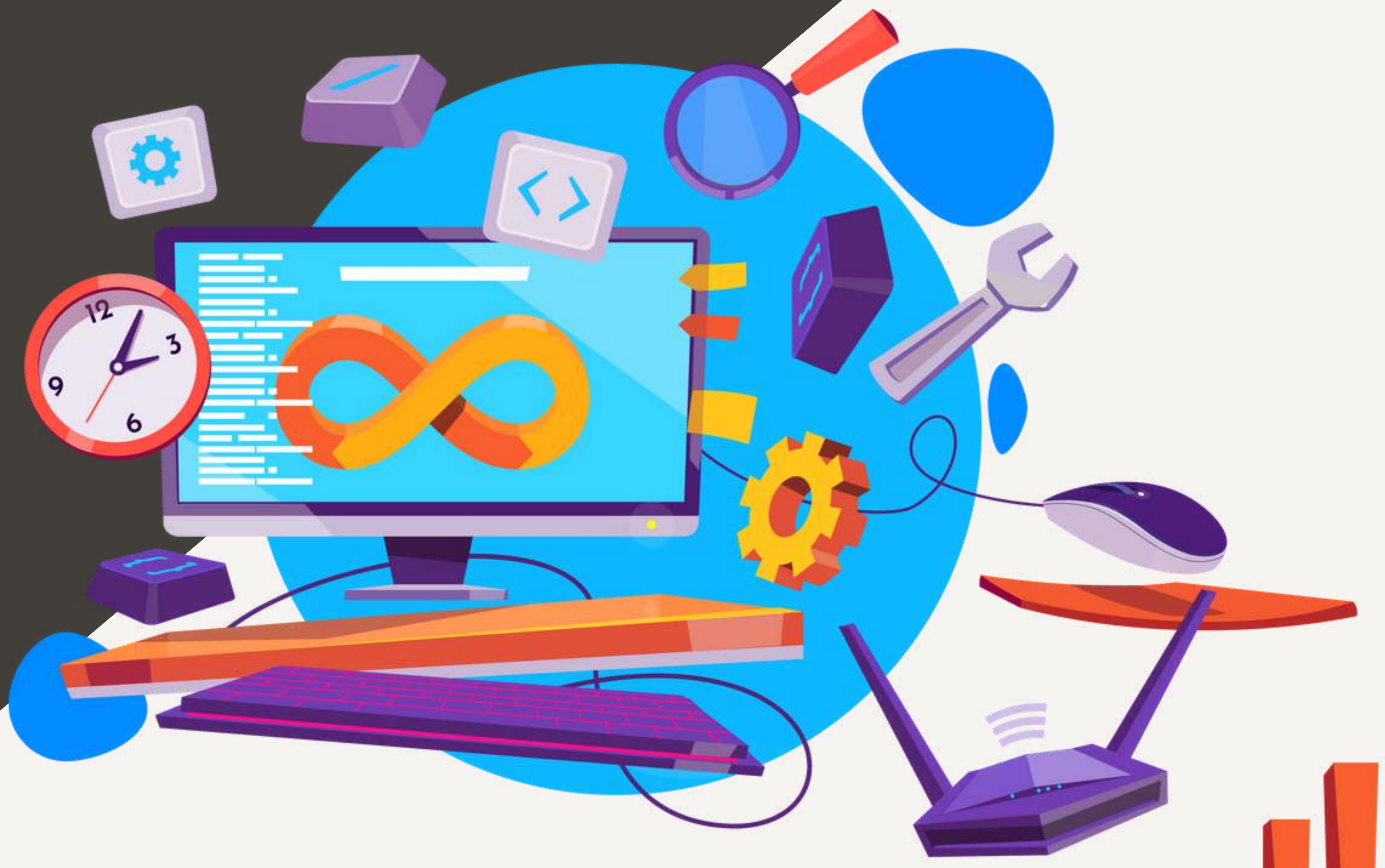
Kampus Merdeka
INDONESIA JAYA

MyEduSolve

TELCO CUSTOMER CHURN



BY GROUP 2

OUR TEAM MEMBER

M. Kamal Jaza

M. Irvan Arfandi

Risma Ashali

Rizal Maulana K.

Shafira Aisyah

TABLE OF CONTENT

WORKFLOW CUSTOMER CHURN ANALYSIS

1. Use Case
2. Bussiness Understanding
3. Data Understanding
4. Data Preparation
5. Data Profiling
6. Data Cleansing
7. Exploratory Data Analysis
8. Preprocessing Data
9. Modelling
10. Evaluation
11. Result
12. Recommendation

USE CASE

USE CASE

OBJECTIVE STATEMENT

- To get insight into what type gender who churn and no churn.
- To gain insight into senior citizens who churn and no churn.
- To gain insight into churn and no churn customers whether the customer has a partner or has dependents.
- To gain insight into churn and no churn customers based on how long the customer have tenure.
- To gain insight into churn and no churn customers based on how many customers use Internet Service, Online Security, Multiple Lines, Online Backup, Device Protection, Tech Support, Streaming TV, Movie Streaming, Payment Methods, and Paperless Billing facilities.
- To get insight about what type based on customer contract churn and no churn.
- To get insight about how long customers churn, monthly subscription fees, and the total cost they spend on services.
- To get insight about customer churn analysis.
- To create modeling with Machine Learning to predict customer churn.

USE CASE

CHALLENGES

- There are some variables containing many missing values
- There are some inappropriate data types
- There is multicolinerity on some variables

METHODOLOGY / ANALYTIC TECHNIQUE

- Descriptive analysis
- Graph analysis
- Modelling using Logistic Regression

BUSINESS BENEFIT

- Helping Business Development Team to create product differentiation based on the characteristic for each customer.
- Know how to treat customers with spesific criteria, especially between churn customers and no churn customers.

USE CASE

EXPECTED OUTCOME

- Know how many customers based on gender type who churn and no churn.
- Know how many customers based on senior citizens who churn and no churn.
- Know how many customers based on churn and no churn customers whether the customer has a partner or has dependents.
- Know how many customers based on churn and no churn customers based on how long the customer has worked in the company.
- Know how many customers based on churn and no churn customers based on how many customers use Internet Service, Online Security, Multiple Lines, Online Backup, Device Protection, Tech Support, Streaming TV, Movie Streaming, Payment Methods, and Paperless Billing facilities.
- Know how many customers based on type customer contract churn and no churn.
- Know how long customers churn, monthly subscription fees, and the total cost they spend on services.
- Know customer churn analysis.
- Create modeling with Machine Learning to predict customer churn.

BUSINESS UNDERSTANDING

BUSINESS UNDERSTANDING

- Data telco is a company engaged in telecommunication and internet services to make easier for consumers to communicate remotely and surf the internet more easily with offers several services such as time contracts, and various types of services.
- This case has some business question using the data:
  - How many customers based on gender type who churn and no churn?
  - How many customers based on senior citizens who churn and no churn?
  - How many customers based on churn and no churn customers whether the customer has a partner or has dependents?
  - How many churn customers and no churn customers based on how long the customers have tenure?
  - How many churn and no churn customers based on how many customers use Internet Service, Online Security, Multiple Lines, Online Backup, Device Protection, Tech Support, Streaming TV, Movie Streaming, Payment Methods, and Paperless Billing facilities?
  - How many customers based on type customer contract churn and no churn?
  - How long customers subscribe, monthly subscription fees, and the total cost they spend on services?
  - How about customer churn analysis?
  - How to create modeling with Machine Learning to predict customers churn?

DATA UNDERSTANDING

DATA UNDERSTANDING

- Data of Telecom Customer with 21 columns and 7043 rows
- Source Code : https://www.kaggle.com/datasets/blastchar/telco-customer-churn
- Data Dictionary :
  - **customerId** : Customer number uniquely assigned to each customer.
  - **gender** : gender of customer
  - **SeniorCitizen** : Whether the customer is a senior citizen or not (1, 0)
  - **partner** : Whether the customer has a partner or not (Yes, No)
  - **Dependets** : Whether the customer has dependents or not (Yes, No)
  - **tenure** : Number of months the customer has stayed with the company
  - **PhoneService** : Whether the customer has a phone service or not (Yes, No)
  - **MultipleLines** : Whether the customer has multiple lines or not (Yes, No, No phone service)
  - **InternetService** : Customer’s internet service provider (DSL, Fiber optic, No)
  - **OnlineSecurity** : Whether the customer has online security or not (Yes, No, No internet service)
  - **OnlineBackup** : Whether the customer has online backup or not (Yes, No, No internet service)
  - **DeviceProtection** : Whether the customer has device protection or not (Yes, No, No internet service)
  - **TechSupport** : Whether the customer has tech support or not (Yes, No, No internet service)
  - **StreamingTV** : Whether the customer has streaming TV or not (Yes, No, No internet service)
  - **StreamingMovies** : Whether the customer has streaming movies or not (Yes, No, No internet service)
  - **Contract** : The contract term of the customer (Month-to-month, One year, Two year)
  - **PaperlessBilling** : Whether the customer has paperless billing or not (Yes, No)
  - **PaymentMethod** : The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
  - **MonthlyCharges** : The amount charged to the customer monthly
  - **TotalCharges** : The total amount charged to the customer
  - **Churn** : The customer churn status (1 - Yes, 0 - No)

DATA PREPARATION

DATA PREPARATION

- Code Used:
  - Python Version: 3.7.15
  - Packages: Pandas, Numpy, Matplotlib, Seaborn, Sklearn, and Warnings

DATA PROFILING

DATA PROFILING

WHAT IS DATA PROFILING?

- Data profiling is the process of reviewing source data, understanding structure, content and interrelationships, and identifying potential for data projects.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split
from sklearn.ensemble import ExtraTreesClassifier

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_auc_score, accuracy_score, precision_score, recall_score, confusion_matrix, roc_curve, auc, log_loss
from imblearn.over_sampling import SMOTE

from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
from sklearn.model_selection import RandomizedSearchCV
from sklearn.model_selection import GridSearchCV

from sklearn.metrics import mean_absolute_percentage_error

import warnings
warnings.filterwarnings('ignore')
```

Import packages

DATA PROFILING

IMPORT DATASET

```
df = pd.read_csv('data_telco.csv')
df.head()
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	NaN	NaN	NaN	NaN	1	No	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	NaN	NaN	NaN	NaN	34	Yes	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	One year	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	NaN	NaN	NaN	NaN	2	Yes	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	NaN	NaN	NaN	NaN	45	No	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	One year	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	NaN	NaN	NaN	NaN	2	Yes	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

5 rows × 21 columns

DATA PROFILING

IMPORT DATASET

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            6034 non-null   object
 2   SeniorCitizen     6034 non-null   float64
 3   Partner           6034 non-null   object
 4   Dependents        6034 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     6034 non-null   object
 8   InternetService   6034 non-null   object
 9   OnlineSecurity    6034 non-null   object
 10  OnlineBackup      6034 non-null   object
 11  DeviceProtection  6034 non-null   object
 12  TechSupport       6034 non-null   object
 13  StreamingTV       6034 non-null   object
 14  StreamingMovies   6034 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(1), object(18)
memory usage: 1.1+ MB
```

df.isna().sum()

```
customerID             0
gender              1009
SeniorCitizen       1009
Partner             1009
Dependents          1009
tenure                 0
PhoneService           0
MultipleLines       1009
InternetService     1009
OnlineSecurity      1009
OnlineBackup        1009
DeviceProtection    1009
TechSupport         1009
StreamingTV         1009
StreamingMovies     1009
Contract               0
PaperlessBilling       0
PaymentMethod          0
MonthlyCharges         0
TotalCharges           0
Churn                  0
dtype: int64
```

DATA CLEANSING

DATA CLEANSING

WHAT IS DATA CLEANSING?

- Data cleansing is the process of identifying and resolving corrupt, inaccurate, or irrelevant data.
- Common inaccuracies in data **include missing values, misplaced entries, and typographical errors**. In some cases, data cleansing requires certain values to be filled in or corrected, while in other instances, the values will need to be removed altogether.

HANDLING MISSING VALUE

- THERE ARE MISSING VALUES IN COLUMN GENDER, SENIORCITIZEN, PARTNER, DEPENDENTS, MULTIPLELINES, INTERNETSERVICE,ONLINESECURITY, ONLINEBACKUP, DEVICEPROTECTION, TECHSUPPORT, STREAMINGTV, AND STREAMING MOVIES.
- IN THIS CASE, WE WILL REPLACE THE MISSING VALUE WITH "UNKNOWN".
- EXAMPLE:

```
df['gender'] = df.gender.fillna("unknown")
df['gender'].value_counts()
```

```
Male       3041
Female     2993
unknown    1009
Name: gender, dtype: int64
```

BEFORE

```
df.isna().sum()
```

```
customerID             0
gender              1009
SeniorCitizen       1009
Partner             1009
Dependents          1009
tenure                 0
PhoneService           0
MultipleLines       1009
InternetService     1009
OnlineSecurity      1009
OnlineBackup        1009
DeviceProtection    1009
TechSupport         1009
StreamingTV         1009
StreamingMovies     1009
Contract               0
PaperlessBilling       0
PaymentMethod          0
MonthlyCharges         0
TotalCharges           0
Churn                  0
dtype: int64
```

AFTER

```
df.isna().sum()
```

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

CHANGE DATA TYPE

- COLUMN TOTAL CHARGES HAS IMPROPER DATA TYPES (OBJECT), SO IT MUST BE CHANGE TO NUMERIC.
- THE SYNTAX IS :

DF['TOTALCHARGES'] = PD.TO_NUMERIC(DF['TOTALCHARGES'],ERRORS="COERCE")

BEFORE

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            6034 non-null   object
 2   SeniorCitizen     6034 non-null   float64
 3   Partner           6034 non-null   object
 4   Dependents        6034 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     6034 non-null   object
 8   InternetService   6034 non-null   object
 9   OnlineSecurity    6034 non-null   object
 10  OnlineBackup      6034 non-null   object
 11  DeviceProtection  6034 non-null   object
 12  TechSupport       6034 non-null   object
 13  StreamingTV       6034 non-null   object
 14  StreamingMovies   6034 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(1), object(18)
memory usage: 1.1+ MB
```

AFTER

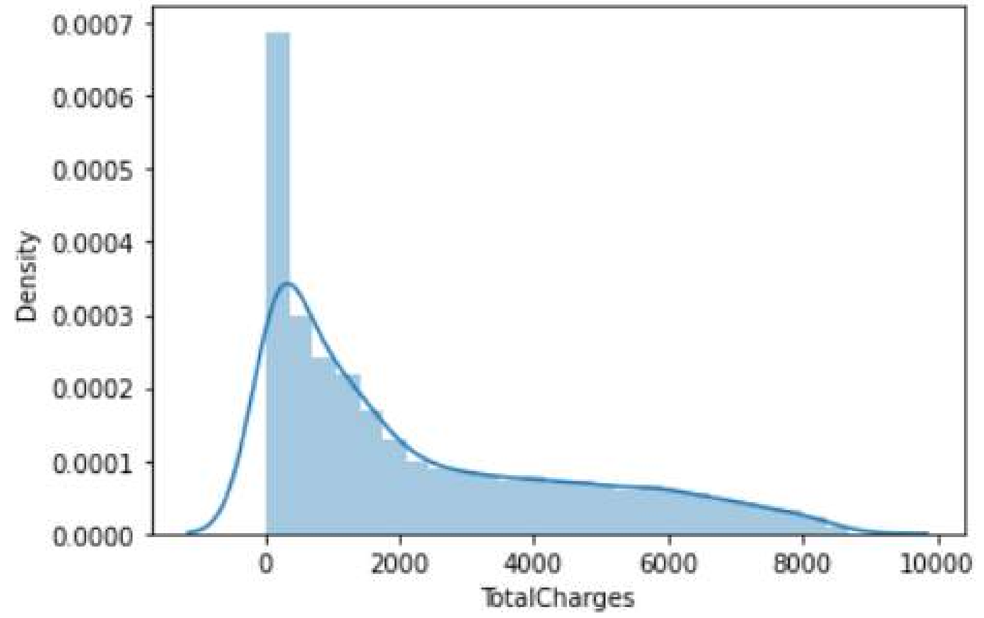
```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   object
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7032 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(1), object(18)
memory usage: 1.1+ MB
```

HANDLING MISSING VALUE (2)

AFTER CHANGING THE DATA TYPE IN COLUMN TOTAL CHARGES, IT FOUND THAT THERE IS SOME MISSING VALUE.

```
sns.distplot(df['TotalCharges'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7825fefa90>
```



Because Total Charges tend to be possitive skewness, to fill in the missing value using the median

```
val = df['TotalCharges'].median()
df['TotalCharges'] = df['TotalCharges'].fillna(val)
```

BEFORE

```
df.isna().sum()
```

```
customerID           0
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

AFTER

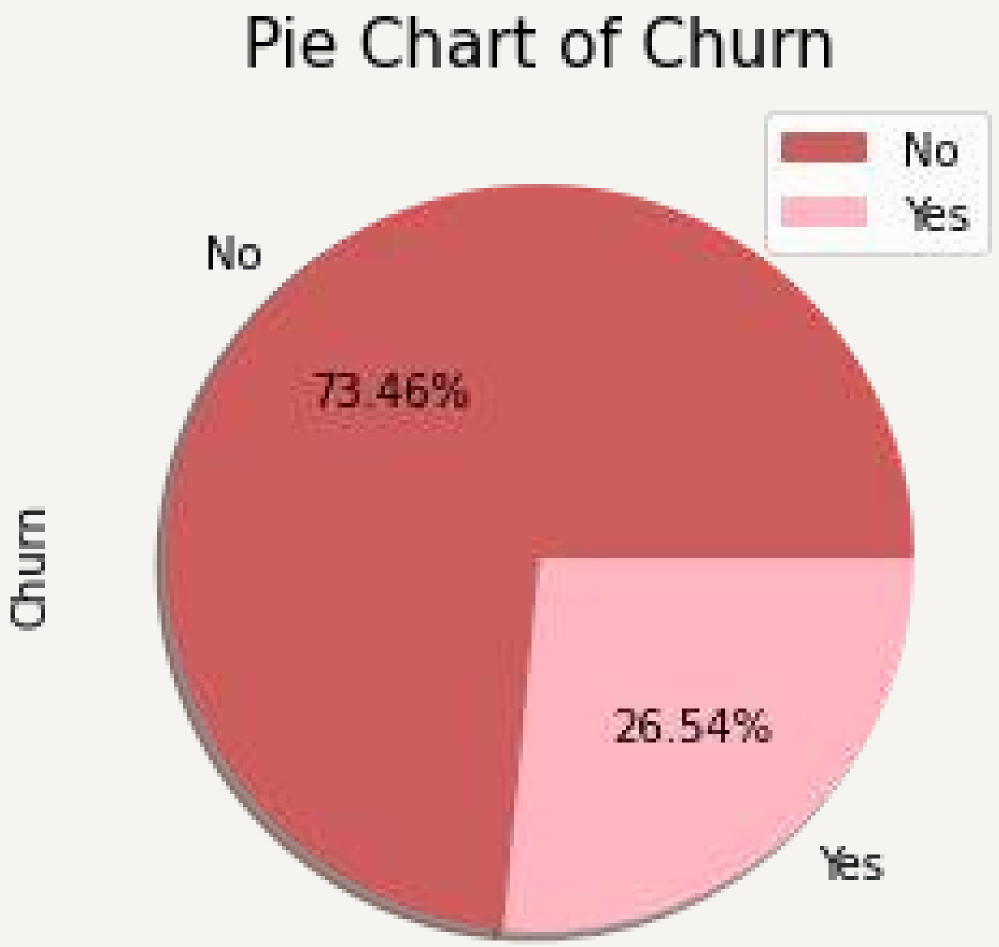
```
df.isna().sum()
```

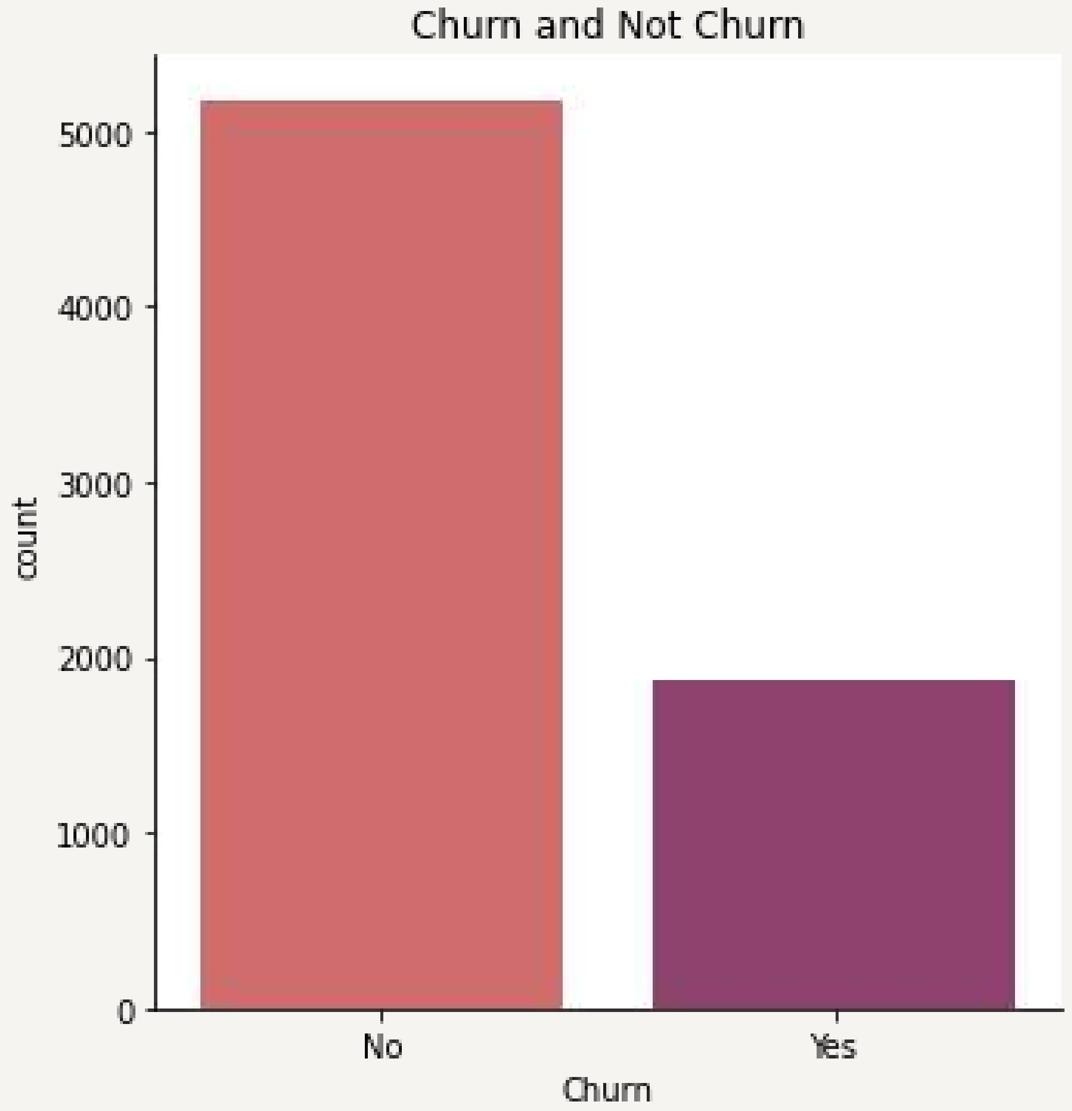
```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

EXPLORATORY DATA ANALYSIS

EDA

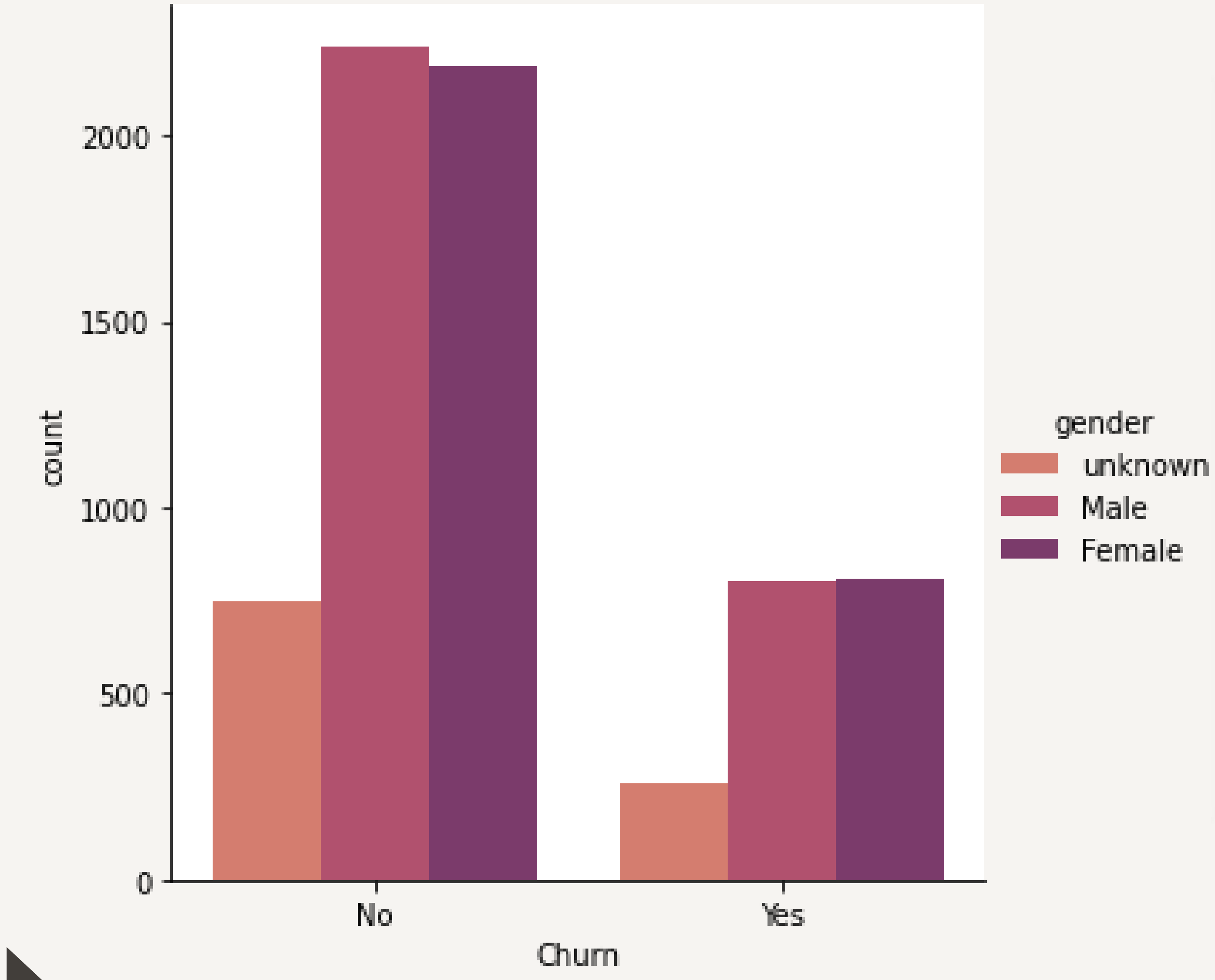
CHURN VS NO CHURN





THERE ARE MORE NO CHURN CUSTOMERS THAN CHURN CUSTOMERS. NO CHURN CUSTOMERS THERE ARE AS MUCH AS 73.46%, WHILE CHURN CUSTOMERS THERE ARE AS MUCH AS 26.54%.

BY GENDER



format in %

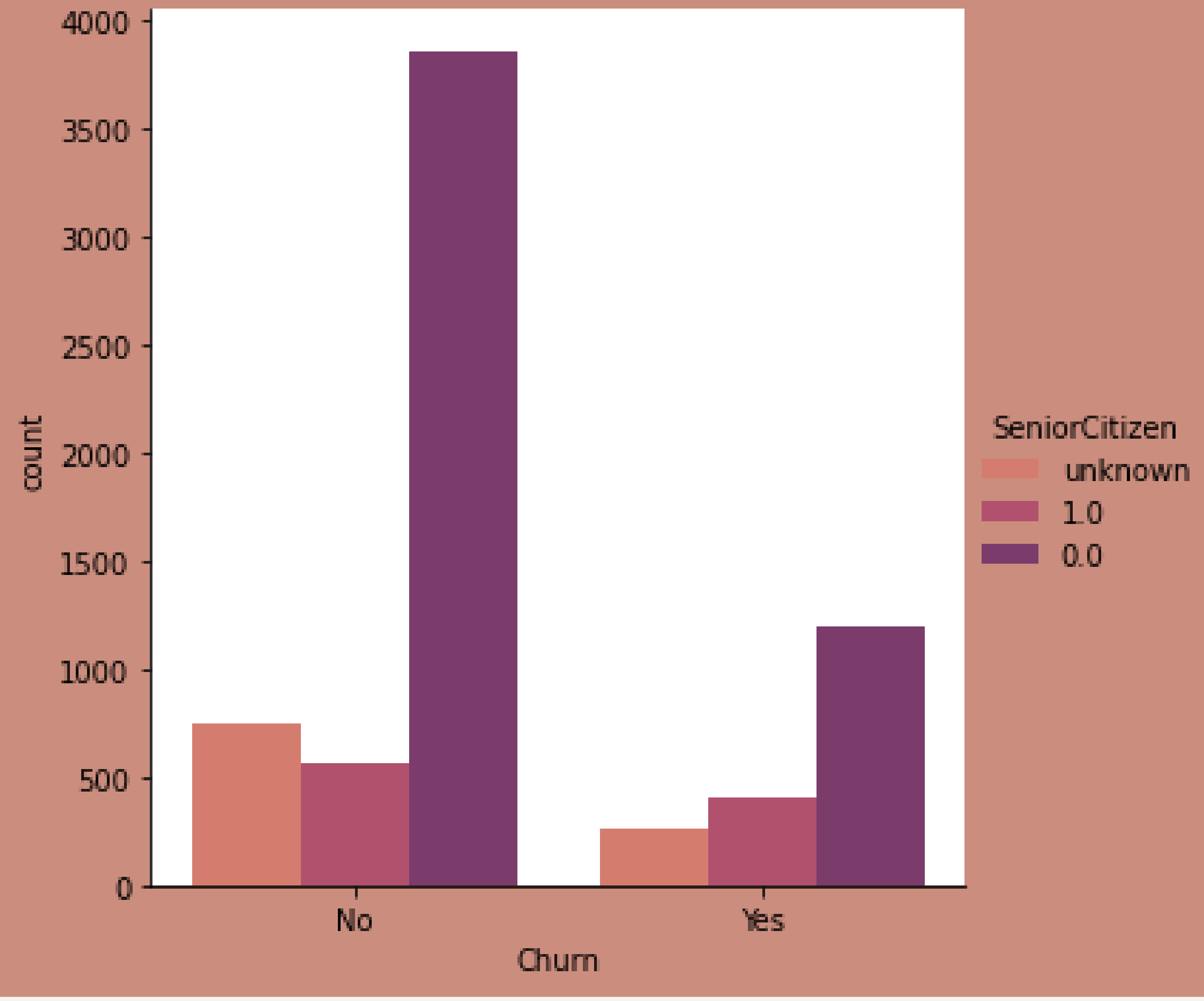
customerID			
gender	Female	Male	unknown
Churn			
No	31.05	31.79	10.62
Yes	11.44	11.39	3.71

customerID			
gender	Female	Male	unknown
Churn			
No	2187	2239	748
Yes	806	802	261
None			

- THERE ARE CUSTOMERS DATA OF UNKNOWN GENDER
- MORE MALE CUSTOMERS THAN FEMALE CUSTOMERS

- THERE ARE 31.79% MALE NO CHURN CUSTOMERS AND 11.39% MALE CHURN CUSTOMERS
- THERE ARE 31.05% FEMALE NO CHURN CUSTOMERS AND 11.44% FEMALE CHURN CUSTOMERS.

BY SENIOR CITIZEN

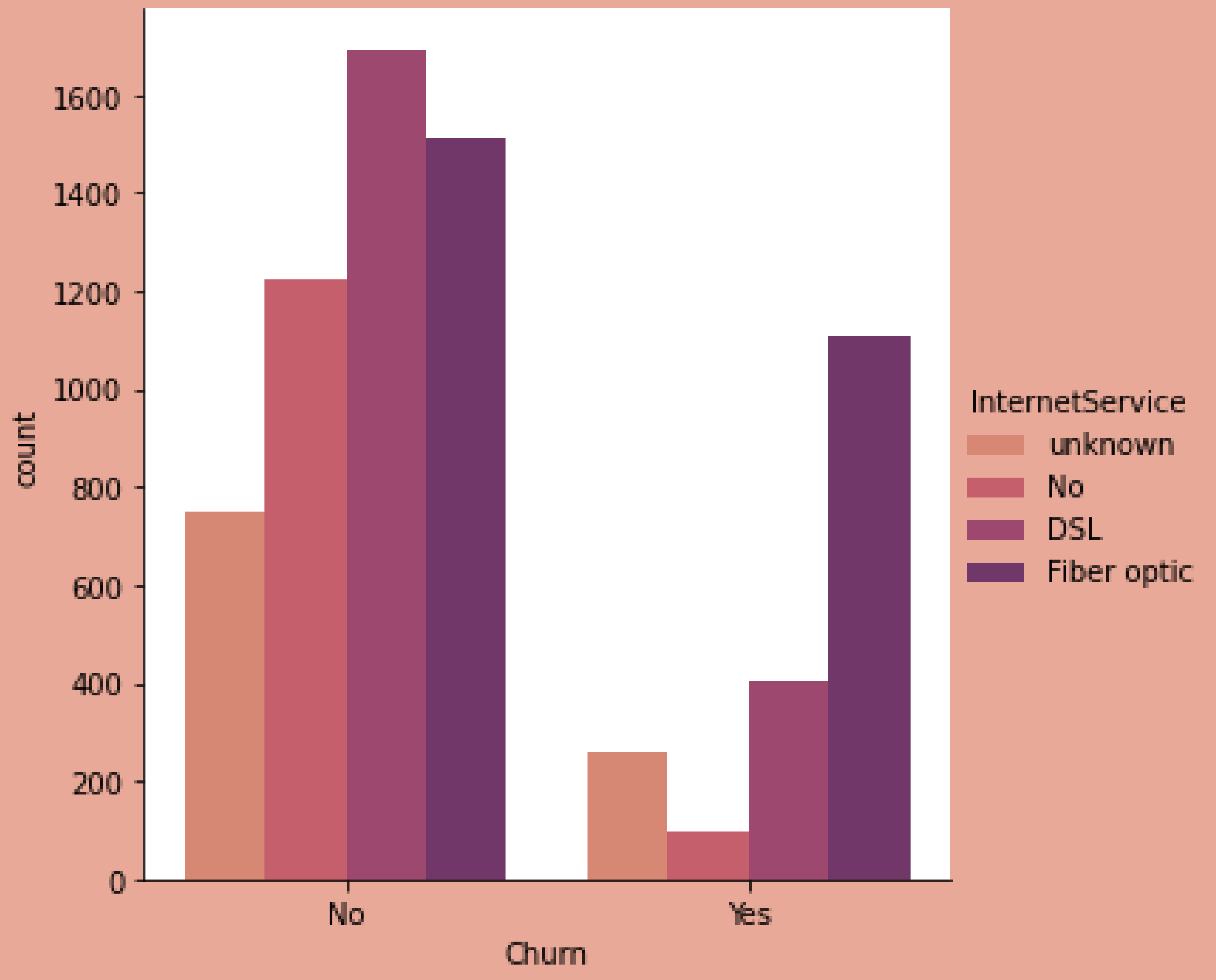



```
0.0        3860
unknown     748
1.0         566
Name: SeniorCitizen, dtype: int64
```

```
0.0        1197
1.0         411
unknown     261
Name: SeniorCitizen, dtype: int64
```

FROM CHURN CUSTOMERS, THE MAJORITY ARE NOT SENIORS, WHICH IS AS MUCH AS 17%. 5.84% OTHER ARE SENIOR CITIZENS AND 3.71% ARE CHURN CUSTOMERS WHO ARE UNKNOWN. WHILE THE REMAINING 73.45% ARE NO CHURN CUSTOMERS.

BY INTERNET SERVICE



format in %

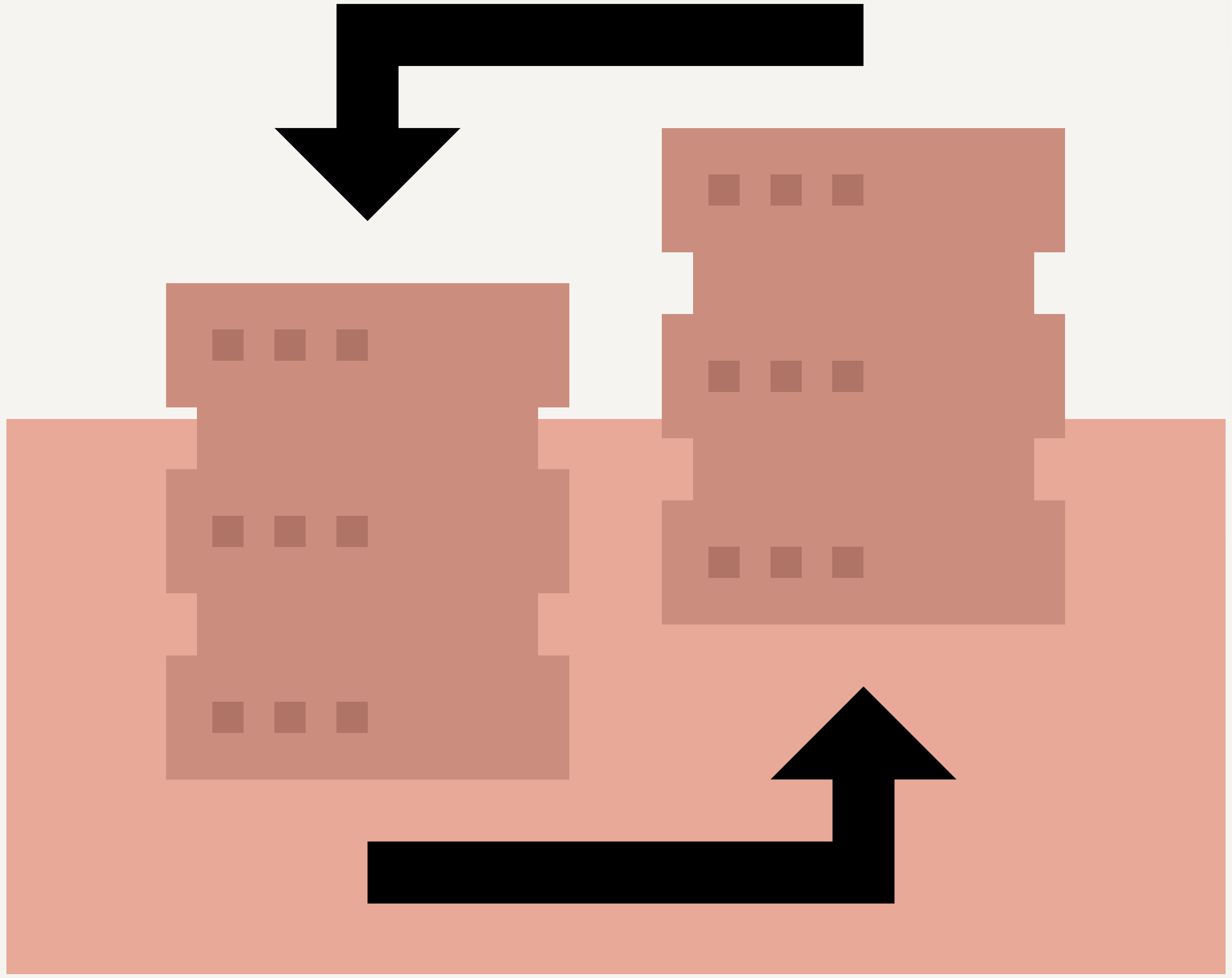
InternetService	customerID DSL	Fiber optic	No	unknown
Churn				
No	24.01	21.47	17.36	10.62
Yes	5.69	15.75	1.39	3.71

InternetService	customerID DSL	Fiber optic	No	unknown
Churn				
No	1691	1512	1223	748
Yes	401	1109	98	261

None

MOST OF THE CHURN CUSTOMERS ARE CUSTOMERS WHO USE FIBER OPTICS INTERNET SERVICES ARE 15.75%. 5.69% DSL USERS, 3.71% UNKNOWN, AND 1.39% DID NOT HAVE INTERNET SERVICES. WHILE THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

BY PAPERLESS BILLING

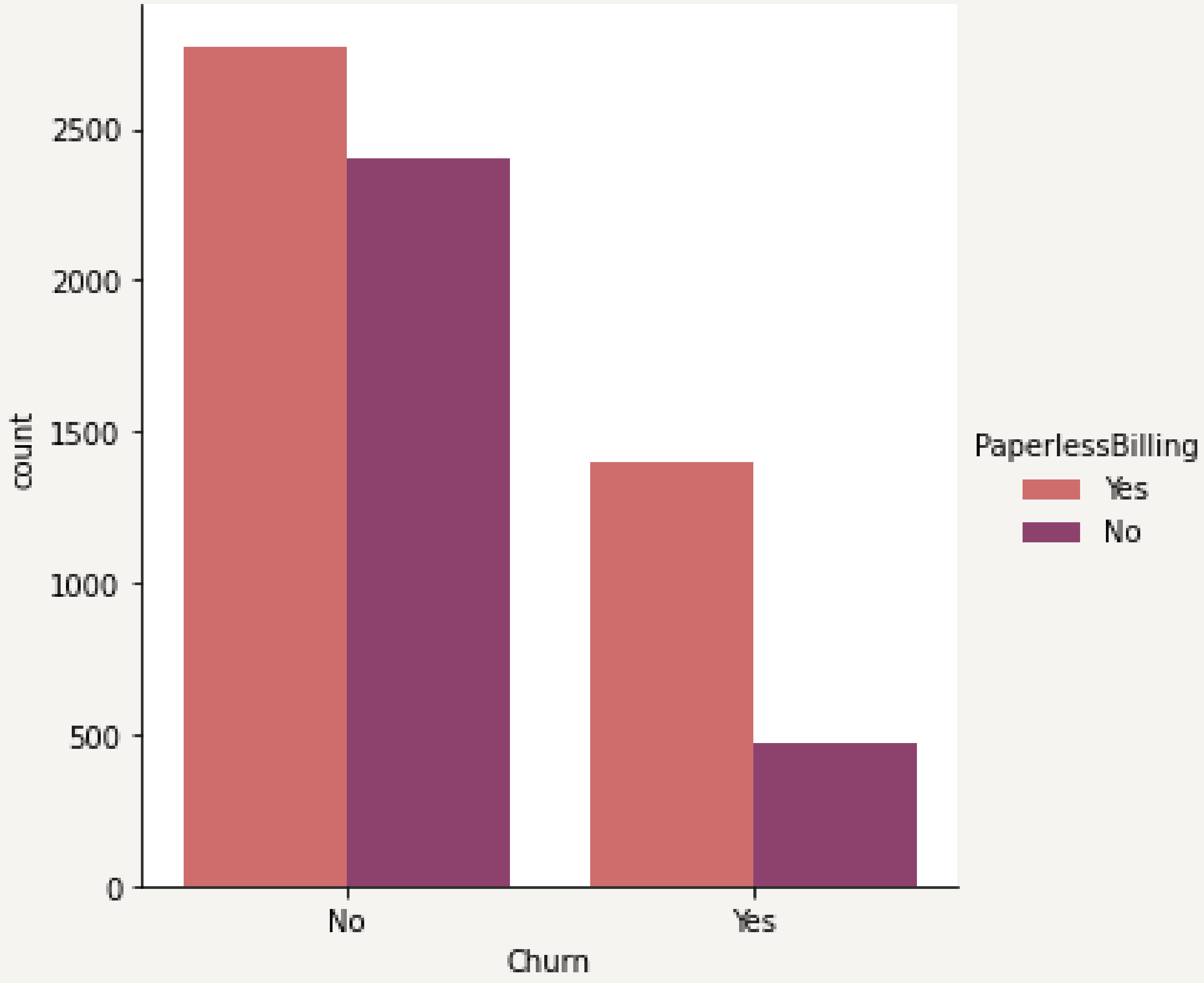


format in %

PaperlessBilling / Churn	customerID No	customerID Yes
No	34.12	39.34
Yes	6.66	19.88

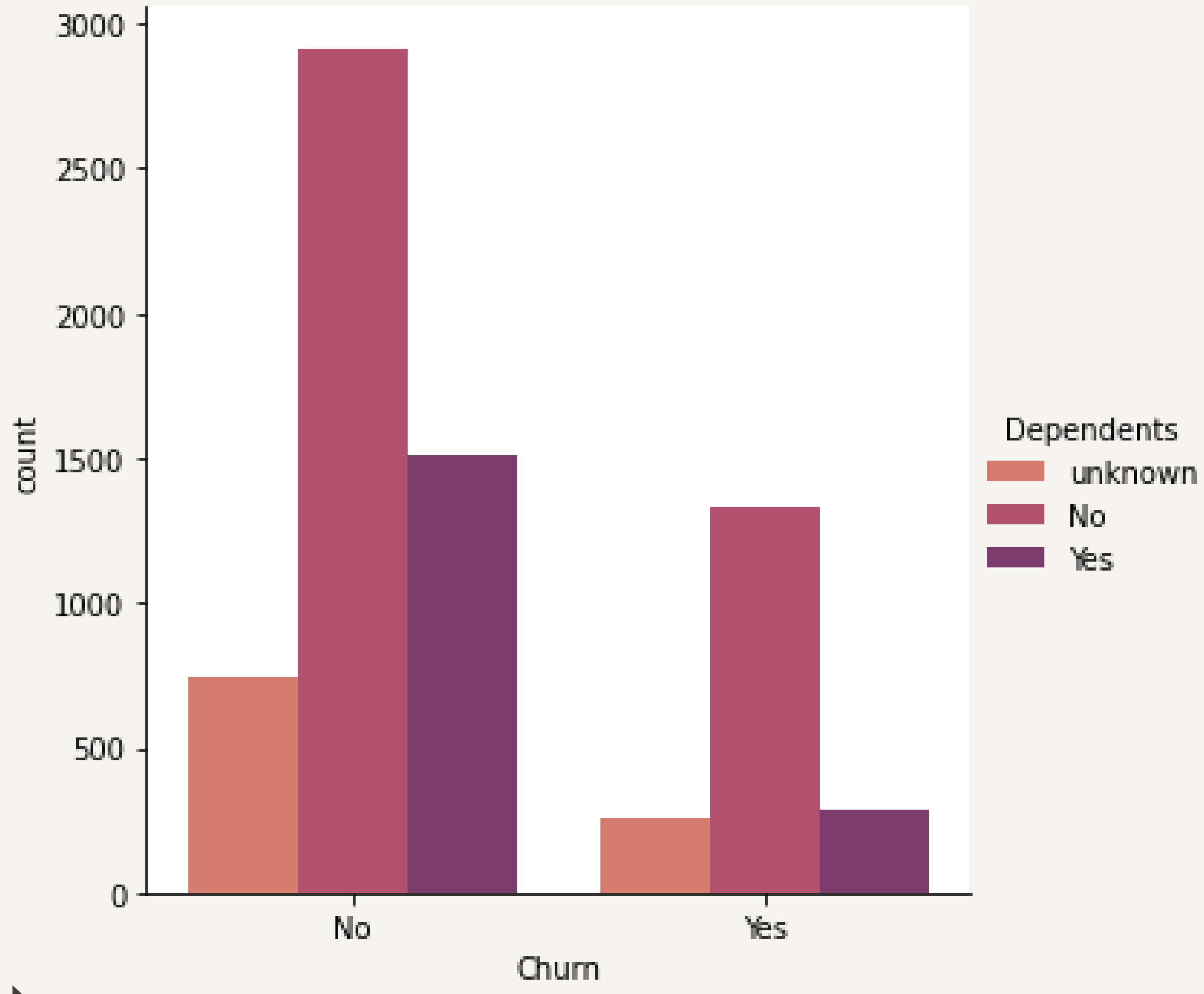
PaperlessBilling / Churn	customerID No	customerID Yes
No	2403	2771
Yes	469	1400

None



Most churn customers are customers who use electronic transactions (paperless billing) that are 19.88% while 6.66% were not paperless billing users. The remaining 73.46% are no churn customers.

BY DEPENDENTS



format in %

Dependents / Churn	customerID No	customerID Yes	customerID unknown
No	41.35	21.5	10.62
Yes	18.83	4.0	3.71

Dependents / Churn	customerID No	customerID Yes	customerID unknown
No	2912	1514	748
Yes	1326	282	261

None

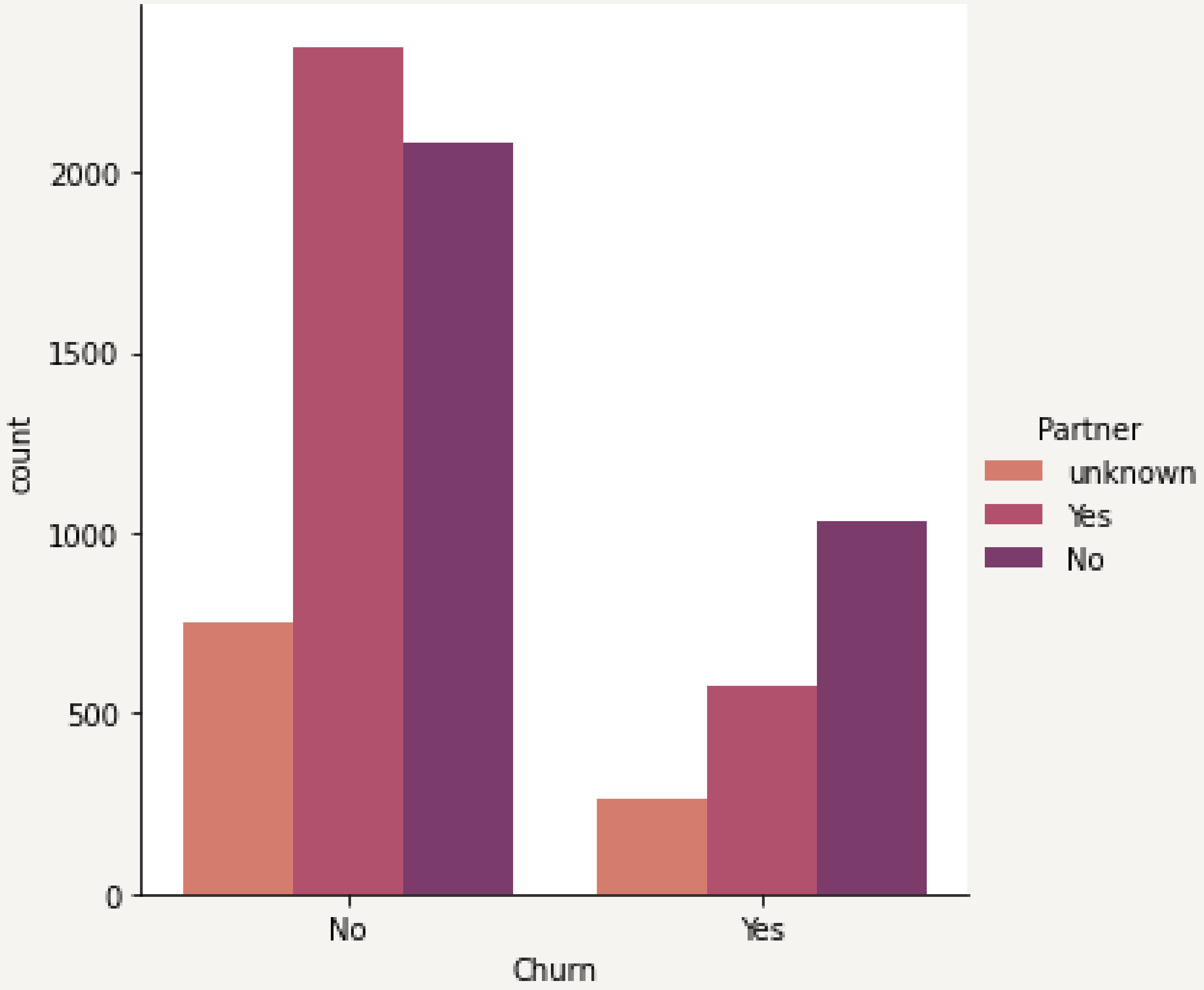
- **FROM THE NO CHURN CUSTOMERS, THE MOST ARE CUSTOMERS WHO HAVE NO DEPENDENTS**
- **FROM THE CHURN CUSTOMERS, THE MOST ARE CUSTOMERS WHO HAVE NO DEPENDENTS**

18.83% OF CUSTOMERS ARE CHURN CUSTOMERS AND HAVE NO DEPENDENTS. WHILE 4% OF THEM HAVE DEPENDENTS AND 3.71% ARE NOT KNOWN. WHILE THE REMAINING 73.47% ARE NO CHURN CUSTOMERS.

BY PARTNER

EDA

Most churn customers are customers who don't have a partner, that is as much as 14.65% of all numbers customers. 8.18% have a partner and 3.71% are not known. While the remaining 73.46% are no churn customers.



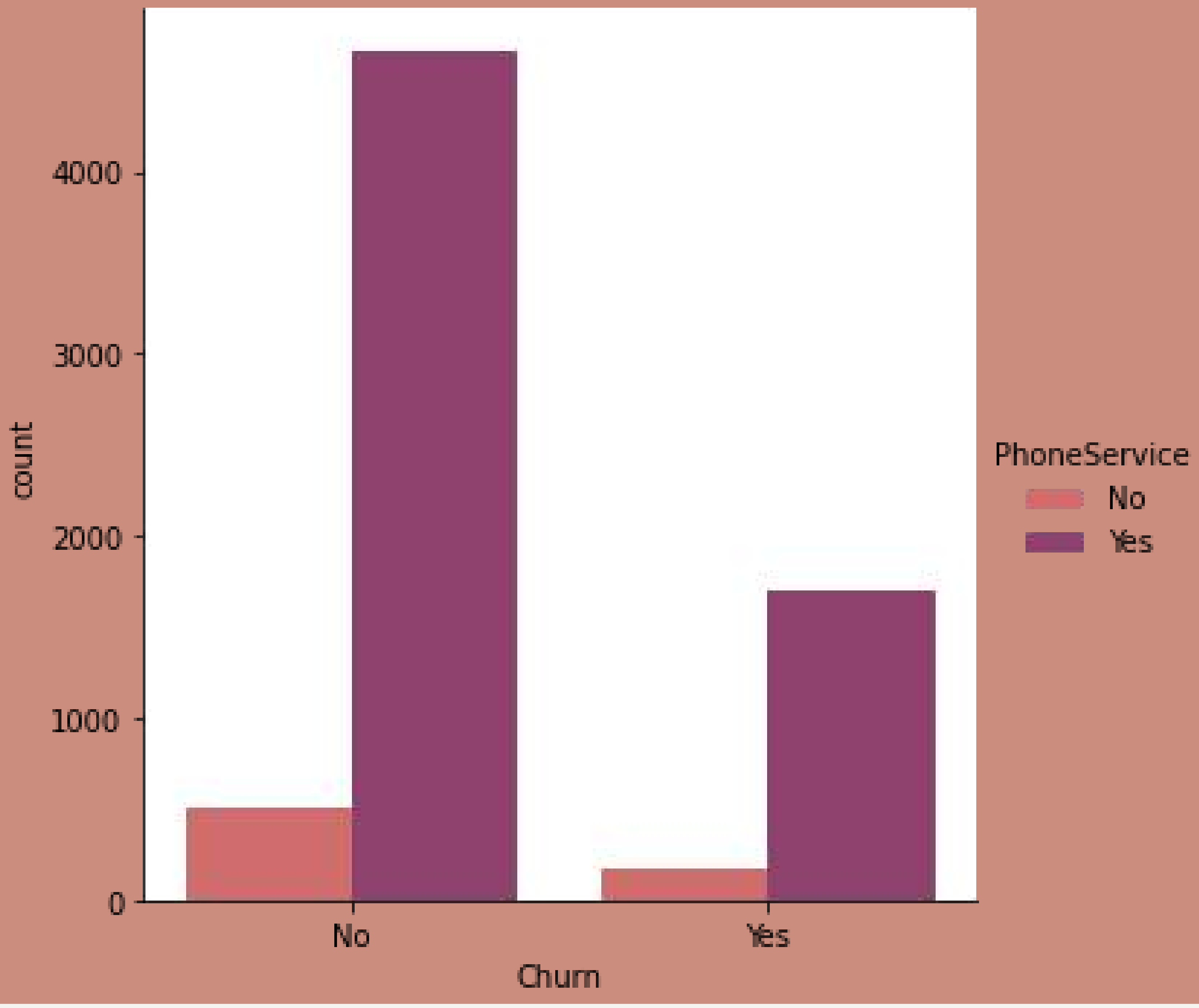
format in %

customerID			
Partner	No	Yes	unknown
Churn			
No	29.52	33.32	10.62
Yes	14.65	8.18	3.71

customerID			
Partner	No	Yes	unknown
Churn			
No	2079	2347	748
Yes	1032	576	261

None

BY PHONE SERVICE



format in %

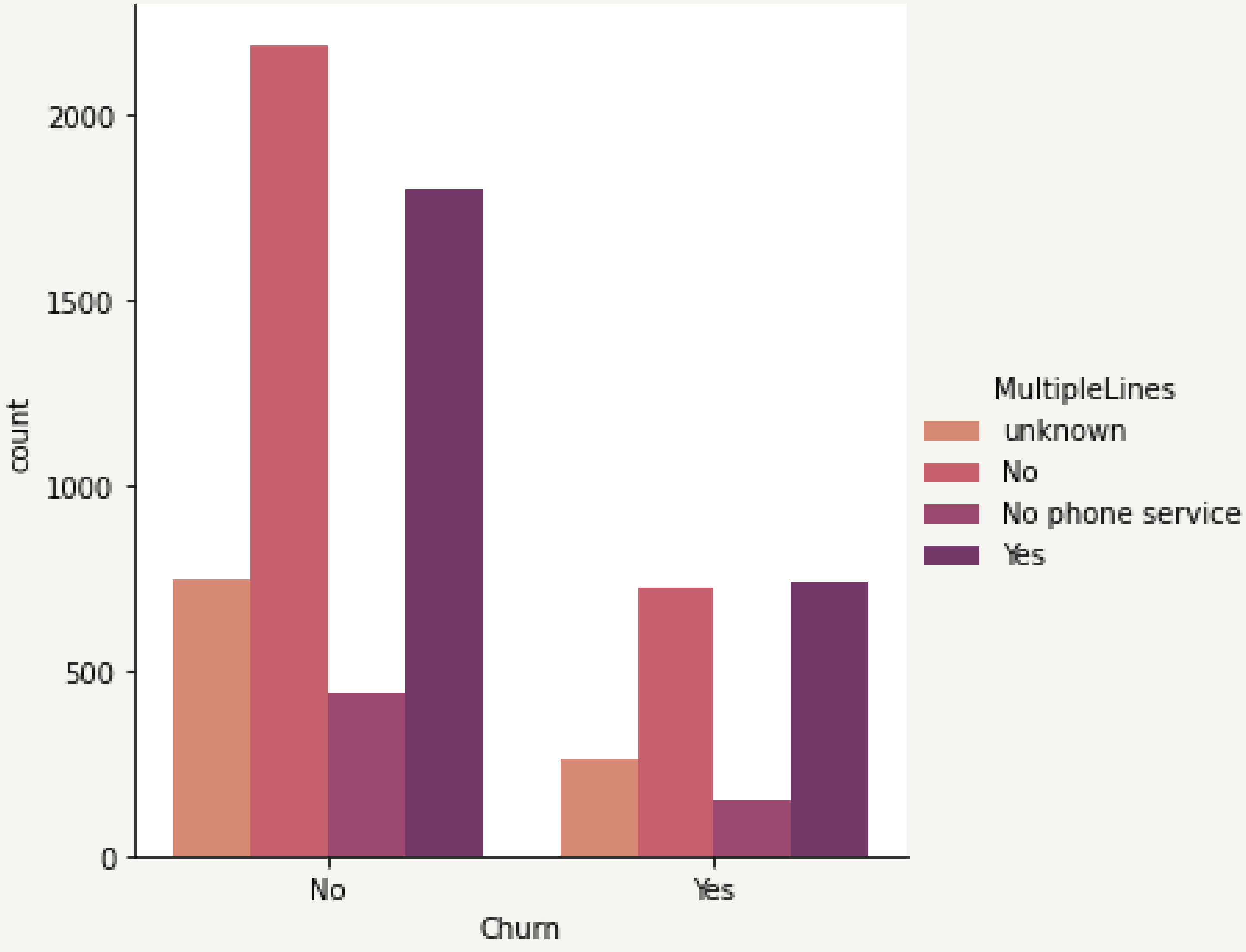
	customerID	
PhoneService	No	Yes
Churn		
No	7.27	66.19
Yes	2.41	24.12

	customerID	
PhoneService	No	Yes
Churn		
No	512	4662
Yes	170	1699

None

- MOST CHURN AND NO CHURN CUSTOMER HAVE PHONE SERVICES THAN THOSE WHO DO NOT.
- 24.12% OF CUSTOMERS ARE CHURN CUSTOMERS AND HAVE PHONE SERVICES. WHILE 2.41% WERE NOT HAVE PHONE SERVICE. THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

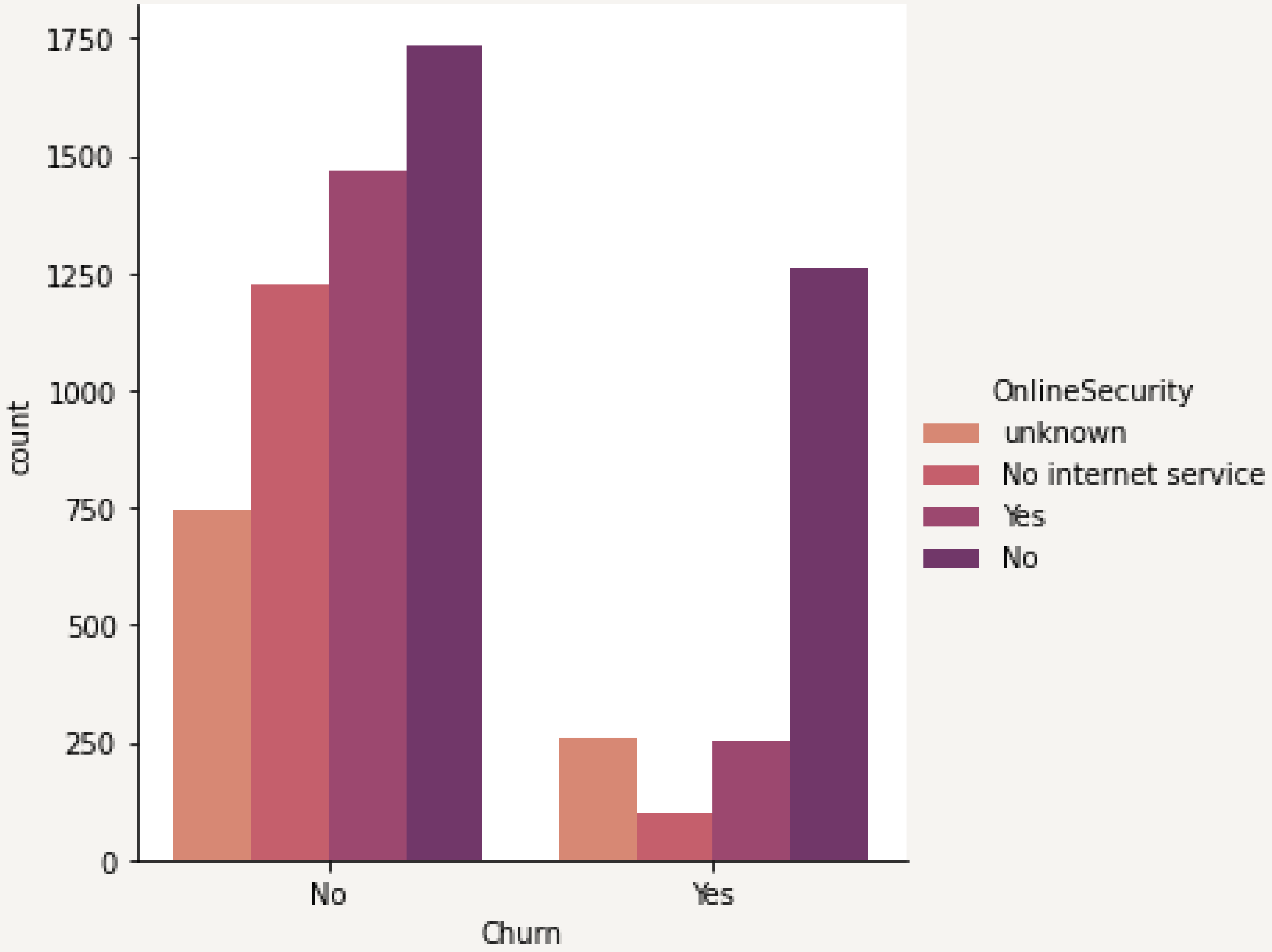
BY MULTIPLE LINES




```
        format in %
             customerID
MultipleLines          No No phone service    Yes unknown
Churn
No                  31.01             6.28  25.56   10.62
Yes                 10.28             2.09  10.46    3.71

             customerID
MultipleLines          No No phone service   Yes unknown
Churn
No                   2184              442  1800     748
Yes                   724              147   737     261
None
```

MOST CHURN CUSTOMERS ARE CUSTOMERS WHO HAVE MULTIPLE LINES, THAT ARE 10.46% OF CUSTOMERS. 10.28% DID NOT HAVE MULTIPLE LINES, 3.71% OF IT IS UNKNOWN, AND 2.09% DID NOT HAVE PHONE SERVICES. WHILE THE REMAINING 73.47% ARE NO CHURN CUSTOMERS.

BY ONLINE SECURITY



format in %

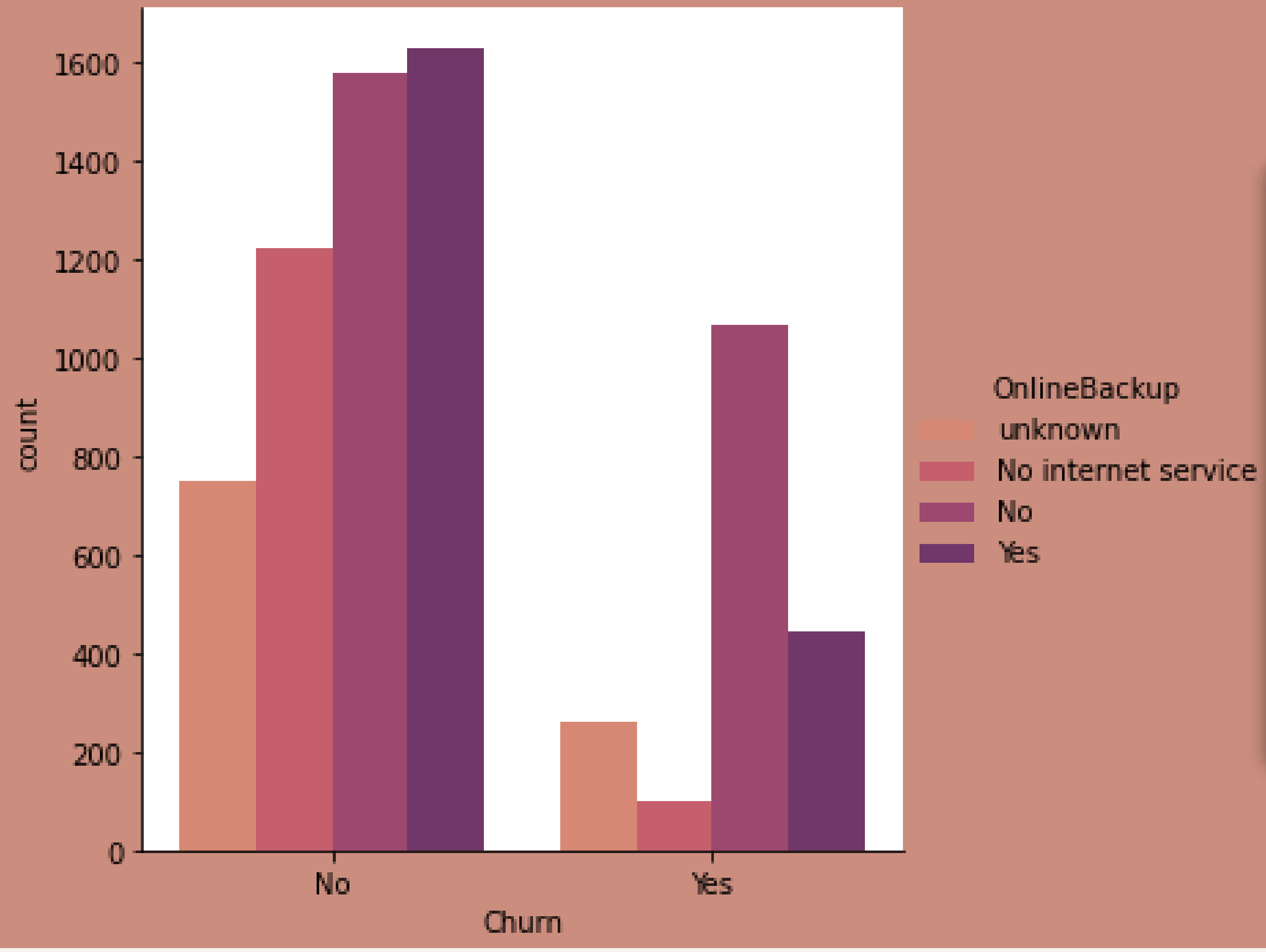
Churn \ OnlineSecurity (customerID)	No	No internet service	Yes	unknown
No	24.62	17.36	20.86	10.62
Yes	17.88	1.39	3.56	3.71

Churn \ OnlineSecurity (customerID)	No	No internet service	Yes	unknown
No	1734	1223	1469	748
Yes	1259	98	251	261

None

Most churn customers are customers who do not have online security, which is 17.88% customers. 3.56% have online security, 3.71% of them are unknown, and 1.39% do not have internet services. The remaining 73.46% are no churn customers.

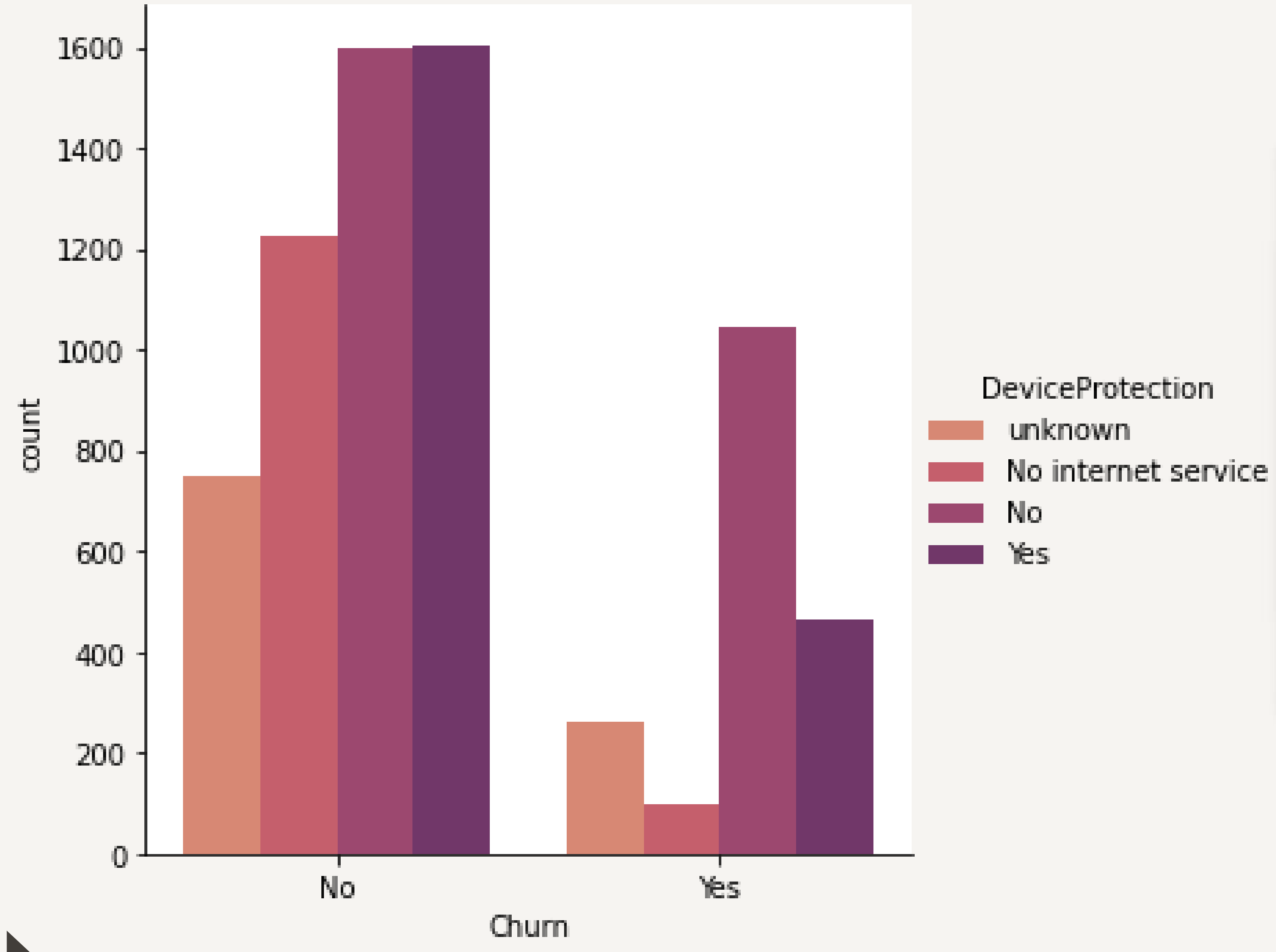
BY ONLINE BACKUP




```
          format in %
              customerID
OnlineBackup          No  No internet service    Yes  unknown
Churn
No                 22.38                17.36  23.10    10.62
Yes                15.15                 1.39   6.29     3.71

              customerID
OnlineBackup          No  No internet service   Yes  unknown
Churn
No                  1576                 1223  1627      748
Yes                 1067                   98   443      261
None
```

MOST CHURN CUSTOMERS ARE CUSTOMERS WHO DO NOT HAVE ONLINE BACKUP, WHICH IS 15.15% CUSTOMERS. 6.29% HAVE ONLINE BACKUP, 3.71% OF THEM ARE UNKNOWN, AND 1.39% DO NOT HAVE INTERNET SERVICES. THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

BY DEVICE PROTECTION



format in %

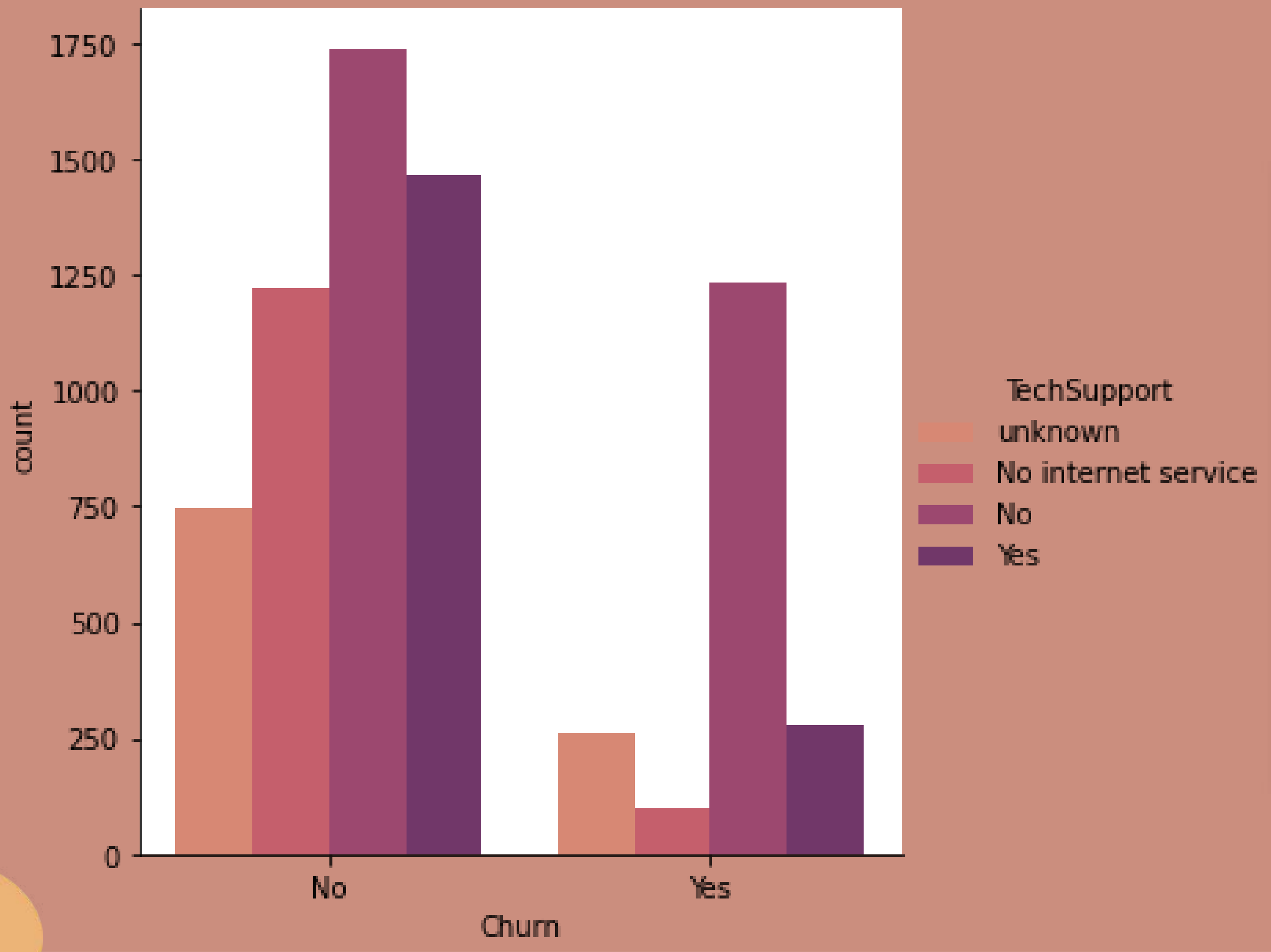
Churn \ DeviceProtection (customerID)	No	No internet service	Yes	unknown
No	22.72	17.36	22.76	10.62
Yes	14.81	1.39	6.63	3.71

Churn \ DeviceProtection (customerID)	No	No internet service	Yes	unknown
No	1600	1223	1603	748
Yes	1043	98	467	261

None

MOST CHURN CUSTOMERS ARE CUSTOMERS WHO HAVE NO DEVICE PROTECTION, WHICH IS 14.81% OF CUSTOMERS. 6.63% OF IT HAVE DEVICE PROTECTION, 3.71% OF THEM ARE UNKNOWN, AND 1.39% DO NOT HAVE INTERNET SERVICES. THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

BY TECH SUPPORT



format in %

customerID

TechSupport	No	No internet service	Yes	unknown
Churn				
No	24.68	17.36	20.8	10.62
Yes	17.54	1.39	3.9	3.71

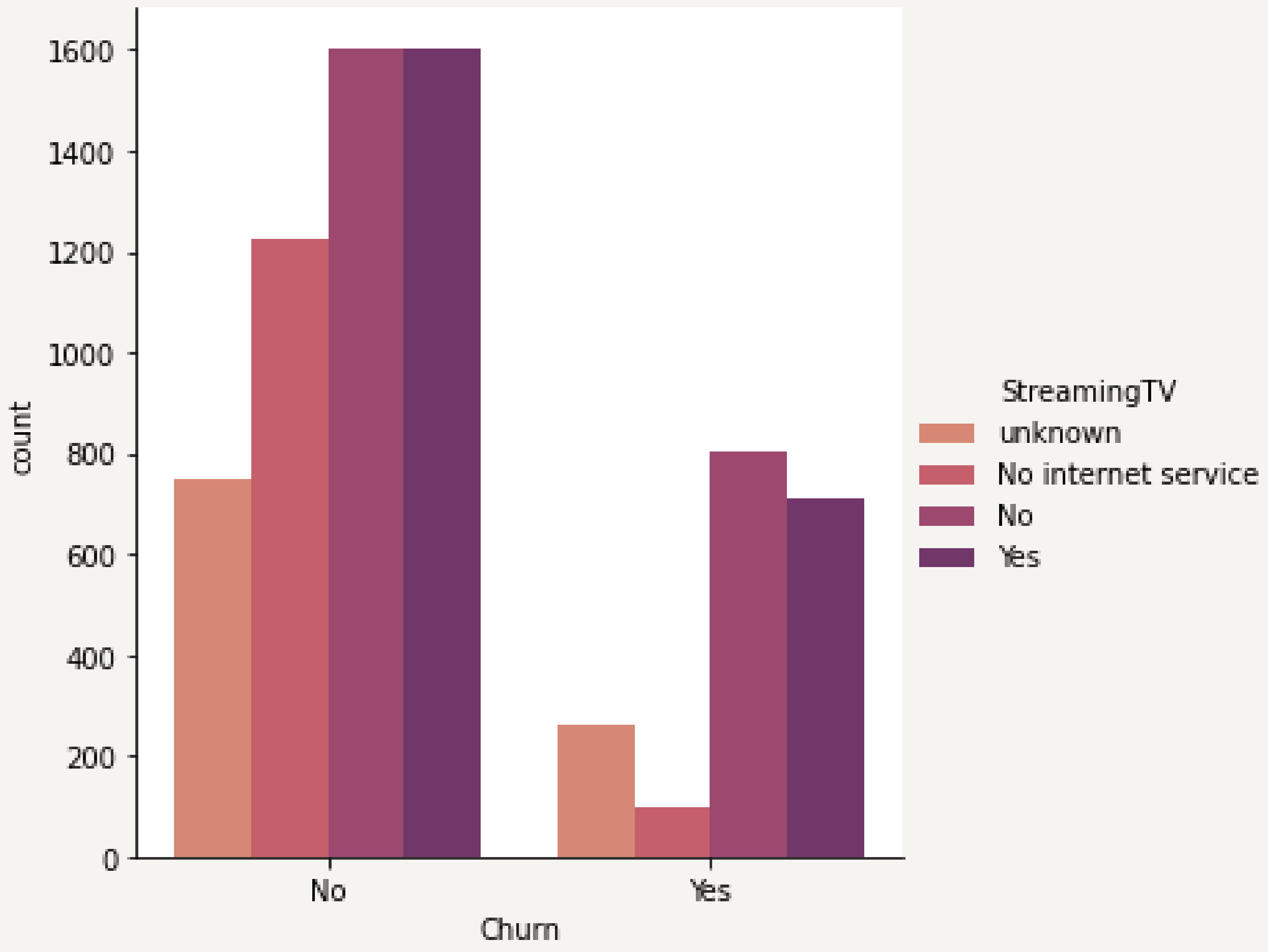
customerID

TechSupport	No	No internet service	Yes	unknown
Churn				
No	1738	1223	1465	748
Yes	1235	98	275	261

None

MOST CHURN CUSTOMERS ARE CUSTOMERS WHO DON'T HAVE TECHNICAL SUPPORT SERVICES, THAT IS 17.54% OF CUSTOMERS. 3.9% OF THEM HAVE A TECHNICAL SUPPORT, 3.71% OF THEM ARE UNKNOWN, AND 1.39% DO NOT HAVE INTERNET SERVICES. THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

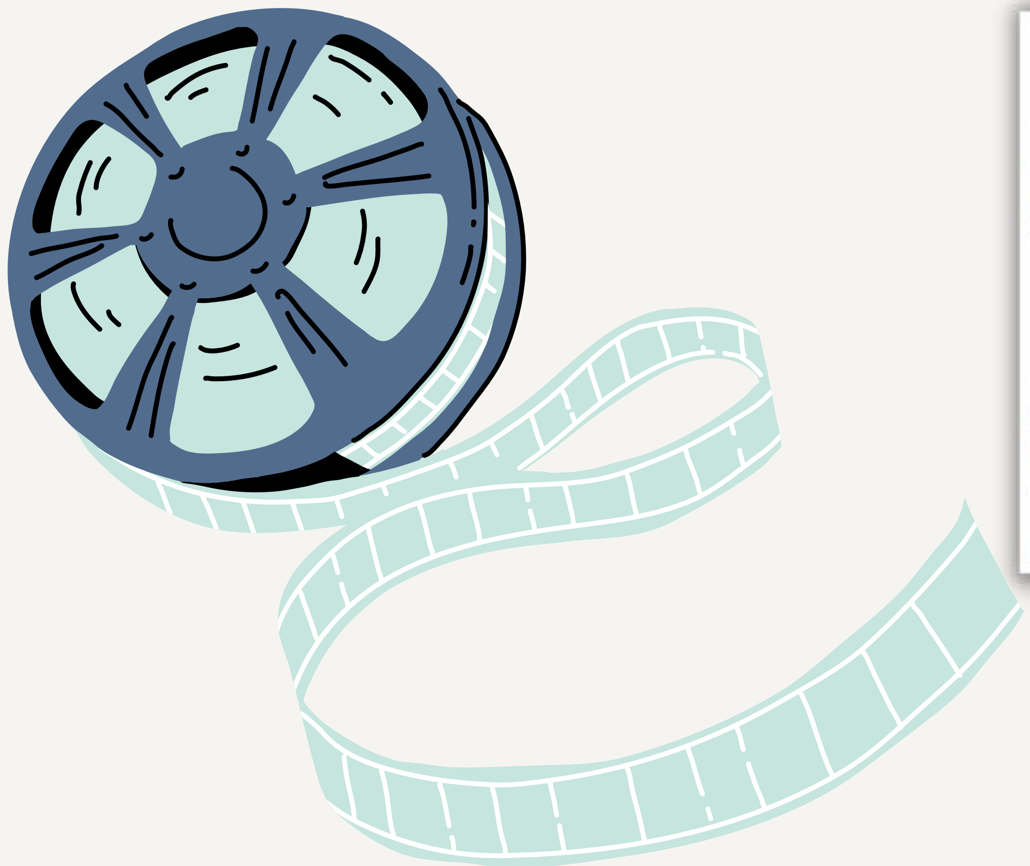
BY STREAMING TV




```
         format in %
            customerID
StreamingTV         No No internet service    Yes unknown
Churn
No               22.73               17.36  22.75   10.62
Yes              11.39                1.39  10.05    3.71

            customerID
StreamingTV         No No internet service   Yes unknown
Churn
No                1601                1223  1602     748
Yes                802                  98   708     261
None
```

Most churn customers are customers who do not use Streaming TV services, which is 11.39% of customers. 10.05% of it uses streaming TV services, 3.71% of which are unknown, and 1.39% have no internet services. The remaining 73.46% are no churn customers.

BY STREAMING MOVIES

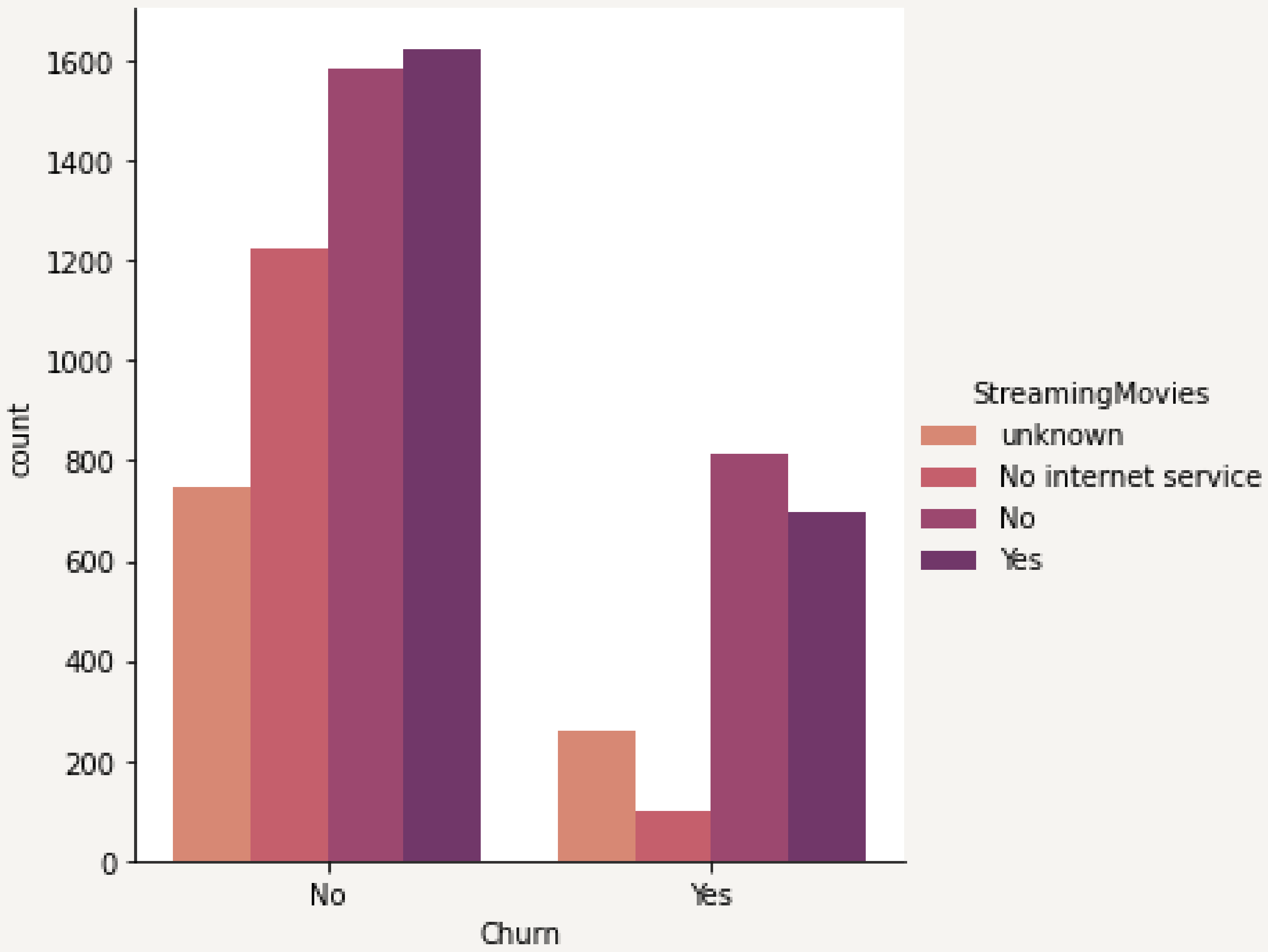


format in %

StreamingMovies / Churn	No	No internet service	Yes	unknown
No	22.45	17.36	23.03	10.62
Yes	11.54	1.39	9.90	3.71

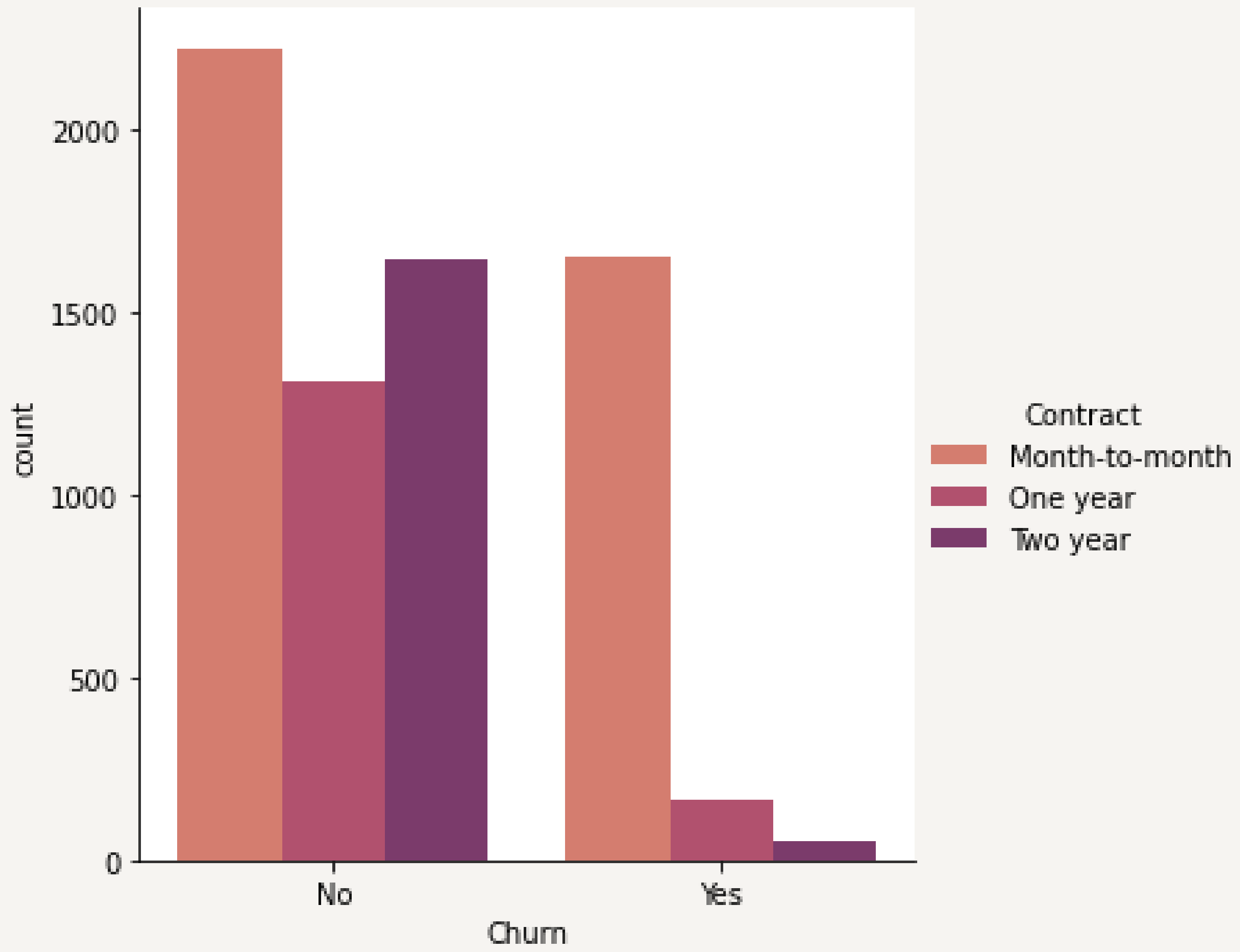
StreamingMovies / Churn	No	No internet service	Yes	unknown
No	1581	1223	1622	748
Yes	813	98	697	261

None



Most churn customers are customers who do not use Streaming Movies service, which is 11.54% of customers. 9.9% of it using Streaming Movies service, 3.71% of it is unknown, and 1.39% have no internet services. The remaining 73.46% are no churn customers.

BY CONTRACT



format in %

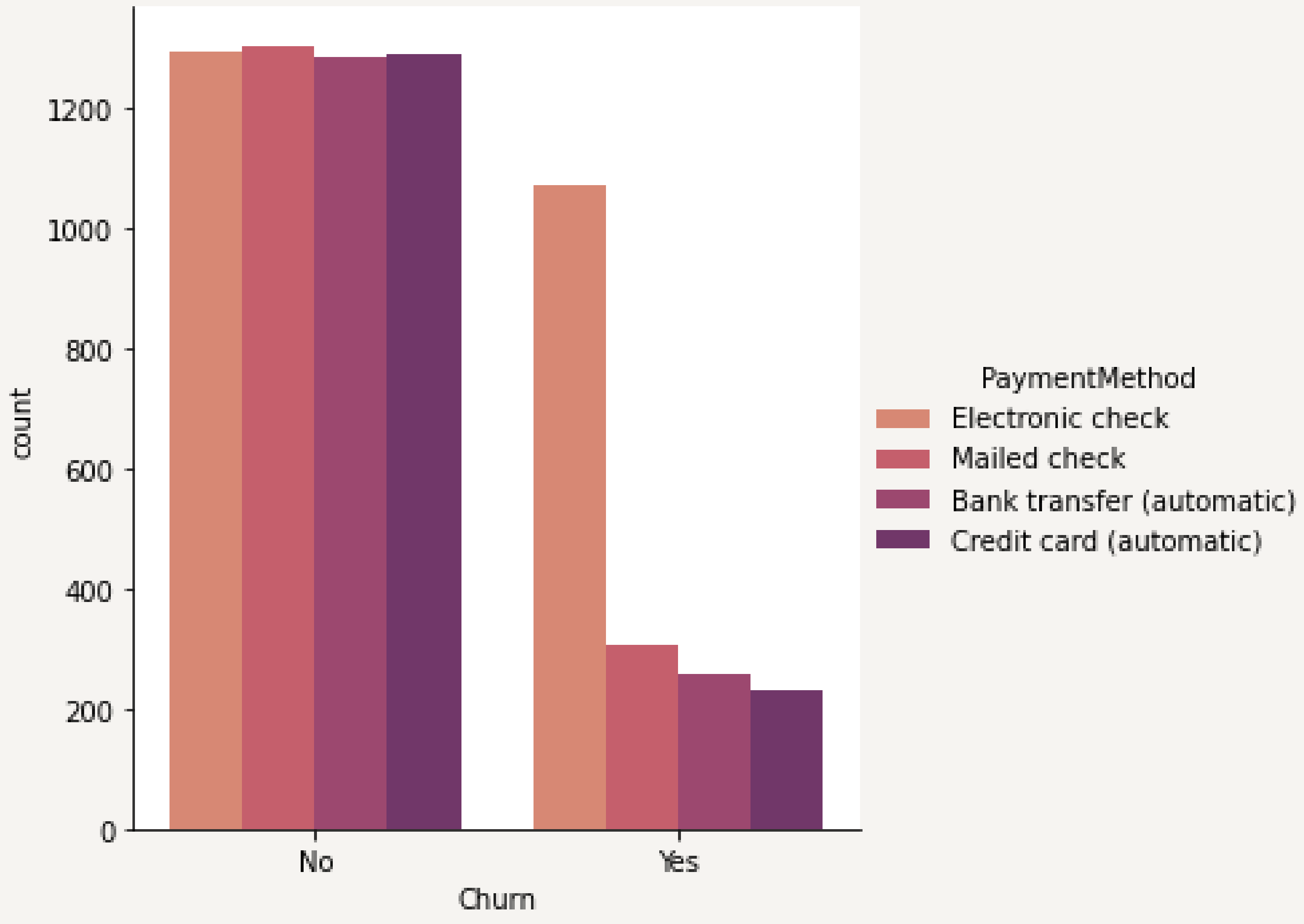
Churn \ Contract (customerID)	Month-to-month	One year	Two year
No	31.52	18.56	23.38
Yes	23.50	2.36	0.68

Churn \ Contract (customerID)	Month-to-month	One year	Two year
No	2220	1307	1647
Yes	1655	166	48

None

MOST CHURN CUSTOMERS ARE CUSTOMERS WHO HAVE MONTHLY CONTRACTS, THAT ARE 23.50% OF CUSTOMERS. 2.36% OF IT HAD A ONE-YEAR CONTRACT, AND 0.68% HAD A TWO-YEAR CONTRACT. THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

BY PAYMENT METHOD

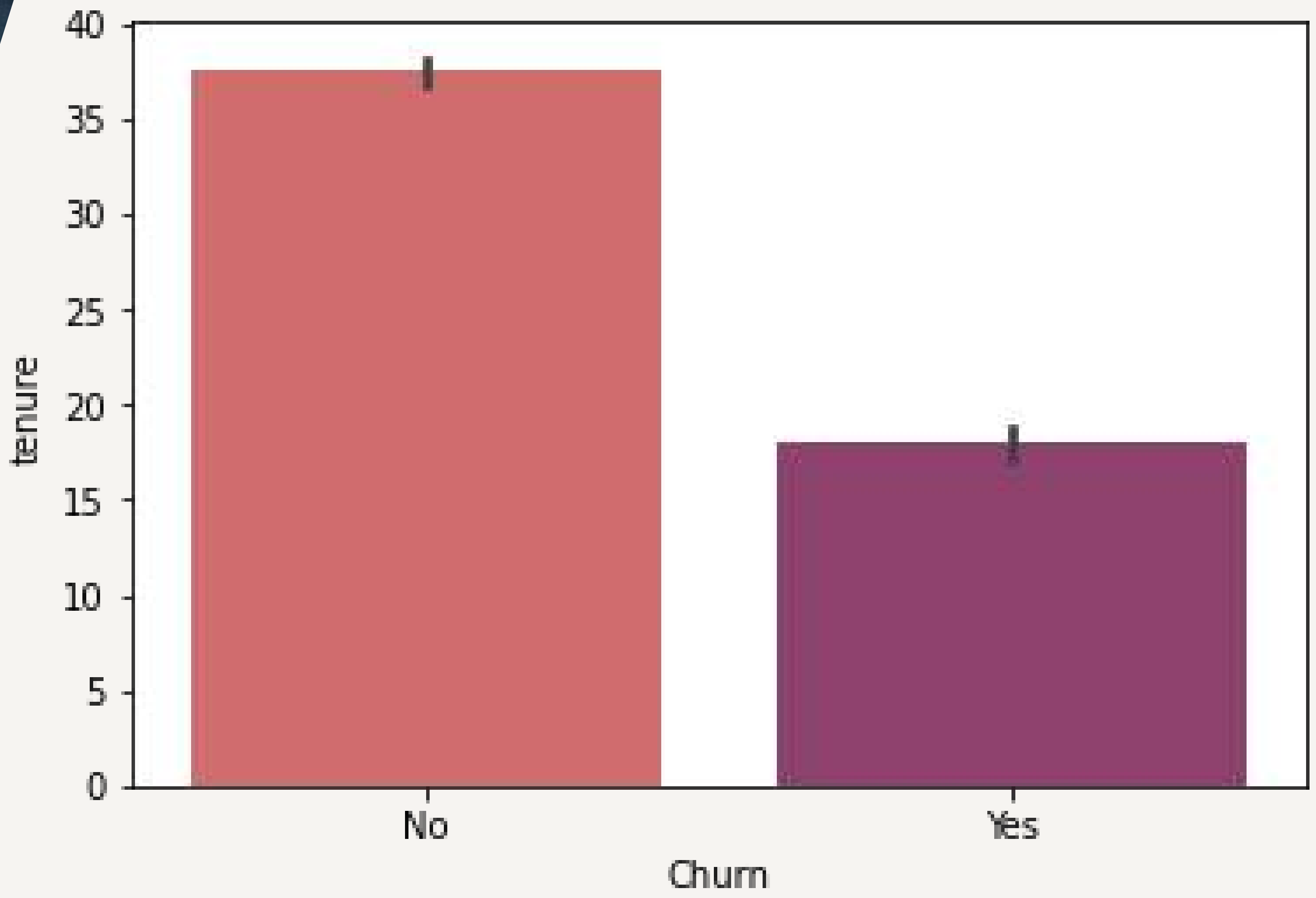



```
          format in %
                                 customerID                          \
PaymentMethod Bank transfer (automatic) Credit card (automatic)
Churn
No                                18.26                    18.32
Yes                                3.66                     3.29


PaymentMethod Electronic check Mailed check
Churn
No                       18.37        18.51
Yes                      15.21         4.37
```

```
                                 customerID                          \
PaymentMethod Bank transfer (automatic) Credit card (automatic)
Churn
No                                 1286                     1290
Yes                                 258                      232


PaymentMethod Electronic check Mailed check
Churn
No                        1294         1304
Yes                       1071          308
None
```

15.21% OF CUSTOMERS ARE CHURN CUSTOMERS AND USE ELEKTRONIC CHECK PAYMENTS METHOD. 4.37% OF THEM USED MAILED CHECK, 3.66% OF THEM USE BANK TRANSFERS, AND 3.29% ARE CHURN CUSTOMERS AND USE CREDIT CARDS. THE REMAINING 73.46% ARE NO CHURN CUSTOMERS.

BY TENURE



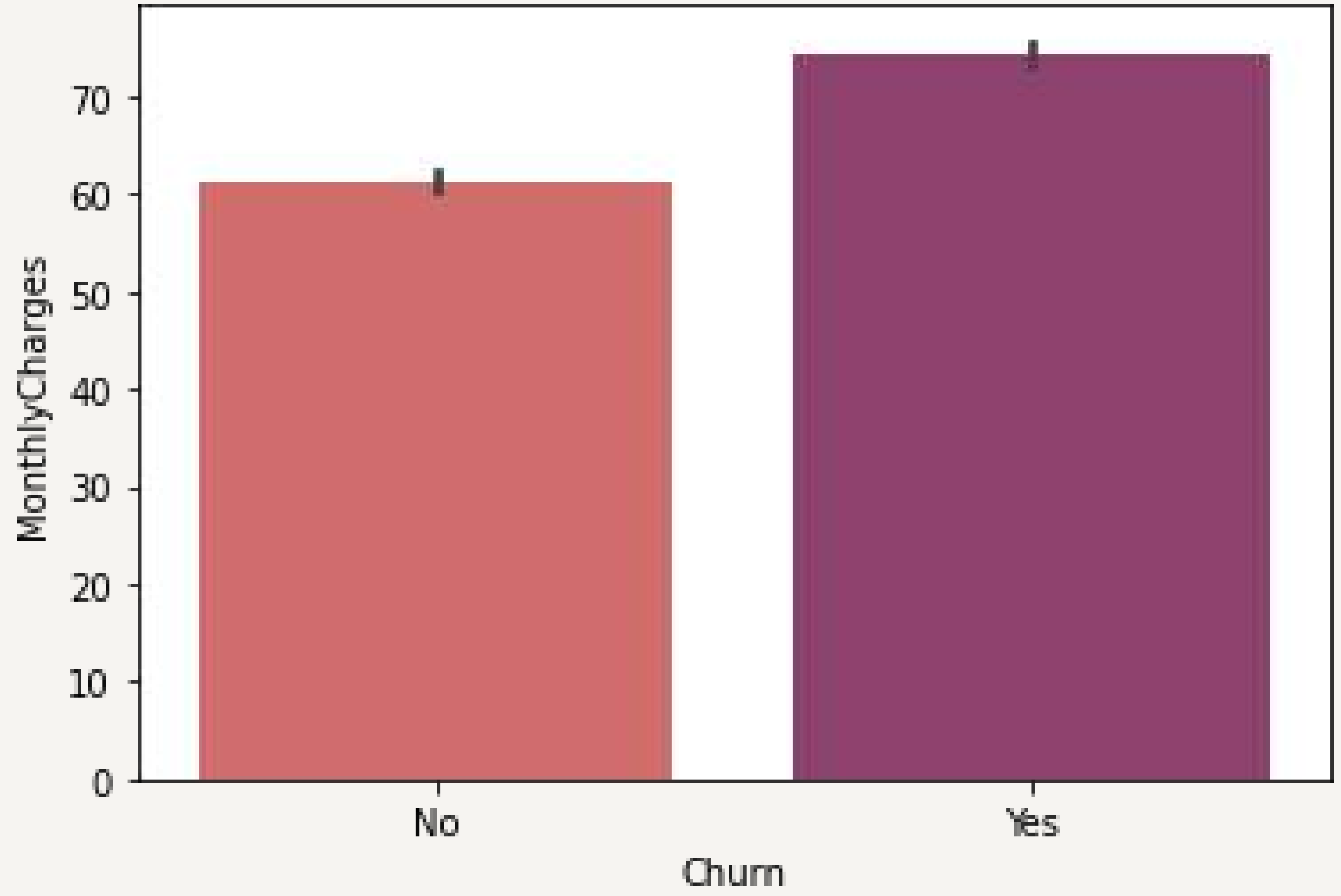
	tenure	MonthlyCharges	TotalCharges
count	1869.000000	1869.000000	1869.000000
mean	17.979133	74.441332	1531.796094
std	19.531123	24.666053	1890.822994
min	1.000000	18.850000	18.850000
25%	2.000000	56.150000	134.500000
50%	10.000000	79.650000	703.550000
75%	29.000000	94.200000	2331.300000
max	72.000000	118.350000	8684.800000

```
Modus No Churn:   72
Modus Churn:   1
```

- The average tenure of no churn customers more than churn customers.
- Most churn customers are customers whose tenure is only 1 month with an average is 17.9 months.

BY MONTHLY CHARGES

EDA

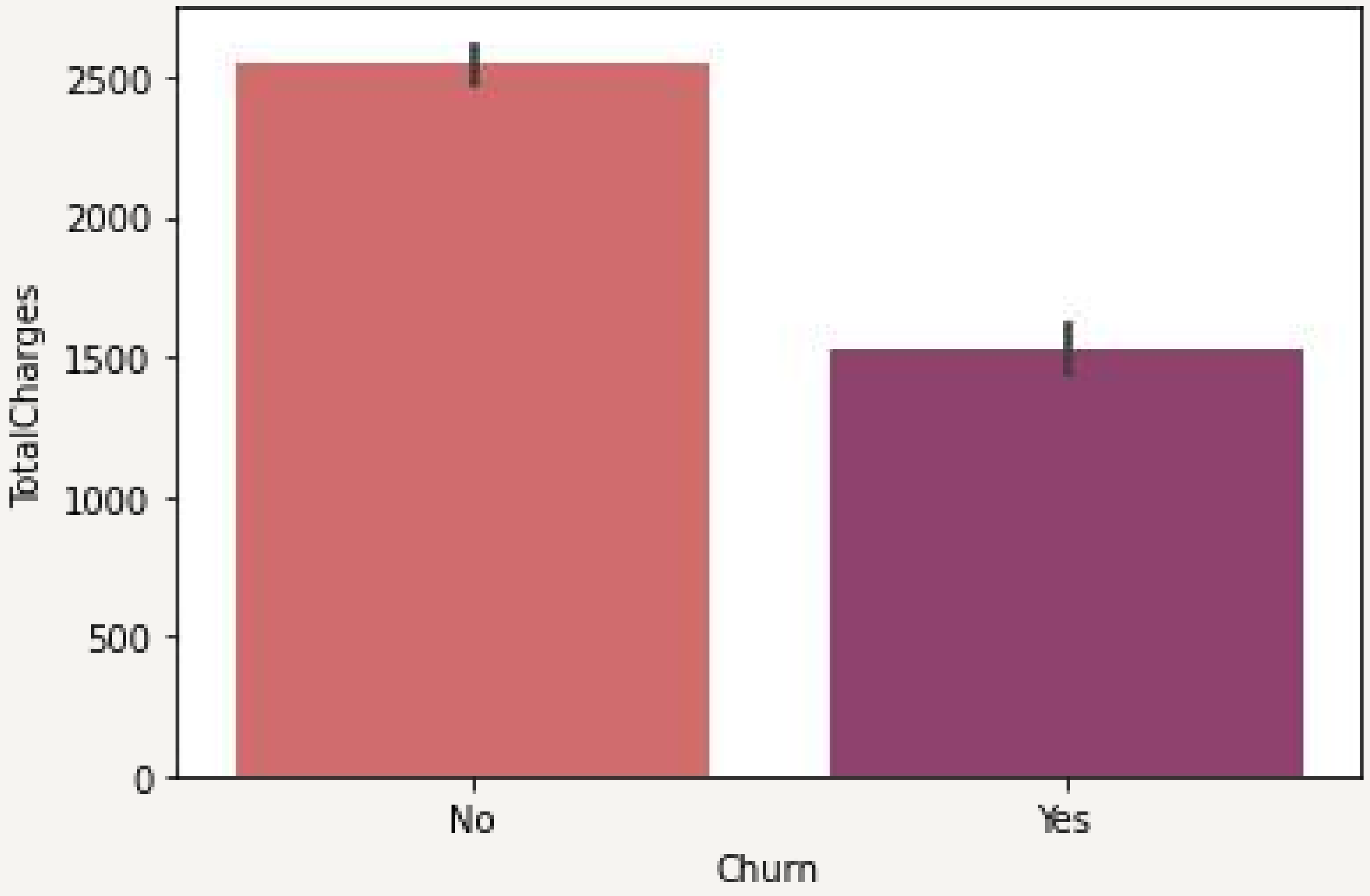


	tenure	MonthlyCharges	TotalCharges
count	1869.000000	1869.000000	1869.000000
mean	17.979133	74.441332	1531.796094
std	19.531123	24.666053	1890.822994
min	1.000000	18.850000	18.850000
25%	2.000000	56.150000	134.500000
50%	10.000000	79.650000	703.550000
75%	29.000000	94.200000	2331.300000
max	72.000000	118.350000	8684.800000

```
Modus No Churn:  20.05
Modus Churn:  74.4
```

- **THE AVERAGE MONTHLY COSTS INCURRED BY CHURN CUSTOMERS ARE MORE THAN NO CHURN CUSTOMERS.**
- **MOST CHURN CUSTOMERS ON AVERAGE ARE CUSTOMERS WHO SPEND 74.4 FOR MONTHLY CHARGES.**

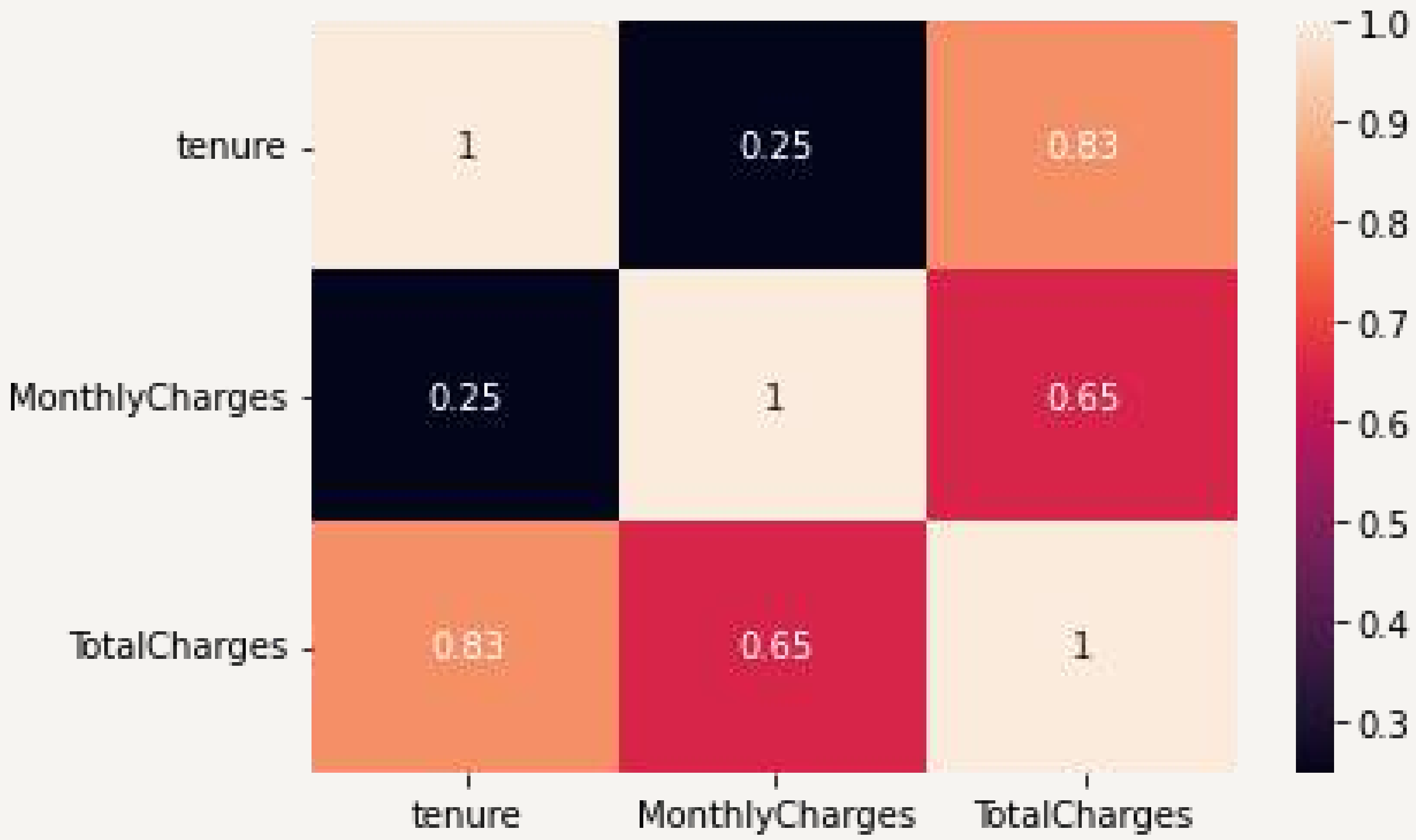
BY TOTAL CHARGES



	tenure	MonthlyCharges	TotalCharges
count	1869.000000	1869.000000	1869.000000
mean	17.979133	74.441332	1531.796094
std	19.531123	24.666053	1890.822994
min	1.000000	18.850000	18.850000
25%	2.000000	56.150000	134.500000
50%	10.000000	79.650000	703.550000
75%	29.000000	94.200000	2331.300000
max	72.000000	118.350000	8684.800000

- THE AVERAGE TOTAL COSTS INCURRED BY NO CHURN CUSTOMERS IS MORE THAN THE CHURN CUSTOMERS.
- THE AVERAGE TOTAL COSTS INCURRED BY NO CHURN CUSTOMERS IS 1531.8 WHERE IT IS SMALLER THAN A NO CHURN CUSTOMERS.

HEATMAP



- There is a multicollinearity between tenure with Total Charges and Monthly Charges with Total Charges.
- Then it must be removed one of the variables from the three variables above before modeling.

FEATURE ENGINEERING

FEATURE ENGINEERING

WHAT IS FEATURE ENGINEERING?

- Feature engineering is the process of **selecting, manipulating, and transforming raw data into features that can be used** in supervised learning. The act of converting raw observations into desired features using statistical or machine learning approaches.
- Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set.
- With the aim **to simplifying and accelerating** data transformation while improving model accuracy.

ONE HOT ENCODING

One-hot encoding is a process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions.

Example:

```
PaymentMethod = pd.get_dummies(df['PaymentMethod'], prefix='PaymentMethod', drop_first=False)
```

After One Hot Encoding

Before One Hot Encoding

PaymentMethod
Electronic check
Mailed check
Mailed check
Bank transfer (automatic)
Electronic check

PaymentMethod_Bank transfer (automatic)	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	0	1	0
0	0	0	1
0	0	0	1
1	0	0	0
0	0	1	0

MAP() FUNCTION

To convert numerical data variables can use the map() function

Column CONTRACT & CHURN is numerical variables so it uses the 'map' method in this feature engineering.

Example:

```
df['Contract'] = df['Contract'].map({"Month-to-month":0, "Two year":1, "One year":2})
```

BEFORE

```
df['Contract'].value_counts()
```

```
Month-to-month    3875
Two year          1695
One year          1473
Name: Contract, dtype: int64
```

AFTER

```
df['Contract'].value_counts()
```

```
0    3875
1    1695
2    1473
Name: Contract, dtype: int64
```

STANDARDSCALER FOR STANDARDIZATION

Standardization is used to center the feature columns at mean 0 with a standard deviation of 1 so that the feature columns have the same parameters as a standard normal distribution.

BEFORE

```
scaler = StandardScaler()
df['tenure'] = scaler.fit_transform(df[['tenure']])
df['MonthlyCharges'] = scaler.fit_transform(df[['MonthlyCharges']])
df['TotalCharges'] = scaler.fit_transform(df[['TotalCharges']])
df['Contract'] = scaler.fit_transform(df[['Contract']])
df.head()
```

AFTER

	customerID	tenure	Contract	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	-1.277445	-0.821752	-1.160323	-0.994242	0
1	5575-GNVDE	0.066327	1.672366	-0.259629	-0.173244	0
2	3668-QPYBK	-1.236724	-0.821752	-0.362660	-0.959674	1
3	7795-CFOCW	0.514251	1.672366	-0.746535	-0.194766	0
4	9237-HQITU	-1.236724	-0.821752	0.197365	-0.940470	1

5 rows × 58 columns

PREPOCESSING DATA

PREPROCESSING DATA

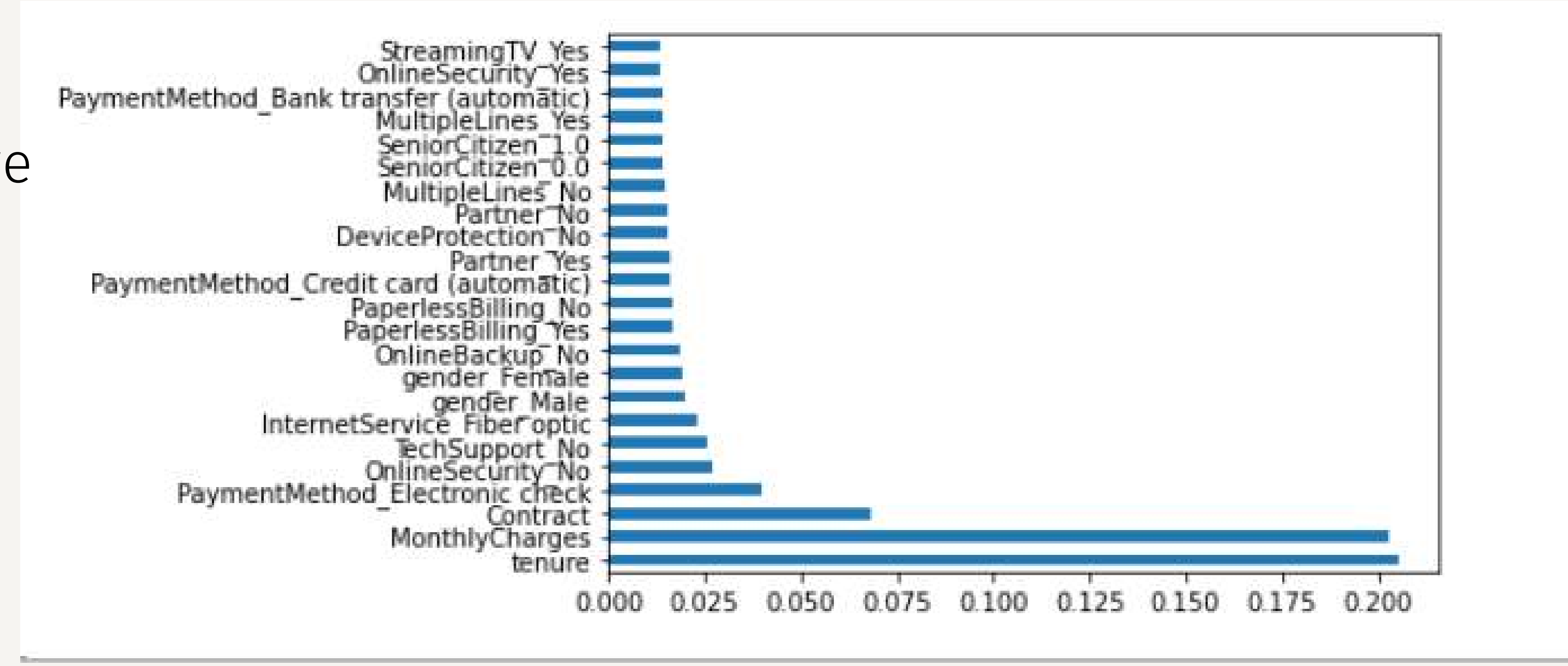
DEFINITION

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

PREPROCESSING DATA

- Remove TotalCharges Column because its multicolinear column.
- Checking Importance data using Feature Importance.
- Split Train and Test data to 80%(Train) and 20%(Test)

Feature Importance



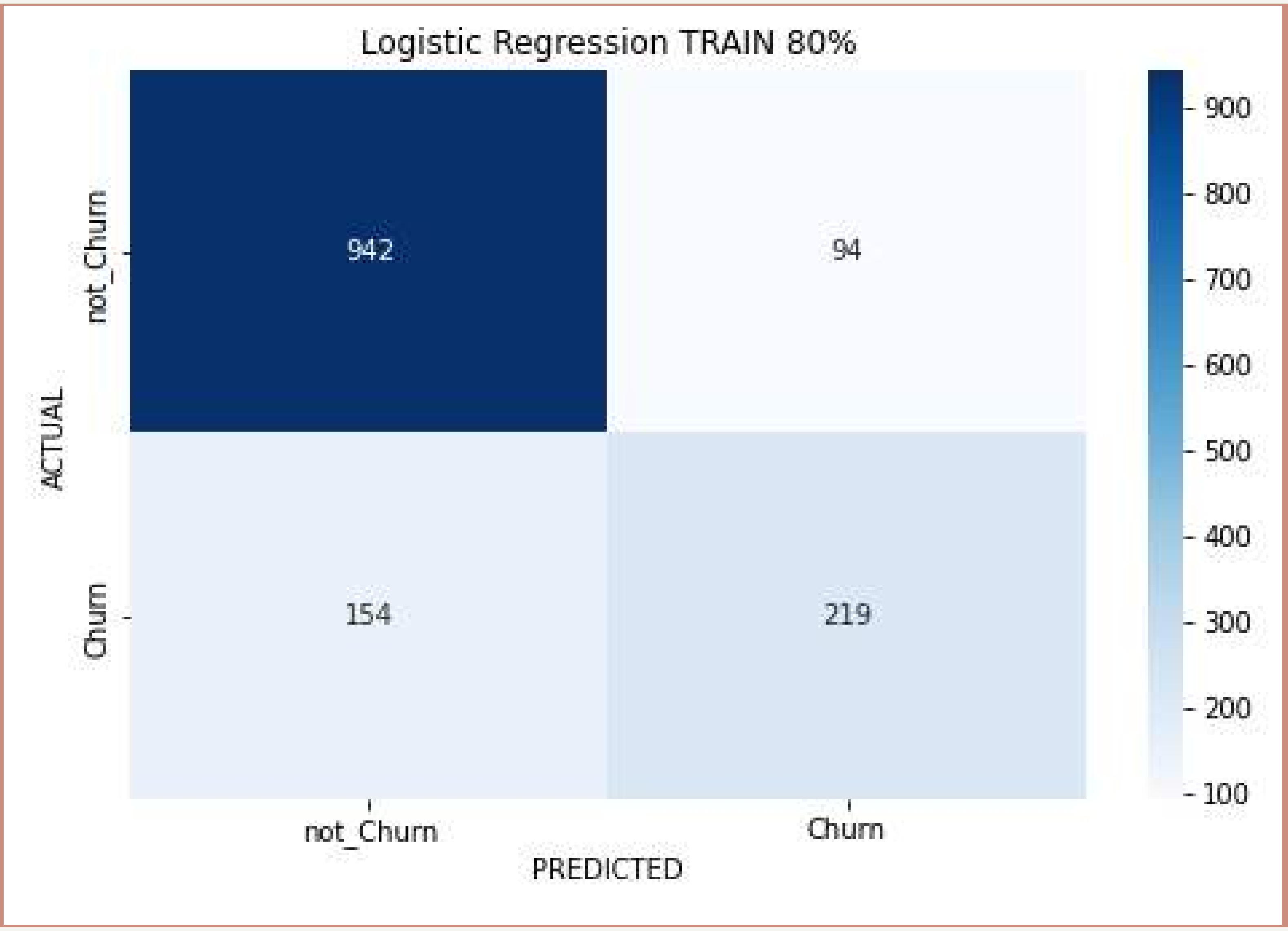
MODELLING

MODELING DATA USING LOGISTIC REGRESSION

DEFINITION

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

LOGISTIC REGRESSION



	precision	recall	f1-score	support
not_Churn	0.86	0.91	0.88	1036
Churn	0.70	0.59	0.64	373
accuracy			0.82	1409
macro avg	0.78	0.75	0.76	1409
weighted avg	0.82	0.82	0.82	1409

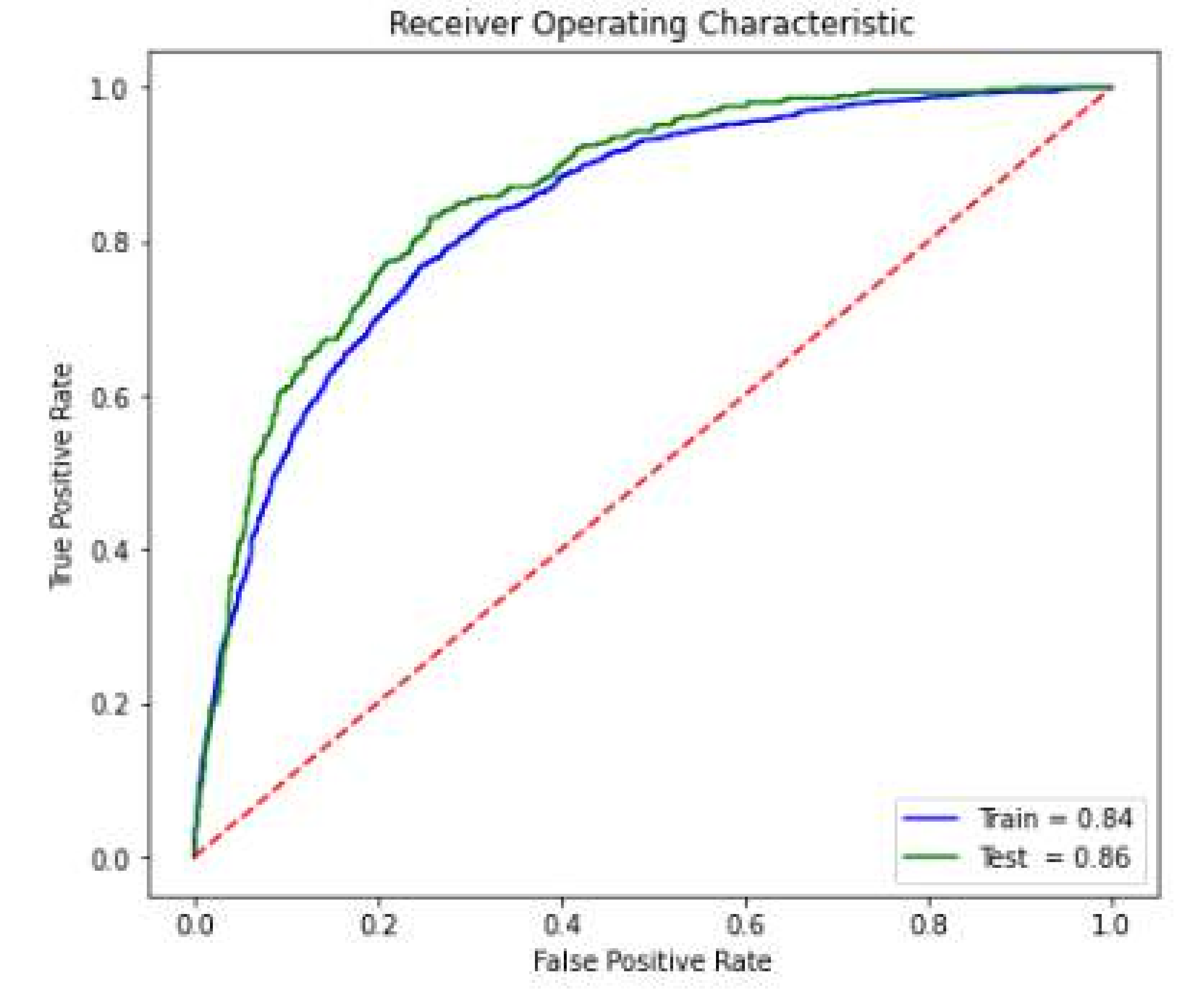
FROM THE CONFUSION MATRIX ABOVE, THE ACCURACY VALUE OF THE LOGISTIC REGRESSION MODEL IS 82%

AUC-ROC

DEFINITION

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

AUC-ROC




```
AUC train & test              : 83.53% & 85.99%

Confusion Matrix Evaluation
Accuracy train & test    : 80.00% & 82.40%
Recall train & test      : 52.01% & 58.71%
Specificity train & test: 90.12% & 90.93%
Precision train & test   : 65.54% & 69.97%
F1 Score train & test    : 57.99% & 63.85%
Log Loss train & test    : 6.909 & 6.0793
```

FROM THE AUC-ROC GRAPH AND CONFUSION MATRIX EVALUATION ABOVE,
WE CAN KNOW THAT THE AUC VALUE IS 86% AND ALSO THE MODEL IS GOOD FIT

BUILDING A MODEL WITH CROSS VALIDATION

WHAT IS CROSS VALIDATION ?

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

```
regressor = LogisticRegression()
```

```
scores = cross_val_score(regressor, X_train, y_train, scoring = 'accuracy', cv=3)
scores
```

```
array([0.79392971, 0.80031949, 0.80244941])
```

HYPERPARAMETER TUNING IN LOGISTIC REGRESSION

WHAT IS THE BEST?

```
regressor.get_params()
parameters = {"penalty": ['l1', 'l2', 'elasticnet', 'none'],
              "solver": ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
              "n_jobs": [None, -1],
              "max_iter": [10, 100, 1000]
             }
```

```
grid = GridSearchCV(estimator = regressor, param_grid = parameters, cv=3)
```

```
best_model = grid.fit(X_train, y_train)
```

```
best_model.best_params_
```

```
{'max_iter': 10, 'n_jobs': None, 'penalty': 'l2', 'solver': 'saga'}
```

```
regres_new = LogisticRegression(max_iter= 10, n_jobs= None, penalty = 'l2', solver= 'saga')
```

```
model_new = regres_new.fit(X_train, y_train)
```

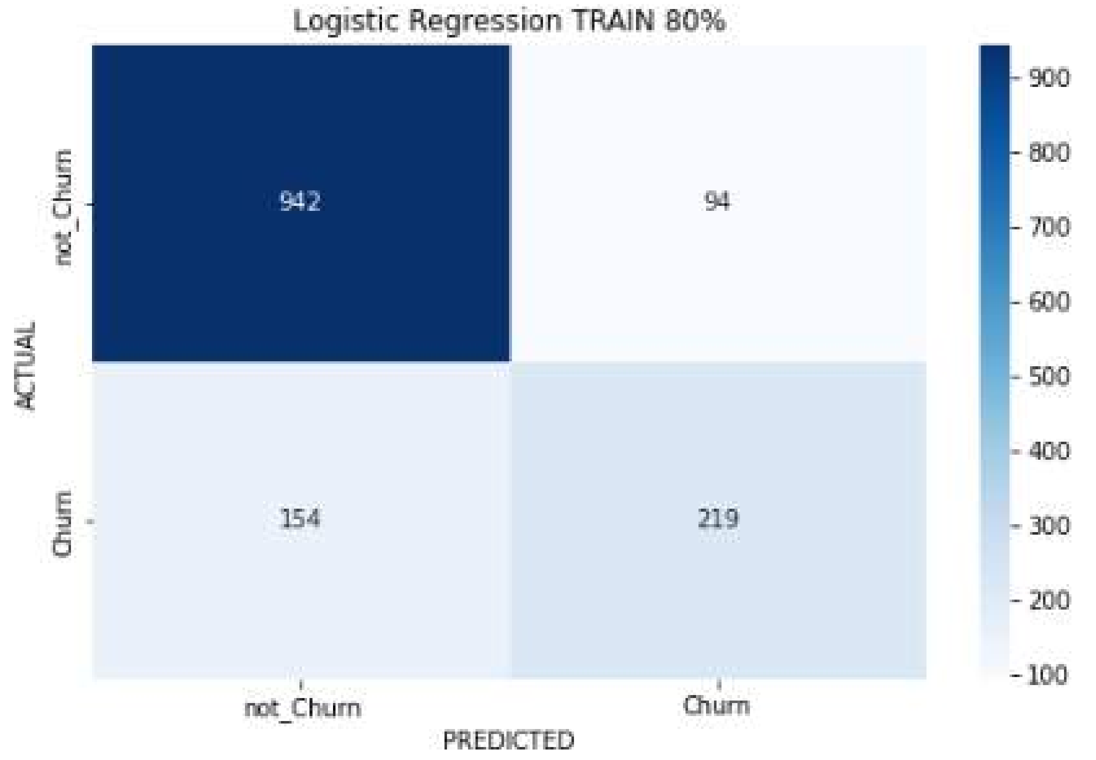
```
y_pred_new = regres_new.predict(X_test)
```

From the Grid Search, it is found that the best Hyperparameters value for max_iter is 10, n_jobs is None, penalty is 'l2', and solver is 'saga'.

EVALUATE MODEL


```
cm_model = confusion_matrix(y_test, y_pred_new)

labels = ['not_Churn', 'Churn']

print(classification_report(y_test, y_pred, target_names = labels))
f, ax = plt.subplots(figsize=(8,5))
sns.heatmap(cm_model, annot=True, fmt=".0f", ax=ax, cmap = 'Blues')

ax.xaxis.set_ticklabels(labels)
ax.yaxis.set_ticklabels(labels)

plt.title('Logistic Regression TRAIN 80%')
plt.xlabel('PREDICTED')
plt.ylabel('ACTUAL')
plt.show()
```

```
              precision    recall  f1-score   support

   not_Churn       0.86      0.91      0.88      1036
       Churn       0.70      0.59      0.64       373

    accuracy                           0.82      1409
   macro avg       0.78      0.75      0.76      1409
weighted avg       0.82      0.82      0.82      1409
```



OVERSAMPLING WITH SMOTE

WHAT IS SMOTE ?

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

```
# Oversampling with SMOTE

X_train_sm, y_train_sm = SMOTE(random_state = False).fit_resample(X_train, y_train)

# Model oversampled

model_sm = LogisticRegression()
model_sm.fit(X_train_sm, y_train_sm)
```

```
▸ LogisticRegression
```

```
# Predic using Logistic Regression oversampled

y_pred_sm = model_sm.predict(X_test)
```

EVALUATE MODEL AFTER OVERSAMPLING WITH SMOTE

SMOTE

```
cm_model = confusion_matrix(y_test, y_pred_sm)

labels = ['not_Churn', 'Churn']

print(classification_report(y_test, y_pred, target_names = labels))
f, ax = plt.subplots(figsize=(8,5))
sns.heatmap(cm_model, annot=True, fmt=".0f", ax=ax, cmap = 'Blues')

ax.xaxis.set_ticklabels(labels)
ax.yaxis.set_ticklabels(labels)

plt.title('Logistic Regression TRAIN 80%')
plt.xlabel('PREDICTED')
plt.ylabel('ACTUAL')
plt.show()
```

	precision	recall	f1-score	support
not_Churn	0.86	0.91	0.88	1036
Churn	0.70	0.59	0.64	373
accuracy			0.82	1409
macro avg	0.78	0.75	0.76	1409
weighted avg	0.82	0.82	0.82	1409

EVALUATE MODEL AFTER OVERSAMPLING WITH SMOTE

It can be seen that after an evaluation using oversampling with SMOTE, the accuracy value increases, indicating this is good and the model is fit

```
# Accuracy Before Evaluate
print(accuracy_score(y_test,y_pred))
```

```
0.8232789212207239
```

```
# Accuracy After Evaluate
print(accuracy_score(y_test,y_pred_new))
```

```
0.8239886444286728
```

RESULT

RESULT

- No Churn customers there are as much as 73.46%, while churn customers there are as much as 26.54%.
- Most churn customers are female who are 11.44%. Male churn customers are as many as 11.39% and 3.71% are churn customers that their gender is not known. While the remaining 73.46% are no churn customers.
- From churn customers, the majority are not seniors, which is as much as 17%. 5.84% other are senior citizens and 3.71% are churn customers who are unknown. While the remaining 73.45% are no churn customers.
- Most of the churn customers are customers who use Fiber Optics internet services are 15.75%. 5.69% DSL users, 3.71% unknown, and 1.39% did not have internet services. While the remaining 73.46% are no churn customers.
- Most churn customers are customers who use electronic transactions (paperless billing) that are 19.88% while 6.66% were not paperless billing users. The remaining 73.46% are no churn customers.

RESULT

- 18.83% of customers are churn customers and have no dependents. While 4% of them have dependents and 3.71% are not known. While the remaining 73.47% are no churn customers.
- Most churn customers are customers who don't have a partner, that is as much as 14.65% of all numbers customers. 8.18% have a partner and 3.71% are not known. While the remaining 73.46% are no churn customers.
- 24.12% of customers are churn customers and have phone services. while 2.41% were not have phone service. The remaining 73.46% are no churn customers.
- Most churn customers are customers who have multiple lines, that are 10.46% of customers. 10.28% did not have multiple lines, 3.71% of it is unknown, and 2.09% did not have phone services. While the remaining 73.47% are no churn customers.
- Most churn customers are customers who do not have online security, which is 17.88% customers. 3.56% have online security, 3.71% of them are unknown, and 1.39% do not have internet services. The remaining 73.46% are no churn customers.

RESULT

- Most churn customers are customers who do not have online backup, which is 15.15% customers. 6.29% have online backup, 3.71% of them are unknown, and 1.39% do not have internet services. The remaining 73.46% are no churn customers.
- Most churn customers are customers who have no device protection, which is 14.81% of customers. 6.63% of it have device protection, 3.71% of them are unknown, and 1.39% do not have internet services. The remaining 73.46% are no churn customers.
- Most churn customers are customers who don't have technical support services, that is 17.54% of customers. 3.9% of them have a technical support, 3.71% of them are unknown, and 1.39% do not have internet services. The remaining 73.46% are no churn customers.
- Most churn customers are customers who do not use Streaming TV services, which is 11.39% of customers. 10.05% of it uses streaming TV services, 3.71% of which are unknown, and 1.39% have no internet services. The remaining 73.46% are no churn customers.

RESULT

- Most churn customers are customers who do not use Streaming Movies service, which is 11.54% of customers. 9.9% of it using Streaming Movies service, 3.71% of it is unknown, and 1.39% have no internet services. The remaining 73.46% are no churn customers.
- Most churn customers are customers who have monthly contracts, that are 23.50% of customers. 2.36% of it had a one-year contract, and 0.68% had a two-year contract. The remaining 73.46% are no churn customers.
- 15.21% of customers are churn customers and use Elektronic Check payments method. 4.37% of them used Mailed check, 3.66% of them use bank transfers, and 3.29% are churn customers and use credit cards. The remaining 73.46% are no churn customers.
- Most churn customers are customers whose tenure is only 1 month with an average is 18 months.
- Most churn customers on average are customers who spend 74.4 per month.
- The average total costs incurred by no churn customers is 1531.8 where it is smaller than a no churn customers.

RESULT

- After creating modeling with logistics regression, then the evaluation is made using the AUC/ROC. It can be concluded that the resulting model does not overfit because the AUC train is obtained by 83.53% and the test earned by 85.99%, where the difference is no more than 0.05.
- After creating modelling with oversampling with SMOTE, it can be seen that the accuracy value increases, indicating this is good and the model is fit

RECOMENDATION

RECOMMENDATION

- Recommendation for “Customer Churn" segment: Focus on increasing customer purchases, such as create marketing campaigns to upsell those currently subscribed to streaming movies and TV services on our other internet services.
- Recommendation for “Device Protection" segment: Improve the Device Protection service in order to prevent a large number of customer churn who use that service.
- Recommendation for “Streaming TV" segment: Improve the Streaming TV service in order to prevent a large number of customer churn who use that service.
- Recommendation for “Streaming Movies" segment: Improve the Streaming Movies service in order to prevent a large number of customer churn who use that service.
- Recommendation for “Internet Service" segment: Improve the internet service with fiber optic in order to prevent a large number of customer churn who use the fiber optic.

RECOMMENDATION

- Recommendation for “Payment Method" segment: Maintain service performance with Mailed Check payment method to prevent customers using that payment method from churn and improve service with Electronic Check payment method so that customers who churn do not get more and more.
- Recommendation for "Contract" segment: Reduce the use of month-to-month contracts, because many customers unsubscribe with month-to-month contracts
- Recommendation for "Technical Support" segment: Have to improve the Technical support service in order to prevent a large number of customer churn who use that service.
- Recommendation for "Partner" segment: Have to increase the number of partners in the company in order to reduce unsubscribed customers
- Recommendation for "Phone Service" segment: must improve phone service in order to reduce unsubscribed customers
- Recommendation for "Multiple line Service" segment:must improve Multiple line service in order to reduce unsubscribed customers

THANK YOU