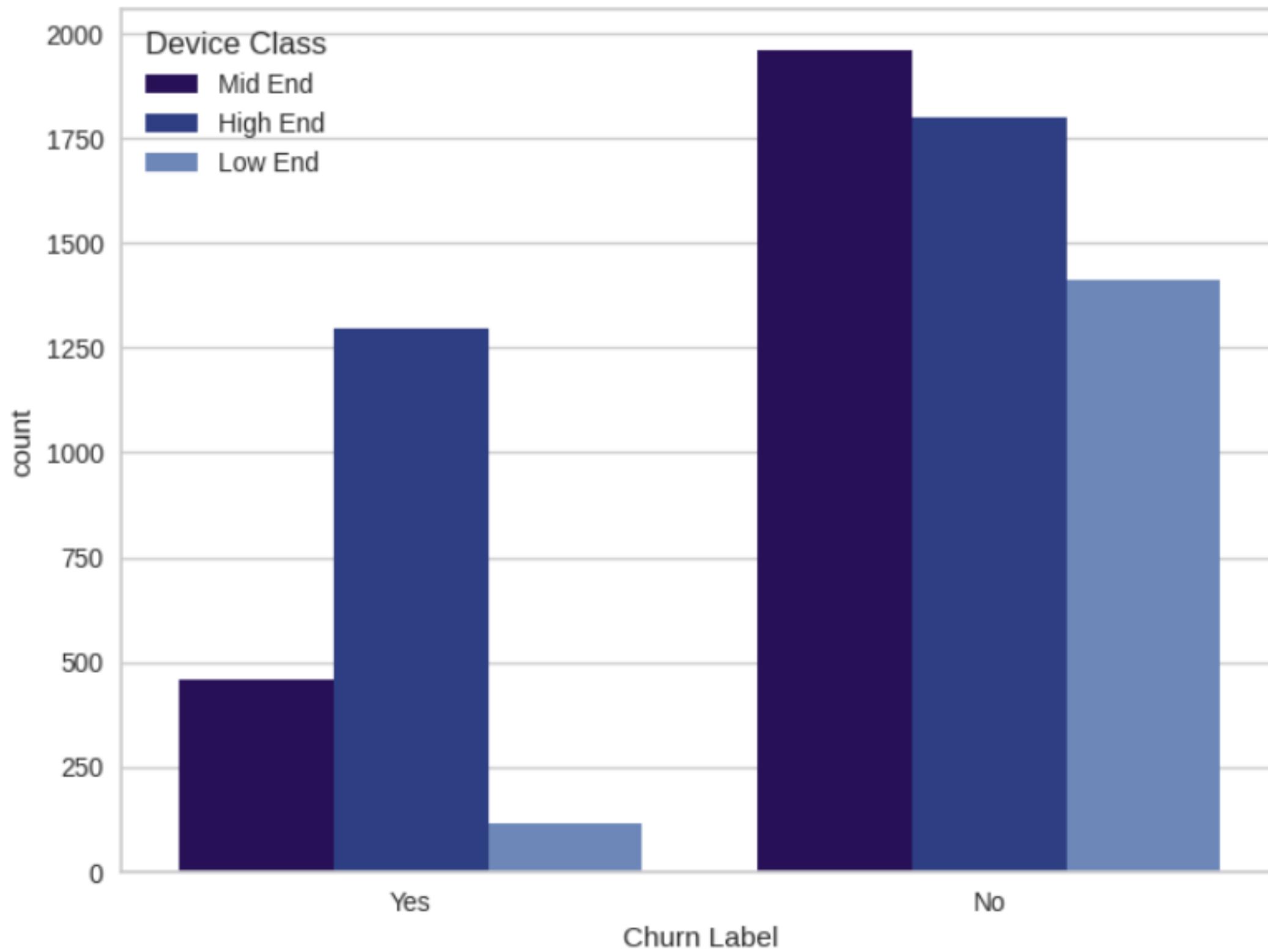
26.5% of 7043
Customers
recognized as
churned

highlighting a noteworthy portion that discontinued our services or subscriptions within third quarter period.



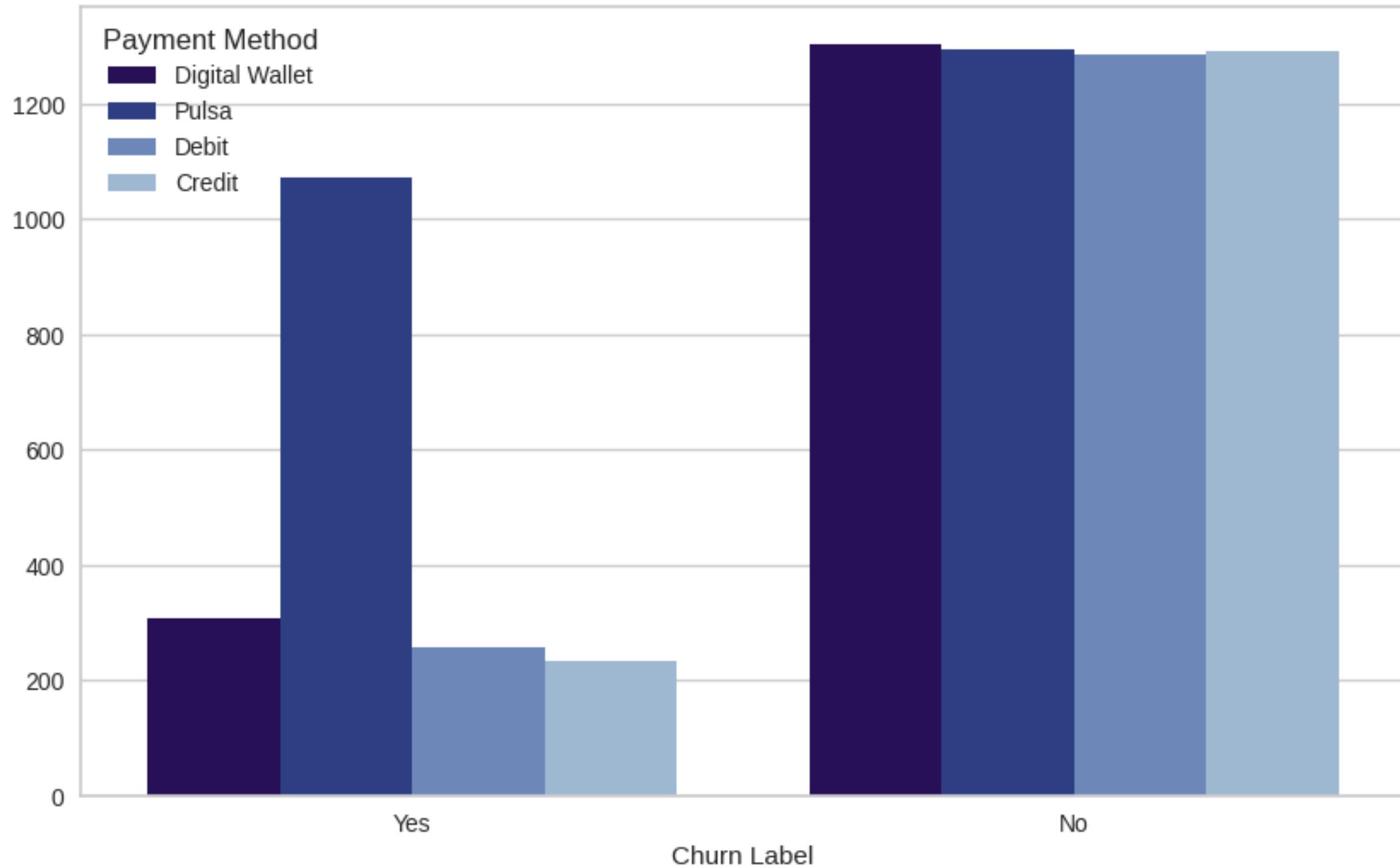
53.5% of churned customers are highly prone to churning within the **initial year of usage**

As time progresses, the percentage of churned customers tends to decrease from year to year in the presentation.



41.8% of churned customers are using high end device

High-end device users exhibit the highest churn, while low-end device users have lower churn, indicating potential product or service issues requiring investigation.

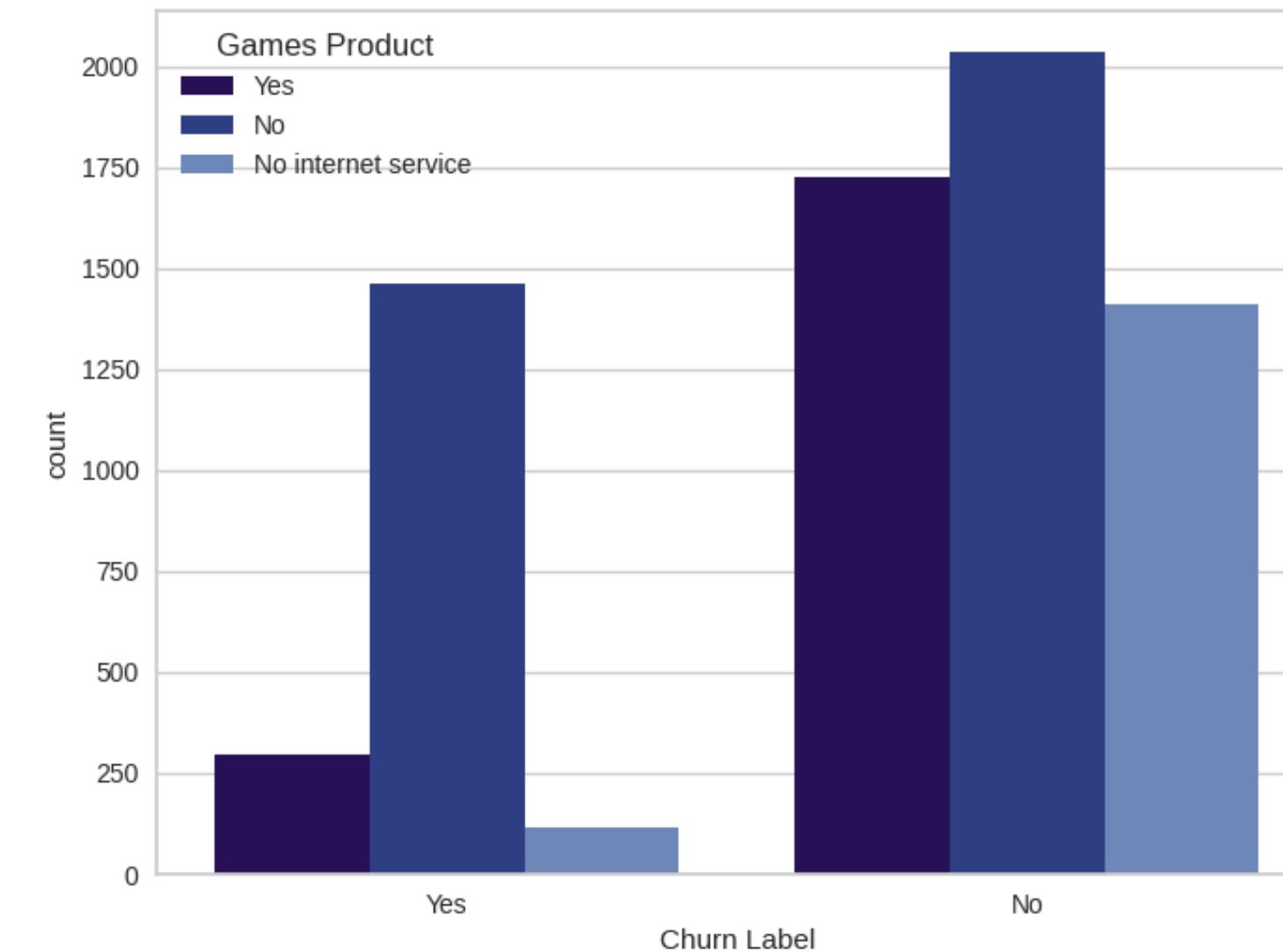
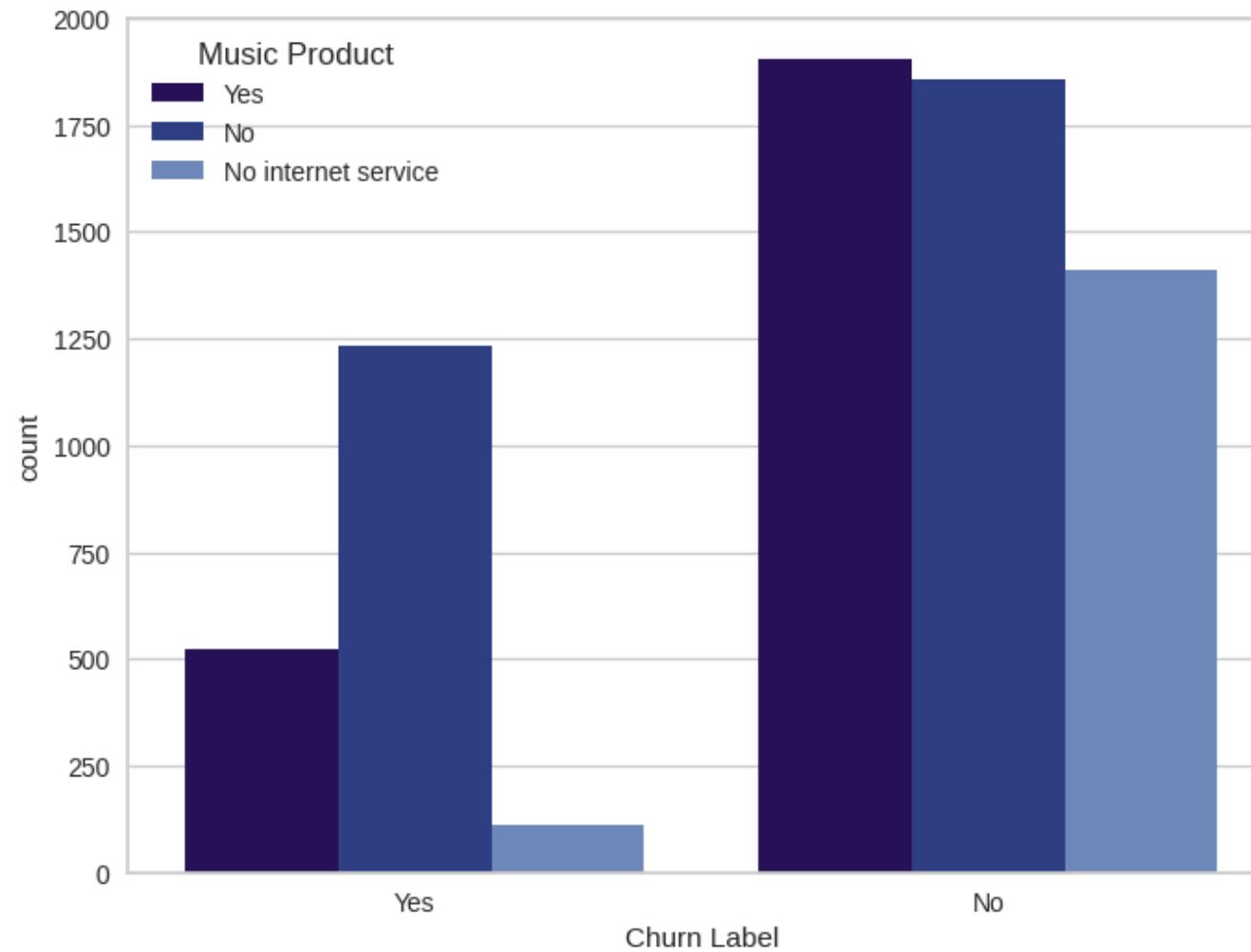


Pulsa stands out with the highest percentage of churned customer in payment methods (**45%**)

The exploration suggests a potential issue, prompting further investigation into the functionality and reliability of the service or system associated with transactions via Pulsa.



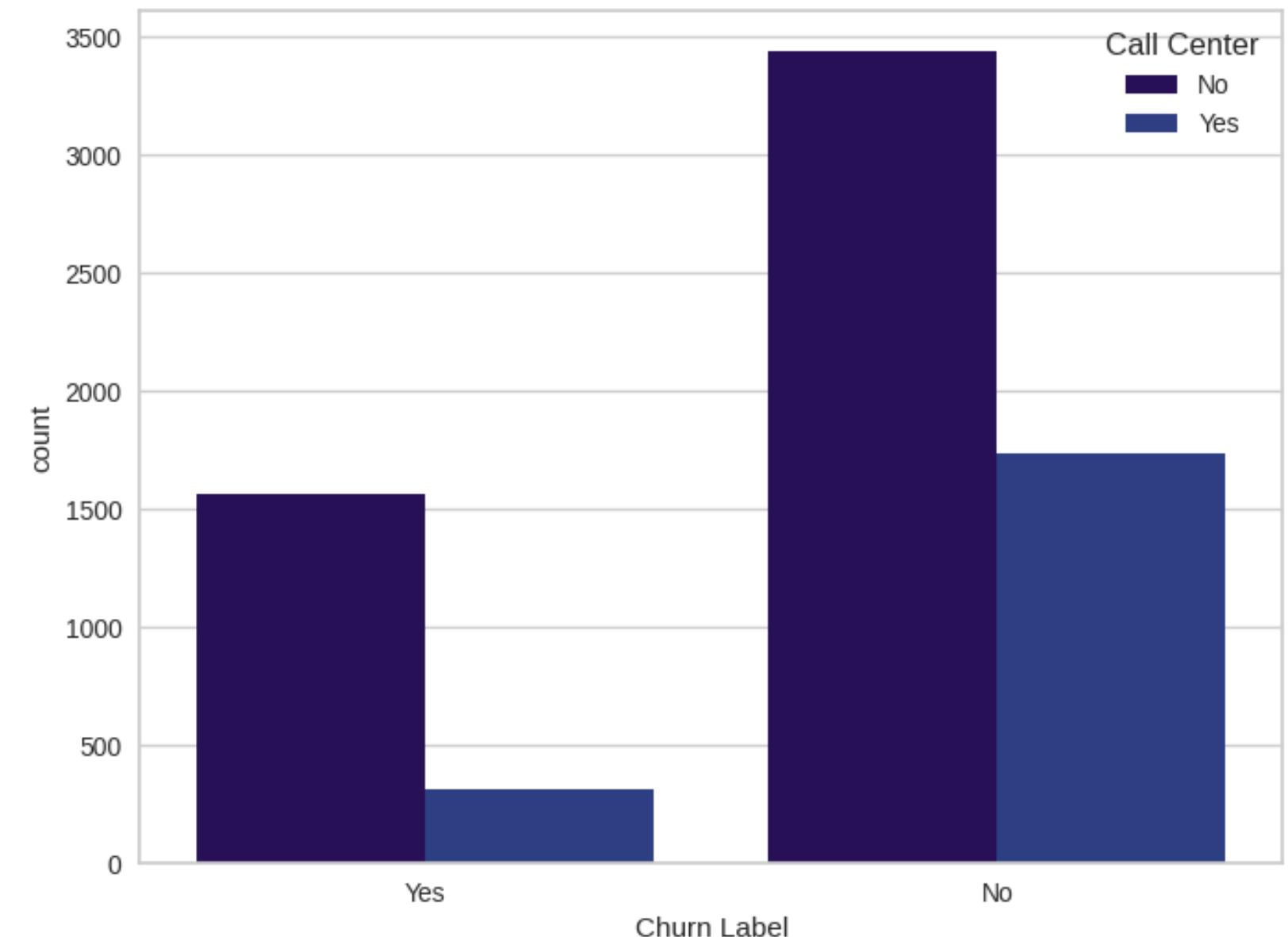
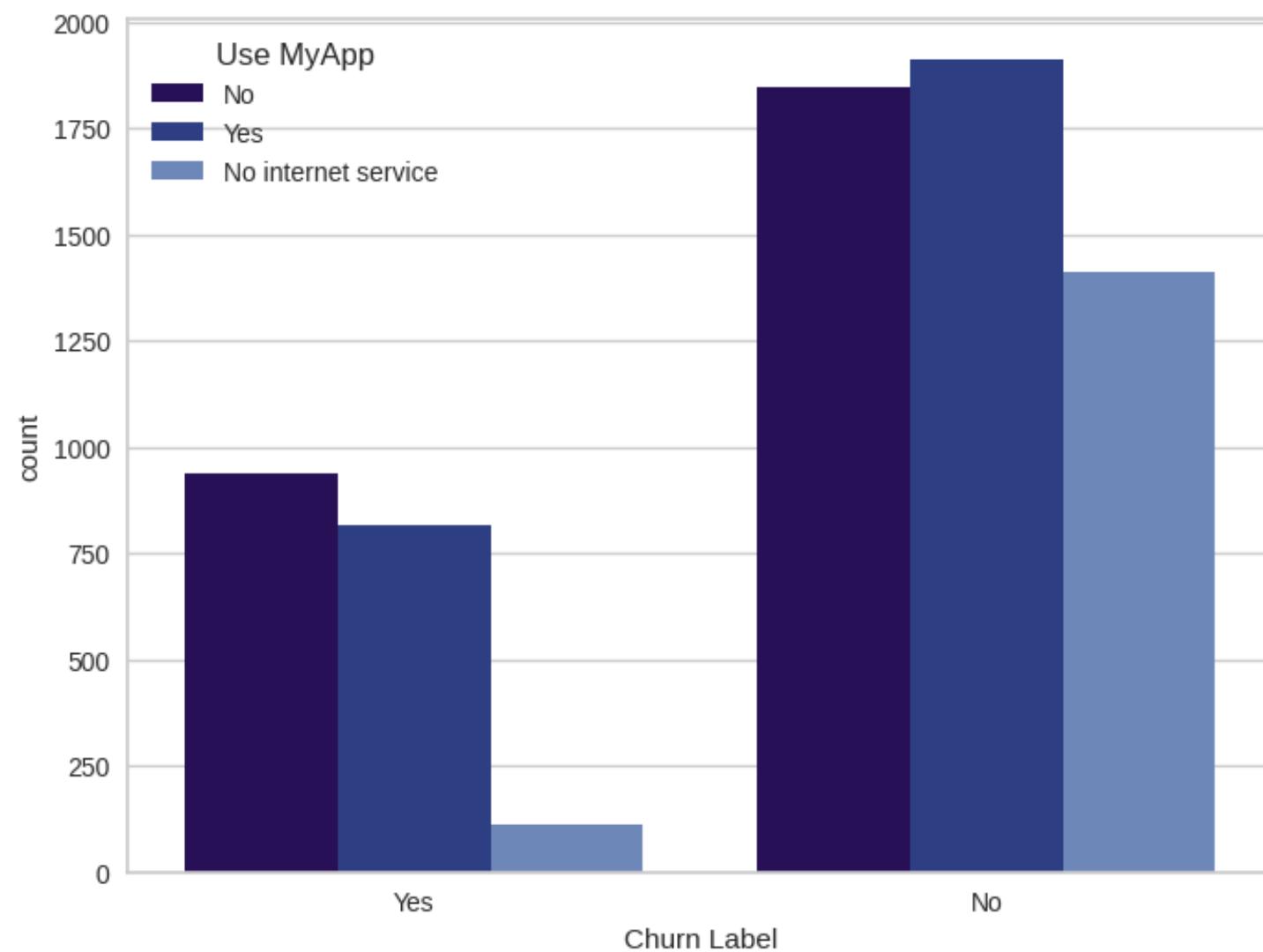
Customers using game and music products tend to stay, while non-users show a significant churn rate.



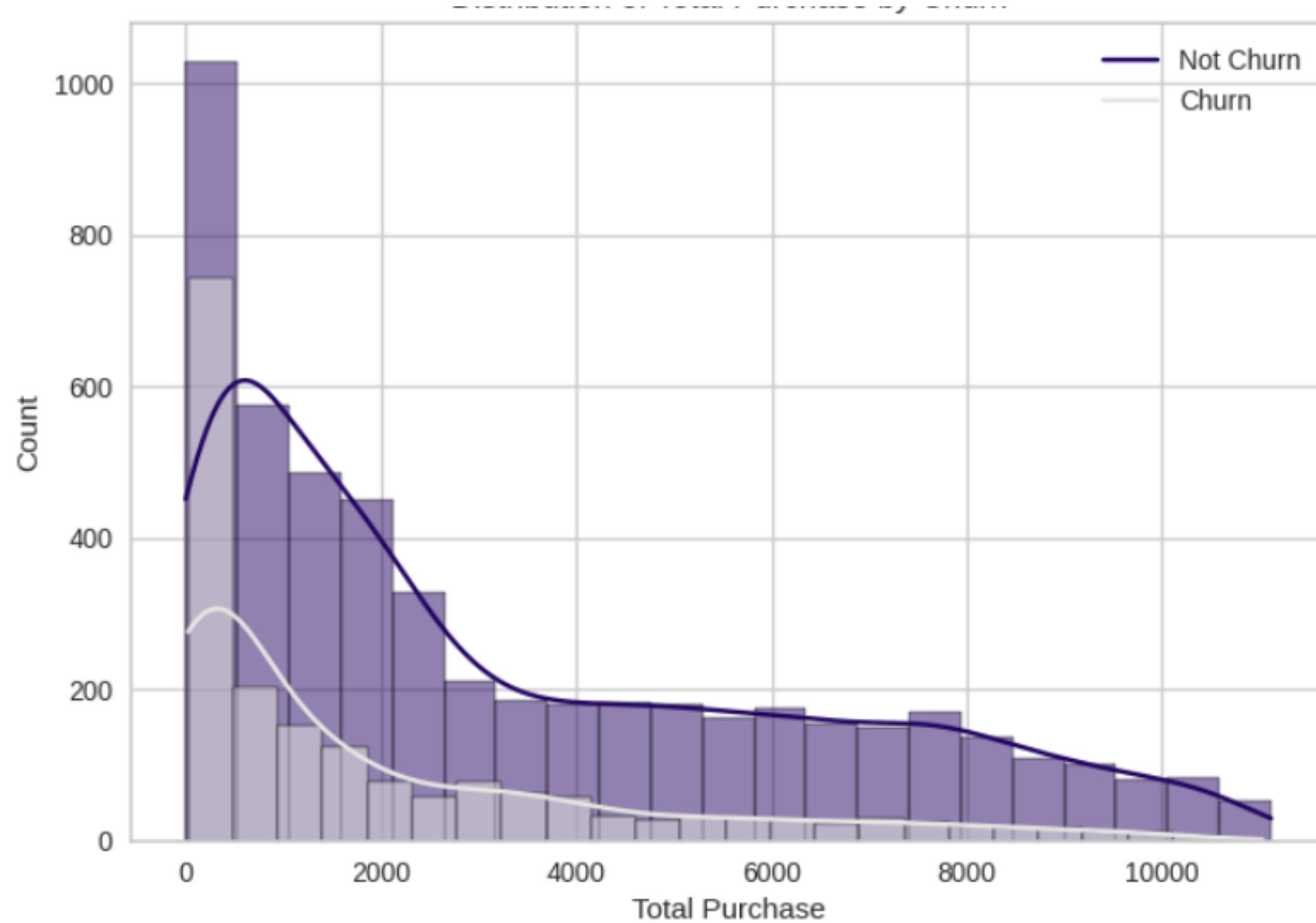
The issue may lie in less appealing promotion or advertising rather than the game and music product quality



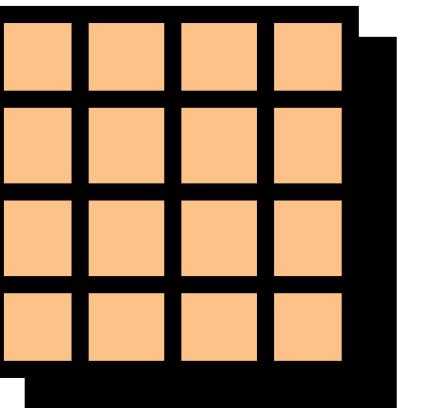
The **effective** call center service, evidenced by an **85% retention rate**, eliminates it as a contributing factor to customer churn



Customers using **MyApp** exhibit an **identical churn rate**, suggesting consistent application quality.



As total purchases **rise**,
the **churn rate decreases**,
suggesting **higher** total purchases
correlate with
greater customer stability



GET TO KNOW OUR CUSTOMER





Our Customers

Cluster 1, identified as 'Royal Customers,' exhibits a low churn rate characterized by high monthly purchases, long-term customer loyalty, frequent product usage, mid-high range devices, and a preference for debit and credit card payments. Personalization strategies for this segment could include tailored loyalty programs, exclusive product offers, and targeted communication highlighting the convenience of using debit or credit cards for transactions.

Cluster 2, representing 'Normal Customers,' shows a high churn rate with characteristics such as high monthly purchases, short-term customer relationships, limited product usage, predominant use of high-end devices, and a preference for pulsa payments. Personalizing strategies for this group may involve targeted promotions to increase product usage, incentives for longer-term commitments, and optimizing payment options compatible with pulsa.



Our Customers

Cluster 3, with a low churn rate, comprises mainly Jakarta residents. They demonstrate high monthly purchases, short-term relationships, limited product usage, prefer mid-range devices, and favor digital wallet payments, indicative of Gen Z traits. Personalization for this group may involve location-based promotions and optimizing the digital wallet experience

Cluster 4, 'Akang Teteh,' shows a high churn rate, all residing in Bandung, with high monthly purchases, short-term relationships, limited product usage, mid-high devices, and a preference for pulsa payments. Personalization can focus on location-based promotions and enhancing the pulsa payment experience

Cluster 5, consists of customers without internet service, using low-end devices, and with monthly purchases

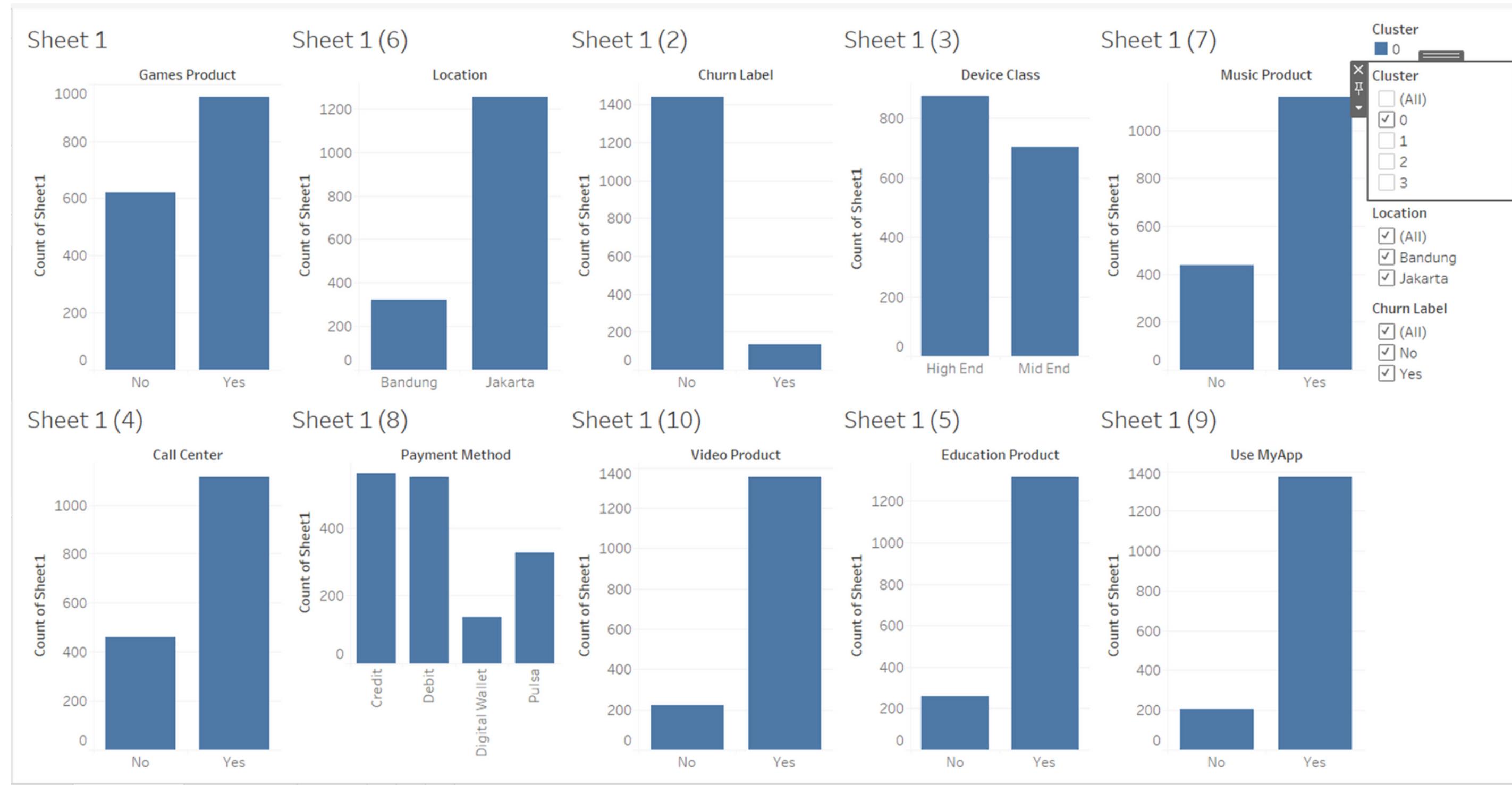


Detailed Information

- Clusters 1 and 3 have low churn rate while clusters 2 and 4 have high churn rate
- Cluster 1 is favourable for the company since they have high monthly purchase, high tenure months and low churn rate, customers in this cluster can be identified in their usage in music product, educational product, video product, MyApp and call center. The majority uses Credit and Debit payment method unlike other clusters. Their usage in educational product can indicate that these customers are studying in school or university and can mean that their ages are the lowest among all the clusters.
- Customers in Cluster 3 tend to not use video products or MyApp. Their device class are in the mid end and they have the highest percentage of digital wallet payment method compared to other clusters. The monthly purchase is relatively low especially compared to Cluster 1.
- The difference in Cluster 2 and 4 is in the location where the location in Cluster 2 is Jakarta while the location in Cluster 4 is Bandung. One other thing to note is the low count of Cluster 2 customers with monthly purchase below 85k IDR.
- -Cluster 5 contains customers who does not have internet service. They use low end device class and have a monthly purchase below 35k IDR.

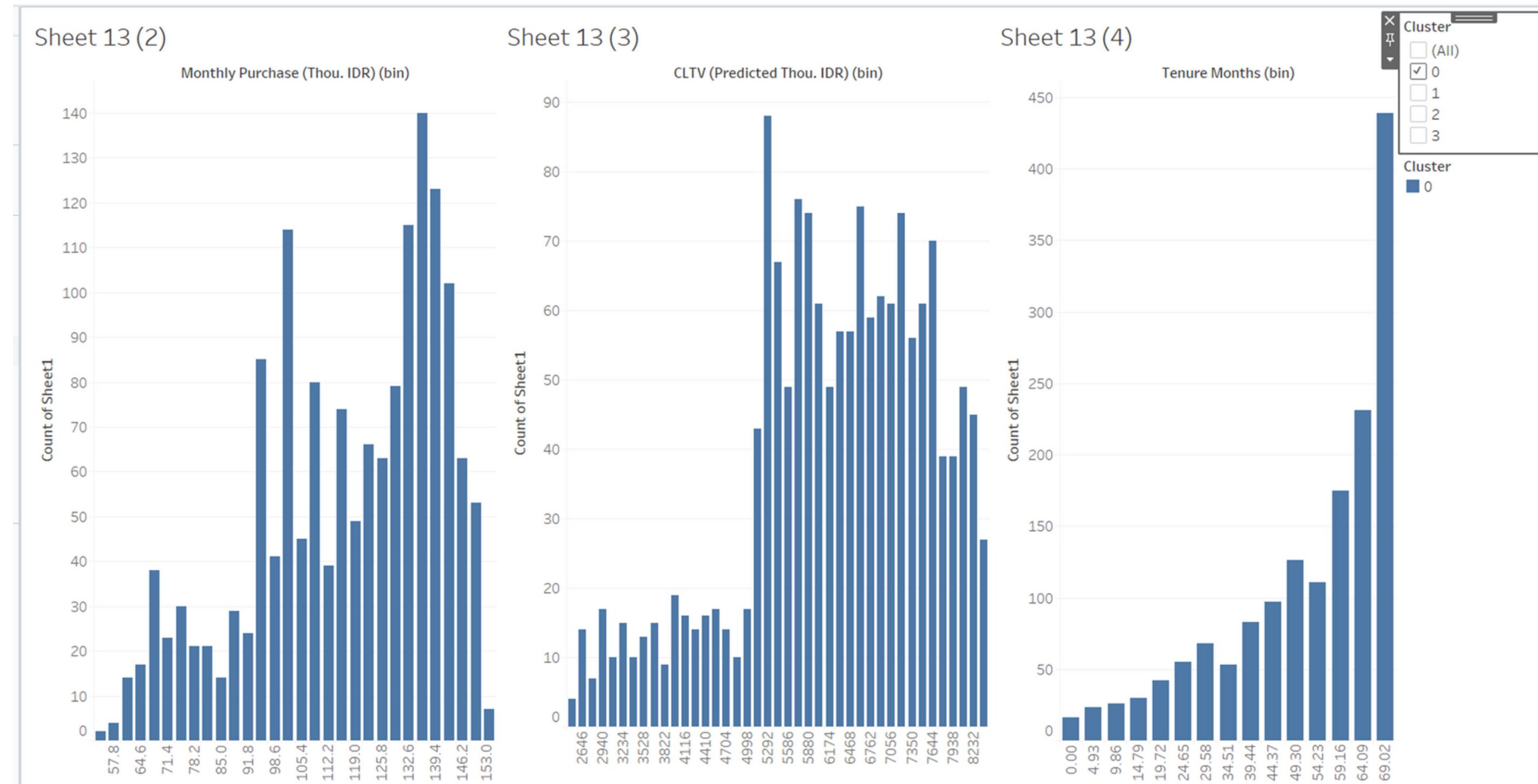


CLUSTER 1



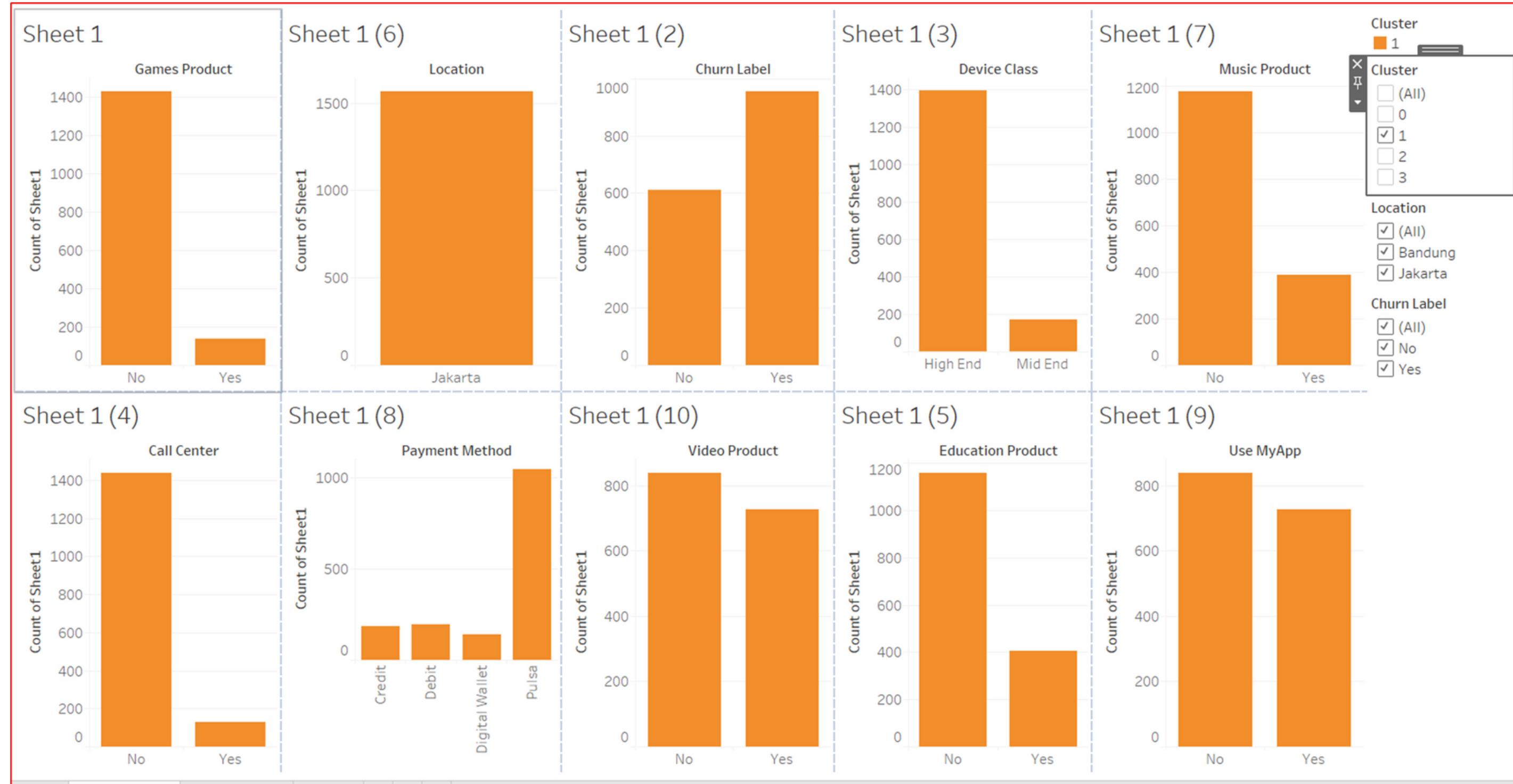


CLUSTER 1





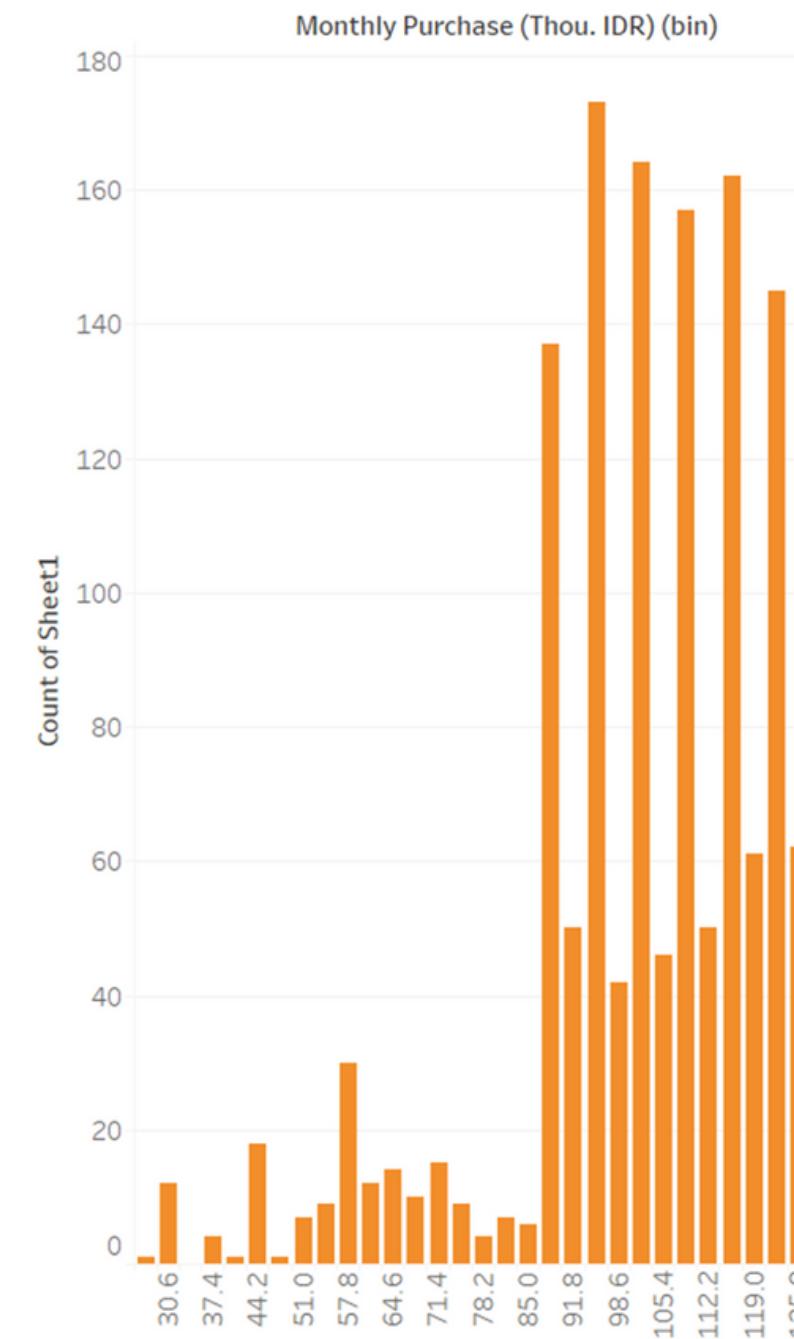
CLUSTER 2



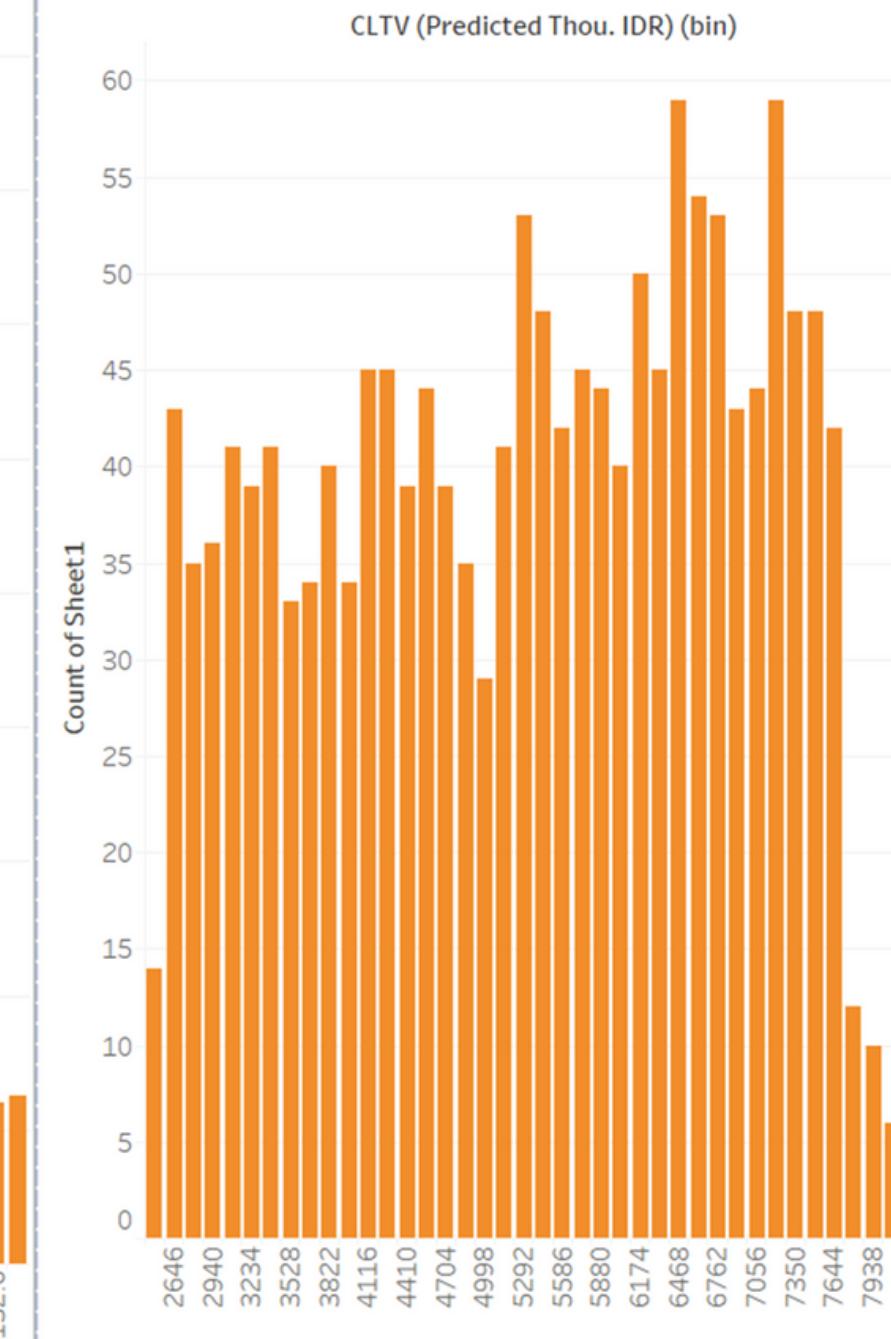


CLUSTER 2

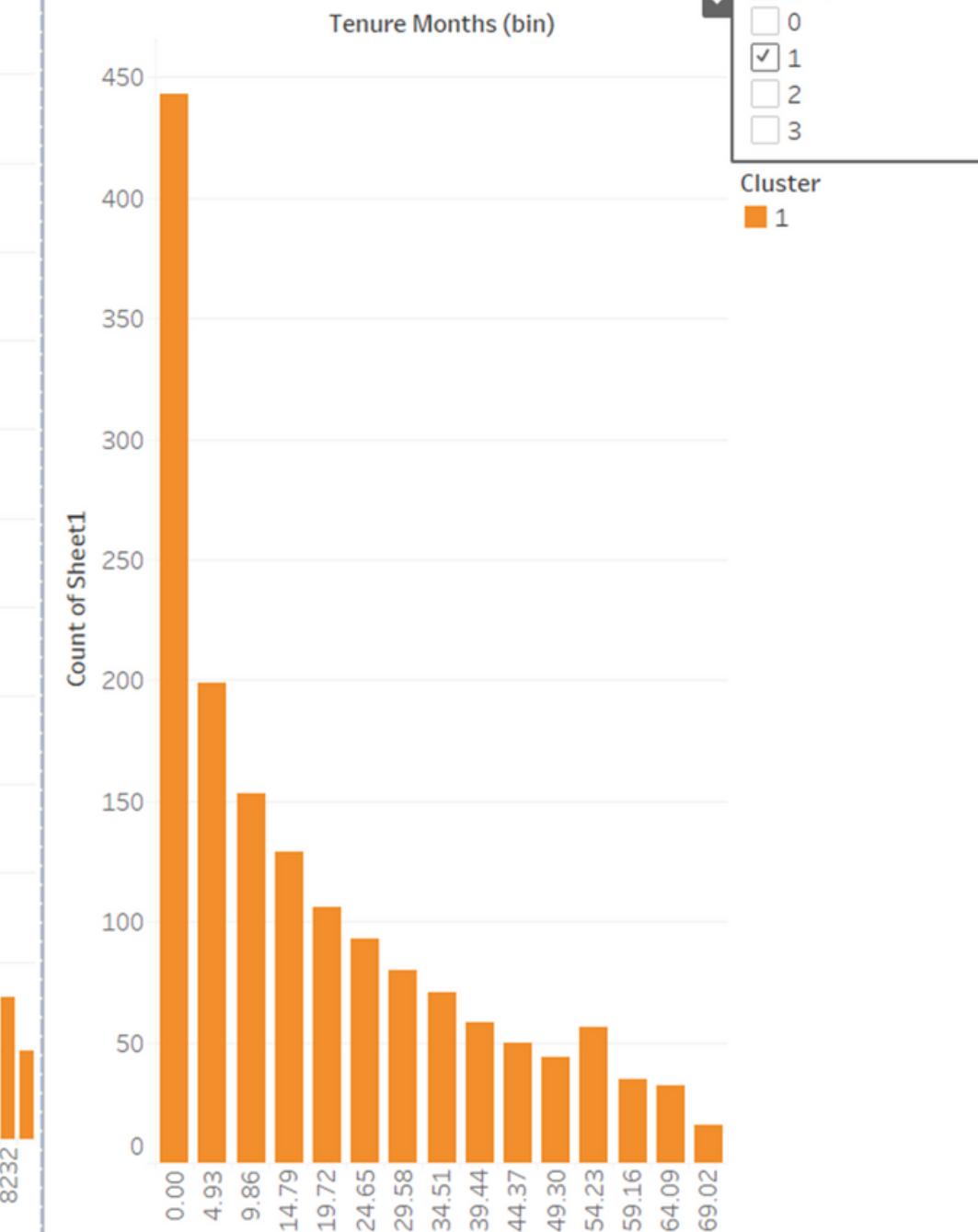
Sheet 13 (2)



Sheet 13 (3)

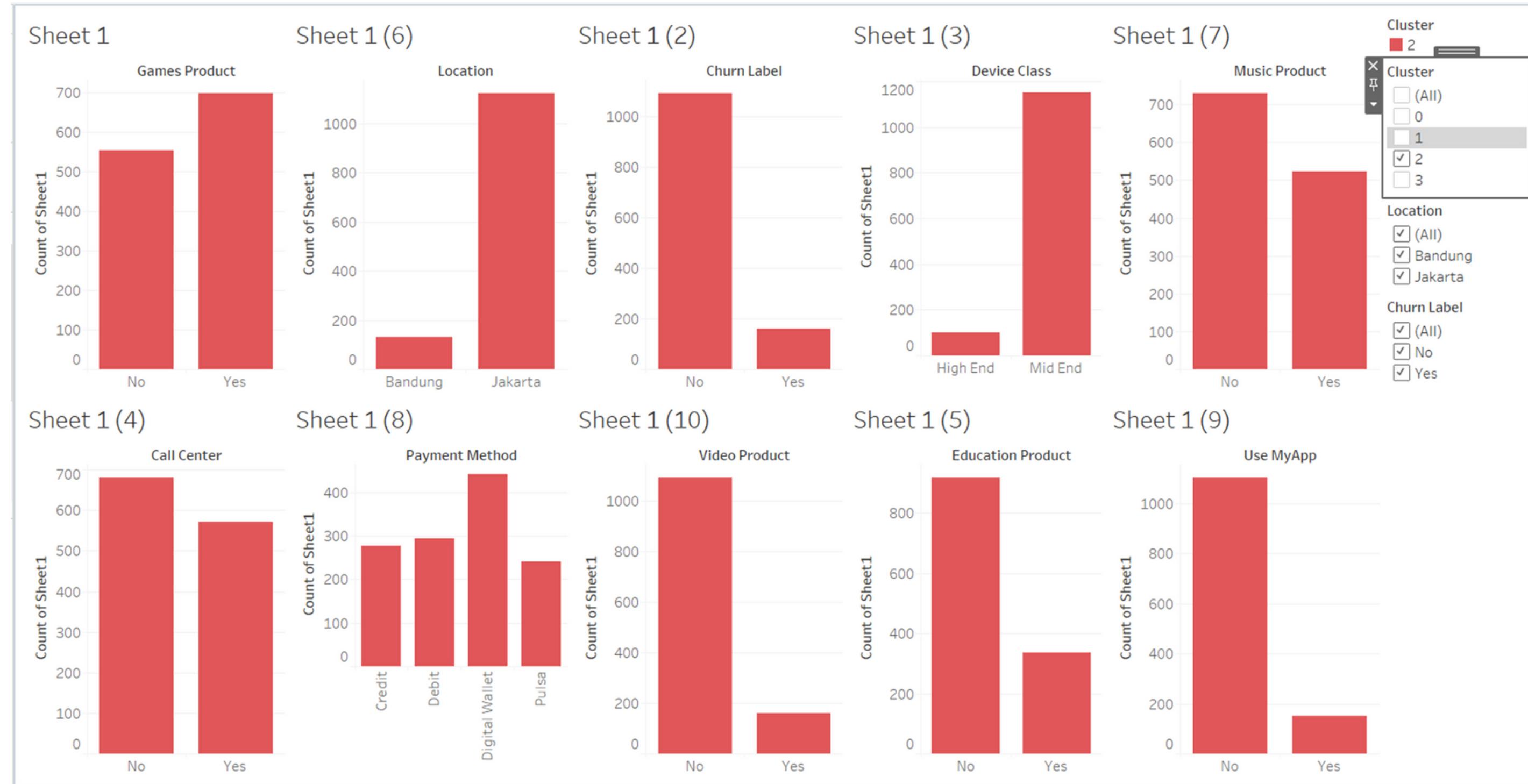


Sheet 13 (4)



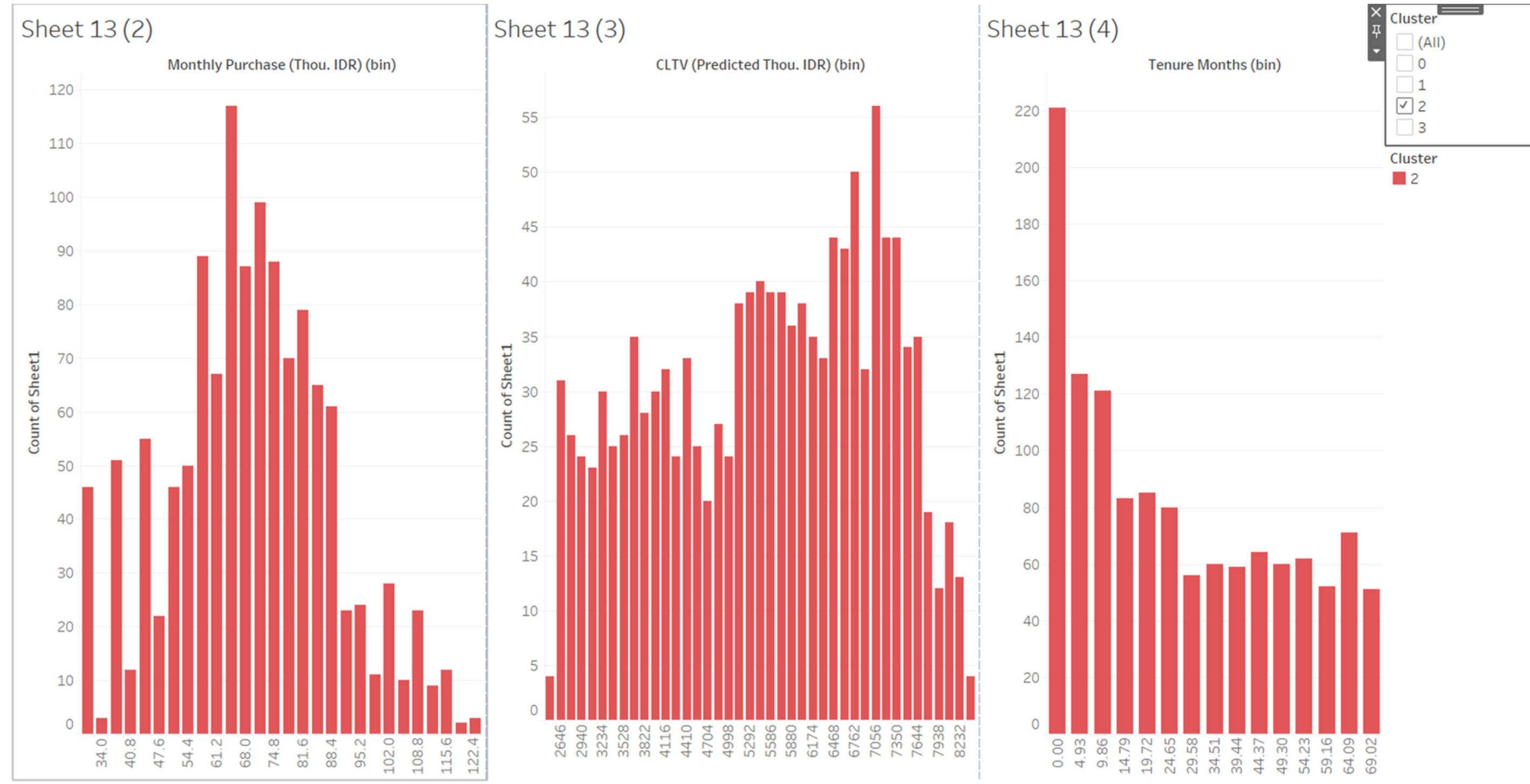


CLUSTER 3



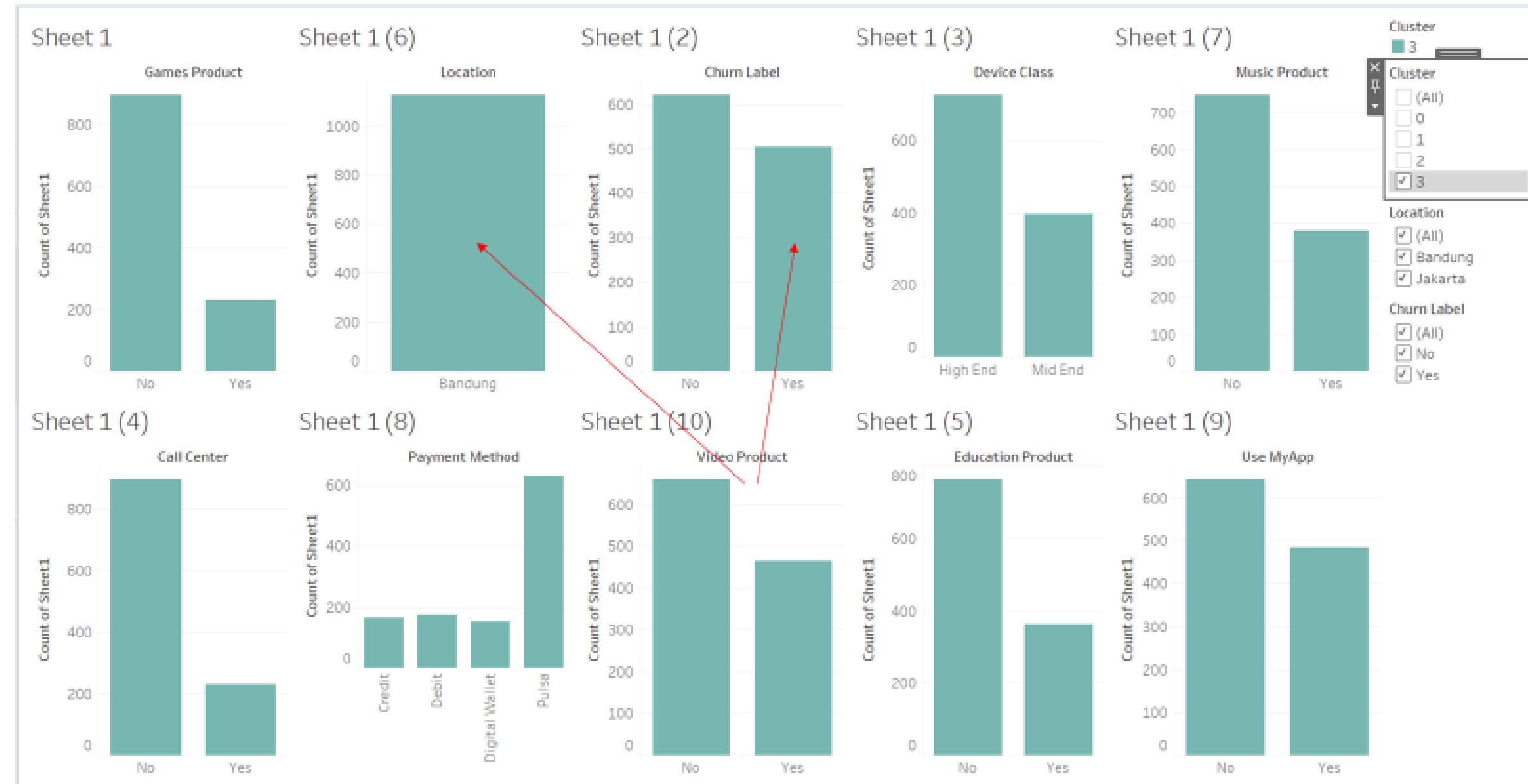


CLUSTER 3



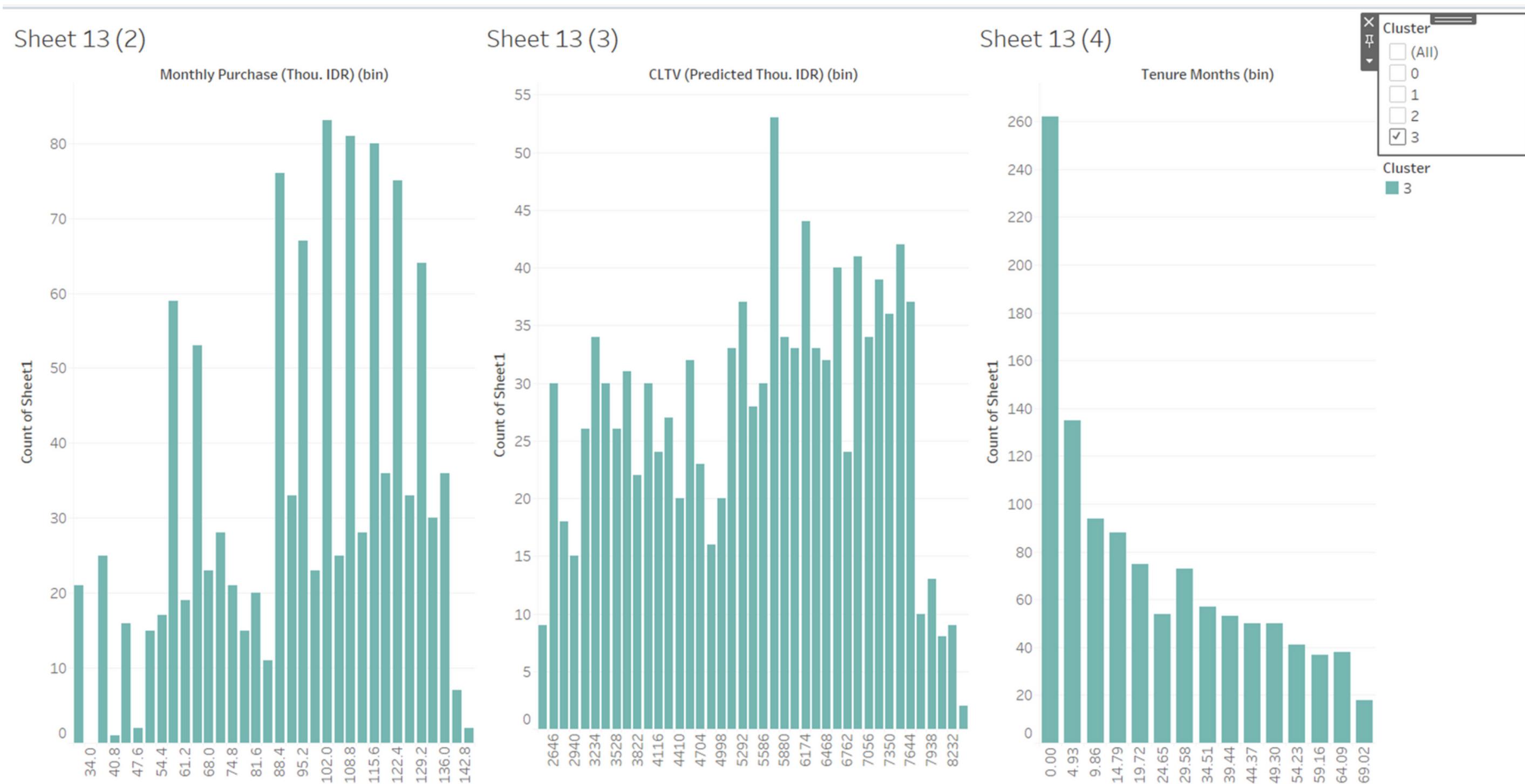


CLUSTER 4



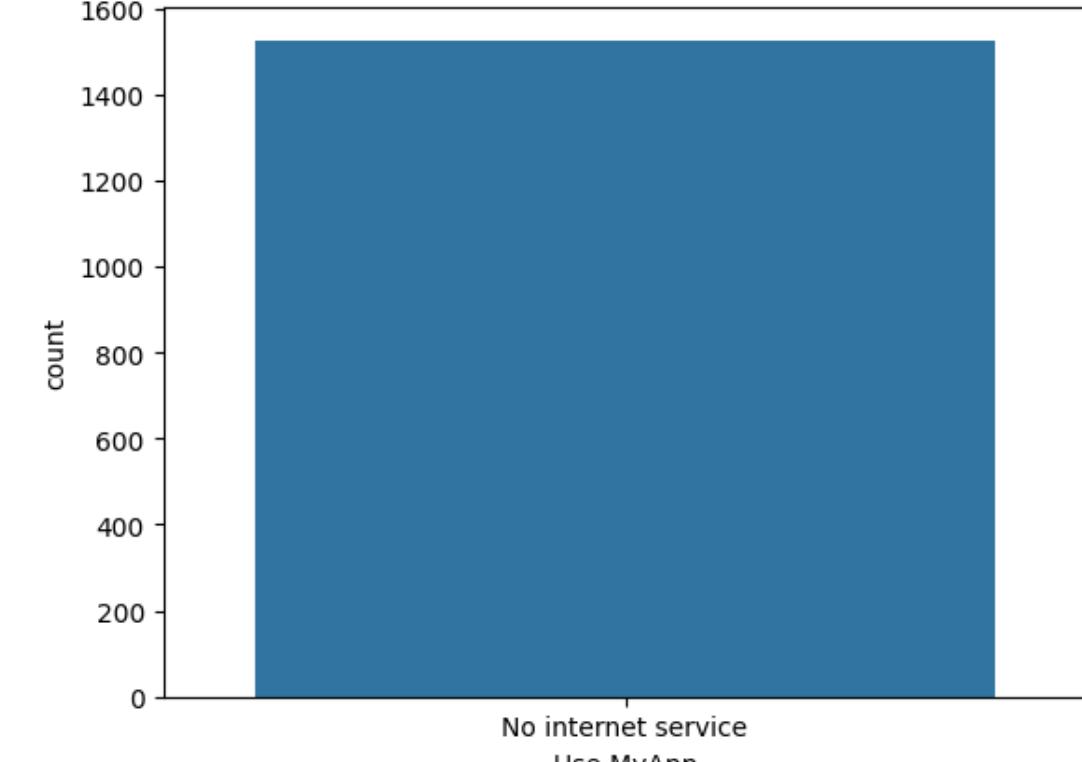
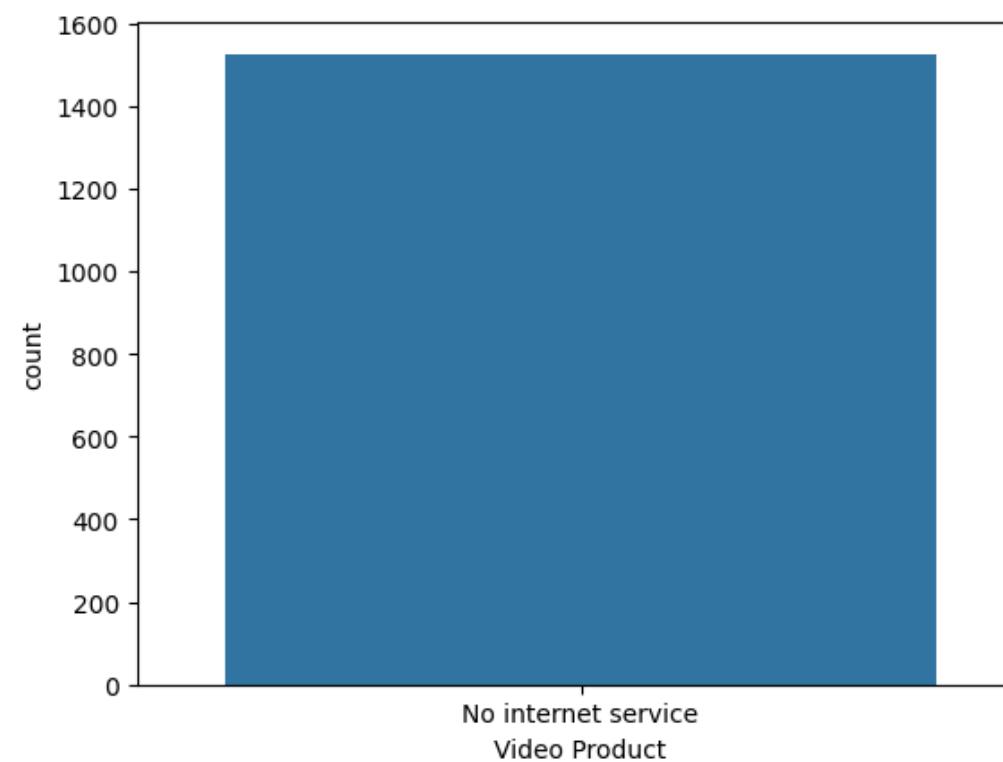
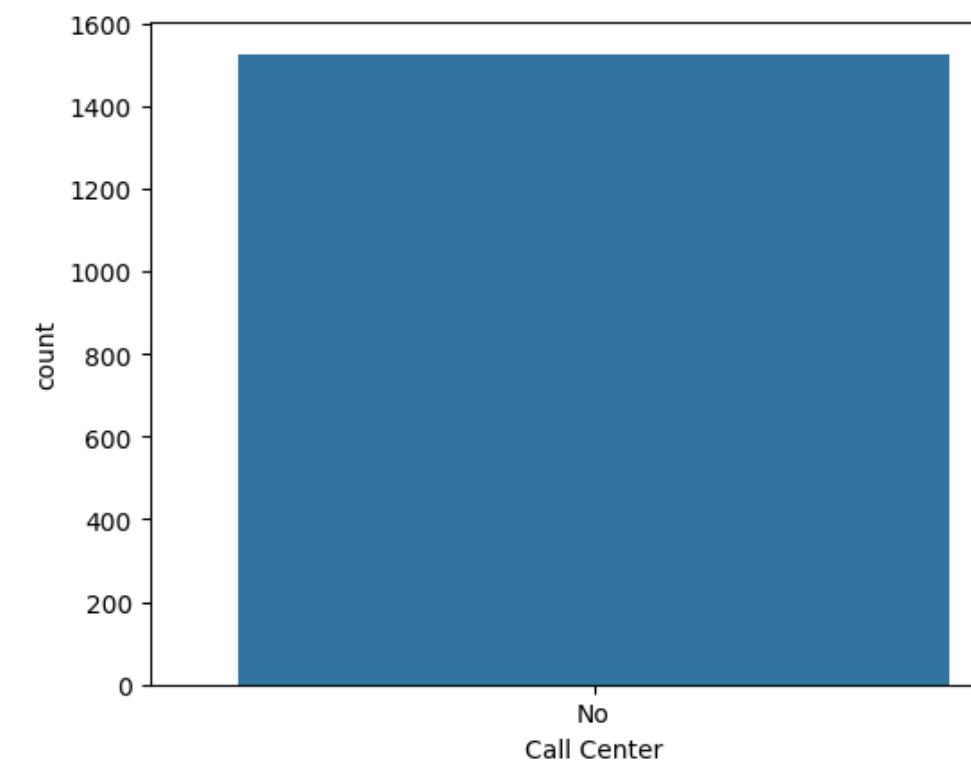
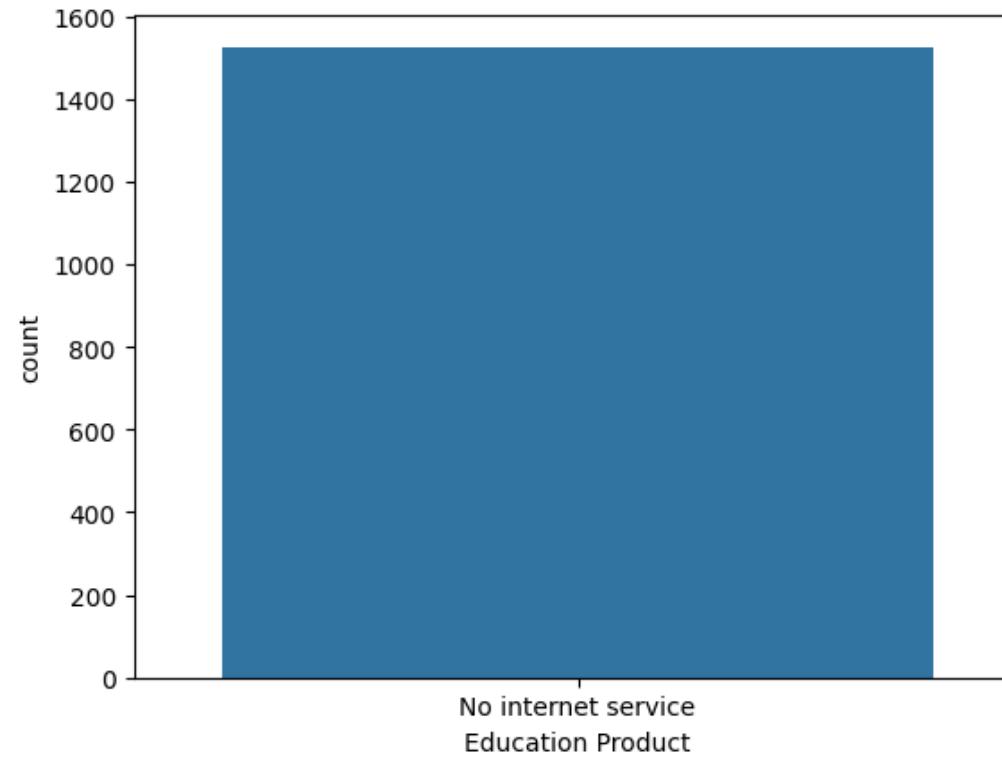
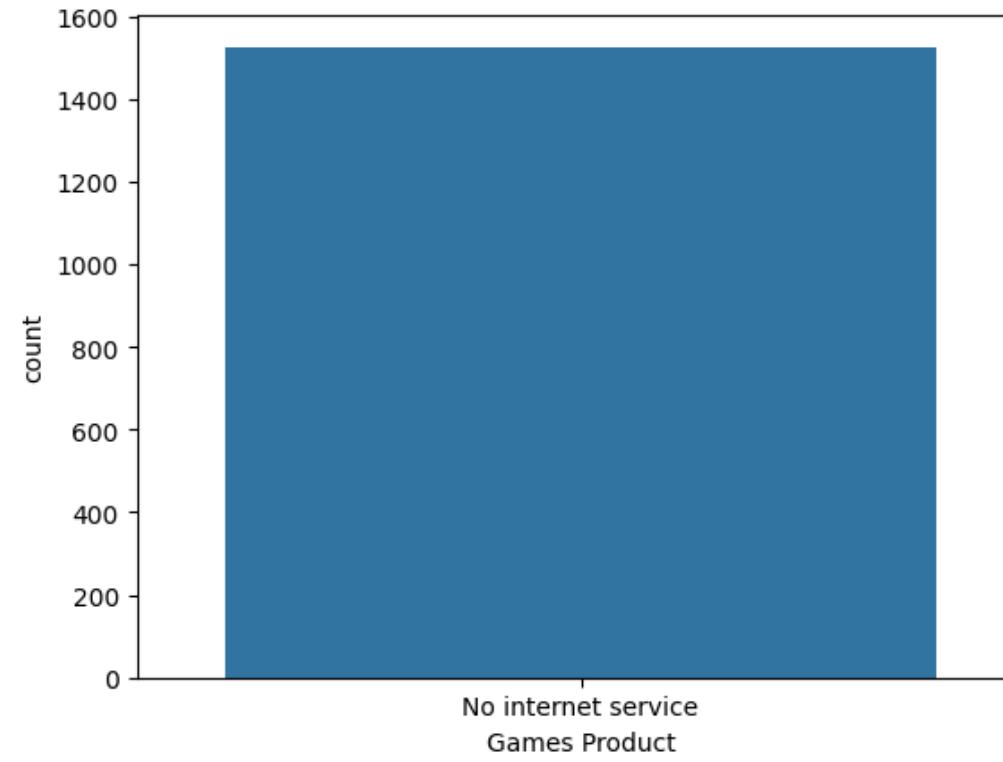
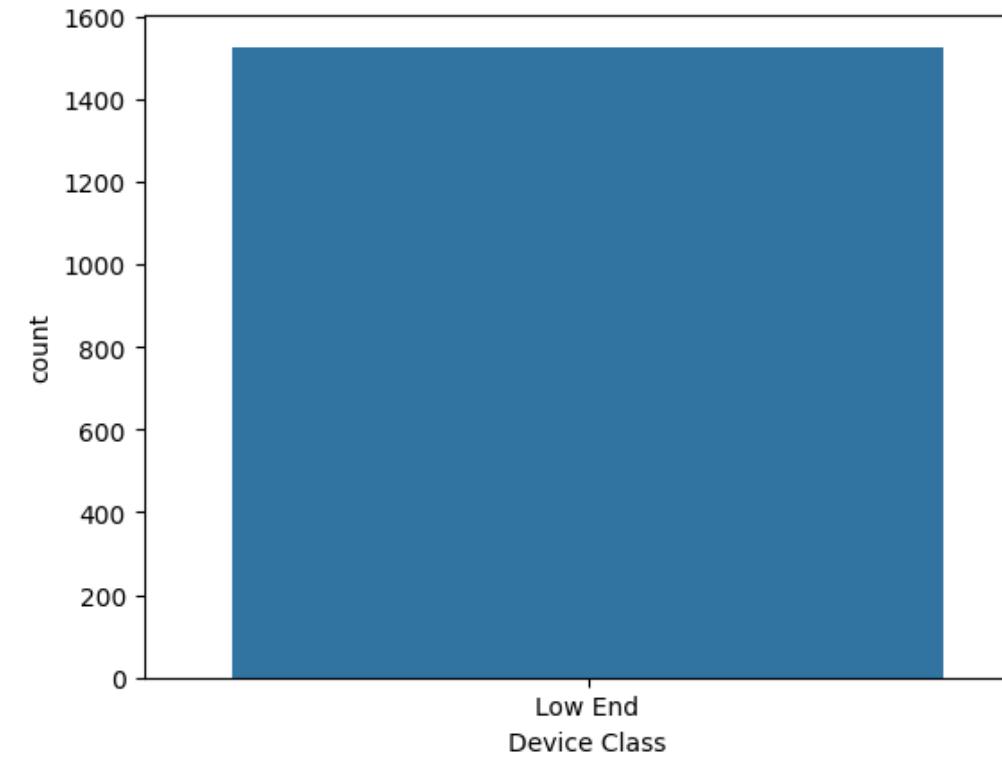


CLUSTER 4



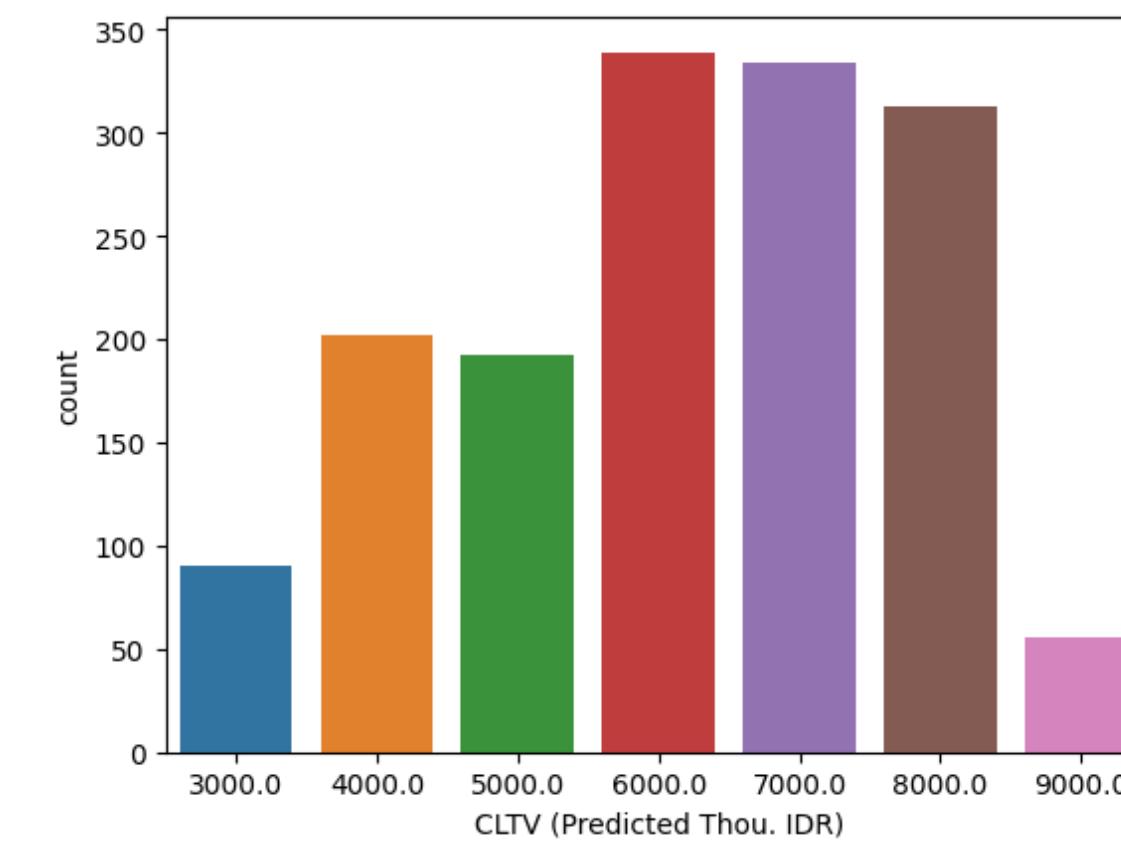
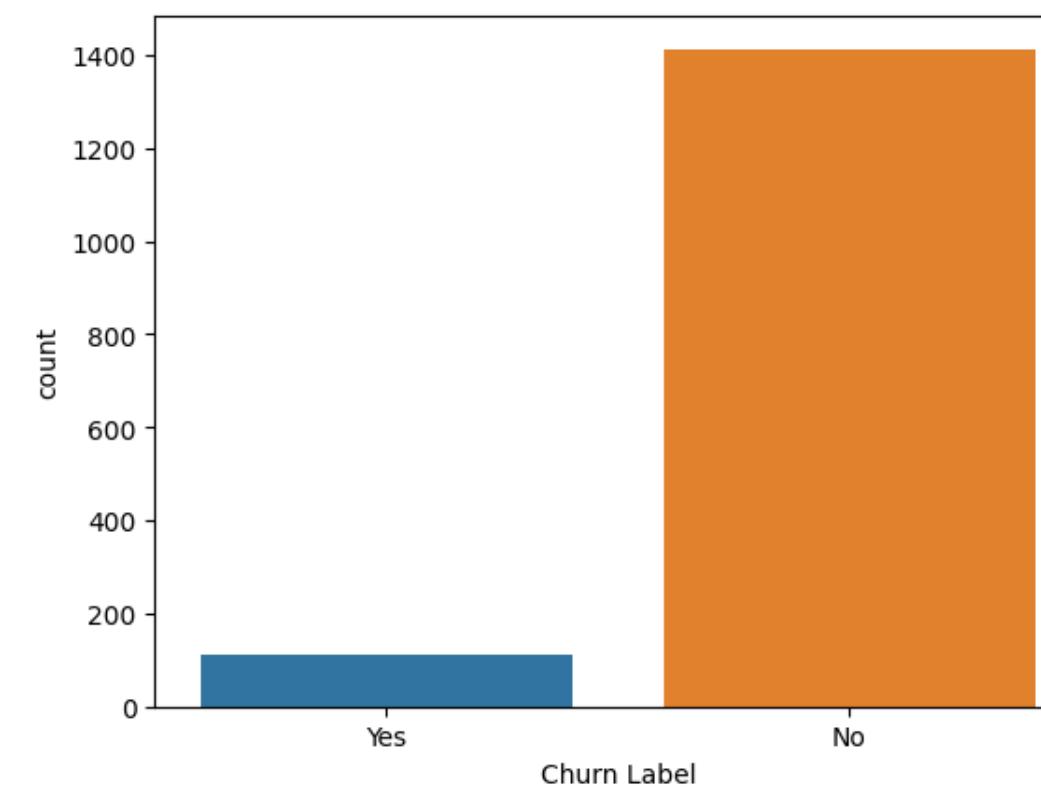
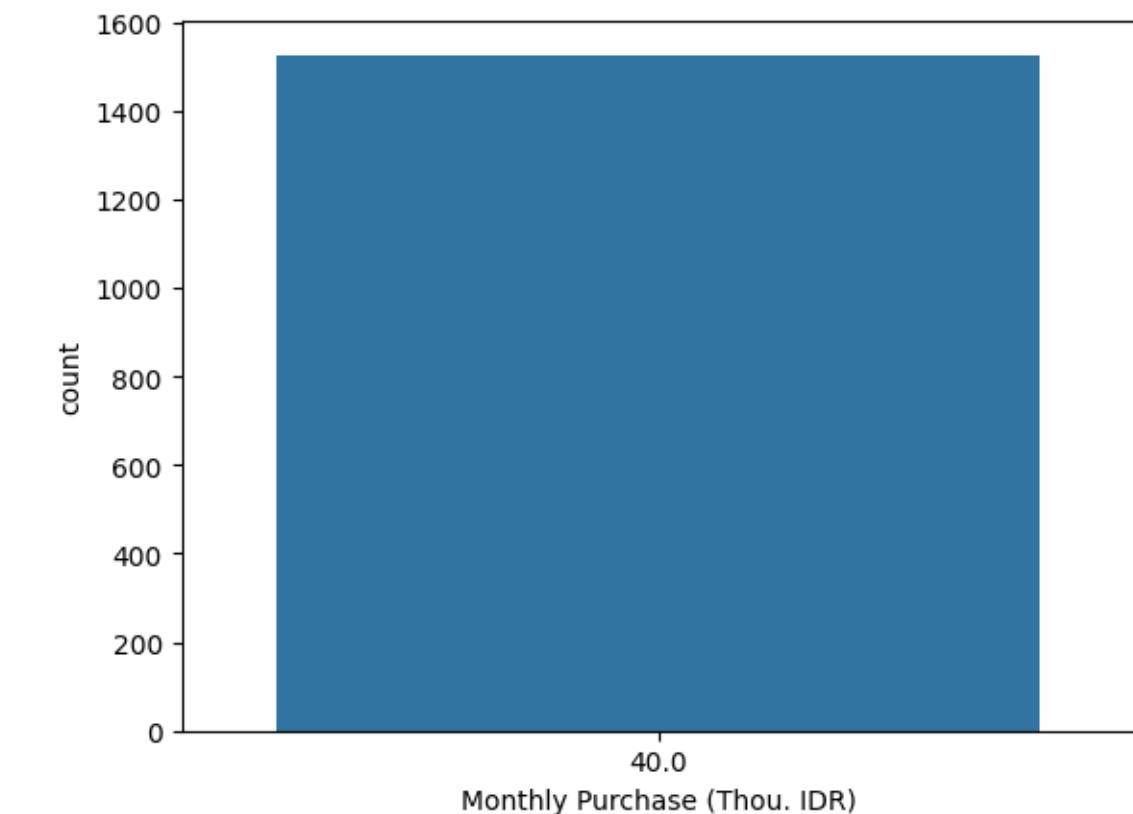
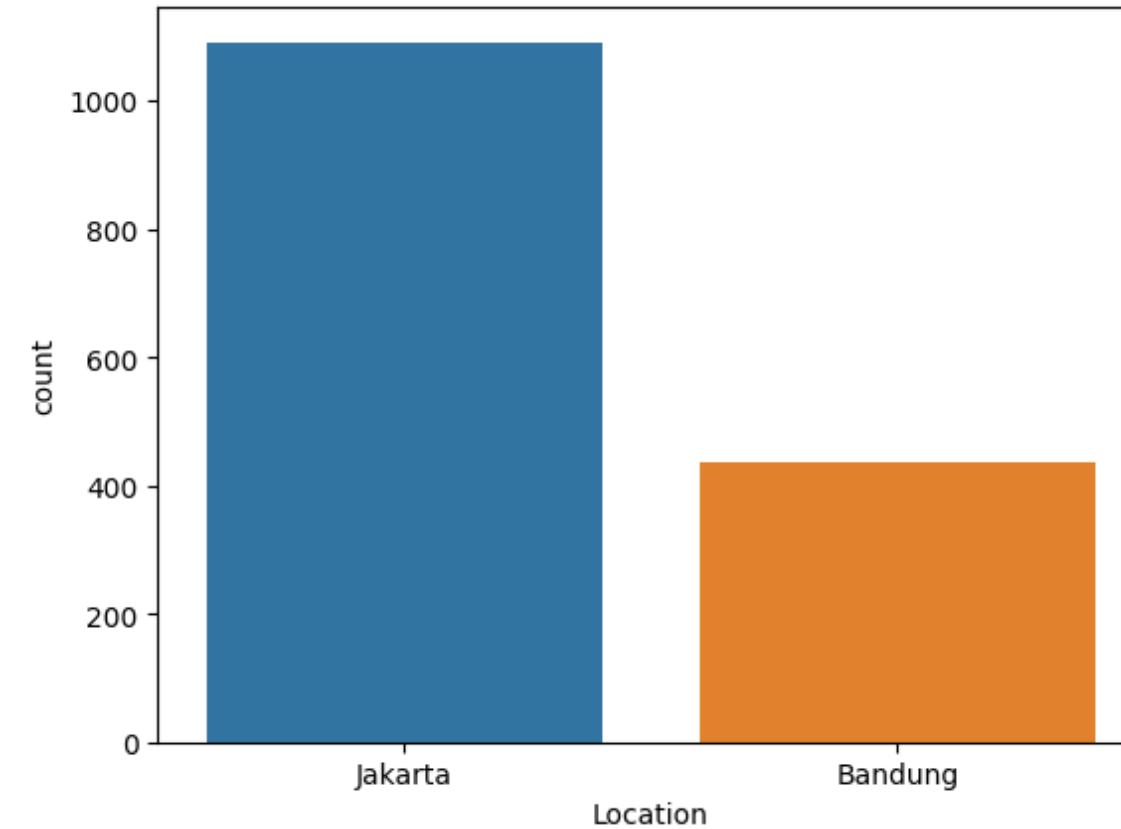
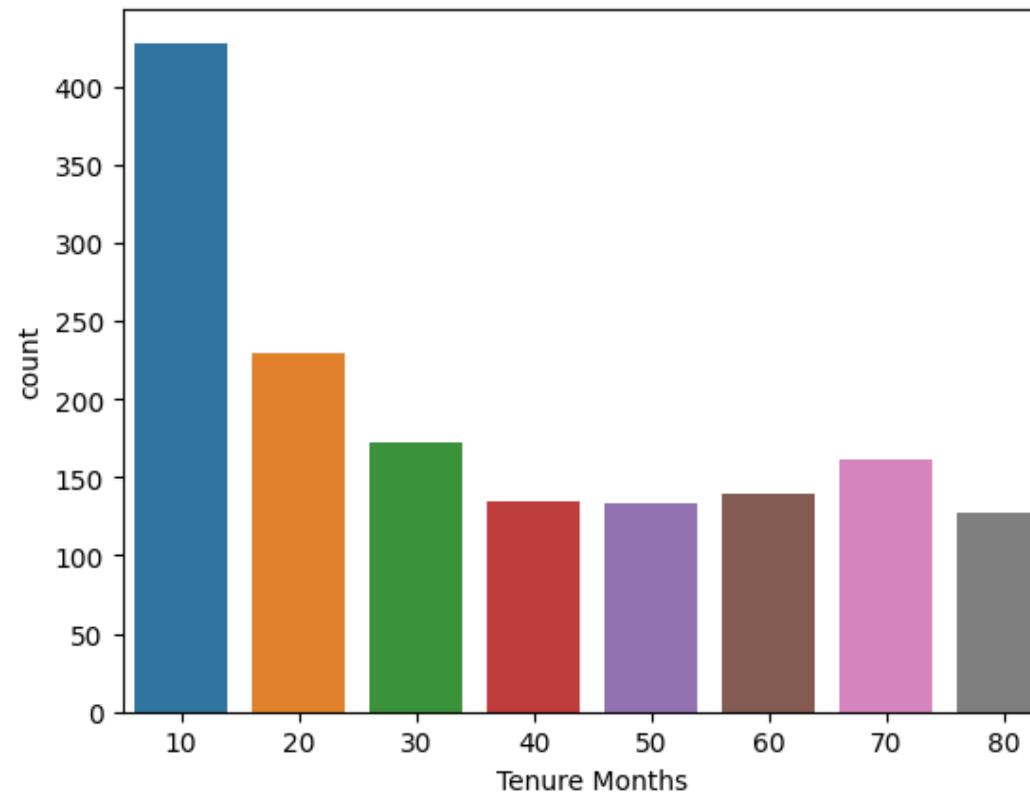


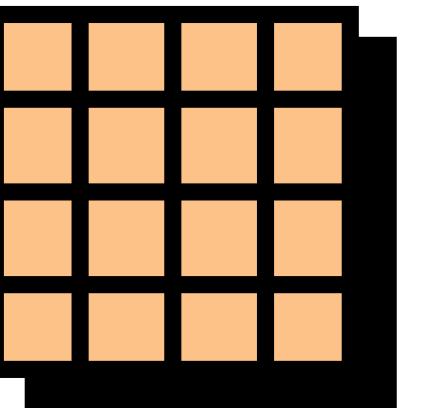
CLUSTER 5 (No internet service customers)





CLUSTER 5 (No internet service customers)





DATA PREPARATION





DATA PREPARATION

NORMALIZATION

MinMaxScaler

DROP DATA

Drop Customer ID, Latitude and Longitude columns

Drop customers with no internet service for a better performing model since those customers have low churn rate anyway

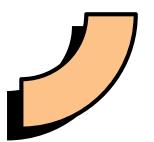
ENCODING

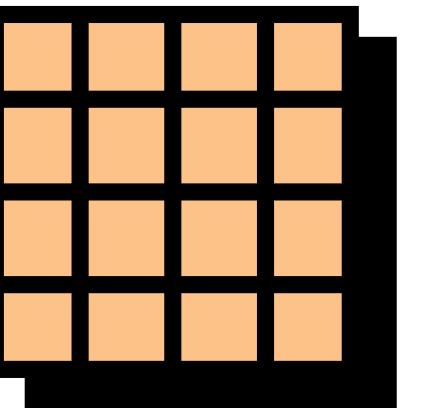
One Hot Encoding (pd.get_dummies)

SPLITTING AND SAMPLING

Test size = 30%

Train data is oversampled using SMOTE





MODELLING





MODELING

Multiple classifiers are used and compared to find the best model:

RANDOM FOREST

SVM

LOGISTIC REGRESSION

K-NEAREST NEIGHBORS

NAIVE BAYES

DECISION TREE

GRADIENT BOOSTING

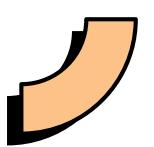
XGBOOST

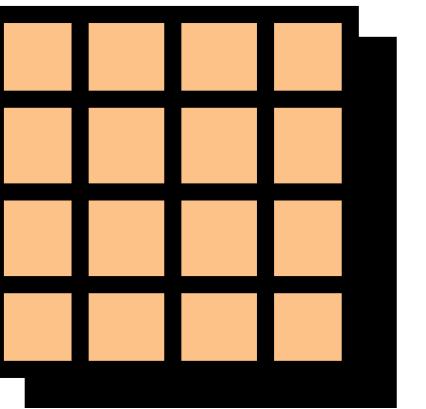
MLPC

LGBM

SGD

ADABOOST





EVALUATION





EVALUATION

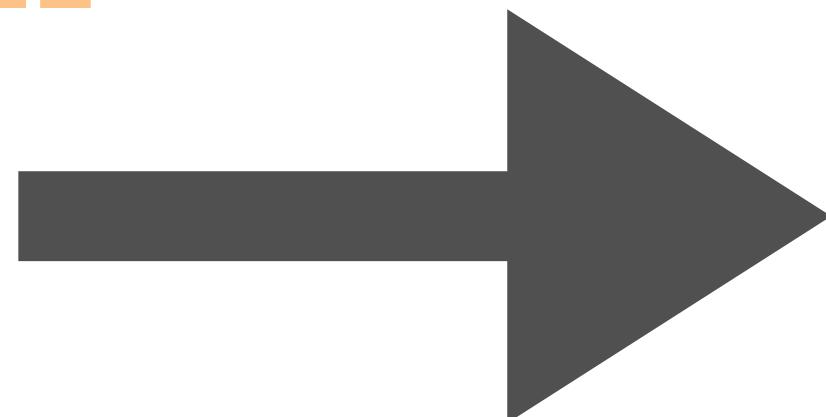
METRICS

Recall to minimize false negatives (churned customers predicted as non-churn)

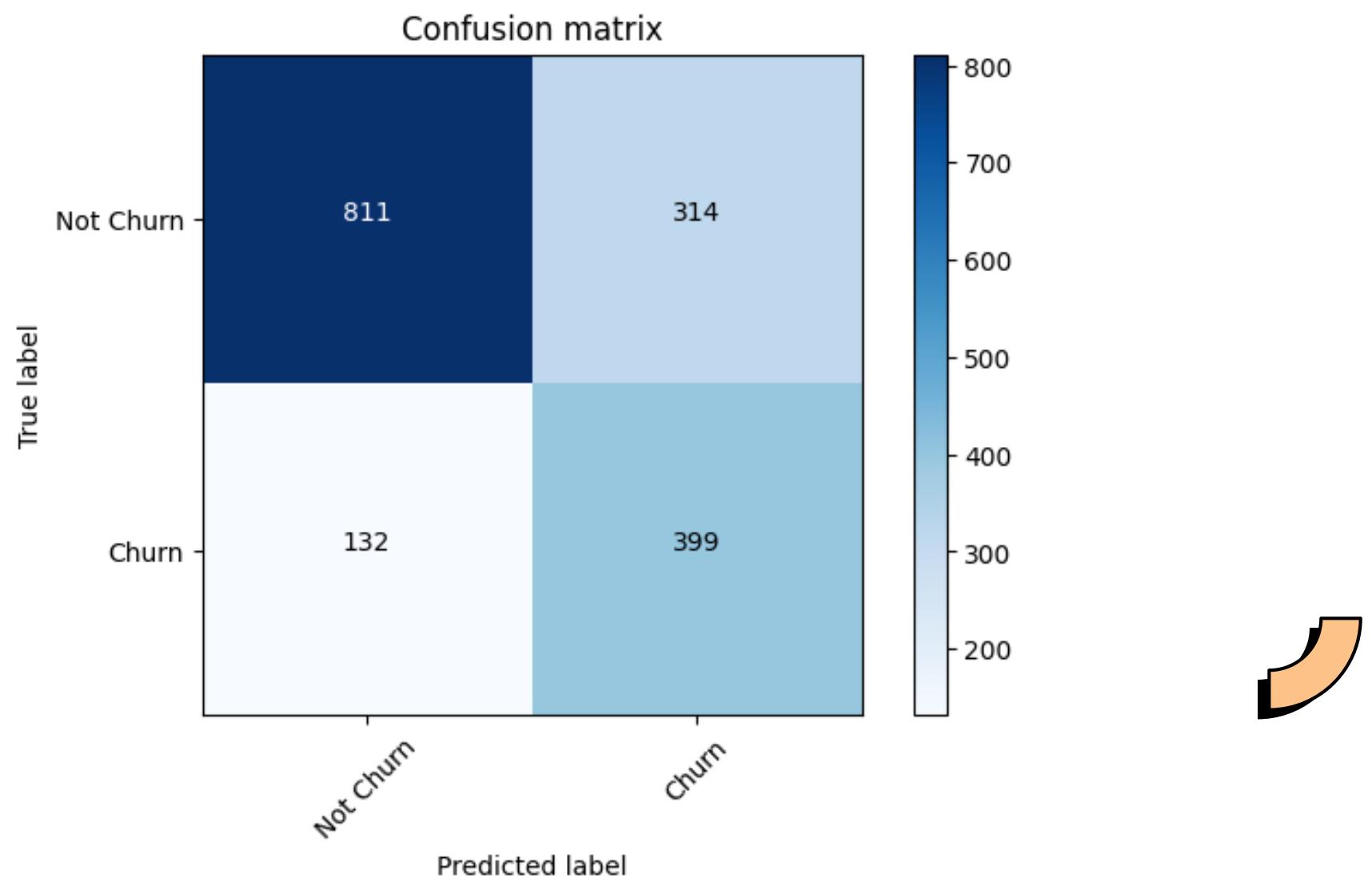
F1-score which gives balance between precision and recall

BEST MODEL

Logistic Regression:



```
Random Forest - F1 Score: 0.8043070811040411, Recall: 0.8262858662526595
SVM - F1 Score: 0.7646030185062405, Recall: 0.8084353976194583
Logistic Regression - F1 Score: 0.7467519043456421, Recall: 0.7647957276752344
K-Nearest Neighbors - F1 Score: 0.7724641349463692, Recall: 0.8387901788281296
Naive Bayes - F1 Score: 0.7397342913518059, Recall: 0.7575922891150595
Decision Tree - F1 Score: 0.7277437757133067, Recall: 0.7401730780288654
Gradient Boosting - F1 Score: 0.7687249612265802, Recall: 0.7951605715600023
XGBoost - F1 Score: 0.7791786302198008, Recall: 0.7963192168362947
MLP Classifier - F1 Score: 0.7677773426474619, Recall: 0.7947774711057444
LGBM Classifier - F1 Score: 0.781348211788362, Recall: 0.8042874187798287
SGD Classifier - F1 Score: 0.7540046078896439, Recall: 0.8000654073946294
AdaBoost Classifier - F1 Score: 0.7622471237932074, Recall: 0.7890891840607211
```



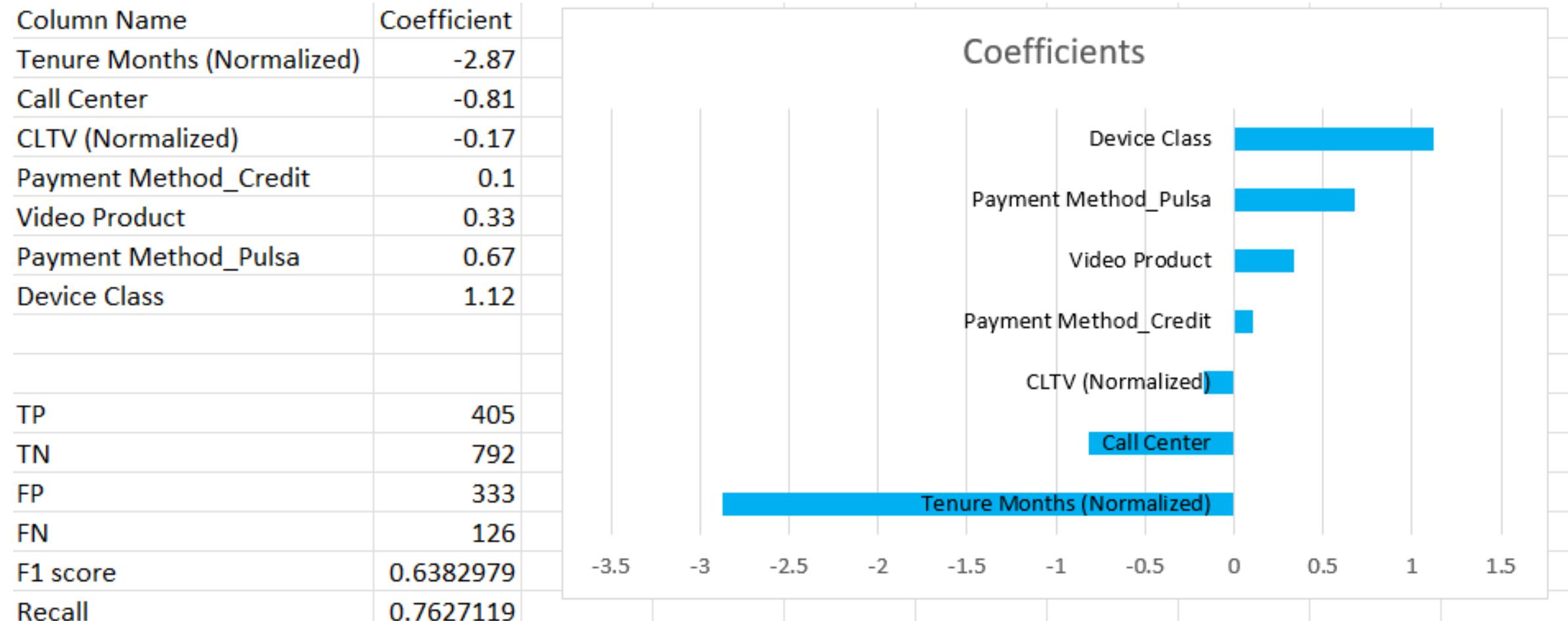


EVALUATION

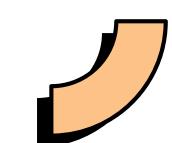
INTERPRETATION

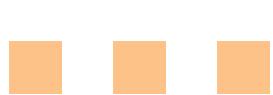
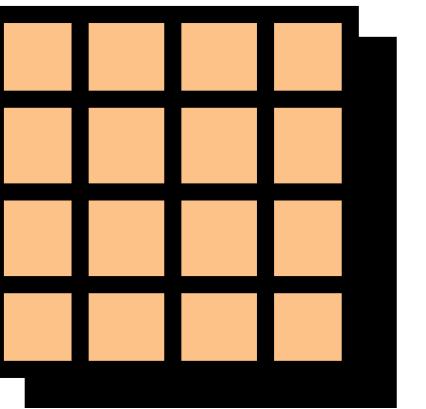
- SequentialFeatureSelection is used to select only features that increases the model's f1-score significantly.
- Logistic regression coefficients can be obtained and used to find characteristics of churned customers. Tenure Months, Device Class and Call Center are the features with the highest absolute coefficients.
- The company has done a great job of retaining loyal customers, but not new customers. Their call center service is good at reducing customer churn.

Column Name	Coefficient
Tenure Months (Normalized)	-2.87
Call Center	-0.81
CLTV (Normalized)	-0.17
Payment Method_Credit	0.1
Video Product	0.33
Payment Method_Pulsa	0.67
Device Class	1.12
TP	405
TN	792
FP	333
FN	126
F1 score	0.6382979
Recall	0.7627119



Logistic Regression with SequentialFeatureSelection,
scoring parameter set to 'f1'





DEPLOYMENT





DEPLOYMENT

DESCRIPTION

This deployment is hosted using Streamlit. It serves the purpose of predicting whether a customer will churn or not. The model utilized for this deployment is a logistic regression model.

INPUT VALUE

Tenure Months, Device Class, Games Product, Music Product, Education Product, Call Center, Video Product, Use MyApp, Payment Method, Monthly Purchase, CLTV

LINK DEPLOYMENT

<https://compdatascienceindo2023-diker.streamlit.app/>

OUTPUT VALUE

Churn Label





DEPLOYMENT

index

xgboost

logistic

Model XGBoost Mengenai Churn

Tenure Months

Lokasi Pengguna ?

0 - + Jakarta

Device Class

have Games Product ?

Low End - No

have Music Product ?

have Education Product ?

No - No

have Call Center ?

have Video Product ?

No - No

have Use MyApp ?

have Payment Method ?

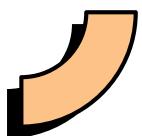
No - Pulsa

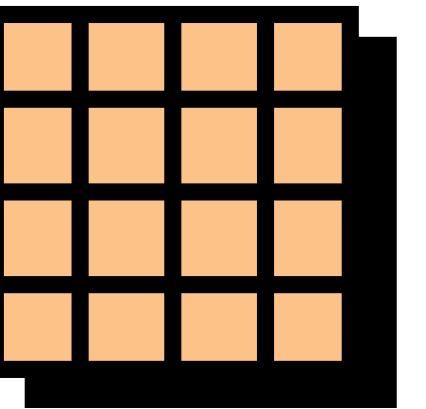
Monthly Purchase (Thou. IDR)

CLTV (Predicted Thou. IDR)

0 - + 0 - +

Bahwa Customer itu churn ? Churn





CONCLUSION





SOLUTIONS

CHURN PREDICTION

Use churn prediction model to retain customers. If a certain customer is predicted to churn, send a notification or text message and give promotions or discounts.

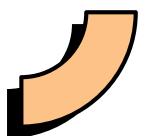
CALL CENTER

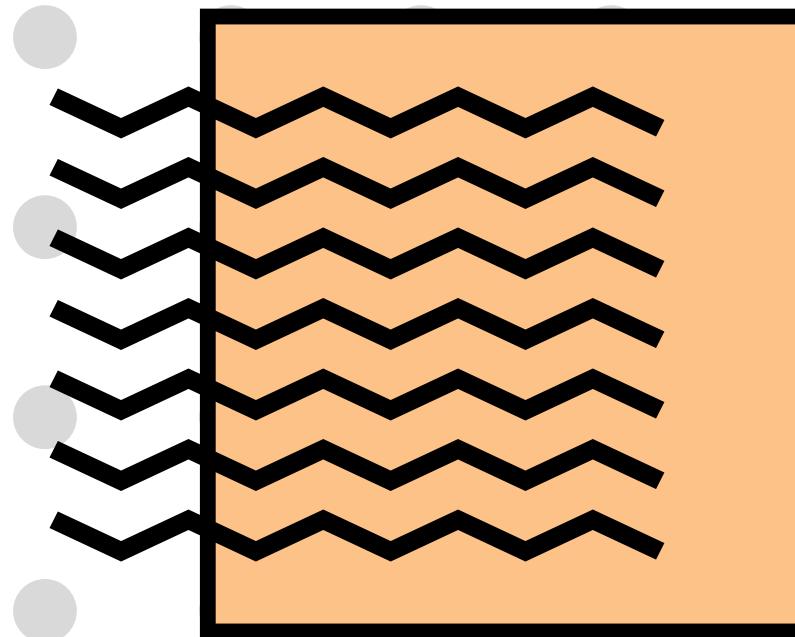
The data has shown that the call centers are effective at retaining customers. Promote this service so that customers keep using the product.

TARGETTING

Cluster 1 contains customers with high monthly consumption, high tenure months and low churn rate.

Users in this cluster use educational product, therefore the company should conduct advertising campaigns targeting students.





**THANKS FOR
WATCHING**



**GOOGLE COLAB
LINK:
[HTTPS://COLAB.RESEARCH.GOOGLE.COM/DRIVE/1A13J9BSIZM9NKGVTAR2BV_Z7MWPICH8?USP=SHARING](https://colab.research.google.com/drive/1A13J9BSIZM9NKGVTAR2BV_Z7MWPICH8?usp=sharing)**