# Customer Churn Analysis – Telecom Dataset

This project analyzes customer churn using Python, aiming to understand customer retention patterns. It explores various factors like demographics, tenure, contract type, and service usage to identify churn drivers.

**Key Libraries Used**: `Pandas`, `NumPy`, `Matplotlib`, `Seaborn`
**Dataset:** [https://app.mavenanalytics.io/datasets?order=-fields.dateUpdated&search=telecom+customer+churn]

## Dataset Overview and Preprocessing

- Loaded the dataset using `pandas`
- Removed unnecessary customer types (e.g., *Joined*)
- Checked for missing values and dataset shape
- Reset index for a clean view

Below is the basic structure and missing value analysis of the dataset.

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore", category=FutureWarning)
df = pd.read_csv(r"telecom_customer_churn.csv")
print("Data Shape:", df.shape)
print("\nColumns:", df.columns.tolist())
print("\nMissing Values:")
print(df.isnull().sum())
df = df[df['Customer Status'].isin(['Stayed', 'Churned'])]
df.reset_index(drop=True, inplace=True)
df.head()
```

```
Data Shape: (7043, 38)

Columns: ['Customer ID', 'Gender', 'Age', 'Married', 'Number of
Dependents', 'City', 'Zip Code', 'Latitude', 'Longitude', 'Number of
Referrals', 'Tenure in Months', 'Offer', 'Phone Service', 'Avg Monthly
Long Distance Charges', 'Multiple Lines', 'Internet Service',
'Internet Type', 'Avg Monthly GB Download', 'Online Security', 'Online
Backup', 'Device Protection Plan', 'Premium Tech Support', 'Streaming
TV', 'Streaming Movies', 'Streaming Music', 'Unlimited Data',
'Contract', 'Paperless Billing', 'Payment Method', 'Monthly Charge',
'Total Charges', 'Total Refunds', 'Total Extra Data Charges', 'Total
Long Distance Charges', 'Total Revenue', 'Customer Status', 'Churn
Category', 'Churn Reason']
```

```
Missing Values:
Customer ID                             0
Gender                                  0
Age                                     0
Married                                 0
Number of Dependents                    0
City                                    0
Zip Code                                0
Latitude                                0
Longitude                               0
Number of Referrals                     0
Tenure in Months                        0
Offer                                3877
Phone Service                           0
Avg Monthly Long Distance Charges     682
Multiple Lines                        682
Internet Service                        0
Internet Type                        1526
Avg Monthly GB Download              1526
Online Security                      1526
Online Backup                        1526
Device Protection Plan               1526
Premium Tech Support                 1526
Streaming TV                         1526
Streaming Movies                     1526
Streaming Music                      1526
Unlimited Data                       1526
Contract                                0
Paperless Billing                       0
Payment Method                          0
Monthly Charge                          0
Total Charges                           0
Total Refunds                           0
Total Extra Data Charges                0
Total Long Distance Charges             0
Total Revenue                           0
Customer Status                         0
Churn Category                       5174
Churn Reason                         5174
dtype: int64

  Customer ID  Gender  Age Married  Number of Dependents          City
\
0  0002-ORFBO  Female   37     Yes                     0  Frazier Park

1  0003-MKNFE    Male   46      No                     0      Glendale

2  0004-TLHLJ    Male   50      No                     0    Costa Mesa

3  0011-IGKFF    Male   78     Yes                     0      Martinez
```

```
4  0013-EXCHZ  Female   75     Yes                     0     Camarillo


   Zip Code   Latitude   Longitude  Number of Referrals  ...   Payment
Method  \
0     93225  34.827662 -118.999073                    2  ...
Credit Card
1     91206  34.162515 -118.203869                    0  ...
Credit Card
2     92627  33.645672 -117.922613                    0  ...  Bank
Withdrawal
3     94553  38.014457 -122.115432                    1  ...  Bank
Withdrawal
4     93010  34.227846 -119.079903                    3  ...
Credit Card

  Monthly Charge Total Charges  Total Refunds Total Extra Data Charges
\
0           65.6        593.30           0.00                        0

1           -4.0        542.40          38.33                       10

2           73.9        280.85           0.00                        0

3           98.0       1237.85           0.00                        0

4           83.9        267.40           0.00                        0


  Total Long Distance Charges Total Revenue  Customer Status   Churn
Category  \
0                      381.51        974.81           Stayed
NaN
1                       96.21        610.28           Stayed
NaN
2                      134.60        415.45          Churned
Competitor
3                      361.66       1599.51          Churned
Dissatisfaction
4                       22.14        289.54          Churned
Dissatisfaction

                    Churn Reason
0                            NaN
1                            NaN
2  Competitor had better devices
3        Product dissatisfaction
4            Network reliability
```

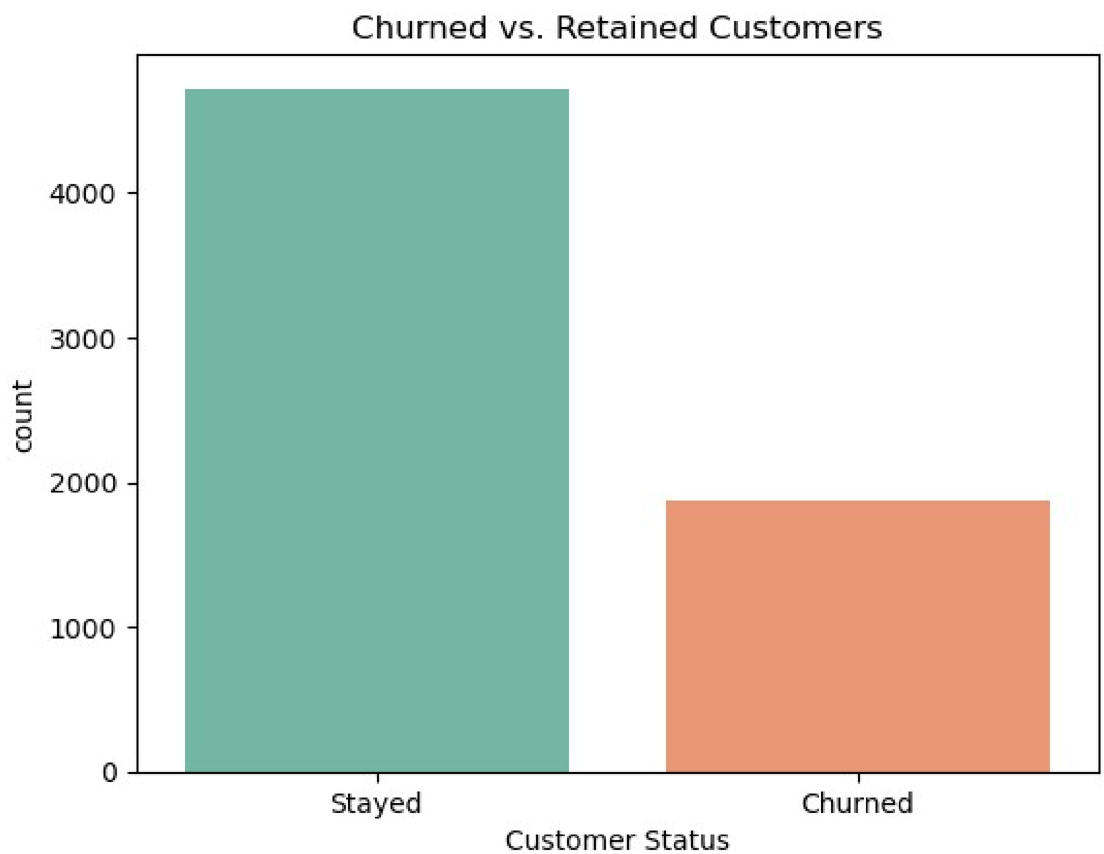
```
[5 rows x 38 columns]

print("Monthly Charge Stats:\n", df['Monthly Charge'].describe())
print("\nTotal Charges Stats:\n", df['Total Charges'].describe())
print("\nTenure Stats:\n", df['Tenure in Months'].describe())

Monthly Charge Stats:
 count    6589.000000
mean       65.030695
std        31.100727
min       -10.000000
25%        35.800000
50%        71.050000
75%        90.400000
max       118.750000
Name: Monthly Charge, dtype: float64

Total Charges Stats:
 count    6589.000000
mean     2432.042243
std      2265.500080
min        18.850000
25%       544.550000
50%      1563.900000
75%      4003.000000
max      8684.800000
Name: Total Charges, dtype: float64

Tenure Stats:
 count    6589.000000
mean       34.499772
std        23.968734
min         1.000000
25%        12.000000
50%        32.000000
75%        57.000000
max        72.000000
Name: Tenure in Months, dtype: float64
```

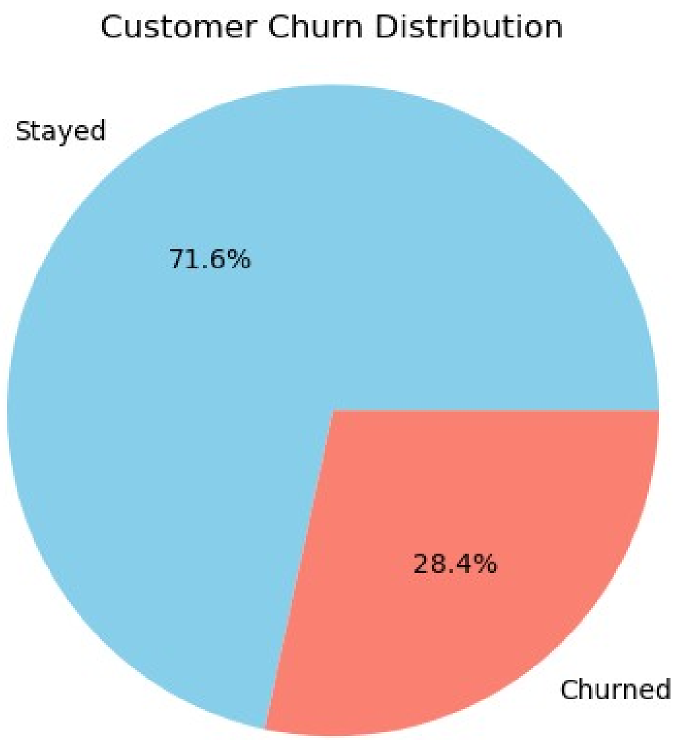
## □ Churned vs. Retained Customers

This section visualizes the distribution of churned and retained customers using bar and pie charts.

```
#1. Analyze the Distribution of Churned vs. Retained Customers
churn_counts = df['Customer Status'].value_counts()
print(churn_counts)
```

```
# Bar plot
sns.countplot(x='Customer Status', data=df, palette='Set2')
plt.title('Churned vs. Retained Customers')
plt.show()

# Pie chart
plt.pie(churn_counts, labels=churn_counts.index, autopct='%1.1f%%',
colors=['skyblue', 'salmon'])
plt.title('Customer Churn Distribution')
plt.axis('equal')
plt.show()

Customer Status
Stayed     4720
Churned    1869
Name: count, dtype: int64
```





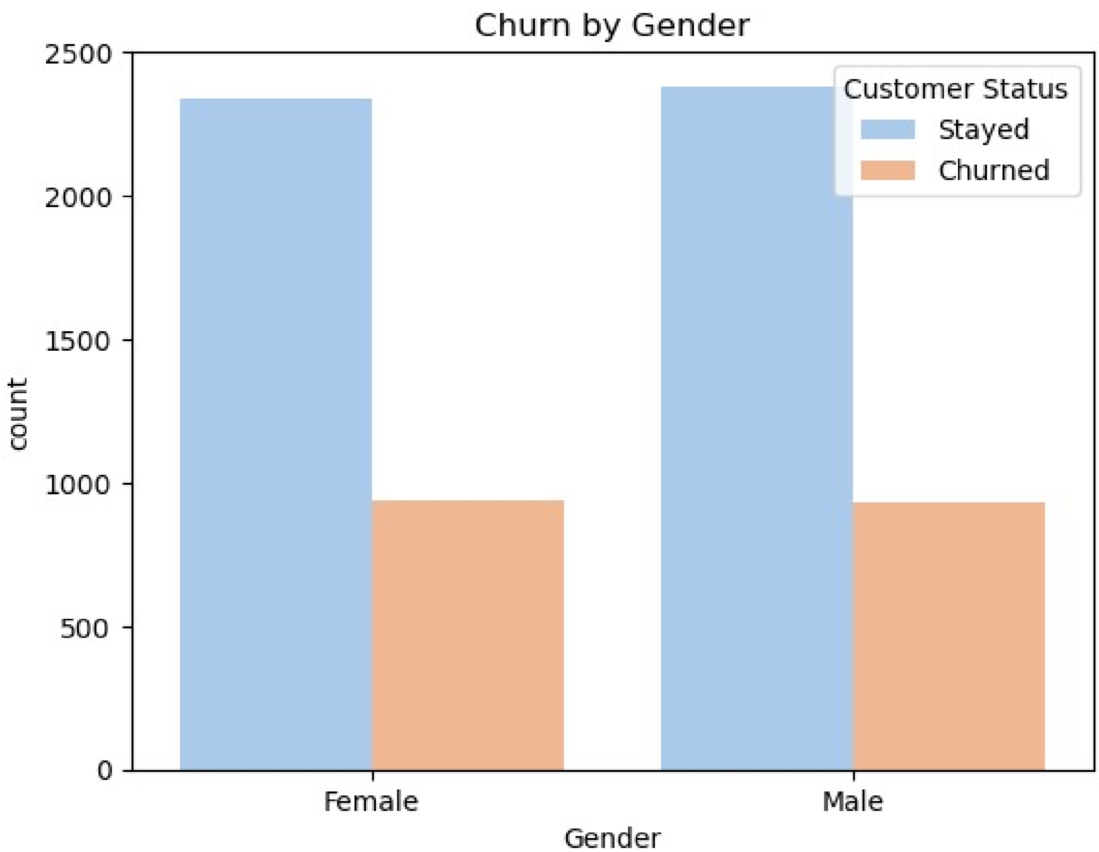

## Demographics and Churn

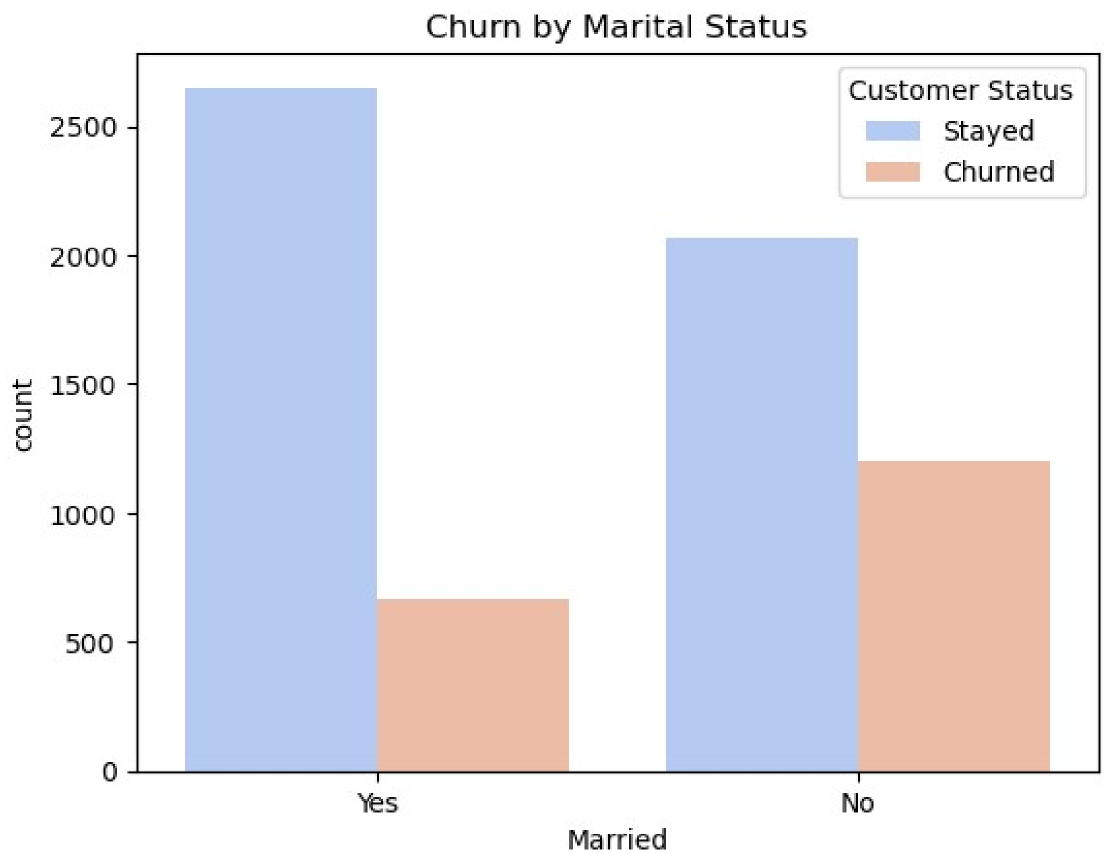
Analyzing churn distribution across demographic categories:

- Gender
- Marital Status
- Tenure (via violin plot)
- Contract type (via boxplot)

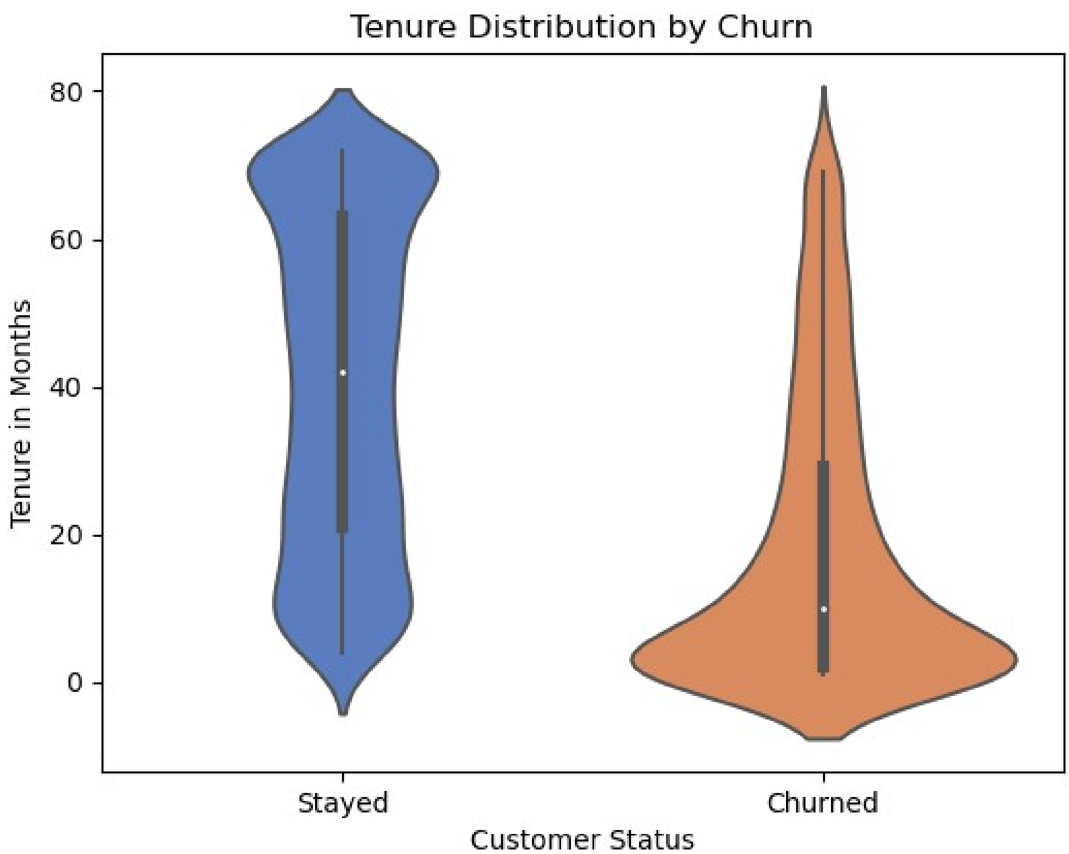
```
#2. Explore Demographic and Subscription-Based Factors
# Gender vs Churn
sns.countplot(x='Gender', hue='Customer Status', data=df,
palette='pastel')
plt.title('Churn by Gender')
plt.show()

# Married vs Churn
sns.countplot(x='Married', hue='Customer Status', data=df,
palette='coolwarm')
plt.title('Churn by Marital Status')
plt.show()

# Violin plot for tenure
```

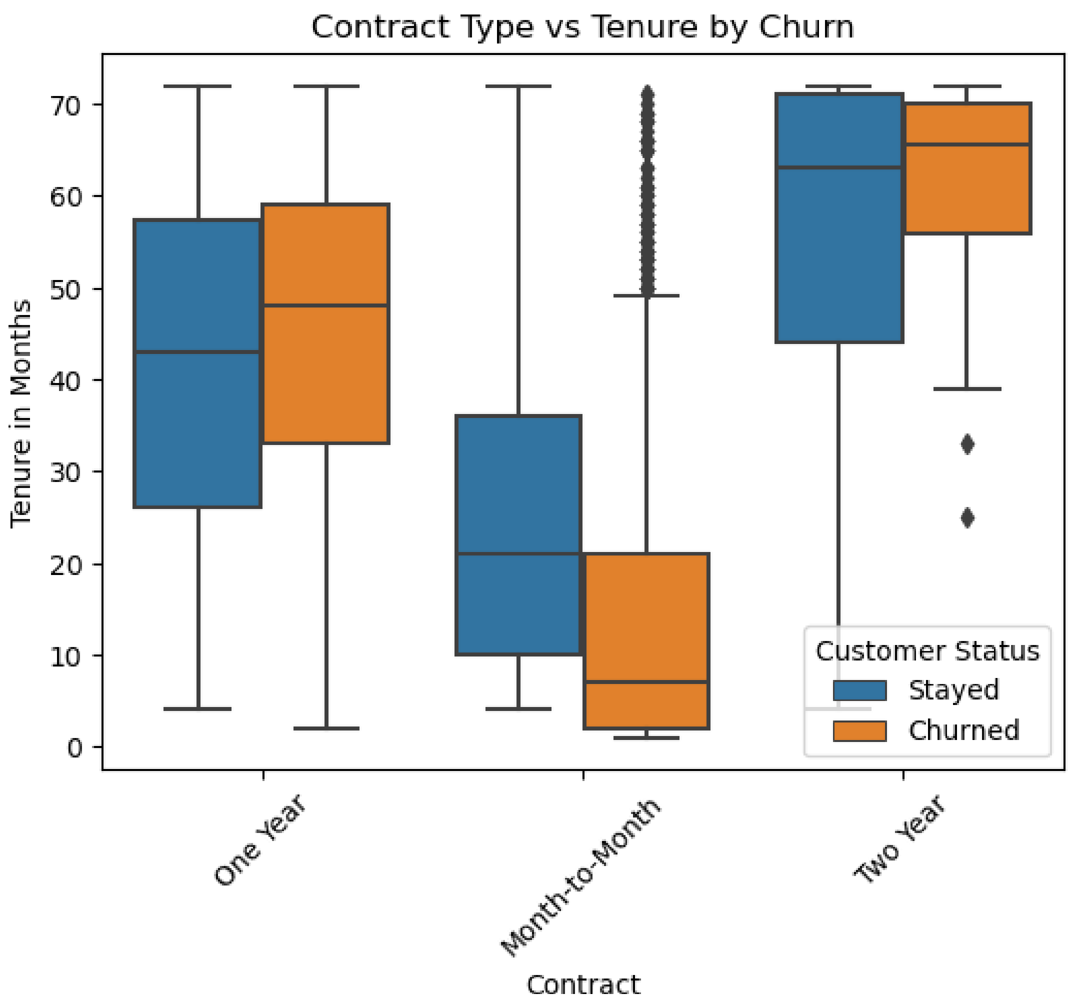
```
sns.violinplot(x='Customer Status', y='Tenure in Months', data=df,
palette='muted')
plt.title('Tenure Distribution by Churn')
plt.show()

# Boxplot for Contract types
sns.boxplot(x='Contract', y='Tenure in Months', hue='Customer Status',
data=df)
plt.xticks(rotation=45)
plt.title('Contract Type vs Tenure by Churn')
plt.show()
```





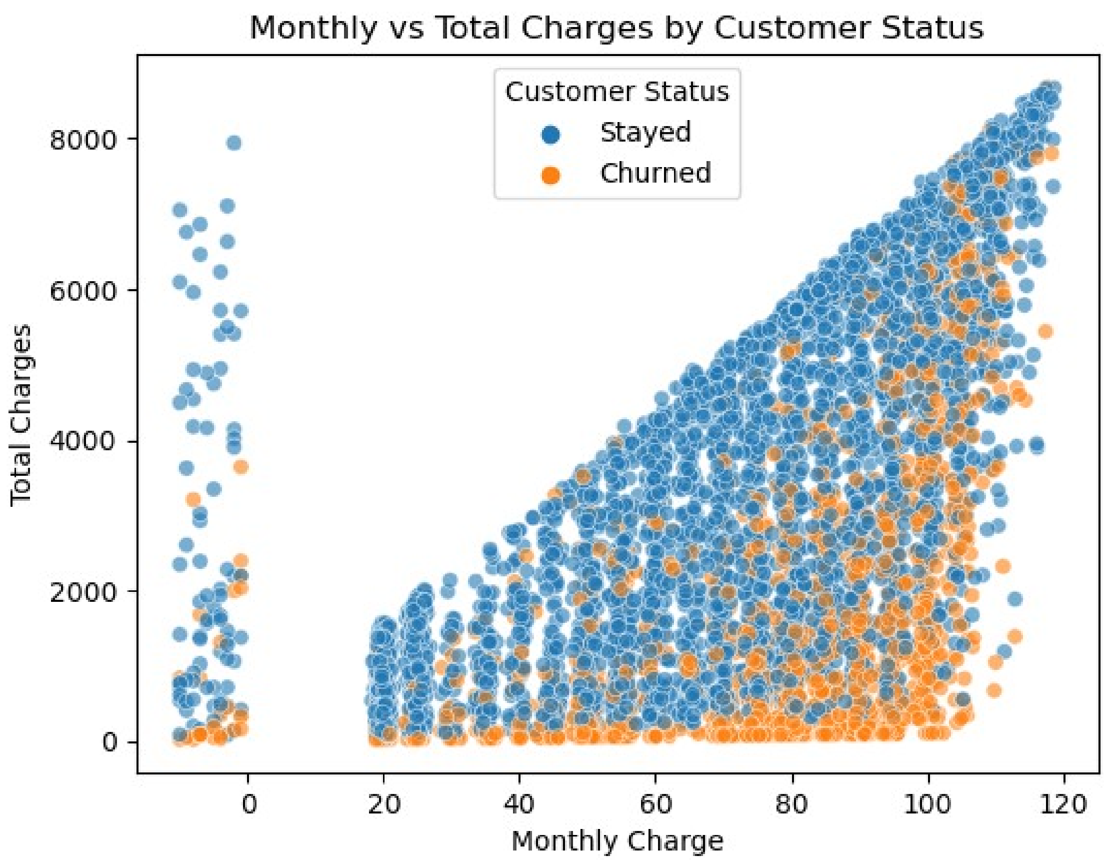
Churn by Marital Status
Customer Status
Stayed
Churned
count
2500
2000
1500
1000
500
0
Yes
No
Married

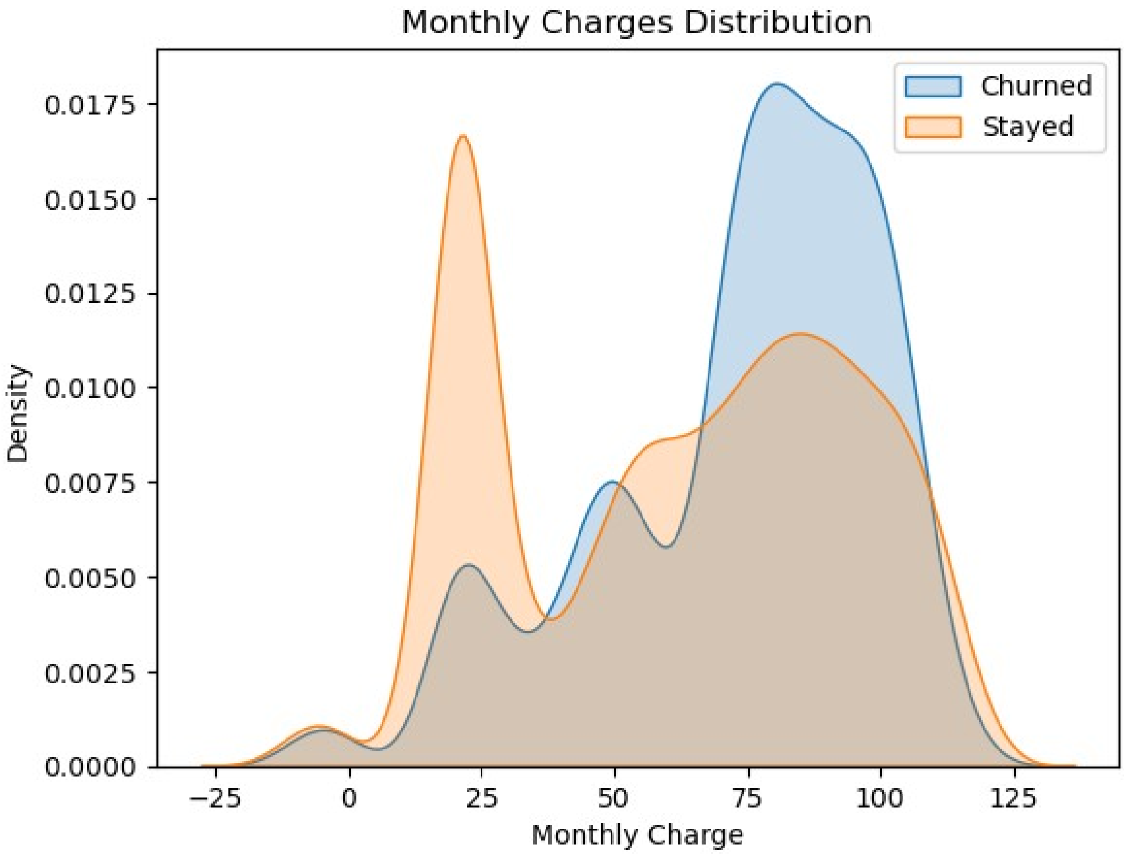

Tenure Distribution by Churn
Tenure in Months
80
60
40
20
0
Stayed
Churned
Customer Status



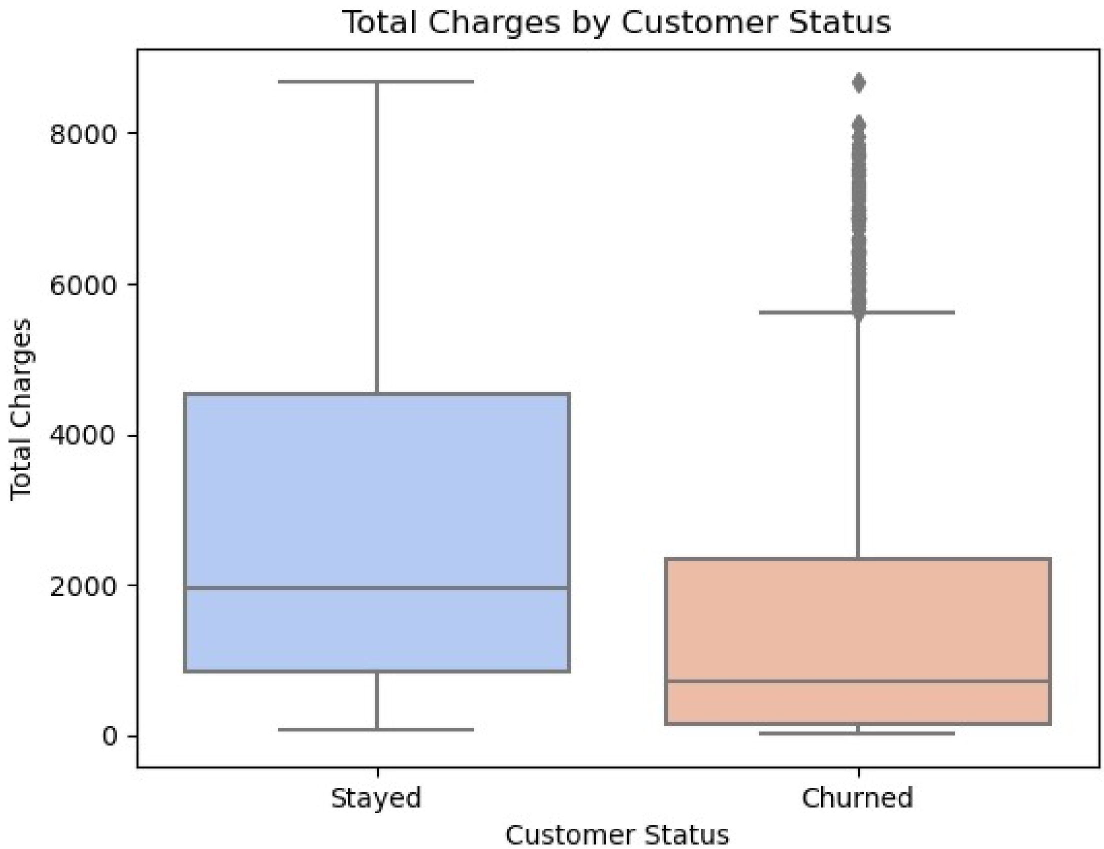

```
#3. Investigate Financial and Service Usage Patterns

# Drop rows with missing Total Charges for accurate scatter plot
df_clean = df.dropna(subset=['Total Charges'])

# Scatter plot: Monthly vs Total Charges
sns.scatterplot(x='Monthly Charge', y='Total Charges', hue='Customer
Status', data=df_clean, alpha=0.6)
plt.title('Monthly vs Total Charges by Customer Status')
plt.show()

# KDE plot: Monthly Charges
sns.kdeplot(data=df[df['Customer Status'] == 'Churned']['Monthly
Charge'], label='Churned', fill=True)
sns.kdeplot(data=df[df['Customer Status'] == 'Stayed']['Monthly
Charge'], label='Stayed', fill=True)
plt.title('Monthly Charges Distribution')
plt.legend()
plt.show()
```

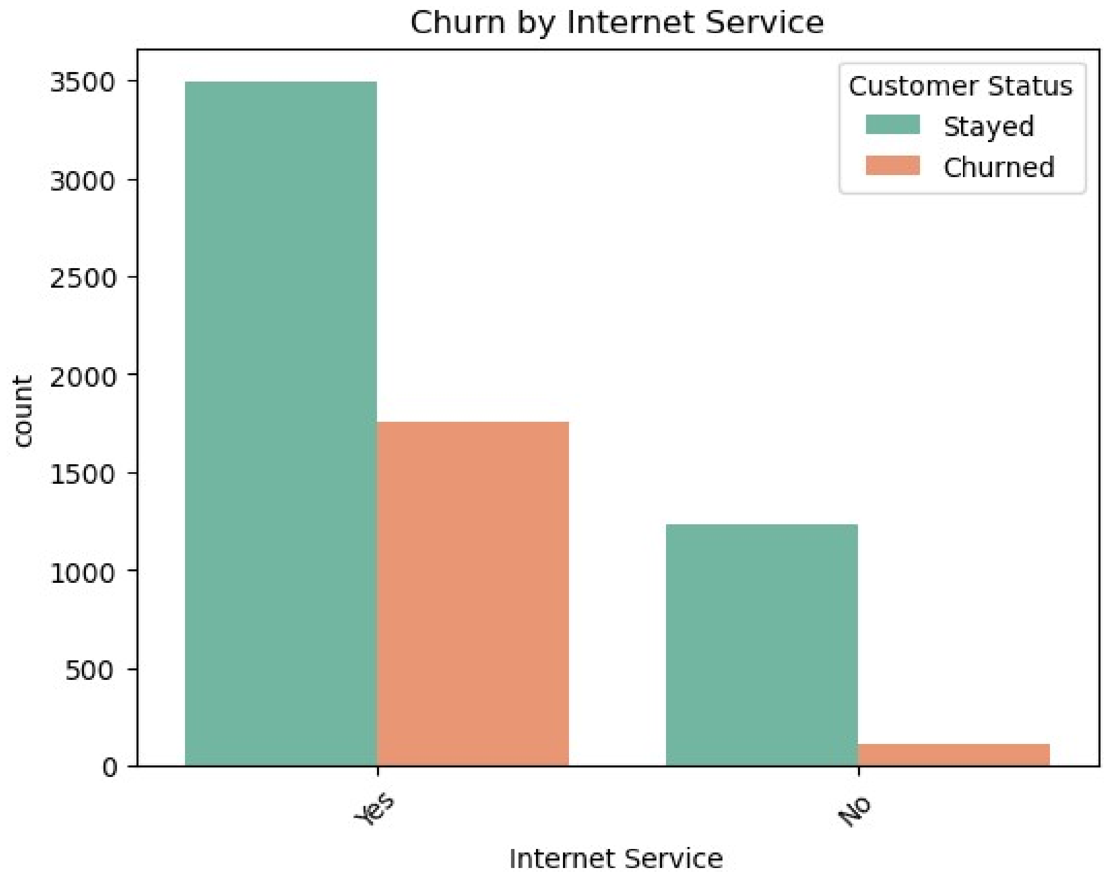
```
# Boxplot for Total Charges
sns.boxplot(x='Customer Status', y='Total Charges', data=df_clean,
palette='coolwarm')
plt.title('Total Charges by Customer Status')
plt.show()# Drop rows with missing Total Charges

# Churn by service features
service_features = ['Internet Service', 'Streaming TV', 'Streaming
Movies']
for feature in service_features:
    sns.countplot(x=feature, hue='Customer Status', data=df,
palette='Set2')
    plt.title(f'Churn by {feature}')
    plt.xticks(rotation=45)
    plt.show()
```

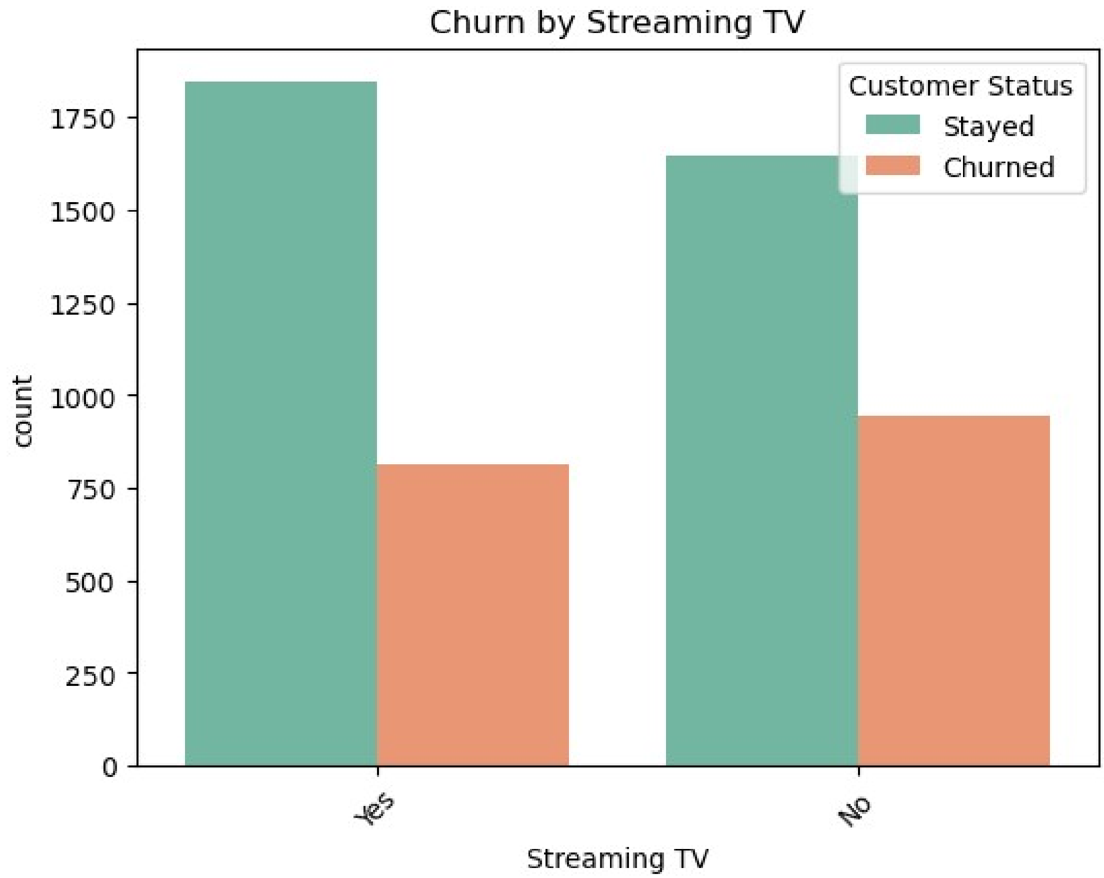


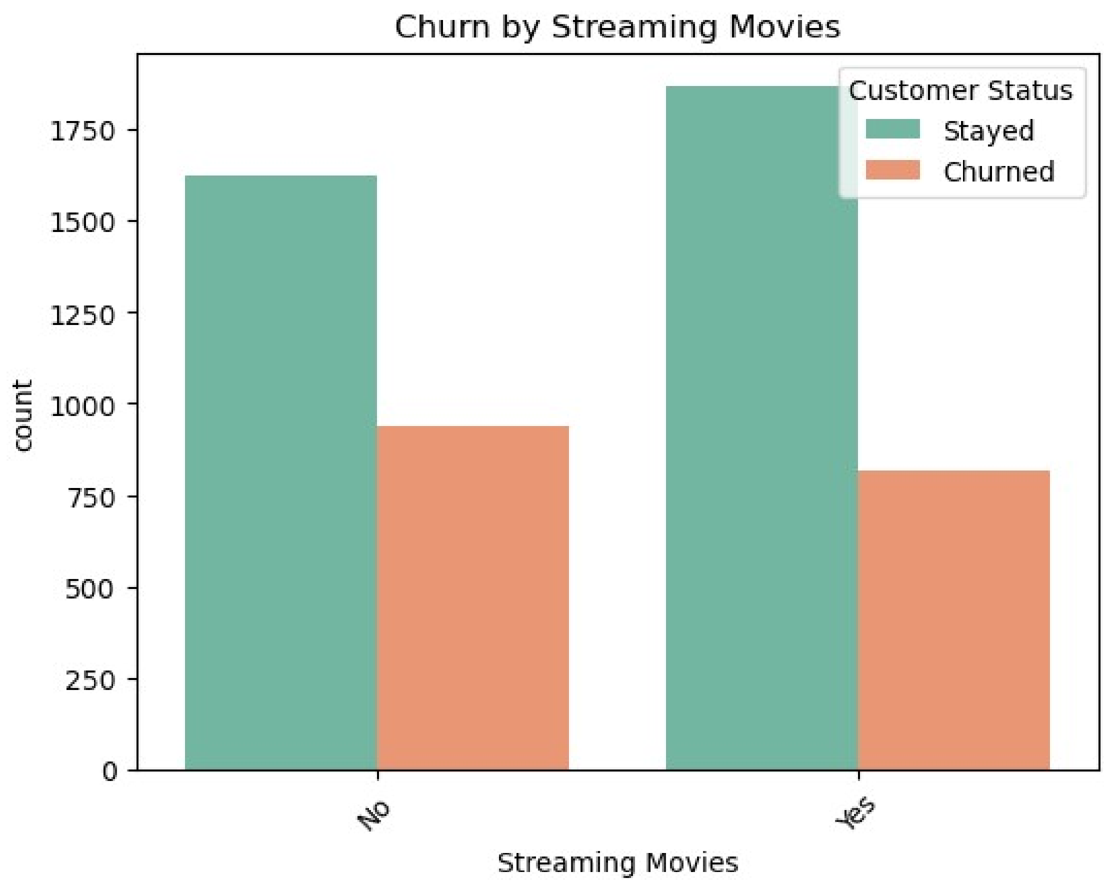


Monthly Charges Distribution
Churned
Stayed
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
−25
0
25
50
75
100
125
Monthly Charge


Total Charges by Customer Status
Total Charges
8000
6000
4000
2000
0
Stayed
Churned
Customer Status

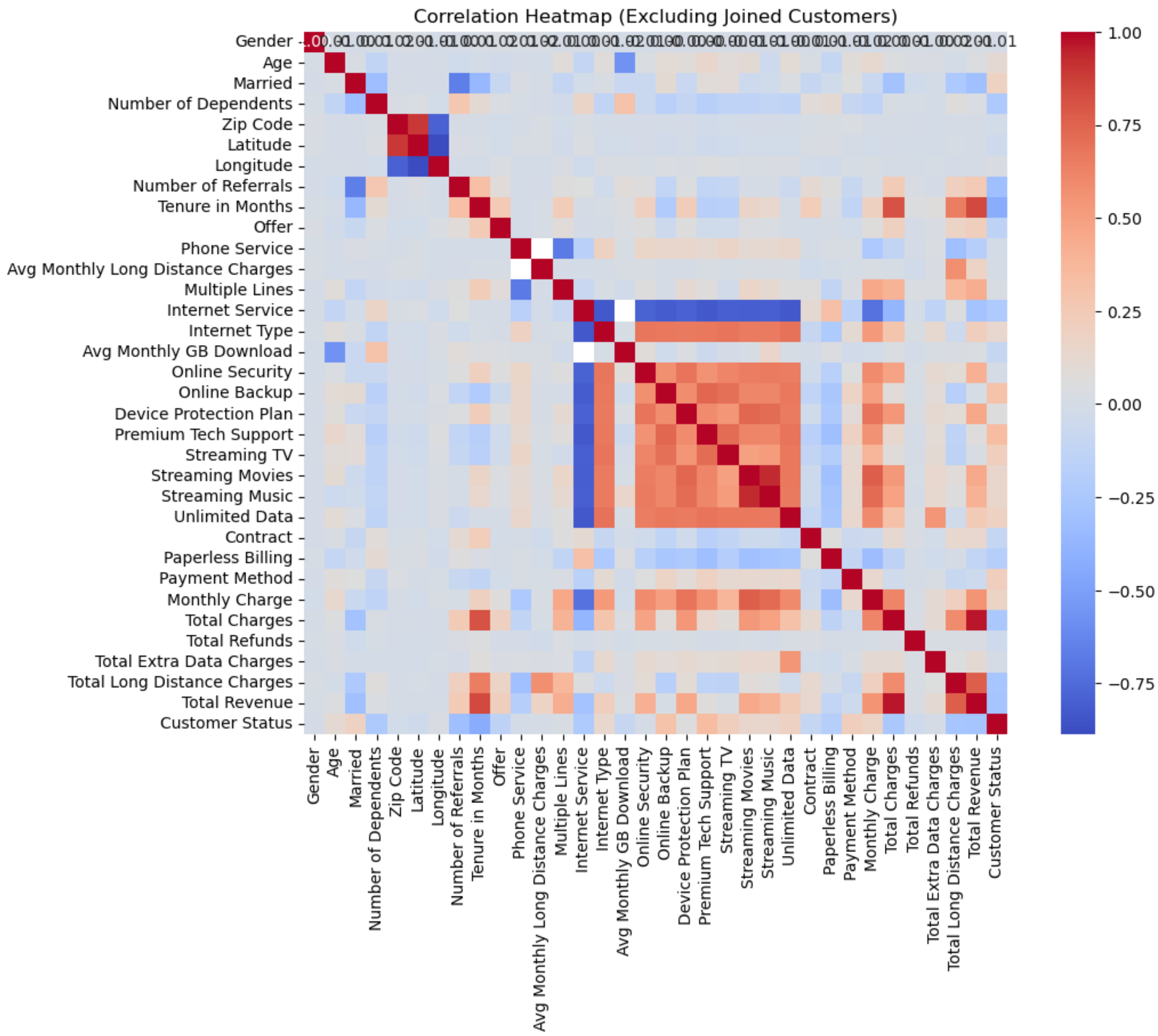

Churn by Internet Service
Customer Status
Stayed
Churned
count
3500
3000
2500
2000
1500
1000
500
0
Yes
No
Internet Service


Churn by Streaming TV
Customer Status
Stayed
Churned
count
1750
1500
1250
1000
750
500
250
0
Yes
No
Streaming TV




```
#4. Check Correlations Between Features

# Encode categorical columns
df_encoded = df.copy()
# Drop irrelevant columns
df_encoded = df.drop(['Customer ID', 'City', 'Churn Reason', 'Churn 
Category'], axis=1, errors='ignore')

# Encode target column (Customer Status): Churned = 1, Stayed = 0
df_encoded['Customer Status'] = df_encoded['Customer 
Status'].map({'Churned': 1, 'Stayed': 0})

# Encode remaining categorical features using pd.factorize
for col in df_encoded.select_dtypes(include='object'):
    df_encoded[col] = pd.factorize(df_encoded[col])[0]

# Compute correlation matrix
corr_matrix = df_encoded.corr()

# Heatmap visualization
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm', 
square=True)
```

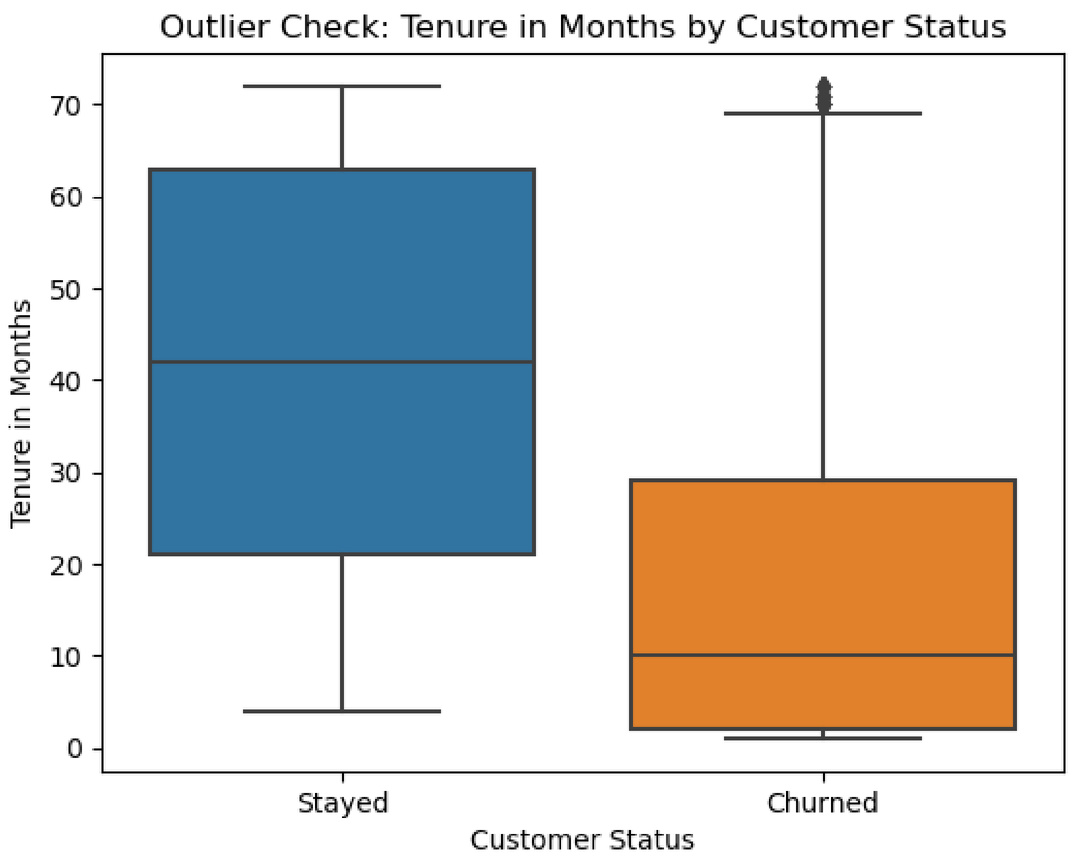
```python
plt.title('Correlation Heatmap (Excluding Joined Customers)')
plt.show()

# Print top correlations with Churn
print("\n Top correlations with 'Customer Status':\n")
print(corr_matrix['Customer Status'].sort_values(ascending=False))
```



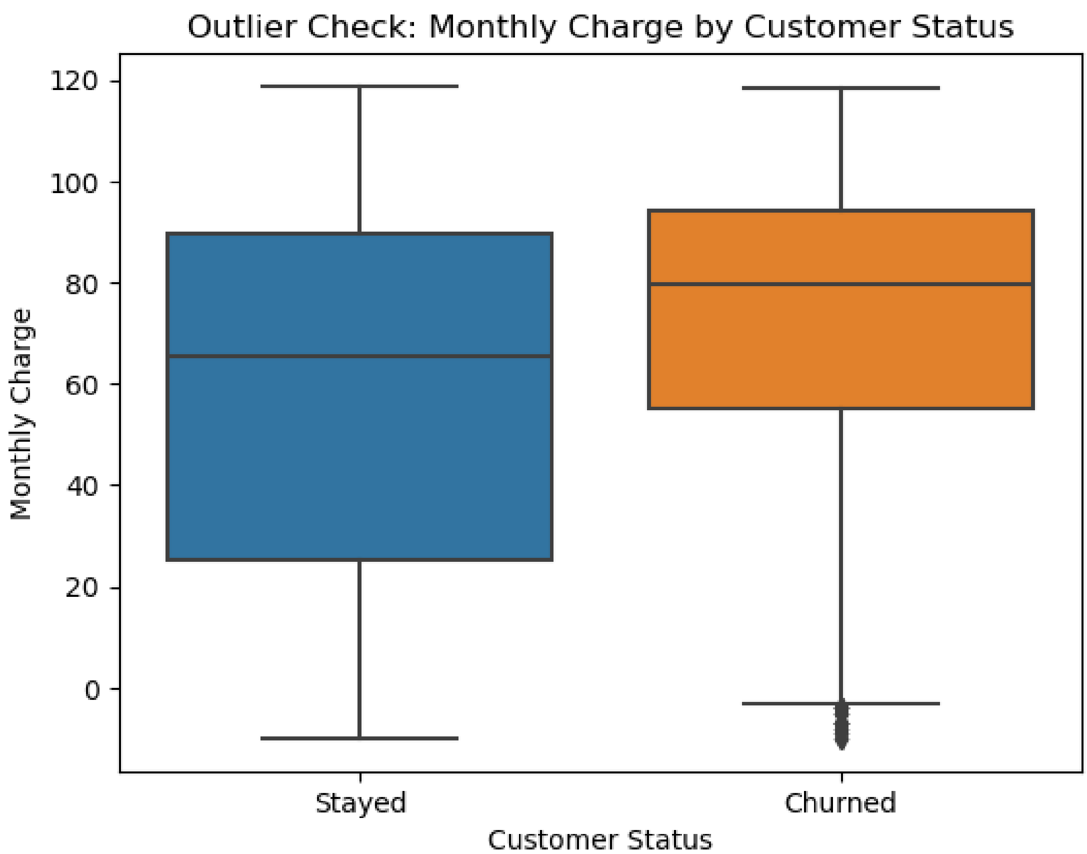

```
 Top correlations with 'Customer Status':

Customer Status               1.000000
Premium Tech Support          0.340860
Online Backup                 0.302242
Streaming TV                  0.214154
Payment Method                0.213091
Unlimited Data                0.193174
```

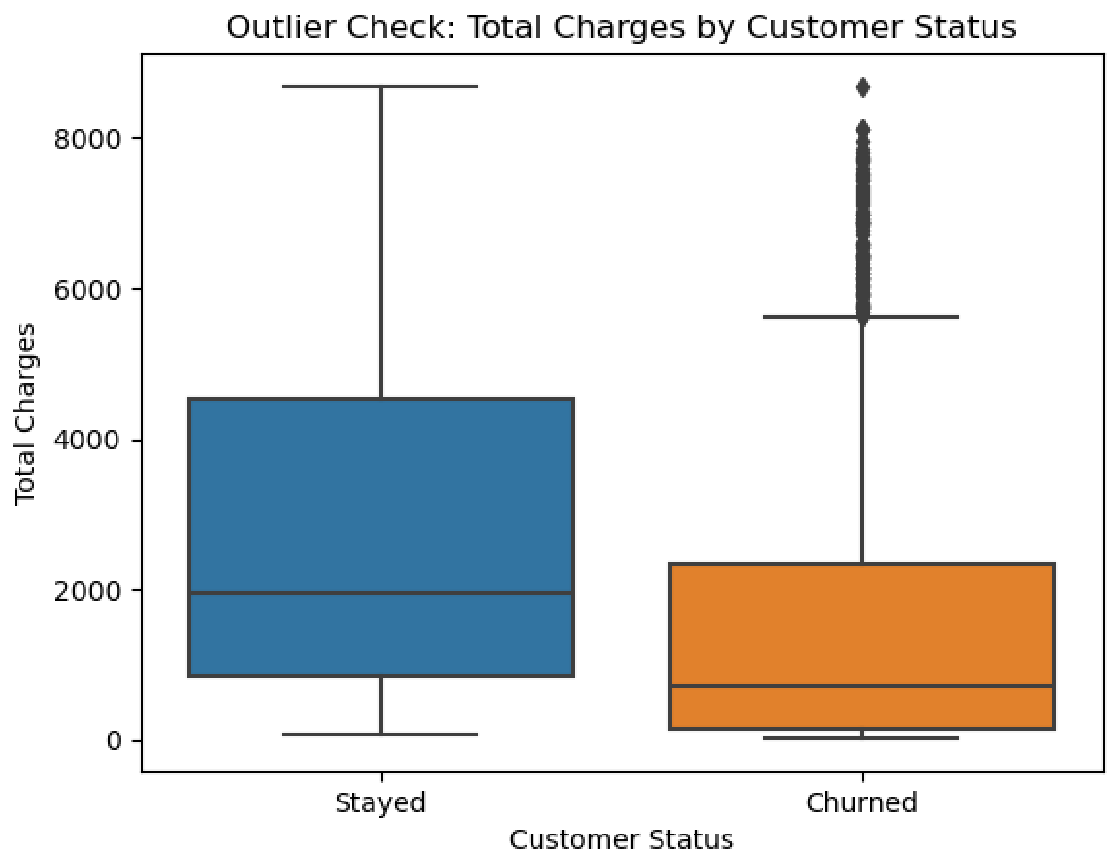
```
Married                                    0.183273
Monthly Charge                             0.168290
Streaming Movies                           0.145033
Streaming Music                            0.139058
Internet Type                              0.134057
Age                                        0.111174
Device Protection Plan                     0.061903
Longitude                                  0.025455
Multiple Lines                             0.016951
Online Security                            0.001715
Total Extra Data Charges                  -0.000259
Avg Monthly Long Distance Charges         -0.000467
Gender                                    -0.006373
Phone Service                             -0.014369
Zip Code                                  -0.018888
Total Refunds                             -0.043525
Latitude                                  -0.044023
Avg Monthly GB Download                   -0.095132
Contract                                  -0.100288
Offer                                     -0.147170
Paperless Billing                         -0.187702
Internet Service                          -0.224121
Number of Dependents                      -0.232525
Total Charges                             -0.250071
Total Long Distance Charges               -0.268430
Total Revenue                             -0.278626
Number of Referrals                       -0.312118
Tenure in Months                          -0.433759
Name: Customer Status, dtype: float64
```

```
#5. Detect Anomalies and Outliers in Customer Behavior

# Boxplots for outlier detection
num_cols = ['Tenure in Months', 'Monthly Charge', 'Total Charges']

# Box plots to check outliers visually
for col in num_cols:
    sns.boxplot(x='Customer Status', y=col, data=df)
    plt.title(f'Outlier Check: {col} by Customer Status')
    plt.show()

# IQR method to identify outliers numerically
for col in num_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    outliers = df[(df[col] < Q1 - 1.5 * IQR) | (df[col] > Q3 + 1.5 *
IQR)]
    print(f"{col}: Found {len(outliers)} potential outliers.")
```



Outlier Check: Tenure in Months by Customer Status
Tenure in Months
70
60
50
40
30
20
10
0
Stayed
Churned
Customer Status


Outlier Check: Monthly Charge by Customer Status
Monthly Charge
120
100
80
60
40
20
0
Stayed
Churned
Customer Status




```
Tenure in Months: Found 0 potential outliers.
Monthly Charge: Found 0 potential outliers.
Total Charges: Found 0 potential outliers.
```

## Key Takeaways

- Customers with shorter tenure and higher charges are more likely to churn.
- Streaming services and contract types play a major role in customer behavior.
- Month-to-month contracts show higher churn.

### Next Steps:

- Build a churn prediction model (Logistic Regression / Random Forest)
- Optimize retention strategies for high-risk customers

Thanks for reading!