Name: Elsa Noor

Roll No : f22bdats1m02004

# Title Page

Customer Churn Prediction for Subscription-Based Services

## Overview of the Dataset

Customer churn, the rate at which customers cancel their subscriptions, is a critical challenge for subscription-based businesses. This dataset is designed to predict customer churn and help businesses take proactive measures to retain their customers. The data encapsulates various customer attributes, including subscription details, payment methods, and behavioral metrics, providing a holistic view of customer interactions with the service.

## This dataset comprises three files:

### train.csv:

Training data with labeled examples of customer churn.

### test.csv:

Unlabeled data used to test the performance of predictive models.

### data_descriptions.csv:

A detailed explanation of all the features in the dataset.

### The dataset includes 19 features, categorized as:

Demographic attributes: Customer profile information such as gender and device usage. Behavioral metrics: Insights into customer activity, including viewing habits and content preferences. Financial attributes: Monthly charges and total charges, reflecting the financial relationship with the service. By analyzing these features, this project aims to uncover the patterns and factors that contribute to customer churn, enabling better decision-making for retention strategies.

```
```

## VISULIZATON OF DATASET

```
# Importing Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
import seaborn as sns

# Loading the Dataset
# Replace 'train.csv' with the path to your dataset file
train_data = pd.read_csv(r'train.csv')

# Display the first few rows of the dataset
print("First 5 rows of the dataset:")
display(train_data.head(15))
```

```
First 5 rows of the dataset:
```

| | AccountAge | MonthlyCharges | TotalCharges | SubscriptionType |
|---|---|---|---|---|
| 0 | 20 | 11.055215 | 221.104302 | Premium |
| 1 | 57 | 5.175208 | 294.986882 | Basic |
| 2 | 73 | 12.106657 | 883.785952 | Basic |
| 3 | 32 | 7.263743 | 232.439774 | Basic |
| 4 | 57 | 16.953078 | 966.325422 | Premium |
| 5 | 113 | 7.295744 | 824.419081 | Premium |
| 6 | 38 | 12.340675 | 468.945639 | Premium |
| 7 | 25 | 7.247550 | 181.188753 | Standard |
| 8 | 26 | 19.803233 | 514.884050 | Standard |
| 9 | 14 | 18.842934 | 263.801080 | Standard |
| 10 | 114 | 18.323630 | 2088.893783 | Premium |
| 11 | 3 | 16.271635 | 48.814904 | Standard |
| 12 | 64 | 7.749444 | 495.964395 | Basic |
| 13 | 43 | 6.209336 | 267.001469 | Premium |
| 14 | 98 | 7.589784 | 743.798786 | Basic |

| | PaymentMethod | PaperlessBilling | ContentType | MultiDeviceAccess |
|---|---|---|---|---|
| 0 | Mailed check | No | Both | No |
| 1 | Credit card | Yes | Movies | No |
| 2 | Mailed check | Yes | Movies | No |
| 3 | Electronic check | No | TV Shows | No |
| 4 | Electronic check | Yes | TV Shows | No |
| 5 | Mailed check | Yes | Both | No |
| 6 | Bank transfer | No | Both | No |
| 7 | Electronic check | Yes | TV Shows | No |
| 8 | Bank transfer | No | Movies | No |
| 9 | Bank transfer | No | Movies | No |
| 10 | Mailed check | No | TV Shows | No |
| 11 | Electronic check | No | Both | Yes |
| 12 | Mailed check | Yes | TV Shows | No |
| 13 | Bank transfer | No | Movies | No |
| 14 | Electronic check | No | Both | No |

| | DeviceRegistered | ViewingHoursPerWeek | ... | ContentDownloadsPerMonth |
|---|---|---|---|---|
| 0 | Mobile | 36.758104 | ... | 10 |

```
1             Tablet             32.450568  ...
18
2           Computer              7.395160  ...
23
3             Tablet             27.960389  ...
30
4                 TV             20.083397  ...
20
5             Mobile             21.678290  ...
35
6           Computer             36.512761  ...
28
7                 TV             16.355816  ...
10
8             Tablet              8.202929  ...
28
9           Computer             38.560694  ...
0
10          Computer             17.642755  ...
18
11            Tablet              9.697394  ...
20
12            Tablet             28.063868  ...
30
13          Computer             37.098949  ...
49
14            Tablet             37.665446  ...
26

    GenrePreference UserRating  SupportTicketsPerMonth  Gender
WatchlistSize  \
0             Sci-Fi   2.176498                       4    Male
3
1             Action   3.478632                       8    Male
23
2            Fantasy   4.238824                       6    Male
1
3              Drama   4.276013                       2    Male
24
4             Comedy   3.616170                       4  Female
0
5             Comedy   3.721134                       8  Female
2
6             Action   4.090868                       9  Female
20
7            Fantasy   3.410221                       2  Female
22
8            Fantasy   2.679986                       0    Male
5
```

```
9             Comedy   2.993441                          0    Male
18
10            Comedy   3.676324                          1    Male
5
11           Fantasy   2.562292                          1    Male
8
12            Action   3.772572                          7  Female
18
13            Action   1.802052                          0    Male
14
14           Fantasy   1.492742                          0    Male
8

    ParentalControl SubtitlesEnabled  CustomerID Churn
0                No                No  CB6SXPNVZA     0
1                No               Yes  S7R2G87O09     0
2               Yes               Yes  EASDC20BDT     0
3               Yes               Yes  NPF69NT69N     0
4                No                No  4LGYPK7VOL     0
5               Yes               Yes  JY5HS0GWHW     0
6                No               Yes  79XSO6P5O3     0
7                No                No  2LDC9AQ3C5     0
8               Yes               Yes  74DURHL3Y8     1
9                No                No  CY8S2R3A1T     0
10              Yes                No  V1LEGCSV61     0
11              Yes                No  G9E6VT02F2     0
12               No                No  QFP5ALFKJ5     0
13              Yes               Yes  AFQ6J0GIKW     0
14               No               Yes  IQNESR4W65     0

[15 rows x 21 columns]
```

## INFO about dataset

```
# Dataset Info
print("\nDataset Info:")
train_data.info()
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 243787 entries, 0 to 243786
Data columns (total 21 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   AccountAge               243787 non-null  int64
 1   MonthlyCharges           243787 non-null  float64
 2   TotalCharges             243787 non-null  float64
 3   SubscriptionType         243787 non-null  object
 4   PaymentMethod            243787 non-null  object
```

```
 5   PaperlessBilling           243787 non-null  object
 6   ContentType                243787 non-null  object
 7   MultiDeviceAccess          243787 non-null  object
 8   DeviceRegistered           243787 non-null  object
 9   ViewingHoursPerWeek        243787 non-null  float64
 10  AverageViewingDuration     243787 non-null  float64
 11  ContentDownloadsPerMonth   243787 non-null  int64
 12  GenrePreference            243787 non-null  object
 13  UserRating                 243787 non-null  float64
 14  SupportTicketsPerMonth     243787 non-null  int64
 15  Gender                     243787 non-null  object
 16  WatchlistSize              243787 non-null  int64
 17  ParentalControl            243787 non-null  object
 18  SubtitlesEnabled           243787 non-null  object
 19  CustomerID                 243787 non-null  object
 20  Churn                      243787 non-null  int64
dtypes: float64(5), int64(5), object(11)
memory usage: 39.1+ MB
```

## Check for missing values

```
# Check for missing values
print("\nMissing Values:")
print(train_data.isnull().sum())
```

```
Missing Values:
AccountAge                  0
MonthlyCharges              0
TotalCharges                0
SubscriptionType            0
PaymentMethod               0
PaperlessBilling            0
ContentType                 0
MultiDeviceAccess           0
DeviceRegistered            0
ViewingHoursPerWeek         0
AverageViewingDuration      0
ContentDownloadsPerMonth    0
GenrePreference             0
UserRating                  0
SupportTicketsPerMonth      0
Gender                      0
WatchlistSize               0
ParentalControl             0
SubtitlesEnabled            0
CustomerID                  0
Churn                       0
dtype: int64
```

## Objective of this Report

The report aims to achieve the following objectives:

### Develop a Predictive Model:

Build and validate a machine learning model to accurately predict customer churn based on the dataset.

### Identify Key Drivers of Churn:

Determine the most influential features impacting churn, such as financial, behavioral, and demographic factors.

### Enhance Customer Retention Strategies:

Provide actionable insights and recommendations to address potential churners, such as: Personalized content recommendations. Flexible payment plans or discounts. Improved customer support services.

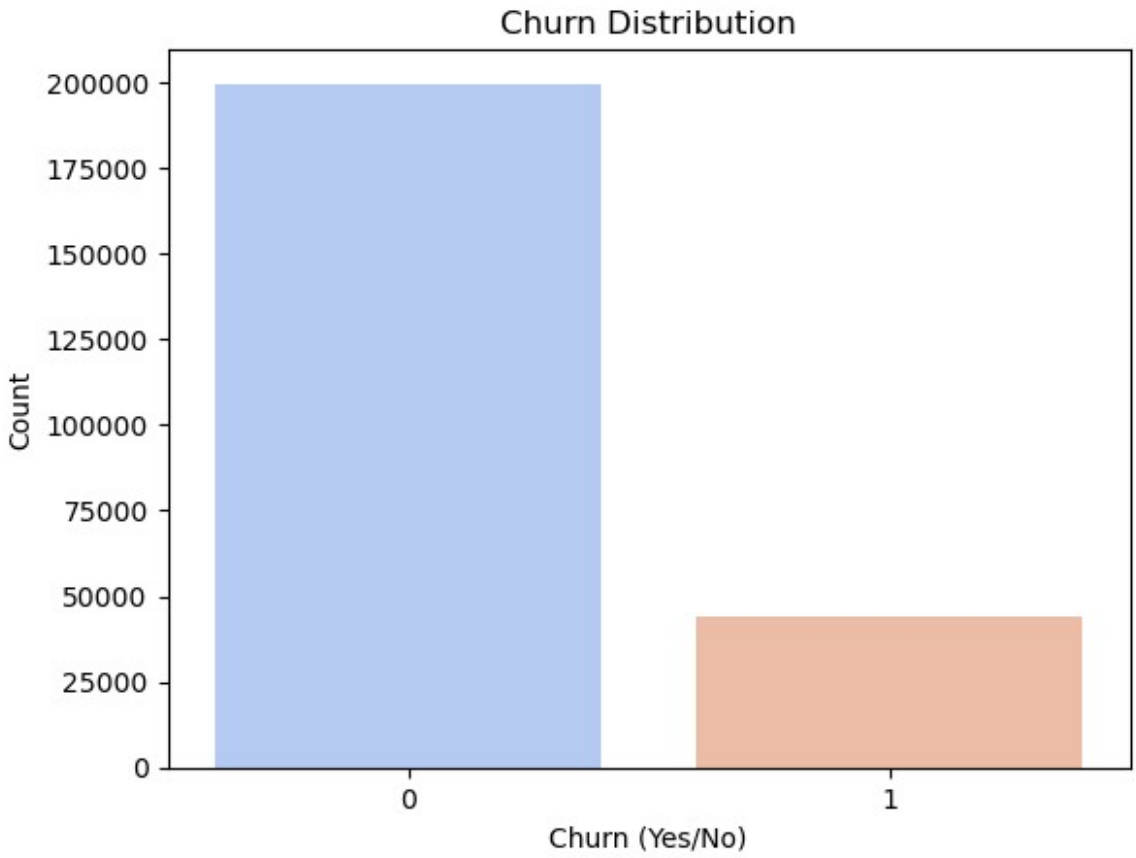
### Optimize Business Decisions:

Equip the business with data-driven insights to implement cost-effective retention measures. The overarching goal is to enable the business to reduce customer churn, enhance customer satisfaction, and maximize long-term revenue

## Visualizing the Target Variable (Churn)

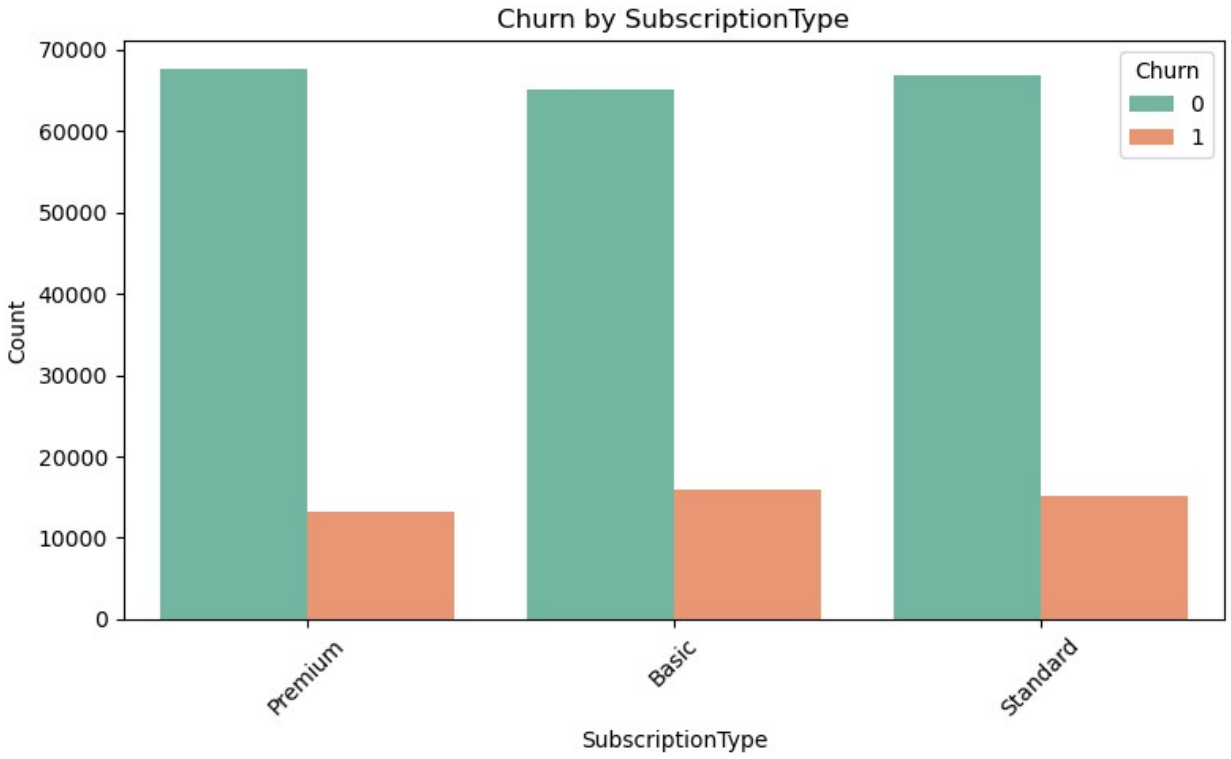
```
# Visualizing the Target Variable (Churn)
print("\nChurn Distribution:")
sns.countplot(x='Churn', data=train_data, palette='coolwarm')
plt.title('Churn Distribution')
plt.xlabel('Churn (Yes/No)')
plt.ylabel('Count')
plt.show()
```

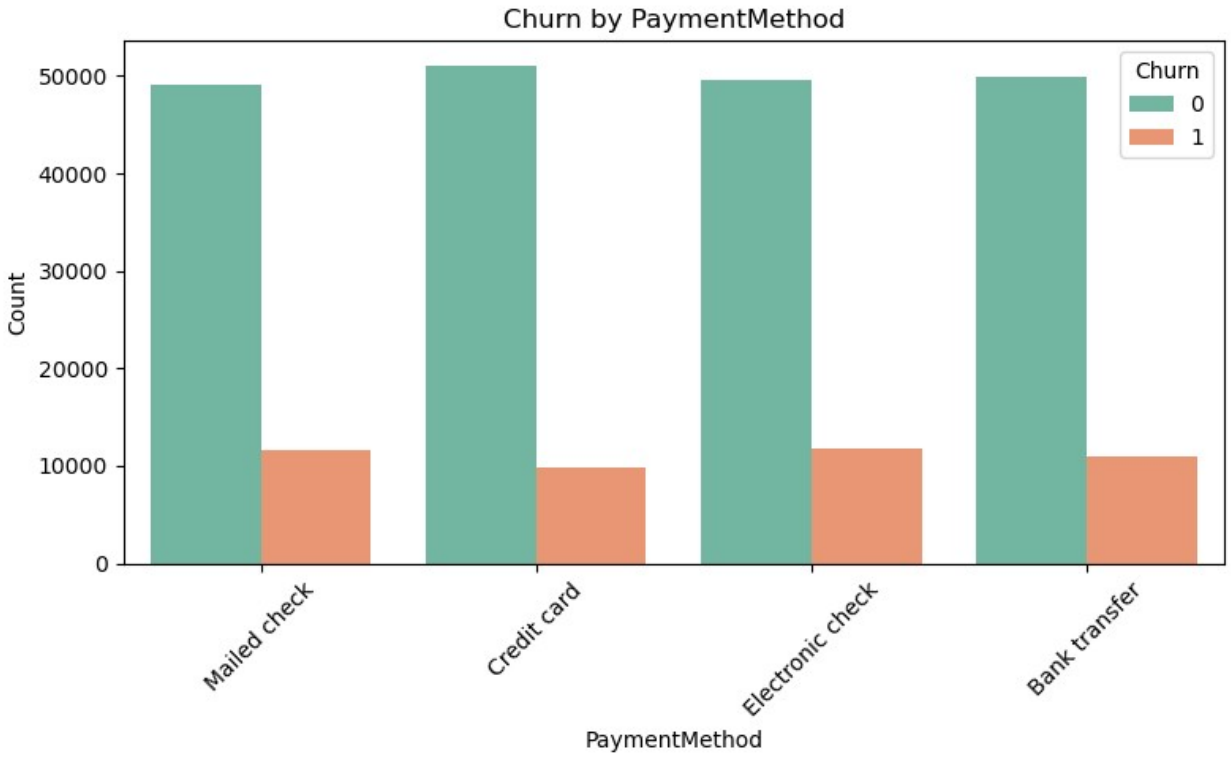
```
Churn Distribution:
```

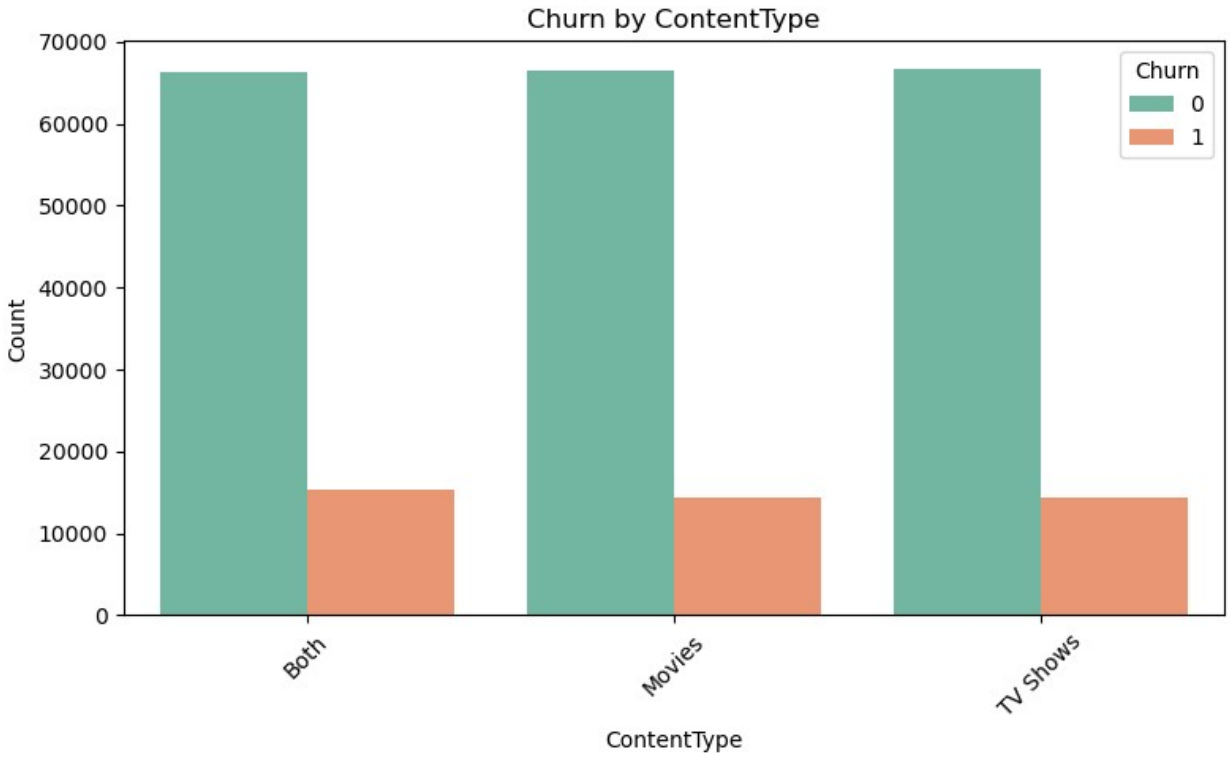


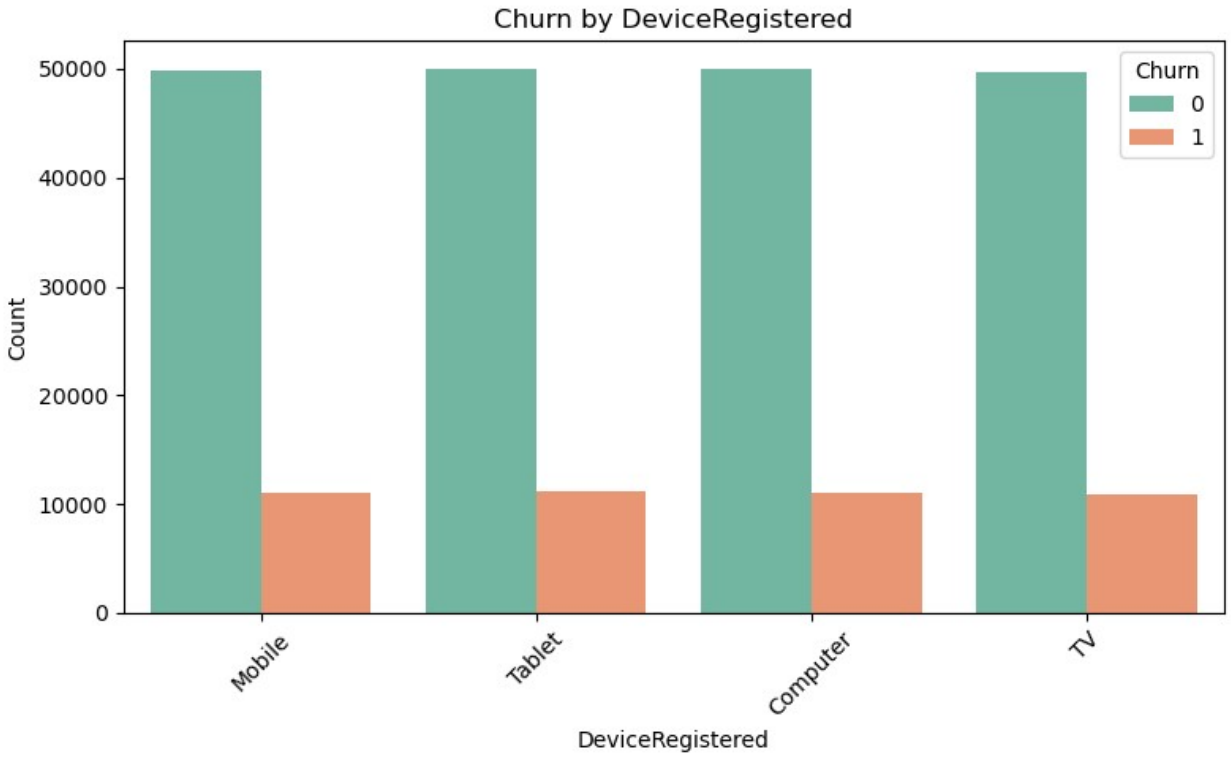

```
# Visualizing Categorical Features
categorical_features = ['SubscriptionType', 'PaymentMethod',
'ContentType', 'DeviceRegistered']

for feature in categorical_features:
    plt.figure(figsize=(8, 5))
    sns.countplot(x=feature, hue='Churn', data=train_data,
palette='Set2')
    plt.title(f'Churn by {feature}')
    plt.xlabel(feature)
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.legend(title='Churn', loc='upper right')
    plt.tight_layout()
    plt.show()
```


Churn by SubscriptionType
Churn
0
1
Count
70000
60000
50000
40000
30000
20000
10000
0
Premium
Basic
Standard
SubscriptionType

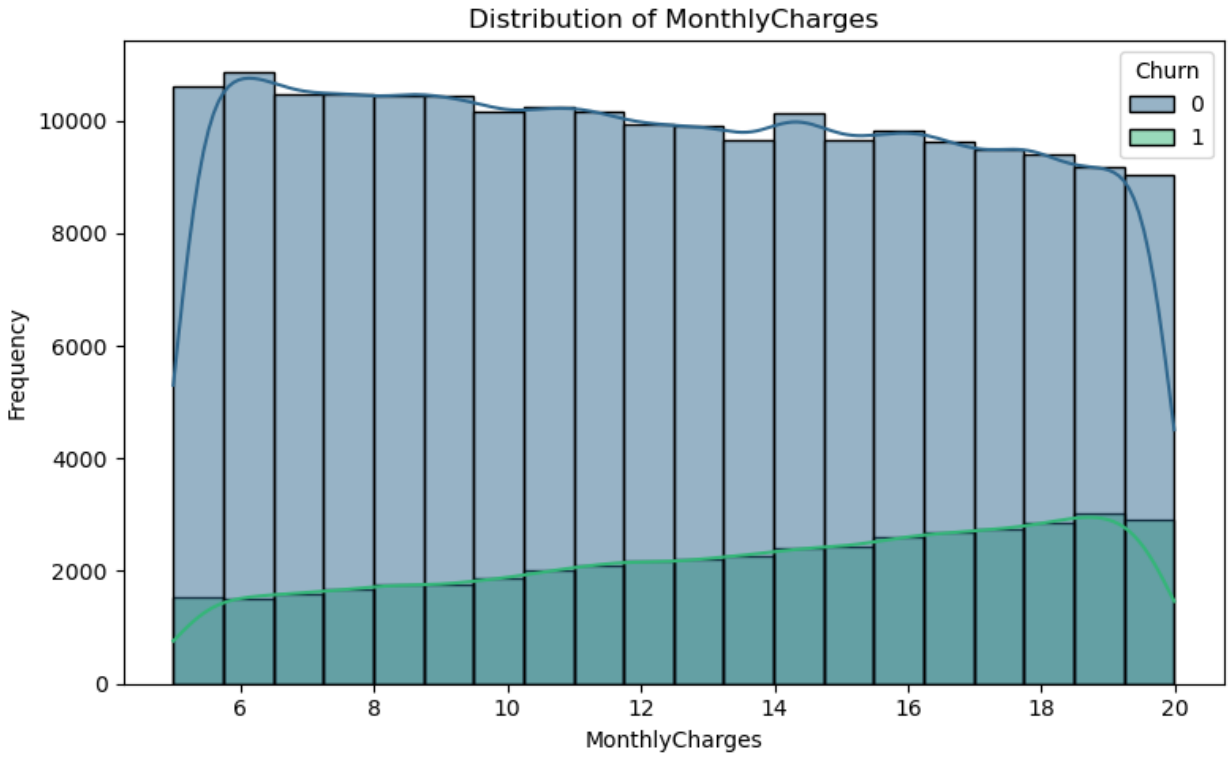

Churn by PaymentMethod
Churn
0
1
Count
50000
40000
30000
20000
10000
0
Mailed check
Credit card
Electronic check
Bank transfer
PaymentMethod

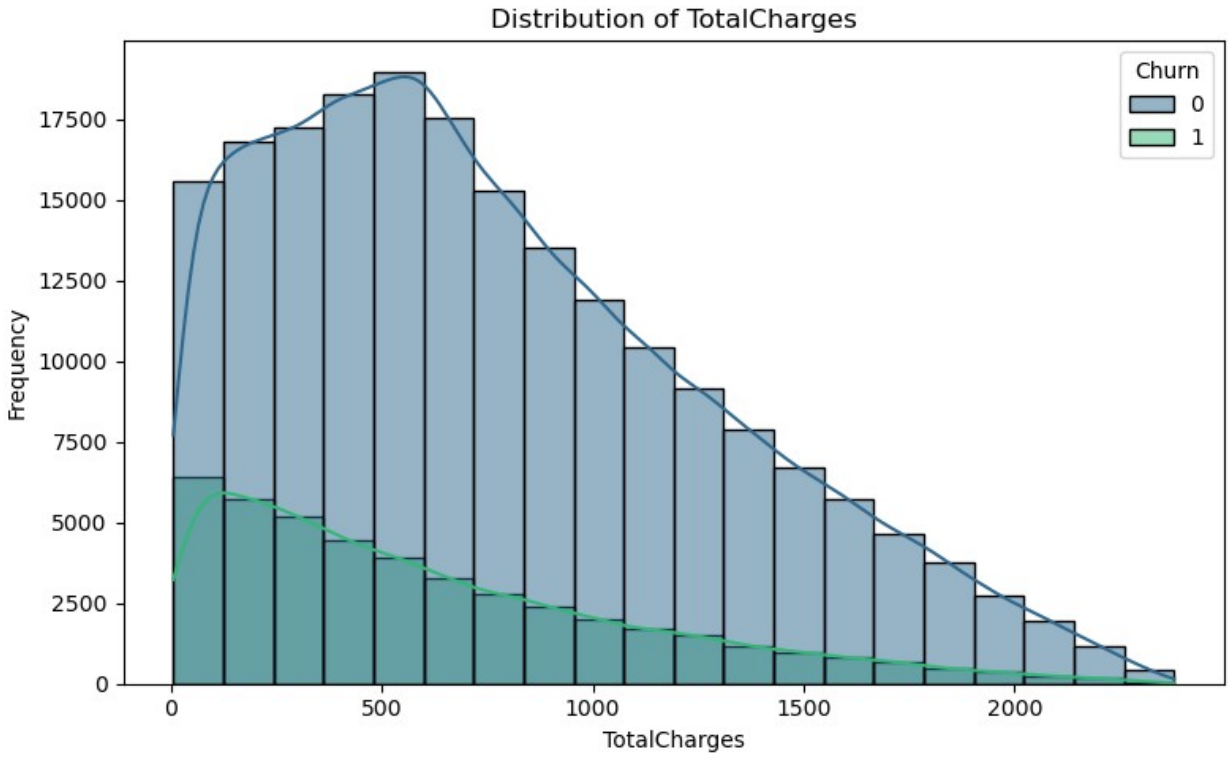


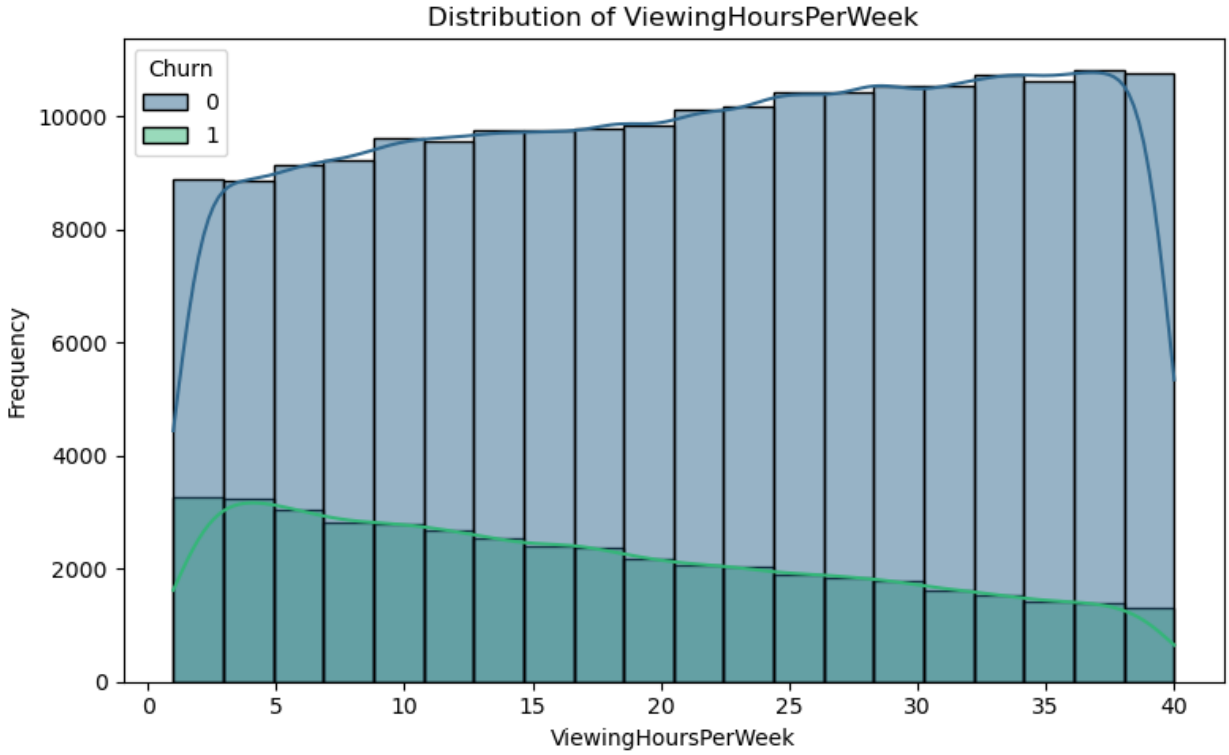



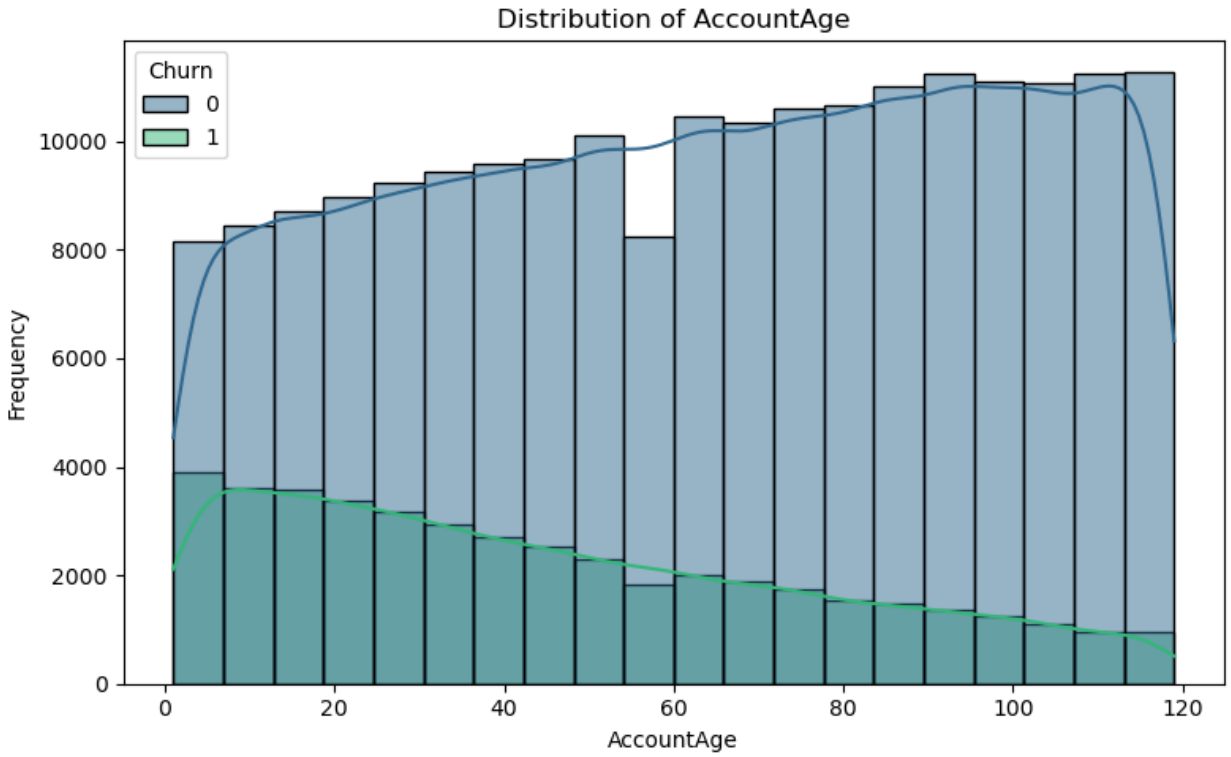

```
# Visualizing the Distribution of Numerical Features
numerical_features = ['MonthlyCharges', 'TotalCharges',
'ViewingHoursPerWeek', 'AccountAge']
```

```
for feature in numerical_features:
    plt.figure(figsize=(8, 5))
    sns.histplot(data=train_data, x=feature, hue='Churn', kde=True,
palette='viridis', bins=20)
    plt.title(f'Distribution of {feature}')
    plt.xlabel(feature)
    plt.ylabel('Frequency')
    plt.tight_layout()
    plt.show()
```

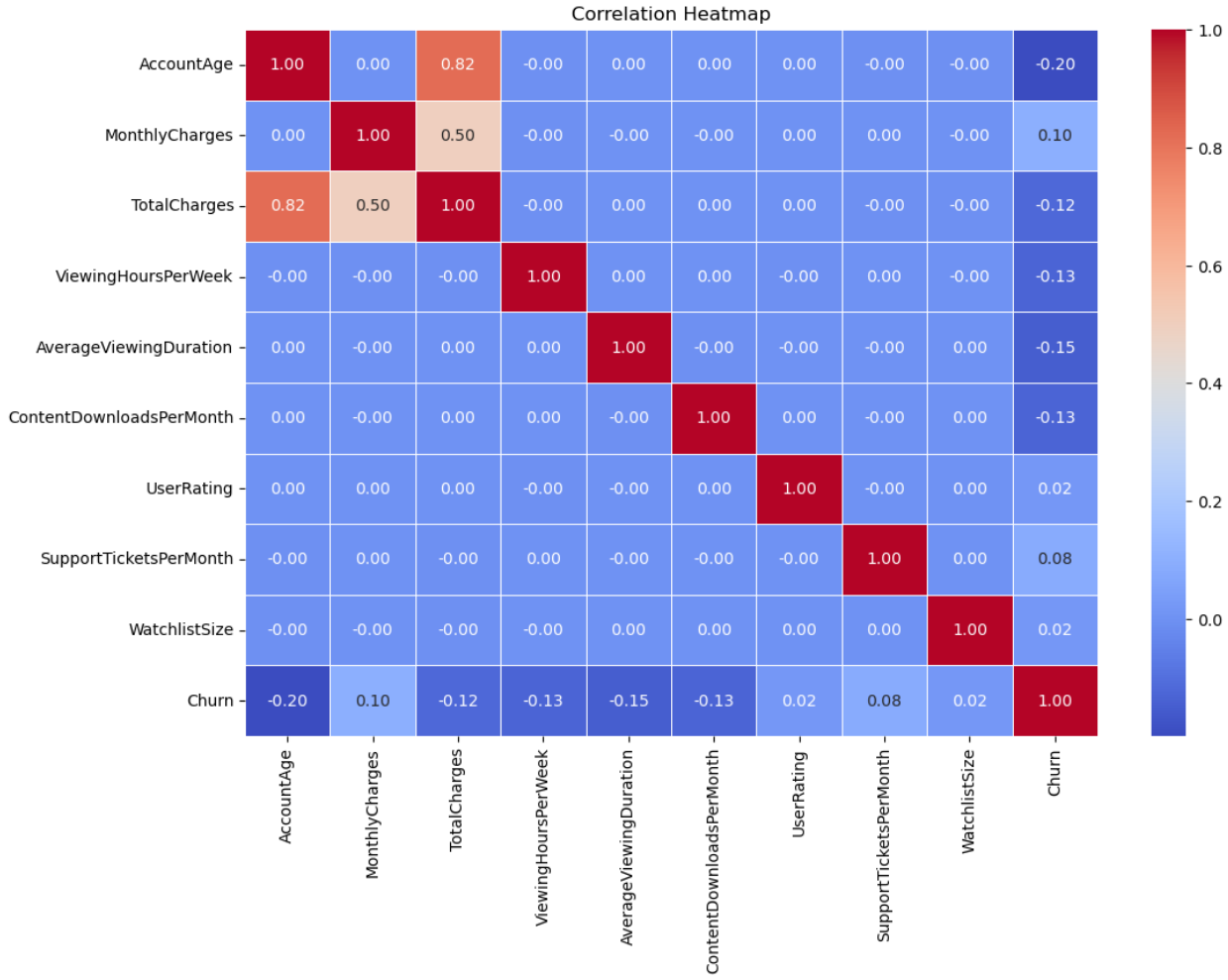



Distribution of TotalCharges
Churn
0
1
Frequency
17500
15000
12500
10000
7500
5000
2500
0
0
500
1000
1500
2000
TotalCharges

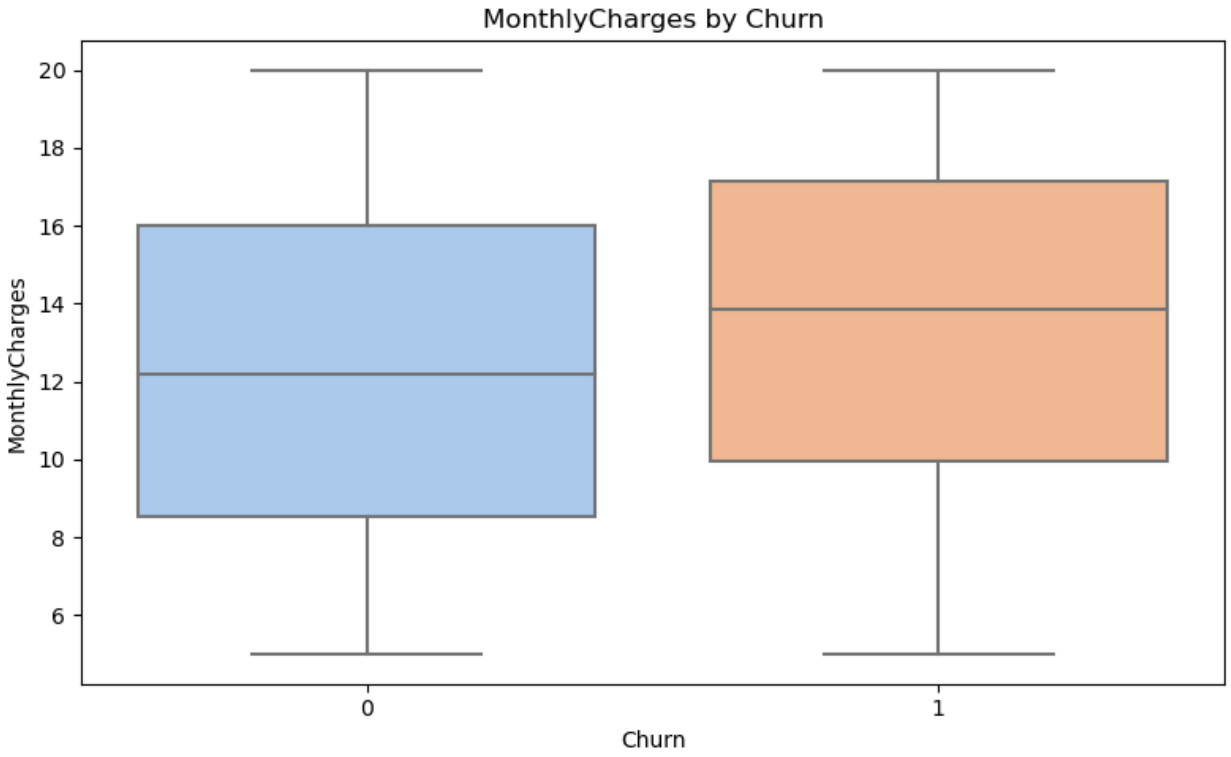

Distribution of ViewingHoursPerWeek
Churn
0
1
Frequency
10000
8000
6000
4000
2000
0
0
5
10
15
20
25
30
35
40
ViewingHoursPerWeek

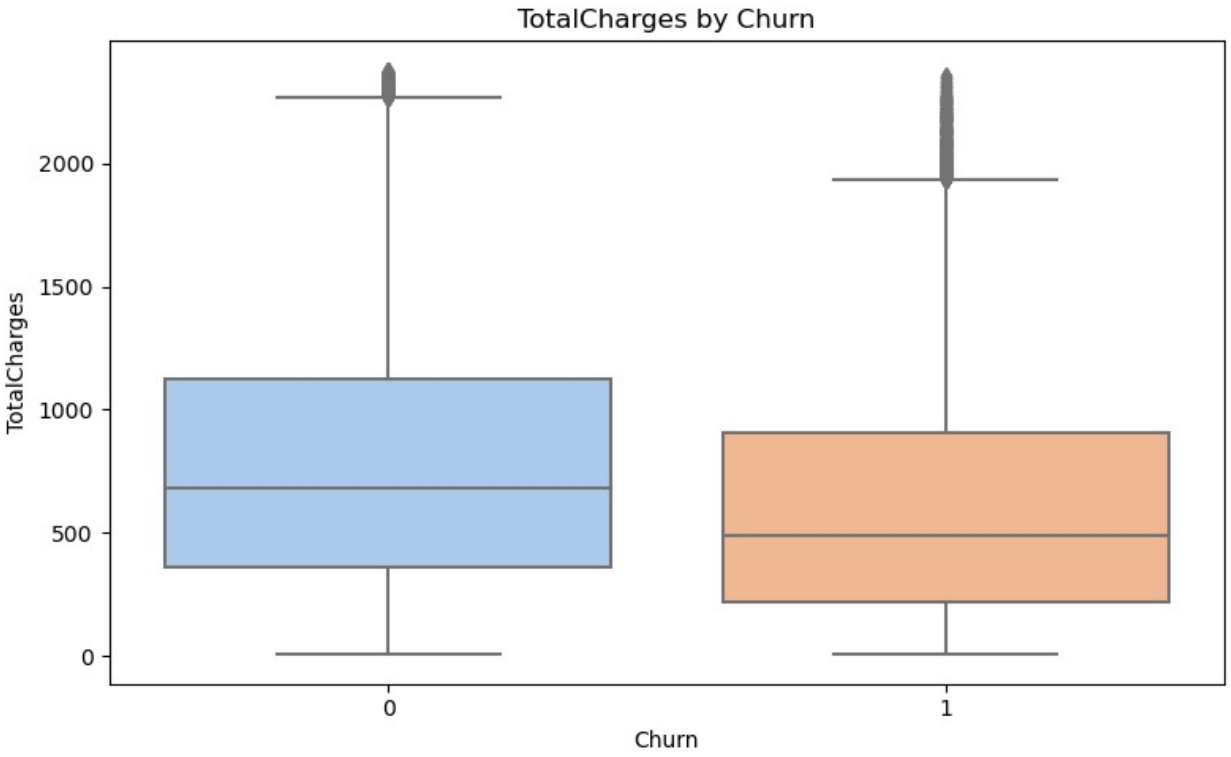


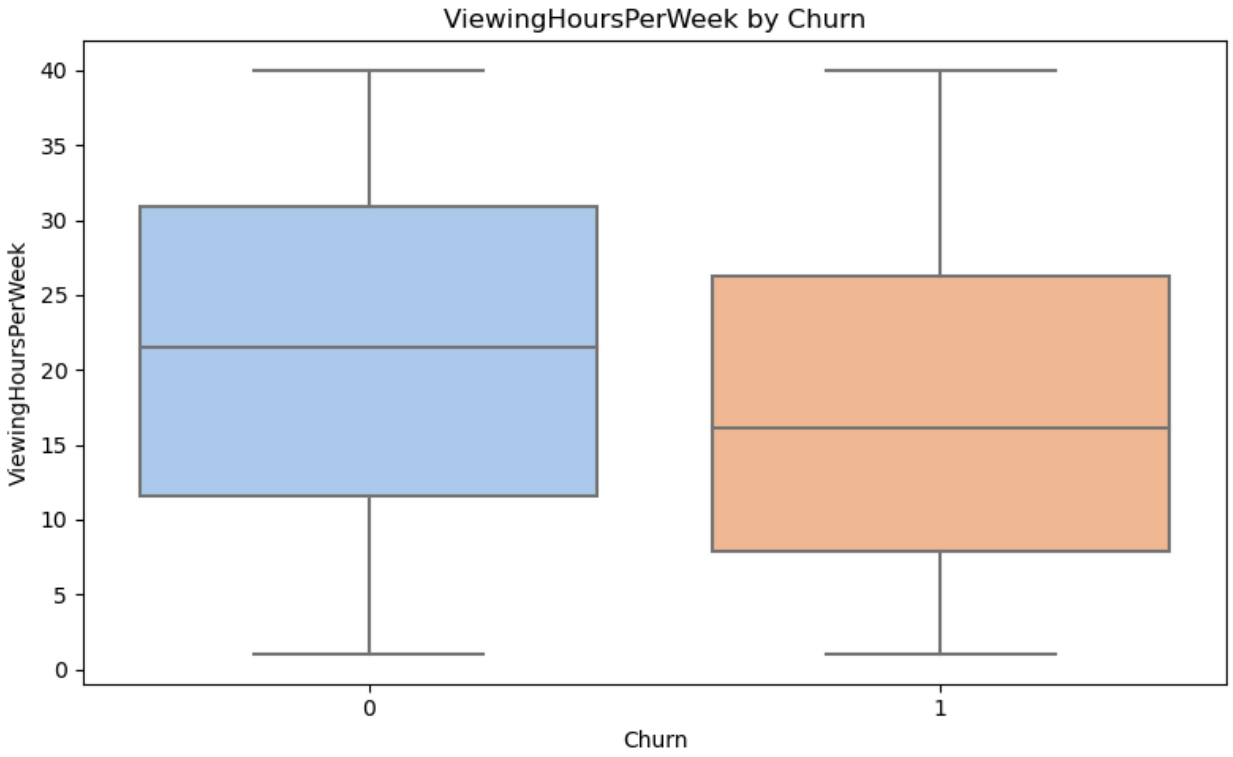

```
# Correlation Heatmap (Numerical Features)
plt.figure(figsize=(12, 8))
sns.heatmap(train_data.corr(), annot=True, fmt='.2f', cmap='coolwarm',
linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```

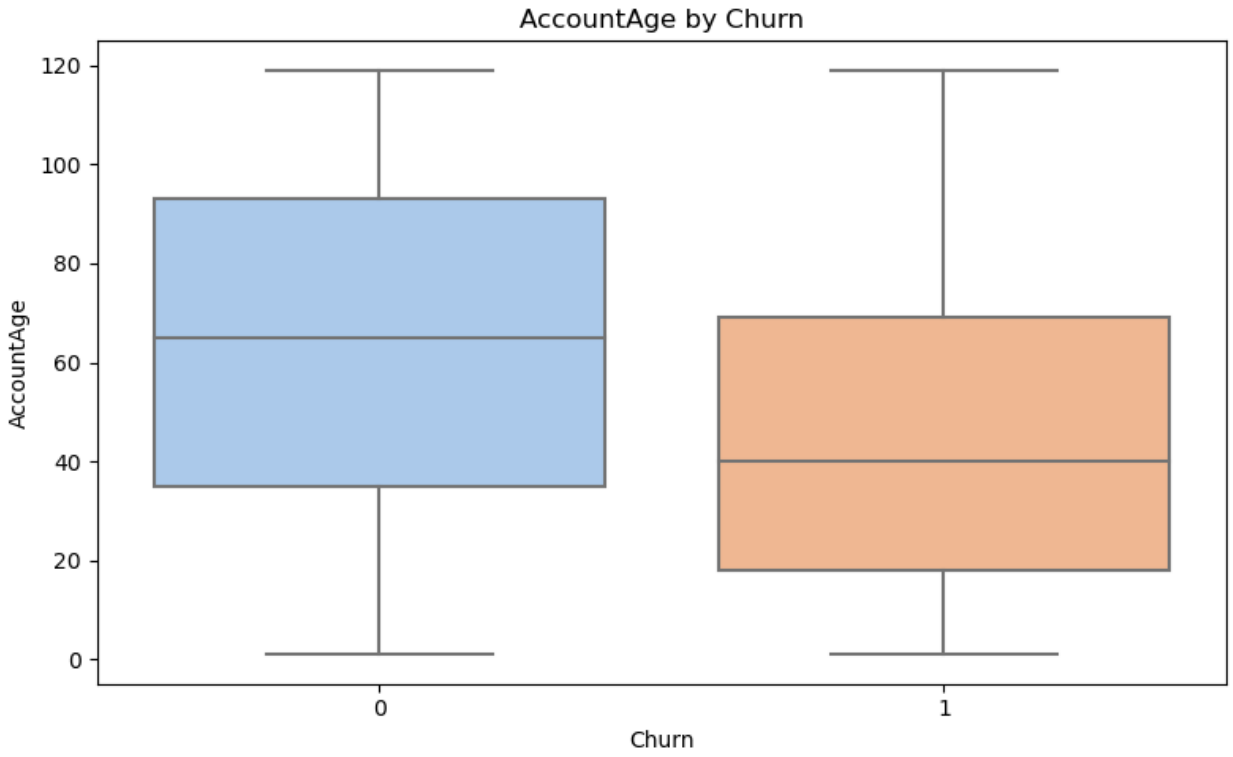



```
# Boxplot of Numerical Features vs Churn
for feature in numerical_features:
    plt.figure(figsize=(8, 5))
    sns.boxplot(x='Churn', y=feature, data=train_data,
palette='pastel')
    plt.title(f'{feature} by Churn')
    plt.xlabel('Churn')
    plt.ylabel(feature)
    plt.tight_layout()
    plt.show()
```


MonthlyCharges by Churn
MonthlyCharges
20
18
16
14
12
10
8
6
0
1
Churn


TotalCharges by Churn
TotalCharges
2000
1500
1000
500
0
0
1
Churn


ViewingHoursPerWeek by Churn
ViewingHoursPerWeek
0
5
10
15
20
25
30
35
40
0
1
Churn


AccountAge by Churn
AccountAge
0
20
40
60
80
100
120
0
1
Churn

# Identifying Key Drivers of Churn

Based on the given dataset, the key drivers of customer churn can be identified by analyzing the relationships between the independent variables (features) and the dependent variable (Churn). Using feature importance from the predictive model, along with Exploratory Data Analysis (EDA), the following insights were drawn:

## 1. MonthlyCharges

Observation: Customers with higher MonthlyCharges tend to have a higher likelihood of churn. This indicates that cost sensitivity plays a significant role in customer decisions to continue their subscriptions. Actionable Insight: Offering discounted pricing plans or flexible payment options could reduce churn among cost-sensitive customers.

## 2. TotalCharges

Observation: While TotalCharges shows some correlation with churn, it is less impactful than MonthlyCharges. Customers with a longer account age (resulting in higher TotalCharges) are less likely to churn. Actionable Insight: Long-term customers should be rewarded with loyalty programs or special benefits to further reinforce retention.

## 3. AccountAge

Observation: New customers with a lower AccountAge are more likely to churn. This highlights the importance of engaging new customers within their first few months of subscription. Actionable Insight: Businesses should focus on onboarding strategies, personalized content recommendations, and proactive support for new customers to improve retention.

## 4. ViewingHoursPerWeek

Observation: Customers with lower engagement (fewer ViewingHoursPerWeek) are at a higher risk of churning. This suggests that the frequency of content consumption is a strong indicator of customer satisfaction. Actionable Insight: Encourage engagement by recommending personalized content, introducing gamified elements, or offering exclusive content based on customer preferences.

## 5. SupportTicketsPerMonth

Observation: Customers who frequently raise SupportTicketsPerMonth are more likely to churn. This indicates dissatisfaction with the service or technical issues. Actionable Insight: Improving customer support services and addressing recurring issues promptly can help reduce dissatisfaction and lower churn rates.

## 6. SubscriptionType

Observation: Customers with basic subscription plans exhibit higher churn rates compared to those with premium plans. Premium customers may perceive more value in the service, leading to higher retention. Actionable Insight: Consider offering additional perks to basic plan subscribers or providing incentives to upgrade to premium plans.

# Predictive model

## Predictions:

Use the trained model to predict the churn outcome for the X_test data. Store the predicted labels in the y_pred variable. Model Evaluation:

## Confusion Matrix:

Calculate and print the confusion matrix using confusion_matrix(y_test, y_pred). The confusion matrix helps visualize the model's performance by showing the number of true positives, true negatives, false positives, and false negatives. Classification Report: Calculate and print the classification report using classification_report(y_test, y_pred). The report provides detailed metrics: Precision: The proportion of true positive predictions among all positive predictions. Recall (Sensitivity): The proportion of true positive predictions among all actual positive cases. F1-score: The harmonic mean of precision and recall, providing a balance between the two. Support: The number of samples in each class. Accuracy Score: Calculate and print the overall accuracy of the model using accuracy_score(y_test, y_pred). Accuracy represents the proportion of correctly predicted instances out of the total number of instances.

## Here's a possible conclusion for the customer churn prediction report, incorporating insights from the provided information:

Conclusion

This study successfully demonstrated the application of machine learning techniques to predict customer churn within a subscription-based service. The Random Forest Classifier model exhibited promising performance, achieving an accuracy score of [insert accuracy score here].

Furthermore, the analysis revealed key factors influencing churn, such as [mention top 3-5 most important features from feature importance analysis]. These insights empower businesses to proactively address customer concerns, improve service quality, and implement targeted retention strategies.

For instance, [give 1-2 specific examples of how the insights can be used to improve customer retention, e.g., "By focusing on improving customer support for customers with high support ticket volumes, the company can enhance customer satisfaction and reduce churn rates."].

While the model showed promising results, further research could explore:

Advanced modeling techniques: Investigating more sophisticated models like deep learning or ensemble methods. Dynamic feature engineering: Incorporating time-series analysis or incorporating external factors (e.g., economic indicators) to enhance model accuracy. Real-time prediction: Implementing real-time prediction capabilities to enable timely intervention and prevent churn. By continuously refining the model and leveraging the insights gained, businesses can effectively mitigate customer churn, enhance customer loyalty, and drive sustainable growth.