AguerosR\_Assignment3.doc

Robert Agueros

5/29/2021

Part A Numerical variable: A numerical variable is a number that holds value to it. A good example for this would be numbers used in math like getting the sum of the age of a group by adding all ages together. What is not a numerical variable would be something like a phone number. While they are numbers, they do not hold any numerical value or weight.

Categorical variable: A categorical variable consists of items like hair colors such as black, brown, blonde… Another categorical variable example would be gender of male, female, trans (male, female, or trans are the variable part of the category). Categorical variables do not have order to them.

Nominal variable: Description or labels with no order to them. Some examples are color and sex.They can be given a numerical code but does not imply any order is set with the numbers assigned. Numbers can not be calculated to produce a mean or numerical value.

Ordinal variable: Ordinal variables are similar to categorical variables, but with ordinal variables, you are able to order the variables. An example of this would be to gather student survey for satisfaction as not satisfied, satisfied, or very satisfied. We are able to order the results.

Binary variable: Describe opposites by assigning numbers such as 0 or 1. some examples of binary variables include true/false, yes/no, male/female, cold/hot, malignant/benign.

Dummy variable: A dummy variable is used to quantify a categorical variables that have no relationship such as different types of color blindness. Example: 1 = protanopia, 2 = deuteranopia, and 3 = tritanopia.

PART B Q1: Download the dataset. Includes 569 examples of cancer biopsies, each with 32 features. Q2: Load the data from your local drive to R.

Cancerdf<- read.csv("cancer\_data.csv")

Q3: Display the first five observations, summary, and structure of the data.

head(Cancerdf, n=5 )

## id diagnosis radius\_mean texture\_mean perimeter\_mean area\_mean  
## 1 87139402 Benign 12.32 12.39 78.85 464.1  
## 2 8910251 Benign 10.60 18.95 69.28 346.4  
## 3 905520 Benign 11.04 16.83 70.92 373.2  
## 4 868871 Benign 11.28 13.39 73.00 384.8  
## 5 9012568 Benign 15.19 13.21 97.65 711.8  
## smoothness\_mean compactness\_mean concavity\_mean points\_mean symmetry\_mean  
## 1 0.10280 0.06981 0.03987 0.03700 0.1959  
## 2 0.09688 0.11470 0.06387 0.02642 0.1922  
## 3 0.10770 0.07804 0.03046 0.02480 0.1714  
## 4 0.11640 0.11360 0.04635 0.04796 0.1771  
## 5 0.07963 0.06934 0.03393 0.02657 0.1721  
## dimension\_mean radius\_se texture\_se perimeter\_se area\_se smoothness\_se  
## 1 0.05955 0.2360 0.6656 1.670 17.43 0.008045  
## 2 0.06491 0.4505 1.1970 3.430 27.10 0.007470  
## 3 0.06340 0.1967 1.3870 1.342 13.54 0.005158  
## 4 0.06072 0.3384 1.3430 1.851 26.33 0.011270  
## 5 0.05544 0.1783 0.4125 1.338 17.72 0.005012  
## compactness\_se concavity\_se points\_se symmetry\_se dimension\_se radius\_worst  
## 1 0.011800 0.01683 0.012410 0.01924 0.002248 13.50  
## 2 0.035810 0.03354 0.013650 0.03504 0.003318 11.88  
## 3 0.009355 0.01056 0.007483 0.01718 0.002198 12.41  
## 4 0.034980 0.02187 0.019650 0.01580 0.003442 11.92  
## 5 0.014850 0.01551 0.009155 0.01647 0.001767 16.20  
## texture\_worst perimeter\_worst area\_worst smoothness\_worst compactness\_worst  
## 1 15.64 86.97 549.1 0.1385 0.1266  
## 2 22.94 78.28 424.8 0.1213 0.2515  
## 3 26.44 79.93 471.4 0.1369 0.1482  
## 4 15.77 76.53 434.0 0.1367 0.1822  
## 5 15.73 104.50 819.1 0.1126 0.1737  
## concavity\_worst points\_worst symmetry\_worst dimension\_worst  
## 1 0.12420 0.09391 0.2827 0.06771  
## 2 0.19160 0.07926 0.2940 0.07587  
## 3 0.10670 0.07431 0.2998 0.07881  
## 4 0.08669 0.08611 0.2102 0.06784  
## 5 0.13620 0.08178 0.2487 0.06766

summary(Cancerdf)

## id diagnosis radius\_mean texture\_mean   
## Min. : 8670 Length:569 Min. : 6.981 Min. : 9.71   
## 1st Qu.: 869218 Class :character 1st Qu.:11.700 1st Qu.:16.17   
## Median : 906024 Mode :character Median :13.370 Median :18.84   
## Mean : 30371831 Mean :14.127 Mean :19.29   
## 3rd Qu.: 8813129 3rd Qu.:15.780 3rd Qu.:21.80   
## Max. :911320502 Max. :28.110 Max. :39.28   
## perimeter\_mean area\_mean smoothness\_mean compactness\_mean   
## Min. : 43.79 Min. : 143.5 Min. :0.05263 Min. :0.01938   
## 1st Qu.: 75.17 1st Qu.: 420.3 1st Qu.:0.08637 1st Qu.:0.06492   
## Median : 86.24 Median : 551.1 Median :0.09587 Median :0.09263   
## Mean : 91.97 Mean : 654.9 Mean :0.09636 Mean :0.10434   
## 3rd Qu.:104.10 3rd Qu.: 782.7 3rd Qu.:0.10530 3rd Qu.:0.13040   
## Max. :188.50 Max. :2501.0 Max. :0.16340 Max. :0.34540   
## concavity\_mean points\_mean symmetry\_mean dimension\_mean   
## Min. :0.00000 Min. :0.00000 Min. :0.1060 Min. :0.04996   
## 1st Qu.:0.02956 1st Qu.:0.02031 1st Qu.:0.1619 1st Qu.:0.05770   
## Median :0.06154 Median :0.03350 Median :0.1792 Median :0.06154   
## Mean :0.08880 Mean :0.04892 Mean :0.1812 Mean :0.06280   
## 3rd Qu.:0.13070 3rd Qu.:0.07400 3rd Qu.:0.1957 3rd Qu.:0.06612   
## Max. :0.42680 Max. :0.20120 Max. :0.3040 Max. :0.09744   
## radius\_se texture\_se perimeter\_se area\_se   
## Min. :0.1115 Min. :0.3602 Min. : 0.757 Min. : 6.802   
## 1st Qu.:0.2324 1st Qu.:0.8339 1st Qu.: 1.606 1st Qu.: 17.850   
## Median :0.3242 Median :1.1080 Median : 2.287 Median : 24.530   
## Mean :0.4052 Mean :1.2169 Mean : 2.866 Mean : 40.337   
## 3rd Qu.:0.4789 3rd Qu.:1.4740 3rd Qu.: 3.357 3rd Qu.: 45.190   
## Max. :2.8730 Max. :4.8850 Max. :21.980 Max. :542.200   
## smoothness\_se compactness\_se concavity\_se points\_se   
## Min. :0.001713 Min. :0.002252 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.005169 1st Qu.:0.013080 1st Qu.:0.01509 1st Qu.:0.007638   
## Median :0.006380 Median :0.020450 Median :0.02589 Median :0.010930   
## Mean :0.007041 Mean :0.025478 Mean :0.03189 Mean :0.011796   
## 3rd Qu.:0.008146 3rd Qu.:0.032450 3rd Qu.:0.04205 3rd Qu.:0.014710   
## Max. :0.031130 Max. :0.135400 Max. :0.39600 Max. :0.052790   
## symmetry\_se dimension\_se radius\_worst texture\_worst   
## Min. :0.007882 Min. :0.0008948 Min. : 7.93 Min. :12.02   
## 1st Qu.:0.015160 1st Qu.:0.0022480 1st Qu.:13.01 1st Qu.:21.08   
## Median :0.018730 Median :0.0031870 Median :14.97 Median :25.41   
## Mean :0.020542 Mean :0.0037949 Mean :16.27 Mean :25.68   
## 3rd Qu.:0.023480 3rd Qu.:0.0045580 3rd Qu.:18.79 3rd Qu.:29.72   
## Max. :0.078950 Max. :0.0298400 Max. :36.04 Max. :49.54   
## perimeter\_worst area\_worst smoothness\_worst compactness\_worst  
## Min. : 50.41 Min. : 185.2 Min. :0.07117 Min. :0.02729   
## 1st Qu.: 84.11 1st Qu.: 515.3 1st Qu.:0.11660 1st Qu.:0.14720   
## Median : 97.66 Median : 686.5 Median :0.13130 Median :0.21190   
## Mean :107.26 Mean : 880.6 Mean :0.13237 Mean :0.25427   
## 3rd Qu.:125.40 3rd Qu.:1084.0 3rd Qu.:0.14600 3rd Qu.:0.33910   
## Max. :251.20 Max. :4254.0 Max. :0.22260 Max. :1.05800   
## concavity\_worst points\_worst symmetry\_worst dimension\_worst   
## Min. :0.0000 Min. :0.00000 Min. :0.1565 Min. :0.05504   
## 1st Qu.:0.1145 1st Qu.:0.06493 1st Qu.:0.2504 1st Qu.:0.07146   
## Median :0.2267 Median :0.09993 Median :0.2822 Median :0.08004   
## Mean :0.2722 Mean :0.11461 Mean :0.2901 Mean :0.08395   
## 3rd Qu.:0.3829 3rd Qu.:0.16140 3rd Qu.:0.3179 3rd Qu.:0.09208   
## Max. :1.2520 Max. :0.29100 Max. :0.6638 Max. :0.20750

str(Cancerdf)

## 'data.frame': 569 obs. of 32 variables:  
## $ id : int 87139402 8910251 905520 868871 9012568 906539 925291 87880 862989 89827 ...  
## $ diagnosis : chr "Benign" "Benign" "Benign" "Benign" ...  
## $ radius\_mean : num 12.3 10.6 11 11.3 15.2 ...  
## $ texture\_mean : num 12.4 18.9 16.8 13.4 13.2 ...  
## $ perimeter\_mean : num 78.8 69.3 70.9 73 97.7 ...  
## $ area\_mean : num 464 346 373 385 712 ...  
## $ smoothness\_mean : num 0.1028 0.0969 0.1077 0.1164 0.0796 ...  
## $ compactness\_mean : num 0.0698 0.1147 0.078 0.1136 0.0693 ...  
## $ concavity\_mean : num 0.0399 0.0639 0.0305 0.0464 0.0339 ...  
## $ points\_mean : num 0.037 0.0264 0.0248 0.048 0.0266 ...  
## $ symmetry\_mean : num 0.196 0.192 0.171 0.177 0.172 ...  
## $ dimension\_mean : num 0.0595 0.0649 0.0634 0.0607 0.0554 ...  
## $ radius\_se : num 0.236 0.451 0.197 0.338 0.178 ...  
## $ texture\_se : num 0.666 1.197 1.387 1.343 0.412 ...  
## $ perimeter\_se : num 1.67 3.43 1.34 1.85 1.34 ...  
## $ area\_se : num 17.4 27.1 13.5 26.3 17.7 ...  
## $ smoothness\_se : num 0.00805 0.00747 0.00516 0.01127 0.00501 ...  
## $ compactness\_se : num 0.0118 0.03581 0.00936 0.03498 0.01485 ...  
## $ concavity\_se : num 0.0168 0.0335 0.0106 0.0219 0.0155 ...  
## $ points\_se : num 0.01241 0.01365 0.00748 0.01965 0.00915 ...  
## $ symmetry\_se : num 0.0192 0.035 0.0172 0.0158 0.0165 ...  
## $ dimension\_se : num 0.00225 0.00332 0.0022 0.00344 0.00177 ...  
## $ radius\_worst : num 13.5 11.9 12.4 11.9 16.2 ...  
## $ texture\_worst : num 15.6 22.9 26.4 15.8 15.7 ...  
## $ perimeter\_worst : num 87 78.3 79.9 76.5 104.5 ...  
## $ area\_worst : num 549 425 471 434 819 ...  
## $ smoothness\_worst : num 0.139 0.121 0.137 0.137 0.113 ...  
## $ compactness\_worst: num 0.127 0.252 0.148 0.182 0.174 ...  
## $ concavity\_worst : num 0.1242 0.1916 0.1067 0.0867 0.1362 ...  
## $ points\_worst : num 0.0939 0.0793 0.0743 0.0861 0.0818 ...  
## $ symmetry\_worst : num 0.283 0.294 0.3 0.21 0.249 ...  
## $ dimension\_worst : num 0.0677 0.0759 0.0788 0.0678 0.0677 ...

Q4: Delete the ID columns since it does not provide useful information for prediction. Review 7.3.2, pg 272

Cancerdf <- Cancerdf[,-1]  
str(Cancerdf)

## 'data.frame': 569 obs. of 31 variables:  
## $ diagnosis : chr "Benign" "Benign" "Benign" "Benign" ...  
## $ radius\_mean : num 12.3 10.6 11 11.3 15.2 ...  
## $ texture\_mean : num 12.4 18.9 16.8 13.4 13.2 ...  
## $ perimeter\_mean : num 78.8 69.3 70.9 73 97.7 ...  
## $ area\_mean : num 464 346 373 385 712 ...  
## $ smoothness\_mean : num 0.1028 0.0969 0.1077 0.1164 0.0796 ...  
## $ compactness\_mean : num 0.0698 0.1147 0.078 0.1136 0.0693 ...  
## $ concavity\_mean : num 0.0399 0.0639 0.0305 0.0464 0.0339 ...  
## $ points\_mean : num 0.037 0.0264 0.0248 0.048 0.0266 ...  
## $ symmetry\_mean : num 0.196 0.192 0.171 0.177 0.172 ...  
## $ dimension\_mean : num 0.0595 0.0649 0.0634 0.0607 0.0554 ...  
## $ radius\_se : num 0.236 0.451 0.197 0.338 0.178 ...  
## $ texture\_se : num 0.666 1.197 1.387 1.343 0.412 ...  
## $ perimeter\_se : num 1.67 3.43 1.34 1.85 1.34 ...  
## $ area\_se : num 17.4 27.1 13.5 26.3 17.7 ...  
## $ smoothness\_se : num 0.00805 0.00747 0.00516 0.01127 0.00501 ...  
## $ compactness\_se : num 0.0118 0.03581 0.00936 0.03498 0.01485 ...  
## $ concavity\_se : num 0.0168 0.0335 0.0106 0.0219 0.0155 ...  
## $ points\_se : num 0.01241 0.01365 0.00748 0.01965 0.00915 ...  
## $ symmetry\_se : num 0.0192 0.035 0.0172 0.0158 0.0165 ...  
## $ dimension\_se : num 0.00225 0.00332 0.0022 0.00344 0.00177 ...  
## $ radius\_worst : num 13.5 11.9 12.4 11.9 16.2 ...  
## $ texture\_worst : num 15.6 22.9 26.4 15.8 15.7 ...  
## $ perimeter\_worst : num 87 78.3 79.9 76.5 104.5 ...  
## $ area\_worst : num 549 425 471 434 819 ...  
## $ smoothness\_worst : num 0.139 0.121 0.137 0.137 0.113 ...  
## $ compactness\_worst: num 0.127 0.252 0.148 0.182 0.174 ...  
## $ concavity\_worst : num 0.1242 0.1916 0.1067 0.0867 0.1362 ...  
## $ points\_worst : num 0.0939 0.0793 0.0743 0.0861 0.0818 ...  
## $ symmetry\_worst : num 0.283 0.294 0.3 0.21 0.249 ...  
## $ dimension\_worst : num 0.0677 0.0759 0.0788 0.0678 0.0677 ...

Q5: Normalize all predictor variables. (We use diagnosis as target/class variable and others as predictor variables) review 7.3.3, pg 273

normalize<- function(x){  
 return((x-min(x))/(max(x)-min(x)))  
}  
  
#some test examples  
normalize(c(1,2,3,4,5))

## [1] 0.00 0.25 0.50 0.75 1.00

normalize(c(1,3,6,7,20))

## [1] 0.0000000 0.1052632 0.2631579 0.3157895 1.0000000

Cancerdf\_n<-as.data.frame(lapply(Cancerdf[2:31], normalize))

Q6: Display summary again to see if all predictors’ values are in the range of 0 to 1.

summary(Cancerdf\_n)

## radius\_mean texture\_mean perimeter\_mean area\_mean   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2233 1st Qu.:0.2185 1st Qu.:0.2168 1st Qu.:0.1174   
## Median :0.3024 Median :0.3088 Median :0.2933 Median :0.1729   
## Mean :0.3382 Mean :0.3240 Mean :0.3329 Mean :0.2169   
## 3rd Qu.:0.4164 3rd Qu.:0.4089 3rd Qu.:0.4168 3rd Qu.:0.2711   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## smoothness\_mean compactness\_mean concavity\_mean points\_mean   
## Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.3046 1st Qu.:0.1397 1st Qu.:0.06926 1st Qu.:0.1009   
## Median :0.3904 Median :0.2247 Median :0.14419 Median :0.1665   
## Mean :0.3948 Mean :0.2606 Mean :0.20806 Mean :0.2431   
## 3rd Qu.:0.4755 3rd Qu.:0.3405 3rd Qu.:0.30623 3rd Qu.:0.3678   
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## symmetry\_mean dimension\_mean radius\_se texture\_se   
## Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.2823 1st Qu.:0.1630 1st Qu.:0.04378 1st Qu.:0.1047   
## Median :0.3697 Median :0.2439 Median :0.07702 Median :0.1653   
## Mean :0.3796 Mean :0.2704 Mean :0.10635 Mean :0.1893   
## 3rd Qu.:0.4530 3rd Qu.:0.3404 3rd Qu.:0.13304 3rd Qu.:0.2462   
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## perimeter\_se area\_se smoothness\_se compactness\_se   
## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.04000 1st Qu.:0.02064 1st Qu.:0.1175 1st Qu.:0.08132   
## Median :0.07209 Median :0.03311 Median :0.1586 Median :0.13667   
## Mean :0.09938 Mean :0.06264 Mean :0.1811 Mean :0.17444   
## 3rd Qu.:0.12251 3rd Qu.:0.07170 3rd Qu.:0.2187 3rd Qu.:0.22680   
## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.00000   
## concavity\_se points\_se symmetry\_se dimension\_se   
## Min. :0.00000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.03811 1st Qu.:0.1447 1st Qu.:0.1024 1st Qu.:0.04675   
## Median :0.06538 Median :0.2070 Median :0.1526 Median :0.07919   
## Mean :0.08054 Mean :0.2235 Mean :0.1781 Mean :0.10019   
## 3rd Qu.:0.10619 3rd Qu.:0.2787 3rd Qu.:0.2195 3rd Qu.:0.12656   
## Max. :1.00000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## radius\_worst texture\_worst perimeter\_worst area\_worst   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.1807 1st Qu.:0.2415 1st Qu.:0.1678 1st Qu.:0.08113   
## Median :0.2504 Median :0.3569 Median :0.2353 Median :0.12321   
## Mean :0.2967 Mean :0.3640 Mean :0.2831 Mean :0.17091   
## 3rd Qu.:0.3863 3rd Qu.:0.4717 3rd Qu.:0.3735 3rd Qu.:0.22090   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## smoothness\_worst compactness\_worst concavity\_worst points\_worst   
## Min. :0.0000 Min. :0.0000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.3000 1st Qu.:0.1163 1st Qu.:0.09145 1st Qu.:0.2231   
## Median :0.3971 Median :0.1791 Median :0.18107 Median :0.3434   
## Mean :0.4041 Mean :0.2202 Mean :0.21740 Mean :0.3938   
## 3rd Qu.:0.4942 3rd Qu.:0.3025 3rd Qu.:0.30583 3rd Qu.:0.5546   
## Max. :1.0000 Max. :1.0000 Max. :1.00000 Max. :1.0000   
## symmetry\_worst dimension\_worst   
## Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.1851 1st Qu.:0.1077   
## Median :0.2478 Median :0.1640   
## Mean :0.2633 Mean :0.1896   
## 3rd Qu.:0.3182 3rd Qu.:0.2429   
## Max. :1.0000 Max. :1.0000

Q7: Randomly split the data into training and testing sets, with 80% observations in the training set and 20% in the testing set.

set.seed(35)  
cancer\_train <-sample(nrow(Cancerdf\_n), floor(nrow(Cancerdf\_n)\*0.8))  
cancer\_train

## [1] 392 135 11 270 146 226 288 201 106 377 539 526 6 157 303 420 138 354  
## [19] 203 170 337 433 439 275 481 295 345 498 347 262 446 562 4 535 208 491  
## [37] 205 386 357 257 397 512 332 550 387 451 309 307 139 506 109 181 2 165  
## [55] 21 188 261 371 18 521 30 487 375 546 513 340 20 435 202 391 554 455  
## [73] 565 44 553 476 118 428 382 360 271 71 211 37 103 115 536 127 310 246  
## [91] 12 445 335 282 85 399 136 91 105 96 84 484 89 95 417 434 426 183  
## [109] 390 29 244 92 75 424 283 567 232 147 467 400 381 475 256 34 26 220  
## [127] 421 166 229 123 427 225 320 349 209 40 306 1 228 493 235 524 438 514  
## [145] 548 292 171 133 293 114 511 31 213 273 63 113 560 184 566 56 489 414  
## [163] 32 552 79 551 108 199 74 43 247 346 516 505 168 441 474 237 444 8  
## [181] 210 94 221 284 507 464 48 15 253 329 152 255 42 264 218 87 66 67  
## [199] 429 119 568 291 88 175 485 443 305 39 141 180 25 51 215 425 131 564  
## [217] 457 50 100 174 344 10 22 482 470 300 276 179 90 458 64 73 469 558  
## [235] 406 352 263 198 68 148 53 185 413 448 430 46 72 163 372 527 557 140  
## [253] 13 162 327 379 395 126 311 301 158 373 83 364 207 374 182 134 16 49  
## [271] 460 154 563 351 3 81 447 219 160 330 289 530 54 370 28 110 384 358  
## [289] 398 323 559 266 477 394 47 436 410 407 517 78 137 355 412 144 500 549  
## [307] 169 227 97 196 52 338 14 463 501 116 223 314 396 59 403 191 216 60  
## [325] 538 402 239 537 38 389 258 490 416 252 452 556 478 190 155 488 93 348  
## [343] 453 353 366 333 231 432 299 312 269 541 294 238 473 45 486 128 279 130  
## [361] 378 367 480 142 17 177 224 35 7 62 5 324 173 523 359 117 145 308  
## [379] 401 472 150 422 318 356 206 249 195 466 156 362 509 143 459 245 322 259  
## [397] 492 510 194 502 334 561 325 408 254 461 153 532 161 57 99 278 431 121  
## [415] 268 454 102 369 204 385 339 405 415 120 350 149 222 107 494 465 547 418  
## [433] 462 189 70 230 172 200 315 112 33 260 419 520 19 555 298 82 531 125  
## [451] 178 187 58 285 519

Q8: Create four data frames for (1) training set predictors, (2) testing set predictors, (3) training set class variable, and (4) testing set class variable. The firs two only have predictor variables and the last two only have class variables.

C\_train.df<- Cancerdf\_n[cancer\_train,]  
C\_test.df<- Cancerdf\_n[-cancer\_train,]  
Cancer\_train\_labels <- Cancerdf[cancer\_train, 1]  
Cancer\_test\_labels <- Cancerdf[-cancer\_train, 1]

Q9: Using KNN algorithm with k=22 to predict class variable labels for the training set.

library(class)  
cancer\_test\_pred<- knn(train = C\_train.df, test = C\_test.df, cl = Cancer\_train\_labels, k=22)  
cancer\_test\_pred

## [1] Benign Benign Benign Benign Benign Malignant Malignant  
## [8] Malignant Malignant Benign Benign Malignant Malignant Benign   
## [15] Benign Benign Malignant Benign Malignant Malignant Malignant  
## [22] Benign Benign Benign Benign Benign Benign Benign   
## [29] Benign Benign Malignant Malignant Benign Benign Benign   
## [36] Malignant Benign Benign Benign Benign Benign Benign   
## [43] Benign Benign Malignant Benign Malignant Benign Benign   
## [50] Benign Malignant Benign Malignant Benign Malignant Malignant  
## [57] Benign Malignant Benign Benign Malignant Benign Benign   
## [64] Malignant Malignant Benign Benign Benign Malignant Malignant  
## [71] Malignant Benign Benign Benign Benign Benign Malignant  
## [78] Benign Benign Malignant Malignant Benign Benign Malignant  
## [85] Benign Benign Benign Malignant Benign Malignant Benign   
## [92] Benign Malignant Benign Benign Benign Malignant Malignant  
## [99] Benign Benign Malignant Malignant Benign Benign Malignant  
## [106] Benign Benign Malignant Benign Benign Benign Malignant  
## [113] Malignant Malignant  
## Levels: Benign Malignant

Q10: Show a confusion matrix to evaluate the performance of the model.

library(gmodels)  
table <- CrossTable(x=Cancer\_test\_labels, y=cancer\_test\_pred, prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 114   
##   
##   
## | cancer\_test\_pred   
## Cancer\_test\_labels | Benign | Malignant | Row Total |   
## -------------------|-----------|-----------|-----------|  
## Benign | 71 | 0 | 71 |   
## | 1.000 | 0.000 | 0.623 |   
## | 0.986 | 0.000 | |   
## | 0.623 | 0.000 | |   
## -------------------|-----------|-----------|-----------|  
## Malignant | 1 | 42 | 43 |   
## | 0.023 | 0.977 | 0.377 |   
## | 0.014 | 1.000 | |   
## | 0.009 | 0.368 | |   
## -------------------|-----------|-----------|-----------|  
## Column Total | 72 | 42 | 114 |   
## | 0.632 | 0.368 | |   
## -------------------|-----------|-----------|-----------|  
##   
##

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

confusionMatrix(as.factor(Cancer\_test\_labels), as.factor(cancer\_test\_pred))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Benign Malignant  
## Benign 71 0  
## Malignant 1 42  
##   
## Accuracy : 0.9912   
## 95% CI : (0.9521, 0.9998)  
## No Information Rate : 0.6316   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.9812   
##   
## Mcnemar's Test P-Value : 1   
##   
## Sensitivity : 0.9861   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 0.9767   
## Prevalence : 0.6316   
## Detection Rate : 0.6228   
## Detection Prevalence : 0.6228   
## Balanced Accuracy : 0.9931   
##   
## 'Positive' Class : Benign   
##