Mini Project

Customer Segmentation

Agenda

- Introduction
- Exploratory Data Analysis
- Build Model
- Conclusion

Introduction

- Customer Segmentation: là bài toán phân tích về khách hàng của một doanh nghiệp. Việc phân tích giúp doanh nghiệp hiểu rõ về khách hàng của mình hơn, giúp họ đưa ra chiến lược quảng cáo, tiếp thị… cho các loại khách hàng khác nhau
- Input: bộ dữ liệu tóm tắt các hành vi hoạt động (18 biến hành vi) của các chủ thẻ tín dụng đang hoạt động trong vòng 6 tháng
- Output: Chia tập khách hàng thành các phân khúc

Introduction

- Tập dữ liệu: Credit Card Dataset từ Kaggle
- Tập dữ liệu gồm 18 thuộc tính: ID khách hàng, số dư tài khoản, tần suất số dư cập nhật, số lần mua hàng sử dụng thẻ, số tiền mua tối đa được thực hiện trong một lần, số tiền mua trả góp, tiền ứng trước, tần suất mua hàng, tần suất mua hàng trả góp, tần suất thanh toán tiền mặt trước, số lượng giao dịch trả trước, hạn mức thẻ tín dụng…

Introduction

Following is the Data Dictionary for Credit Card dataset :-

CUST_ID : Identification of Credit Card holder (Categorical)

BALANCE : Balance amount left in their account to make purchases (

BALANCE_FREQUENCY : How frequently the Balance is updated, score between 0 and **1** (**1** = frequently updated, 0 = not frequently updated)

PURCHASES : Amount of purchases made from account

ONEOFF_PURCHASES : Maximum purchase amount done in one-go

INSTALLMENTS_PURCHASES : Amount of purchase done in installment

CASH_ADVANCE : Cash in advance given by the user

PURCHASES_FREQUENCY : How frequently the Purchases are being made, score between 0 and **1** (**1** = frequently purchased, 0 = not frequently purchased)

ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (**1** = frequently purchased, 0 = not frequently purchased)

PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (**1** = frequently done, 0 = not frequently done)

CASHADVANCEFREQUENCY : How frequently the cash in advance being paid

CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"

PURCHASES_TRX : Numbe of purchase transactions made

CREDIT_LIMIT : Limit of Credit Card for user

PAYMENTS : Amount of Payment done by user

MINIMUM_PAYMENTS : Minimum amount of payments made by user

PRCFULLPAYMENT : Percent of full payment paid by user

TENURE : Tenure of credit card service for user

Exploratory Data Analysis

- Thực hiện một số khảo sát về dữ liệu
  - Missing value
  - Scatter
  - Distribution

  …

Exploratory Data Analysis

Tập dữ liệu chứa 8950 bản ghi và 18 cột thuộc tính.

Các loại dữ liệu của các thuộc tính bao gồm 3 thuộc tính kiểu số nguyên rời rạc,14 thuộc tính số thực liên tục và 1 thuộc tính kiểu đối tượng.

Sử dụng không gian bộ nhớ ít nhất là 1.2 megabyte (MB).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8950 entries, 0 to 8949
Data columns (total 18 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   CUST_ID                           8950 non-null   object
 1   BALANCE                           8950 non-null   float64
 2   BALANCE_FREQUENCY                 8950 non-null   float64
 3   PURCHASES                         8950 non-null   float64
 4   ONEOFF_PURCHASES                  8950 non-null   float64
 5   INSTALLMENTS_PURCHASES            8950 non-null   float64
 6   CASH_ADVANCE                      8950 non-null   float64
 7   PURCHASES_FREQUENCY               8950 non-null   float64
 8   ONEOFF_PURCHASES_FREQUENCY        8950 non-null   float64
 9   PURCHASES_INSTALLMENTS_FREQUENCY  8950 non-null   float64
 10  CASH_ADVANCE_FREQUENCY            8950 non-null   float64
 11  CASH_ADVANCE_TRX                  8950 non-null   int64
 12  PURCHASES_TRX                     8950 non-null   int64
 13  CREDIT_LIMIT                      8949 non-null   float64
 14  PAYMENTS                          8950 non-null   float64
 15  MINIMUM_PAYMENTS                  8637 non-null   float64
 16  PRC_FULL_PAYMENT                  8950 non-null   float64
 17  TENURE                            8950 non-null   int64
dtypes: float64(14), int64(3), object(1)
memory usage: 1.2+ MB
```

Exploratory Data Analysis

(8950, 18)

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLME
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166667	0.000000	
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000000	0.000000	
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.000000	1.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.083333	0.083333	
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.083333	0.083333	

Bản ghi trong tập dữ liệu

Exploratory Data Analysis

	Name of Col	Num of Null	Dtype	N_Unique
0	CUST_ID	0	object	8950
1	BALANCE	0	float64	8871
2	BALANCE_FREQUENCY	0	float64	43
3	PURCHASES	0	float64	6203
4	ONEOFF_PURCHASES	0	float64	4014
5	INSTALLMENTS_PURCHASES	0	float64	4452
6	CASH_ADVANCE	0	float64	4323
7	PURCHASES_FREQUENCY	0	float64	47
8	ONEOFF_PURCHASES_FREQUENCY	0	float64	47
9	PURCHASES_INSTALLMENTS_FREQUENCY	0	float64	47
10	CASH_ADVANCE_FREQUENCY	0	float64	54
11	CASH_ADVANCE_TRX	0	int64	65
12	PURCHASES_TRX	0	int64	173
13	CREDIT_LIMIT	1	float64	205
14	PAYMENTS	0	float64	8711
15	MINIMUM_PAYMENTS	313	float64	8636
16	PRC_FULL_PAYMENT	0	float64	47
17	TENURE	0	int64	7

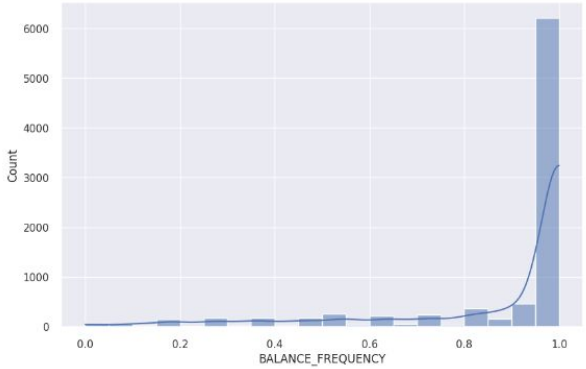
Thống kê số giá trị null

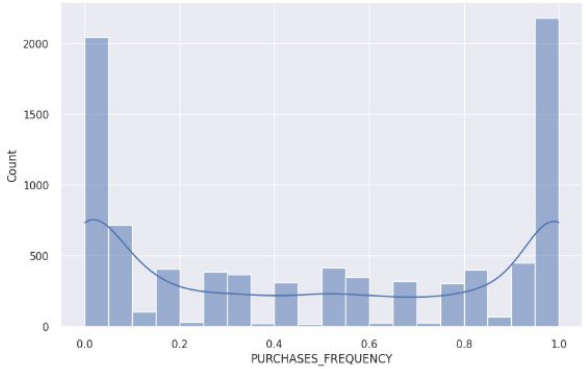
Exploratory Data Analysis

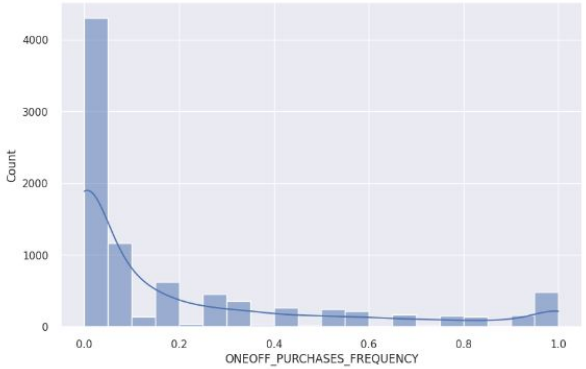
- Xóa bỏ cột thuộc tính CUST_ID
- Xử lý các bản ghi chứa thuộc tính null

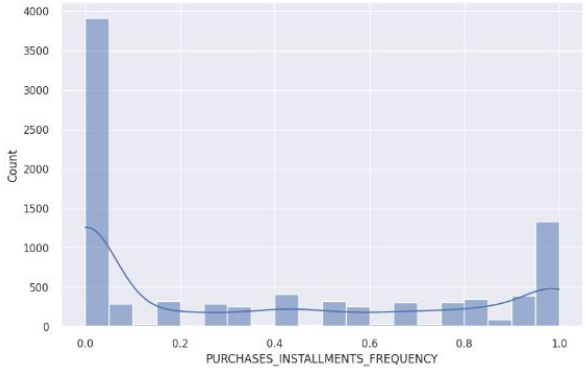
```
data.drop(columns='CUST_ID', inplace=True)

data['MINIMUM_PAYMENTS'].fillna( data['MINIMUM_PAYMENTS'].median(), inplace = True )
data.dropna(subset=['CREDIT_LIMIT'], inplace=True)
```

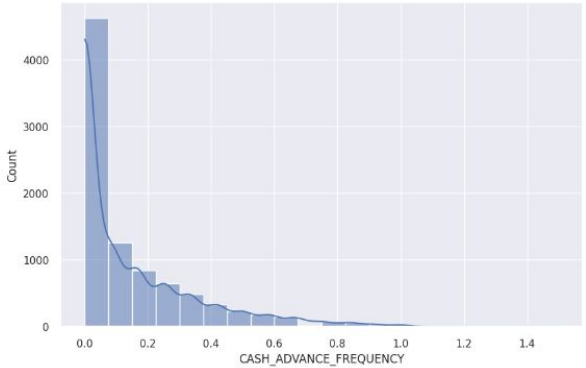
Exploratory Data Analysis

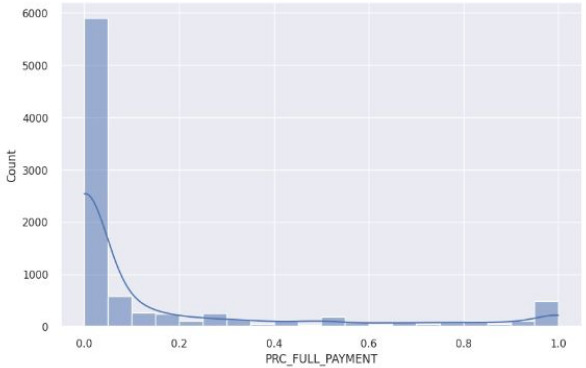




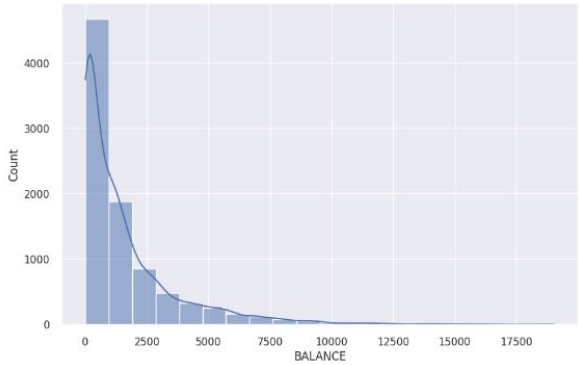


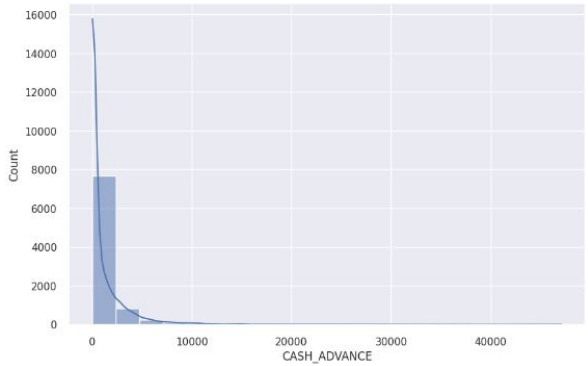


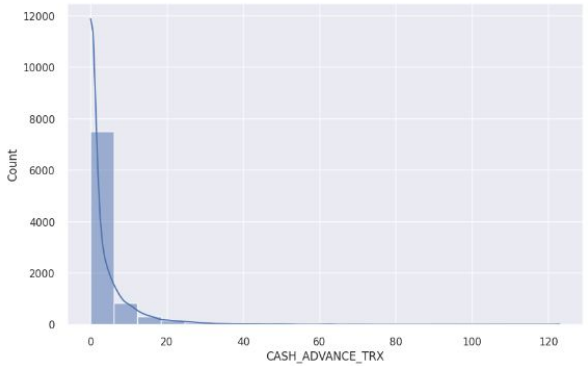


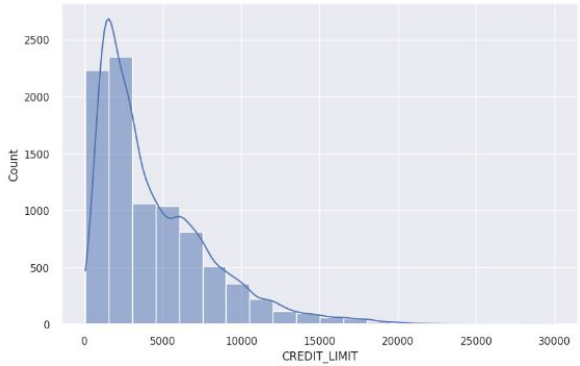


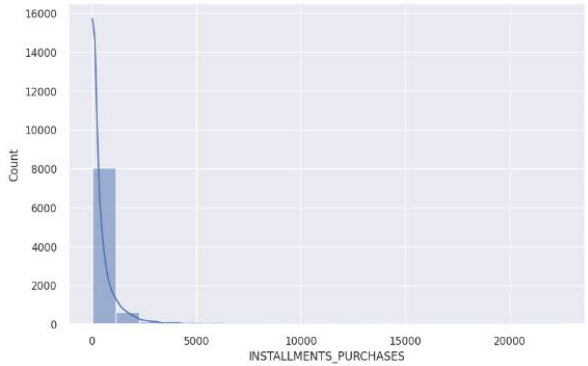
Exploratory Data Analysis

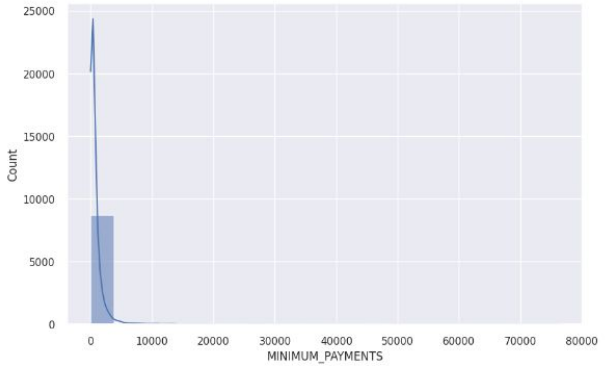




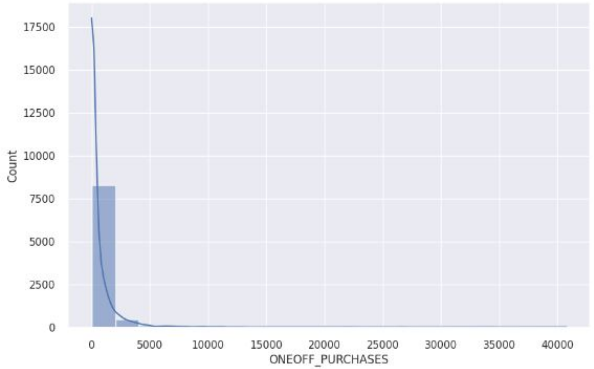


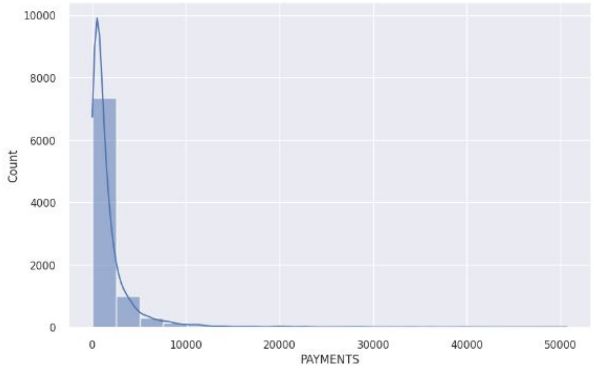


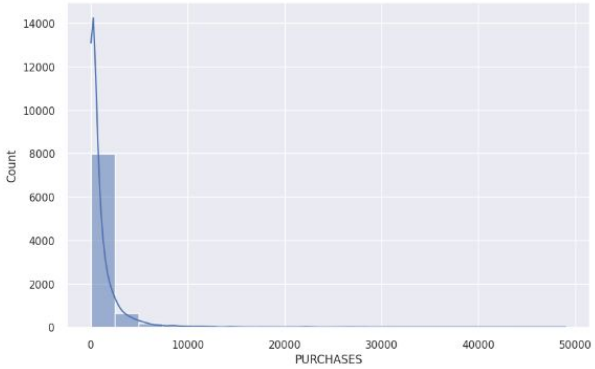


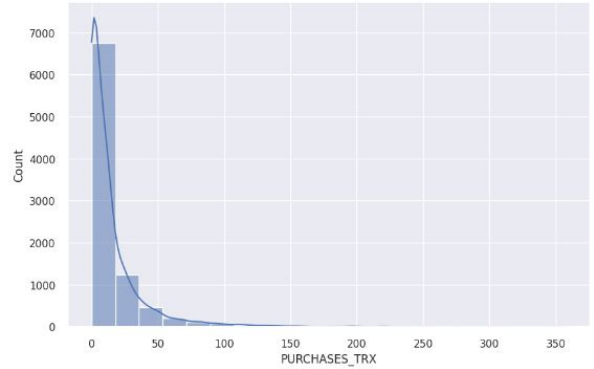


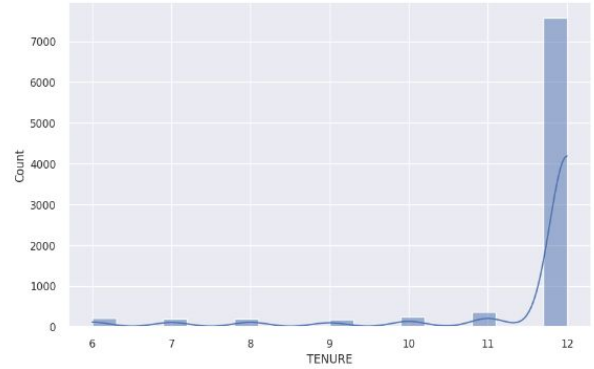
Exploratory Data Analysis



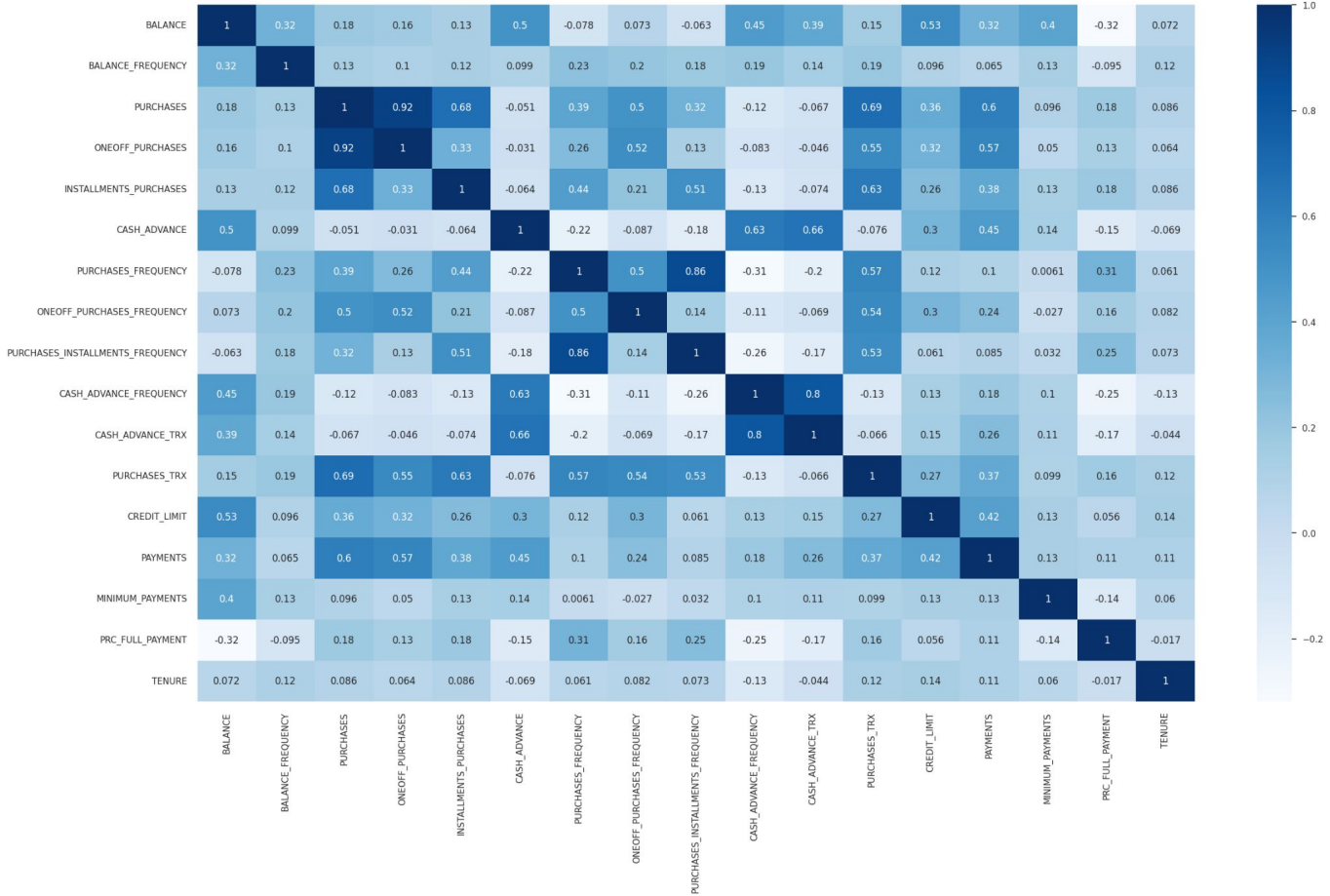








Exploratory Data Analysis

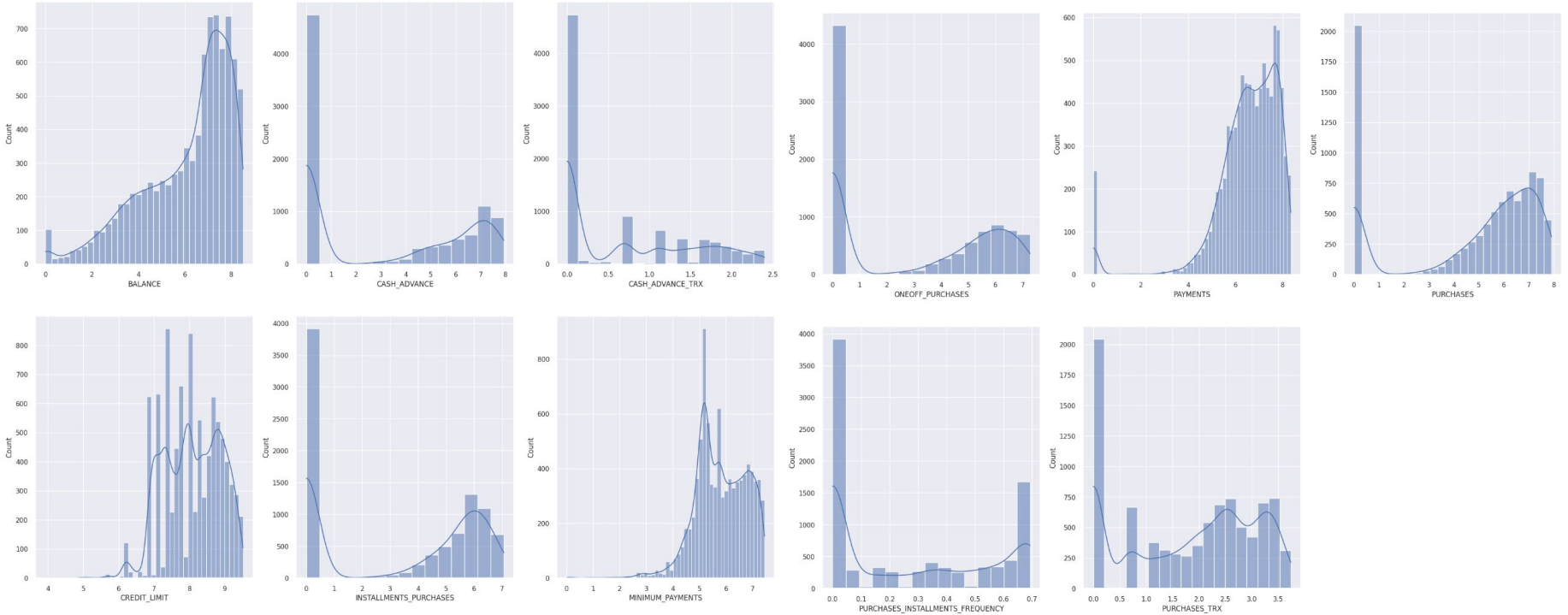


Exploratory Data Analysis

- Phát hiện & xử lý các điểm ngoại lai: sử dụng hàm Quantile và KNNImputer

```
BALANCE                             695
BALANCE_FREQUENCY                  1492
PURCHASES                           808
ONEOFF_PURCHASES                   1013
INSTALLMENTS_PURCHASES              867
CASH_ADVANCE                       1030
PURCHASES_FREQUENCY                   0
ONEOFF_PURCHASES_FREQUENCY          782
PURCHASES_INSTALLMENTS_FREQUENCY      0
CASH_ADVANCE_FREQUENCY              525
CASH_ADVANCE_TRX                    804
PURCHASES_TRX                       766
CREDIT_LIMIT                        248
PAYMENTS                            808
MINIMUM_PAYMENTS                    909
PRC_FULL_PAYMENT                   1474
TENURE                             1365
dtype: int64
```

Exploratory Data Analysis



Exploratory Data Analysis

- Chuẩn hóa dữ liệu: StandardScaler

```
from sklearn.preprocessing import StandardScaler
Scaler = StandardScaler()
df_scaled = Scaler.fit_transform(trans_df)
df_scaled = pd.DataFrame(df_scaled, columns=df.columns)
df_scaled
```

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVAN
0	-1.218028	-2.647658	-0.075795	-0.996248	0.471989	-0.915486	-0.806649	-0.744814	-0.674357	
1	1.026043	-1.140357	-1.713215	-0.996248	-1.103750	1.202662	-1.221928	-0.744814	-0.958359	
2	0.896944	0.366944	0.670885	1.270692	-1.103750	-0.915486	1.269742	2.115946	-0.958359	
3	0.688240	0.065483	0.907951	0.959792	-1.103750	0.703721	-1.014290	-0.378051	-0.958359	
4	0.320093	0.366944	-0.697749	-0.030685	-1.103750	-0.915486	-1.014290	-0.378051	-0.958359	
...	...	...	...	...	...	...	...	...	...	
8944	-1.399726	0.366944	0.321565	-0.996248	0.854380	-0.915486	1.269742	-0.744814	1.192296	
8945	-1.596010	0.366944	0.332298	-0.996248	0.864709	-0.915486	1.269742	-0.744814	1.192296	
8946	-1.497859	-2.396449	0.071507	-0.996248	0.613742	-0.915486	0.854463	-0.744814	0.854123	
8947	-1.768653	-2.396449	-1.713215	-0.996248	-1.103750	0.185683	-1.221928	-0.744814	-0.958359	
8948	-0.085732	0.366944	0.794909	1.388621	-1.103750	0.558145	0.439186	2.189301	-0.958359	

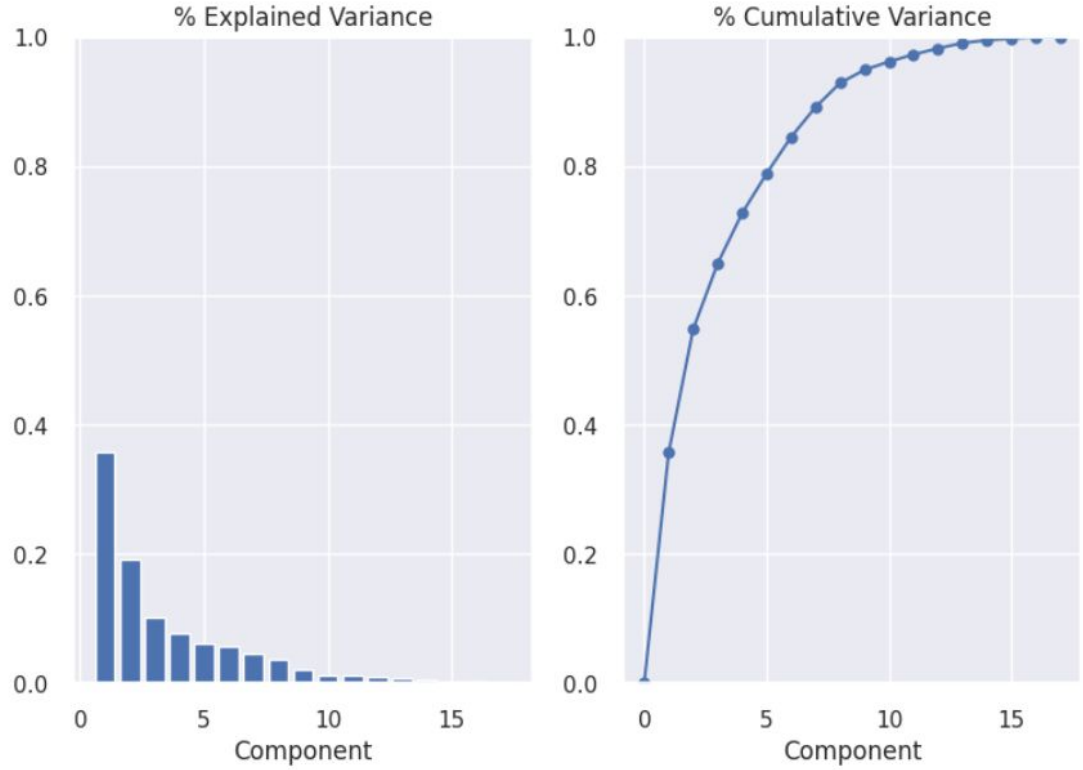
Exploratory Data Analysis

- Giảm chiều dữ liệu: sử dụng PCA

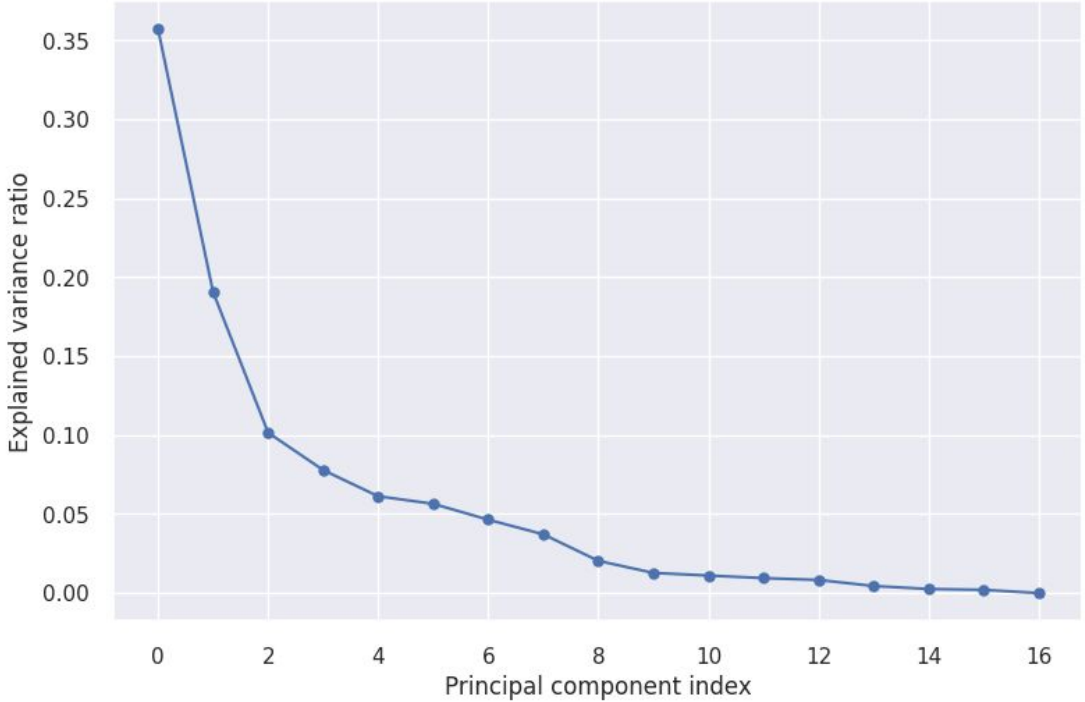
```
component_names = [f"PC{i+1}" for i in range(X_pca.shape[1])]

X_pca = pd.DataFrame(X_pca, columns=component_names)

X_pca.head()
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
0	0.069796	3.130811	0.006396	0.663919	-0.068158	2.462334	0.613468	0.140814	-0.188834	0.281248	-0.019296	-0.618209	0.614137	-0.021790	0.063023	0.121120	0.0
1	3.672927	-1.013897	0.494409	2.506673	-2.158713	-0.485405	0.574291	-1.199969	0.254733	0.103307	-0.124536	-0.051441	-0.193520	-0.103747	-0.006581	0.022669	-0.0
2	-1.523284	-1.488468	-2.741011	-0.786700	-0.156101	-0.084165	0.047017	-0.590857	1.089029	0.304061	-0.012100	0.869735	0.153590	-0.091977	0.106437	0.578654	0.0
3	1.377533	0.859298	-1.746111	-1.311830	1.696081	0.603446	-1.822839	-3.132442	-1.127105	-1.117890	0.549913	0.737677	0.497196	0.287129	0.661442	-0.138471	-0.0
4	1.074717	1.564840	-1.307795	-0.927415	-0.375704	-0.111225	1.387779	0.467646	-0.041009	-0.111425	0.472472	0.110265	-0.121659	-0.099385	-0.048007	-0.056304	0.0

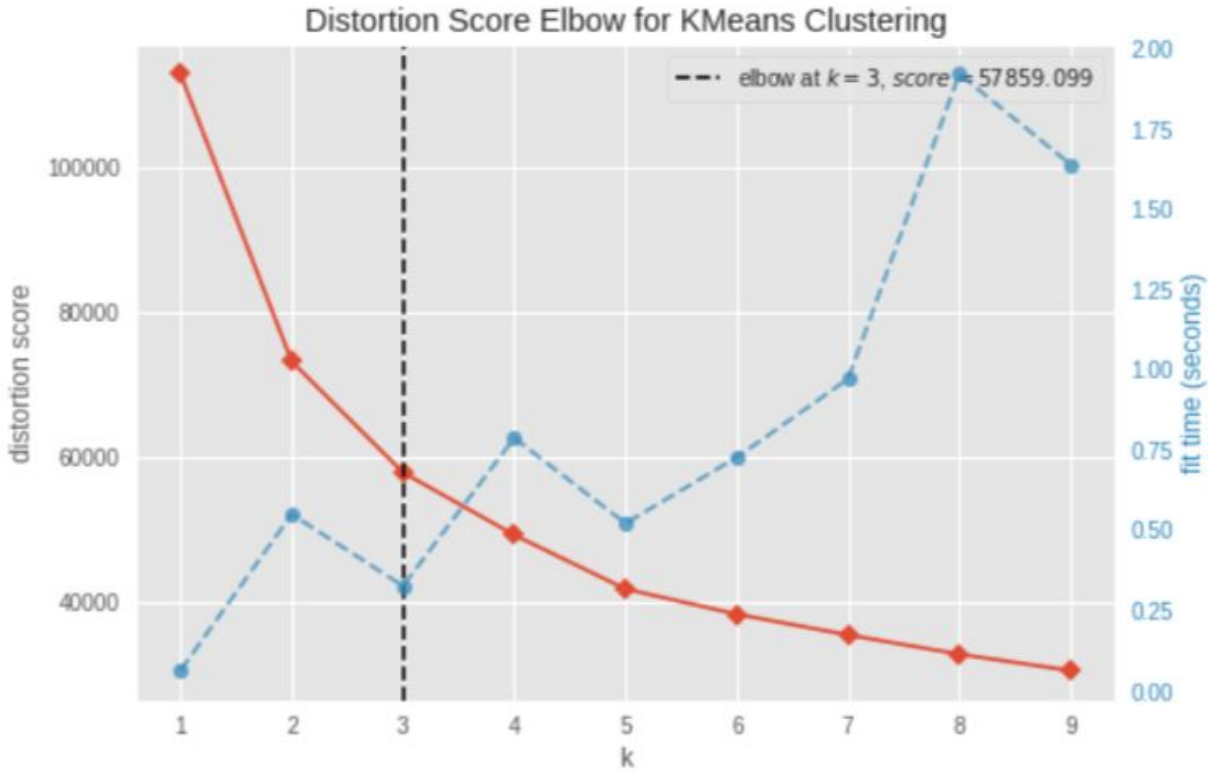
Exploratory Data Analysis



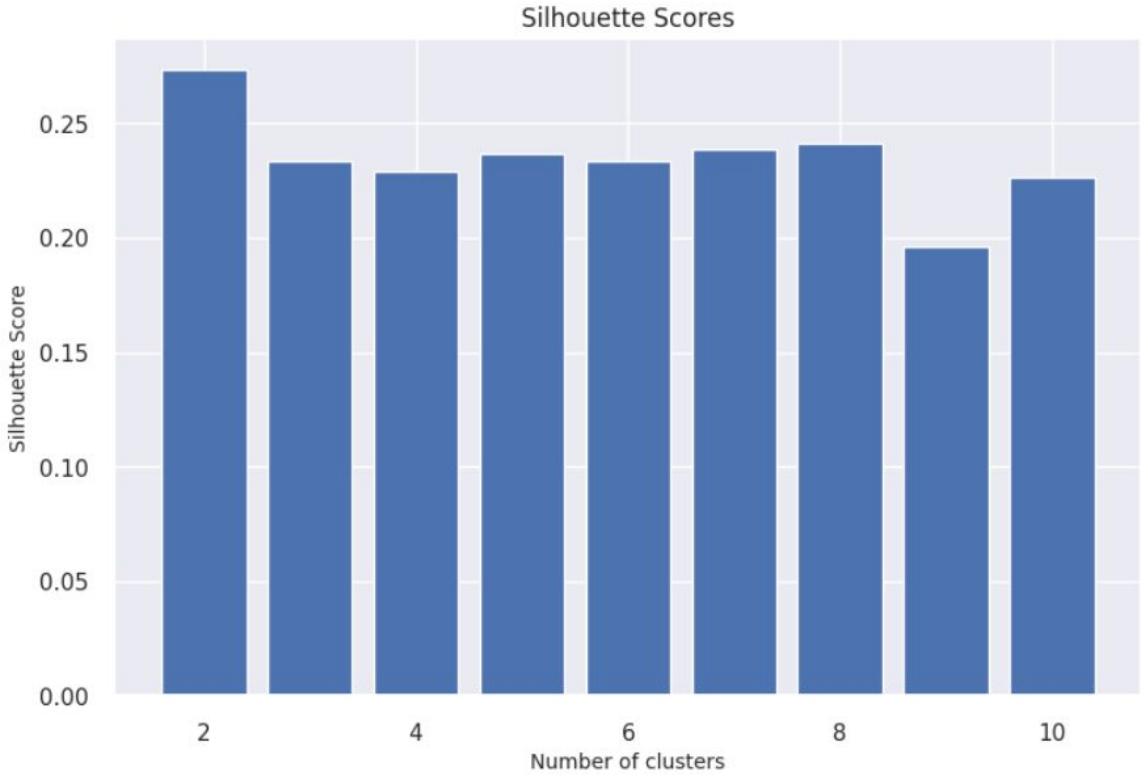
Exploratory Data Analysis



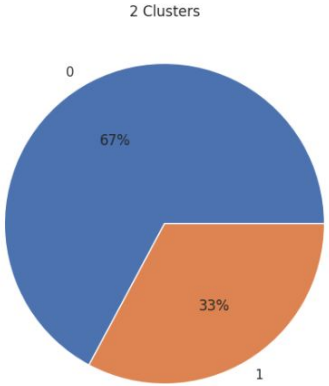
Build Model: KMeans

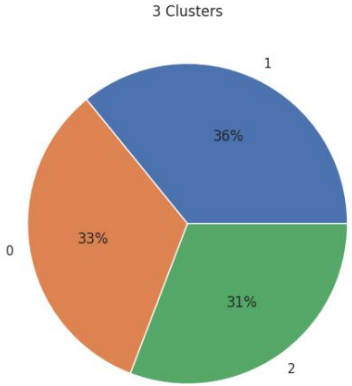


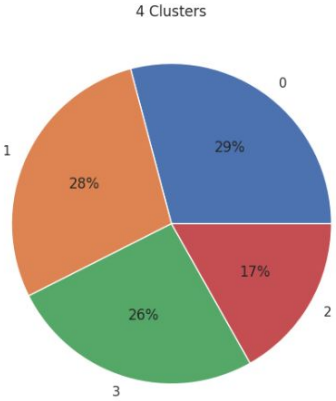
Build Model: KMeans

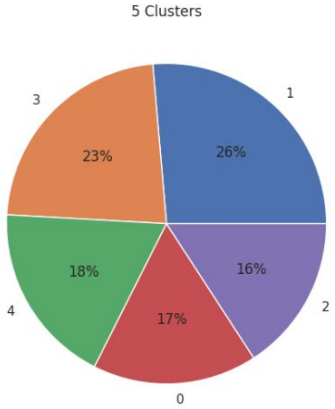


Build Model: KMeans

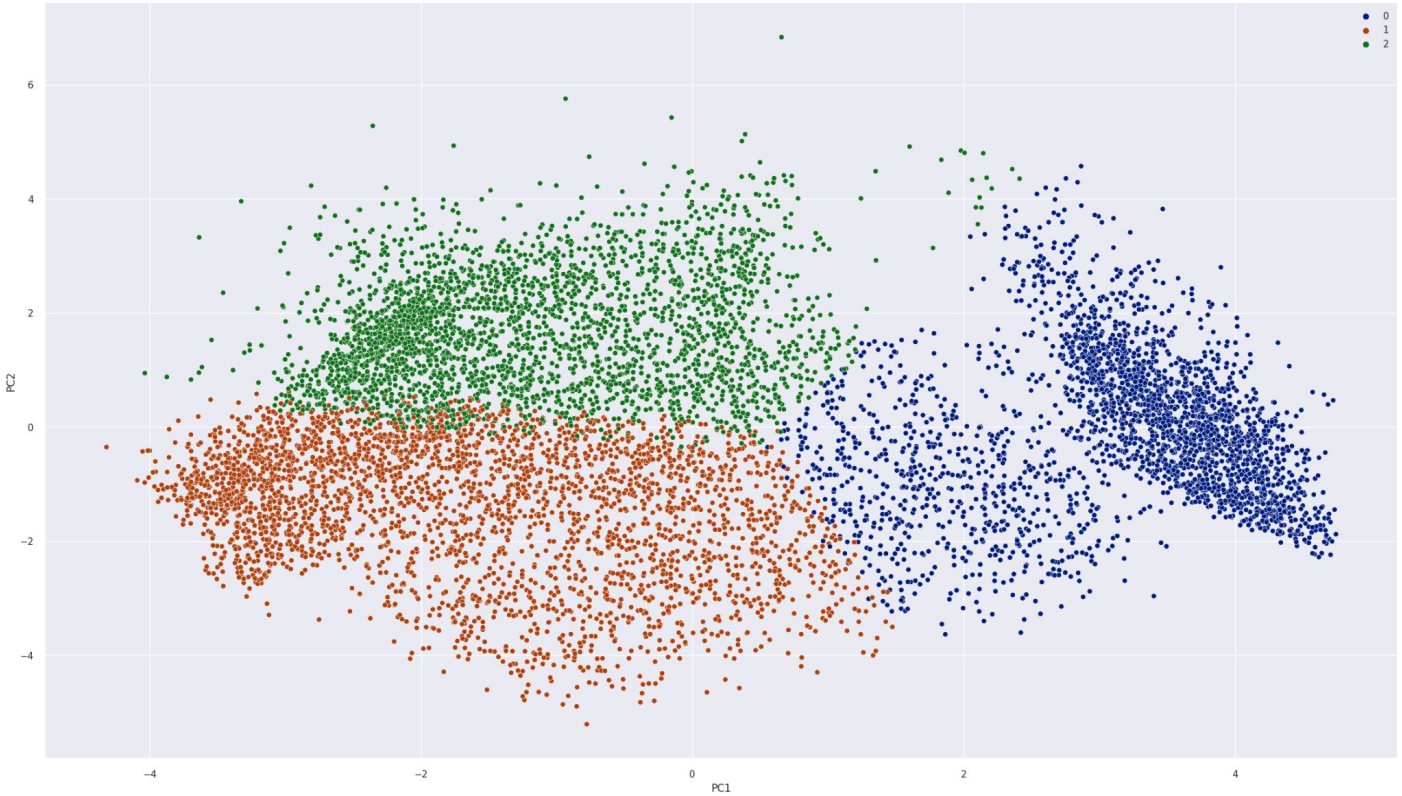








Build Model: KMeans



Build Model: KMeans

- Cụm 1: Các khách hàng có số dư tài khoản thấp, cập nhật số dư thường xuyên. Họ không thích thanh toán trả trước và có nhu cầu mua hàng trả góp. Nhóm này thường xuyên mua hàng với khoản chi thấp và có hạn mức tín dụng thấp
- Cụm 2: Các khách hàng có số dư tài khoản ở mức trung bình, tần suất cập nhật số dư thường xuyên hơn Cụm 1. Họ ưu tiên mua hàng trả góp và chi tiêu ở mức trung bình. Nhóm này thường xuyên mua hàng với khoản chi cao và có hạn mức tín dụng cao
- Cụm 3: Các khách hàng có số dư tài khoản ở mức trên trung bình. Họ thường thanh toán trả trước, và không có nhu cầu mua hàng trả góp. Nhóm này không mua hàng thường xuyên, và khi mua hàng thường chi tiêu mức trung bình và hạn mức tín dụng trung bình

Thank you