TELECOM CHURN ANALYSIS

Name: **Purnima Agarwal**

Dataset: **Telecom Churn Dataset**

Aim: We want to know the reasons that lead people to churn or retain by studying the dataset. This will tell us in which area we should focus to reduce the churn rate.

Step 1**: Importing necessary libraries**

```
import numpy as np
# linear algebra
import pandas as pd
# data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
# for data visualization
import seaborn as sns
# for data visualization
import warnings
warnings.filterwarnings('ignore')
```

Step 2: **Data Description**

There are 3333 rows and 17 columns in the dataset. Our target variable is churn rate which is percentage of subscriber who discontinue the subscription which is indicated by true. There are 4 categorical, 2 discrete, 10 continuous and 1 boolean variable.

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day calls	total day charge	total eve calls	total eve charge	total night calls	total night charge	total intl calls	total intl charge	customer service calls	churn
0	KS	128	415	382-4657	no	yes	25.0	110.0	45.07	99.0	16.78	91.0	11.01	3.0	2.70	1.0	False
1	OH	107	415	371-7191	no	yes	26.0	123.0	27.47	103.0	16.62	103.0	11.45	3.0	3.70	1.0	False
2	NJ	137	415	358-1921	no	no	0.0	114.0	41.38	110.0	10.30	104.0	7.32	5.0	3.29	0.0	False
3	OH	84	408	375-9999	yes	no	0.0	71.0	50.90	88.0	5.26	89.0	8.86	7.0	1.78	2.0	False
4	OK	75	415	330-6626	yes	no	0.0	113.0	28.34	122.0	12.61	121.0	8.41	3.0	2.73	3.0	False

Type of each attribute

```
df.dtypes
```

```
state                    object
account length            int64
area code                 int64
phone number             object
international plan       object
voice mail plan          object
number vmail messages   float64
total day calls         float64
total day charge        float64
total eve calls         float64
total eve charge        float64
total night calls       float64
total night charge      float64
total intl calls        float64
total intl charge       float64
customer service calls  float64
churn                      bool
dtype: object
```

Step 3: **Data cleaning Missing value treatment**

The table shows number of missing values in the dataset and their respective percentage.

	Total	Percent
total eve charge	4	0.0012
total day charge	3	0.0009
total intl charge	3	0.0009
total night calls	3	0.0009
total intl calls	2	0.0006
total eve calls	2	0.0006
total day calls	2	0.0006
total night charge	1	0.0003
customer service calls	1	0.0003
number vmail messages	1	0.0003
phone number	0	0.0000
account length	0	0.0000
area code	0	0.0000
churn	0	0.0000
international plan	0	0.0000
voice mail plan	0	0.0000
state	0	0.0000

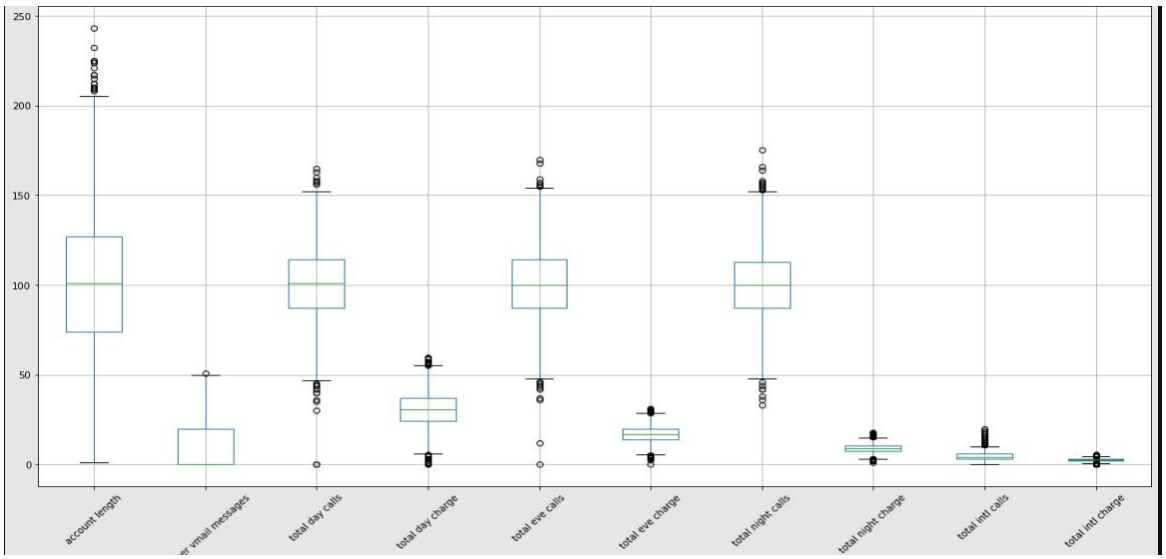
Since the percentage of missing value is less than 5%, we will remove the missing value instead of replacing them.

Variable phone number has no significance therefore we are dropping it.

Treatment of outlier

An outlier is an observation that diverges well -structured data. Root cause of outlier can be an error in measurement or data collection error. We can either delete the outliers or replace them with average values. They are replaced with mean or median in quantitative data and mode in case of qualitative data.

We have used whisker box to identify the outlier value.



Outlier value are detected in account length, number vmail messages, total day calls, total day charge, total eve calls, total eve charge, total night calls, total night charge, total intl calls, total intl charge. We will replace these outlier value with median value of concerned variables.

Step 4**: Explanatory Data Analysis**

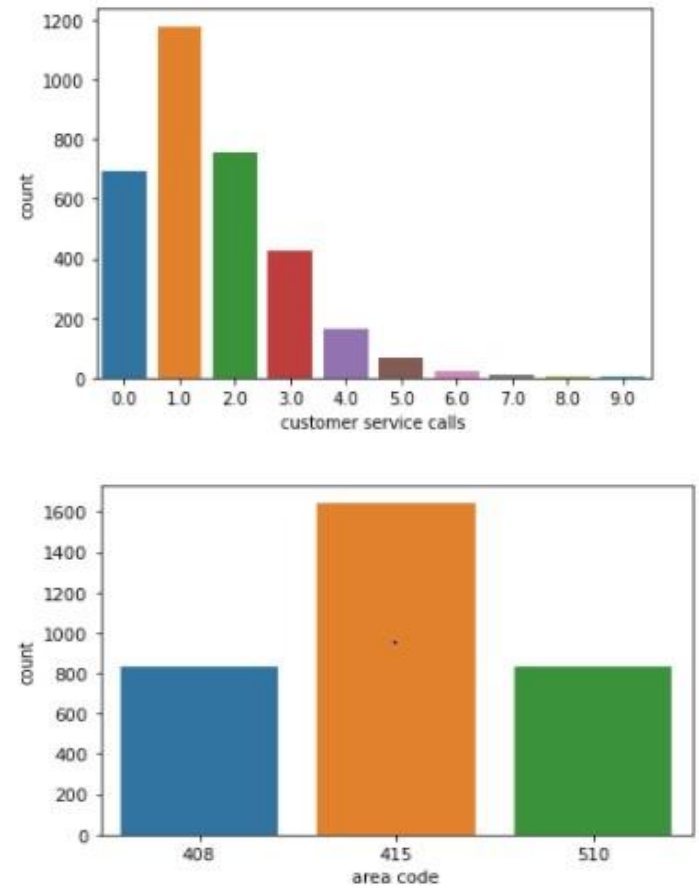
Under EDA we employ graphical tools to analyse the data to identify patterns and relationship. We will do three types of analysis here: univariate analysis bivariate analysis and multivariate analysis.

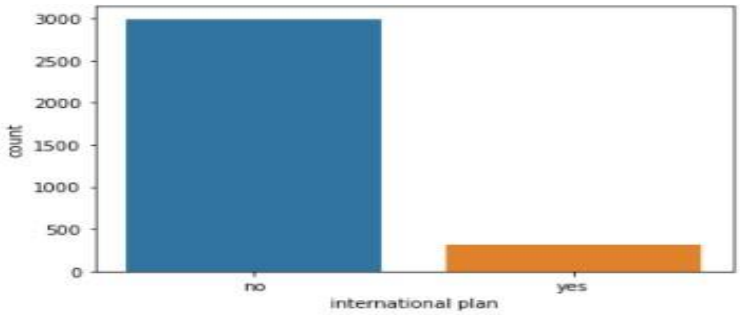
Univariate Analysis of Categorical Variables Code-

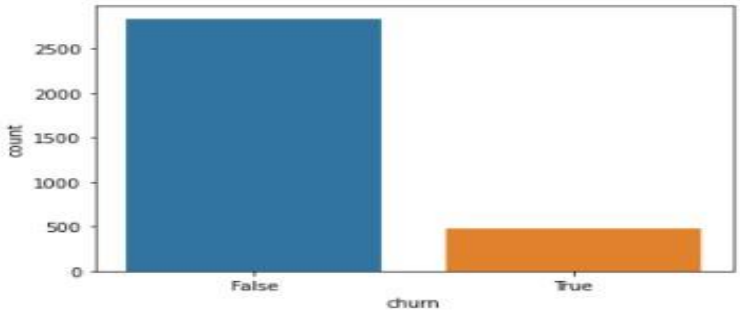
```
sns.countplot(df['international plan'])
plt.show()
sns.countplot(df['churn'])
plt.show()
sns.countplot(df['voice mail plan'])
plt.show()
sns.countplot(df['area code'])
plt.show()
```

```
df['area code'].value_counts()
sns.countplot(df['customer service calls'])
plt.show()
df['customer service calls'].value_counts()
```

Output-





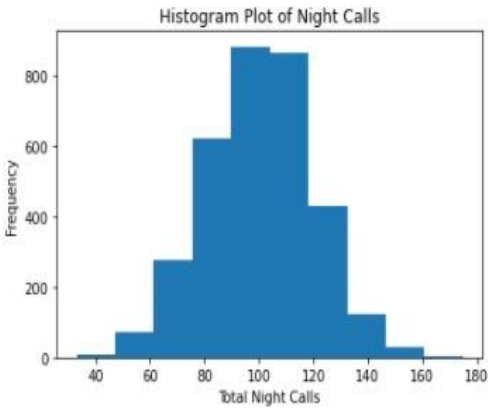


Observation-

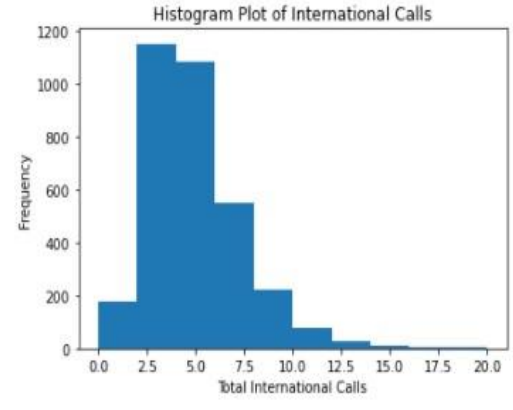
- 14% of customers have discontinued telecom services i.e. churn
- 49% of customers belong to Area Code 415
- 36% of customers gave only one customer service call
- 9% of customers use international plan
- 27% of customers use voice mail plan

Univariate Analysis of Quantitative Variables Code with their respective output

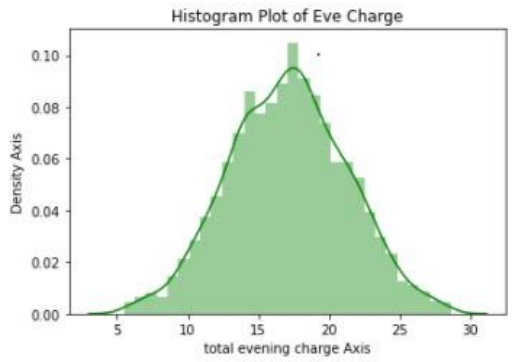
```
plt.hist(df['total night calls'], bins=10)
plt.xlabel("Total Night Calls")
plt.ylabel("Frequency")
plt.title("Histogram Plot of Night Calls")
plt.show()
```



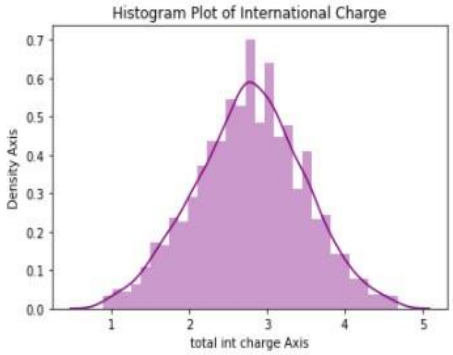

```
plt.hist(df['total intl calls'], bins=10)
plt.xlabel("Total International Calls")
plt.ylabel("Frequency")
plt.title("Histogram Plot of International Calls")
plt.show()
```



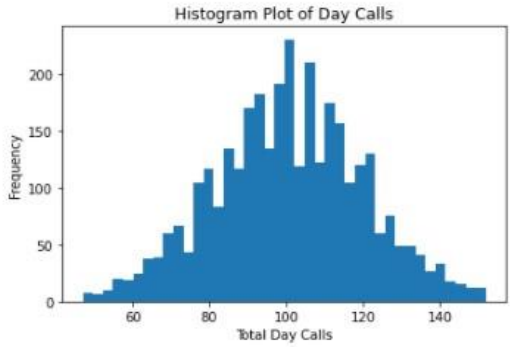
```
sns.distplot(df['total eve charge'],color="green")
plt.xlabel("total evening charge Axis")
plt.ylabel("Density Axis")
plt.title("Histogram Plot of Eve Charge")
plt.show()
```



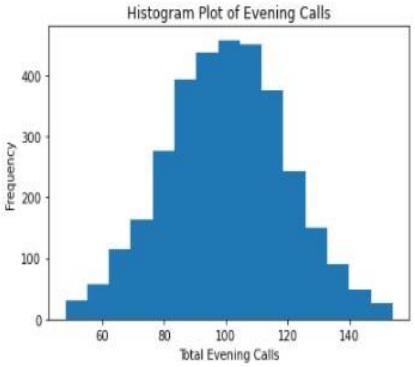

```
sns.distplot(df['total intl charge'],color="purple")
plt.xlabel("total int charge Axis")
plt.ylabel("Density Axis")
plt.title("Histogram Plot of International Charge")
plt.show()
```




```
plt.hist(df['total day calls'], bins=40)
plt.xlabel("Total Day Calls")
plt.ylabel("Frequency")
plt.title("Histogram Plot of Day Calls")
plt.show()
```




```
plt.hist(df['total eve calls'], bins=15)
plt.xlabel("Total Evening Calls")
plt.ylabel("Frequency")
plt.title("Histogram Plot of Evening Calls")
plt.show()
```

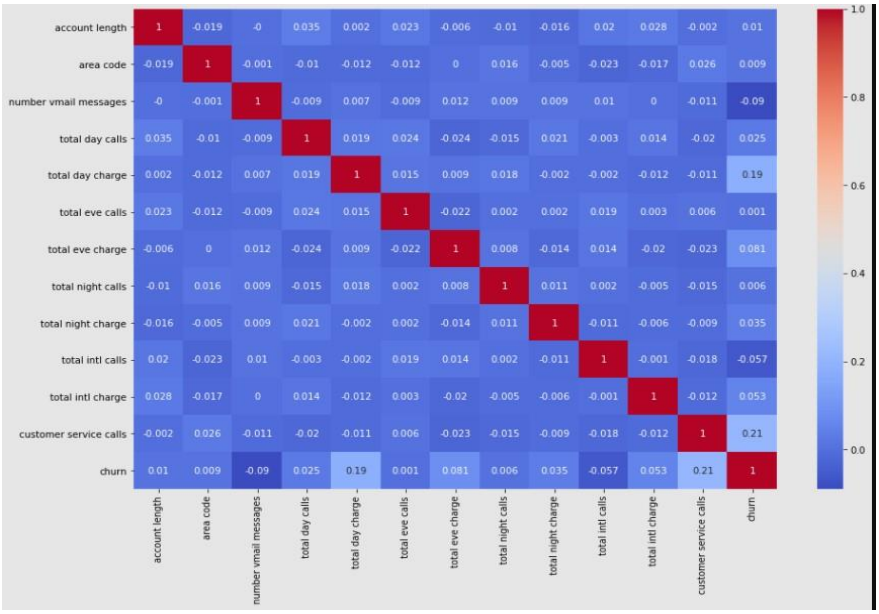


Observation-

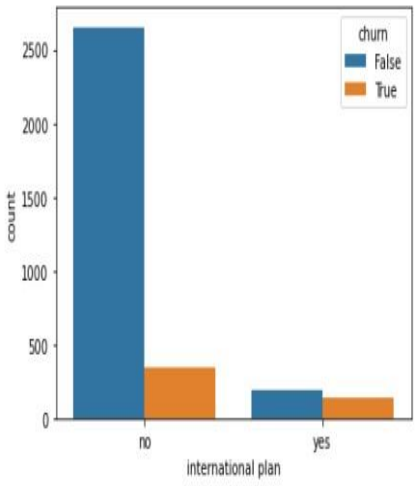
Total eve charge, total eve calls, total day calls and total day charge columns are uniformly distributed. Total night calls and total intl charge are moderately left skewed. Total intl calls are rightly skewed.

Bivariate Analysis Code and Output-

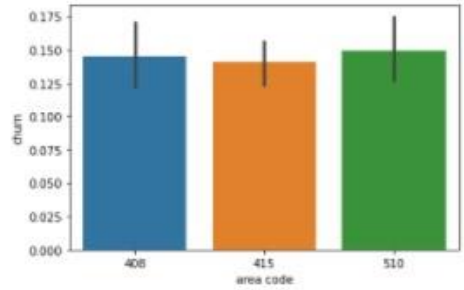
```
cor = df.corr()
plt.figure(figsize=(15,10))
sns.heatmap(cor.round(3),annot=True,cmap='coolwarm')
plt.show()
```



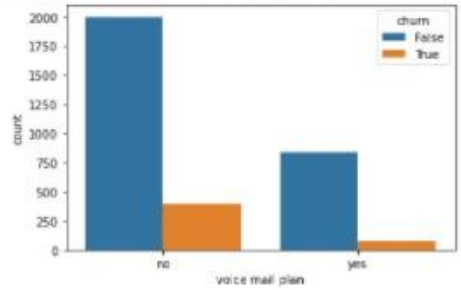

```
sns.countplot(x='international plan', hue='churn', data=df, order = df['international plan'].value_counts().index);
```



```
sns.barplot(df['area code'], df['churn'])
plt.show()
```



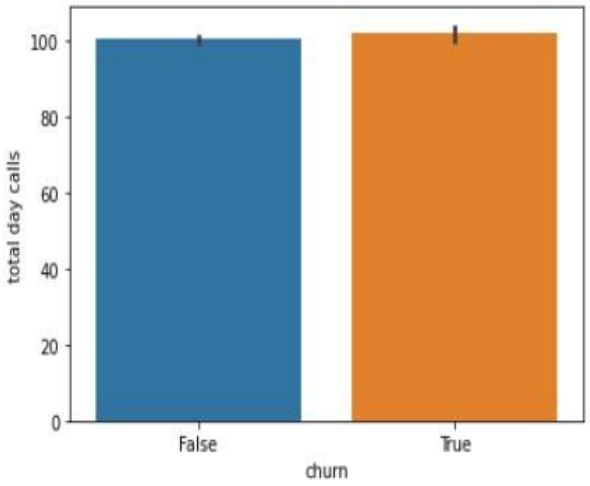

```
sns.countplot(x='voice mail plan', hue='churn', data=df, order = df['voice mail plan'].value_counts().index);
```




```
df.groupby('area code').aggregate({'total day charge':'sum','total eve charge':'sum',
                                   'total night charge':'sum'})
```

area code	total day charge	total eve charge	total night charge
408	25213.68	14285.43	7490.04
415	50846.12	28110.65	14953.71
510	25356.51	14275.57	7530.25

```
sns.barplot(df['churn'], df['total day calls'])
plt.show()
```

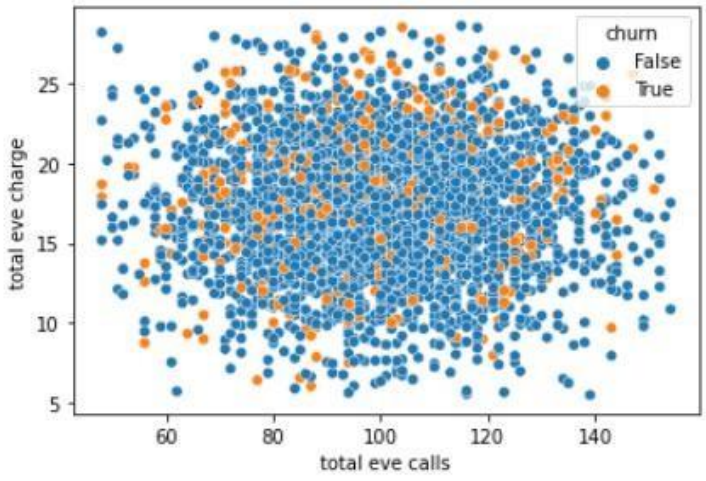


Observation-

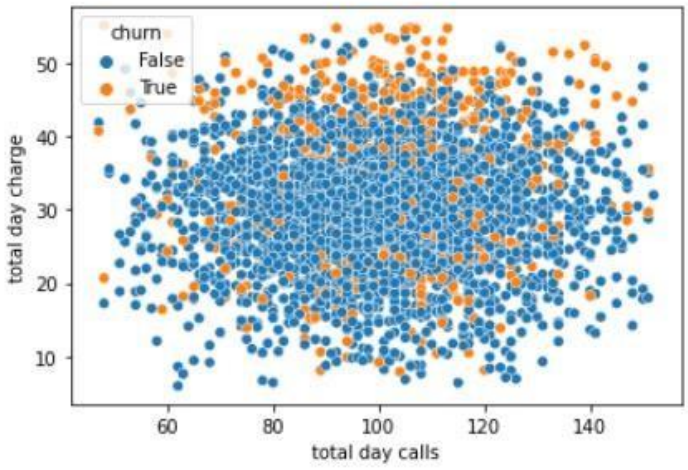
- Total day charge, total eve charge and customer service call influence churn.
- Account length, area code, total day calls, total night calls and total intl charge has moderately positive relationship with churn.
- Number of voice mail messages, total night charge and total intl charge has moderately negative relationship with churn.
- 15% of customer who don't have voice mail plan discontinued the telecom services i.e. churn
- 12% of customer who have voice mail plan discontinued the telecom services i.e. churn
- Churn rate for area code 415,510 and 408 is 14.28%,14.28% and 15.71% respectively.
- 10% of customer who don't have international plan discontinued the telecom services i.e. churn.
- 90% of customer who have international plan discontinued the telecom services i.e. churn.
- Telecom company generates double revenue from area code 415 in comparison to 408 and 510 area code.
- Possible reason for this could be more customer in area code 415.
- Customer who churn and don't churn makes 110 and 100 calls respectively, which indicate that total number of calls doesn't affect churn rate.

Multivariate Analysis Code and Output-

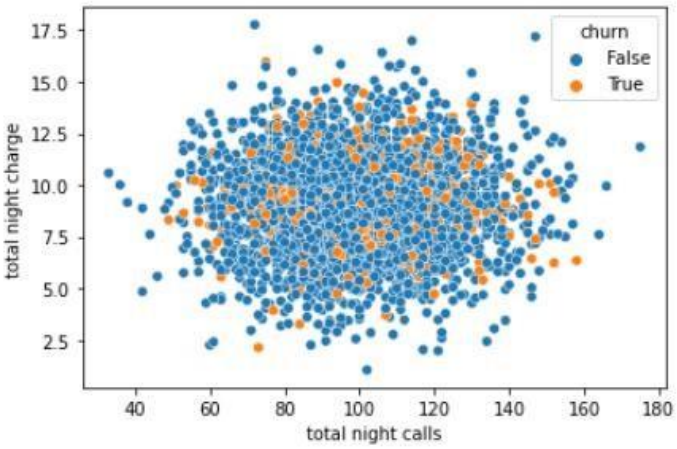
```
sns.scatterplot(df["total eve calls"], df["total eve charge"], hue=df["churn"])
plt.show()
```



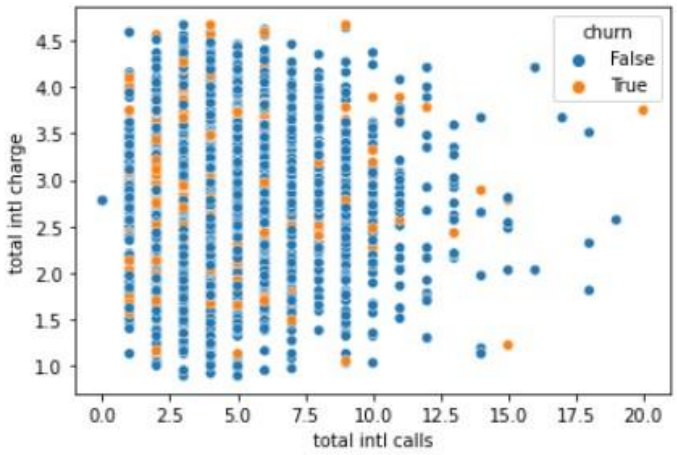
```
sns.scatterplot(df["total day calls"], df["total day charge"], hue=df["churn"])
plt.show()
```

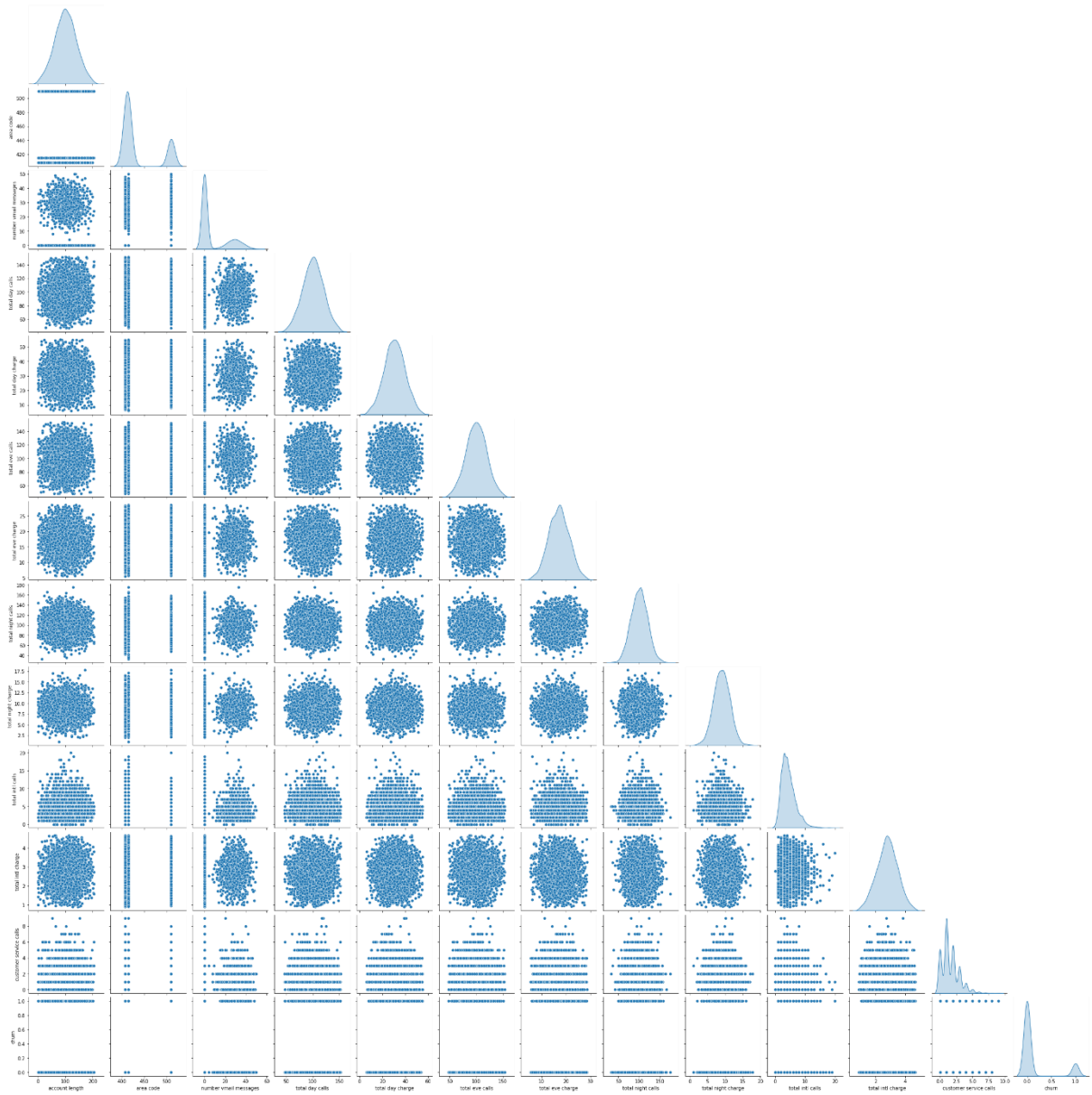



```
sns.scatterplot(df["total night calls"], df["total night charge"], hue=df["churn"])
plt.show()
```




```
sns.scatterplot(df["total intl calls"], df["total intl charge"], hue=df["churn"])
plt.show()
```





Observation-

- There seems to less churners among the evening callers.
- Churn rate increases at higher rate when the total day charge is more than 40.
- There seems to less churners among the night callers. • Customers who make 2 international calls are most
- likely to churn.
- Customers who make 7 international calls are least likely to churn.

CONCLUSION

At the end we will conclude our analysis with major findings and recommendation. 49% of customer resides in area code 415 and the company generates half of it's revenue from here, thus it should focus more on area code 415. While California is the most populous state in the U.S, there are not as many customers from California in our dataset. Arizona (AZ), for example, has 64 customers, 4 of whom ended up churning. In comparison, California has a higher number (and percentage) of customers who churned. This is useful information for a company. Proportion of churn rate is higher in more customer calls. Churn rate in international plans is high(90%).

Recommendation-

Survey International plan customers to understand pain points and identify root causes for churn intent. Then take steps to address those concerns and ensure that appropriate service is provided.

Try to solve all customer problem within one customer call to ensure that any issues that the customer faces is fixed before the next call. Proactively check on customers to confirm that their issue is fixed.