

MVA3.r

Yshah

Fri Feb 15 14:12:58 2019

```
wisc_bc_df <- read.csv("C://Users//Yshah//Downloads//Rutgers Sem 2//MVA//wisc_bc_data.csv")  
head(wisc_bc_df)
```

```
##      id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1 87139402      B      12.32      12.39      78.85      464.1
## 2 8910251      B      10.60      18.95      69.28      346.4
## 3 905520      B      11.04      16.83      70.92      373.2
## 4 868871      B      11.28      13.39      73.00      384.8
## 5 9012568      B      15.19      13.21      97.65      711.8
## 6 906539      B      11.57      19.04      74.20      409.7
## smoothness_mean compactness_mean concavity_mean points_mean
## 1      0.10280      0.06981      0.03987      0.03700
## 2      0.09688      0.11470      0.06387      0.02642
## 3      0.10770      0.07804      0.03046      0.02480
## 4      0.11640      0.11360      0.04635      0.04796
## 5      0.07963      0.06934      0.03393      0.02657
## 6      0.08546      0.07722      0.05485      0.01428
## symmetry_mean dimension_mean radius_se texture_se perimeter_se area_se
## 1      0.1959      0.05955      0.2360      0.6656      1.670      17.43
## 2      0.1922      0.06491      0.4505      1.1970      3.430      27.10
## 3      0.1714      0.06340      0.1967      1.3870      1.342      13.54
## 4      0.1771      0.06072      0.3384      1.3430      1.851      26.33
## 5      0.1721      0.05544      0.1783      0.4125      1.338      17.72
## 6      0.2031      0.06267      0.2864      1.4400      2.206      20.30
## smoothness_se compactness_se concavity_se points_se symmetry_se
## 1      0.008045      0.011800      0.01683      0.012410      0.01924
## 2      0.007470      0.035810      0.03354      0.013650      0.03504
## 3      0.005158      0.009355      0.01056      0.007483      0.01718
## 4      0.011270      0.034980      0.02187      0.019650      0.01580
## 5      0.005012      0.014850      0.01551      0.009155      0.01647
## 6      0.007278      0.020470      0.04447      0.008799      0.01868
## dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1      0.002248      13.50      15.64      86.97      549.1
## 2      0.003318      11.88      22.94      78.28      424.8
## 3      0.002198      12.41      26.44      79.93      471.4
## 4      0.003442      11.92      15.77      76.53      434.0
## 5      0.001767      16.20      15.73      104.50      819.1
## 6      0.003339      13.07      26.98      86.43      520.5
## smoothness_worst compactness_worst concavity_worst points_worst
## 1      0.1385      0.1266      0.12420      0.09391
## 2      0.1213      0.2515      0.19160      0.07926
## 3      0.1369      0.1482      0.10670      0.07431
## 4      0.1367      0.1822      0.08669      0.08611
## 5      0.1126      0.1737      0.13620      0.08178
## 6      0.1249      0.1937      0.25600      0.06664
## symmetry_worst dimension_worst
## 1      0.2827      0.06771
## 2      0.2940      0.07587
## 3      0.2998      0.07881
## 4      0.2102      0.06784
## 5      0.2487      0.06766
## 6      0.3035      0.08284
```

```
#Renmaing the dataset
cancer<-wisc_bc_df
library("ggplot2")
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```
library("reshape")
```

```
#number of obsevation and data type
str(cancer)
```

```
## 'data.frame': 569 obs. of 32 variables:
## $ id : int 87139402 8910251 905520 868871 9012568 906539 925291 87880 862989
89827 ...
## $ diagnosis : Factor w/ 2 levels "B","M": 1 1 1 1 1 1 1 2 1 1 ...
## $ 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 ...
```

```
#gives you the summary of the dataset
summary(cancer)
```

```

##          id          diagnosis radius_mean    texture_mean
## Min.      :    8670    B:357      Min.      : 6.981    Min.      : 9.71
## 1st Qu.:   869218    M:212      1st Qu.:11.700    1st Qu.:16.17
## Median :    906024                      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
## Min.      :0.001713    Min.      :0.002252    Min.      :0.00000
## 1st Qu.:0.005169    1st Qu.:0.013080    1st Qu.:0.01509
## Median :0.006380    Median :0.020450    Median :0.02589
## Mean      :0.007041    Mean      :0.025478    Mean      :0.03189
## 3rd Qu.:0.008146    3rd Qu.:0.032450    3rd Qu.:0.04205
## Max.      :0.031130    Max.      :0.135400    Max.      :0.39600
## points_se    symmetry_se    dimension_se    radius_worst
## Min.      :0.000000    Min.      :0.007882    Min.      :0.0008948    Min.      : 7.93
## 1st Qu.:0.007638    1st Qu.:0.015160    1st Qu.:0.0022480    1st Qu.:13.01
## Median :0.010930    Median :0.018730    Median :0.0031870    Median :14.97
## Mean      :0.011796    Mean      :0.020542    Mean      :0.0037949    Mean      :16.27
## 3rd Qu.:0.014710    3rd Qu.:0.023480    3rd Qu.:0.0045580    3rd Qu.:18.79
## Max.      :0.052790    Max.      :0.078950    Max.      :0.0298400    Max.      :36.04
## texture_worst    perimeter_worst    area_worst    smoothness_worst
## Min.      :12.02    Min.      : 50.41    Min.      : 185.2    Min.      :0.07117
## 1st Qu.:21.08    1st Qu.: 84.11    1st Qu.: 515.3    1st Qu.:0.11660
## Median :25.41    Median : 97.66    Median : 686.5    Median :0.13130
## Mean      :25.68    Mean      :107.26    Mean      : 880.6    Mean      :0.13237
## 3rd Qu.:29.72    3rd Qu.:125.40    3rd Qu.:1084.0    3rd Qu.:0.14600
## Max.      :49.54    Max.      :251.20    Max.      :4254.0    Max.      :0.22260
## compactness_worst    concavity_worst    points_worst    symmetry_worst
## Min.      :0.02729    Min.      :0.0000    Min.      :0.00000    Min.      :0.1565
## 1st Qu.:0.14720    1st Qu.:0.1145    1st Qu.:0.06493    1st Qu.:0.2504
## Median :0.21190    Median :0.2267    Median :0.09993    Median :0.2822

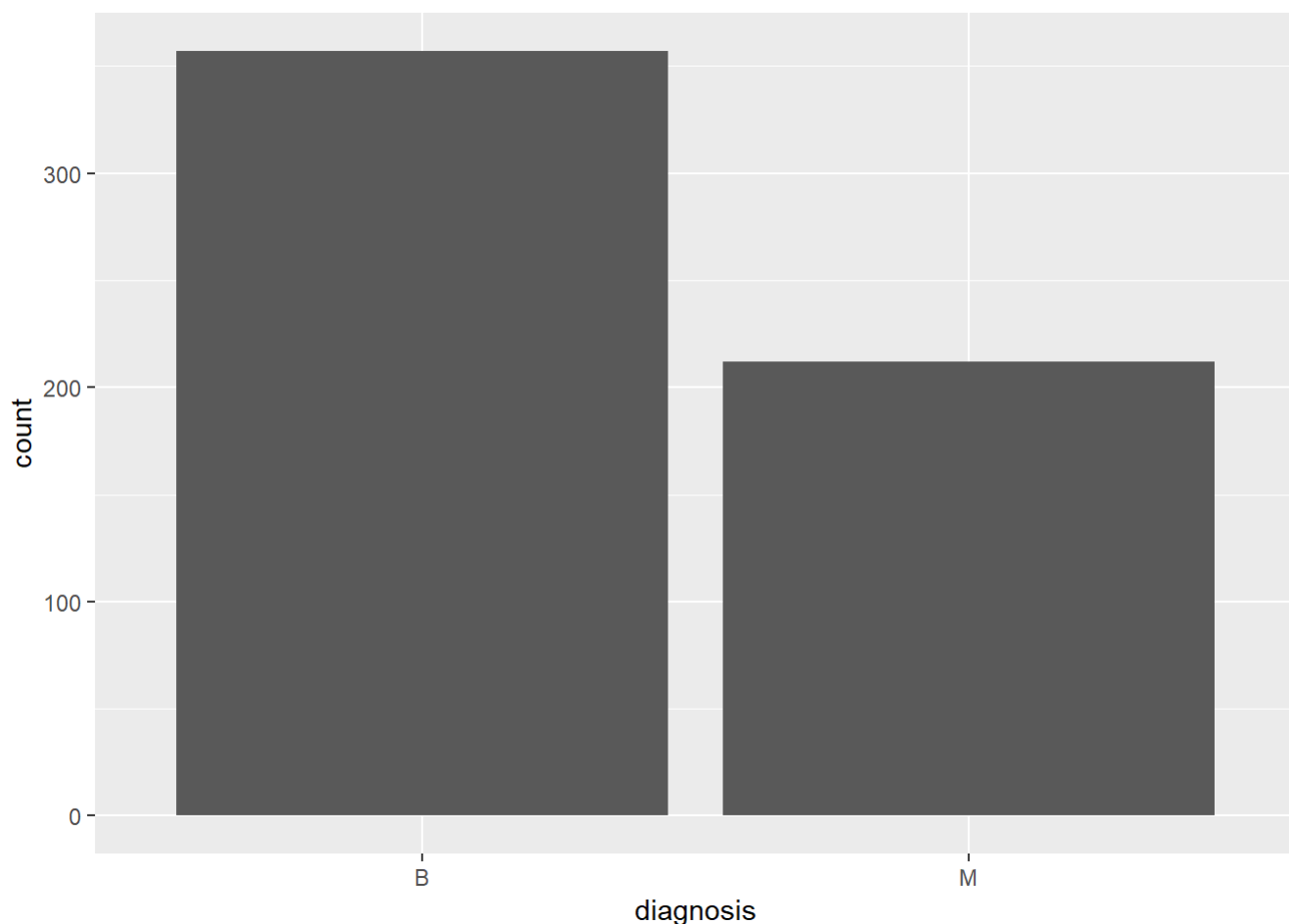
```

```
## Mean :0.25427 Mean :0.2722 Mean :0.11461 Mean :0.2901
## 3rd Qu.:0.33910 3rd Qu.:0.3829 3rd Qu.:0.16140 3rd Qu.:0.3179
## Max. :1.05800 Max. :1.2520 Max. :0.29100 Max. :0.6638
## dimension_worst
## Min. :0.05504
## 1st Qu.:0.07146
## Median :0.08004
## Mean :0.08395
## 3rd Qu.:0.09208
## Max. :0.20750
```

```
#Gives you frequency table
diagnosis.table <- table(cancer$diagnosis)
diagnosis.table
```

```
##
## B M
## 357 212
```

```
#Bar plot
ggplot(data=cancer, aes(x=diagnosis)) + geom_bar(stat = "count")
```

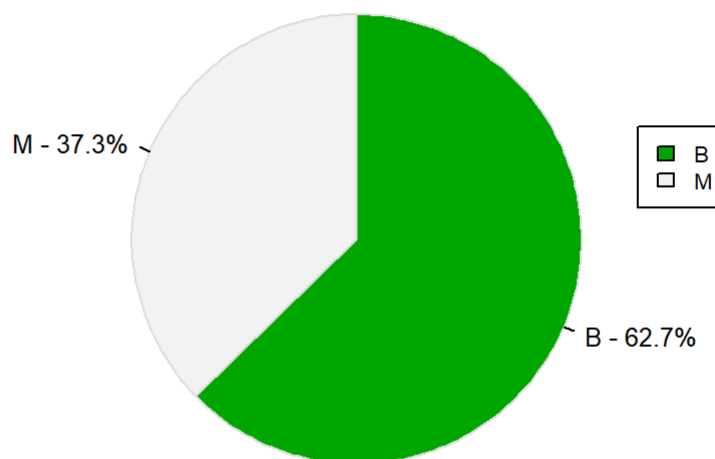


```

#Pie chart represented in frequency
diagnosis.prop.table <- prop.table(diagnosis.table)*100
diagnosis.prop.df <- as.data.frame(diagnosis.prop.table)
pielabels <- sprintf("%s - %3.1f%s", diagnosis.prop.df[,1], diagnosis.prop.table, "%")
colors <- terrain.colors(2)
pie(diagnosis.prop.table,
    labels=pielabels,
    clockwise=TRUE,
    col=colors,
    border="gainsboro",
    radius=0.8,
    cex=0.8,
    main="frequency of cancer diagnosis")
legend(1, .4, legend=diagnosis.prop.df[,1], cex = 0.7, fill = colors)

```

frequency of cancer diagnosis

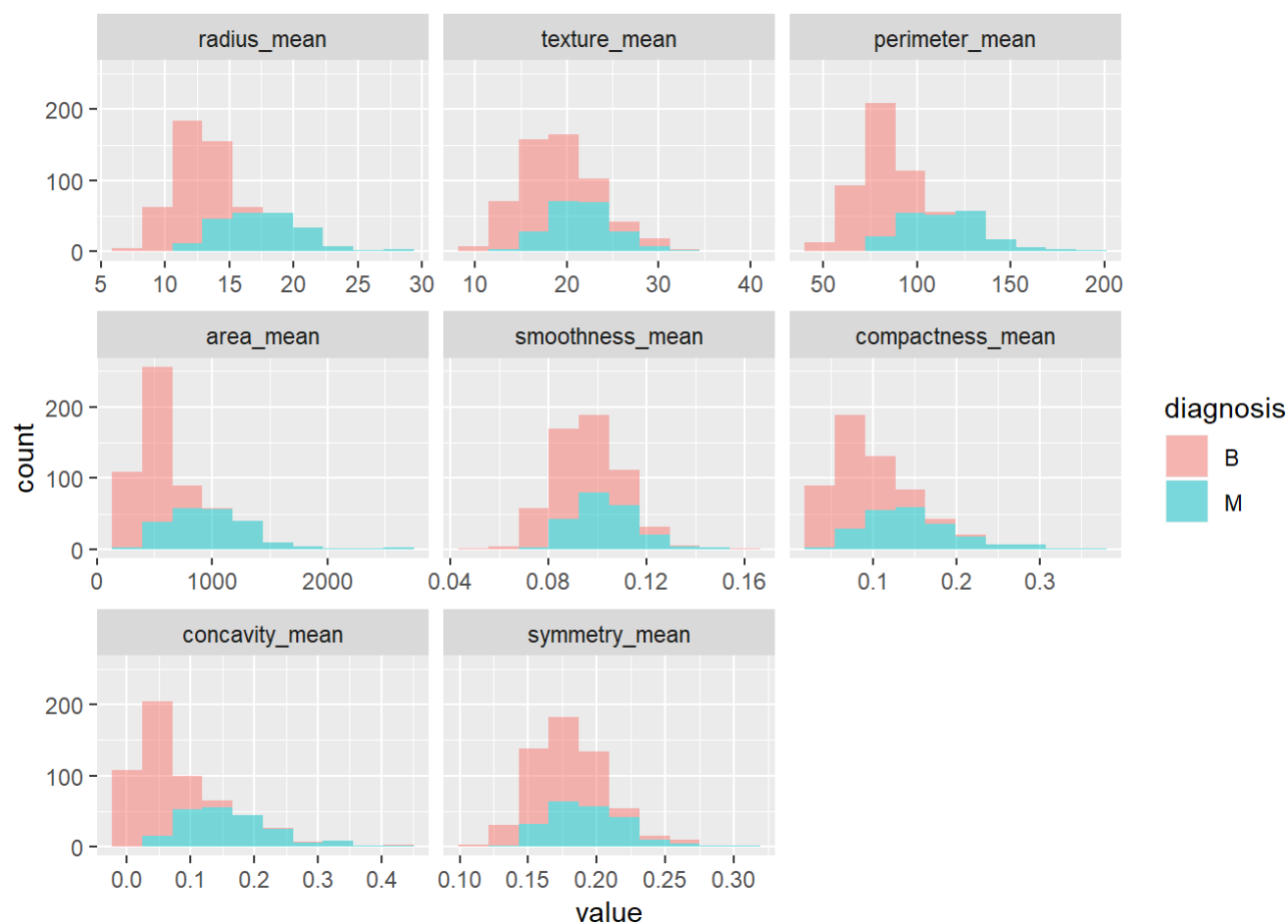


```

#Plot histograms of "mean" variables group by diagnosis
data_mean <- cancer[ ,c("diagnosis", "radius_mean", "texture_mean", "perimeter_mean", "area_mean",
    "smoothness_mean", "compactness_mean", "concavity_mean", "symmetry_mean" )]

#Plot histograms
ggplot(data = melt(data_mean, id.var = "diagnosis"), mapping = aes(x = value)) +
    geom_histogram(bins = 10, aes(fill=diagnosis), alpha=0.5) + facet_wrap(~variable, scales = 'free_x')

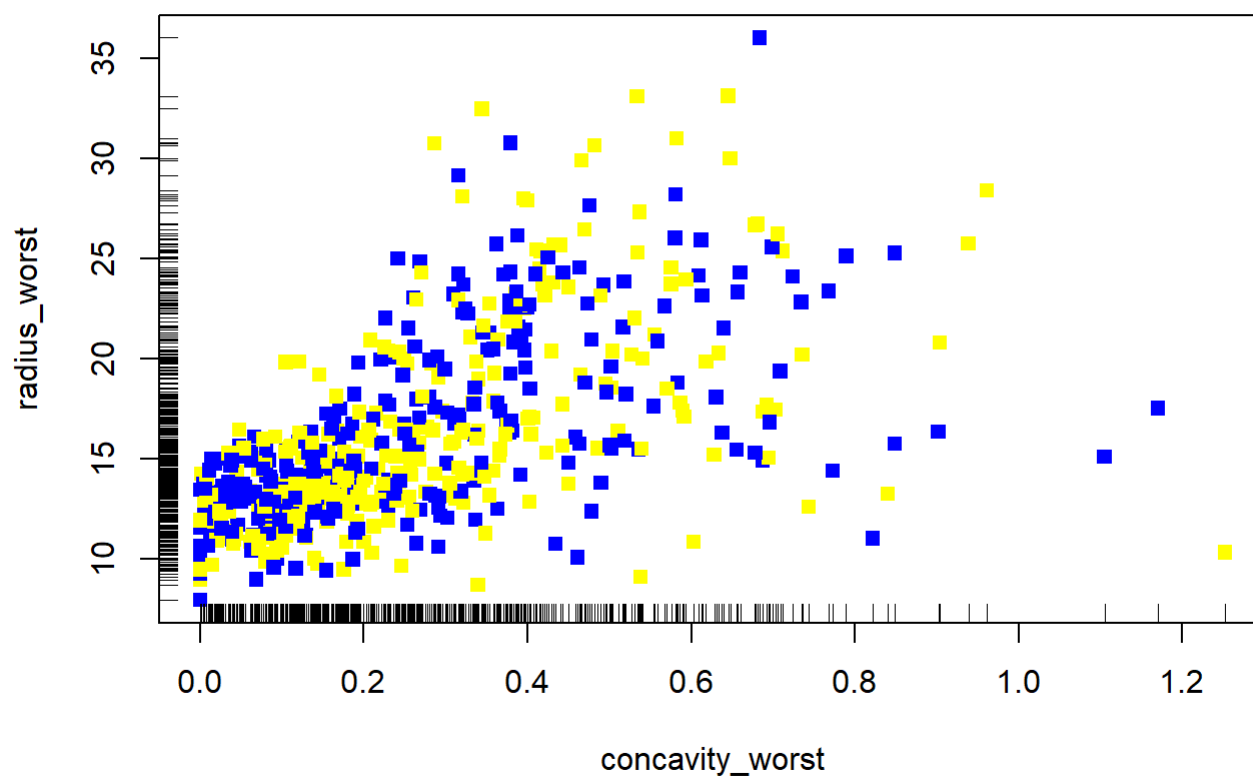
```



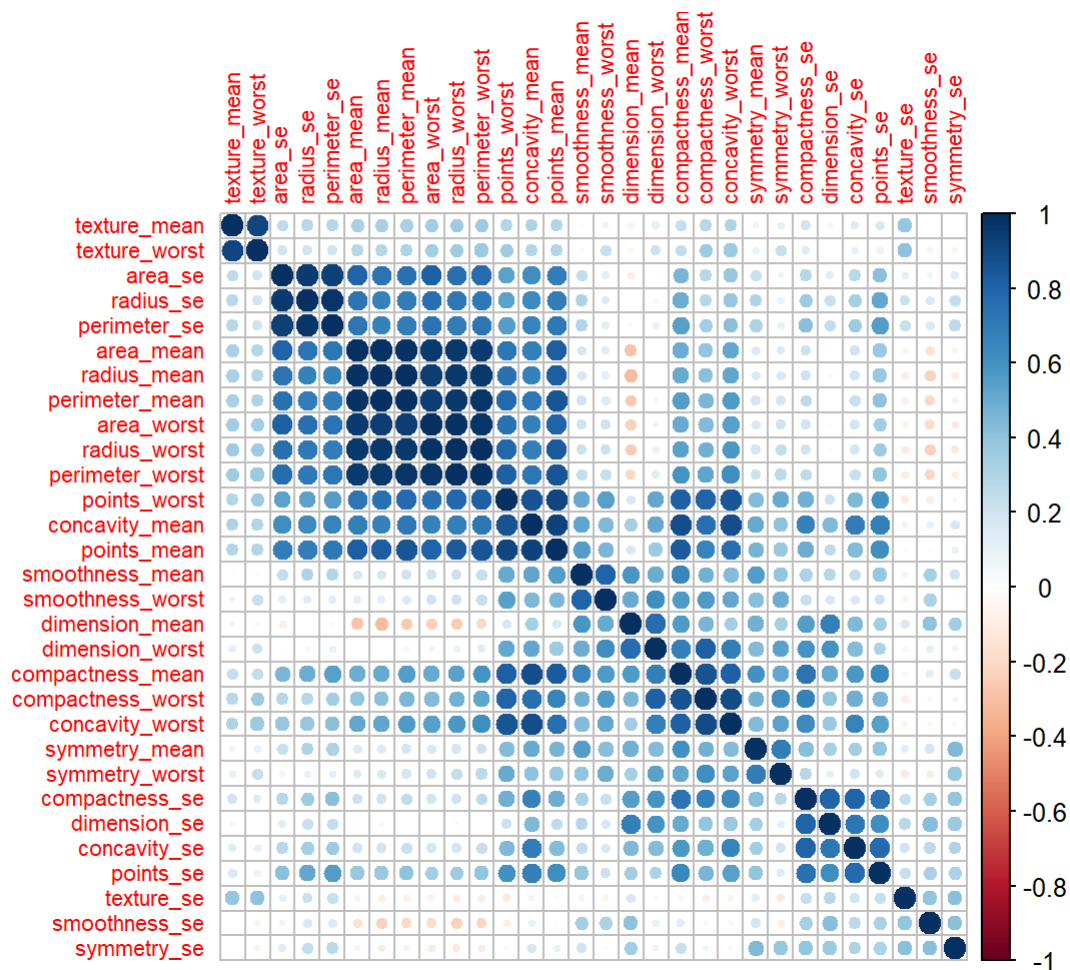
```
#Scatter plot of two variable (concavity again radius)
a <- cancer[,c('concavity_worst','radius_worst')]
plot(x = cancer$concavity_worst,y = cancer$radius_worst,
     xlab = "concavity_worst",
     ylab = "radius_worst",
     main = "Concavity_worst vs radius_worst",
     pch=15,col=c("blue","yellow"))
rug(cancer$concavity_worst, side = 1)

rug(cancer$radius_worst, side = 2)
```

Concavity_worst vs radius_worst



```
#Corelation Matrix of columns  
corMatMy <- cor(cancer[,3:32])  
corrplot(corMatMy, order = "hclust", tl.cex = 0.7)
```

```
#Scatterplot Matrix
```

```
pairs(~radius_mean+perimeter_mean+area_mean+compactness_mean+concavity_mean,data = cancer,main =  
"Scatterplot Matrix",col=c("blue","green","yellow","red"))
```

Scatterplot Matrix

