In [2]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from math import sqrt
from pathlib import Path
```

In [3]:

```python
df = pd.read_csv('Telecom_churn_data.csv',na_values=[" "])
```

In [4]:

```python
df
```

Out[4]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes
...	...	...	...	...	...	...	...
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes
7040	4801-JZAZL	Female	0	Yes	Yes	11	No
7041	8361-LTMKD	Male	1	Yes	No	4	Yes
7042	3186-AJIEK	Male	0	No	No	66	Yes

7043 rows × 21 columns

In [5]:

```python
yn_cols = [
    "Partner","Dependents","PhoneService","PaperlessBilling","Churn",
    "MultipleLines","OnlineSecurity","OnlineBackup","DeviceProtection",
    "TechSupport","StreamingTV","StreamingMovies"
]
```

In [6]:

```python
for c in yn_cols:
    if c in df.columns:
        df[c] = df[c].replace({"No internet service":"No","No phone service"
```

In [7]:
```
df["SeniorCitizen"] = pd.to_numeric(df["SeniorCitizen"], errors="coerce").fi
```

In [8]:
```
for col in ["tenure","MonthlyCharges","TotalCharges"]:
    if col in df.columns:
        df[col] = pd.to_numeric(df[col], errors="coerce")
```

In [9]:
```
df["TotalCharges"] = df["TotalCharges"].fillna(df["TotalCharges"].median())
```

In [10]:
```
df["ChurnFlag"] = (df["Churn"].str.lower() == "yes").astype(int)
```

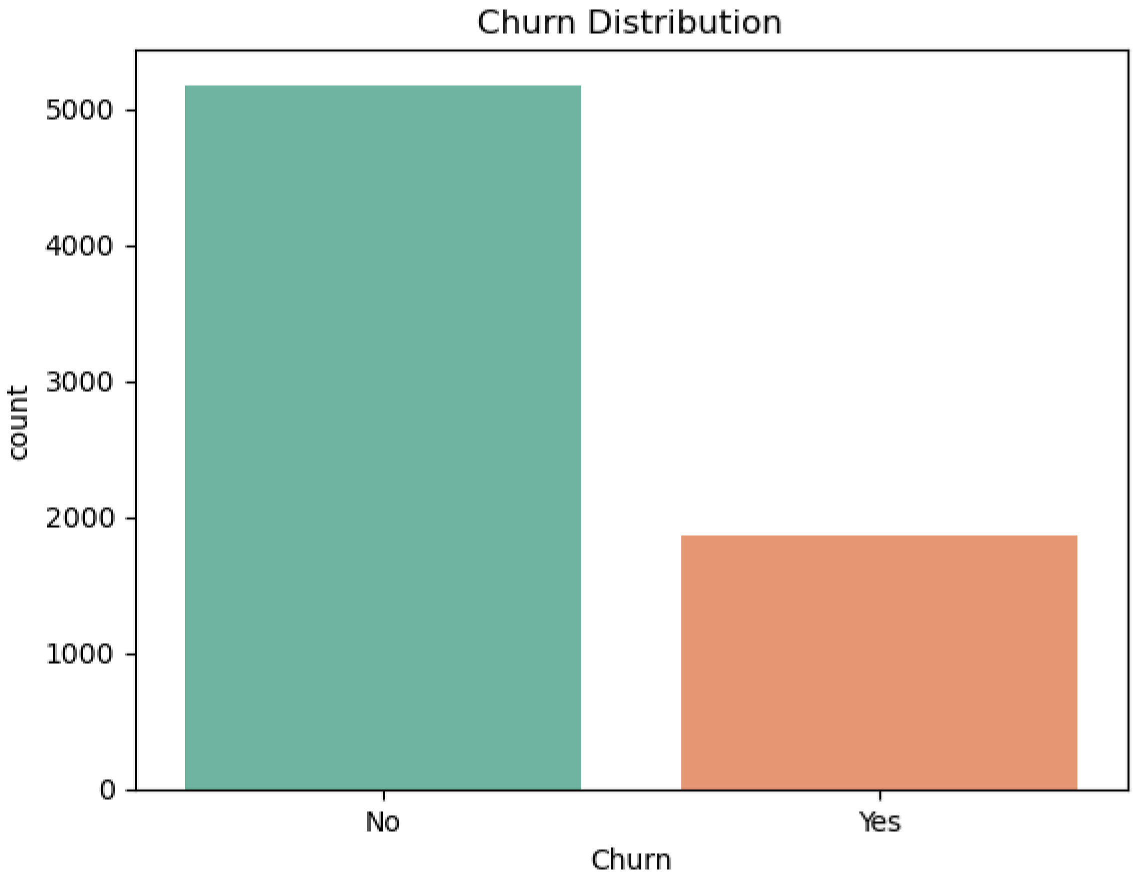
In [11]:
```
print("\nNull counts after cleaning:\n", df.isna().sum()) #checking counts c
```

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

In [12]:
```
print("\nOverall churn rate (%):", round(df["ChurnFlag"].mean()*100,2))
```

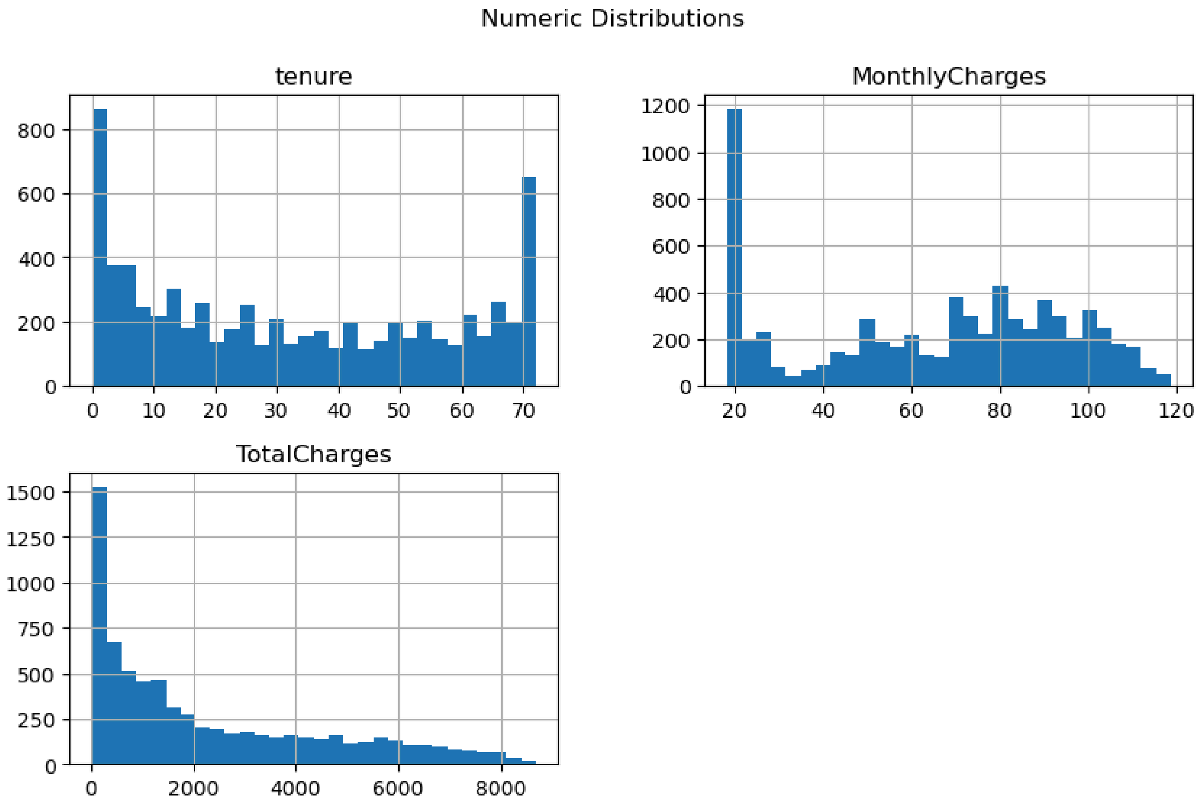
```
Overall churn rate (%): 26.54
```

In [13]:
```
plt.figure()
sns.countplot(x="Churn", data=df, palette="Set2", hue = 'Churn')
plt.title("Churn Distribution")
plt.show()
```



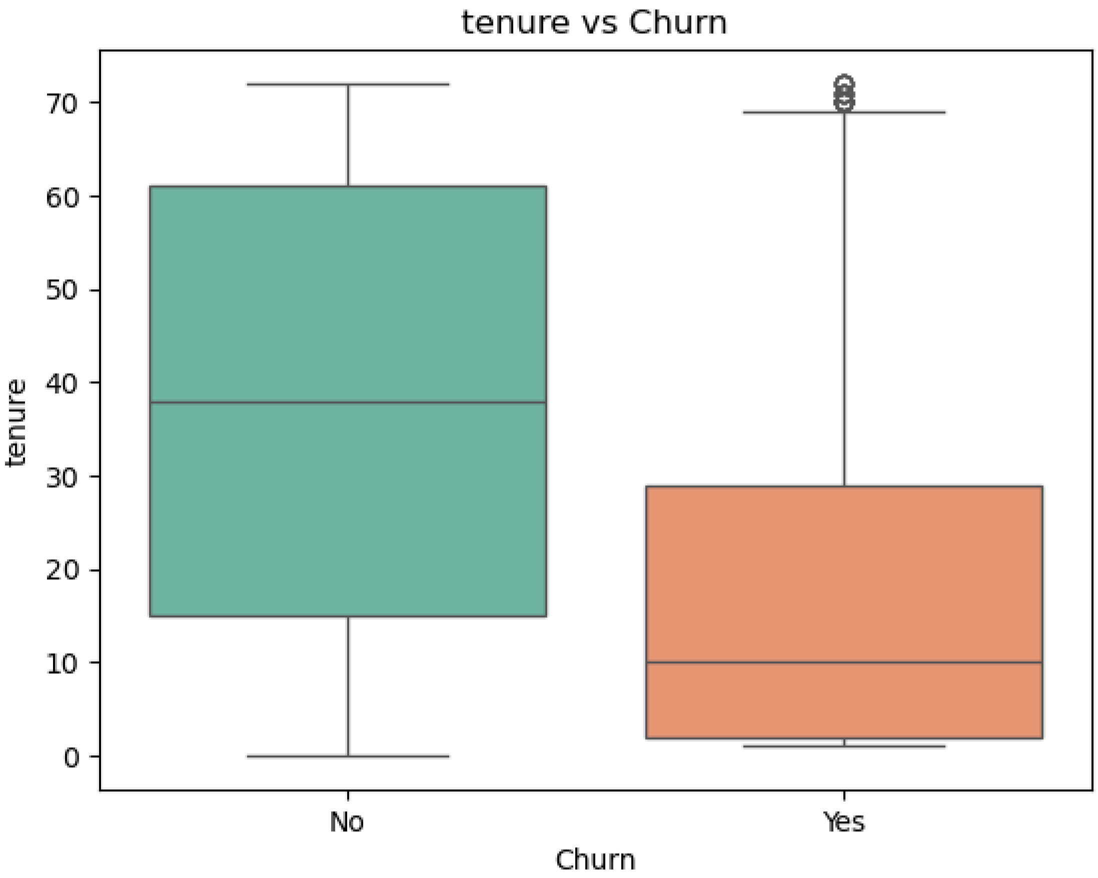
In [14]:

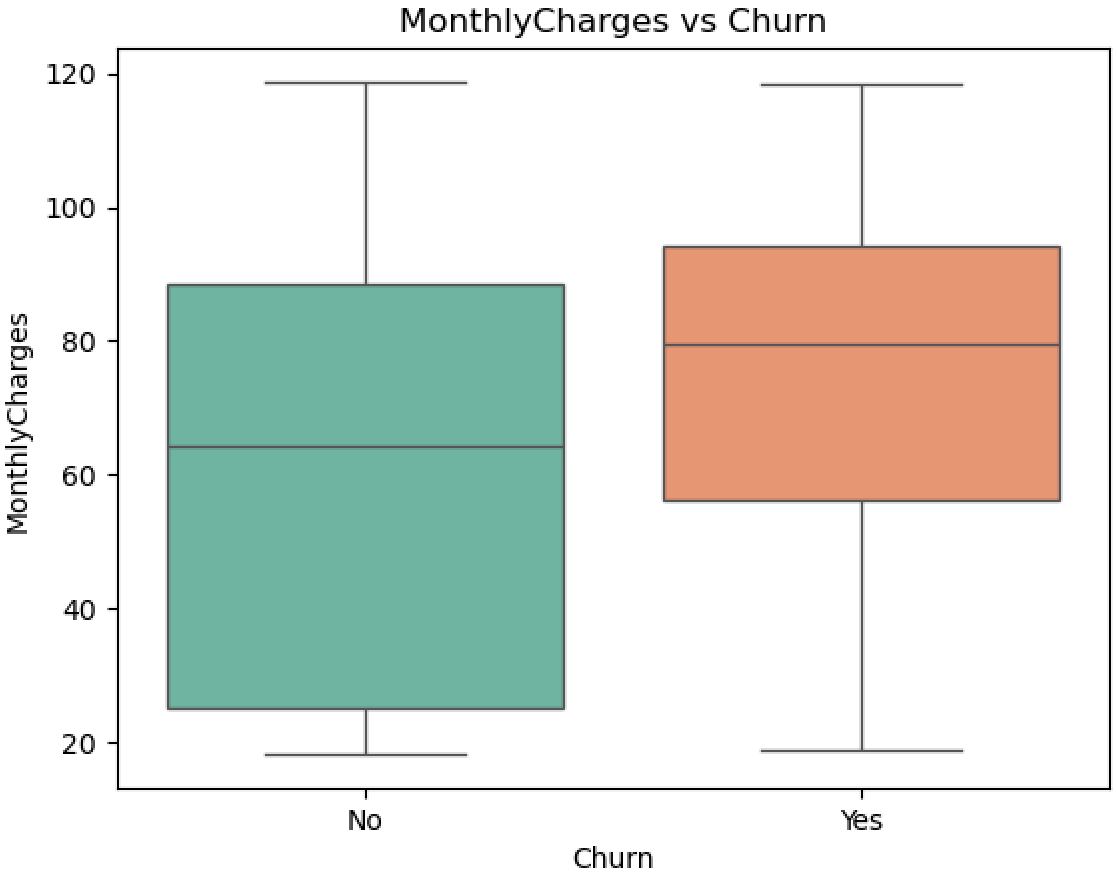
```
num_cols = ["tenure","MonthlyCharges","TotalCharges"]
df[num_cols].hist(bins=30, figsize=(10,6))
plt.suptitle("Numeric Distributions")
plt.show()
```

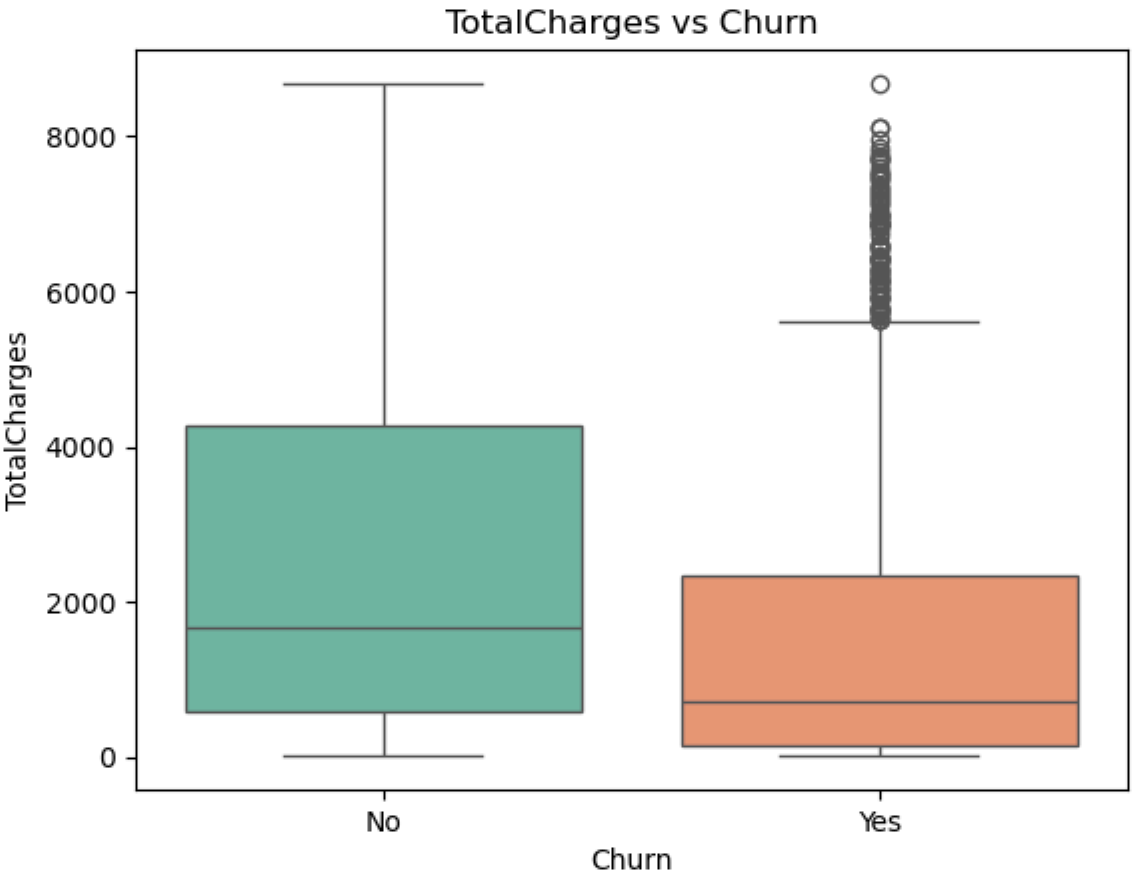


In [14]:

```python
for col in num_cols:
    plt.figure()
    sns.boxplot(x="Churn", y=col, data=df, palette="Set2", hue = "Churn")
    plt.title(f"{col} vs Churn")
    plt.show()
```


tenure vs Churn
tenure
70
60
50
40
30
20
10
0
No
Yes
Churn

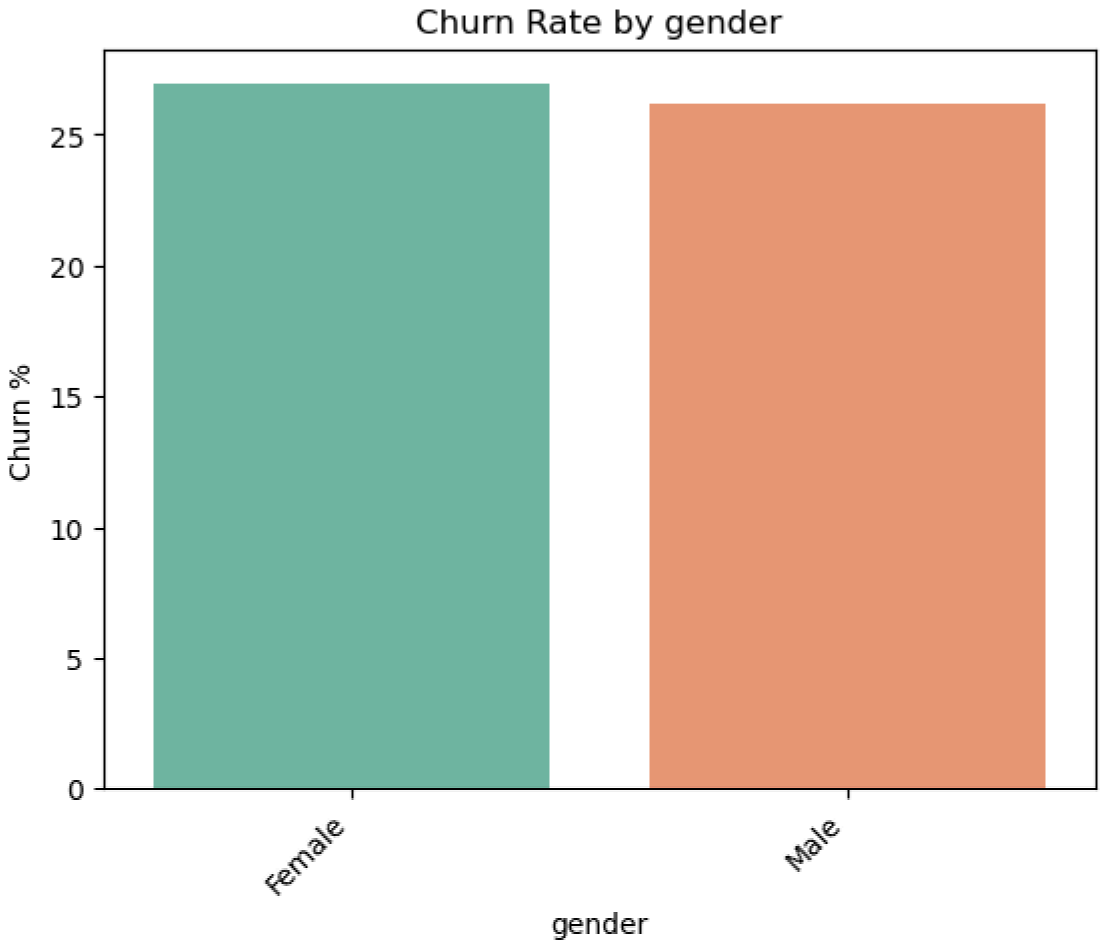

MonthlyCharges vs Churn
MonthlyCharges
120
100
80
60
40
20
No
Yes
Churn

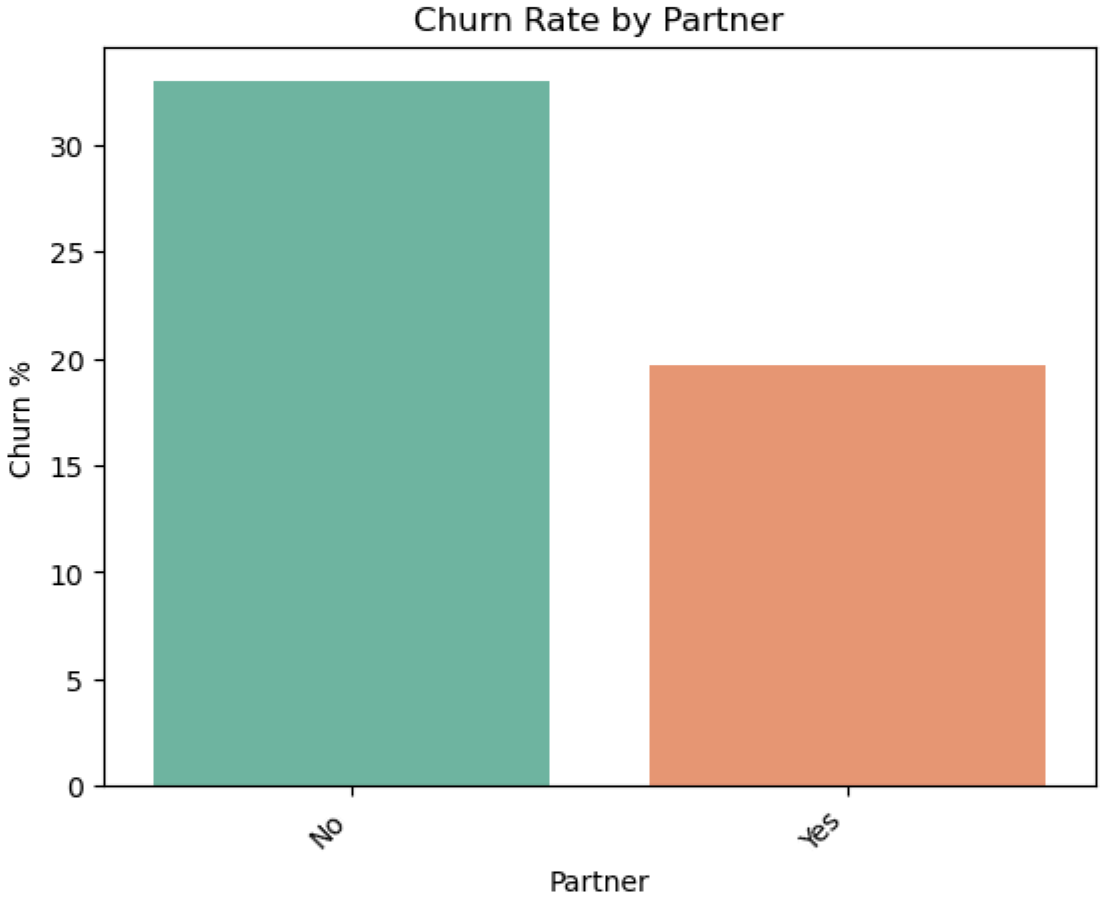


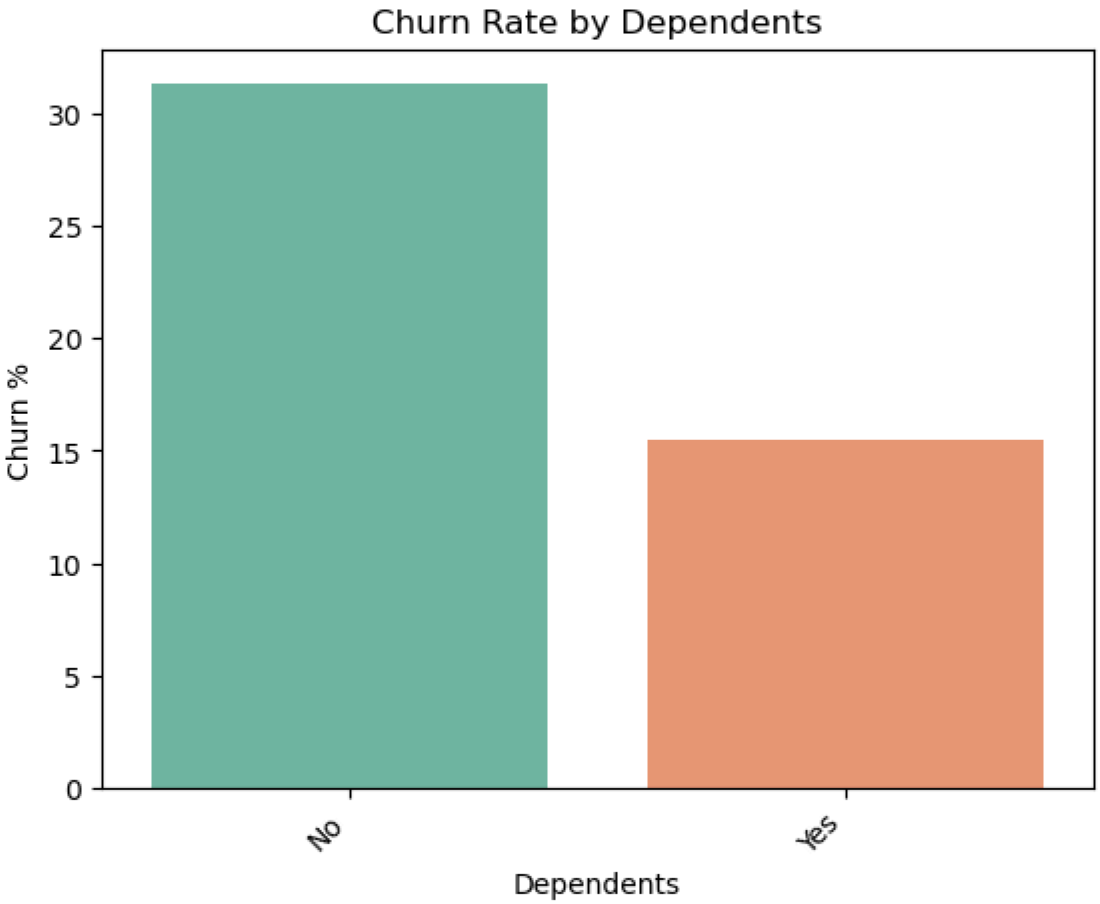
In [15]:
```
def churn_rate_by(df, col):
    return (df.groupby(col)["ChurnFlag"].mean()*100).sort_values(ascending=F
```

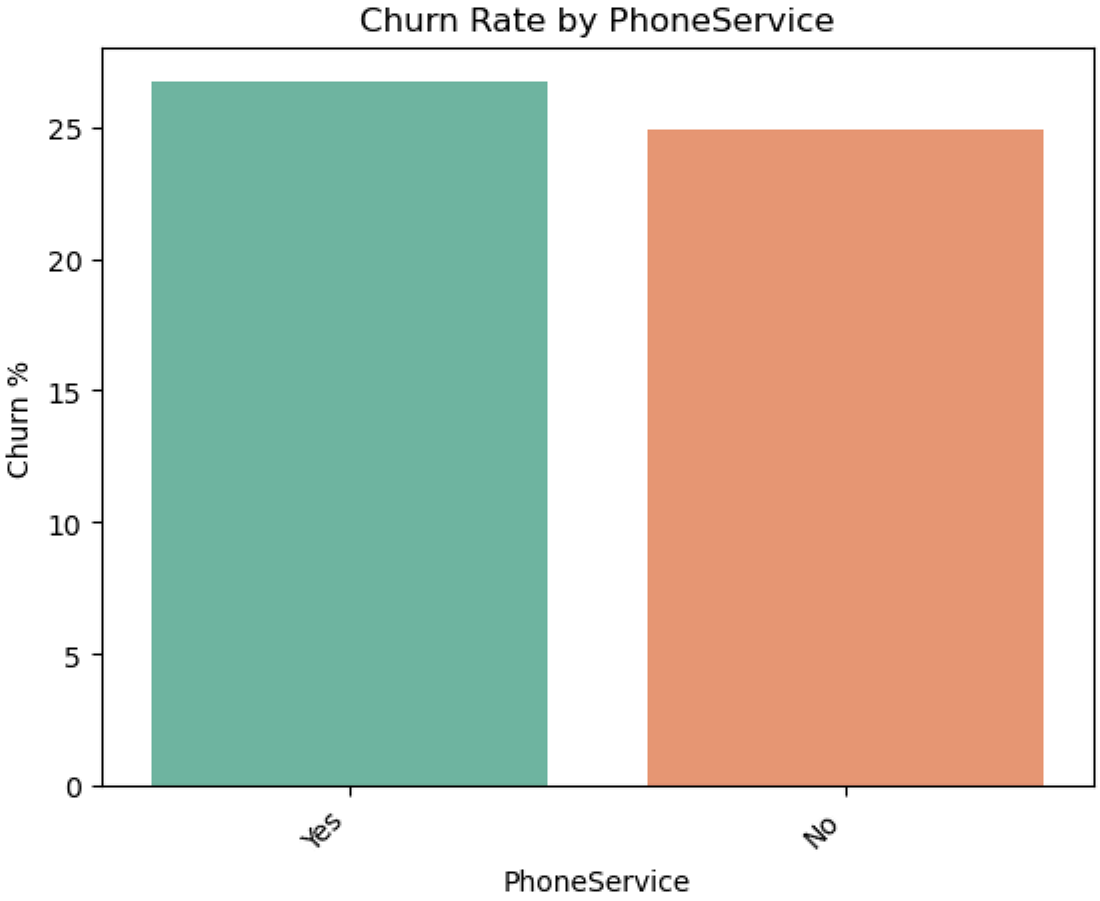
In [16]:
```
cat_cols = [c for c in df.columns if df[c].dtype=="object" and c not in ["cu
```

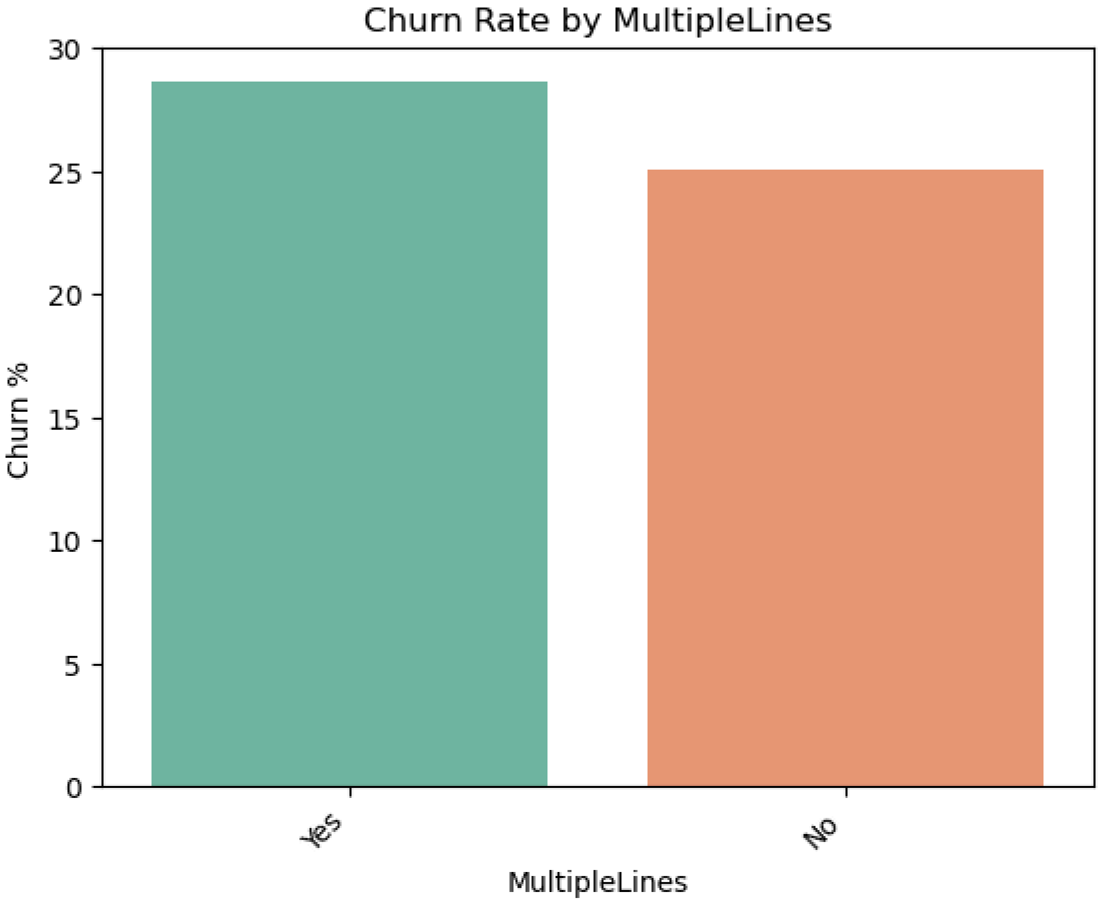
In [17]:
```
for c in cat_cols:
    rates = churn_rate_by(df, c)
    plt.figure()
    sns.barplot(x=rates.index, y=rates.values, hue=rates.index, palette="Set
    plt.title(f"Churn Rate by {c}")
    plt.ylabel("Churn %")
    plt.xticks(rotation=45, ha="right")
    plt.show()
```

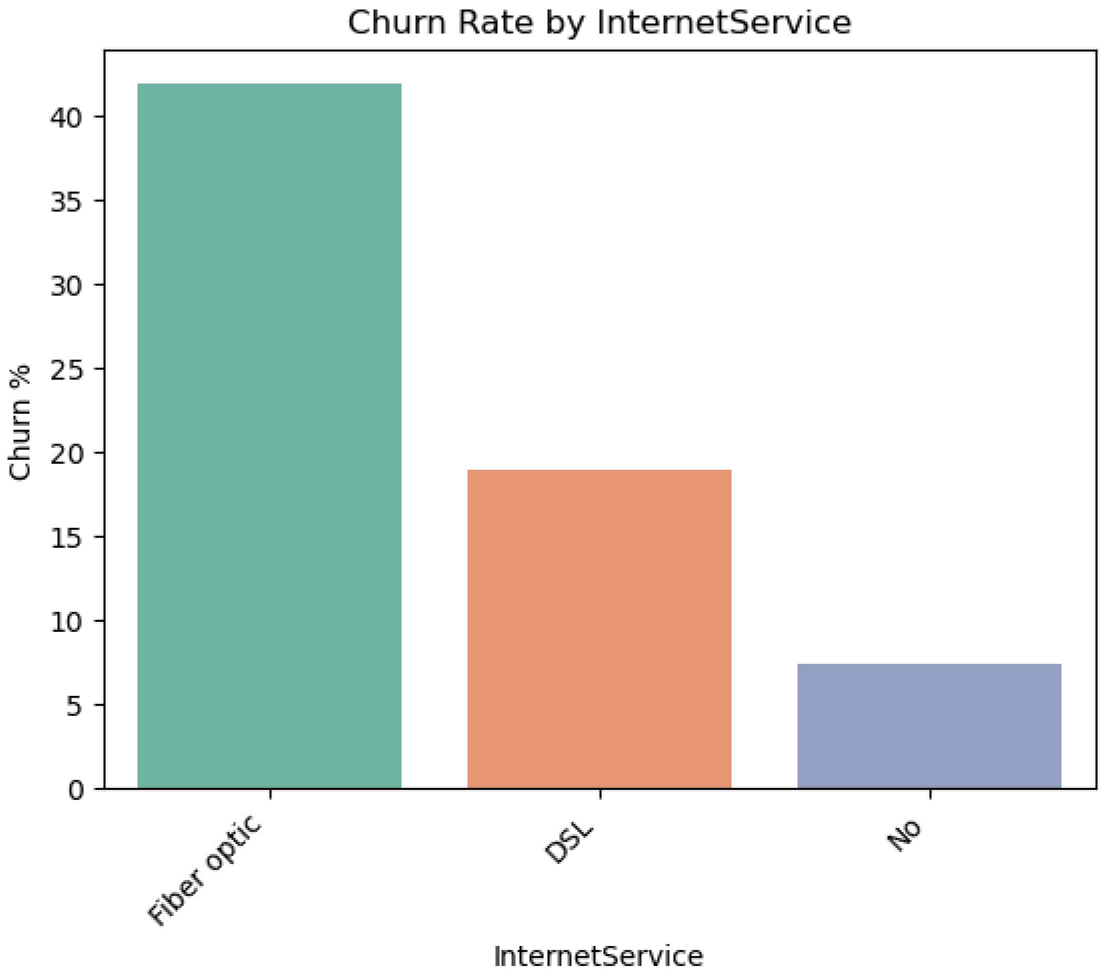

Churn Rate by gender
Churn %
0
5
10
15
20
25
Female
Male
gender

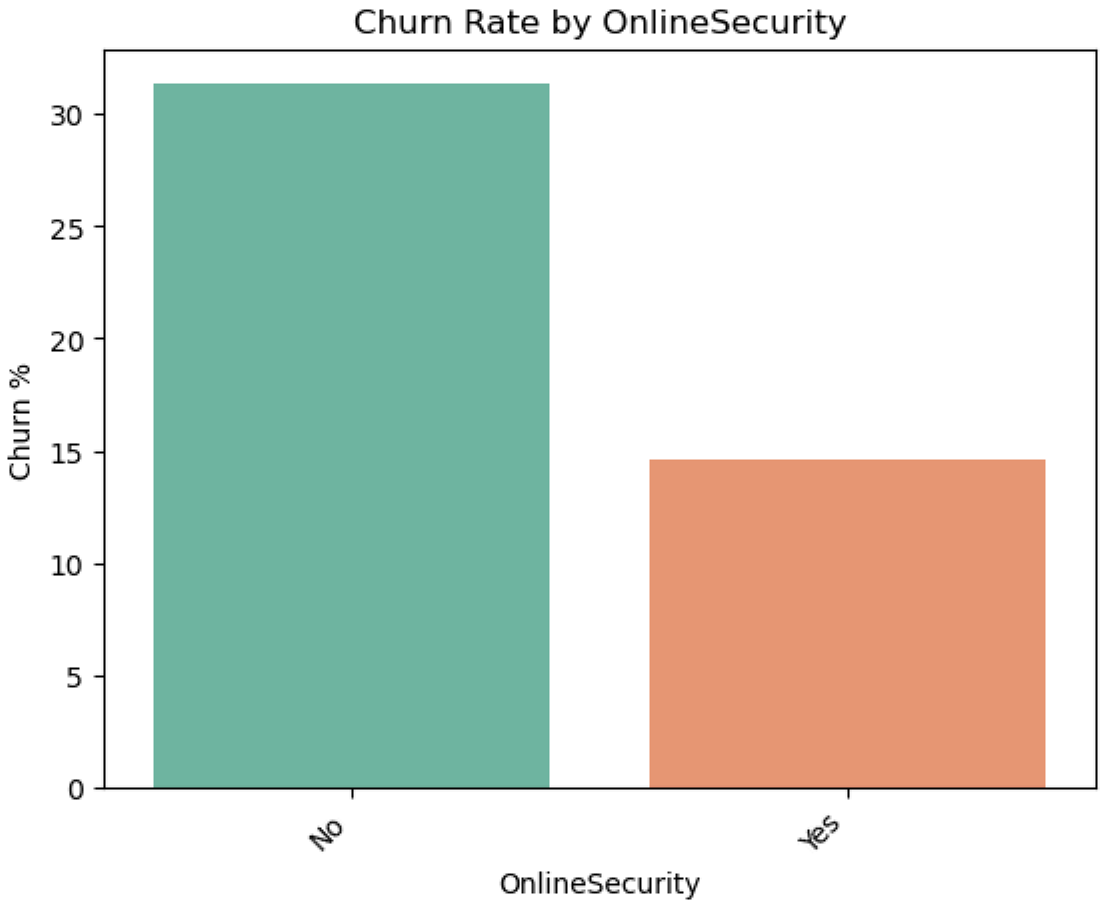

Churn Rate by Partner
Churn %
0
5
10
15
20
25
30
No
Yes
Partner

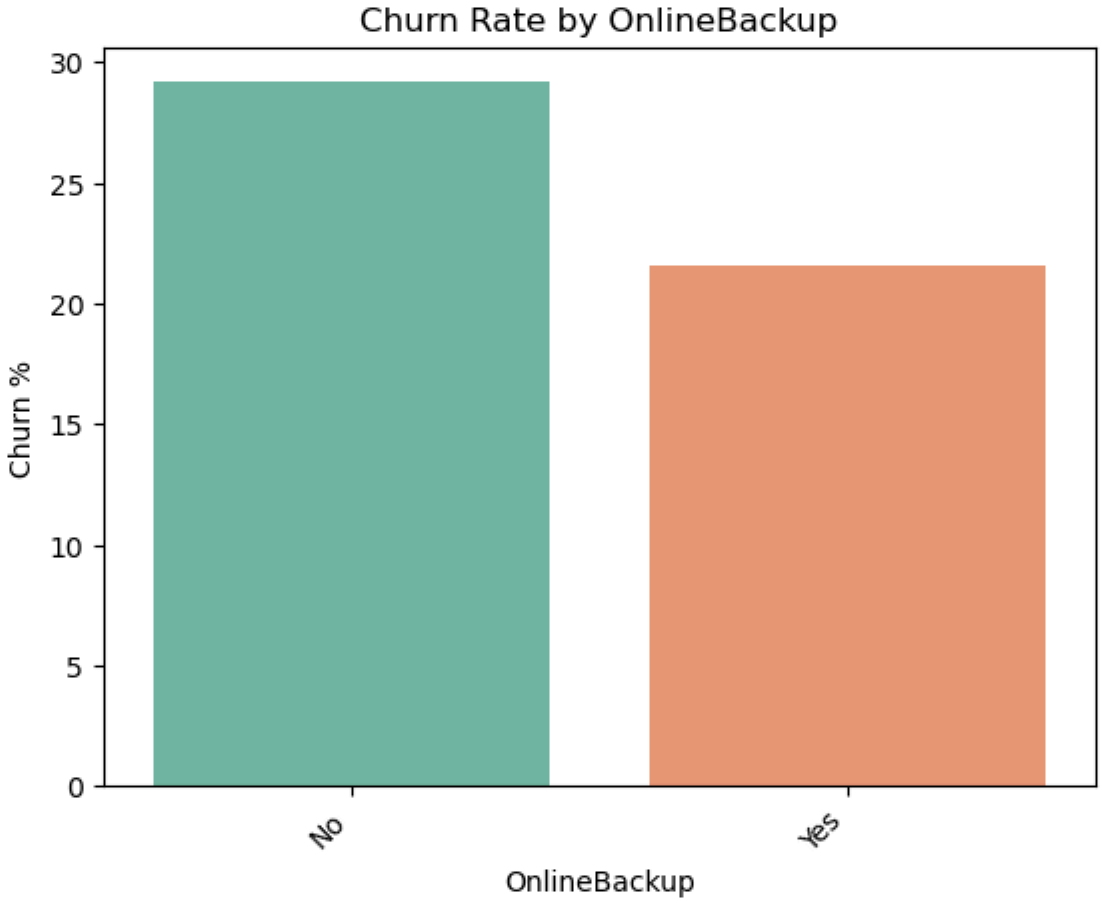

Churn Rate by Dependents

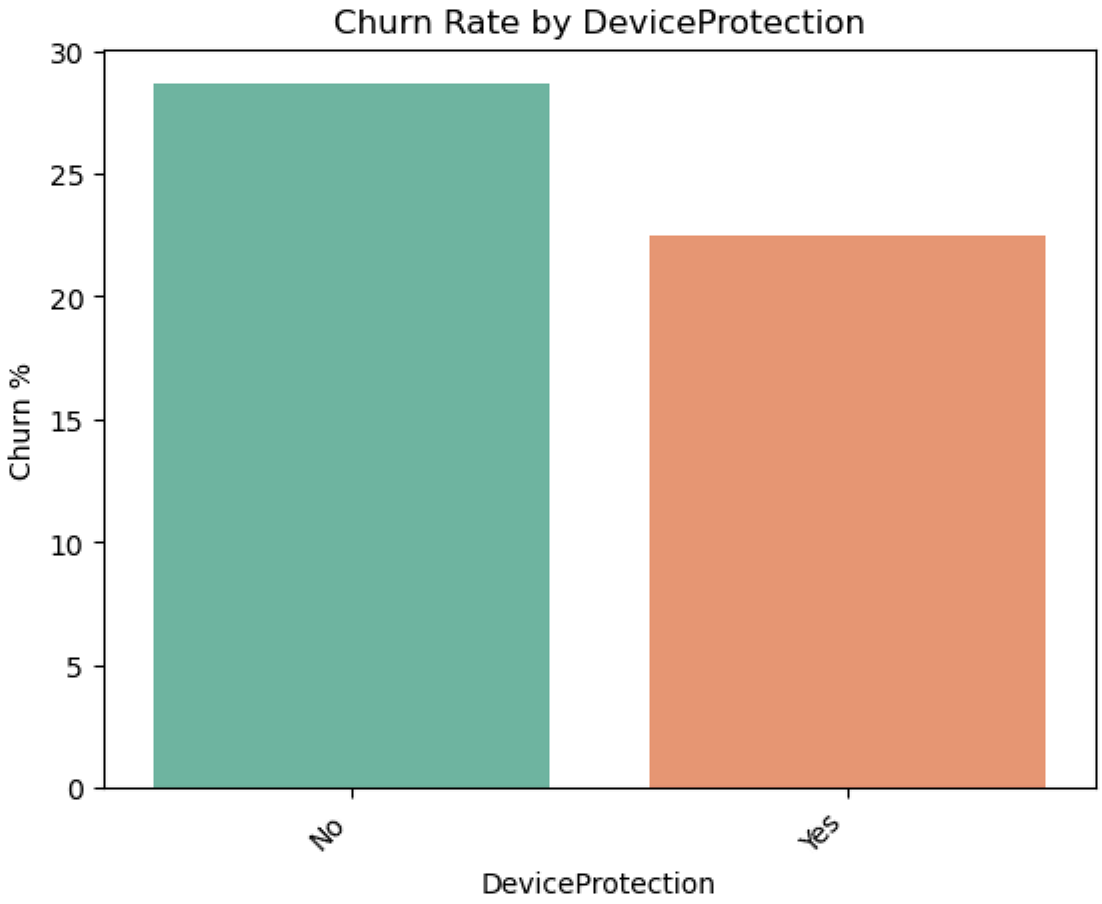

Churn Rate by PhoneService

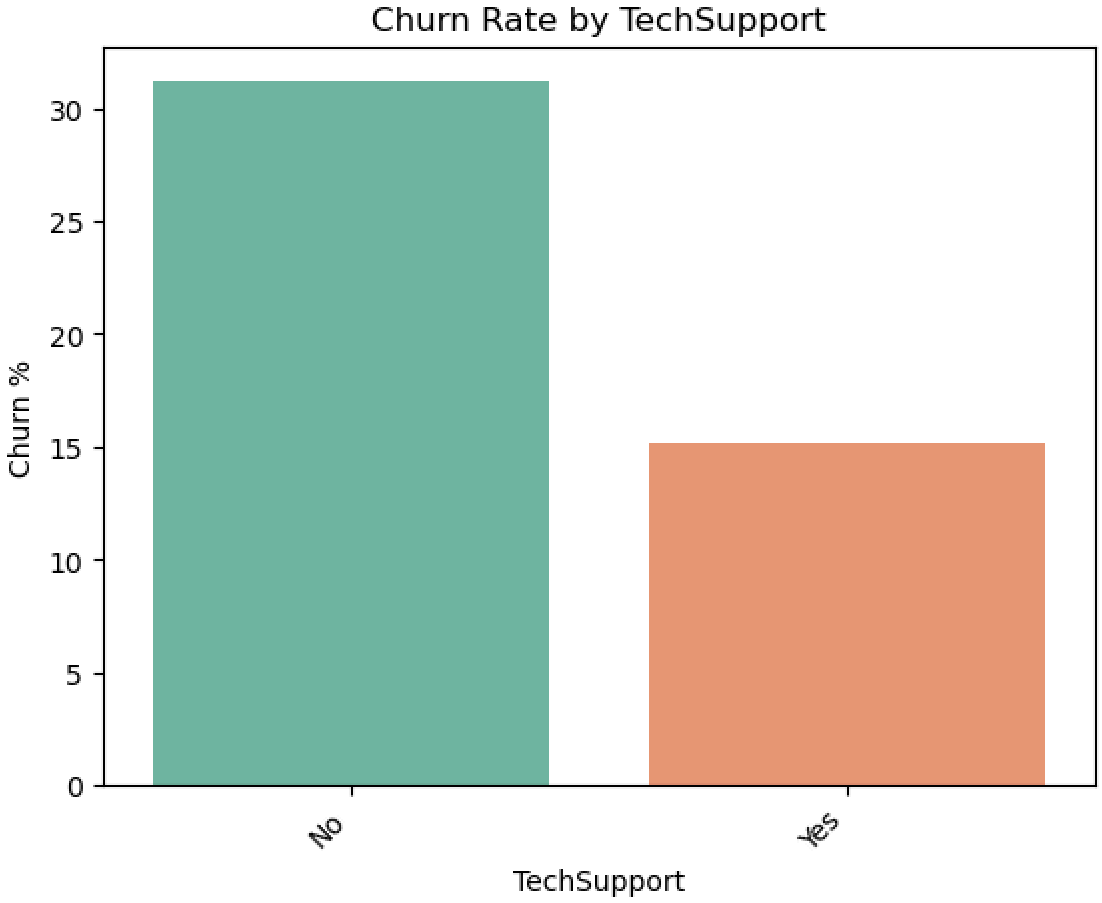

Churn Rate by MultipleLines
Churn %
30
25
20
15
10
5
0
Yes
No
MultipleLines

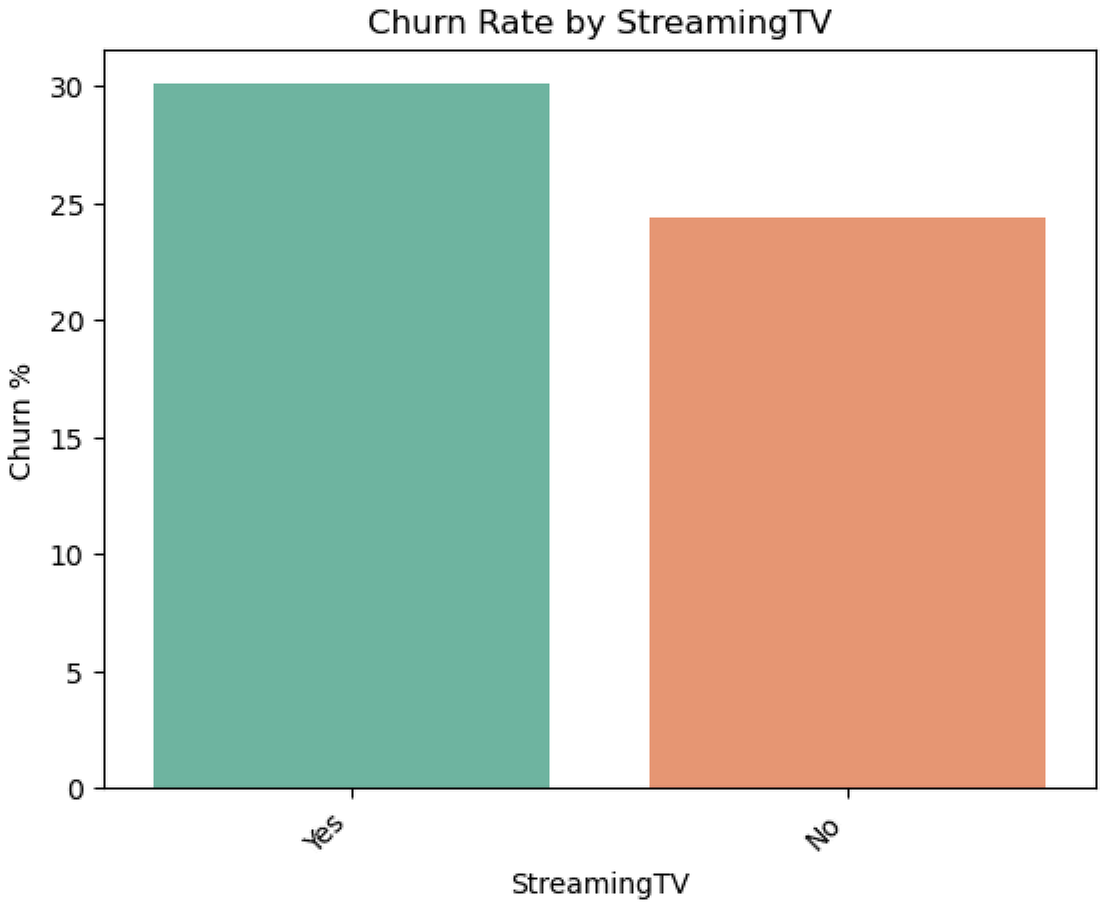

Churn Rate by InternetService
Churn %
0
5
10
15
20
25
30
35
40
Fiber optic
DSL
No
InternetService

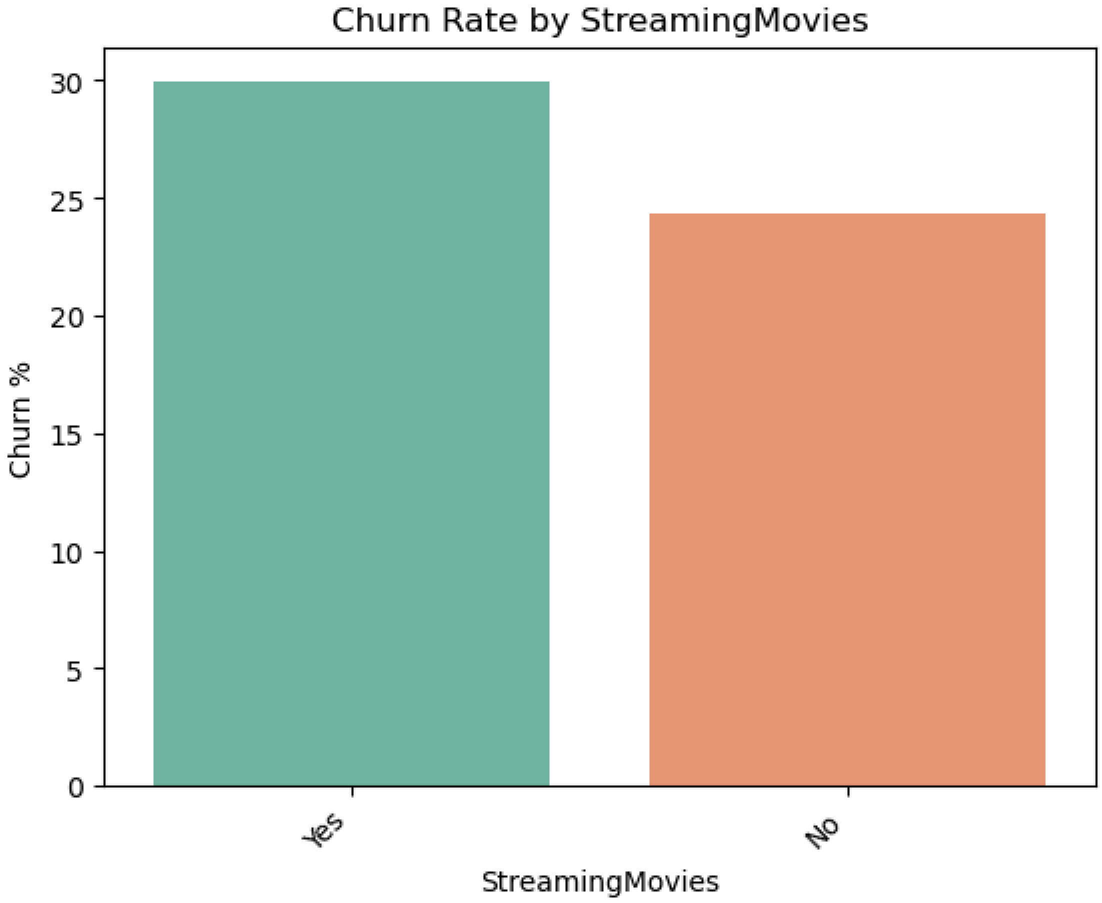

Churn Rate by OnlineSecurity
Churn %
0
5
10
15
20
25
30
No
Yes
OnlineSecurity

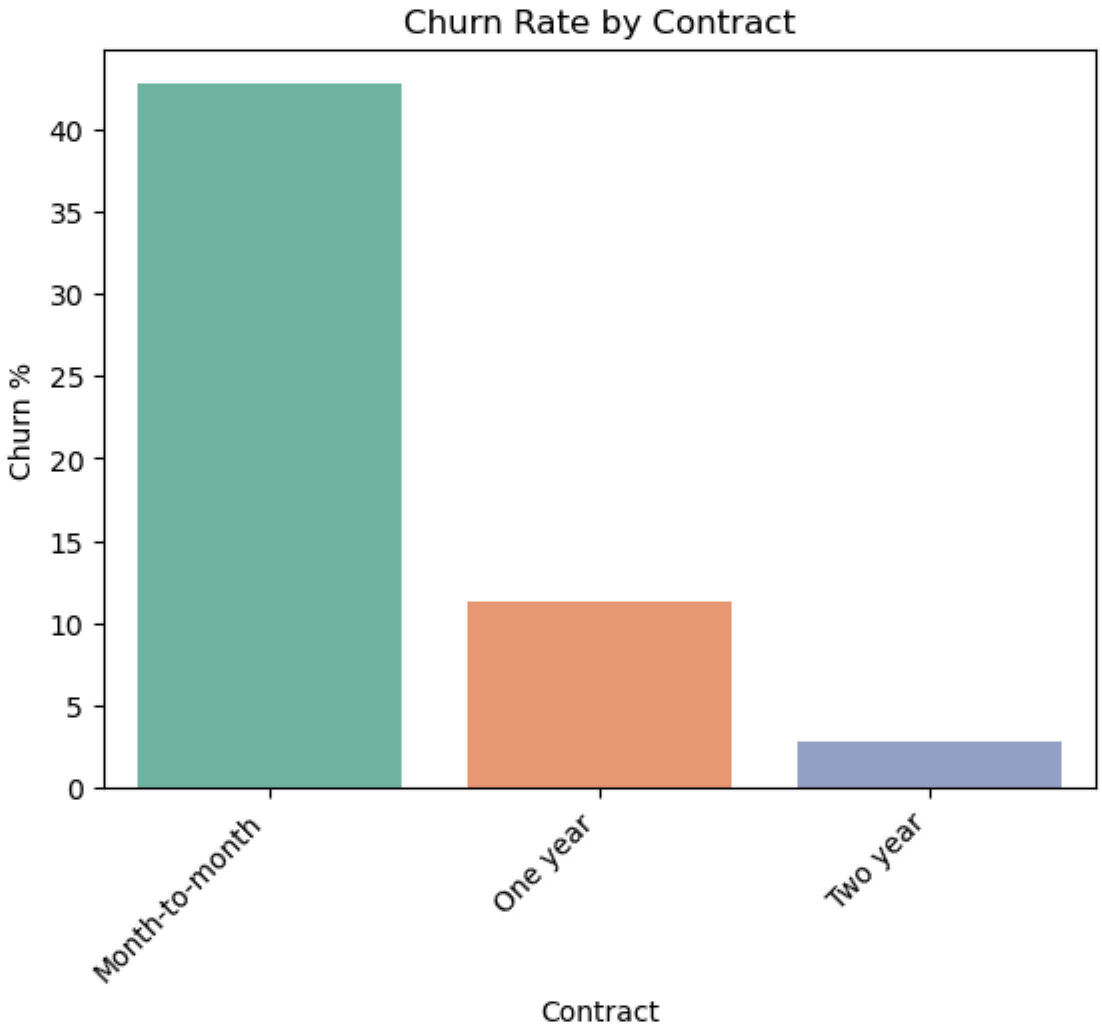

Churn Rate by OnlineBackup
Churn %
0
5
10
15
20
25
30
No
Yes
OnlineBackup


Churn Rate by DeviceProtection
Churn %
0
5
10
15
20
25
30
No
Yes
DeviceProtection

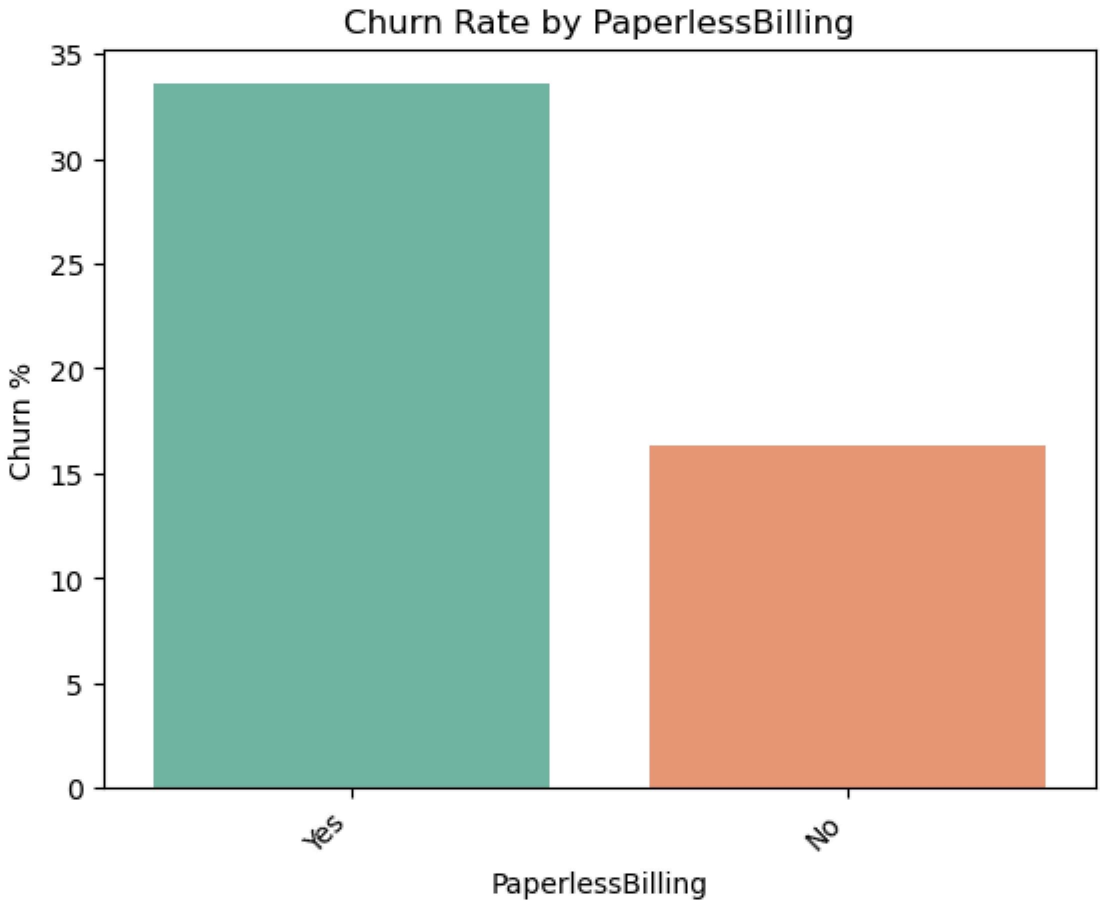

Churn Rate by TechSupport
Churn %
0
5
10
15
20
25
30
No
Yes
TechSupport

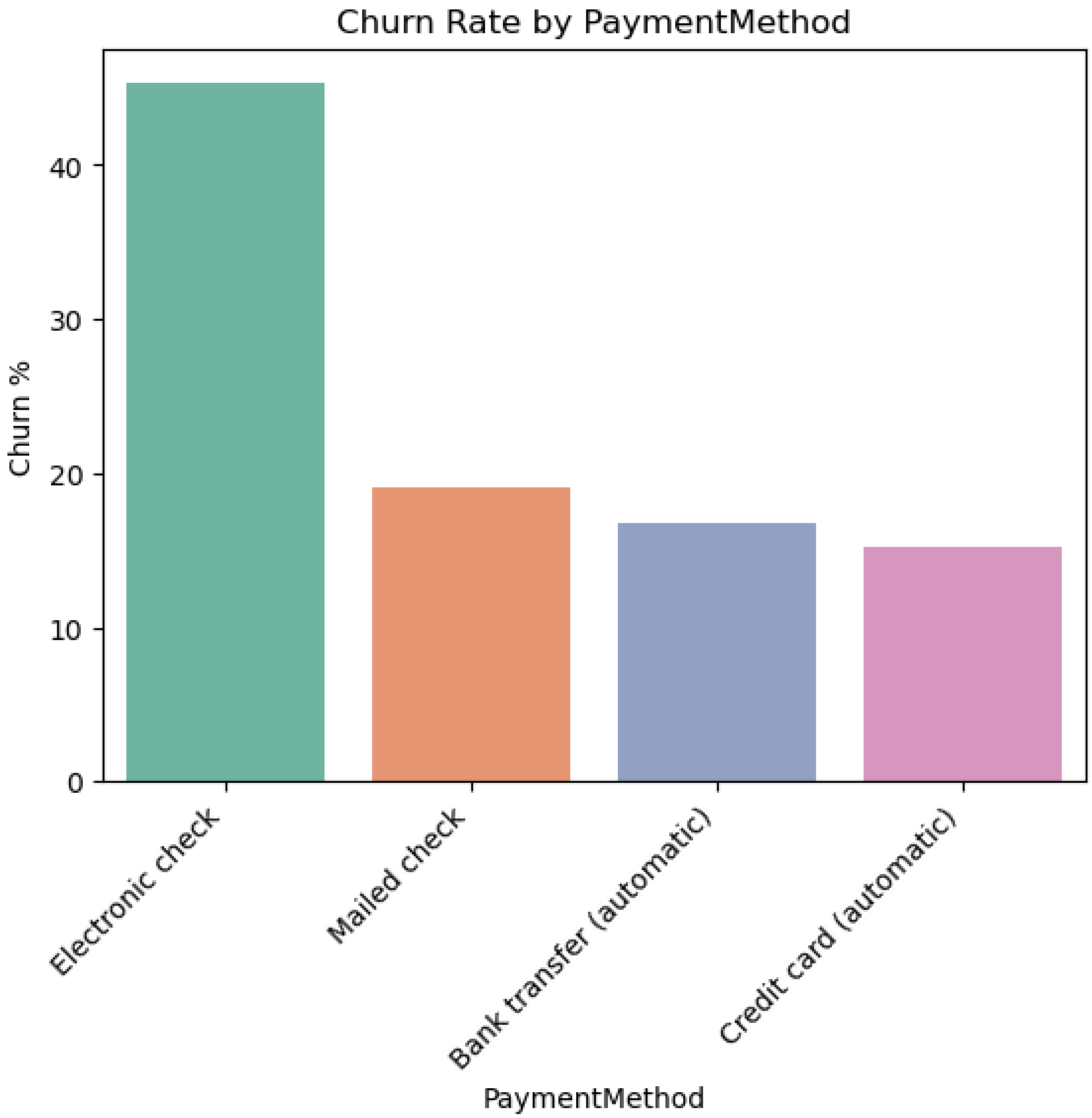

Churn Rate by StreamingTV
Churn %
0
5
10
15
20
25
30
Yes
No
StreamingTV


Churn Rate by StreamingMovies
Churn %
0
5
10
15
20
25
30
Yes
No
StreamingMovies



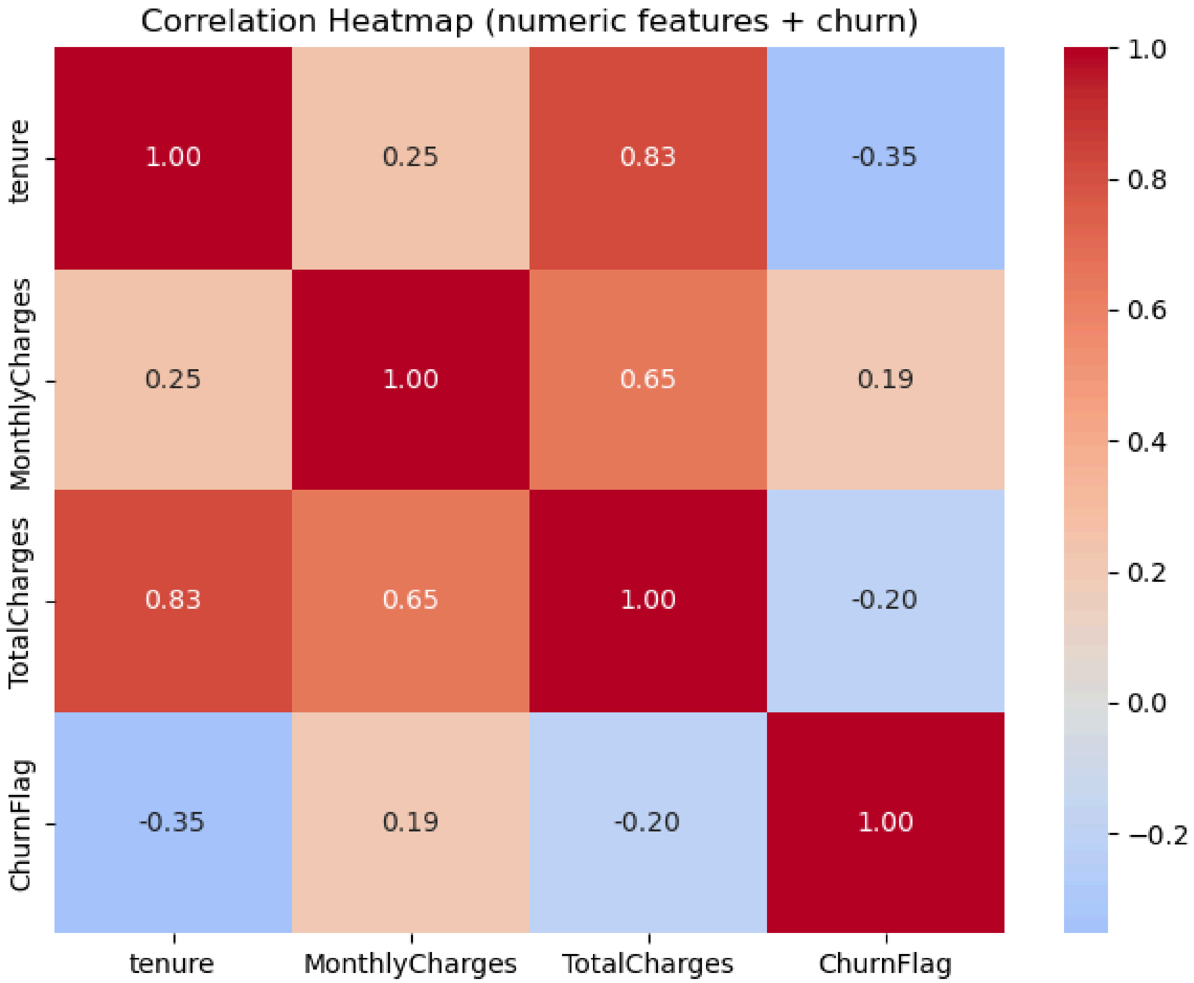
Churn Rate by Contract
Churn %
0
5
10
15
20
25
30
35
40
Month-to-month
One year
Two year
Contract


Churn Rate by PaperlessBilling
Churn %
35
30
25
20
15
10
5
0
Yes
No
PaperlessBilling



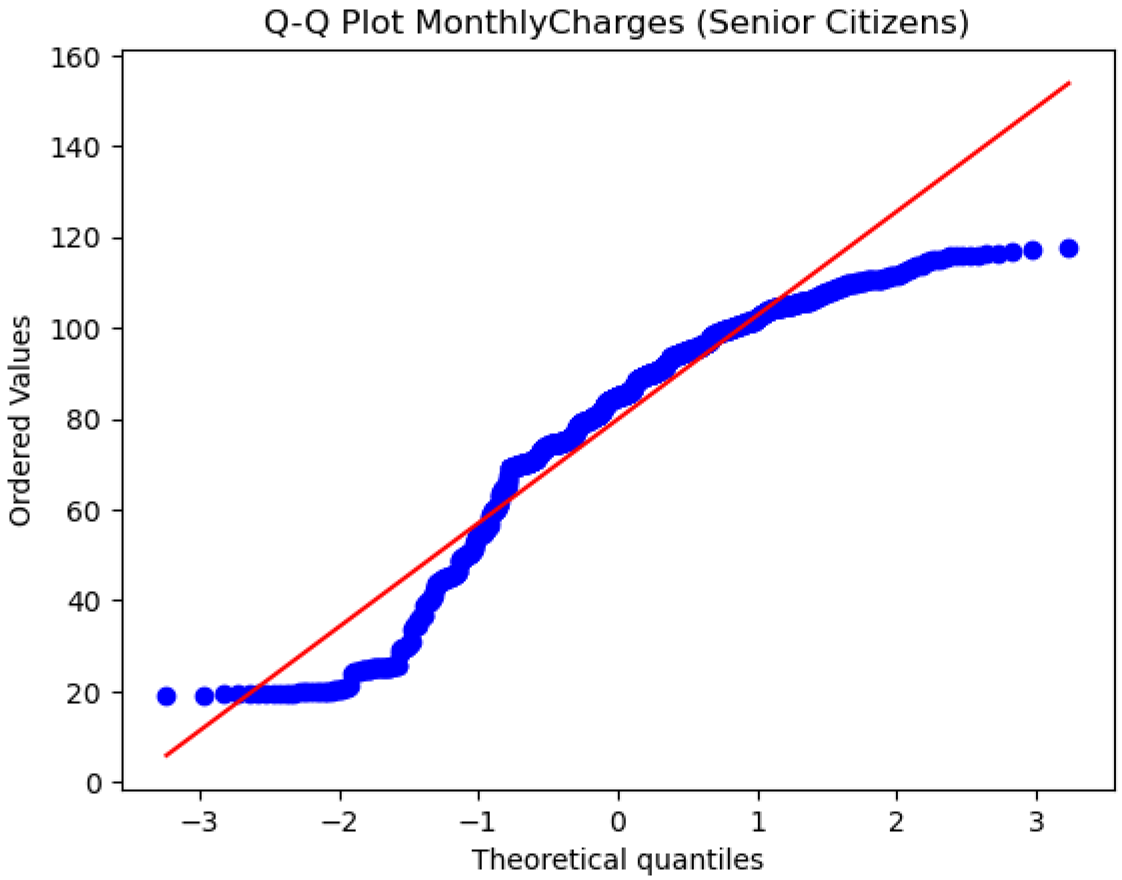
In [18]:

```
# Correlation heatmap (numerics + churn)
corr = df[num_cols+["ChurnFlag"]].corr()
plt.figure(figsize=(8,6))
sns.heatmap(corr, annot=True, fmt=".2f", cmap="coolwarm", center=0)
plt.title("Correlation Heatmap (numeric features + churn)")
plt.show()
```

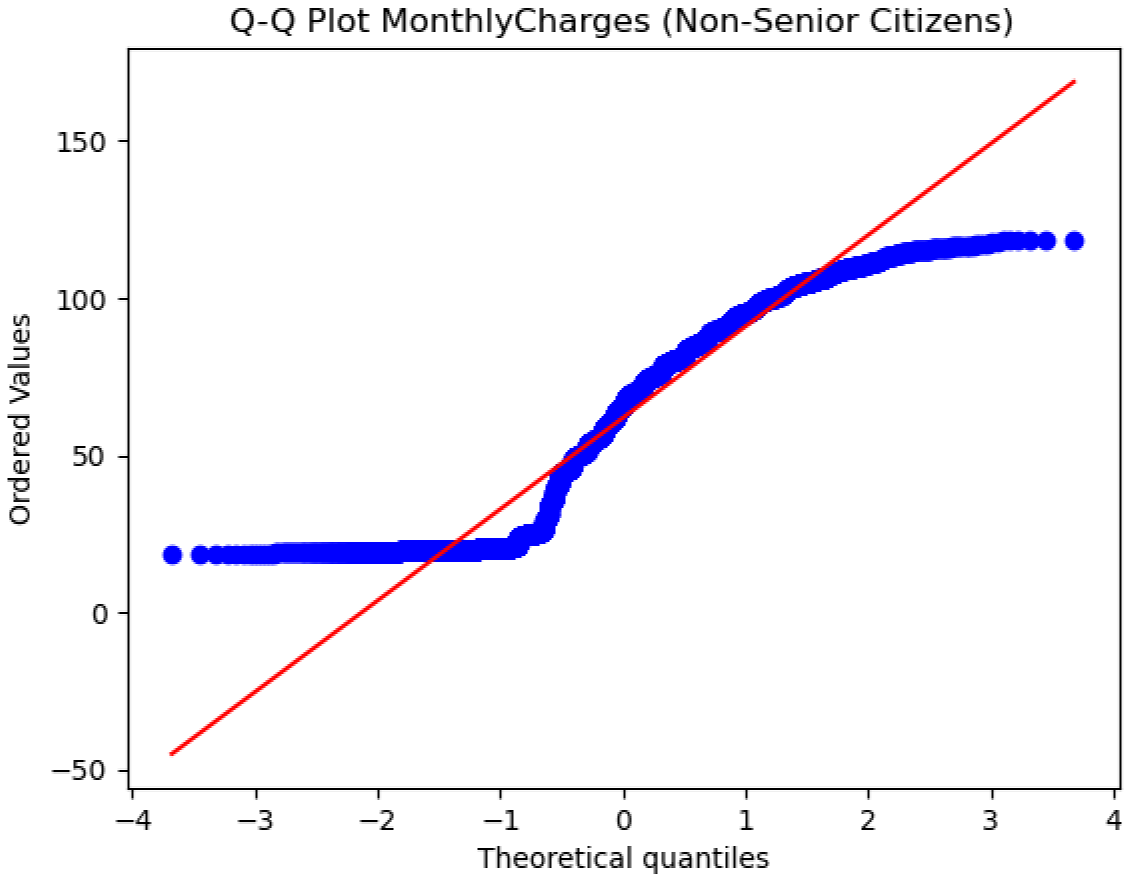


In [21]:

```
sc = df.loc[df["SeniorCitizen"]==1,"MonthlyCharges"]
nsc = df.loc[df["SeniorCitizen"]==0,"MonthlyCharges"]
plt.figure(); stats.probplot(sc, dist="norm", plot=plt)
plt.title("Q-Q Plot MonthlyCharges (Senior Citizens)")
plt.show()
plt.figure(); stats.probplot(nsc, dist="norm", plot=plt)
plt.title("Q-Q Plot MonthlyCharges (Non-Senior Citizens)")
plt.show()
```

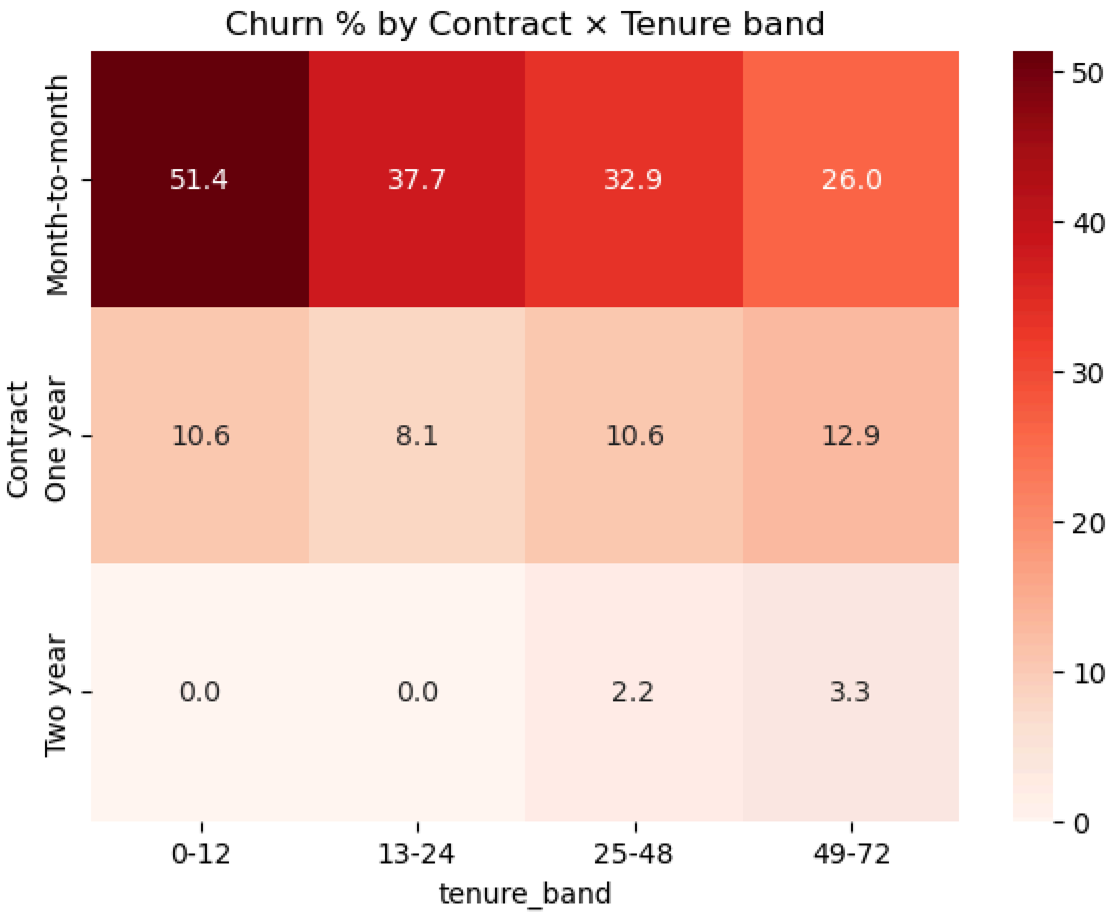


Q-Q Plot MonthlyCharges (Senior Citizens)



Q-Q Plot MonthlyCharges (Non-Senior Citizens)

```
In [19]: if "Contract" in df.columns and "tenure" in df.columns:
             df["tenure_band"] = pd.cut(df["tenure"],bins=[0,12,24,48,72,np.inf],labe
             pivot = pd.crosstab(df["Contract"], df["tenure_band"], values=df["ChurnF
             plt.figure(figsize=(7,5))
             sns.heatmap(pivot, annot=True, fmt=".1f", cmap="Reds")
             plt.title("Churn % by Contract × Tenure band")
             plt.show()
```



```
In [21]: print("\n------------------------------ Insights from the EDA -------------
         print("1. Senior citizens are leaving at nearly twice the rate of younger cu
         print("2. Contract type is the biggest churn driver: month-to-month customer
         print("3. Tenure shows a strong trend: new customers (within the first year)
         print("4. Internet service type also matters: fiber optic users churn the mo
         print("5. Customers who don't take add-on services like Tech Support, Online
         print("6. Streaming services (TV, Movies) don't make much difference—nice to
         print("7. Billing and payment show some clear patterns: electronic check use
         print("8. Monthly charges have a strong effect: customers paying over $90 a
         print("9. Total charges also tell a story: customers with low totals (often
         print("10. Some combinations are especially risky: month-to-month + short te
         print("11. Fiber optic plus high monthly charges is a particularly bad mix,
         print("12. Overall, the groups at highest risk are: senior citizens, month-t
         print("\n--------------------------------------------------------------------
------------------------------ Insights from the EDA -----------------------
-------

1. Senior citizens are leaving at nearly twice the rate of younger customers
(about 42% vs 24%). Age clearly plays a role, but gender doesn’t seem to mat
ter much.
2. Contract type is the biggest churn driver: month-to-month customers leave
the most (>40%), while one-year contracts have much lower churn (~11%) and t
wo-year contracts almost never churn (<5%).
3. Tenure shows a strong trend: new customers (within the first year) churn
heavily, but once they stay beyond 2 years, churn becomes very rare.
4. Internet service type also matters: fiber optic users churn the most (~4
1%), DSL users are more stable, and people without internet service churn th
e least.
5. Customers who don’t take add-on services like Tech Support, Online Securi
ty, or Device Protection are much more likely to leave (~45% churn) compared
to those who do (<15%).
6. Streaming services (TV, Movies) don’t make much difference—nice to have,
but not really a retention factor.
7. Billing and payment show some clear patterns: electronic check users chur
n the most (~45%), while auto-pay users (bank transfer/credit card) churn th
e least (<15%). Paperless billing users leave a bit more than those with mai
led bills.
8. Monthly charges have a strong effect: customers paying over $90 a month a
re much more likely to churn, while those with lower bills tend to stay.
9. Total charges also tell a story: customers with low totals (often new sig
n-ups) churn heavily, but those who’ve spent more over time (loyal, long-ten
ure customers) rarely leave.
10. Some combinations are especially risky: month-to-month + short tenure cu
stomers churn the most (~60%), while long-tenure + two-year contract custome
rs almost never leave (<2%).
11. Fiber optic plus high monthly charges is a particularly bad mix, with ve
ry high churn. But fiber optic users who have Tech Support churn far less.
12. Overall, the groups at highest risk are: senior citizens, month-to-month
customers, new/short-tenure customers, fiber optic users with high bills, el
ectronic check payers, and those without Tech Support or security add-ons.

----------------------------------------------------------------------------
-----
```

```
In [ ]:
```