Q1

```
#IE 500 SMLE HW5 Q1
#part 1
wbdc <- read.table("C:/Users/ppill/Desktop/R files/wdbc.data", header = FALSE,sep = ",")
colnames(wbdc) <- c('ID','Diagnosis','Mean_Radius','Mean_Texture','Mean_Perimeter','Mean_Area',
                    'Mean_Smoothness','Mean_Compactness','Mean_Concavity', 'Mean_ConcavePoints',
                    'Mean_Symmetry','Mean_FractalDimesnsion','Radius_SE','Texture_SE','Perimeter_SE',
                    'Area_SE','Smoothness_SE','Compactness_SE','Concavity_SE','ConcavePoints_SE',
                    'Symmetry_SE','FractalDimension_SE','Worst_Radius','Worst_Texture','Worst_Perimeter',
                    'Worst_Area','Worst_Smoothness','Worst_Compactness','Worst_Concavity','Worst_ConcavePoints',
                    'Worst_Symmetry','Worst_FractalDimesnsion')
wbdc <- na.omit(wbdc)
wbdc$Diagnosis <- ifelse(wbdc$Diagnosis=="M",2,1)
wbdc$Diagnosis <- as.numeric(as.factor(wbdc$Diagnosis))
hist(wbdc$Diagnosis)
# part 2
#k-means clustering
set.seed(125)
wbdc_12 <- wbdc[,3:32]
kcluster <- kmeans(wbdc_12,2, nstart = 125)
kcluster

#hierarchical clustering
str(wbdc_12)
summary(wbdc_12)
any(is.na(wbdc_12))
hcluster <- hclust(dist(wbdc_12), method = "complete")
cutree_hcluster <- cutree(hcluster, k=2)
hcluster
table(wbdc$Diagnosis,kcluster$cluster)
table(wbdc$Diagnosis,cutree_hcluster)

#principle component analysis
wbdc_pc <- prcomp(wbdc_12, center = TRUE, scale. = TRUE)
attributes(wbdc_pc)
print(wbdc_pc)
summary(wbdc_pc)
wbdc_pc$center
wbdc_pc$scale
wbdc_pc$rotation
head(wbdc_pc$x)
kcluster_pca <- kmeans(wbdc_pc$x[,1:17],2, nstart = 125)
table(kcluster_pca$cluster,wbdc$Diagnosis)
hcluster_pca <- hclust(dist(wbdc_pc$x[,1:17]),method = "complete")
hcluster_clusters <- cutree(hcluster_pca, k=2 )
table(hcluster_clusters,wbdc$Diagnosis)

#data visualization

plot(wbdc_pc$x[,c(1,2)],col = (wbdc$Diagnosis+1),xlab = "PC1",ylab = "PC2")
plot(wbdc_pc$x[,c(1,4)],col = (wbdc$Diagnosis+1),xlab = "PC1",ylab = "PC4")
plot(wbdc_pc$x[,c(1,6)],col = (wbdc$Diagnosis+1),xlab = "PC1",ylab = "PC6")
plot(hcluster_pca)
```

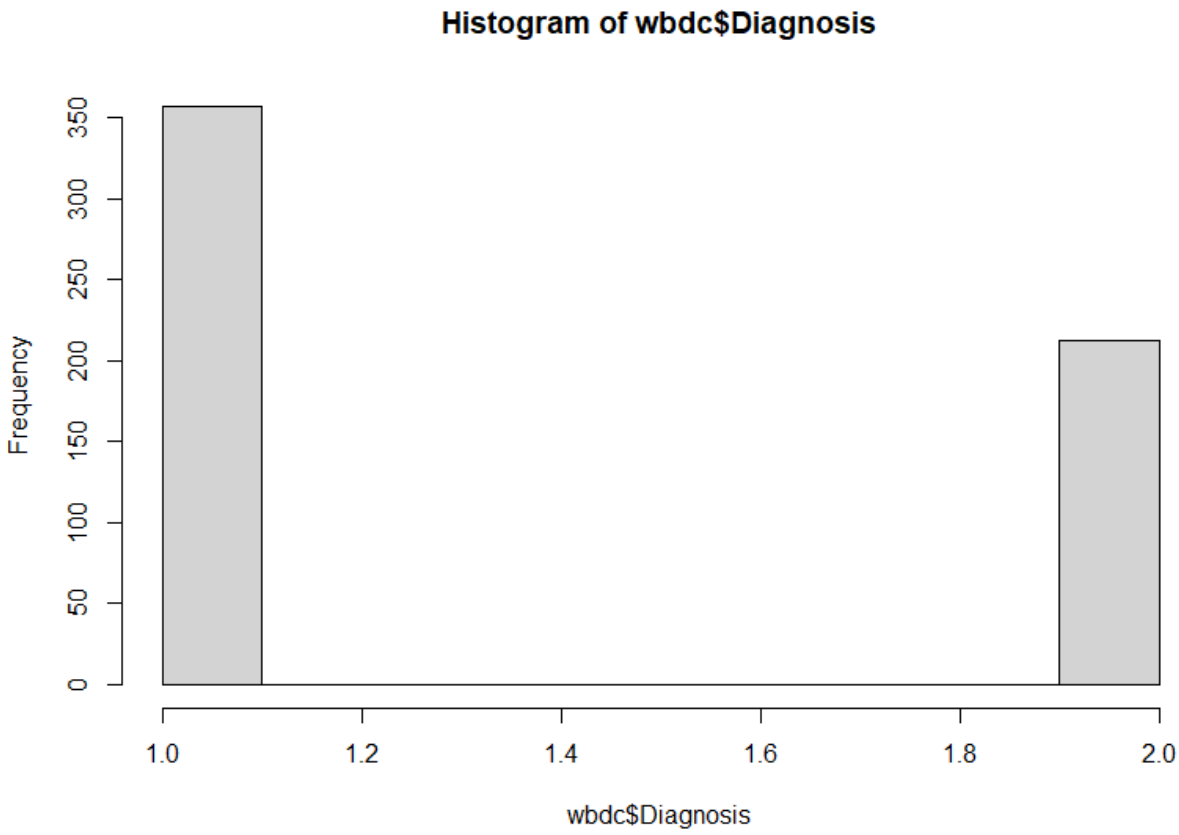
```
> kcluster
K-means clustering with 2 clusters of sizes 131, 438

Cluster means:
  Mean_Radius Mean_Texture Mean_Perimeter Mean_Area Mean_Smoothness Mean_Compactness Mean_Concavity Mean_ConcavePoints
1    19.37992     21.69458      128.23130 1185.9298       0.1012946       0.14861298     0.17693947         0.10069878
2    12.55630     18.57037       81.12347  496.0619       0.0948845       0.09109982     0.06243776         0.03343254
  Mean_Symmetry Mean_FractalDimesnsion Radius_SE Texture_SE Perimeter_SE  Area_SE Smoothness_SE Compactness_SE Concavity_SE
1     0.1915397             0.06060290 0.7428038   1.222538     5.250580 95.67817   0.006598687     0.03217669   0.04241977
2     0.1780580             0.06345402 0.3041909   1.215153     2.152881 23.78529   0.007173263     0.02347469   0.02874551
  ConcavePoints_SE Symmetry_SE FractalDimension_SE Worst_Radius Worst_Texture Worst_Perimeter Worst_Area Worst_Smoothness
1       0.01567398  0.02030397         0.003953389     23.70947      28.91267       158.49618  1753.0229         0.1404247
2       0.01063632  0.02061358         0.003747503     14.04390      24.70954        91.93751   619.6479         0.1299591
  Worst_Compactness Worst_Concavity Worst_ConcavePoints Worst_Symmetry Worst_FractalDimesnsion
1         0.3577577       0.4493061          0.19243107      0.3118817              0.08616550
2         0.2233118       0.2192149          0.09132984      0.2835537              0.08328194
```

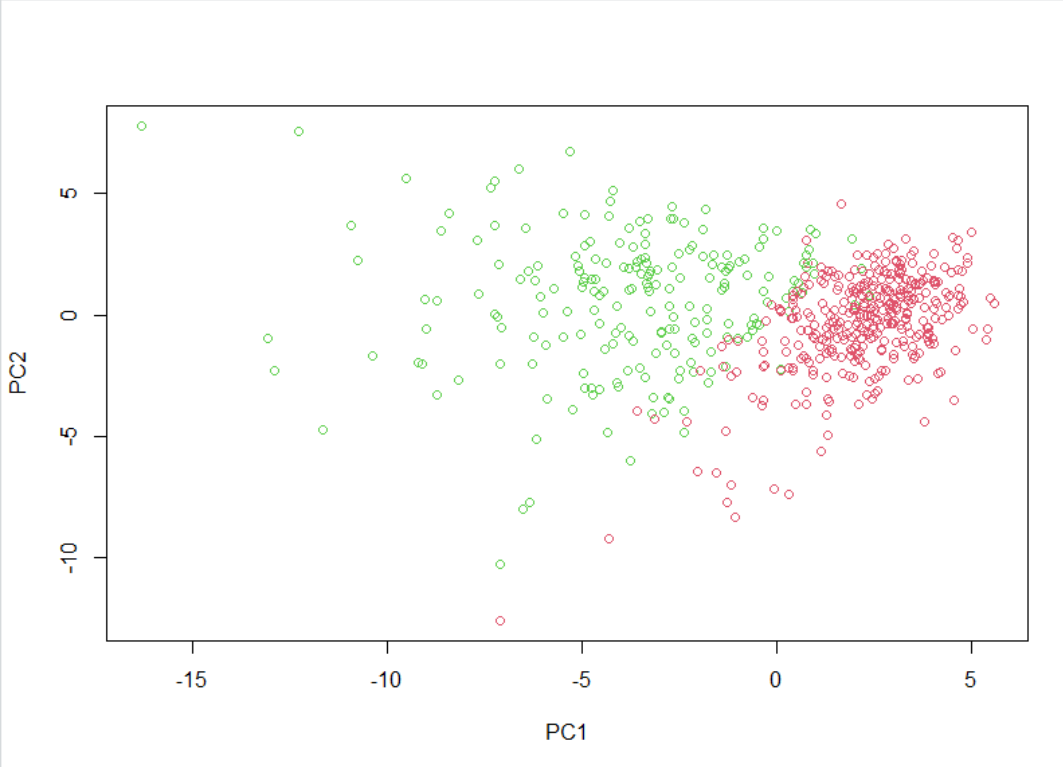
```
Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32
  1   1   1   2   1   2   1   2   2   2   2   1   1   2   2   2   2   1   1   2   2   2   2   1   1   1   2   1   1   1   1   2
 33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
  1   1   1   1   2   2   2   2   2   2   1   2   2   1   2   2   2   2   2   2   2   1   2   2   1   2   2   2   2   2   2   2
 65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96
  2   2   2   2   2   2   1   2   1   2   2   1   2   1   1   2   2   2   1   1   2   1   2   1   2   2   2   2   2   2   2   1
 97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128
  2   2   2   2   2   2   2   2   2   2   2   2   1   2   2   2   2   2   2   2   2   2   1   1   2   1   1   2   2   2   2   1
129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
  2   1   2   2   2   2   1   2   2   2   2   2   2   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1   2   2   2
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192
  2   1   1   2   1   2   2   1   1   2   2   2   2   2   2   2   2   2   2   2   2   1   1   1   2   2   2   1   2   2   2   2
193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224
  2   2   2   2   2   1   1   2   2   1   1   2   2   2   2   1   2   2   1   2   1   2   2   2   2   2   1   1   2   2   2   2
225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256
  2   2   2   2   2   2   1   2   2   1   2   2   1   1   2   1   2   2   2   2   1   2   2   2   2   2   1   2   1   1   1   2
257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
  1   2   1   2   1   1   1   2   1   1   2   2   2   2   2   2   1   2   1   2   2   1   2   2   1   2   1   2   2   2   2   2
289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320
  2   2   2   2   2   2   2   2   2   2   2   2   1   2   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1   2   2
321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352
  2   1   2   1   2   2   2   2   2   2   2   2   2   2   2   1   2   1   2   1   2   2   2   1   2   2   2   2   2   2   2   2
353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384
  1   2   2   2   2   2   2   2   2   2   2   2   2   1   1   2   1   1   2   2   1   1   2   2   2   2   2   2   2   2   2   2
385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416
  2   2   2   2   2   1   2   2   1   1   2   2   2   2   2   2   1   2   2   2   2   2   2   2   1   2   2   2   2   2   2   2
417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448
  2   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1   1   2   2   2   2   2   2   2   1   2   2   1   2   1   2
449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480
  2   1   2   1   2   2   2   2   2   2   2   2   1   1   2   2   2   2   2   2   1   2   2   2   2   2   2   2   2   2   2   2
481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512
  2   2   2   2   2   2   2   1   2   2   2   1   1   2   2   2   2   2   1   1   2   2   2   1   2   2   2   2   2   2   2   2
513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544
  2   2   2   2   1   1   2   2   2   1   2   2   2   2   2   2   2   2   2   2   2   1   2   1   2   2   2   2   2   2   2   2
545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   1   1   1   2   1   2

Within cluster sum of squares by cluster:
[1] 49383423 28559677
 (between_SS / total_SS =  69.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

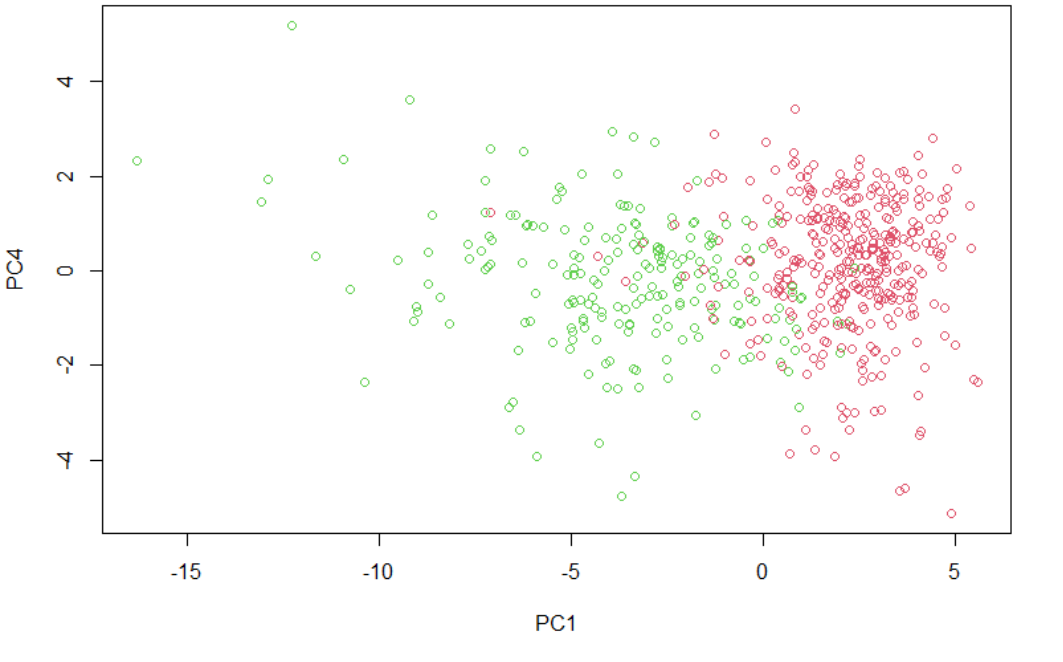
```
> hcluster <- hclust(dist(wbdc_12), method = "complete")
> cutree_hcluster <- cutree(hcluster, k=2)
> hcluster

Call:
hclust(d = dist(wbdc_12), method = "complete")

Cluster method   : complete
Distance         : euclidean
Number of objects: 569

> table(wbdc$Diagnosis,kcluster$cluster)

      1   2
  1   1 356
  2 130  82
> table(wbdc$Diagnosis,cutree_hcluster)
   cutree_hcluster
      1   2
  1 357   0
  2 192  20
>
> #principle component analysis
> wbdc_pc <- prcomp(wbdc_12, center = TRUE, scale. = TRUE)
> attributes(wbdc_pc)
$names
[1] "sdev"     "rotation" "center"   "scale"    "x"

$class
[1] "prcomp"
```

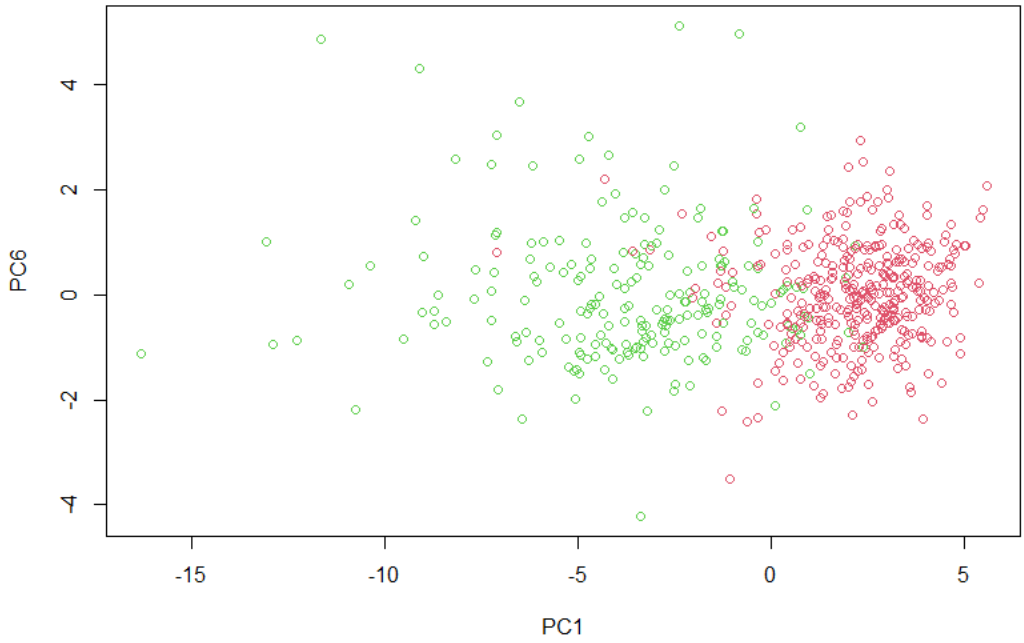
```
> kcluster_pca <- kmeans(wbdc_pc$x[,1:17],2, nstart = 125)
> table(kcluster_pca$cluster,wbdc$Diagnosis)

     1   2
  1 343  37
  2  14 175
> hcluster_pca <- hclust(dist(wbdc_pc$x[,1:17]),method = "complete")
> hcluster_clusters <- cutree(hcluster_pca, k=2 )
> table(hcluster_clusters,wbdc$Diagnosis)

hcluster_clusters   1   2
                1 357 210
                2   0   2
```

Histogram for Part 1

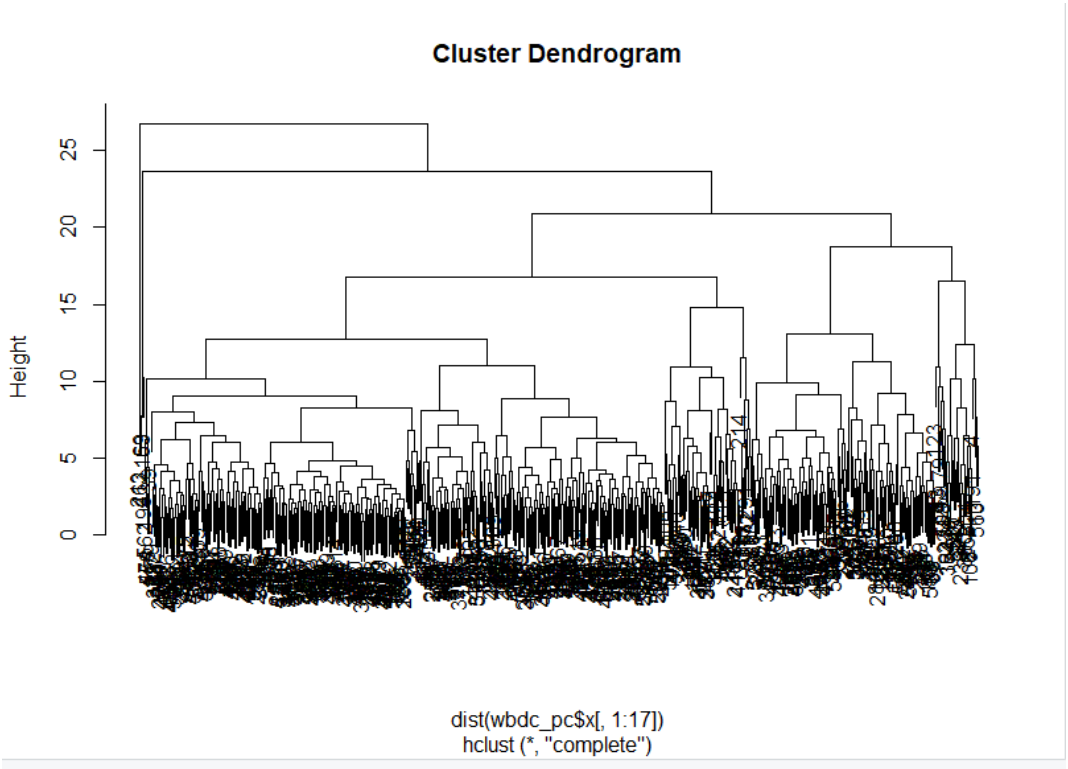



PC2
5
0
-5
-10
-15
-10
-5
0
5
PC1


PC4
4
2
0
-2
-4
-15
-10
-5
0
5
PC1



The red clusters are towards the right and green are to the left and the more close the clusters are the more significant is the PCA is towards the clusters and its influencing the factors of breast cancer.



Q2

```
#IE 500 SMLE HW 5 Q2
library(tidyverse)
library(caret)
library(e1071)
library(caTools)
library(ggplot2)
library(lattice)
library(tibble)
library(MASS)

faults <- read.table("C:/Users/ppill/Desktop/R files/Faults",header = FALSE,sep = "")
colnames(faults) <- c('X_Minimum','X_Maximum','Y_Minimum','Y_Maximum',
                      'Pixels_Area','X_Perimeter','Y_Perimeter','Sum_of_Luminosity',
                      'Minimum_of_Luminosity','Maximum_of_Luminosity','Length_of_Conveyer',
                      'TypeofSteel_A300','TypeofSteel_A400','Steel_Plate_Thickness',
                      'Edges_Index','Empty_Index','Square_Index','Outside_X_Index',
                      'Edges_X_Index','Egdes_Y_Index','Outside_Global_Index','LogOfAreas',
                      'Log_X_Index','Log_Y_Index','Orientation_Index','Luminosity_Index',
                      'SigmoidOfAreas','Pastry','Z_Scratch','K_Scratch','Stains','Dirtiness','Bumps',
                      'Other Faults')
#part 1
set.seed(1029)
#unifying classes into a vector
for(i in 28:34)
{
  for(j in 1:nrow(faults))
    if(faults[j,i]==1)
      faults[j,i] <- colnames(faults[i])
}

faults <- add_column(faults,0)
colnames(faults)[35] <- c("type")

for(i in 28:34)
{
  for(j in 1:nrow(faults))
    if(faults[j,i]!=0)
      faults[j,35]<- faults[j,i]
}

faults <- faults[,-c(28:34)]
faults[,28] <- as.factor(faults[,28])

#splitting data
split_faults <- sample.split(faults$type, SplitRatio = 0.5 )
train <- subset(faults, split = TRUE)
test <- subset(faults,split = FALSE)

svm_faults <- svm(faults$type~. , data = train)
svm_faults
summary(svm_faults)
predict_test <- predict(svm_faults, data = test)
summary(predict_test)
test_tab <- table(predict_test,faults$type)
confusionMatrix(test_tab, positive = "Yes")
```

```r
#neural network

library(nnet)

nnet_faults <- multinom(faults$type~. , data = train)
nnet_faults
summary(nnet_faults)
predict_testnn <- predict(nnet_faults, data = test)
test_nntab <- table(predict_testnn,faults$type)
confusionMatrix(test_nntab,positive = "Yes")

#random forests

library(randomForest)

rf_faults <- randomForest(faults$type~. , data = train)
rf_faults
summary(rf_faults)
predict_testrf <- predict(rf_faults,data = test)
test_rftab <- table(predict_testrf,faults$type)
confusionMatrix(test_rftab,positive = "Yes")
```

```r
#part 2
#splitting data
library(tidyverse)
library(caret)
library(e1071)
library(caTools)
library(ggplot2)
library(lattice)
library(tibble)
library(MASS)
set.seed(1029)
split_faults2 <- sample.split(faults$type, SplitRatio = 0.7 )
train2 <- subset(faults, split = TRUE)
test2 <- subset(faults,split = FALSE)

svm_faults2 <- svm(faults$type~. , data = train2)
svm_faults2
summary(svm_faults2)
predict_test2 <- predict(svm_faults2,data=test2)
test_tab2 <- table(predict_test2,faults$type)
confusionMatrix(test_tab2,positive = "Yes")


#neural network

library(nnet)

nnet_faults2 <- multinom(faults$type~. , data = train2)
nnet_faults2
summary(nnet_faults2)
predict_testnn2 <- predict(nnet_faults2, data=test)
test_nntab2 <- table(predict_testnn2,faults$type)
confusionMatrix(test_nntab2,positive = "Yes")


#random forests

library(randomForest)

rf_faults2 <- randomForest(faults$type~. , data = train2)
rf_faults2
summary(rf_faults2)
predict_testrf2 <- predict(rf_faults2,data = test2)
test_rftab2 <- table(predict_testrf2,faults$type)
confusionMatrix(test_rftab2,positive = "Yes")
```

```
> svm_faults

Call:
svm(formula = faults$type ~ ., data = train)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1241

> summary(svm_faults)

Call:
svm(formula = faults$type ~ ., data = train)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1241

 ( 146 102 86 37 49 320 501 )


Number of Classes:  7

Levels:
 Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch



> predict_test <- predict(svm_faults, data = test)
> summary(predict_test)
       Bumps     Dirtiness     K_Scratch Other Faults       Pastry       Stains     Z_Scratch
         411           51           373           711           127           72           196

> confusionMatrix(test_tab, positive = "Yes")
Confusion Matrix and Statistics


predict_test   Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
  Bumps          285         3         2           97     18      2         4
  Dirtiness        2        43         0            6      0      0         0
  K_Scratch        0         0       369            4      0      0         0
  Other Faults    95         7        20          527     40      3        19
  Pastry          11         2         0           15     99      0         0
  Stains           0         0         0            5      0     67         0
  Z_Scratch        9         0         0           19      1      0       167

Overall Statistics

               Accuracy : 0.8022
                 95% CI : (0.7837, 0.8197)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7441

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bumps Class: Dirtiness Class: K_Scratch Class: Other Faults Class: Pastry Class: Stains
Sensitivity                0.7090          0.78182           0.9437              0.7831       0.62658       0.93056
Specificity                0.9181          0.99576           0.9974              0.8549       0.98430       0.99732
Pos Pred Value             0.6934          0.84314           0.9893              0.7412       0.77953       0.93056
Neg Pred Value             0.9235          0.99365           0.9860              0.8813       0.96748       0.99732
Prevalence                 0.2071          0.02834           0.2014              0.3467       0.08140       0.03709
Detection Rate             0.1468          0.02215           0.1901              0.2715       0.05100       0.03452
Detection Prevalence       0.2117          0.02628           0.1922              0.3663       0.06543       0.03709
Balanced Accuracy          0.8135          0.88879           0.9706              0.8190       0.80544       0.96394
                     Class: Z_Scratch
Sensitivity                   0.87895
Specificity                   0.98344
Pos Pred Value                0.85204
Neg Pred Value                0.98682
Prevalence                    0.09789
Detection Rate                0.08604
Detection Prevalence          0.10098
Balanced Accuracy             0.93119
```

```
> nnet_faults <- multinom(faults$type~. , data = train)
# weights:  203 (168 variable)
initial  value 3777.011599
iter  10 value 3145.865778
iter  20 value 2789.992222
iter  30 value 2657.156367
iter  40 value 2546.955090
iter  50 value 2531.529231
iter  60 value 2409.547388
iter  70 value 2299.079861
iter  80 value 1980.349066
iter  90 value 1641.100561
iter 100 value 1503.285612
final  value 1503.285612
stopped after 100 iterations
```

```
> nnet_faults
Call:
multinom(formula = faults$type ~ ., data = train)

Coefficients:
            (Intercept)     X_Minimum     X_Maximum   Y_Minimum  Y_Maximum   Pixels_Area  X_Perimeter Y_Perimeter
Dirtiness      0.8146088  0.0126496482 -0.0121898412 -0.03323921 0.03323923  0.0008504645  0.008979494 -0.02666439
K_Scratch     -2.4297647  0.0001662720 -0.0008884680 -0.03880953 0.03880934  0.0006659034  0.008024554 -0.02081541
Other Faults  -1.1456323 -0.0087206831  0.0086932552 -0.05521092 0.05521082 -0.0003270031 -0.004615200 -0.02766587
Pastry        -0.4161188 -0.0008145865  0.0007751533 -0.04465182 0.04465184  0.0004663432  0.006816089 -0.02741656
Stains         1.8760103 -0.0006136417  0.0007059892 -0.05769128 0.05769101  0.0013340640  0.010935575 -0.04729510
Z_Scratch      5.2638483  0.0035807715 -0.0047841543 -0.04819798 0.04819751 -0.0011349346  0.002815174 -0.02423072
            Sum_of_Luminosity Minimum_of_Luminosity Maximum_of_Luminosity Length_of_Conveyer TypeofSteel_A300 TypeofSteel_A400
Dirtiness       -4.720325e-06           0.021828965          -0.001099485      -0.0008134161       -0.2330571       1.04766596
K_Scratch       -5.259492e-06           0.004251805          -0.003034929       0.0010495363       -1.9685951      -0.46116954
Other Faults     2.826981e-06           0.022697235          -0.007534332       0.0014365211       -1.1235573      -0.02207445
Pastry          -1.918754e-06          -0.014396096           0.012910807       0.0037916996       -0.8890079       0.47288962
Stains          -5.093091e-06          -0.004149393           0.022248067      -0.0008405408       -0.3581541       2.23416447
Z_Scratch        1.148049e-05           0.041026822          -0.032871795      -0.0061056433        3.4091715       1.85467734
            Steel_Plate_Thickness Edges_Index Empty_Index Square_Index Outside_X_Index Edges_X_Index Egdes_Y_Index
Dirtiness           0.0038762040  0.48123099   -0.198886  -4.09429958     -0.01094891    -2.4250466     0.6241579
K_Scratch          -0.0078357487 -1.38768182   -3.404790   0.09462122     -0.08665126     2.6289564    -1.2210349
Other Faults        0.0095290221 -0.56799792    1.556426  -1.18232778     -0.00290764     1.9168168    -3.7338685
Pastry              0.0052286067 -0.92518606   -1.998320  -3.07802291      0.03823603     1.3920406    -1.0221248
Stains             -0.0019223756 -0.01933685    4.741920  -1.06227059      0.01234804     0.9541186     0.9976447
Z_Scratch          -0.0007758232 -1.72919971    2.119733  -0.39064251      0.06114043    -2.0600541     0.8502296
            Outside_Global_Index LogOfAreas Log_X_Index Log_Y_Index Orientation_Index Luminosity_Index SigmoidOfAreas
Dirtiness           -0.88941528 -0.2593389   -1.709619  0.81350385       -0.3945058       -0.8865580     -0.2804113
K_Scratch            0.54173894  1.2753784   -2.348874  1.95136529       -2.5150895        5.6727785      0.2120718
Other Faults        -1.01766682 -0.3317425    3.718203 -2.18964011        3.5201798       -1.9090640     -1.1287589
Pastry              -1.06062790 -0.9044638   -1.191954 -1.16467873        2.7813644        0.3736122      0.4413293
Stains              -0.72855801 -4.2821954   -1.416389  0.07516077       -0.5461076       -0.9959822      2.0485879
Z_Scratch           -0.09654471  0.5140868    1.219210 -0.53569819       -0.7183470       -4.0260140     -1.7379447

Residual Deviance: 3006.571
AIC: 3330.571
```

```
> confusionMatrix(test_nntab,positive = "Yes")
Confusion Matrix and Statistics


predict_testnn Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
  Bumps          217         1         6          93     15      2         6
  Dirtiness        1        25         0           5      0      0         0
  K_Scratch        1         0       360          15      1      0         1
  Other Faults   171        25        20         510     63      4        25
  Pastry           6         4         2          30     73      0         0
  Stains           0         0         2           1      0     66         0
  Z_Scratch        6         0         1          19      6      0       158

Overall Statistics

               Accuracy : 0.7259
                 95% CI : (0.7055, 0.7457)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6404

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bumps Class: Dirtiness Class: K_Scratch Class: Other Faults Class: Pastry Class: Stains
Sensitivity                0.5398          0.45455           0.9207               0.7578       0.46203       0.91667
Specificity                0.9201          0.99682           0.9884               0.7571       0.97644       0.99839
Pos Pred Value             0.6382          0.80645           0.9524               0.6235       0.63478       0.95652
Neg Pred Value             0.8844          0.98429           0.9802               0.8549       0.95345       0.99679
Prevalence                 0.2071          0.02834           0.2014               0.3467       0.08140       0.03709
Detection Rate             0.1118          0.01288           0.1855               0.2628       0.03761       0.03400
Detection Prevalence       0.1752          0.01597           0.1947               0.4214       0.05925       0.03555
Balanced Accuracy          0.7299          0.72568           0.9546               0.7574       0.71923       0.95753
                     Class: Z_Scratch
Sensitivity                   0.83158
Specificity                   0.98172
Pos Pred Value                0.83158
Neg Pred Value                0.98172
Prevalence                    0.09789
Detection Rate                0.08140
Detection Prevalence          0.09789
Balanced Accuracy             0.90665
```

```
> rf_faults <- randomForest(faults$type~. , data = train)
> rf_faults

Call:
 randomForest(formula = faults$type ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 21.17%
Confusion matrix:
             Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch class.error
Bumps          275         1         0          112     13      0         1  0.31592040
Dirtiness        2        46         0            6      1      0         0  0.16363636
K_Scratch        1         0       371           19      0      0         0  0.05115090
Other Faults    96         4         6          529     23      3        12  0.21396731
Pastry          16         0         0           57     82      0         3  0.48101266
Stains           2         0         0            5      0     65         0  0.09722222
Z_Scratch        0         0         2           26      0      0       162  0.14736842
```

```
> confusionMatrix(test_rftab,positive = "Yes")
Confusion Matrix and Statistics


predict_testrf Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
  Bumps          275         2         1          96     16      2         0
  Dirtiness        1        46         0           4      0      0         0
  K_Scratch        0         0       371           6      0      0         2
  Other Faults   112         6        19         529     57      5        26
  Pastry          13         1         0          23     82      0         0
  Stains           0         0         0           3      0     65         0
  Z_Scratch        1         0         0          12      3      0       162

Overall Statistics

               Accuracy : 0.7883
                 95% CI : (0.7694, 0.8062)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7243

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bumps Class: Dirtiness Class: K_Scratch Class: Other Faults Class: Pastry Class: Stains
Sensitivity                0.6841          0.83636          0.9488              0.7860       0.51899       0.90278
Specificity                0.9240          0.99735          0.9948              0.8226       0.97925       0.99839
Pos Pred Value             0.7015          0.90196          0.9789              0.7016       0.68908       0.95588
Neg Pred Value             0.9180          0.99524          0.9872              0.8787       0.95829       0.99626
Prevalence                 0.2071          0.02834          0.2014              0.3467       0.08140       0.03709
Detection Rate             0.1417          0.02370          0.1911              0.2725       0.04225       0.03349
Detection Prevalence       0.2020          0.02628          0.1953              0.3885       0.06131       0.03503
Balanced Accuracy          0.8040          0.91686          0.9718              0.8043       0.74912       0.95059
                     Class: Z_Scratch
Sensitivity                   0.85263
Specificity                   0.99086
Pos Pred Value                0.91011
Neg Pred Value                0.98412
Prevalence                    0.09789
Detection Rate                0.08346
Detection Prevalence          0.09171
Balanced Accuracy             0.92175
`
```

```
> svm_faults2 <- svm(faults$type~. , data = train2)
> svm_faults2

Call:
svm(formula = faults$type ~ ., data = train2)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1241

> summary(svm_faults2)

Call:
svm(formula = faults$type ~ ., data = train2)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  1241

 ( 146 102 86 37 49 320 501 )


Number of Classes:  7

Levels:
 Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
```

```
> confusionMatrix(test_tab2,positive = "Yes")
Confusion Matrix and Statistics


predict_test2  Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
  Bumps          285         3         2           97     18      2         4
  Dirtiness        2        43         0            6      0      0         0
  K_Scratch        0         0       369            4      0      0         0
  Other Faults    95         7        20          527     40      3        19
  Pastry          11         2         0           15     99      0         0
  Stains           0         0         0            5      0     67         0
  Z_Scratch        9         0         0           19      1      0       167

Overall Statistics

               Accuracy : 0.8022
                 95% CI : (0.7837, 0.8197)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7441

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bumps Class: Dirtiness Class: K_Scratch Class: Other Faults Class: Pastry Class: Stains
Sensitivity                0.7090          0.78182           0.9437              0.7831       0.62658       0.93056
Specificity                0.9181          0.99576           0.9974              0.8549       0.98430       0.99732
Pos Pred Value             0.6934          0.84314           0.9893              0.7412       0.77953       0.93056
Neg Pred Value             0.9235          0.99365           0.9860              0.8813       0.96748       0.99732
Prevalence                 0.2071          0.02834           0.2014              0.3467       0.08140       0.03709
Detection Rate             0.1468          0.02215           0.1901              0.2715       0.05100       0.03452
Detection Prevalence       0.2117          0.02628           0.1922              0.3663       0.06543       0.03709
Balanced Accuracy          0.8135          0.88879           0.9706              0.8190       0.80544       0.96394
                     Class: Z_Scratch
Sensitivity                   0.87895
Specificity                   0.98344
Pos Pred Value                0.85204
Neg Pred Value                0.98682
Prevalence                    0.09789
Detection Rate                0.08604
Detection Prevalence          0.10098
Balanced Accuracy             0.93119
>
```

```
> nnet_faults2 <- multinom(faults$type~. , data = train2)
# weights:  203 (168 variable)
initial  value 3777.011599
iter  10 value 3145.865778
iter  20 value 2789.992222
iter  30 value 2657.156367
iter  40 value 2546.955090
iter  50 value 2531.529231
iter  60 value 2409.547388
iter  70 value 2299.079861
iter  80 value 1980.349066
iter  90 value 1641.100561
iter 100 value 1503.285612
final  value 1503.285612
stopped after 100 iterations
> nnet_faults2
Call:
multinom(formula = faults$type ~ ., data = train2)

Coefficients:
            (Intercept)     X_Minimum     X_Maximum   Y_Minimum  Y_Maximum   Pixels_Area  X_Perimeter Y_Perimeter
Dirtiness      0.8146088  0.0126496482 -0.0121898412 -0.03323921 0.03323923  0.0008504645  0.008979494 -0.02666439
K_Scratch     -2.4297647  0.0001662720 -0.0008884680 -0.03880953 0.03880934  0.0006659034  0.008024554 -0.02081541
Other Faults  -1.1456323 -0.0087206831  0.0086932552 -0.05521092 0.05521082 -0.0003270031 -0.004615200 -0.02766587
Pastry        -0.4161188 -0.0008145865  0.0007751533 -0.04465182 0.04465184  0.0004663432  0.006816089 -0.02741656
Stains         1.8760103 -0.0006136417  0.0007059892 -0.05769128 0.05769101  0.0013340640  0.010935575 -0.04729510
Z_Scratch      5.2638483  0.0035807715 -0.0047841543 -0.04819798 0.04819751 -0.0011349346  0.002815174 -0.02423072
            Sum_of_Luminosity Minimum_of_Luminosity Maximum_of_Luminosity Length_of_Conveyer TypeofSteel_A300 TypeofSteel_A400
Dirtiness       -4.720325e-06           0.021828965          -0.001099485      -0.0008134161       -0.2330571       1.04766596
K_Scratch       -5.259492e-06           0.004251805          -0.003034929       0.0010495363       -1.9685951      -0.46116954
Other Faults     2.826981e-06           0.022697235          -0.007534332       0.0014365211       -1.1235573      -0.02207445
Pastry          -1.918754e-06          -0.014396096           0.012910807       0.0037916996       -0.8890079       0.47288962
Stains          -5.093091e-06          -0.004149393           0.022248067      -0.0008405408       -0.3581541       2.23416447
Z_Scratch        1.148049e-05           0.041026822          -0.032871795      -0.0061056433        3.4091715       1.85467734
            Steel_Plate_Thickness Edges_Index Empty_Index Square_Index Outside_X_Index Edges_X_Index Egdes_Y_Index
Dirtiness             0.0038762040  0.48123099   -0.198886  -4.09429958     -0.01094891    -2.4250466     0.6241579
K_Scratch            -0.0078357487 -1.38768182   -3.404790   0.09462122     -0.08665126     2.6289564    -1.2210349
Other Faults          0.0095290221 -0.56799792    1.556426  -1.18232778     -0.00290764     1.9168168    -3.7338685
Pastry                0.0052286067 -0.92518606   -1.998320  -3.07802291      0.03823603     1.3920406    -1.0221248
Stains               -0.0019223756 -0.01933685    4.741920  -1.06227059      0.01234804     0.9541186     0.9976447
Z_Scratch            -0.0007758232 -1.72919971    2.119733  -0.39064251      0.06114043    -2.0600541     0.8502296
            Outside_Global_Index LogOfAreas Log_X_Index Log_Y_Index Orientation_Index Luminosity_Index SigmoidOfAreas
Dirtiness             -0.88941528 -0.2593389   -1.709619  0.81350385        -0.3945058       -0.8865580     -0.2804113
K_Scratch              0.54173894  1.2753784   -2.348874  1.95136529        -2.5150895        5.6727785      0.2120718
Other Faults          -1.01766682 -0.3317425    3.718203 -2.18964011         3.5201798       -1.9090640     -1.1287589
Pastry                -1.06062790 -0.9044638   -1.191954 -1.16467873         2.7813644        0.3736122      0.4413293
Stains                -0.72855801 -4.2821954   -1.416389  0.07516077        -0.5461076       -0.9959822      2.0485879
Z_Scratch             -0.09654471  0.5140868    1.219210 -0.53569819        -0.7183470       -4.0260140     -1.7379447

Residual Deviance: 3006.571
AIC: 3330.571
```

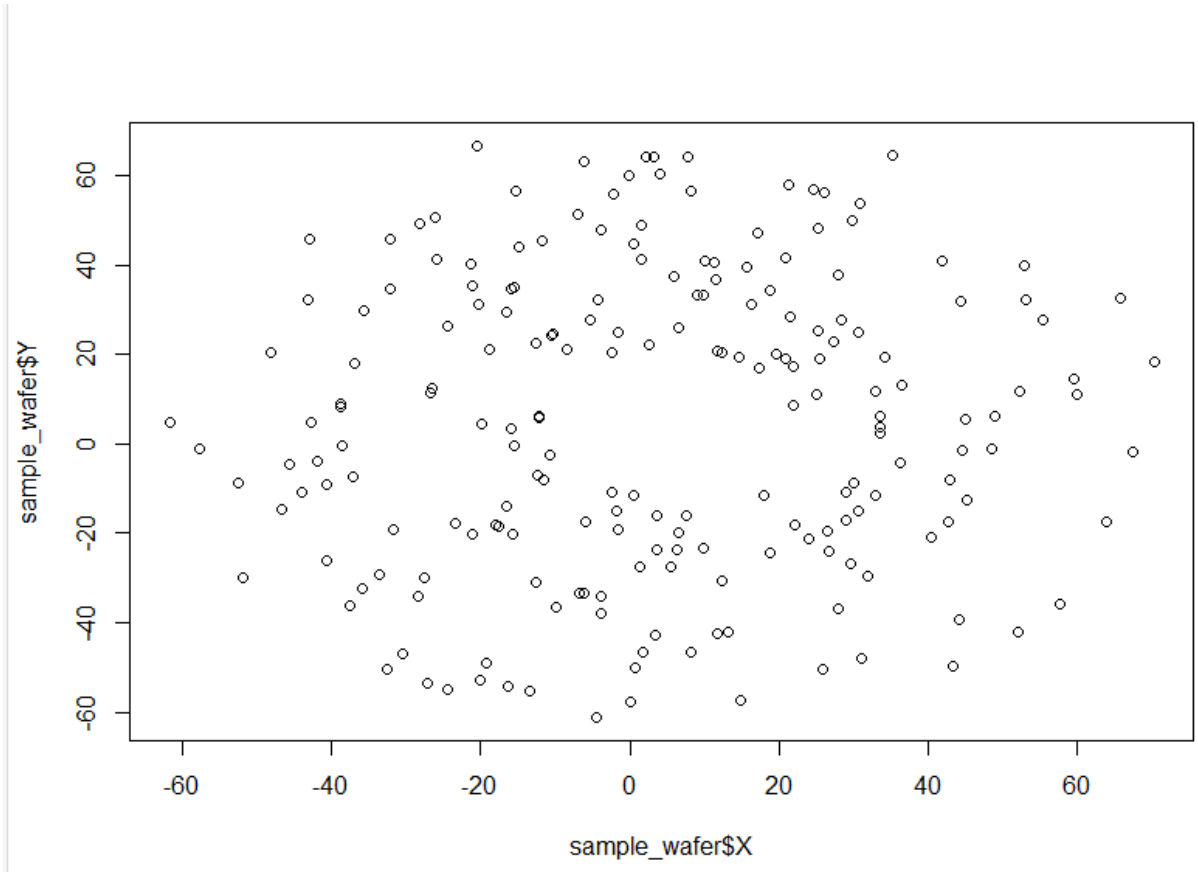
```
> confusionMatrix(test_nntab2,positive = "Yes")
Confusion Matrix and Statistics


predict_testnn2 Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
   Bumps          217         1         6           93     15      2         6
   Dirtiness        1        25         0            5      0      0         0
   K_Scratch        1         0       360           15      1      0         1
   Other Faults   171        25        20          510     63      4        25
   Pastry           6         4         2           30     73      0         0
   Stains           0         0         2            1      0     66         0
   Z_Scratch        6         0         1           19      6      0       158

Overall Statistics

               Accuracy : 0.7259
                 95% CI : (0.7055, 0.7457)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6404

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bumps Class: Dirtiness Class: K_Scratch Class: Other Faults Class: Pastry Class: Stains
Sensitivity                0.5398          0.45455           0.9207               0.7578       0.46203        0.91667
Specificity                0.9201          0.99682           0.9884               0.7571       0.97644        0.99839
Pos Pred Value             0.6382          0.80645           0.9524               0.6235       0.63478        0.95652
Neg Pred Value             0.8844          0.98429           0.9802               0.8549       0.95345        0.99679
Prevalence                 0.2071          0.02834           0.2014               0.3467       0.08140        0.03709
Detection Rate             0.1118          0.01288           0.1855               0.2628       0.03761        0.03400
Detection Prevalence       0.1752          0.01597           0.1947               0.4214       0.05925        0.03555
Balanced Accuracy          0.7299          0.72568           0.9546               0.7574       0.71923        0.95753
                     Class: Z_Scratch
Sensitivity                   0.83158
Specificity                   0.98172
Pos Pred Value                0.83158
Neg Pred Value                0.98172
Prevalence                    0.09789
Detection Rate                0.08140
Detection Prevalence          0.09789
Balanced Accuracy             0.90665
```

```
> rf_faults2 <- randomForest(faults$type~. , data = train2)
> rf_faults2

Call:
 randomForest(formula = faults$type ~ ., data = train2)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 21.17%
Confusion matrix:
             Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch class.error
Bumps          275         1         0          112     13      0         1  0.31592040
Dirtiness        2        46         0            6      1      0         0  0.16363636
K_Scratch        1         0       371           19      0      0         0  0.05115090
Other Faults    96         4         6          529     23      3        12  0.21396731
Pastry          16         0         0           57     82      0         3  0.48101266
Stains           2         0         0            5      0     65         0  0.09722222
Z_Scratch        0         0         2           26      0      0       162  0.14736842
> summary(rf_faults2)
               Length Class  Mode
call              3   -none- call
type              1   -none- character
predicted      1941   factor numeric
err.rate       4000   -none- numeric
confusion        56   -none- numeric
votes         13587   matrix numeric
oob.times      1941   -none- numeric
classes           7   -none- character
importance       27   -none- numeric
importanceSD      0   -none- NULL
localImportance   0   -none- NULL
proximity         0   -none- NULL
ntree             1   -none- numeric
mtry              1   -none- numeric
forest           14   -none- list
y              1941   factor numeric
test              0   -none- NULL
inbag             0   -none- NULL
terms             3   terms  call
> predict_testrf2 <- predict(rf_faults2,data = test2)
```

```
> confusionMatrix(test_rftab2,positive = "Yes")
Confusion Matrix and Statistics


predict_testrf2 Bumps Dirtiness K_Scratch Other Faults Pastry Stains Z_Scratch
   Bumps          275         2         1           96     16      2         0
   Dirtiness        1        46         0            4      0      0         0
   K_Scratch        0         0       371            6      0      0         2
   Other Faults   112         6        19          529     57      5        26
   Pastry          13         1         0           23     82      0         0
   Stains           0         0         0            3      0     65         0
   Z_Scratch        1         0         0           12      3      0       162

Overall Statistics

               Accuracy : 0.7883
                 95% CI : (0.7694, 0.8062)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7243

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bumps Class: Dirtiness Class: K_Scratch Class: Other Faults Class: Pastry Class: Stains
Sensitivity                0.6841          0.83636           0.9488               0.7860       0.51899       0.90278
Specificity                0.9240          0.99735           0.9948               0.8226       0.97925       0.99839
Pos Pred Value             0.7015          0.90196           0.9789               0.7016       0.68908       0.95588
Neg Pred Value             0.9180          0.99524           0.9872               0.8787       0.95829       0.99626
Prevalence                 0.2071          0.02834           0.2014               0.3467       0.08140       0.03709
Detection Rate             0.1417          0.02370           0.1911               0.2725       0.04225       0.03349
Detection Prevalence       0.2020          0.02628           0.1953               0.3885       0.06131       0.03503
Balanced Accuracy          0.8040          0.91686           0.9718               0.8043       0.74912       0.95059
                     Class: Z_Scratch
Sensitivity                   0.85263
Specificity                   0.99086
Pos Pred Value                0.91011
Neg Pred Value                0.98412
Prevalence                    0.09789
Detection Rate                0.08346
Detection Prevalence          0.09171
Balanced Accuracy             0.92175
```

Q3 Bonus

```
# IE 500 SMLE HW 5 Q3
#part 1
library(dplyr)
library(GauPro)
library(tidyverse)
library(caTools)

wafer <- read.csv("C://Users/ppill/Desktop/R files/Wafer+Data.csv",header = FALSE)
names(wafer) <- c('X','Y','T')
wafer <- na.omit(wafer)

set.seed(200)
sample_pts<- wafer[,1:2]
split_samplewafer <- sample.split(sample_pts, SplitRatio = 0.036)
train_wafer <- subset(sample_pts, split = TRUE)
test_wafer <- subset(sample_pts,split = FALSE)
plot(sample_wafer$X,sample_wafer$Y)
gp_wafer <- GauPro(train_wafer$X,train_wafer$Y)

#part 2

library(Metrics)


lm_wafer <- lm(train_wafer$Y ~train_wafer$X , data = train_wafer)
summary(lm_wafer)
predict_lm <- predict(lm_wafer, data = test_wafer)
rmse(test_wafer$Y,predict_lm)
rmse(train_wafer$Y,predict_lm)
rmse(test_wafer$X,predict_lm)
rmse(train_wafer$X,predict_lm)
```


sample_wafer$Y
60
40
20
0
-20
-40
-60
-60
-40
-20
0
20
40
60
sample_wafer$X

```
> lm_wafer <- lm(train_wafer$Y ~train_wafer$X , data = train_wafer)
> summary(lm_wafer)

Call:
lm(formula = train_wafer$Y ~ train_wafer$X, data = train_wafer)

Residuals:
    Min      1Q  Median      3Q     Max
-66.249 -23.898   0.042  24.226  63.545

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.447e+00  4.240e-01   10.49   <2e-16 ***
train_wafer$X -9.907e-12  1.319e-02    0.00        1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.21 on 5551 degrees of freedom
Multiple R-squared:  1.017e-22, Adjusted R-squared:  -0.0001801
F-statistic: 5.643e-19 on 1 and 5551 DF,  p-value: 1

> predict_lm <- predict(lm_wafer, data = test_wafer)
> rmse(test_wafer$Y,predict_lm)
[1] 31.20205
> rmse(train_wafer$Y,predict_lm)
[1] 31.20205
> rmse(test_wafer$X,predict_lm)
[1] 31.75864
> rmse(train_wafer$X,predict_lm)
[1] 31.75864
>
```