Customer Churn Prediction

Data Science Assignment 3 Fall 2023 BSCS 7A

Group Members:

1) 20L-0915 Manal Rizwan Qureshi (LEAD)

2) 20L-0921 Aisha Muhammad Nawaz

3) 20L-2156 Mariyam Ali

Exploratory Data Analysis and Visualization

In [1]:

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

In [2]:

```python
df = pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')
df.head()
```

Out[2]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service
1	5575-GNVDE	Male	0	No	No	34	Yes	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service
4	9237-HQITU	Female	0	No	No	2	Yes	No

5 rows × 21 columns

```
In [3]: df.keys()
```

```
Out[3]: Index(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
               'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
               'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
               'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',
               'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'],
              dtype='object')
```

Univariate Analysis

Begin by analyzing each variable individually. Use summary statistics and visualizations such as line charts, histograms, box plots, bar/pie charts, word clouds, and density plots to understand the distribution, central tendency, and variability of each variable.

Column 1: Customer ID (Has been dropped as it has no use)

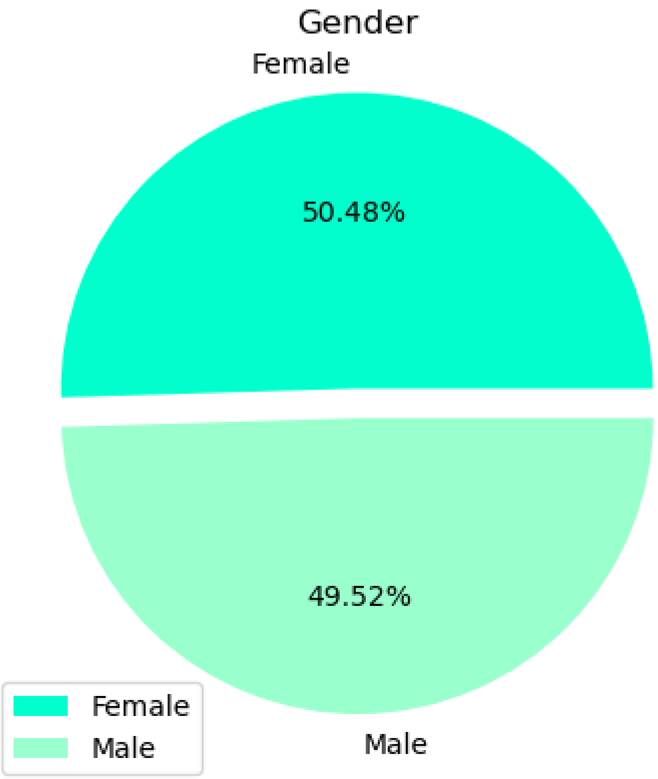
Column 2 : Gender

Tells whether the customer is Female or Male

```
In [4]: df['gender'].unique()
```

```
Out[4]: array(['Female', 'Male'], dtype=object)
```

```
In [5]: plt.pie(df['gender'].value_counts(),labels=['Female','Male'],autopct='%1.2f%%',explode=[0.1,0],colors=['#00ffcc','#99ffcc'])
        plt.title('Gender')
        plt.legend()
        plt.show()
```



Descriptive Analysis:

The Descriptive Analysis shows that 50.48% of customers are Female while 49.52% are Male.

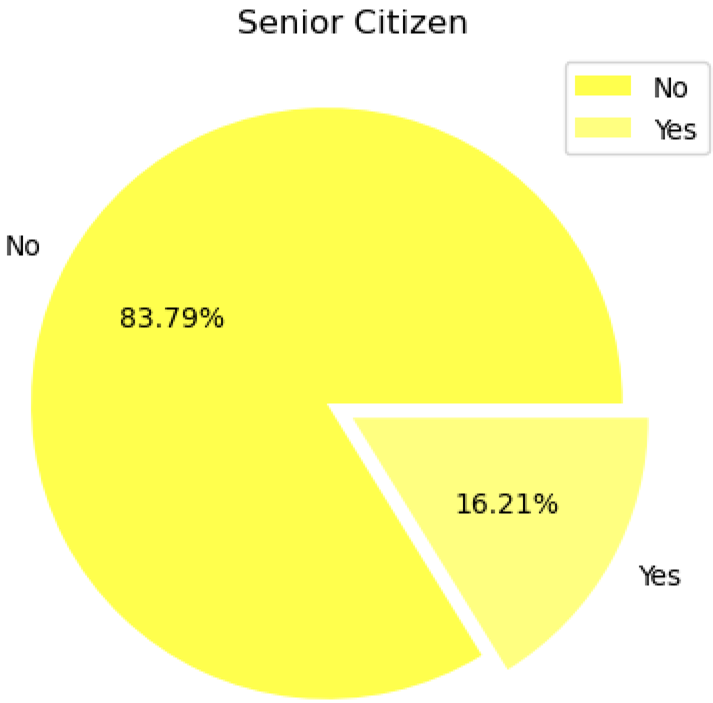
Column 3 : SeniorCitizen

Tells whether the customer is a senior citizen or not

```
In [6]: df['SeniorCitizen'].unique() # 0 stands for No and 1 stands for Yes
```

```
Out[6]: array([0, 1], dtype=int64)
```

```
In [7]: plt.pie(df['SeniorCitizen'].value_counts(),labels=['No','Yes'],autopct='%1.2f%%',explode=[0.1,0],colors=['#ffff4d','#ffff80'])
        plt.title('Senior Citizen')
        plt.legend()
        plt.show()
```



Descriptive Analysis:

The Descriptive Analysis shows that only 16.21% of the customers are senior citizens while the rest 83.79% are not.

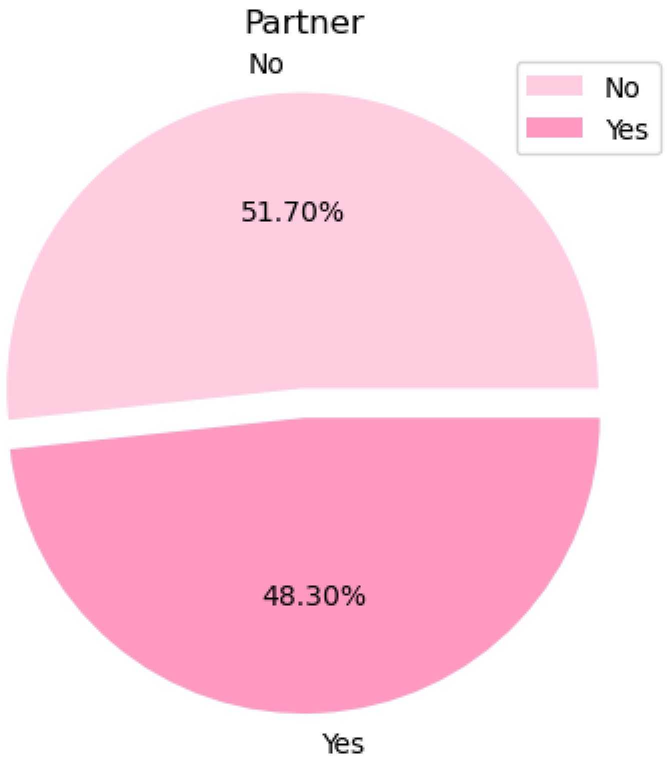
Column 4: Partner

Tells whether the customer has a partner or not

```
In [8]: df['Partner'].unique()
```

```
Out[8]: array(['Yes', 'No'], dtype=object)
```

In [9]:
```
plt.pie(df['Partner'].value_counts(),labels=df['Partner'].value_counts().keys
(),autopct='%1.2f%%',explode=[0.1,0],colors=['#ffcce0','#ff99c2'])
plt.title('Partner')
plt.legend()
plt.show()
```



Descriptive Analysis:

The above pie chart shows that 48.30% of the customers have a partner while 51.70% don't.

Column 5: Dependents

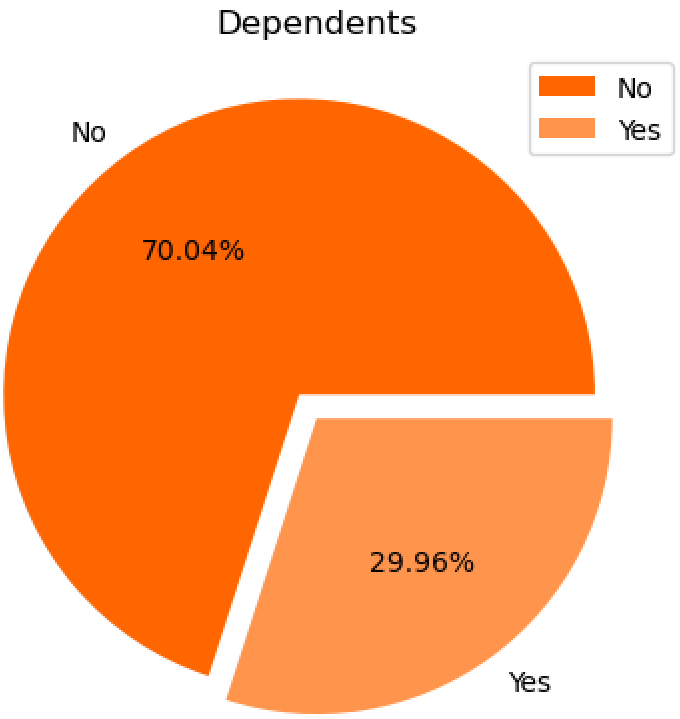
Tells whether the customer has dependents or not.

In [10]:
```
df['Dependents'].unique()
```

Out[10]:
```
array(['No', 'Yes'], dtype=object)
```

In [11]:

```
plt.pie(df['Dependents'].value_counts(),labels=df['Dependents'].value_counts
().keys(),autopct='%1.2f%%',explode=[0.1,0],colors=['#ff6600','#ff944d'])
plt.title('Dependents')
plt.legend()
plt.show()
```



Descriptive Analysis:

The visualization of the 'Dependents' column shows that 70.04% of the customers don't have dependents while 29.96% do.

Column 6: Tenure

Tells how long the customer has been with the company.

In [12]:

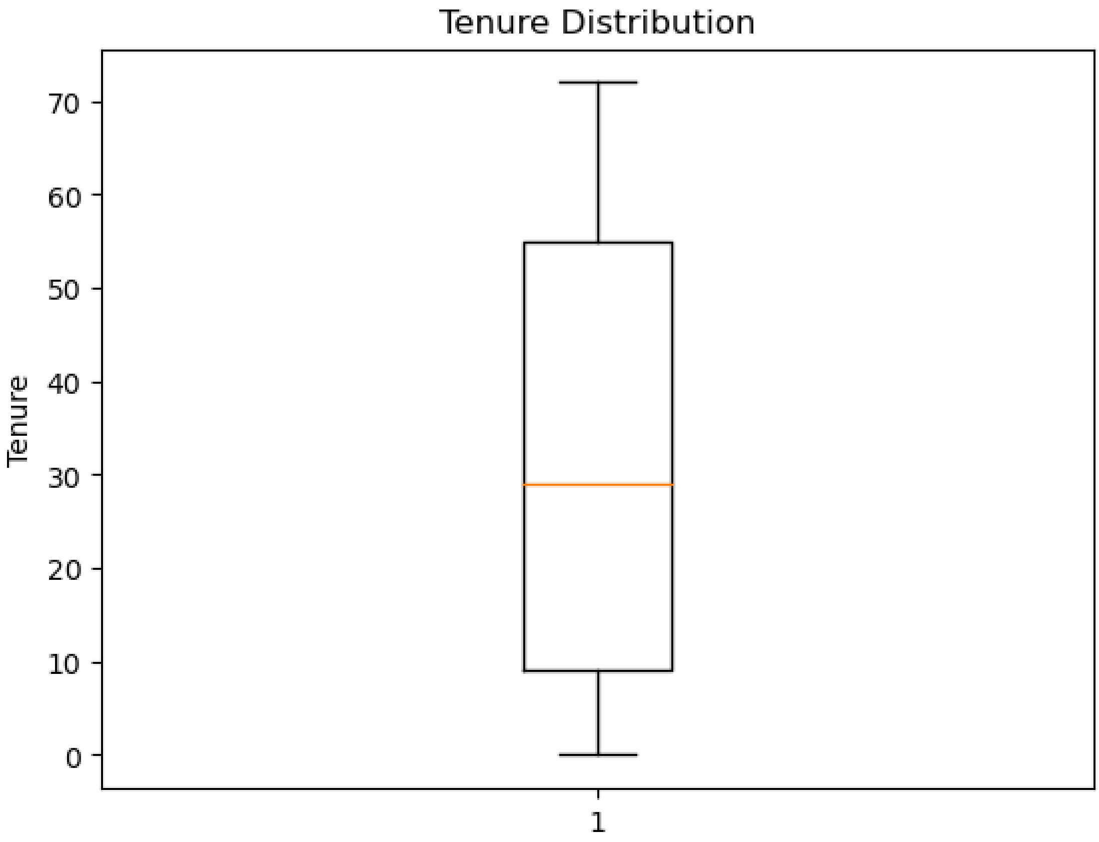
```
df['tenure'].describe()
```

Out[12]:

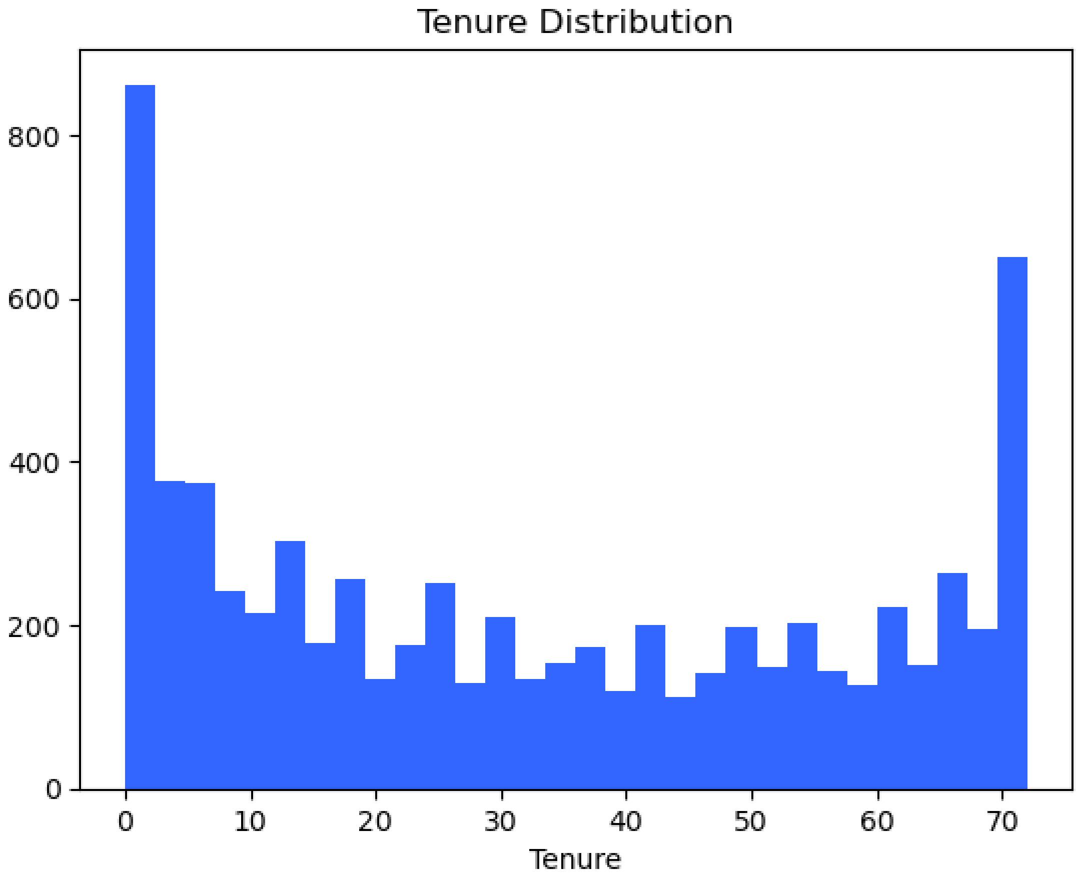
```
count    7043.000000
mean       32.371149
std        24.559481
min         0.000000
25%         9.000000
50%        29.000000
75%        55.000000
max        72.000000
Name: tenure, dtype: float64
```

In [13]:

```
plt.boxplot(df['tenure'])
plt.title('Tenure Distribution')
plt.ylabel('Tenure')
plt.show()
```



```
In [14]: plt.hist(df['tenure'],color='#3366ff',bins=30)
         plt.title('Tenure Distribution')
         plt.xlabel('Tenure')
         plt.show()
```



Descriptive Analysis:

The box plot and the histogram of the tenure column show that the data is slightly right-skewed which means that majority of the customers have been with the company for a short period of time. The average tenure is 32 months.

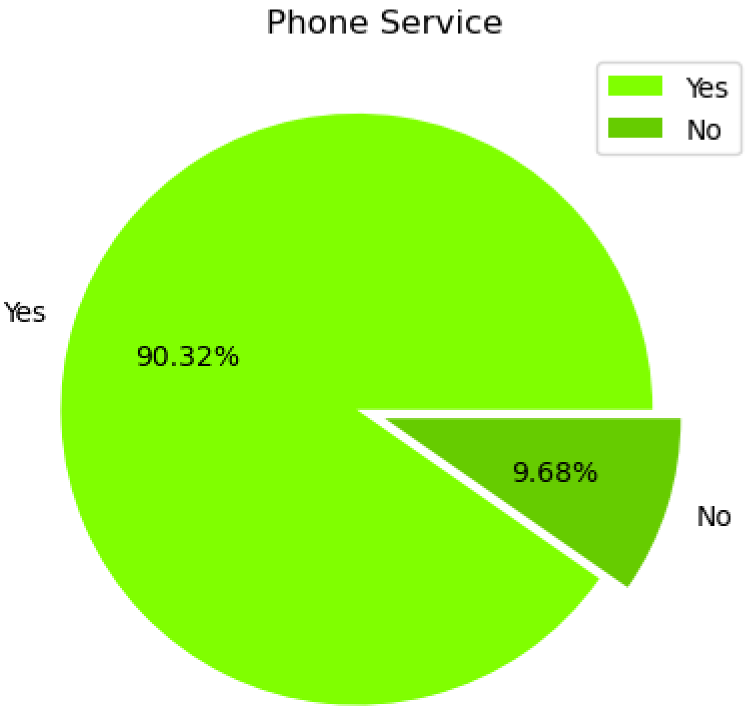
Column 7: PhoneService

Tells whether the customer has phone service or not.

```
In [15]: df['PhoneService'].unique()

Out[15]: array(['No', 'Yes'], dtype=object)
```

In [16]:

```
plt.pie(df['PhoneService'].value_counts(),labels=df['PhoneService'].value_counts().keys(),autopct='%1.2f%%',explode=[0.1,0],colors=['#80ff00','#66cc00'])
plt.title('Phone Service')
plt.legend()
plt.show()
```



Descriptive Analysis

The descriptive analysis shows that 90.32% of the customers have phone service while the rest 9.68% don't.

Column 8: MultipleLines

Tells whether the customer has multiple lines or not (Yes,No,No phone service).

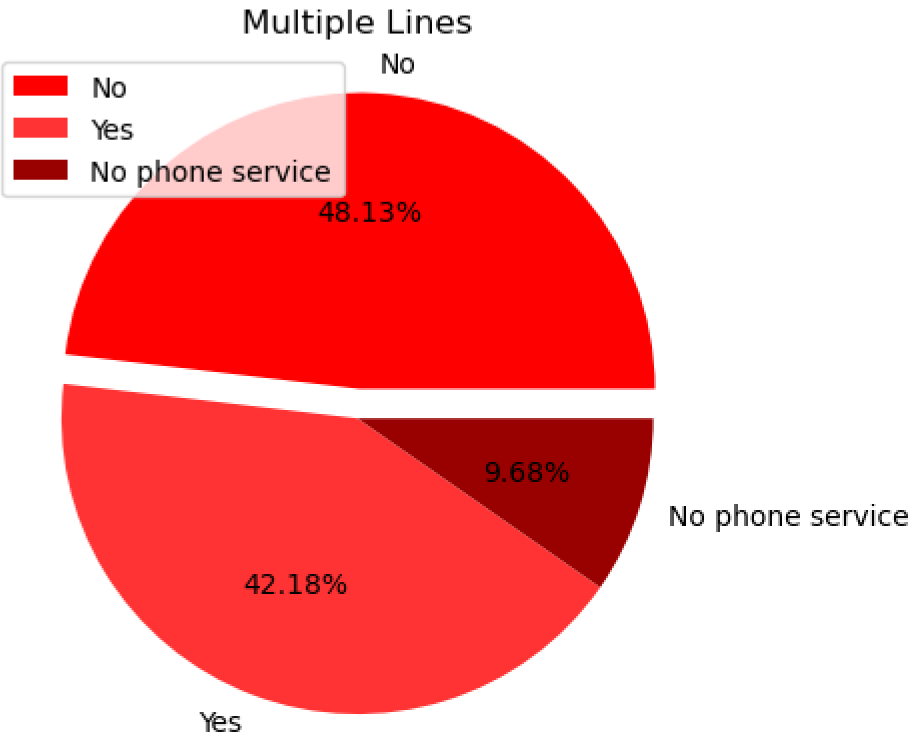
In [17]:

```
df['MultipleLines'].unique()
```

Out[17]:

```
array(['No phone service', 'No', 'Yes'], dtype=object)
```

In [18]:
```
plt.pie(df['MultipleLines'].value_counts(),labels=df['MultipleLines'].value_counts().keys(),autopct='%1.2f%%',explode=[0.1,0,0],colors=['#ff0000','#ff3333','#990000'])
plt.title('Multiple Lines')
plt.legend()
plt.show()
```



Descriptive Analysis

The above pie chart shows that 48.13% of customers have a single line, 42.18% have multiple lines while the rest 9.68% don't even have phone service

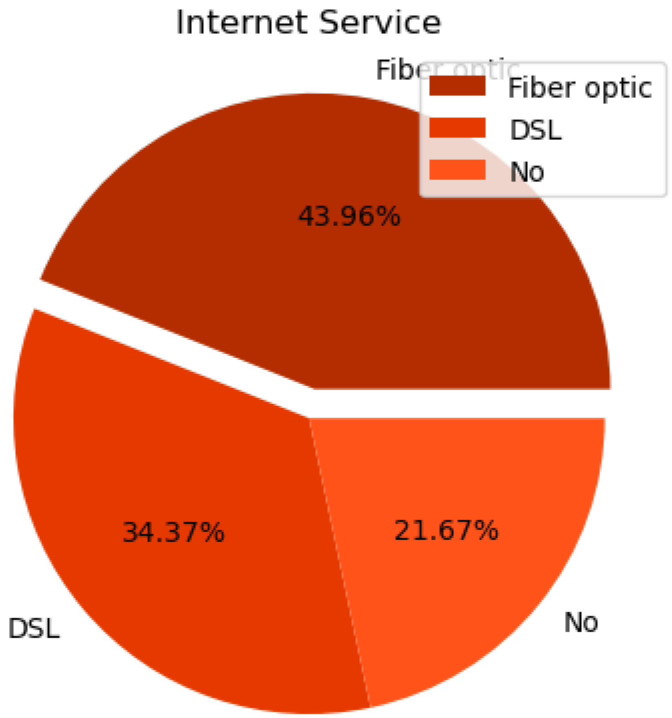
Column 9: InternetService

Tells which type of Internet Service the customer has (DSL,Fiber optic,No).

In [19]:
```
df['InternetService'].unique()
```

Out[19]:
```
array(['DSL', 'Fiber optic', 'No'], dtype=object)
```

In [20]:
```
plt.pie(df['InternetService'].value_counts(),labels=df['InternetService'].value_counts().keys(),autopct='%1.2f%%',explode=[0.1,0,0],colors=['#b32d00','#e63900','#ff531a'])
plt.title('Internet Service')
plt.legend()
plt.show()
```



Descriptive Analysis:

The visualization of the 'InternetService' column shows that 43.96% of the customers prefer Fiber optic, 34.37% prefer DSL while 21.67% don't even have internet service.

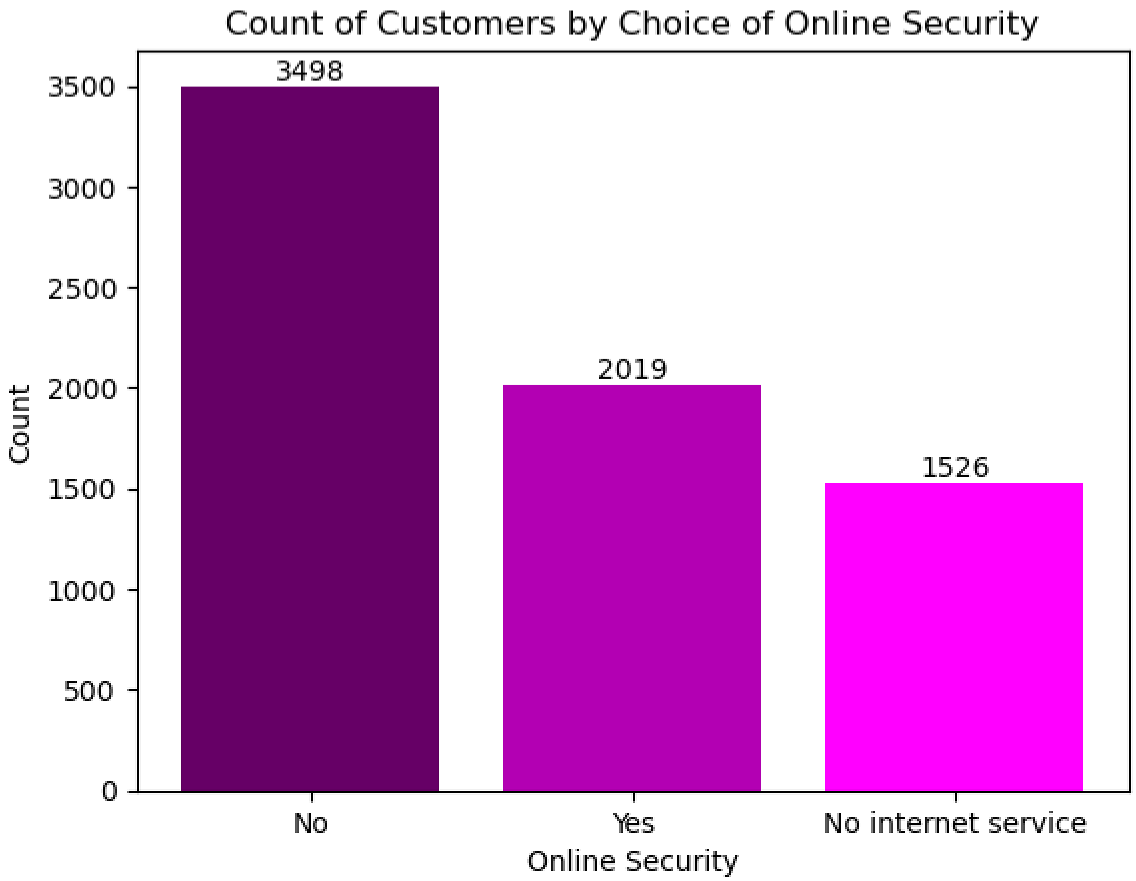
Column 10: OnlineSecurity

Tells whether the customer has Online Security or not.(No,Yes,No internet service)

In [21]:
```
df['OnlineSecurity'].unique()
```

Out[21]:
```
array(['No', 'Yes', 'No internet service'], dtype=object)
```

In [22]:
```
plt.bar(df['OnlineSecurity'].value_counts().keys(),df['OnlineSecurity'].value_counts(),color=['#660066','#b300b3','#ff00ff'])
plt.title('Count of Customers by Choice of Online Security')
plt.xlabel('Online Security')
plt.ylabel('Count')

for x, y in zip(df['OnlineSecurity'].value_counts().index, df['OnlineSecurity'].value_counts().values):
    plt.text(x, y, f'{y}', ha='center', va='bottom')

plt.show()
```



Descriptive Analysis:

The above bar chart shows that 3498 customers don't have online security, 2019 customers prefer online security while the rest 1526 customers don't even have internet service.

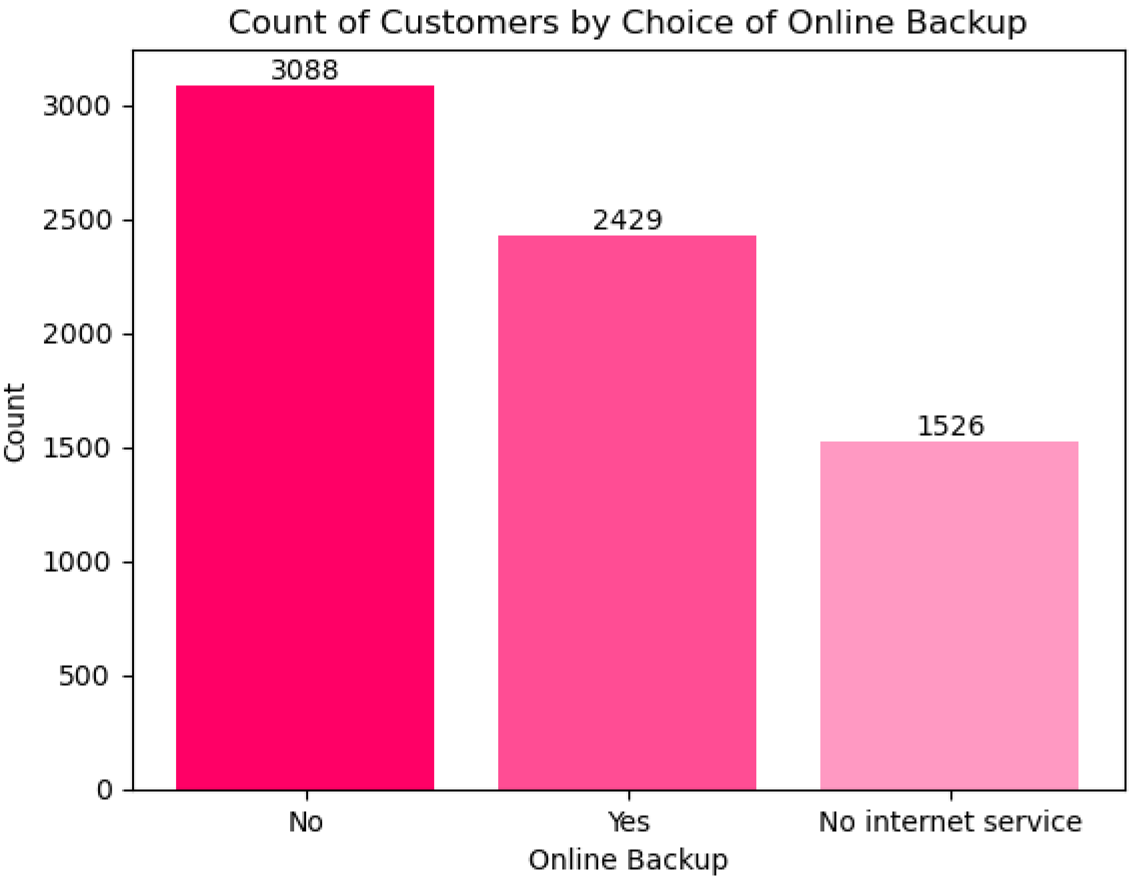
Column 11: OnlineBackup

Tells whether the customer has Online Backup or not.(No,Yes,No internet service)

```
In [23]: df['OnlineBackup'].unique()
```

```
Out[23]: array(['Yes', 'No', 'No internet service'], dtype=object)
```

```
In [24]: plt.bar(df['OnlineBackup'].value_counts().keys(),df['OnlineBackup'].value_coun
ts(),color=['#ff0066','#ff4d94','#ff99c2'])
plt.title('Count of Customers by Choice of Online Backup')
plt.xlabel('Online Backup')
plt.ylabel('Count')

for x, y in zip(df['OnlineBackup'].value_counts().index, df['OnlineBackup'].va
lue_counts().values):
    plt.text(x, y, f'{y}', ha='center', va='bottom')

plt.show()
```



Descriptive Analysis:

The descriptive analysis shows that 3088 customers don't have online backup, 2429 customers prefer online backup while the rest 1526 customers don't even have internet service.

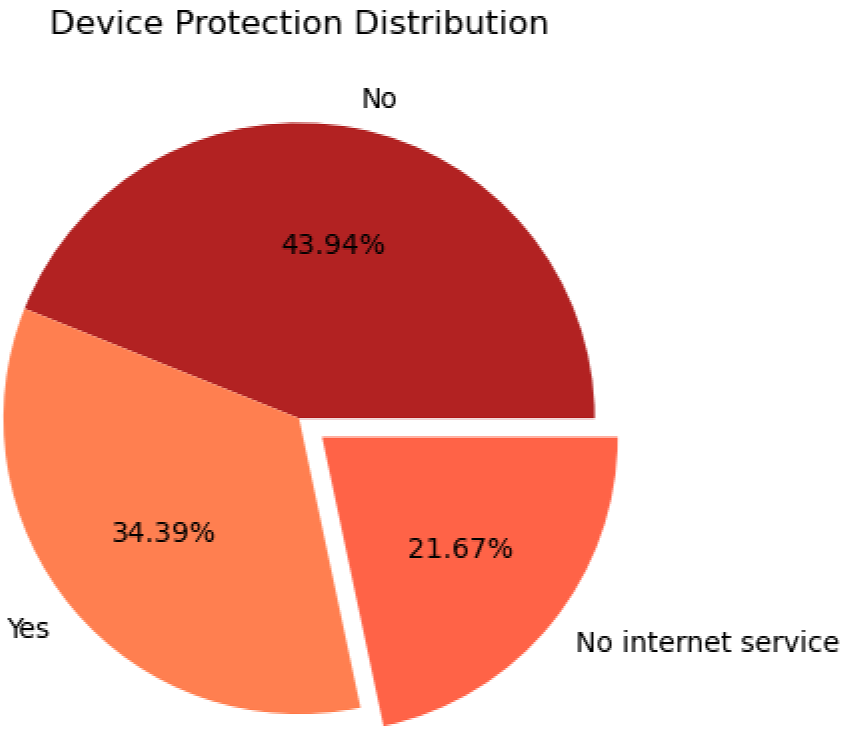
Column 12: DeviceProtection

Tells whether the customer has device protection or not (Yes, No, No internet service)

```
In [25]: df['DeviceProtection'].unique()
```

```
Out[25]: array(['No', 'Yes', 'No internet service'], dtype=object)
```

```
In [26]: plt.pie(df['DeviceProtection'].value_counts(),labels=df['DeviceProtection'].va
lue_counts().keys(),autopct='%0.2f%%',colors=['firebrick','coral','tomato'],ex
plode=[0.0,0.0,0.1])
plt.title('Device Protection Distribution')
plt.show()
```



Descriptive Analysis:

We found that a sizable majority of customers, or about 43.94%, skip Device Protection services in our thorough examination of customer preferences. Additionally, a sizeable minority (21.67%) choose not to use internet services. As an illustration of the wide range of options available to customers, a sizable portion of customers—roughly 34.39%—actively embrace device protection.

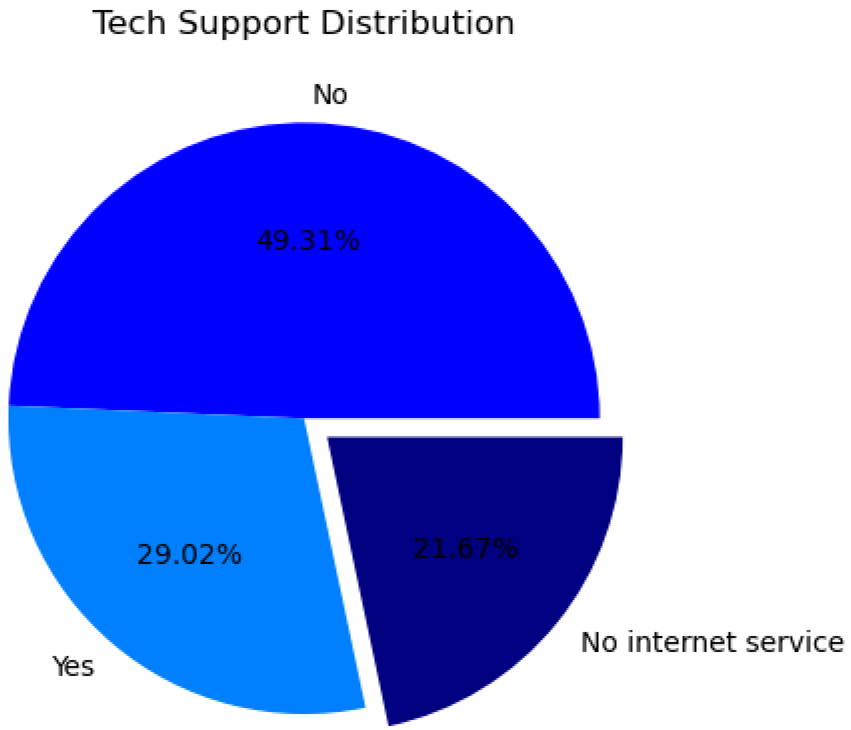
Column 13: TechSupport

Tells whether the customer has tech support or not (Yes, No, No internet service).

```
In [27]: df['TechSupport'].unique()
```

```
Out[27]: array(['No', 'Yes', 'No internet service'], dtype=object)
```

```
In [28]: plt.pie(df['TechSupport'].value_counts(),labels=df['TechSupport'].value_counts
         ().keys(),autopct='%0.2f%%',colors = ['#0000FF', '#007FFF',  '#000080'] ,explo
         de=[0.0,0.0,0.1])
         plt.title('Tech Support Distribution')
         plt.show()
```



Descriptive Analysis:

In our Descriptive Analysis, we found that a sizable fraction of clients, at 49.31%, do not use tech support services. Meanwhile, 29.02% of users actively interact with tech assistance, highlighting its applicability. Furthermore, about 21.67% of the customer base is not an internet service user.

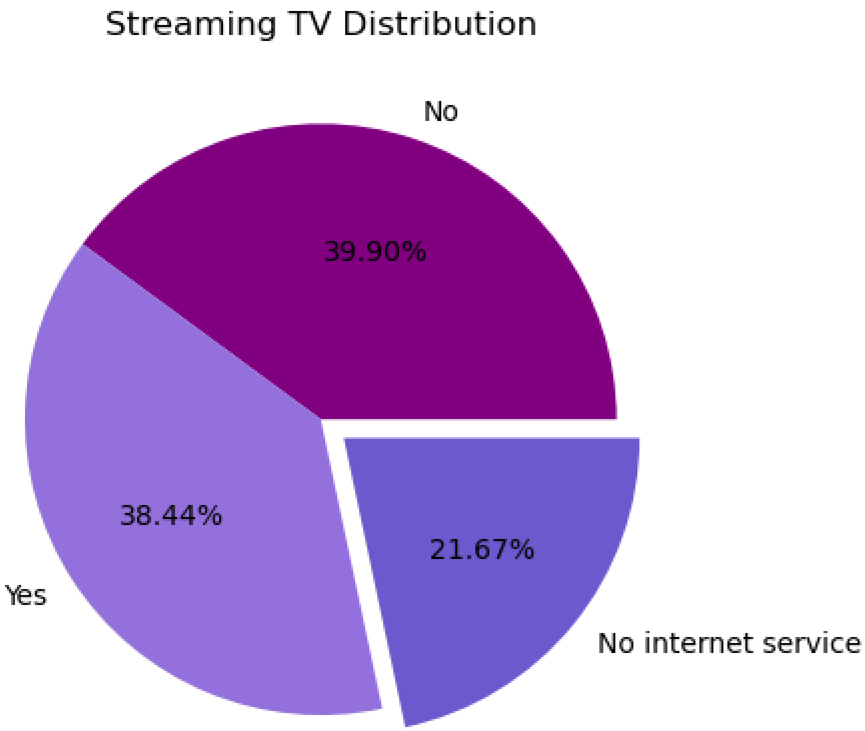
Column 14: StreamingTV

Tells whether the customer has streaming TV or not (Yes, No, No internet service).

```
In [29]: df['StreamingTV'].unique()
```

```
Out[29]: array(['No', 'Yes', 'No internet service'], dtype=object)
```

```
In [30]: plt.pie(df['StreamingTV'].value_counts(),labels=df['StreamingTV'].value_counts().keys(),autopct='%0.2f%%',colors = ['#800080', '#9370DB', '#6A5ACD'] ,explode=[0.0,0.0,0.1])
         plt.title('Streaming TV Distribution')
         plt.show()
```



Descriptive Analysis:

Our Descriptive Analysis shows that 38.44% of the consumers actively use streaming TV services, compared to a greater number of 39.90% of the customers who do not subscribe to these services. Additionally, a significant portion of the clientele—21.67% of them—does not have internet service.

Column 15: StreamingMovies

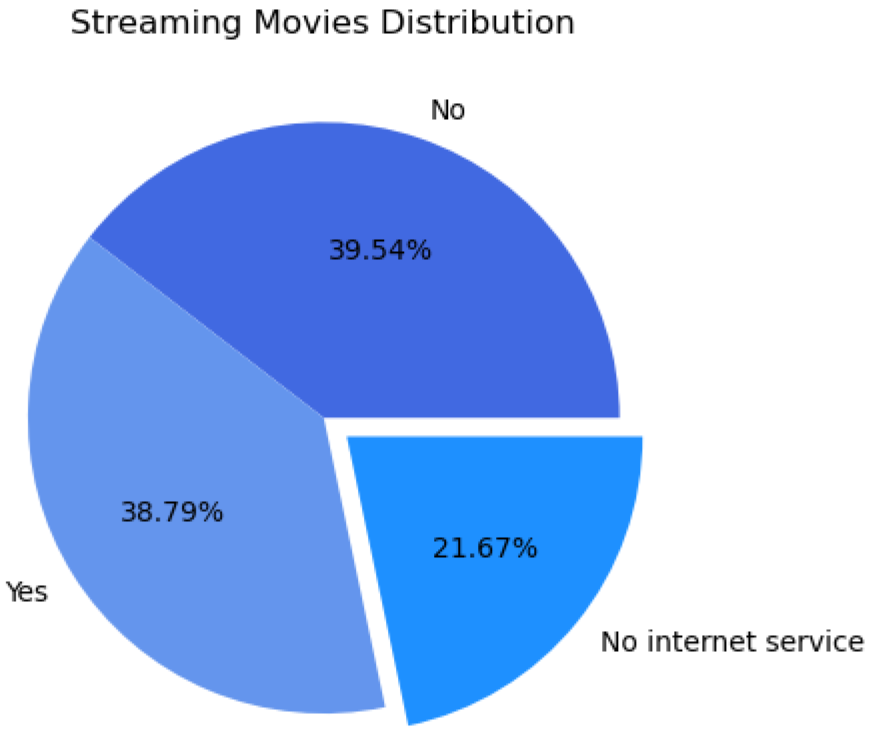
Tells whether the customer has streaming movies or not (Yes, No, No internet service)

```
In [31]: df['StreamingMovies'].unique()
```

```
Out[31]: array(['No', 'Yes', 'No internet service'], dtype=object)
```

In [32]:

```
plt.pie(df['StreamingMovies'].value_counts(),labels=df['StreamingMovies'].value_counts().keys(),autopct='%0.2f%%',colors = ['#4169E1', '#6495ED', '#1E90FF'],explode=[0.0,0.0,0.1])
plt.title('Streaming Movies Distribution')
plt.show()
```



Descriptive Analysis:

In our Descriptive Analysis, we found that 38.79% of customers have streaming movie subscriptions, compared to a greater number of 39.54% who do not. Furthermore, a significant portion of the clientele—21.67% of them—does not have internet service.

Column 16: Contract

Tells the contract term of the customer (Month-to-month, One year, Two years)

In [33]:

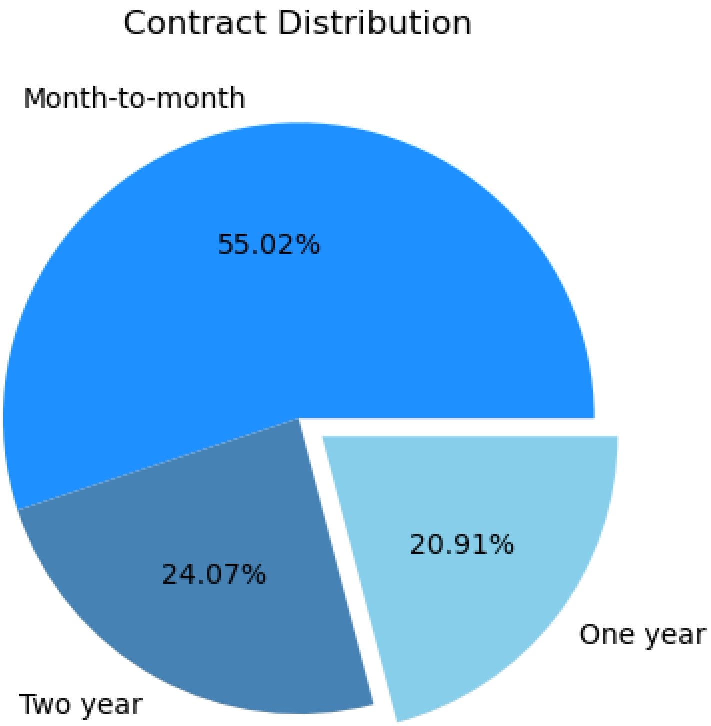
```
df['Contract'].unique()
```

Out[33]:

```
array(['Month-to-month', 'One year', 'Two year'], dtype=object)
```

In [34]:

```
plt.pie(df['Contract'].value_counts(),labels=df['Contract'].value_counts().keys(),autopct='%0.2f%%',colors = ['#1E90FF', '#4682B4', '#87CEEB'] ,explode=[0.0,0.0,0.1])
plt.title('Contract Distribution')
plt.show()
```



Descriptive Analysis:

In our Descriptive Analysis, we found that a majority of clients, around 55.02%, have month-to-month contract making it the most desirable among customers. Meanwhile, 24.07% of users have Two year contracts, highlighting its applicability. Furthermore, about 20.91% of the customer base have one year contract - the least famous type of contract.

Column 17 : PaperlessBilling

Tells whether the customer has paperless billing or not (Yes, No).

In [35]:

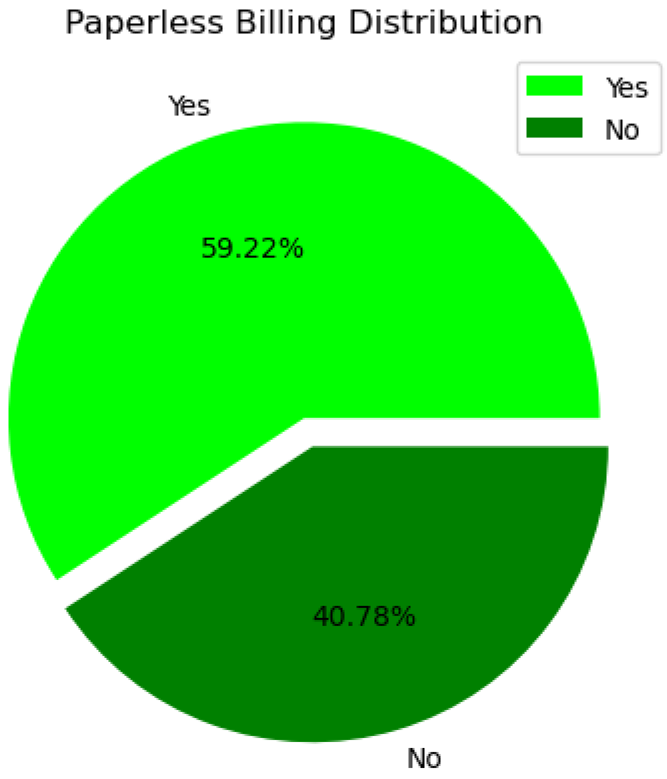
```
df['PaperlessBilling'].unique()
```

Out[35]:

```
array(['Yes', 'No'], dtype=object)
```

In [36]:

```
plt.pie(df['PaperlessBilling'].value_counts(),labels=df['PaperlessBilling'].value_counts().keys(),autopct='%0.2f%%',colors = ['#00FF00', '#008000'] ,explode=[0.0,0.1])
plt.title('Paperless Billing Distribution')
plt.legend()
plt.show()
```



Descriptive Analysis:

According to our Descriptive Analysis, a sizable majority of the customers—59.22%—accept paperless billing, demonstrating a preference for digital convenience. In contrast, 40.78% of our clients use traditional billing methods, demonstrating the wide range of options available to them.

Column 18: PaymentMethod

Tells the customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).

In [37]:

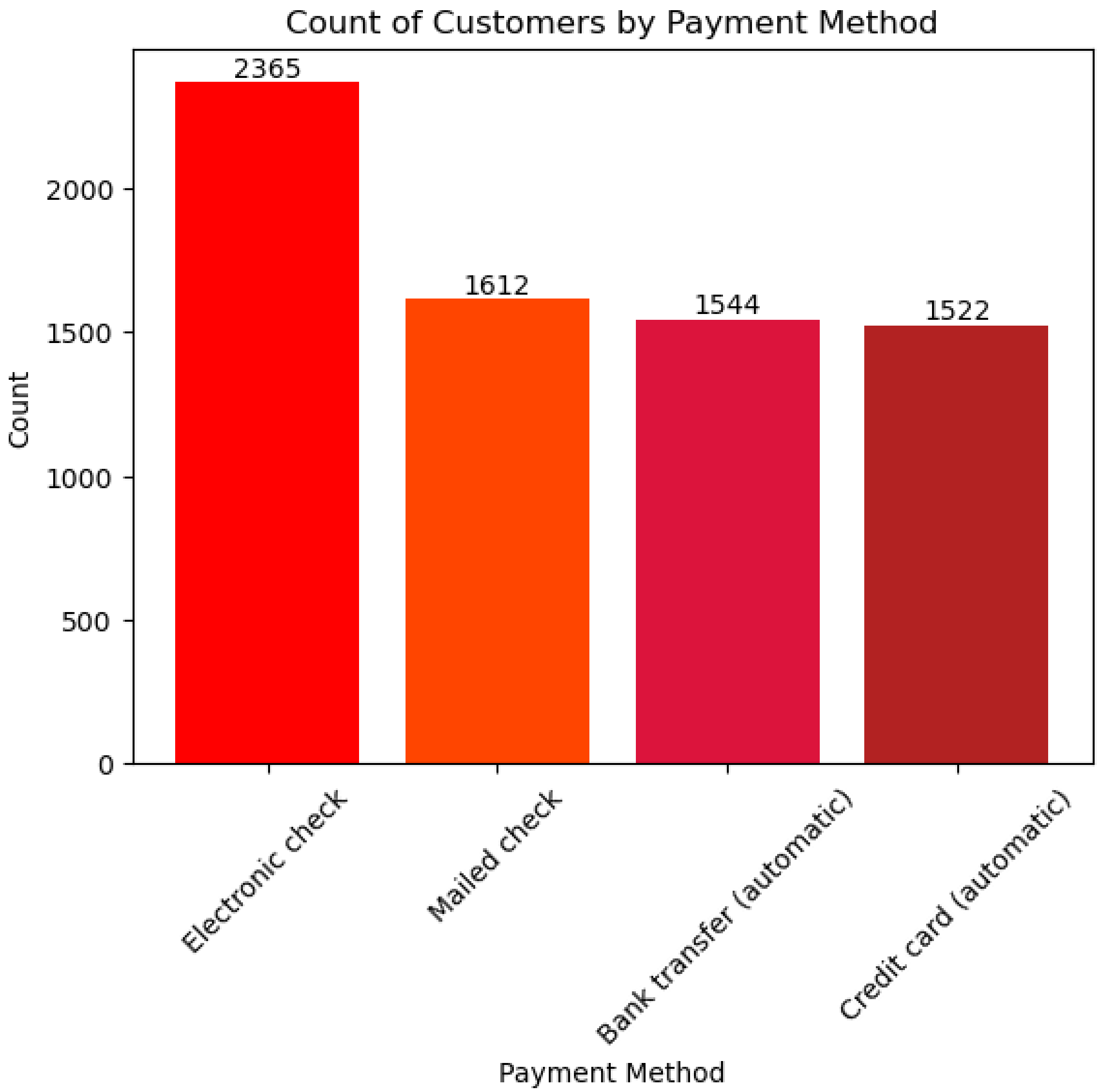
```
df['PaymentMethod'].unique()
```

Out[37]:

```
array(['Electronic check', 'Mailed check', 'Bank transfer (automatic)',
       'Credit card (automatic)'], dtype=object)
```

In [38]:

```
methods = df['PaymentMethod'].value_counts()
plt.bar(methods.index, methods.values,color=['#FF0000', '#FF4500', '#DC143C',
'#B22222'])
plt.xticks(rotation=45)
plt.xlabel('Payment Method')
plt.ylabel('Count')
plt.title('Count of Customers by Payment Method')

# Adding annotations
for x, y in zip(methods.index, methods.values):
    plt.text(x, y, f'{y}', ha='center', va='bottom')
plt.show()
```



Descriptive Analysis:

With 2,365 people picking it as their preferred method of payment, electronic checks are the most popular payment method among the customers. This implies a preference for instantaneous payment processing and digital transactions. The fact that Mailed check is the second most popular choice (1612), however, shows that a sizeable proportion of consumers still use conventional paper checks to make payments. Additionally popular but with significantly fewer users are Bank Transfer (automatic),1544 customers use it and Credit Card (automatic), 1522 customers use it, demonstrating the wide range of payment preferences among the customers.

Column 19: MonthlyCharges

Tells the amount charged to the customer monthly

In [39]:

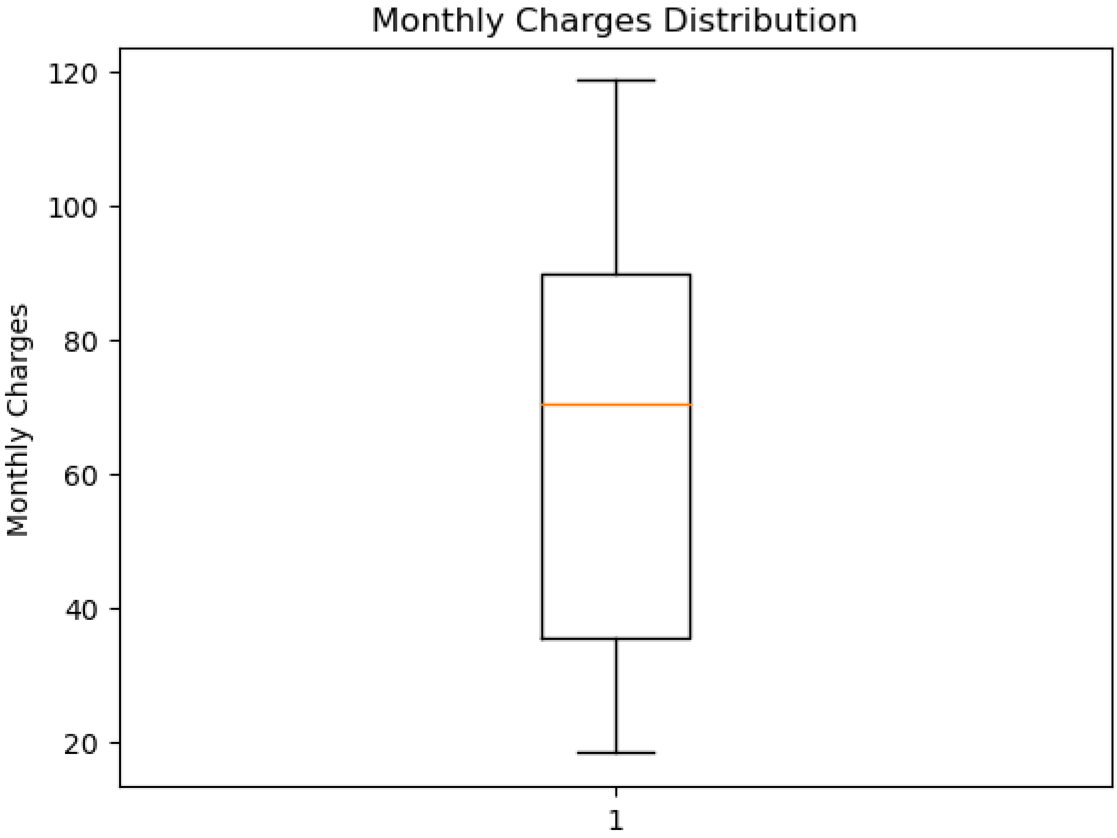
```
df.MonthlyCharges.describe()
```

Out[39]:

```
count    7043.000000
mean       64.761692
std        30.090047
min        18.250000
25%        35.500000
50%        70.350000
75%        89.850000
max       118.750000
Name: MonthlyCharges, dtype: float64
```

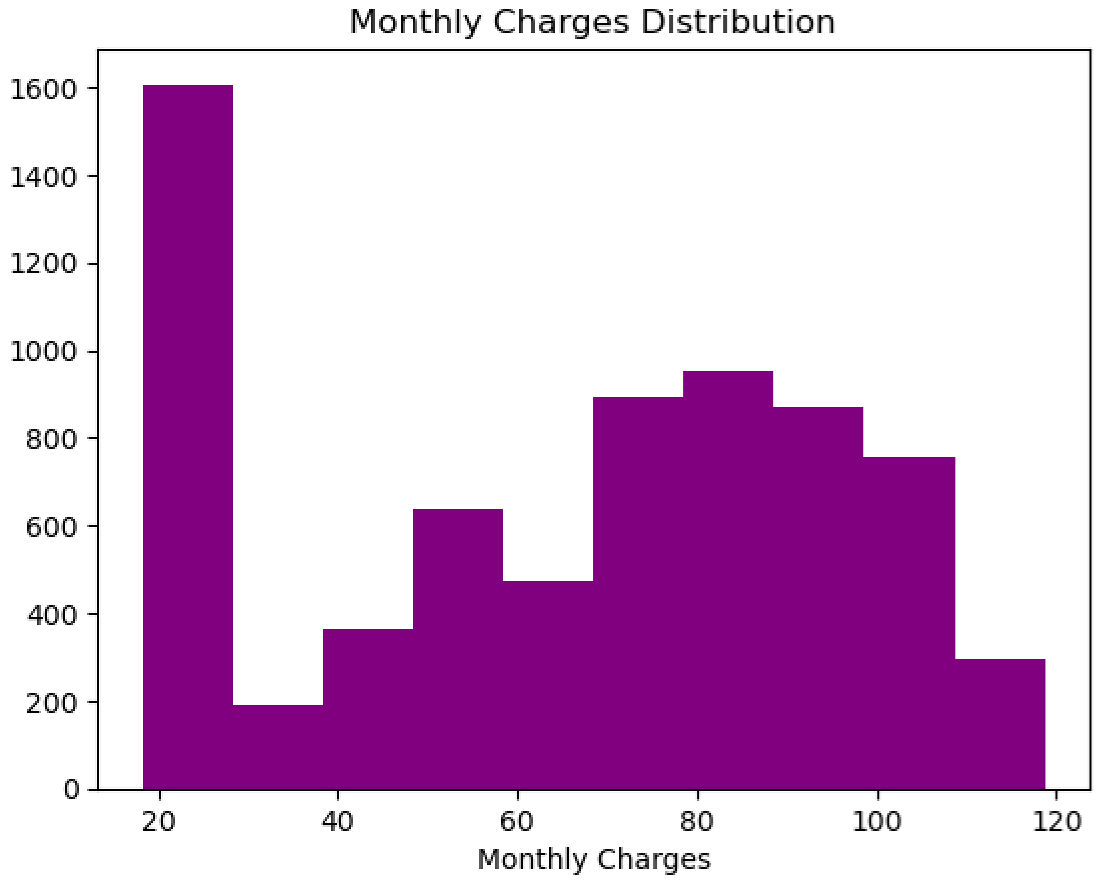
In [40]:

```
plt.boxplot(df['MonthlyCharges'])
plt.title('Monthly Charges Distribution')
plt.ylabel('Monthly Charges')
plt.show()
```



In [41]:

```
plt.hist(df['MonthlyCharges'],color='purple')
plt.title('Monthly Charges Distribution')
plt.xlabel('Monthly Charges')
plt.show()
```



Descriptive Analysis:

In our examination of monthly charges, we found that the data had a distribution that was slightly right-skewed and concentrated on lower monthly charge values. The following are the monthly charge summary statistics: This dataset contains 7,043 records with an average monthly charge of 64.76 dollars and a standard deviation of 30.09 dollars. The 25th percentile (Q1) is at 35.50 dollars, while the minimum monthly fee is 18.25 dollars. The 50th percentile median charge is 70.35 dollars, and the 75th percentile (Q3) median charge is 89.85 dollars. The highest monthly fee that has been noted is $118.75. These figures show the variety of monthly expenses among the customers and provide useful insights on the distribution and summary features of monthly charges.

Column 20: TotalCharges

Tells the total amount charged to the customer

In [42]:

```
df.TotalCharges=pd.to_numeric(df.TotalCharges,errors='coerce')
```

In [43]:

```
df.dropna(inplace=True)
```

In [44]:

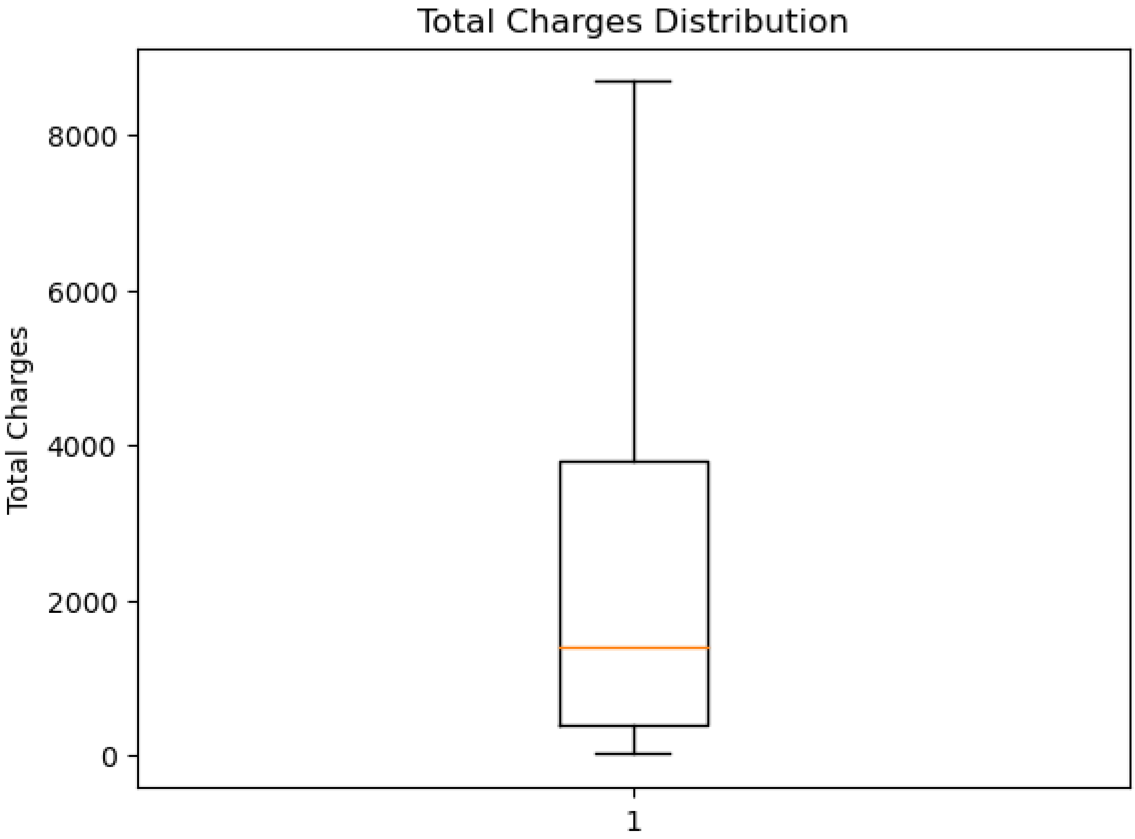
```
df.TotalCharges.describe()
```

Out[44]:

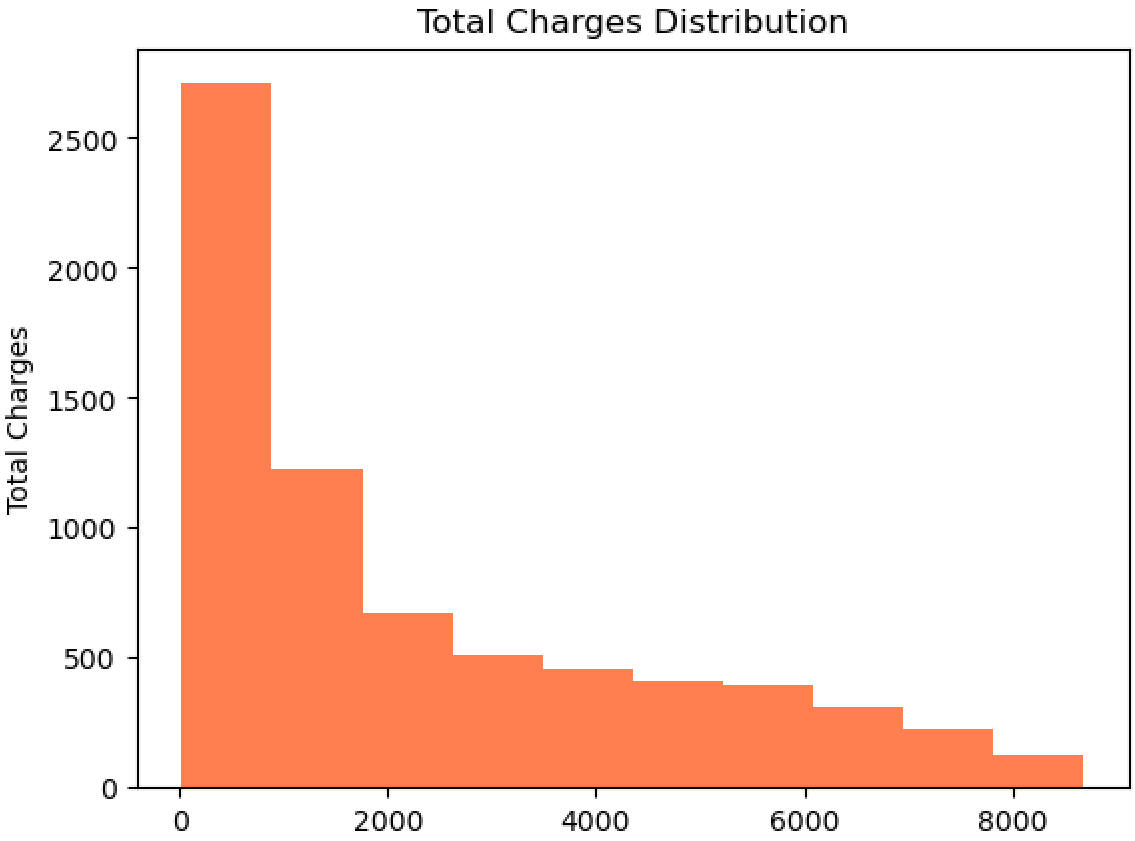
```
count    7032.000000
mean     2283.300441
std      2266.771362
min        18.800000
25%       401.450000
50%      1397.475000
75%      3794.737500
max      8684.800000
Name: TotalCharges, dtype: float64
```

In [45]:

```
plt.boxplot(df['TotalCharges'])
plt.title('Total Charges Distribution')
plt.ylabel('Total Charges')
plt.show()
```



In [46]:
```
plt.hist(df['TotalCharges'],color='coral')
plt.title('Total Charges Distribution')
plt.ylabel('Total Charges')
plt.show()
```



Descriptive Analysis:

In our descriptive analysis, we see that there is a concentration of lower values and that the distribution of total costs for consumers is favorably skewed. According to this skewed distribution, the bulk of clients likely have significantly lower overall charges. The distribution of the data is not typical.

The number of 7,032 consumers, the mean charge of 2,283.30 dollars and the standard deviation of 2,266.77 dollars are the key figures for total charges. The lowest total charge is dollar 18.80, and the 25 percent (Q1) average is 401.45 dollars. Costs range from 1,397.48 dollars at the median (50th percentile) to 3,794.74 dollars at the 75th percentile (Q3). 8,684.80 dollars is the largest total charge that has been noted. This data gives us insights into the total charges for the customers base's distribution and summary statistics.

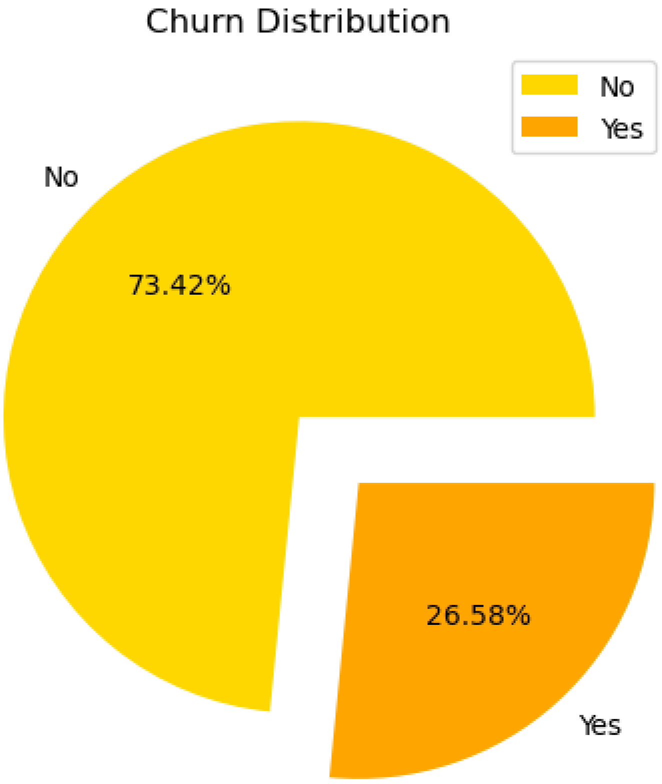
Column 21: Churn

Target Variable. Tells whether the customer churned or not (Yes or No)

```
In [47]: df['Churn'].unique()
```

```
Out[47]: array(['No', 'Yes'], dtype=object)
```

```
In [48]: plt.pie(df['Churn'].value_counts(),labels=df['Churn'].value_counts().keys(),au
         topct='%0.2f%%',colors = ['#FFD700', '#FFA500'] ,explode=[0.0,0.3])
         plt.title('Churn Distribution')
         plt.legend()
         plt.show()
```



Descriptive Analysis:

The bulk of the customers, or about 73.46%, did not churn, indicating a high customer retention rate, according to our Descriptive Analysis. However, a portion of the customers did choose to churn, accounting for 26.54% of the total customer turnover.

Bivariate Analysis

Performing data visualisation on pairs of variables, to observe relationship of each column with the Churn column.

In [53]:

```
""" Reading data from the file containing preprocessed data, for correlation a
nd scatter plot analyses"""

df2= pd.read_csv('Churn_Data_Cleaned.csv')
df2.drop(columns=['Unnamed: 0'],axis=1,inplace=True)
df2
```

Out[53]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	Internet
0	0	0	1	0	0.013889	0	0	
1	1	0	0	0	0.472222	1	1	
2	1	0	0	0	0.027778	1	1	
3	1	0	0	0	0.625000	0	0	
4	0	0	0	0	0.027778	1	1	
...	...	...	...	...	...	...	...	
7027	1	0	1	1	0.333333	1	2	
7028	0	0	1	1	1.000000	1	2	
7029	0	0	1	1	0.152778	0	0	
7030	1	1	1	0	0.055556	1	2	
7031	1	0	0	0	0.916667	1	1	

7032 rows × 25 columns

In [158]:

```
"""
Copying all the label encoded columns from df2 into df3, and applying label en
coding on Payment Method and Contract columns,
as they had been previously one-hot encoded. Also, dropping any null rows.
"""

df3 = pd.DataFrame()

df3['gender'] = df2['gender']
df3['SeniorCitizen'] = df2['SeniorCitizen']
df3['Partner'] = df2['Partner']
df3['Dependents'] = df2['Dependents']
df3['tenure'] = df2['tenure']
df3['PhoneService'] = df2['PhoneService']
df3['MultipleLines'] = df2['MultipleLines']
df3['InternetService'] = df2['InternetService']
df3['OnlineSecurity'] = df2['OnlineSecurity']
df3['OnlineBackup'] = df2['OnlineBackup']
df3['DeviceProtection'] = df2['DeviceProtection']
df3['TechSupport'] = df2['TechSupport']
df3['StreamingTV'] = df2['StreamingTV']
df3['StreamingMovies'] = df2['StreamingMovies']
df3['PaperlessBilling'] = df2['PaperlessBilling']

df3['PaymentMethod'] = df.PaymentMethod.astype('category').cat.codes
df3['Contract'] = df.Contract.astype('category').cat.codes

df3['MonthlyCharges'] = df2['MonthlyCharges']
df3['TotalCharges'] = df2['TotalCharges']
df3['Churn'] = df2['Churn']

df3.dropna(inplace=True)
df3.head()
```

Out[158]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetSer
0	0	0	1	0	0.013889	0	0	
1	1	0	0	0	0.472222	1	1	
2	1	0	0	0	0.027778	1	1	
3	1	0	0	0	0.625000	0	0	
4	0	0	0	0	0.027778	1	1	

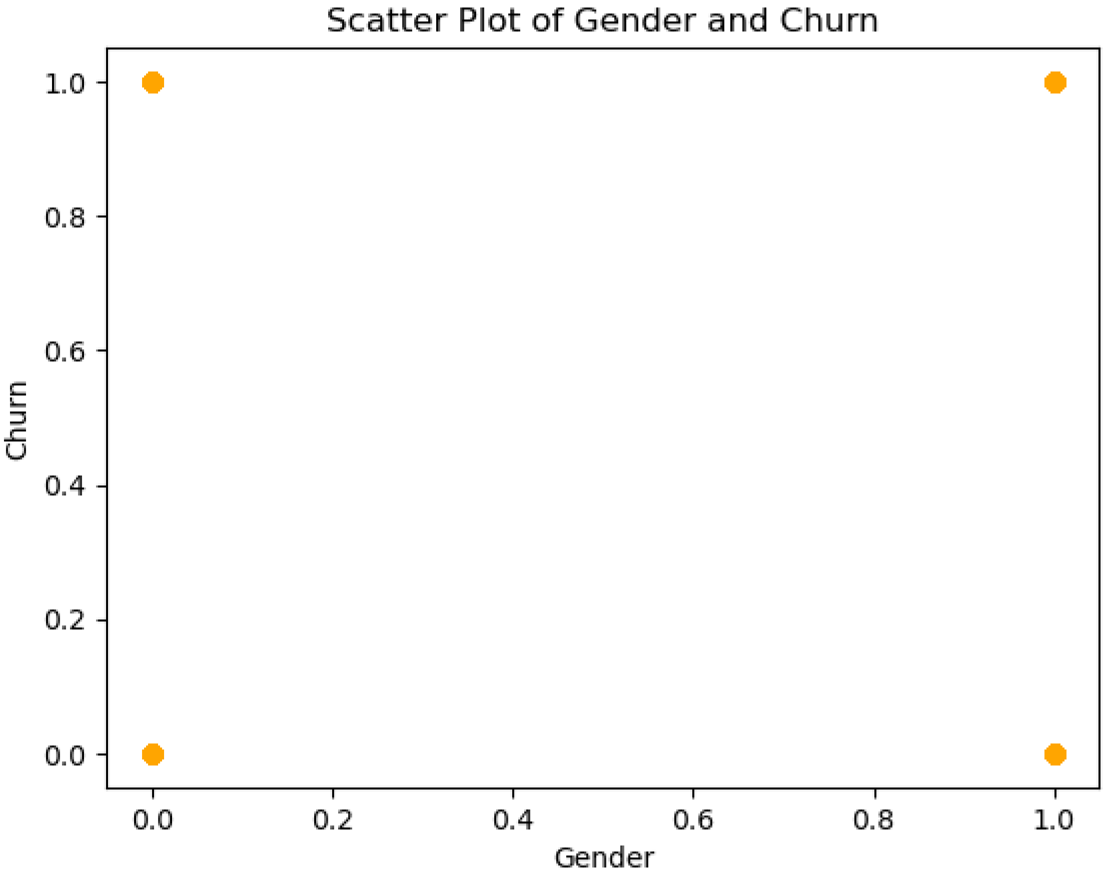
1. Gender and Churn

In [55]:
```
df_subset = df3[['gender', 'Churn']]
df_subset.corr()
```

Out[55]:

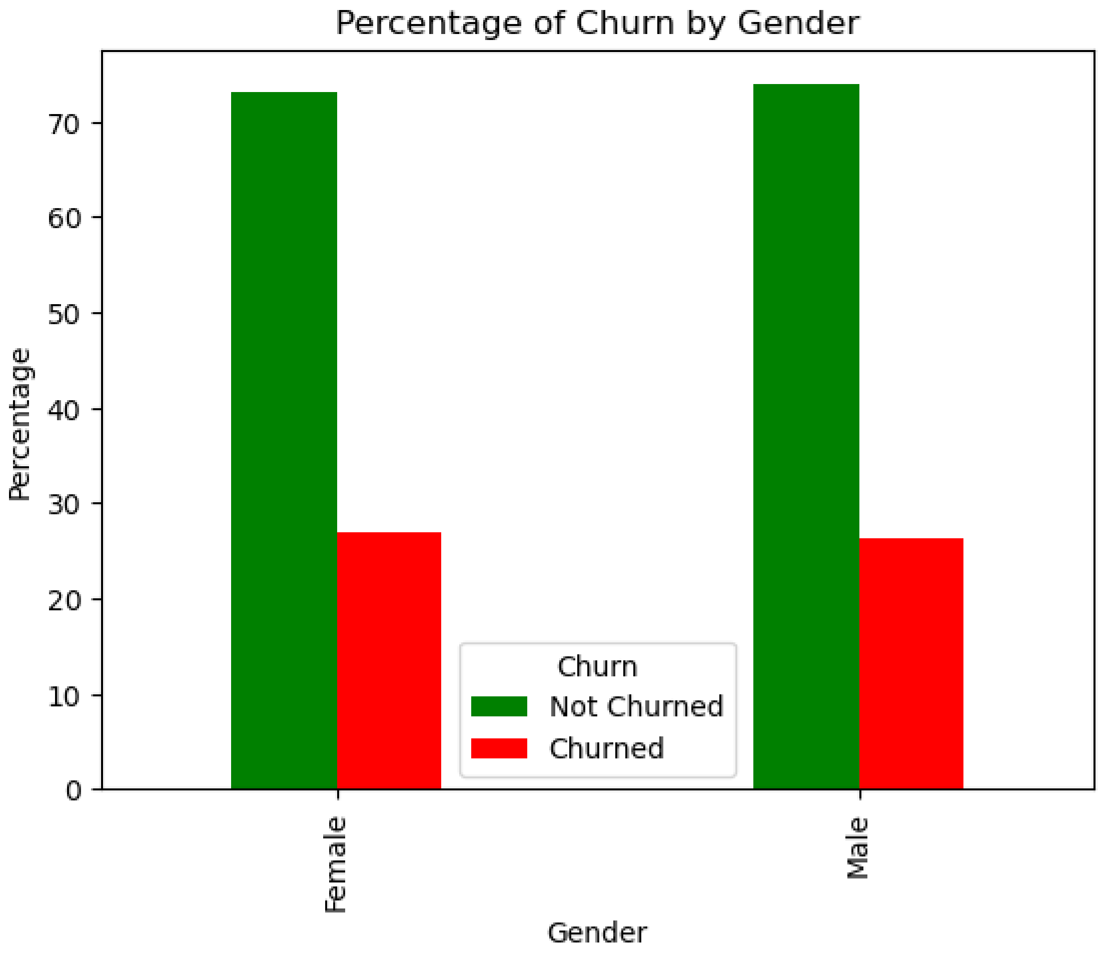
	gender	Churn
gender	1.000000	-0.008813
Churn	-0.008813	1.000000

In [56]:
```
plt.scatter(x =df3['gender'], y=df3['Churn'], c='orange')
plt.xlabel("Gender")
plt.ylabel("Churn")
plt.title("Scatter Plot of Gender and Churn")
plt.show()
```



In [59]:

```
grouped_data = pd.crosstab(df['gender'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('Gender')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Gender')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak negative correlation, or no correlation, between the Gender and Churn columns. From the bar graph, we can observe that around 70% of the female customers did not churn, while the remaining churned. Similarly, around 70% of the male customers also churned, while the remaining did not churn. As reflected in the univariate analysis of the Churn column already, almost 70% of the customers churned, and the statistic is repeated for each gender value as well.

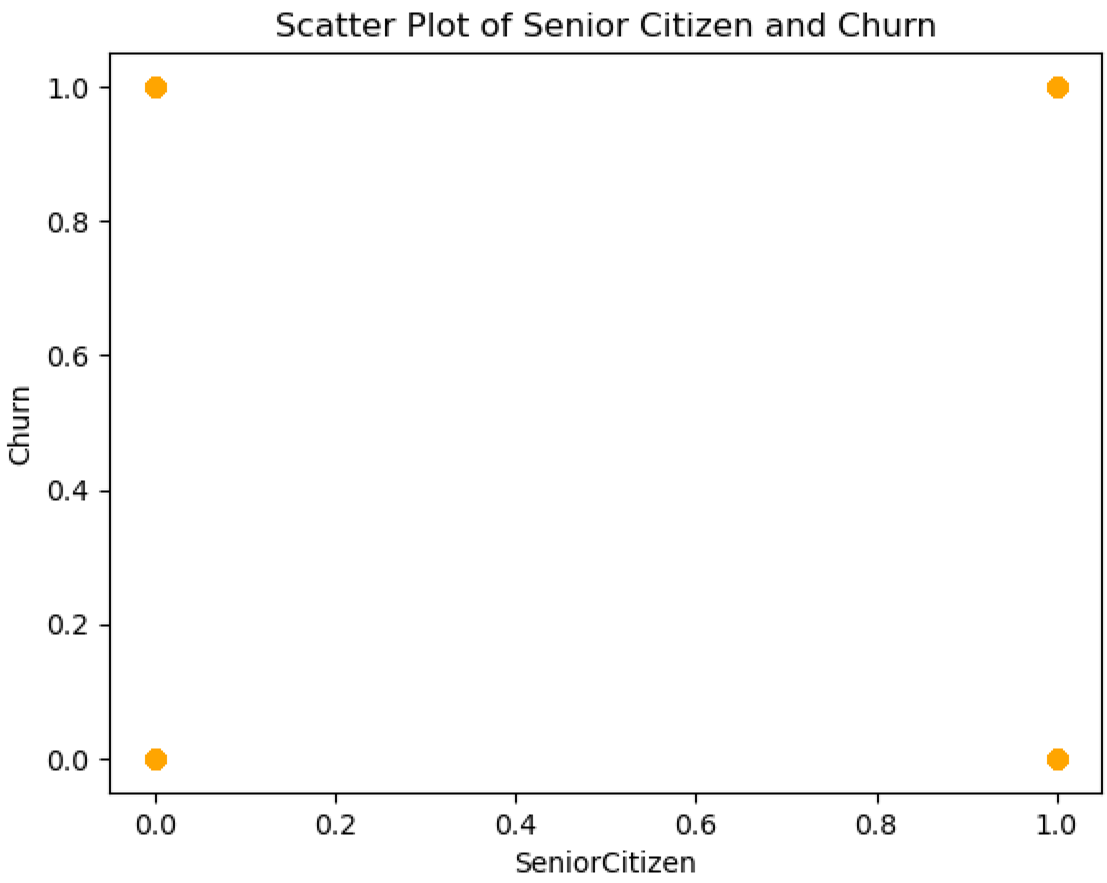
2. SeniorCitizen and Churn

In [66]:
```
df_subset = df3[['SeniorCitizen', 'Churn']]
df_subset.corr()
```

Out[66]:

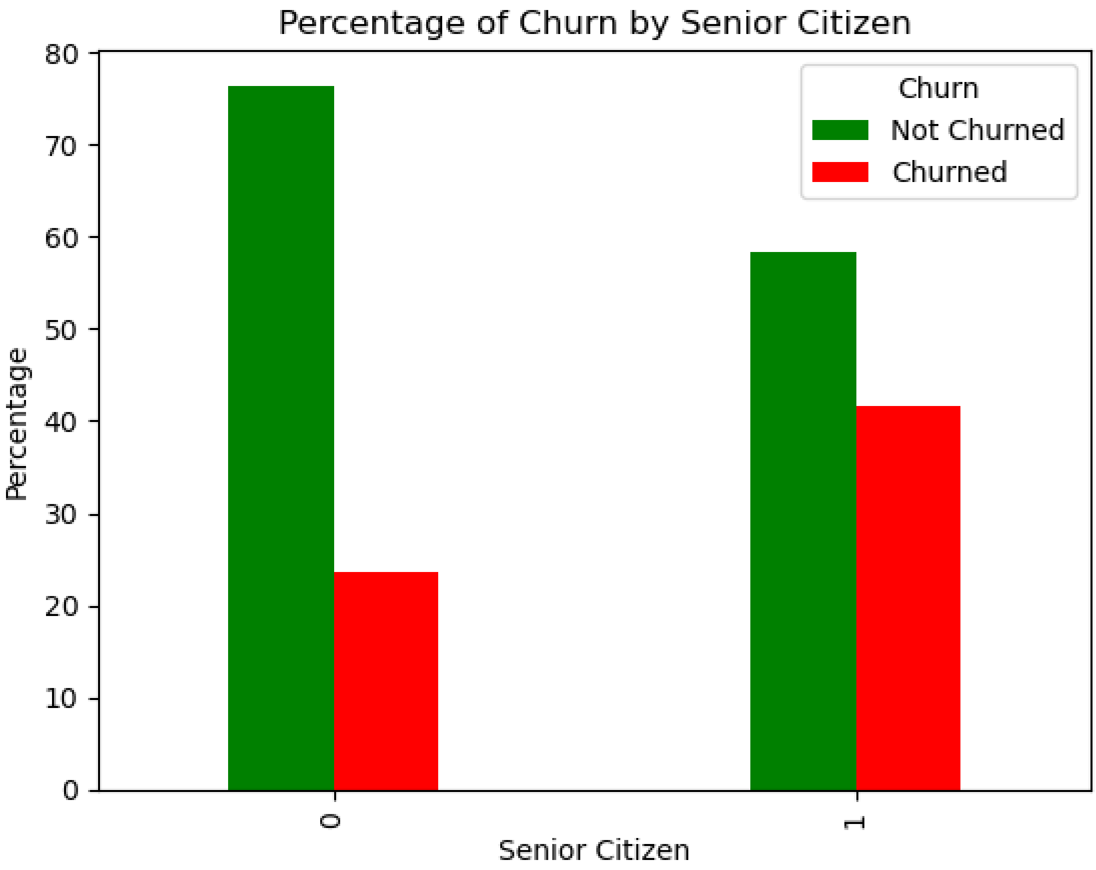
	gender	Churn
gender	1.000000	-0.008813
Churn	-0.008813	1.000000

In [67]:
```
plt.scatter(x =df3['SeniorCitizen'], y=df3['Churn'], c='orange')
plt.xlabel("SeniorCitizen")
plt.ylabel("Churn")
plt.title("Scatter Plot of Senior Citizen and Churn")
plt.show()
```



In [69]:

```
grouped_data = pd.crosstab(df['SeniorCitizen'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('Senior Citizen')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Senior Citizen')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak negative correlation, or no correlation, between the SeniorCitizen and Churn columns. From the bar graph, we can observe that around 75% of the non-senior citizens did not churn, while the remaining 25% did. On the other hand, almost 55% of the senior citizens did not churn, while nearly 45% did.

3. Partner and Churn

In [70]:

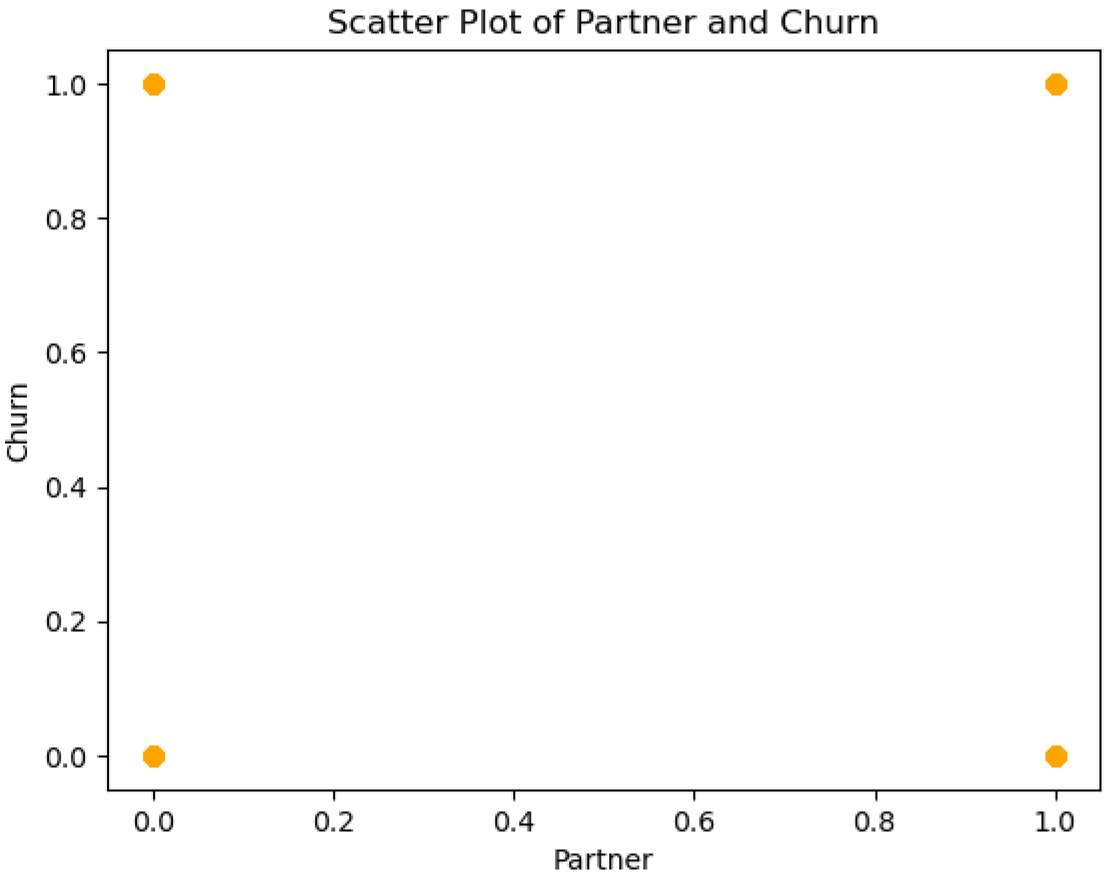
```
df_subset = df3[['Partner', 'Churn']]
df_subset.corr()
```

Out[70]:

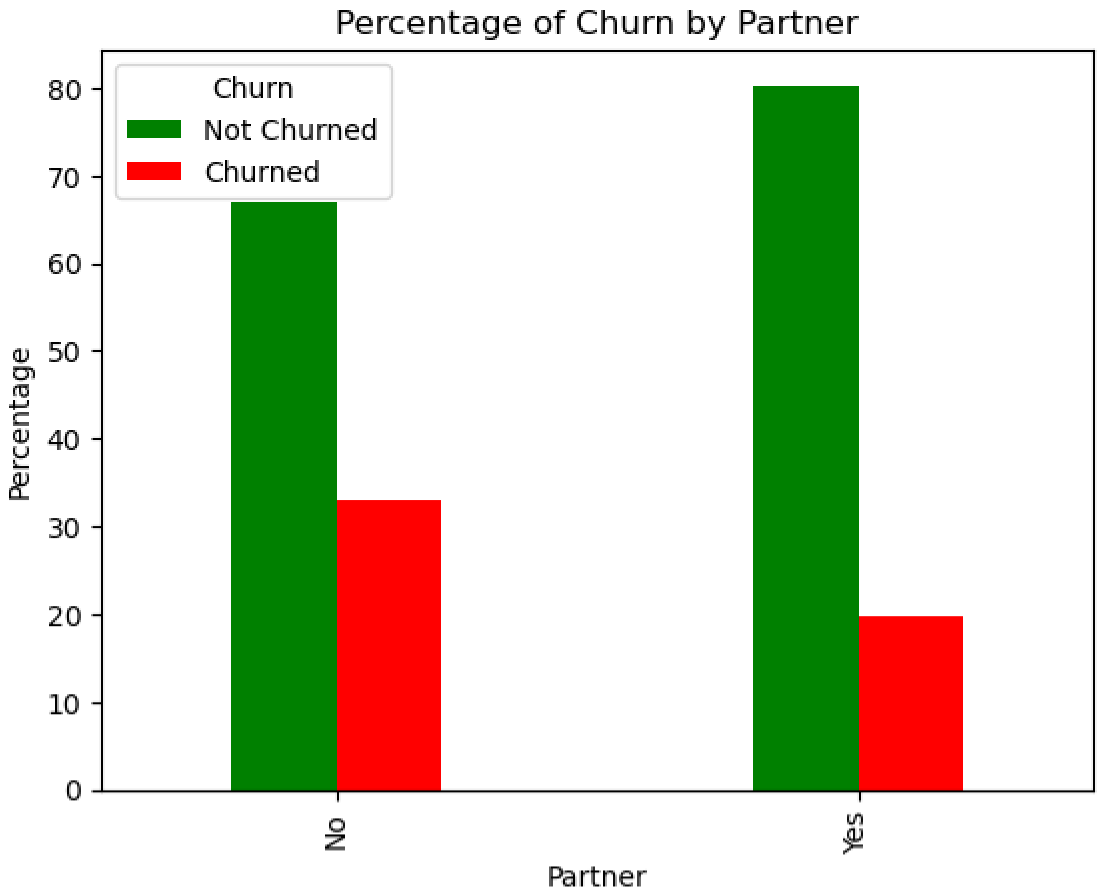
	Partner	Churn
Partner	1.000000	-0.149561
Churn	-0.149561	1.000000

In [71]:

```
plt.scatter(x =df3['Partner'], y=df3['Churn'], c='orange')
plt.xlabel("Partner")
plt.ylabel("Churn")
plt.title("Scatter Plot of Partner and Churn")
plt.show()
```



In [72]:
```
grouped_data = pd.crosstab(df['Partner'], df['Churn'], normalize='index') * 10
0
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('Partner')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Partner')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak negative correlation between the Partner and Churn columns. From the bar graph, we can observe that around 65% of the customers without partners did not churn, while the remaining 35% did. On the other hand, almost 80% of the customers with partners did not churn, while nearly 20% did.

4. Dependents and Churn

In [73]:

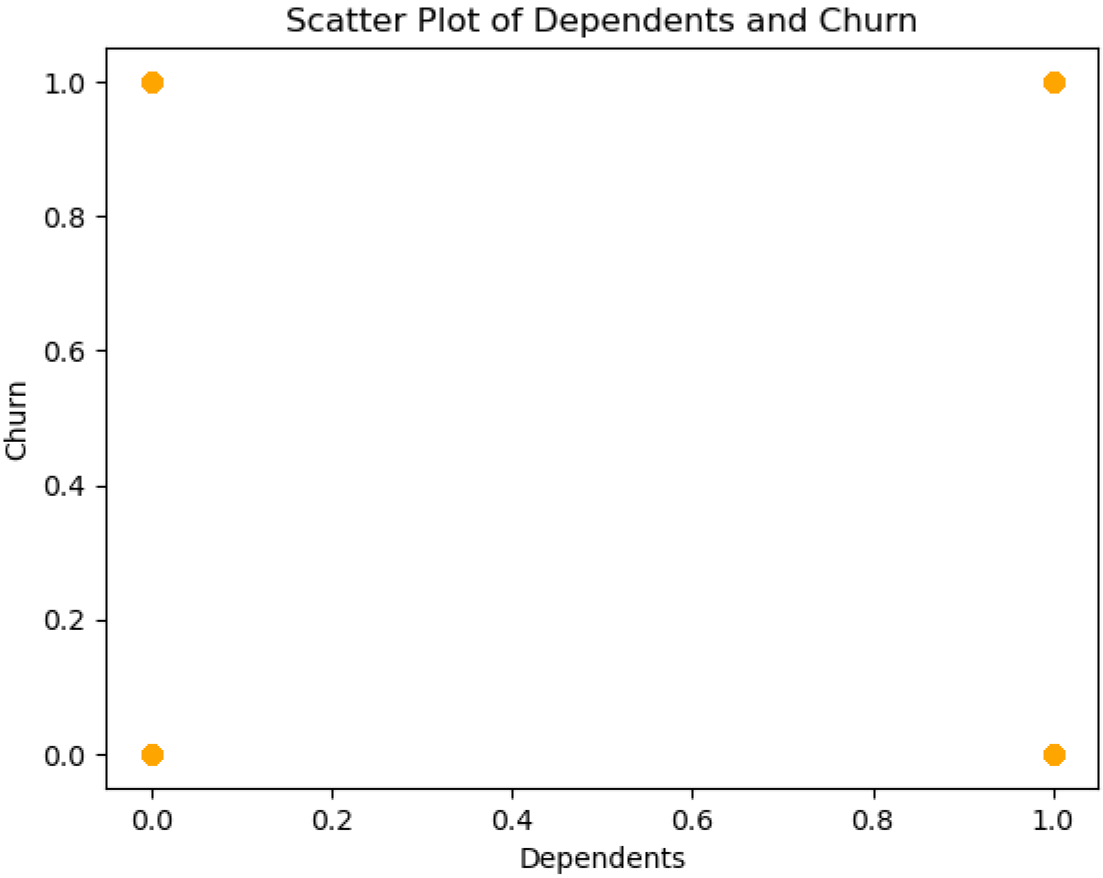
```
df_subset = df3[['Dependents', 'Churn']]
df_subset.corr()
```

Out[73]:

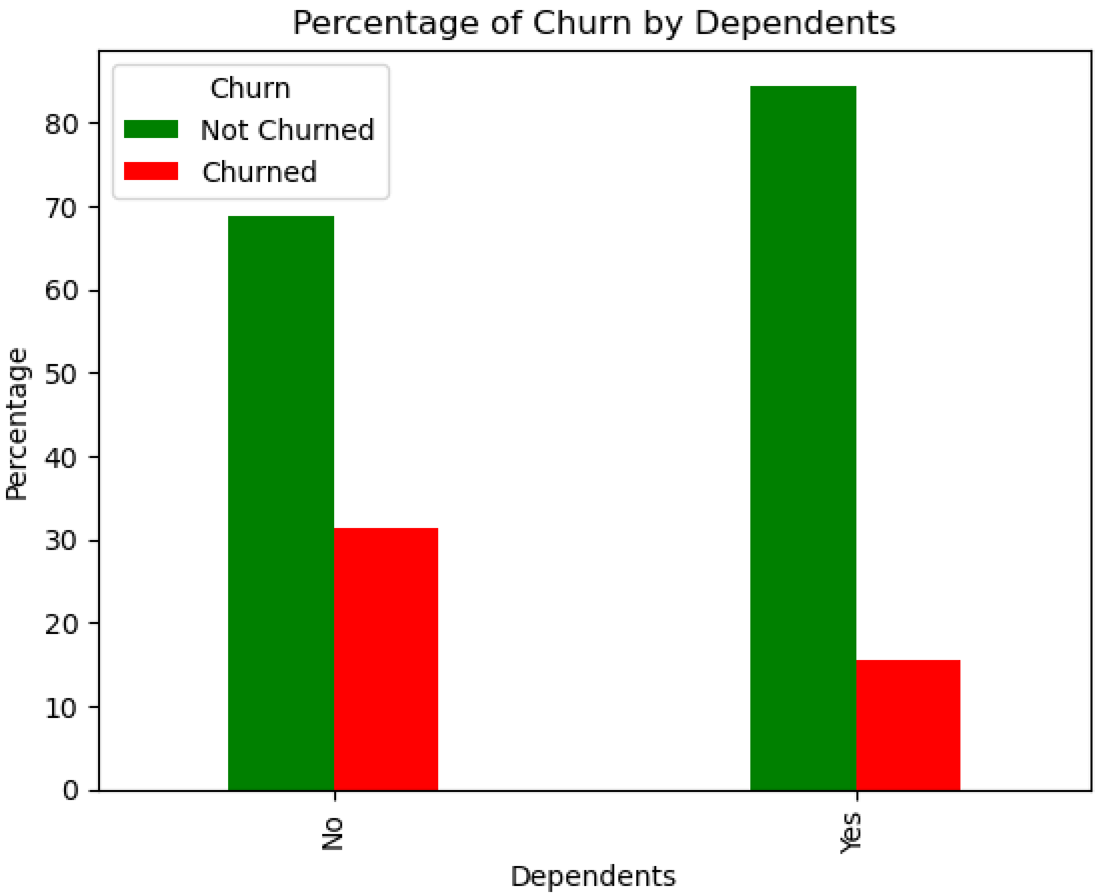
	Dependents	Churn
Dependents	1.000000	-0.164029
Churn	-0.164029	1.000000

In [74]:

```
plt.scatter(x =df3['Dependents'], y=df3['Churn'], c='orange')
plt.xlabel("Dependents")
plt.ylabel("Churn")
plt.title("Scatter Plot of Dependents and Churn")
plt.show()
```



```
In [75]: grouped_data = pd.crosstab(df['Dependents'], df['Churn'], normalize='index') * 100
         colors = ['green', 'red']
         chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

         plt.xlabel('Dependents')
         plt.ylabel('Percentage')
         plt.title('Percentage of Churn by Dependents')
         plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
         plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak negative correlation between the Dependents and Churn columns. From the bar graph, we can observe that around 68% of the customers without any dependents did not churn, while the remaining 32% did. On the other hand, almost 85% of the customers with dependents did not churn, while the remaining 15% did.

5. Tenure and Churn

In [76]:

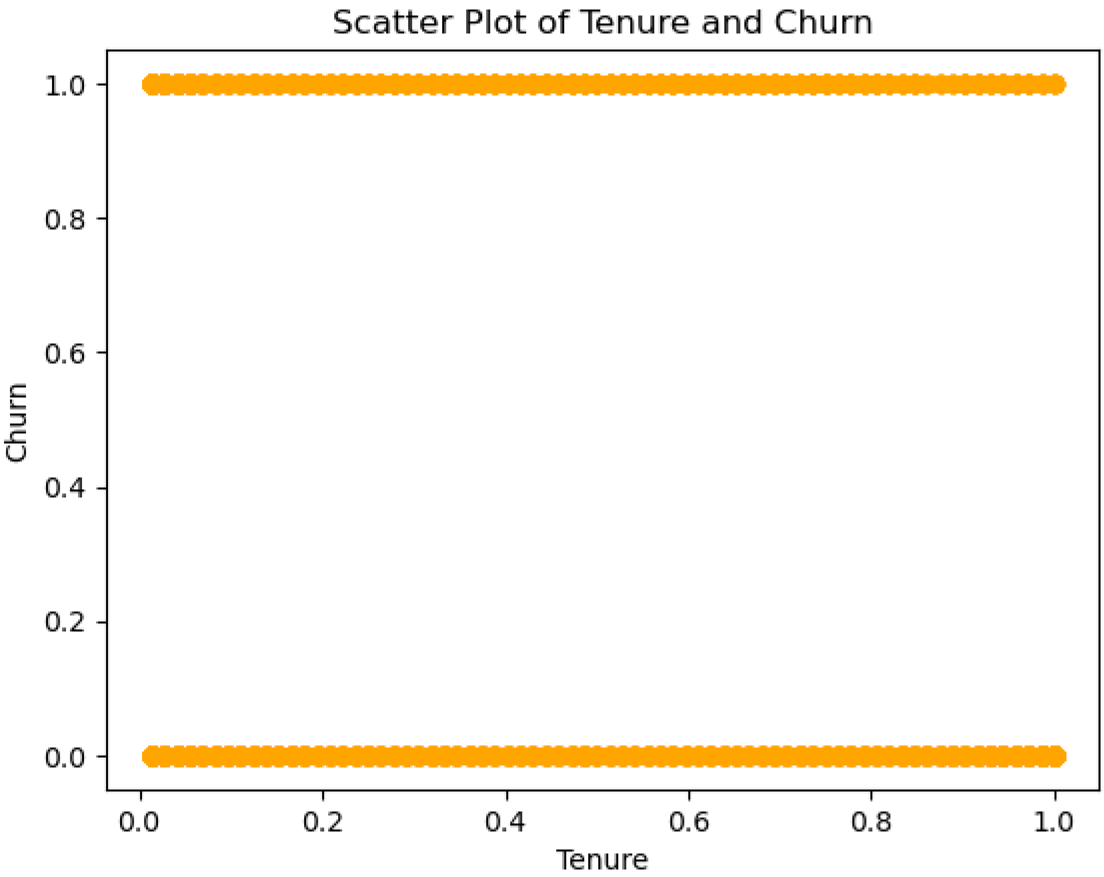
```
df_subset = df3[['tenure', 'Churn']]
df_subset.corr()
```

Out[76]:

	tenure	Churn
tenure	1.000000	-0.354315
Churn	-0.354315	1.000000

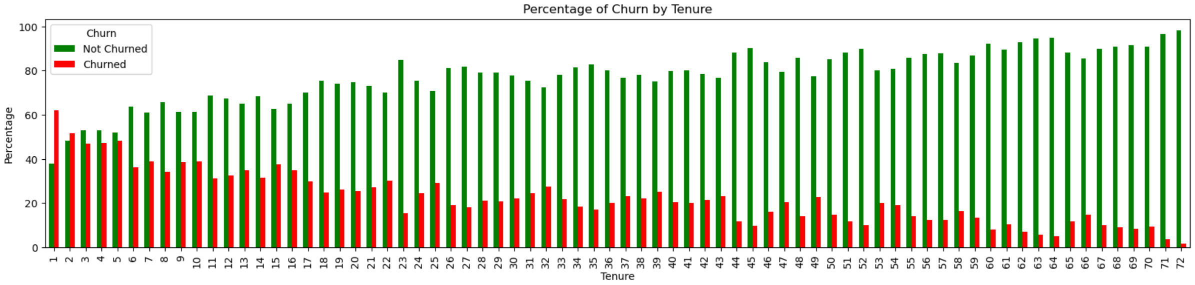
In [77]:

```
plt.scatter(x =df3['tenure'], y=df3['Churn'], c='orange')
plt.xlabel("Tenure")
plt.ylabel("Churn")
plt.title("Scatter Plot of Tenure and Churn")
plt.show()
```



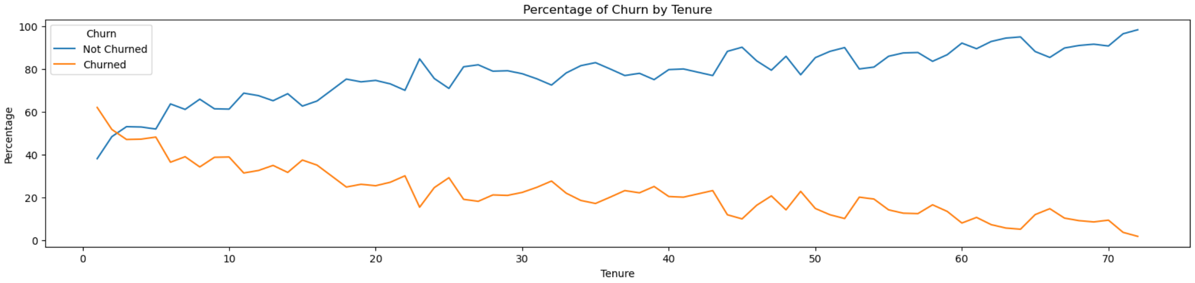
In [82]:

```
grouped_data = pd.crosstab(df['tenure'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.6, color=colors, figsize=(20,4))

plt.xlabel('Tenure')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Tenure')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



In [153]:

```
grouped_data = pd.crosstab(df['tenure'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='line', figsize=(20,4))

plt.xlabel('Tenure')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Tenure')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a weak negative correlation between the Tenure and Churn columns. From the bar graph, and the line graph, we can observe a general trend of churn percentage decreasing with the increasing tenure, with the highest value of churn percentage being 60% at tenure value 1, and the lowest value of churn percentage being less than 5% at 72, the highest value of tenure.

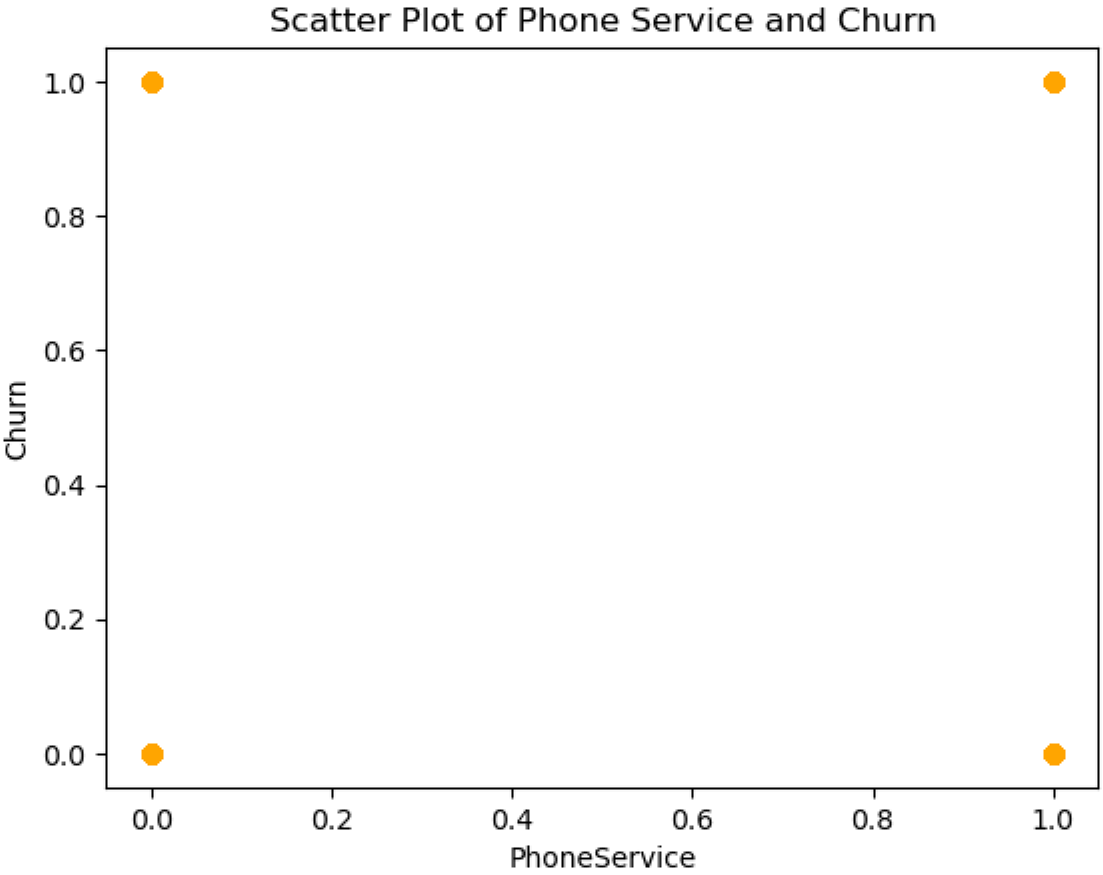
6. PhoneService and Churn

In [83]:
```
df_subset = df3[['PhoneService', 'Churn']]
df_subset.corr()
```

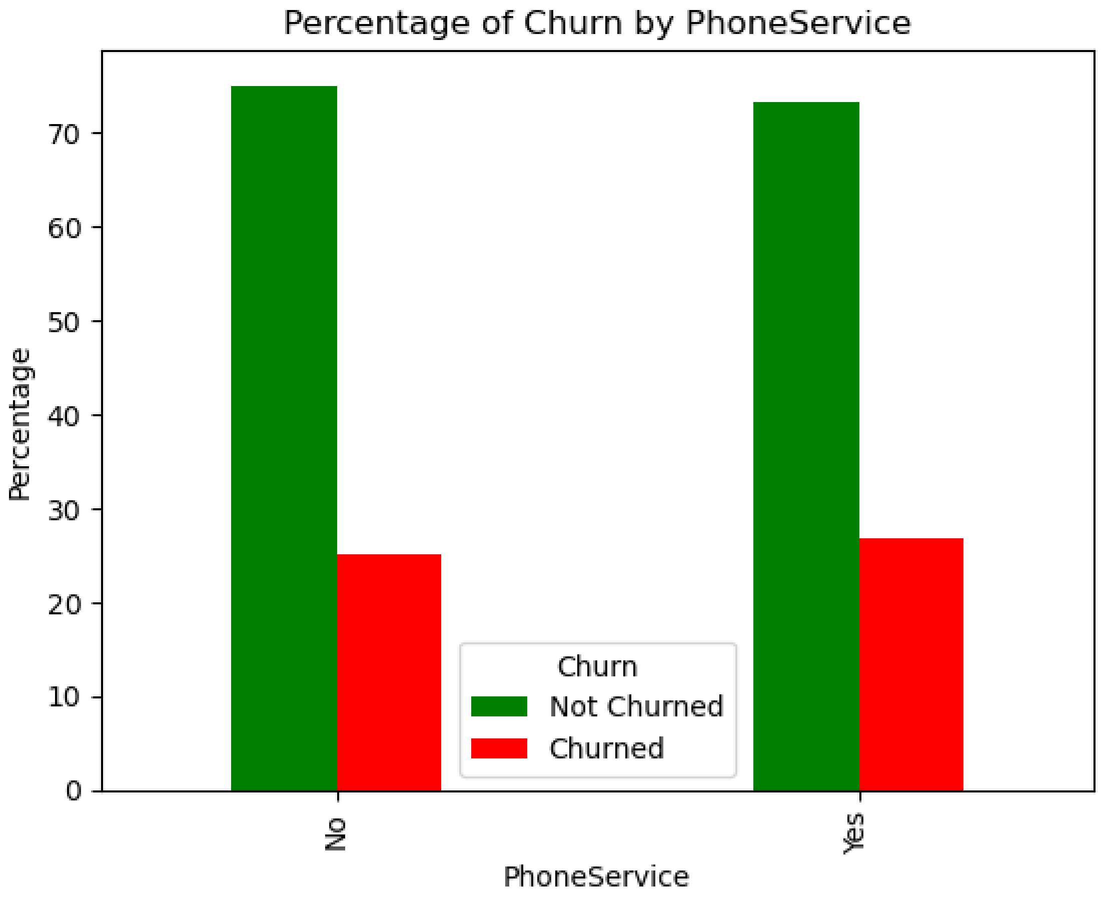
Out[83]:

	PhoneService	Churn
PhoneService	1.00000	0.01159
Churn	0.01159	1.00000

In [84]:
```
plt.scatter(x =df3['PhoneService'], y=df3['Churn'], c='orange')
plt.xlabel("PhoneService")
plt.ylabel("Churn")
plt.title("Scatter Plot of Phone Service and Churn")
plt.show()
```



```
In [85]: grouped_data = pd.crosstab(df['PhoneService'], df['Churn'], normalize='index') * 100
         colors = ['green', 'red']
         chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

         plt.xlabel('PhoneService')
         plt.ylabel('Percentage')
         plt.title('Percentage of Churn by PhoneService')
         plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
         plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak postive correlation, or no correlation, between the PhoneService and Churn columns. From the bar graph, we can observe that nearly 75% of the customers without a phone service did not churn, while the remaining 25% did. Similarly, nearly 72% of the customers with a phone service did not churn, while the remaining 28% did.

7. MultipleLines and Churn

In [86]:

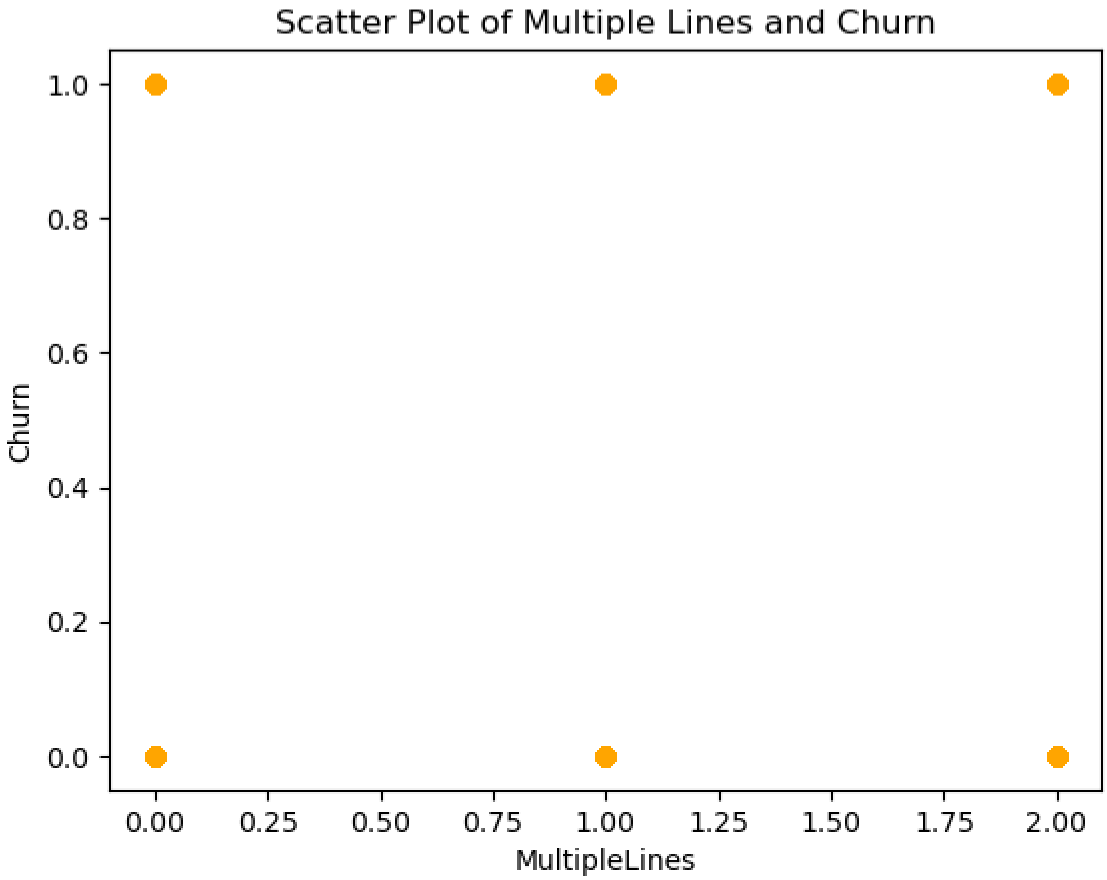
```
df_subset = df3[['MultipleLines', 'Churn']]
df_subset.corr()
```

Out[86]:

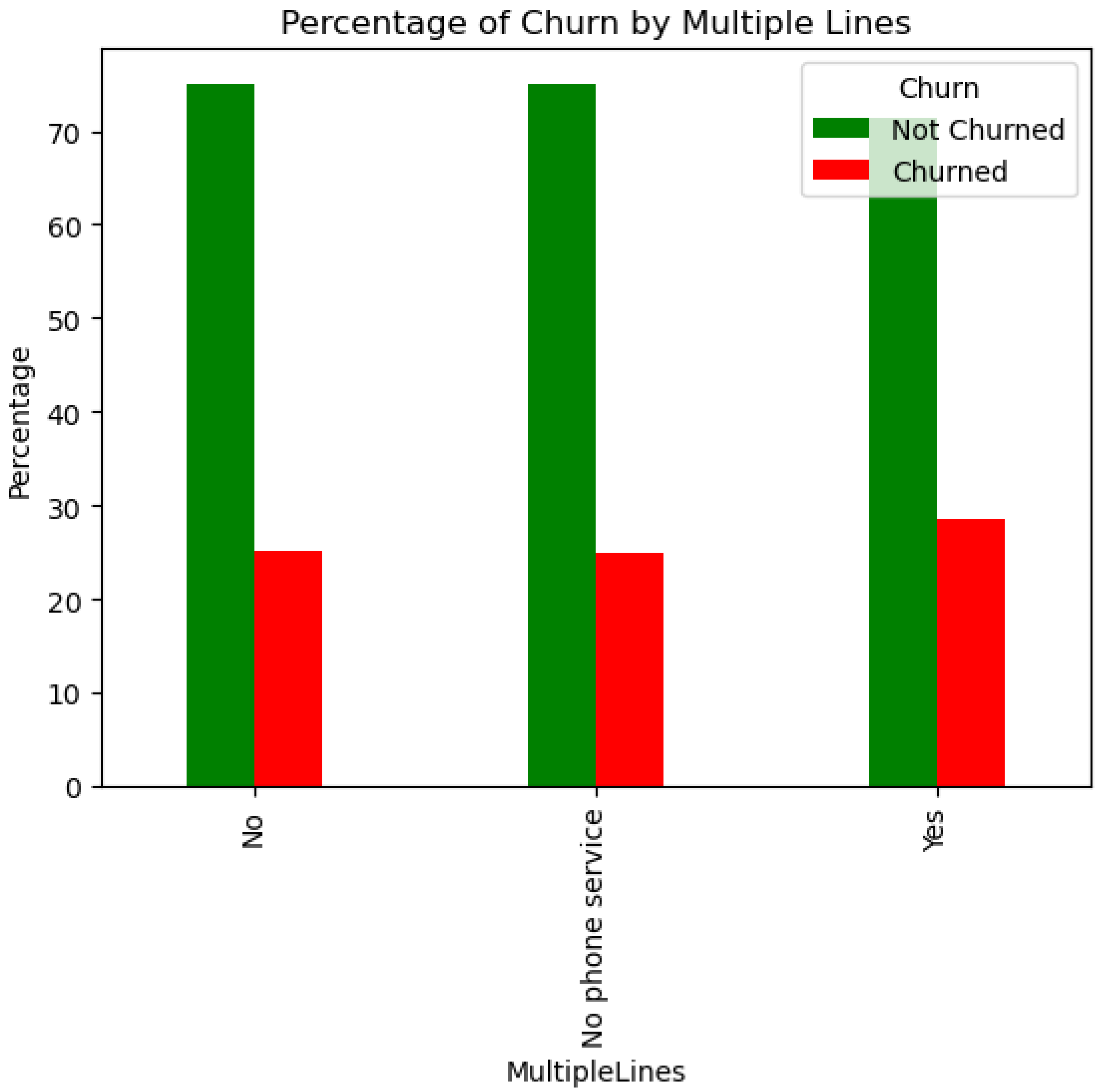
	MultipleLines	Churn
MultipleLines	1.000000	0.036311
Churn	0.036311	1.000000

In [87]:

```
plt.scatter(x =df3['MultipleLines'], y=df3['Churn'], c='orange')
plt.xlabel("MultipleLines")
plt.ylabel("Churn")
plt.title("Scatter Plot of Multiple Lines and Churn")
plt.show()
```



In [88]:
```
grouped_data = pd.crosstab(df['MultipleLines'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('MultipleLines')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Multiple Lines')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the PhoneService and Churn columns. From the bar graph, we can observe that nearly 75% of the customers without multiple lines, or without a phone service altogether, did not churn, while the remaining 25% did. Similarly, nearly 72% of the customers with multiple lines did not churn, while the remaining 28% did.

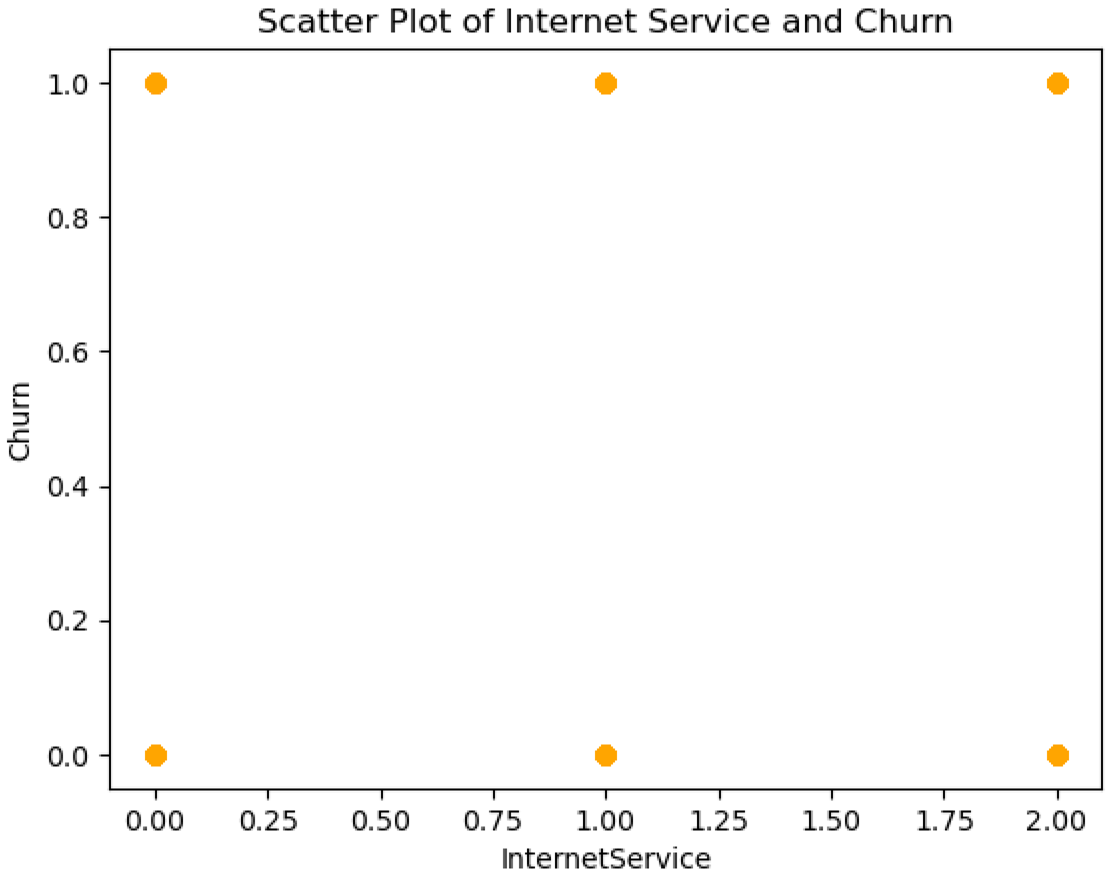
8. InternetService and Churn

In [89]:
```
df_subset = df3[['InternetService', 'Churn']]
df_subset.corr()
```

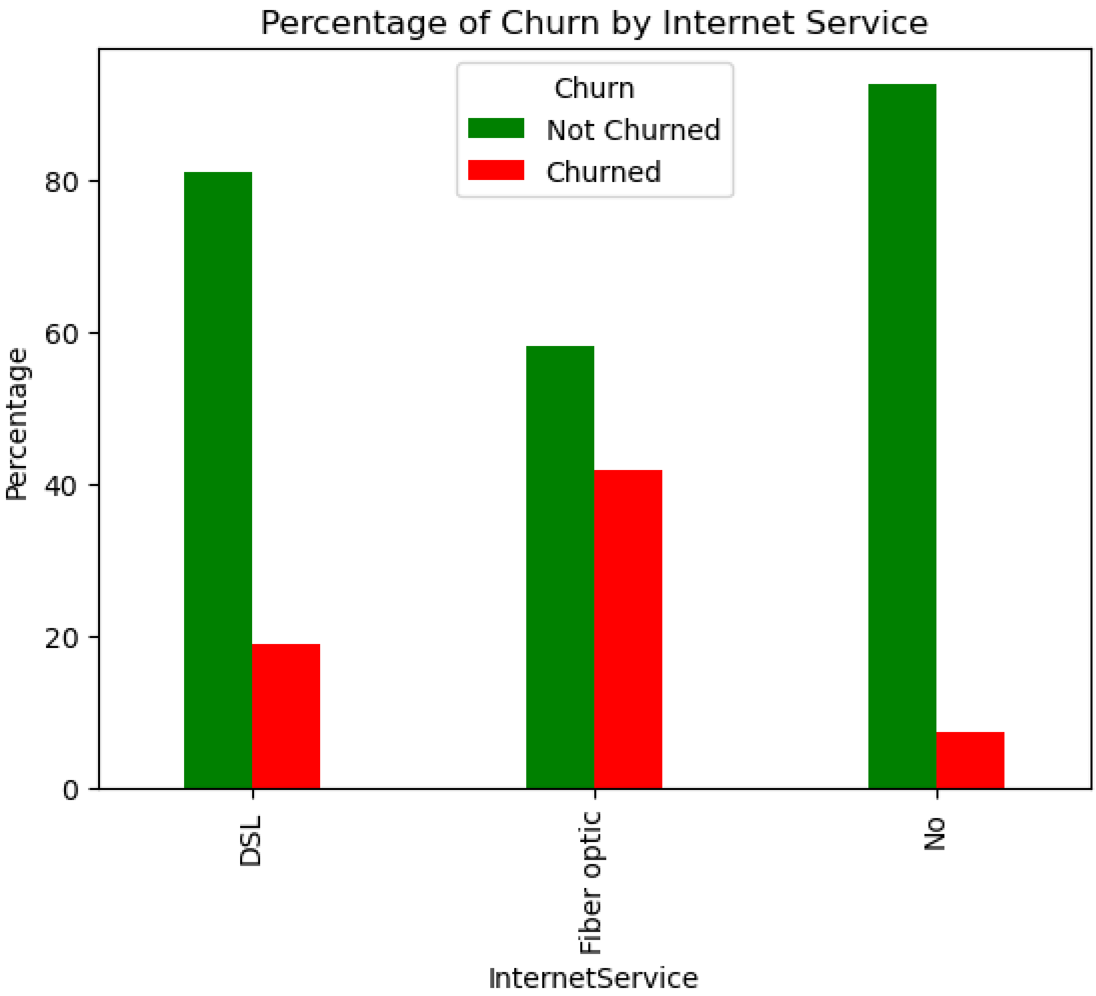
Out[89]:

	InternetService	Churn
InternetService	1.000000	0.315865
Churn	0.315865	1.000000

In [90]:
```
plt.scatter(x =df3['InternetService'], y=df3['Churn'], c='orange')
plt.xlabel("InternetService")
plt.ylabel("Churn")
plt.title("Scatter Plot of Internet Service and Churn")
plt.show()
```



In [91]:
```
grouped_data = pd.crosstab(df['InternetService'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('InternetService')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Internet Service')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a weak positive correlation between the InternetService and Churn columns. From the bar graph, we can observe that nearly 80% of the customers with DSL Internet Service did not churn, while the remaining 20% did. 60% of the customers with Fiber Optic service did not churn, while the remaining 40% did. Almost 90% of the customers without any Internet Service did not churn, while the remaining 10% did.

9. OnlineSecurity and Churn

In [93]:

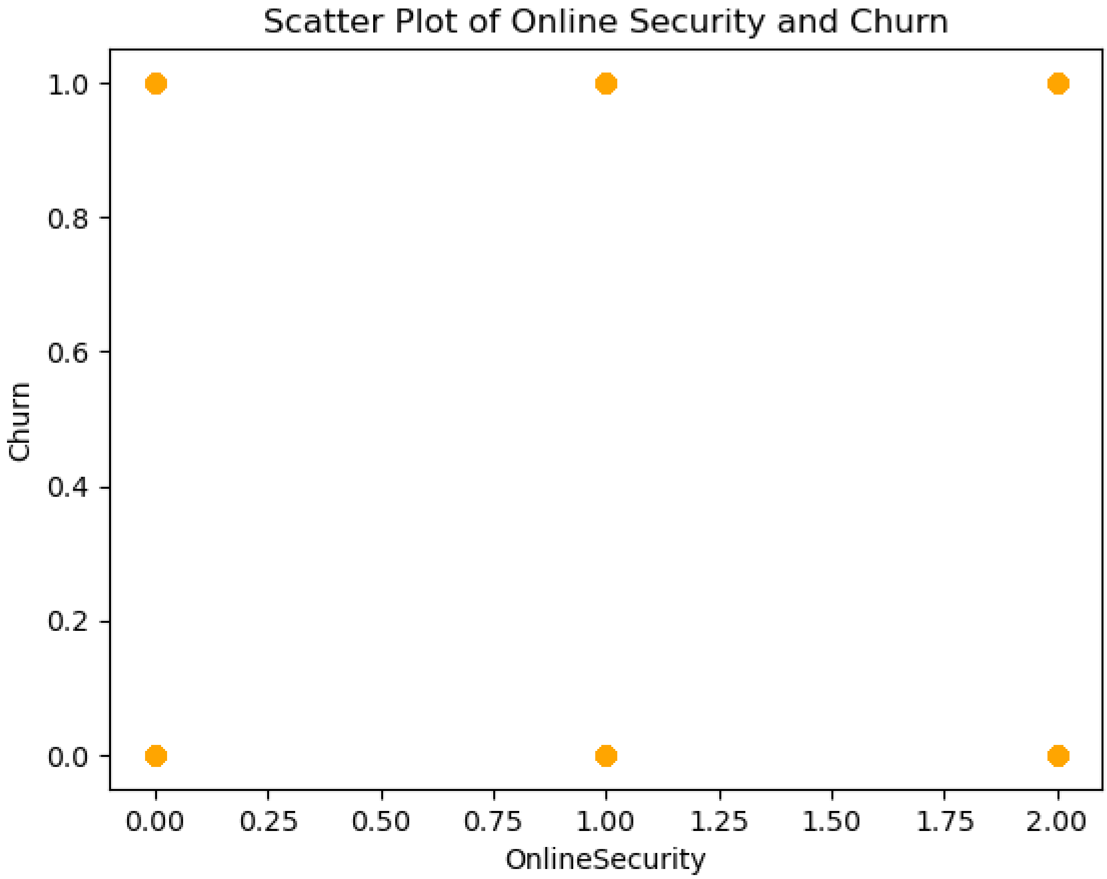
```
df_subset = df3[['OnlineSecurity', 'Churn']]
df_subset.corr()
```

Out[93]:

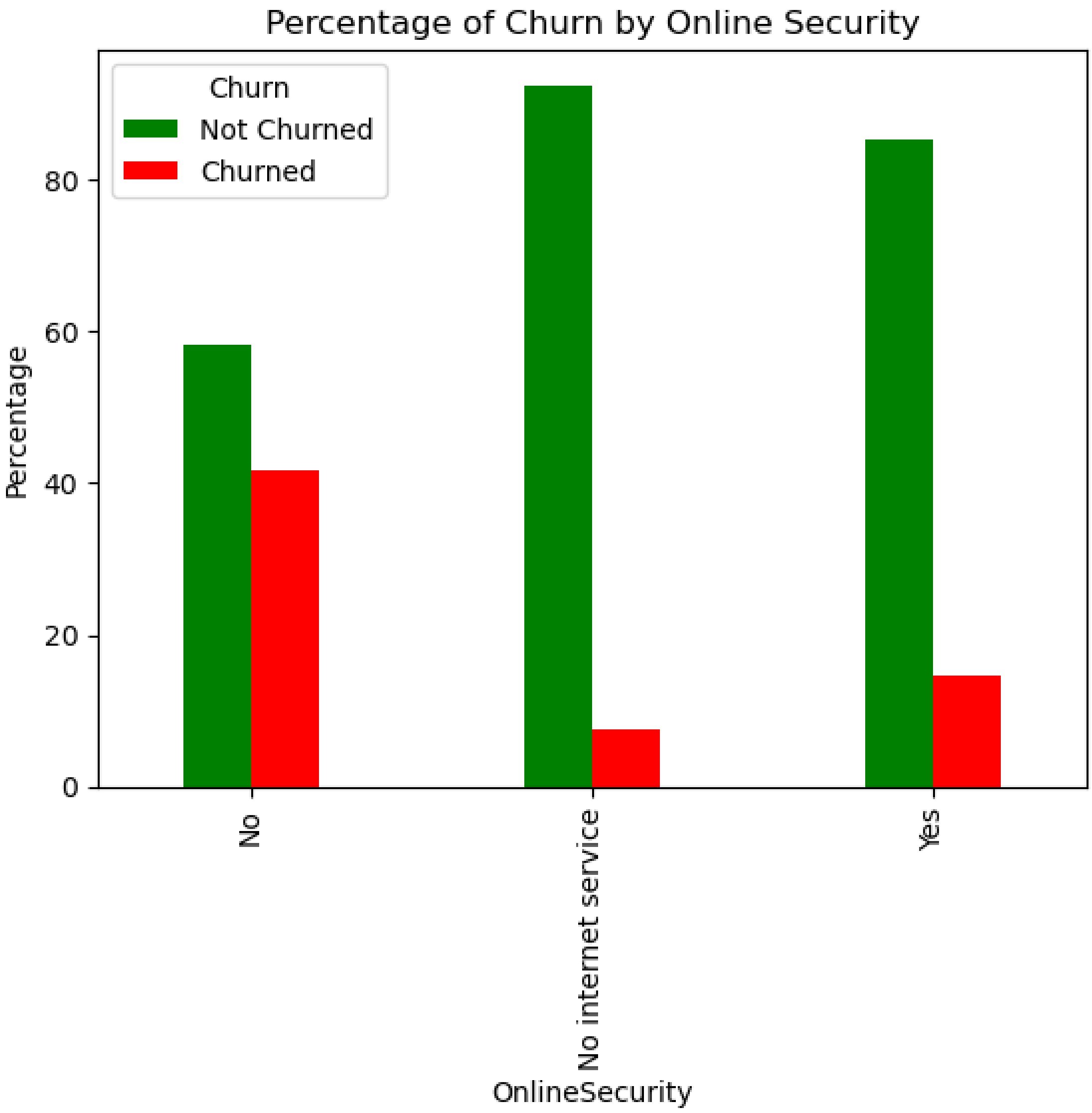
	OnlineSecurity	Churn
OnlineSecurity	1.000000	0.022718
Churn	0.022718	1.000000

In [92]:

```
plt.scatter(x =df3['OnlineSecurity'], y=df3['Churn'], c='orange')
plt.xlabel("OnlineSecurity")
plt.ylabel("Churn")
plt.title("Scatter Plot of Online Security and Churn")
plt.show()
```



In [94]:
```
grouped_data = pd.crosstab(df['OnlineSecurity'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('OnlineSecurity')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Online Security')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the OnlineSecurity and Churn columns. From the bar graph, we can observe that nearly 60% of the customers with No Online Security did not churn, while the remaining 40% did. Almost 90% of the customers without any Internet Service did not churn, while the remaining 10% did. Almost 85% of the customers with Online Security did not churn, while the remaining 15% did.

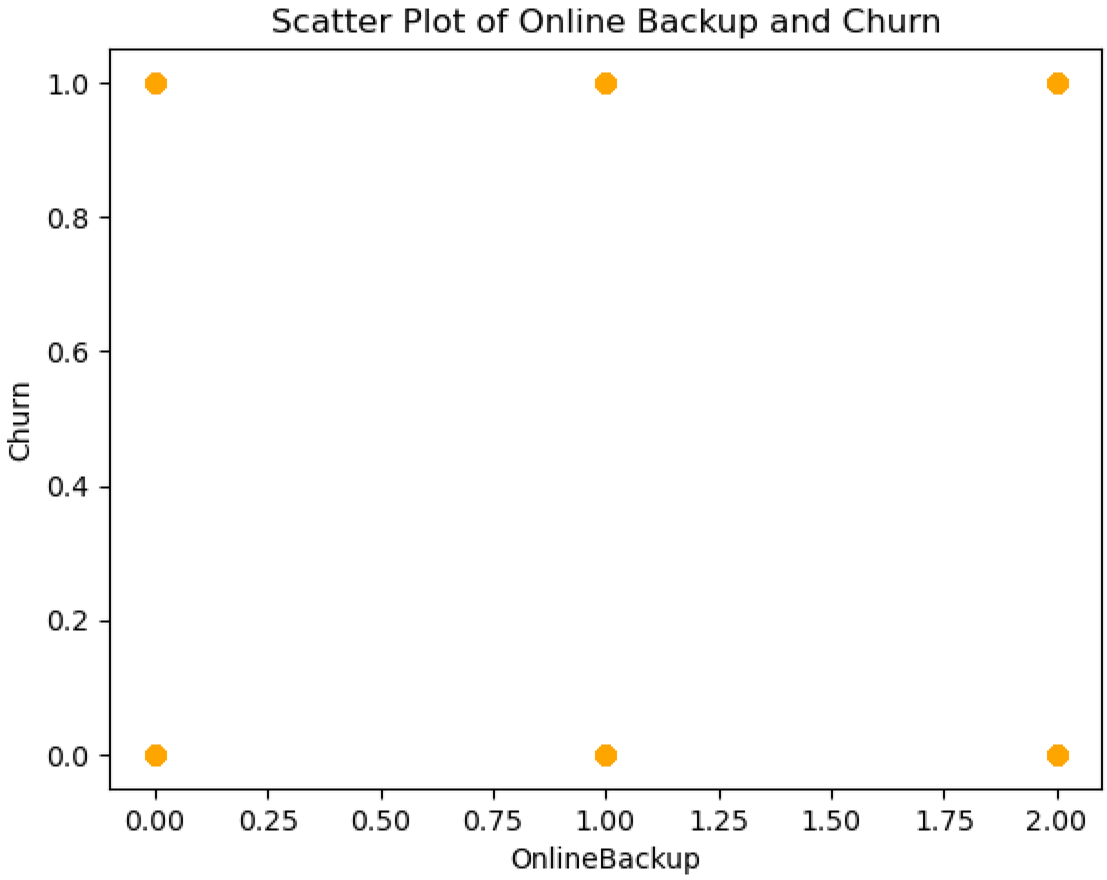
10. OnlineBackup and Churn

In [95]:
```
df_subset = df3[['OnlineBackup', 'Churn']]
df_subset.corr()
```

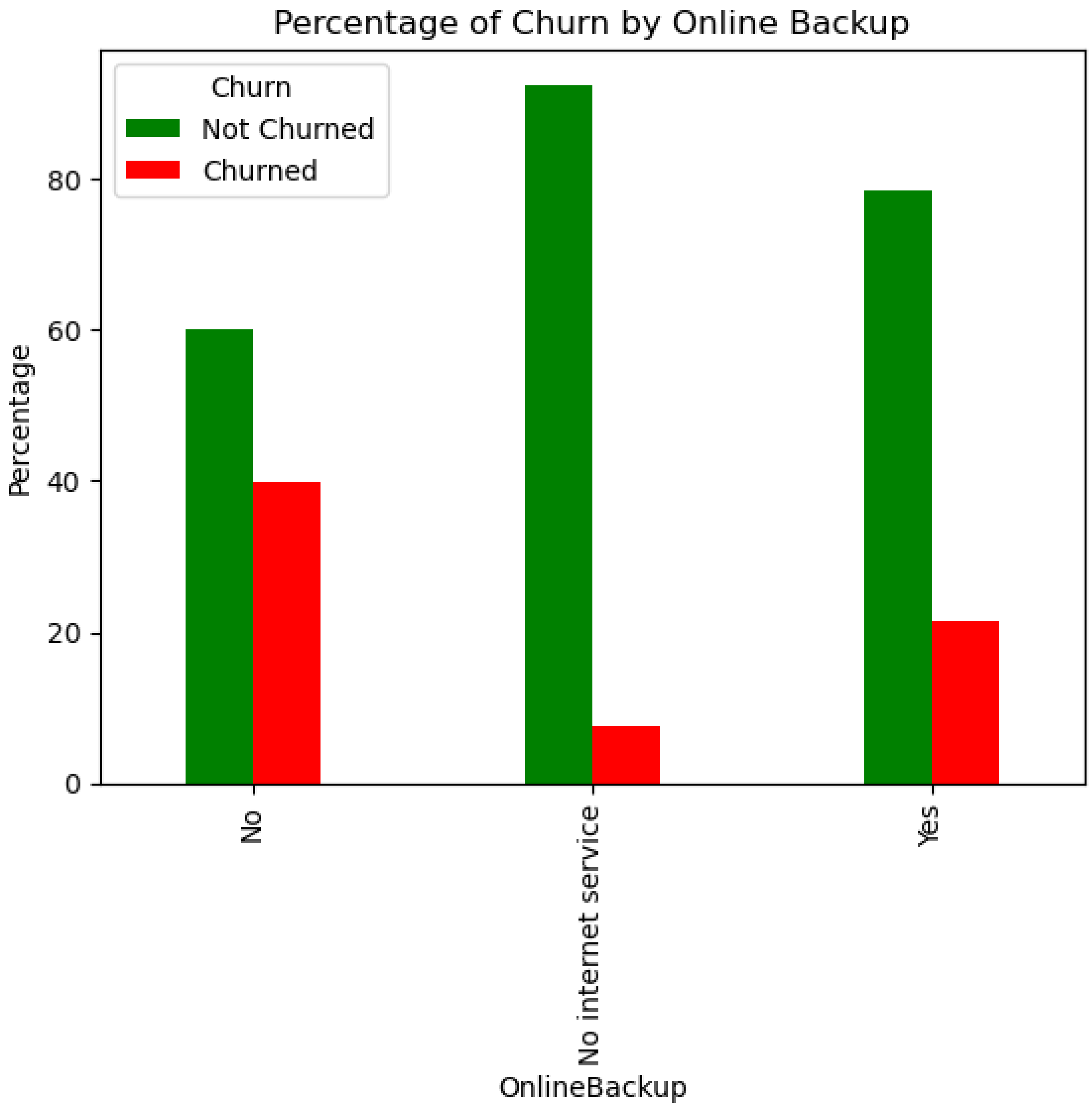
Out[95]:

	OnlineBackup	Churn
OnlineBackup	1.000000	0.073443
Churn	0.073443	1.000000

In [96]:
```
plt.scatter(x =df3['OnlineBackup'], y=df3['Churn'], c='orange')
plt.xlabel("OnlineBackup")
plt.ylabel("Churn")
plt.title("Scatter Plot of Online Backup and Churn")
plt.show()
```



In [97]:
```
grouped_data = pd.crosstab(df['OnlineBackup'], df['Churn'], normalize='index')
* 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('OnlineBackup')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Online Backup')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the OnlineBackup and Churn columns. From the bar graph, we can observe that nearly 60% of the customers with No Backup did not churn, while the remaining 40% did. Almost 90% of the customers without any Internet Service did not churn, while the remaining 10% did. Almost 80% of the customers with Online Backup did not churn, while the remaining 20% did.

11. DeviceProtection and Churn

In [98]:

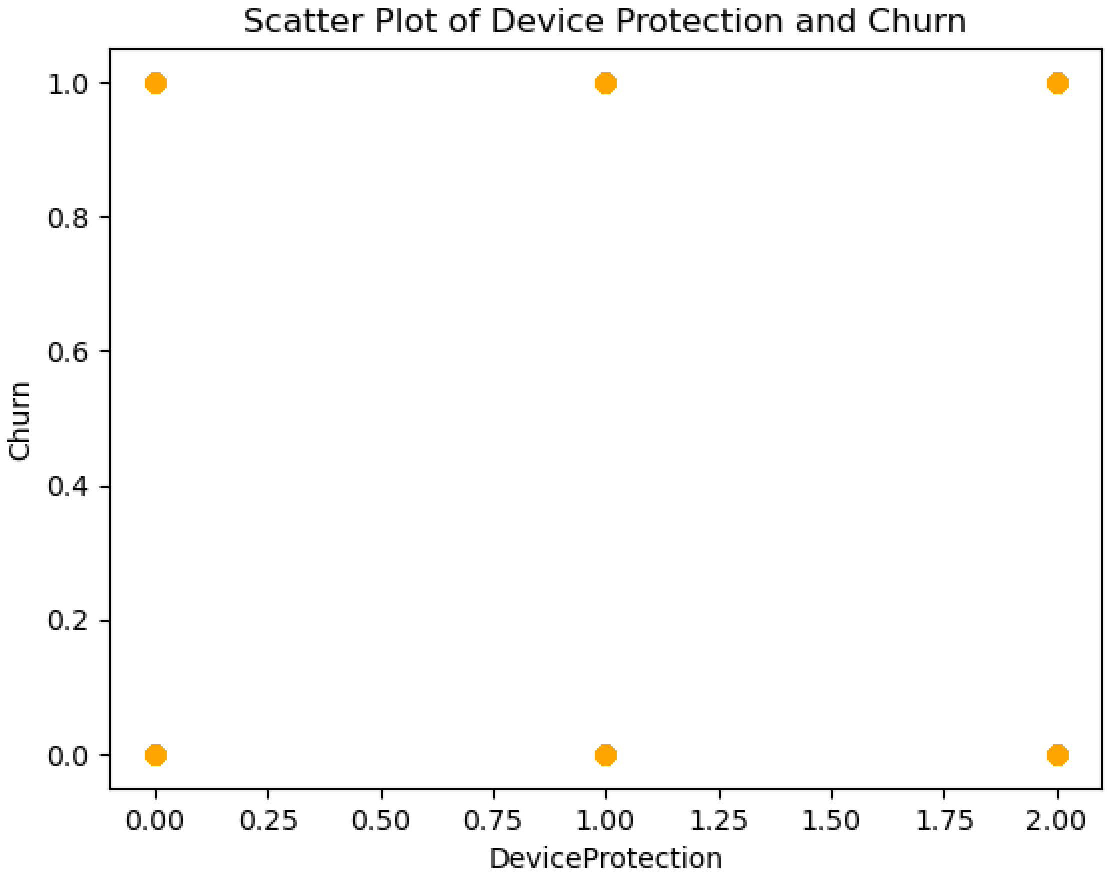
```
df_subset = df3[['DeviceProtection', 'Churn']]
df_subset.corr()
```

Out[98]:

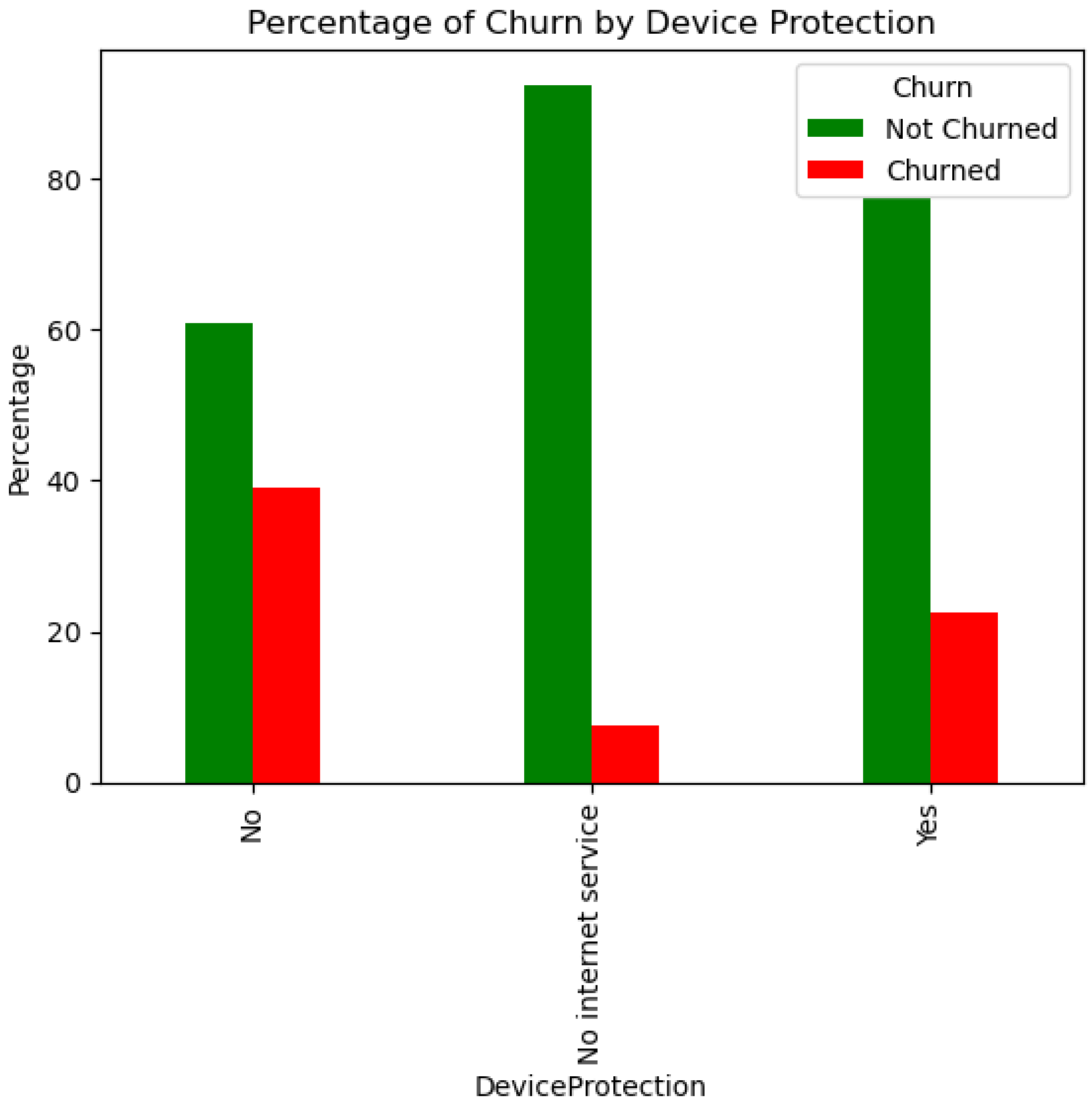
	DeviceProtection	Churn
DeviceProtection	1.000000	0.083862
Churn	0.083862	1.000000

In [99]:

```
plt.scatter(x =df3['DeviceProtection'], y=df3['Churn'], c='orange')
plt.xlabel("DeviceProtection")
plt.ylabel("Churn")
plt.title("Scatter Plot of Device Protection and Churn")
plt.show()
```



In [100]:
```
grouped_data = pd.crosstab(df['DeviceProtection'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('DeviceProtection')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Device Protection')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the DeviceProtection and Churn columns. From the bar graph, we can observe that nearly 60% of the customers with No Device Protection did not churn, while the remaining 40% did. Almost 95% of the customers without any Internet Service did not churn, while the remaining 5% did. Almost 80% of the customers with Device Protection did not churn, while the remaining 20% did.

12. TechSupport and Churn

In [101]:

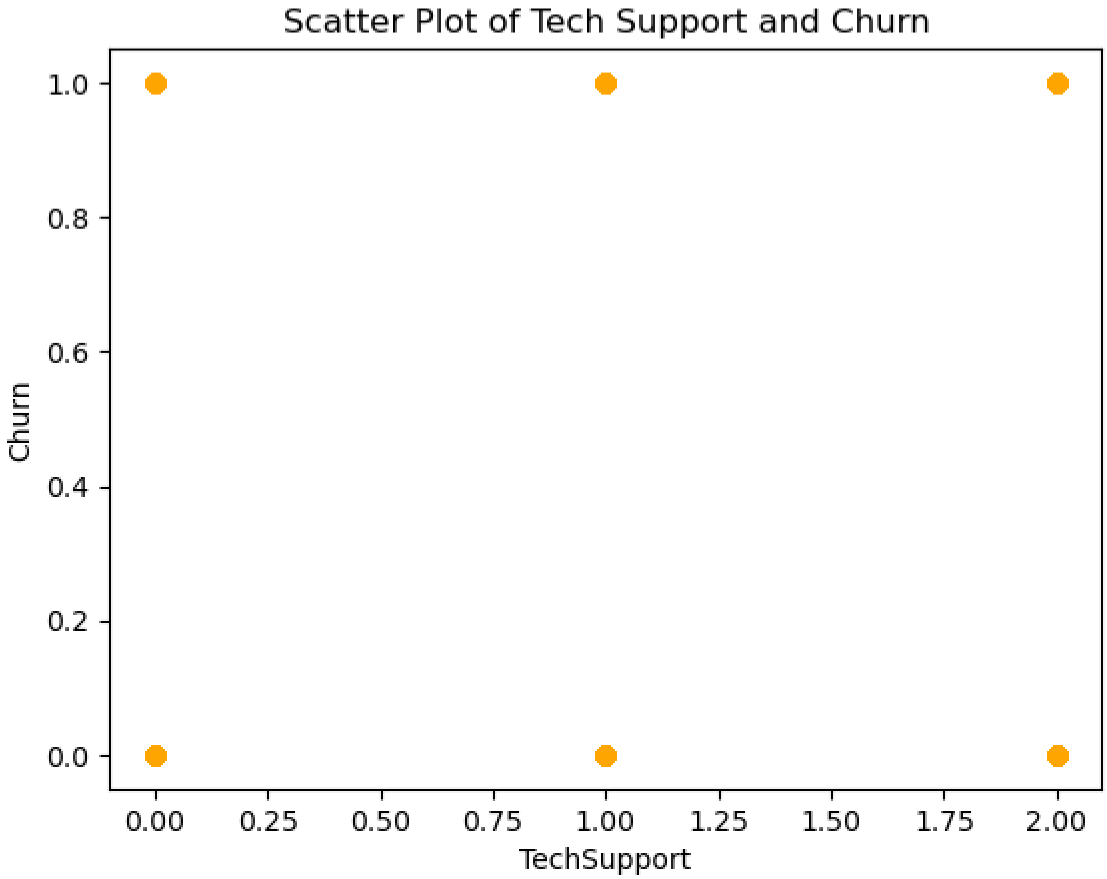
```
df_subset = df3[['TechSupport', 'Churn']]
df_subset.corr()
```

Out[101]:

	TechSupport	Churn
TechSupport	1.000000	0.026126
Churn	0.026126	1.000000

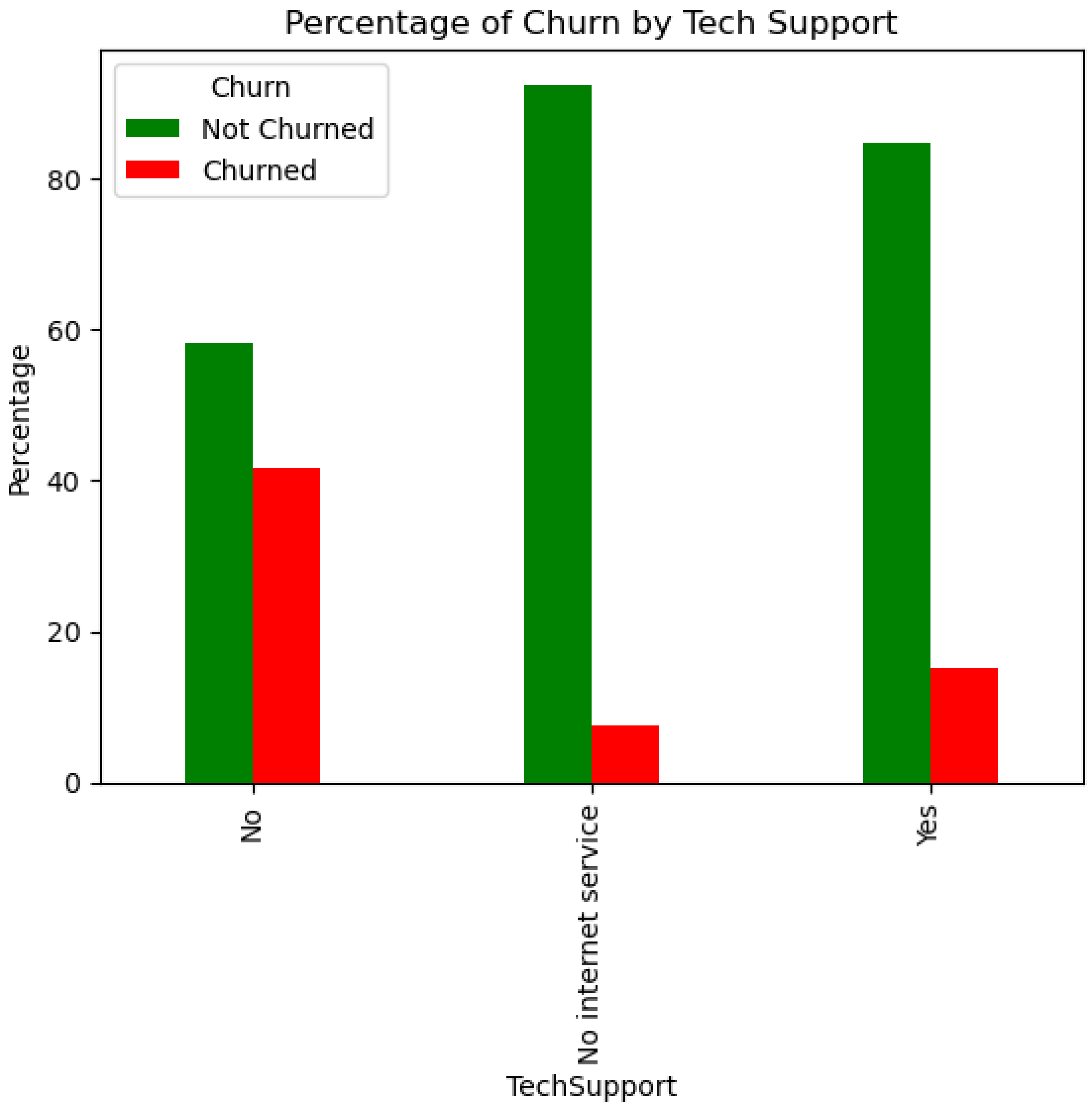
In [104]:

```
plt.scatter(x =df3['TechSupport'], y=df3['Churn'], c='orange')
plt.xlabel("TechSupport")
plt.ylabel("Churn")
plt.title("Scatter Plot of Tech Support and Churn")
plt.show()
```



In [103]:

```
grouped_data = pd.crosstab(df['TechSupport'], df['Churn'], normalize='index')
* 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('TechSupport')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Tech Support')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the TechSupport and Churn columns. From the bar graph, we can observe that nearly 60% of the customers with No Tech Support did not churn, while the remaining 40% did. Almost 95% of the customers without any Internet Service did not churn, while the remaining 5% did. Almost 85% of the customers with Tech Support did not churn, while the remaining 15% did.

13. StreamingTV and Churn

In [105]:

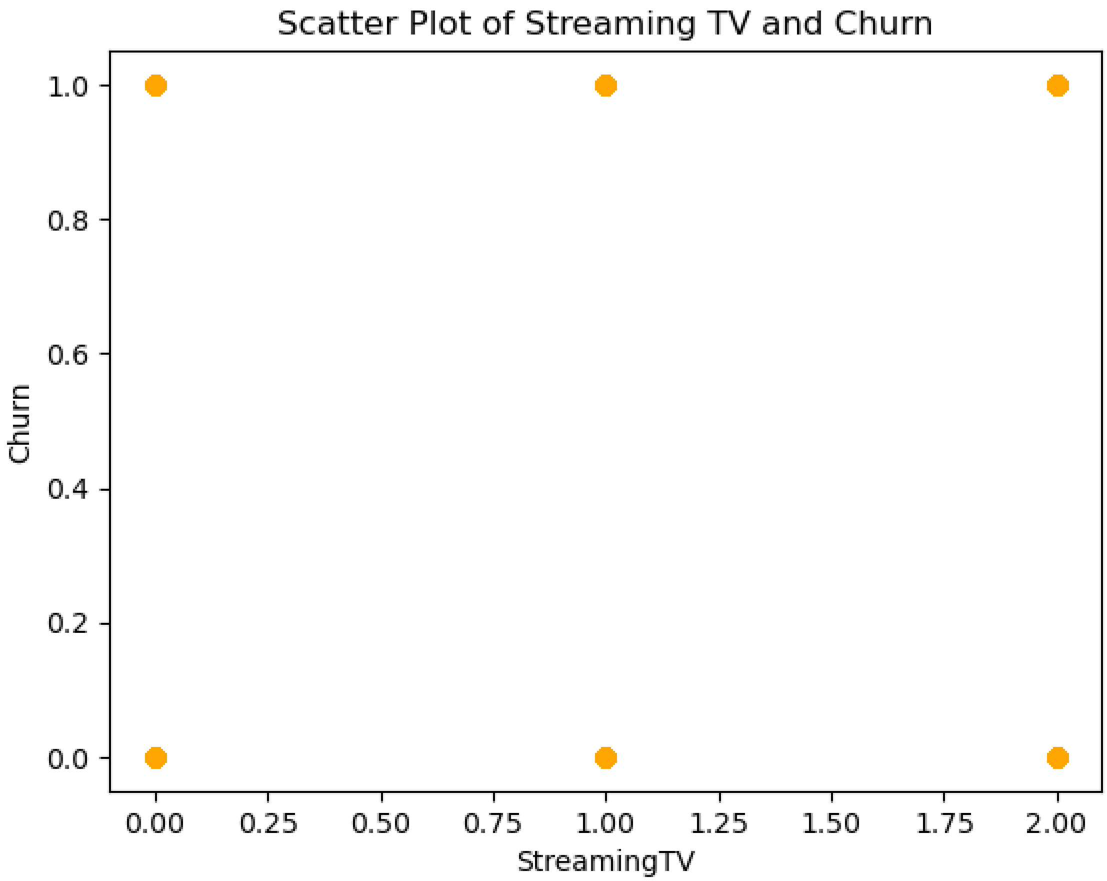
```
df_subset = df3[['StreamingTV', 'Churn']]
df_subset.corr()
```

Out[105]:

	StreamingTV	Churn
StreamingTV	1.000000	0.163977
Churn	0.163977	1.000000

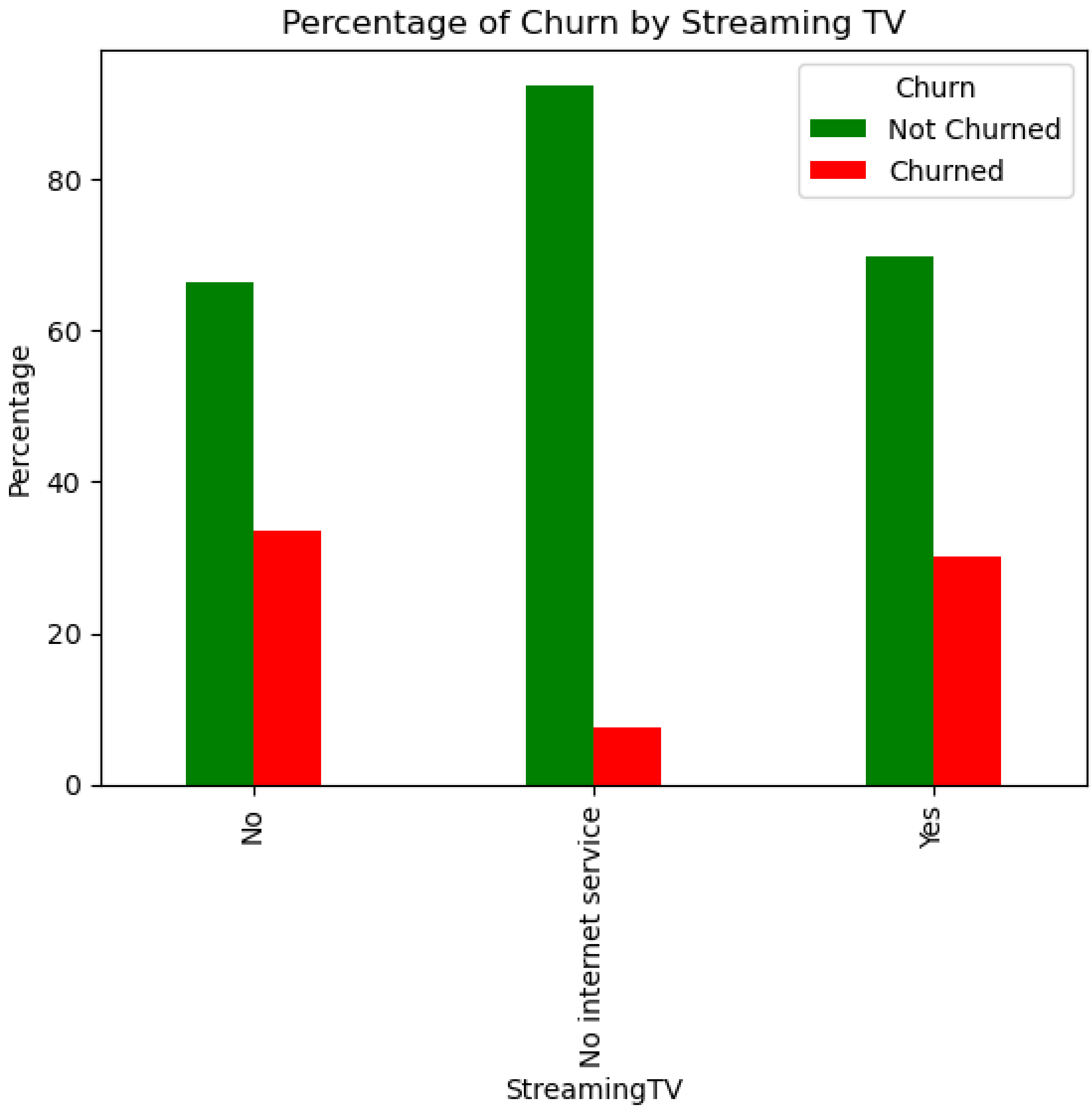
In [106]:

```
plt.scatter(x =df3['StreamingTV'], y=df3['Churn'], c='orange')
plt.xlabel("StreamingTV")
plt.ylabel("Churn")
plt.title("Scatter Plot of Streaming TV and Churn")
plt.show()
```



In [107]:

```
grouped_data = pd.crosstab(df['StreamingTV'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('StreamingTV')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Streaming TV')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the StreamingTV and Churn columns. From the bar graph, we can observe that nearly 65% of the customers with No Streaming TV service did not churn, while the remaining 35% did. Almost 95% of the customers without any Internet Service did not churn, while the remaining 5% did. Almost 70% of the customers with Streaming TV did not churn, while the remaining 30% did.

14. StreamingMovies and Churn

In [108]:

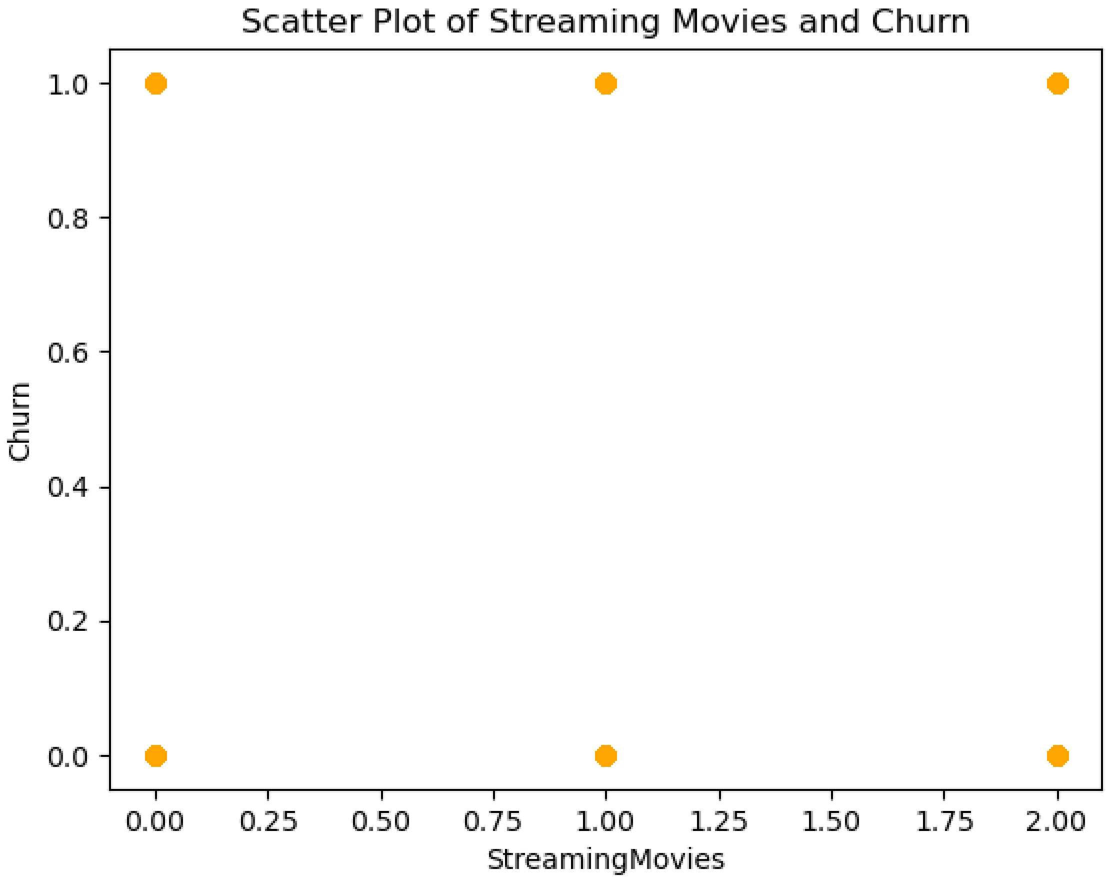
```
df_subset = df3[['StreamingMovies', 'Churn']]
df_subset.corr()
```

Out[108]:

	StreamingMovies	Churn
StreamingMovies	1.000000	0.162437
Churn	0.162437	1.000000

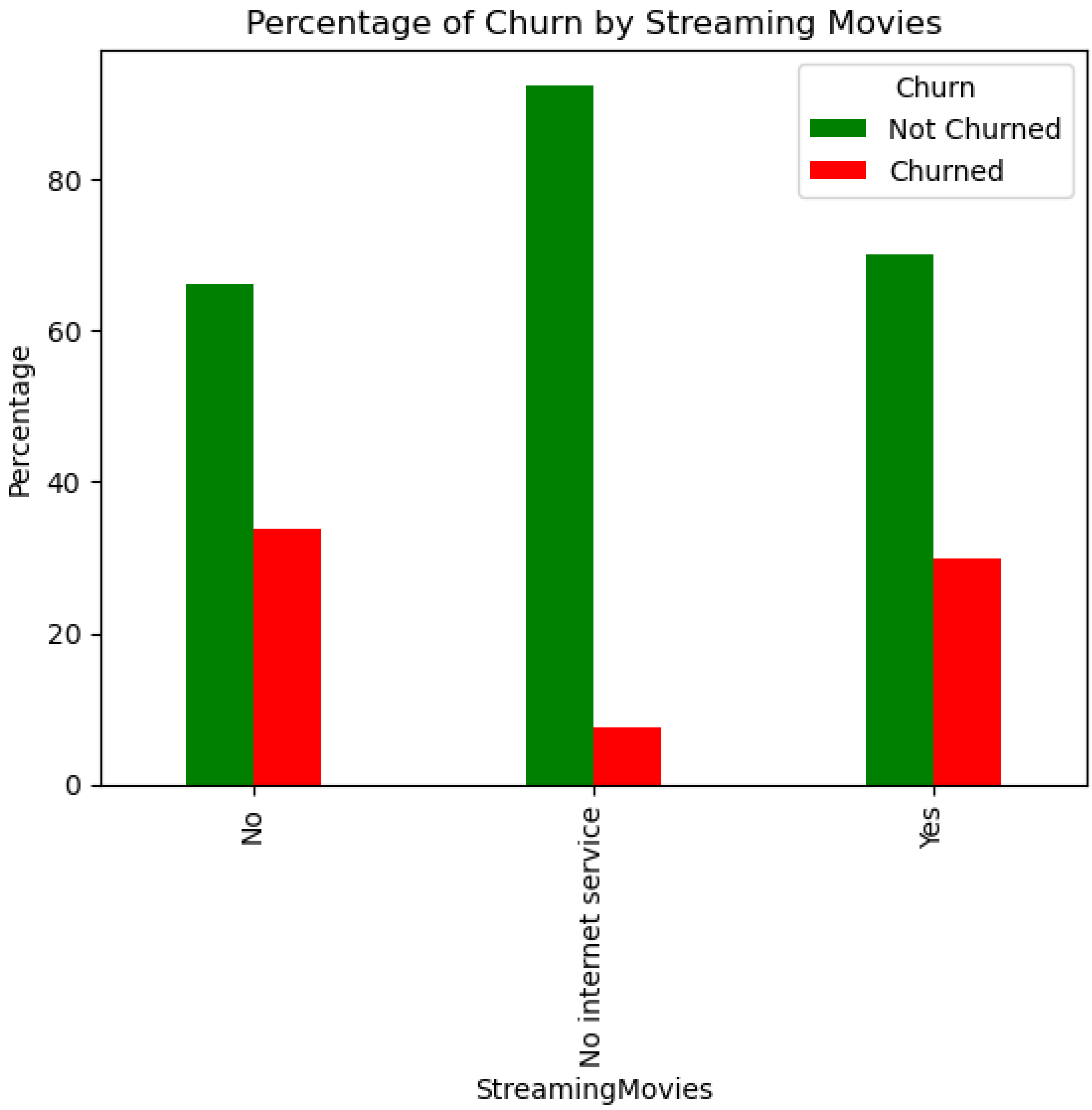
In [109]:

```
plt.scatter(x =df3['StreamingMovies'], y=df3['Churn'], c='orange')
plt.xlabel("StreamingMovies")
plt.ylabel("Churn")
plt.title("Scatter Plot of Streaming Movies and Churn")
plt.show()
```



In [110]:

```
grouped_data = pd.crosstab(df['StreamingMovies'], df['Churn'], normalize='inde
x') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('StreamingMovies')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Streaming Movies')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the StreamingMovies and Churn columns. From the bar graph, we can observe that nearly 65% of the customers with No Streaming Movies service did not churn, while the remaining 35% did. Almost 95% of the customers without any Internet Service did not churn, while the remaining 5% did. Almost 70% of the customers with Streaming Movies did not churn, while the remaining 30% did.

15. Contract and Churn

In [111]:

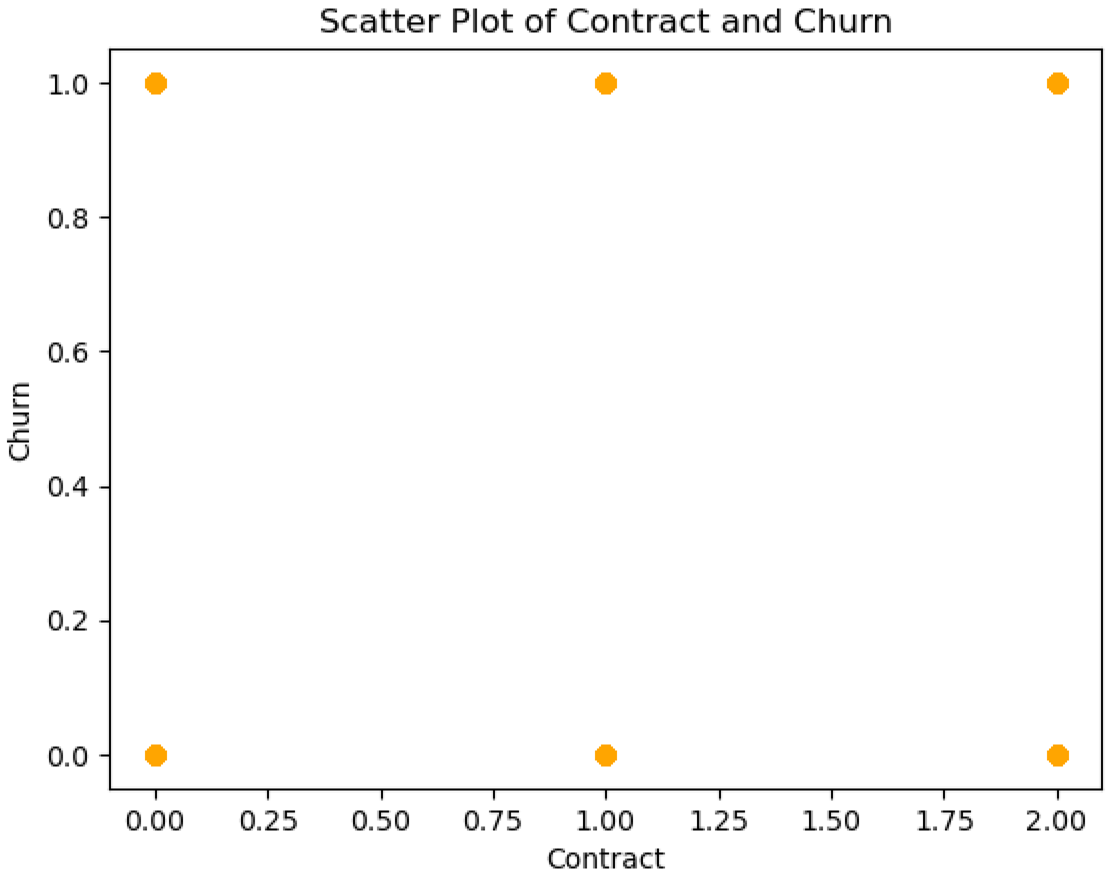
```
df_subset = df3[['Contract', 'Churn']]
df_subset.corr()
```

Out[111]:

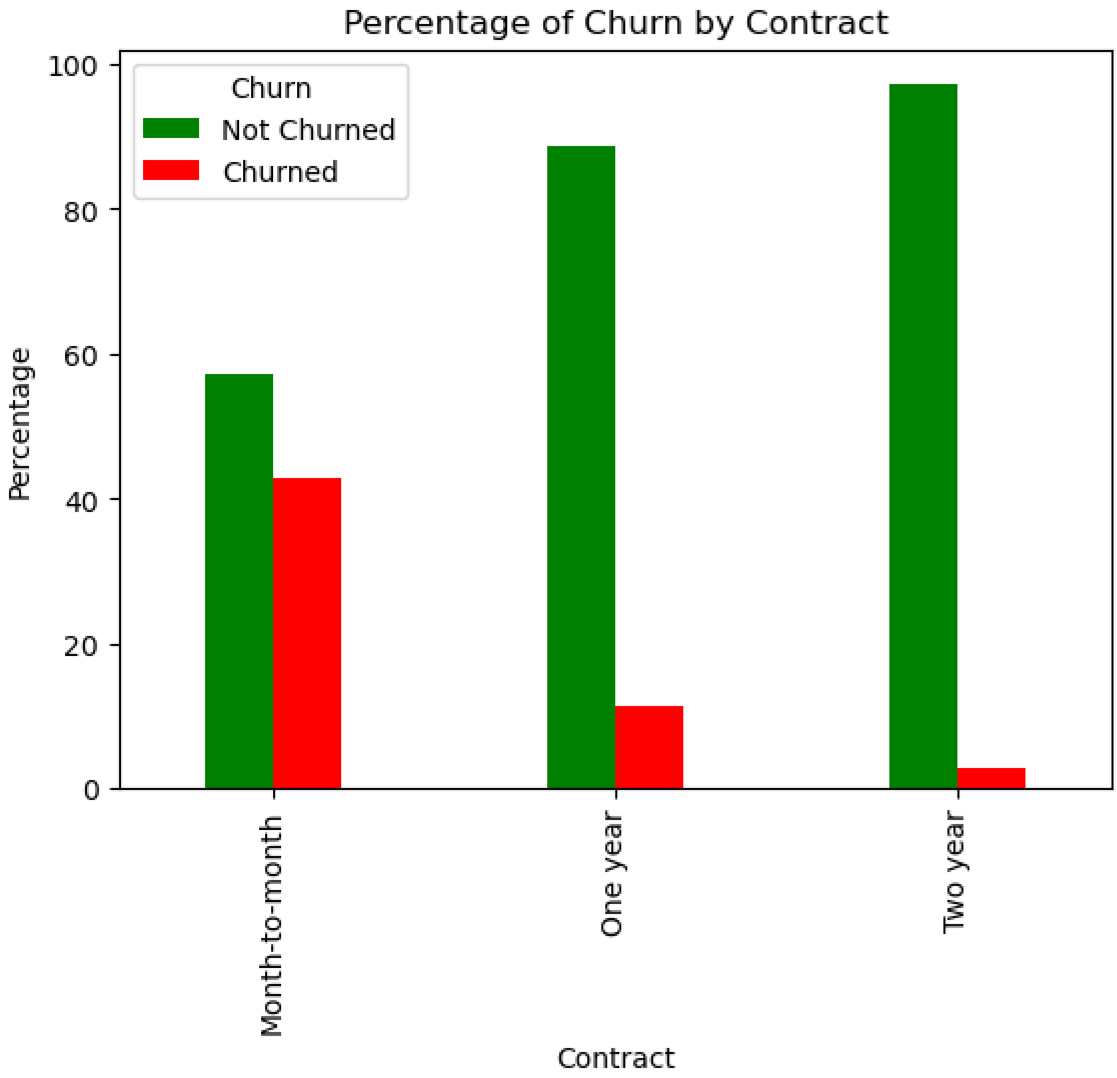
	Contract	Churn
Contract	1.000000	-0.031874
Churn	-0.031874	1.000000

In [112]:

```
plt.scatter(x =df3['Contract'], y=df3['Churn'], c='orange')
plt.xlabel("Contract")
plt.ylabel("Churn")
plt.title("Scatter Plot of Contract and Churn")
plt.show()
```



```
In [113]: grouped_data = pd.crosstab(df['Contract'], df['Churn'], normalize='index') * 1
          00
          colors = ['green', 'red']
          chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

          plt.xlabel('Contract')
          plt.ylabel('Percentage')
          plt.title('Percentage of Churn by Contract')
          plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
          plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak negative correlation, or no correlation, between the Contract and Churn columns. From the bar graph, we can observe that nearly 58% of the customers with a Month-to-month contract did not churn, while the remaining 42% did. Almost 88% of the customers with a 1-year contract did not churn, while the remaining 12% did. Almost 95% of the customers with a 2-year contract did not churn, while the remaining 5% did.

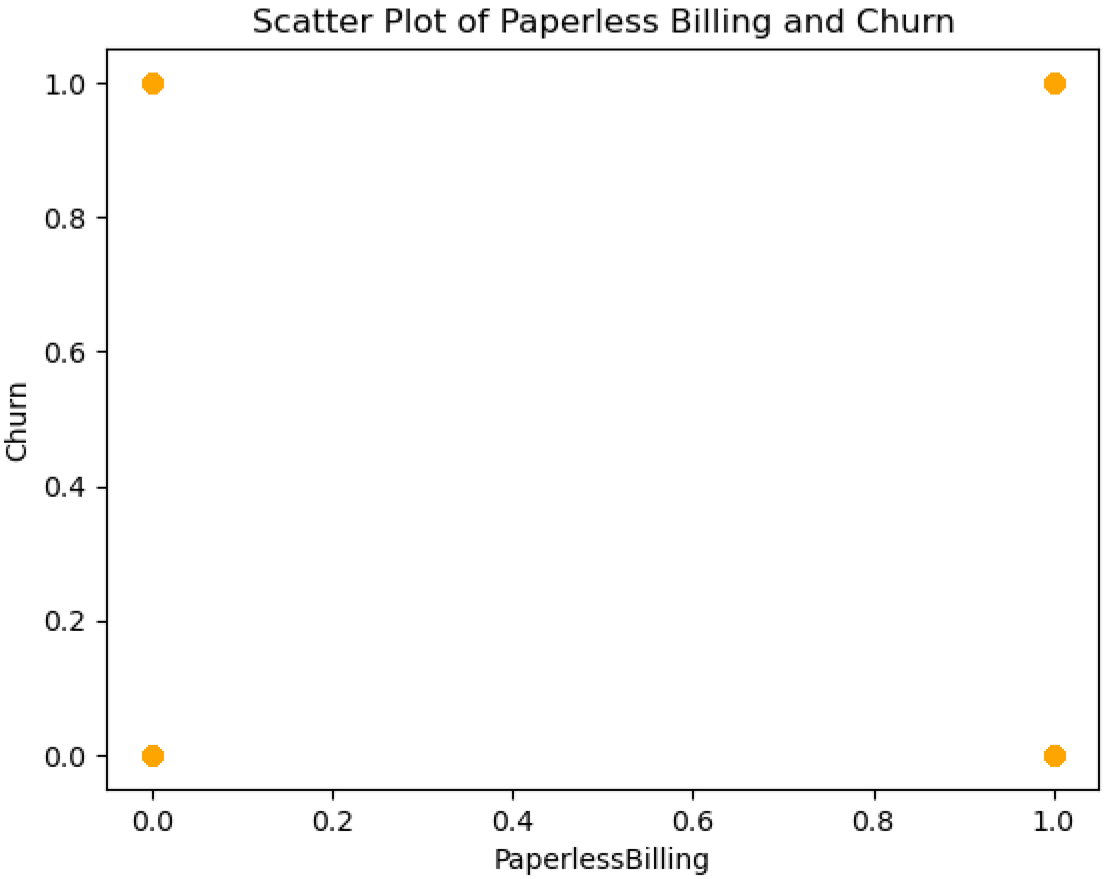
16. PaperlessBilling and Churn

In [114]:
```
df_subset = df3[['PaperlessBilling', 'Churn']]
df_subset.corr()
```

Out[114]:

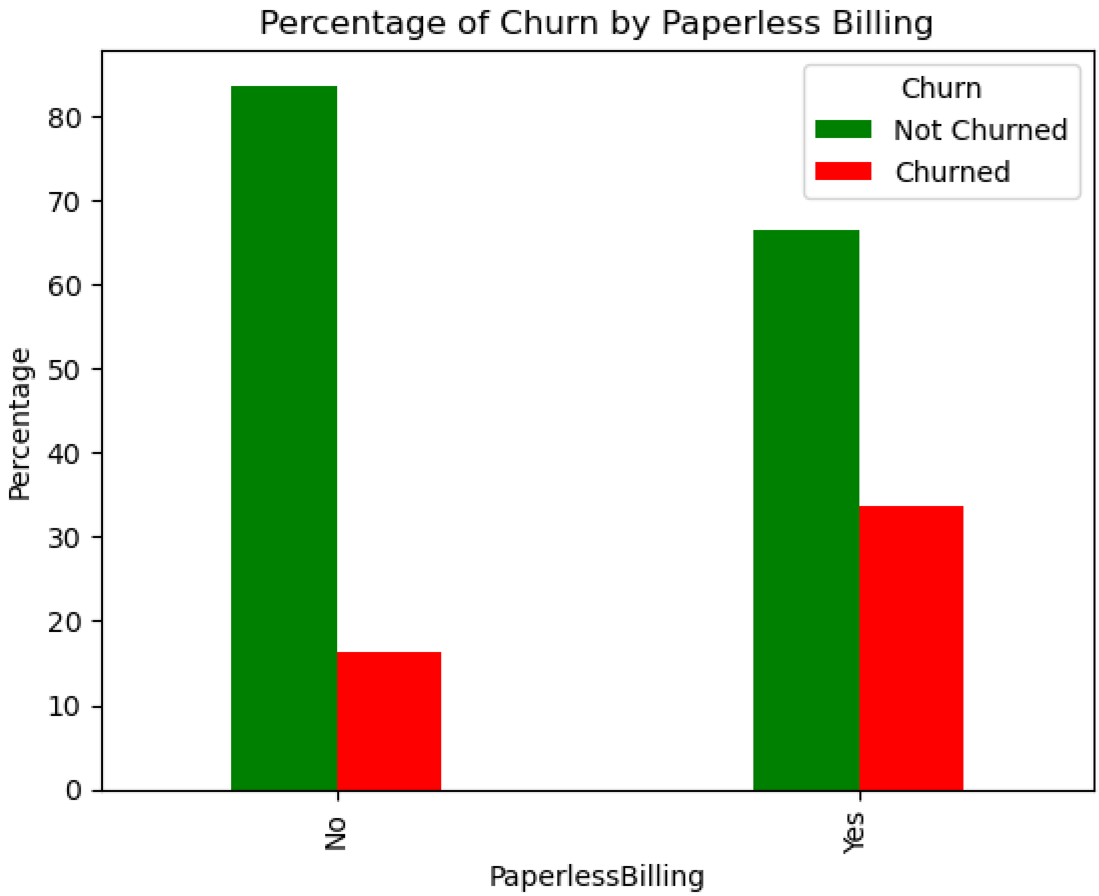
	PaperlessBilling	Churn
PaperlessBilling	1.000000	0.191291
Churn	0.191291	1.000000

In [115]:
```
plt.scatter(x =df3['PaperlessBilling'], y=df3['Churn'], c='orange')
plt.xlabel("PaperlessBilling")
plt.ylabel("Churn")
plt.title("Scatter Plot of Paperless Billing and Churn")
plt.show()
```



In [116]:

```
grouped_data = pd.crosstab(df['PaperlessBilling'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('PaperlessBilling')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Paperless Billing')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the PaperlessBilling and Churn columns. From the bar graph, we can observe that nearly 85% of the customers without Paperless Billing did not churn, while the remaining 15% did. Almost 68% of the customers with Paperless Billing did not churn, while the remaining 32% did.

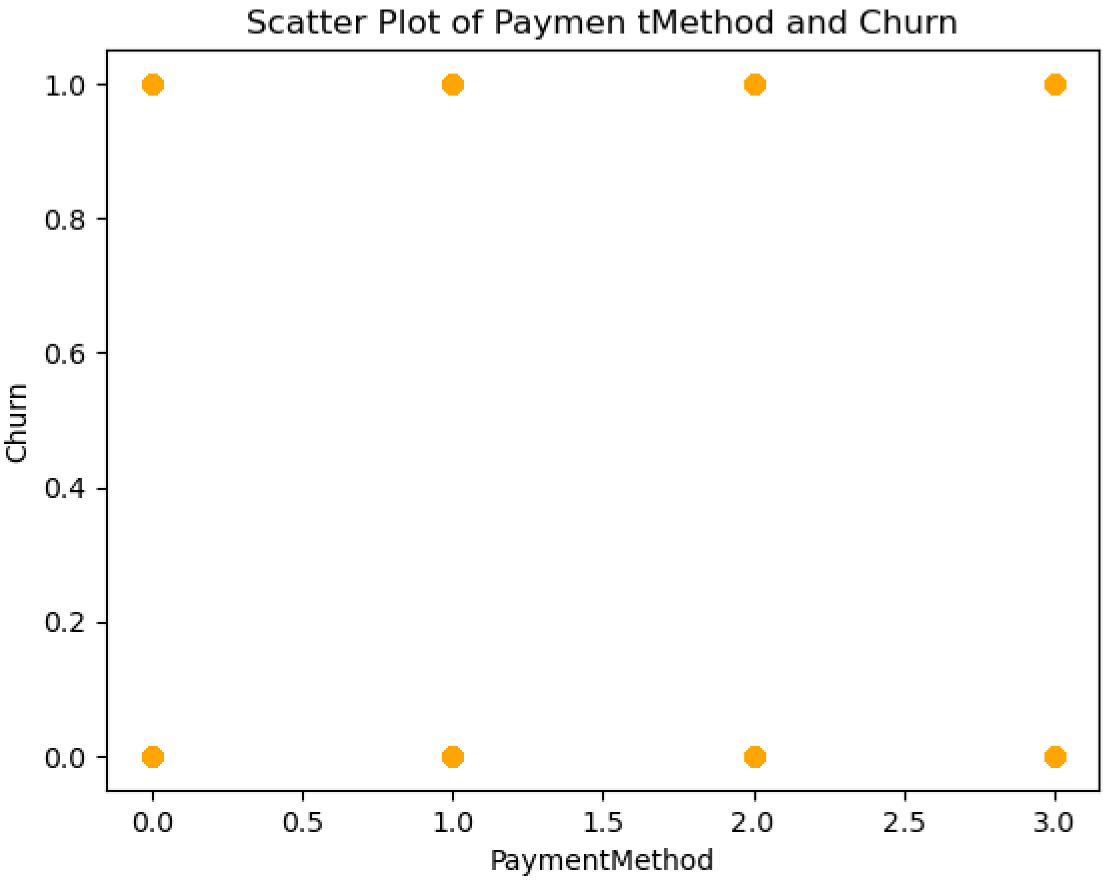
17. PaymentMethod and Churn

In [117]:
```
df_subset = df3[['PaymentMethod', 'Churn']]
df_subset.corr()
```

Out[117]:

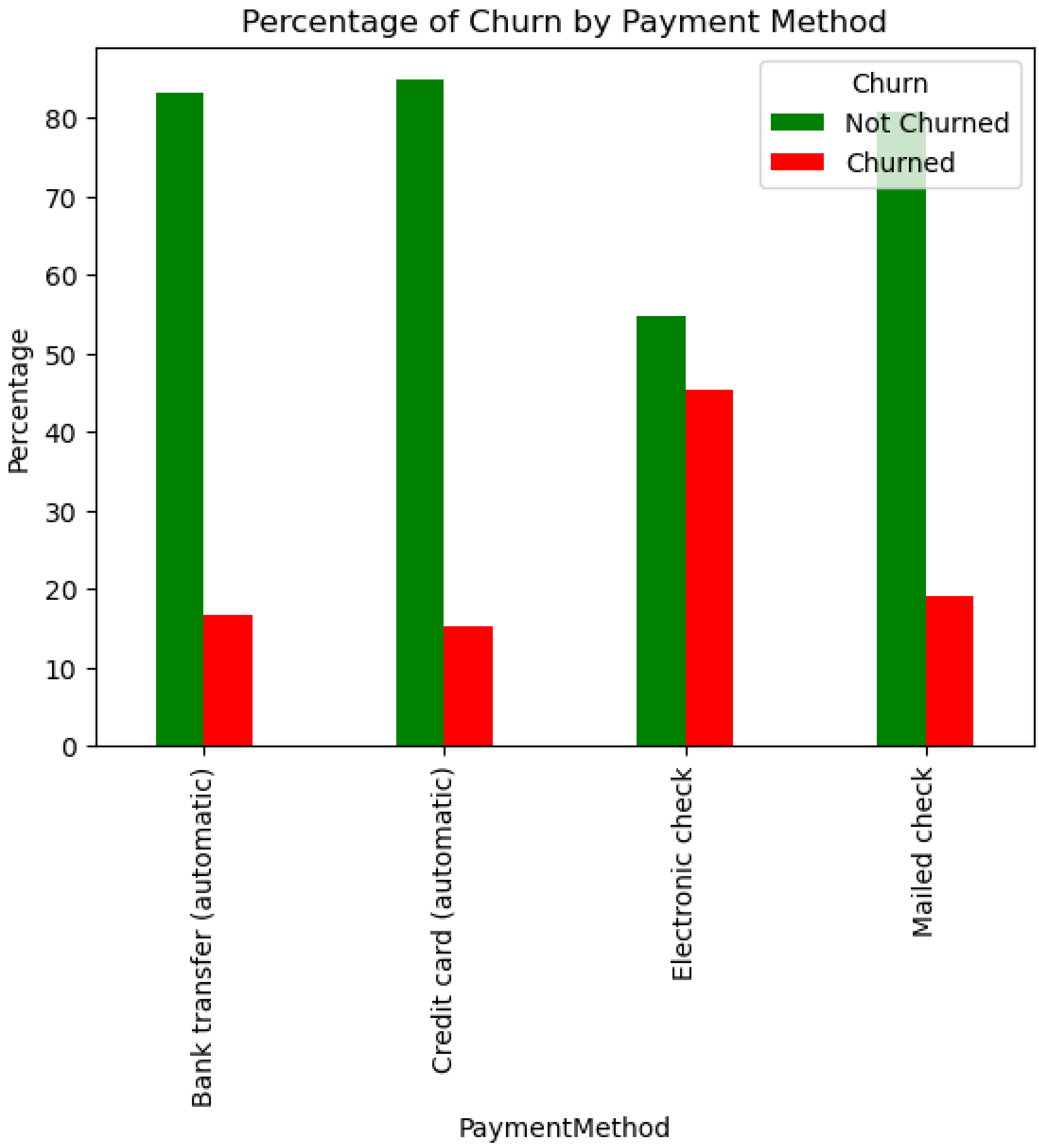
	PaymentMethod	Churn
PaymentMethod	1.000000	0.024865
Churn	0.024865	1.000000

In [118]:
```
plt.scatter(x =df3['PaymentMethod'], y=df3['Churn'], c='orange')
plt.xlabel("PaymentMethod")
plt.ylabel("Churn")
plt.title("Scatter Plot of Paymen tMethod and Churn")
plt.show()
```



In [119]:

```
grouped_data = pd.crosstab(df['PaymentMethod'], df['Churn'], normalize='index') * 100
colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.4, color=colors)

plt.xlabel('PaymentMethod')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Payment Method')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the PaymentMethod and Churn columns. From the bar graph, we can observe that nearly 83% of the customers with Bank transfer payment method did not churn, while the remaining 17% did. Almost 85% of the customers with Credit card Payment method did not churn, while the remaining 15% did. Almost 55% of the customers with Electronic check Payment method did not churn, while the remaining 45% did. Almost 80% of the customers with Mailed check Payment method did not churn, while the remaining 20% did.

18. MonthlyCharges and Churn

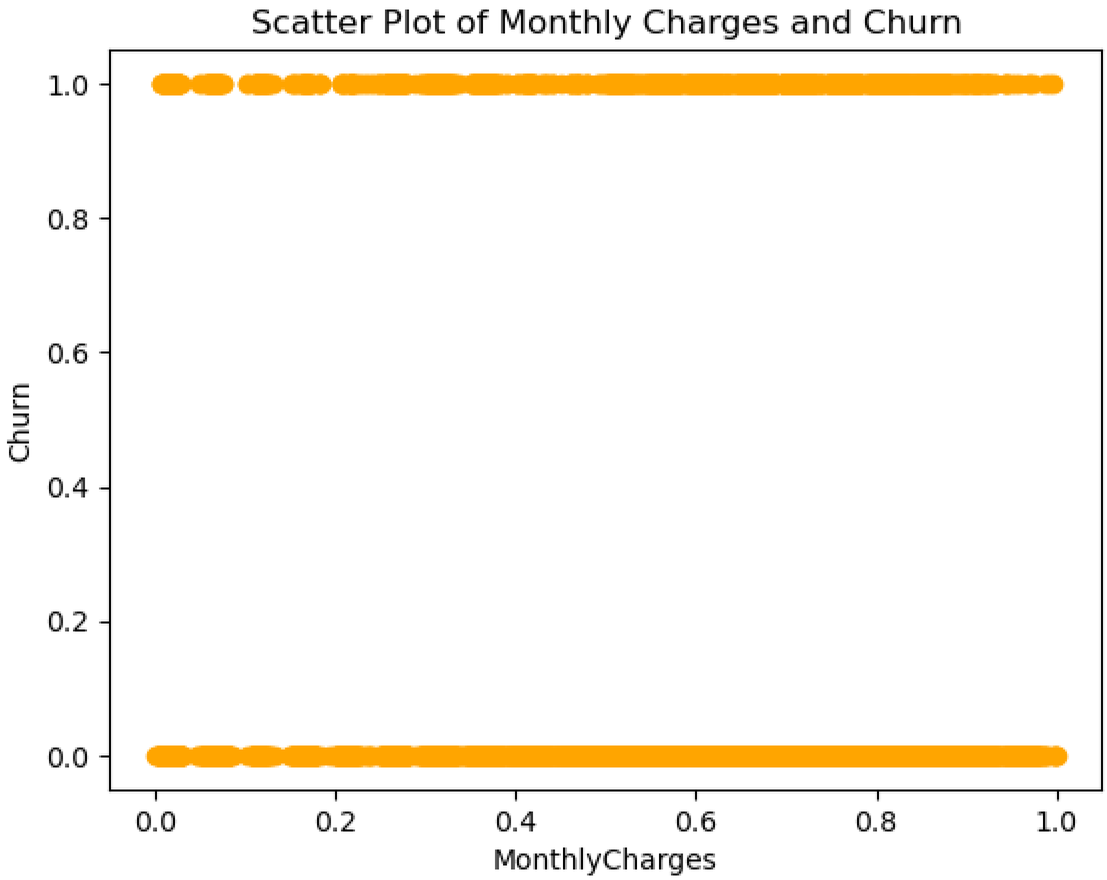
In [120]:

```
df_subset = df3[['MonthlyCharges', 'Churn']]
df_subset.corr()
```

Out[120]:

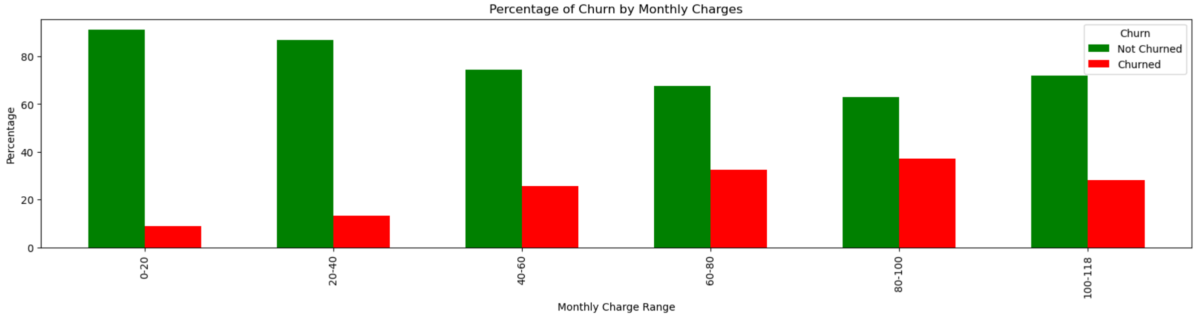
	MonthlyCharges	Churn
MonthlyCharges	1.000000	0.192313
Churn	0.192313	1.000000

In [121]:
```
plt.scatter(x =df3['MonthlyCharges'], y=df3['Churn'], c='orange')
plt.xlabel("MonthlyCharges")
plt.ylabel("Churn")
plt.title("Scatter Plot of Monthly Charges and Churn")
plt.show()
```



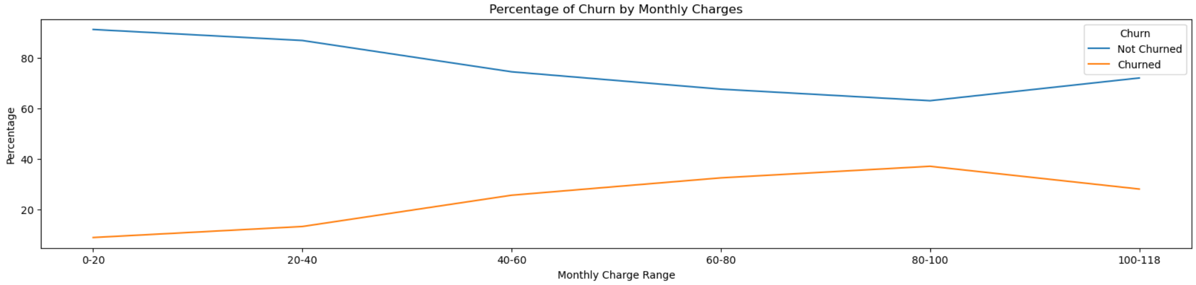
In [154]:

```
bin_range = [0, 20, 40, 60, 80, 100, df['MonthlyCharges'].max()]

# Labels for the bins
labels = [f"{int(bin_range[i-1])}-{int(bin_range[i])}" for i in range(1, len(bin_range))]
df['MonthlyChargeRange'] = pd.cut(df['MonthlyCharges'], bins=bin_range, labels=labels)
grouped_data = pd.crosstab(df['MonthlyChargeRange'], df['Churn'], normalize='index') * 100

colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.6, color=colors, figsize=(20, 4))
plt.xlabel('Monthly Charge Range')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Monthly Charges')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



In [155]:

```
bin_range = [0, 20, 40, 60, 80, 100, df['MonthlyCharges'].max()]

# Labels for the bins
labels = [f"{int(bin_range[i-1])}-{int(bin_range[i])}" for i in range(1, len(bin_range))]
df['MonthlyChargeRange'] = pd.cut(df['MonthlyCharges'], bins=bin_range, labels=labels)
grouped_data = pd.crosstab(df['MonthlyChargeRange'], df['Churn'], normalize='index') * 100

colors = ['green', 'red']
chart = grouped_data.plot(kind='line', figsize=(20, 4))
plt.xlabel('Monthly Charge Range')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Monthly Charges')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak positive correlation, or no correlation, between the MonthlyCharges and Churn columns. From the bar and line graphs, we cannot identify a steady trend. The churn percentage is the lowest (5%) at very low monthly charges. With increasing monthly charges, the churn percentage steadily increases, reaching 10% at the 20-40 interval, 25% at the 40-60 interval, 35% at the 60-80 interval, and finally its highest value of 40% at the 80-100 interval. After that churn percentage steadily decreases, reaching almost 30% at the maximum monthly charges.

19. TotalCharges and Churn

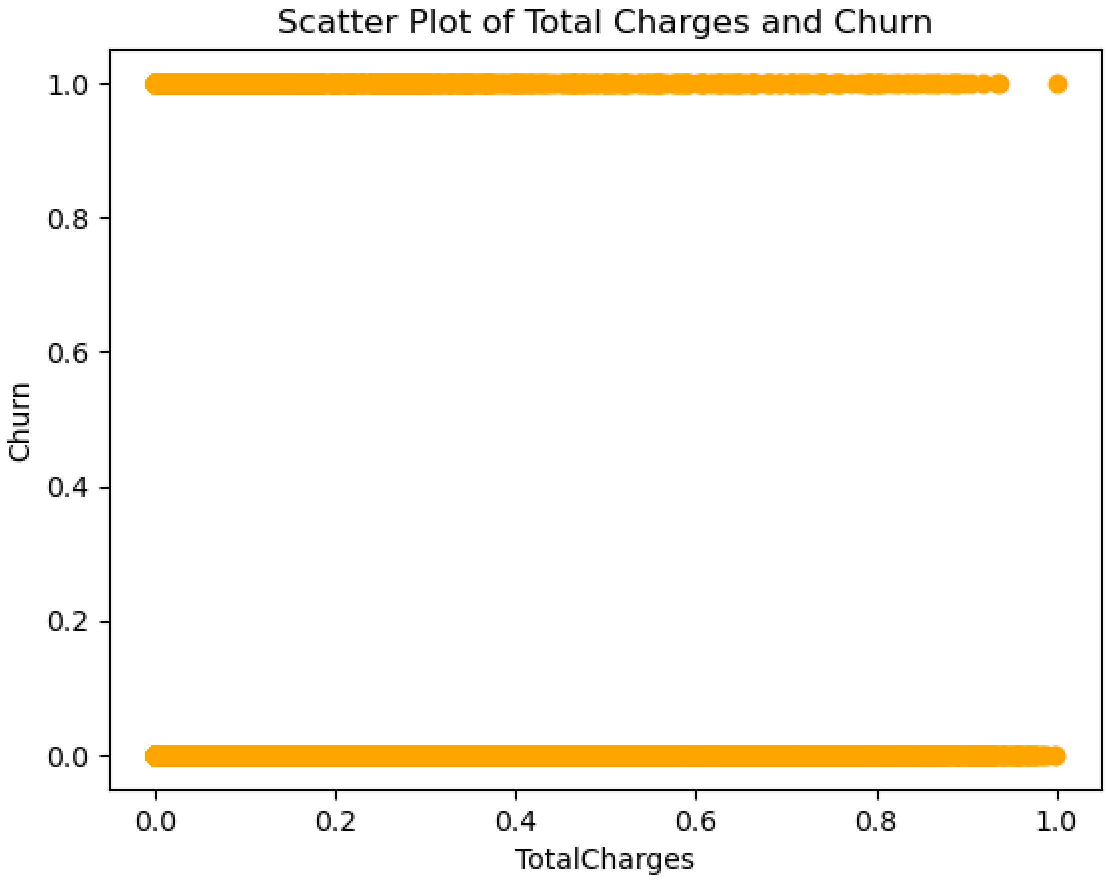
In [142]:

```
df_subset = df3[['TotalCharges', 'Churn']]
df_subset.corr()
```

Out[142]:

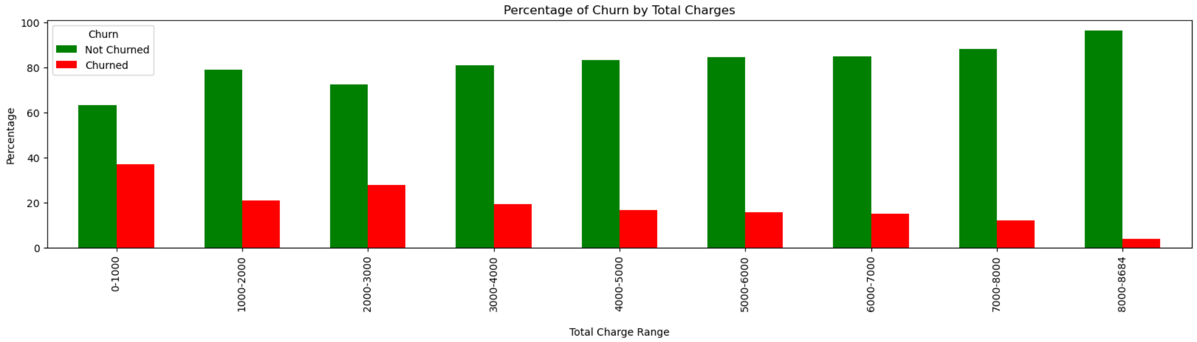
	TotalCharges	Churn
TotalCharges	1.000000	-0.199843
Churn	-0.199843	1.000000

```
In [143]: plt.scatter(x =df3['TotalCharges'], y=df3['Churn'], c='orange')
          plt.xlabel("TotalCharges")
          plt.ylabel("Churn")
          plt.title("Scatter Plot of Total Charges and Churn")
          plt.show()
```



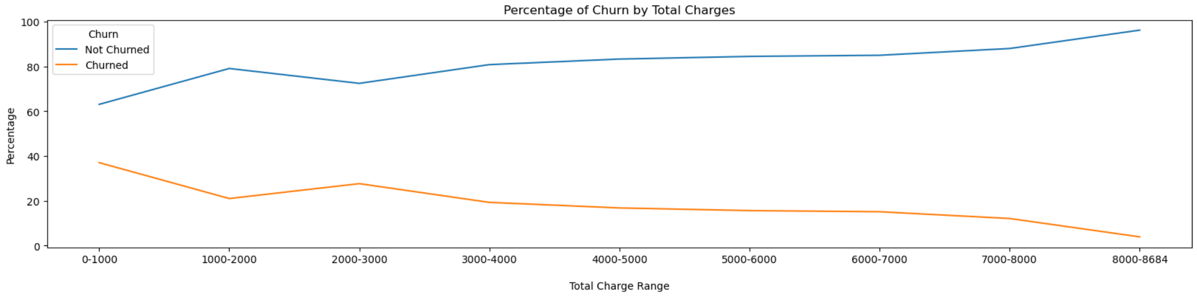
In [156]:

```
# Defining the bin edges for monthly charges
bin_edges = []
num=0
for i in range(9):
    bin_edges.append(num)
    num+=1000
bin_edges.append(df['TotalCharges'].max())

# Creating labels for the bins
labels = [f"{int(bin_edges[i-1])}-{int(bin_edges[i])}" for i in range(1, len(bin_edges))]

# Adding a new column to DataFrame with the labels
df['TotalChargeRange'] = pd.cut(df['TotalCharges'], bins=bin_edges, labels=labels)

# Using crosstab with the new column
grouped_data = pd.crosstab(df['TotalChargeRange'], df['Churn'], normalize='index') * 100

colors = ['green', 'red']
chart = grouped_data.plot(kind='bar', width=0.6, color=colors, figsize=(20, 4))

plt.xlabel('\nTotal Charge Range')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Total Charges')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



In [157]:

```
# Defining the bin edges for monthly charges
bin_edges = []
num=0
for i in range(9):
    bin_edges.append(num)
    num+=1000
bin_edges.append(df['TotalCharges'].max())

# Creating labels for the bins
labels = [f"{int(bin_edges[i-1])}-{int(bin_edges[i])}" for i in range(1, len(bin_edges))]

# Adding a new column to DataFrame with the labels
df['TotalChargeRange'] = pd.cut(df['TotalCharges'], bins=bin_edges, labels=labels)

# Using crosstab with the new column
grouped_data = pd.crosstab(df['TotalChargeRange'], df['Churn'], normalize='index') * 100

colors = ['green', 'red']
chart = grouped_data.plot(kind='line', figsize=(20, 4))

plt.xlabel('\nTotal Charge Range')
plt.ylabel('Percentage')
plt.title('Percentage of Churn by Total Charges')
plt.legend(title='Churn', labels=['Not Churned', 'Churned'])
plt.show()
```



Descriptive Analysis

From the correlation matrix and the scatter plot, we can observe a very weak negative correlation, or no correlation, between the TotalCharges and Churn columns. From the bar and line graphs, we can identify that he churn percentage is its highest (almost 40%) at lowest total charges. At the start of the 1000-2000 interval, it decreases to nearly 20%. At the start of the next interval, 2000-3000, churn percentage increases to nearly 30%. At the beginnning of the 3000-4000 interval, it again decreases to almost 20%. At the start of the 4000-5000 interval, churn percentage further decreases to around 15%, and changes very slightly till the beginning of the 7000-8000 interval, where it reaches nearly 10%. By the final value of total charges, churn percentage reaches its minimum value of less than 5%.

Summary of Multivariate Analysis

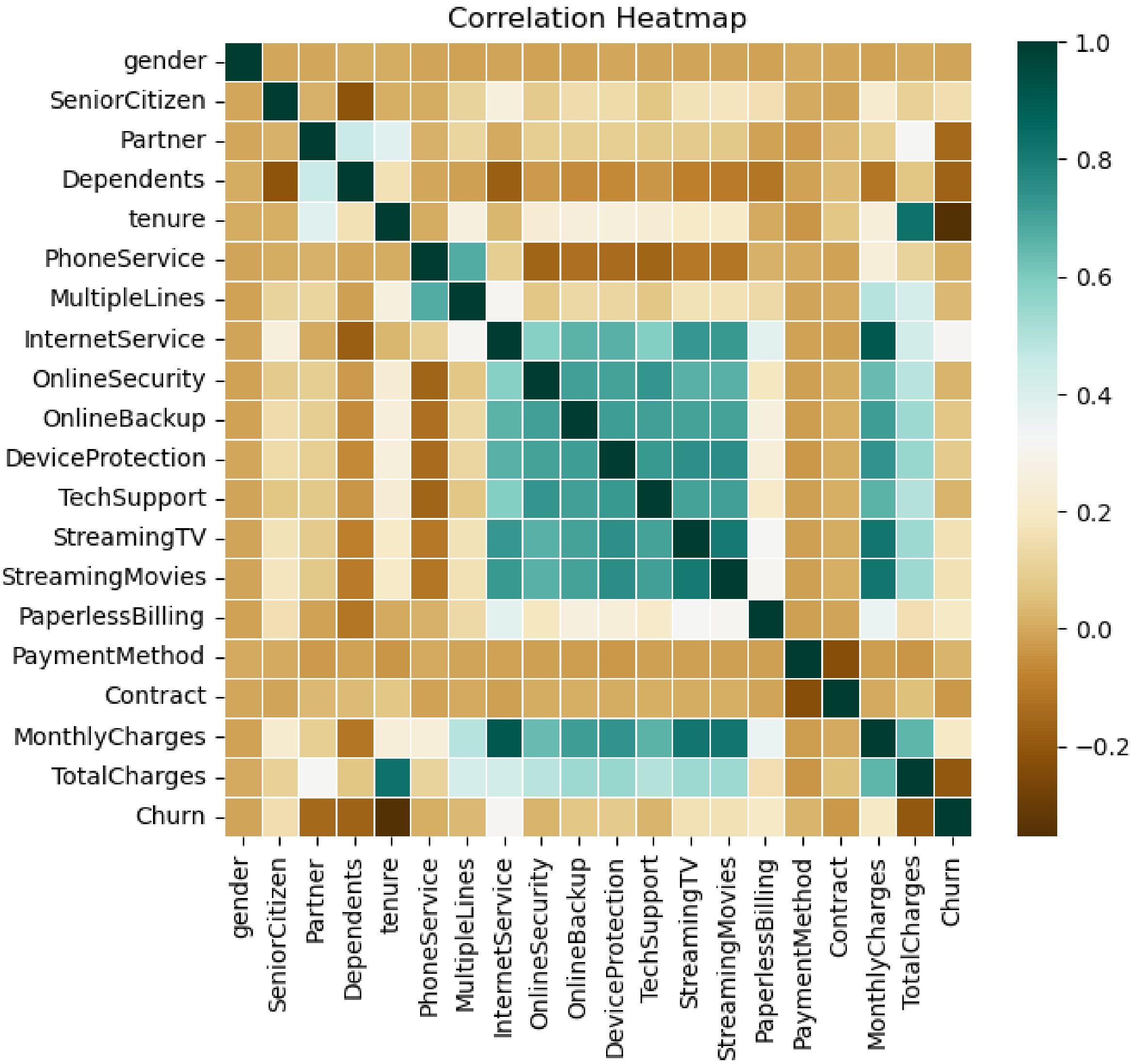
In [145]:

```
corr_matrix = df3.corr()
corr_matrix
```

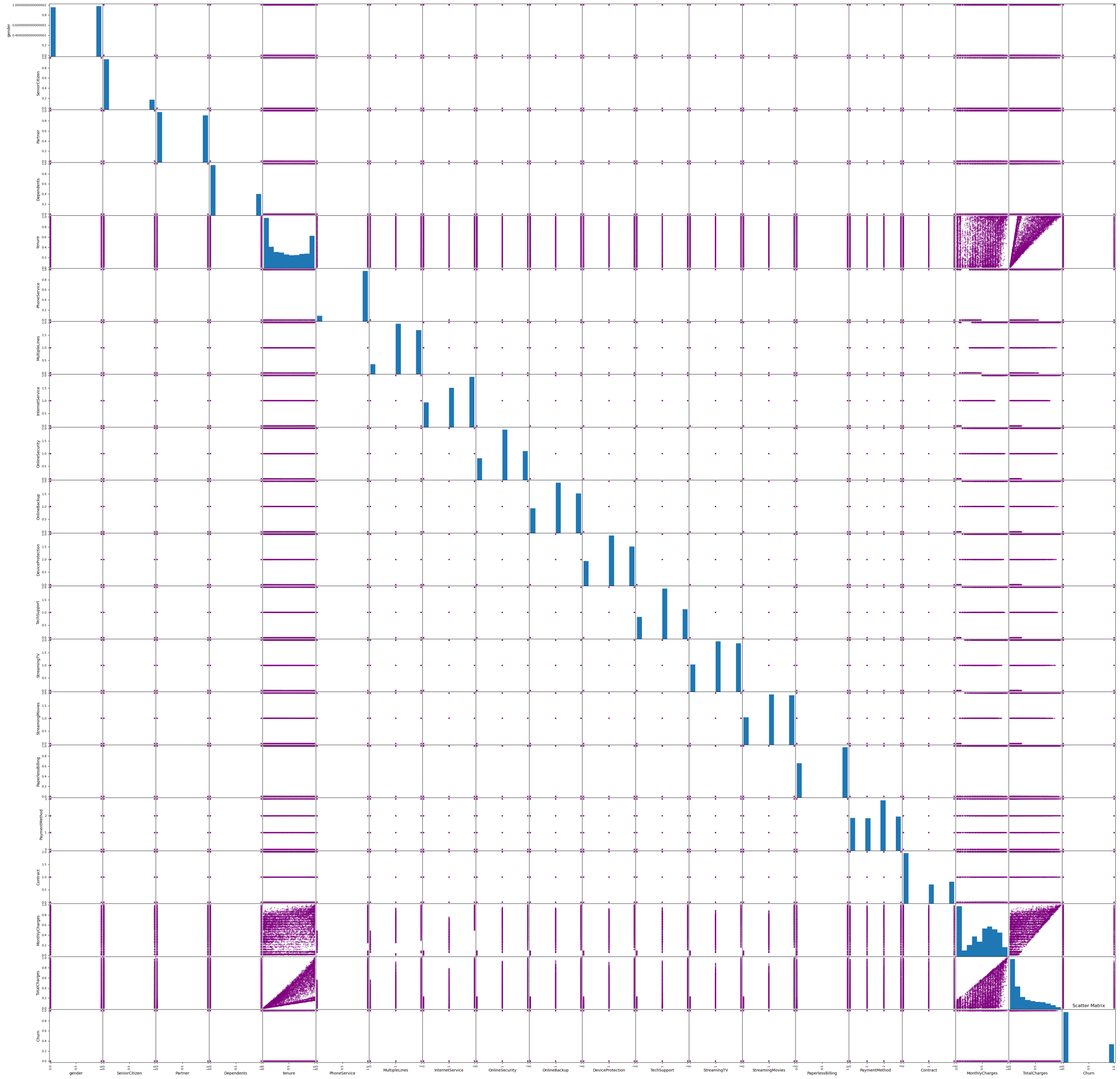
Out[145]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	Multi
gender	1.000000	-0.002010	-0.001233	0.010396	0.005791	-0.007377	-0
SeniorCitizen	-0.002010	1.000000	0.017315	-0.210423	0.016004	0.008373	0
Partner	-0.001233	0.017315	1.000000	0.452609	0.381241	0.018721	0
Dependents	0.010396	-0.210423	0.452609	1.000000	0.163011	-0.001050	-0
tenure	0.005791	0.016004	0.381241	0.163011	1.000000	0.008060	0
PhoneService	-0.007377	0.008373	0.018721	-0.001050	0.008060	1.000000	0
MultipleLines	-0.010729	0.114062	0.118334	-0.019206	0.259673	0.675052	1
InternetService	-0.010146	0.258873	0.001375	-0.178052	0.031766	0.094195	0
OnlineSecurity	-0.013905	0.081618	0.091874	-0.028963	0.232553	-0.158840	0
OnlineBackup	-0.011564	0.144820	0.091510	-0.062321	0.253725	-0.129307	0
DeviceProtection	-0.003828	0.140049	0.099265	-0.068298	0.254124	-0.141170	0
TechSupport	-0.008680	0.067249	0.077272	-0.040384	0.230583	-0.160877	0
StreamingTV	-0.007834	0.166685	0.080684	-0.085713	0.201209	-0.107433	0
StreamingMovies	-0.009936	0.175608	0.076354	-0.099328	0.204501	-0.114923	0
PaperlessBilling	-0.011858	0.156598	-0.014597	-0.111174	0.004094	0.016760	0
PaymentMethod	0.000457	0.001135	-0.026959	-0.012011	-0.039232	0.000913	-0
Contract	-0.004192	-0.010228	0.035174	0.040577	0.068938	-0.015448	0
MonthlyCharges	-0.014330	0.219618	0.098275	-0.112357	0.247464	0.248178	0
TotalCharges	0.000316	0.102645	0.318672	0.064510	0.825958	0.113230	0
Churn	-0.008813	0.149685	-0.149561	-0.164029	-0.354315	0.011590	0

```
In [148]: plt.figure(figsize=(7, 6))
          sns.heatmap(corr_matrix, annot=False, cmap="BrBG", linewidths=.5)

          plt.title("Correlation Heatmap")
          plt.show()
```



```
In [149]: pd.plotting.scatter_matrix(df3,figsize=(50, 50), c='purple')
          plt.title("Scatter Matrix")
          plt.show()
```



Descriptive Analysis

The correlation matrix above shows the correlation between all the columns in the dataframe. These correlation values are then represented via a heatmap. Finally, the scatter matrix shows the correlation between all the columns via scatter plots.

Key findings and insights from the descriptive analysis

1) The customer base exhibits nearly equal numbers of both genders, with an equally balanced ratio of churned and non-churned customers. This suggests that the company has effectively attracted and retained customers of both genders.

2) The majority of the company's customers fall under the non-senior citizen category, accounting for approximately 83.79%. Senior citizens are more likely to churn (55%) compared to non-senior citizens, indicating that the company struggles to both attract and retain senior citizens.

3) A significant portion of the company's customers are single, making up around 51.70% of the total. However, customers without partners are more prone to churn (35%) than those with partners (20%), indicating that while the company has attracted more single individuals, it has also experienced higher attrition in this group.

4) Approximately 70.04% of the customers do not have dependents, and those with dependents are less likely to churn (15%) compared to those without dependents (32%). This implies that the company has successfully attracted customers with dependents and retained them more effectively.

5) Most customers have a relatively short tenure with the company, averaging around 32 months. There is a very weak negative correlation between customer tenure and churn, indicating that tenure has a minimal impact on churn.

6) The majority of customers have phone service (90.32%), and customers with phone service are slightly more likely to churn (28%) compared to those without (25%).

7) A significant portion of customers do not have multiple lines (48.13%), and those with multiple lines are most likely to churn (28%), while those with no phone service or no multiple lines have nearly equal churn rates (25%).

8) The majority of customers use fiber optic as their internet service (43.96%), and these customers are also more likely to churn (40%), while those with no internet service are least likely to churn (10%).

9) Most customers do not have online security (3498), and these customers are most likely to churn (40%), while those with no internet service are least likely to churn (10%).

10) The majority of customers have no online backup (3088), and these customers are most likely to churn (40%), while those with no internet service are least likely to churn (10%).

11) A significant number of customers have no device protection (43.94%), and these customers are most likely to churn (40%), while those with no internet

service are least likely to churn (5%).

12) A large portion of customers have no device protection (43.94%), and these customers are most likely to churn (40%), while those with no internet service are least likely to churn (5%).

13) The majority of customers lack tech support (49.31%), and these customers are most likely to churn (40%), while those with no internet service are least likely to churn (5%).

14) Most customers do not have streaming TV (39.90%), and these customers are most likely to churn (35%), while those with no internet service are least likely to churn (5%).

15) A significant portion of customers have no movie streaming (39.54%), and these customers are most likely to churn (35%), while those with no internet service are least likely to churn (5%).

16) The majority of customers have a month-to-month contract (55.02%), and these customers are most likely to churn (42%), while those with a two-year contract are least likely to churn (5%).

17) The majority of customers prefer paperless billing (59.22%), and these customers are more likely to churn (32%), while those without paperless billing are less likely to churn (15%).

18) Most customers use electronic check as their method of payment (2365), and these customers are most likely to churn (45%), while those who choose to pay with credit card (automatic) are least likely to churn (15%).

19) Customers have an average monthly charge of approximately $64, and there is a weak positive correlation between monthly charges and customer churn.

20) Customers have an average total charge of around $2283, with most customers having lower charges due to a right-skewed distribution. There is a weak negative correlation or no correlation between total charges and customer churn.

21) The majority of customers, approximately 73.46%, did not churn, indicating a high customer retention rate. However, 26.54% of customers did choose to churn. Features with a positive correlation with churn include monthly charges, payment method, paperless billing, streaming movies, streaming TV, tech support, device protection, online backup, online security, internet service, multiple lines, and phone service, while the rest have negative correlations.