Assignment 2: Feature Engineering, Statistical Analysis and Machine Learning

Eunan Diamond (40293751)

# Introduction

## How to Use report

-Run each chunk in order every time you use this rmd file ## goals The goals for this assignment are to perform feature engineering, statistical analysis, Regression and Machine Learning on a data set of symbols of letters and non letters. Using r, I should develop code, and use my report to clearly show that I understand all the methods I used in each section to explain the results I get from my analysis.

# Section 1

I created the images using GIMP. After that I used read.delim to read them into R code, and then used then used R code to change the values with 0 (black pixels) to 1 and values with 255 (white pixels) to 0. I then created a matrix with the table we imported in, only from line 3 to last line as first 2 lines were headers, and the matrix has 18 rows and columns. I exported the matrix using write.table. I used separator as “,” and saved the file as .csv. I did this for all 140 of my images. # Section 2

source("./section2\_code.r")

## Loading required package: sp

## Loading required namespace: igraph

For Section 2, I used one function to create all the features.Using one big function was easier as it made it easier to create the features vector in there and adding it to the matrix which would then be converted to csv file at the end. I converted the csv files to matrices so it would be easier to use for loops and if statements to attain the features I needed. I struggled with the eyes function, I worked around this by converting the 1’s into 0’s and 0’s into ones, this lead the areas of connected 0’s between 1’s using the 4 directions. I then took 2 away from the number of columns in the freq table, as 1 is for none and the other is for connected 0 (the converted 1’s). for my custom feature, I went for number with row with most black pixels as I thought they would help distinguish between symbols as faces would have more than a’s etc. Based on features data this has proved to be the case with the sad and smiley having big largest row of black pixel’s. And the ones expected to have less such as ! and a have a lot smaller, I believe this will help me distinguish between symbols in section 3.

csvF<-read.csv(file ="40293751\_features.csv" ,header = TRUE)  
print(csvF)

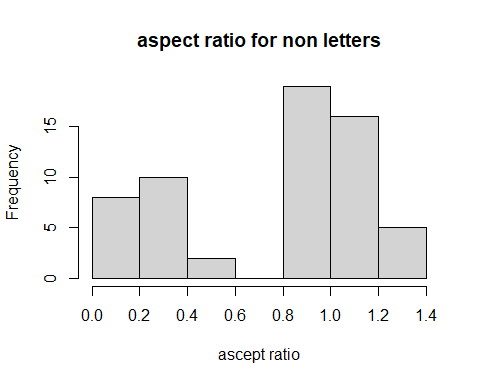
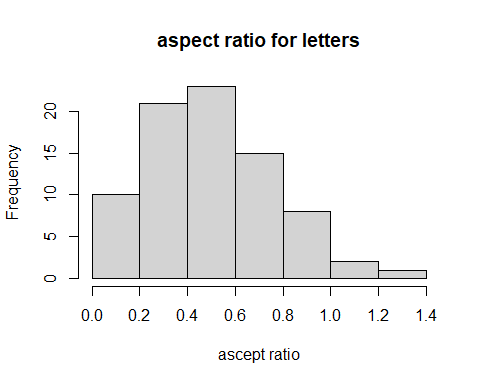
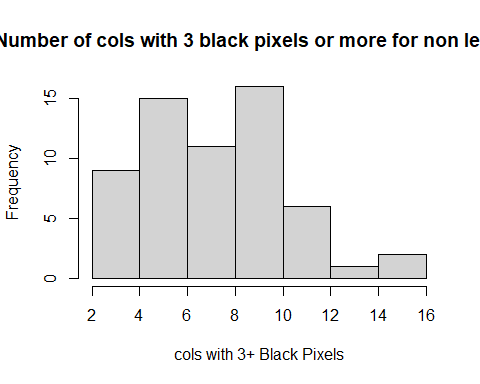
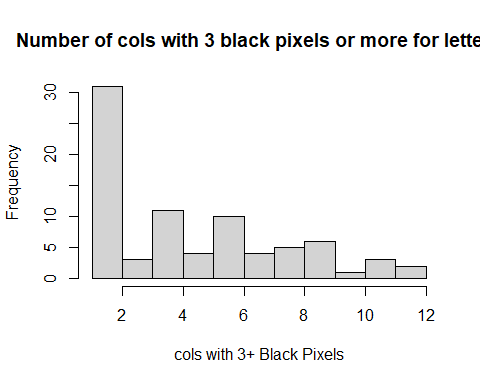
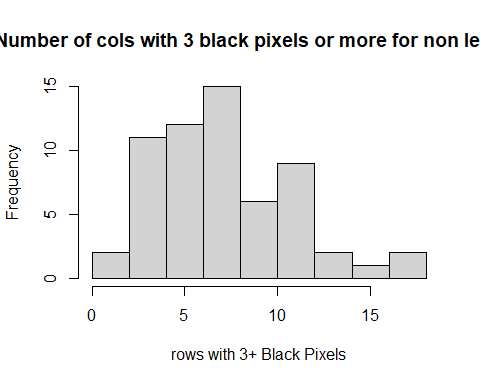
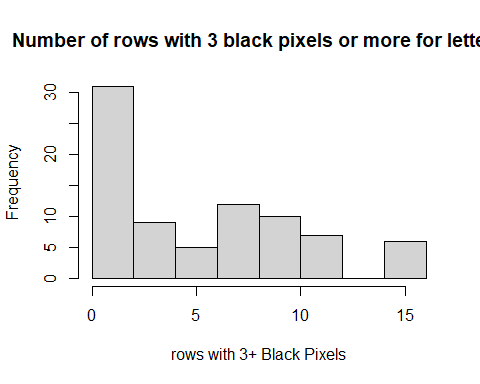
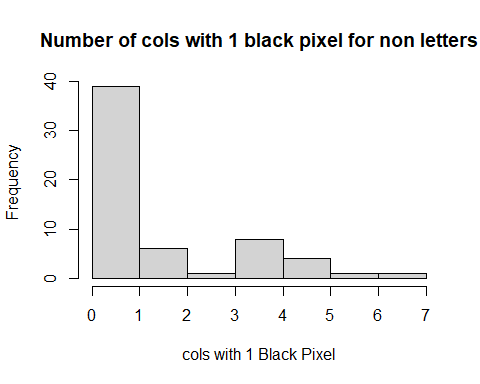
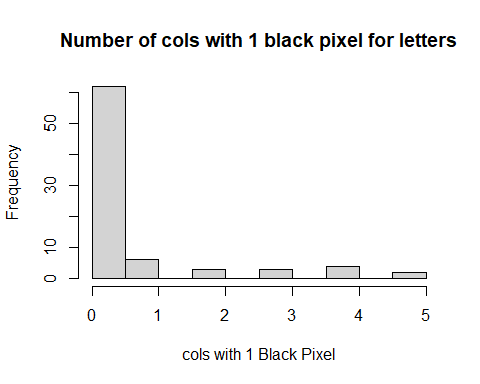
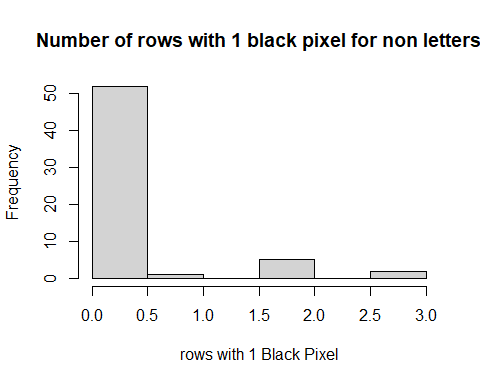
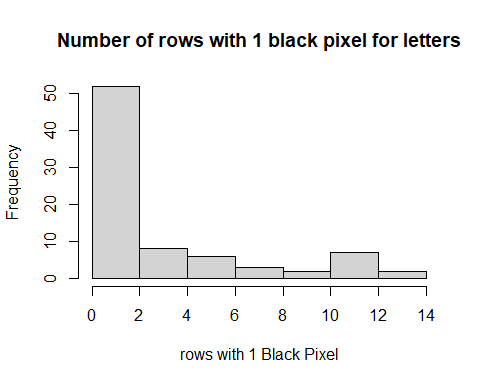
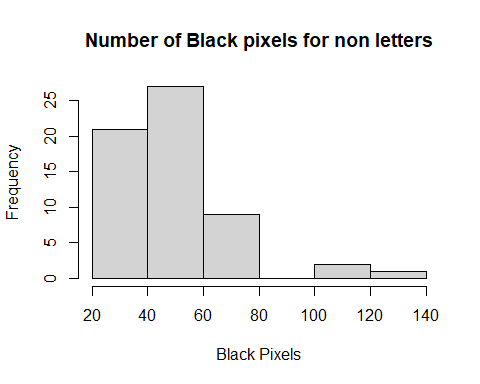
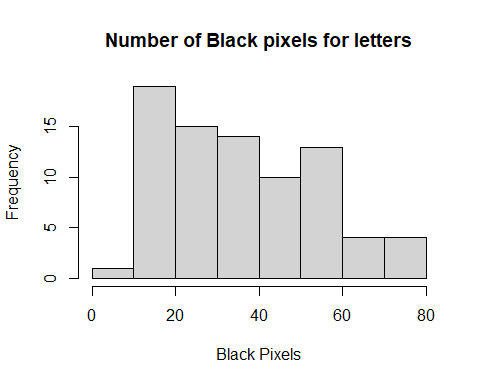
## label index nr\_pix rows\_with\_1 cols\_with\_1 rows\_with\_3p cols\_with\_3p  
## 1 a 1 35 1 0 8 7  
## 2 a 2 33 2 0 8 6  
## 3 a 3 53 0 0 10 7  
## 4 a 4 31 1 0 8 6  
## 5 a 5 62 0 0 11 8  
## 6 a 6 76 0 0 12 9  
## 7 a 7 80 0 0 12 9  
## 8 a 8 62 0 0 10 8  
## 9 b 1 25 7 0 2 2  
## 10 b 2 36 6 0 6 4  
## 11 b 3 50 0 0 8 6  
## 12 b 4 69 0 0 12 8  
## 13 b 5 23 5 0 0 2  
## 14 b 6 46 0 0 10 6  
## 15 b 7 14 4 1 0 1  
## 16 b 8 27 0 0 5 4  
## 17 c 1 20 4 0 2 1  
## 18 c 2 52 0 0 8 9  
## 19 c 3 40 0 0 8 7  
## 20 c 4 55 0 0 11 11  
## 21 c 5 59 0 0 11 11  
## 22 c 6 47 3 0 8 9  
## 23 c 7 54 0 0 8 9  
## 24 c 8 14 6 0 2 1  
## 25 d 1 22 6 0 2 2  
## 26 d 2 46 0 0 9 6  
## 27 d 3 36 5 0 9 5  
## 28 d 4 53 0 0 12 7  
## 29 d 5 18 4 1 0 1  
## 30 d 6 24 4 0 6 4  
## 31 d 7 42 0 0 8 5  
## 32 d 8 52 0 0 8 6  
## 33 e 1 20 1 0 2 6  
## 34 e 2 33 1 0 4 9  
## 35 e 3 71 0 0 10 12  
## 36 e 4 74 0 0 10 12  
## 37 e 5 24 2 0 2 8  
## 38 e 6 33 1 0 6 8  
## 39 e 7 35 1 0 4 10  
## 40 e 8 61 1 0 8 11  
## 41 f 1 19 11 4 2 1  
## 42 f 2 39 0 0 4 4  
## 43 f 3 44 0 0 4 4  
## 44 f 4 20 11 4 1 1  
## 45 f 5 51 0 0 15 4  
## 46 f 6 55 0 0 15 5  
## 47 f 7 60 0 0 16 6  
## 48 f 8 16 11 3 1 1  
## 49 g 1 19 11 4 2 1  
## 50 g 2 39 0 0 4 4  
## 51 g 3 44 0 0 4 4  
## 52 g 4 20 11 4 1 1  
## 53 g 5 51 0 0 15 4  
## 54 g 6 55 0 0 15 5  
## 55 g 7 60 0 0 16 6  
## 56 g 8 16 11 3 1 1  
## 57 h 1 18 4 3 2 2  
## 58 h 2 37 0 2 6 4  
## 59 h 3 43 0 0 7 6  
## 60 h 4 22 5 5 2 2  
## 61 h 5 43 0 0 9 3  
## 62 h 6 49 0 0 9 4  
## 63 h 7 25 3 5 3 2  
## 64 h 8 31 3 0 4 2  
## 65 i 1 12 12 0 0 1  
## 66 i 2 24 0 0 0 2  
## 67 i 3 14 14 0 0 1  
## 68 i 4 28 0 0 0 2  
## 69 i 5 9 9 0 0 1  
## 70 i 6 20 0 0 0 2  
## 71 i 7 24 0 0 0 2  
## 72 i 8 13 13 0 0 1  
## 73 j 1 12 8 1 0 1  
## 74 j 2 22 0 1 2 2  
## 75 j 3 28 0 0 3 3  
## 76 j 4 13 8 2 1 1  
## 77 j 5 24 0 1 2 2  
## 78 j 6 32 0 1 10 3  
## 79 j 7 11 9 2 0 1  
## 80 j 8 23 0 0 1 2  
## 81 sad 1 32 0 5 5 9  
## 82 sad 2 50 0 2 7 10  
## 83 sad 3 78 0 0 7 16  
## 84 sad 4 28 0 5 3 5  
## 85 sad 5 54 0 0 8 8  
## 86 sad 6 48 0 4 5 12  
## 87 sad 7 40 0 2 6 10  
## 88 sad 8 52 0 4 7 12  
## 89 sad 9 42 0 2 8 6  
## 90 sad 10 56 0 3 8 8  
## 91 sad 11 60 0 1 10 9  
## 92 sad 12 30 0 6 4 6  
## 93 sad 13 42 0 1 5 8  
## 94 sad 14 46 0 4 5 10  
## 95 sad 15 24 2 7 3 5  
## 96 sad 16 38 0 4 6 9  
## 97 sad 17 40 0 4 6 10  
## 98 sad 18 64 0 0 7 12  
## 99 sad 19 38 0 2 6 10  
## 100 sad 20 72 0 0 10 11  
## 101 smiley 1 64 0 0 7 12  
## 102 smiley 2 41 0 1 5 10  
## 103 smiley 3 54 0 0 7 10  
## 104 smiley 4 30 0 4 3 6  
## 105 smiley 5 38 0 4 4 8  
## 106 smiley 6 54 0 0 8 8  
## 107 smiley 7 60 0 2 6 10  
## 108 smiley 8 34 0 2 6 6  
## 109 smiley 9 26 2 5 3 5  
## 110 smiley 10 26 2 5 3 5  
## 111 smiley 11 44 0 4 4 10  
## 112 smiley 12 38 2 0 4 7  
## 113 smiley 13 40 0 0 4 8  
## 114 smiley 14 42 0 0 4 8  
## 115 smiley 15 60 0 0 5 14  
## 116 smiley 16 50 0 0 8 10  
## 117 smiley 17 80 0 0 8 12  
## 118 smiley 18 56 0 0 7 10  
## 119 smiley 19 78 0 0 7 16  
## 120 smiley 20 52 0 0 7 10  
## 121 xclaim 1 77 0 0 11 7  
## 122 xclaim 2 55 0 0 11 5  
## 123 xclaim 3 65 0 0 13 5  
## 124 xclaim 4 112 0 0 16 7  
## 125 xclaim 5 119 0 0 17 7  
## 126 xclaim 6 28 0 0 0 2  
## 127 xclaim 7 34 0 0 0 2  
## 128 xclaim 8 45 0 0 9 5  
## 129 xclaim 9 50 0 0 10 5  
## 130 xclaim 10 55 0 0 11 5  
## 131 xclaim 11 54 0 0 12 5  
## 132 xclaim 12 62 0 0 14 5  
## 133 xclaim 13 48 0 0 10 4  
## 134 xclaim 14 56 0 0 12 4  
## 135 xclaim 15 58 0 0 12 4  
## 136 xclaim 16 133 0 0 17 9  
## 137 xclaim 17 31 1 0 10 3  
## 138 xclaim 18 35 2 0 11 3  
## 139 xclaim 19 36 3 0 11 3  
## 140 xclaim 20 39 3 0 12 3  
## aspect\_ratio neigh\_1 no\_neigh\_above no\_neigh\_below no\_neigh\_left  
## 1 0.60000000 1 6 7 7  
## 2 0.50000000 0 6 7 7  
## 3 0.63636364 0 8 9 8  
## 4 0.55555556 0 7 7 6  
## 5 0.63636364 0 9 10 7  
## 6 0.72727273 0 10 11 8  
## 7 0.72727273 0 9 12 7  
## 8 0.63636364 0 9 10 8  
## 9 0.28571429 2 1 2 16  
## 10 0.31250000 1 3 4 19  
## 11 0.42857143 0 2 4 16  
## 12 0.43750000 0 4 6 19  
## 13 0.30769231 1 1 1 15  
## 14 0.35714286 0 0 3 15  
## 15 0.37500000 1 1 1 9  
## 16 0.50000000 0 3 3 9  
## 17 0.77777778 2 7 7 6  
## 18 0.81818182 0 10 10 6  
## 19 1.00000000 0 9 9 5  
## 20 1.00000000 0 12 12 7  
## 21 1.00000000 0 14 14 9  
## 22 0.80000000 0 14 14 13  
## 23 0.80000000 0 13 13 13  
## 24 0.71428571 2 7 7 4  
## 25 0.33333333 1 2 2 10  
## 26 0.38461538 0 3 3 11  
## 27 0.33333333 1 1 2 14  
## 28 0.40000000 0 1 3 14  
## 29 0.40000000 1 1 1 6  
## 30 0.40000000 1 2 2 8  
## 31 0.38461538 0 3 3 12  
## 32 0.38461538 0 5 5 14  
## 33 0.71428571 1 5 7 4  
## 34 1.25000000 1 16 18 4  
## 35 1.20000000 0 16 20 7  
## 36 1.20000000 0 17 21 7  
## 37 0.87500000 1 9 10 3  
## 38 0.87500000 1 11 11 5  
## 39 1.00000000 1 14 16 4  
## 40 1.00000000 0 16 18 5  
## 41 0.50000000 4 5 5 9  
## 42 0.58333333 0 6 7 9  
## 43 0.53846154 0 8 7 12  
## 44 0.46153846 4 4 4 10  
## 45 0.42857143 0 3 4 11  
## 46 0.50000000 0 5 6 11  
## 47 0.46666667 1 5 6 13  
## 48 0.33333333 2 1 2 9  
## 49 0.50000000 4 5 5 9  
## 50 0.58333333 0 6 7 9  
## 51 0.53846154 0 8 7 12  
## 52 0.46153846 4 4 4 10  
## 53 0.42857143 0 3 4 11  
## 54 0.50000000 0 5 6 11  
## 55 0.46666667 1 5 6 13  
## 56 0.33333333 2 1 2 9  
## 57 0.44444444 3 3 2 14  
## 58 0.77777778 0 6 4 14  
## 59 0.77777778 0 6 4 14  
## 60 0.54545455 3 4 3 16  
## 61 0.58333333 1 5 4 18  
## 62 0.66666667 0 5 5 18  
## 63 0.63636364 3 2 3 17  
## 64 0.63636364 3 4 3 17  
## 65 0.08333333 2 2 2 12  
## 66 0.08333333 2 4 4 12  
## 67 0.07142857 2 2 2 14  
## 68 0.07142857 2 4 4 14  
## 69 0.10000000 2 2 2 9  
## 70 0.09090909 2 4 4 10  
## 71 0.08333333 0 4 4 12  
## 72 0.07692308 2 2 2 13  
## 73 0.20000000 1 2 3 8  
## 74 0.30000000 2 4 5 8  
## 75 0.25000000 0 4 5 10  
## 76 0.30000000 1 3 4 8  
## 77 0.40000000 2 5 6 8  
## 78 0.40000000 2 6 7 8  
## 79 0.18181818 2 3 3 9  
## 80 0.25000000 2 5 4 10  
## 81 1.30000000 2 21 19 10  
## 82 1.08333333 2 20 22 9  
## 83 1.15384615 0 22 20 10  
## 84 0.90909091 2 16 14 12  
## 85 0.84615385 2 16 18 14  
## 86 1.15384615 2 22 20 10  
## 87 1.00000000 0 22 20 10  
## 88 1.15384615 0 20 16 9  
## 89 0.84615385 2 14 16 13  
## 90 1.00000000 2 18 20 13  
## 91 0.92857143 0 19 19 14  
## 92 1.00000000 0 16 16 11  
## 93 1.00000000 0 18 18 10  
## 94 1.08333333 2 20 18 10  
## 95 1.10000000 4 13 13 9  
## 96 1.09090909 2 17 17 11  
## 97 1.30000000 2 20 20 9  
## 98 1.25000000 2 26 24 15  
## 99 1.00000000 0 20 18 10  
## 100 0.92857143 0 21 21 14  
## 101 0.91666667 2 18 16 12  
## 102 1.00000000 2 19 17 8  
## 103 1.08333333 0 18 16 7  
## 104 1.10000000 2 16 18 10  
## 105 1.00000000 2 18 20 12  
## 106 0.91666667 2 18 16 14  
## 107 1.18181818 2 22 20 11  
## 108 1.00000000 2 16 14 9  
## 109 1.00000000 4 13 11 9  
## 110 1.00000000 4 13 11 9  
## 111 1.08333333 2 20 22 15  
## 112 1.10000000 2 17 17 8  
## 113 1.10000000 0 18 18 8  
## 114 1.10000000 0 18 20 9  
## 115 1.30000000 0 20 22 10  
## 116 1.09090909 0 20 20 10  
## 117 0.91666667 0 18 22 16  
## 118 1.08333333 0 18 18 8  
## 119 1.25000000 0 20 22 10  
## 120 1.00000000 0 20 22 11  
## 121 0.50000000 0 14 14 11  
## 122 0.33333333 0 10 10 11  
## 123 0.28571429 0 10 10 13  
## 124 0.37500000 0 14 7 15  
## 125 0.35294118 0 7 7 15  
## 126 0.07142857 0 4 4 14  
## 127 0.05882353 0 2 2 15  
## 128 0.40000000 0 10 10 9  
## 129 0.36363636 0 10 10 10  
## 130 0.36363636 0 10 10 11  
## 131 0.33333333 0 8 8 12  
## 132 0.28571429 0 8 8 14  
## 133 0.21428571 0 6 6 14  
## 134 0.18750000 0 2 6 15  
## 135 0.17647059 0 2 4 15  
## 136 0.47058824 0 5 9 15  
## 137 0.16666667 0 4 4 11  
## 138 0.14285714 2 4 4 13  
## 139 0.14285714 2 4 4 14  
## 140 0.13333333 2 1 4 14  
## no\_neigh\_right no\_neigh\_horiz no\_neigh\_vert connected\_areas eyes custom  
## 1 10 10 7 1 1 7  
## 2 9 8 5 1 1 6  
## 3 10 1 1 1 1 7  
## 4 10 8 8 1 1 6  
## 5 12 0 2 1 1 8  
## 6 11 0 1 1 1 9  
## 7 11 0 0 1 1 9  
## 8 11 0 1 1 1 8  
## 9 12 21 6 1 1 3  
## 10 17 20 4 1 1 4  
## 11 10 0 2 1 1 5  
## 12 17 0 1 1 1 7  
## 13 11 19 6 1 1 2  
## 14 12 3 1 1 1 4  
## 15 4 10 5 1 1 2  
## 16 6 1 2 1 1 4  
## 17 6 12 14 1 0 4  
## 18 6 0 2 1 0 8  
## 19 7 2 8 1 0 6  
## 20 9 0 0 1 0 8  
## 21 11 0 0 1 0 8  
## 22 11 7 0 1 0 8  
## 23 11 0 0 1 0 8  
## 24 4 6 10 1 0 4  
## 25 14 16 6 1 1 3  
## 26 15 0 0 1 1 4  
## 27 18 16 2 1 1 3  
## 28 18 2 1 1 1 4  
## 29 11 14 7 1 1 2  
## 30 11 10 2 1 1 3  
## 31 14 2 2 1 1 4  
## 32 16 0 0 1 1 6  
## 33 3 6 12 1 1 6  
## 34 3 8 25 1 1 11  
## 35 4 0 0 1 1 13  
## 36 4 0 0 1 1 13  
## 37 3 10 19 1 1 8  
## 38 4 4 13 1 1 8  
## 39 3 10 25 1 1 11  
## 40 4 1 1 1 1 11  
## 41 9 11 8 1 0 5  
## 42 10 0 0 1 0 6  
## 43 11 0 0 1 0 6  
## 44 10 13 8 1 0 5  
## 45 12 1 0 1 0 5  
## 46 12 1 0 1 0 7  
## 47 14 2 0 1 0 7  
## 48 9 13 5 1 0 3  
## 49 9 11 8 1 0 5  
## 50 10 0 0 1 0 6  
## 51 11 0 0 1 0 6  
## 52 10 13 8 1 0 5  
## 53 12 1 0 1 0 5  
## 54 12 1 0 1 0 7  
## 55 14 2 0 1 0 7  
## 56 9 13 5 1 0 3  
## 57 12 14 3 1 0 3  
## 58 11 0 2 1 0 6  
## 59 11 0 0 1 0 8  
## 60 14 17 5 1 0 4  
## 61 16 6 0 1 0 7  
## 62 15 0 0 1 0 7  
## 63 15 20 5 1 0 4  
## 64 15 16 0 1 0 6  
## 65 12 12 1 2 0 1  
## 66 12 0 2 2 0 2  
## 67 14 14 1 2 0 1  
## 68 14 0 2 2 0 2  
## 69 9 9 1 2 0 1  
## 70 10 0 2 2 0 2  
## 71 12 0 0 2 0 2  
## 72 13 13 1 2 0 1  
## 73 9 8 2 2 0 2  
## 74 9 0 3 2 0 3  
## 75 11 0 0 2 0 4  
## 76 9 8 3 2 0 3  
## 77 9 0 3 2 0 4  
## 78 9 0 4 2 0 4  
## 79 9 9 3 2 0 2  
## 80 9 0 2 2 0 3  
## 81 10 4 12 4 0 14  
## 82 9 2 4 4 0 12  
## 83 10 2 0 4 0 16  
## 84 12 6 8 4 0 10  
## 85 14 0 6 4 0 8  
## 86 10 4 14 4 0 16  
## 87 10 2 10 4 0 12  
## 88 9 2 10 4 0 12  
## 89 13 0 6 4 0 8  
## 90 13 0 6 4 0 10  
## 91 14 2 4 4 0 10  
## 92 11 4 10 4 0 10  
## 93 10 2 2 4 0 10  
## 94 10 4 12 4 0 14  
## 95 9 8 10 4 0 8  
## 96 11 6 10 4 0 8  
## 97 9 4 14 4 0 10  
## 98 15 6 0 4 0 16  
## 99 10 2 8 4 0 10  
## 100 14 2 2 4 0 10  
## 101 12 4 2 4 0 10  
## 102 8 0 4 4 0 11  
## 103 7 0 2 4 0 12  
## 104 10 4 8 4 0 10  
## 105 12 4 10 4 0 12  
## 106 14 0 6 4 0 8  
## 107 11 2 4 4 0 12  
## 108 9 0 6 4 0 8  
## 109 9 8 8 4 0 8  
## 110 9 8 8 4 0 8  
## 111 15 6 12 4 0 14  
## 112 8 4 0 4 0 12  
## 113 8 2 0 4 0 12  
## 114 9 2 0 4 0 12  
## 115 10 2 0 4 0 14  
## 116 10 0 2 4 0 10  
## 117 16 2 0 4 0 12  
## 118 8 0 2 4 0 12  
## 119 10 2 0 4 0 16  
## 120 11 2 0 4 0 12  
## 121 11 0 0 2 0 7  
## 122 11 0 0 2 0 5  
## 123 13 0 0 2 0 5  
## 124 15 0 0 2 0 7  
## 125 15 0 0 2 0 7  
## 126 14 0 0 2 0 2  
## 127 15 0 0 2 0 2  
## 128 9 0 0 2 0 5  
## 129 10 0 0 2 0 5  
## 130 11 0 0 2 0 5  
## 131 12 0 0 2 0 5  
## 132 14 0 0 2 0 5  
## 133 14 0 0 2 0 4  
## 134 15 0 0 2 0 4  
## 135 15 0 0 2 0 4  
## 136 15 0 0 2 0 9  
## 137 11 1 1 2 0 3  
## 138 13 2 0 2 0 3  
## 139 14 3 0 2 0 3  
## 140 14 3 0 2 0 3

# Section 3

## Section 3.1

## [1] ".GlobalEnv" "package:psych" "package:ggplot2"   
## [4] "package:data.table" "package:knitr" "package:rstudioapi"  
## [7] "package:raster" "package:sp" "package:stats"   
## [10] "package:graphics" "package:grDevices" "package:utils"   
## [13] "package:datasets" "package:methods" "Autoloads"   
## [16] "package:base"

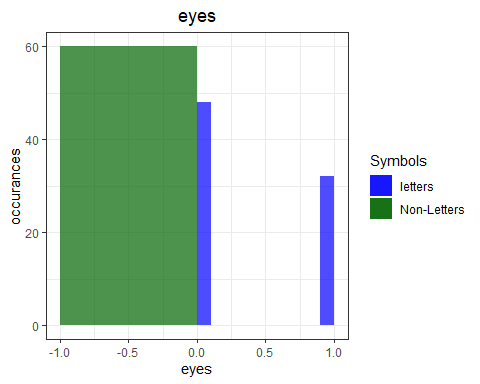
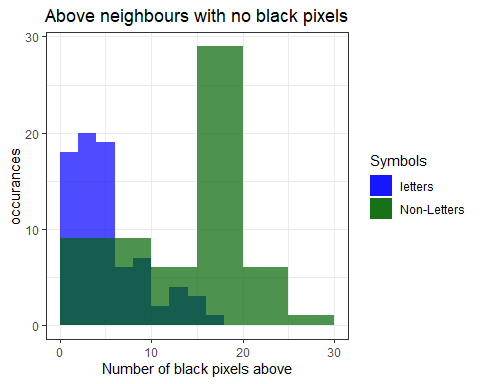
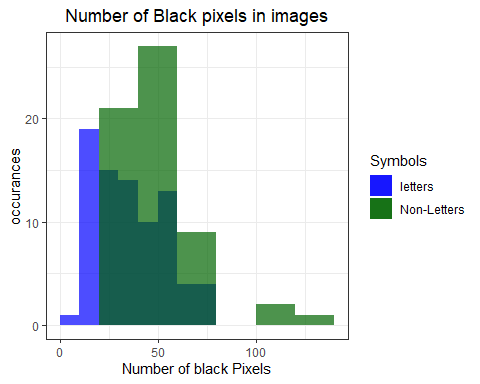
## [1] "C:/Users/eunan/Documents/CSC2062 Assignment 2"



Number of rows and cols with 1 blc px can be seen as positvely or right skewed with tail on right side of distribution, this shows there is a similarly between letters and non letters for these and may. For the number of black pixels for letters there is a middle heavy distribution shape, the non letters one is not too dissimilar but also has a bigger range and the populated more distrubuted to the right. for 3 or more black pixels in rows or cols , the distribution is interesting. letters for rows has a right skewed distribution shape where as non letter has middle heavy distribution, showing it has more rows that has 3 or more blc pixels where letters have lot of images which have none or very little, making this a very useful to discriminate between the two. For cols it follows a similar pattern as rows. For aspect ratio the distribution shape is almost bell shape for letters and almost has two peaks for non letters with the larger peak on the left side.

## Section 3.2

## [1] "Summary stats for no. of black pixels in letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 9.00 21.50 33.00 36.15 51.00 80.00   
## [1] "Summary stats for no. of black pixels in non letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 24.00 38.00 49.00 51.55 58.50 133.00   
## [1] "Summary stats for no. of rows with 1 black pixel in letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 2.737 4.250 14.000   
## [1] "Summary stats for no. of rows with 1 black pixel in non letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.2833 0.0000 3.0000   
## [1] "Summary stats for no. of cols with 1 black pixel in letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.5875 0.0000 5.0000   
## [1] "Summary stats for no. of cols with 1 black pixel in non letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 1.383 2.250 7.000   
## [1] "Summary stats for no. of rows with 3+ black pixel in letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 4.500 5.713 9.000 16.000   
## [1] "Summary stats for no. of rows with 3+ black pixel in letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 5.000 7.000 7.583 10.000 17.000   
## [1] "Summary stats for no. of cols with 3+ black pixel in letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 4.000 4.588 7.000 12.000   
## [1] "Summary stats for no. of cols with 3+ black pixel in nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 5.000 8.000 7.733 10.000 16.000   
## [1] "Summary stats for ascept ratio for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.07143 0.33333 0.48333 0.52129 0.67857 1.25000   
## [1] "Summary stats for ascept ratio for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.05882 0.36364 1.00000 0.79514 1.09091 1.30000   
## [1] "Summary stats for all neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 1 1 2 4   
## [1] "Summary stats for all neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.9667 2.0000 4.0000   
## [1] "Summary stats for above neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 3.00 5.00 5.65 7.25 17.00   
## [1] "Summary stats for above neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 10.0 17.0 14.6 20.0 26.0   
## [1] "Summary stats for below neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 3.0 5.0 6.2 7.0 21.0   
## [1] "Summary stats for below neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 10.00 16.00 14.53 20.00 24.00   
## [1] "Summary stats for left neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 8.00 10.00 10.35 13.00 19.00   
## [1] "Summary stats for left neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.0 10.0 11.0 11.5 14.0 16.0   
## [1] "Summary stats for right neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 9.00 11.00 10.45 12.00 18.00   
## [1] "Summary stats for right neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.0 10.0 11.0 11.5 14.0 16.0   
## [1] "Summary stats for horiz neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 2.000 5.688 10.250 21.000   
## [1] "Summary stats for horiz neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 2.00 2.05 4.00 8.00   
## [1] "Summary stats for vert neighs with no black pixels for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 2.0 3.6 5.0 25.0   
## [1] "Summary stats for vert neighs with no black pixels for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 2.000 3.717 8.000 14.000   
## [1] "Summary stats for connected areas for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 1.0 1.0 1.2 1.0 2.0   
## [1] "Summary stats for connected areas for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 2.000 4.000 3.333 4.000 4.000   
## [1] "Summary stats for eyes for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 0.4 1.0 1.0   
## [1] "Summary stats for eyes for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 0 0 0   
## [1] "Summary stats for widest row for letters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 5.000 5.312 7.000 13.000   
## [1] "Summary stats for widest row for nonletters"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 5.000 10.000 9.033 12.000 16.000

 Above we can see the summary statistics for the features for letters and non letters. These will help me to see which features are most useful in distinguishing between letters and non letters.

for no. of black pixels there is a clear difference between letters and non letters. These differences are shown With a difference of 15 between min, 53 between max, this shows there is a big contrast and it could 3 difference in 1be used to distinguish between letters and non letters.

Between rows and cols with 1 pixels there isn’t much difference between letters and non letters, except between max between rows with 14 for letters and 3 for non letters. This does show that these features may not be much use to distinguish between letters and non letters. For rows with 3+ black pixels they are again very similar, with 1 between max the min the same and only 2 between the means, however there is st quarter, but there isn’t much difference overall . The stats tell us, it is expected that letters have more rows and cols with 1 pixel.

For cols with 3+ black pixels, it is very similar to rows with the mean just a bit bigger and 4 difference between max and the same difference of 1st quadrant between rows with 3+ blc pixels.

Aspect ratio for all summary stats are very similar between letters and non letters.

For no. black pixels above the letters and non letters have almost exactly the same summary stats.The rest of the neighouring features, they are not so dissimilar between each other, e.g. for letters no. neighbour above and below the median is 5 for each, as well as non letter for these two with median 17 for letters and 16 for non letters as well as the means being within 0.007 of each other. for Non letters the left and right neighbours have exactly the same values, this shows the relationship between each of the neighbours of the same type. However for not the same type there can be difference, which could help distinguish between letter and non letter, such as above neighbour with 12 difference in median, 9 in the mean and 8 in max. The left and right neighbours features are less useful in distinguishing, with left only have small difference between all the summary stats.. All 4 of the right and left neighbours have near the same values for all summary stats. With the horizontal neighbours, there is a slight difference between letters and non letters, although the min, 1st quadrant are the same with 0, and the median is same as 2, there is a difference in mean of around 3 and 3rd quadrant differs by 6 and the max difference is 13, these differences may mean the horiz neighbours may be used to see differences. In terms of vertical neighbours almost all the summary stats are very similar.

With connected areas, non letters has the slightly bigger stats meaning it probably has more connected areas than letters. The difference is stats is not much as it isn’t expected for any one symbol to have a lot of different connected areas. This is similar to eyes but it is significant that non letters have no eyes, meaning eyes can be a big factor in telling if it is a letter or not because if it has an eye it will be a letter. With the final feature, widest row, non letters has bigger for everything, telling us if the wider the row the more like it is to be a non letter.

The 3 features I chose to represent in histograms:

1)number of black pixels because of the summary stats having such a wide difference with non letters having the higher values. it happens to be the case in the histogram, with letters (blue) having a high proportion in below 50. some of non letters fall below that range as well, but there is high number over 50, going the whole way to over 100.

1. No. Black pixels in above neighbour. I chose this as there was I big difference between the two in the summary stats and I think that will make it useful to distinguish between letters and non letters. And as expected from the summary stats the non letters (green) have far higher values. with letters having no image which has more than 20 pixels than have no pixels above when non letters have about 7 over 20 and up to 30, as well as having the majority over 15 pixels. this illustrate shows why it is a good feature to use.

3)Eyes. This is based off the fact that non letters has no eyes, . As see all the non letters are at 0, as well as about the 32 letter images which have an eye. No images has an more than 1 eye, so therefore if a image has an eye we know it is 100% a letter.Therefore this will be a key feature in finding a difference.

## Section 3.3

I will use statistical Analysis techniques to test if this which features are useful in telling difference between letter and non letters. After my analysis I will interpret my findings and validity my assumptions. I expect there will be features which will stand. I am making these assumptions based off the first two parts of section 3.

To start my analysis I calculated the skews for each feature for letters and non letters. I felt like this was a good stat to analysis as it will give me an idea of the distribution for letters and non letters for each feature.

skewStats<-read.csv(file ="SkewStats.csv" ,header = TRUE)  
 print(skewStats)

## letters Non.letters  
## NA NA  
## number of pixels 0.4776982 1.71719691  
## rows with 1 black pixel 1.2998730 2.44061601  
## cols with 1 black pixel 2.1517397 1.09240056  
## rows with 3 or more black pixels 0.4510288 0.47823148  
## cols with 3 or more black pixels 0.5841886 0.34679509  
## Aspect ratio 0.5365861 -0.60642373  
## neighbours with 1 black pixel 0.9651017 0.77004570  
## no above neighbours with black pixels 1.1132016 -0.66083843  
## no below neighbours with black pixels 1.4344206 -0.48729171  
## no left neighbours with black pixels 0.3059638 0.17425679  
## no right neighbours with black pixels -0.2858862 0.17425679  
## no horiz neighbours with black pixels 0.7214405 1.00656173  
## no vert neighbours with black pixels 2.3692287 0.80311748  
## connected areas 1.4719631 -0.68950297  
## eyes 0.4006176 NA  
## custom 0.6048080 -0.08176291

As shown above is the table of the skews for letters and non letters for each feature. I believe if the skews differ a lot between letters and non letters it could be a factor in telling if the feature can be used to tell difference between letters hand non letters.

To start number of black pixels skew differs between letters and non letters. letters is only moderately skewed with 0.4 whereas non letters is highly skewed with 1.7, this shows the distribution shape is massively difference. This supports our summary stats as this was one of the features I had suggested that could be used to differ between the two.

With The 1 black pixel in cols and rows there is a big difference in skews again. With rows having 1.2 difference, both having a highly skewed, this tells us that they both have a high count in the smaller values, such as 0, which was proven to be the case in 3.1, Cols is very similar.

for 3+ rows and cols for both letters and non letters the skews are very similar with both of them having approximately symmetric skew with around 0.5. Therefore these features do not differ much and would not be a good measure to find difference based on their skews.

For aspect ratio there is a 0.9 difference. for letters it approximately symmetric skew with around 0.5 as well as is non letters but with high values.Therefore this could be used to say there is a difference between letters and non letters.

for 1 black pixel neighbour it supports the summary stats where the they were very similar between letters and non letters.

In terms of the above and below neighs with no black pixels,the skews are very different for both , as expected based off summary stats, the above one 1.3 difference in skews and below 1.8, this again strengths these features can be used to tell difference between letters and non letters.

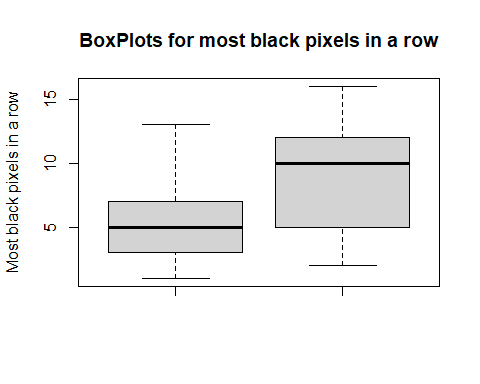
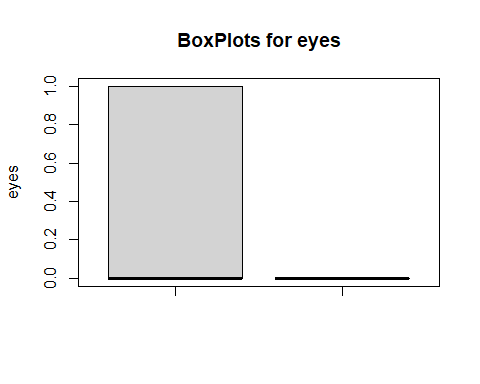
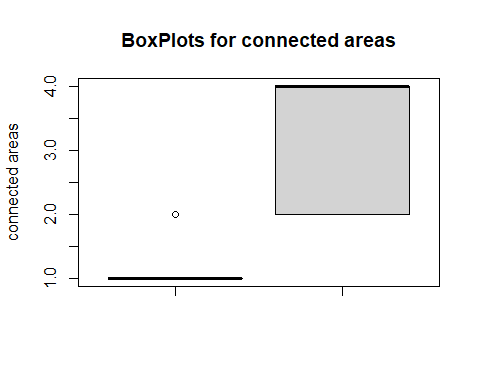
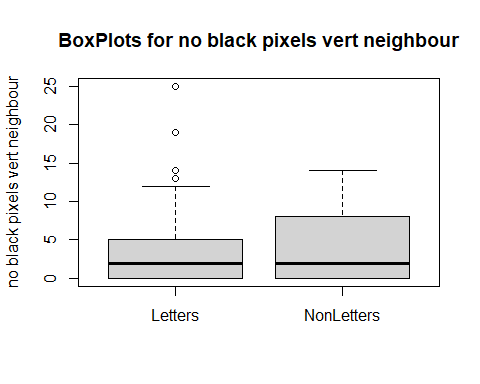
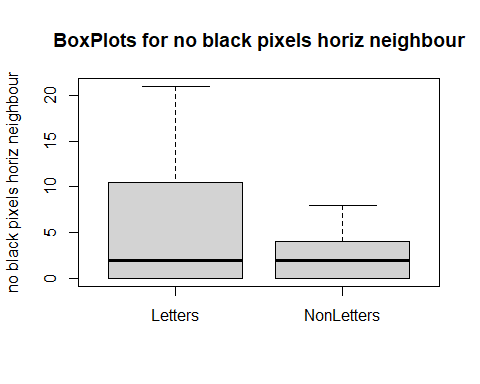
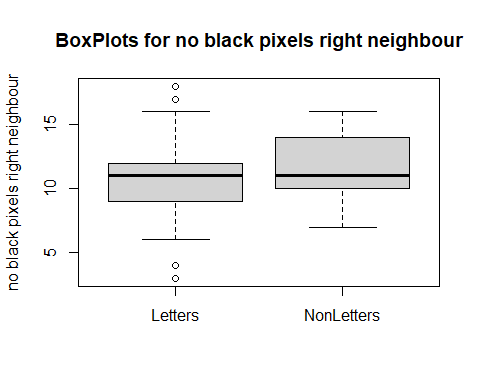
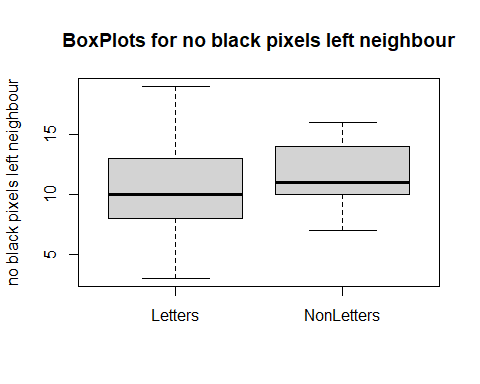
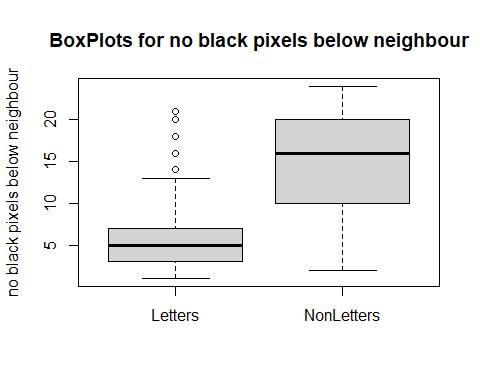
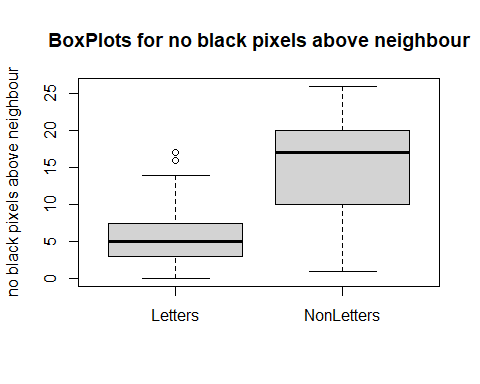
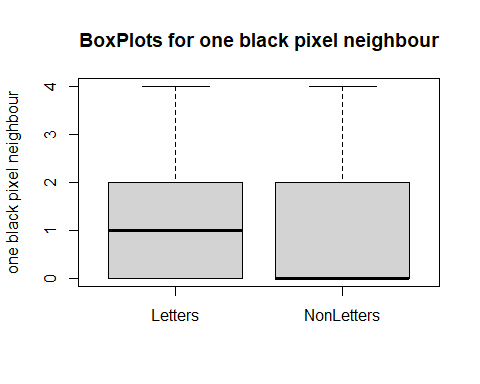
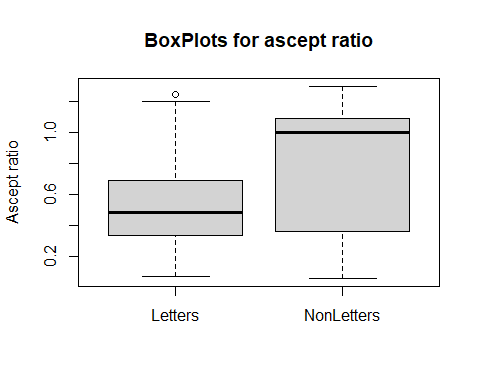
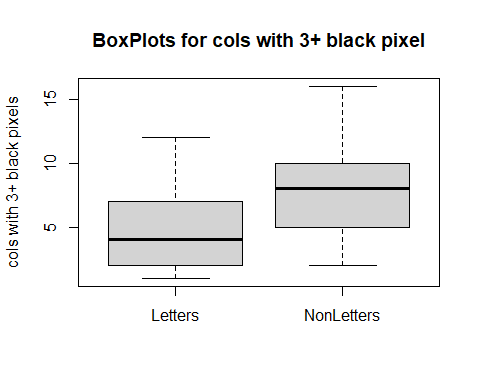
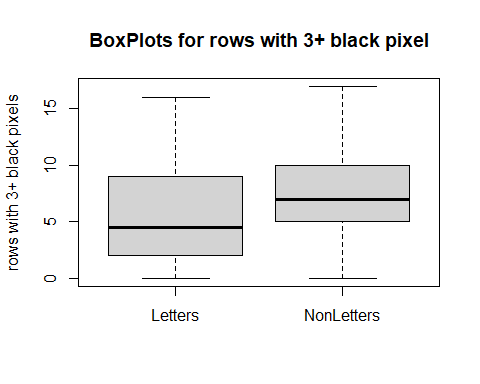
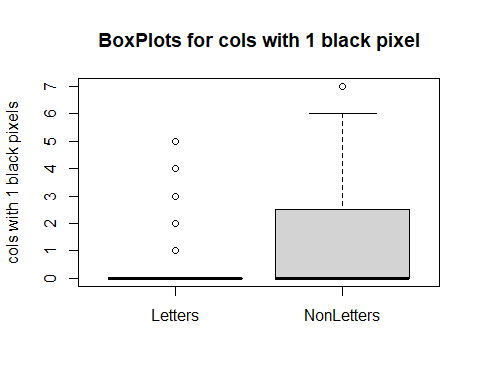
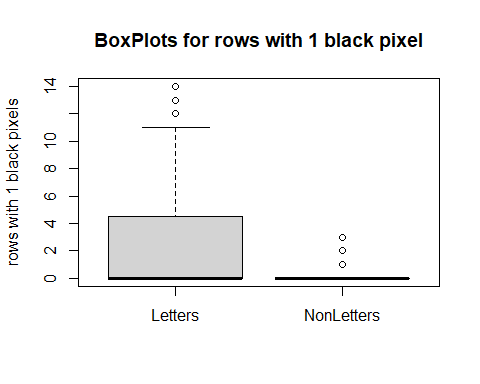
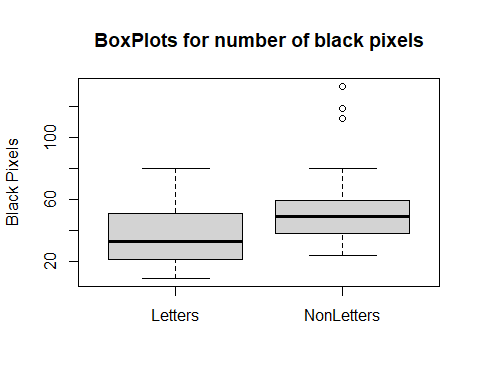
for left and right neighbours no black pixels, they are very similar with skews in a range of 0.3(letters) and 0.17(non letters) for left and -0.28(letters) and 0.17(non letters). Therefore as the same as summary stats these are suggested.

For horiz neighbours there is only 0.3 between skewness meaning there distribution is similar therefore it can not be used. The same can not be said for vert.

for Connected areas there is a massive difference of 1.8, telling us that the shape of population distribution is massively difference between letters and non letters for this feature.

For eyes as expected based off summary stats, non letters will have no skewness because all values are 0, while letters are 0.4, this shown on the summary stats. The fact it has no Skewness, makes it a key feature in telling the difference as stated in the summary stats section

For the row with most black pixels there is a difference of 1.2 skewness. With letters having 0.6, meaning it has more values which are smaller where as non letters has -0.8 meaning it has more values which are higher making it’s skewness towards left side of the graph. Meaning these could be used to differ between the two.

In next part of my analysis I will use box plots and T-tests to see if the analysis I have shown before with the skews and summary stats are proven to be correct. 

##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 3] and csvF[81:140, 3]  
## t = -4.5153, df = 113.99, p-value = 1.55e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -22.156448 -8.643552  
## sample estimates:  
## mean of x mean of y   
## 36.15 51.55   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 4] and csvF[81:140, 4]  
## t = 5.3288, df = 86.453, p-value = 7.763e-07  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.538691 3.369643  
## sample estimates:  
## mean of x mean of y   
## 2.7375000 0.2833333   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 5] and csvF[81:140, 5]  
## t = -2.699, df = 94.909, p-value = 0.008233  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.3812125 -0.2104542  
## sample estimates:  
## mean of x mean of y   
## 0.587500 1.383333   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 6] and csvF[81:140, 6]  
## t = -2.599, df = 137.43, p-value = 0.01037  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.2941904 -0.4474763  
## sample estimates:  
## mean of x mean of y   
## 5.712500 7.583333   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 7] and csvF[81:140, 7]  
## t = -5.667, df = 125.09, p-value = 9.499e-08  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.244463 -2.047204  
## sample estimates:  
## mean of x mean of y   
## 4.587500 7.733333   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 8] and csvF[81:140, 8]  
## t = -4.5882, df = 100.45, p-value = 1.294e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3922658 -0.1554463  
## sample estimates:  
## mean of x mean of y   
## 0.5212857 0.7951417   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 9] and csvF[81:140, 9]  
## t = 0.16567, df = 125.2, p-value = 0.8687  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3648635 0.4315302  
## sample estimates:  
## mean of x mean of y   
## 1.0000000 0.9666667   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 10] and csvF[81:140, 10]  
## t = -9.4505, df = 93.87, p-value = 2.742e-15  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -10.830412 -7.069588  
## sample estimates:  
## mean of x mean of y   
## 5.65 14.60   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 11] and csvF[81:140, 11]  
## t = -8.8796, df = 103.77, p-value = 2.157e-14  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -10.194420 -6.472247  
## sample estimates:  
## mean of x mean of y   
## 6.20000 14.53333   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 12] and csvF[81:140, 12]  
## t = -2.1636, df = 133.72, p-value = 0.03227  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.20125518 -0.09874482  
## sample estimates:  
## mean of x mean of y   
## 10.35 11.50   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 13] and csvF[81:140, 13]  
## t = -2.0556, df = 135.97, p-value = 0.04174  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.06014868 -0.03985132  
## sample estimates:  
## mean of x mean of y   
## 10.45 11.50   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 14] and csvF[81:140, 14]  
## t = 4.7033, df = 104.39, p-value = 7.87e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 2.103915 5.171085  
## sample estimates:  
## mean of x mean of y   
## 5.6875 2.0500   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 15] and csvF[81:140, 15]  
## t = -0.1435, df = 134.61, p-value = 0.8861  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.724575 1.491242  
## sample estimates:  
## mean of x mean of y   
## 3.600000 3.716667   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 16] and csvF[81:140, 16]  
## t = -16.318, df = 74.918, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.393772 -1.872894  
## sample estimates:  
## mean of x mean of y   
## 1.200000 3.333333   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 17] and csvF[81:140, 17]  
## t = 7.2572, df = 79, p-value = 2.399e-10  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.2902907 0.5097093  
## sample estimates:  
## mean of x mean of y   
## 0.4 0.0   
##   
##   
## Welch Two Sample t-test  
##   
## data: csvF[1:80, 18] and csvF[81:140, 18]  
## t = -6.3872, df = 101.81, p-value = 5.115e-09  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -4.876327 -2.565340  
## sample estimates:  
## mean of x mean of y   
## 5.312500 9.033333

Above you can see the boxplots, which I have decided to use to further validate my early findings. It graphically shows which features differ for letters and non letters, such as first feature number of black pixels. Which with my early findings I found to be be a feature that could be used to differ between letters and non letters, these box plots shows the difference in ranges, and they may not be so different as first expected with only a few massive valeus for non letters which may of helped raised the mean, but the overall range isn’t too different to different between the two. and With eyes, this shows graphically what the skews showed, that all non letters have no eyes, as everything is 0 for it in the boxplot, this is not the same for letters, therefore it shows us what we assumed previously.

I think these box plots has helped to further prove the difference or similarities between letters and non letters for each feature. As mean is not included I will now use T-tests to test the mean between letters and non letters for each feature and come to a conclusion which features are most useful in terms of telling the difference between the letters and non letters. I will use welch t-tests as the populations are not the same size to do paired t tests.

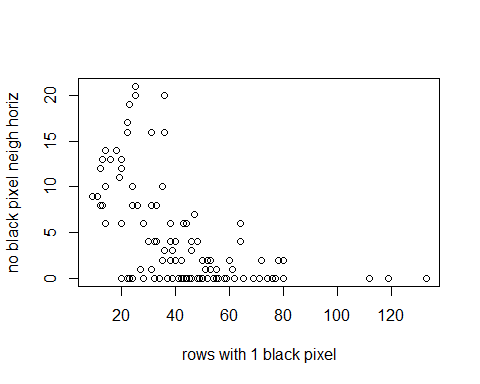
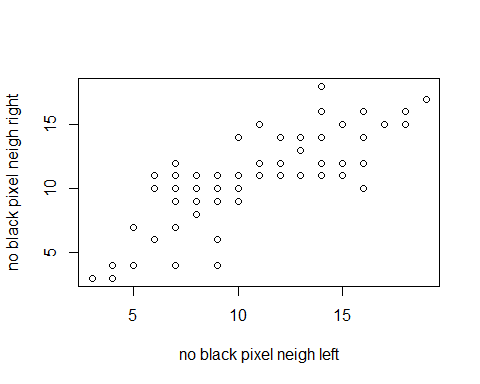
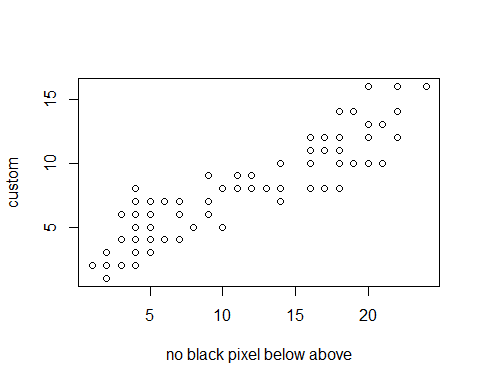
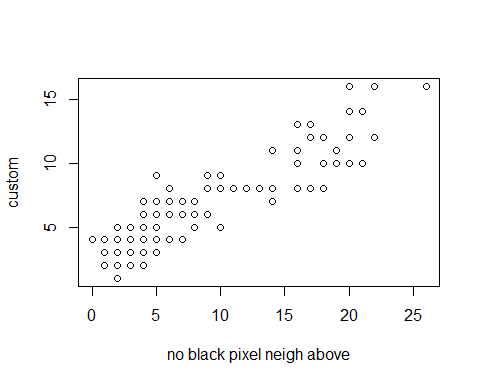
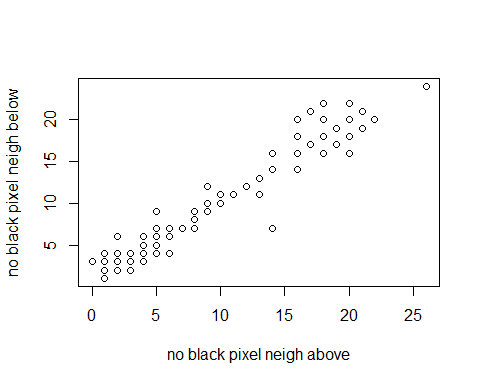
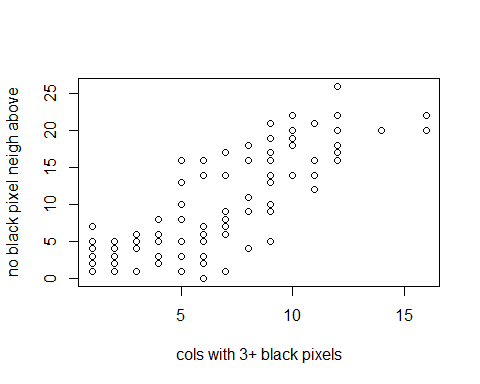
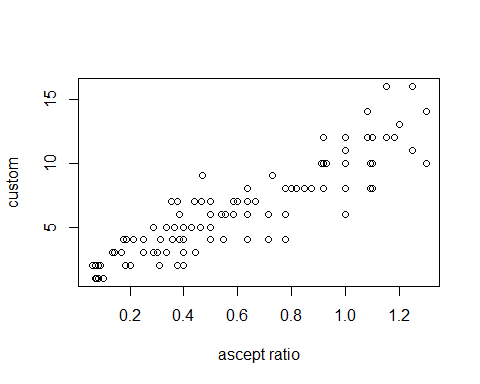
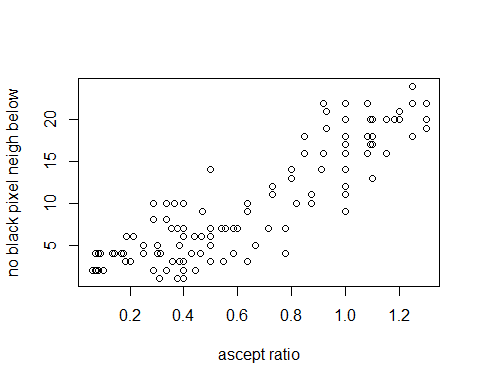
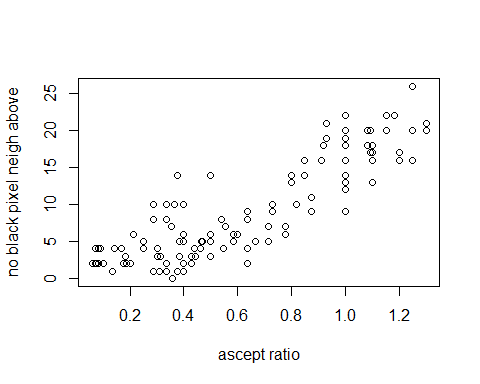
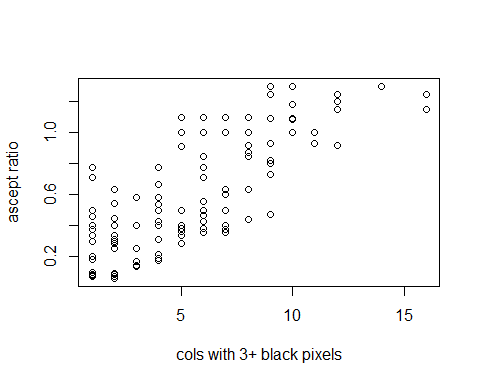
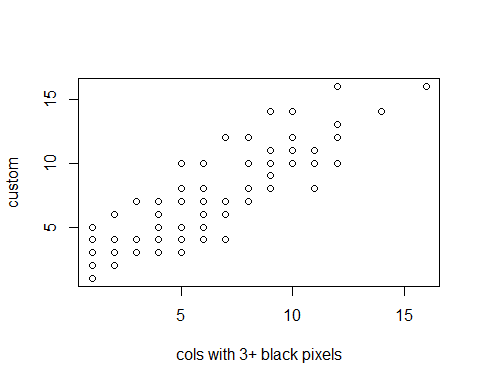
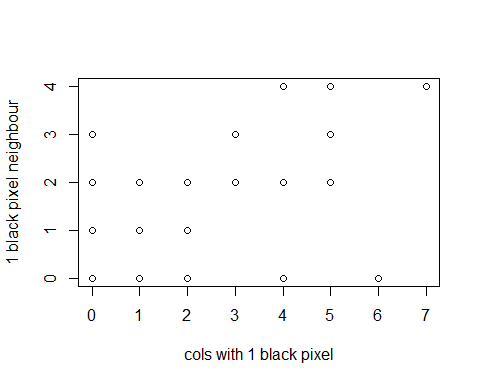
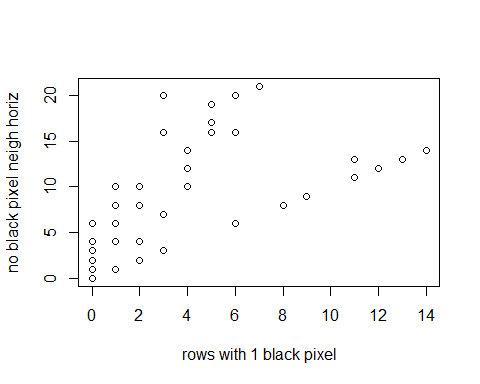
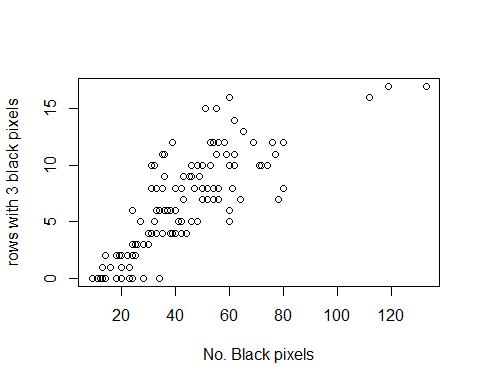
T-tests are good for statistical analyses as it’s used to compare the means of two groups, which is perfect for the analysis we want to perform. There is null hypothesis and alternative hypothesis. The null hypothesis is the difference in group means is zero and the alternative hypothesis is the difference in group means is different from zero. It is used to test if there is a difference between two groups. It also calculates a p value. A p value is the probability that the results from your data occurred by chance. They range from 0% to 1% and written in decimal. If the p value is less than 0.05 it is strongly significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis. A p-value higher than 0.05 is not statistically significant and indicates strong evidence for the null hypothesis. Therefore the p value can be used to help us verify our assumptions, As you can see from the t-tests the p values are behind 5% for all them except for feature 7, 1 black pixel neighbour with 8.687% and feature 13 vert no black pixel neighbour with 8.8%. This tells us that there is a significant difference between letters and non letters for the rest of the features. The t-test shows differences in means which tell us that there is a difference between letters and non letters. The biggest difference in mean is 15.4 for feature 1, which backs up our earlier summary which suggested that it would be a good feature to tell difference. Feature 8 no\_neigh\_above has a mean difference of 8.95 between letters and non letters which backs up earlier analysis which suggested it was a good feature to tell difference, the same can be said for no\_neigh\_below, which difference in mean is about 8.3. As you can see above the means differ slightly between other features which have a signicant difference between means but based off the t tests as well as the box plots and skews I believe these 3 features are the best suited to tell the difference between letters and non letters as well as eyes because it’s all 0’s for non letters. ## Section 3.4

cors<-read.csv(file ="correlations.csv" ,header = TRUE)  
 print(cors)

## nr\_pix rows\_with\_1 cols\_with\_1 rows\_with\_3p cols\_with\_3p  
## nr\_pix 1.000000000 -0.56077283 -0.31520252 0.79119224 0.65531196  
## rows\_with\_1 -0.560772832 1.00000000 0.20703546 -0.51041636 -0.56455712  
## cols\_with\_1 -0.315202522 0.20703546 1.00000000 -0.35075351 -0.05731039  
## rows\_with\_3p 0.791192236 -0.51041636 -0.35075351 1.00000000 0.40239633  
## cols\_with\_3p 0.655311961 -0.56455712 -0.05731039 0.40239633 1.00000000  
## aspect\_ratio 0.269571627 -0.37857492 0.32782637 0.04574902 0.79054621  
## neigh\_1 -0.527050327 0.52358786 0.62621567 -0.51484769 -0.36605708  
## no\_neigh\_above 0.389727091 -0.44101965 0.23872825 0.10674969 0.82412908  
## no\_neigh\_below 0.424176500 -0.45358925 0.19011901 0.14678716 0.85109076  
## no\_neigh\_left 0.231560393 -0.03236569 -0.02514277 0.21728960 -0.15465860  
## no\_neigh\_right 0.280120977 -0.05539138 -0.07145362 0.32795443 -0.13685876  
## no\_neigh\_horiz -0.553150627 0.74253454 0.31558577 -0.51460288 -0.41145479  
## no\_neigh\_vert -0.358053170 0.12313225 0.49619393 -0.37686593 0.08633827  
## connected\_areas 0.145949593 -0.27010351 0.38230658 -0.09012643 0.48762501  
## eyes -0.001636818 -0.01002933 -0.28344259 0.05213987 0.07174159  
## custom 0.501925418 -0.47140542 0.19978182 0.22321649 0.89063760  
## aspect\_ratio neigh\_1 no\_neigh\_above no\_neigh\_below  
## nr\_pix 0.26957163 -0.527050327 0.3897271 0.42417650  
## rows\_with\_1 -0.37857492 0.523587860 -0.4410196 -0.45358925  
## cols\_with\_1 0.32782637 0.626215671 0.2387283 0.19011901  
## rows\_with\_3p 0.04574902 -0.514847695 0.1067497 0.14678716  
## cols\_with\_3p 0.79054621 -0.366057083 0.8241291 0.85109076  
## aspect\_ratio 1.00000000 0.013853963 0.8742152 0.87366851  
## neigh\_1 0.01385396 1.000000000 -0.0576068 -0.09989739  
## no\_neigh\_above 0.87421524 -0.057606799 1.0000000 0.97331030  
## no\_neigh\_below 0.87366851 -0.099897385 0.9733103 1.00000000  
## no\_neigh\_left -0.32708801 -0.004245014 -0.1976094 -0.21184056  
## no\_neigh\_right -0.38154687 -0.061983461 -0.2231341 -0.23449750  
## no\_neigh\_horiz -0.14660995 0.549348544 -0.3246704 -0.34165713  
## no\_neigh\_vert 0.40274009 0.411588829 0.2538882 0.24009345  
## connected\_areas 0.56714810 0.166609793 0.7772364 0.74155466  
## eyes -0.06090808 -0.241570293 -0.2559332 -0.18827842  
## custom 0.90665137 -0.132159742 0.9054797 0.91623642  
## no\_neigh\_left no\_neigh\_right no\_neigh\_horiz no\_neigh\_vert  
## nr\_pix 0.231560393 0.280120977 -0.553150627 -0.35805317  
## rows\_with\_1 -0.032365688 -0.055391384 0.742534541 0.12313225  
## cols\_with\_1 -0.025142766 -0.071453618 0.315585769 0.49619393  
## rows\_with\_3p 0.217289604 0.327954431 -0.514602884 -0.37686593  
## cols\_with\_3p -0.154658599 -0.136858763 -0.411454790 0.08633827  
## aspect\_ratio -0.327088011 -0.381546866 -0.146609946 0.40274009  
## neigh\_1 -0.004245014 -0.061983461 0.549348544 0.41158883  
## no\_neigh\_above -0.197609353 -0.223134106 -0.324670415 0.25388820  
## no\_neigh\_below -0.211840557 -0.234497501 -0.341657129 0.24009345  
## no\_neigh\_left 1.000000000 0.859115751 0.051164044 -0.41864146  
## no\_neigh\_right 0.859115751 1.000000000 -0.003444361 -0.46270991  
## no\_neigh\_horiz 0.051164044 -0.003444361 1.000000000 0.37844732  
## no\_neigh\_vert -0.418641465 -0.462709913 0.378447318 1.00000000  
## connected\_areas 0.036491280 0.031327999 -0.233734071 0.15021337  
## eyes -0.218656912 -0.164471996 0.216314783 0.17728945  
## custom -0.179786216 -0.211130296 -0.297906256 0.22938021  
## connected\_areas eyes custom  
## nr\_pix 0.14594959 -0.001636818 0.50192542  
## rows\_with\_1 -0.27010351 -0.010029326 -0.47140542  
## cols\_with\_1 0.38230658 -0.283442586 0.19978182  
## rows\_with\_3p -0.09012643 0.052139865 0.22321649  
## cols\_with\_3p 0.48762501 0.071741592 0.89063760  
## aspect\_ratio 0.56714810 -0.060908078 0.90665137  
## neigh\_1 0.16660979 -0.241570293 -0.13215974  
## no\_neigh\_above 0.77723638 -0.255933209 0.90547975  
## no\_neigh\_below 0.74155466 -0.188278424 0.91623642  
## no\_neigh\_left 0.03649128 -0.218656912 -0.17978622  
## no\_neigh\_right 0.03132800 -0.164471996 -0.21113030  
## no\_neigh\_horiz -0.23373407 0.216314783 -0.29790626  
## no\_neigh\_vert 0.15021337 0.177289447 0.22938021  
## connected\_areas 1.00000000 -0.481481481 0.60884042  
## eyes -0.48148148 1.000000000 -0.09157427  
## custom 0.60884042 -0.091574267 1.00000000

For this part, I used the cor function in r, which calculates a value between -1 and 1, values close to -1 mean a negative correlation between two features and a value near 1 means a positive correlation between the two features.

In the table above it shows all the correlations between each of the feauteres. There isn’t many negative correlations with the biggest being -0.553150627 between number of black pixels and and no black pixels horiz neighbours, but this isn’t a big enough correlation to document as a degree of linear assosciation.

Through this analysis I found 14 significant degrees of correlation and this and shown below in scatter graphs to prove the relationships between the features. The highest cor value was 0.973310298355819 with no neighbour above and no neighbour left.  As you can see with the first graph the correlation of 0.79119224 is clearly shown in the scatter plot as for the most part the bigger the number of black pixels the more rows with 3+ black pixels you have. I believe this is because the more black pixels you have, the more you will have on each row, therefore you are more likely to have 3+ on each row, meaning that there is a positive correlation and a clear degree of linear association between the two.

For the second graph, rows with 1 black pixel and and no black pixel neighbour horiz, the correlation is less than define but still sigincant with 0.742534541, there are a number of plots which go against the correlation as seen in graph above with highest rows with 1 black pixel coming in at around 7 no neigh horiz which is about half way through it’s range, but overall there is a degree of linear assoication between the feauteres, I think this is the case because if it is a row with 1 black pixel it is expected to not have any neighbours horiz as that would therefore be on the same row as them, but what lets this degree of linear association down is the fact that the black pixel on same row could not be a horiz neighbour to a black pixel, which of course the no black pixel horiz neighbour does not calculate.

For the 4th scatter plot, between widest black pixel row and cols with 3+ black pixel, it is clear by just looking at the scatter plot distrubution there is a degree of linear assosciation between the two, this is the case as the wider the row the more like there is to be more rows and cols like that, e.g. over 3+ black pixels hence giving a lot of 3+ black pixel cols, the graph clearly shows this

In terms of the aspect ratio scatter plots, it has a lot of features which it has a high degree of linear association with, these are cols with 3+ which has 0.79054621 cor value, no neigh above 0.8742152 cor value , no neigh below cor value 0.87366851 and widest row with black pixel 0.90665137 cor value. These plots are shown above and expected out of all them row with most black pixels and aspect ratio has the most significant distributions to show a degree of linear association, this is expected as because it uses the leftmost and rightmost black pixel and using the widest row therefore, this row that the custom feature is finding, therefore there is a significant association between the two.

The 10th scatter plot, is the highest degree of linear association between all the features, and looking at the scatter graph this is the case, as the distribution is nearly a perfect diagonal line and certainly the most perfect distribution we see out of these scatter plots as expected from the cor value. This relationship is so significant, because if a black pixel does not have above it, it could mean there is another black pixel with the same pixel space that does have a below neighbour.

For the second last feature, the left and right no black pixel has a cor value of 0.859115751. The significant degree of linear association is clear in the scatter plot, with the distrubution pattern showing that there is a relationship to the two, and it is simple why there. If a black pixel does not have a left neighbours another black pixel near it will not have a right neighbours, and vice versa leading to this relationship

For the last feature the cor is 0.742534541 between rows with 1 black pixel and no neigh horiz, one of the smallest value for the significant correlations and this is clear on the scatter plot as the distribution does not show one of relationship between the two, although there is a high cor value for the two, they may not have a high degree of linear association between the two features.

To conclude I believe I have highlighted the most significant relationships and provided scatter plots which I have used to describe why there may be a degree of linear association between the two features.

# Section 4

## Section 4.1

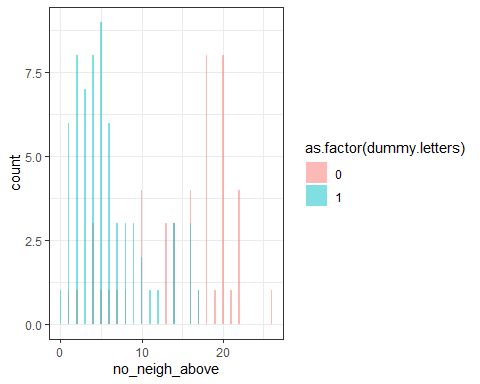
## [1] ".GlobalEnv" "package:MASS" "package:psych"   
## [4] "package:ggplot2" "package:data.table" "package:knitr"   
## [7] "package:rstudioapi" "package:raster" "package:sp"   
## [10] "package:stats" "package:graphics" "package:grDevices"   
## [13] "package:utils" "package:datasets" "package:methods"   
## [16] "Autoloads" "package:base"

## [1] "C:/Users/eunan/Documents/CSC2062 Assignment 2"

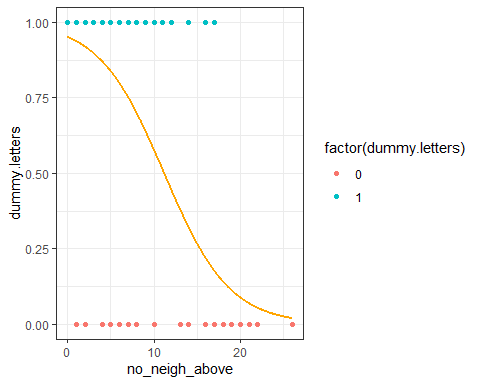
##   
## Call:  
## lm(formula = aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 +   
## rows\_with\_3p + cols\_with\_3p + neigh\_1 + no\_neigh\_above +   
## no\_neigh\_below + no\_neigh\_left + no\_neigh\_right + no\_neigh\_horiz +   
## no\_neigh\_vert + connected\_areas + eyes + custom, data = csvF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.279393 -0.065511 -0.008668 0.068580 0.300786   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.341097 0.055373 6.160 9.33e-09 \*\*\*  
## nr\_pix -0.002922 0.001030 -2.838 0.005301 \*\*   
## rows\_with\_1 -0.018600 0.005234 -3.554 0.000539 \*\*\*  
## cols\_with\_1 0.027333 0.008551 3.196 0.001766 \*\*   
## rows\_with\_3p 0.003366 0.003930 0.857 0.393349   
## cols\_with\_3p 0.020738 0.008520 2.434 0.016351 \*   
## neigh\_1 0.013820 0.012934 1.069 0.287368   
## no\_neigh\_above 0.013544 0.006918 1.958 0.052499 .   
## no\_neigh\_below 0.007007 0.006783 1.033 0.303579   
## no\_neigh\_left -0.001529 0.005697 -0.268 0.788869   
## no\_neigh\_right -0.014832 0.006363 -2.331 0.021360 \*   
## no\_neigh\_horiz 0.011913 0.003546 3.360 0.001037 \*\*   
## no\_neigh\_vert -0.002262 0.003027 -0.747 0.456318   
## connected\_areas -0.042928 0.016035 -2.677 0.008430 \*\*   
## eyes -0.026158 0.033712 -0.776 0.439277   
## custom 0.044436 0.008250 5.386 3.47e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1075 on 124 degrees of freedom  
## Multiple R-squared: 0.9198, Adjusted R-squared: 0.9101   
## F-statistic: 94.81 on 15 and 124 DF, p-value: < 2.2e-16  
##   
## Start: AIC=-609.55  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + rows\_with\_3p +   
## cols\_with\_3p + neigh\_1 + no\_neigh\_above + no\_neigh\_below +   
## no\_neigh\_left + no\_neigh\_right + no\_neigh\_horiz + no\_neigh\_vert +   
## connected\_areas + eyes + custom  
##   
## Df Sum of Sq RSS AIC  
## - no\_neigh\_left 1 0.00083 1.4329 -611.47  
## - no\_neigh\_vert 1 0.00645 1.4385 -610.93  
## - eyes 1 0.00695 1.4390 -610.88  
## - rows\_with\_3p 1 0.00847 1.4405 -610.73  
## - no\_neigh\_below 1 0.01233 1.4444 -610.35  
## - neigh\_1 1 0.01319 1.4452 -610.27  
## <none> 1.4321 -609.55  
## - no\_neigh\_above 1 0.04427 1.4763 -607.29  
## - no\_neigh\_right 1 0.06276 1.4948 -605.55  
## - cols\_with\_3p 1 0.06843 1.5005 -605.02  
## - connected\_areas 1 0.08277 1.5148 -603.69  
## - nr\_pix 1 0.09304 1.5251 -602.74  
## - cols\_with\_1 1 0.11799 1.5500 -600.47  
## - no\_neigh\_horiz 1 0.13035 1.5624 -599.36  
## - rows\_with\_1 1 0.14583 1.5779 -597.98  
## - custom 1 0.33508 1.7671 -582.12  
##   
## Step: AIC=-611.47  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + rows\_with\_3p +   
## cols\_with\_3p + neigh\_1 + no\_neigh\_above + no\_neigh\_below +   
## no\_neigh\_right + no\_neigh\_horiz + no\_neigh\_vert + connected\_areas +   
## eyes + custom  
##   
## Df Sum of Sq RSS AIC  
## - no\_neigh\_vert 1 0.00604 1.4389 -612.88  
## - eyes 1 0.00633 1.4392 -612.86  
## - rows\_with\_3p 1 0.00977 1.4427 -612.52  
## - no\_neigh\_below 1 0.01216 1.4450 -612.29  
## - neigh\_1 1 0.01310 1.4460 -612.20  
## <none> 1.4329 -611.47  
## - no\_neigh\_above 1 0.04509 1.4780 -609.14  
## - cols\_with\_3p 1 0.06988 1.5028 -606.81  
## - connected\_areas 1 0.08245 1.5153 -605.64  
## - nr\_pix 1 0.09942 1.5323 -604.08  
## - cols\_with\_1 1 0.11781 1.5507 -602.41  
## - no\_neigh\_horiz 1 0.13206 1.5649 -601.13  
## - rows\_with\_1 1 0.14727 1.5801 -599.78  
## - no\_neigh\_right 1 0.15844 1.5913 -598.79  
## - custom 1 0.33425 1.7671 -584.12  
##   
## Step: AIC=-612.88  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + rows\_with\_3p +   
## cols\_with\_3p + neigh\_1 + no\_neigh\_above + no\_neigh\_below +   
## no\_neigh\_right + no\_neigh\_horiz + connected\_areas + eyes +   
## custom  
##   
## Df Sum of Sq RSS AIC  
## - rows\_with\_3p 1 0.00881 1.4477 -614.03  
## - eyes 1 0.00998 1.4489 -613.92  
## - neigh\_1 1 0.01078 1.4497 -613.84  
## - no\_neigh\_below 1 0.01152 1.4505 -613.77  
## <none> 1.4389 -612.88  
## - no\_neigh\_above 1 0.04226 1.4812 -610.83  
## - cols\_with\_3p 1 0.06998 1.5089 -608.24  
## - connected\_areas 1 0.08012 1.5191 -607.30  
## - nr\_pix 1 0.09466 1.5336 -605.96  
## - cols\_with\_1 1 0.11267 1.5516 -604.33  
## - no\_neigh\_horiz 1 0.12677 1.5657 -603.06  
## - rows\_with\_1 1 0.14139 1.5803 -601.76  
## - no\_neigh\_right 1 0.15529 1.5942 -600.54  
## - custom 1 0.34233 1.7813 -585.01  
##   
## Step: AIC=-614.03  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + cols\_with\_3p +   
## neigh\_1 + no\_neigh\_above + no\_neigh\_below + no\_neigh\_right +   
## no\_neigh\_horiz + connected\_areas + eyes + custom  
##   
## Df Sum of Sq RSS AIC  
## - neigh\_1 1 0.01081 1.4586 -614.99  
## - eyes 1 0.01139 1.4591 -614.93  
## - no\_neigh\_below 1 0.01282 1.4606 -614.79  
## <none> 1.4477 -614.03  
## - no\_neigh\_above 1 0.03937 1.4871 -612.27  
## - cols\_with\_3p 1 0.07531 1.5231 -608.93  
## - nr\_pix 1 0.09322 1.5410 -607.29  
## - connected\_areas 1 0.09538 1.5431 -607.10  
## - cols\_with\_1 1 0.11856 1.5663 -605.01  
## - no\_neigh\_horiz 1 0.12121 1.5690 -604.77  
## - no\_neigh\_right 1 0.14758 1.5953 -602.44  
## - rows\_with\_1 1 0.15398 1.6017 -601.88  
## - custom 1 0.33603 1.7838 -586.81  
##   
## Step: AIC=-614.99  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + cols\_with\_3p +   
## no\_neigh\_above + no\_neigh\_below + no\_neigh\_right + no\_neigh\_horiz +   
## connected\_areas + eyes + custom  
##   
## Df Sum of Sq RSS AIC  
## - eyes 1 0.01343 1.4720 -615.70  
## - no\_neigh\_below 1 0.01344 1.4720 -615.70  
## <none> 1.4586 -614.99  
## - no\_neigh\_above 1 0.03949 1.4981 -613.25  
## - cols\_with\_3p 1 0.06760 1.5261 -610.65  
## - connected\_areas 1 0.08938 1.5479 -608.66  
## - nr\_pix 1 0.09617 1.5547 -608.05  
## - no\_neigh\_horiz 1 0.14474 1.6033 -603.74  
## - rows\_with\_1 1 0.14569 1.6042 -603.66  
## - no\_neigh\_right 1 0.15125 1.6098 -603.17  
## - cols\_with\_1 1 0.17522 1.6338 -601.10  
## - custom 1 0.34357 1.8021 -587.37  
##   
## Step: AIC=-615.7  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + cols\_with\_3p +   
## no\_neigh\_above + no\_neigh\_below + no\_neigh\_right + no\_neigh\_horiz +   
## connected\_areas + custom  
##   
## Df Sum of Sq RSS AIC  
## - no\_neigh\_below 1 0.01009 1.4821 -616.75  
## <none> 1.4720 -615.70  
## - cols\_with\_3p 1 0.05434 1.5263 -612.63  
## - no\_neigh\_above 1 0.05850 1.5305 -612.25  
## - connected\_areas 1 0.07897 1.5510 -610.39  
## - nr\_pix 1 0.08883 1.5608 -609.50  
## - rows\_with\_1 1 0.13227 1.6043 -605.66  
## - no\_neigh\_horiz 1 0.13234 1.6043 -605.65  
## - no\_neigh\_right 1 0.14041 1.6124 -604.95  
## - cols\_with\_1 1 0.18495 1.6569 -601.13  
## - custom 1 0.35750 1.8295 -587.27  
##   
## Step: AIC=-616.75  
## aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 + cols\_with\_3p +   
## no\_neigh\_above + no\_neigh\_right + no\_neigh\_horiz + connected\_areas +   
## custom  
##   
## Df Sum of Sq RSS AIC  
## <none> 1.4821 -616.75  
## - cols\_with\_3p 1 0.06474 1.5468 -612.76  
## - connected\_areas 1 0.07206 1.5541 -612.10  
## - nr\_pix 1 0.08870 1.5708 -610.61  
## - rows\_with\_1 1 0.12861 1.6107 -607.10  
## - no\_neigh\_horiz 1 0.12940 1.6115 -607.03  
## - no\_neigh\_right 1 0.15241 1.6345 -605.04  
## - cols\_with\_1 1 0.17811 1.6602 -602.86  
## - no\_neigh\_above 1 0.18983 1.6719 -601.87  
## - custom 1 0.40278 1.8849 -585.09  
##   
## Call:  
## lm(formula = aspect\_ratio ~ nr\_pix + rows\_with\_1 + cols\_with\_1 +   
## cols\_with\_3p + no\_neigh\_above + no\_neigh\_right + no\_neigh\_horiz +   
## connected\_areas + custom, data = csvF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.302730 -0.065542 -0.008587 0.066007 0.311728   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3215079 0.0454991 7.066 8.68e-11 \*\*\*  
## nr\_pix -0.0023211 0.0008321 -2.789 0.006075 \*\*   
## rows\_with\_1 -0.0155955 0.0046432 -3.359 0.001028 \*\*   
## cols\_with\_1 0.0290180 0.0073415 3.953 0.000126 \*\*\*  
## cols\_with\_3p 0.0177960 0.0074680 2.383 0.018621 \*   
## no\_neigh\_above 0.0190536 0.0046693 4.081 7.79e-05 \*\*\*  
## no\_neigh\_right -0.0144056 0.0039399 -3.656 0.000370 \*\*\*  
## no\_neigh\_horiz 0.0099421 0.0029510 3.369 0.000993 \*\*\*  
## connected\_areas -0.0375398 0.0149319 -2.514 0.013155 \*   
## custom 0.0470077 0.0079086 5.944 2.41e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1068 on 130 degrees of freedom  
## Multiple R-squared: 0.917, Adjusted R-squared: 0.9113   
## F-statistic: 159.6 on 9 and 130 DF, p-value: < 2.2e-16

## Section 4.2

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0



##   
## Call:  
## glm(formula = dummy.letters ~ no\_neigh\_above, family = "binomial",   
## data = trainingdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3525 -0.5521 0.4094 0.5949 1.8619   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.96740 0.51887 5.719 1.07e-08 \*\*\*  
## no\_neigh\_above -0.26508 0.04555 -5.819 5.91e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 152.360 on 111 degrees of freedom  
## Residual deviance: 97.038 on 110 degrees of freedom  
## AIC: 101.04  
##   
## Number of Fisher Scoring iterations: 4  
##   
## 1 2 3   
## 0.8378073 0.5784977 0.0883317



## [1] "Prop correct for training data"  
## [1] 0.8125  
## [1] "Prop incorrect for training data"  
## [1] 0.1875  
## [1] "Prop correct for test data"  
## [1] 0.7857143  
## [1] "Prop incorrect for test data"  
## [1] 0.2142857

The results are shown on the graph above as well as the accuracy of the model. The plot shows that the lower the no\_neigh\_above the more chance it is a letter. Based on the predictions I did of 5 10 20, which came out as 0.83…, 0.57… and 0.0883… respectfully, this is proven to be the case on the plot. It is 81.25% accurate for training data and 78.57% for test data, It is expected that training data is higher % as it was used to fit the model.

## Section 4.3 and 4.4

## Split1 Split2 Split3  
## Letters 0.3875 0.375 0.5375  
## Faces 0.6250 1.000 0.5750  
## Exclamation Marks 0.7000 0.000 0.1500  
## [1] "Naive bayes classifer for all 1s in letters"  
## [1] 0.0446317  
## [1] "Naive bayes classifer for all 0s in letters"  
## [1] 0.1011719  
## [1] "Naive bayes classifer for all 1s in faces"  
## [1] 0.1026786  
## [1] "Naive bayes classifer for all 0s in faces"  
## [1] 0  
## [1] "Naive bayes classifer for all 1s in xclaim"  
## [1] 0  
## [1] "Naive bayes classifer for all 0s in xclaim"  
## [1] 0.03642857

For 4.4 the results for the naive bayes classifer is shown above for 1 for all 3 splits and then 0 for all 3. faces has the most chance of having above median for all 3 with 1.026786% and xclaim has lowest for all 1s with 0, as it has aspect ratio above the median. For 0s for all 3, letters having most chance with 1.011719%. The least chance is faces with 0, as all it’s values for aspect ratio are bigger than the median, therefore it is not possible for faces to have any images with all 0’s.