**ANAND INSTITUTE OF HIGHER TECHNOLOGY OLD MAHABALIPURAM ROAD, KALASALINGAM NAGAR, KAZHIPATTUR – 603103**

****

**CUSTOMER CHURN PREDICTION**

**WITH**

**DATA ANALYTICS USING COGNOS**

**PHASE – 3**

**NAME : NOKITHA V M**

**REG No. : 310121104069**

**BRANCH : COMPUTER SCIENCE & ENGINEERING**

**YEAR/SEM : III / V**

**IBM DATA ANALYTICS WITH COGNOS**

**TEAM NAME** **:** Proj\_229798\_Team\_1

**PROJECT :** 3101-Customer Churn Prediction

**TEAM MEMBERS :**

1. Nokitha V M
2. Mounika V
3. Yasmeen U
4. Roshini

**LEADER :** Nokitha V M

**CUSTOMER CHURN PREDICTION**

**INTRODUCTION**

* Churn prediction is predicting which customers are at high risk of leaving your company or canceling a subscription to a service, based on their behavior with your product.
* To predict churn effectively, you’ll want to synthesize and utilize key indicators defined by your team to signal when a customer has a probability of churning so that your company can take action.
* At a high level, predicting customer churn requires a detailed grasp of your clientele. Both qualitative and quantitative customer data are usually needed to start building an effective churn prediction model. To ensure that predictions aren’t being made by arbitrary human guesses, these models are often built by a data scientist using machine learning.
* In a churn prediction model case, the target variable would be the indicator signifying whether a customer is likely to churn–(yes/no) or (1/0). To obtain this variable, you would need to use historical data of existing customers and previous customers that denotes whether this customer left or stayed; this could be a subscription cancellation, a closed contract, etc.



**PHASE - 3 : DEVELOPMENT PART 1**

* Start building the customer churn prediction using IBM Cognos for visualization. Define the analysis objectives and collect customer data from the source shared.
* Process and clean the collected data to ensure its quality and accuracy.
* This process involves collecting, cleaning, transforming, reduction of null values, visualization, scalability, efficiency and structuring raw data to make it suitable for analysis.
* Both qualitative and quantitative customer data are usually needed to start building an effective churn prediction model. To ensure that predictions aren't being made by arbitrary human guesses, these models are often built by a data scientist using machine learning.
* The goal of Customer Churn Prediction in this Phase3 is to prepare and begin building your project by loading and preprocessing the dataset.



**DATA COLLECTION:**

CUSTOMER CHURN PREDICTION is done by using the Dataset of “Telco Customer Churn” provided by the dataset site [www.Kaggle.com](http://www.Kaggle.com)



**Dataset Link:**

**<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>**

**DATASET AND ITS DETAILS :**

**TITLE:** CUSTOMER CHURN PREDICTION

**LIBRARIES:** sklearn, Matplotlib, pandas, seaborn, and NumPy

**Context:**

The dataset “CUSTOMER CHURN PREDICTION” on Kaggle is a collection of data related to the a Machine Learning Model That Can Predict Customers Who Will Leave The Company "Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs."

**Content:**

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

**The data set includes information about:**

**Gender** -- Whether the customer is a male or a female

**SeniorCitizen** -- Whether a customer is a senior citizen or not

**Partner** -- Whether the customer has a partner or not (Yes, No)

**Dependents** -- Whether the customer has dependents or not (Yes, No)

**Tenure** -- Number of months the customer has stayed with the company

**Phone Service** -- Whether the customer has a phone service or not (Yes, No)

**MultipleLines** -- Whether the customer has multiple lines or not

**InternetService** -- Customer's internet service provider (DSL, Fiber Optic, No)

**OnlineSecurity** -- Whether the customer has online security or not (Yes, No, No Internet)

**OnlineBackup** -- Whether the customer has online backup or not (Yes, No, No Internet)

**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)

**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)

**Payment Method** -- 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

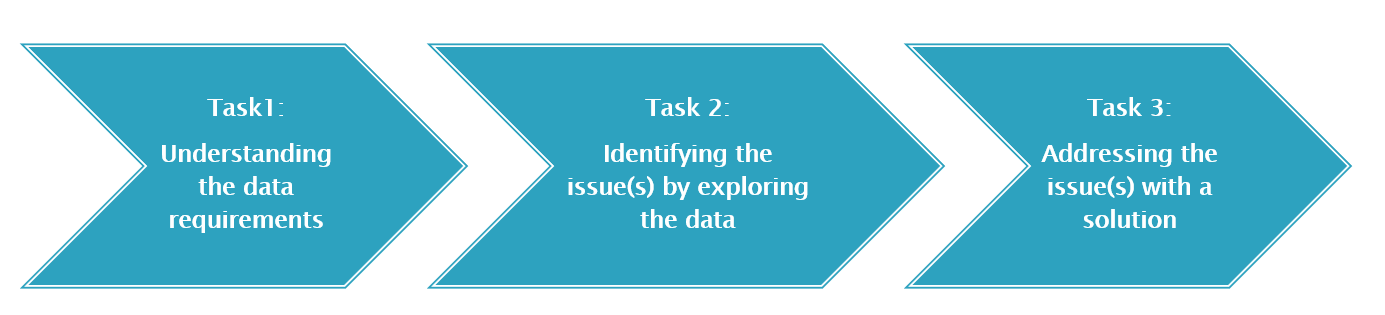
**Churn** -- Whether the customer churned or not (Yes or No)

**SIGNIFICANCE OF LOADING AND PREPROCESSING THE DATASET:**

Data Loading is defined as copying data from one electronic file or database into another. Data loading implies converting from one format into another; for example, from one type of production database into a decision support database from a different vendor.

Data preprocessing is essential before its actual use. Data preprocessing is the concept of changing the raw data into a clean data set. The dataset is preprocessed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm.

*Tasks under data preprocessing:*



**CHALLENGES INVOLVED IN LOADING AND PREPROCESSING:**

**1. Irrelevant data:**

The most basic yet unavoidable issue that requires data cleaning is the presence of attributes in the data set that are irrelevant to the problem we are trying to solve.

**2. Duplicate data:**

Integrating data from different sources may result in redundant columns and rows in the data set.

**3. Incorrect data type:**

A data-set would generally store different types of data such as integer, float, and string. However, the type in which a data column is stored may be wrong.

**4. Missing values:**

It is seldom a case where values are not missing in a data set.

**5. Outliers:**

Outliers are extreme data points compared to the rest of the data. They can be detected numerically by calculating the data distribution’s inter-quartile range and standard deviation. They can also be visualized using box plots, histograms, and scatter plots.

**6. Unacceptable format:**

The data may be in a format that is not acceptable to the machine learning algorithm to be applied later in the data mining stage.

**7. Too many categories:**

In a categorical variable, there can be many categories. For example, the column storing the name of ‘Locality’, to which a person belongs, can lead to too many categories. This issue can be observed through descriptive statistics like a bar chart showing the frequency of data category-wise.

**8. Class Imbalance**: In customer churn prediction, it's common to have an imbalanced dataset, where the number of customers who stayed far exceeds those who left. This imbalance can lead to model bias,where the model may have difficulty in correctly identifying the minority class (churned customers).

Solution -Utilized models like Random Forest (with tuned hyperparameters) that are robust and less prone to overfitting, making the model more reliable when dealing with imbalanced datasets.

**IMPORT AND LOAD THE DATASET :**

Use Pandas to read the dataset file you downloaded into a DataFrame:

**CODING AND ITS OUTPUT:**

**LOAD DATA:**

**import** numpy **as** np

**import** pandas **as** pd

*# Visualization*

**import** plotly.express **as** px

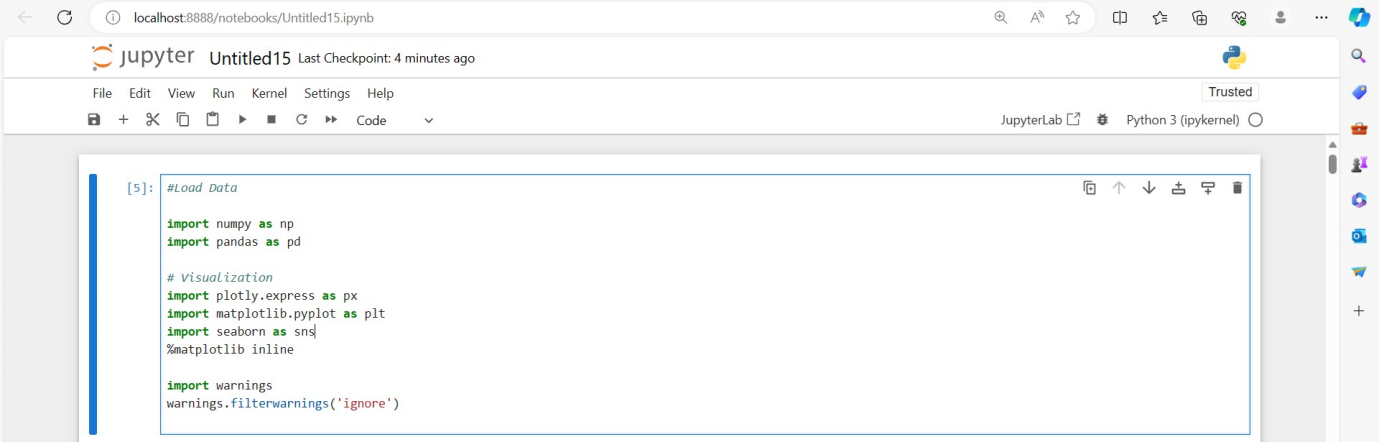
**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**%matplotlib** inline

**import** warnings

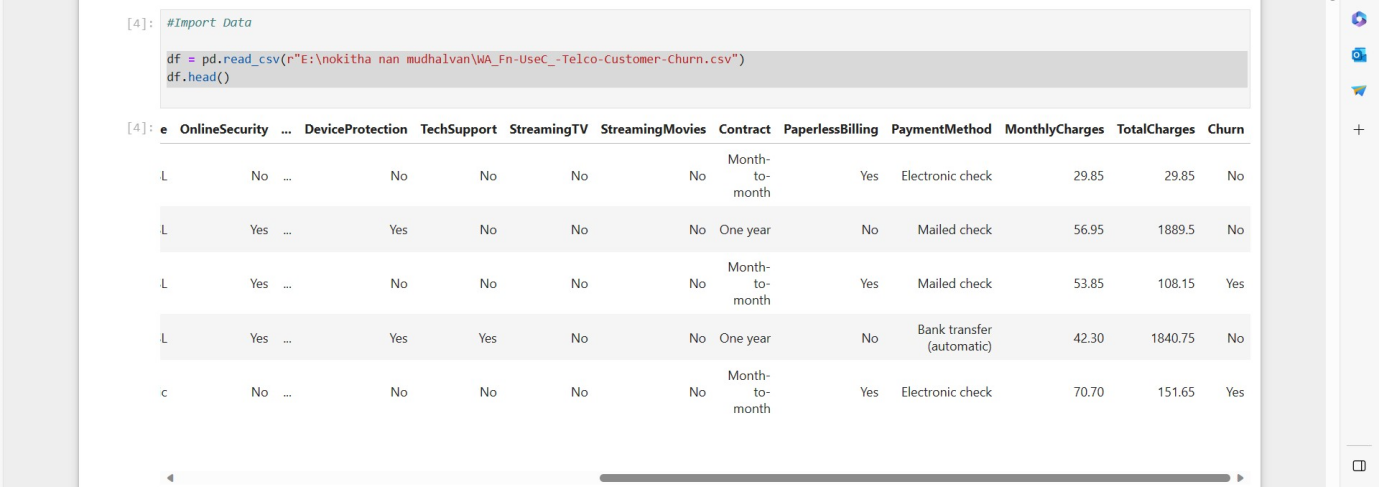
warnings**.**filterwarnings('ignore')



**IMPORT DATA:**

df = pd.read\_csv(r"E:\nokitha nan mudhalvan\WA\_Fn-UseC\_-Telco-Customer-Churn.csv")

df.head()



**DATA UNDERSTANDING:**

df.info()



df['TotalCharges'] **=** pd**.**to\_numeric(df['TotalCharges'], errors**=**'coerce')

**Check the Duplicate:**

print(df.duplicated().value\_counts())

**Check the missing values:**

df**.**isnull()**.**values**.**any()



**Overview:**

round(df**.**describe(include**=**'all'),2)

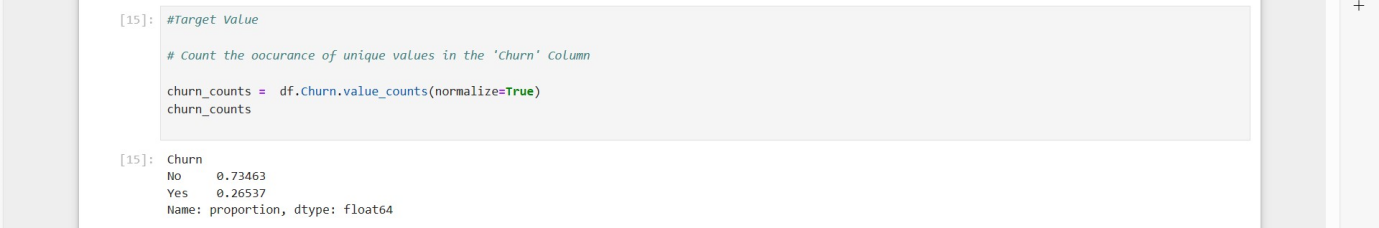


**Target value:**

*# Count the oocurance of unique values in the 'Churn' Column*

churn\_counts **=** df**.**Churn**.**value\_counts(normalize**=True**)

churn\_counts



*# Calculate the percentage of 'Yes' and 'No' label*

total\_count **=** churn\_counts**.**sum()

percentage\_yes **=** (churn\_counts['Yes']**/** total\_count) **\*** 100

percentage\_no **=** (churn\_counts['No']**/** total\_count) **\***100

*# Plot the target value*

ax **=** churn\_counts**.**plot(kind**=**'bar')

*# Annotate the bars with percentages*

**for** i, count **in** enumerate(churn\_counts):

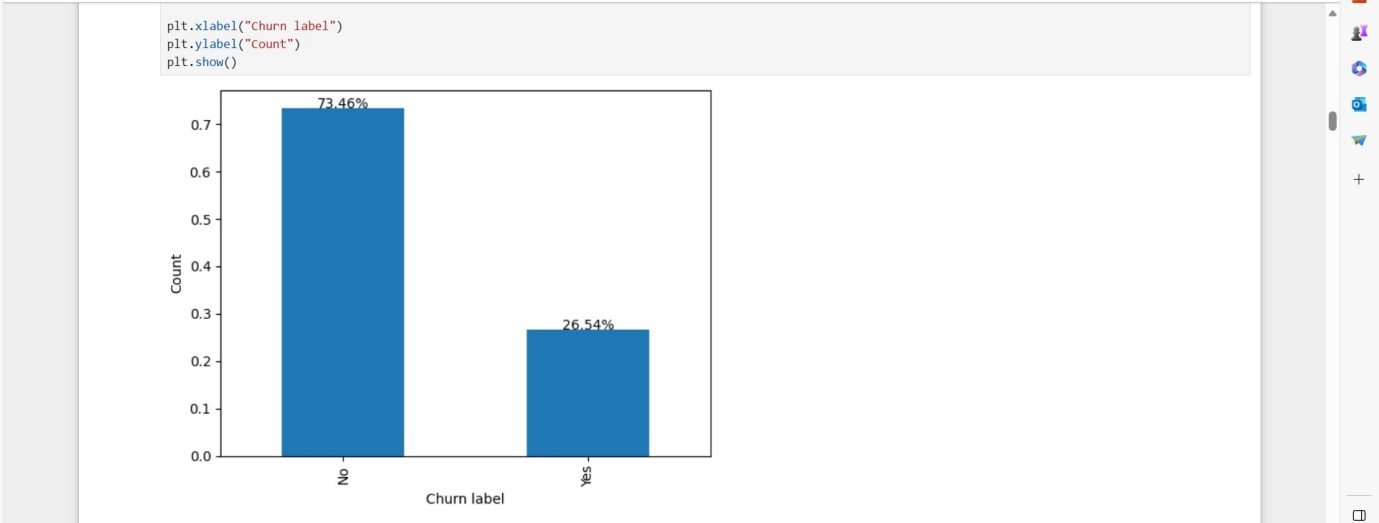
percentage **=** percentage\_yes **if** i **==** 1 **else** percentage\_no

ax**.**annotate(f'{percentage:.2f}%', xy**=**(i, count), ha**=**'center')

plt**.**xlabel("Churn label")

plt**.**ylabel("Count")

plt**.**show()



**EDA - Exploratory Data Analysis on each feature:**

**It is observed that the dataset exhibits a significant class imbalance, with a larger amount of data representing non-churners.**

**1 .** Is there a correlation between churn and factors such as monthly charges and total charges?

df[['MonthlyCharges','TotalCharges']]

df.groupby(['MonthlyCharges','TotalCharges'])['Churn'].size()

sns.kdeplot(data=df, x="MonthlyCharges",hue="Churn",multiple="stack")

# Customize the plot appearance

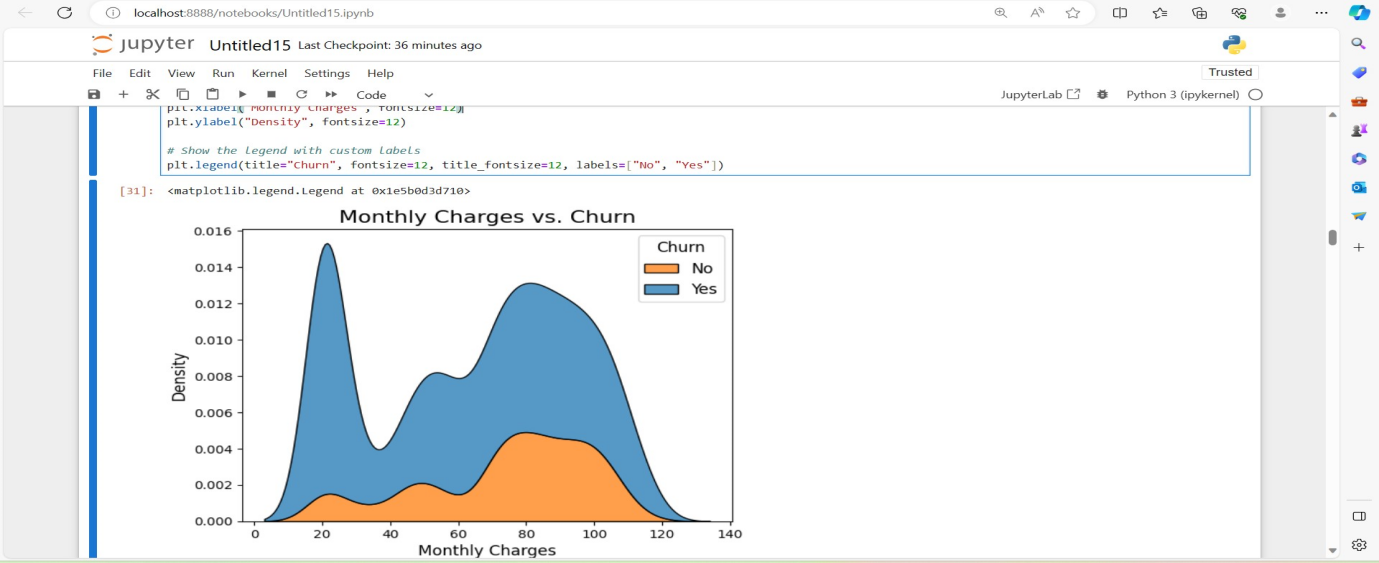
plt.title("Monthly Charges vs. Churn", fontsize=16)

plt.xlabel("Monthly Charges", fontsize=12)

plt.ylabel("Density", fontsize=12)

# Show the legend with custom labels

plt.legend(title="Churn", fontsize=12, title\_fontsize=12, labels=["No", "Yes"])



* Note: it's noticeable that as monthly charges increase within the range of 60 to 120, the density also rises. This trend indicates a higher rate of churn as monthly charges increase.

df['TotalCharges'] = pd.to\_numeric(df['TotalCharges'], errors='coerce')

df['TotalCharges']

print(df['TotalCharges'].dtype)



sns.kdeplot(data=df, x="TotalCharges",hue="Churn",multiple="stack")

# Customize the plot appearance

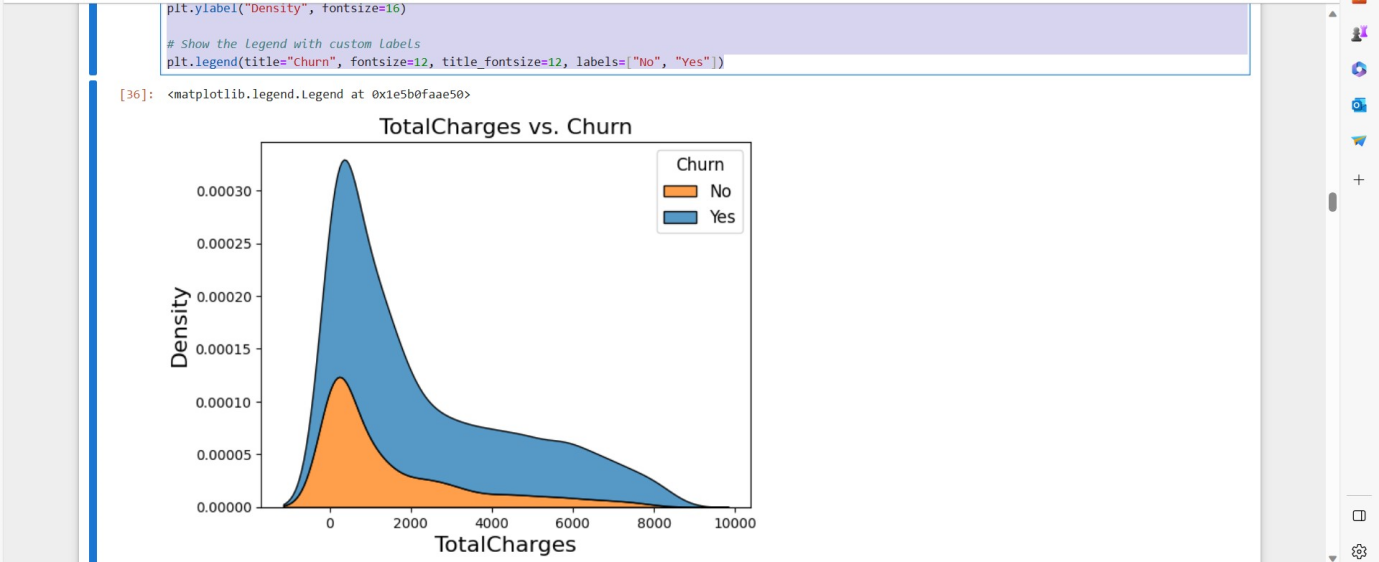
plt.title("TotalCharges vs. Churn", fontsize=16)

plt.xlabel("TotalCharges", fontsize=16)

plt.ylabel("Density", fontsize=16)

# Show the legend with custom labels

plt.legend(title="Churn", fontsize=12, title\_fontsize=12, labels=["No", "Yes"])



Note: High churn rates are associated with lower total charges, with the highest churning occurring in the 0-2000 total charges range.

**2.** How does the length of a customer's tenure with the company influence their likelihood of churning?

grouped\_data = df.groupby(['tenure', 'Churn']).size().reset\_index(name='count')

grouped\_data[20:30]



grouped\_data = df.groupby(['tenure', 'Churn']).size().reset\_index(name='count')

grouped\_data[30:40]

# Separate churn and non-churn counts

churn\_data = grouped\_data[grouped\_data['Churn'] == 'Yes']

non\_churn\_data = grouped\_data[grouped\_data['Churn'] == 'No']

# Create a line chart for churn and non-churn counts

plt.figure(figsize=(5, 4))

plt.plot(churn\_data['tenure'], churn\_data['count'], label='Churn', marker='\*')

plt.plot(non\_churn\_data['tenure'], non\_churn\_data['count'], label='Non-Churn', marker='.')

plt.xlabel('Tenure')

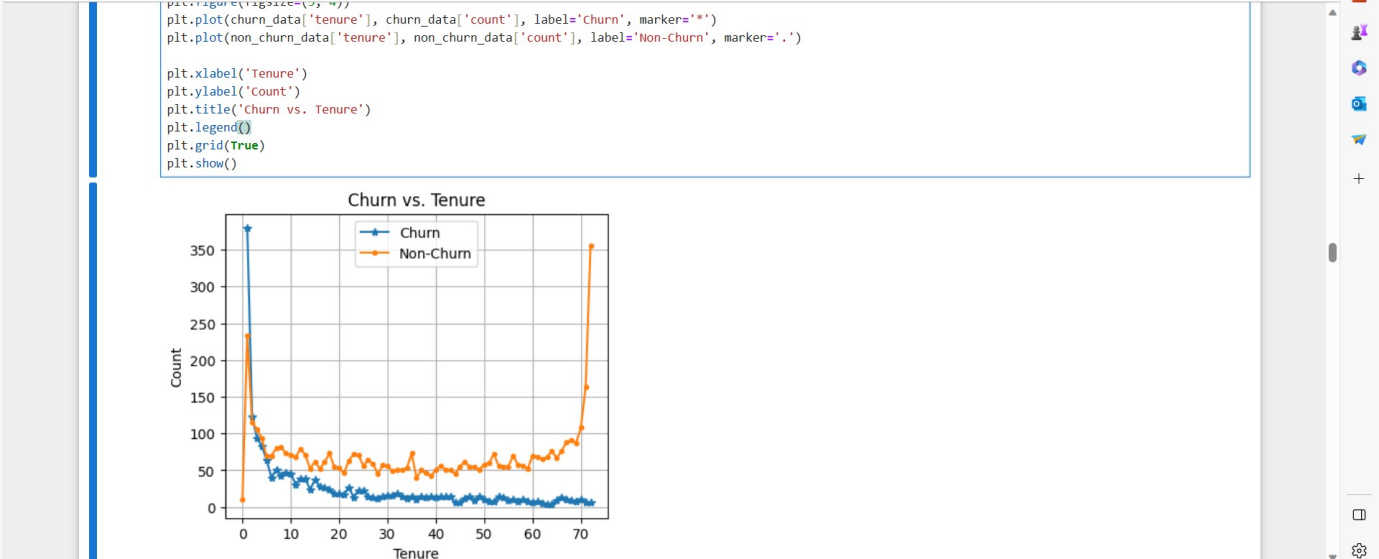
plt.ylabel('Count')

plt.title('Churn vs. Tenure')

plt.legend()

plt.grid(True)

plt.show()



Note:

Customers with the low tenure **"0-10”** has the highest rate of churning,these range can be crucial for business decisions.

In general the **Non-Churn** line has small fluctuation and remains **relatively stable**, with longer tenure tend to stay with the company.

The Churn line decreases or remains relatively stable as tenure increases, it suggests that **customer loyalty** increases with longer tenure

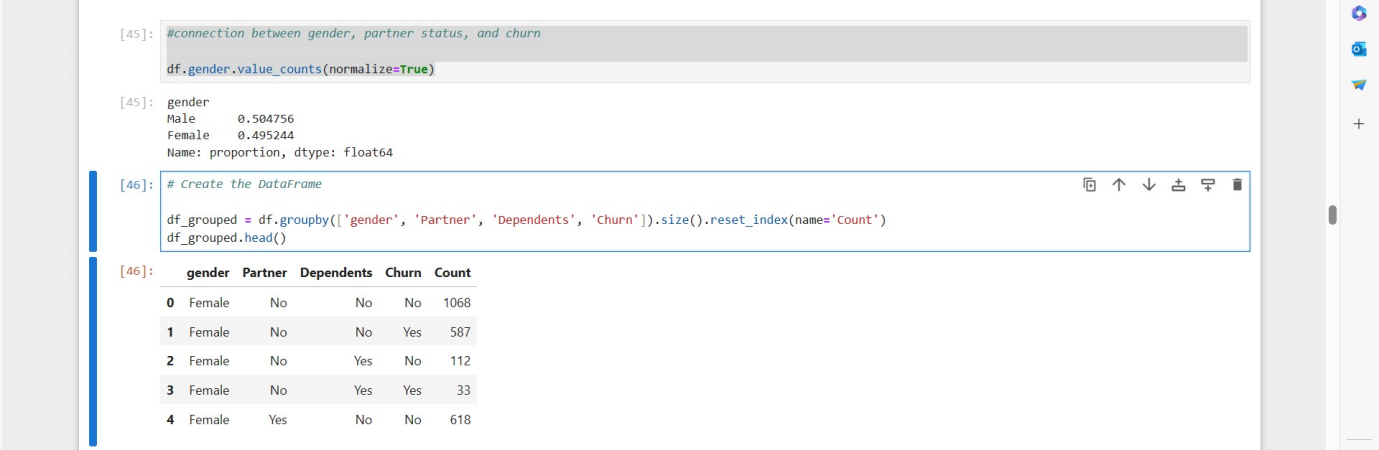
**3.** connection between gender, partner status, and churn

df.gender.value\_counts(normalize=True)

# Create the DataFrame

df\_grouped = df.groupby(['gender', 'Partner', 'Dependents', 'Churn']).size().reset\_index(name='Count')

df\_grouped.head()



# Create the DataFrame

df\_grouped = df.groupby(['gender', 'Partner', 'Dependents', 'Churn']).size().reset\_index(name='Count')

# Create a subplot with two subplots (one for Dependents and one for Partner)

fig = px.bar(df\_grouped, x='Count', y='Churn', color='Dependents', facet\_col='Partner',

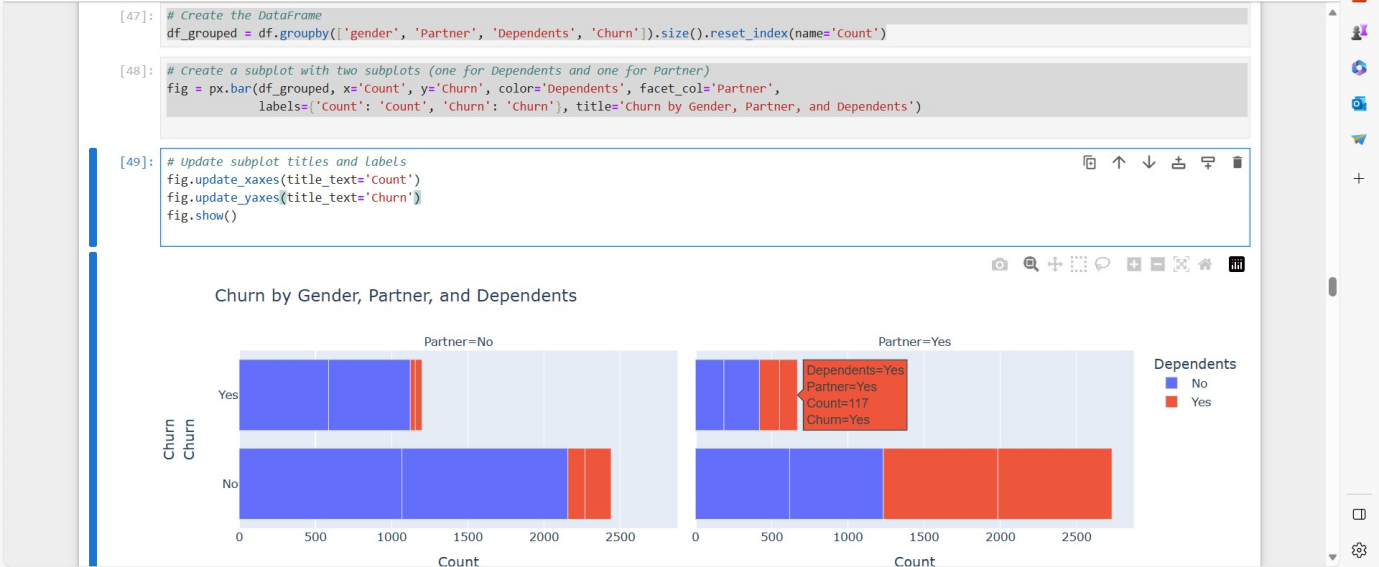
labels={'Count': 'Count', 'Churn': 'Churn'}, title='Churn by Gender, Partner, and Dependents')

# Update subplot titles and labels

fig.update\_xaxes(title\_text='Count')

fig.update\_yaxes(title\_text='Churn')

fig.show()



Note:

Regardless of gender, individuals who lack a partner or dependents are at a higher risk of churning.

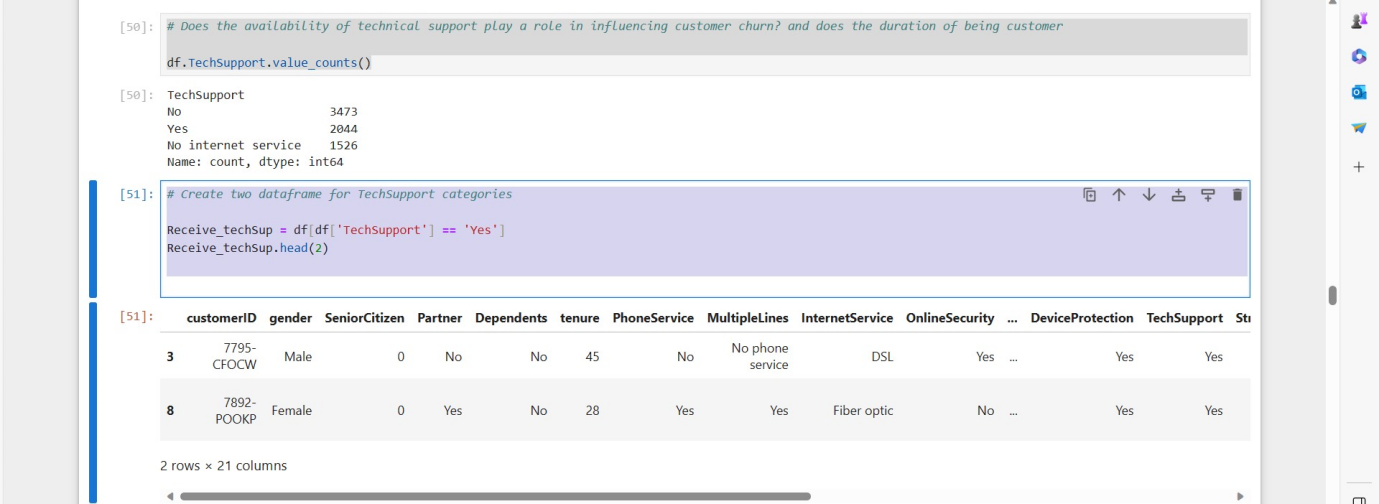
**4**. Does the availability of technical support play a role in influencing customer churn? and does the duration of being customer

df.TechSupport.value\_counts()

# Create two dataframe for TechSupport categories

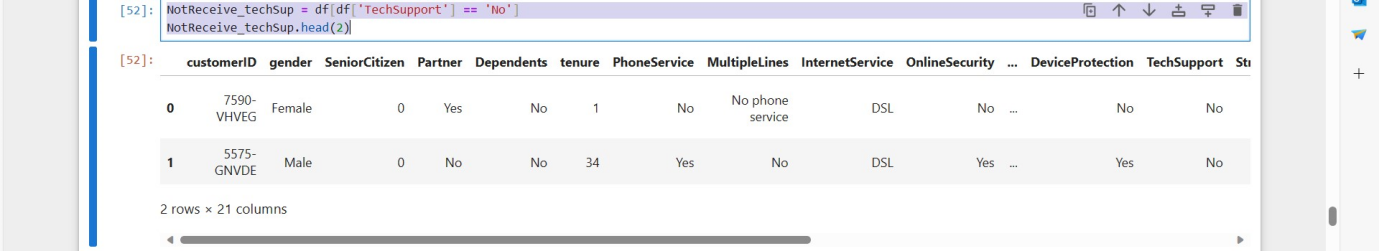
Receive\_techSup = df[df['TechSupport'] == 'Yes']

Receive\_techSup.head(2)



NotReceive\_techSup = df[df['TechSupport'] == 'No']

NotReceive\_techSup.head(2)



# Create subplots to compare the distribution

fig, axes = plt.subplots(1, 2, figsize=(10, 5))

# Plot the distribution of 'TechSupport' for 'Yes' category

sns.countplot(data=Receive\_techSup, x='TechSupport', hue='Churn', ax=axes[0])

axes[0].set\_title('TechSupport = Yes')

axes[0].set\_xlabel('TechSupport')

axes[0].set\_ylabel('Count')

# Plot the distribution of 'TechSupport' for 'No' category

sns.countplot(data=NotReceive\_techSup, x='TechSupport', hue='Churn', ax=axes[1])

axes[1].set\_title('TechSupport = No')

axes[1].set\_xlabel('TechSupport')

axes[1].set\_ylabel('Count')

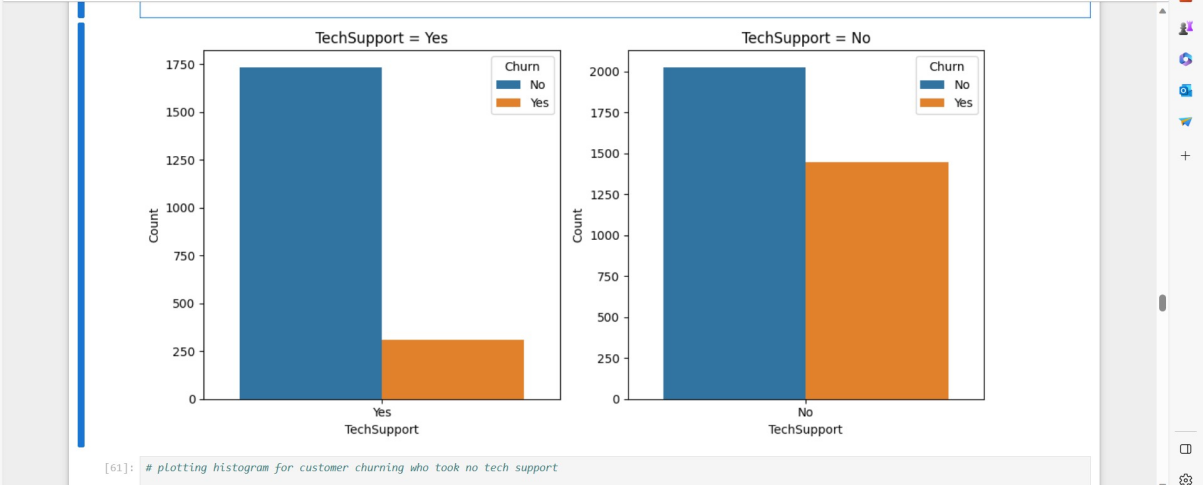
# Adjust layout

plt.tight\_layout()

plt.show()

Note:

Customers who do not receive tech support are more likely to churn.



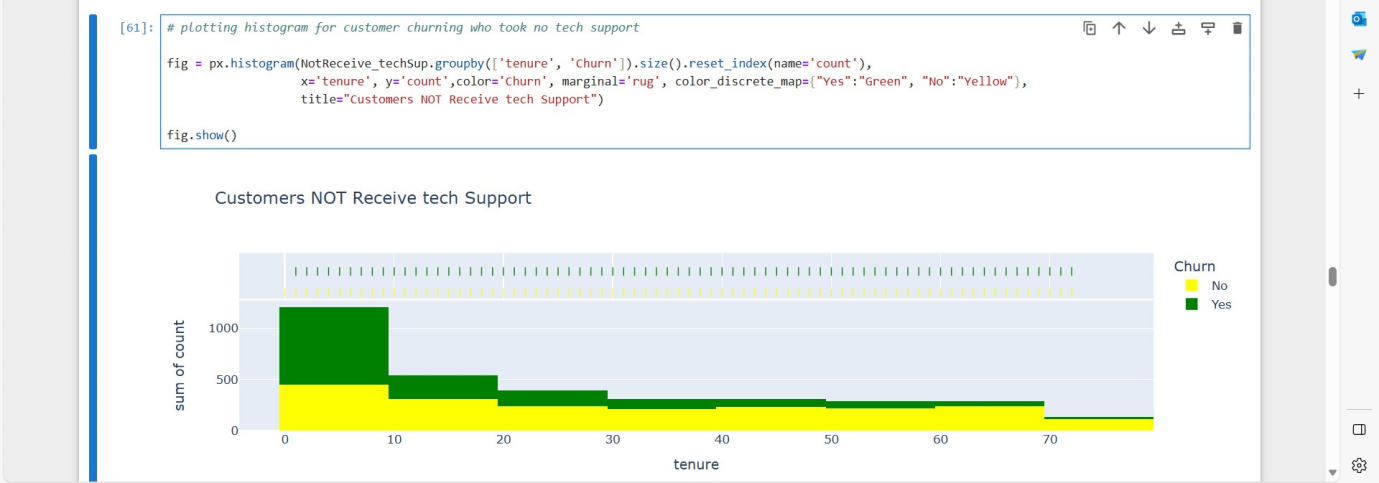
# plotting histogram for customer churning who took no tech support

fig = px.histogram(NotReceive\_techSup.groupby(['tenure', 'Churn']).size().reset\_index(name='count'),

x='tenure', y='count',color='Churn', marginal='rug', color\_discrete\_map={"Yes":"Green", "No":"Yellow"},

title="Customers NOT Receive tech Support")

fig.show()



# plotting histogram for customer churning who took no tech support

fig = px.histogram(Receive\_techSup.groupby(['tenure', 'Churn']).size().reset\_index(name='count'),

x='tenure', y='count',color='Churn', marginal='rug', color\_discrete\_map={"Yes":"Green", "No":"Yellow"},

title="Customers NOT Receive tech Support")

fig.show()



Note:

The data indicates that churn rates are highest within the first year of service, especially among customers without tech support. Conversely, longer tenure is associated with increased customer loyalty.

**5** .Which aspect of the contract has the most significant impact on the business?

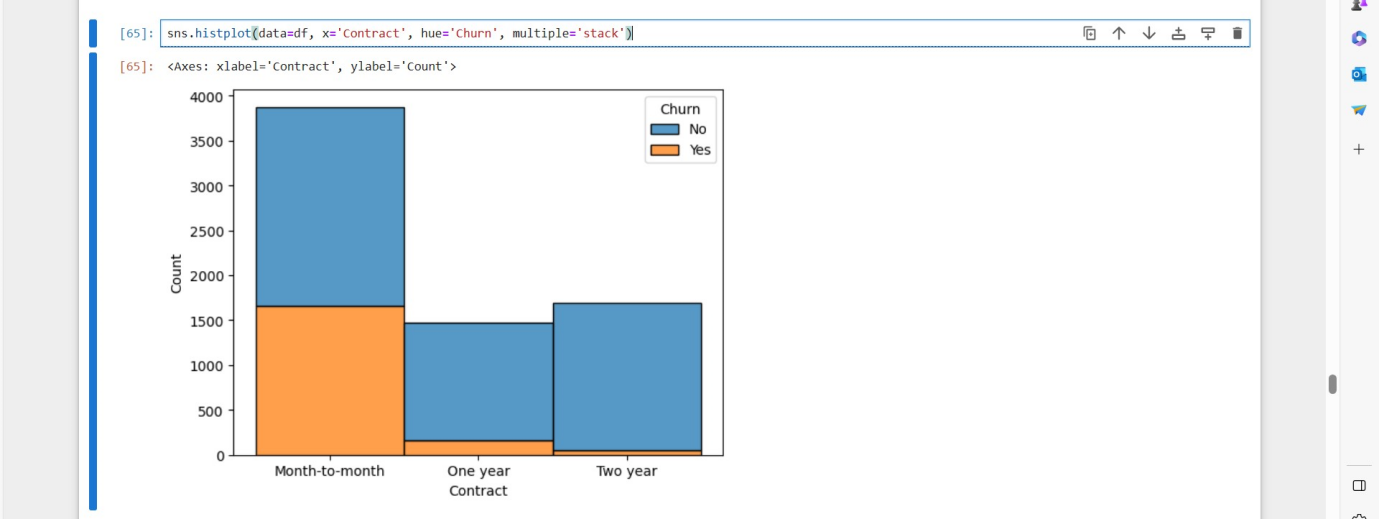
df.Contract.value\_counts(normalize =True)

Contract\_condition = df.groupby(['Churn','Contract']).size().reset\_index(name='count')

Contract\_condition



sns.histplot(data=df, x='Contract', hue='Churn', multiple='stack')



Note:

It is evident that customers with month-to-month contracts have the highest churn rates.

**6** .How does the quality of service differ for customers who have opted for streaming services?

df.StreamingTV.value\_counts()

Streming = df.groupby (['Churn','StreamingTV']).size().reset\_index(name='count')

Streming



ig, axes = plt.subplots(1,2,figsize = (10,5))

# Plot the distribution of 'StreamingTV'

sns.histplot(data=df, x='StreamingTV', hue='Churn', multiple='stack',ax=axes[0])

axes[0].set\_title('StreamingTV',fontsize=12)

axes[0].set\_xlabel('StreamingTV')

axes[0].set\_ylabel('Count')

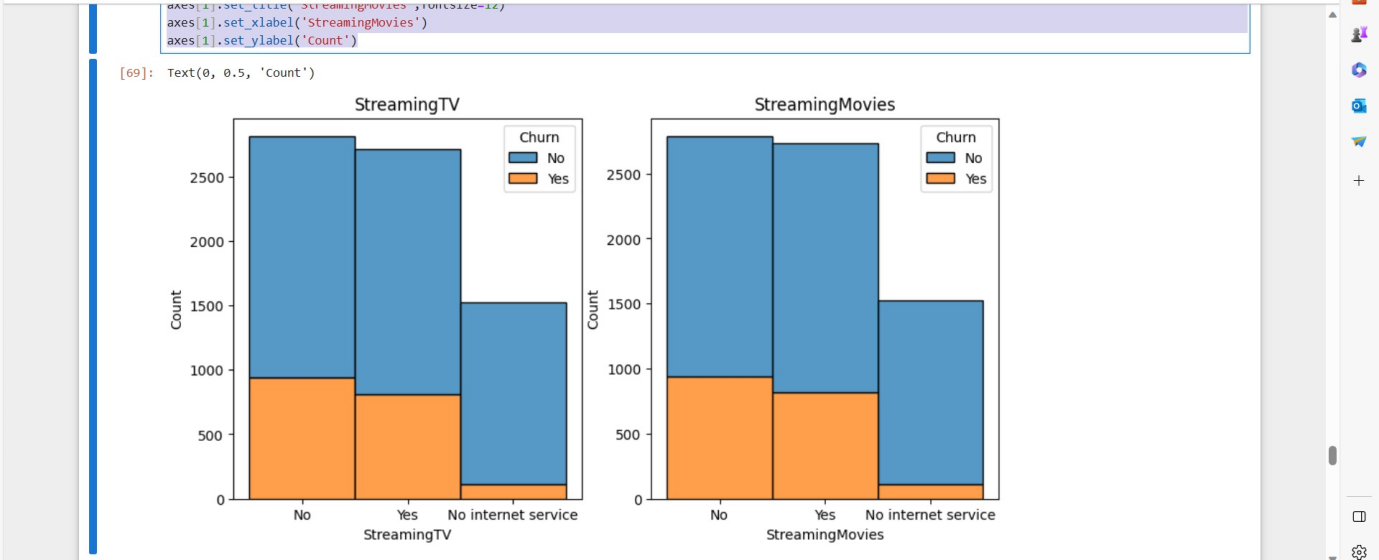
# Plot the distribution of 'StreamingMovies'

sns.histplot(data=df, x='StreamingMovies', hue='Churn', multiple='stack',ax=axes[1])

axes[1].set\_title('StreamingMovies',fontsize=12)

axes[1].set\_xlabel('StreamingMovies')

axes[1].set\_ylabel('Count')



Note: Churn rates are similar for both the 'Yes' and 'No' groups in terms of whether customers are connected to StreamingTV and StreamingMovies.

7. Given that the dataset pertains to the telecom industry, what insights can we uncover regarding phone and internet services?

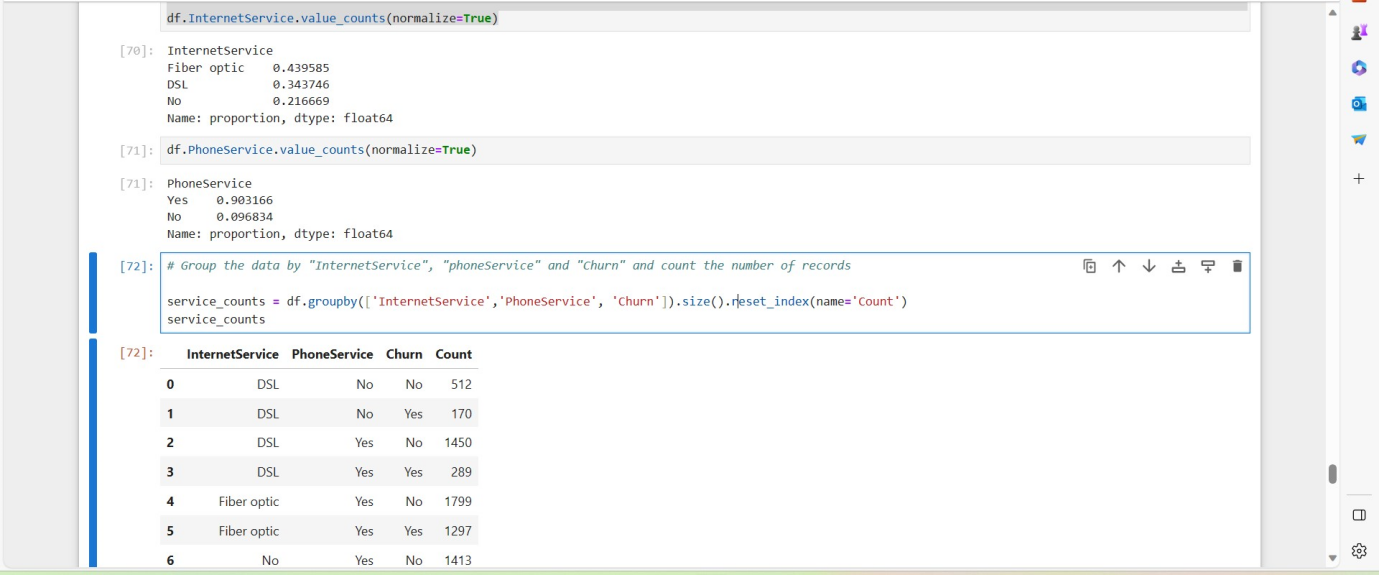
df.InternetService.value\_counts(normalize=True)

df.PhoneService.value\_counts(normalize=True)

# Group the data by "InternetService", "phoneService" and "Churn" and count the number of records

service\_counts = df.groupby(['InternetService','PhoneService', 'Churn']).size().reset\_index(name='Count')

service\_counts



plt.figure(figsize=(10, 5))

# Create two subplots

plt.subplot(121)

# 1 row, 2 columns, the first plot for Phone Service

sns.barplot(data=service\_counts[service\_counts['InternetService'] == 'DSL'], x='PhoneService', y='Count', hue='Churn', palette='viridis')

plt.xlabel('Phone Service')

plt.ylabel('Count')

plt.title('Counts by Phone Service (DSL Internet)')

plt.subplot(122)

# 1 row, 2 columns, the second plot for Internet Service

sns.barplot(data=service\_counts[service\_counts['InternetService'] != 'No'], x='InternetService', y='Count', hue='Churn', palette='viridis')

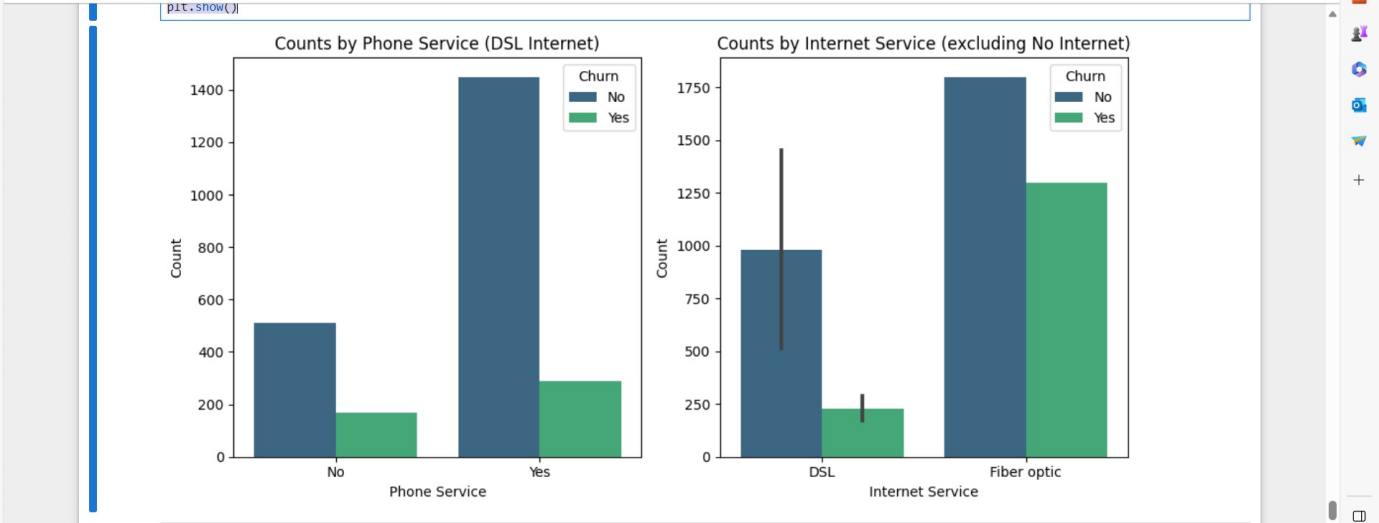
plt.xlabel('Internet Service')

plt.ylabel('Count')

plt.title('Counts by Internet Service (excluding No Internet)')

plt.tight\_layout()

plt.show()



Note:

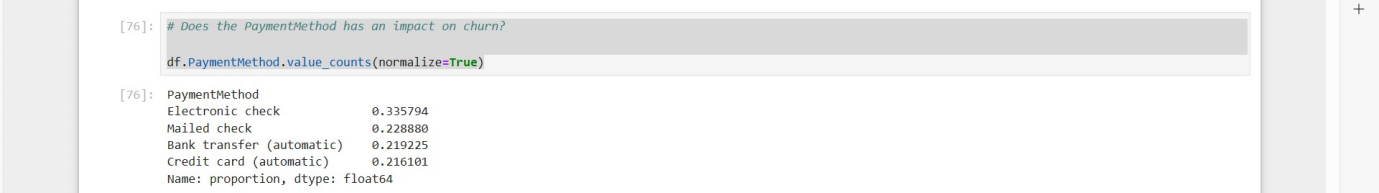
The plot suggests that the customers with phone service are loyal and have not churned.(the blue bar is significantly taller than green).

For customers without phone service, the blue bar for non-churn is still higher than the green bar. This indicate that even among those without phone service, a significant portion has not churned, and the absence of phone service is associated with a lower churn rate.

In summary, the chart illustrates that customers with phone service, as well as some customers without phone service, are less likely to churn, This suggests that both groups have a relatively lower churn rate, with a particularly strong retention rate among customers with phone service.

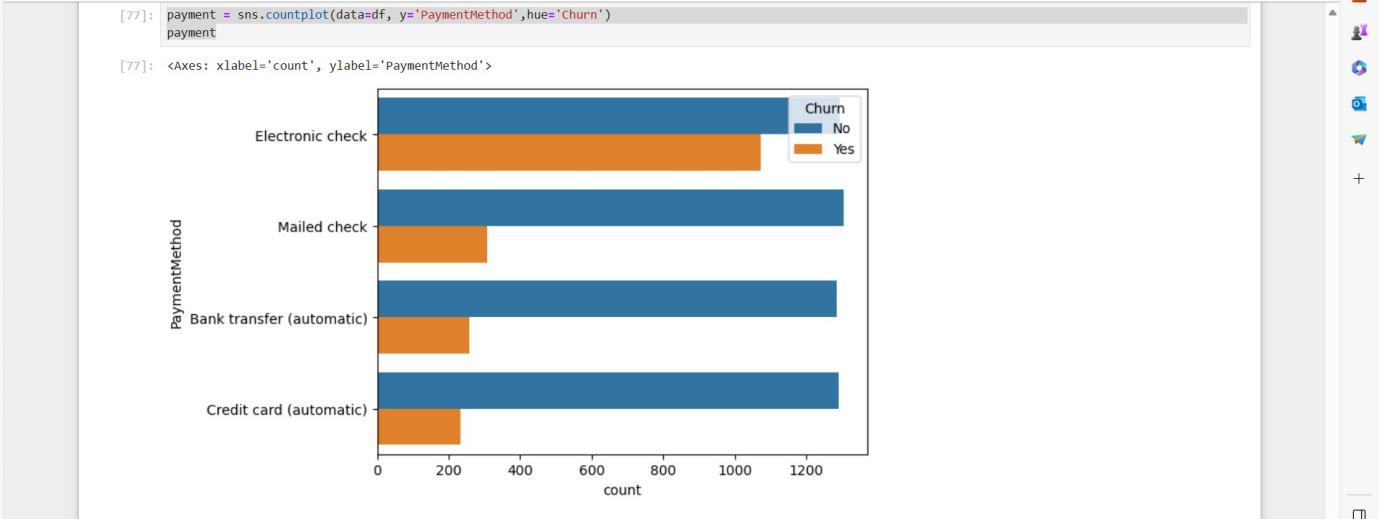
8. Does the PaymentMethod has an impact on churn?

df.PaymentMethod.value\_counts(normalize=True)



payment = sns.countplot(data=df, y='PaymentMethod',hue='Churn')

payment



**CONCLUSION:**

**EDA Outcome:**

**Observation:**

* The data reveals that the highest churn rate occurs within the first 0-10 years of customer tenure, suggesting that customers are less inclined to switch telecom providers as they become more familiar with their current one.
* Churning tends to increase with rising monthly charges, particularly in the 60-120 range, indicating a correlation between higher charges and increased churn.
* In contrast to monthly charges, the most significant churning occurs in the early phases, with the 0-2000 total charges bracket experiencing the highest churn rate.
* Regardless of gender, customers without partners or dependents are more likely to churn.
* Churning is more prevalent within the first 10 years of tenure for customers with or without tech support, with a higher rate among those without tech support.
* Customers who have both Phone services (yes) and 'Fiber optic' Internet Service are more prone to churn.
* The presence of StreamingTV, whether 'Yes' or 'No,' does not significantly affect the rate of churn; it appears to be relatively consistent
* The availability of StreamingMovies (either 'Yes' or 'No') does not appear to have a substantial impact on churn.
* Customers with month-to-month contracts are the most frequent churners, highlighting the importance of contract type in predicting churn.
* Senior citizens exhibit a lower likelihood of churning compared to non-senior citizens. It's worth noting that the dataset comprises a significantly higher proportion of non-senior citizens at a 5:1 ratio.