Exploratory Data Analysis on the Telecom Dataset

Introduction

The data contains information about a telecoms company. It details raw data about activities of the consumers like how many minutes users spend on a call, the status of the consumers and the general information of loyal vs disloyal consumers. It has 3333 entries and 20 columns or features. The data is composed of numerical data and categorical data. It's divided into integers, objects and floats.

The goal is to perform an EDA on the data to extract insight from the data. The primary goal among others is to evaluate to what extent the churn rate is influenced by other features within the data set. This insight can help the company make business decision with regards to retaining users and improving their services to retain customers

Exploration of the data

The features of the data fall within categorical, quantitative and ordinal forms of data. To explore statistics various functions like describe were used to see the statistical characteristics of our numerical data. This helped us understand the mean, standard deviation, quartiles and the count which was helpful in understanding the data.

Handling Inconsistences

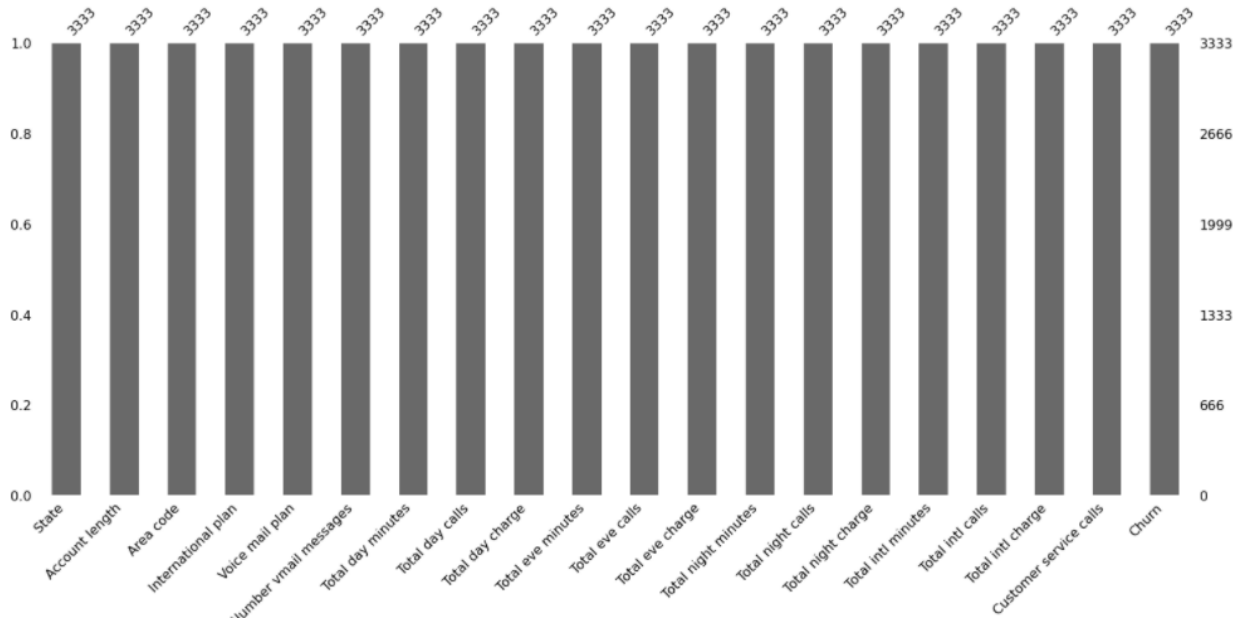
As explained above, the data contains integers, floats, Booleans and objects. Only one feature (churn) was changed from Boolean to an integer. This transformed Boolean True to 1 and Boolean 2 to False. The purpose of this transformation makes it easy to compute and perform an analysis on this feature called churn.

Besides this, there were not inconsistences within the data for the features we wanted to perform analysis on.

Another important thing is duplication. We checked our dataset to make sure that there were no duplicates. Using the duplicated function we discovered that there is no duplication.

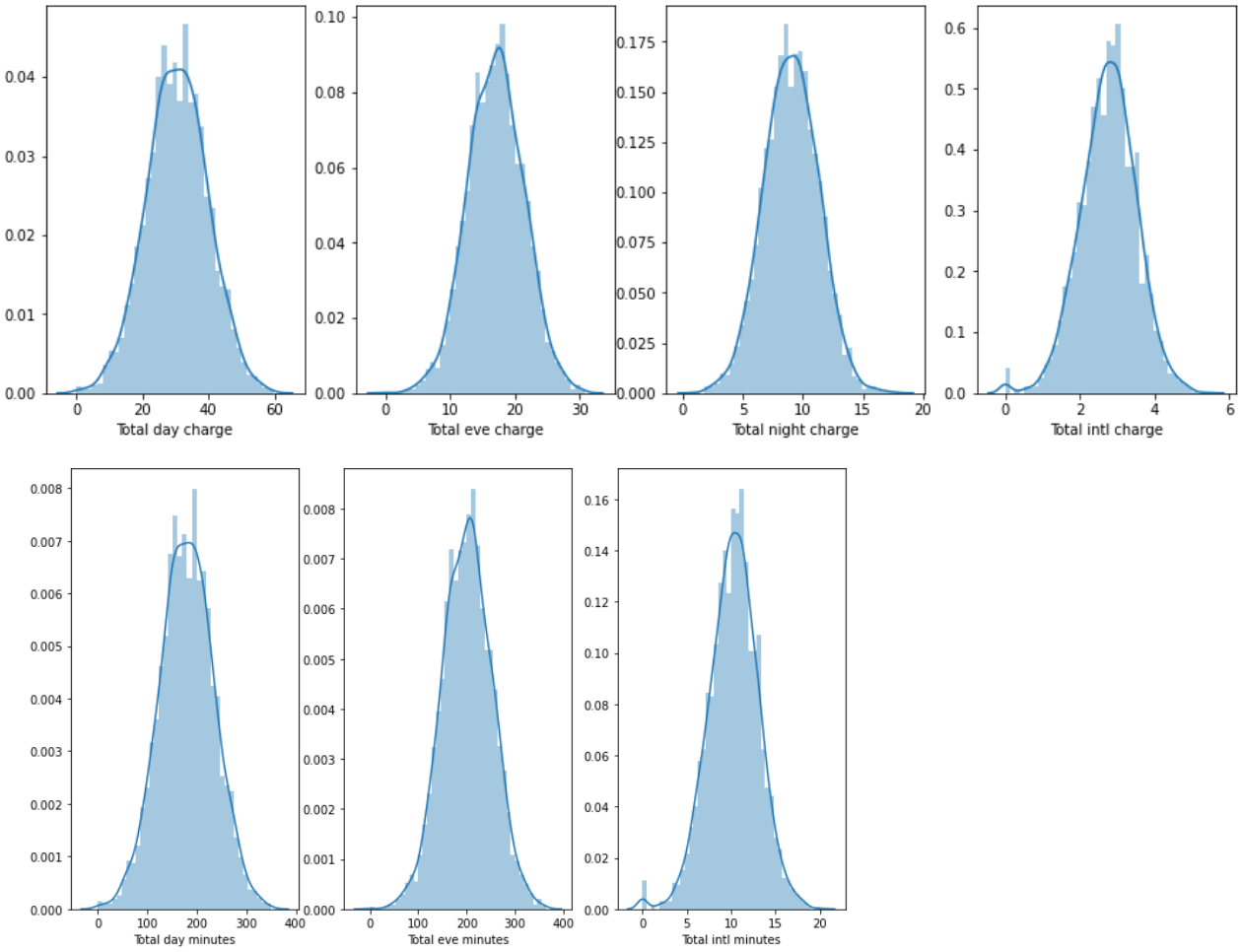
Handling Missing data.

Creating a model with missing data may cause some bias and lead to wrong inferences and conclusions. Before I performed the analysis I cleaned the data. The missingno module was used in order to visualize some null values or missing values in the data frame. Upon use, no data was missing. This was further confirmed by calculating the percentage of null values in the data, this also proved that there is no missing data.



Data Analysis

We start by understanding the data distribution of the features we want to extract information from. We do this by plotting some histograms, from these histograms we deduced some statistical information. This is shown below.



So what form of observation or insight do we derive from these histograms?

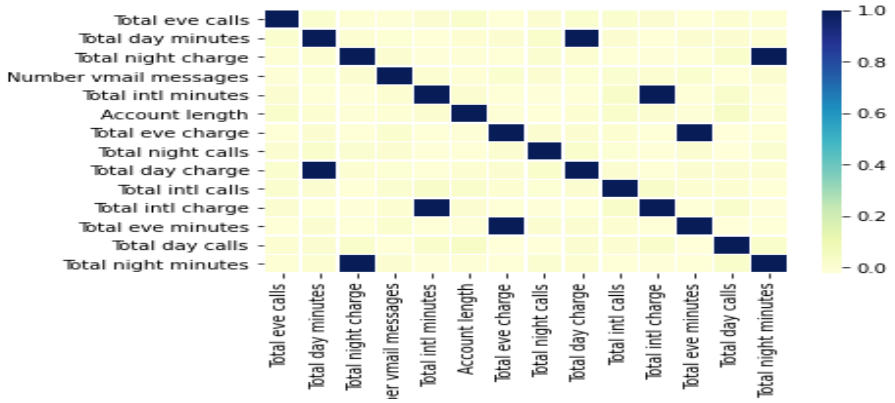
The minutes spent on the phone during the day, in the evening and on international calls follows a normal distribution. This means that our mode is closer to our mean and median. On average 200 minutes are spent on a call during the day by a consumer. This values is the same in the evening. On average 10 minutes are spent on international calls. We have less action when it comes to international calls. Most calls seem to be local not international.

Correlation

It's easy to notice that is a strong relationship between the total evening minutes and the total evening charges. This is the same for total day charges and the total night charges. We soon realize that there are some features which

are not related. This help us select the features which have a strong relationship and can help us predict outcomes. In this case our target variable is churn.

Let’s get a clear picture of the relationships within our features. The darker the color the stronger the relationship. We soon realize that there is a strong relationship between features like total eve minutes and total eve charges, this holds true for total day minutes as well.

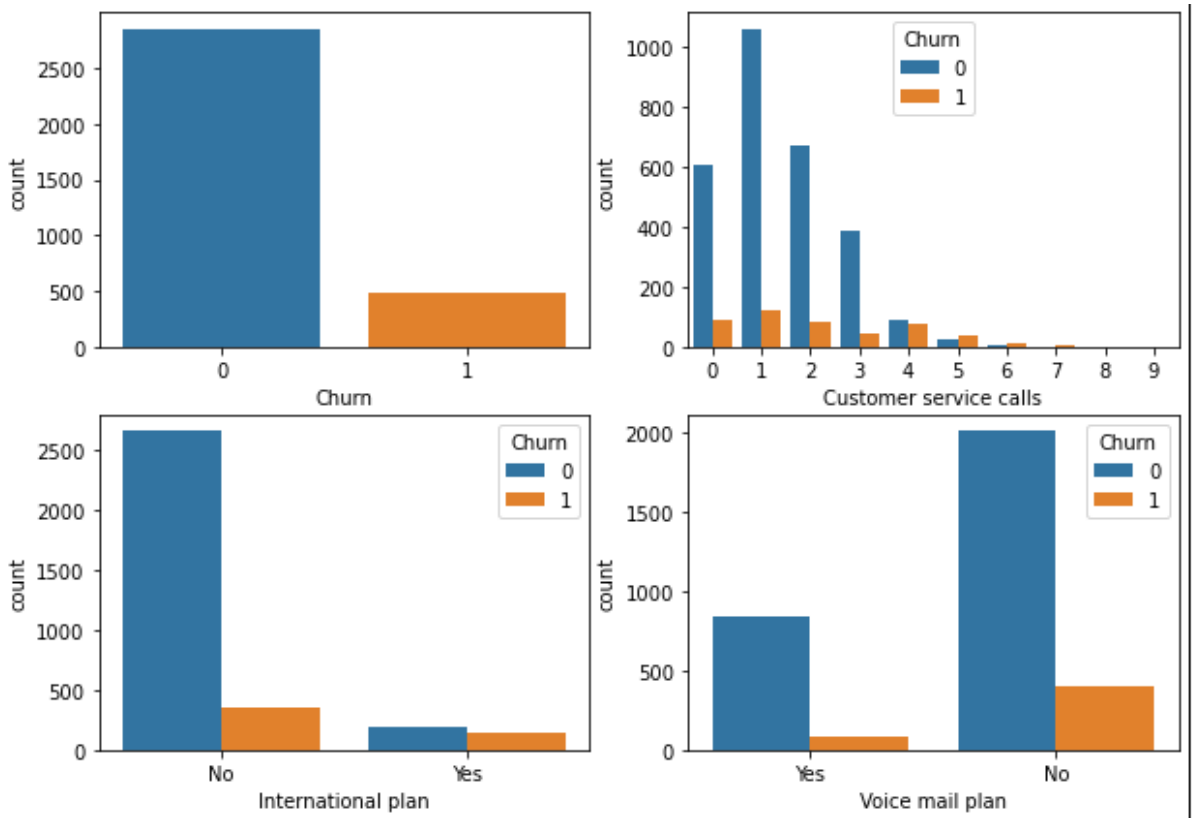


More Insights:

There are over 2500 loyal customers and about 500 churned. This information is useful in a business.

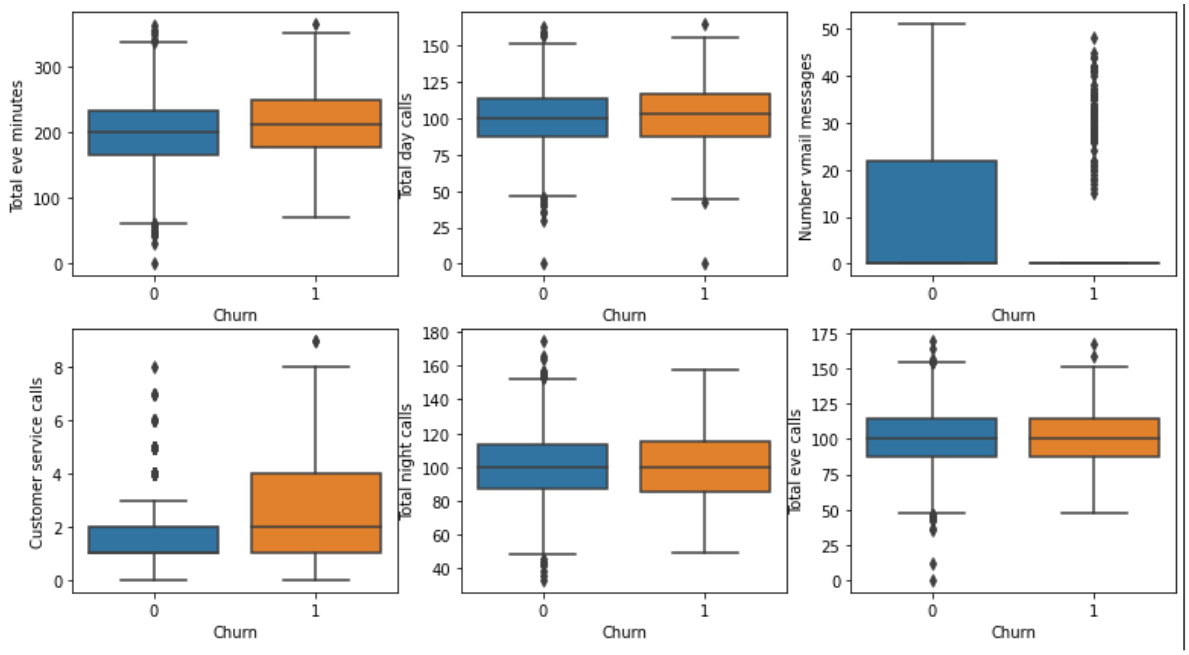
Loyal customers do not have international plans, most of them. And when it comes to voicemail plans, most churned customers have no voicemail plans and a fair number of loyal customers have voice mail plans. When International Plan is enabled, the churn rate is much higher, thus international is one of the features affecting churn rate.

Churn rate increases significantly after 4 or more calls to customer service as seen below.



More Insight:

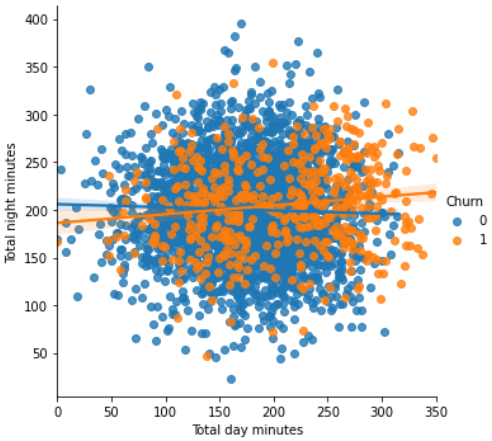
The first observation, on average customers that discontinue their contracts are more active users of communication services. Most churned users call a lot during the evening and during the day compared to loyal users (those who stick with us). Most churned users barely use our voice mail messages whereas most of our loyal users are huge fans of voice mail messages. See below ;



More Insight:

Here we try to analysis how churned users use the product in relation to minutes spent during the day and night. It appears that a portion of the users spent more minutes both during the day and night. A fair number of customers spend more time during the night.

Here we also try to understand the total day calls for churned users in relation to total evening calls. We realize that we have some outliers. Users who spent most of time calling in the evening than during the day. We also see that most churned users spent fairly the same amount time in the evening and during the day.



Conclusion

In conclusion, the goal was to discover how other features in the dataset affect the status of our users. By status is meant, whether a customer is churned or loyal. We discovered that on average the business lose about 14 % of its customers after using the product. This a huge number. We have 2850 loyal customers (those who continue to use our services) and 483 churned (those who have left us)

The churn rate increases significantly after 4 or more calls to customer service.

Customers that discontinue (churned) their contracts are more active users of communication services

On average the business lose about 14 % of its customers.

Disloyal customers tend to spend more time on the phone during both day and night.

Most churned users barely use our voice mail messages whereas most of our loyal users are huge fans of voice mail messages

When International Plan is enabled, the churn rate is much higher.