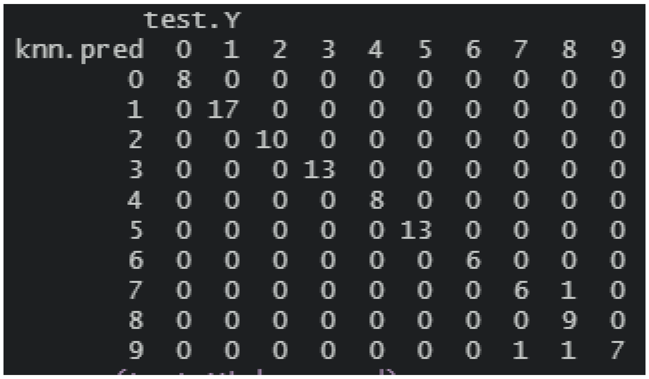
Neha Maddali

Problem 1:

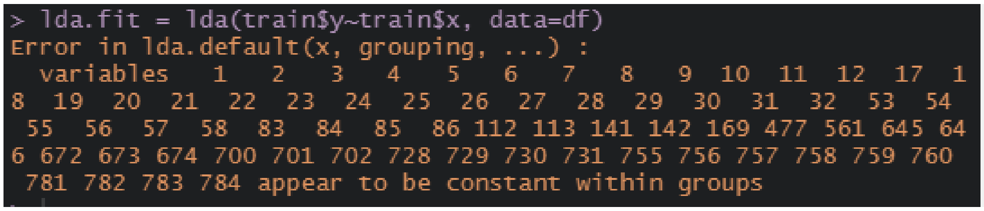
a) Optimal K = 5



```
         test.Y
knn.pred  0  1  2  3  4  5  6  7  8  9
       0  8  0  0  0  0  0  0  0  0  0
       1  0 17  0  0  0  0  0  0  0  0
       2  0  0 10  0  0  0  0  0  0  0
       3  0  0  0 13  0  0  0  0  0  0
       4  0  0  0  0  8  0  0  0  0  0
       5  0  0  0  0  0 13  0  0  0  0
       6  0  0  0  0  0  0  6  0  0  0
       7  0  0  0  0  0  0  0  6  1  0
       8  0  0  0  0  0  0  0  0  9  0
       9  0  0  0  0  0  0  0  1  1  7
```

Misclassifcation rate: 0.03

b)

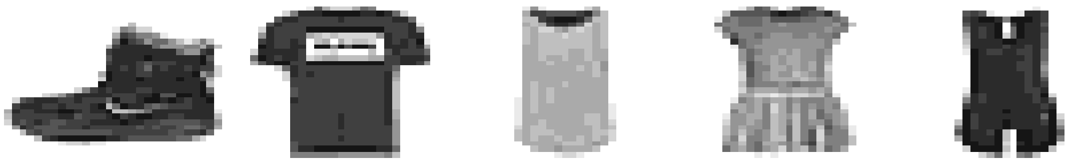


```
> lda.fit = lda(train$y~train$x, data=df)
Error in lda.default(x, grouping, ...) :
  variables   1   2   3   4   5   6   7   8   9  10  11  12  17  1
8  19  20  21  22  23  24  25  26  27  28  29  30  31  32  53  54
 55  56  57  58  83  84  85  86 112 113 141 142 169 477 561 645 64
6 672 673 674 700 701 702 728 729 730 731 755 756 757 758 759 760
 781 782 783 784 appear to be constant within groups
```

The error means that the variance for these predictors are either zero or close to zero

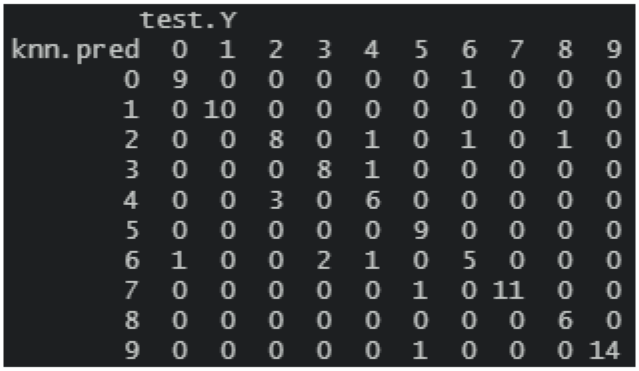
c) KNN doesn’t make any assumptions about the data since it non-parametric. KNN doesn’t cause an error when there is not much variation in the variables while LDA does. LDA only works with continuous and categorical variables.

Problem 2:



a)

b) Optimal K = 5



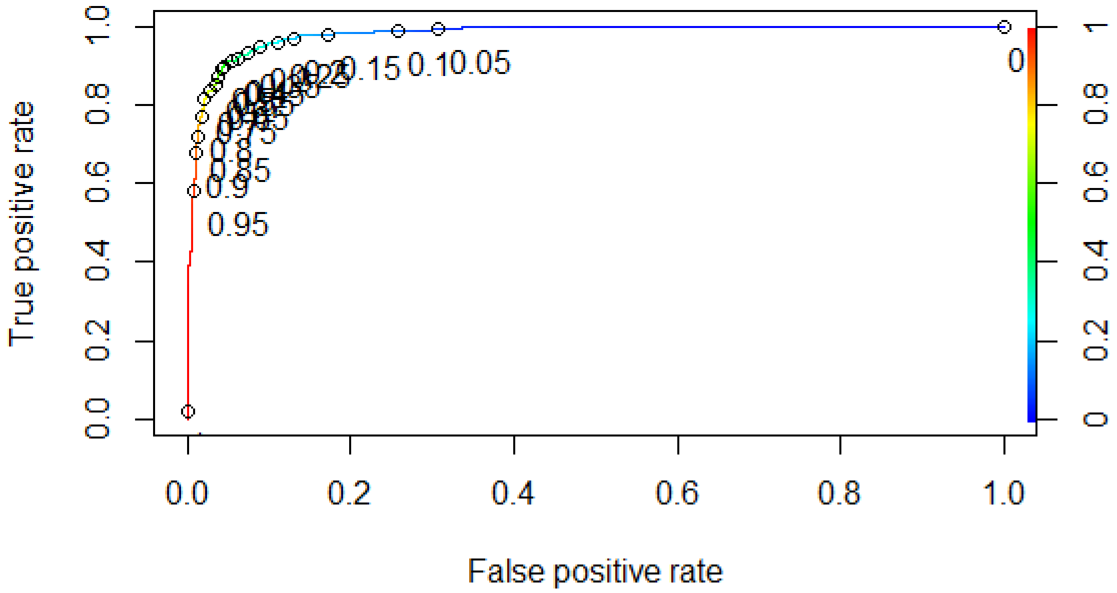
```
         test.Y
knn.pred  0  1  2  3  4  5  6  7  8  9
       0  9  0  0  0  0  0  1  0  0  0
       1  0 10  0  0  0  0  0  0  0  0
       2  0  0  8  0  1  0  1  0  1  0
       3  0  0  0  8  1  0  0  0  0  0
       4  0  0  3  0  6  0  0  0  0  0
       5  0  0  0  0  0  9  0  0  0  0
       6  1  0  0  2  1  0  5  0  0  0
       7  0  0  0  0  0  1  0 11  0  0
       8  0  0  0  0  0  0  0  0  6  0
       9  0  0  0  0  0  1  0  0  0 14
```

Misclassification rate: 0.14

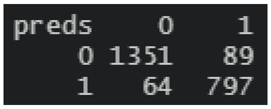
Compared to 1a, this misclassification is larger so the model in 1a has better results.

Problem 3:

a. False positives seem highly problematic. I do not want a potentially important email to be marked as spam. Therefore, I can tune the threshold for logistic regression that my spam filter is more conservative and makes it harder to mark emails as spam.



b.



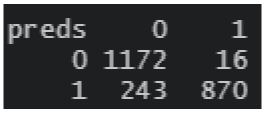
```
preds    0    1
    0 1351   89
    1   64  797
```

c.

False positive rate: 0.06180556

False negative rate: 0.07433217

d. Threshold of 0.15



```
preds    0    1
    0 1172   16
    1  243  870
```

False positive rate: 0.01346801

False negative rate: 0.2183288