Data 621 - Final Project

Anthony Munoz

5/22/2020

Contents

Abstract

Breast cancer its one of the most common cancer nowadays in our society. For this project, I going to work with a dataset from Kaggle and also available on the UIC repository. In this project analysis we going to use regression modeling and try to replicate the outcome of the prediction which its this case we determine the change of cancer by declaring the results if they are Benign or Malignant. The modeling will try to come with the best accuracy prediction outcome. on this dataset we most will be working with the dimension and spect of the cancer cell.

Introduction

In this paper analysis, we going to work on getting a results product prediction by finding some insight into the dataset. most of our predictor variables are cancer cell dimensions for which just will allow us just to

predict those base on those features. This analysis won't support other aspects that may be important in the analysis of cancer prediction such as, people age, use or drug or alcohol, and more.

*Key works: Cancer, cells,regression.

Literature Review

When it comes to research on the type of cancer, breast cancer it's one with a big notorious amount of research because it one of the most dangerous illnesses that affect thousands of women worldwide.

Breast cancer its one of the leading causes of death and compassion with other types of cancer. There have been many approaches to do analysis and predict the risk of breast cancer. many of these approaches are using logistic regression, Machine learning, SVM, and others modeling. LR, SVM and KNN approaches (Madhu Kumaria, Vijendra Singhb,2018).

Kumaria and Singh worked on different modeling techniques and were able to obtain an accuracy of 99.28. The methodology on how they worked with the dataset it's similar on how I work with, for example, data selection, data processing/splitting the data, work with different model and select the most accurate one to test it on the evaluation dataset.

Breast cancer affects a range of different ages in a woman but according to this journal paper we see that more 89% percent of cancer was diagnosed with women older than 50 years old and just in 2017 more than 40,000 thousand women die because of this type of cancer(DeSantis, Jiemin Ma,Ann Goding Sauer, Newman,Ahmedin Jemal 2017).

Methodology

The mythology for this project is to work with regression modeling on a dataset which contains 570 rows and 33 columns. this dataset source comes from Kaggle and UIC repository. the data will be split and 2 datasets first by having the training dataset and the evaluation dataset. Later apply the GLM regression model to the training sample data and select the most accurate ones to test it with the evaluation dataset.

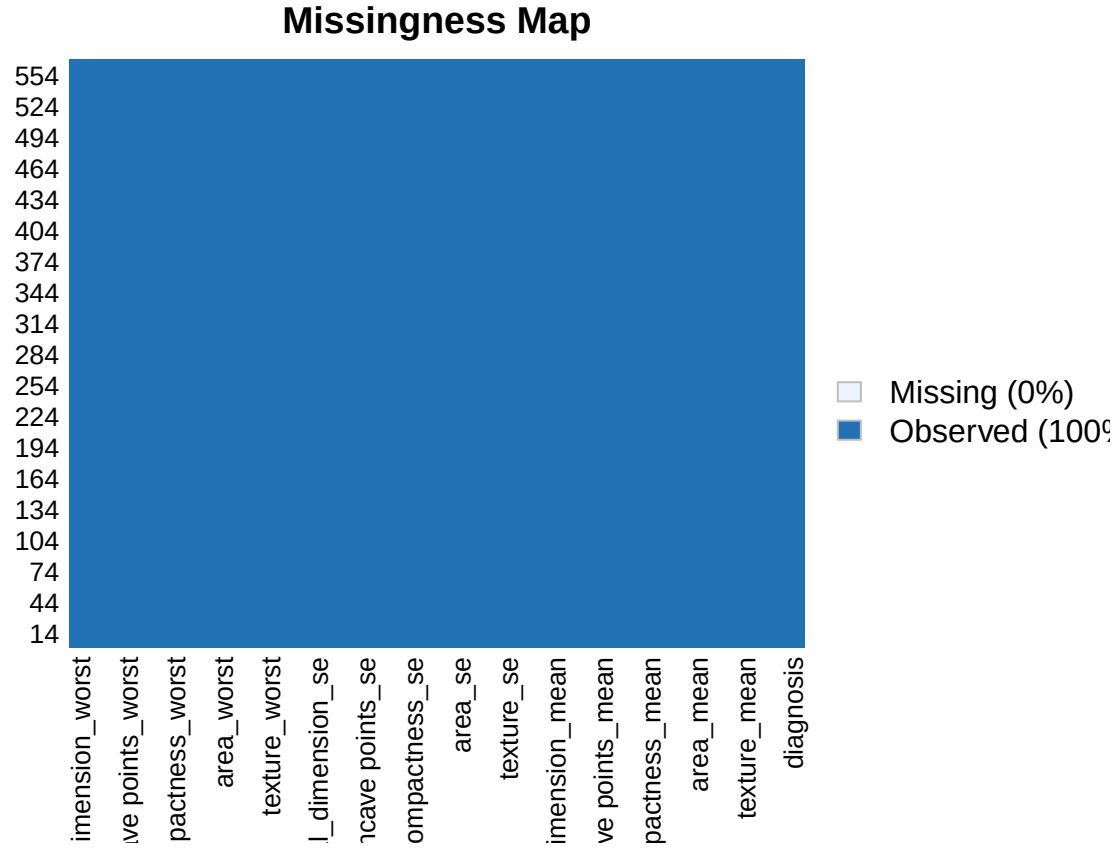
Experiment and results

Data Exploratory

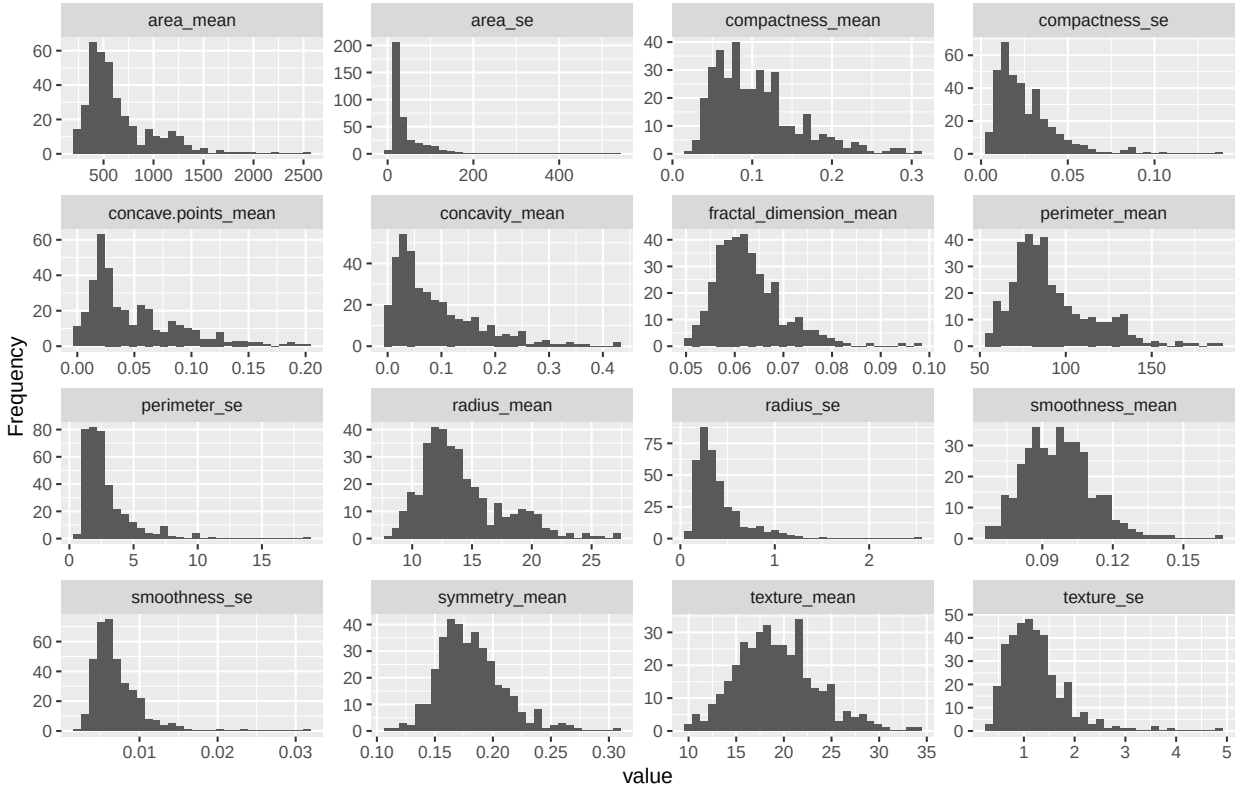
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    368 obs. of  31 variables:
##  $ diagnosis              : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  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 ...
```

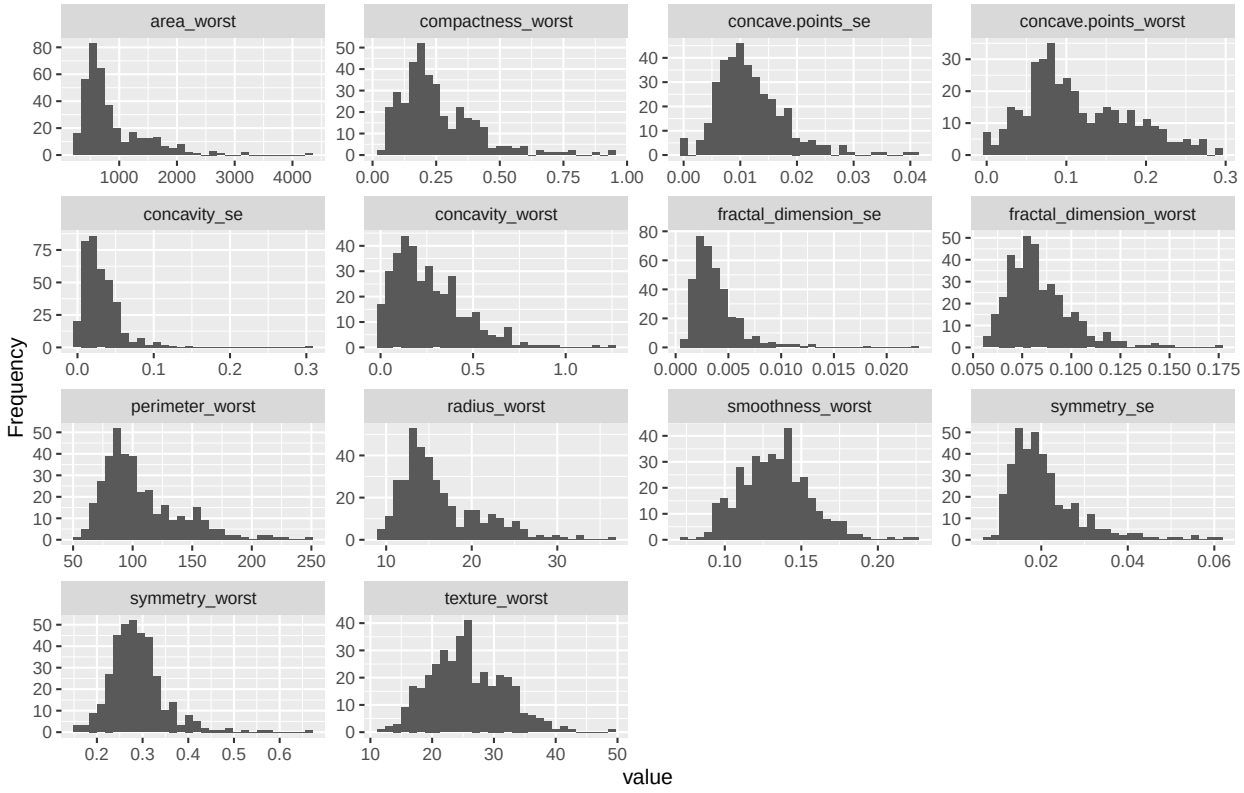
```
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ 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 ...
##  - 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'" ...

##    diagnosis       radius_mean      texture_mean   perimeter_mean
##  Min.   :0.0000   Min.   : 8.219   Min.   : 9.71   Min.   : 53.27
##  1st Qu.:0.0000   1st Qu.:11.705   1st Qu.:16.33   1st Qu.: 75.41
##  Median :0.0000   Median :13.325   Median :18.82   Median : 86.04
##  Mean   :0.3696   Mean   :14.115   Mean   :19.21   Mean   : 91.89
##  3rd Qu.:1.0000   3rd Qu.:15.715   3rd Qu.:21.70   3rd Qu.:103.45
##  Max.   :1.0000   Max.   :27.420   Max.   :33.81   Max.   :186.90
##    area_mean      smoothness_mean   compactness_mean  concavity_mean
##  Min.   : 203.9   Min.   :0.06576   Min.   :0.02344   Min.   :0.00000
##  1st Qu.: 422.4   1st Qu.:0.08660   1st Qu.:0.06363   1st Qu.:0.02964
##  Median : 548.8   Median :0.09671   Median :0.09403   Median :0.06071
##  Mean   : 651.7   Mean   :0.09675   Mean   :0.10473   Mean   :0.08814
##  3rd Qu.: 761.4   3rd Qu.:0.10540   3rd Qu.:0.13000   3rd Qu.:0.12662
##  Max.   :2501.0   Max.   :0.16340   Max.   :0.31140   Max.   :0.42680
##  concave points_mean symmetry_mean    fractal_dimension_mean   radius_se
##  Min.   :0.00000     Min.   :0.1060   Min.   :0.05024        Min.   :0.1144
##  1st Qu.:0.01969     1st Qu.:0.1620   1st Qu.:0.05828        1st Qu.:0.2315
##  Median :0.03367     Median :0.1783   Median :0.06158        Median :0.3320
##  Mean   :0.04914     Mean   :0.1811   Mean   :0.06285        Mean   :0.4013
##  3rd Qu.:0.06616     3rd Qu.:0.1958   3rd Qu.:0.06620        3rd Qu.:0.4704
##  Max.   :0.20120     Max.   :0.3040   Max.   :0.09744        Max.   :2.5470
##    texture_se      perimeter_se       area_se        smoothness_se
##  Min.   :0.3602   Min.   : 0.757   Min.   :  6.802   Min.   :0.001713
##  1st Qu.:0.8398   1st Qu.: 1.601   1st Qu.: 17.793   1st Qu.:0.005078
##  Median :1.1500   Median : 2.303   Median : 24.530   Median :0.006176
##  Mean   :1.2297   Mean   : 2.829   Mean   : 39.635   Mean   :0.006975
##  3rd Qu.:1.4785   3rd Qu.: 3.228   3rd Qu.: 44.852   3rd Qu.:0.008104
##  Max.   :4.8850   Max.   :18.650   Max.   :542.200   Max.   :0.031130
##  compactness_se      concavity_se     concave points_se   symmetry_se
##  Min.   :0.003012   Min.   :0.00000   Min.   :0.000000   Min.   :0.007882
```

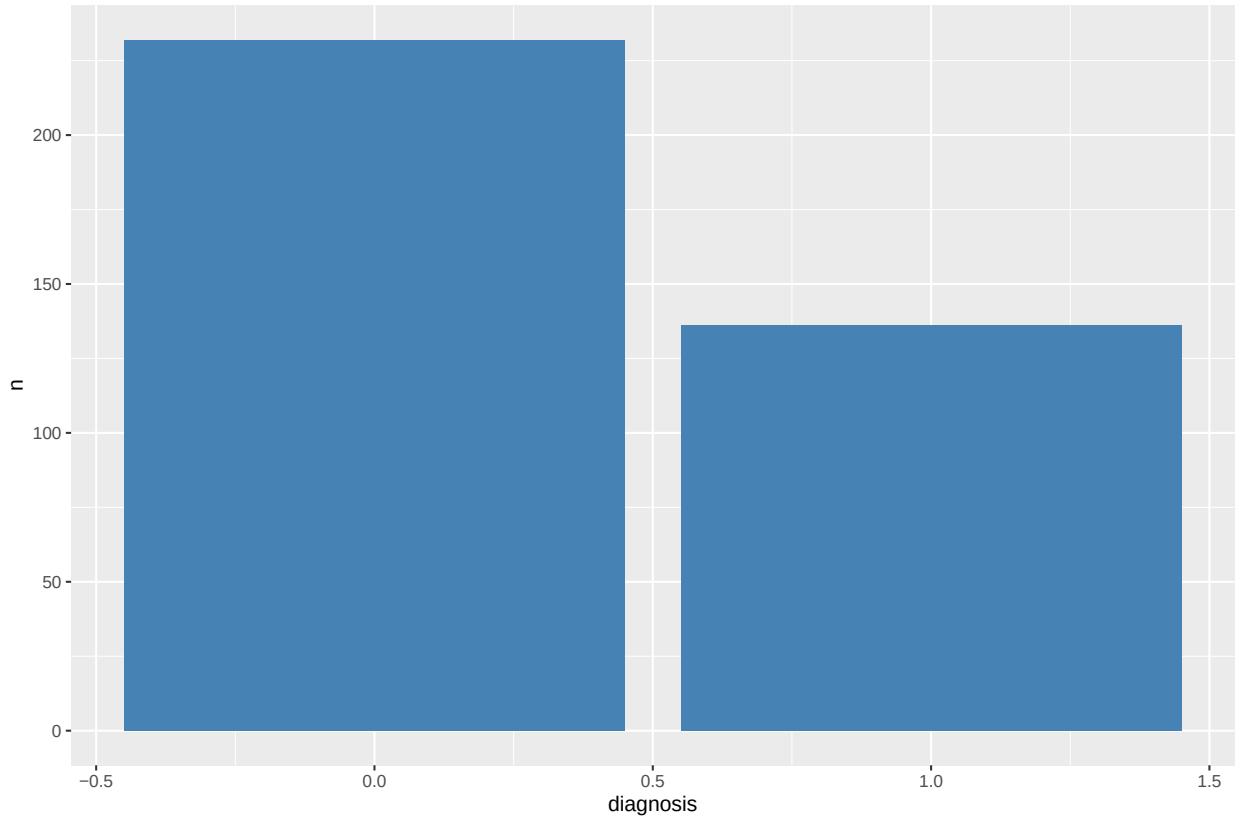
```
##  1st Qu.:0.013695   1st Qu.:0.01456   1st Qu.:0.007616   1st Qu.:0.015128
##  Median :0.021075   Median :0.02581   Median :0.010680   Median :0.018765
##  Mean   :0.025632   Mean   :0.03123   Mean   :0.011784   Mean   :0.020627
##  3rd Qu.:0.032745   3rd Qu.:0.04253   3rd Qu.:0.015012   3rd Qu.:0.023398
##  Max.   :0.135400   Max.   :0.30380   Max.   :0.040900   Max.   :0.061460
##  fractal_dimension_se  radius_worst    texture_worst   perimeter_worst
##  Min.   :0.0008948    Min.   : 8.952   Min.   :12.02   Min.   : 56.65
##  1st Qu.:0.0022870    1st Qu.:13.045   1st Qu.:21.29   1st Qu.: 84.28
##  Median :0.0032625    Median :14.875   Median :25.21   Median : 97.78
##  Mean   :0.0037782    Mean   :16.203   Mean   :25.60   Mean   :106.80
##  3rd Qu.:0.0045625    3rd Qu.:18.370   3rd Qu.:30.06   3rd Qu.:123.58
##  Max.   :0.0228600    Max.   :36.040   Max.   :49.54   Max.   :251.20
##    area_worst     smoothness_worst  compactness_worst concavity_worst
##  Min.   : 240.1   Min.   :0.07117   Min.   :0.02729   Min.   :0.0000
##  1st Qu.: 515.7   1st Qu.:0.11675   1st Qu.:0.15245   1st Qu.:0.1210
##  Median : 673.5   Median :0.13135   Median :0.21375   Median :0.2248
##  Mean   : 868.0   Mean   :0.13239   Mean   :0.25240   Mean   :0.2663
##  3rd Qu.:1037.2   3rd Qu.:0.14522   3rd Qu.:0.33930   3rd Qu.:0.3798
##  Max.   :4254.0   Max.   :0.22260   Max.   :0.93790   Max.   :1.2520
##  concave points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.06557      1st Qu.:0.2511   1st Qu.:0.07213
##  Median :0.09885      Median :0.2826   Median :0.08071
##  Mean   :0.11405      Mean   :0.2902   Mean   :0.08368
##  3rd Qu.:0.15890      3rd Qu.:0.3185   3rd Qu.:0.09169
##  Max.   :0.29100      Max.   :0.6638   Max.   :0.17300
```



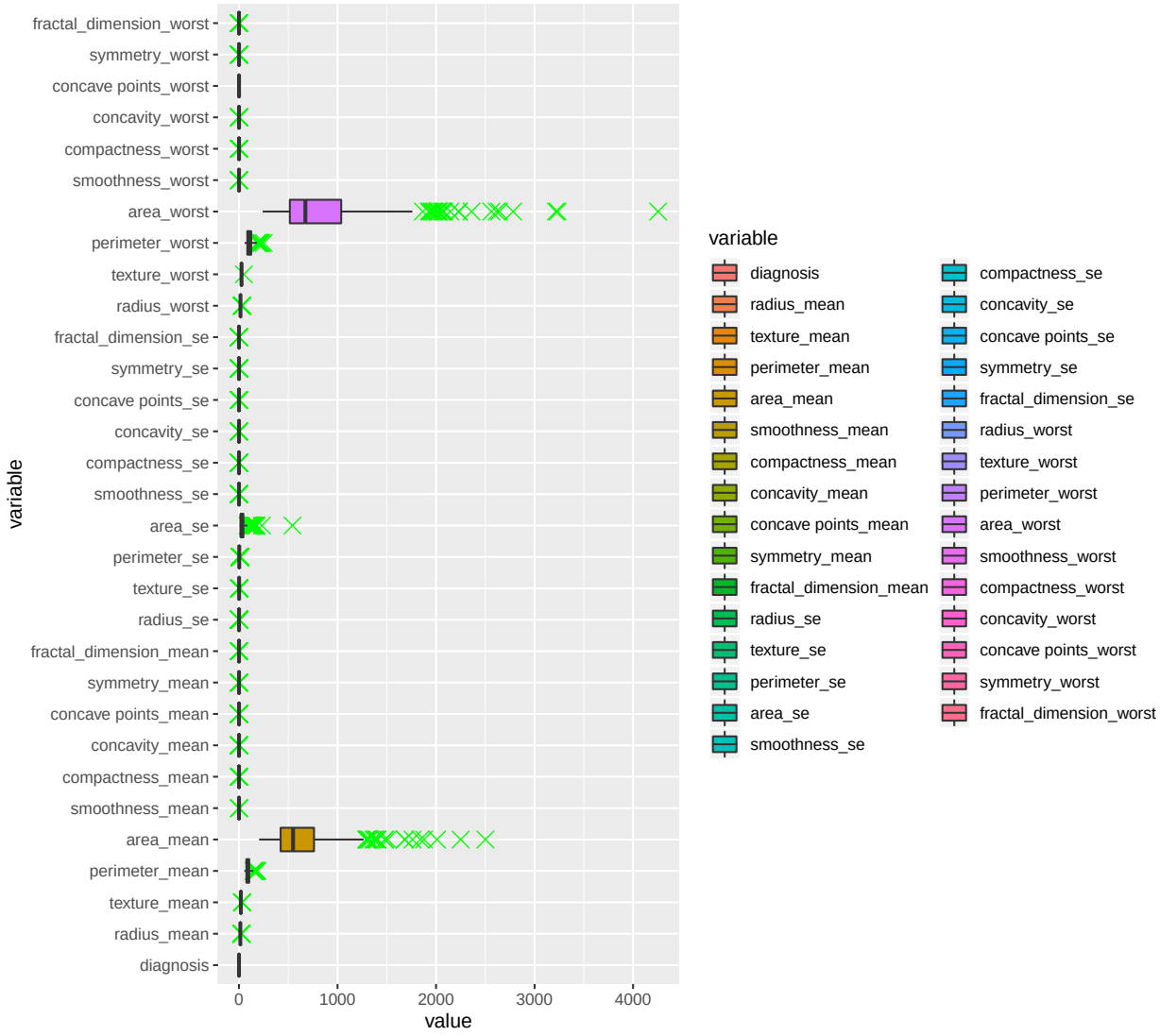
The first insight we can observe its that our dataset it's small with just 570 rows but all the observation is complete and we don't have missing values. in the histogram plot we can see that most of the variables nearly close to normal distributed especially the most significant variables one and others somehow a little skewed.



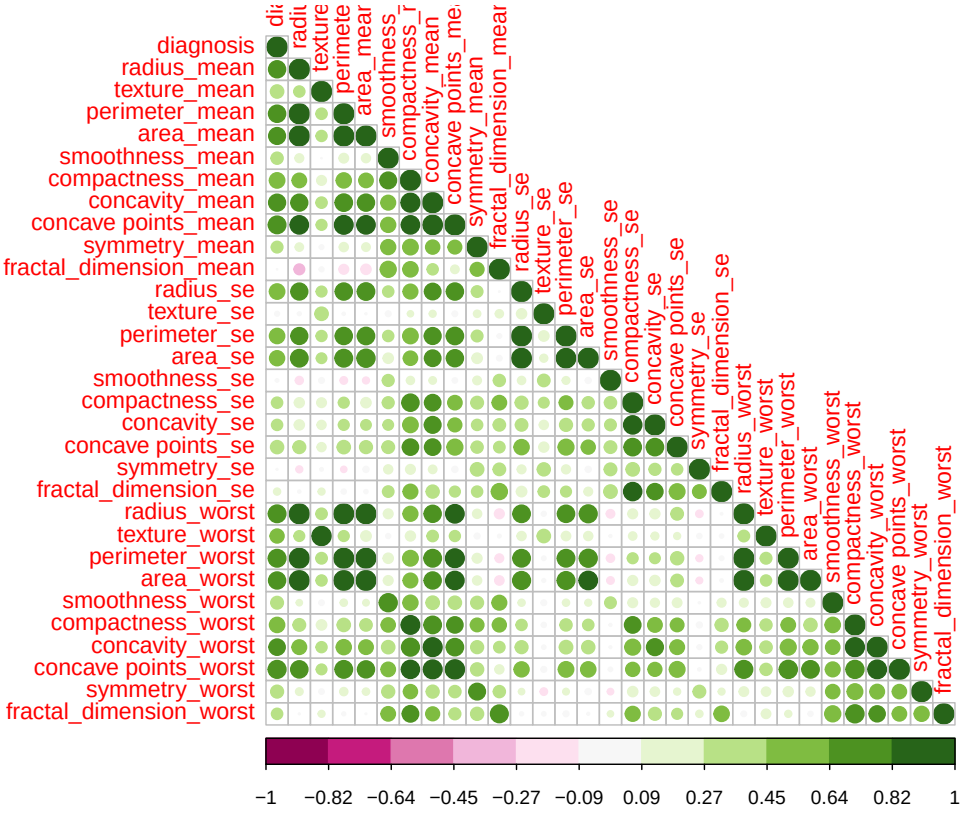





```
## No id variables; using all as measure variables
```




```
## Warning in brewer.pal(n = 12, name = "PiYG"): n too large, allowed maximum for palette PiYG is 11
## Returning the palette you asked for with that many colors
```



In the correlation plot we can observe they are many correlate variables for which this can be an issue and affect our modeling design. we going to use GLM modeling regression. we going to try to run a model with all the variables and then we going to implement a stepwise function in order to remove the correlate variables and enhance the model by obtaining one with just the most significant variables.

Modeling

```
##
## Call:
## lm(formula = diagnosis ~ ., data = df.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54827 -0.16468 -0.02668  0.14195  0.64714
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.564e+00  5.300e-01  -4.838    2e-06 ***
## radius_mean              -2.436e-01  2.245e-01  -1.085  0.27860
## texture_mean              1.840e-03  9.748e-03   0.189  0.85038
## perimeter_mean            4.169e-02  3.288e-02   1.268  0.20567
## area_mean                -3.697e-04  7.886e-04  -0.469  0.63948
## smoothness_mean           3.587e+00  2.475e+00   1.449  0.14824
## compactness_mean         -5.850e+00  1.679e+00  -3.484  0.00056 ***
## concavity_mean            1.662e+00  1.278e+00   1.301  0.19422
## `concave points_mean`     1.116e+00  2.411e+00   0.463  0.64373
## symmetry_mean             7.962e-01  8.724e-01   0.913  0.36208
```

```
## fractal_dimension_mean  -5.306e+00  7.049e+00  -0.753  0.45215
## radius_se                3.032e-01  3.914e-01   0.775  0.43914
## texture_se              -3.591e-02  4.438e-02  -0.809  0.41904
## perimeter_se             1.954e-02  5.104e-02   0.383  0.70215
## area_se                 -9.152e-04  1.849e-03  -0.495  0.62099
## smoothness_se            9.547e+00  8.079e+00   1.182  0.23816
## compactness_se          -6.434e-01  3.358e+00  -0.192  0.84817
## concavity_se            -3.081e+00  1.917e+00  -1.608  0.10882
## `concave points_se`      9.357e+00  6.617e+00   1.414  0.15828
## symmetry_se              7.871e+00  3.456e+00   2.277  0.02340 *
## fractal_dimension_se    -1.302e+01  1.639e+01  -0.794  0.42773
## radius_worst             2.033e-01  8.146e-02   2.496  0.01304 *
## texture_worst            1.401e-02  8.301e-03   1.687  0.09246 .
## perimeter_worst         -9.516e-03  7.826e-03  -1.216  0.22482
## area_worst              -8.728e-04  5.122e-04  -1.704  0.08930 .
## smoothness_worst         8.510e-02  1.761e+00   0.048  0.96149
## compactness_worst        3.241e-01  5.578e-01   0.581  0.56160
## concavity_worst          2.779e-01  3.887e-01   0.715  0.47519
## `concave points_worst`   9.876e-01  1.161e+00   0.850  0.39572
## symmetry_worst          -2.545e-01  6.025e-01  -0.422  0.67298
## fractal_dimension_worst  7.663e+00  3.326e+00   2.304  0.02184 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.232 on 337 degrees of freedom
## Multiple R-squared:  0.7884, Adjusted R-squared:  0.7696
## F-statistic: 41.86 on 30 and 337 DF,  p-value: < 2.2e-16


##
## Call:
## glm(formula = diagnosis ~ ., family = "binomial", data = df.train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -7.138e-05  -2.100e-08  -2.100e-08   2.100e-08   9.689e-05
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.119e+03  1.359e+06  -0.001    0.999
## radius_mean              1.687e+02  3.492e+05   0.000    1.000
## texture_mean            -5.011e+00  7.406e+03  -0.001    0.999
## perimeter_mean          -1.243e+01  4.432e+04   0.000    1.000
## area_mean               -7.243e-01  2.725e+03   0.000    1.000
## smoothness_mean          1.034e+03  3.243e+06   0.000    1.000
## compactness_mean        -8.985e+02  3.593e+06   0.000    1.000
## concavity_mean           2.069e+03  2.645e+06   0.001    0.999
## `concave points_mean`   -8.579e+02  3.952e+06   0.000    1.000
## symmetry_mean            7.357e+01  1.029e+06   0.000    1.000
## fractal_dimension_mean   2.785e+03  9.669e+06   0.000    1.000
## radius_se               -3.781e+02  8.915e+05   0.000    1.000
## texture_se              -1.166e+02  8.913e+04  -0.001    0.999
## perimeter_se             7.782e+01  6.248e+04   0.001    0.999
## area_se                  1.222e+00  8.794e+03   0.000    1.000
## smoothness_se            9.697e+03  2.109e+07   0.000    1.000
```

```
## compactness_se           1.550e+03  4.939e+06   0.000    1.000
## concavity_se            -1.294e+03  4.094e+06   0.000    1.000
## `concave points_se`      1.816e+03  2.095e+07   0.000    1.000
## symmetry_se             -1.167e+03  6.000e+06   0.000    1.000
## fractal_dimension_se    -2.962e+04  7.371e+07   0.000    1.000
## radius_worst             3.565e+01  2.494e+05   0.000    1.000
## texture_worst            1.353e+01  7.632e+03   0.002    0.999
## perimeter_worst         -1.197e+01  6.140e+03  -0.002    0.998
## area_worst               3.430e-01  2.424e+03   0.000    1.000
## smoothness_worst        -1.068e+03  3.669e+06   0.000    1.000
## compactness_worst       -2.014e+02  1.299e+06   0.000    1.000
## concavity_worst         -1.157e+02  4.475e+05   0.000    1.000
## `concave points_worst`   6.814e+02  1.698e+06   0.000    1.000
## symmetry_worst           4.278e+02  8.251e+05   0.001    1.000
## fractal_dimension_worst  3.088e+03  9.668e+06   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.8482e+02  on 367  degrees of freedom
## Residual deviance: 6.0596e-08  on 337  degrees of freedom
## AIC: 62
##
## Number of Fisher Scoring iterations: 25
```

```
##             radius_mean            texture_mean          perimeter_mean
##             15748.44212               75.69624            13264.43906
##               area_mean         smoothness_mean        compactness_mean
##              8456.51404               95.32661             1537.85717
##          concavity_mean   `concave points_mean`           symmetry_mean
##              1046.81925              539.75379               33.87372
##  fractal_dimension_mean               radius_se              texture_se
##               202.81641             1477.82063              150.36683
##            perimeter_se                area_se           smoothness_se
##               523.96968             1720.22186              106.34432
##          compactness_se            concavity_se     `concave points_se`
##               535.84836              674.75609              662.88494
##             symmetry_se    fractal_dimension_se            radius_worst
##                95.34593             1697.15462             9311.91355
##           texture_worst         perimeter_worst              area_worst
##               172.18668              369.64718             9618.43288
##        smoothness_worst       compactness_worst         concavity_worst
##               293.49460             1883.51196              302.71959
##  `concave points_worst`          symmetry_worst fractal_dimension_worst
##               243.96246               89.87959             1819.15228
```

```
##
## Call:
## glm(formula = diagnosis ~ ., family = binomial(link = "logit"),
##     data = df.train, trace = F)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -7.138e-05  -2.100e-08  -2.100e-08   2.100e-08   9.689e-05
##
```

```
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.119e+03  1.359e+06  -0.001    0.999
## radius_mean              1.687e+02  3.492e+05   0.000    1.000
## texture_mean            -5.011e+00  7.406e+03  -0.001    0.999
## perimeter_mean          -1.243e+01  4.432e+04   0.000    1.000
## area_mean               -7.243e-01  2.725e+03   0.000    1.000
## smoothness_mean          1.034e+03  3.243e+06   0.000    1.000
## compactness_mean        -8.985e+02  3.593e+06   0.000    1.000
## concavity_mean           2.069e+03  2.645e+06   0.001    0.999
## `concave points_mean`   -8.579e+02  3.952e+06   0.000    1.000
## symmetry_mean            7.357e+01  1.029e+06   0.000    1.000
## fractal_dimension_mean   2.785e+03  9.669e+06   0.000    1.000
## radius_se               -3.781e+02  8.915e+05   0.000    1.000
## texture_se              -1.166e+02  8.913e+04  -0.001    0.999
## perimeter_se             7.782e+01  6.248e+04   0.001    0.999
## area_se                  1.222e+00  8.794e+03   0.000    1.000
## smoothness_se            9.697e+03  2.109e+07   0.000    1.000
## compactness_se           1.550e+03  4.939e+06   0.000    1.000
## concavity_se            -1.294e+03  4.094e+06   0.000    1.000
## `concave points_se`      1.816e+03  2.095e+07   0.000    1.000
## symmetry_se             -1.167e+03  6.000e+06   0.000    1.000
## fractal_dimension_se    -2.962e+04  7.371e+07   0.000    1.000
## radius_worst             3.565e+01  2.494e+05   0.000    1.000
## texture_worst            1.353e+01  7.632e+03   0.002    0.999
## perimeter_worst         -1.197e+01  6.140e+03  -0.002    0.998
## area_worst               3.430e-01  2.424e+03   0.000    1.000
## smoothness_worst        -1.068e+03  3.669e+06   0.000    1.000
## compactness_worst       -2.014e+02  1.299e+06   0.000    1.000
## concavity_worst         -1.157e+02  4.475e+05   0.000    1.000
## `concave points_worst`   6.814e+02  1.698e+06   0.000    1.000
## symmetry_worst           4.278e+02  8.251e+05   0.001    1.000
## fractal_dimension_worst  3.088e+03  9.668e+06   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.8482e+02  on 367  degrees of freedom
## Residual deviance: 6.0596e-08  on 337  degrees of freedom
## AIC: 62
##
## Number of Fisher Scoring iterations: 25

##             radius_mean            texture_mean          perimeter_mean
##             15748.44212                75.69624             13264.43906
##               area_mean         smoothness_mean        compactness_mean
##              8456.51404                95.32661              1537.85717
##          concavity_mean   `concave points_mean`           symmetry_mean
##              1046.81925               539.75379                33.87372
##  fractal_dimension_mean              radius_se             texture_se
##               202.81641             1477.82063              150.36683
##            perimeter_se                area_se           smoothness_se
##               523.96968             1720.22186              106.34432
##          compactness_se            concavity_se     `concave points_se`
##               535.84836              674.75609               662.88494
```

```
##             symmetry_se    fractal_dimension_se            radius_worst
##                95.34593              1697.15462              9311.91355
##           texture_worst         perimeter_worst              area_worst
##              172.18668               369.64718              9618.43288
##        smoothness_worst       compactness_worst         concavity_worst
##              293.49460              1883.51196               302.71959
##  `concave points_worst`          symmetry_worst fractal_dimension_worst
##              243.96246               89.87959              1819.15228
```

```
##
## Call:
## glm(formula = diagnosis ~ concavity_mean + perimeter_se + smoothness_se +
##     concavity_se + texture_worst + perimeter_worst + area_worst +
##     `concave points_worst` + symmetry_worst, family = "binomial",
##     data = df.train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.982e-03  -2.000e-08  -2.000e-08   2.000e-08   2.647e-03
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3212.28   49333.92  -0.065    0.948
## concavity_mean          12627.95  195002.93   0.065    0.948
## perimeter_se              660.75   10764.30   0.061    0.951
## smoothness_se           64635.17  935717.40   0.069    0.945
## concavity_se           -24581.21  392285.53  -0.063    0.950
## texture_worst             136.11    2100.46   0.065    0.948
## perimeter_worst          -218.98    3352.27  -0.065    0.948
## area_worst                 17.78     269.14   0.066    0.947
## `concave points_worst`  22976.17  346316.75   0.066    0.947
## symmetry_worst          12729.76  195183.52   0.065    0.948
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4.8482e+02  on 367  degrees of freedom
## Residual deviance: 2.4451e-05  on 358  degrees of freedom
## AIC: 20
##
## Number of Fisher Scoring iterations: 25
```

```
##         concavity_mean           perimeter_se          smoothness_se
##              2224.6820              2535.5272               293.1970
##           concavity_se         texture_worst        perimeter_worst
##               927.1817              2652.3247             39019.2171
##             area_worst `concave points_worst`         symmetry_worst
##             19257.4881              3398.6951              3983.2079
```

```
##
## Call:
## glm(formula = diagnosis ~ texture_mean + compactness_mean + concavity_mean +
##     radius_se + area_se + smoothness_se + compactness_se + concavity_se +
##     `concave points_se` + radius_worst + area_worst + compactness_worst +
```

```
##     symmetry_worst + fractal_dimension_worst, data = df.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.57997  -0.15296  -0.03476   0.14902   0.71117
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.6018212  0.2363349 -11.009  < 2e-16 ***
## texture_mean             0.0170717  0.0032534   5.247 2.67e-07 ***
## compactness_mean        -4.6537004  0.8423352  -5.525 6.42e-08 ***
## concavity_mean           3.2240935  0.6128373   5.261 2.49e-07 ***
## radius_se                0.3925925  0.1904650   2.061 0.040013 *
## area_se                 -0.0008786  0.0013780  -0.638 0.524127
## smoothness_se           15.3960482  5.3463758   2.880 0.004223 **
## compactness_se          -3.8564944  1.7949419  -2.149 0.032351 *
## concavity_se            -3.5806993  1.2005270  -2.983 0.003057 **
## `concave points_se`     15.4361616  4.0337759   3.827 0.000154 ***
## radius_worst             0.1626584  0.0215023   7.565 3.40e-13 ***
## area_worst              -0.0010507  0.0002027  -5.184 3.66e-07 ***
## compactness_worst        0.6501844  0.2625606   2.476 0.013743 *
## symmetry_worst           1.0502906  0.2609726   4.025 6.99e-05 ***
## fractal_dimension_worst  5.4127475  1.5755455   3.435 0.000662 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.05445245)
##
##     Null deviance: 85.739  on 367  degrees of freedom
## Residual deviance: 19.222  on 353  degrees of freedom
## AIC: -10.013
##
## Number of Fisher Scoring iterations: 2

##            texture_mean        compactness_mean          concavity_mean
##                1.242884               13.257475               15.970499
##               radius_se                 area_se           smoothness_se
##                17.138708               23.583979                1.752302
##          compactness_se            concavity_se     `concave points_se`
##                6.801459                6.823981                4.037250
##            radius_worst              area_worst       compactness_worst
##                68.319526               83.556326               10.574976
##          symmetry_worst fractal_dimension_worst
##                1.795425                4.736530

##
## Call:
## glm(formula = diagnosis ~ radius_mean + perimeter_mean + compactness_mean +
##     `concave points_mean` + fractal_dimension_mean + radius_se +
##     perimeter_se + compactness_se + fractal_dimension_se + texture_worst +
##     perimeter_worst + concavity_worst + symmetry_worst, family = "binomial",
##     data = df.train)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -1.56369  -0.00022   0.00000   0.00000   2.51818
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -113.3328    55.2772  -2.050   0.0403 *
## radius_mean              12.9620    14.0446   0.923   0.3561
## perimeter_mean           -2.1005     2.1926  -0.958   0.3381
## compactness_mean       -181.9062   115.1628  -1.580   0.1142
## `concave points_mean`   469.9835   228.3478   2.058   0.0396 *
## fractal_dimension_mean  -52.6371   478.6441  -0.110   0.9124
## radius_se                11.1797    19.4267   0.575   0.5650
## perimeter_se              2.0698     3.0294   0.683   0.4945
## compactness_se         -362.8651   220.7315  -1.644   0.1002
## fractal_dimension_se   2059.0926  1384.5538   1.487   0.1370
## texture_worst             1.0678     0.4380   2.438   0.0148 *
## perimeter_worst           0.4992     0.3086   1.617   0.1058
## concavity_worst          28.7160    13.9103   2.064   0.0390 *
## symmetry_worst           91.3814    43.8143   2.086   0.0370 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 484.821  on 367  degrees of freedom
## Residual deviance:  22.415  on 354  degrees of freedom
## AIC: 50.415
##
## Number of Fisher Scoring iterations: 13
```

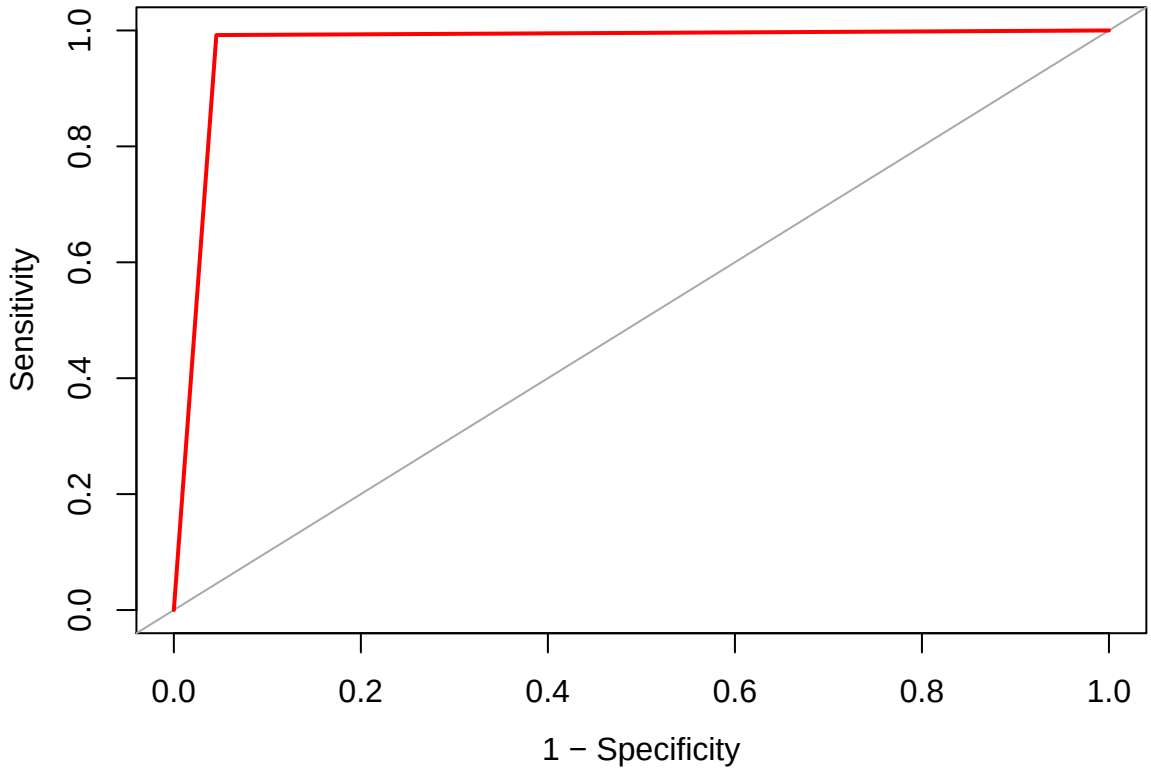
```
##            radius_mean         perimeter_mean       compactness_mean
##             2072.09528             2239.85188                88.40949
##  `concave points_mean` fractal_dimension_mean              radius_se
##               49.11191               33.62067               70.46288
##           perimeter_se         compactness_se   fractal_dimension_se
##               71.09107              131.23234              106.17452
##          texture_worst        perimeter_worst        concavity_worst
##               25.98067               53.25954               38.11723
##         symmetry_worst
##               35.90475
```

Select Model

For the modeling part, I work with 5 different GLM models on model 1 I did stepwise in order to see if I can just select the most significant variables. taking in consideration the AIC results I choose model number 4 because this has the lowest AIC values (29.489) for which means it has the most accurate training modeling.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 231  11
##          1   1 125
```

```
##
##                Accuracy : 0.9674
##                  95% CI : (0.9437, 0.983)
##     No Information Rate : 0.6304
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9289
##
##  Mcnemar's Test P-Value : 0.009375
##
##             Sensitivity : 0.9957
##             Specificity : 0.9191
##          Pos Pred Value : 0.9545
##          Neg Pred Value : 0.9921
##              Prevalence : 0.6304
##          Detection Rate : 0.6277
##    Detection Prevalence : 0.6576
##       Balanced Accuracy : 0.9574
##
##        'Positive' Class : 0
##
```




```
##
## Call:
## roc.default(response = df.train$target.pred, predictor = df.train$diagnosis,     plot = TRUE, asp = N
##
```

```
## Data: df.train$diagnosis in 242 controls (df.train$target.pred 0) < 126 cases (df.train$target.pred 1
## Area under the curve: 0.9733
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 70  1
##          1  1 41
##
##                Accuracy : 0.9823
##                  95% CI : (0.9375, 0.9978)
##     No Information Rate : 0.6283
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9621
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9859
##             Specificity : 0.9762
##          Pos Pred Value : 0.9859
##          Neg Pred Value : 0.9762
##              Prevalence : 0.6283
##          Detection Rate : 0.6195
##    Detection Prevalence : 0.6283
##       Balanced Accuracy : 0.9811
##
##        'Positive' Class : 0
##
```

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

```
## [1] 0.5224567 0.9463788 0.7905980 1.0065903 0.5402072 1.1607534
```

Discussion and Conclusion

As the final result, we conclude with a Glm model for which prediction rates of 95 %. we see that the most significant variables such as (texture_mean,compactness_mean,radius_worst,area_worst) to mention some has a big impact of diagnosed breast cancer by jus observing the cancer cell shape and dimensions. To further this analysis work could be very interesting to compare with other datasets or biggest numbers of observations dataset that contains other predictors variables such as (Age, countries, alcohol, etc) to have a more understanding of this illness.

References

1. Carol E. DeSantis MPH Jiemin Ma PhD Ann Goding Sauer MSPH Lisa A. Newman MD, MPH Ahmedin Jemal DVM, PhD , (03 October 2017). Retrieved from https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21412

2. Madhu Kumaria , Vijendra Singhb (2018) Breast Cancer Prediction system, retireved from https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050918X00088/1-s2.0-S1877050918309323/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEPH%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCICTgNtCQXc%2FkfBPOwfIkLlcuvmzEFHbFbfT4N5AyRvzRAiEAvpGq4DgUM%2BBAMgbmwkfBqbTDNbKCMkSUEnEZQTjY8eAqtAMISRADGgwwNTkwMDM1NDY4NjUiDCmf1UEGQgOS1Z%2BTACqRAwTi2DfRhDwHZ7EOaQn3dkOQF6RLEM000BYYQk6%2BEp4rldE0MdmHsPznoz%2BxrX85x2Dsb1OMCQwig%2FPr4VQe1ZWoPuX1zrmGHRdwLktlUsW1lCBxVCtsZlXkOfj3cUzMky9%2BB5u74hswWYiGUcMKibNsE4HnqQJteeC1FxKFvHpBuzDlsc1ykKqAEYHCGWDdbBKojOOaDqL2R3gbdjOYPUy2BfNIk53DDAZatcywb%2BFLK6OUgHlZtWnWUQ6iB1Ztv1afgrtDDMufEFuISoI23Vy3sOAYXX%2FYMFdXgwDzqDhMNtdIav8JOVz8%2BSqTiXwha06hd65ZZaqHNQDkmZwUC%2B64p4BidmawbFSTFCxWCDGYy2ByVfM0S876ADEpKgIiDmf%2BPOgfGI%2F4eb1tXfw2wtpBxWUoxc3gg1SSZJpNfdvwT1Msz6a%2F4bVTWiG5tPZipODwq0Kr%2BfZOm07X27CNAxwOjecXTfcuYfMLAmlwIhbQNfF13CcA4r4F0r%2Fdee87N7CbgPSkyOrsO08bk0nMJPzn%2FYFOusByU%2B9E%2FbkOS2rW5tRU%2FOsadeF7mfWntpc7GwtsV0V3rJ2FWP2U3bYyh13uLkQ9UnAVGIkHOKflYYTblJJmx2I2YKjr0jstpIZBMz1Bx7A4QpshwFvjKOayZT104OmmY%2BibjZY1u0%2FgVG69jEhlRLtKlZtk5y02OScOfRQR8cSx%2Fj9H%2Fqc6FolOOazFY%2BK2zO9oBzYIW6k0peJi96yTJ2FWQtuC%2BIW5WUHHM0le%2Ftk40m2kCMuwuvdECybybqlE72ONgXxF3WjQMFed8taflbhHIfJE6dkAahYHJSEA3D%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20200522T164047Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYZMTOQFXP%2F20200522%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=dbc61c841fbd5845aa6d120f880645d3f7b26b0010hash=ab881c72b5aba7662aa695aa75722923bfbfd50f8aa6f75380e6816ca4b6de42&host=68042c943591013ac2b2430a89b270pii=S1877050918309323&tid=spdf-788c59c6-5339-4dd6-b2bd-3d4ec3e6feef&sid=5541cdbe9d5bb64c1a8a63432e8af3b0btype=client

3. Dataset source retrieved from Kaggle/ UCI

.https://www.kaggle.com/vikasp/loadpred

.https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Appendix

```
knitr::opts_chunk$set(echo=FALSE)

library(caret)
library(dplyr)
library(psych)
library(corrplot)
library(tidyr)
library(ggplot2)
library(tidyverse)
library(DataExplorer)
library(RColorBrewer)
library(Amelia)
library(MASS)
library(car)
library(pROC)




data <- read_csv('data.csv')[,c(-1,-33)]
```

```r
index <-  data$diagnosis %>% createDataPartition(p = 0.8, list = FALSE, times = 1)


df <- data[index,]
eval.data <- data[-index,] # for final evaluation data

df$diagnosis <- ifelse(df$diagnosis=="M",1,0)
diagnosis <- data.matrix(df[,2])

index2 <-  diagnosis %>% createDataPartition(p = 0.8, list = FALSE, times = 1)

df.train <- df[index2,]
df.test <- df[-index2,]
diagnosis.train <- diagnosis[index2,]
diagnosis.test <- diagnosis[-index2,]

str(df.train)
summary(df.train)


missmap(data)

plot_histogram(df.train)
plt <- df.train %>% group_by(diagnosis) %>% count()

plt <- as.data.frame(plt)

p <- ggplot(data=plt, aes(x=diagnosis, y=n)) +
  geom_bar(stat="identity", fill="steelblue")
p


ggplot(data = reshape2::melt(df.train) , aes(x=variable, y=value)) +
  geom_boxplot(outlier.colour="green", outlier.shape=4, outlier.size=4,aes(fill=variable)) +
  coord_flip()
corrplot(cor(df.train, use = "na.or.complete"), type="lower",
         col=brewer.pal(n=12, name="PiYG"))


model0 <- lm(diagnosis ~.,data = df.train)
summary(model0)


model1 <-  glm(diagnosis ~., data = df.train,family = 'binomial')
summary(model1)
vif(model1)

model2 <- glm(diagnosis ~.,data = df.train, family = binomial(link = "logit"),  trace = F)
summary(model2)
vif(model2)

model3 <- stepAIC(model1, trace = F)
summary(model3)
```

```r
vif(model3)


model4 <- glm(formula = diagnosis ~ texture_mean + compactness_mean + concavity_mean +
    radius_se + area_se + smoothness_se + compactness_se + concavity_se +
    `concave points_se` + radius_worst + area_worst + compactness_worst +
    symmetry_worst + fractal_dimension_worst, data = df.train)

summary(model4)
vif(model4)



model5 <- glm(formula = diagnosis ~ radius_mean + perimeter_mean + compactness_mean +
    `concave points_mean` + fractal_dimension_mean + radius_se +
    perimeter_se + compactness_se + fractal_dimension_se + texture_worst +
    perimeter_worst + concavity_worst + symmetry_worst, family = "binomial",
    data = df.train)
summary(model5)
vif(model5)









model4 <- glm(formula = diagnosis ~ texture_mean + compactness_mean + concavity_mean +
    radius_se + area_se + smoothness_se + compactness_se + concavity_se +
    `concave points_se` + radius_worst + area_worst + compactness_worst +
    symmetry_worst + fractal_dimension_worst, data = df.train)


df.train$pred <- predict(model4, df.train, interval="response")
df.train$target.pred<- ifelse(df.train$pred >= 0.5, 1, 0)
confusionMatrix(factor(df.train$target.pred),factor(df.train$diagnosis))

roc.value <- roc(df.train$target.pred, df.train$diagnosis, plot=TRUE, asp=NA,
                 legacy.axes=TRUE, col="red")

roc.value

## Test model with evaluation data


eval.data$diagnosis <- ifelse(eval.data$diagnosis=="M",1,0)


eval.data$pre <- predict(model4, newdata = eval.data, interval="response")
eval.data$target.pred <- ifelse(eval.data$pre >= 0.5, 1, 0)
```

```r
confusionMatrix(factor(eval.data$target.pred),factor(eval.data$diagnosis))

head(eval.data$target.pred)
head(eval.data$pre)



library(caret)
library(dplyr)
library(psych)
library(corrplot)
library(tidyr)
library(ggplot2)
library(tidyverse)
library(DataExplorer)
library(RColorBrewer)
library(Amelia)
library(MASS)
library(car)
library(pROC)




data <- read_csv('data.csv')[,c(-1,-33)]

index <-  data$diagnosis %>% createDataPartition(p = 0.8, list = FALSE, times = 1)


df <- data[index,]
eval.data <- data[-index,] # for final evaluation data

df$diagnosis <- ifelse(df$diagnosis=="M",1,0)
diagnosis <- data.matrix(df[,2])

index2 <-  diagnosis %>% createDataPartition(p = 0.8, list = FALSE, times = 1)

df.train <- df[index2,]
df.test <- df[-index2,]
diagnosis.train <- diagnosis[index2,]
diagnosis.test <- diagnosis[-index2,]

str(df.train)
summary(df.train)


missmap(data)

plot_histogram(df.train)


plt <- df.train %>% group_by(diagnosis) %>% count()
```

```r
plt <- as.data.frame(plt)

p <- ggplot(data=plt, aes(x=diagnosis, y=n)) +
  geom_bar(stat="identity", fill="steelblue")
p

ggplot(data = reshape2::melt(df.train) , aes(x=variable, y=value)) +
  geom_boxplot(outlier.colour="green", outlier.shape=4, outlier.size=4,aes(fill=variable)) +
  coord_flip()


corrplot(cor(df.train, use = "na.or.complete"), type="lower",
         col=brewer.pal(n=12, name="PiYG"))




model0 <- lm(diagnosis ~.,data = df.train)
summary(model0)


model1 <-  glm(diagnosis ~., data = df.train,family = 'binomial')
summary(model1)
vif(model1)

model2 <- glm(diagnosis ~.,data = df.train, family = binomial(link = "logit"),  trace = F)
summary(model2)
vif(model2)

model3 <- stepAIC(model1, trace = F)
summary(model3)
vif(model3)


model4 <- glm(formula = diagnosis ~ texture_mean + compactness_mean + concavity_mean +
    radius_se + area_se + smoothness_se + compactness_se + concavity_se +
    `concave points_se` + radius_worst + area_worst + compactness_worst +
    symmetry_worst + fractal_dimension_worst, data = df.train)

summary(model4)
vif(model4)



model5 <- glm(formula = diagnosis ~ radius_mean + perimeter_mean + compactness_mean +
    `concave points_mean` + fractal_dimension_mean + radius_se +
    perimeter_se + compactness_se + fractal_dimension_se + texture_worst +
    perimeter_worst + concavity_worst + symmetry_worst, family = "binomial",
    data = df.train)
summary(model5)
vif(model5)
```

```r
model4 <- glm(formula = diagnosis ~ texture_mean + compactness_mean + concavity_mean +
    radius_se + area_se + smoothness_se + compactness_se + concavity_se +
    `concave points_se` + radius_worst + area_worst + compactness_worst +
    symmetry_worst + fractal_dimension_worst, data = df.train)


df.train$pred <- predict(model4, df.train, interval="response")
df.train$target.pred<- ifelse(df.train$pred >= 0.5, 1, 0)
confusionMatrix(factor(df.train$target.pred),factor(df.train$diagnosis))

roc.value <- roc(df.train$target.pred, df.train$diagnosis, plot=TRUE, asp=NA,
                 legacy.axes=TRUE, col="red")

roc.value

## Test model with evaluation data


eval.data$diagnosis <- ifelse(eval.data$diagnosis=="M",1,0)


eval.data$pre <- predict(model4, newdata = eval.data, interval="response")
eval.data$target.pred <- ifelse(eval.data$pre >= 0.5, 1, 0)

confusionMatrix(factor(eval.data$target.pred),factor(eval.data$diagnosis))

head(eval.data$target.pred)
head(eval.data$pre)
```