Customer churn Analysis in Orange S.A telecom

Abstract:

Customer churn is a major problem and the most concern for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customers to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn.

1. Problem Statement

Given data consists of customer activities of a French multinational telecommunications company, named Orange S.A. Data provided is of 51 states of the USA with 3 area codes. The activities of customers such as total day , night calls made for a particular day are provided.

The company provides a range of services to customers like **International Plan**, **Voice mail plan** etc.

But then also, in the last few years, the company has seen that some customers are not satisfied with their service and has left the subscription and has switched to a different service provider.

The main objective is to analyze the main factors for which churn is increasing which will, in turn, assist the telecom service provider in predicting the customers who are most likely subject to churn and take necessary measures in retaining them.

We have the following variables in the dataset:

1. **State** - 51 Unique States in United States of America (Categorical)
2. **Account Length** - Length of The Account
3. **Area Code** - 415 relates to San Francisco,408 is of San Jose and 510 is of City of Oakland (Categorical)
4. **International Plan** - Yes Indicate International Plan is Present and No Indicates no subscription for International Plan (Categorical)
5. **Voice mail plan** - Yes Indicates Voice Mail Plan is Present and No Indicates no subscription for Voice Mail Plan (Categorical)
6. **Number vmail messages** - Number of Voice Mail Messages ranging from 0 to 50
7. **Total day minutes** - Total Number of Minutes Spent by Customers in Morning
8. **Total day calls** - Total Number of Calls made by Customer in Morning
9. **Total day charge** - Total Charge to the Customers in Morning
10. **Total eve minutes** - Total Number of Minutes Spent by Customers in Evening
11. **Total eve calls** - Total Number of Calls made by Customer in Evening
12. **Total eve charge** - Total Charge to the Customers in Evening
13. **Total night minutes** - Total Number of Minutes Spent by Customers in the Night
14. **Total night calls** - Total Number of Calls made by Customer in Night
15. **Total night charge** - Total Charge to the Customers in Night

16. **Total intl minutes** - Total Number of Minutes Spent by Customers on international calls
17. **Total intl calls**- Total Number of international Calls made by Customer
18. **Total intl charge** - Total Charge to the Customers on international calls
19. **Customer service calls** - Total numbers of calls to customer service
20. **Churn** - True indicate that customer doesn't continue with the subscription and False indiacte customer continue with the subscription. (Categorical)

The dataset contains nineteen columns (independent variables) that indicate the characteristics of the clients of a telecommunications corporation. The Churn column (response variable) indicates whether the customer left the subscription or not.

2.Introduction:

The dataset is from Orange S.A., a telecom based company. There are 19 independent variables and 1 dependent variable (Churn). We have to explore all the independent variables and their effect on churn. The primary objective is to analyze the main factors for which churn is increasing will, in turn, assist the telecom service provider in predicting the customers who are most likely subject to churn and take necessary measures in retaining them.

3. Steps involved:

Step 1: Gather, Assess the data.

Step 2: Clean the data – Check for Missing values, check correlation among the variables and drop unnecessary columns.

Step 3: Take some Assumptions on the dataset.

Step 4: Conduct exploratory data analysis - check each variable and its dependency on the Churn variable and create visualizations.

Step 5: Understand limitations and sort out possible solutions

Step 6: Summarize

4. Data Cleaning:

- **Missing values and data types**:

At the beginning of EDA, we want to know as much information as possible about the data, this is when the pandas.DataFrame.info method comes in handy. This method prints a **concise summary of the data frame**, including the column names and their data types, the number of non-null values, and the amount of memory used by the data frame.

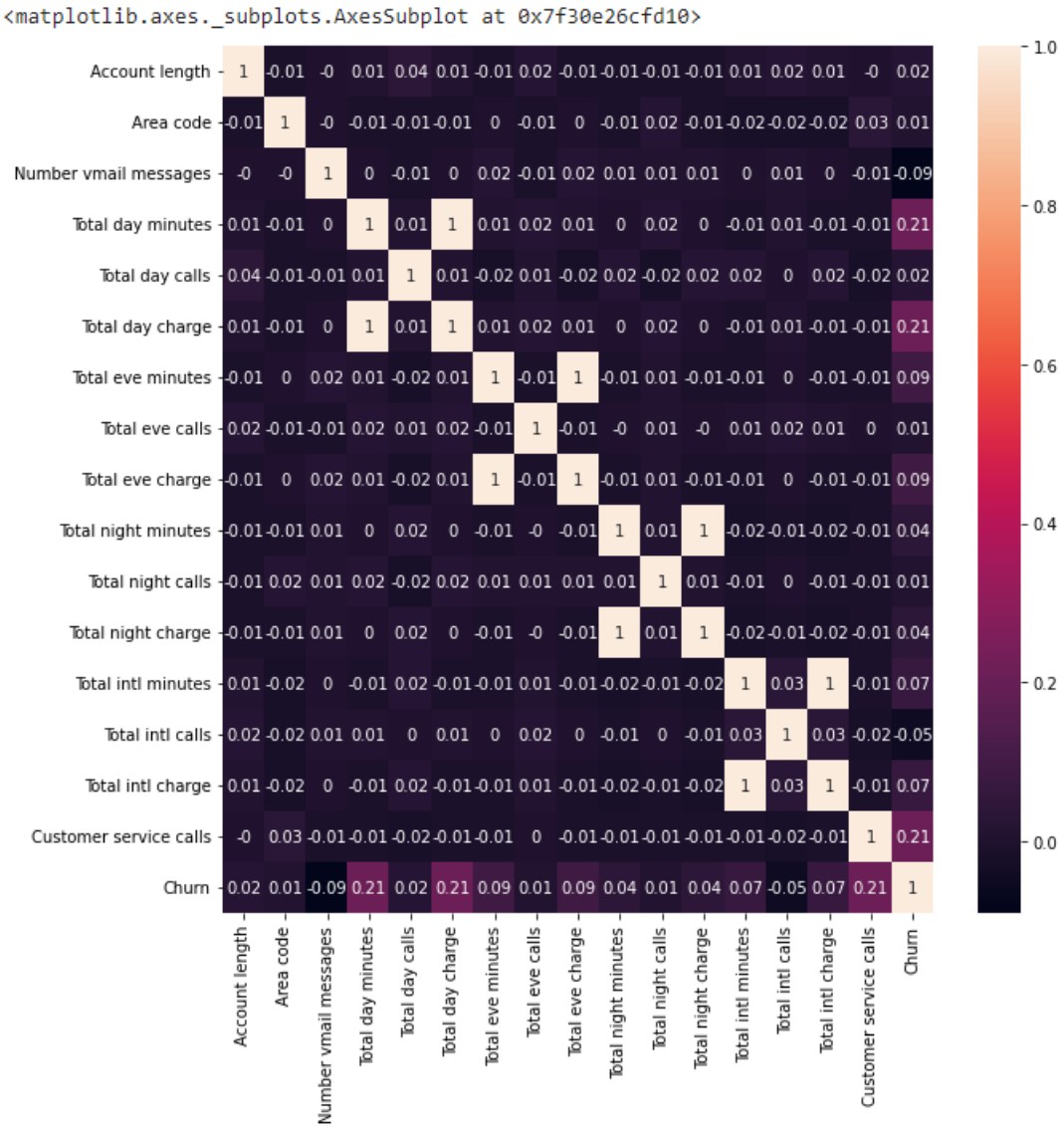
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   State                   3333 non-null   object 
 1   Account length          3333 non-null   int64  
 2   Area code               3333 non-null   int64  
 3   International plan      3333 non-null   object 
 4   Voice mail plan         3333 non-null   object 
 5   Number vmail messages   3333 non-null   int64  
 6   Total day minutes       3333 non-null   float64
 7   Total day calls         3333 non-null   int64  
 8   Total day charge        3333 non-null   float64
 9   Total eve minutes       3333 non-null   float64
 10  Total eve calls         3333 non-null   int64  
 11  Total eve charge        3333 non-null   float64
 12  Total night minutes     3333 non-null   float64
 13  Total night calls       3333 non-null   int64  
 14  Total night charge      3333 non-null   float64
 15  Total intl minutes      3333 non-null   float64
 16  Total intl calls        3333 non-null   int64  
 17  Total intl charge       3333 non-null   float64
 18  Customer service calls  3333 non-null   int64  
 19  Churn                   3333 non-null   bool   
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 498.1+ KB
```

As shown above, the data set contains **3333 observations** and **20 variables**. Apparently, there are **no null** values on the data set.

Also, we will consider 3 categorical columns namely- State, International plan and Voice mail plan since its datatype is object. We will separately analyse these variables and find out the respective Churn rate. Rest are continuous variables.

- **Correlation**:

The correlation between the columns is shown below:



A value of exactly 1 means **there is a perfect positive relationship between the two variables**.
So, From the heat map, we see that, there is **strong correlation between the variables:**

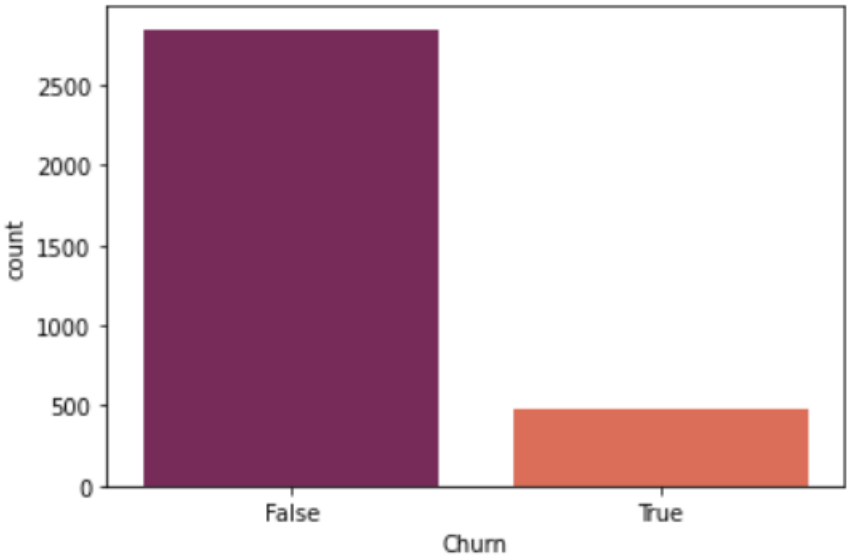
1. Total day minutes vs total day charge

2. Total eve minutes vs total eve charge
3. Total night minutes vs total night charge
4. Total intl minutes vs total intl charge

Therefore we can choose any one variable from each pair for analysing the churn and drop the other variable.

5. Exploratory data analysis and Data Wrangling:

- **Analysing dependant variable CHURN:**



From this chart, we see that the true values of Churn (customers discontinuing the subscription) is comparatively much lesser than the False value of Churn(customers continuing the subscription).

Around 14% customers are leaving the subscription.

So we can say that the service of the Orange telecom company is more or less good, but they may need slight modifications on the operations or the existing plans.

Let’s analyse the factors on which the company can take necessary actions to reduce Churn.

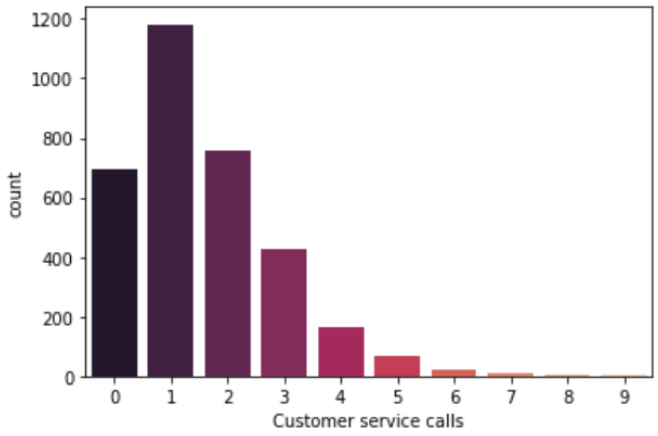
- **Correlation between Churn and other continuous variables:**

```
df.corr()['Churn'][:-1].sort_values(ascending=False)
```

```
Customer service calls    0.208750
Total day minutes         0.205151
Total day charge          0.205151
Total eve minutes         0.092796
Total eve charge          0.092786
Total intl charge         0.068259
Total intl minutes        0.068239
Total night charge        0.035496
Total night minutes       0.035493
Total day calls           0.018459
Account length            0.016541
Total eve calls           0.009233
Area code                 0.006174
Total night calls         0.006141
Total intl calls         -0.052844
Number vmail messages    -0.089728
Name: Churn, dtype: float64
```

From the above we can see that the variables **Customer service calls** and **Total day minutes (Total day charge)** are slightly related to **Churn**. As for the other variables, the correlation is negligible. So we will analyze only these two continuous variables.

i) **Customer service calls:**



Above chart shows the number of counts (customers), customer service calls were made. We can see that most Customers calls customer service only one time. There are less customers who call customer service multiple times.

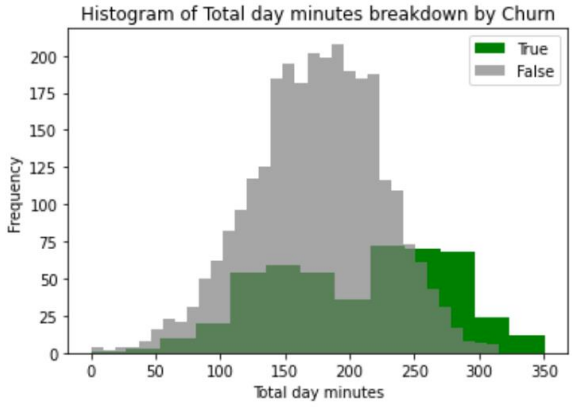
	calls	total_customers	churn	churn_per
7	9	2	2	100.000000
8	6	22	14	63.636364
5	5	66	40	60.606061
6	7	9	5	55.555556
9	8	2	1	50.000000
4	4	166	76	45.783133
1	0	697	92	13.199426
2	2	759	87	11.462451
0	1	1181	122	10.330229
3	3	429	44	10.256410

By getting into further deep insights and linking the service calls with Churn rate, we found that:

1. The customers who have called more than 3 calls have a minimum of 45% chance of leaving the subscription.
2. The customers who have called less than or equal to 3 calls have a maximum of 13% chance of leaving the subscription.

So, we can conclude that churn rate increases with increase in customer service calls.

ii) **'Total day minutes'**



From the above histogram plot, we can clearly observe that customers who has longer talk time in the morning has a high chance of discontinue the subscription.

- **Analyzing the categorical variables:**

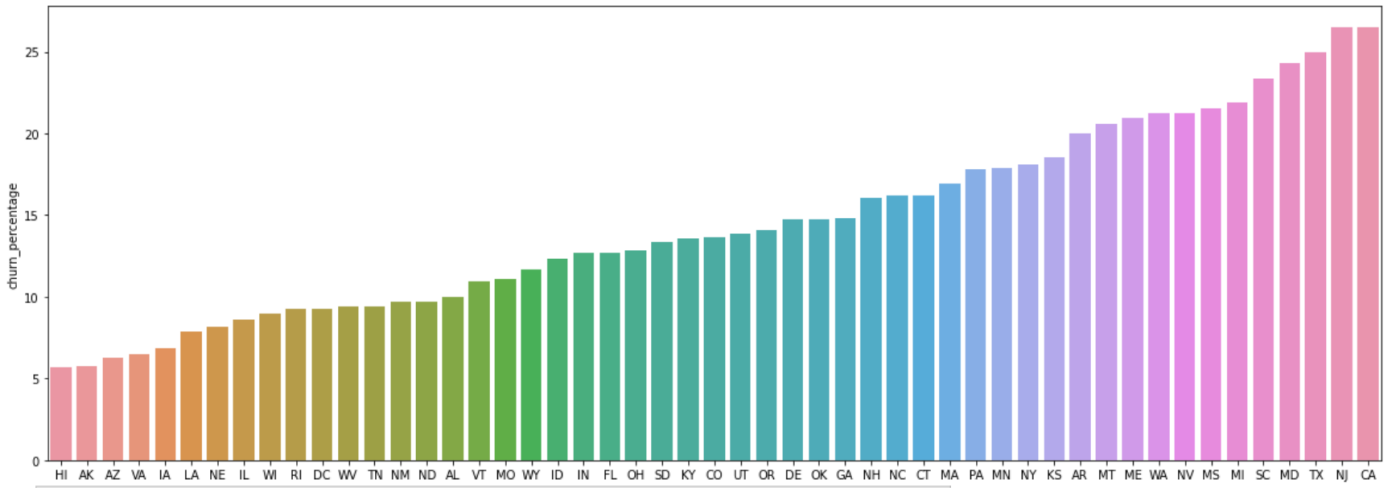
i) State:

By linking churn with state and customers, we deduce top 5 states with highest Churn rate:

```
top_churn=state_wise_customer_churn.sort_values('churn_percentage', ascending=False)
top_churn.head()
```

	State	Total Customers	Churn	churn_percentage
50	CA	34	9	26.47
19	NJ	68	18	26.47
13	TX	72	18	25.00
16	MD	70	17	24.29
36	SC	60	14	23.33

Plotting all the states with churn rate:



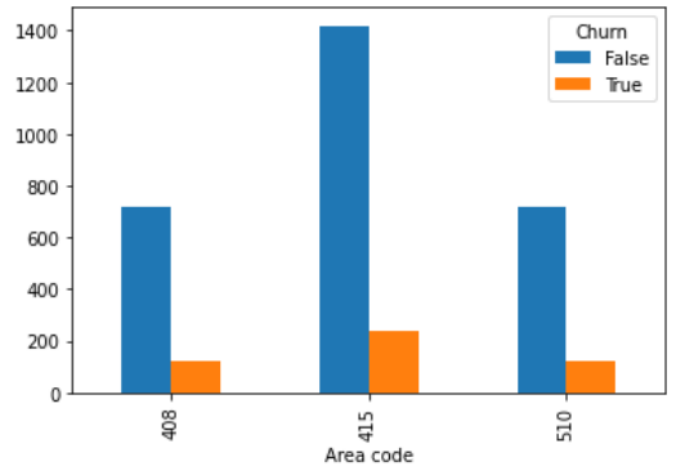
It is clear that CA(California) and NJ (New Jersey) has the highest churn percentage and HI(Hawaii) & AK(Alaska) has the lowest churn percentage. It may be due to good network or lack of competition in HI & AK compared to CA & NJ.

ii) Area:

Note: In the given dataset, since there are only 3 unique area codes, we can consider it as categorical variable. Inside one area code, there are many states.

By plotting the churn rate with area code**:**

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f58b0abd710>
```




```
area_wise_churn
```

Churn	Area code	False	True	churn_per
0	408	716	122	14.558473
1	415	1419	236	14.259819
2	510	715	125	14.880952

Since churn percentage is almost same in all the area codes, we can conclude that Churn rate is not dependent on Area code. So, we can drop the area code.

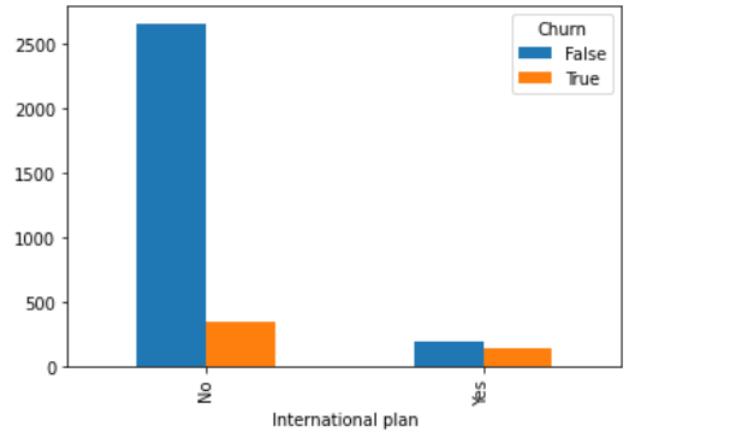
3) International plan:

Plotting the churn with each International plan:

Churn	False	True
International plan		
No	2664	346
Yes	186	137

```
international_plan.plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f72ca484bd0>
```



We can see that the customers who have international plan, have high chance of leaving the subscription. It may be because customers are not happy with the international plan.

By getting more deep, we observe that the per minute international call charge is same for both international and non- international plan.

Even if customer takes international plan the call charge is same as customers without international plan. Thus customers did not get any benefit even after taking international plan. This may be the reason why there is a high chance of churn for customers with international plan, because they may be expecting less charge for international plan(discount) or some other benefits.

Let's check the churn rate for customer who spent more time on international calls:

```
df[(df['International plan'] == 'Yes')]['Total intl minutes'].describe()
```

```
count    323.000000
mean      10.628173
std        2.697787
min        1.300000
25%        9.000000
50%       10.800000
75%       12.200000
max       20.000000
Name: Total intl minutes, dtype: float64
```

```
df[(df['International plan'] == 'Yes') & (df['Total intl minutes'] >= 12)
```

```
70.96774193548387
```

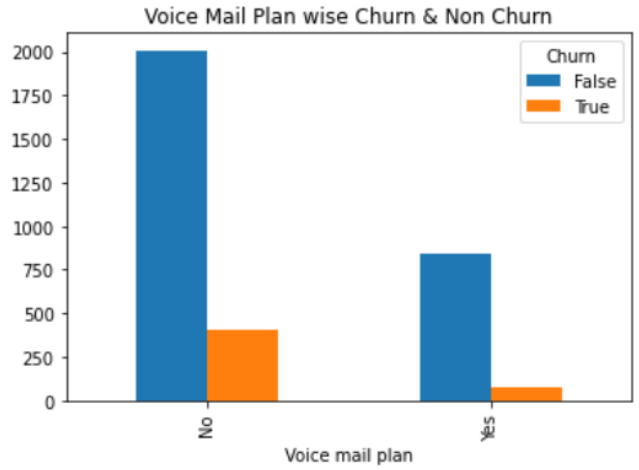
```
df[(df['International plan'] == 'Yes') & (df['Total intl minutes'] < 12) &
```

```
30.869565217391305
```

For the customers having International Plan those with talk time greater than 75% (>=12 min) have a higher chance of leaving than those with less talk time.

4) Voice mail plan:

Plotting the churn rate with voice mail plan:




```
# churn rate for no voice mail plan
voice_plan_churn.iloc[0,1]/(voice_plan_churn.iloc[0,0]+voice_plan_churn.iloc[0,1])*100
```

```
16.71505599336375
```

```
# churn rate for voice mail plan
voice_plan_churn.iloc[1,1]/(voice_plan_churn.iloc[1,0]+voice_plan_churn.iloc[1,1])*100
```

```
8.676789587852495
```

We can see that for the customers with no voice mail plan, there is some chance of leaving the subscription.

6. Observations:

1. Churn rate increases with an increase in customer service calls.
2. Customers who have longer talk time in the morning has a high chance of discontinuing the subscription.
3. High population area has a high churn rate.
4. Customers with an international plan, have a high chance of leaving the subscription. It increases even more for those with longer international calls.
5. For customers with no voice mail plan, the chance of discontinuing the subscription is greater than with a voice mail plan.

6. Conclusion:

From the observations, we can conclude that the company needs to work on the following fields for customer retention:

i) Offer better service in most populated areas.
ii) Provide better long talk time plans especially for day time customers.
iii) Take feedback and suggestions on a regular basis, try to implement it and strive for better communication.
iv) Company can offer better International plans.
v) Offer more incentives in the form of discounts and cashbacks to old customers to retain them.