AnalysisScript - Breast Cancer Wisconsin (Diagnostic)

Thanh La, Son Luong

10/17/2021

```
workfile <- read.csv("data.csv")
numeric_workfile <- workfile[,3:32 ]
```

workfile Exploration

```
colnames(workfile)
```

```
##  [1] "id"                      "diagnosis"
##  [3] "radius_mean"             "texture_mean"
##  [5] "perimeter_mean"          "area_mean"
##  [7] "smoothness_mean"         "compactness_mean"
##  [9] "concavity_mean"          "concave.points_mean"
## [11] "symmetry_mean"           "fractal_dimension_mean"
## [13] "radius_se"               "texture_se"
## [15] "perimeter_se"            "area_se"
## [17] "smoothness_se"           "compactness_se"
## [19] "concavity_se"            "concave.points_se"
## [21] "symmetry_se"             "fractal_dimension_se"
## [23] "radius_worst"            "texture_worst"
## [25] "perimeter_worst"         "area_worst"
## [27] "smoothness_worst"        "compactness_worst"
## [29] "concavity_worst"         "concave.points_worst"
## [31] "symmetry_worst"          "fractal_dimension_worst"
```

```
summary(numeric_workfile)
```

```
##   radius_mean      texture_mean   perimeter_mean     area_mean
##  Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5
##  1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3
##  Median :13.370   Median :18.84   Median : 86.24   Median : 551.1
##  Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9
##  3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7
##  Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0
##  smoothness_mean   compactness_mean  concavity_mean    concave.points_mean
##  Min.   :0.05263   Min.   :0.01938   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031
##  Median :0.09587   Median :0.09263   Median :0.06154   Median :0.03350
```

```
##  Mean   :0.09636   Mean   :0.10434   Mean   :0.08880   Mean   :0.04892
##  3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400
##  Max.   :0.16340   Max.   :0.34540   Max.   :0.42680   Max.   :0.20120
##  symmetry_mean    fractal_dimension_mean   radius_se        texture_se
##  Min.   :0.1060   Min.   :0.04996        Min.   :0.1115   Min.   :0.3602
##  1st Qu.:0.1619   1st Qu.:0.05770        1st Qu.:0.2324   1st Qu.:0.8339
##  Median :0.1792   Median :0.06154        Median :0.3242   Median :1.1080
##  Mean   :0.1812   Mean   :0.06280        Mean   :0.4052   Mean   :1.2169
##  3rd Qu.:0.1957   3rd Qu.:0.06612        3rd Qu.:0.4789   3rd Qu.:1.4740
##  Max.   :0.3040   Max.   :0.09744        Max.   :2.8730   Max.   :4.8850
##   perimeter_se       area_se        smoothness_se      compactness_se
##  Min.   : 0.757   Min.   :  6.802   Min.   :0.001713   Min.   :0.002252
##  1st Qu.: 1.606   1st Qu.: 17.850   1st Qu.:0.005169   1st Qu.:0.013080
##  Median : 2.287   Median : 24.530   Median :0.006380   Median :0.020450
##  Mean   : 2.866   Mean   : 40.337   Mean   :0.007041   Mean   :0.025478
##  3rd Qu.: 3.357   3rd Qu.: 45.190   3rd Qu.:0.008146   3rd Qu.:0.032450
##  Max.   :21.980   Max.   :542.200   Max.   :0.031130   Max.   :0.135400
##   concavity_se     concave.points_se   symmetry_se       fractal_dimension_se
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.007882   Min.   :0.0008948
##  1st Qu.:0.01509   1st Qu.:0.007638   1st Qu.:0.015160   1st Qu.:0.0022480
##  Median :0.02589   Median :0.010930   Median :0.018730   Median :0.0031870
##  Mean   :0.03189   Mean   :0.011796   Mean   :0.020542   Mean   :0.0037949
##  3rd Qu.:0.04205   3rd Qu.:0.014710   3rd Qu.:0.023480   3rd Qu.:0.0045580
##  Max.   :0.39600   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
##  Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2822   Median :0.08004
##  Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :0.6638   Max.   :0.20750
```

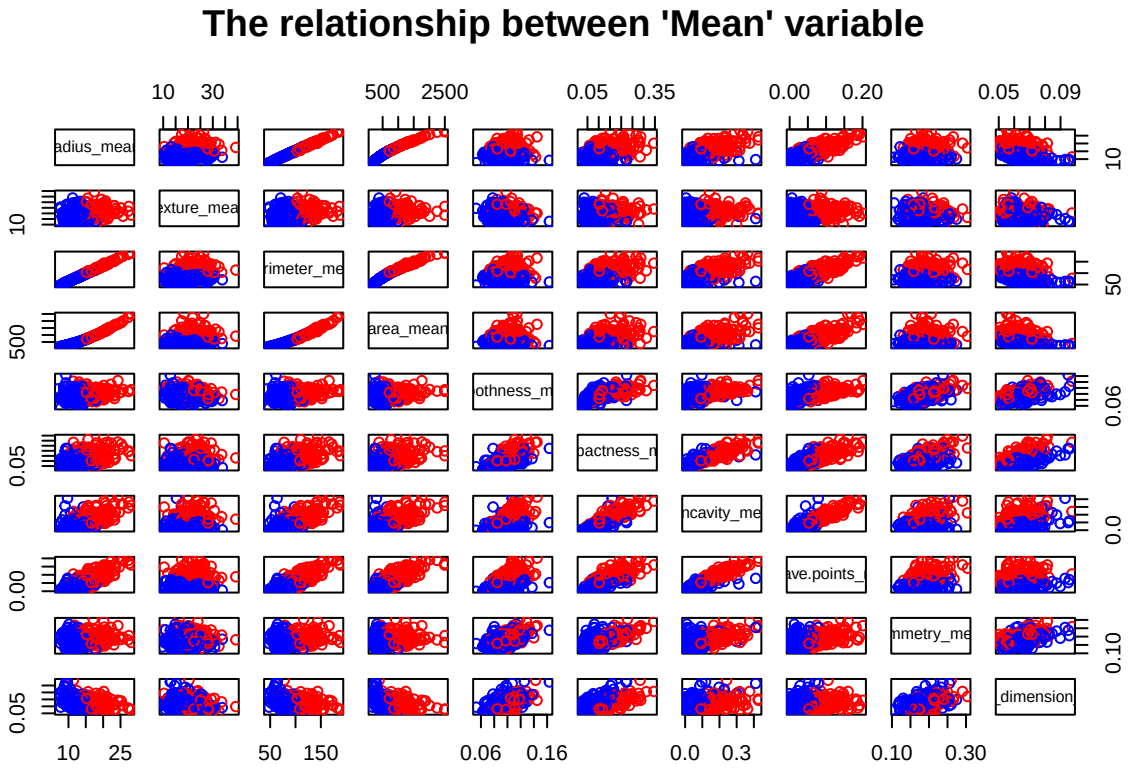
```
all(is.na(workfile))
```

```
## [1] FALSE
```

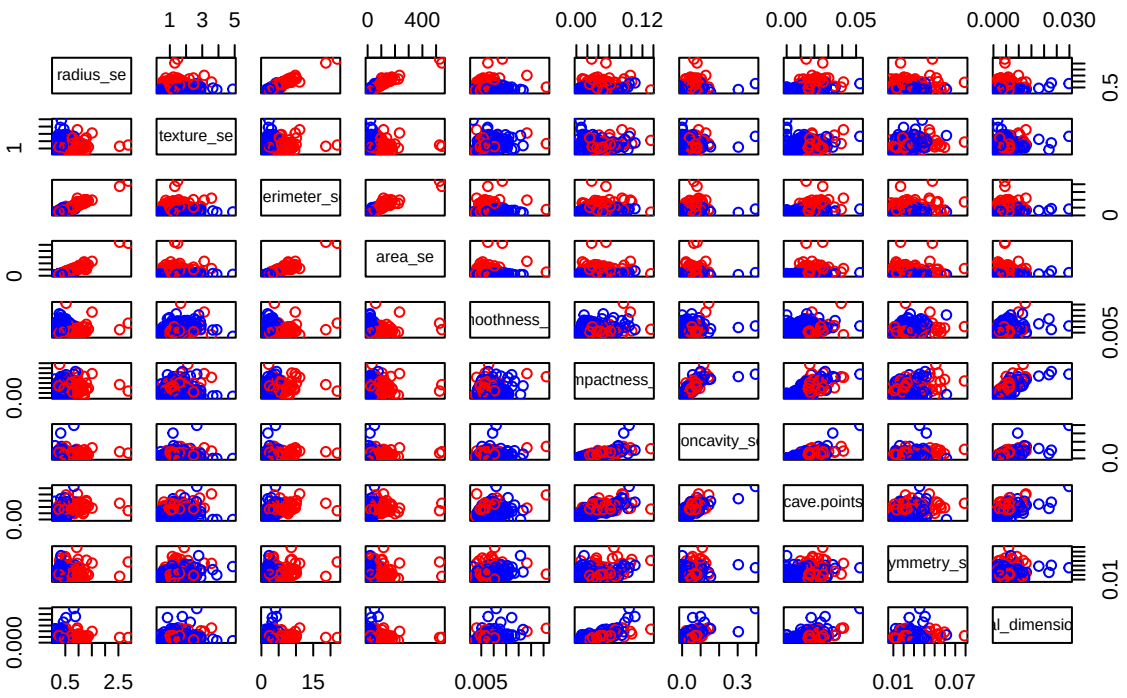
workfile Visualization

I. Relationship Between Variable

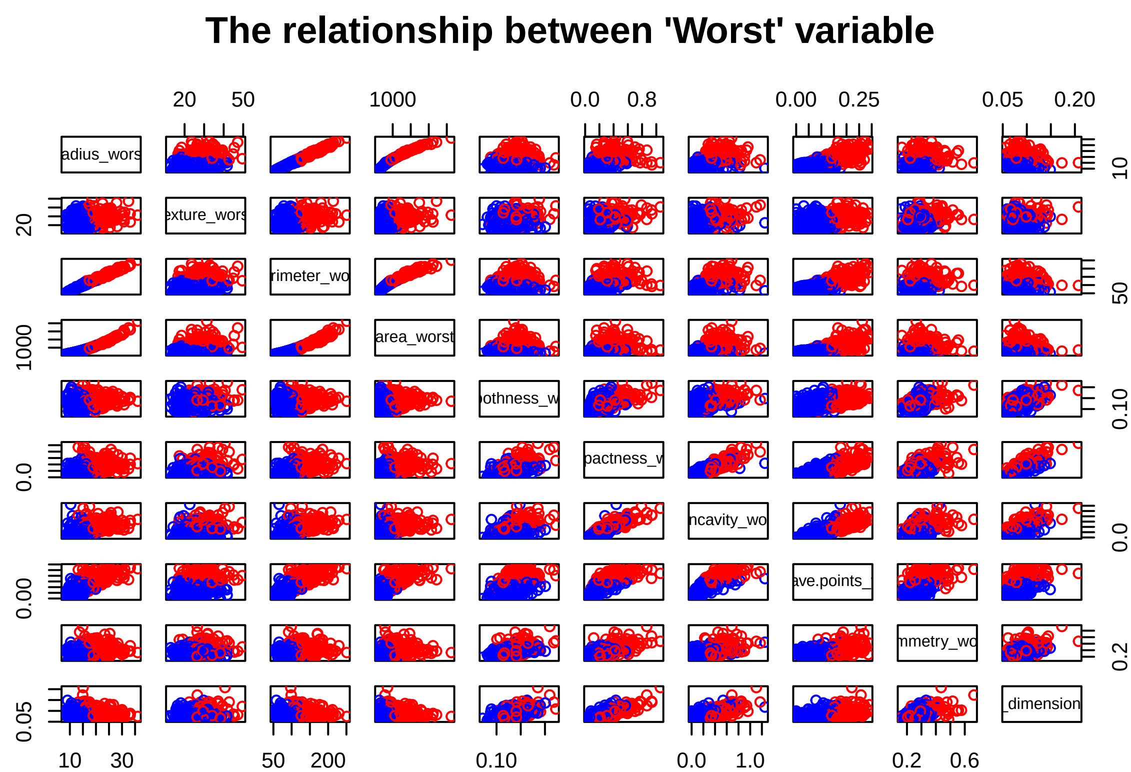
```
group = NA
group[workfile$diagnosis == "B"] = 1
group[workfile$diagnosis == "M"] = 2


workfile %>%
  select_if(is.numeric) %>%
  dplyr::select(contains("mean")) %>%
  pairs(col = c("blue", "red")[group], main = "The relationship between 'Mean' variable")
```




```
workfile %>%
  select_if(is.numeric) %>%
  dplyr::select(contains("SE")) %>%
  pairs(col = c("blue", "red")[group], main = "The relationship between 'Standard Error' variable")
```




```
workfile %>%
  select_if(is.numeric) %>%
  dplyr::select(contains("worst")) %>%
  pairs(col = c("blue", "red")[group], main = "The relationship between 'Worst' variable")
```



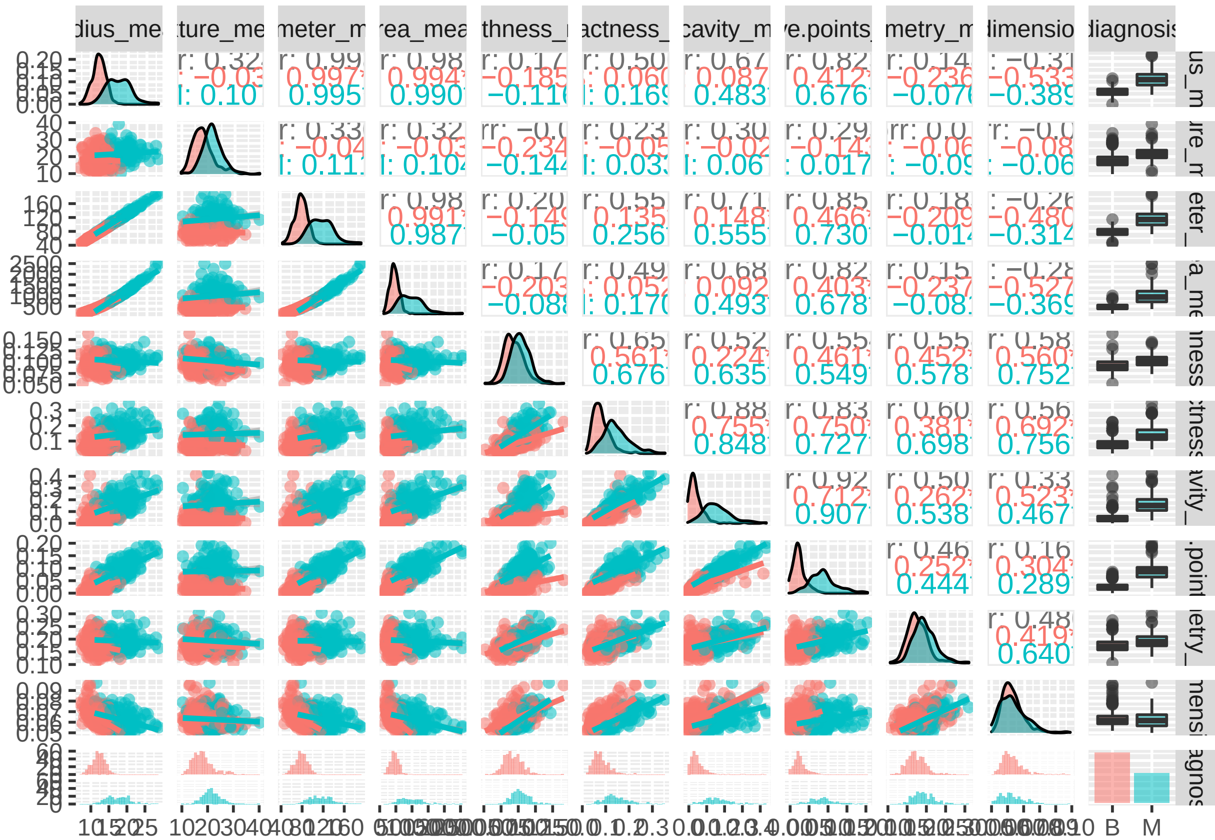
II. Differently Distributed Between Two Group of Diagnosis

look at how the variables are differently distributed between the two groups

these variables below are greatly different among 2 group of diagnosis

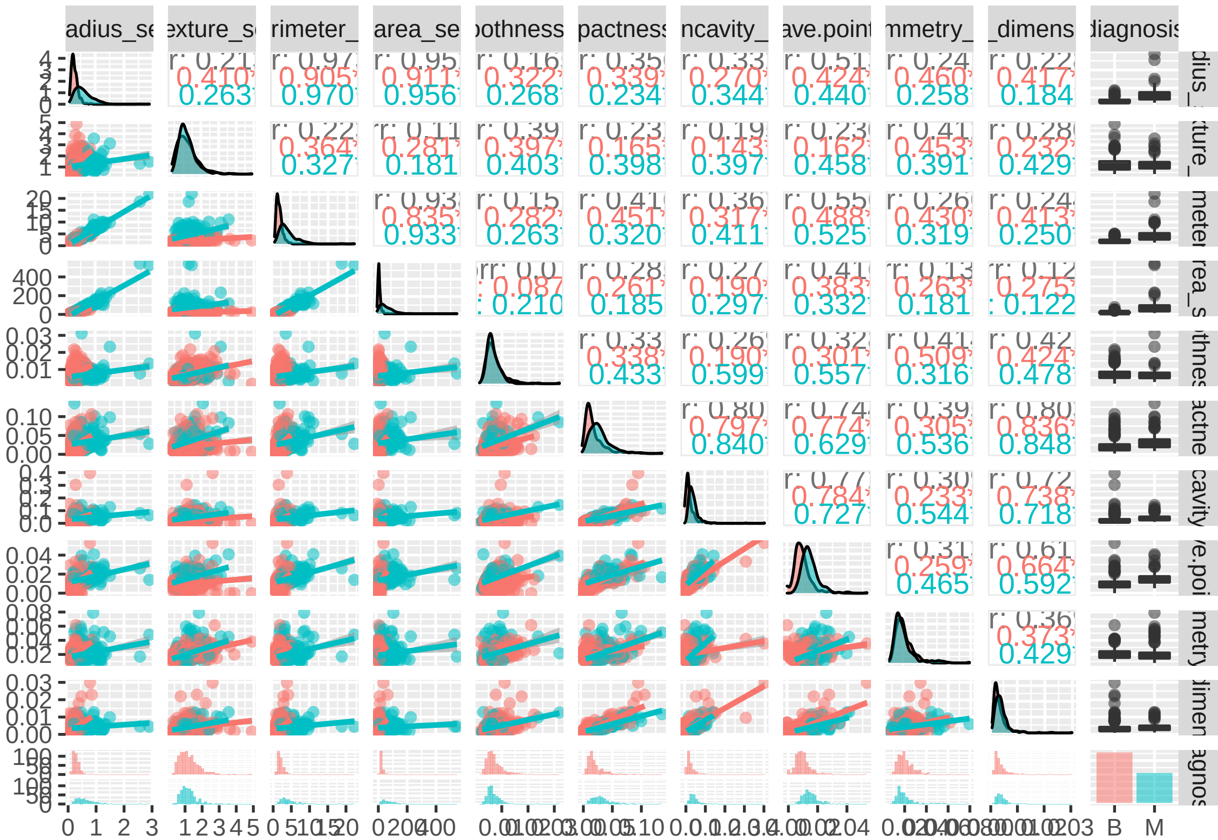
```
ggpairs(workfile[,c(3:12,2)], aes(color=diagnosis, alpha=0.5), lower=list(continuous="smooth"))+
theme(plot.title=element_text(face='bold',color='black',hjust=0.5,size=12))
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



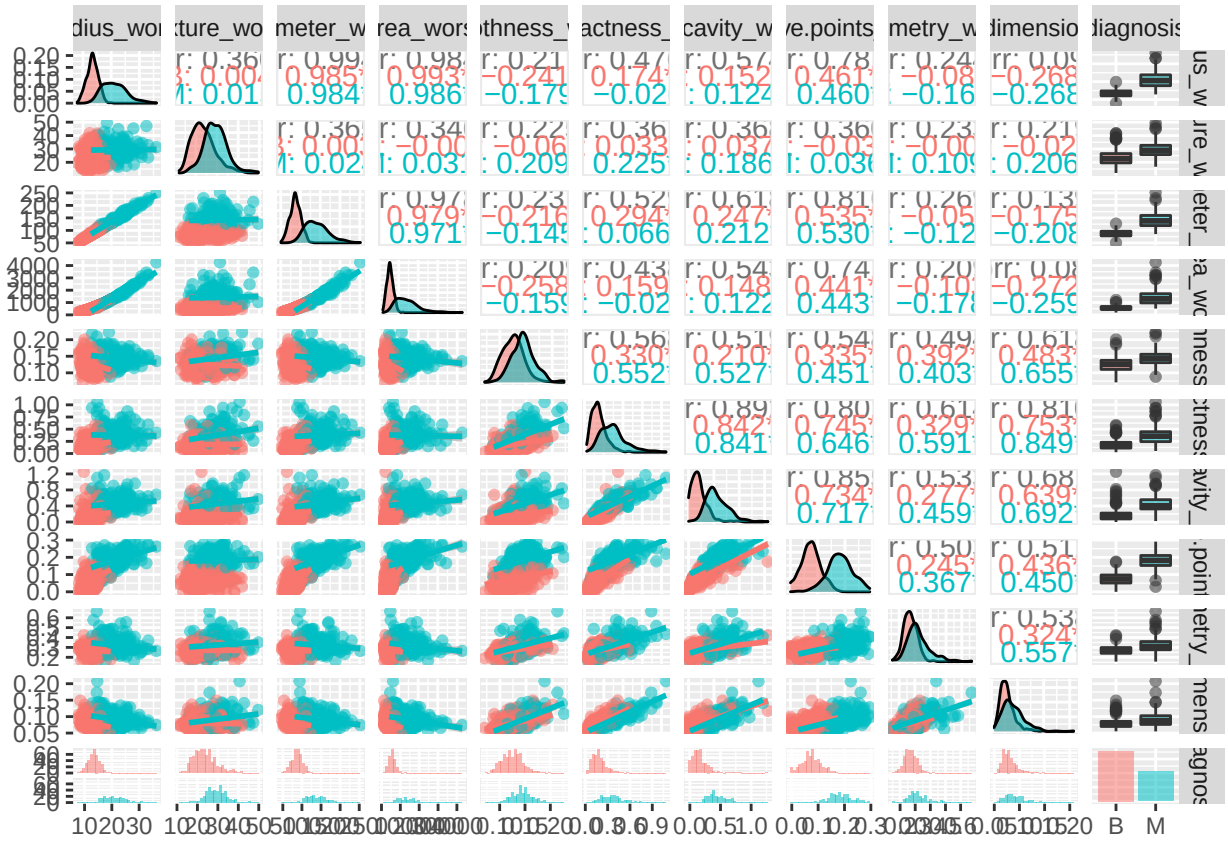

```
ggpairs(workfile[,c(13:22,2)], aes(color=diagnosis, alpha=0.5), lower=list(continuous="smooth"))+
theme(plot.title=element_text(face='bold',color='black',hjust=0.5,size=12))
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



```r
ggpairs(workfile[,c(23:32,2)], aes(color=diagnosis, alpha=0.5), lower=list(continuous="smooth"))+
theme(plot.title=element_text(face='bold',color='black',hjust=0.5,size=12))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



cor()# cov()# https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

III. Logistic Regression Model

```
workfile$diagnosis = as.factor(workfile$diagnosis)
workfile = subset(workfile, select = -c(id))
```

```
str(workfile)
```

```
## 'data.frame':    569 obs. of  31 variables:
##  $ diagnosis               : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ 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 ...
```

```
head(workfile)
```

```
##   diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
## 1         M       17.99        10.38         122.80    1001.0         0.11840
## 2         M       20.57        17.77         132.90    1326.0         0.08474
## 3         M       19.69        21.25         130.00    1203.0         0.10960
## 4         M       11.42        20.38          77.58     386.1         0.14250
## 5         M       20.29        14.34         135.10    1297.0         0.10030
## 6         M       12.45        15.70          82.57     477.1         0.12780
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1          0.27760         0.3001             0.14710        0.2419
## 2          0.07864         0.0869             0.07017        0.1812
## 3          0.15990         0.1974             0.12790        0.2069
## 4          0.28390         0.2414             0.10520        0.2597
## 5          0.13280         0.1980             0.10430        0.1809
## 6          0.17000         0.1578             0.08089        0.2087
##   fractal_dimension_mean radius_se texture_se perimeter_se area_se
## 1                0.07871    1.0950     0.9053        8.589  153.40
## 2                0.05667    0.5435     0.7339        3.398   74.08
## 3                0.05999    0.7456     0.7869        4.585   94.03
## 4                0.09744    0.4956     1.1560        3.445   27.23
## 5                0.05883    0.7572     0.7813        5.438   94.44
## 6                0.07613    0.3345     0.8902        2.217   27.19
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1      0.006399        0.04904      0.05373           0.01587     0.03003
## 2      0.005225        0.01308      0.01860           0.01340     0.01389
## 3      0.006150        0.04006      0.03832           0.02058     0.02250
## 4      0.009110        0.07458      0.05661           0.01867     0.05963
## 5      0.011490        0.02461      0.05688           0.01885     0.01756
## 6      0.007510        0.03345      0.03672           0.01137     0.02165
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1             0.006193        25.38         17.33          184.60     2019.0
## 2             0.003532        24.99         23.41          158.80     1956.0
## 3             0.004571        23.57         25.53          152.50     1709.0
## 4             0.009208        14.91         26.50           98.87      567.7
```

```
## 5             0.005115        22.54         16.67          152.20     1575.0
## 6             0.005082        15.47         23.75          103.40      741.6
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1           0.1622            0.6656          0.7119               0.2654
## 2           0.1238            0.1866          0.2416               0.1860
## 3           0.1444            0.4245          0.4504               0.2430
## 4           0.2098            0.8663          0.6869               0.2575
## 5           0.1374            0.2050          0.4000               0.1625
## 6           0.1791            0.5249          0.5355               0.1741
##   symmetry_worst fractal_dimension_worst
## 1         0.4601                 0.11890
## 2         0.2750                 0.08902
## 3         0.3613                 0.08758
## 4         0.6638                 0.17300
## 5         0.2364                 0.07678
## 6         0.3985                 0.12440
```

Apply Stepwise Regression to find Correlattion of variable

```
# backward stepwise regression
full_model = glm(diagnosis~., family = "binomial", data = workfile)
backward_model = step(full_model, direction = "backward")
```

```
## Start:  AIC=32068.76
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + radius_se + texture_se +
##     perimeter_se + area_se + smoothness_se + compactness_se +
##     concavity_se + concave.points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     concave.points_worst + symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance   AIC
## - concave.points_worst     1        0    60
## - radius_se                1        0    60
## - perimeter_worst          1        0    60
## - texture_se               1        0    60
## - smoothness_se            1        0    60
## - smoothness_worst         1        0    60
## - perimeter_mean           1        0    60
## - texture_mean             1        0    60
## - smoothness_mean          1        0    60
## - radius_worst             1        0    60
## - texture_worst            1        0    60
## - concave.points_mean      1        0    60
## - fractal_dimension_mean   1        0    60
## - perimeter_se             1        0    60
## - area_mean                1       26    86
## - symmetry_se              1       27    87
## - fractal_dimension_worst  1       27    87
```

```
## - symmetry_mean            1       27    87
## - compactness_se           1       29    89
## - compactness_mean         1       31    91
## - symmetry_worst           1       31    91
## - concave.points_se        1       33    93
## - fractal_dimension_se     1       34    94
## - concavity_se             1       36    96
## - area_worst               1      721   781
## - compactness_worst        1      793   853
## - concavity_mean           1      865   925
## - area_se                  1      937   997
## - concavity_worst          1     1009  1069
## - radius_mean              1     1442  1502
## <none>                          32007 32069
##
## Step:  AIC=60
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + radius_se + texture_se +
##     perimeter_se + area_se + smoothness_se + compactness_se +
##     concavity_se + concave.points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance     AIC
## - radius_se                1     0.00   58.00
## - perimeter_mean           1     0.00   58.00
## - perimeter_worst          1     0.00   58.00
## - smoothness_se            1     0.00   58.00
## - texture_se               1     0.00   58.00
## - smoothness_worst         1     0.00   58.00
## - area_worst               1     0.00   58.00
## - radius_worst             1     0.00   58.00
## - texture_mean             1     0.00   58.00
## - concave.points_mean      1     0.00   58.00
## - fractal_dimension_mean   1     0.00   58.00
## - texture_worst            1     0.00   58.00
## - area_se                  1     0.00   58.00
## - perimeter_se             1     0.00   58.00
## - smoothness_mean          1     0.00   58.00
## <none>                           0.00   60.00
## - radius_mean              1    20.79   78.79
## - symmetry_se              1    26.69   84.69
## - area_mean                1    26.96   84.96
## - symmetry_mean            1    27.50   85.50
## - fractal_dimension_worst  1    28.70   86.70
## - compactness_se           1    28.78   86.78
## - compactness_mean         1    30.67   88.67
## - symmetry_worst           1    30.82   88.82
## - fractal_dimension_se     1    34.01   92.01
## - concavity_se             1    35.95   93.95
## - concave.points_se        1    39.33   97.33
## - concavity_mean           1   648.79  706.79
```

```
## - compactness_worst        1   865.05  923.05
## - concavity_worst          1  1009.22 1067.22
##
## Step:  AIC=58
## diagnosis ~ radius_mean + texture_mean + perimeter_mean + area_mean +
##     smoothness_mean + compactness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + texture_se + perimeter_se +
##     area_se + smoothness_se + compactness_se + concavity_se +
##     concave.points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance     AIC
## - perimeter_mean           1     0.00   56.00
## - texture_se               1     0.00   56.00
## - perimeter_worst          1     0.00   56.00
## - smoothness_worst         1     0.00   56.00
## - smoothness_se            1     0.00   56.00
## - texture_worst            1     0.00   56.00
## - texture_mean             1     0.00   56.00
## - fractal_dimension_mean   1     0.00   56.00
## - concave.points_mean      1     0.00   56.00
## - concavity_worst          1     0.00   56.00
## - radius_worst             1     0.00   56.00
## - smoothness_mean          1     0.05   56.05
## <none>                           0.00   58.00
## - radius_mean              1    23.23   79.23
## - perimeter_se             1    24.34   80.34
## - symmetry_se              1    27.06   83.06
## - symmetry_mean            1    27.50   83.50
## - compactness_se           1    29.12   85.12
## - area_se                  1    29.99   85.99
## - area_mean                1    30.00   86.00
## - symmetry_worst           1    30.90   86.90
## - fractal_dimension_worst  1    31.05   87.05
## - compactness_mean         1    31.10   87.10
## - fractal_dimension_se     1    34.76   90.76
## - concavity_se             1    36.14   92.14
## - concave.points_se        1    39.58   95.58
## - area_worst               1   865.05  921.05
## - concavity_mean           1  1297.57 1353.57
## - compactness_worst        1  1441.75 1497.75
##
## Step:  AIC=56
## diagnosis ~ radius_mean + texture_mean + area_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + texture_se + perimeter_se +
##     area_se + smoothness_se + compactness_se + concavity_se +
##     concave.points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     smoothness_worst + compactness_worst + concavity_worst +
##     symmetry_worst + fractal_dimension_worst
##
```

```
##                           Df Deviance     AIC
## - smoothness_worst         1     0.00   54.00
## - texture_se               1     0.00   54.00
## - texture_worst            1     0.00   54.00
## - smoothness_se            1     0.00   54.00
## - perimeter_worst          1     0.00   54.00
## - concave.points_mean      1     0.00   54.00
## - fractal_dimension_mean   1     0.00   54.00
## - compactness_worst        1     0.00   54.00
## - radius_worst             1     0.00   54.00
## <none>                           0.00   56.00
## - area_worst               1    21.06   75.06
## - concavity_mean           1    24.23   78.23
## - perimeter_se             1    24.40   78.40
## - smoothness_mean          1    25.25   79.25
## - symmetry_se              1    27.19   81.19
## - symmetry_mean            1    27.61   81.61
## - area_se                  1    30.04   84.04
## - compactness_se           1    30.06   84.06
## - area_mean                1    30.18   84.18
## - symmetry_worst           1    31.20   85.20
## - fractal_dimension_worst  1    31.21   85.21
## - radius_mean              1    31.96   85.96
## - compactness_mean         1    33.05   87.05
## - fractal_dimension_se     1    35.15   89.15
## - concavity_se             1    37.41   91.41
## - concave.points_se        1    40.19   94.19
## - texture_mean             1   576.70  630.70
## - concavity_worst          1  1225.48 1279.48
##
## Step:  AIC=54
## diagnosis ~ radius_mean + texture_mean + area_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + texture_se + perimeter_se +
##     area_se + smoothness_se + compactness_se + concavity_se +
##     concave.points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + texture_worst + perimeter_worst + area_worst +
##     compactness_worst + concavity_worst + symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - texture_se               1    0.000 52.000
## - perimeter_worst          1    0.000 52.000
## - texture_worst            1    0.000 52.000
## - concave.points_mean      1    0.000 52.000
## - compactness_worst        1    0.000 52.000
## <none>                           0.000 54.000
## - fractal_dimension_mean   1   21.389 73.389
## - concavity_worst          1   21.620 73.620
## - smoothness_se            1   23.542 75.542
## - area_worst               1   24.151 76.151
## - texture_mean             1   24.439 76.439
## - concavity_mean           1   25.381 77.381
## - perimeter_se             1   25.627 77.627
## - radius_worst             1   27.369 79.369
```

```
## - symmetry_mean            1   28.009 80.009
## - smoothness_mean          1   28.788 80.788
## - symmetry_se              1   28.927 80.927
## - compactness_se           1   30.065 82.065
## - area_mean                1   30.690 82.690
## - area_se                  1   31.272 83.272
## - symmetry_worst           1   32.185 84.185
## - fractal_dimension_worst  1   32.417 84.417
## - radius_mean              1   32.575 84.575
## - fractal_dimension_se     1   35.282 87.282
## - compactness_mean         1   35.306 87.306
## - concavity_se             1   37.431 89.431
## - concave.points_se        1   40.995 92.995
##
## Step:  AIC=52
## diagnosis ~ radius_mean + texture_mean + area_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + perimeter_se + area_se +
##     smoothness_se + compactness_se + concavity_se + concave.points_se +
##     symmetry_se + fractal_dimension_se + radius_worst + texture_worst +
##     perimeter_worst + area_worst + compactness_worst + concavity_worst +
##     symmetry_worst + fractal_dimension_worst
##
##                           Df Deviance    AIC
## - compactness_worst        1    0.000 50.000
## - texture_worst            1    0.000 50.000
## <none>                          0.000 52.000
## - fractal_dimension_mean   1   22.823 72.823
## - smoothness_se            1   23.627 73.627
## - perimeter_worst          1   24.913 74.913
## - area_worst               1   25.059 75.059
## - concavity_mean           1   25.383 75.383
## - concavity_worst          1   27.830 77.830
## - radius_worst             1   28.254 78.254
## - perimeter_se             1   28.531 78.531
## - symmetry_se              1   29.094 79.094
## - symmetry_mean            1   29.147 79.147
## - smoothness_mean          1   29.839 79.839
## - concave.points_mean      1   29.989 79.989
## - compactness_se           1   30.175 80.175
## - texture_mean             1   31.019 81.019
## - area_mean                1   32.081 82.081
## - area_se                  1   32.478 82.478
## - fractal_dimension_worst  1   32.531 82.531
## - symmetry_worst           1   33.456 83.456
## - radius_mean              1   34.428 84.428
## - compactness_mean         1   35.862 85.862
## - fractal_dimension_se     1   37.141 87.141
## - concavity_se             1   37.610 87.610
## - concave.points_se        1   41.959 91.959
##
## Step:  AIC=50
## diagnosis ~ radius_mean + texture_mean + area_mean + smoothness_mean +
##     compactness_mean + concavity_mean + concave.points_mean +
```

```
##     symmetry_mean + fractal_dimension_mean + perimeter_se + area_se +
##     smoothness_se + compactness_se + concavity_se + concave.points_se +
##     symmetry_se + fractal_dimension_se + radius_worst + texture_worst +
##     perimeter_worst + area_worst + concavity_worst + symmetry_worst +
##     fractal_dimension_worst
##
##                           Df Deviance    AIC
## <none>                          0.000 50.000
## - smoothness_se            1   23.696 71.696
## - area_worst               1   25.061 73.061
## - perimeter_worst          1   25.093 73.093
## - fractal_dimension_mean   1   27.129 75.129
## - radius_worst             1   28.390 76.390
## - perimeter_se             1   28.714 76.714
## - concavity_mean           1   28.747 76.747
## - concavity_worst          1   28.973 76.973
## - symmetry_se              1   29.102 77.102
## - symmetry_mean            1   29.227 77.227
## - texture_worst            1   30.276 78.276
## - smoothness_mean          1   30.474 78.474
## - concave.points_mean      1   30.640 78.640
## - compactness_se           1   30.876 78.876
## - texture_mean             1   31.204 79.204
## - area_mean                1   32.145 80.145
## - area_se                  1   32.513 80.513
## - fractal_dimension_worst  1   32.852 80.852
## - symmetry_worst           1   33.504 81.504
## - radius_mean              1   34.507 82.507
## - fractal_dimension_se     1   37.143 85.143
## - concavity_se             1   37.623 85.623
## - compactness_mean         1   40.345 88.345
## - concave.points_se        1   42.693 90.693
```

```
# forward stepwise regression
null_model = glm(diagnosis~1, family = "binomial", data = workfile)

forwar_model = step(null_model,
                    scope = list(lower = formula(null_model),
                                 uper = formula(null_model)),
                    direction = "forward")
```

```
## Start:  AIC=753.44
## diagnosis ~ 1
```

Apply logistic regression to data

```
train_control = trainControl(method = "repeatedcv", number = 10, repeats = 3)
```

To make the logistic regression model more efficient (or improve the probability of the prediction), I will separate the characteristic variable to different part such as: size, shape, surface

- Every variable

```
by_all = function(workfile){
  workfile_glm = train(diagnosis~.,
                       data = workfile,
                       trControl = train_control,
                       method = "glm",
                       family = binomial())
  return(workfile_glm)
}
```

- Size: Radius, perimeter, area, compactness, fractal dimension

```
by_size = function(workfile){
  workfile_glm = train(diagnosis~radius_mean + radius_se + radius_worst +
                         perimeter_mean + perimeter_se + perimeter_worst +
                         area_mean + area_se + area_worst +
                         compactness_mean + compactness_se + compactness_worst +
                         fractal_dimension_mean +  fractal_dimension_se + fractal_dimension_worst
                       ,data = workfile,
                       trControl = train_control,
                       method = "glm",
                       family = binomial())

  return(workfile_glm)
}
```

- Shape, surface: texture, smoothness, concavity, concave.points, symmetry, fractal dimension

```
by_shape = function(workfile){
  workfile_glm = train(diagnosis~ texture_mean + texture_se + texture_worst +
                         smoothness_mean + smoothness_se + smoothness_worst +
                         concavity_mean + concavity_se + concavity_worst +
                         concave.points_mean + concave.points_se + concave.points_worst +
                         symmetry_mean + symmetry_se + symmetry_worst +
                         fractal_dimension_mean +  fractal_dimension_se + fractal_dimension_worst,
                       data = workfile,
                       method = "glm",
                       family = binomial())
  return(workfile_glm)
}
```

- Stepwise Backward suggest

```
by_stepwise_backward = function(workfile){
  workfile_glm = train(diagnosis ~ radius_mean + texture_mean + area_mean + smoothness_mean +
    compactness_mean + concavity_mean + concave.points_mean +
    symmetry_mean + fractal_dimension_mean + perimeter_se + area_se +
    smoothness_se + compactness_se + concavity_se + concave.points_se +
    symmetry_se + fractal_dimension_se + radius_worst + texture_worst +
    perimeter_worst + area_worst + concavity_worst + symmetry_worst +
    fractal_dimension_worst,
```

```
                      data = workfile,
                      method = "glm",
                      family = binomial())

  return(workfile_glm)
}
```

- Stepwise Forward Suggest

Build a Train - Test of Logistic Regression

Train - test by size

```
workfile_glm = by_stepwise_backward(train_workfile)
summary(workfile_glm)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.003508   0.000000   0.000000   0.000000   0.003994
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.122e+03  1.349e+05  -0.045    0.964
## radius_mean              -6.659e+03  5.920e+04  -0.112    0.910
## texture_mean              1.941e+02  1.638e+03   0.118    0.906
## area_mean                 6.102e+01  5.500e+02   0.111    0.912
## smoothness_mean           3.969e+04  2.893e+05   0.137    0.891
## compactness_mean         -8.694e+04  6.526e+05  -0.133    0.894
## concavity_mean            2.885e+04  2.111e+05   0.137    0.891
## concave.points_mean       5.881e+04  7.105e+05   0.083    0.934
## symmetry_mean            -1.993e+04  1.573e+05  -0.127    0.899
## fractal_dimension_mean    1.644e+05  1.181e+06   0.139    0.889
## perimeter_se             -1.259e+03  1.029e+04  -0.122    0.903
## area_se                   1.571e+02  1.275e+03   0.123    0.902
## smoothness_se            -9.834e+04  8.223e+05  -0.120    0.905
## compactness_se            9.333e+04  7.580e+05   0.123    0.902
## concavity_se             -8.184e+04  6.379e+05  -0.128    0.898
## concave.points_se         4.420e+05  3.676e+06   0.120    0.904
## symmetry_se              -1.041e+05  1.050e+06  -0.099    0.921
## fractal_dimension_se     -1.103e+06  8.432e+06  -0.131    0.896
## radius_worst              2.248e+03  1.637e+04   0.137    0.891
## texture_worst             7.174e+01  1.479e+03   0.049    0.961
## perimeter_worst           1.274e+02  9.244e+02   0.138    0.890
## area_worst               -1.644e+01  1.129e+02  -0.146    0.884
## concavity_worst           6.759e+03  5.774e+04   0.117    0.907
## symmetry_worst            2.215e+04  1.837e+05   0.121    0.904
## fractal_dimension_worst   5.928e+04  5.380e+05   0.110    0.912
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.7139e+02  on 509  degrees of freedom
## Residual deviance: 1.4737e-04  on 485  degrees of freedom
## AIC: 50
##
## Number of Fisher Scoring iterations: 25
```

```
# now ewe predict on the test data
y_hat = predict(workfile_glm, newdata = test_workfile)
confusionMatrix(y_hat, test_workfile$diagnosis)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 35  0
##          M  0 24
##
##                Accuracy : 1
##                  95% CI : (0.9394, 1)
##     No Information Rate : 0.5932
##     P-Value [Acc > NIR] : 4.166e-14
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0000
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 1.0000
##              Prevalence : 0.5932
##          Detection Rate : 0.5932
##    Detection Prevalence : 0.5932
##       Balanced Accuracy : 1.0000
##
##        'Positive' Class : B
##
```