Question 1:
We first plot the relation between the target variables and each of the predictor variables (Univariate Analysis).
A few points to be noted:

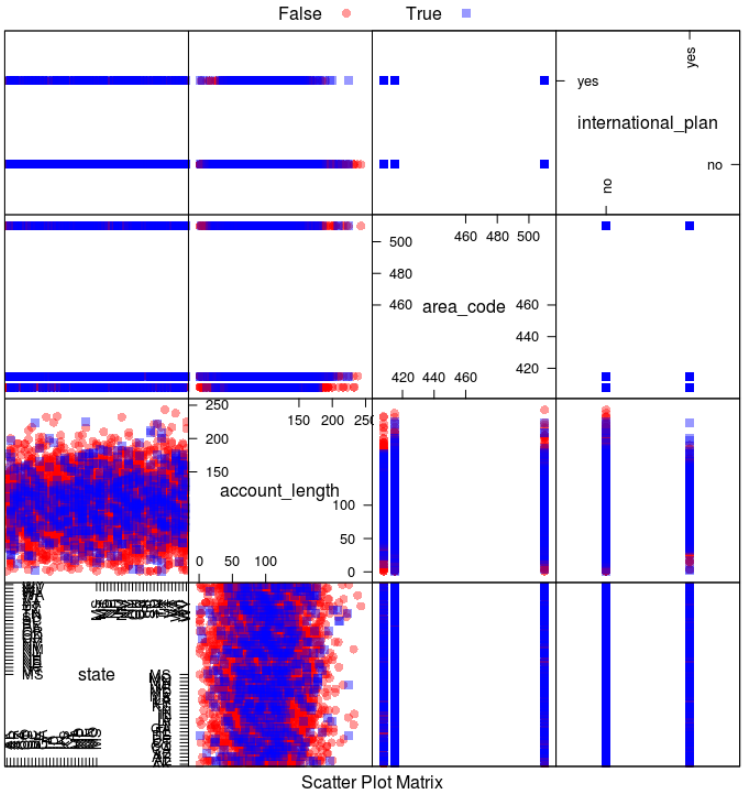
1. We discard the variable “Phone Number” because this data has cannot have any relation with target variable whatsoever.
2. No missing data or NaN data found.

A preliminary analysis found the following things:

1. “Account length”, “State”, “Area code”, “total day calls”, “total eve calls”, “total night minutes”, “total night charges”, “total intl minutes”, “total int minutes”, “total intl charge” has no effect.
2. “International plan”, “Voice mail plan”, “Number_vmail_messages”, “Total day minutes”, “total day charge”, “total eve minutes”, “total night call”, “total intl calls”, “number customer service calls”, might have an effect on the target variable.

This analysis was done, using the same method as discussed in the class. As an example, I shall explain, how I classified “Account Length” as insignificant.

We calculate the means of “Account Length” of the dataset with churn as True and with False. Now, the difference of means, is significant if the p-value of the dfference, in the t-statistics, is greater than 0.95, otherwise, it is a case of mere chance that the difference between the lengths arose. I did this for all the variables which I classified as insignificant. However, I didn’t set an hard-bound of 0.95. I included those as significant, even if they had p-values in a high range but not 0.95. Because, in case of regression, if those values were not actually related, their coefficients would be less, but not including a value will have greater errors, in terms of loss of data. This was also visually verified by drawing scatter plots.



Question 2,3,4:
These have been documented in the form of codes. Please see the corresponding R file.