Capstone Project - 1 - EDA

Telecom Churn Analysis

By

Subhajit Ganguly

Data Science Trainee, Almabetter

Problem in Hand:

- We need to find the reasons behind customer churns of Orange S.A. by simply performing some exploratory analyses.
- Some intuitive guesses that drive customer churn in a telecom company : pricing, network problems, issues with data speed etc.

Variables in hand :

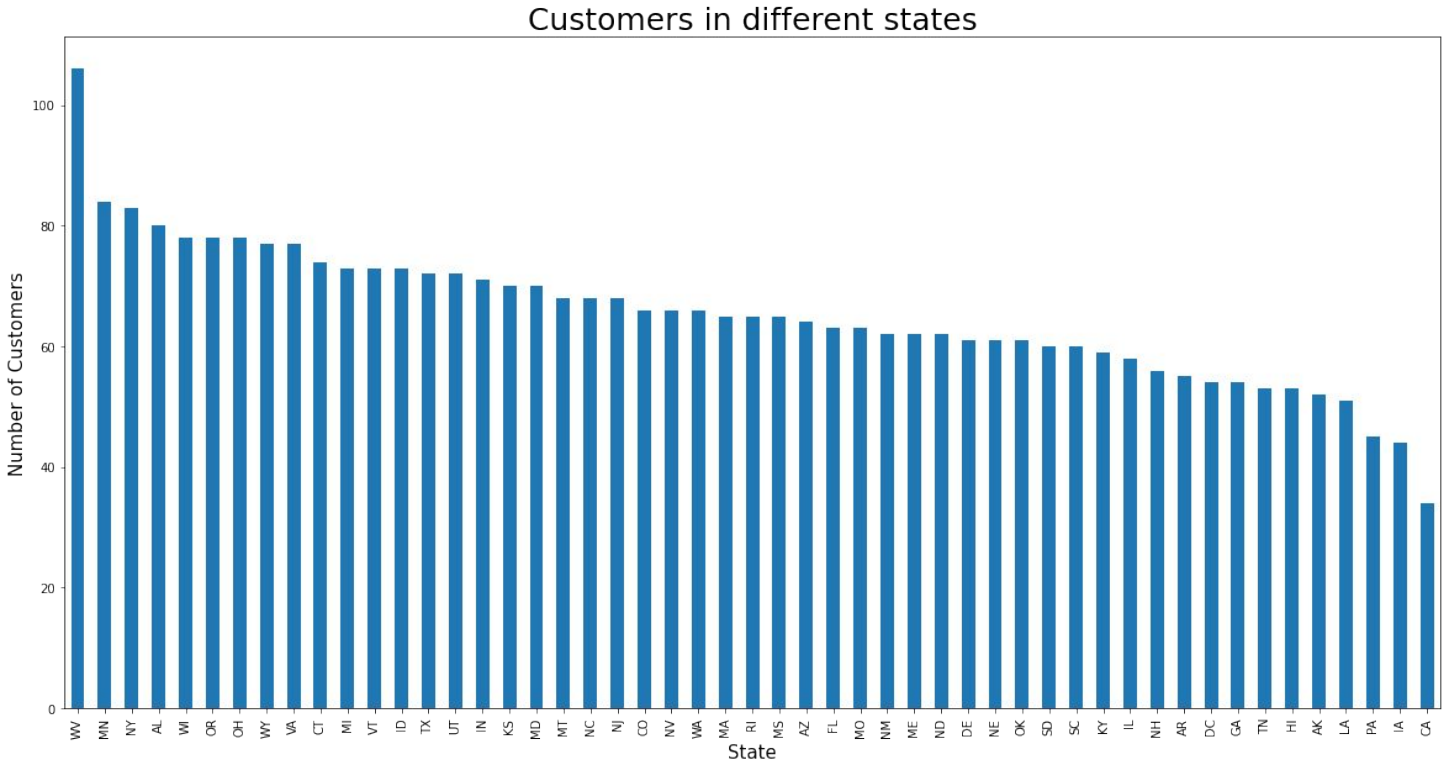
- Categorical variables :
  1. State
  2. Area code
  3. International plan
  4. Voice mail plan
  5. Churn

- Numerical variables :
  1. Account length
  2. Number vmail messages
  3. Total day minutes
  4. Total day calls
  5. Total day charge
  6. Total eve minutes
  7. Total eve calls
  8. Total eve charge
  9. Total night minutes
  10. Total night calls
  11. Total night charge
  12. Total intl minutes
  13. Total intl calls
  14. Total intl charge
  15. Customer service calls

Approaching the problem :

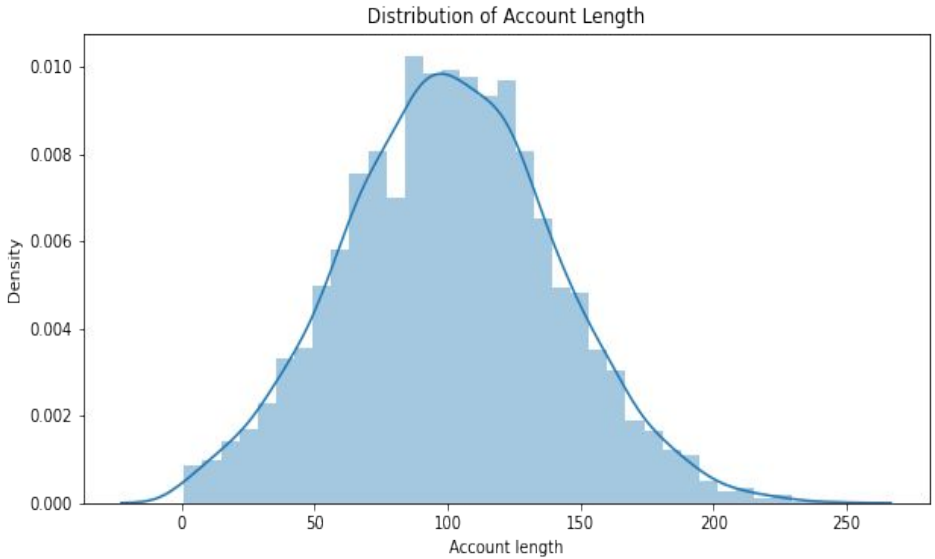
1. **Data Cleaning :**
   The first and most crucial step towards any analytics problem is to clean the data first. Cleaning can be done in various steps like : removing or replacing missing values and outliers. As we are only performing EDA, we can keep aside the outlier removal part for now.
2. **Univariate Analysis :**
   For the second step, we shall perform a univariate analysis on all the available features to get some insights about them.
3. **Bivariate Analysis :**
   For the third step, we shall perform some bivariate analysis to evaluate the relationship between the pairs of features. We want to do it especially with 'Churn' feature to evaluate what other features are driving it.
4. **Multivariate Analysis :**
   For the final step, we shall try to evaluate relationship between multiple variables and try to conclude how they are driving the churns.

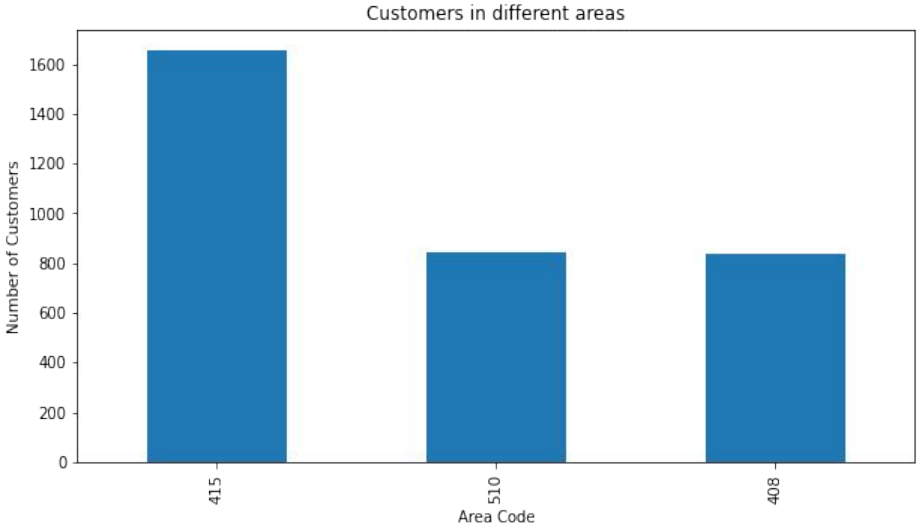
Univariate Analysis - State



- **WV has highest number of customers and CA has lowest number of customers.**

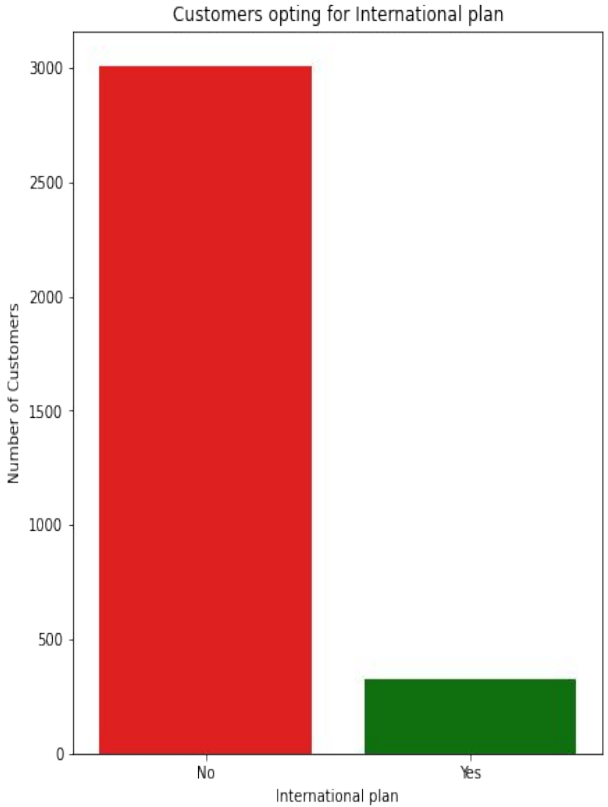
Univariate Analysis - Account length, Area code

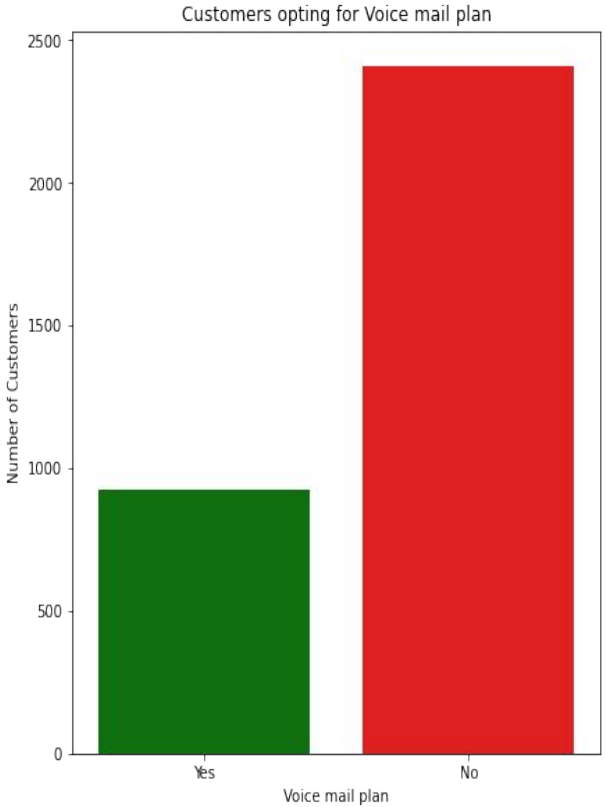




- Account length of customers have an almost Normal Distribution.
- People with very old or very new accounts are very few in numbers.
- The area with code 415 has most number of customers.
- The areas with codes 408 and 510 have almost half number of customers than the area 415.

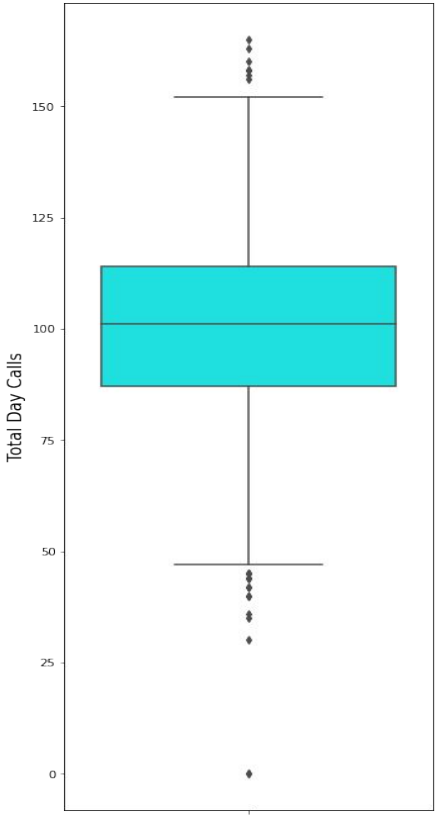
Univariate Analysis - International and voice mail plan

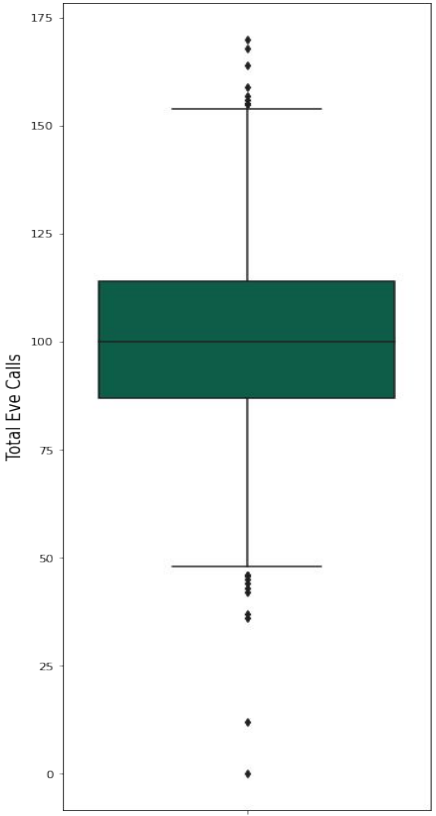


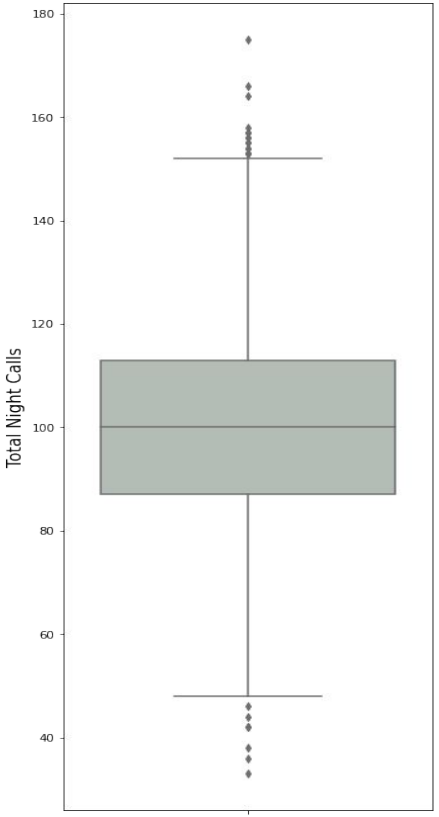


- More people didn't choose to recharge with International or Voice mail plans.
- But in case of Voice mail plan, the ratio between Yes and No is a bit higher.

Univariate Analysis - Day, Eve, Night Calls

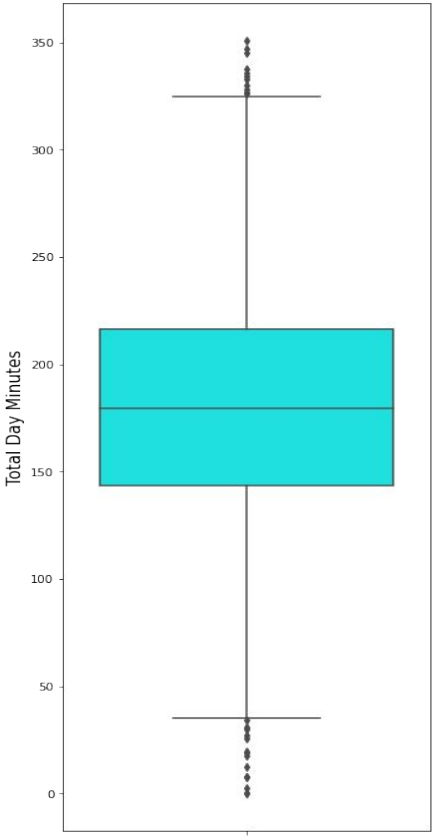


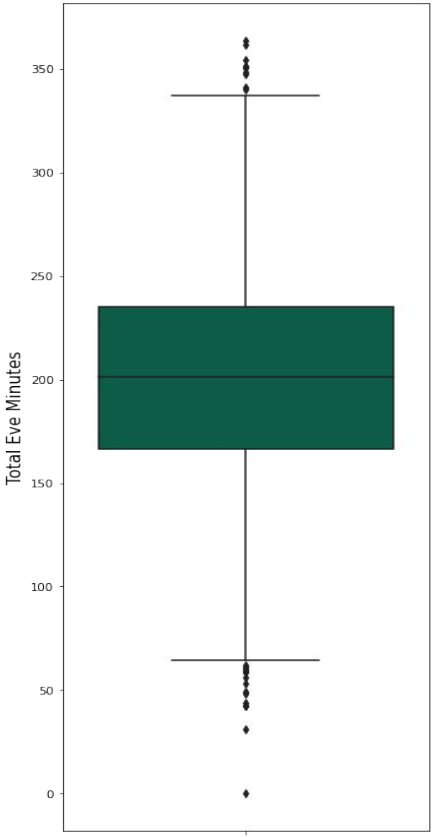


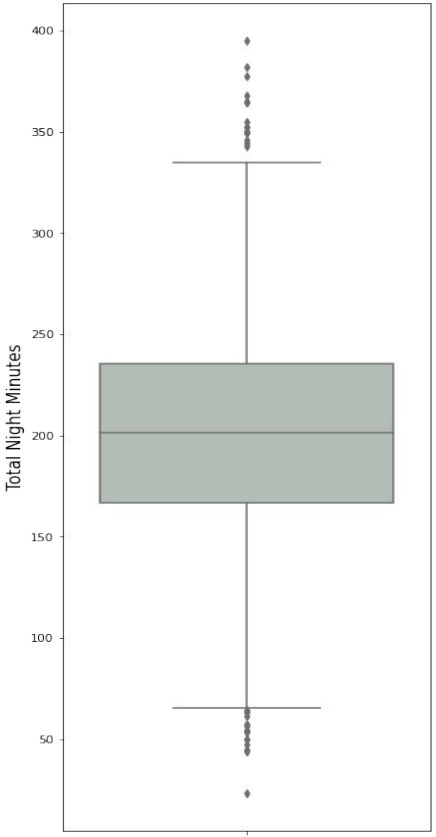


- Most calls were made during daytime and least calls were made in the night.
- There are few too small outlier points in Total Day Calls and Total Eve Calls and a large outlier point in Total Night Calls which need to be treated before feeding the data into an ML algorithm.

Univariate Analysis - Day, Eve, Night Minutes

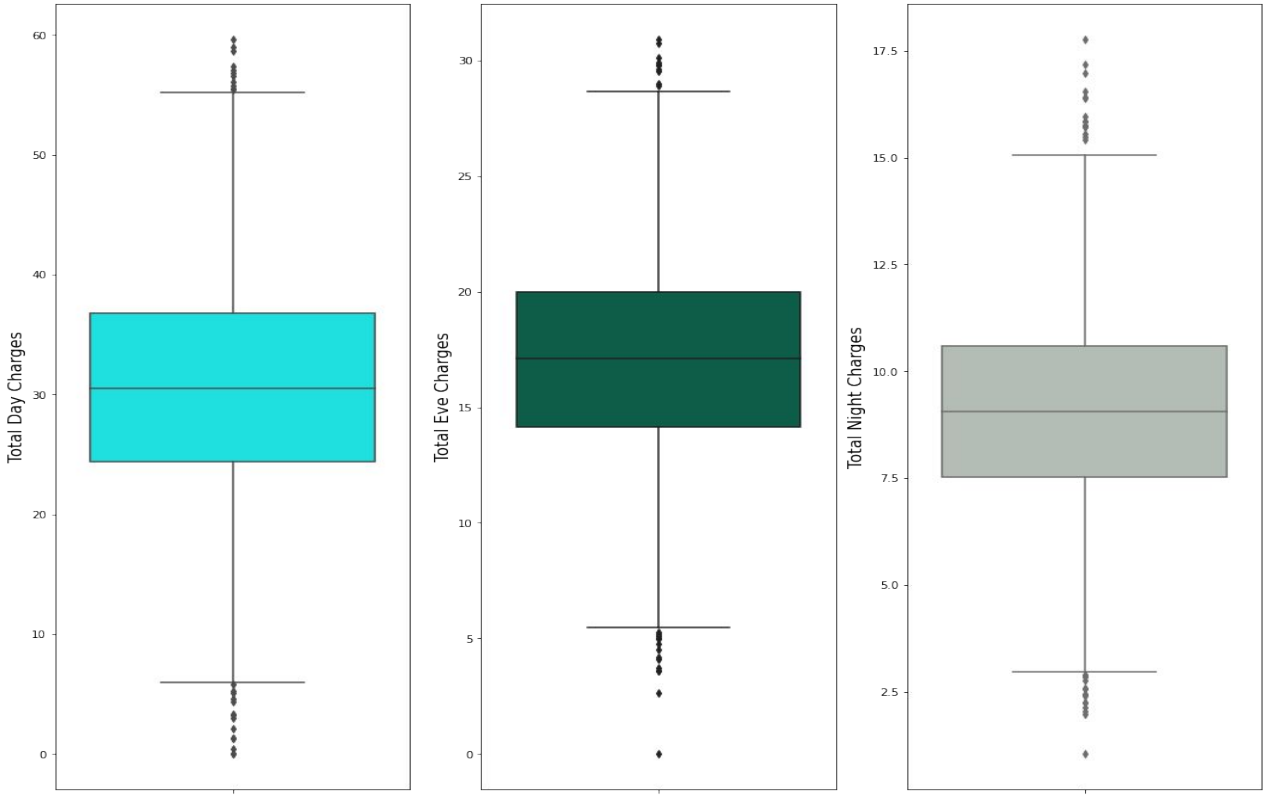






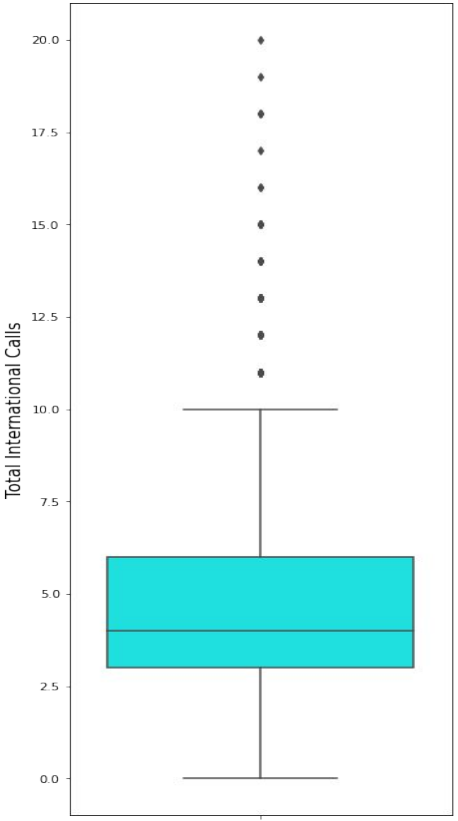
- Customers talked more during evening and less during night.
- Despite the fact that most calls were made during the daytime, customers did not spend much time talking over phone.
- Some outlier treatment is necessary for the extreme points in Total Eve Minutes and Total Night Minutes.

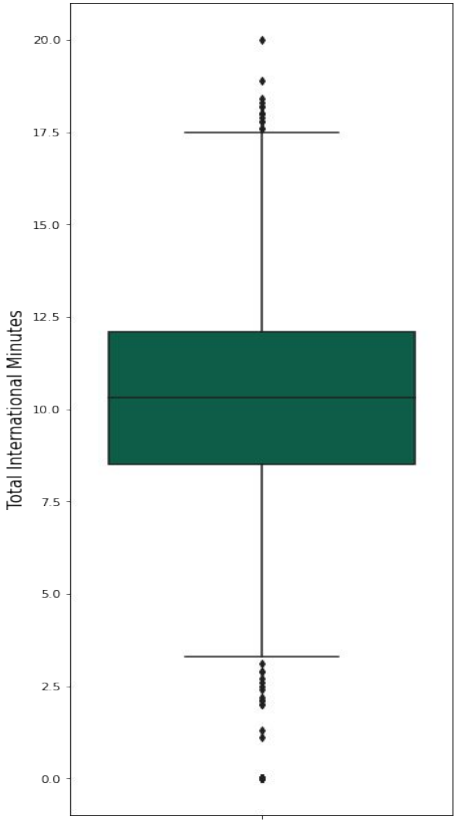
Univariate Analysis - Day, Eve, Night Charge

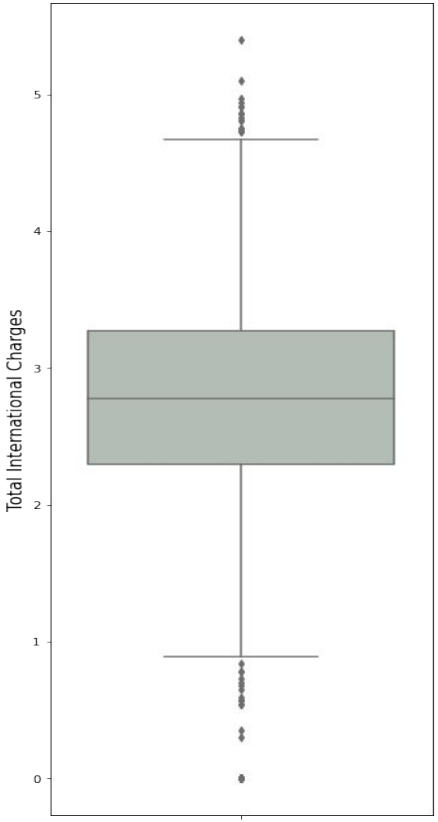


- The charges are maximum in the evening time and lowest in the night time.
- These plots are in conjunction with the minutes spoken.

Univariate Analysis - International Calls, Minutes, Charge

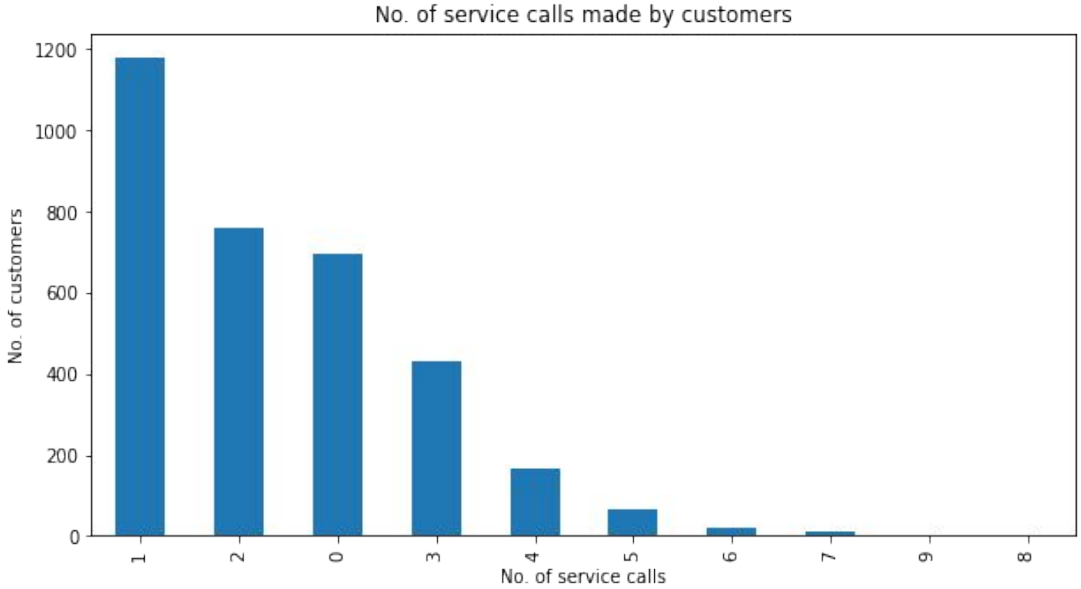


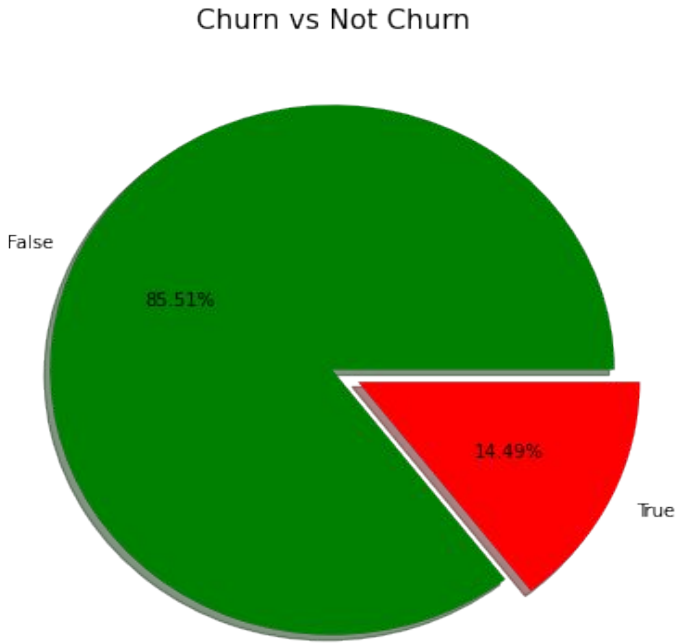




- Number of average International calls made is too low.
- Range of all three metrics are much lesser than domestic day, eve and night calls.

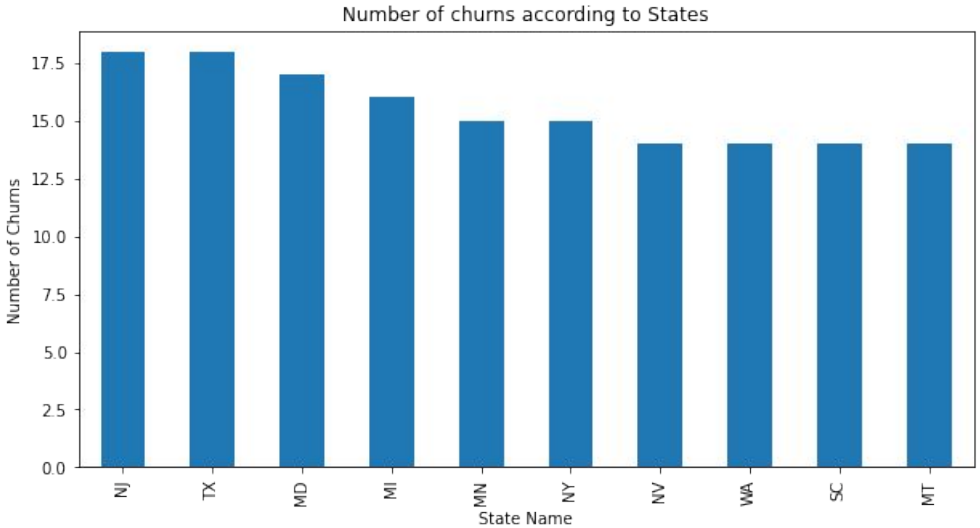
Univariate Analysis - Customer service calls, Churn

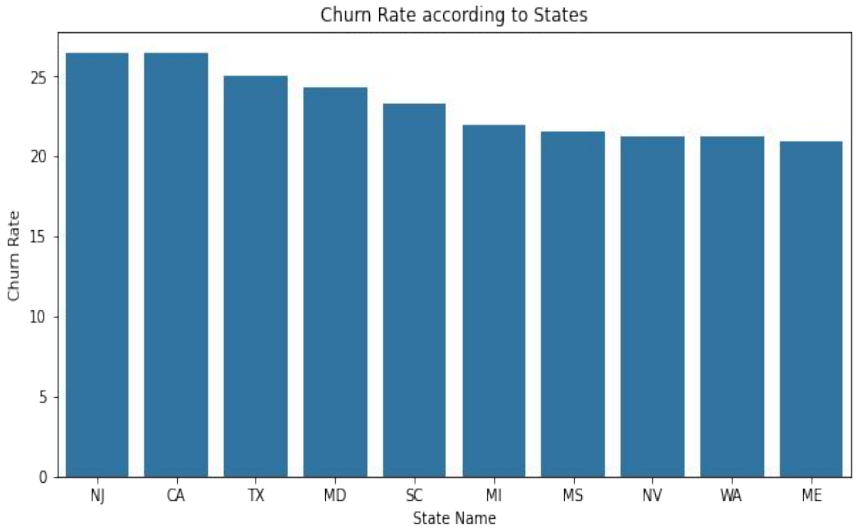




- **Probability of making more than 5 service calls is minuscule.**
- **Almost 14.49% of total customers have left the service.**

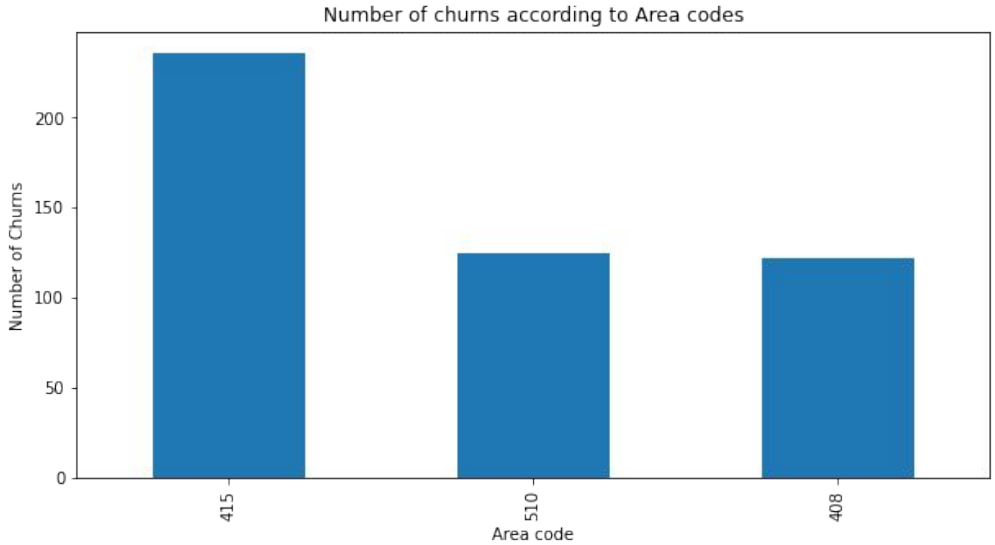
Bivariate Analysis - State vs Churn

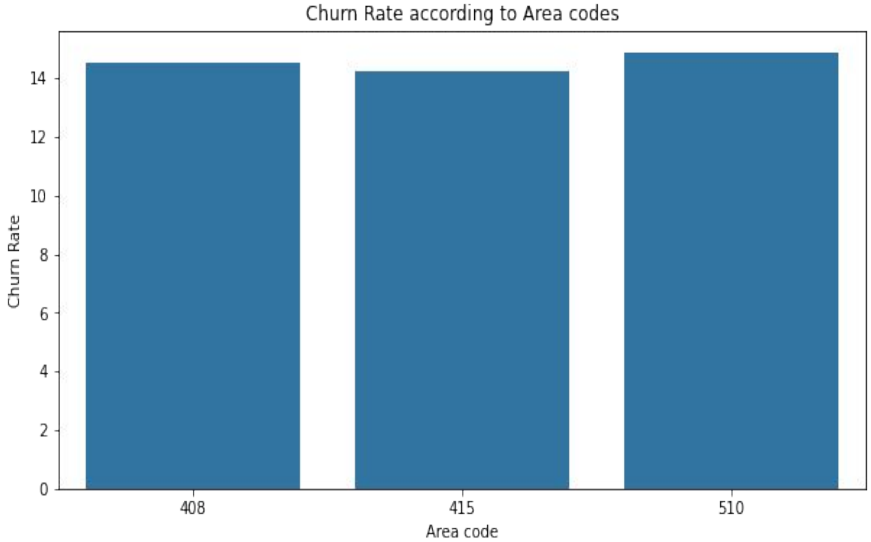




- NJ has highest number of churns as well as highest churn rate.
- Although CA doesn't have many customer churns or many customers, it has high churn rate.
- Though NY and MN have high customer churns, they don't have high churn rate.

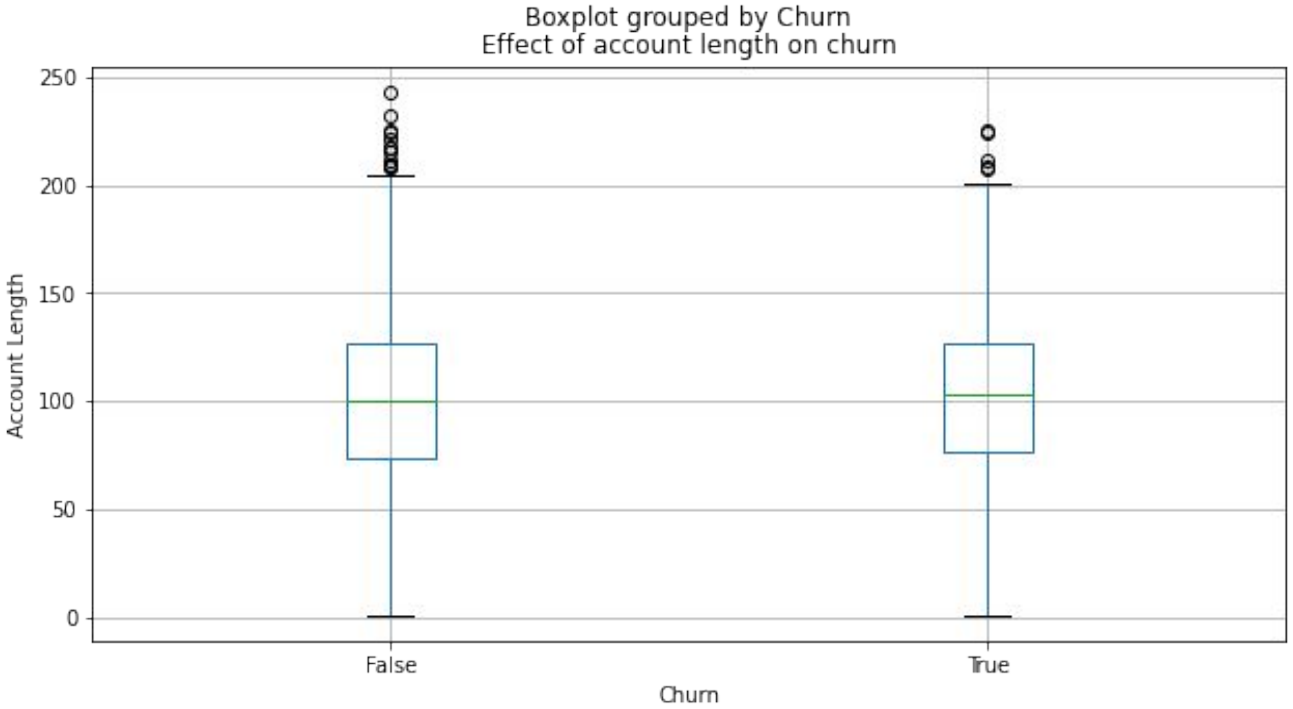
Bivariate Analysis - Area vs Churn





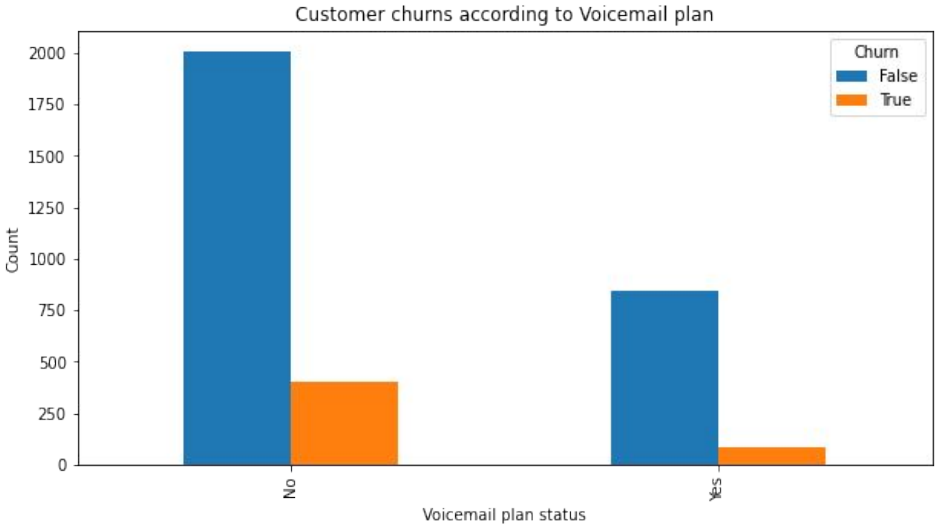
- **Area 415 has highest number of churns.**
- **But the churn rates for all 3 areas are almost similar. Which means customers are likely to churn in any area.**

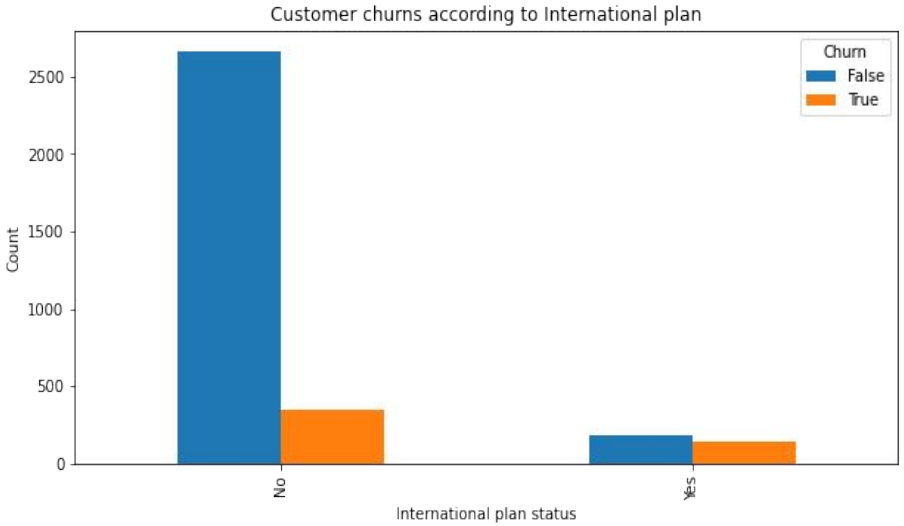
Bivariate Analysis - Account length vs Churn



- **It is clear from the above graph that customers have similar probability to churn irrespective of the length they are using the service.**

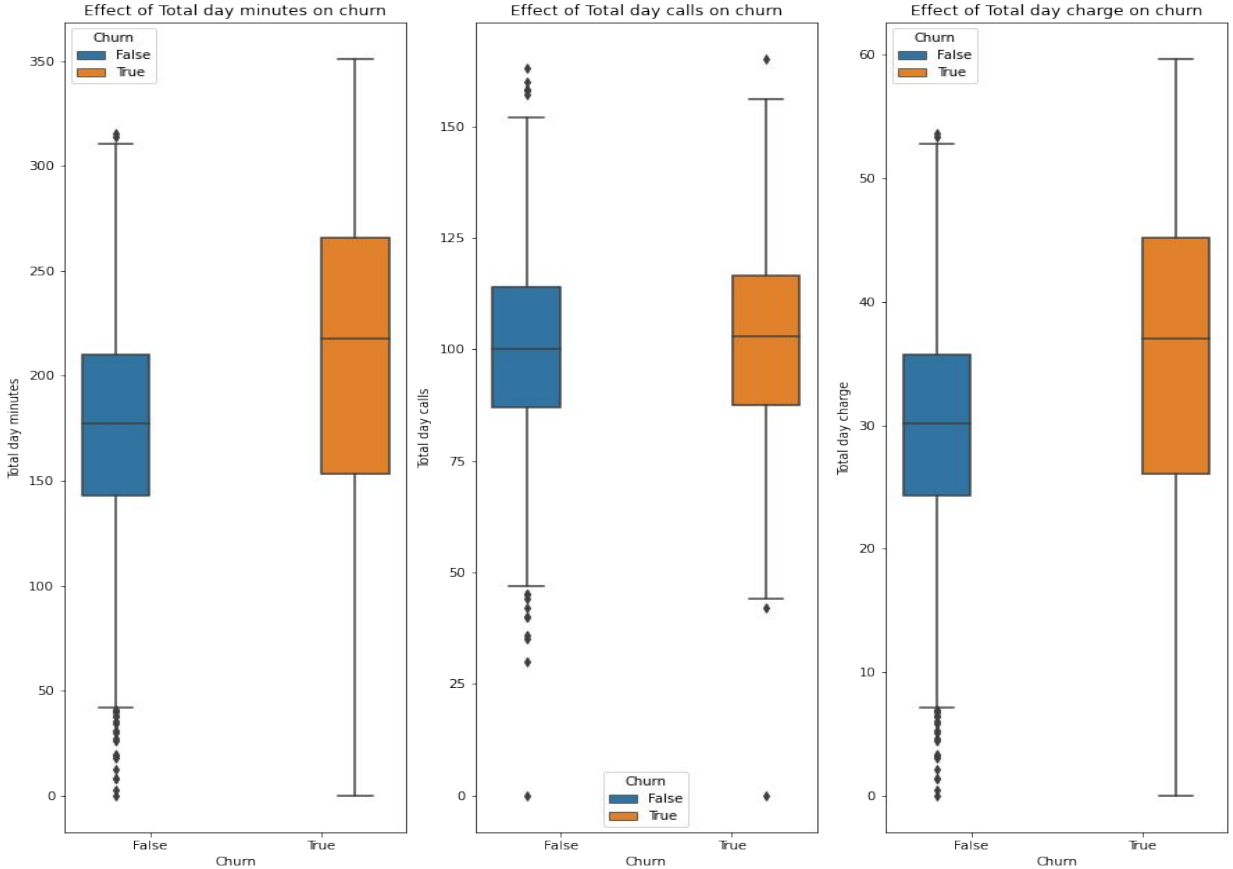
Bivariate Analysis - International plan, Voice mail plan vs Churn





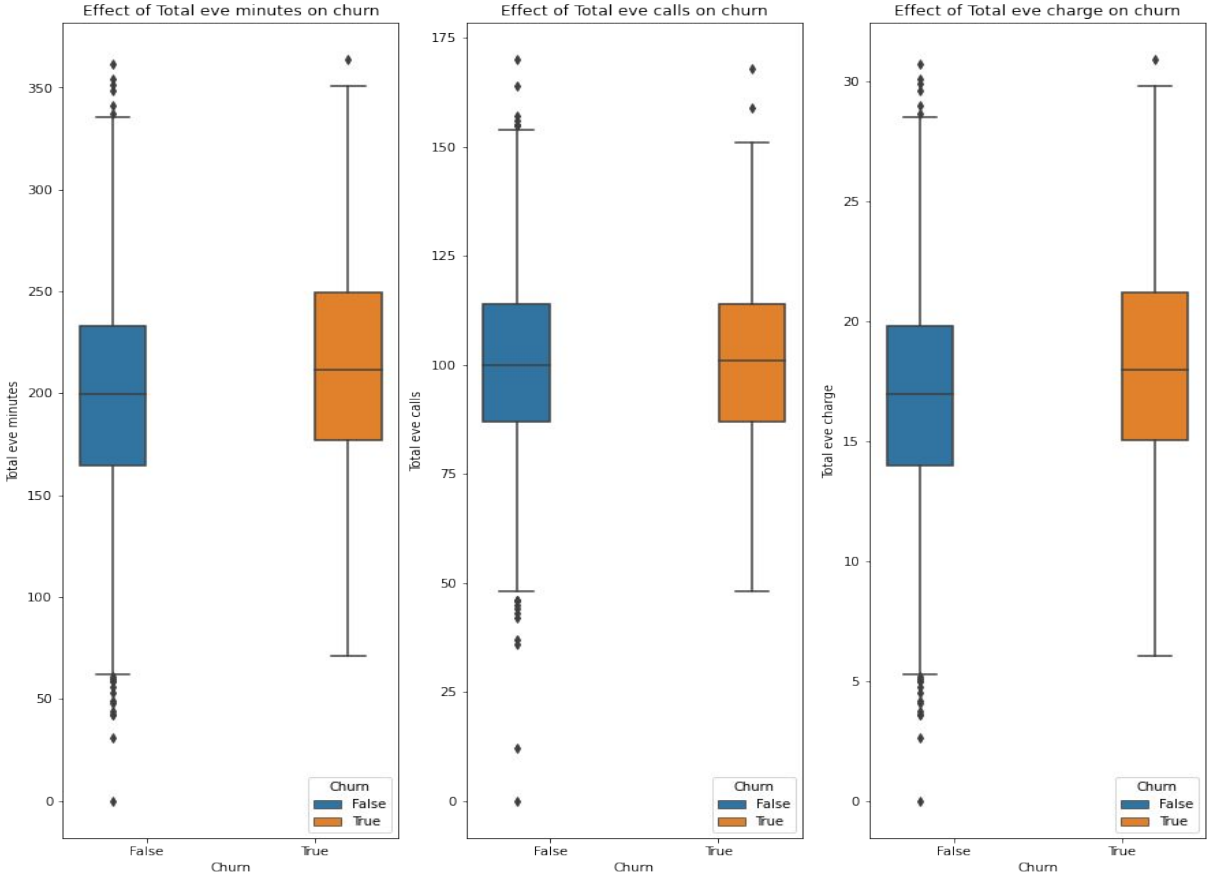
- Customers with or without voicemail plan are equally likely to churn.
- Although customers with international plans are lesser in number, they are more likely to churn.

Bivariate Analysis - Day minutes, calls, charge vs Churn



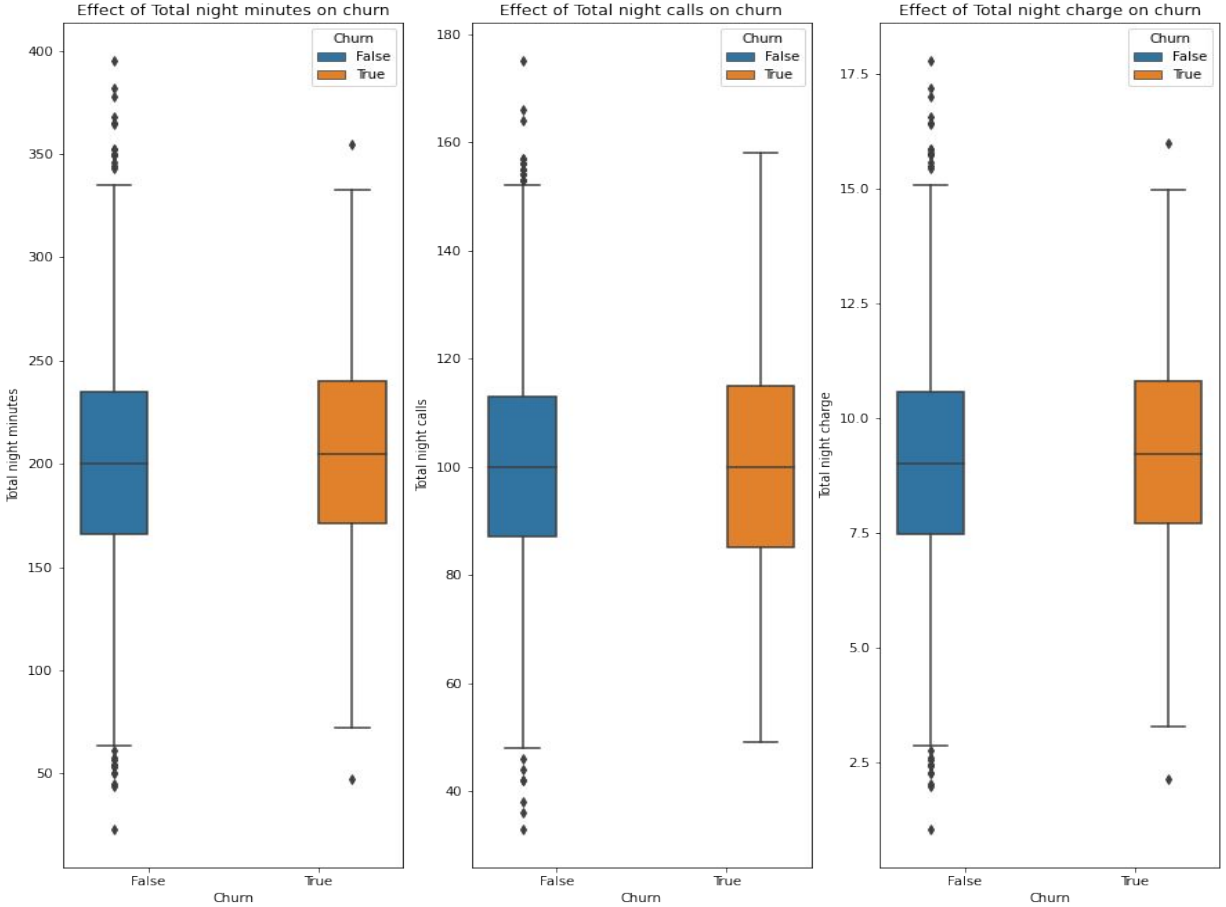
- People who are talking more during daytime and are being charged more, are churning more.

Bivariate Analysis - Eve minutes, calls, charge vs Churn



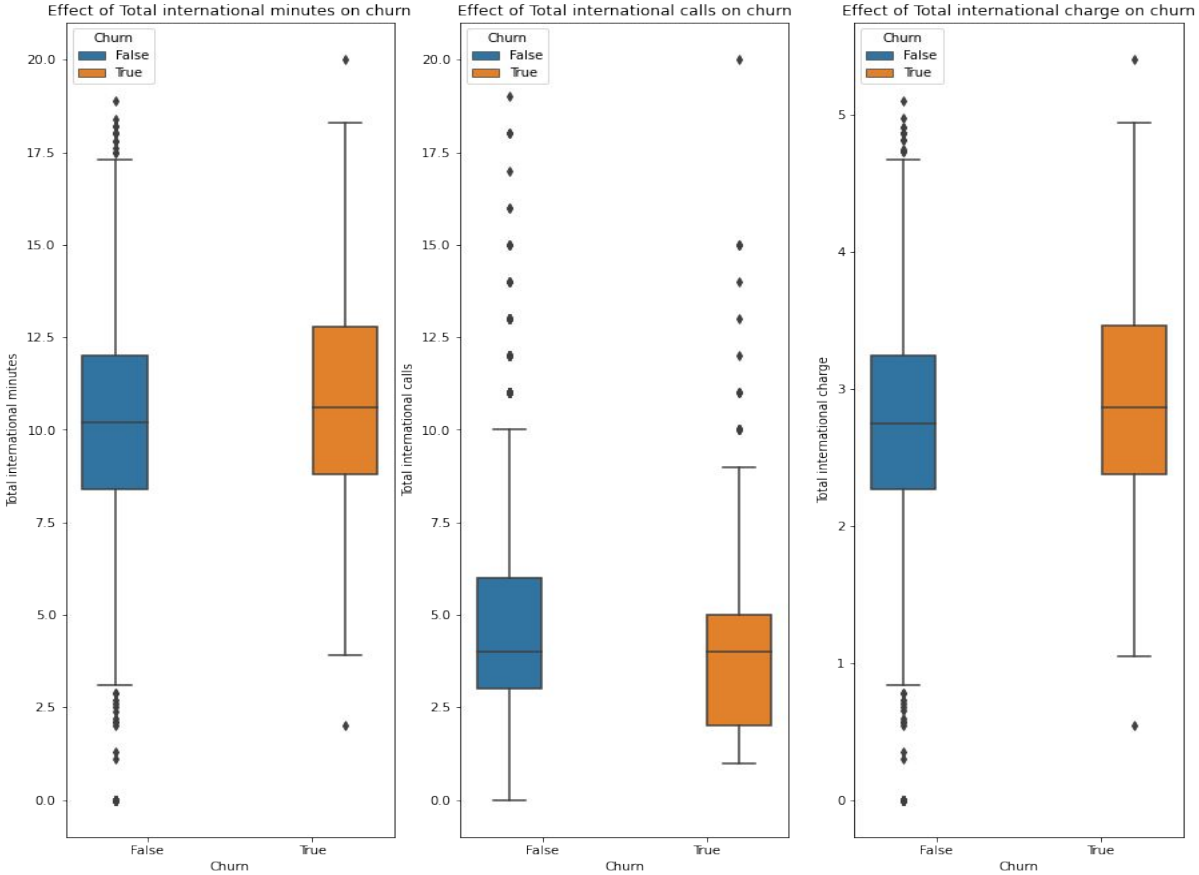
- People who are talking more during evening and are being charged more, are churning more.

Bivariate Analysis - Night minutes, calls, charge vs Churn



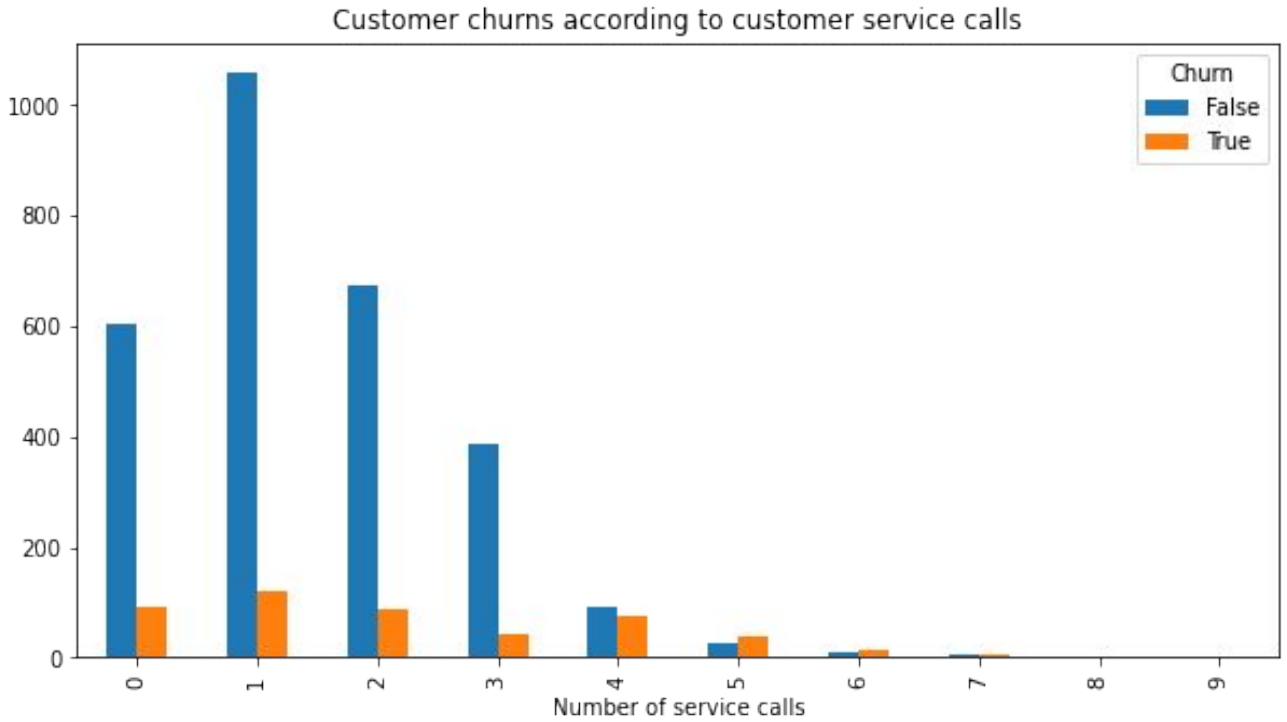
- **There is almost no effect of night metrics over churn.**

Bivariate Analysis - International minutes, calls, charge vs Churn



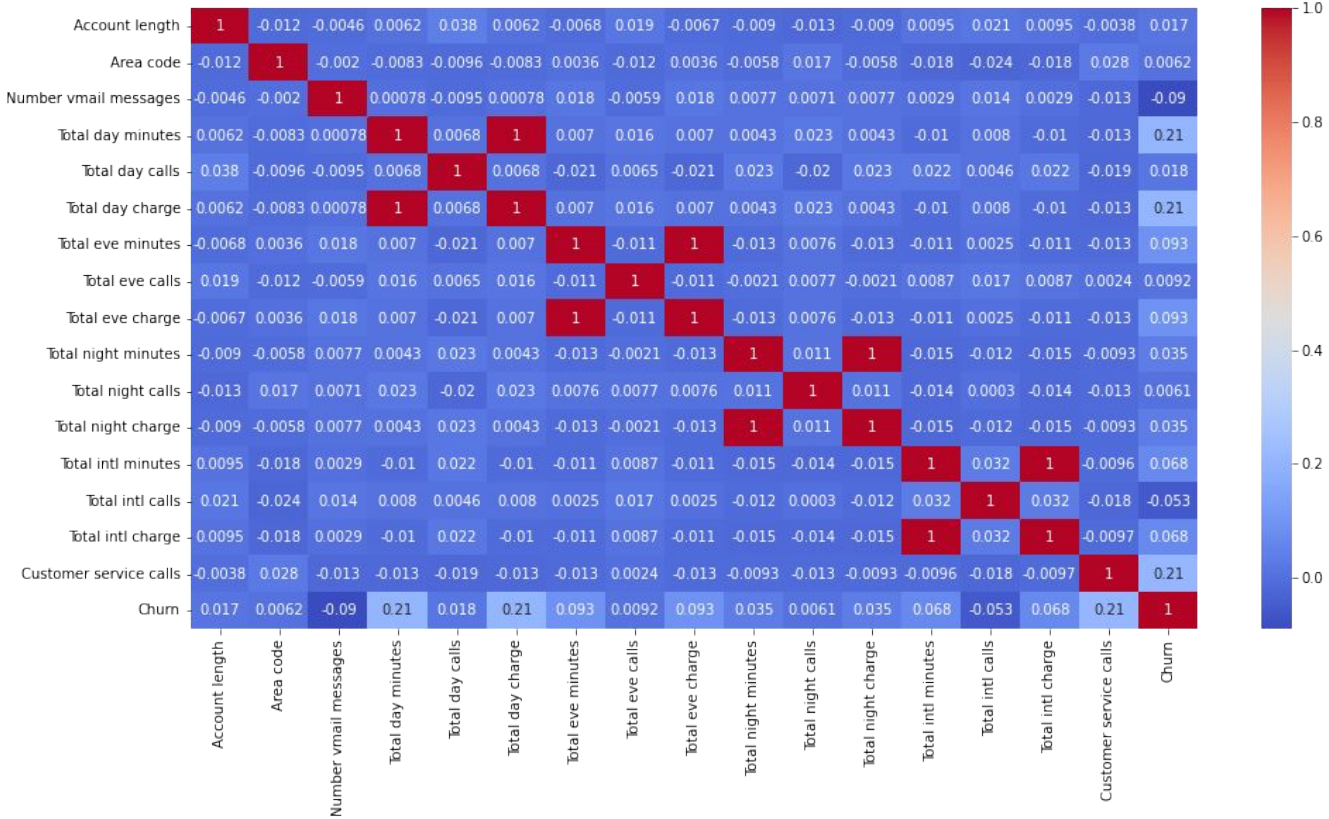
- People who are talking more internationally and are being charged more, have slightly higher chance of churning.

Bivariate Analysis - Customer service calls vs Churn



- **If more than 3 service calls are made, customers are more likely to churn.**

Multivariate Analysis - 1



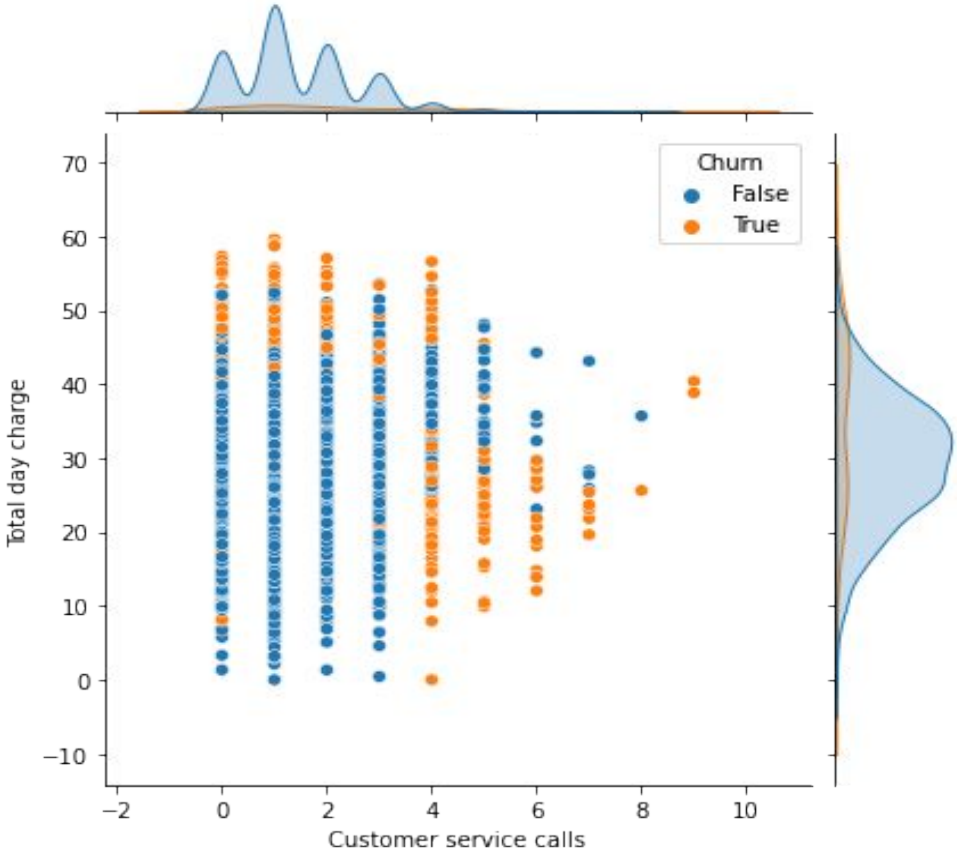
We can see that there is a perfect positive correlation between talking minutes and charges for day, eve, night and international calls. It means that prices charged for talking is perfectly balanced throughout any time of the day or international calls.

- **Note :**
- **If we want to fit a logistic regression model to the data, we have to take care of this high correlations or multicollinearity.**

Some other notable relations are:

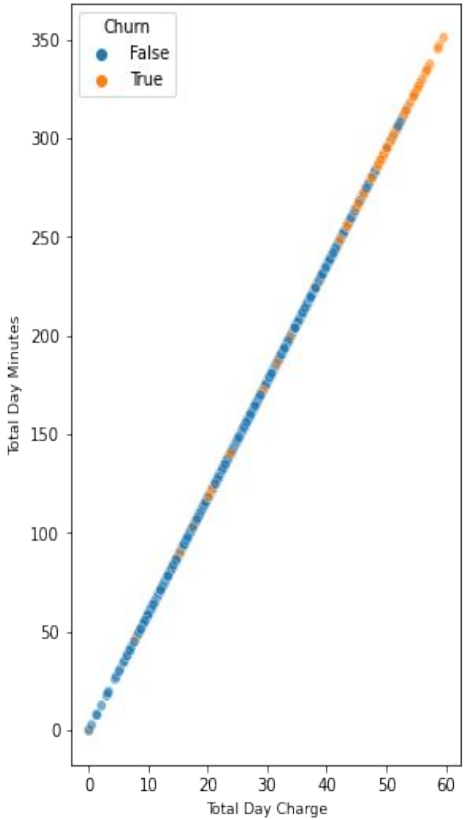
- **Total day minutes to Churn at 0.21**
- **Total day charge to Churn at 0.21**
- **Customer service calls to Churn at 0.21**

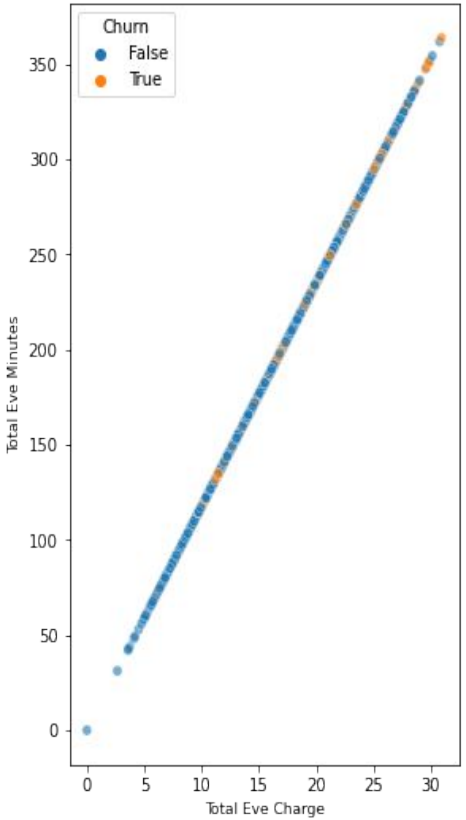
Multivariate Analysis - 2

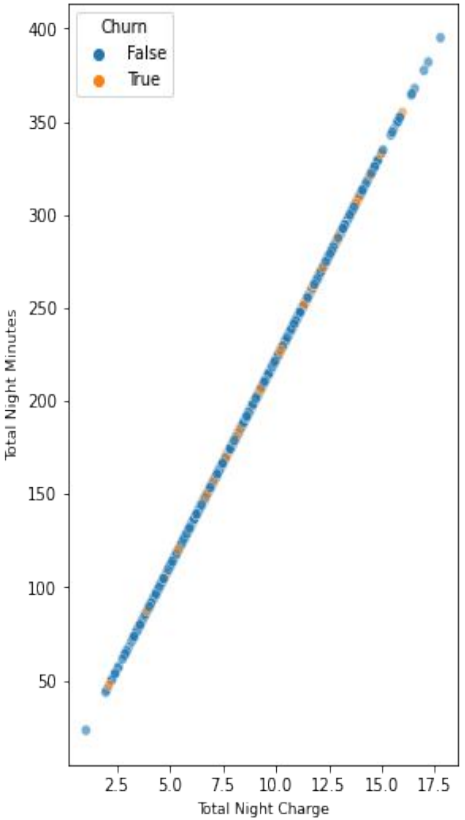


- If service calls are made more than 8 times, customers are bound to churn.
- If service calls are made less than 5 times and customers are being charged more, they have high probability of churning.

Multivariate Analysis - 3







- We can observe from the three scatterplots that total minutes talked and charges have most impact on the churn rate during the Day.
- In the evening and night we can see that the churn rate is not so relevant.

Final Verdicts :

Important Features:

1. Based on my observation I can see that most people who leave the service are the ones who use the service in the day mostly.
2. It can also be observed that most people who use the service in the morning, speak for shorter amounts of time but make more calls.

- From above two points, it is clear that daytime metrics play important role in deciding churns.
- Opting for International plans is a driver for churns.
- Number of customer service calls has a positive impact on churns.
- Area codes have a good impact on churns but again not on churn rates.

According to my analysis, these are the most important features that are driving the churns mostly.

Recommendations to retain customers:

- Introducing plans which minimize costs for more number of calls can be used.
- Decreasing the prices as the talk-time increases can be an effective way to reduce the churn.
- Improvement in the customer service can be done to reduce the number of calls which cause the churn.
- Improving International plan benefits might help.
- Improving services for high churning areas might help.