## What are the key factors influencing customer churn in the telecommunication industry?[¶](#What-are-the-key-factors-influencing-cu)

## How can predictive models help in identifying and preventing churn?[¶](#How-can-predictive-models-help-in-ident)

In the telecommunication industry, several factors can influence customer churn. Some key factors that we can explore in this dataset are:

International Plan: Customers with international plans may have different behaviors and needs, which can impact churn rates. Analyzing the relationship between having an international plan and churn can provide insights.

Voice Mail Plan: Similarly, customers who have a voice mail plan may have different engagement patterns, and it's worth investigating if this feature has any correlation with churn.

Usage Patterns: Features such as "total day minutes," "total eve minutes," "total night minutes," and "total intl minutes" can provide insights into how customers are utilizing the telecommunication services. High usage or low usage compared to the average customer might be indicative of potential churn.

Customer Service Calls: The number of customer service calls can reflect customer satisfaction and could be an important factor influencing churn. Higher customer service calls might be associated with a higher probability of churn.

Here's a brief explanation of each column name in the telecommunication churn dataset:

1. **'state':** This column represents each state included in this dataset
2. **'account length':** This column represents the duration of the customer's account or how long they have been using the telecommunication services. It is usually measured in some unit of time, such as days or months.
3. **'area code':** This column represents the area code associated with the customer's telephone number. It is used to identify the geographic region or location of the customer.
4. **'phone number':** This column contains the phone number associated with each customer. It uniquely identifies individual customers within the dataset.
5. **'international plan':** This column indicates whether a customer has an international plan or not. An international plan allows customers to make international calls or use international services, typically at an additional cost.
6. **'voice mail plan':** This column indicates whether a customer has a voice mail plan or not. A voice mail plan enables customers to receive and store voice messages in their mailbox when they are unavailable to answer calls.
7. **'number vmail messages':** This column represents the number of voice mail messages received by the customer. It indicates the level of engagement or utilization of the voice mail service.
8. **'total day minutes':** This column represents the total number of minutes the customer spent on calls during the daytime or specific daytime period.
9. **'total day calls':** This column represents the total number of calls made by the customer during the daytime or specific daytime period.
10. **'total day charge':** This column represents the total charges incurred by the customer for calls made during the daytime or specific daytime period.
11. **'total eve minutes':** This column represents the total number of minutes the customer spent on calls during the evening or specific evening period.
12. **'total eve calls':** This column represents the total number of calls made by the customer during the evening or specific evening period.
13. **'total eve charge':** This column represents the total charges incurred by the customer for calls made during the evening or specific evening period.
14. **'total night minutes':** This column represents the total number of minutes the customer spent on calls during the nighttime or specific nighttime period.
15. **'total night calls':** This column represents the total number of calls made by the customer during the nighttime or specific nighttime period.
16. **'total night charge':** This column represents the total charges incurred by the customer for calls made during the nighttime or specific nighttime period.
17. **'total intl minutes':** This column represents the total number of minutes the customer spent on international calls or international services.
18. **'total intl calls':** This column represents the total number of international calls made by the customer.
19. **'total intl charge':** This column represents the total charges incurred by the customer for international calls or international services.
20. **'customer service calls':** This column represents the total number of customer service calls made by the customer. It indicates the level of interaction or engagement with customer service representatives.
21. **'churn':** This column indicates whether a customer has churned (cancelled their subscription or discontinued the service) or not. It typically takes binary values, with '1' representing churn and '0' representing no churn.

In [1]:

#python libraries that are needed

import pandas as pd

import numpy as np

In [2]:

df = pd.read\_csv('telecom\_churn.csv')

df.columns

Out[2]:

Index(['state', 'account length', 'area code', 'phone number',

'international plan', 'voice mail plan', 'number vmail messages',

'total day minutes', 'total day calls', 'total day charge',

'total eve minutes', 'total eve calls', 'total eve charge',

'total night minutes', 'total night calls', 'total night charge',

'total intl minutes', 'total intl calls', 'total intl charge',

'customer service calls', 'churn'],

dtype='object')

In [3]:

df.isnull().sum()

Out[3]:

state 0

account length 0

area code 0

phone number 0

international plan 0

voice mail plan 0

number vmail messages 0

total day minutes 0

total day calls 0

total day charge 0

total eve minutes 0

total eve calls 0

total eve charge 0

total night minutes 0

total night calls 0

total night charge 0

total intl minutes 0

total intl calls 0

total intl charge 0

customer service calls 0

churn 0

dtype: int64

**Observation:** There are no missing values in the dataset, therefore the dataset is clean

### How does the presence of an international plan relate to customer churn? Are customers with international plans more likely to churn compared to those without international plans?[¶](#How-does-the-presence-of-an-internation)

In [4]:

import matplotlib.pyplot as plt

import seaborn as sns

# Group the data by 'international plan' and calculate churn rate

international\_churn\_rate = df.groupby('international plan')['churn'].mean()

# Plot the churn rate based on international plan

plt.figure(figsize=(8, 6))

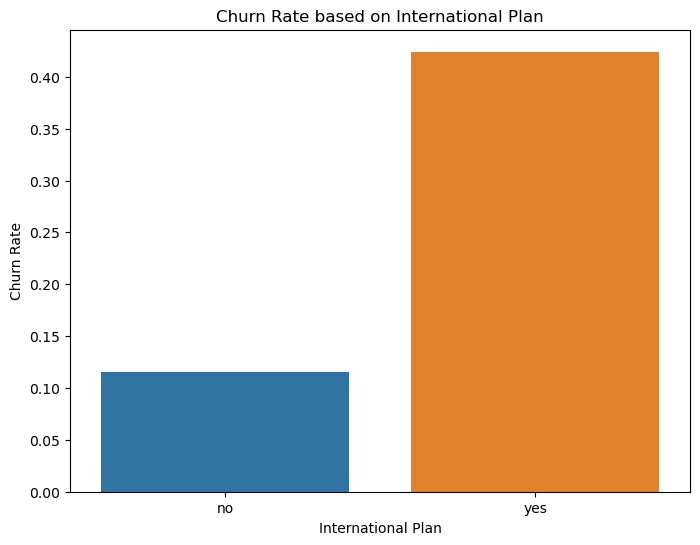
sns.barplot(x=international\_churn\_rate.index, y=international\_churn\_rate.values)

plt.xlabel('International Plan')

plt.ylabel('Churn Rate')

plt.title('Churn Rate based on International Plan')

plt.show()



**Observation:** From the bar chart above, it seems that customers using international plan tends to churn more

### Is there a significant difference in churn rates between customers who have a voice mail plan and those who don't? Does the usage of the voice mail plan impact customer retention?[¶](#Is-there-a-significant-difference-in-ch)

In [5]:

# Group the data by 'voice mail plan' and calculate churn rate

voicemail\_churn\_rate = df.groupby('voice mail plan')['churn'].mean()

# Plot the churn rate based on voice mail plan

plt.figure(figsize=(8, 6))

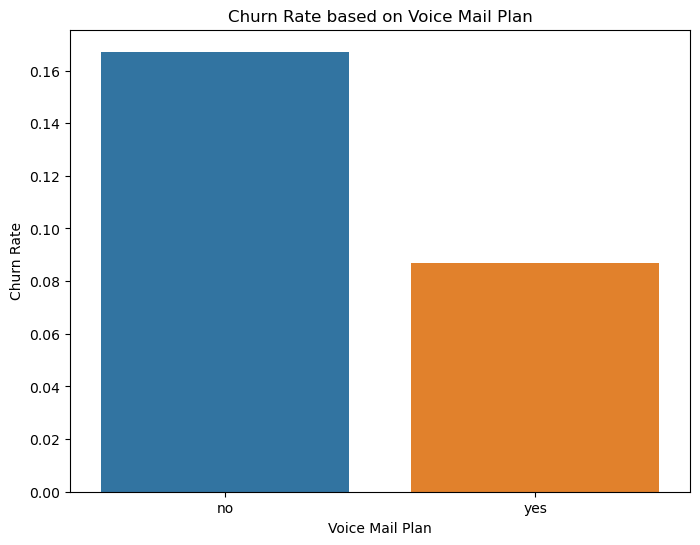
sns.barplot(x=voicemail\_churn\_rate.index, y=voicemail\_churn\_rate.values)

plt.xlabel('Voice Mail Plan')

plt.ylabel('Churn Rate')

plt.title('Churn Rate based on Voice Mail Plan')

plt.show()



**Observation:** It seems that a voice mail plan is beneficial in reducing customers churn rate

# What is the average total day minutes used by customers based on their international plan status?[¶](#What-is-the-average-total-day-minutes-u)

The Aim Of This Question:

* To analyze if there's a significant difference in average total day minutes used between customers with an international plan and those without, which could help in understanding the impact of international plans on usage patterns.

In [6]:

avg\_day = df.pivot\_table(values='total day minutes', index='international plan', aggfunc='mean')

avg\_day

Out[6]:

|  | **total day minutes** |
| --- | --- |
| **international plan** |  |
| **no** | 178.893887 |
| **yes** | 187.986997 |

**Observations:**

* There's a difference in average total day minutes between cutomers using international plans and those not using an international plan

# Do customers who make more customer service calls have a higher likelihood of churn? Is there a threshold beyond which the number of customer service calls becomes a strong predictor of churn?[¶](#Do-customers-who-make-more-customer-ser)

Aim:

* To explore if there's a correlation between the number of customer service calls made and customer churn. This analysis could provide insights into whether higher levels of customer service calls are indicative of potential churn and guide strategies to reduce churn.

In [7]:

avg\_da = df.pivot\_table(values='churn', index='customer service calls', aggfunc='mean')

avg\_da

Out[7]:

|  | **churn** |
| --- | --- |
| **customer service calls** |  |
| **0** | 0.131994 |
| **1** | 0.103302 |
| **2** | 0.114625 |
| **3** | 0.102564 |
| **4** | 0.457831 |
| **5** | 0.606061 |
| **6** | 0.636364 |
| **7** | 0.555556 |
| **8** | 0.500000 |
| **9** | 1.000000 |

In [8]:

# Create a bar plot to visualize it

avg\_da.plot(kind='bar', color='blue', alpha=0.7)

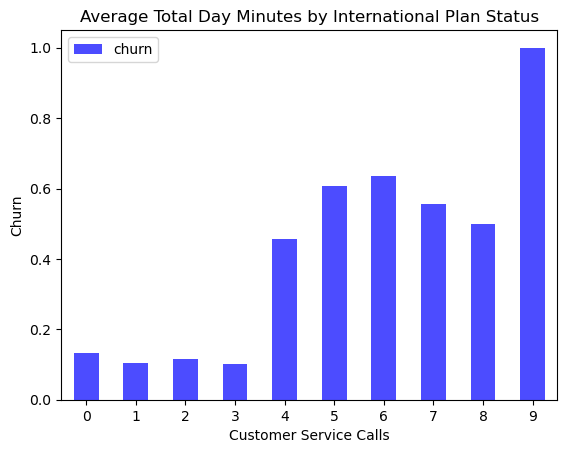
plt.title('Average Total Day Minutes by International Plan Status')

plt.xlabel('Customer Service Calls')

plt.ylabel('Churn')

plt.xticks(rotation=0)

plt.show()



**Observation:** The 'customer service calls' column in the dataset, ranging from 0 to 9, represents the total number of customer service calls made by each customer.

Here's what each value in the column typically signifies: 0: The customer did not make any customer service calls. 1-9: The customer made one or more customer service calls, with the value indicating the specific count of calls made.

In this case, the column represents a discrete numerical feature that captures the engagement level of customers with the customer service department. A higher value in this column suggests that the customer has had more interactions with customer service, potentially indicating concerns, issues, or queries they had regarding the telecommunication services.

Customers with a higher number of customer service calls may be more likely to churn due to unresolved issues or dissatisfaction

# What's the average total day charge for customers who churned and those who didn't, based on their international plan status?[¶](#What's-the-average-total-day-charge-for)

Aim:

* To explore if there's a relationship between churn, international plan usage and average total day charge. This analysis can help identify whether customers with international plans who churned had significantly different day charges compared to those who didn't churn

In [9]:

churn\_custom = df.pivot\_table(values= 'total day charge',

index= ['churn', 'international plan'],

aggfunc= 'mean').reset\_index()

churn\_custom

Out[9]:

|  | **churn** | **international plan** | **total day charge** |
| --- | --- | --- | --- |
| **0** | False | no | 29.677646 |
| **1** | False | yes | 31.252419 |
| **2** | True | no | 36.070405 |
| **3** | True | yes | 32.916861 |

In [10]:

import seaborn as sns

sns.barplot(

x='churn',

y='total day charge',

hue= 'international plan',

data = churn\_custom

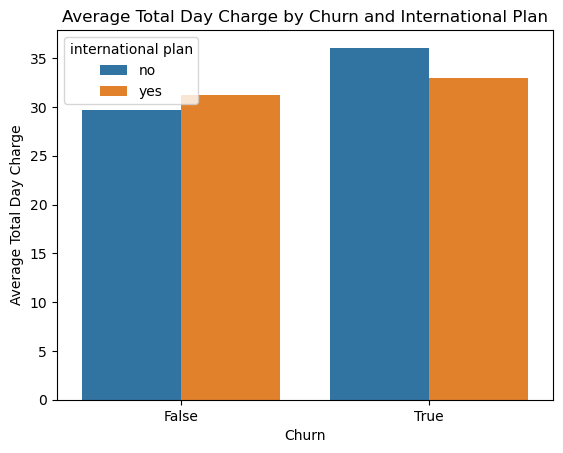
)

plt.xlabel('Churn')

plt.ylabel('Average Total Day Charge')

plt.title('Average Total Day Charge by Churn and International Plan')

plt.show()



**Observations:**

* From the barplot above, it seems that customers who did not use international plans seemed to churn and have a higher total day charge when compared to customers that used international plans and churned
* However, the difference in average total day charge between customers who have churned and customers who have not churned is smaller for customers with an international plan than for customers without an international plan.

### How does the average 'total day charge' vary across different 'area code' regions?[¶](#How-does-the-average-'total-day-charge')

In [11]:

# Calculate the average total day charge for each area code

avg\_day\_charge\_by\_area = df.groupby('area code')['total day charge'].mean().reset\_index()

# Plot the average total day charge by area code

plt.figure(figsize=(10, 6))

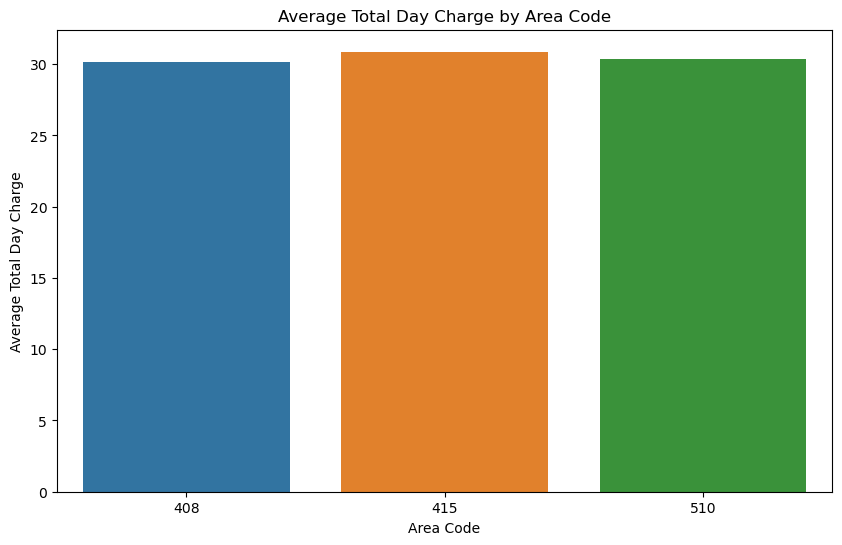
sns.barplot(x='area code', y='total day charge', data=avg\_day\_charge\_by\_area)

plt.xlabel('Area Code')

plt.ylabel('Average Total Day Charge')

plt.title('Average Total Day Charge by Area Code')

plt.show()



Observations

* The average total day charge is highest for area code 415.
* The average total day charge is lowest for area code 408 and 510.
* The difference in the average total day charge between the three area codes may be due to different factors, such as the cost of living in the area, the demographics of the population, or the competition from other telecommunications providers.
* The company could use this information to target its marketing efforts to specific area codes. For example, the company could offer discounts to customers in area code 408, or they could focus on increasing customer satisfaction in area code 510.

### What is the distribution of the 'number vmail messages' feature among customers who have a voice mail plan?[¶](#What-is-the-distribution-of-the-'number)

In [12]:

# Subset the dataset for customers with a voice mail plan

voicemail\_df = df[df['voice mail plan'] == 'yes']

# Plot the distribution of number voice mail messages

plt.figure(figsize=(8, 6))

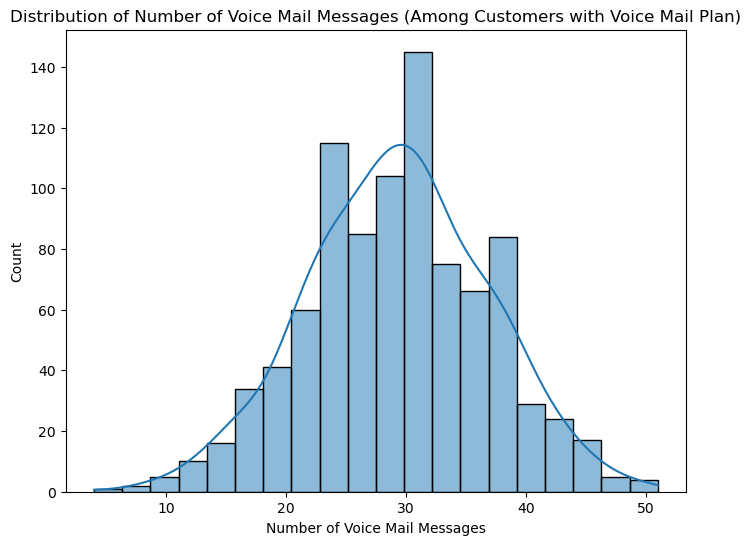
sns.histplot(voicemail\_df['number vmail messages'], bins=20, kde=True)

plt.xlabel('Number of Voice Mail Messages')

plt.ylabel('Count')

plt.title('Distribution of Number of Voice Mail Messages (Among Customers with Voice Mail Plan)')

plt.show()



Observations:

* The most number of customers have 30 voicemail messages, that's approx 140 customers.
* The number of customers decreases as the number of voicemail messages increases.
* The distribution of the number of voicemail messages may be due to different factors, such as the type of phone plan customers have, the frequency of calls they make, or their personal preferences.

# The Relationship between total intl calls and total intl minutes[¶](#The-Relationship-between-total-intl-cal)

### What is the distribution of 'total night calls' for churned and retained customers?[¶](#What-is-the-distribution-of-'total-nigh)

In [13]:

# Subset the dataset for churned and retained customers

churned\_df = df[df['churn'] == 1]

retained\_df = df[df['churn'] == 0]

# Plot the distribution of total night calls for churned and retained customers

plt.figure(figsize=(8, 6))

sns.kdeplot(churned\_df['total night calls'], label='Churned')

sns.kdeplot(retained\_df['total night calls'], label='Retained')

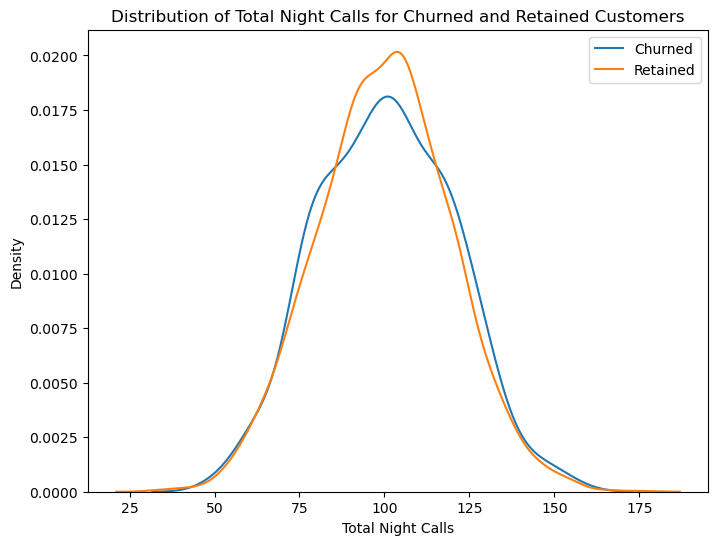
plt.xlabel('Total Night Calls')

plt.ylabel('Density')

plt.title('Distribution of Total Night Calls for Churned and Retained Customers')

plt.legend()

plt.show()



Observations:

* The distribution of the total night calls for retained customers is more spread out than the distribution for churned customers.
* There is a higher concentration of churned customers in the lower total night calls range.
* There is a higher concentration of retained customers in the higher total night calls range.
* This suggests that churned customers are more likely to call during the night, while retained customers are more likely to call during the day.
* The company could use this information to improve its customer service for night time callers.
* For example, the company could offer a way for customers to make night time calls more easily or to get help with night time billing.
* The company could also target its marketing campaigns to customers who are more likely to call during the night.
* For example, the company could offer discounts on night time calls or data plans to these customers.

### How does the 'customer service calls' vary based on whether a customer has an 'international plan' or a 'voice mail plan'?[¶](#How-does-the-'customer-service-calls'-v)

In [14]:

# Plot the relationship between customer service calls and international plan

plt.figure(figsize=(8, 6))

sns.boxplot(x='customer service calls', y='international plan', data=df)

plt.xlabel('Customer Service Calls')

plt.ylabel('International Plan')

plt.title('Relationship between Customer Service Calls and International Plan')

plt.show()

# Plot the relationship between customer service calls and voice mail plan

plt.figure(figsize=(8, 6))

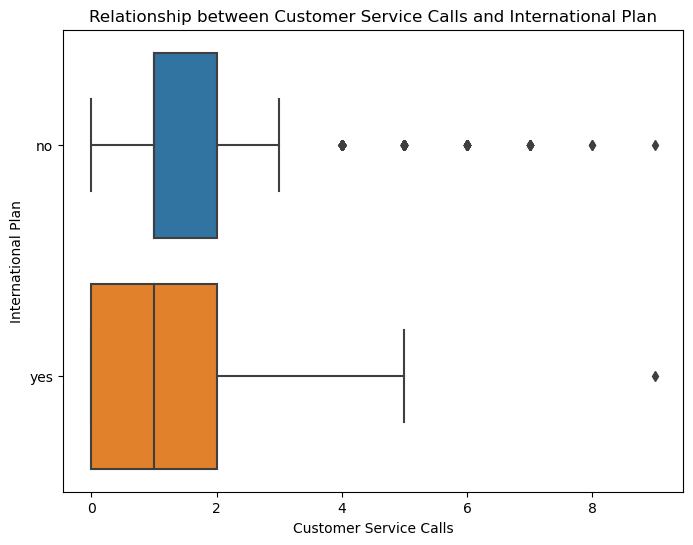
sns.boxplot(x='customer service calls', y='voice mail plan', data=df)

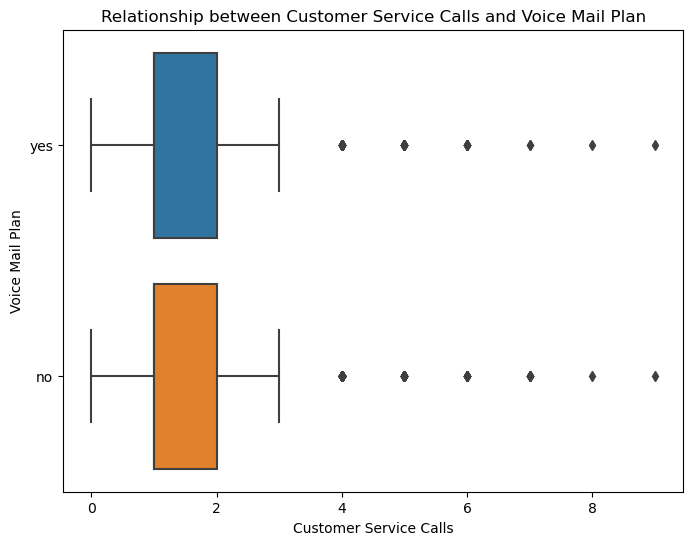
plt.xlabel('Customer Service Calls')

plt.ylabel('Voice Mail Plan')

plt.title('Relationship between Customer Service Calls and Voice Mail Plan')

plt.show()





Observations:

* There are a few outliers in the data.
* These outliers may be due to customers who made a large number of customer service calls for reasons unrelated to their international plan, or they may be due to errors in the data
* The difference in the number of customer service calls between customers with and without an international plan may be due to different factors, such as the complexity of the international plan, the level of customer support offered, or the number of international calls made.
* The company could use this information to improve its customer service for customers with international plans.
* For example, the company could provide more training to customer service representatives on international plans, or it could offer a way for customers to make international calls more easily.
* There are more customers with a voice mail plan than without a voice mail plan.
* The number of customer service calls is lower for customers with a voice mail plan than for customers without a voice mail plan.
* This suggests that voice mail plans may be effective in reducing the number of customer service calls.
* The difference in the number of customer service calls between customers with and without a voice mail plan may be due to different factors, such as the complexity of the voice mail plan, the level of customer support offered, or the number of voicemail messages left.
* The company could also use this information to improve its marketing campaigns for voice mail plans.

### Which states have the highest churn rate?[¶](#Which-states-have-the-highest-churn-rat)

In [15]:

# Calculate churn rate for each state

state\_churn\_rate = df.groupby('state')['churn'].mean().reset\_index()

# Sort states by churn rate in descending order

state\_churn\_rate = state\_churn\_rate.sort\_values('churn', ascending=False)

# Plot the states with highest churn rates

top\_states = state\_churn\_rate.head(10) # Change the number to show more or fewer states

plt.figure(figsize=(10, 6))

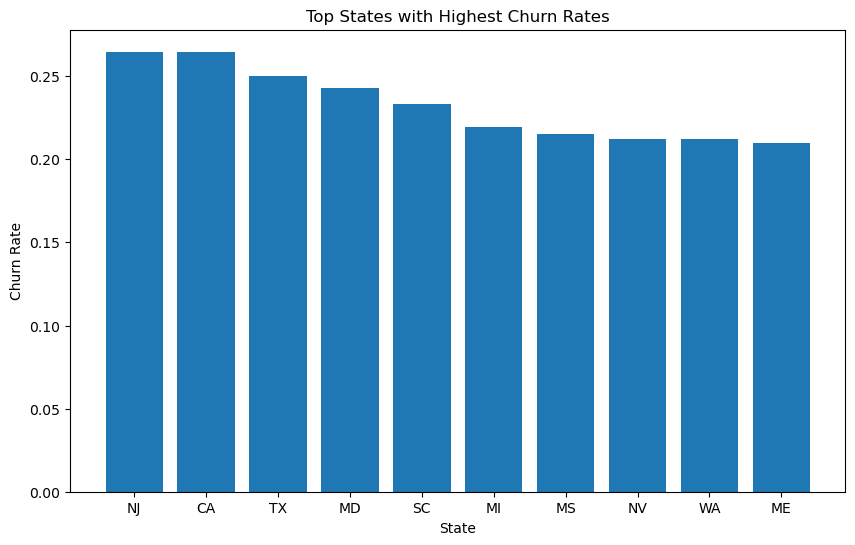
plt.bar(top\_states['state'], top\_states['churn'])

plt.xlabel('State')

plt.ylabel('Churn Rate')

plt.title('Top States with Highest Churn Rates')

plt.show()



**Observation:** New Jersey and California are the states with the highes churn rate.

* I suggest that the company should focus on improving their services in these states and this could lead to a reduction in the churn rate

### Are there any regional patterns in customer churn?[¶](#Are-there-any-regional-patterns-in-cust)

In [16]:

# Define the regions and their corresponding states

regions = {

'Northeast': ['CT', 'ME', 'MA', 'NH', 'RI', 'VT', 'NY', 'NJ', 'PA'],

'Midwest': ['IL', 'IN', 'MI', 'OH', 'WI', 'IA', 'KS', 'MN', 'MO', 'NE', 'ND', 'SD'],

'South': ['DE', 'FL', 'GA', 'MD', 'NC', 'SC', 'VA', 'DC', 'WV', 'AL', 'KY', 'MS', 'TN', 'AR', 'LA', 'OK', 'TX'],

'West': ['AZ', 'CO', 'ID', 'MT', 'NV', 'NM', 'UT', 'WY', 'AK', 'CA', 'HI', 'OR', 'WA']

}

# Map the states to their corresponding regions

df['region'] = df['state'].map({state: region for region, states in regions.items() for state in states})

# Calculate the churn rate for each region

region\_churn\_rate = df.groupby('region')['churn'].mean().reset\_index()

# Sort regions by churn rate in descending order

region\_churn\_rate = region\_churn\_rate.sort\_values('churn', ascending=False)

# Plot the churn rate for each region

plt.figure(figsize=(10, 6))

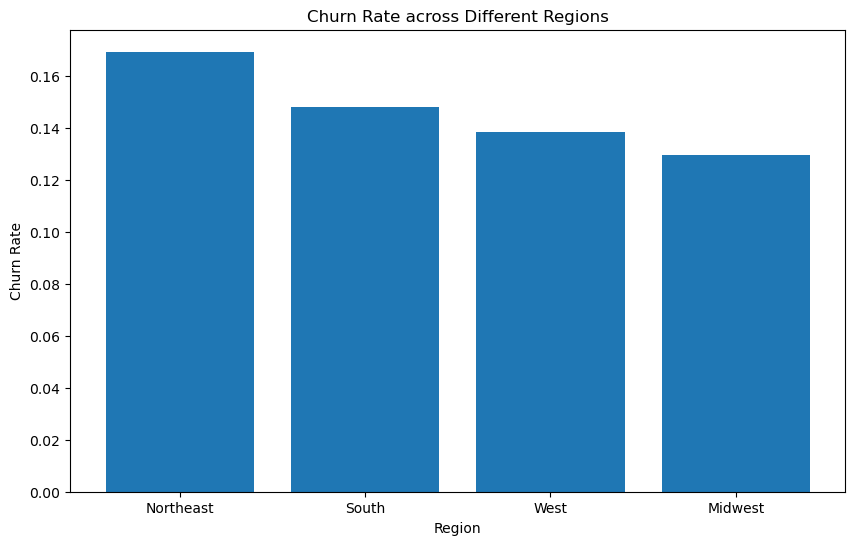
plt.bar(region\_churn\_rate['region'], region\_churn\_rate['churn'])

plt.xlabel('Region')

plt.ylabel('Churn Rate')

plt.title('Churn Rate across Different Regions')

plt.show()



**Observation:** The Northeast regions has a higher rate of churn. I suggest that the telecommunication company should make this region a priority and engage its customers as this might reduce the rate of customer churn

In [ ]:

In [ ]: