Problem 1

In this experiment, we will do exploratory data analysis to understand the data better. The dataset contains the record of telecom customers along with the label "churn". Churn = "true" signifies that the customer has left the company, and churn = "false" signifies that the customer is still loyal to the company.
Answer the following questions:

1. How many records are there in the dataset?
   3333

2. How many features are there? Name each feature and assign it as binary, discrete, or continuous.
   21 features

   1. state discrete
   2. account length discrete
   3. area code discrete
   4. phone number discrete
   5. international plan binary
   6. voice mail plan binary
   7. number vmail messages discrete
   8. total day minutes continuous
   9. total day calls discrete
   10. total day charge continuous
   11. total eve minutes continuous
   12. total eve calls discrete
   13. total eve charge continuous
   14. total night minutes continuous
   15. total night calls discrete
   16. total night charge continuous
   17. total intl minutes continuous
   18. total intl calls discrete
   19. total intl charge continuous
   20. customer service calls discrete
   21. churn binary

3. As a data scientist, your job is to build a model that identifies customers intending to leave your company. To do that, we prepare our data for the machine learning model. We can have the most advanced algorithm, but our results will be poor if our training data is terrible. According to your intuition, which features are irrelevant? Briefly explain your reasoning.
   I believe that ['state', 'account length', 'area code', 'phone number'] are irelevant features. Regarding all of these features are either random or have nothing to do with the company's service.

4. Are there any missing values in the data?
   No

5. What are the average, median, maximum, minimum, and standard deviation values for the continuous features?

	total day minutes	total day charge	total eve minutes	total eve charge	total night minutes	total night charge	total intl minutes	total intl charge
mean	179.78	30.56	200.98	17.08	200.87	9.04	10.24	2.76
std	54.47	9.26	50.71	4.31	50.57	2.28	2.79	0.75
min	0.00	0.00	0.00	0.00	23.20	1.04	0.00	0.00
median	179.40	30.50	201.40	17.12	201.20	9.05	10.30	2.78
max	350.80	59.64	363.70	30.91	395.00	17.77	20.00	5.40

6. What is the average number of customer service calls a customer makes to the company?
   1.56

7. In our dataset, data comes from how many states?
   51 states

8. What's the distribution of the "Churn" feature? Is this feature skewed?
   "Churn" is a binary attribute is used as a class variable, where one of the categories occurs 86% of the time, while the other categories together occur 14% of the time. It is not equally distributed. The feature is slightly skewed.

9. What are the customers' highest and lowest "total day charge"? If we sort the dataset in ascending and descending order by "total day charge", what observation can you make regarding the connection between "total day charge" and "churn" rate?
   The highest is 59.64, the lowest is 0. Top 10 "total day charge" custorms all have Ture churn value. Customers with lower "total day charge" tends to have churn = False.

10. What's the average number of customer service calls made by the user who has churned out of the company? Compare and contrast it with the average number of customer service calls made by the user who is still with the company.
    The average number of customer service calls made by the user who has churned out of the company is 2.23. The average number of customer service calls made by the user who is still with the company is 1.45

11. Compare and contrast the average values of numerical features for churned and non- churned users. As a data scientist, what strategy will you recommend to the company to retain more customers?
    I would recommend the company to adjust the price for long day call's, and improve the quality of customer service calls. By constrastin the average value of numerical features for churned and non-churned users. The total day charge and the number of customer service calls are where two class of users differ most.

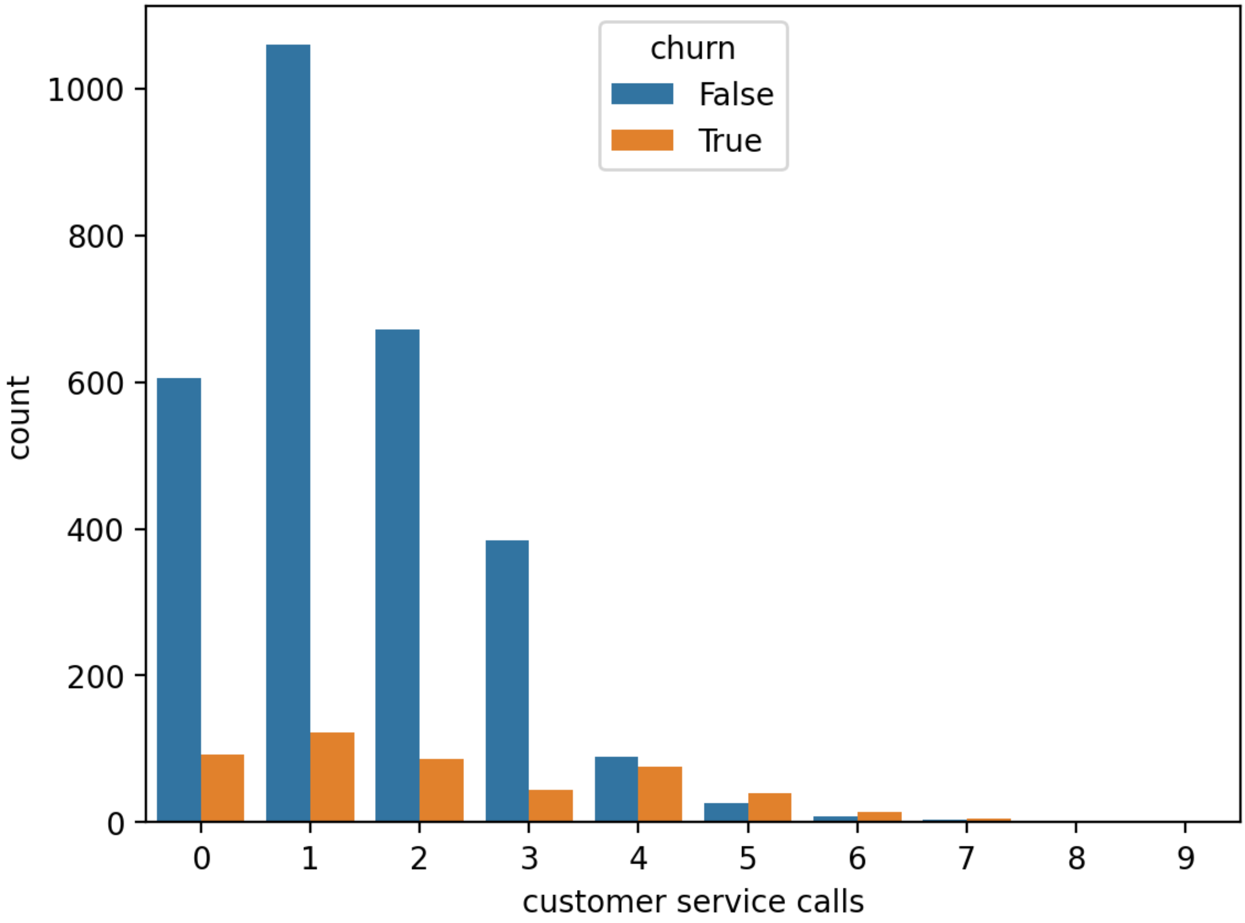
12. Assume you have devised a model which states that if "international plan" = 'no', then the customer will not churn (i.e., "churn" = False). Report accuracy, precision, and recall concerning the "churned" class.
    precision= 0.8850498338870432
    recall= 0.9347368421052632
    accuracy= 0.8403840384038403

13. Calculate P (churn = True | international plan = ‘yes’), P (churn = False | international plan = ‘yes’), P (churn= True | international plan = ‘no’), P (churn = False | international plan = ‘no’). For a customer who has churned, what are the probabilities that the customer has opted/not opted for the international plan? Similarly, given that the customer has not churned, what are the probabilities that the customer has opted/not opted for the international plan?
    P (churn = True | international plan = ‘yes’) = 0.4241486068111455
    P (churn = False | international plan = ‘yes’) = 0.5758513931888545
    P (churn= True | international plan = ‘no’) = 0.11495016611295682
    P (churn = False | international plan = ‘no’) = 0.8850498338870432

    p (international plan = ‘yes’ |churn = True) = 0.2836438923395445
    p (international plan = ‘no’ |churn = True) = 0.7163561076604554
    p (international plan = ‘yes’ |churn = False) = 0.06526315789473684
    p (international plan = ‘no’ |churn = False) = 0.9347368421052632

14. Calculate the probability of customers leaving the company, given that he has not made any customer service call. Compare and contrast it with the customer making 1,2,3,4,5,6,7,8,9 customer service calls. Plot the probability of customers leaving the company as customer service calls increase.
    p(churn=True|customer service call=0) = 0.13199426111908177
    p(churn=True|customer service call=1) = 0.10330228619813717
    p(churn=True|customer service call=2) = 0.11462450592885376
    p(churn=True|customer service call=3) = 0.10256410256410256
    p(churn=True|customer service call=4) = 0.4578313253012048
    p(churn=True|customer service call=5) = 0.6060606060606061
    p(churn=True|customer service call=6 = 0.6363636363636364
    p(churn=True|customer service call=7= 0.5555555555555556
    p(churn=True|customer service call=8= 0.5

p(churn=True|customer service call=9 = 1.0



15. Assume you have devised a model which states that if "international plan" = 'yes' and the number of calls to the service center is greater than 3, then the customer will churn (i.e., "churn" = True). Report accuracy, precision, and recall concerning the "churned" class.
    precision = 0.6785714285714286
    recall = 0.039337474120082816
    accuracy= 0.858085808580858