

Analisis Churn Dalam Pemasaran

Tim Kamboja:
Azis Rahmanto
Fahmi Ramadhan Putra
Ridho Ardia Rahman
Umi Purnamasari



Daftar Isi

Poin-poin Diskusi

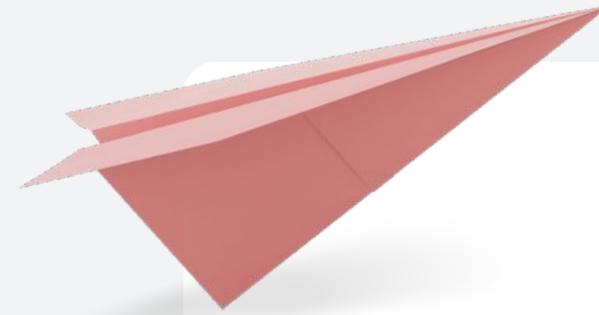
Stage 1 - Understanding

Stage 2 - Identify

Stage 3 - Exploratory Data Analysis and Visualization

Stage 4 - Data Preprocessing

Stage 5 - Modelling



Stage 1

Understanding



DigitalSkola

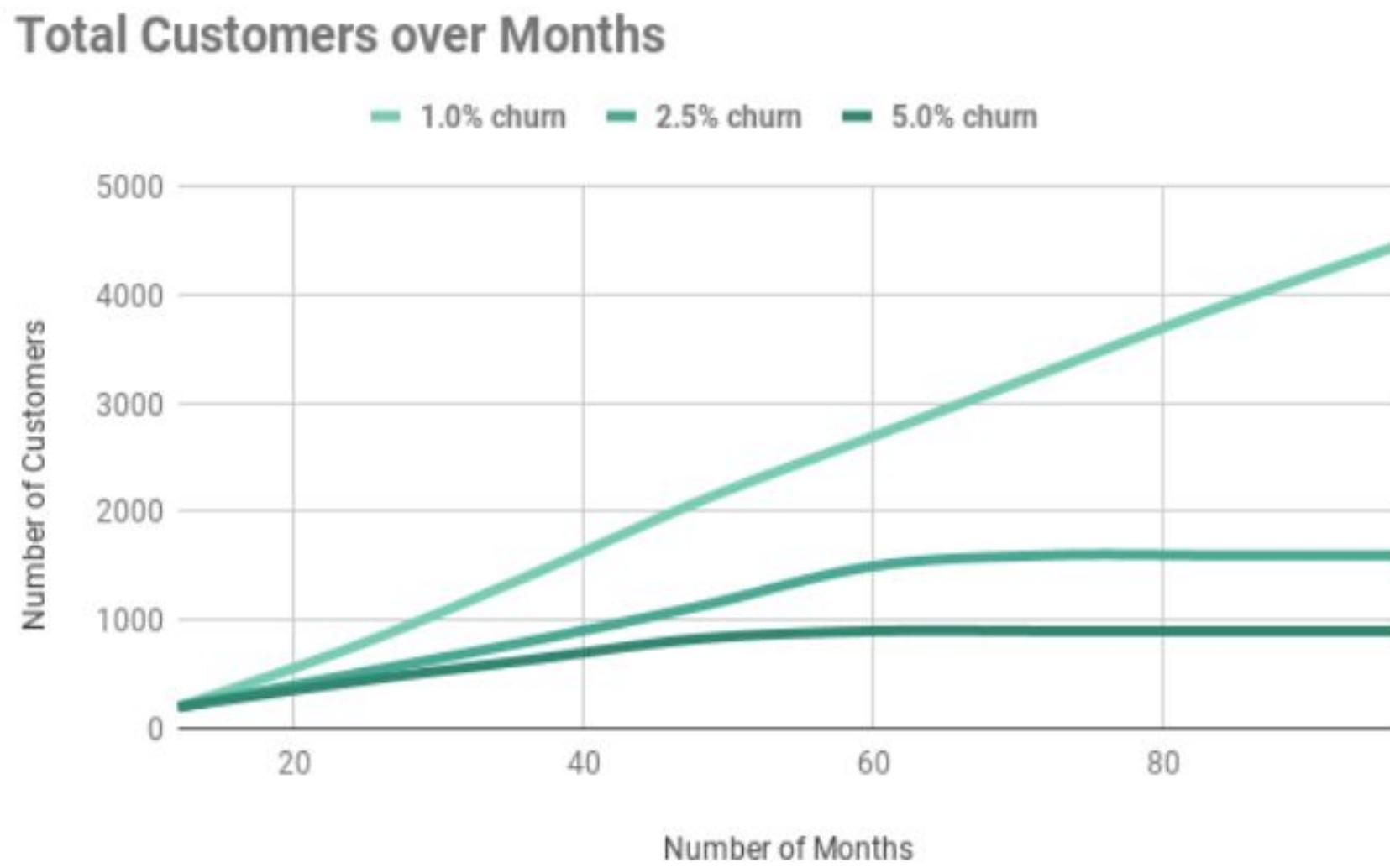
Apakah **analisis Churn** untuk pelanggan?

Analisis churn adalah evaluasi tingkat kehilangan atau perpindahan pelanggan perusahaan, sehingga dapat dipertimbangkan upaya untuk menguranginya.

Ini juga disebut sebagai **tingkat atrisi pelanggan**, churn dapat diminimalkan dengan menilai produk Anda dan bagaimana orang menggunakannya.



Contoh Analisis Churn



*contoh bersumber dari <https://www.profitwell.com/customer-churn/analysis>

Tingkat churn yang tinggi memaksa bisnis untuk bersaing dengan tekanan dan kesulitan membawa cukup banyak pelanggan baru. Bahkan pada tingkat kecil, peningkatan satu angka dalam tingkat churn (%) dapat dengan cepat memiliki efek negatif yang besar pada kemampuan perusahaan untuk tumbuh.

Macam-macam bentuk

Customer Churn

Berhenti Berlangganan

Langgan yang dibatalkan mungkin merupakan jenis churn pertama, dan itu dapat dimotivasi oleh sejumlah alasan berbeda. Seperti: **ketidakcocokan pelanggan, fungsionalitas yang hilang, dan kegagalan untuk mencapai hasil.**

Tidak Memperbaharui Langganan

Beberapa churn bukanlah hasil dari ketidakpuasan aktif terhadap produk atau layanan, tetapi hasil dari pemeliharaan hubungan pelanggan yang tidak tepat. Seringkali, pelanggan yang tidak cukup terlibat hanya akan menjauh dari suatu produk.

Berpindah ke Kompetitor

Aspek-aspek tertentu dari analisis churn mengharuskan kita untuk fokus pada operasi perusahaan itu sendiri. Namun, ada elemen kompetitif juga..

Penutupan Akun

Bahkan jika pelanggan meninggalkan bisnis kita, senang dengan layanan yang kita berikan dan kebutuhan mereka terpenuhi, itu masih merupakan bentuk churn.

Kenapa Analisis Churn Penting?

Mengapa memiliki pemahaman tentang analisis churn sangat penting?. Memahami berbagai alasan di balik churn pelanggan adalah langkah mendasar dalam mengatasi dan mengurangi rate.



Mengapa perlu menganalisis *Churn* secara berkala ?

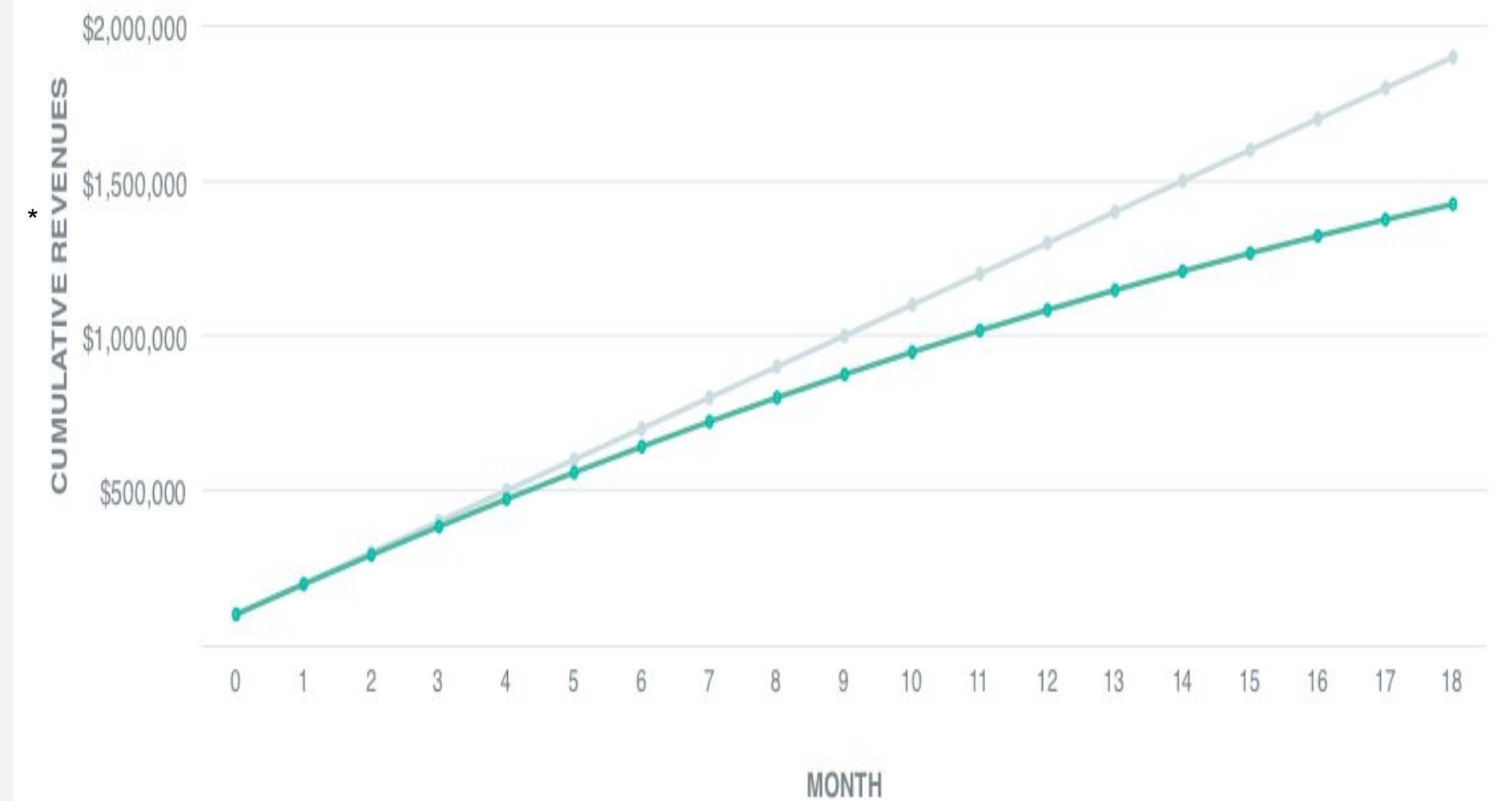
Poin diskusi

Churn adalah statistik yang sangat berpengaruh di seluruh bisnis SaaS*. Ini adalah metrik di mana bisnis, tua dan muda, hidup atau mati. Membiarkan tingkat churn merayap lebih tinggi dapat menyebabkan sejumlah masalah terkait.

COMPARING REVENUES WITH & WITHOUT DELINQUENT CHURN

Involuntary churn reduces revenue

Revenues fall significantly over the customer lifecycle when nothing is done to fight involuntary churn



*contoh bersumber dari <https://www.profitwell.com/customer-churn/analysis>

Analisis Data Churn



Analisis Data Churn

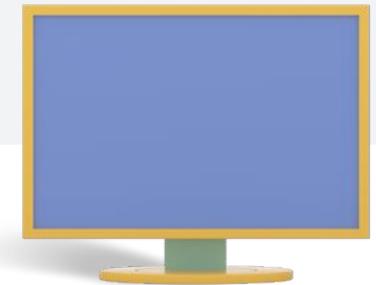


Seperti halnya bentuk analisis apa pun, analisis churn yang efektif mengharuskan kita untuk melacak data yang benar. Menetapkan tujuan berorientasi KPI* yang tepat dapat membantu kita melihat lebih dekat apa yang mempengaruhi churn kita.



BENTUK-BENTUK KPI

- Pelibatan pelanggan dan penggunaan.
- Harga pesaing.
- Kemungkinan untuk *upgrade* dan improvisasi.



TAHAPAN ANALISIS LANJUTAN SETELAH MENGETAHUI KPI:

- Pola perilaku pelanggan
- Segmentasi pelanggan
- Perilaku titik sentuh (*Touch point behaviour*)

Yang diperlukan untuk Analisis Churn

Kita harus mengambil
pendekatan holistik untuk
berbagai jenis potensi churn.



MEMPREDIKSI CHURN

Solusi yang bermanfaat akan dapat memprediksi kemungkinan penyebab churn dan menandai setiap pelanggan yang berisiko.

ANALISIS TINGKATAN HARGA

Tingkat penetapan harga yang tepat sangat penting untuk konversi serta churn minimal. Pastikan bahwa tingkatan harga kita difokuskan pada persona pembeli, bukan persepsi kita tentang fitur kita sendiri.

MENDAPATKAN METRIK (CONTOH:ARPU, MRR, dan ARR)*

Memahami churn sangat mendasar karena efek churn dapat terjadi pada metrik utama yang lain.

*ARPU: Average Revenue per User, MRR: Monthly Recurring Revenue, ARR:Annual Recurring Revenue

Bentuk Churn yang baik

Sebenarnya, hanya ada satu jenis churn yang baik, dan itu adalah "**churn negatif**".

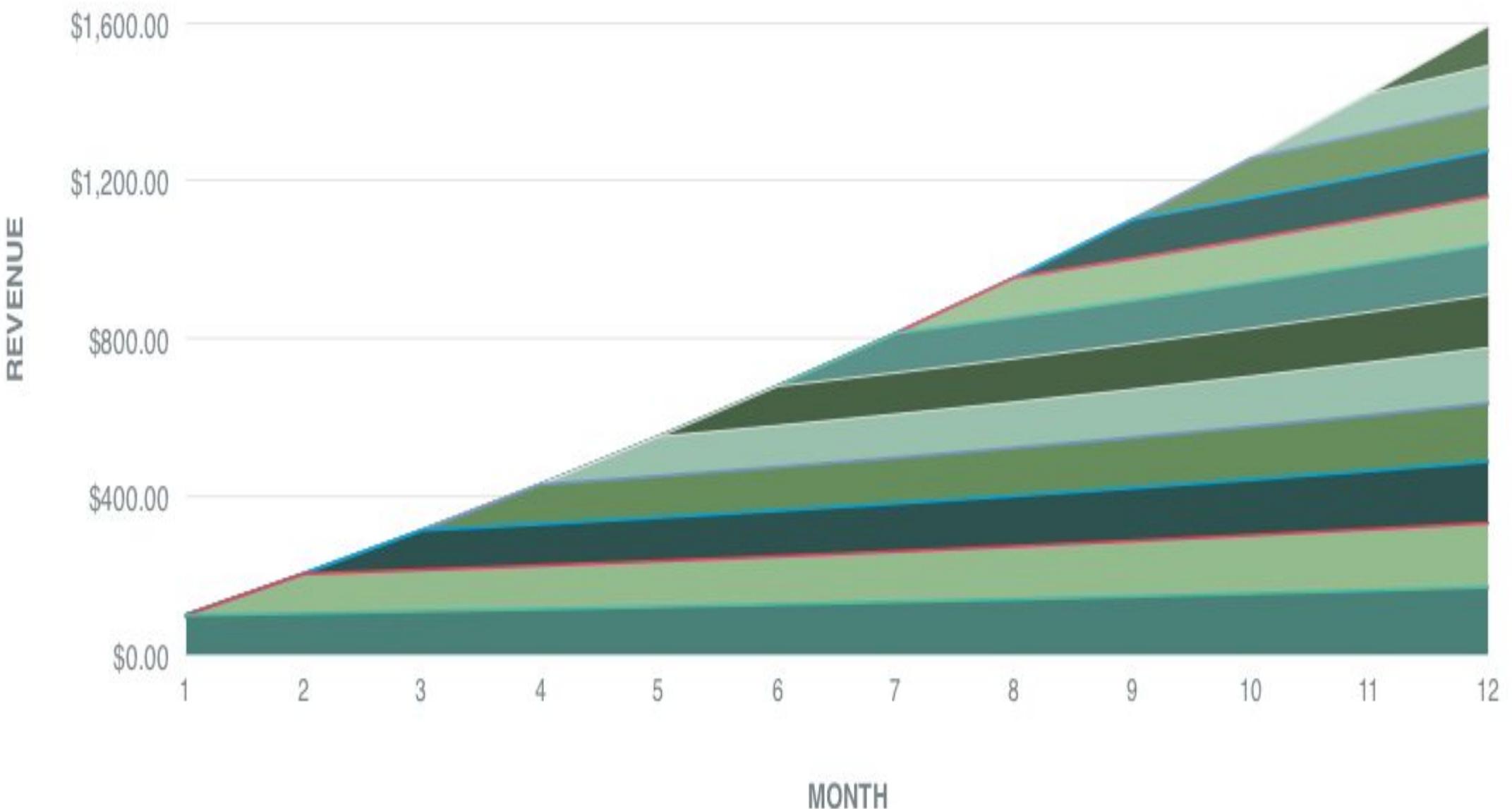
Memahami arah dari mana risiko churn berasal, dan bahwa masing-masing mewakili **peluang** untuk meningkatkan perusahaan kita dan **memperkuat hubungan pelanggan** kita, adalah **langkah pertama**.



THE POWER OF NEGATIVE CHURN

Growth success with 5% negative churn

When you are growing each revenue cohort through expansions and upsells, this grows the foundation of your business, so each new cohort grows on top of the previous.



*contoh bersumber dari <https://www.profitwell.com/customer-churn/analysis>



Identifikasi Dataset Churn

Sumber :

<https://www.kaggle.com/shubh0799/churn-modelling>

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
---  --  
 0   RowNumber        10000 non-null   int64  
 1   CustomerId       10000 non-null   int64  
 2   Surname          10000 non-null   object  
 3   CreditScore      10000 non-null   int64  
 4   Geography         10000 non-null   object  
 5   Gender            10000 non-null   object  
 6   Age               10000 non-null   int64  
 7   Tenure            10000 non-null   int64  
 8   Balance           10000 non-null   float64 
 9   NumOfProducts     10000 non-null   int64  
 10  HasCrCard         10000 non-null   int64  
 11  IsActiveMember    10000 non-null   int64  
 12  EstimatedSalary   10000 non-null   float64 
 13  Exited            10000 non-null   int64  
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

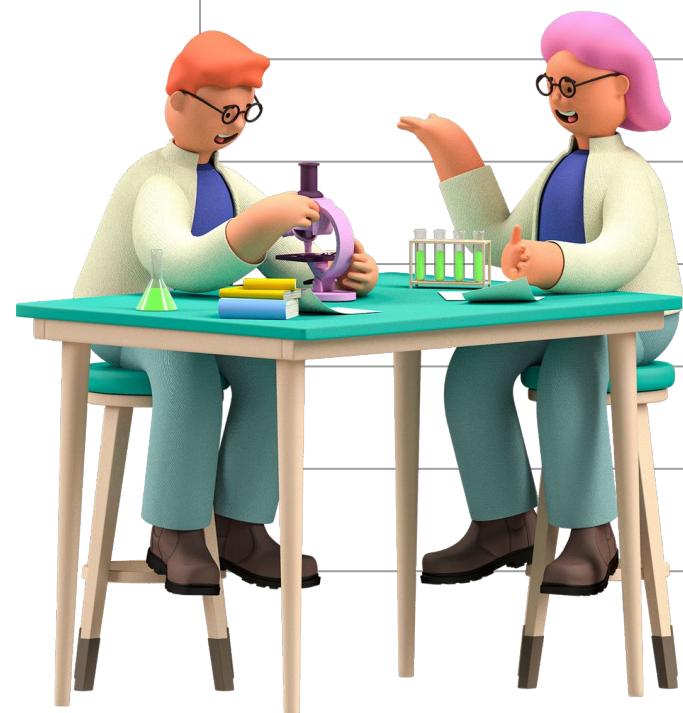


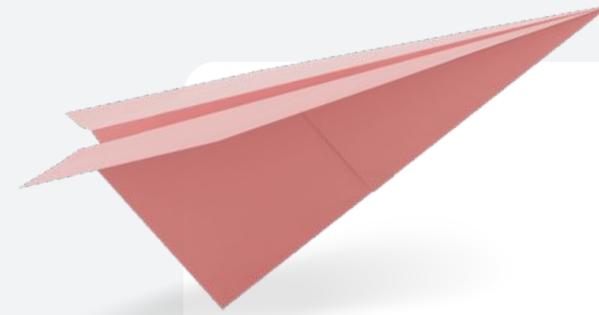
Deskripsi Dataset Churn

- ❑ Dataset memiliki 14 kolom. Kolom terakhir adalah variabel dependen ['Exited']. Angka 1 di kolom itu memberitahu kita bahwa pelanggan telah meninggalkan bank [churn].
- ❑ Kita dapat melihat bahwa bank beroperasi di 3 negara [Prancis, Spanyol dan Jerman].
- ❑ Kolom ['NumOfProducts'] mengacu pada jumlah layanan yang telah dimanfaatkan oleh pelanggan oleh bank, misalnya pinjaman, kartu kredit, rekening tabungan, dll.
- ❑ ['Tenure'] atau jangka waktu mengacu pada jumlah tahun pelanggan telah bersama dengan bank tersebut.
- ❑ Kita dapat melihat bahwa ada jenis objek data di kolom ['Gender'] dan ['Geography']. Kemungkinan kita harus merubahnya menjadi numerik.

Tipe Dataset Churn

KOLOM	TIPE DATASET
RowNumber	Integer
CustomerId	Integer
Surname	Object
CreditScore	Integer
Geography	Object
Gender	Object
Age	Integer
Tenure	Integer
Balance	Float
NumOfProducts	Integer
HasCrCard	Integer
IsActiveMember	Integer
EstimatedSalary	Float
Exited	Integer





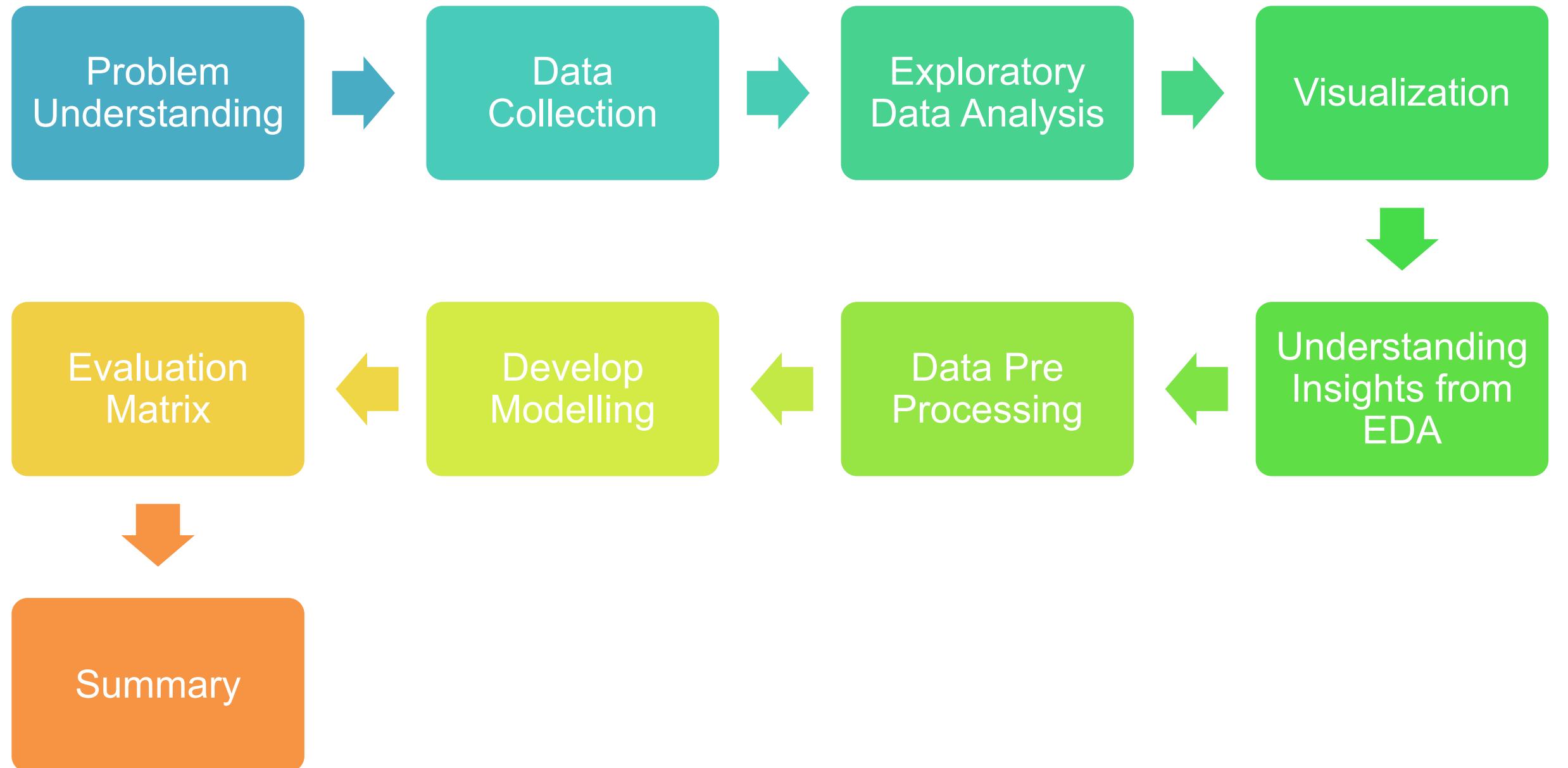
Stage 2

Identify

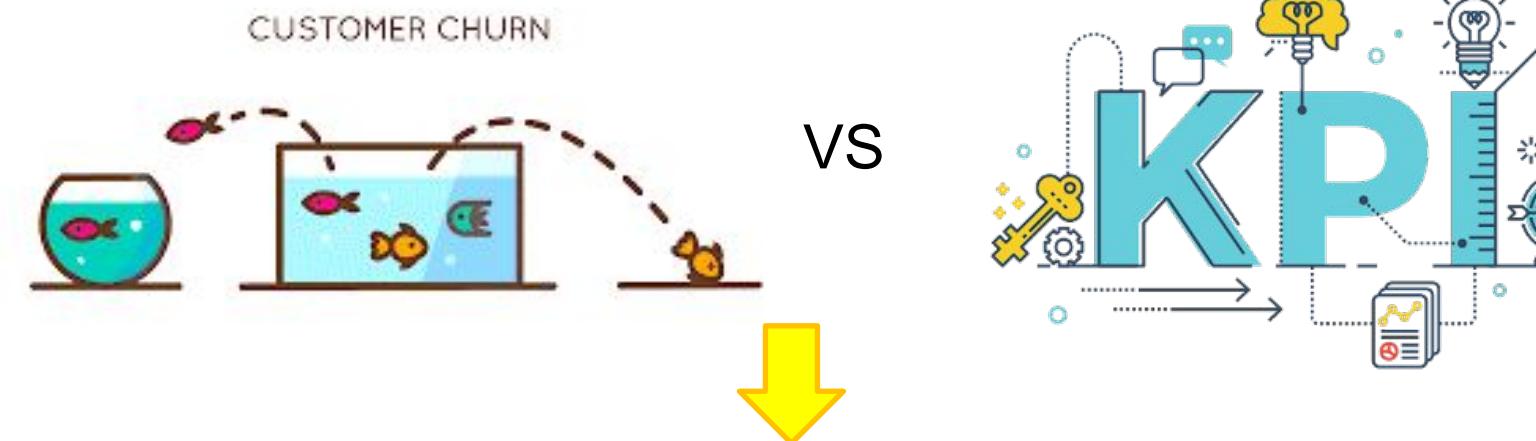


DigitalSkola

Strategy



Conceptual Model



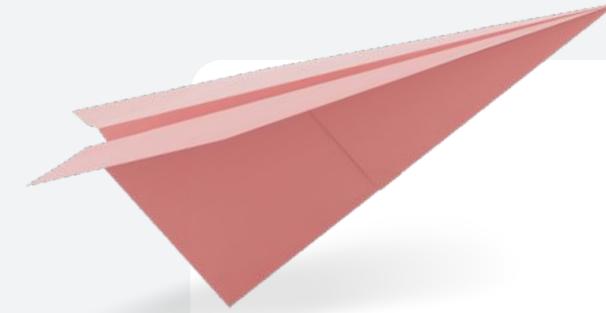
Analisis Churn dan
Membandingkan dengan KPI

```
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count Dtype
 ---  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography         10000 non-null  object
 5   Gender            10000 non-null  object
 6   Age               10000 non-null  int64
 7   Tenure            10000 non-null  int64
 8   Balance           10000 non-null  float64
 9   NumOfProducts     10000 non-null  int64
 10  HasCrCard         10000 non-null  int64
 11  IsActiveMember    10000 non-null  int64
 12  EstimatedSalary   10000 non-null  float64
 13  Exited            10000 non-null  int64
 dtypes: float64(2), int64(9), object(3)
```

Data Analysis & Modelling



Strategic Initiatives



Stage 3

Exploratory Data Analysis & Visualization



DigitalSkola

Check for Missing Value

Dengan menggunakan fungsi `.info()`, akan diketahui gambaran data yang diteliti. Dari hasil di samping terlihat semua kolom berisi **10000 non-null data** yang mana dapat diartikan tidak ada data kosong atau **no missing value**, sehingga tidak perlu dilakukan *handling missing value*.

RangeIndex: 10000 entries, 0 to 9999			
Data columns (total 14 columns):			
#	Column	Non-Null Count	Dtype
---	-----	-----	-----
0	RowNumber	10000 non-null	int64
1	CustomerId	10000 non-null	int64
2	Surname	10000 non-null	object
3	CreditScore	10000 non-null	int64
4	Geography	10000 non-null	object
5	Gender	10000 non-null	object
6	Age	10000 non-null	int64
7	Tenure	10000 non-null	int64
8	Balance	10000 non-null	float64
9	NumOfProducts	10000 non-null	int64
10	HasCrCard	10000 non-null	int64
11	IsActiveMember	10000 non-null	int64
12	EstimatedSalary	10000 non-null	float64
13	Exited	10000 non-null	int64

dtypes: float64(2), int64(9), object(3)

Descriptive Statistic

Statistika deskriptif dilakukan dengan menggunakan fungsi **.describe()**. Disini akan ditampilkan **rata-rata**, **minimum**, dan **maksimum** tiap kolom yang bertipe data numerik. Berikut beberapa kesimpulan dari statistika deskriptif.

Umur pelanggan rata-rata 39 tahun dengan usia termuda 18 tahun dan tertua 92 tahun.

Gaji pelanggan rata-rata \$100,090 dengan gaji terendah \$11 dan tertinggi \$199,992.

Sebanyak 71% pelanggan memiliki credit card.

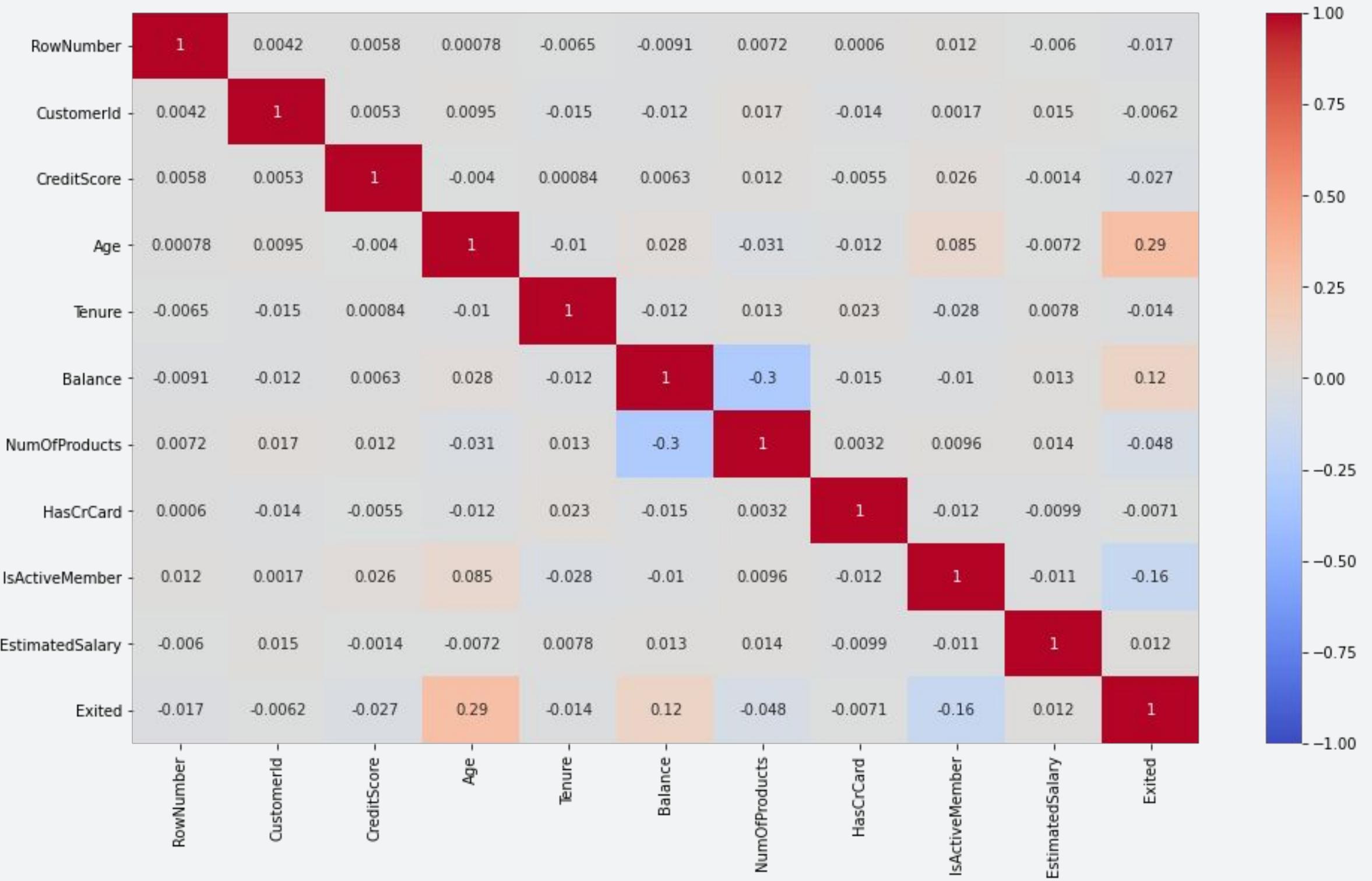
Sebanyak 20% pelanggan memilih untuk berhenti berlangganan.

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
mean	5000.5	1.569094e+07	650.5288	38.9218	5.0128	76485.889288	1.5302	0.7055	0.5151	100090.239881	0.2037
min	1.0	1.556570e+07	350.0000	18.0000	0.0000	0.000000	1.0000	0.0000	0.0000	11.580000	0.0000
max	10000.0	1.581569e+07	850.0000	92.0000	10.0000	250898.090000	4.0000	1.0000	1.0000	199992.480000	1.0000

Correlation

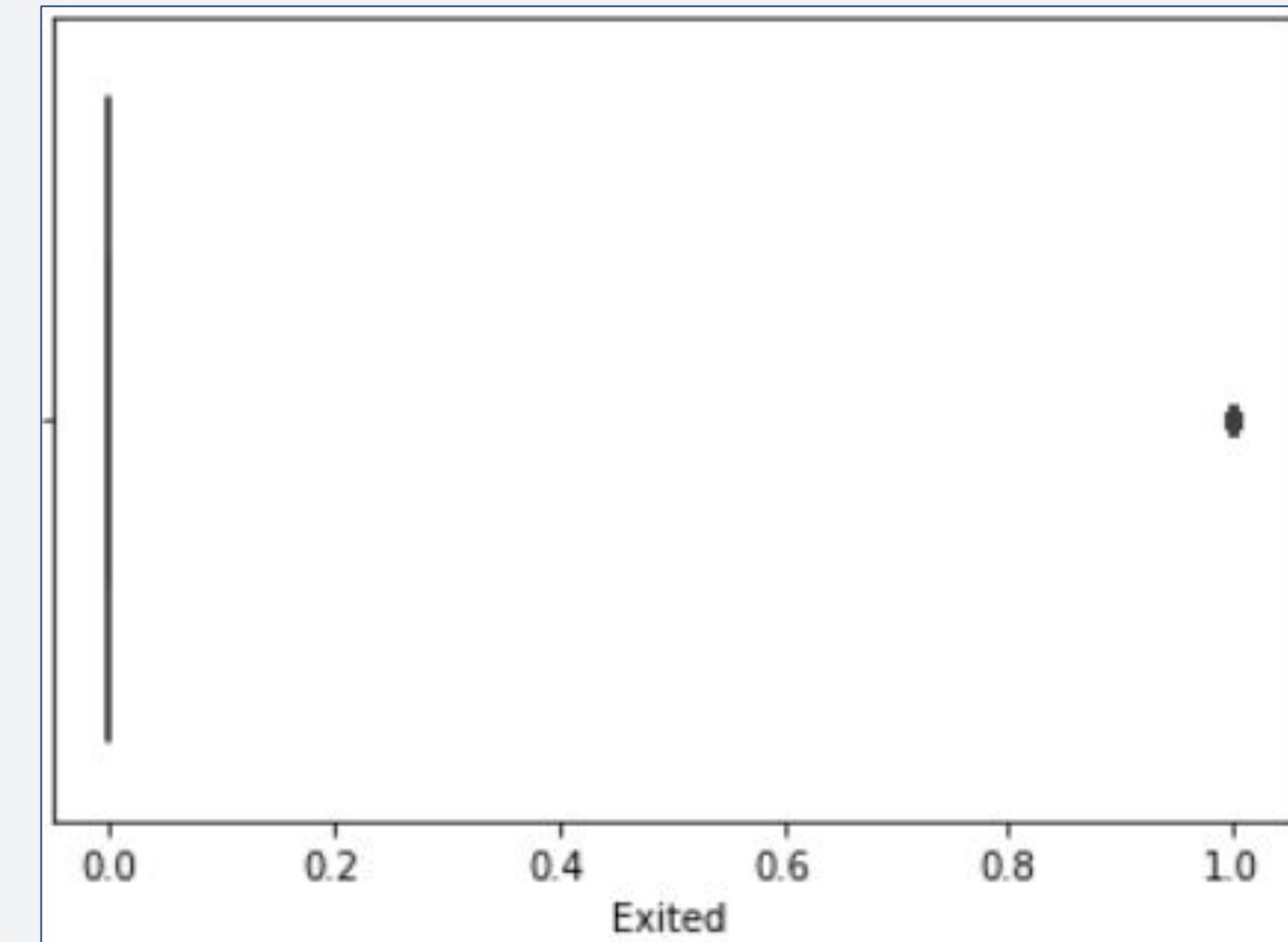
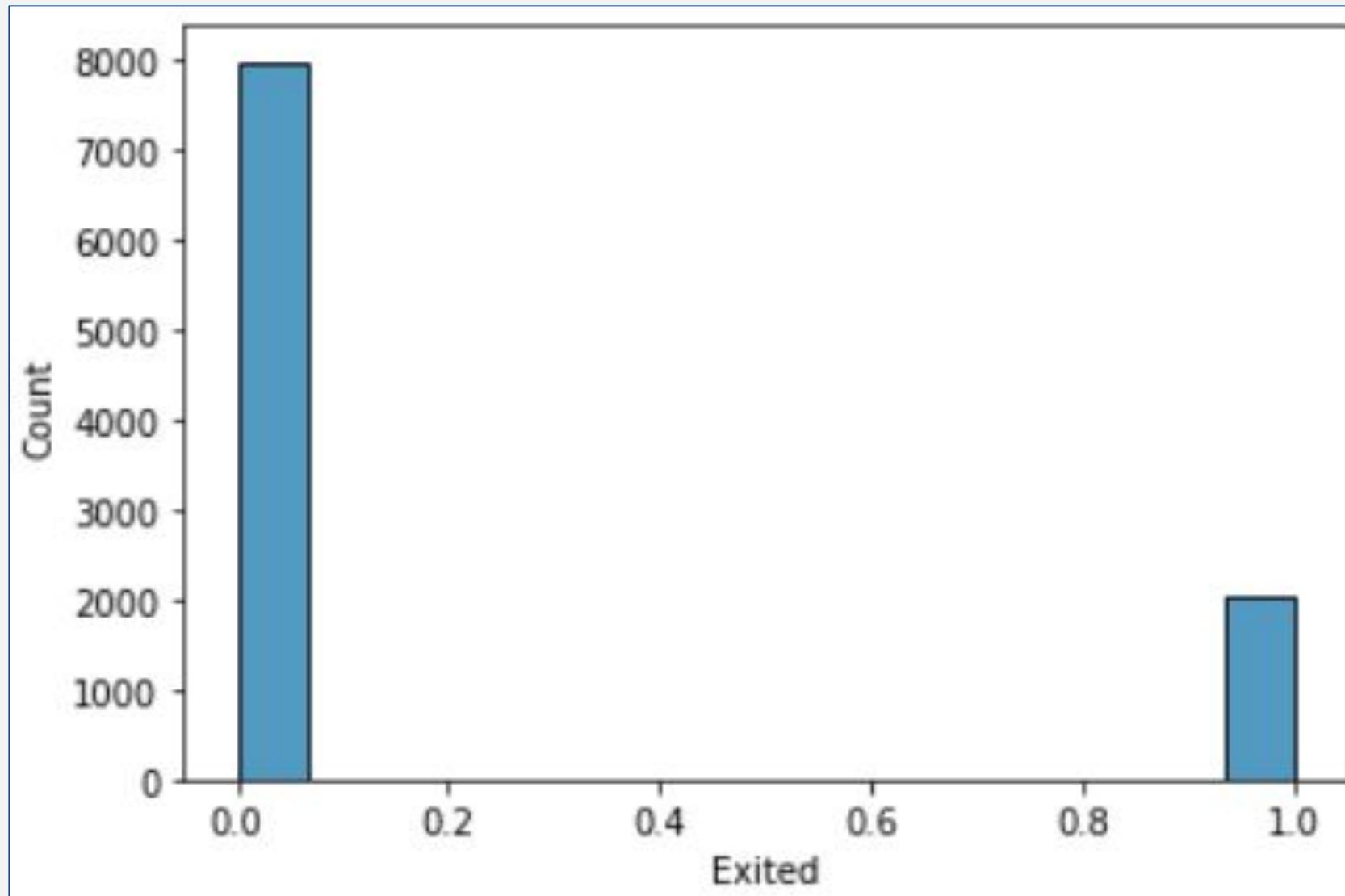
Dataset ini memiliki **jumlah data yang besar** dan memiliki ukuran parameter seperti **mean** dan **standar deviasi** populasi, serta **data numerik** yang **bertipe rasio**. Maka, korelasi yang cocok digunakan adalah **Korelasi Pearson**.

Dari hasil perhitungan korelasi di samping, **tidak ada kolom-kolom yang memiliki korelasi yang kuat**.



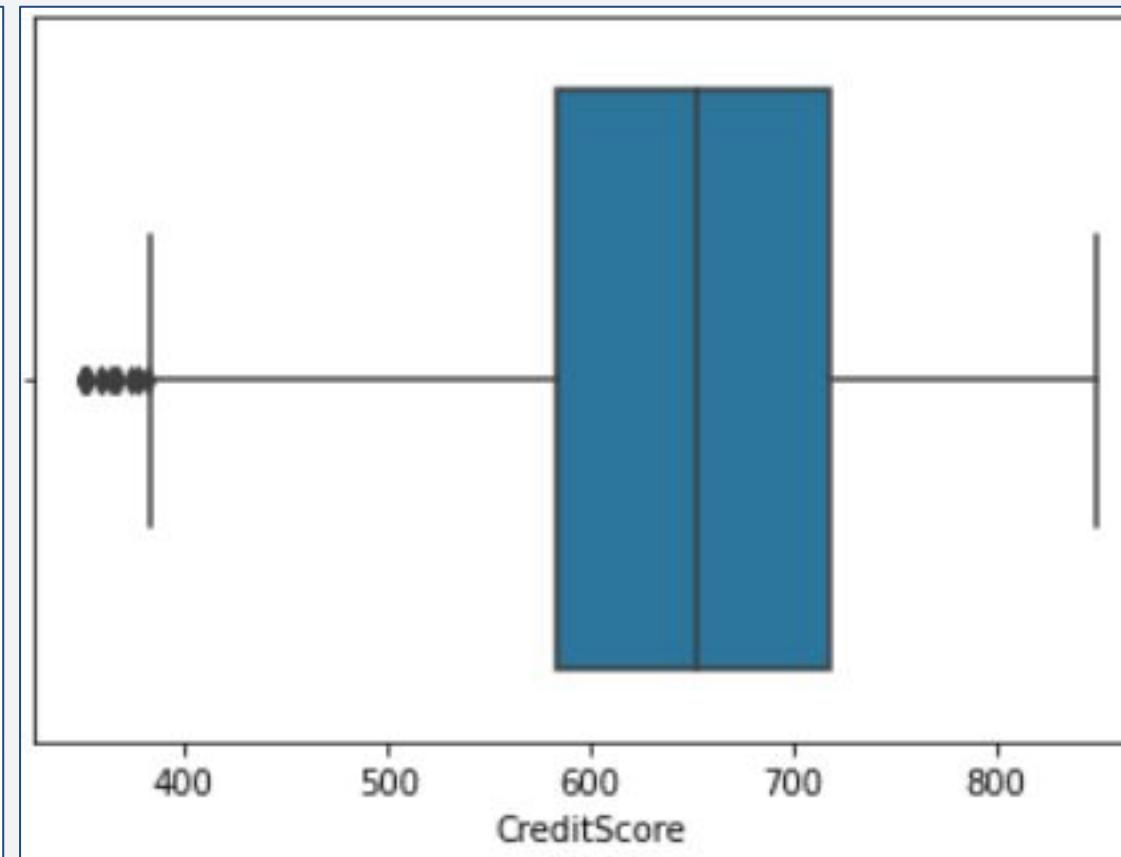
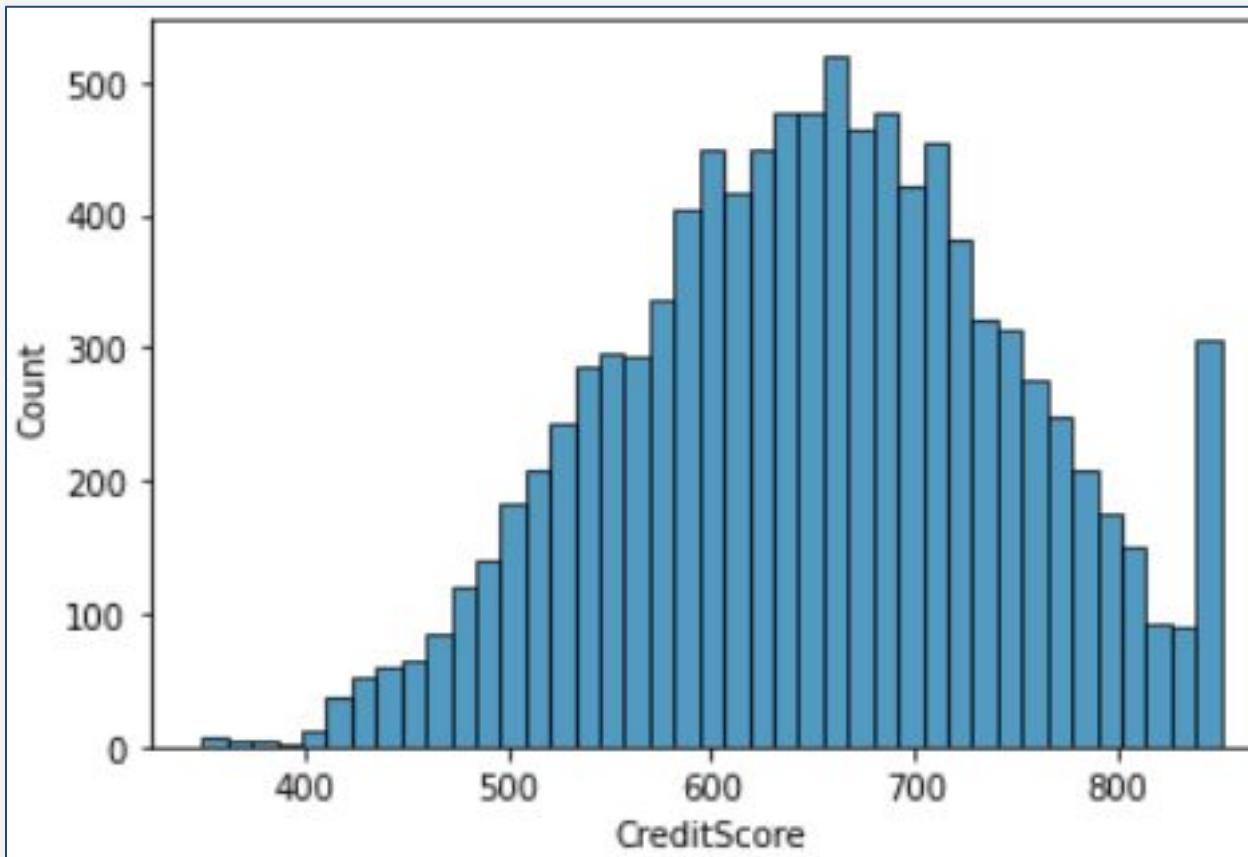
Distribution & Comparison

Target: Exited

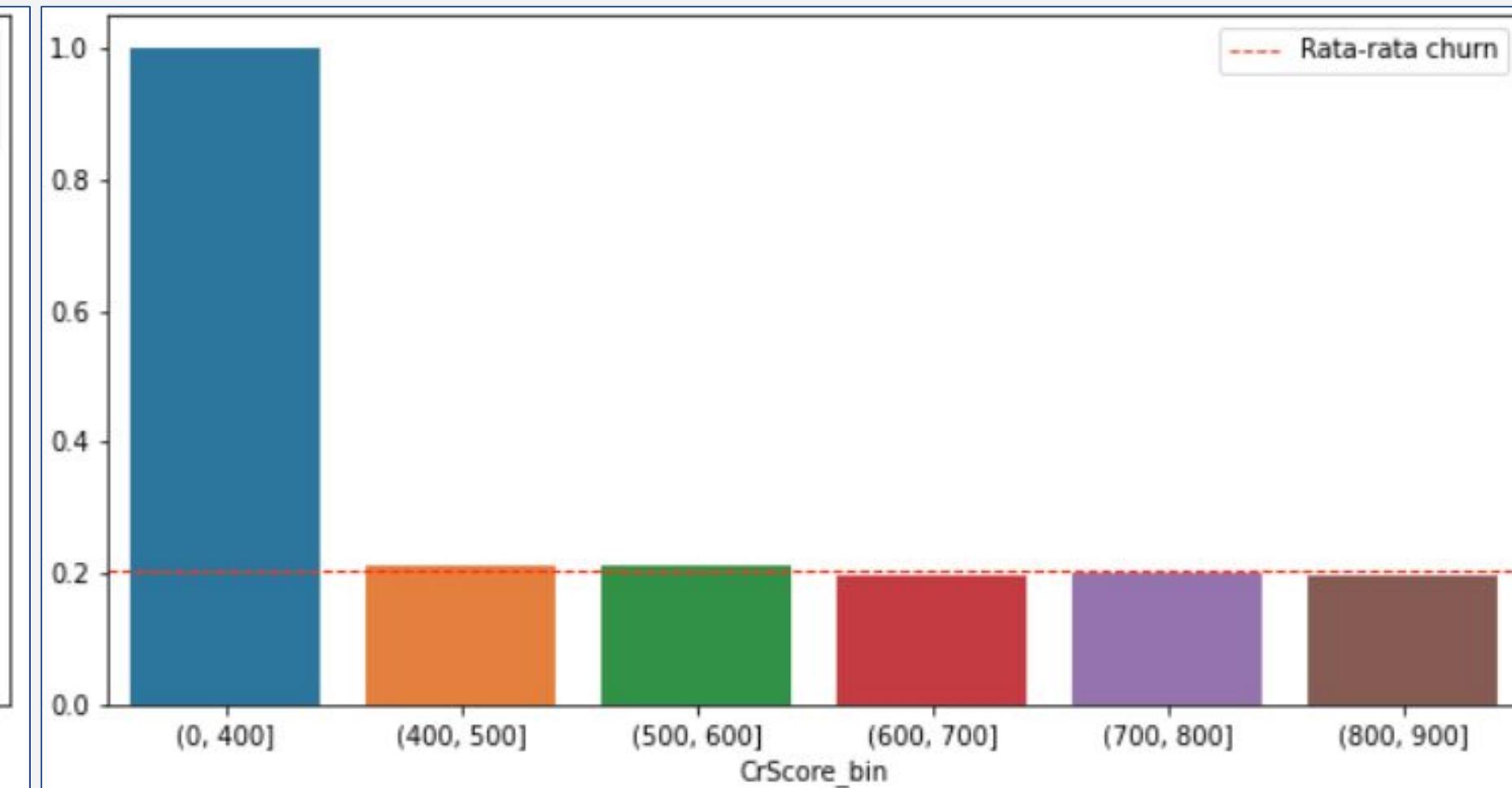
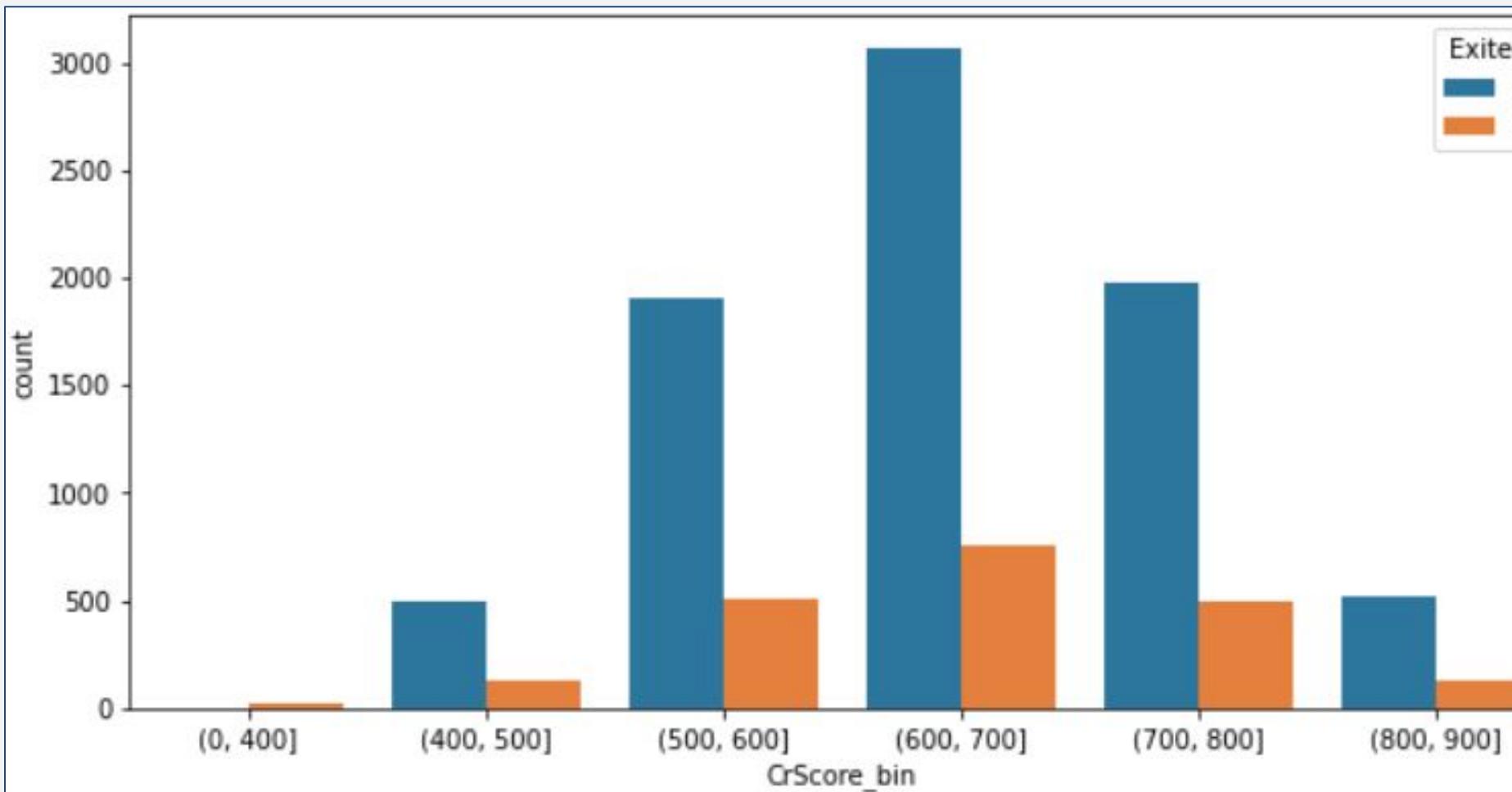


Jumlah pelanggan yang meninggalkan layanan sekitar 2000 orang, sedangkan yang tetap menggunakan layanan ada sekitar hampir 8000 orang.

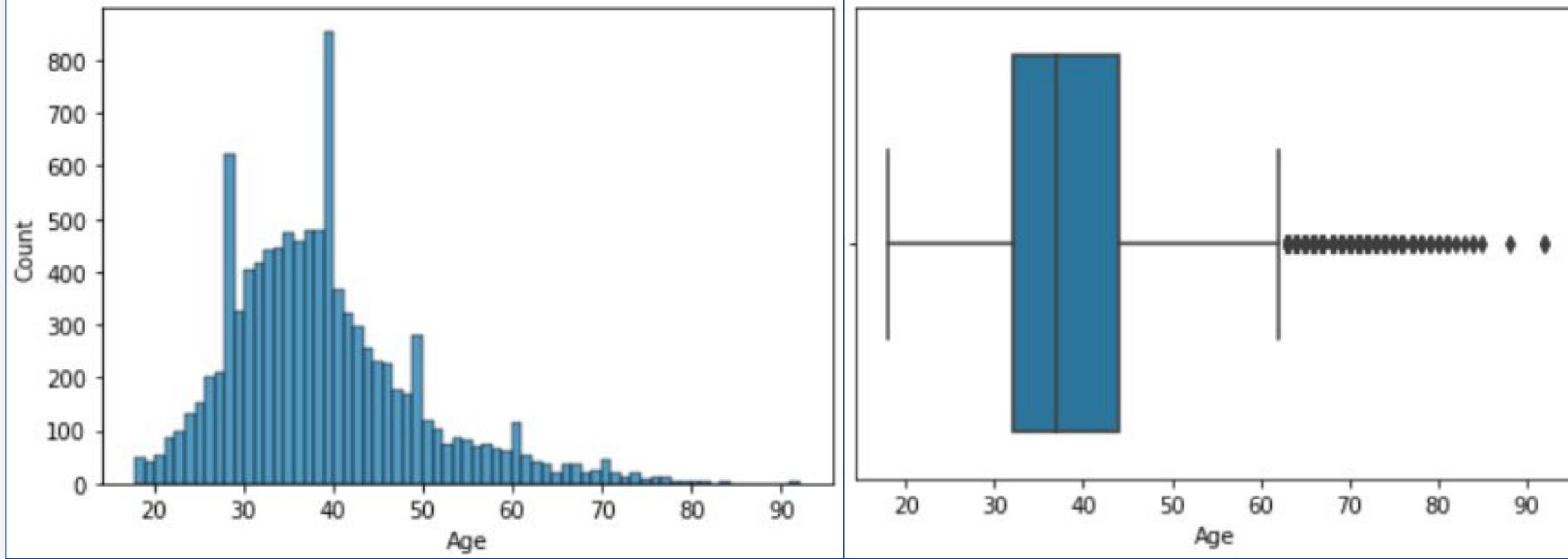
CreditScore



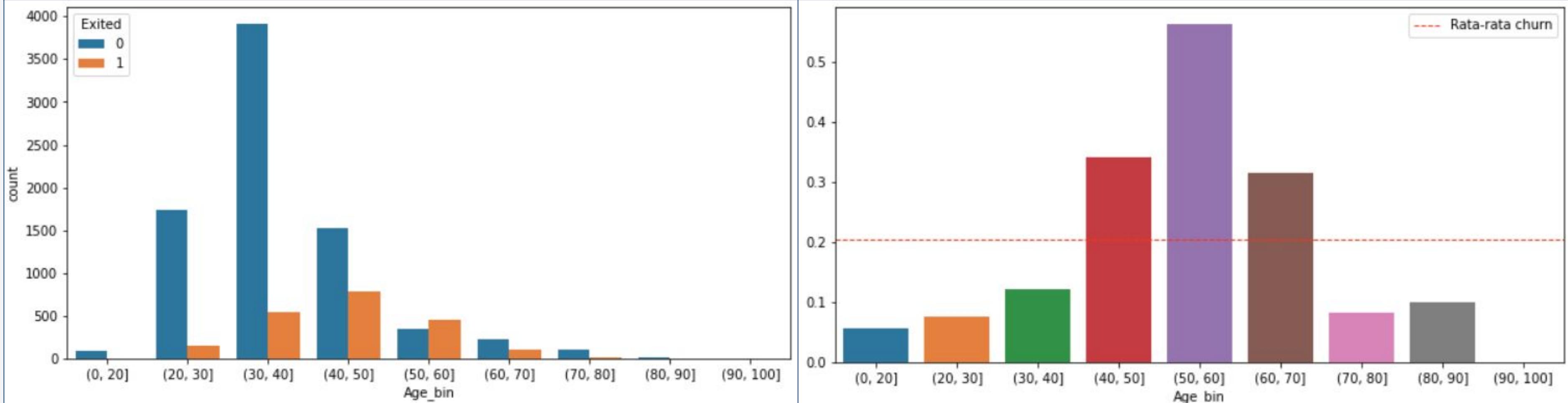
Secara umum data tersebar dan berkumpul di nilai Credit Score 600-700. Namun terdapat beberapa data yang jumlahnya sangat sedikit dibawah nilai 400. Terdapat beberapa outlier di bawah nilai 400.



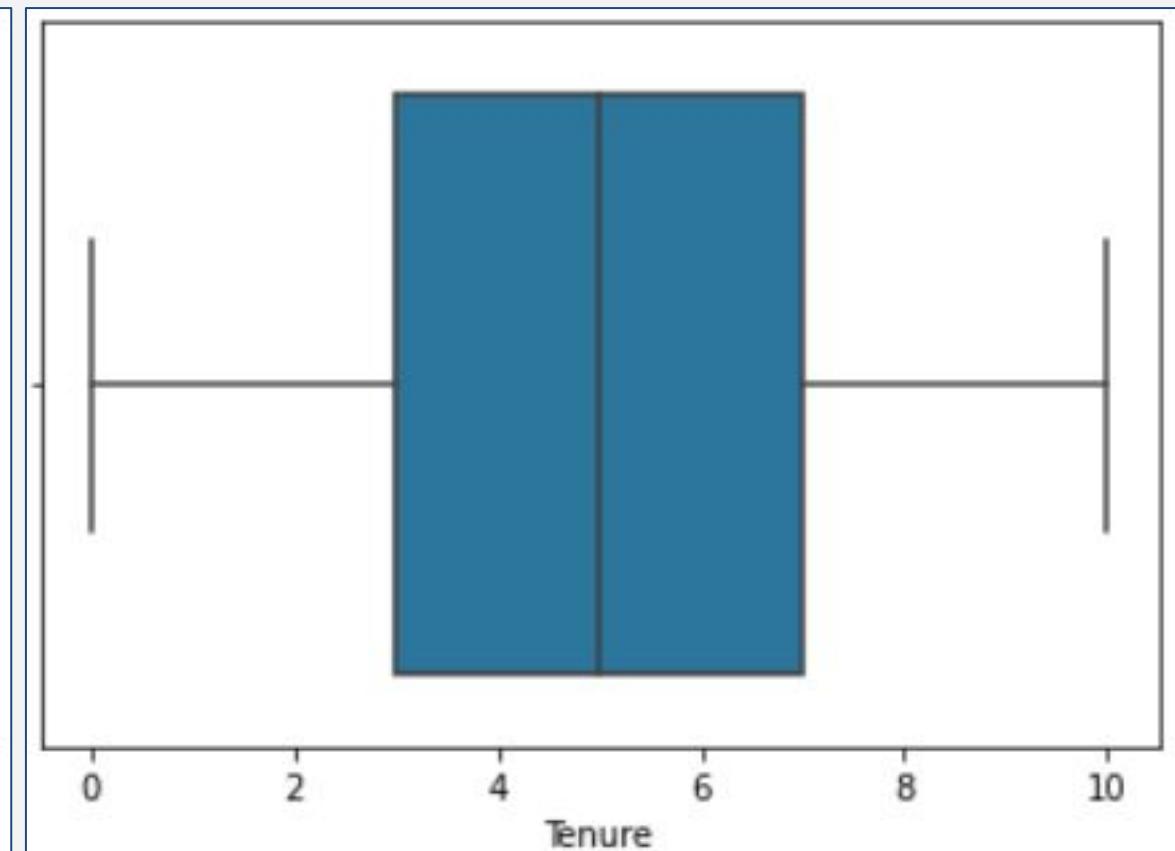
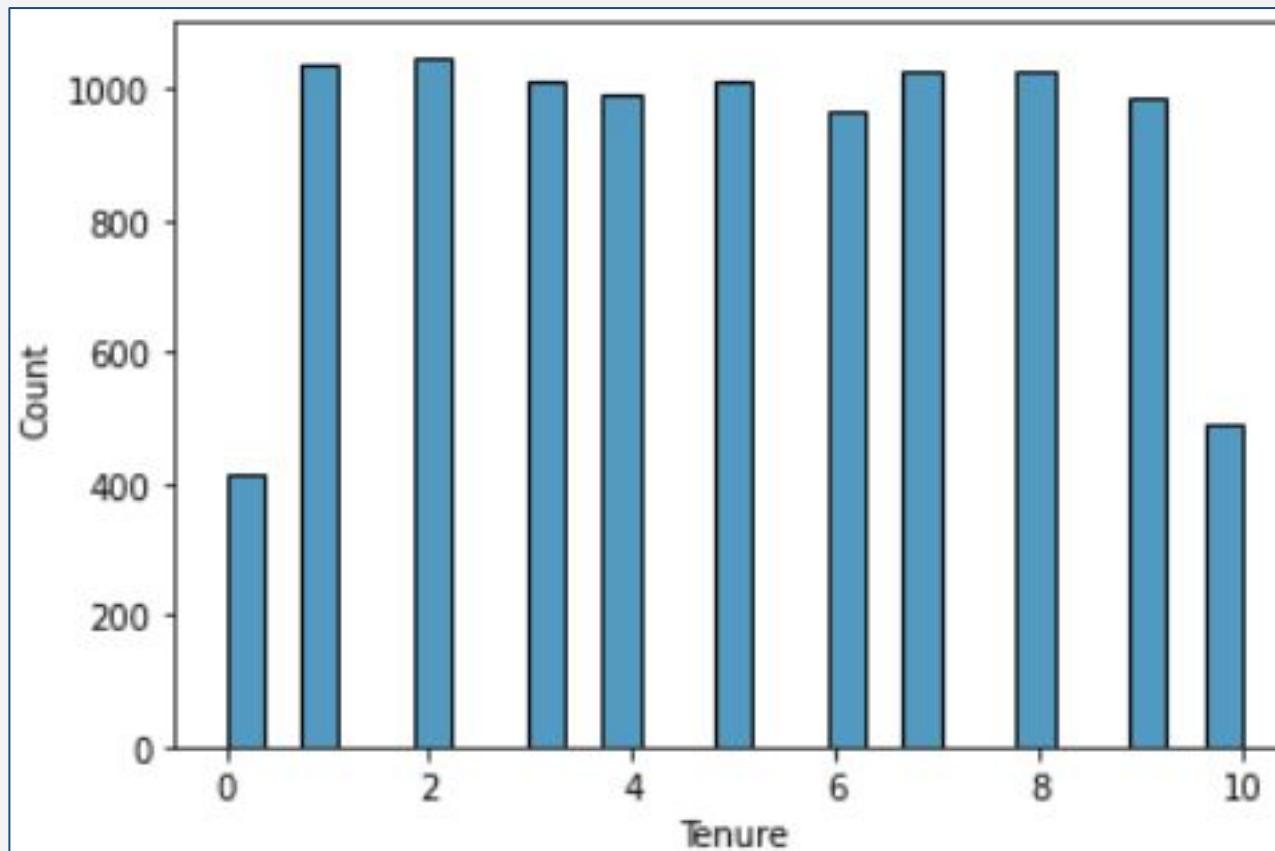
Age



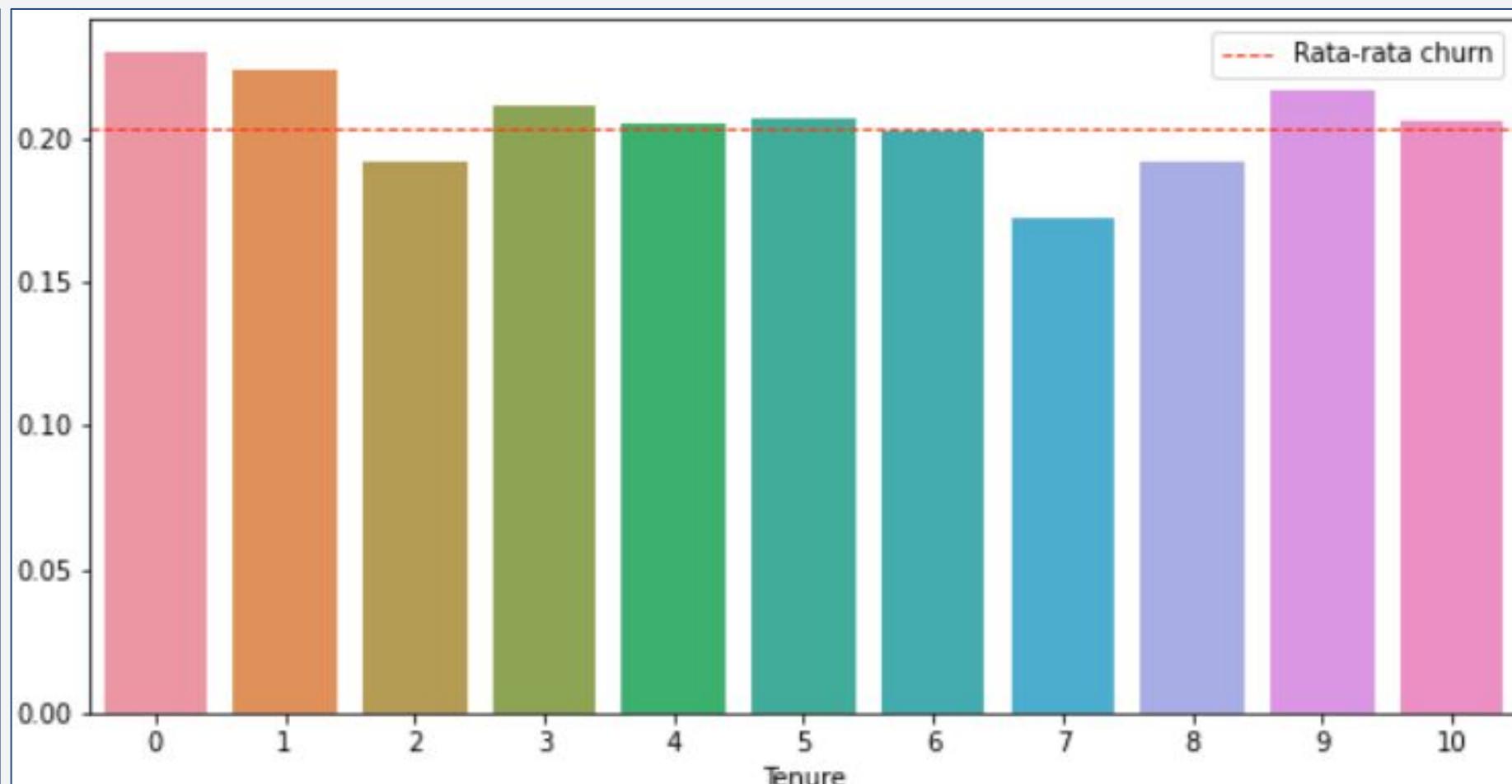
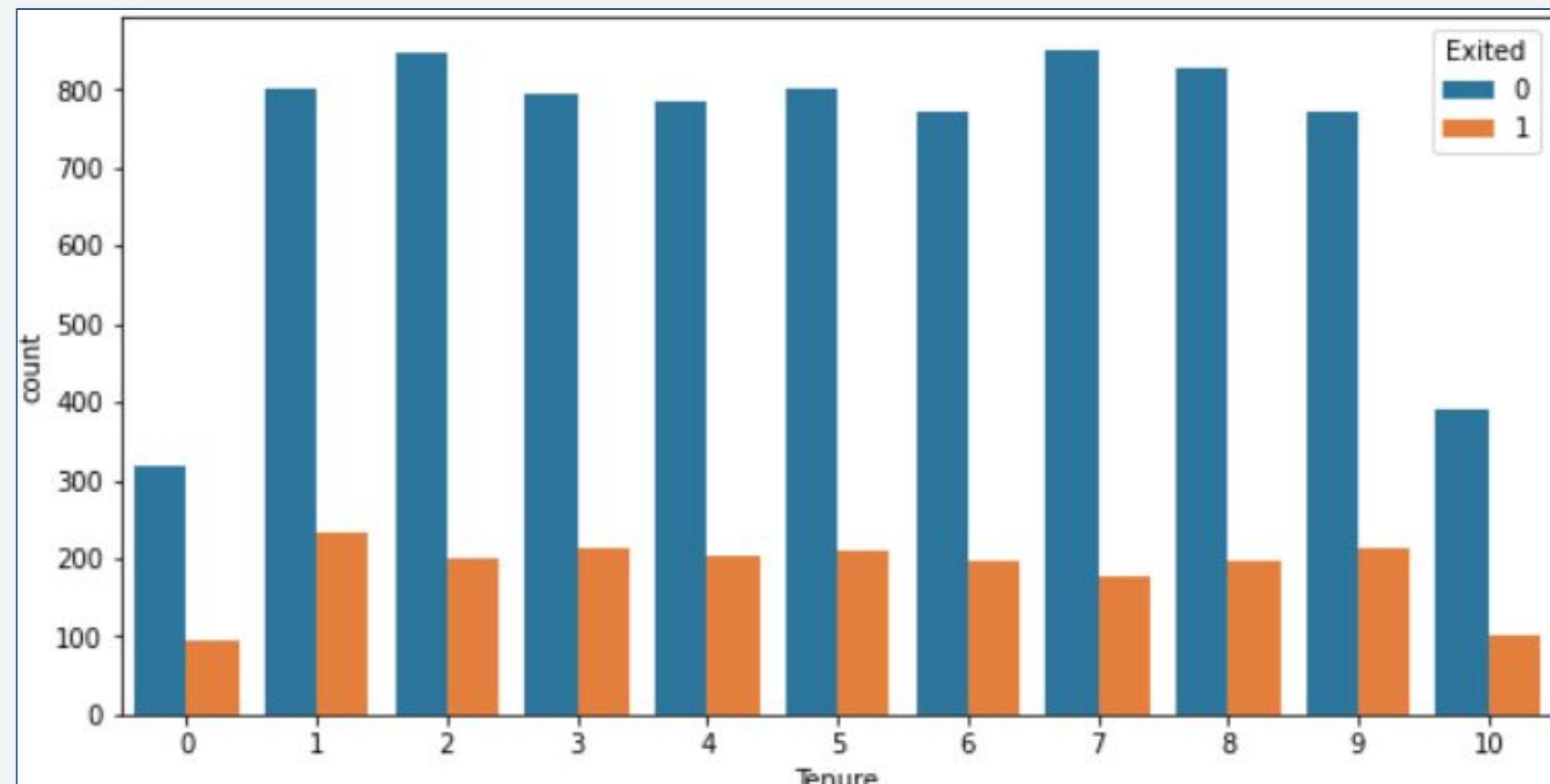
Rentang usia pelanggan cukup beragam, namun kebanyakan di rentang usia sekitar 30-45 tahun. Terlihat pada boxplot, bahwa terdapat banyak outlier yang berusia > 60 tahun.



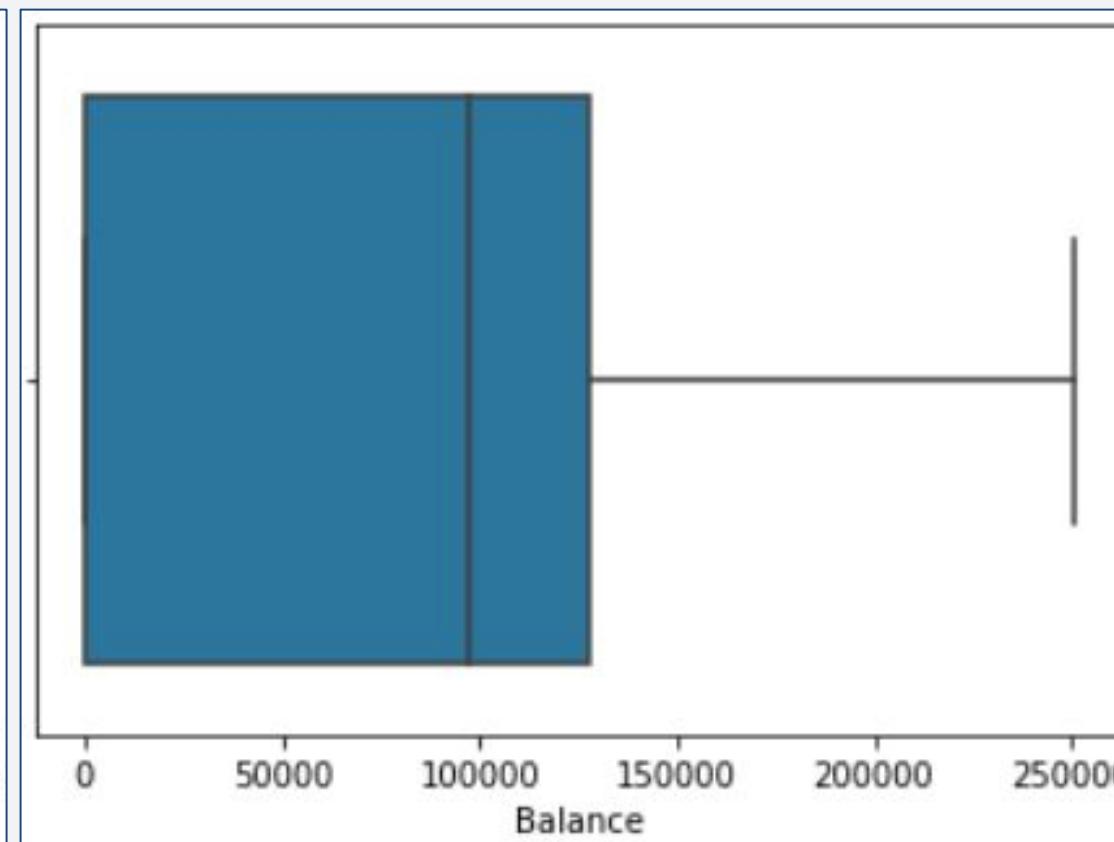
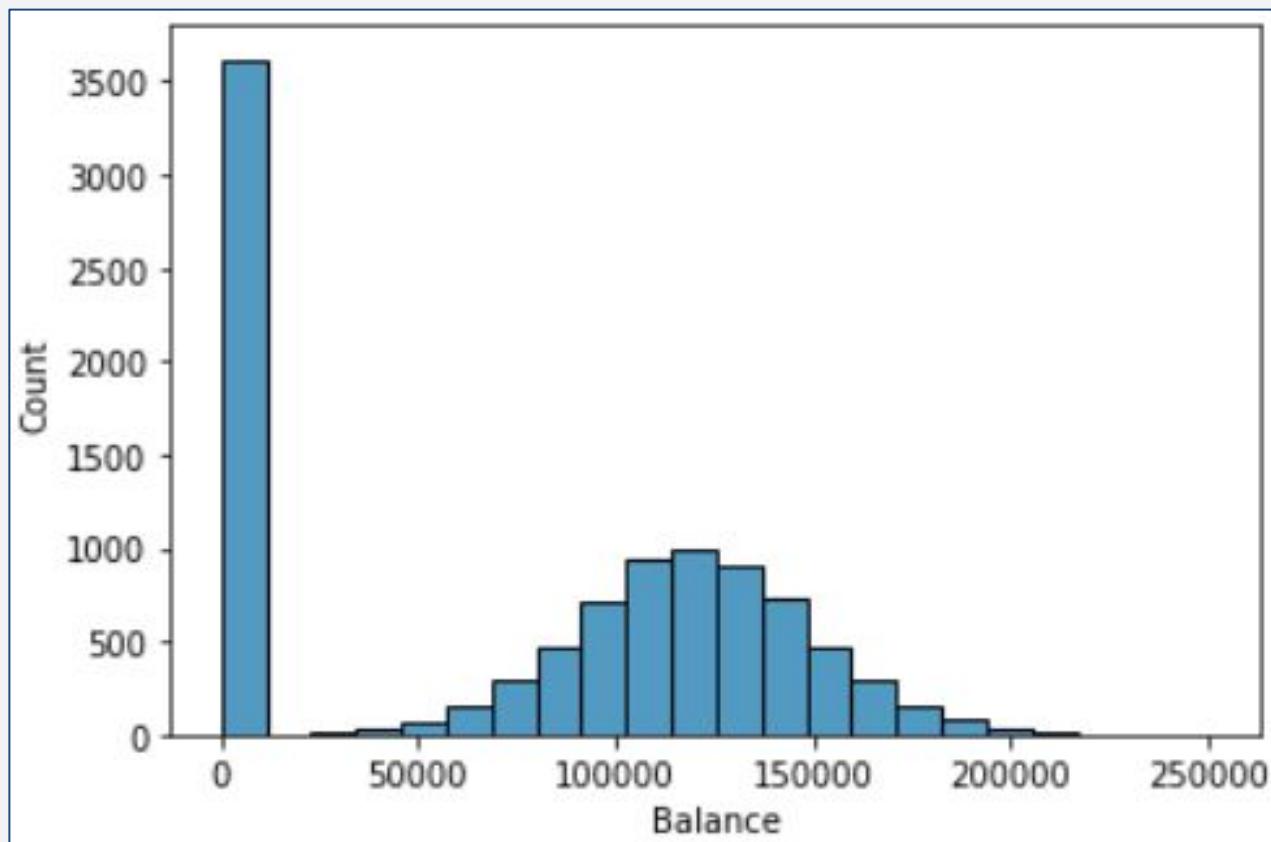
Tenure



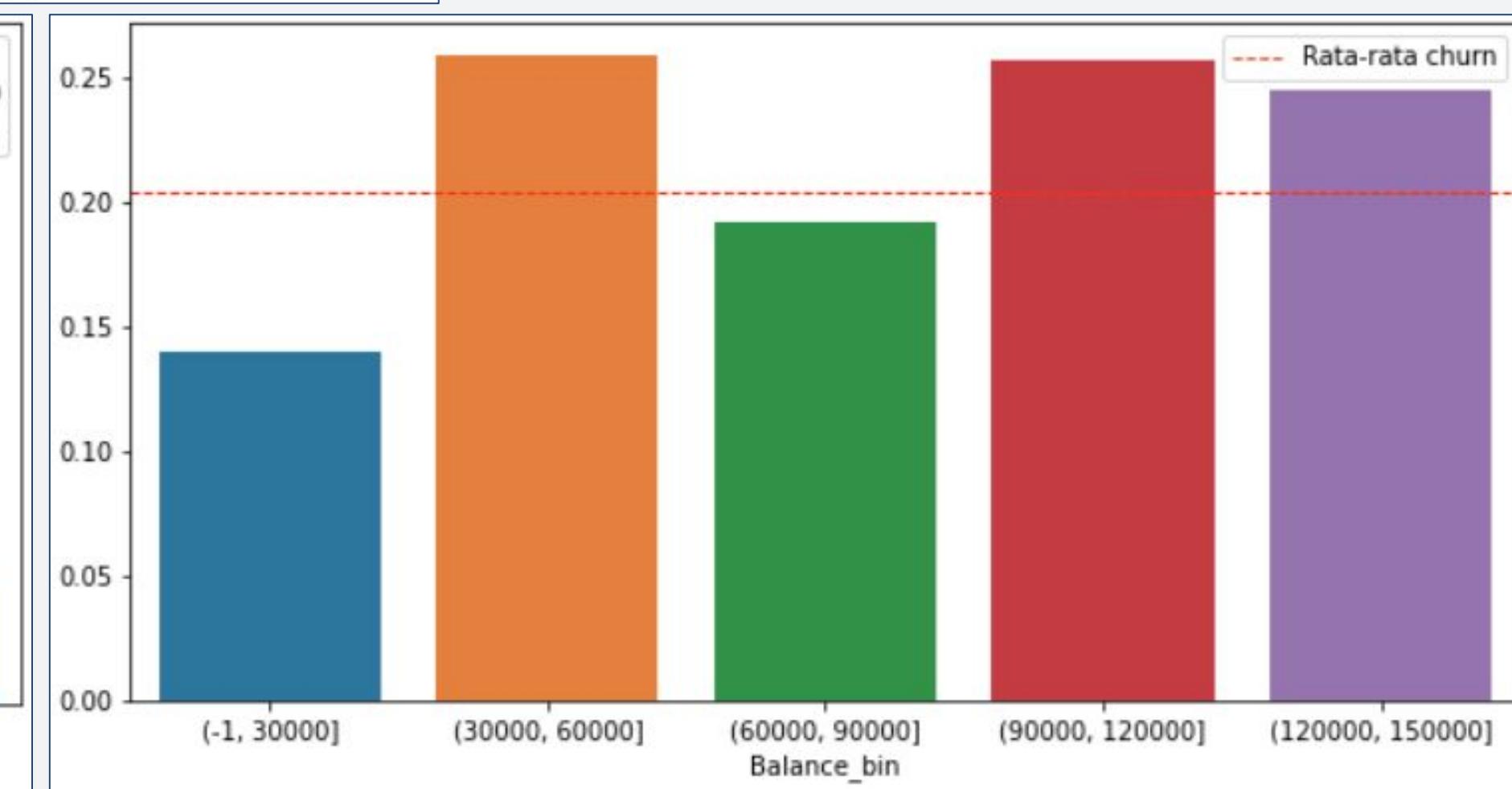
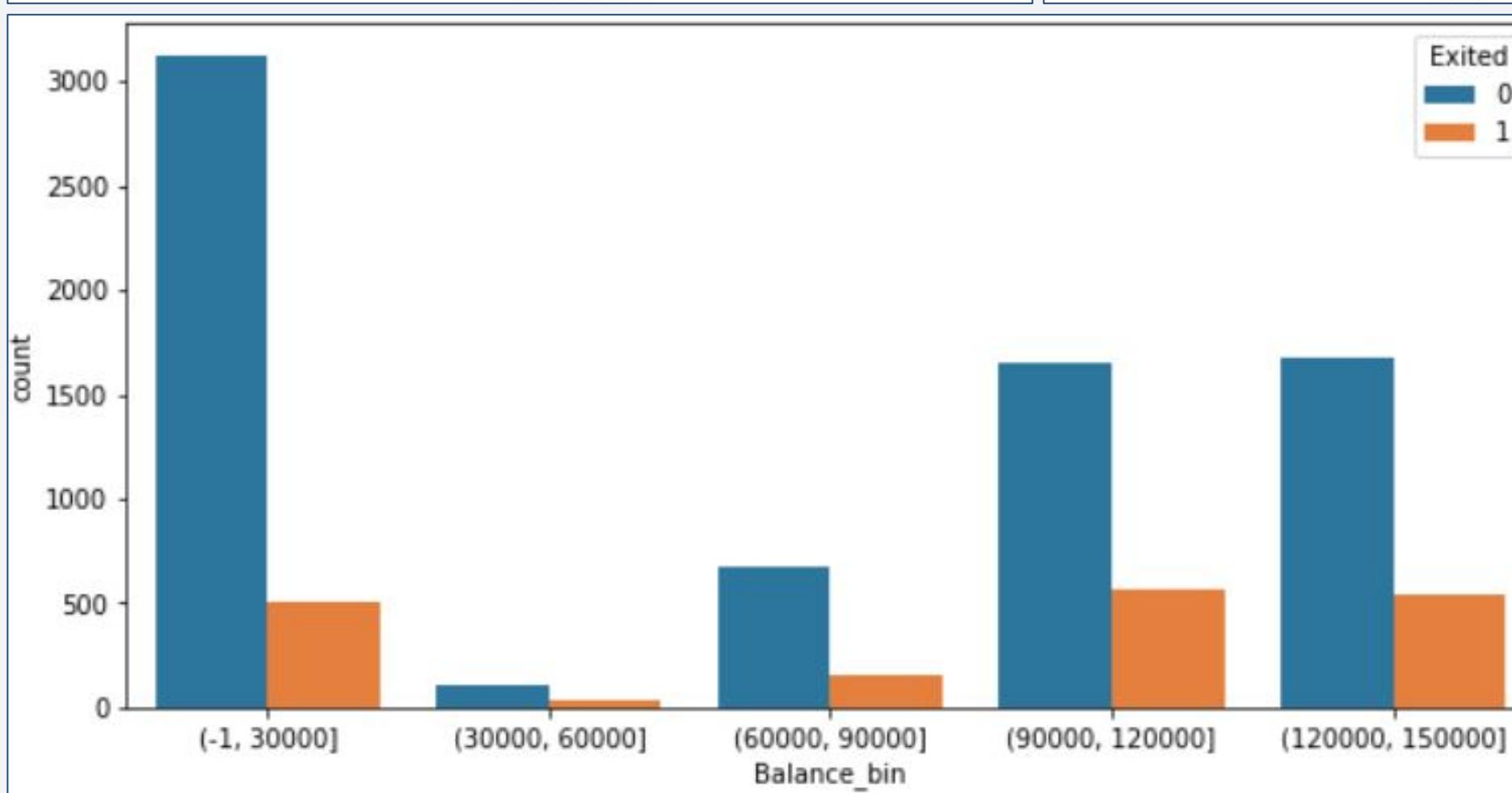
Secara umum data tersebar merata di masa berlanggan 0 - 10 tahun. tidak ditemukan outlier.



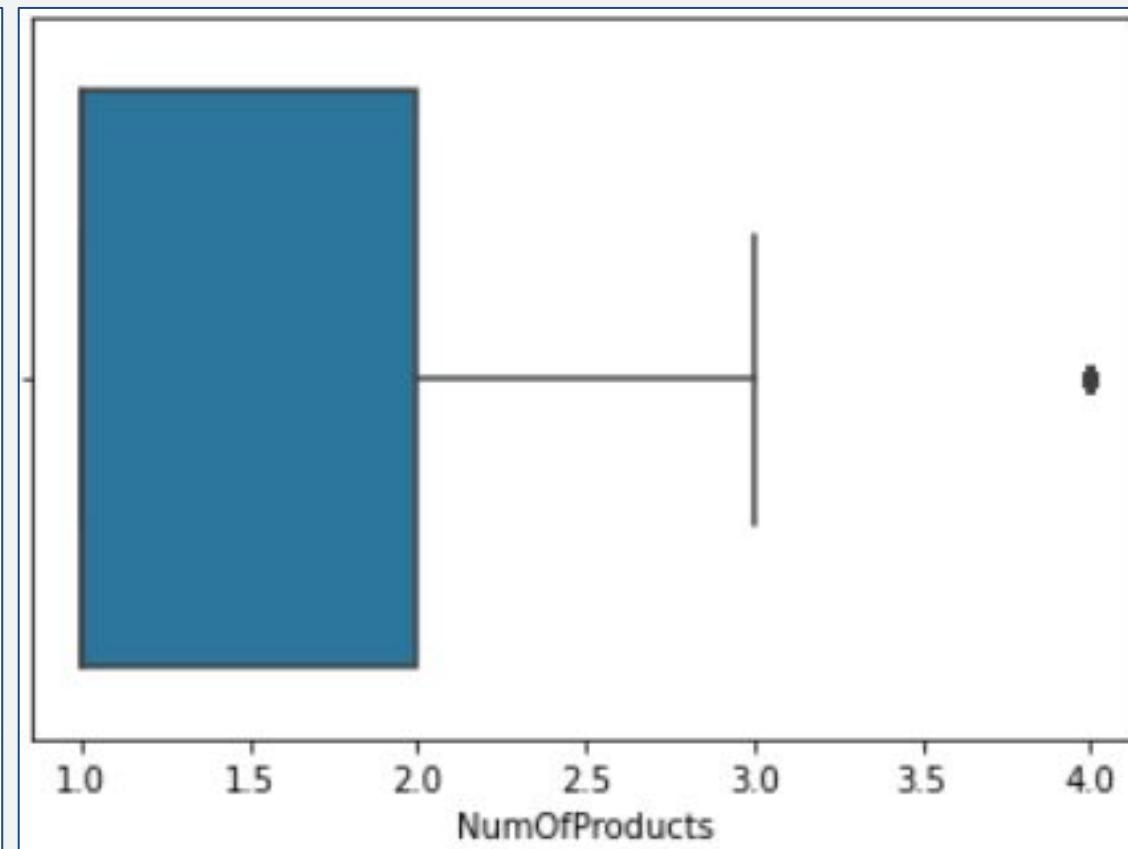
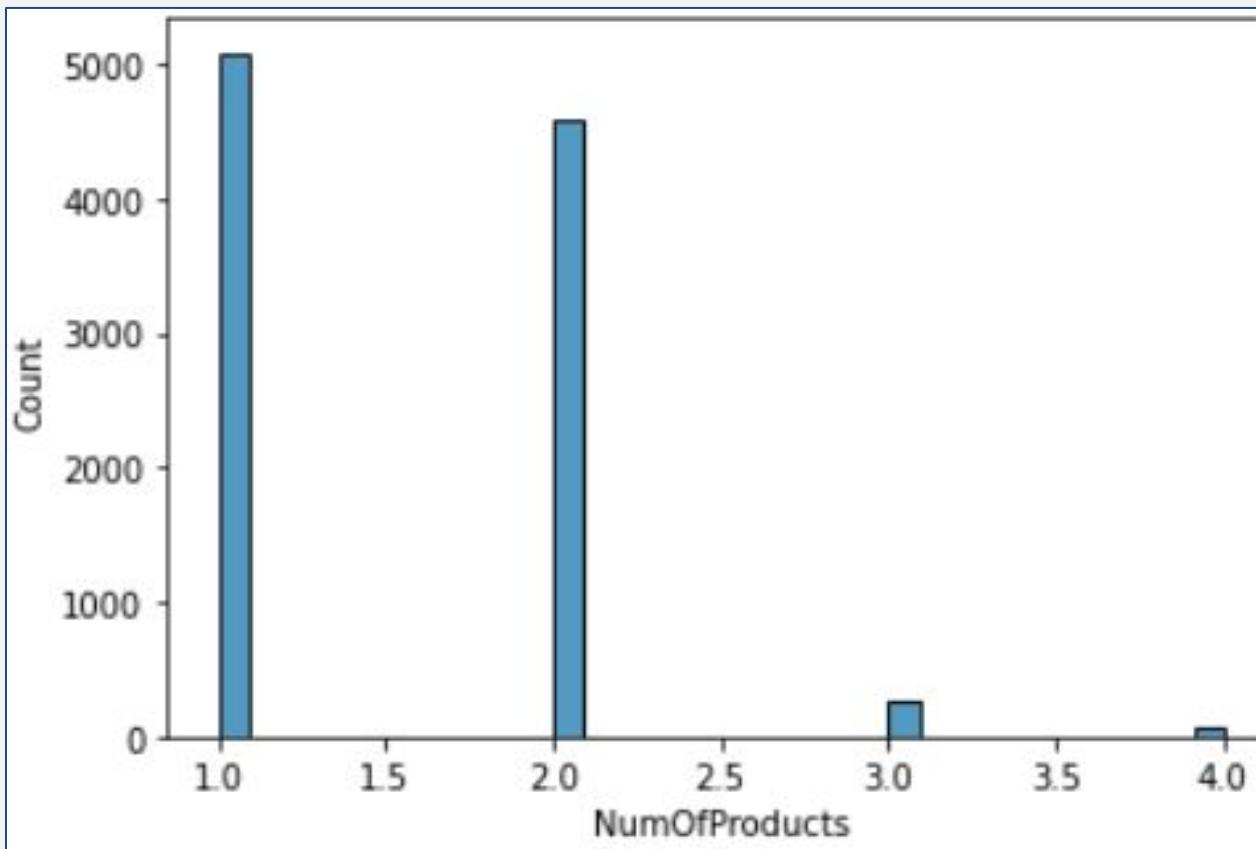
Balance



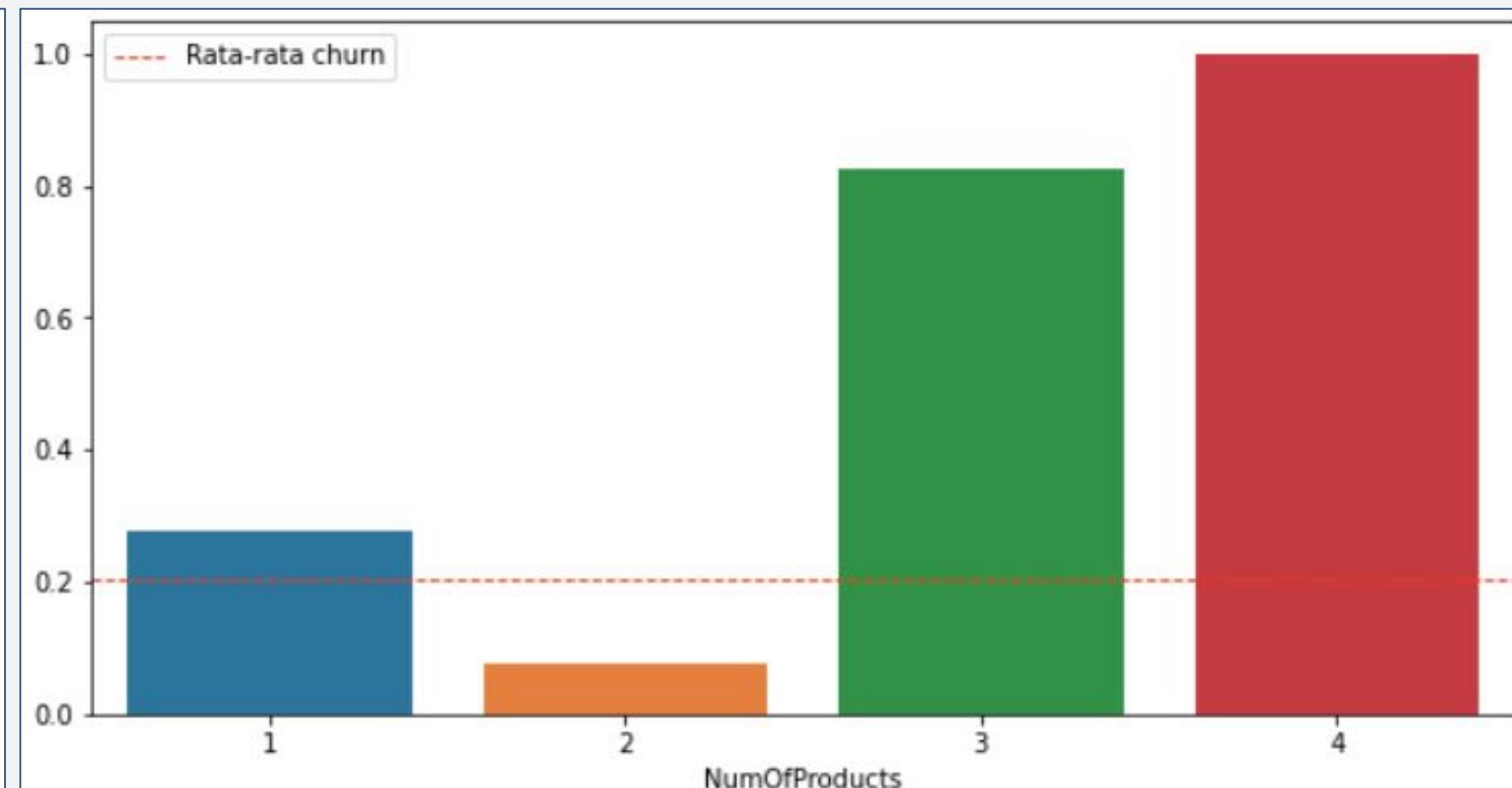
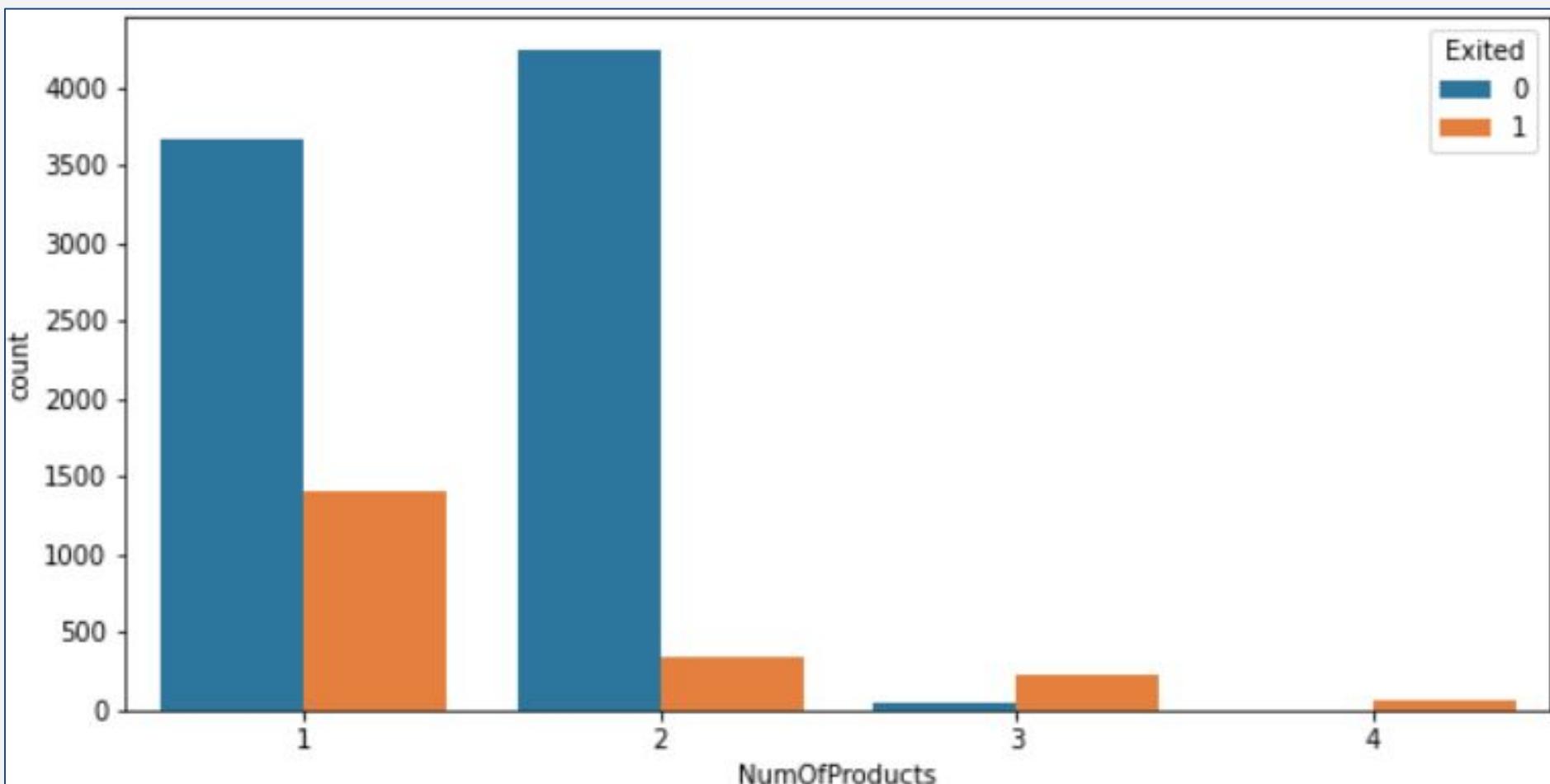
Pelanggan paling banyak tidak memiliki tabungan ($balance = 0$), lalu tersebar di rentang jumlah tabungan 100.000 sampai 150.000.



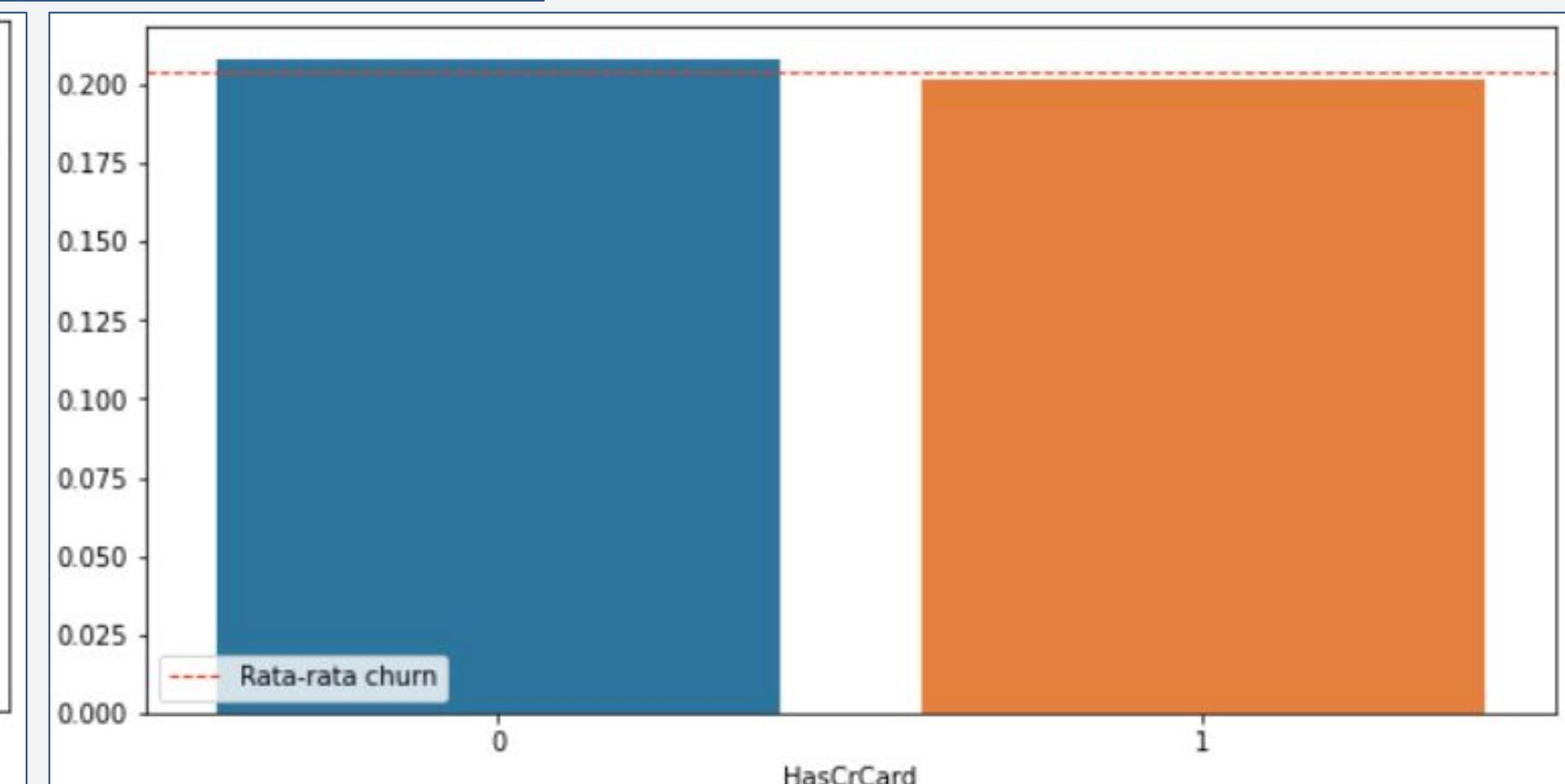
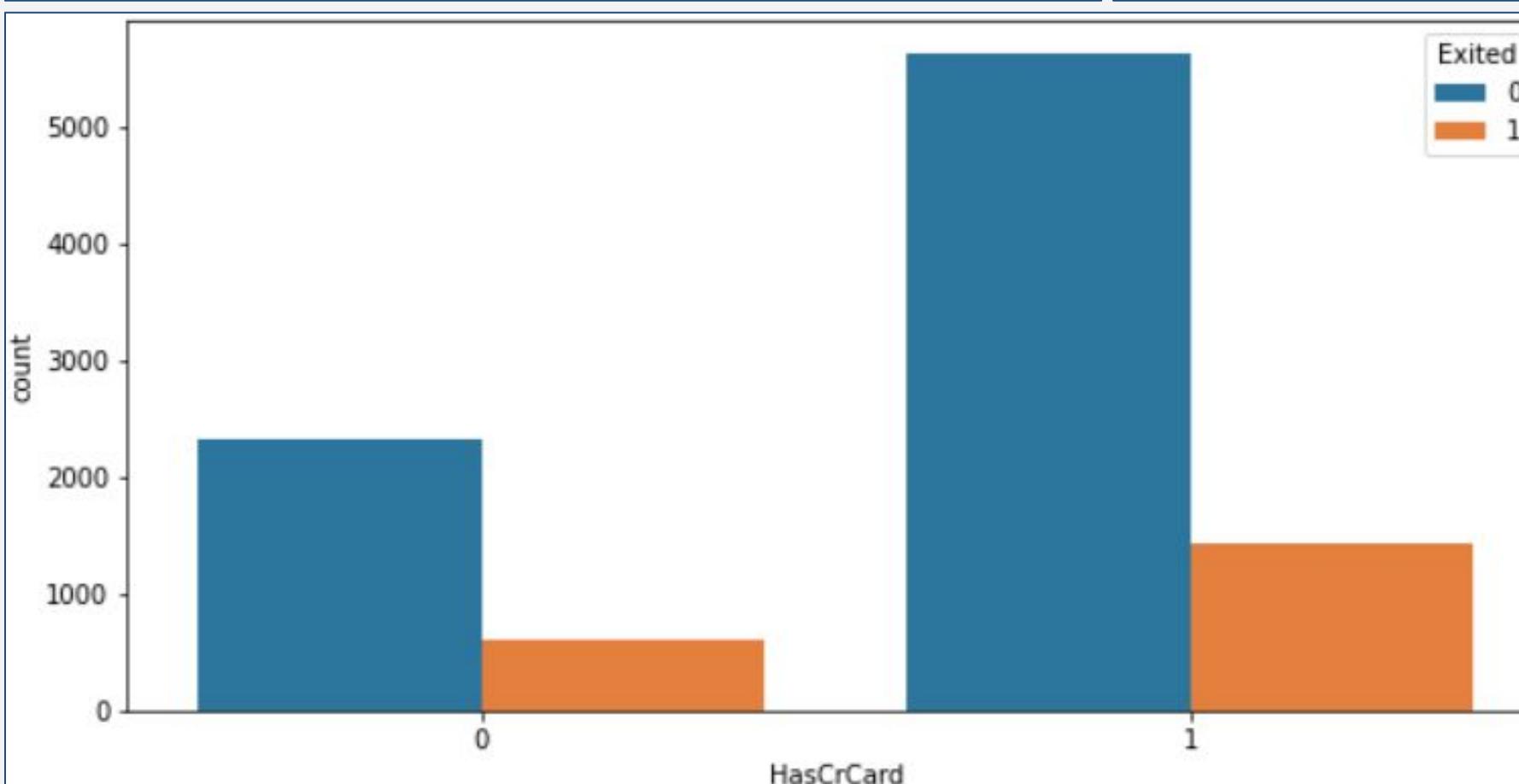
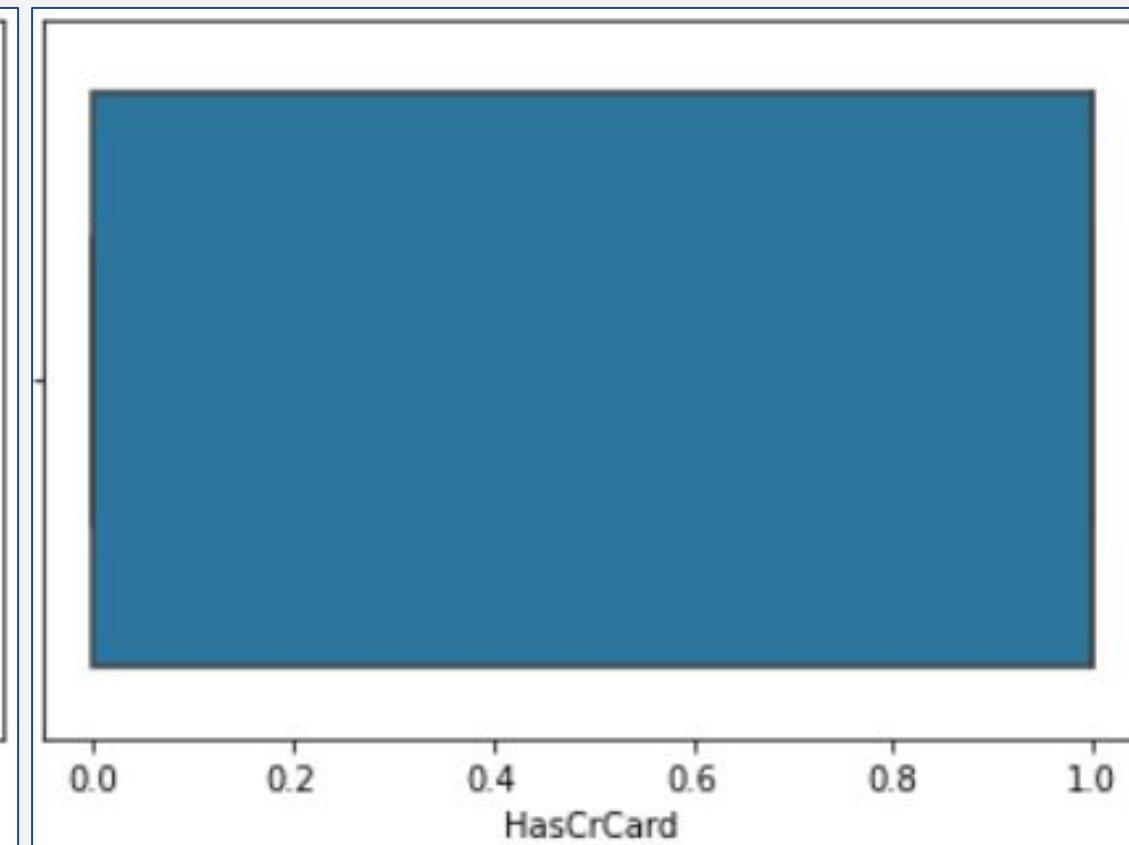
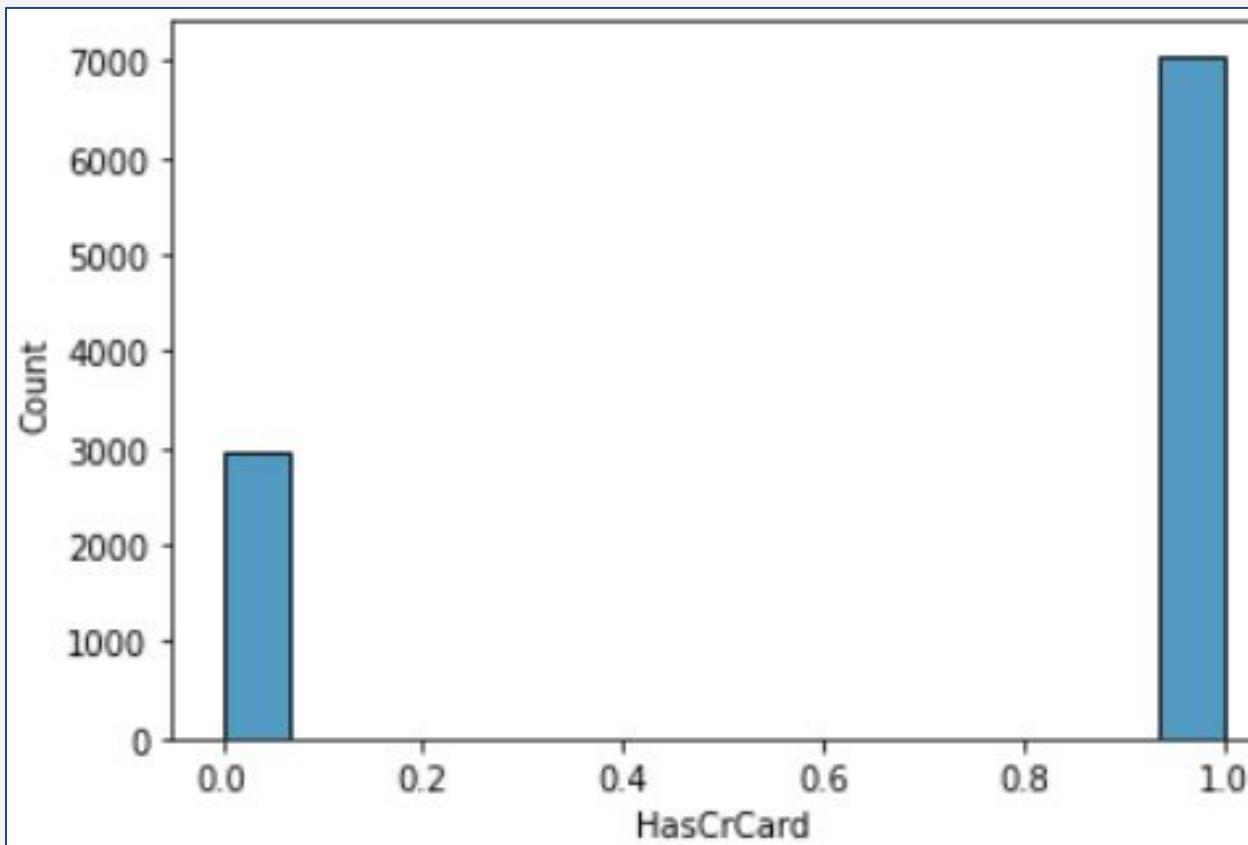
NumOfProducts



Pelanggan paling banyak memiliki 1 atau 2 produk. Terdapat data outlier yaitu pelanggan yang memiliki 4 produk langganan.

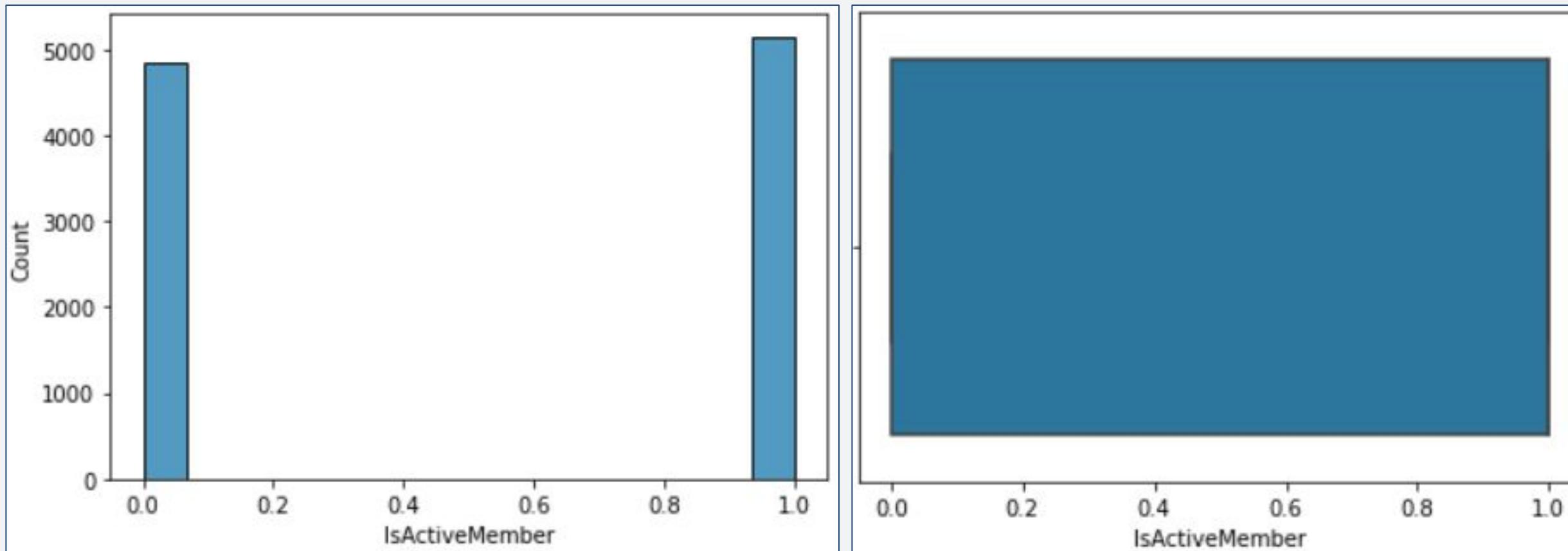


HasCrCard

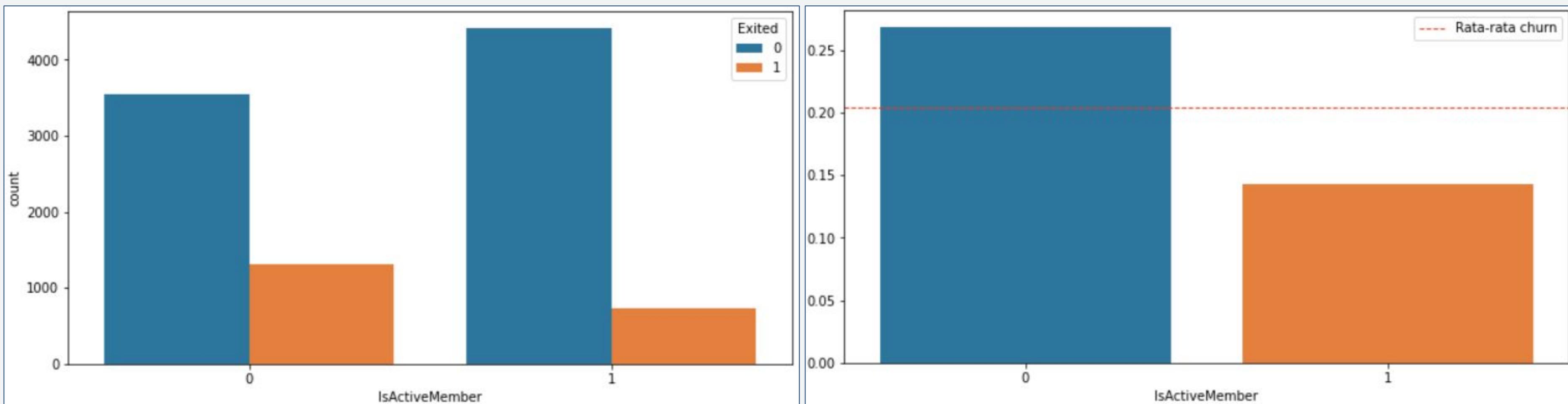


Sekitar 7000 pelanggan memiliki kartu kredit, sisanya hampir 3000 pelanggan tidak memiliki kartu kredit.

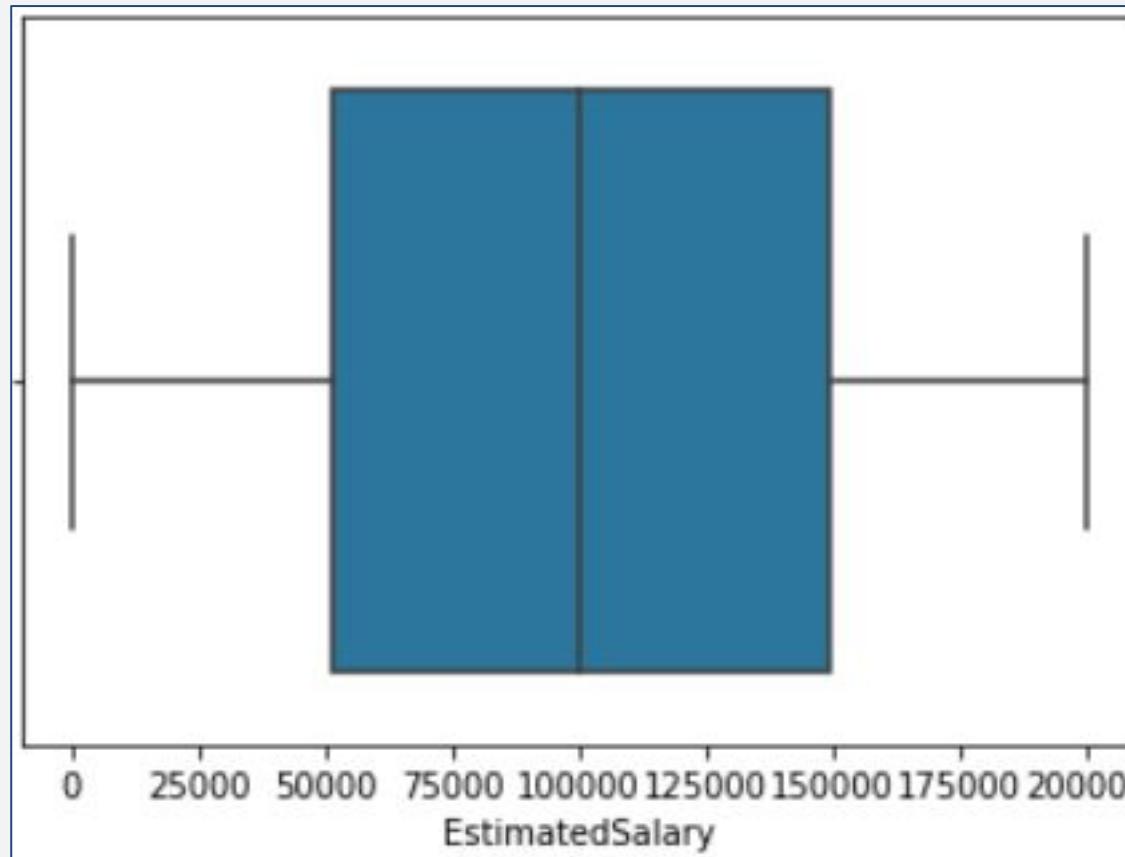
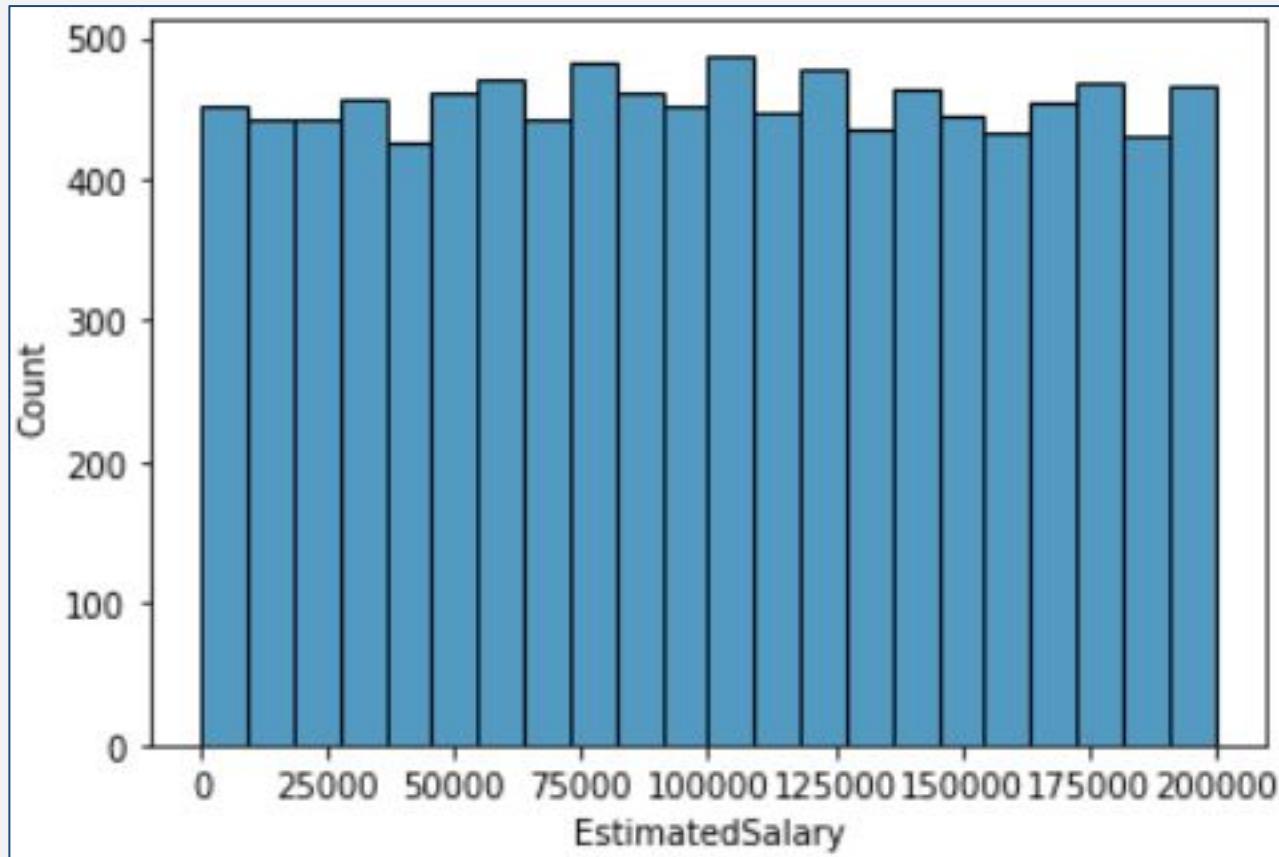
IsActiveMember



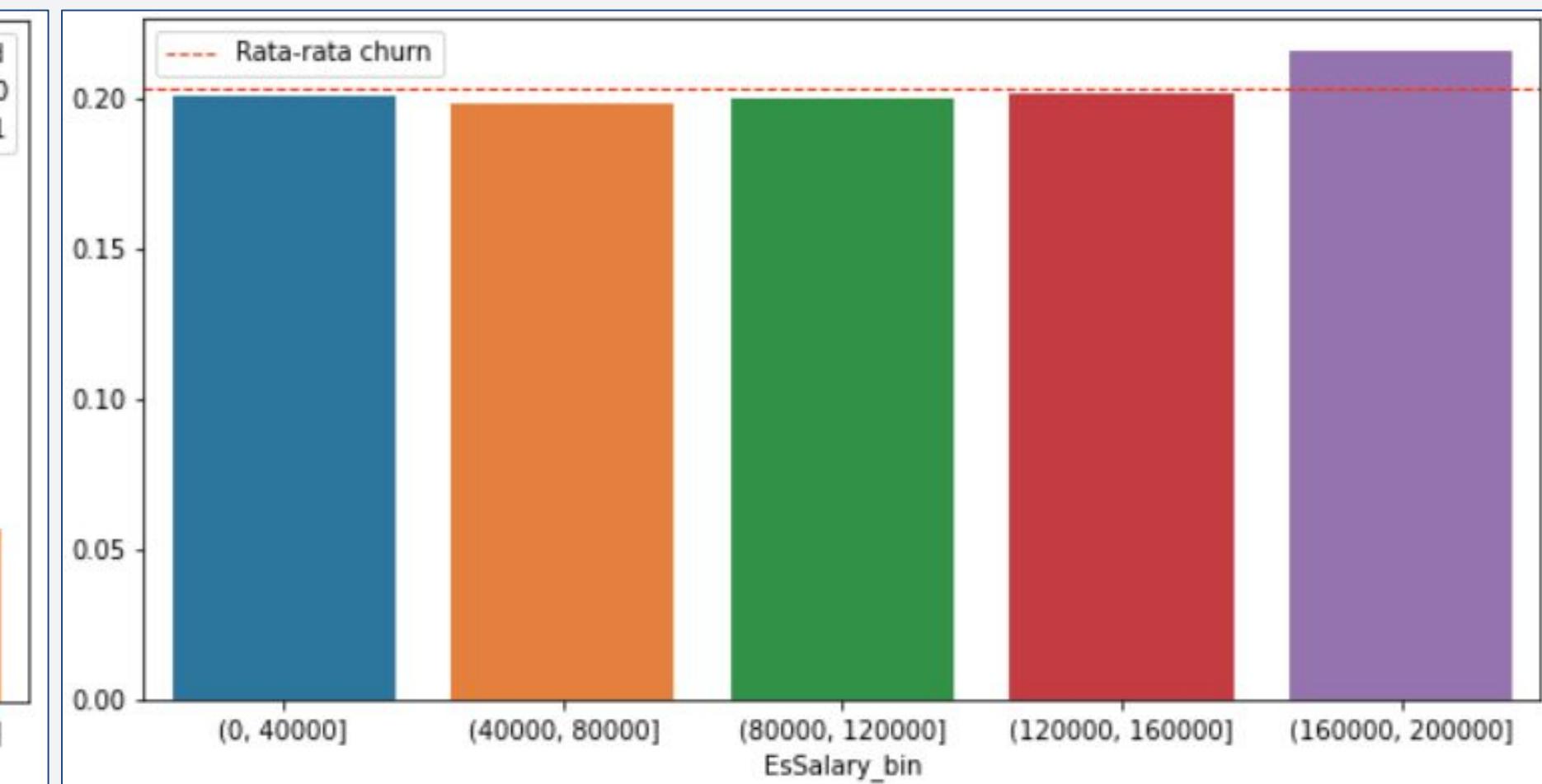
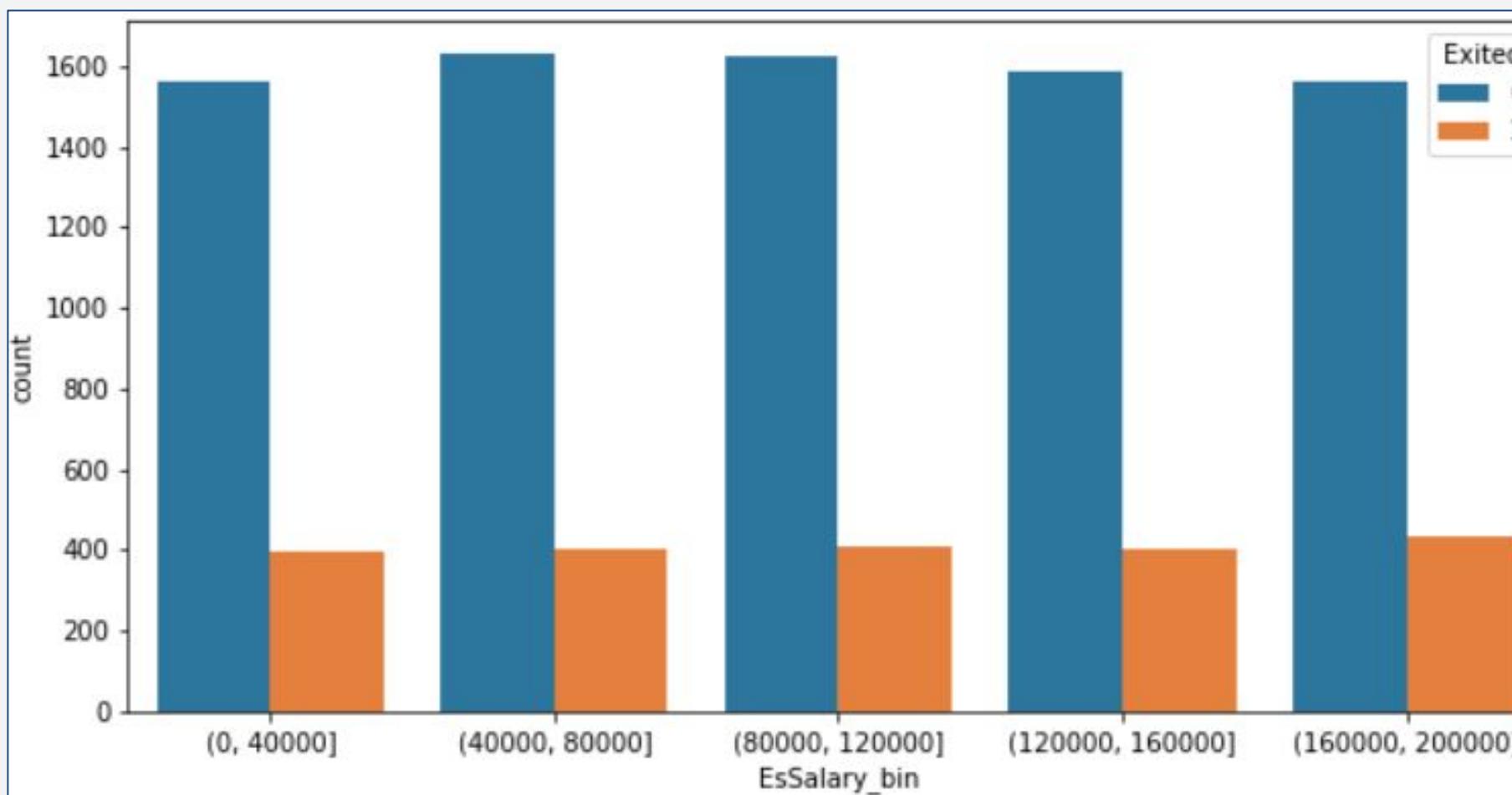
Data menunjukkan bahwa pelanggan aktif sedikit lebih banyak dibandingkan pelanggan tidak aktif, namun sebarannya cukup merata.



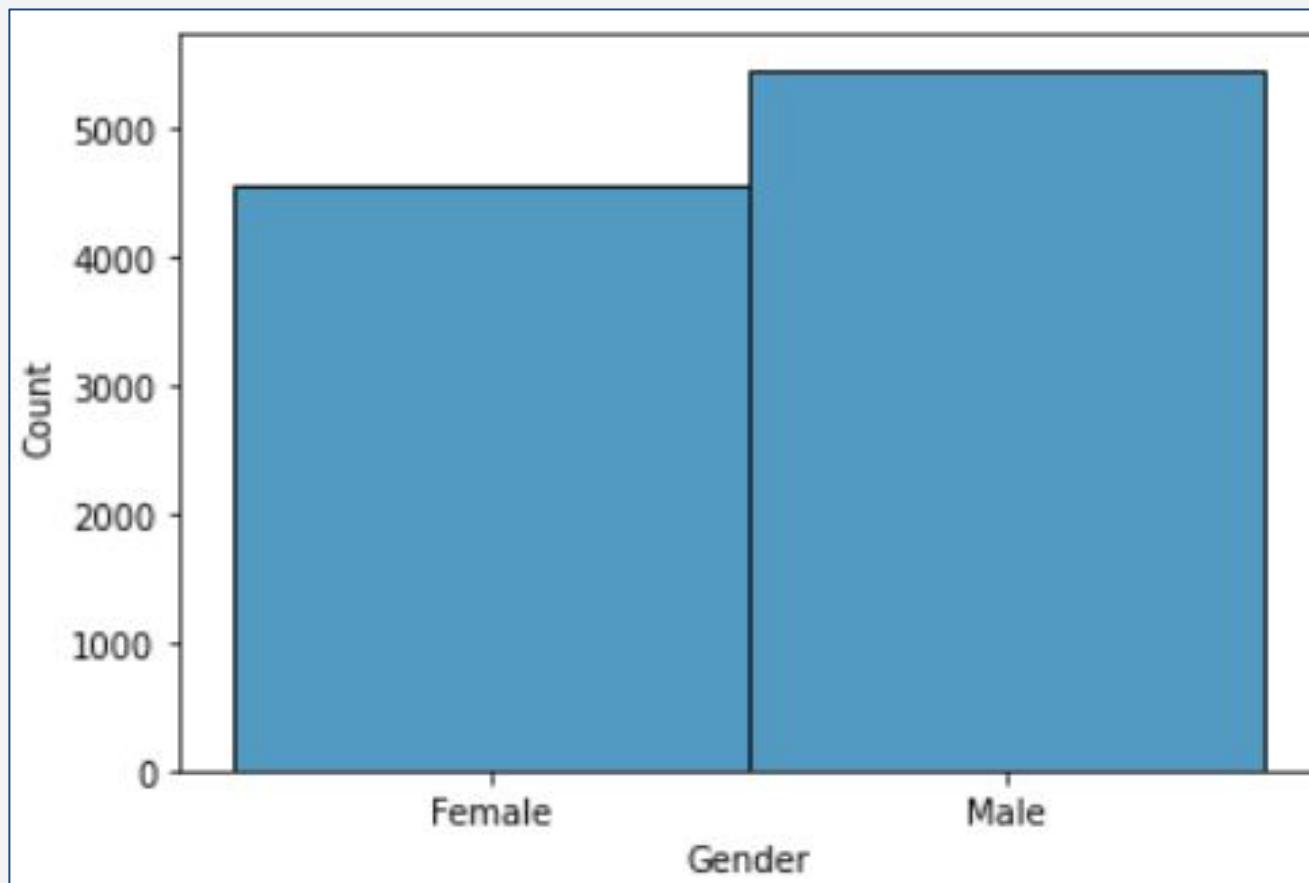
EstimatedSalary



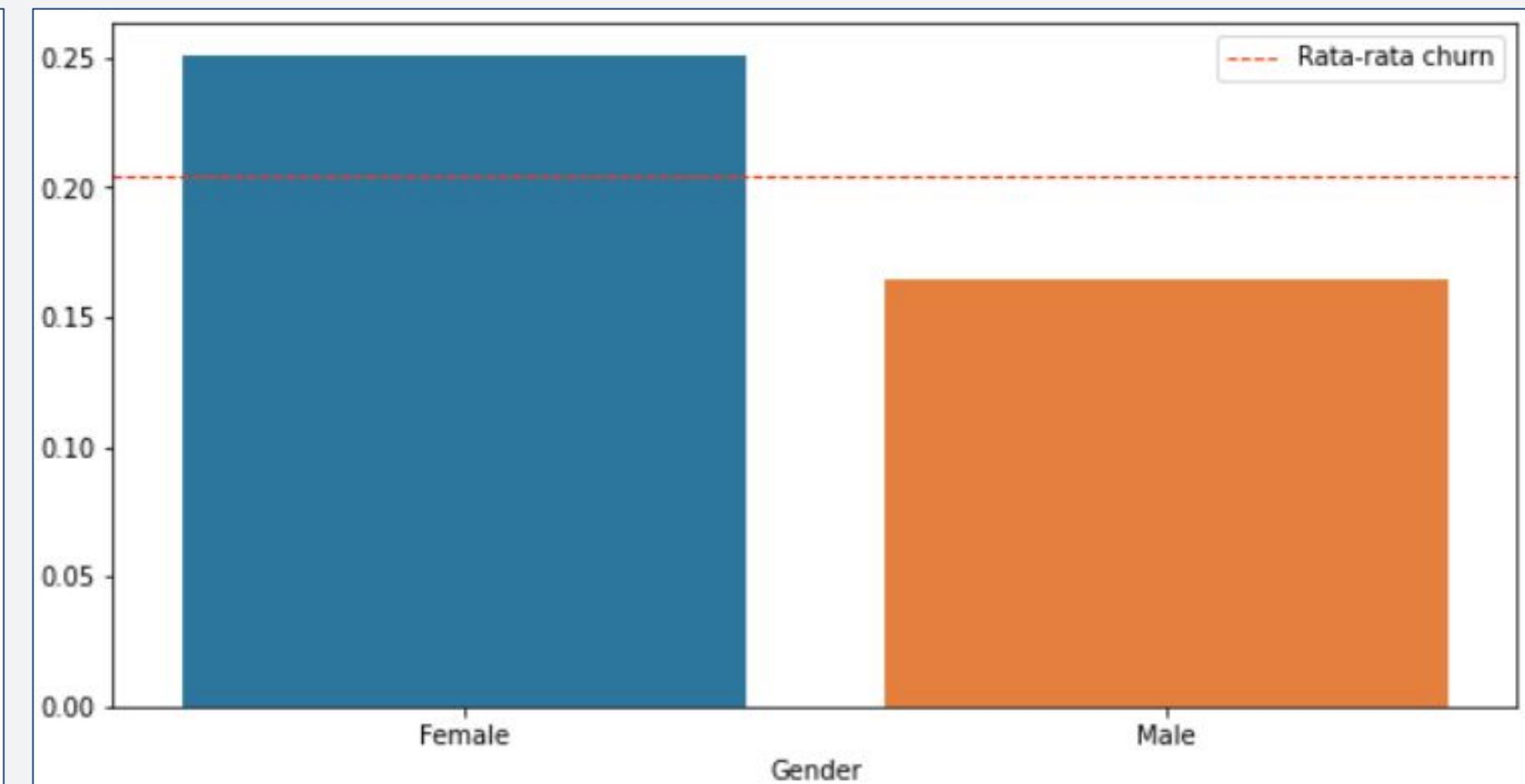
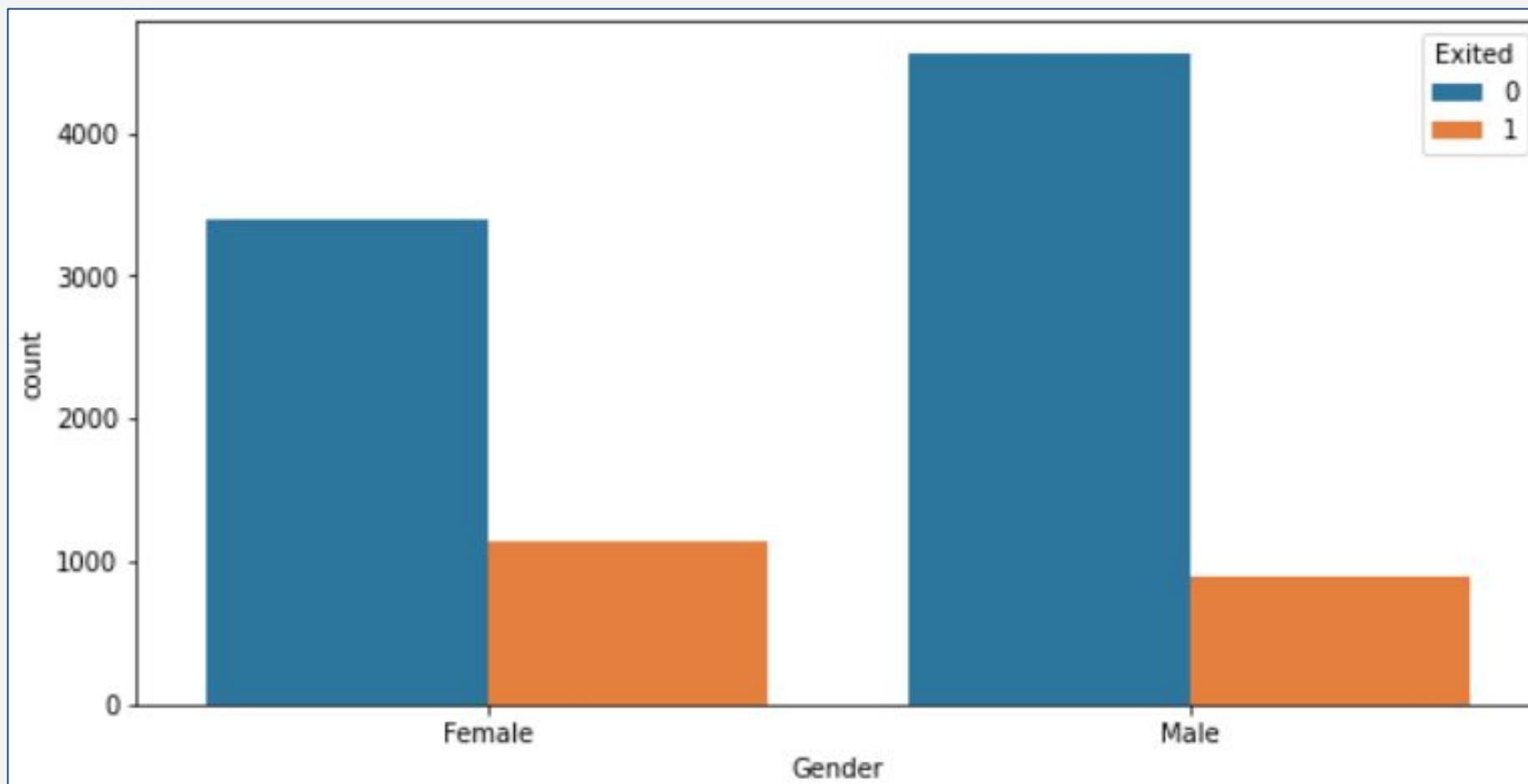
Sebaran data estimasi gaji cukup merata di rentang 0 sampai 200.000.



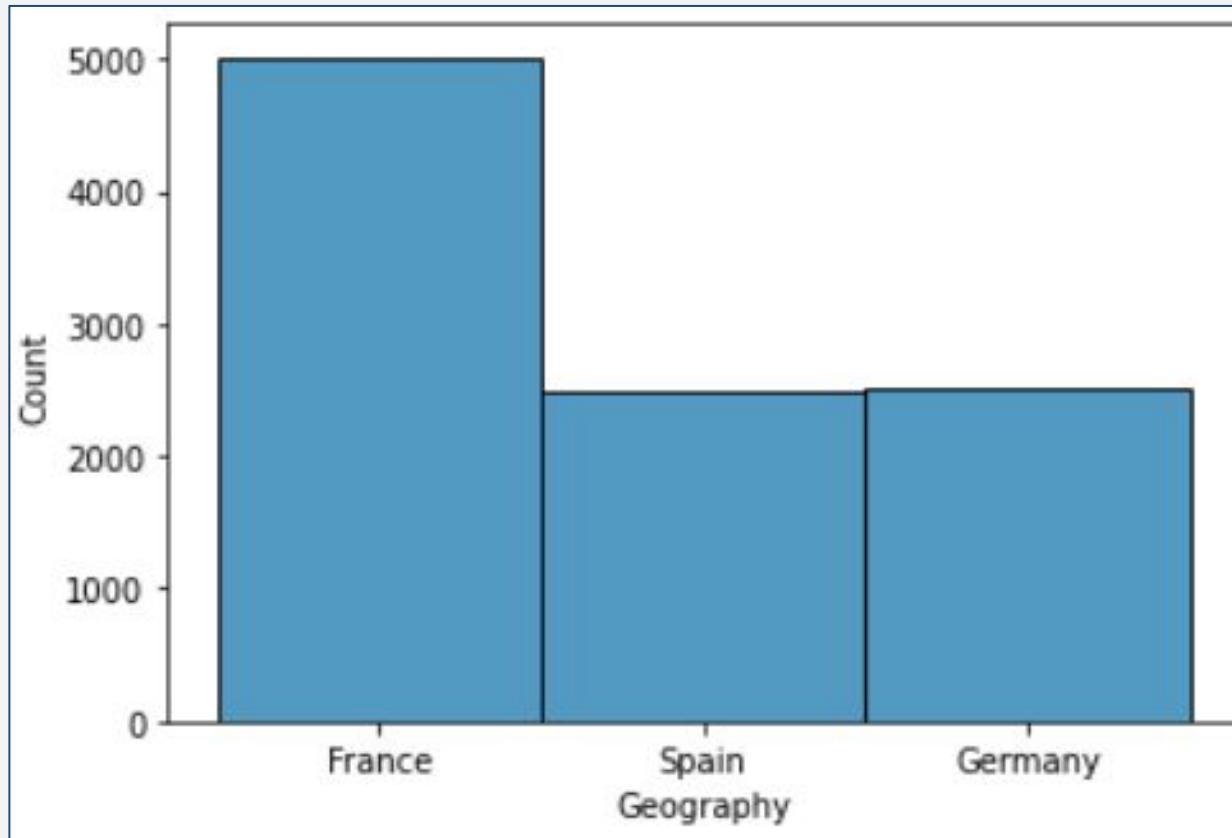
Gender



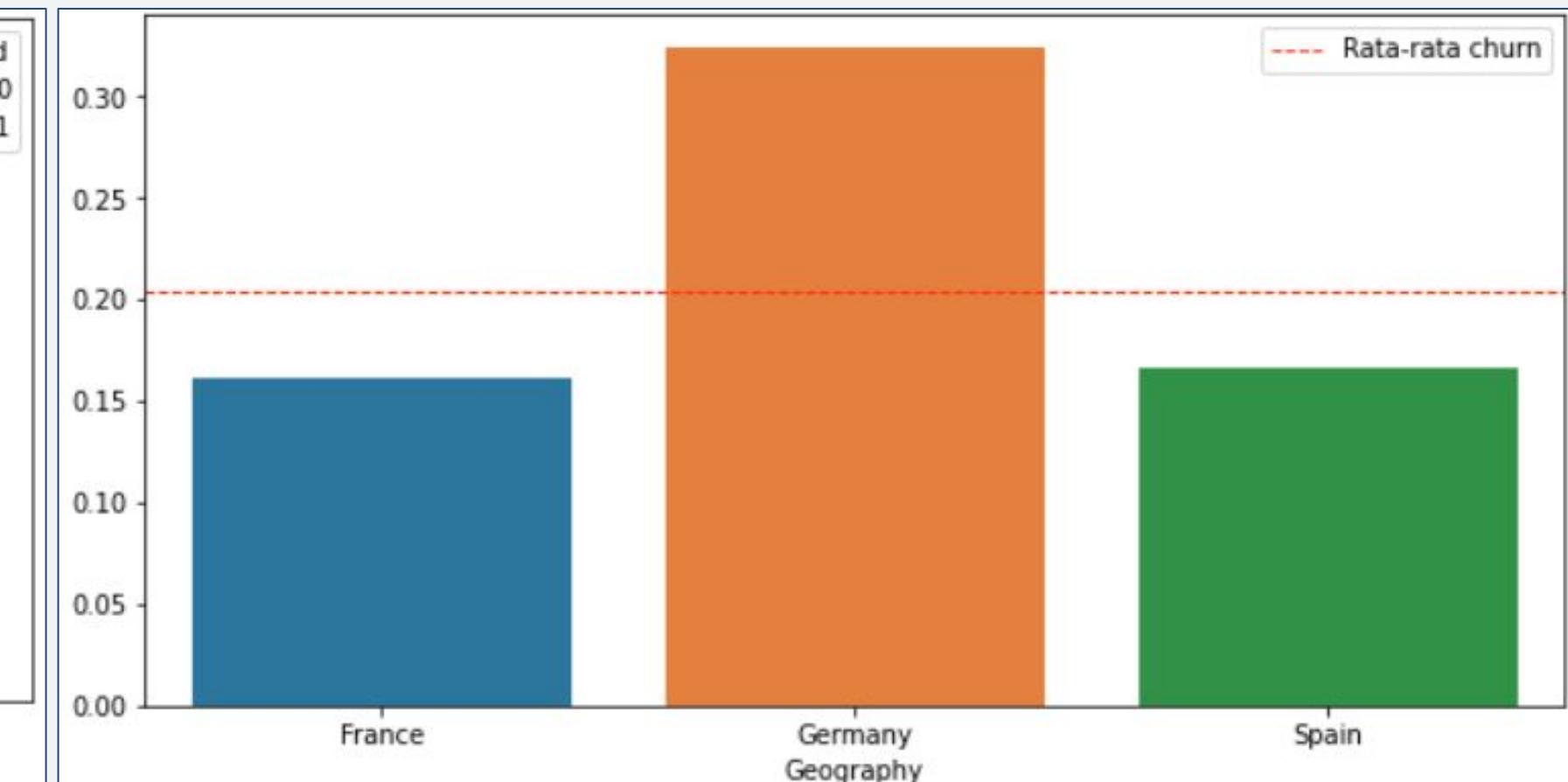
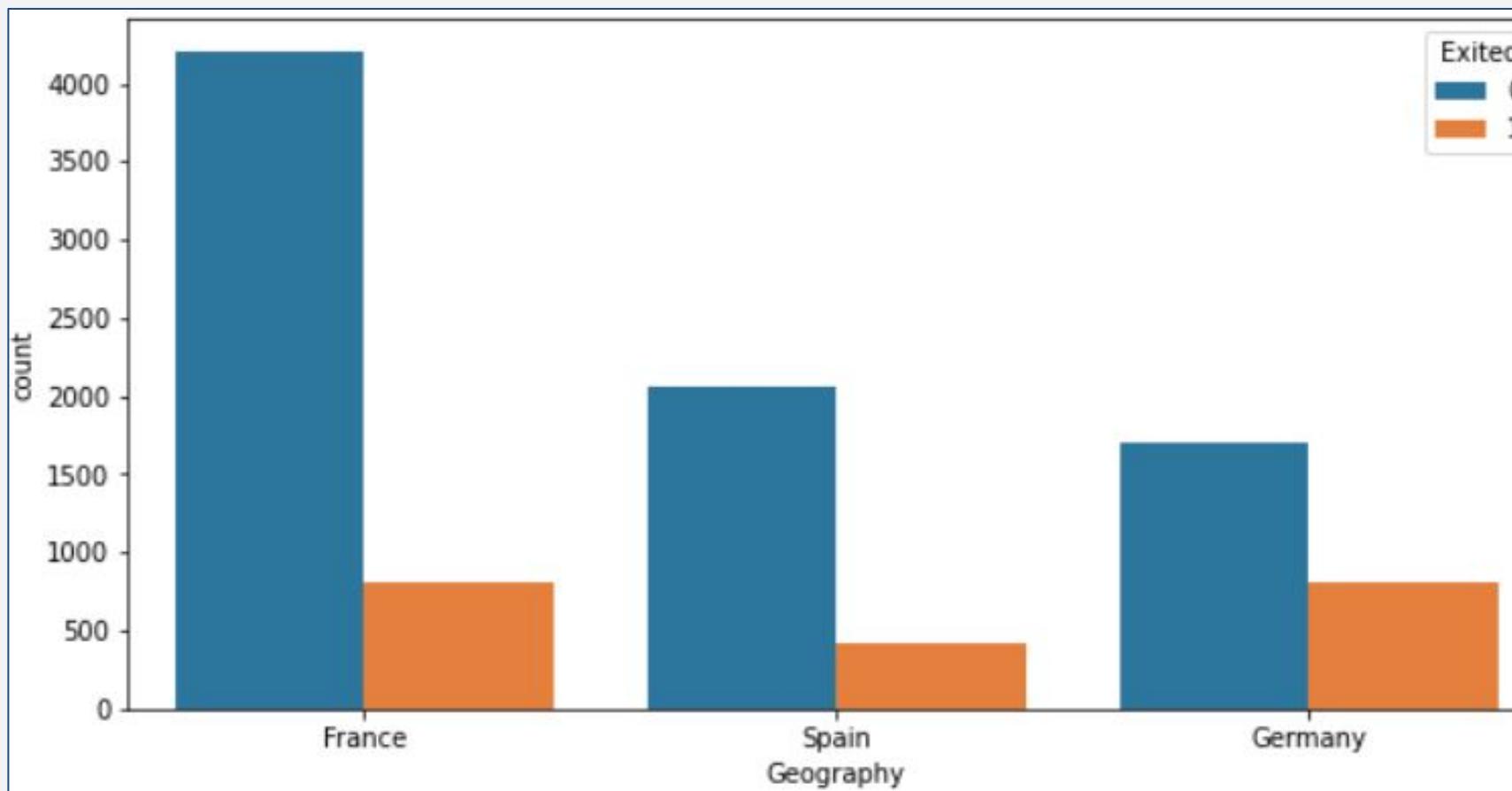
Pelanggan laki-laki lebih banyak dibanding pelanggan perempuan.

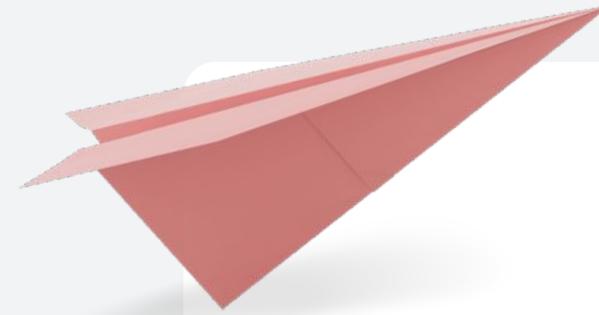


Geography



Perancis memiliki pelanggan lebih banyak dibanding Spanyol dan Jerman.





Stage 4

Data Preprocessing



DigitalSkola

One Hot Encoder untuk Analisis Churn: Kenapa?



PENAFIAN (*DISCLAIMER*)

Teknik **label encoding** bisa disalahartikan bahwa *categorical data* mempunyai urutan padahal bobot setiap value adalah sama. Ada teknik lainnya yang disebut dengan **one hot encoder**.

ONE HOT ENCODER

One hot encoder adalah teknik yang merubah setiap nilai di dalam kolom menjadi kolom baru dan mengisinya dengan nilai biner yaitu 0 dan 1.

One Hot Encoder

CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...
15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1
15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

Dataset sebelum dilakukan One Hot Encoder pada kolom yang bersifat kategorikal yaitu pada Kolom Geography dan Gender

One Hot Encoder (2)

```
▶ df = pd.get_dummies(data, columns=['Gender'], drop_first=True)
  df = pd.get_dummies(df, columns=['Geography'])
  df
```

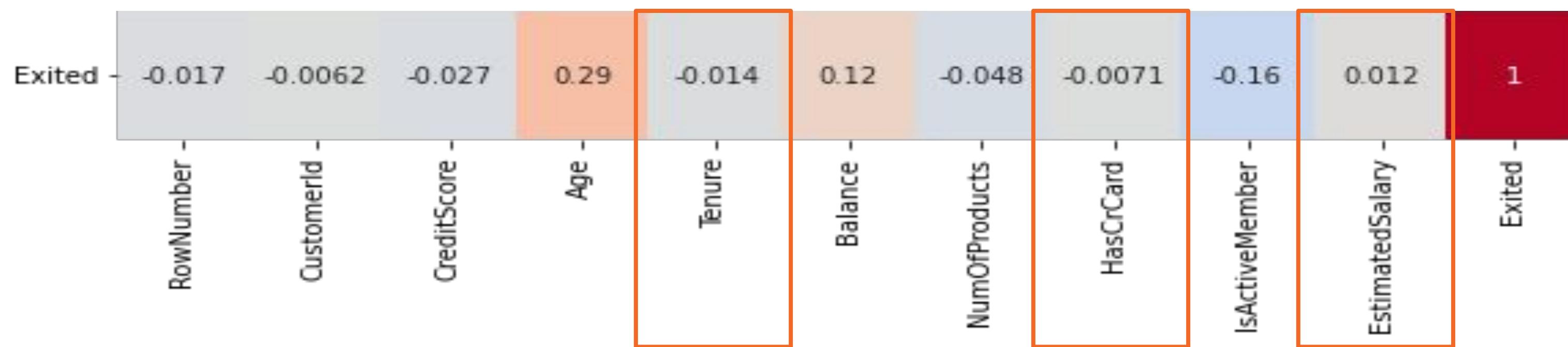
Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Gender_Male	Geography_France	Geography_Germany	Geography_Spain
0.00	1	1	1	101348.88	1	0	1	0	0
33807.86	1	0	1	112542.58	0	0	0	0	1
59660.80	3	1	0	113931.57	1	0	1	0	0
0.00	2	0	0	93826.63	0	0	1	0	0
25510.82	1	1	1	79084.10	0	0	0	0	1
...
0.00	2	1	0	96270.64	0	1	1	0	0
57369.61	1	1	1	101699.77	0	1	1	0	0
0.00	1	0	1	42085.58	1	0	1	0	0
75075.31	2	1	0	92888.52	1	1	0	1	0
30142.79	1	1	0	38190.78	0	0	1	0	0

Pada kasus Customer Churn ini, One Hot Encoder digunakan untuk mengubah gender male → 1 dan gender female → 0.
One Hot Encoder juga digunakan untuk mengubah Geography

Features Selection

Kolom “CustomerId”, “RowNumber” dan “Surname” merupakan atribut *identifier* yang tidak berpengaruh kepada target, sehingga dapat dihapus.

Sedangkan berdasarkan heatmap, terlihat kolom “Tenure”, “HasCrCard”, dan “EstimatedSalary” memiliki nilai korelasi yang kecil mendekati nol, sehingga kolom ini di drop dari Features karena dinilai tidak berkorelasi dengan nilai target “Exited”.



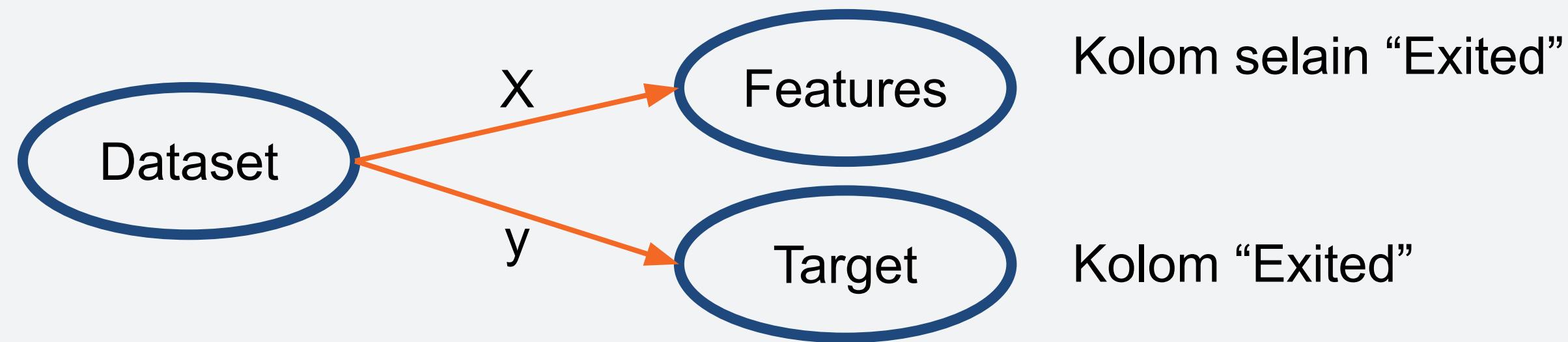
Features Selection (2)

```
df = pd.get_dummies(df.drop(columns = ["CustomerId", "RowNumber", "Surname", "Tenure", "HasCrCard", "EstimatedSalary"]))
df
```

	CreditScore	Age	Balance	NumOfProducts	IsActiveMember	Exited	Gender_Male	Geography_France	Geography_Germany	Geography_Spain
0	619	42	0.00		1	1	1	0	1	0
1	608	41	83807.86		1	1	0	0	0	1
2	502	42	159660.80		3	0	1	0	1	0
3	699	39	0.00		2	0	0	0	1	0
4	850	43	125510.82		1	1	0	0	0	1
...
9995	771	39	0.00		2	0	0	1	1	0
9996	516	35	57369.61		1	1	0	1	1	0
9997	709	36	0.00		1	1	1	0	1	0
9998	772	42	75075.31		2	0	1	1	0	1
9999	792	28	130142.79		1	0	0	0	1	0

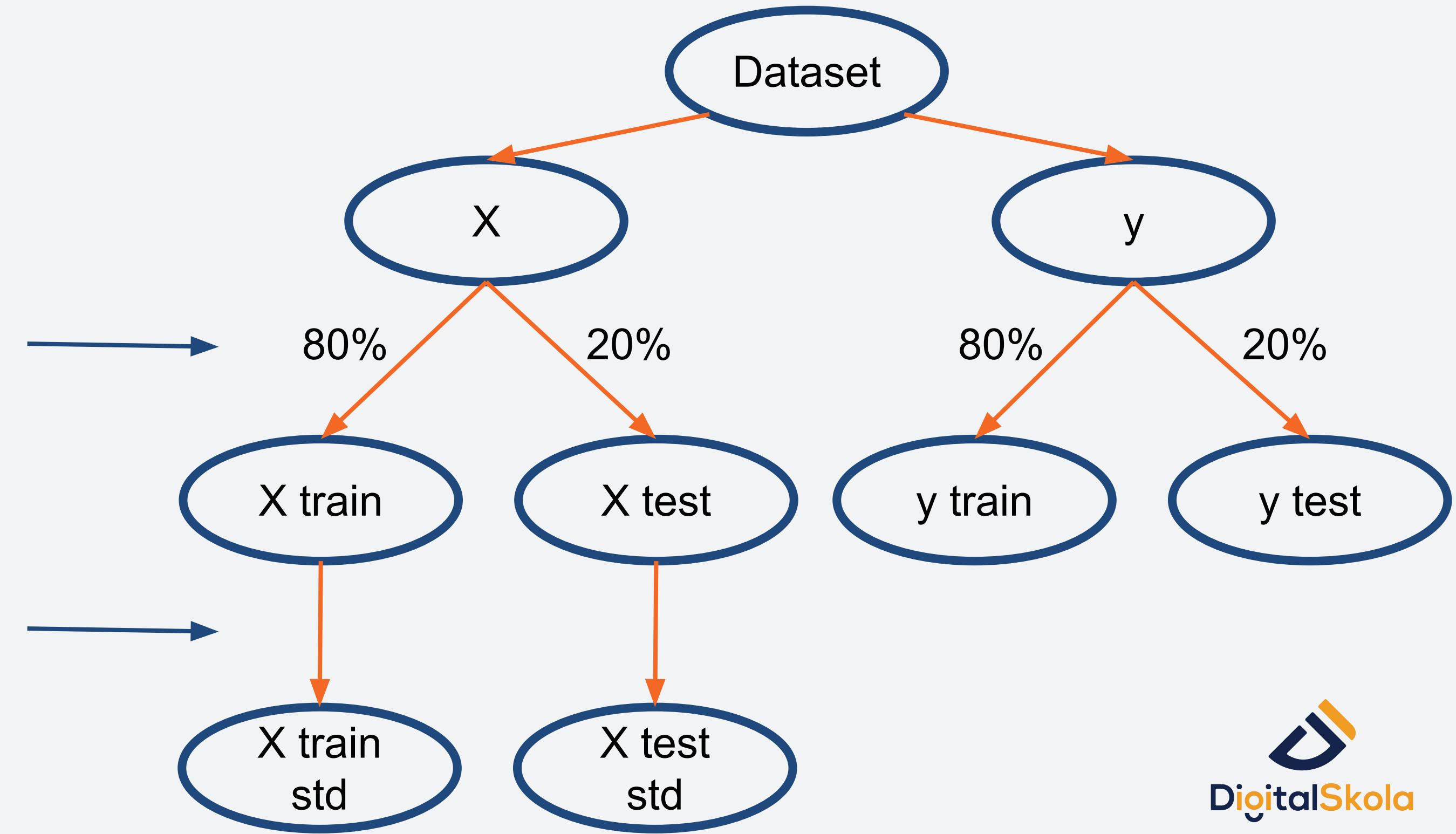
10000 rows × 10 columns

Splitting Features and Target Variables



Splitting Train-Test Dataset and Scaling

Proporsi split dataset:
- 80% train dataset
- 20% test dataset



Feature Scaling:
- StandardScaler

Hasil Scaling

```
from sklearn.preprocessing import StandardScaler  
scaler = StandardScaler()  
X_train = scaler.fit_transform(X_train)  
X_test = scaler.transform(X_test)  
  
X_train = pd.DataFrame(X_train, columns=X.columns)  
X_test = pd.DataFrame(X_test, columns=X.columns)  
  
X_train
```

	CreditScore	Age	Balance	NumOfProducts	IsActiveMember	Gender_Male	Geography_France	Geography_Germany	Geography_Spain
0	0.468234	-0.571396	-1.223126	0.806566	0.964625	-1.090038	1.00075	-0.577350	-0.577928
1	0.591686	-0.192392	1.239155	-0.912732	0.964625	0.917399	-0.99925	1.732051	-0.577928
2	-1.002902	0.376114	0.144766	0.806566	0.964625	0.917399	-0.99925	1.732051	-0.577928
3	-0.735423	-0.571396	0.677660	-0.912732	0.964625	-1.090038	1.00075	-0.577350	-0.577928
4	-0.056437	0.849870	0.231648	-0.912732	0.964625	-1.090038	1.00075	-0.577350	-0.577928
...
7995	2.052535	0.186612	0.967035	-0.912732	0.964625	0.917399	1.00075	-0.577350	-0.577928
7996	0.900316	-0.381894	0.410379	-0.912732	-1.036672	0.917399	-0.99925	1.732051	-0.577928
7997	-1.023477	-1.234654	1.092933	-0.912732	-1.036672	-1.090038	1.00075	-0.577350	-0.577928
7998	0.756289	-0.760898	1.361677	0.806566	0.964625	0.917399	-0.99925	1.732051	-0.577928
7999	-0.591395	0.660368	0.006645	-0.912732	0.964625	-1.090038	-0.99925	-0.577350	1.730320

Imbalance Dataset

```
Counter(y_train)  
Counter({0: 6365, 1: 1635})
```

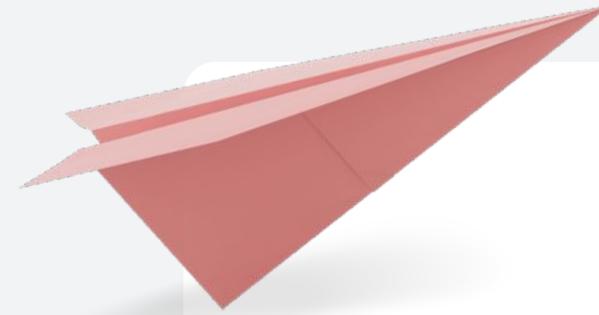
BEFORE

Pada data target (`y_train`), terlihat perbandingan yang cukup jauh antara pengguna yang tidak keluar (0) dan yang keluar (1). Agar target data training lebih seimbang kita akan melakukan balancing dataset menggunakan metode SMOTE.

```
from imblearn.over_sampling import SMOTE  
X_dummy = pd.get_dummies(X_train)  
X_dummy = X_dummy.fillna(0)  
smote = SMOTE(sampling_strategy = 0.5)  
X_smote, y_smote = smote.fit_resample(X_dummy, y_train)
```

AFTER

```
Counter(y_smote)  
Counter({0: 6365, 1: 3182})
```



Stage 5

Modelling



DigitalSkola

ANN (Data Preprocessing)

1. One Hot Encoder untuk Fitur Gender dan Geography
2. Membagi data input (X) dan target (y):
 $X = \text{CreditScore, Age, Balance, NumOfProducts, IsActiveMember, Gender, Geography}$
 $y = \text{Exited}$
3. Train Test Split
4. Standard Scaling
5. Balancing Dataset
undersampling sebanyak 50% lalu oversampling SMOTE sebanyak 100%

ANN (Modeling)

1. Menggunakan 4 layer dengan rule sebagai berikut:

Layer	Jumlah Node	Activation	Dropout
Layer I	8	Relu	0,2
Layer II	7	Relu	0,2
Layer III	3	Relu	0,2
Layer IV	1	Sigmoid	0,2

2. Menggunakan Adam Optimizer
3. Epoch = 100, batch_size = 8

ANN (Performance)

Training Performance

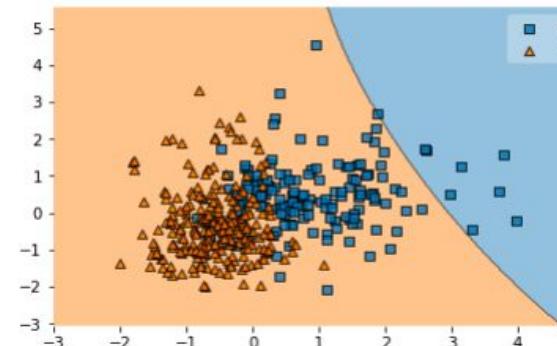
Accuracy : 79.04 %
Precision : 77.89 %
Recall : 81.08 %
Specificity: 76.99 %
NPV : 80.28 %

Testing Performance

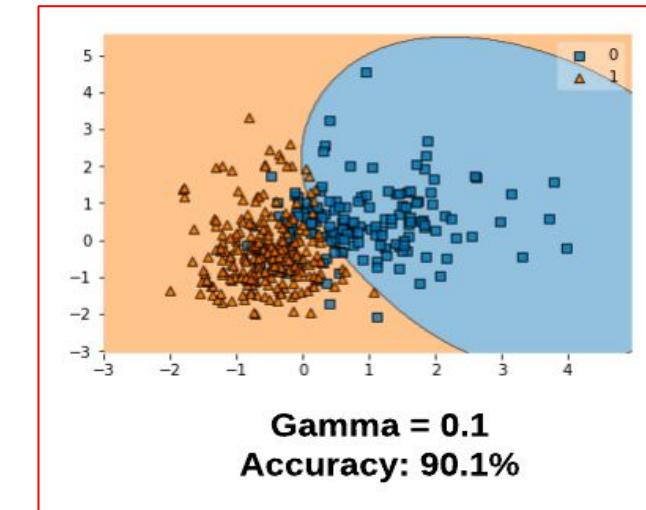
Accuracy : 76.9 %
Precision : 43.59 %
Recall : 76.66 %
Specificity: 76.96 %
NPV : 93.42 %

SVM-Radial Basis Function

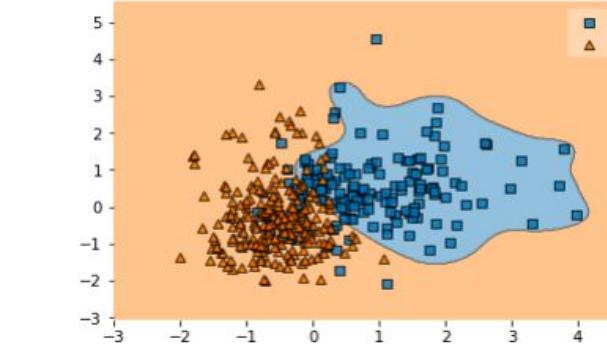
1. Support Vector Machine-Radial Basis Function (SVM-RBF) dipilih karena dataset tidak linear dan *inseparable*.
2. Parameter yang digunakan dalam SVM-RBF adalah Gamma dan C.
3. Nilai Gamma merepresentasikan seberapa jauh capaian pengaruh satu data training. Nilai Gamma yang tinggi belum tentu akurat, begitu sebaliknya, nilai yang kecil juga belum tentu akurat. **Accuracy** yang optimal justru di dapat dari nilai Gamma intermediate.



Gamma = 0.008
Accuracy: 63.7%

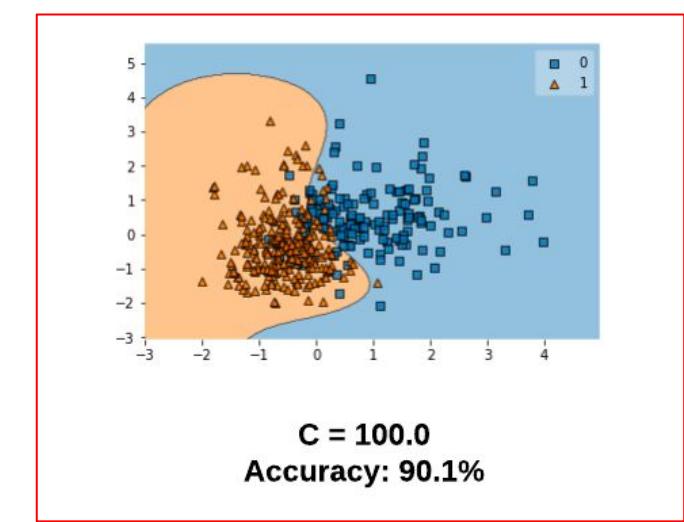


Gamma = 0.1
Accuracy: 90.1%



Gamma = 3.0
Accuracy: 88.9%

4. Parameter C digunakan untuk mentoleransi misklasifikasi pada data points untuk mengecilkan eror. Semakin besar nilai C, maka nilai **accuracy**-nya baik.
5. Nilai parameter yang dipakai Gamma = 0.1 dan C = 150



C = 100.0
Accuracy: 90.1%

SVM-RBF (Data Preprocessing)

1. One Hot Encoder untuk Fitur Gender dan Geography
2. Membagi data input (X) dan target (y):
X = CreditScore, Age, Balance, NumOfProducts, IsActiveMember, Gender, Geography
y = Exited
3. Train Test Split
4. Standard Scaling
5. Balancing Dataset
oversampling SMOTE sebanyak 50%

SVM-RBF (Performance)

Training Performance

Accuracy : 87.78 %
Precision : 70.73 %
Recall : 68.56 %
Specificity: 92.71 %
NPV : 91.99 %

Testing Performance

Accuracy : 83.3 %
Precision : 58.67 %
Recall : 57.21 %
Specificity: 89.86 %
NPV : 89.3 %

Catatan: Nilai parameter Gamma = 0.1 dan C = 150

XGBoost (Data Preprocessing)

1. One Hot Encoder untuk Fitur Gender dan Geography
2. Membagi data input (X) dan target (y):
 $X = \text{CreditScore, Age, Balance, NumOfProducts, IsActiveMember, Gender, Geography}$
 $y = \text{Exited}$
3. Train Test Split
4. Standard Scaling
5. Balancing Dataset
undersampling sebanyak 50% lalu oversampling SMOTE sebanyak 100%

XGBoost (Performance)

Training Performance

Accuracy : 91.8 %
Precision : 91.55 %
Recall : 92.11 %
Specificity: 91.5 %
NPV : 92.06 %

Testing Performance

Accuracy : 79.05 %
Precision : 48.55 %
Recall : 70.65 %
Specificity: 81.16 %
NPV : 91.66 %

Matriks Evaluasi

		Predicted
		True Negative
Actual	Predicted	False Positive
	False Negative	True Positive



		Predicted
		True Stay
Actual	Predicted	False Exit
	False Stay	True Exit

Nilai **Accuracy** merepresentasikan nilai prediksi True Negative dan True Positive. Artinya jumlah yang benar-benar stay (True Stay) dan jumlah yang benar-benar exit (True Exit) dapat dilihat diukur dari nilai akurasi ini.

False Positive merepresentasikan jumlah customer yang exit, tetapi ternyata masih stay (False Exit). Kita menginginkan hasil minimal untuk customer yang diprediksi **Stay** namun sebenarnya **Exit**, sehingga selain **Accuracy**, kita mempertimbangkan nilai **Recall** yang dapat menunjukkan False Negative. **Semakin kecil False Negative maka semakin kecil pula customer yang diprediksi stay, tetapi ternyata exit (False Stay).**

Summary

ANN

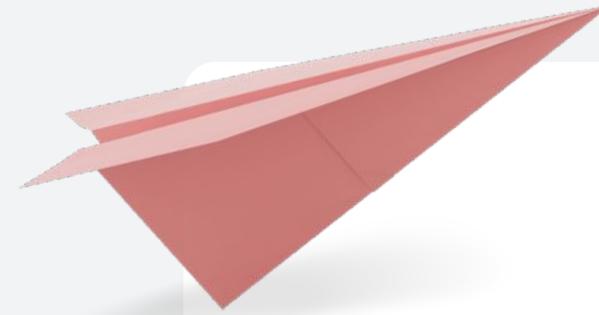
# Testing Performance	
Accuracy	: 76.9 %
Precision	: 43.59 %
Recall	: 76.66 %
Specificity	: 76.96 %
NPV	: 93.42 %

SVM-RBF

# Testing Performance	
Accuracy	: 83.3 %
Precision	: 58.67 %
Recall	: 57.21 %
Specificity	: 89.86 %
NPV	: 89.3 %

XGBoost

# Testing Performance	
Accuracy	: 79.05 %
Precision	: 48.55 %
Recall	: 70.65 %
Specificity	: 81.16 %
NPV	: 91.66 %



Terima Kasih

Send it to us! We hope you learned something new.



DigitalSkola