Problem Statement:

A Telecom company is very concerned about its customers discontinuing the service and opting to move to the one provided by the competitors. It is a major phenomenon is several service industries and is called 'Churn'. The company thinks that there are early indications available in the way a customer uses its service in predicting whether he/she is likely to churn.

The **file Churn.csv** contains some of the measures tracked by the company to see if it can predict the customers who are likely to churn and then take proactive action.

The dataset is about telecom industry which tells about the number of customers who churned the service. It consists of 3333 observations having 21 variables. We have to predict which customer is going to churn the service.

Account.Length: how long account has been active.

VMail.Message: Number of voicemail messages sent by the customer.

Day.Mins: Time spent on day calls.

Eve.Mins: Time spent on evening calls.

Night.Mins: Time spent on night calls.

Intl. Mins: Time spent on international calls.

Day.Calls: Number of day calls by customers.

Eve.Calls: Number of evening calls by customers.

Intl.Calls: Number of international calls.

Night.Calls: Number of night calls by the customer.

Day.Charge: Charges of Day Calls.

Night.Charge: Charges of Night Calls.

Eve.Charge: Charges of evening Calls.

Intl.Charge: Charges of international calls.

VMail.Plan: Voicemail plan taken by the customer or not.

State: State in Area of study.

Phone: Phone number of the customer.

Area.Code: Area Code of the customer.

Int.l.Plan: Does the customer have an international plan or not.

CustServ.Calls: Number of customer service calls by the customer.

Churn: Customers who churned the telecom service or who doesn't(0="Churner", 1=" Non-Churner")

Your task is to perform exploratory data analysis on this data and ascertain if the data still has the power to predict churn. If indeed such power exists, which variables have the capability to predict churn.

Follow these steps to explore it --

1. Create and Check Spark Context for Pyspark shell.
2. Load necessary libraries
3. Check the information provided about data.
4. Import the data files provided from HDFS (Churn.csv and Churntest.csv).
5. Display the data in Spark Dataframe. (Note:: In pyspark, dataframe index the rows from 0 instead of 1)
6. Do data pre-processing required? (Hint - We have some variables which should be of categorical datatype but they are of type integer. Convert them)
7. Do exploratory data analysis.
   7.1 - Describe the data using describe the function and state your insights.
   7.2 - Create Histogram for Day minutes spent by customers for churn=0 and 1 values.
   7.3 - Create count plots for Number of customers opt voicemail plan with Churn values.
   7.4 - Create count plots for International Plan opt by the customer with Churn values.
   7.5 - Plot Area Wise churner and non-churner.
   7.6 - Get correlation matrix using corr() function.
8. Get the correlation between Predicting Variable and independent variable and state your insights. (Now that we want to predict which customer is going to churn, let's see what columns might be interesting for our prediction. One way is to find the correlation between "Churn" and each of the other columns. This will show us which other columns might predict "Churn" the best.)
9. Applying Machine Learning Model
   9.1 - Import necessary libraries
   9.2 - Create vectors of all independent variables (Hint - use VectorAssembler)
   9.3 - Apply Decision Tree Classifier using dependent and independent variables.
   9.4 - Create a pipeline to build the classifier.
   9.5 - Use stratified sampling to get a sample of data.
   9.6 - Split the data into train and test dataset.
   9.7 - Make predictions and validate your model by calculating the accuracy score.
   9.8 - Calculate recall and precision score.
   9.9 - Test the model using test data and calculate accuracy, recall, and precision.
   9.10- Repeat steps from 9.3 to 9.9 for Random-forest and Gradient-Boost Classifiers.
10. State your insights and conclusions from the above analysis.