Breast Cancer Wisconsin (Diagnostic) Data Set Visualization

Vipul Sharma

# REMINDER

## What is the data?

The data set I used is called Breast Cancer Wisconsin (Diagnostic) Data Set. The  
features are computed from a digitized image of a fine needle aspirate (FNA) of a  
beast mass which describes characteristics of the cell nuclei present in the image.

## How was the data collected?

Data was obtained using Multisurface Method-Tree (MSM-T), a classification method  
which uses linear programming to construct a decision tree.

## Who collected the data?

William H. Wolberg, General Surgery Dept.  
University of Wisconsin, Clinical Sciences Center  
Madison, WI 53792  
wolberg '@' eagle.surgery.wisc.edu

W. Nick Street, Computer Sciences Dept.  
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706  
Street \*\*'@'\*\* cs.wisc.edu 608-262-

Olvi L. Mangasarian, Computer Sciences Dept.  
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706  
olvi \*\*'@'\*\* cs.wisc.edu

# DATA DESCRIPTION

## Shape

dim(dataset)

## [1] 569 33

Therefore there are 569 rows and 33 columns in the dataset.

## Structure

## Classes 'tbl\_df', 'tbl' and 'data.frame': 569 obs. of 33 variables:  
## $ id : int 842302 842517 84300903 84348301 84358402 843786 844359 84458202 844981 84501001 ...  
## $ diagnosis : chr "M" "M" "M" "M" ...  
## $ radius\_mean : num 18 20.6 19.7 11.4 20.3 ...  
## $ texture\_mean : num 10.4 17.8 21.2 20.4 14.3 ...  
## $ perimeter\_mean : num 122.8 132.9 130 77.6 135.1 ...  
## $ area\_mean : num 1001 1326 1203 386 1297 ...  
## $ smoothness\_mean : num 0.1184 0.0847 0.1096 0.1425 0.1003 ...  
## $ compactness\_mean : num 0.2776 0.0786 0.1599 0.2839 0.1328 ...  
## $ concavity\_mean : num 0.3001 0.0869 0.1974 0.2414 0.198 ...  
## $ concave points\_mean : num 0.1471 0.0702 0.1279 0.1052 0.1043 ...  
## $ symmetry\_mean : num 0.242 0.181 0.207 0.26 0.181 ...  
## $ fractal\_dimension\_mean : num 0.0787 0.0567 0.06 0.0974 0.0588 ...  
## $ radius\_se : num 1.095 0.543 0.746 0.496 0.757 ...  
## $ texture\_se : num 0.905 0.734 0.787 1.156 0.781 ...  
## $ perimeter\_se : num 8.59 3.4 4.58 3.44 5.44 ...  
## $ area\_se : num 153.4 74.1 94 27.2 94.4 ...  
## $ smoothness\_se : num 0.0064 0.00522 0.00615 0.00911 0.01149 ...  
## $ compactness\_se : num 0.049 0.0131 0.0401 0.0746 0.0246 ...  
## $ concavity\_se : num 0.0537 0.0186 0.0383 0.0566 0.0569 ...  
## $ concave points\_se : num 0.0159 0.0134 0.0206 0.0187 0.0188 ...  
## $ symmetry\_se : num 0.03 0.0139 0.0225 0.0596 0.0176 ...  
## $ fractal\_dimension\_se : num 0.00619 0.00353 0.00457 0.00921 0.00511 ...  
## $ radius\_worst : num 25.4 25 23.6 14.9 22.5 ...  
## $ texture\_worst : num 17.3 23.4 25.5 26.5 16.7 ...  
## $ perimeter\_worst : num 184.6 158.8 152.5 98.9 152.2 ...  
## $ area\_worst : num 2019 1956 1709 568 1575 ...  
## $ smoothness\_worst : num 0.162 0.124 0.144 0.21 0.137 ...  
## $ compactness\_worst : num 0.666 0.187 0.424 0.866 0.205 ...  
## $ concavity\_worst : num 0.712 0.242 0.45 0.687 0.4 ...  
## $ concave points\_worst : num 0.265 0.186 0.243 0.258 0.163 ...  
## $ symmetry\_worst : num 0.46 0.275 0.361 0.664 0.236 ...  
## $ fractal\_dimension\_worst: num 0.1189 0.089 0.0876 0.173 0.0768 ...  
## $ X33 : chr NA NA NA NA ...  
## - attr(\*, "problems")=Classes 'tbl\_df', 'tbl' and 'data.frame': 569 obs. of 5 variables:  
## ..$ row : int 1 2 3 4 5 6 7 8 9 10 ...  
## ..$ col : chr NA NA NA NA ...  
## ..$ expected: chr "33 columns" "33 columns" "33 columns" "33 columns" ...  
## ..$ actual : chr "32 columns" "32 columns" "32 columns" "32 columns" ...  
## ..$ file : chr "'data.csv'" "'data.csv'" "'data.csv'" "'data.csv'" ...  
## - attr(\*, "spec")=List of 2  
## ..$ cols :List of 33  
## .. ..$ id : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_integer" "collector"  
## .. ..$ diagnosis : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## .. ..$ radius\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ texture\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ perimeter\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ area\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ smoothness\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ compactness\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ concavity\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ concave points\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ symmetry\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ fractal\_dimension\_mean : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ radius\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ texture\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ perimeter\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ area\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ smoothness\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ compactness\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ concavity\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ concave points\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ symmetry\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ fractal\_dimension\_se : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ radius\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ texture\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ perimeter\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ area\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ smoothness\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ compactness\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ concavity\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ concave points\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ symmetry\_worst : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ fractal\_dimension\_worst: list()  
## .. .. ..- attr(\*, "class")= chr "collector\_double" "collector"  
## .. ..$ X33 : list()  
## .. .. ..- attr(\*, "class")= chr "collector\_character" "collector"  
## ..$ default: list()  
## .. ..- attr(\*, "class")= chr "collector\_guess" "collector"  
## ..- attr(\*, "class")= chr "col\_spec"

## Summary

## 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 concave points\_mean symmetry\_mean   
## Min. :0.00000 Min. :0.00000 Min. :0.1060   
## 1st Qu.:0.02956 1st Qu.:0.02031 1st Qu.:0.1619   
## Median :0.06154 Median :0.03350 Median :0.1792   
## Mean :0.08880 Mean :0.04892 Mean :0.1812   
## 3rd Qu.:0.13070 3rd Qu.:0.07400 3rd Qu.:0.1957   
## Max. :0.42680 Max. :0.20120 Max. :0.3040   
## fractal\_dimension\_mean radius\_se texture\_se perimeter\_se   
## Min. :0.04996 Min. :0.1115 Min. :0.3602 Min. : 0.757   
## 1st Qu.:0.05770 1st Qu.:0.2324 1st Qu.:0.8339 1st Qu.: 1.606   
## Median :0.06154 Median :0.3242 Median :1.1080 Median : 2.287   
## Mean :0.06280 Mean :0.4052 Mean :1.2169 Mean : 2.866   
## 3rd Qu.:0.06612 3rd Qu.:0.4789 3rd Qu.:1.4740 3rd Qu.: 3.357   
## Max. :0.09744 Max. :2.8730 Max. :4.8850 Max. :21.980   
## area\_se smoothness\_se compactness\_se concavity\_se   
## Min. : 6.802 Min. :0.001713 Min. :0.002252 Min. :0.00000   
## 1st Qu.: 17.850 1st Qu.:0.005169 1st Qu.:0.013080 1st Qu.:0.01509   
## Median : 24.530 Median :0.006380 Median :0.020450 Median :0.02589   
## Mean : 40.337 Mean :0.007041 Mean :0.025478 Mean :0.03189   
## 3rd Qu.: 45.190 3rd Qu.:0.008146 3rd Qu.:0.032450 3rd Qu.:0.04205   
## Max. :542.200 Max. :0.031130 Max. :0.135400 Max. :0.39600   
## concave points\_se symmetry\_se fractal\_dimension\_se  
## Min. :0.000000 Min. :0.007882 Min. :0.0008948   
## 1st Qu.:0.007638 1st Qu.:0.015160 1st Qu.:0.0022480   
## Median :0.010930 Median :0.018730 Median :0.0031870   
## Mean :0.011796 Mean :0.020542 Mean :0.0037949   
## 3rd Qu.:0.014710 3rd Qu.:0.023480 3rd Qu.:0.0045580   
## Max. :0.052790 Max. :0.078950 Max. :0.0298400   
## radius\_worst texture\_worst perimeter\_worst area\_worst   
## Min. : 7.93 Min. :12.02 Min. : 50.41 Min. : 185.2   
## 1st Qu.:13.01 1st Qu.:21.08 1st Qu.: 84.11 1st Qu.: 515.3   
## Median :14.97 Median :25.41 Median : 97.66 Median : 686.5   
## Mean :16.27 Mean :25.68 Mean :107.26 Mean : 880.6   
## 3rd Qu.:18.79 3rd Qu.:29.72 3rd Qu.:125.40 3rd Qu.:1084.0   
## Max. :36.04 Max. :49.54 Max. :251.20 Max. :4254.0   
## smoothness\_worst compactness\_worst concavity\_worst concave points\_worst  
## Min. :0.07117 Min. :0.02729 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.11660 1st Qu.:0.14720 1st Qu.:0.1145 1st Qu.:0.06493   
## Median :0.13130 Median :0.21190 Median :0.2267 Median :0.09993   
## Mean :0.13237 Mean :0.25427 Mean :0.2722 Mean :0.11461   
## 3rd Qu.:0.14600 3rd Qu.:0.33910 3rd Qu.:0.3829 3rd Qu.:0.16140   
## Max. :0.22260 Max. :1.05800 Max. :1.2520 Max. :0.29100   
## symmetry\_worst fractal\_dimension\_worst X33   
## Min. :0.1565 Min. :0.05504 Length:569   
## 1st Qu.:0.2504 1st Qu.:0.07146 Class :character   
## Median :0.2822 Median :0.08004 Mode :character   
## Mean :0.2901 Mean :0.08395   
## 3rd Qu.:0.3179 3rd Qu.:0.09208   
## Max. :0.6638 Max. :0.20750

## Changes made in dataset

If we look at the structure of the dataset, we find that there are two columns which have character datatypes. The first column which has the character datatype can just be ignored since it is the last column and is filled with NA's only. Therefore the dataset will have 32 columns now instead of 33. The second factor column is the 'diagnosis' column which has two labels 'M' for Malignant and 'B' which stands for Benign. We are going to change this column into numeric datatype by classifying 'M' as 1 and 'B' as 0. This results in 357 1's and 212 0's.

dataset <- dataset[, 1:32 ]  
dim(dataset)

## [1] 569 32

dataset$diagnosis <- as.integer(dataset$diagnosis == 'M')  
head(dataset$diagnosis)

## [1] 1 1 1 1 1 1

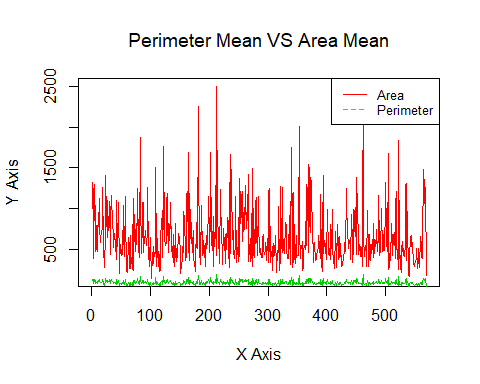
## Graph Visualizations

### Mean Graphs

Graph plots means of three real-valued features are computed for each cell nucleus.   
Note : Click on the graph if the graph does not show up.

This graph plots the radius mean - mean of distances from center to points on the perimeter, texture - standard deviation of gray-scale values and the perimeter mean.

From this we can plot the area vs perimeter graph.



This graph plots the indiviual means of smoothness, compactness and concativity variables.

### Inital Observation

Ten real-valued features are computed for each cell nucleus:  
  
a) radius (mean of distances from center to points on the perimeter)   
b) texture (standard deviation of gray-scale values)   
c) perimeter   
d) area   
e) smoothness (local variation in radius lengths)   
f) compactness (perimeter^2 / area - 1.0)   
g) concavity (severity of concave portions of the contour)   
h) concave points (number of concave portions of the contour)   
i) symmetry   
j) fractal dimension ("coastline approximation" - 1)  
  
The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.   
For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.  
  
All feature values are recoded with four significant digits.  
  
Missing attribute values: none  
  
Class distribution: 357 benign, 212 malignant