01_eda

October 21, 2024

1 Exploratory Data Analysis of SyriaTel Customer Churn Dataset

In this notebook, I will explore the customer churn dataset to understand the structure of the data, investigate patterns, and derive insights that will guide the preprocessing and modeling steps in later stages.

They key objectives of this EDA include: - Understanding the distribution of the target variable (`churn`). - Investigating the relationships between features and churn. - Identifying any missing data or outliers. - Exploring potential features to include or exclude from the final model.

2 Loading and Understanding the Dataset

I will begin by loading the data set and reviewing its structure. This includes checking the data types, looking for missing values, and getting a summary of numerical features.

[1]:
```
import pandas as pd

pd.set_option('display.max_columns', None)

data_path = '../data/raw/telecom_churn_dataset.csv'
df = pd.read_csv(data_path)

df
```

[1]:
```
     state  account length  area code phone number international plan  \
0       KS             128        415     382-4657                 no
1       OH             107        415     371-7191                 no
2       NJ             137        415     358-1921                 no
3       OH              84        408     375-9999                yes
4       OK              75        415     330-6626                yes
…    …              …         …           …                   …
3328    AZ             192        415     414-4276                 no
3329    WV              68        415     370-3271                 no
3330    RI              28        510     328-8230                 no
3331    CT             184        510     364-6381                yes
3332    TN              74        415     400-4344                 no

     voice mail plan  number vmail messages  total day minutes  \
```

```
0                yes                      25              265.1
1                yes                      26              161.6
2                 no                       0              243.4
3                 no                       0              299.4
4                 no                       0              166.7
…               …                        …               …
3328             yes                      36              156.2
3329              no                       0              231.1
3330              no                       0              180.8
3331              no                       0              213.8
3332             yes                      25              234.4

      total day calls  total day charge  total eve minutes  total eve calls  \
0                 110             45.07              197.4               99
1                 123             27.47              195.5              103
2                 114             41.38              121.2              110
3                  71             50.90               61.9               88
4                 113             28.34              148.3              122
…                 …                …                  …                  …
3328               77             26.55              215.5              126
3329               57             39.29              153.4               55
3330              109             30.74              288.8               58
3331              105             36.35              159.6               84
3332              113             39.85              265.9               82

      total eve charge  total night minutes  total night calls  \
0                16.78                244.7                 91
1                16.62                254.4                103
2                10.30                162.6                104
3                 5.26                196.9                 89
4                12.61                186.9                121
…                …                    …                    …
3328             18.32                279.1                 83
3329             13.04                191.3                123
3330             24.55                191.9                 91
3331             13.57                139.2                137
3332             22.60                241.4                 77

      total night charge  total intl minutes  total intl calls  \
0                  11.01                10.0                 3
1                  11.45                13.7                 3
2                   7.32                12.2                 5
3                   8.86                 6.6                 7
4                   8.41                10.1                 3
…                   …                   …                   …
3328               12.56                 9.9                 6
3329                8.61                 9.6                 4
```

```
3330                8.64                14.1                    6
3331                6.26                 5.0                   10
3332               10.86                13.7                    4

      total intl charge  customer service calls  churn
0                  2.70                       1  False
1                  3.70                       1  False
2                  3.29                       0  False
3                  1.78                       2  False
4                  2.73                       3  False
…                     …                       …      …
3328               2.67                       2  False
3329               2.59                       3  False
3330               3.81                       2  False
3331               1.35                       2  False
3332               3.70                       0  False

[3333 rows x 21 columns]
```

[2]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   state                   3333 non-null   object
 1   account length          3333 non-null   int64
 2   area code               3333 non-null   int64
 3   phone number            3333 non-null   object
 4   international plan      3333 non-null   object
 5   voice mail plan         3333 non-null   object
 6   number vmail messages   3333 non-null   int64
 7   total day minutes       3333 non-null   float64
 8   total day calls         3333 non-null   int64
 9   total day charge        3333 non-null   float64
 10  total eve minutes       3333 non-null   float64
 11  total eve calls         3333 non-null   int64
 12  total eve charge        3333 non-null   float64
 13  total night minutes     3333 non-null   float64
 14  total night calls       3333 non-null   int64
 15  total night charge      3333 non-null   float64
 16  total intl minutes      3333 non-null   float64
 17  total intl calls        3333 non-null   int64
 18  total intl charge       3333 non-null   float64
 19  customer service calls  3333 non-null   int64
 20  churn                   3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(4)
```

```
memory usage: 524.2+ KB
```

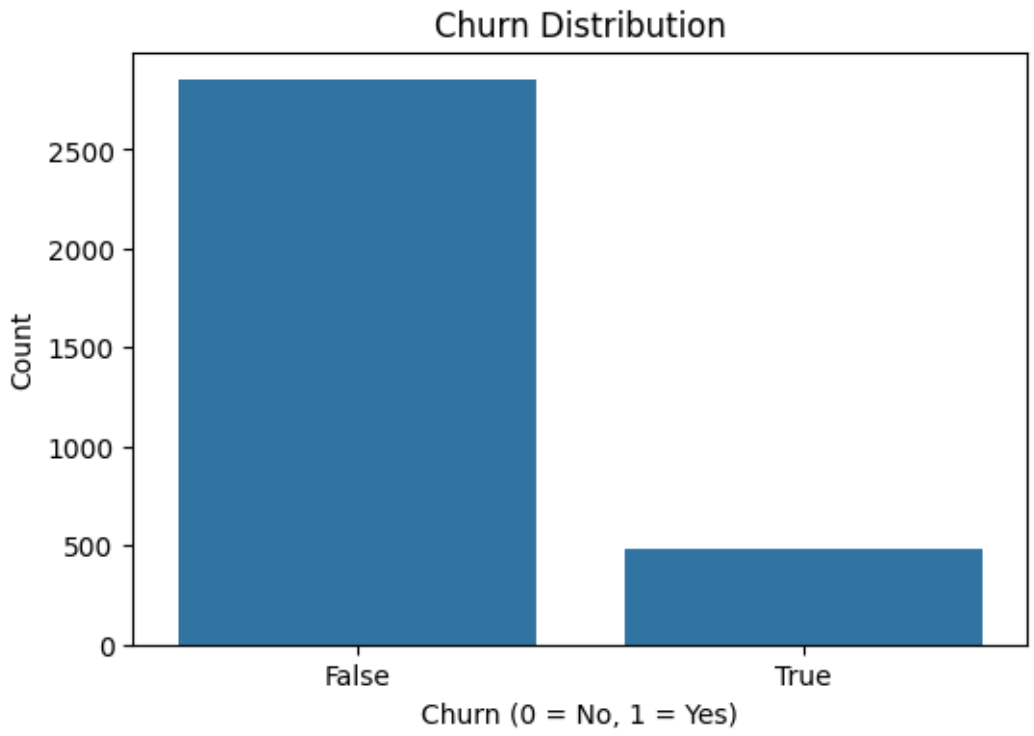
[3]:
```
df.isna().sum()
```

[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
```

2.1 Churn Distribution Chart

Plotting a Churn Distribution Chart to better visualize if any class imbalance is present.

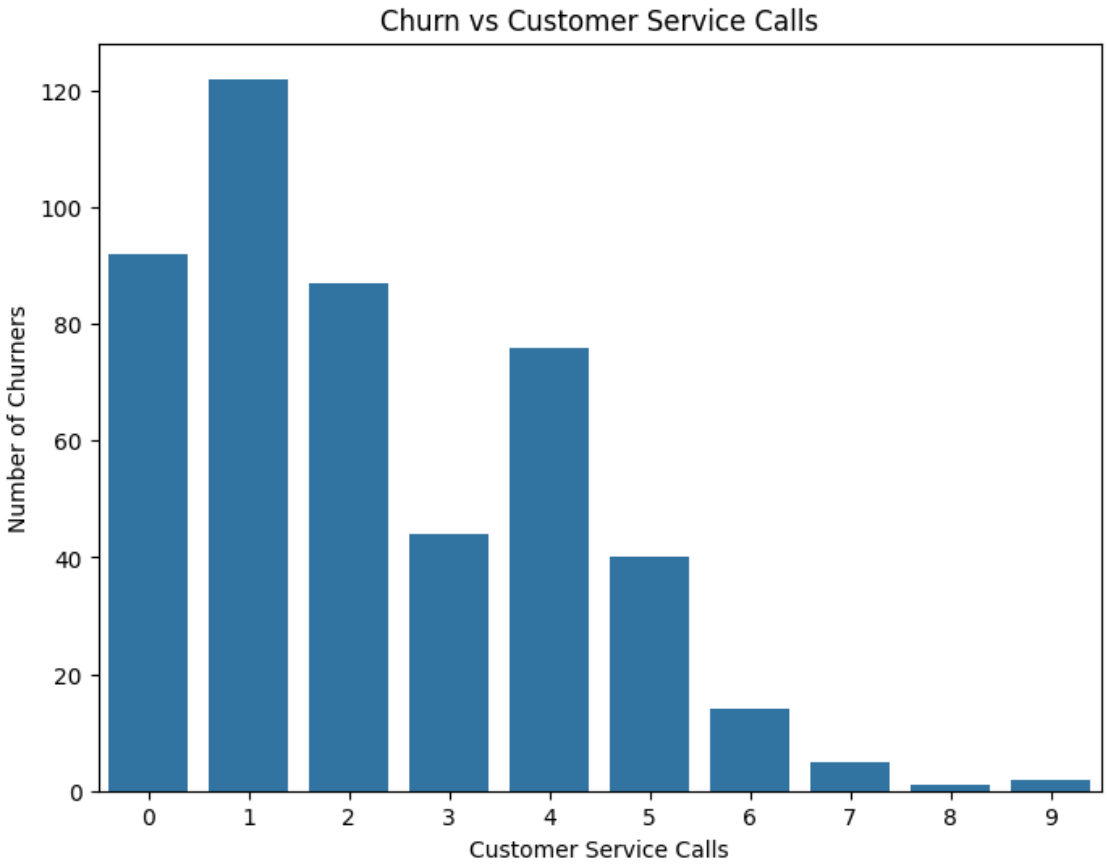
[4]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# plotting churn distribution
plt.figure(figsize=(6, 4))
sns.countplot(x='churn', data=df)
plt.title('Churn Distribution')
plt.xlabel('Churn (0 = No, 1 = Yes)')
plt.ylabel('Count')
plt.show()
```



2.2 Churn vs. Customer Service Calls

I wanted to look into how customer service call frequency correlates with churn, providing insights into risk factors.

[5]:
```
plt.figure(figsize=(8, 6))
sns.barplot(x='customer service calls', y='churn', data=df, estimator=sum,␣
 ↪errorbar=None)
plt.title('Churn vs Customer Service Calls')
plt.ylabel('Number of Churners')
plt.xlabel('Customer Service Calls')
plt.show()
```



2.3 Exploring Categorical Variables

I plan to investigate the categorical variables in the dataset, such as `international_plan` and `voice_mail_plan`, to understand their distributions and potential relationships with customer churn.

Currently, the `state` and `phone_number` columns are also included in the categorical variables. However, these columns are not relevant to predicting customer churn. Therefore, they will be excluded during the preprocessing phase, as they serve no meaningful purpose in modeling and might introduce noise into the dataset.

These features will later be encoded to numerical values during the preprocessing phase.

[6]:
```
def get_uniques(df, columns):
    return {column: list(df[column].unique()) for column in columns}
```

[7]:
```
def get_categorical_columns(df):
    return [column for column in df.columns if df.dtypes[column] == 'object']
```

[8]:
```
get_uniques(df, get_categorical_columns(df))
```

```
[8]: {'state': ['KS',
        'OH',
        'NJ',
        'OK',
        'AL',
        'MA',
        'MO',
        'LA',
        'WV',
        'IN',
        'RI',
        'IA',
        'MT',
        'NY',
        'ID',
        'VT',
        'VA',
        'TX',
        'FL',
        'CO',
        'AZ',
        'SC',
        'NE',
        'WY',
        'HI',
        'IL',
        'NH',
        'GA',
        'AK',
        'MD',
        'AR',
        'WI',
        'OR',
        'MI',
        'DE',
        'UT',
        'CA',
        'MN',
        'SD',
        'NC',
        'WA',
        'NM',
        'NV',
        'DC',
        'KY',
        'ME',
        'MS',
```

```
 'TN',
 'PA',
 'CT',
 'ND'],
'phone number': ['382-4657',
 '371-7191',
 '358-1921',
 '375-9999',
 '330-6626',
 '391-8027',
 '355-9993',
 '329-9001',
 '335-4719',
 '330-8173',
 '329-6603',
 '344-9403',
 '363-1107',
 '394-8006',
 '366-9238',
 '351-7269',
 '350-8884',
 '386-2923',
 '356-2992',
 '373-2782',
 '396-5800',
 '393-7984',
 '358-1958',
 '350-2565',
 '343-4696',
 '331-3698',
 '357-3817',
 '418-6412',
 '353-2630',
 '410-7789',
 '416-8428',
 '370-3359',
 '383-1121',
 '360-1596',
 '395-2854',
 '362-1407',
 '341-9764',
 '353-3305',
 '402-1381',
 '332-9891',
 '372-9976',
 '383-6029',
 '353-7289',
```

```
'390-7274',
'352-1237',
'353-3061',
'363-5450',
'364-1995',
'398-1294',
'405-7146',
'413-4957',
'420-5645',
'349-4396',
'404-3211',
'353-3759',
'363-5947',
'340-5121',
'370-7574',
'403-9733',
'355-7251',
'359-5893',
'405-3371',
'344-5117',
'332-8160',
'359-4081',
'352-8305',
'329-9847',
'365-9011',
'338-9472',
'374-8042',
'359-1231',
'413-7170',
'415-2935',
'399-4246',
'362-5889',
'350-8921',
'374-5353',
'360-1171',
'355-8887',
'333-1967',
'354-4577',
'331-7425',
'419-2637',
'411-1530',
'395-3026',
'388-6441',
'402-1251',
'412-9997',
'346-7302',
'358-9095',
```

```
'400-9770',
'334-1275',
'340-4953',
'400-9510',
'387-6103',
'366-4467',
'370-3450',
'327-3954',
'355-6291',
'362-9748',
'379-6506',
'347-7741',
'354-3783',
'401-7594',
'397-4976',
'334-2577',
'400-3637',
'383-4361',
'371-4306',
'403-4298',
'409-3786',
'337-4697',
'383-1509',
'359-9794',
'407-7035',
'363-1069',
'391-4652',
'355-6837',
'409-1244',
'328-3266',
'352-7072',
'370-7550',
'369-5526',
'329-4391',
'408-4195',
'354-4445',
'335-4858',
'414-8718',
'409-5939',
'331-4902',
'353-6870',
'355-2909',
'390-6101',
'400-3446',
'411-5859',
'387-2919',
'374-8525',
```

```
'379-5592',
'345-8237',
'422-6690',
'346-2359',
'374-3534',
'381-4756',
'390-2805',
'390-2390',
'419-9097',
'386-7281',
'380-3561',
'390-8760',
'366-6730',
'395-5285',
'354-3436',
'336-7600',
'383-6293',
'362-4596',
'401-3926',
'370-9116',
'328-6289',
'350-9994',
'351-4616',
'360-5779',
'417-4885',
'406-4710',
'409-8743',
'335-4584',
'361-9845',
'366-5699',
'329-9364',
'390-7434',
'404-9680',
'338-9398',
'394-2445',
'381-2709',
'397-5060',
'415-2393',
'377-1765',
'409-2111',
'401-3170',
'405-5681',
'411-4582',
'355-5009',
'372-3750',
'405-2888',
'361-3337',
```

```
'350-1639',
'333-3221',
'422-1471',
'399-7865',
'373-4819',
'338-6981',
'418-4365',
'359-5461',
'375-3586',
'407-8376',
'408-6496',
'385-7688',
'332-6934',
'378-3625',
'353-7292',
'399-6786',
'358-3261',
'377-9932',
'397-4030',
'367-1062',
'341-8467',
'339-9453',
'344-3388',
'375-8013',
'408-4142',
'386-3671',
'411-2284',
'346-7795',
'333-5609',
'405-1842',
'366-6345',
'337-9345',
'328-6770',
'380-7321',
'375-1476',
'356-1567',
'422-6685',
'336-1090',
'343-2095',
'345-3934',
'338-8050',
'388-9568',
'402-6591',
'403-6419',
'386-9790',
'378-5692',
'360-3324',
```

```
'410-3719',
'352-4221',
'327-6179',
'359-6196',
'374-9107',
'357-4078',
'366-5780',
'393-9619',
'355-9295',
'400-5751',
'338-1027',
'405-8867',
'336-5616',
'335-1697',
'331-5138',
'385-8240',
'348-1359',
'354-7339',
'349-1687',
'380-2558',
'365-2153',
'345-6043',
'349-2808',
'411-1715',
'385-2488',
'377-7177',
'342-1099',
'386-4170',
'413-1269',
'396-4460',
'334-2730',
'340-3182',
'377-8608',
'417-3676',
'417-6774',
'411-9554',
'420-3192',
'389-1475',
'343-7734',
'410-3390',
'344-6495',
'331-6229',
'337-7501',
'339-9631',
'369-4384',
'416-3915',
'339-3049',
```

```
'361-7998',
'355-4842',
'387-6440',
'369-2625',
'389-7073',
'370-8463',
'362-7318',
'412-1194',
'355-9508',
'352-8202',
'335-5882',
'352-6976',
'393-6733',
'335-1838',
'355-6930',
'387-5860',
'343-2605',
'350-6759',
'371-1514',
'346-9317',
'398-4313',
'412-4399',
'330-1835',
'416-1676',
'329-7347',
'360-6868',
'405-6641',
'393-2373',
'419-1714',
'336-3819',
'341-3464',
'413-5310',
'366-7912',
'399-8845',
'368-2583',
'360-6309',
'359-5890',
'332-2462',
'381-9196',
'329-3222',
'363-5819',
'413-9269',
'330-7483',
'403-7775',
'360-2479',
'394-3791',
'384-2632',
```

```
'359-8466',
'331-8909',
'359-5160',
'330-9833',
'362-2314',
'338-8478',
'387-5453',
'380-3437',
'365-8779',
'407-2750',
'396-8265',
'397-4304',
'333-2611',
'409-8814',
'336-5406',
'343-6940',
'361-9923',
'350-6639',
'376-4300',
'349-6567',
'333-7749',
'408-6089',
'375-2165',
'400-6999',
'420-7823',
'366-5241',
'413-3412',
'406-2752',
'337-8078',
'402-1942',
'371-7917',
'343-6374',
'385-8730',
'393-7892',
'407-6748',
'341-4463',
'351-2587',
'421-9752',
'356-4001',
'328-9869',
'343-5709',
'334-1872',
'350-1040',
'369-3214',
'385-6778',
'383-7689',
'385-5722',
```

```
'357-1909',
'364-3567',
'422-4241',
'370-2957',
'329-6562',
'363-3515',
'374-7787',
'345-2931',
'373-5732',
'348-7437',
'332-9460',
'355-6531',
'336-9390',
'346-8581',
'363-8824',
'353-3351',
'360-4320',
'417-6252',
'393-4949',
'401-3156',
'338-6283',
'352-9017',
'405-5305',
'376-9249',
'339-7139',
'328-6011',
'378-1303',
'402-5155',
'333-5430',
'365-9696',
'410-4023',
'411-7649',
'338-4065',
'421-9401',
'343-9658',
'332-5521',
'349-4369',
'351-4288',
'422-5874',
'396-2324',
'416-5662',
'363-9663',
'410-9477',
'352-4418',
'361-6563',
'417-4404',
'372-8048',
```

```
'356-3646',
'351-9604',
'355-9581',
'396-5189',
'356-9187',
'394-5537',
'408-2712',
'404-4486',
'355-1113',
'411-4674',
'376-4519',
'365-5979',
'382-2879',
'420-1383',
'411-7390',
'383-8848',
'387-9301',
'399-3164',
'385-8997',
'352-6573',
'408-3384',
'419-6033',
'336-2090',
'343-7242',
'376-1713',
'381-5878',
'390-5470',
'414-4803',
'382-5478',
'333-7637',
'341-1647',
'411-4232',
'339-2616',
'327-3850',
'328-7209',
'405-6189',
'418-6737',
'366-2212',
'356-1420',
'343-1323',
'361-8239',
'384-1621',
'360-3525',
'392-3813',
'337-6898',
'366-8036',
'352-8327',
```

```
'334-9505',
'336-5702',
'392-2381',
'369-6880',
'416-8697',
'345-3451',
'379-2514',
'418-6651',
'421-3528',
'329-9046',
'406-3890',
'403-8904',
'393-4086',
'400-1367',
'377-9473',
'396-3068',
'331-9293',
'347-1914',
'395-1962',
'401-5485',
'355-6560',
'363-3911',
'345-1998',
'361-5277',
'376-7145',
'375-2975',
'376-8573',
'366-7360',
'347-7898',
'390-7328',
'356-5491',
'373-3251',
'343-1965',
'378-8019',
'386-1548',
'397-1649',
'366-7247',
'402-9691',
'334-9806',
'378-7733',
'407-2248',
'405-3916',
'407-2081',
'397-9148',
'415-4857',
'354-7314',
'346-5611',
```

```
'349-4703',
'411-7778',
'421-1469',
'420-5990',
'389-4780',
'357-2735',
'409-4791',
'380-5286',
'394-8402',
'392-1616',
'364-1969',
'390-4152',
'367-7039',
'391-6607',
'379-6652',
'384-1833',
'403-2455',
'391-1348',
'408-4174',
'366-4334',
'406-5059',
'373-6784',
'408-3532',
'350-8680',
'398-9870',
'343-3356',
'415-4609',
'404-5387',
'415-8151',
'416-2778',
'393-3300',
'391-7661',
'339-4317',
'418-6455',
'378-3508',
'390-9359',
'364-2495',
'364-7719',
'421-1189',
'419-8987',
'402-9980',
'376-5908',
'400-3150',
'336-1749',
'420-1259',
'339-7541',
'378-9029',
```

```
'342-7514',
'422-4956',
'389-8606',
'406-7261',
'417-7973',
'390-2891',
'385-7387',
'362-2776',
'329-1955',
'344-3160',
'406-2454',
'335-3913',
'355-7705',
'410-7108',
'419-1674',
'351-7369',
'349-9566',
'333-3421',
'393-8199',
'388-4879',
'353-6007',
'416-1557',
'356-7217',
'350-2012',
'420-9838',
'373-6379',
'355-7293',
'406-4588',
'345-1524',
'375-8493',
'361-2924',
'359-6163',
'411-8140',
'381-9049',
'344-4478',
'360-2690',
'410-7383',
'356-1889',
'341-2360',
'370-3021',
'336-4656',
'386-2810',
'350-1354',
'346-5707',
'405-8370',
'373-2053',
'369-5222',
```

```
'347-7420',
'392-6856',
'371-5556',
'334-5337',
'334-8817',
'339-1405',
'380-7742',
'329-6191',
'340-3075',
'416-5849',
'334-7443',
'394-9121',
'383-1657',
'347-4112',
'362-8280',
'402-9982',
'392-8905',
'392-5512',
'351-6384',
'348-8015',
'374-6966',
'328-2236',
'372-6497',
'417-8617',
'361-9621',
'421-2723',
'327-9341',
'383-5474',
'328-8147',
'373-5438',
'333-9253',
'347-9421',
'419-3167',
'414-4162',
'416-5341',
'368-8600',
'336-6085',
'377-1479',
'360-2107',
'405-9384',
'420-5179',
'331-3174',
'411-5958',
'333-8180',
'357-2679',
'396-2867',
'341-9443',
```

```
'341-4103',
'416-8701',
'397-6109',
'392-5587',
'392-9342',
'368-2845',
'405-4920',
'348-7484',
'338-5207',
'418-7846',
'358-8729',
'349-1943',
'368-8283',
'345-1419',
'358-8025',
'383-8695',
'370-7565',
'401-6162',
'386-5303',
'351-6552',
'345-5338',
'330-2849',
'364-6801',
'375-3003',
'383-8878',
'384-2372',
'377-7107',
'361-1581',
'417-7888',
'383-8364',
'396-2335',
'408-4530',
'408-6621',
'393-7522',
'338-7120',
'357-4265',
'398-8801',
'346-2347',
'343-2741',
'420-8242',
'402-7746',
'332-1494',
'388-6223',
'404-9539',
'341-7332',
'338-7886',
'332-5596',
```

```
'348-9945',
'407-1896',
'398-9408',
'369-8005',
'346-2530',
'400-5984',
'351-1007',
'345-5980',
'368-8964',
'358-1912',
'379-3132',
'340-9910',
'396-2719',
'369-6204',
'420-9971',
'410-3782',
'404-4481',
'383-9255',
'418-9385',
'360-9676',
'327-3587',
'385-4715',
'414-2695',
'331-5999',
'337-7739',
'388-6658',
'405-6943',
'382-4084',
'352-8249',
'353-8363',
'416-2825',
'342-6696',
'338-9210',
'328-1768',
'406-5870',
'398-3834',
'330-7754',
'414-9054',
'350-2832',
'414-9027',
'337-1225',
'394-6577',
'359-6995',
'377-7561',
'380-6631',
'390-8876',
'413-2201',
```

```
'374-2073',
'417-1272',
'358-1129',
'394-1211',
'327-1319',
'399-4413',
'393-9985',
'401-8377',
'331-3202',
'358-5953',
'380-7624',
'416-5261',
'417-5067',
'345-6515',
'406-1349',
'360-9038',
'348-3444',
'370-2892',
'383-4061',
'391-7937',
'389-4083',
'410-9633',
'418-9502',
'339-6637',
'356-3403',
'371-9457',
'391-8087',
'392-6420',
'399-4094',
'378-4013',
'386-6306',
'359-3618',
'340-8875',
'330-2693',
'403-7627',
'342-3678',
'344-9943',
'390-5686',
'358-5826',
'393-7826',
'335-9786',
'368-9860',
'416-9522',
'416-7307',
'397-6789',
'335-9501',
'388-1250',
```

```
'386-1374',
'346-8112',
'364-9040',
'366-3944',
'331-9861',
'330-2881',
'402-9558',
'341-3180',
'371-8598',
'398-8385',
'333-4154',
'330-3589',
'417-1477',
'327-5525',
'363-8244',
'419-9104',
'363-1560',
'341-4075',
'366-9074',
'367-1424',
'341-1191',
'342-9480',
'355-9360',
'343-1538',
'335-2331',
'335-7257',
'332-2137',
'352-2998',
'346-6941',
'400-2203',
'421-2955',
'331-6629',
'343-6314',
'414-6638',
'350-5883',
'409-4447',
'376-4856',
'335-1874',
'397-6255',
'381-5047',
'403-6850',
'375-3658',
'341-2603',
'342-5062',
'354-1558',
'351-9537',
'401-1252',
```

```
'366-9538',
'364-7622',
'393-9918',
'376-4484',
'410-6791',
'343-2392',
'408-4323',
'395-1718',
'354-9492',
'367-8168',
'340-3500',
'369-4962',
'334-8967',
'402-2377',
'366-3917',
'333-3531',
'333-8954',
'374-9203',
'334-3289',
'353-7822',
'350-1422',
'385-8406',
'380-7277',
'352-1798',
'385-7922',
'353-7730',
'337-7163',
'348-5567',
'420-9575',
'366-3358',
'359-9972',
'387-3332',
'354-6960',
'405-3335',
'379-4257',
'355-4992',
'383-6373',
'382-7993',
'422-5865',
'410-9961',
'343-9946',
'357-7060',
'355-9541',
'378-4145',
'386-2317',
'335-8146',
'377-2235',
```

```
'386-9141',
'416-5623',
'327-3053',
'395-6195',
'397-8772',
'346-7656',
'343-2350',
'372-4722',
'399-8615',
'379-8248',
'414-6219',
'387-1343',
'370-5527',
'393-8736',
'402-7626',
'370-2688',
'418-9036',
'417-2035',
'355-3602',
'393-4027',
'418-5141',
'406-1247',
'402-3892',
'332-2965',
'377-1218',
'355-2464',
'404-4611',
'373-2339',
'410-3503',
'410-4739',
'365-3562',
'387-7641',
'385-9744',
'398-5006',
'408-3977',
'334-2729',
'334-7685',
'350-9228',
'407-5774',
'413-4039',
'343-2077',
'336-5661',
'355-4143',
'366-5918',
'368-7555',
'408-2119',
'329-2789',
```

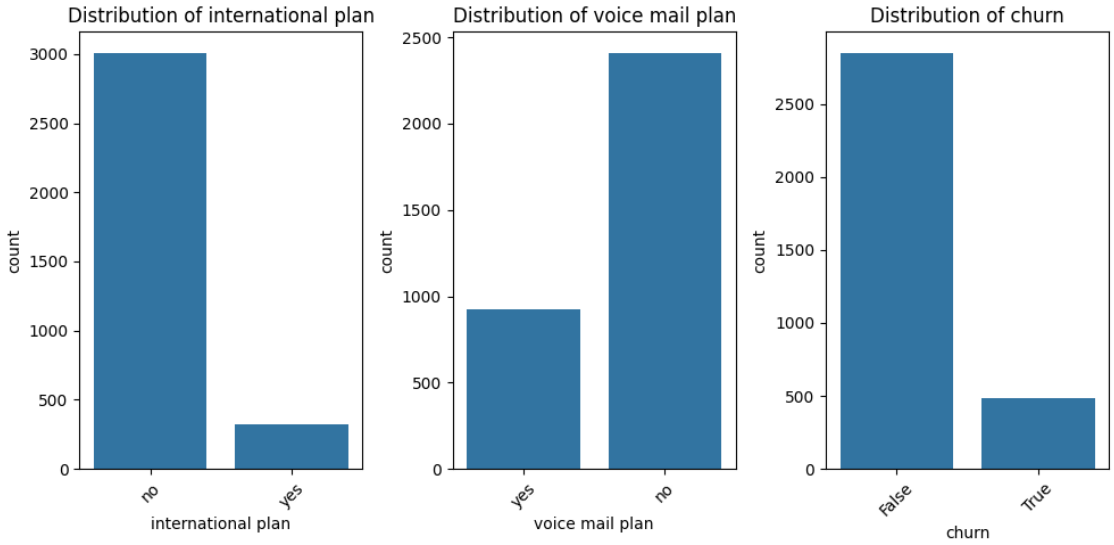
```
'334-1508',
'337-1506',
'396-8400',
'349-2654',
'417-2716',
'402-1725',
'403-3229',
'353-6056',
'399-9802',
'366-7069',
'332-5949',
'393-6376',
'354-7025',
'351-6585',
'330-5462',
'407-2292',
'340-3011',
'345-9153',
'379-5503',
'389-6790',
'408-3610',
'375-8238',
'378-5633',
'389-9120',
'380-2758',
'402-2728',
'354-9062',
'417-2054',
'353-1941',
'328-1522',
'408-4529',
'417-5320',
'370-9755',
'372-4835',
'334-6605',
'399-5564',
'392-2887',
'328-2110',
'383-5976',
'332-6181',
'330-5255',
'413-5190',
'394-7447',
'353-7096',
'395-6002',
'372-9816',
'400-7253',
```

```
  '344-7470',
  '378-8572',
  '345-9140',
  '340-5460',
  '369-8024',
  '395-8595',
  '359-4587',
  '375-5439',
  '361-3779',
  '375-8934',
  '395-4757',
  '421-7205',
  '379-8805',
  '348-2150',
  '417-9128',
  '351-4226',
  '330-6630',
 …],
 'international plan': ['no', 'yes'],
 'voice mail plan': ['yes', 'no']}
```

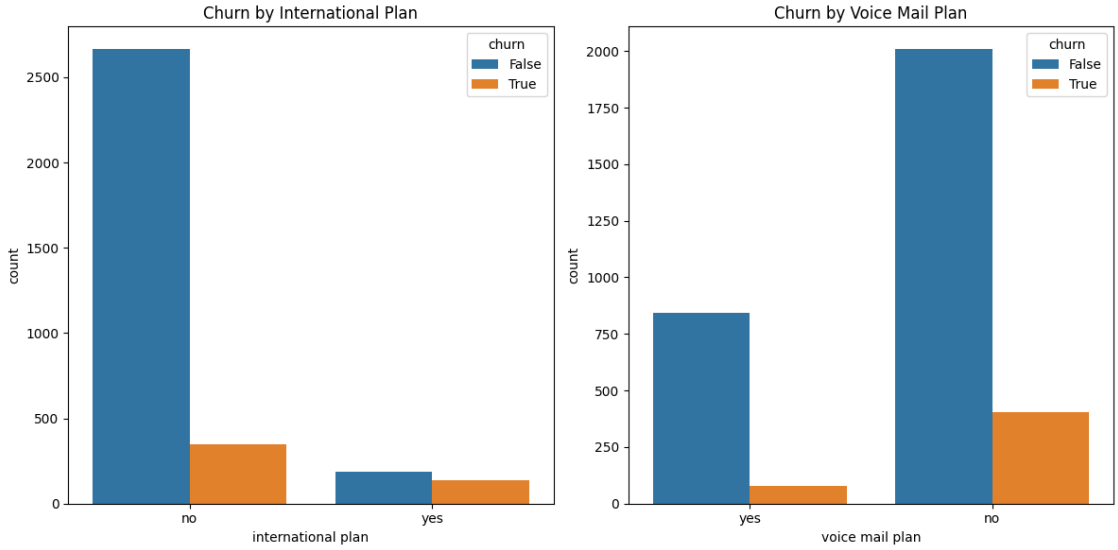
[9]:
```
# list of categorical columns
categorical_columns = ['international plan', 'voice mail plan', 'churn']

# verifying the unique values in each categorical variable
for column in categorical_columns:
    print(f'{column}:\n', df[column].value_counts(), '\n')
```

```
international plan:
 international plan
no     3010
yes     323
Name: count, dtype: int64

voice mail plan:
 voice mail plan
no     2411
yes     922
Name: count, dtype: int64

churn:
 churn
False    2850
True      483
Name: count, dtype: int64
```

[10]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# plotting bar charts for each categorical variable
plt.figure(figsize=(10, 5))

for i, column in enumerate(categorical_columns, 1):
    plt.subplot(1, 3, i)
    sns.countplot(x=df[column])
    plt.title(f'Distribution of {column}')
    plt.xticks(rotation=45)

plt.tight_layout()
plt.show()
```



[11]:
```
# plotting distribution of churn across international_plan and voice_mail_plan
plt.figure(figsize=(12, 6))

plt.subplot(1, 2, 1)
sns.countplot(x='international plan', hue='churn', data=df)
plt.title('Churn by International Plan')

plt.subplot(1, 2, 2)
sns.countplot(x='voice mail plan', hue='churn', data=df)
plt.title('Churn by Voice Mail Plan')

plt.tight_layout()
plt.show()
```



[12]:
```
# churn rate by international_plan
churn_rate_international = df.groupby('international plan')['churn'].mean()
print('Churn Rate by International Plan:\n', churn_rate_international)

# churn rate by voice_mail_plan
churn_rate_voicemail = df.groupby('voice mail plan')['churn'].mean()
print('Churn Rate by Voice Mail Plan:\n', churn_rate_voicemail)
```

```
Churn Rate by International Plan:
 international plan
no     0.114950
yes    0.424149
Name: churn, dtype: float64
Churn Rate by Voice Mail Plan:
 voice mail plan
no     0.167151
yes    0.086768
Name: churn, dtype: float64
```

2.3.1 Summary of Exploring Categorical Variables

Churn Rate by Categorical Variables I calculated the churn rate for customers based on whether they have an international plan or a voice mail plan.

Churn Rate by International Plan:

- **Customers without an international plan**: 11.5% churn rate.
- **Customers with an international plan**: 42.4% churn rate.

Having an international plan is associated with a significantly higher churn rate. This could indicate

that customers with international plans are more likely to leave, possibly due to unmet expectations or service issues.

Churn Rate by Voice Mail Plan:

- **Customers without a voice mail plan**: 16.7% churn rate.
- **Customers with a voice mail plan**: 8.7% churn rate.

Having a voice mail plan appears to reduce the likelihood of churn, suggesting that this feature may play a positive role in retaining customers.

2.4 Exploring Numerical Features

Here, I explore the numerical features (e.g., `total_day_minutes`, `customer_service_calls`) to understand their distributions and check for any potential outliers or patterns. I also analyze correlations between these features and the target variable (`churn`).

[13]:
```
df.describe()
```

[13]:

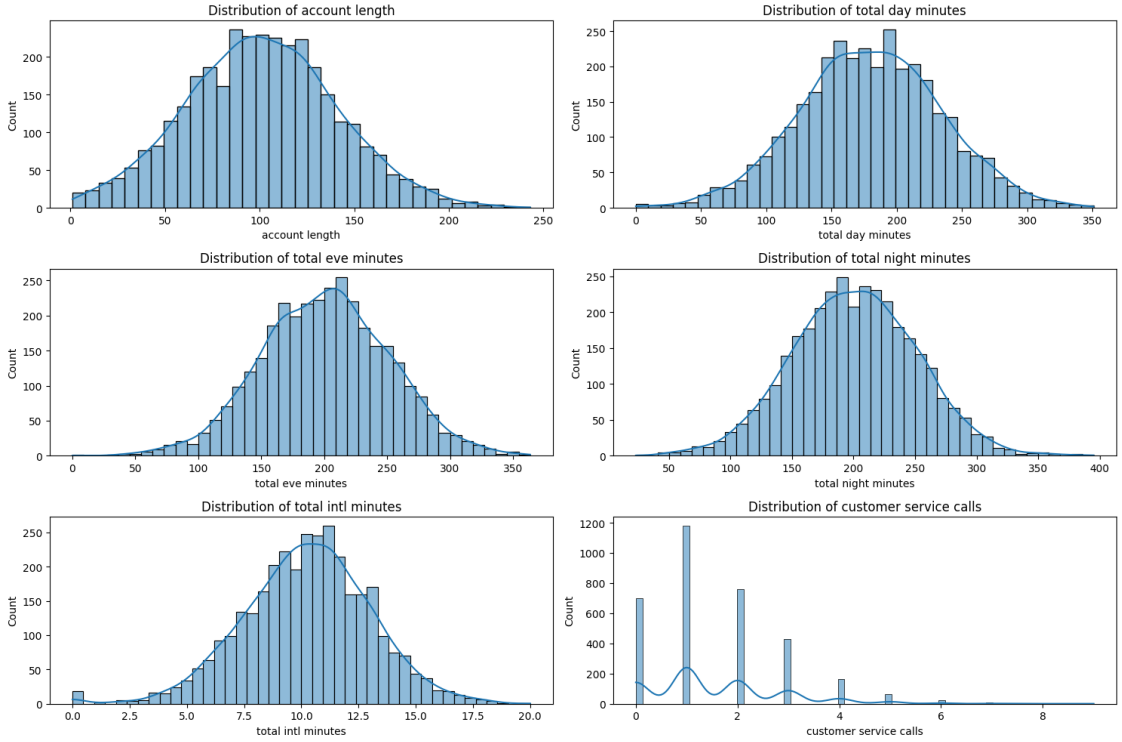
	account length	area code	number vmail messages	total day minutes	\
count	3333.000000	3333.000000	3333.000000	3333.000000	
mean	101.064806	437.182418	8.099010	179.775098	
std	39.822106	42.371290	13.688365	54.467389	
min	1.000000	408.000000	0.000000	0.000000	
25%	74.000000	408.000000	0.000000	143.700000	
50%	101.000000	415.000000	0.000000	179.400000	
75%	127.000000	510.000000	20.000000	216.400000	
max	243.000000	510.000000	51.000000	350.800000	

	total day calls	total day charge	total eve minutes	total eve calls	\
count	3333.000000	3333.000000	3333.000000	3333.000000	
mean	100.435644	30.562307	200.980348	100.114311	
std	20.069084	9.259435	50.713844	19.922625	
min	0.000000	0.000000	0.000000	0.000000	
25%	87.000000	24.430000	166.600000	87.000000	
50%	101.000000	30.500000	201.400000	100.000000	
75%	114.000000	36.790000	235.300000	114.000000	
max	165.000000	59.640000	363.700000	170.000000	

	total eve charge	total night minutes	total night calls	\
count	3333.000000	3333.000000	3333.000000	
mean	17.083540	200.872037	100.107711	
std	4.310668	50.573847	19.568609	
min	0.000000	23.200000	33.000000	
25%	14.160000	167.000000	87.000000	
50%	17.120000	201.200000	100.000000	
75%	20.000000	235.300000	113.000000	
max	30.910000	395.000000	175.000000	

```
       total night charge  total intl minutes  total intl calls  \
count         3333.000000         3333.000000       3333.000000
mean             9.039325           10.237294          4.479448
std              2.275873            2.791840          2.461214
min              1.040000            0.000000          0.000000
25%              7.520000            8.500000          3.000000
50%              9.050000           10.300000          4.000000
75%             10.590000           12.100000          6.000000
max             17.770000           20.000000         20.000000

       total intl charge  customer service calls
count        3333.000000             3333.000000
mean            2.764581                1.562856
std             0.753773                1.315491
min             0.000000                0.000000
25%             2.300000                1.000000
50%             2.780000                1.000000
75%             3.270000                2.000000
max             5.400000                9.000000
```

[14]:
```
import matplotlib.pyplot as plt
import seaborn as sns

numerical_columns = ['account length', 'total day minutes', 'total eve minutes',
                     'total night minutes', 'total intl minutes', 'customer␣
 ↪service calls']

plt.figure(figsize=(15, 10))

for i, column in enumerate(numerical_columns, 1):
    plt.subplot(3, 2, i)
    sns.histplot(df[column], kde=True)
    plt.title(f'Distribution of {column}')

plt.tight_layout()
plt.show()
```



2.4.1 Summary Statistics of Numerical Features

The table above summarizes the distribution of numerical features in the dataset. It provides key metrics such as mean, standard deviation, and percentiles.

Key Observations:

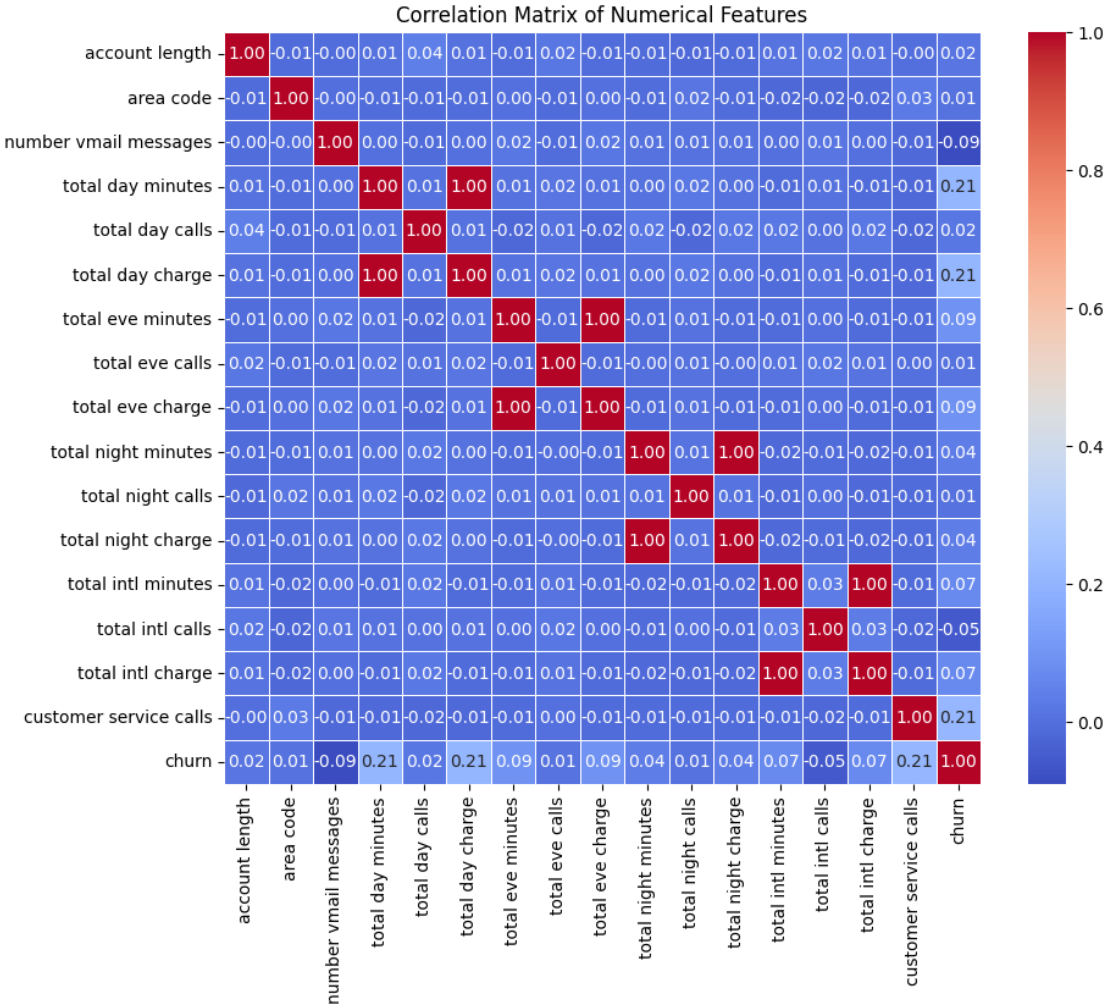
- **Account Length**: The average customer has an account length of approximately 101 days, with a standard deviation of 40 days. The maximum account length is 243 days, and the minimum is 1 day.
- **Total Day, Evening, and Night Minutes**: Customers tend to use a similar amount of minutes during the day, evening, and night, with means of around 180-200 minutes in each category. However, there are outliers, as seen from the maximum values.
- **Total International Minutes**: The average usage of international minutes is low, at around 10 minutes, with a maximum of 20 minutes.
- **Customer Service Calls**: Most customers make an average of 1-2 customer service calls, though some outliers have made up to 9 calls. This feature may be interesting to explore further as it could be related to customer churn.

I will use this information to identify potential outliers, scale the data if necessary, and select relevant features for modeling.

2.5 Correlation Analysis

In this section, I compute the correlation matrix to investigate the relationships between the numerical features and the target variable (**churn**). This will help me identify which features are strongly correlated with churn, guiding the feature selection for modeling.

[15]:
```
# temp selecting only numerical columns without modifying the original DataFrame
numeric_df = df.select_dtypes(include=['float64', 'int64', 'bool']).copy()

# converting boolean columns to integers, without affecting the original␣
 ↪DataFrame
numeric_df['churn'] = numeric_df['churn'].astype(int)
corr_matrix = numeric_df.corr()

# plotting the heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Matrix of Numerical Features')
plt.show()
```



2.5.1 Correlation Matrix Summary

- **Perfect Correlation**: Features like `total_day_minutes` and `total_day_charge` (and similar pairs) are perfectly correlated. I may drop one feature from each pair to avoid multicollinearity.
- **Customer Service Calls**: This feature shows a moderate positive correlation with churn (0.21), indicating that customers who make more service calls are more likely to churn.
- **Other Features**: Most features show weak correlations with churn, suggesting that they may not be strong individual predictors but could still contribute when combined with other features in the model.

Note: Churn was temporarily encoded as 1 (True) and 0 (False) for this correlation analysis.

3 Insights and Conclusions

3.1 1. Class Imbalance:

- The dataset shows a **class imbalance** with only **14.5% of customers churning** and **85.5% not churning**. This imbalance may require resampling techniques or special consideration in the model to avoid bias towards the majority class.

3.2 2. Categorical Features:

- **International Plan**: Customers with an international plan have a significantly higher churn rate (**42.4%**) compared to those without one (**11.5%**). This suggests that having an international plan may be a strong indicator of customer dissatisfaction or unmet expectations, making it a key feature for predicting churn.
- **Voice Mail Plan**: Customers without a voice mail plan have a higher churn rate (**16.7%**) compared to those with one (**8.7%**). This implies that offering voice mail plans may help retain customers, and this feature is likely to be important for the prediction model.

3.3 3. Numerical Features:

- **Customer Service Calls**: This feature has a positive correlation with churn (**0.21**). Customers who make more service calls are more likely to churn, suggesting that frequent interaction with customer service could be a sign of dissatisfaction.
- **Highly Correlated Features**: Several pairs of features (e.g., `total_day_minutes` and `total_day_charge`) are perfectly correlated. These pairs essentially represent the same information, so I may consider dropping one feature from each pair during the preprocessing phase to avoid multicollinearity.

3.4 4. General Distribution Patterns:

- Most numerical features (e.g., `total_day_minutes`, `total_eve_minutes`) follow a normal distribution, while **customer service calls** has a more skewed distribution, with many customers having zero or very few calls but a few customers making frequent calls.

- **Weak Correlations with Churn**: Most features have weak correlations with churn, indicating that no single feature strongly predicts churn on its own. However, combinations of features may still provide valuable information for churn prediction.

3.5 Conclusion:

The EDA revealed key insights that will guide the next steps of the project: - **International Plan** and **Voice Mail Plan** are important categorical features related to churn, with clear differences in churn rates between their respective groups. - **Customer Service Calls** has a meaningful relationship with churn and should be considered as a crucial feature for the prediction model. - To handle multicollinearity, I will likely remove one feature from highly correlated pairs like `total_day_minutes` and `total_day_charge`. - Addressing class imbalance will be important during model building to ensure I accurately predict churners despite the skewed distribution of the target variable.

The next steps will involve preprocessing the data, encoding categorical variables, addressing class imbalance, and selecting features for the churn prediction model.