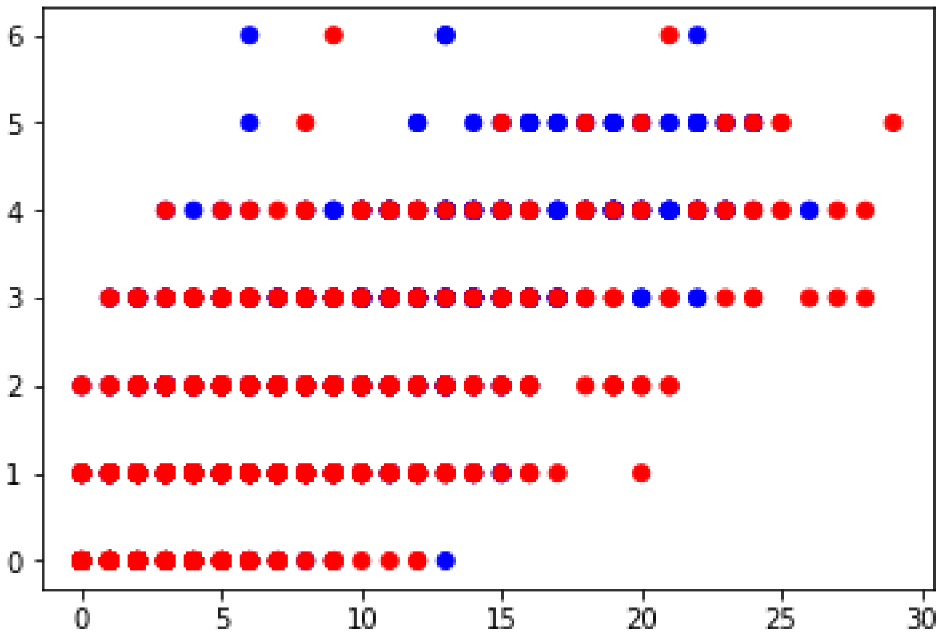
Question 1. A

For the logistic regression function, we set the value of C which is equal to 1/(lambda). So, for this example, I took five values of C = [0.01, 0.1, 1, 10, 100]. This values correspond to the following values of lambda = [100, 10, 1, 0.1, 0.01]. Then comparing the 5-fold cross validation errors, the best values of lambda were selected for every preprocessing types discussed. This was based on the value of lambda, which gave the least error rate. Thus, the error rates obtained were as follows:

Method	Best Value of lambda	Cross Validation train error rate	Cross Validation test error rate	Train Error Rate	Test Error Rate
Standardization	0.01	0.06435563	0.07862969	0.0701468189233	0.0865885416667
Log Transform	0.1	0.05089723	0.05448613	0.051223491	0.05859375
Binary Transform	1	0.06362153	0.06688418	0.063295269	0.072265625

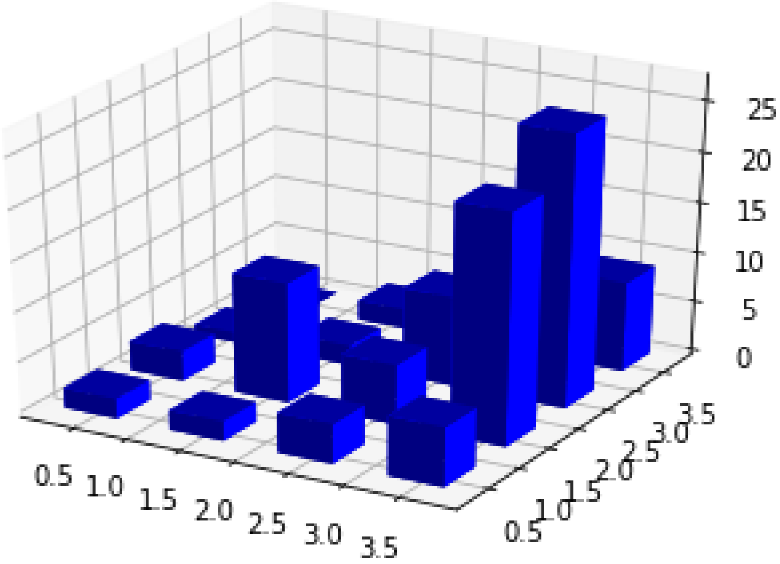
Question 1. B

(i) Scatter Plot of all testing points:



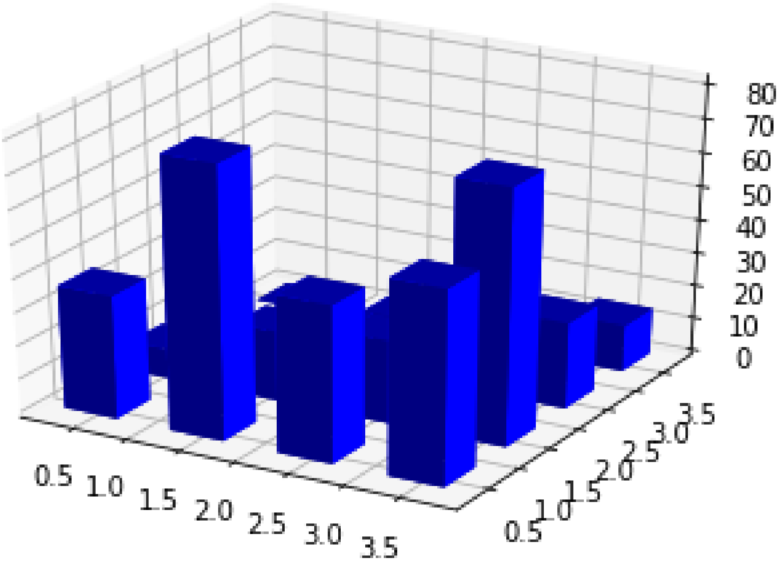
Here, Blue points denote spam mail and Red denotes non-spam. The data points are overwritten in the scatter plot and hence cannot see all the points.

(ii) 3D histogram for emails labeled spam:



Given test data, we used sum of features 1-48 (total count of keywords in percentage) as x axis, and sum of features 49-54 (total count of special characters in percentage) as y axis.

(iii) 3D histogram for emails labeled non-spam:



Given test data, we used sum of features 1-48 (total count of keywords in percentage) as x axis, and sum of features 49-54 (total count of special characters in percentage) as y axis.

(iv) Yes, there’s significant difference between the two histograms. As we can see, for spam emails, the bars are concentrated near the zero value. In the case of non-spam mail, it’s the opposite with the bars being concentrated in the opposite direction and more spread out. Also, the bars in the histogram have a higher z axis value for non-spam case.