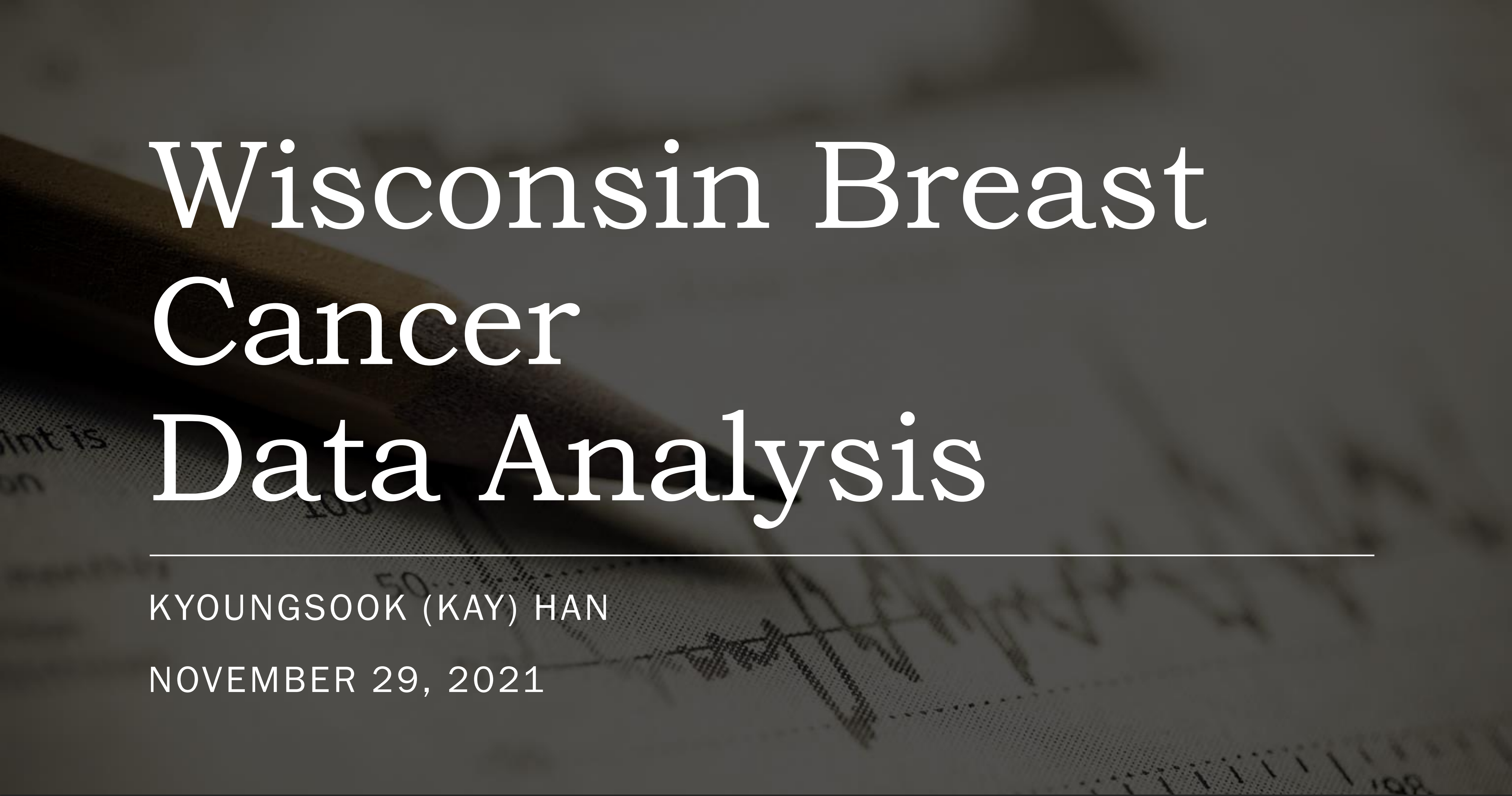


Wisconsin Breast Cancer Data Analysis

KYOUNGSOOK (KAY) HAN

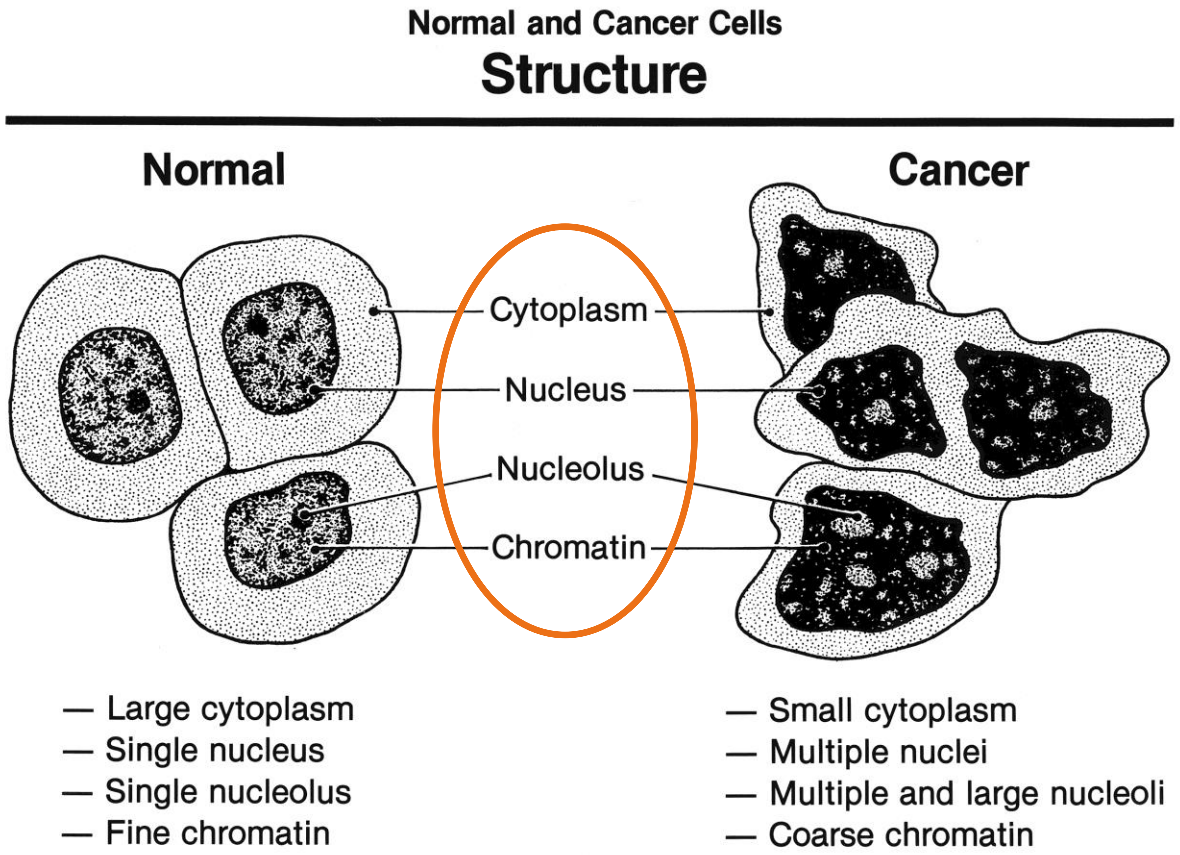
NOVEMBER 29, 2021

Background Information

About the Dataset

- The University of Wisconsin Hospitals in 1992
- Collected 699 observations consisting of 11 attributes
- The features are computed from a digitized image

About the Dataset



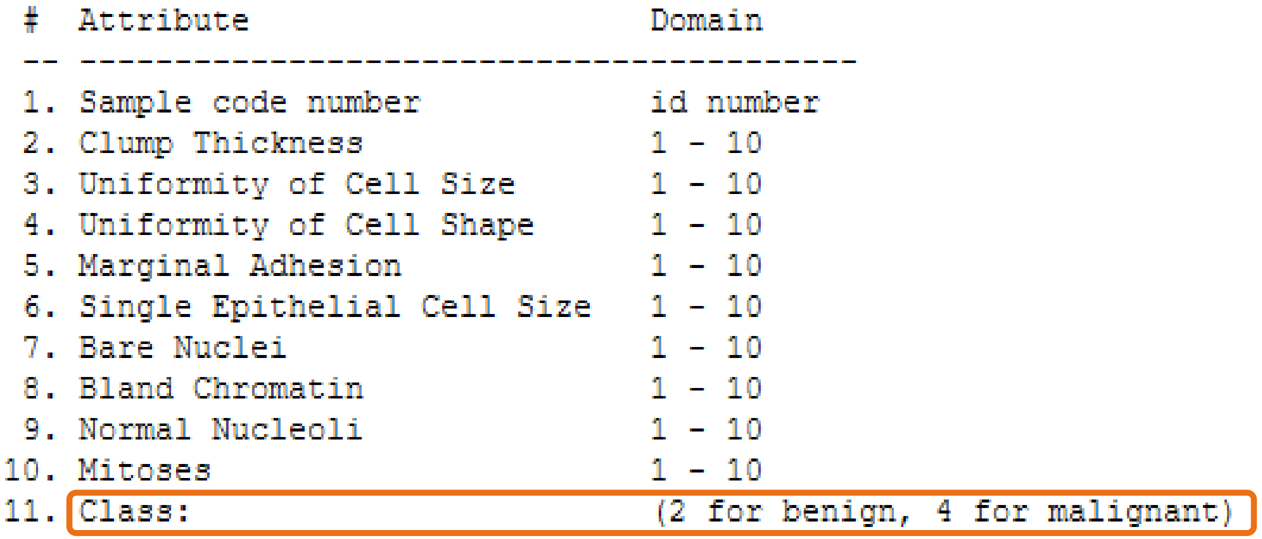
Normal and Cancer Cells

Structure

- Large cytoplasm
- Single nucleus
- Single nucleolus
- Fine chromatin

- Small cytoplasm
- Multiple nuclei
- Multiple and large nucleoli
- Coarse chromatin

<Source> *https://visualsonline.cancer.gov/*



```
 #  Attribute                     Domain
 -- -----------------------------------------
 1. Sample code number            id number
 2. Clump Thickness               1 - 10
 3. Uniformity of Cell Size       1 - 10
 4. Uniformity of Cell Shape      1 - 10
 5. Marginal Adhesion             1 - 10
 6. Single Epithelial Cell Size   1 - 10
 7. Bare Nuclei                   1 - 10
 8. Bland Chromatin               1 - 10
 9. Normal Nucleoli               1 - 10
10. Mitoses                       1 - 10
11. Class:                        (2 for benign, 4 for malignant)
```

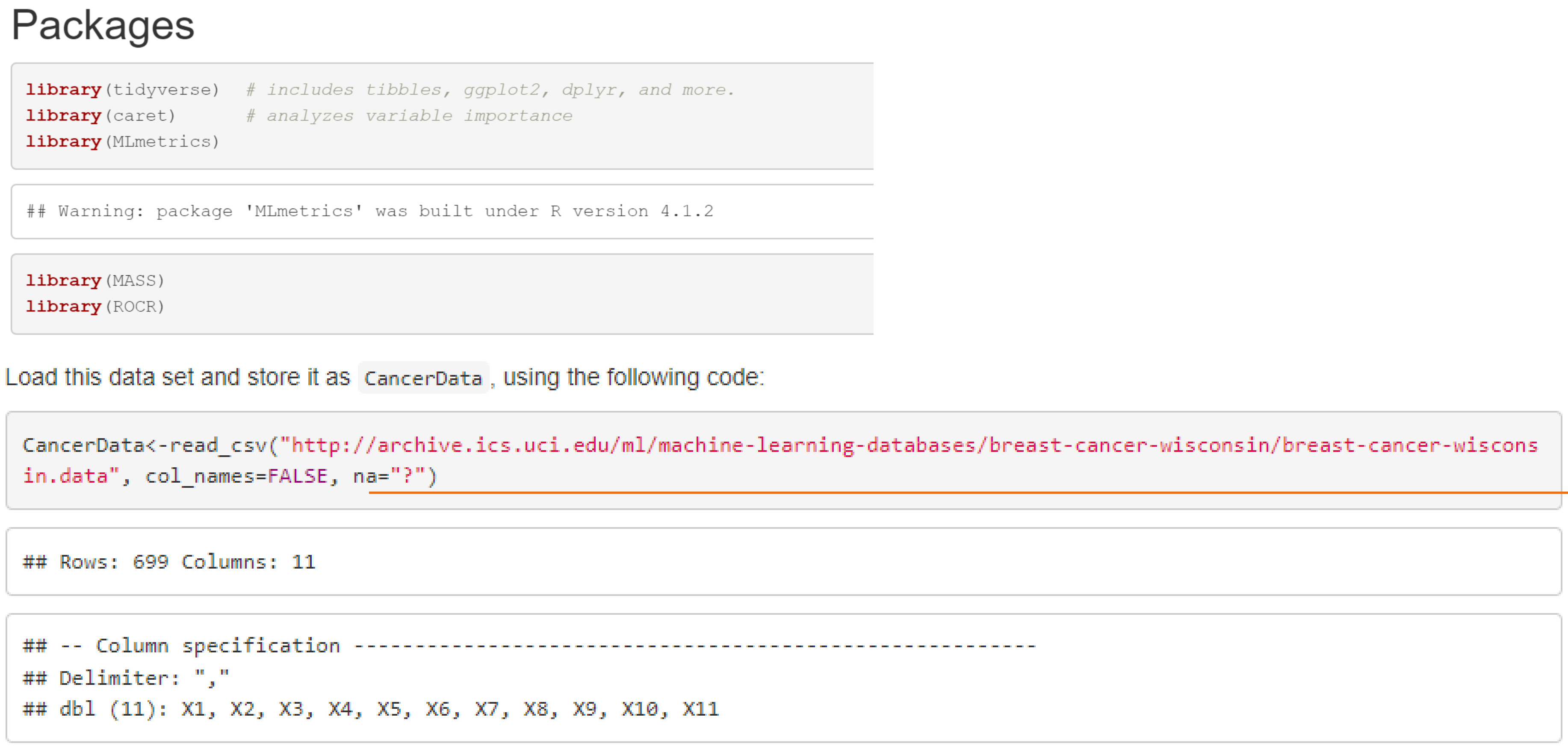
<Source> *https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)*

About the Dataset

- Except class, all other attributes are numeric type
- Class attribute indicates Benign or Malignant
  - **Benign:** A tumor which can usually be removed without serious complications and will not be fatal to the patient.
  - **Malignant:** Cancerous cells have ability to spread to other sites in the body.

An Exploratory Data Analysis

Step (1) Prepares the Dataset



Packages

```
library(tidyverse)  # includes tibbles, ggplot2, dplyr, and more.
library(caret)      # analyzes variable importance
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 4.1.2
```

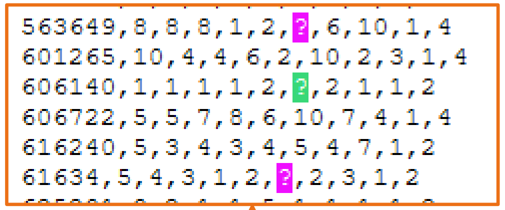
```
library(MASS)
library(ROCR)
```

Load this data set and store it as `CancerData`, using the following code:

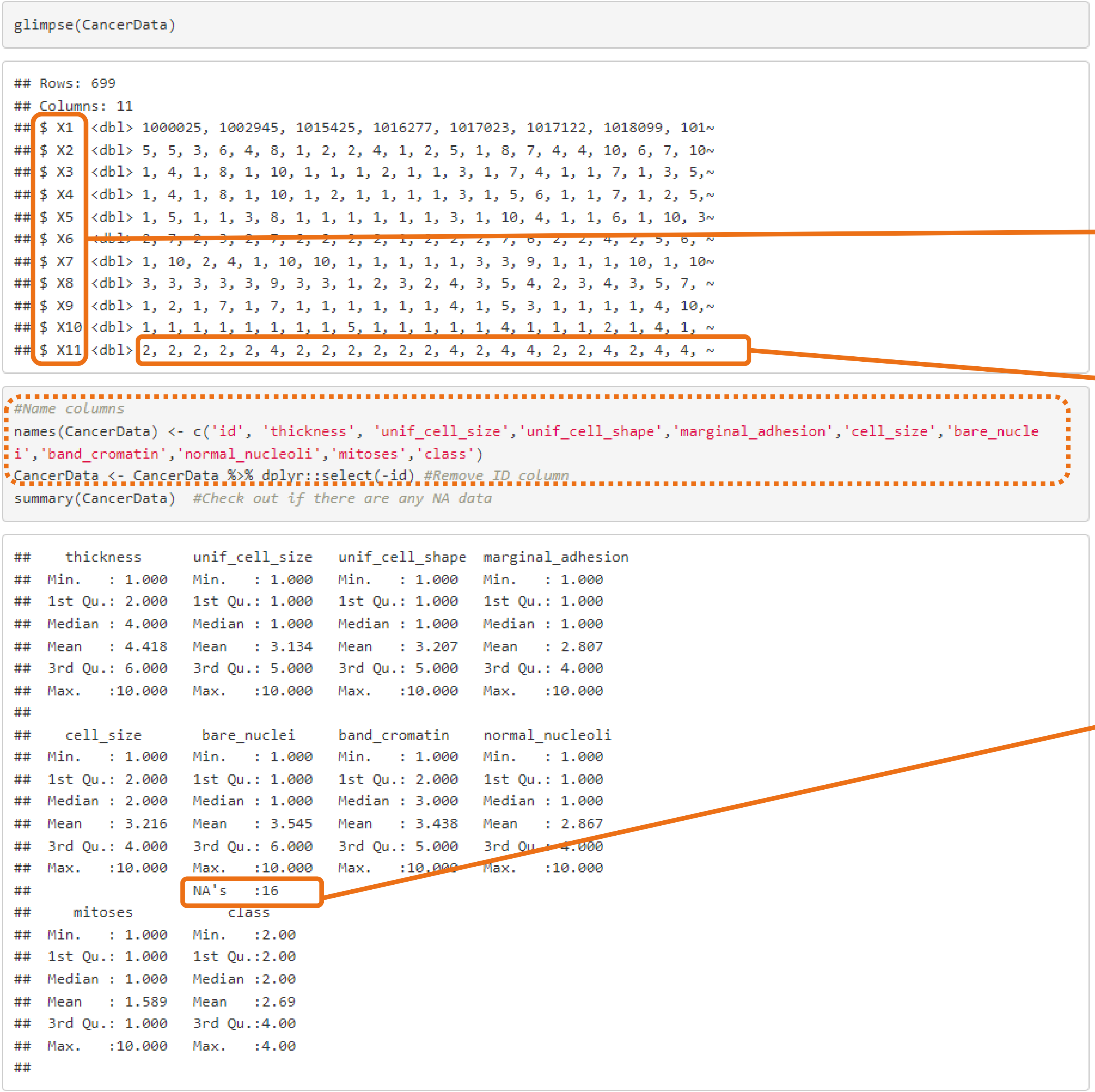
```
CancerData<-read_csv("http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wiscons
in.data", col_names=FALSE, na="?")
```

```
## Rows: 699 Columns: 11
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (11): X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11
```



```
563649,8,8,8,1,2,?,6,10,1,4
601265,10,4,4,6,2,10,2,3,1,4
606140,1,1,1,1,2,?,2,1,1,2
606722,5,5,7,8,6,10,7,4,1,4
616240,5,3,4,3,4,5,4,7,1,2
61634,5,4,3,1,2,?,2,3,1,2
```



```
glimpse(CancerData)
```

```
## Rows: 699
## Columns: 11
## $ X1  <dbl> 1000025, 1002945, 1015425, 1016277, 1017023, 1017122, 1018099, 101~
## $ X2  <dbl> 5, 5, 3, 6, 4, 8, 1, 2, 2, 4, 1, 2, 5, 1, 8, 7, 4, 4, 10, 6, 7, 10~
## $ X3  <dbl> 1, 4, 1, 8, 1, 10, 1, 1, 1, 2, 1, 1, 3, 1, 7, 4, 1, 1, 7, 1, 3, 5,~
## $ X4  <dbl> 1, 4, 1, 8, 1, 10, 1, 2, 1, 1, 1, 1, 3, 1, 5, 6, 1, 1, 7, 1, 2, 5,~
## $ X5  <dbl> 1, 5, 1, 1, 3, 8, 1, 1, 1, 1, 1, 1, 3, 1, 10, 4, 1, 1, 6, 1, 10, 3~
## $ X6  <dbl> 2, 7, 2, 3, 2, 7, 2, 2, 2, 2, 1, 2, 2, 2, 7, 6, 2, 2, 4, 2, 5, 6, ~
## $ X7  <dbl> 1, 10, 2, 4, 1, 10, 10, 1, 1, 1, 1, 1, 3, 3, 9, 1, 1, 1, 10, 1, 10~
## $ X8  <dbl> 3, 3, 3, 3, 3, 9, 3, 3, 1, 2, 3, 2, 4, 3, 5, 4, 2, 3, 4, 3, 5, 7, ~
## $ X9  <dbl> 1, 2, 1, 7, 1, 7, 1, 1, 1, 1, 1, 1, 4, 1, 5, 3, 1, 1, 1, 1, 4, 10,~
## $ X10 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, 1, 1, 4, 1, 1, 1, 2, 1, 4, 1, ~
## $ X11 <dbl> 2, 2, 2, 2, 2, 4, 2, 2, 2, 2, 2, 2, 4, 2, 4, 4, 2, 2, 4, 2, 4, 4, ~
```

```
#Name columns
names(CancerData) <- c('id', 'thickness', 'unif_cell_size','unif_cell_shape','marginal_adhesion','cell_size','bare_nucle
i','band_cromatin','normal_nucleoli','mitoses','class')
CancerData <- CancerData %>% dplyr::select(-id) #Remove ID column
summary(CancerData)  #Check out if there are any NA data
```

```
##    thickness      unif_cell_size   unif_cell_shape  marginal_adhesion
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000   
##  Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.000   
##  Mean   : 4.418   Mean   : 3.134   Mean   : 3.207   Mean   : 2.807   
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000   
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   
##                                                                      
##    cell_size       bare_nuclei      band_cromatin    normal_nucleoli 
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   
##  Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.000   
##  Mean   : 3.216   Mean   : 3.545   Mean   : 3.438   Mean   : 2.867   
##  3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.000   
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   
##                   NA's   :16                                         
##     mitoses           class     
##  Min.   : 1.000   Min.   :2.00  
##  1st Qu.: 1.000   1st Qu.:2.00  
##  Median : 1.000   Median :2.00  
##  Mean   : 1.589   Mean   :2.69  
##  3rd Qu.: 1.000   3rd Qu.:4.00  
##  Max.   :10.000   Max.   :4.00  
## 
```

1 Columns

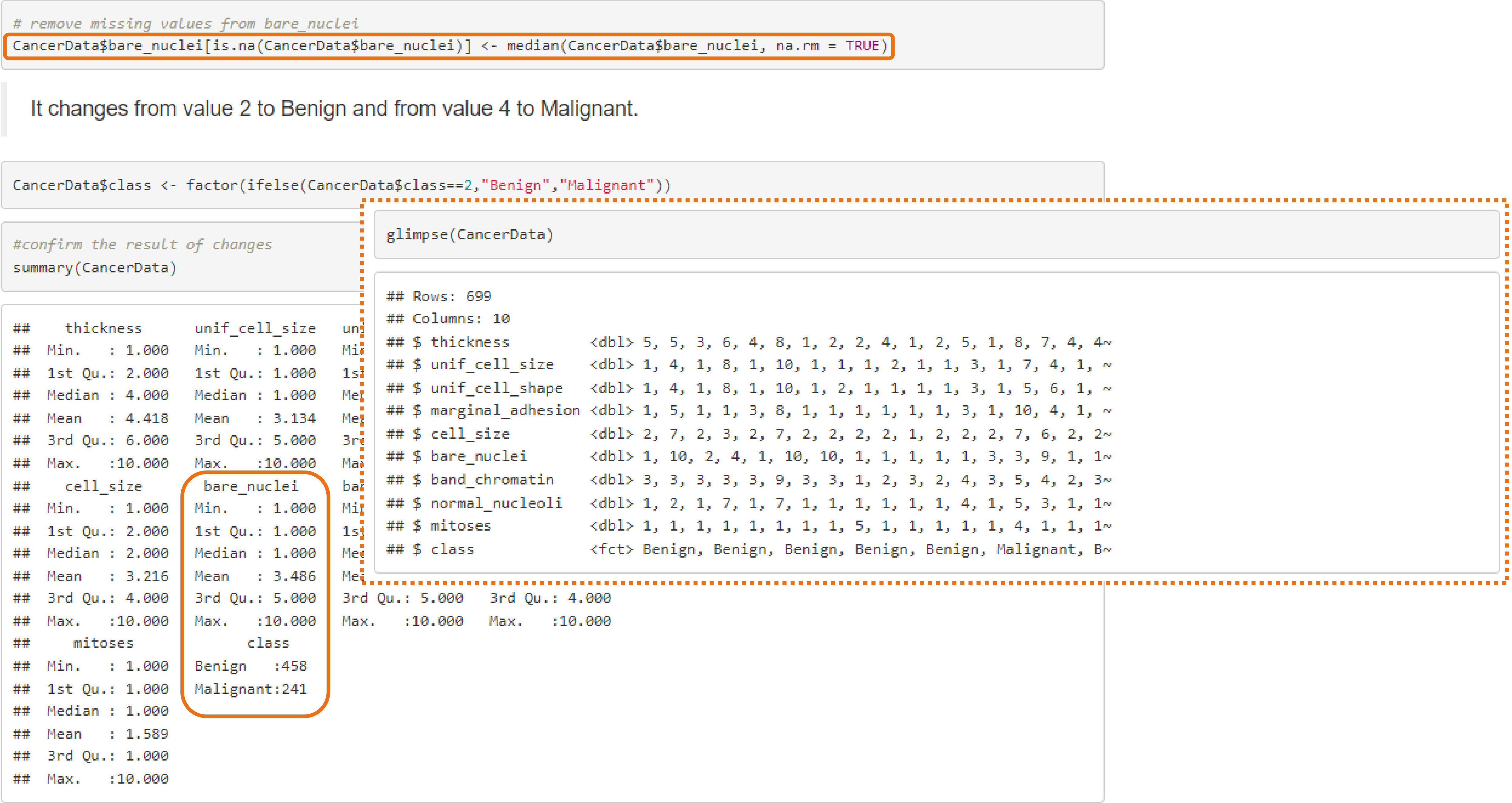
- Original data has no column names

2 Class

- Benign: 2
- Malignant: 4

3 Bare nuclei

- 16 missing values in the column (NA's 16)



```r
# remove missing values from bare_nuclei
CancerData$bare_nuclei[is.na(CancerData$bare_nuclei)] <- median(CancerData$bare_nuclei, na.rm = TRUE)
```

> It changes from value 2 to Benign and from value 4 to Malignant.

```r
CancerData$class <- factor(ifelse(CancerData$class==2,"Benign","Malignant"))
```

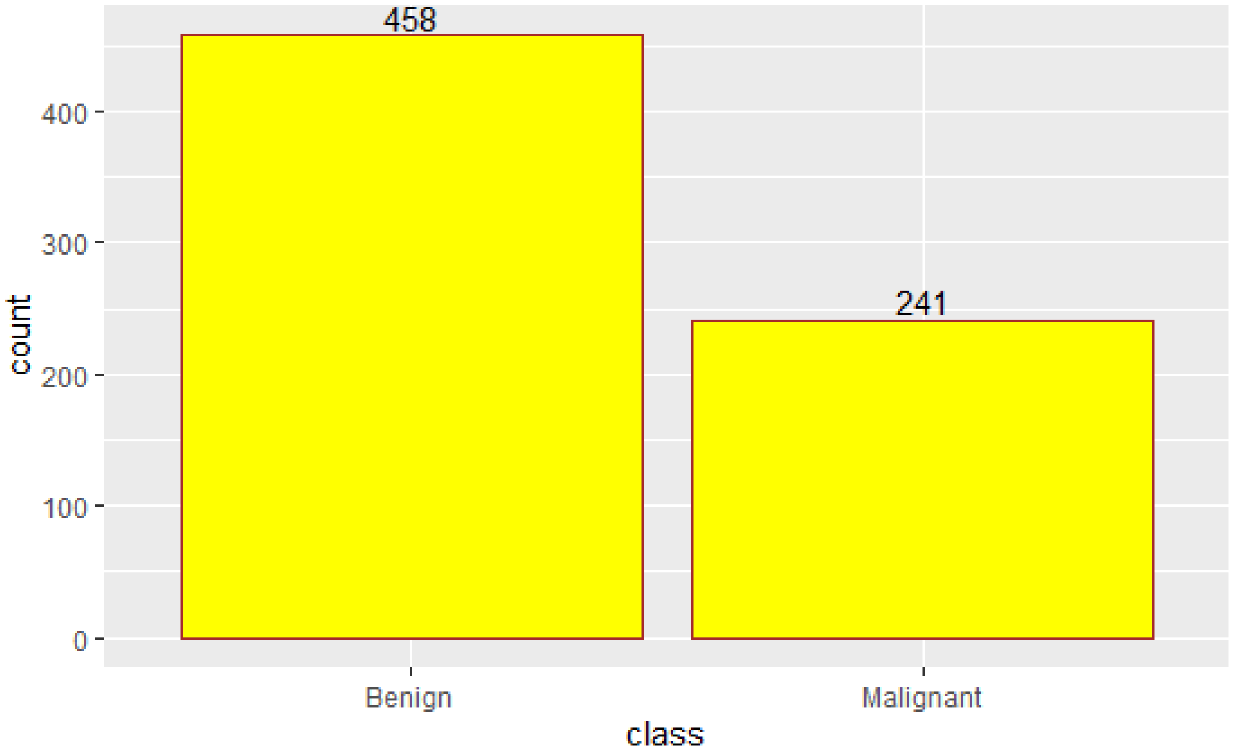
```r
#confirm the result of changes
summary(CancerData)
```

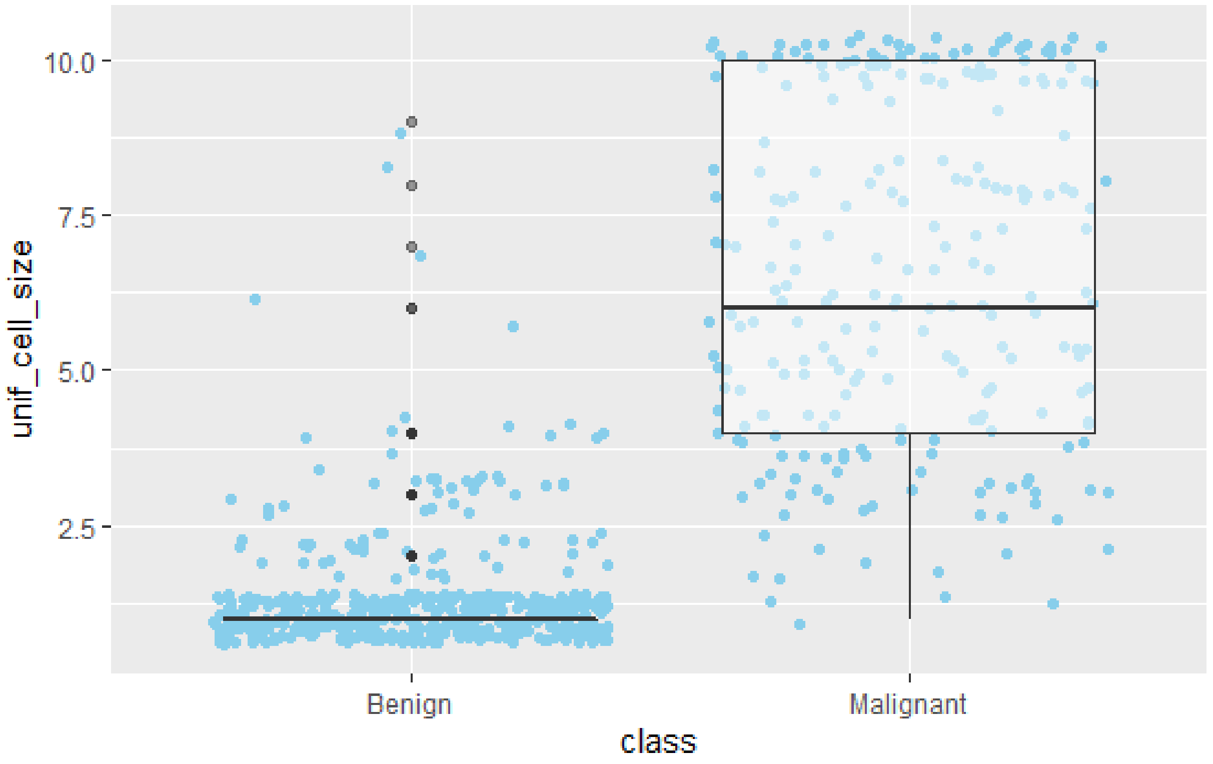
```
##    thickness      unif_cell_size   un
##  Min.   : 1.000   Min.   : 1.000   Mi
##  1st Qu.: 2.000   1st Qu.: 1.000   1s
##  Median : 4.000   Median : 1.000   Me
##  Mean   : 4.418   Mean   : 3.134   Me
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3r
##  Max.   :10.000   Max.   :10.000   Ma
##    cell_size       bare_nuclei     ba
##  Min.   : 1.000   Min.   : 1.000   Mi
##  1st Qu.: 2.000   1st Qu.: 1.000   1s
##  Median : 2.000   Median : 1.000   Me
##  Mean   : 3.216   Mean   : 3.486   Me
##  3rd Qu.: 4.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##     mitoses             class
##  Min.   : 1.000   Benign   :458
##  1st Qu.: 1.000   Malignant:241
##  Median : 1.000
##  Mean   : 1.589
##  3rd Qu.: 1.000
##  Max.   :10.000
```

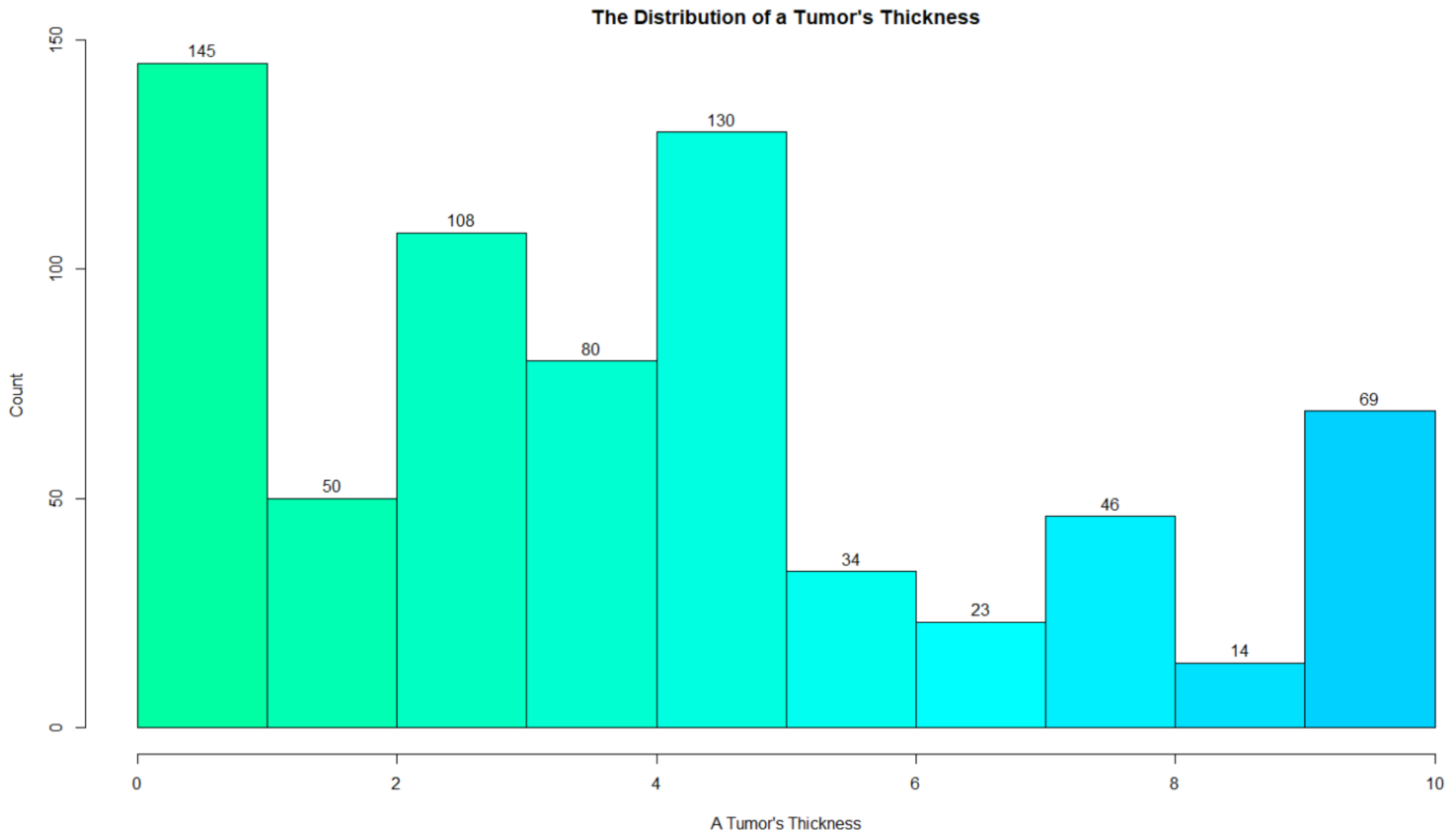
```r
glimpse(CancerData)
```

```
## Rows: 699
## Columns: 10
## $ thickness         <dbl> 5, 5, 3, 6, 4, 8, 1, 2, 2, 4, 1, 2, 5, 1, 8, 7, 4, 4~
## $ unif_cell_size    <dbl> 1, 4, 1, 8, 1, 10, 1, 1, 1, 2, 1, 1, 3, 1, 7, 4, 1, ~
## $ unif_cell_shape   <dbl> 1, 4, 1, 8, 1, 10, 1, 2, 1, 1, 1, 1, 3, 1, 5, 6, 1, ~
## $ marginal_adhesion <dbl> 1, 5, 1, 1, 3, 8, 1, 1, 1, 1, 1, 1, 3, 1, 10, 4, 1, ~
## $ cell_size         <dbl> 2, 7, 2, 3, 2, 7, 2, 2, 2, 2, 1, 2, 2, 2, 7, 6, 2, 2~
## $ bare_nuclei       <dbl> 1, 10, 2, 4, 1, 10, 10, 1, 1, 1, 1, 1, 3, 3, 9, 1, 1~
## $ band_chromatin    <dbl> 3, 3, 3, 3, 3, 9, 3, 3, 1, 2, 3, 2, 4, 3, 5, 4, 2, 3~
## $ normal_nucleoli   <dbl> 1, 2, 1, 7, 1, 7, 1, 1, 1, 1, 1, 1, 4, 1, 5, 3, 1, 1~
## $ mitoses           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, 1, 1, 4, 1, 1, 1~
## $ class             <fct> Benign, Benign, Benign, Benign, Benign, Malignant, B~
```

Step (2) Data Visualization







Step (2) Data Visualization (Cont.)

Hypothesis Testing

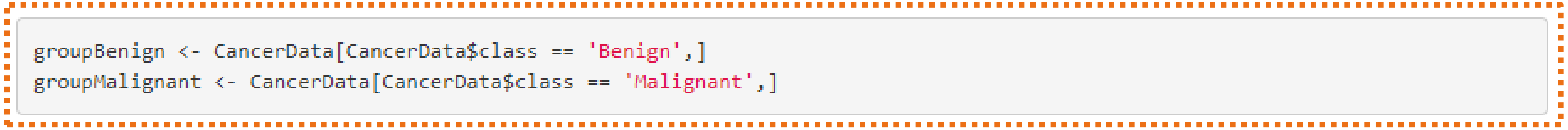
Step (1) Set Up a Null Hypothesis

$H_0 : \mu_{1=}\mu_2$ $H_1 : \mu_{1\neq}\mu_2$

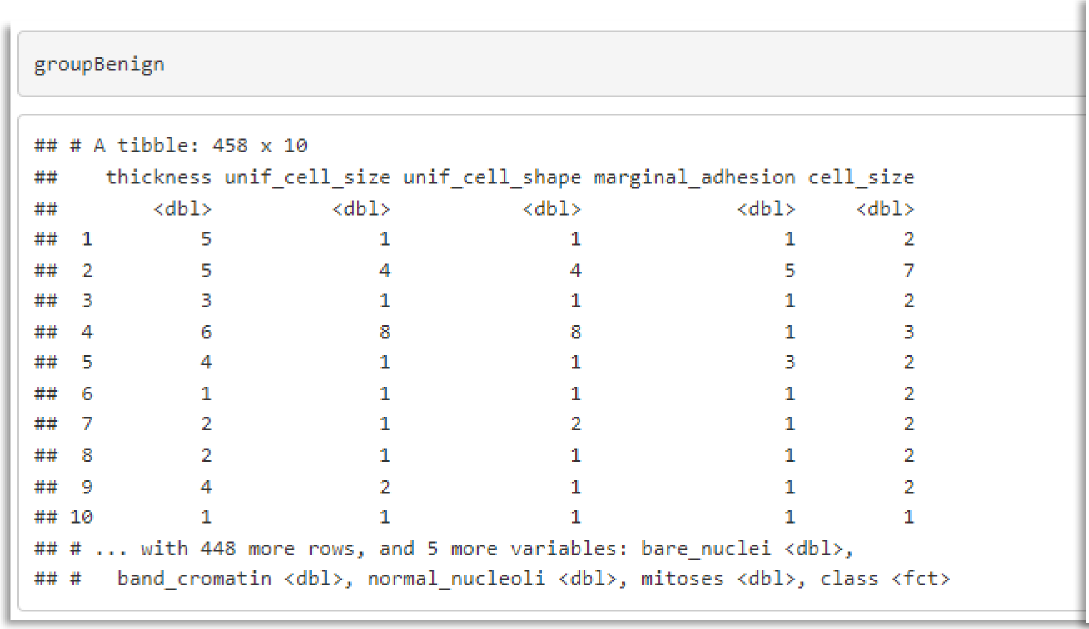
- The means of thickness from Benign and Malignant groups
- H_0: The two means of thickness are same.
- H_1: The two means of thickness are not same.

Step (2) Data Extract

- Divided into two groups by class attribute

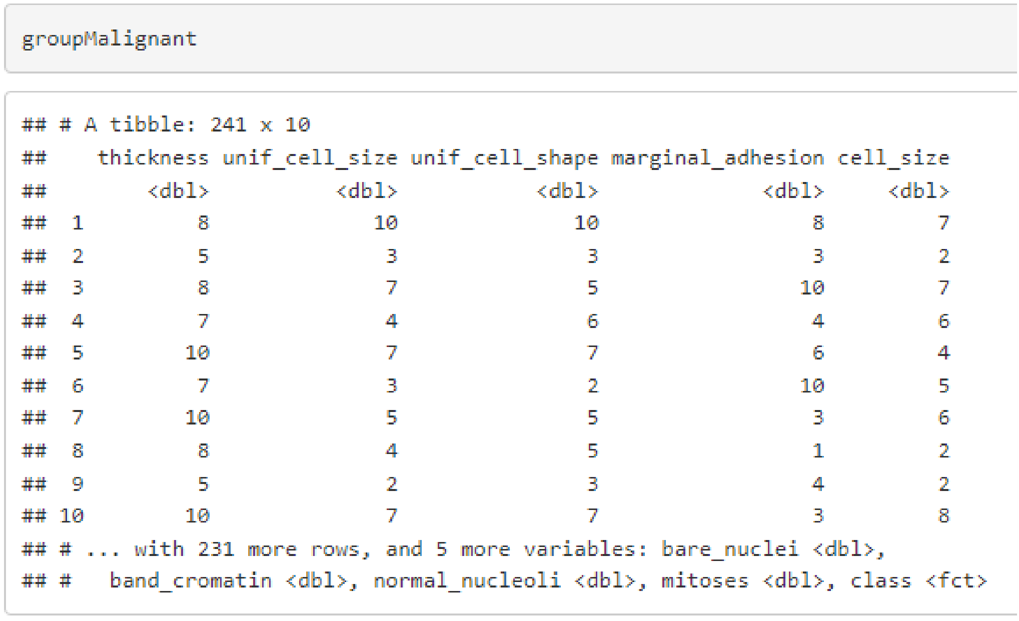


```
groupBenign <- CancerData[CancerData$class == 'Benign',]
groupMalignant <- CancerData[CancerData$class == 'Malignant',]
```



```
groupBenign
```

```
## # A tibble: 458 x 10
##    thickness unif_cell_size unif_cell_shape marginal_adhesion cell_size
##        <dbl>          <dbl>           <dbl>             <dbl>     <dbl>
##  1         5              1               1                 1         2
##  2         5              4               4                 5         7
##  3         3              1               1                 1         2
##  4         6              8               8                 1         3
##  5         4              1               1                 3         2
##  6         1              1               1                 1         2
##  7         2              1               2                 1         2
##  8         2              1               1                 1         2
##  9         4              2               1                 1         2
## 10         1              1               1                 1         1
## # ... with 448 more rows, and 5 more variables: bare_nuclei <dbl>,
## #   band_cromatin <dbl>, normal_nucleoli <dbl>, mitoses <dbl>, class <fct>
```

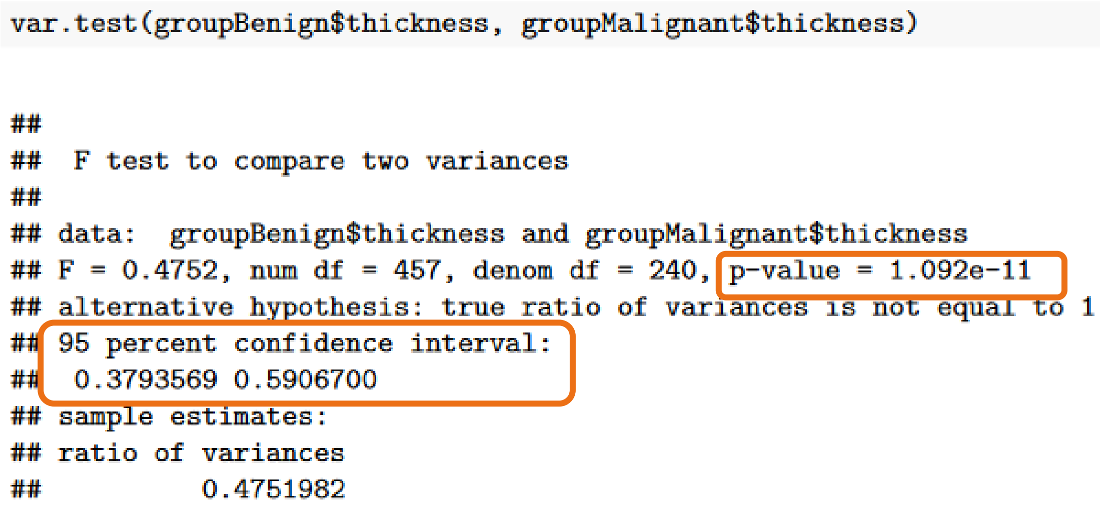


```
groupMalignant
```

```
## # A tibble: 241 x 10
##    thickness unif_cell_size unif_cell_shape marginal_adhesion cell_size
##        <dbl>          <dbl>           <dbl>             <dbl>     <dbl>
##  1         8             10              10                 8         7
##  2         5              3               3                 3         2
##  3         8              7               5                10         7
##  4         7              4               6                 4         6
##  5        10              7               7                 6         4
##  6         7              3               2                10         5
##  7        10              5               5                 3         6
##  8         8              4               5                 1         2
##  9         5              2               3                 4         2
## 10        10              7               7                 3         8
## # ... with 231 more rows, and 5 more variables: bare_nuclei <dbl>,
## #   band_cromatin <dbl>, normal_nucleoli <dbl>, mitoses <dbl>, class <fct>
```

Step (3) F-test

- Assumed the variances are unknown and equal.
- Conducted a F-test first to compare two variances at $\alpha = 0.05$



```
var.test(groupBenign$thickness, groupMalignant$thickness)
```

```
##
##  F test to compare two variances
##
## data:  groupBenign$thickness and groupMalignant$thickness
## F = 0.4752, num df = 457, denom df = 240, p-value = 1.092e-11
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3793569 0.5906700
## sample estimates:
## ratio of variances
##          0.4751982
```

P-value < α=0.05

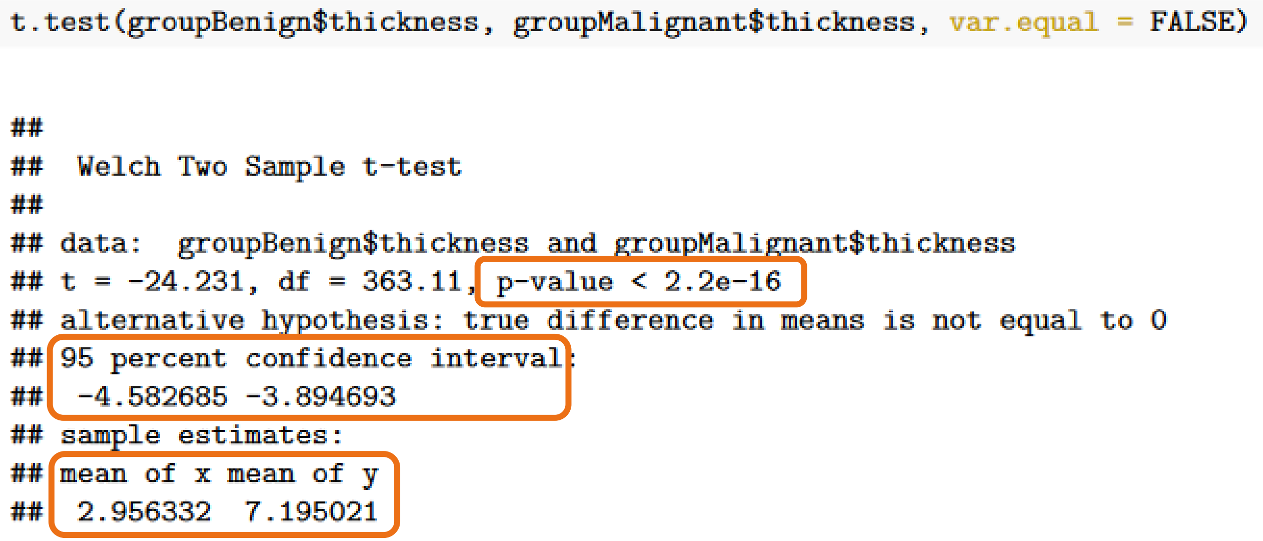
- 1.092×10^{-11}
- 0.00000000001092

Reject H_0 and Select H_1

- Select the alternative hypothesis
- The confidence interval (0.38, 0.59)
- There is a difference in the two variances

Step (4) T-test

- Assumed the variances are not equal based on the F-test.
- Assumed the two samples are normally distributed and independent.



```
t.test(groupBenign$thickness, groupMalignant$thickness, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  groupBenign$thickness and groupMalignant$thickness
## t = -24.231, df = 363.11, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.582685 -3.894693
## sample estimates:
## mean of x mean of y
##  2.956332  7.195021
```

P-value $< \alpha = 0.05$

- 2.2×10^{-16} (0.00000000000000022)

Reject H_0 *and Select* H_1

- Select the alternative hypothesis
- The confidence interval (-4.58, -3.89)
- The mean of Benign is 2.96
- The mean of Malignant is 7.20

Step (5) Conclusion

- Reject H_0 and Select H_1
- Malignant cells are much thicker than Benign cells.

Linear Regression & Prediction

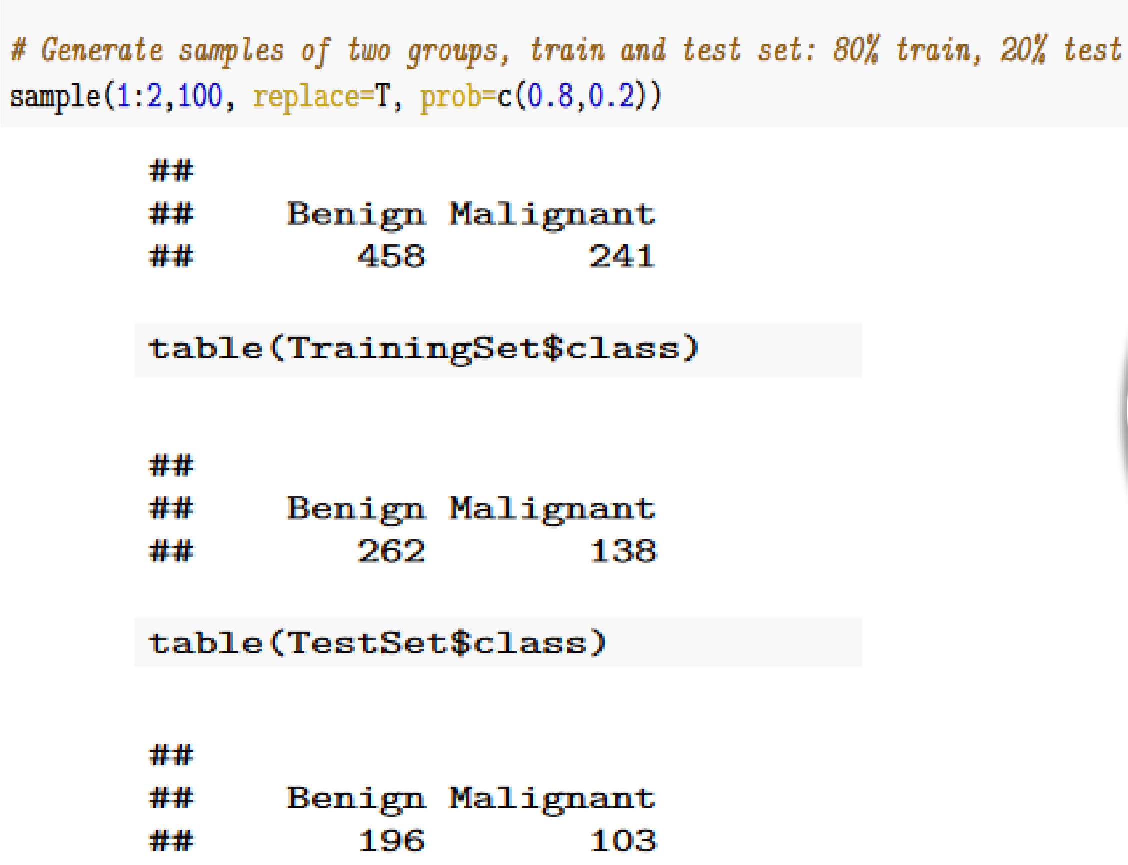
Step (1) Concept of the Regression

- Measures the relationship between the dependent variable and the one or more independent variables.

Step (2) Regression Function

- $f(x) = E(Y|X = x)$
- $class \approx f(all\ other\ features)$
- class = a response, target or outcome
    - Benign or Malignant
    - family=binomial
- thickness or bare_nuclei = a predictor or input

Step (3) Generates Train & Test Sets



```
# Generate samples of two groups, train and test set: 80% train, 20% test
sample(1:2,100, replace=T, prob=c(0.8,0.2))
```

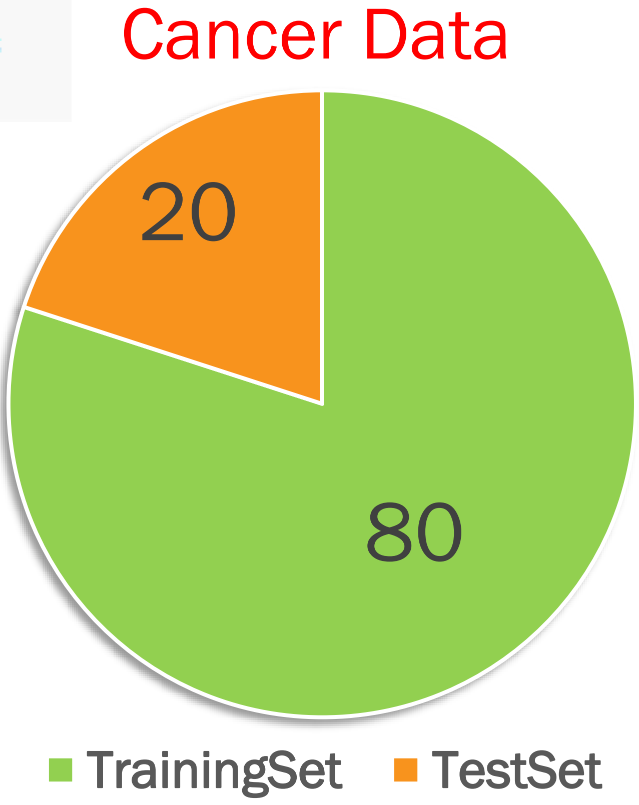
```
##
##     Benign Malignant
##        458       241
```

```
table(TrainingSet$class)
```

```
##
##     Benign Malignant
##        262       138
```

```
table(TestSet$class)
```

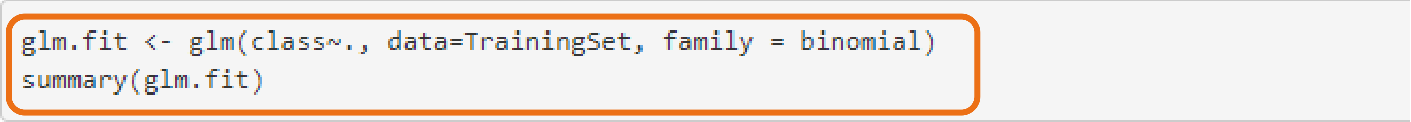
```
##
##     Benign Malignant
##        196       103
```



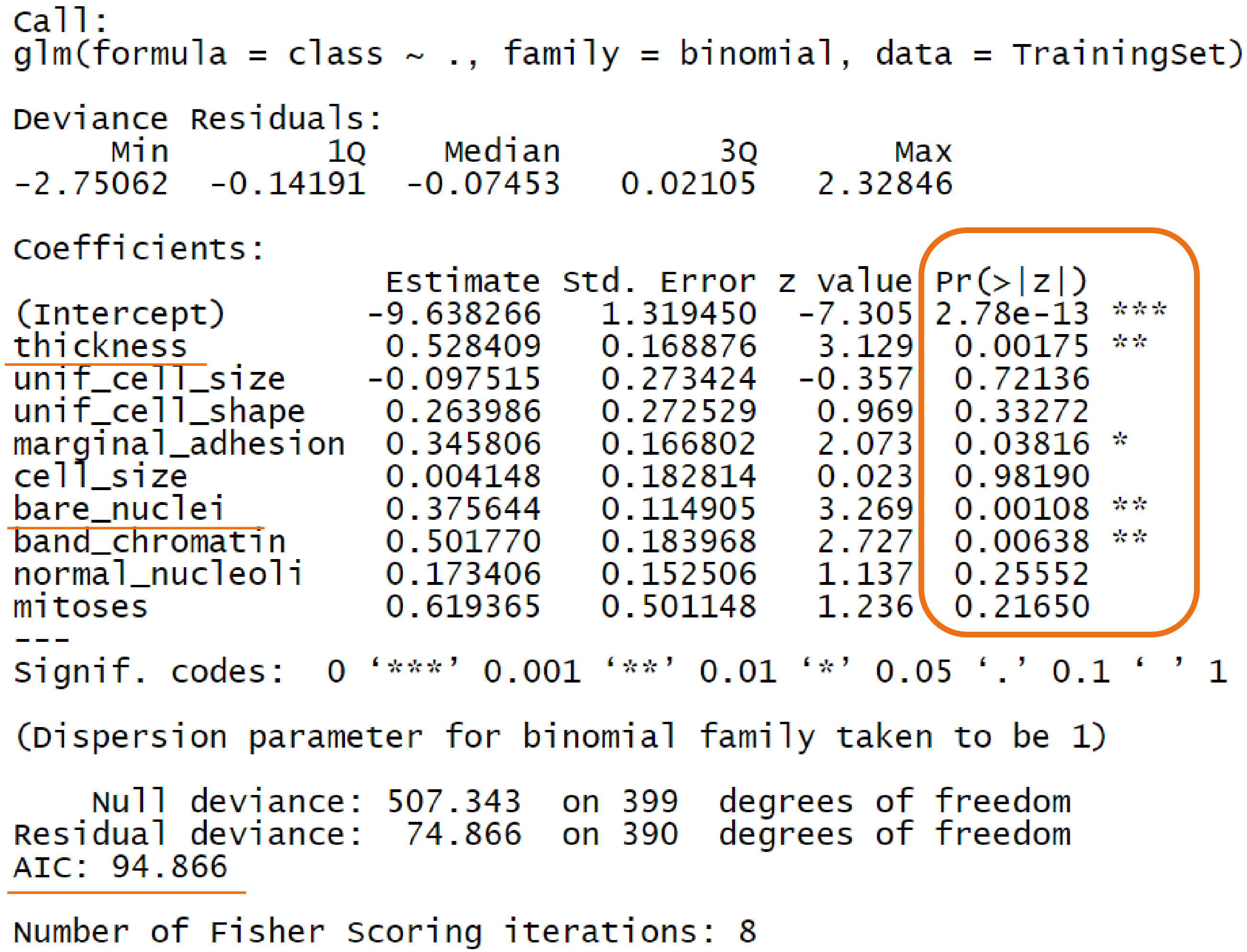
Step (3) Train & Test Sets (Cont.)

	Original	TrainingSet	TestSet
Benign	66 %	67 %	64 %
Malignant	34 %	33 %	36 %

Step (4) Linear Regression



```
glm.fit <- glm(class~., data=TrainingSet, family = binomial)
summary(glm.fit)
```



```
Call:
glm(formula = class ~ ., family = binomial, data = TrainingSet)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.75062  -0.14191  -0.07453   0.02105   2.32846  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -9.638266   1.319450  -7.305 2.78e-13 ***
thickness          0.528409   0.168876   3.129  0.00175 ** 
unif_cell_size    -0.097515   0.273424  -0.357  0.72136    
unif_cell_shape    0.263986   0.272529   0.969  0.33272    
marginal_adhesion  0.345806   0.166802   2.073  0.03816 *  
cell_size          0.004148   0.182814   0.023  0.98190    
bare_nuclei        0.375644   0.114905   3.269  0.00108 ** 
band_chromatin     0.501770   0.183968   2.727  0.00638 ** 
normal_nucleoli    0.173406   0.152506   1.137  0.25552    
mitoses            0.619365   0.501148   1.236  0.21650    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 507.343  on 399  degrees of freedom
Residual deviance:  74.866  on 390  degrees of freedom
AIC: 94.866

Number of Fisher Scoring iterations: 8
```

P-Value is close to 0

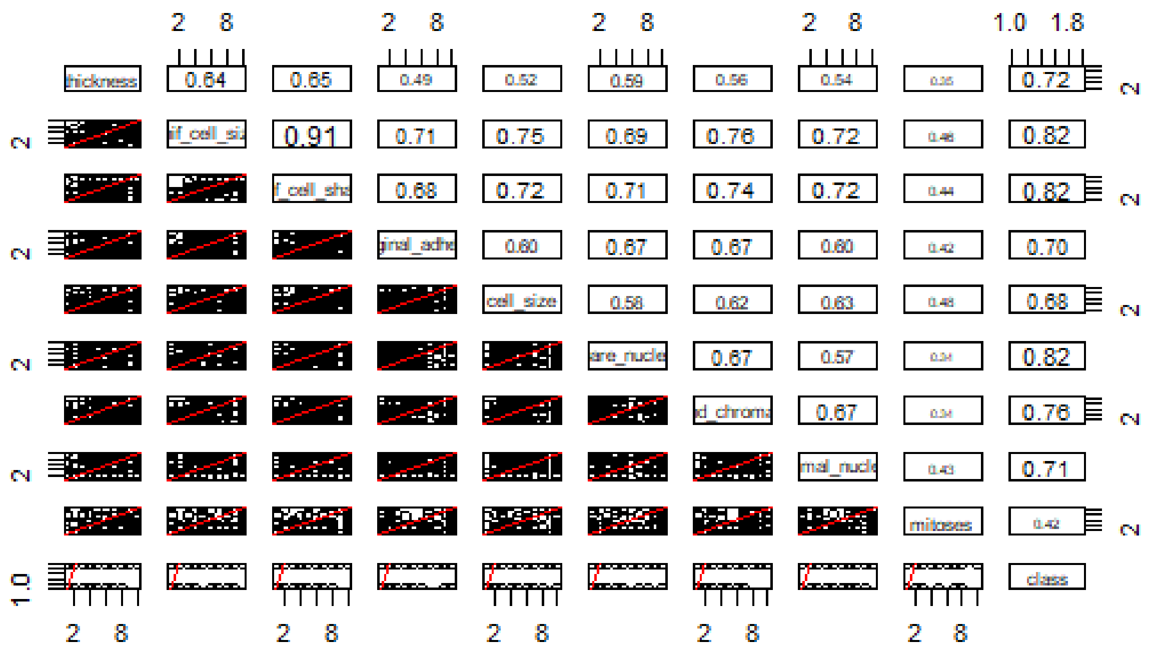
chromatin

Step (4) Linear Regression (Cont.)

- Is there a relationship between predictors and response?

Yes, there is a significant relationship between predictors and response.

- The scatter plots matrix about the relationship between predictors and response.



chromatin

Step (4) Linear Regression (Cont.)

- The statistically significant variables

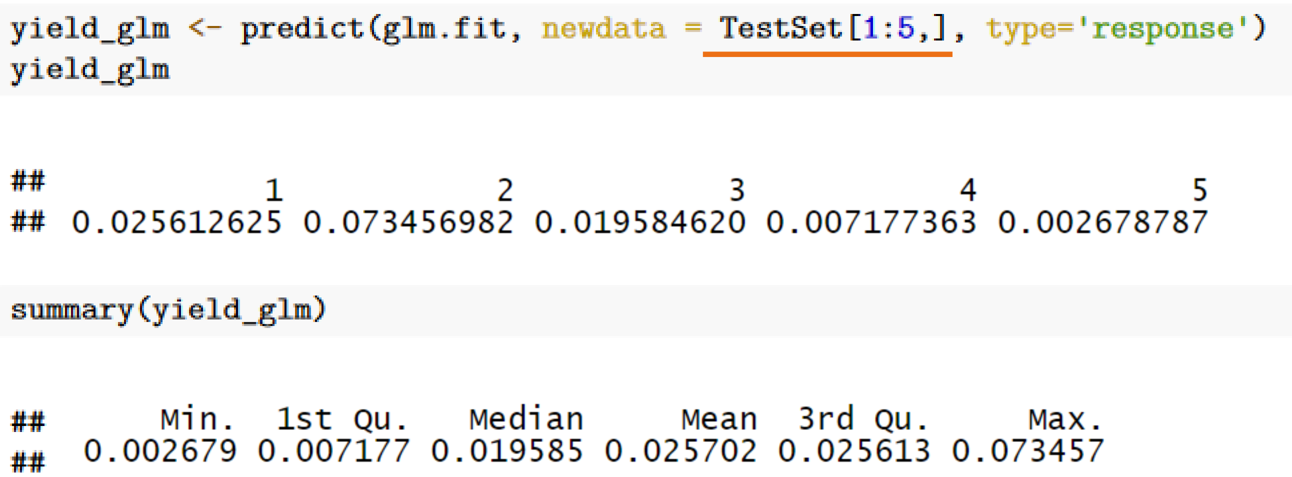
thickness

Bare_nuclei

- The coefficient for thickness

0.52. Hence, it is a positive relationship between class and thickness.

Step (5) Prediction to the Response



```
yield_glm <- predict(glm.fit, newdata = TestSet[1:5,], type='response')
yield_glm
```

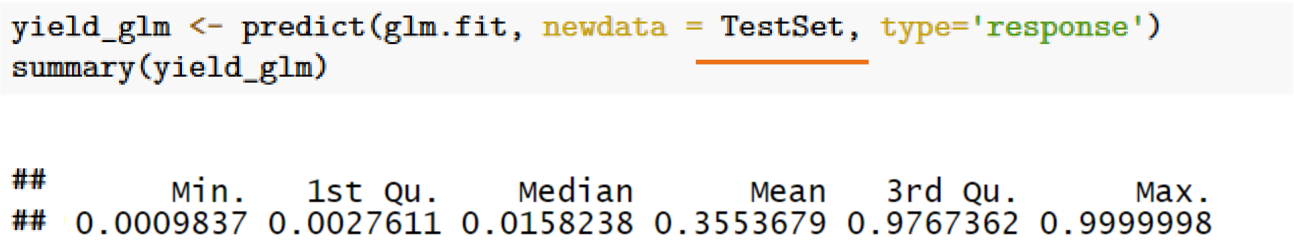
```
##           1           2           3           4           5 
## 0.025612625 0.073456982 0.019584620 0.007177363 0.002678787
```

```
summary(yield_glm)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## 0.002679 0.007177 0.019585 0.025702 0.025613 0.073457
##
```

Each value of yield_glm has the probability placing between 0 and 1
which means how much close to Malignant (1)

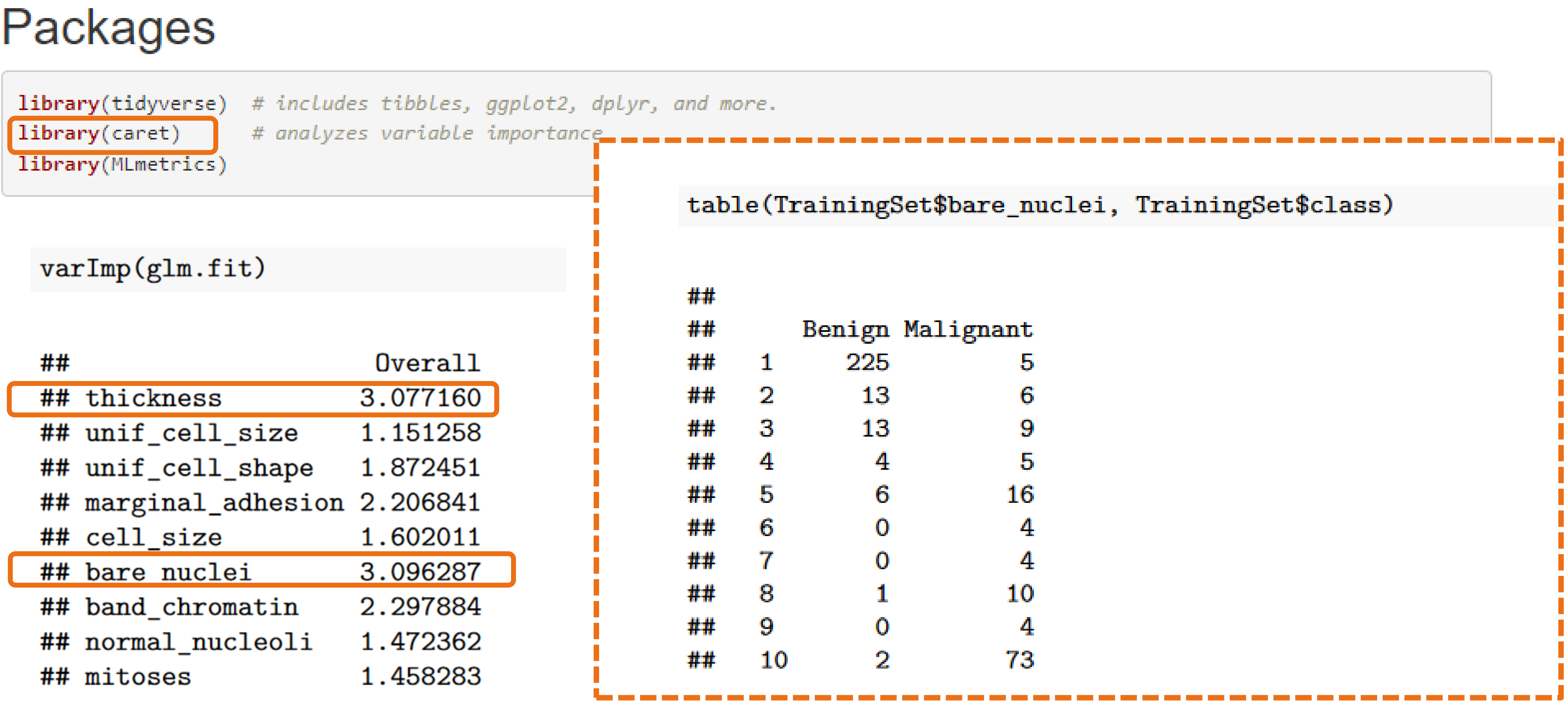
Let's apply the model to a New Data, TestSet



```
yield_glm <- predict(glm.fit, newdata = TestSet, type='response')
summary(yield_glm)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 0.0009837 0.0027611 0.0158238 0.3553679 0.9767362 0.9999998
```

Step (6) The Importance of Variables



Packages

```
library(tidyverse)  # includes tibbles, ggplot2, dplyr, and more.
library(caret)      # analyzes variable importance
library(MLmetrics)
```

```
varImp(glm.fit)
```

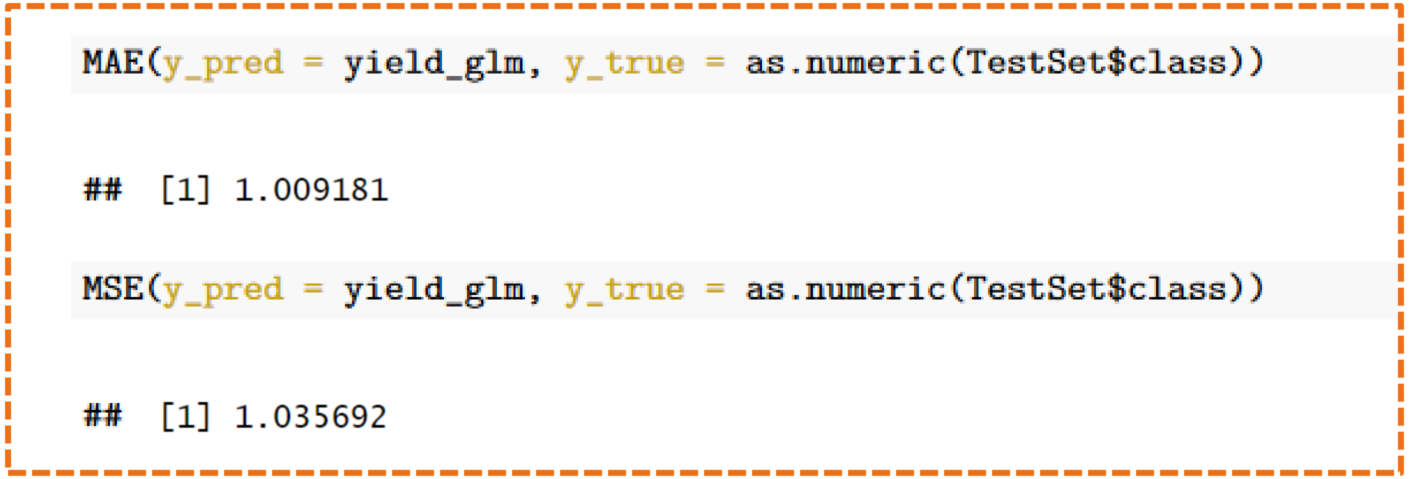
```
##                  Overall
## thickness       3.077160
## unif_cell_size  1.151258
## unif_cell_shape 1.872451
## marginal_adhesion 2.206841
## cell_size       1.602011
## bare_nuclei     3.096287
## band_chromatin  2.297884
## normal_nucleoli 1.472362
## mitoses         1.458283
```

```
table(TrainingSet$bare_nuclei, TrainingSet$class)
```

```
##
##      Benign Malignant
##   1     225         5
##   2      13         6
##   3      13         9
##   4       4         5
##   5       6        16
##   6       0         4
##   7       0         4
##   8       1        10
##   9       0         4
##   10      2        73
```

The Model Evaluation

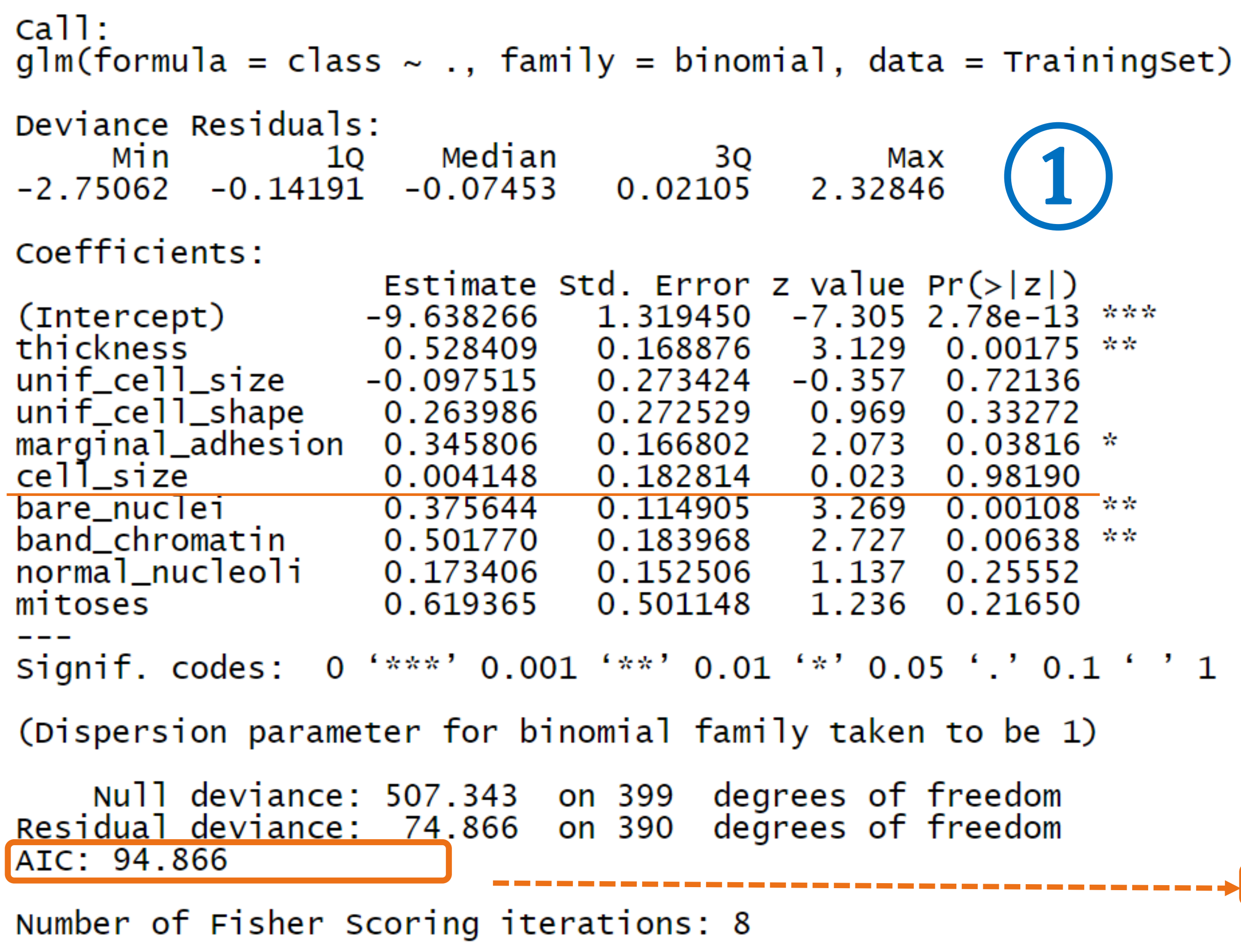
Step (1) MAE & MSE



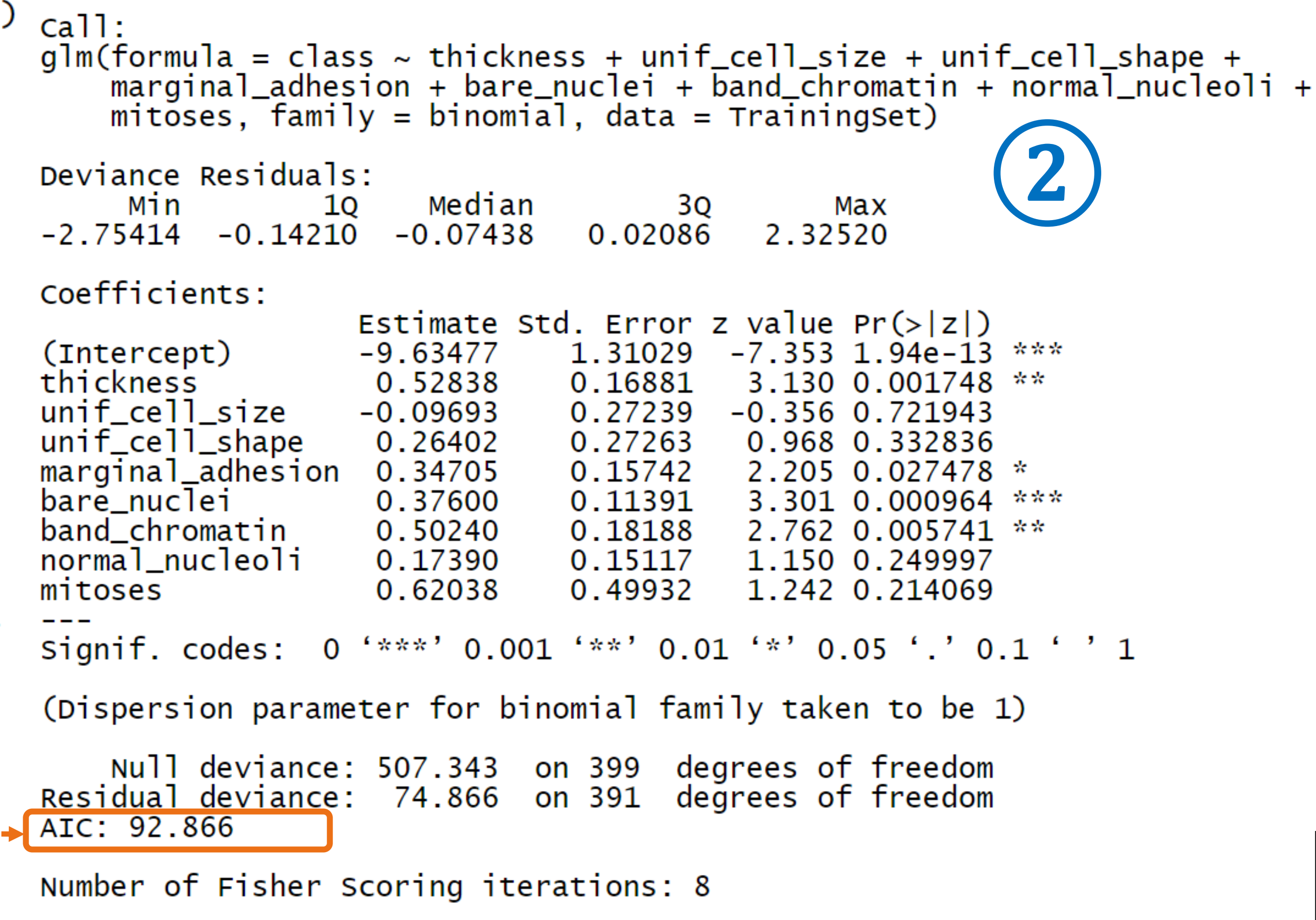
```
MAE(y_pred = yield_glm, y_true = as.numeric(TestSet$class))

## [1] 1.009181

MSE(y_pred = yield_glm, y_true = as.numeric(TestSet$class))

## [1] 1.035692
```

- The Mean Absolute Error (MAE)
- The Mean Squared Error (MSE)
- MAE and MSE are good to close to 0

Step (2) AIC Evaluation
: Stepwise Regression - Backward

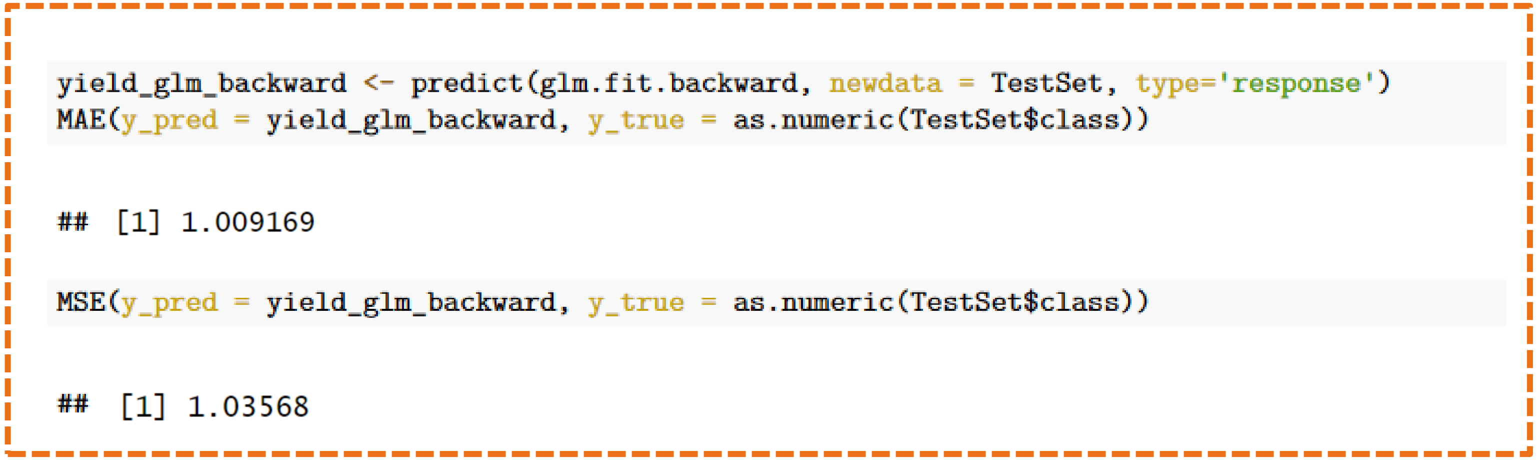


```
Call:
glm(formula = class ~ ., family = binomial, data = TrainingSet)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.75062  -0.14191  -0.07453   0.02105   2.32846  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -9.638266   1.319450  -7.305 2.78e-13 ***
thickness          0.528409   0.168876   3.129  0.00175 ** 
unif_cell_size    -0.097515   0.273424  -0.357  0.72136    
unif_cell_shape    0.263986   0.272529   0.969  0.33272    
marginal_adhesion  0.345806   0.166802   2.073  0.03816 *  
cell_size          0.004148   0.182814   0.023  0.98190    
bare_nuclei        0.375644   0.114905   3.269  0.00108 ** 
band_chromatin     0.501770   0.183968   2.727  0.00638 ** 
normal_nucleoli    0.173406   0.152506   1.137  0.25552    
mitoses            0.619365   0.501148   1.236  0.21650    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 507.343  on 399  degrees of freedom
Residual deviance:  74.866  on 390  degrees of freedom
AIC: 94.866

Number of Fisher Scoring iterations: 8
```



```
Call:
glm(formula = class ~ thickness + unif_cell_size + unif_cell_shape + 
    marginal_adhesion + bare_nuclei + band_chromatin + normal_nucleoli + 
    mitoses, family = binomial, data = TrainingSet)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.75414  -0.14210  -0.07438   0.02086   2.32520  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -9.63477    1.31029  -7.353 1.94e-13 ***
thickness          0.52838    0.16881   3.130 0.001748 ** 
unif_cell_size    -0.09693    0.27239  -0.356 0.721943    
unif_cell_shape    0.26402    0.27263   0.968 0.332836    
marginal_adhesion  0.34705    0.15742   2.205 0.027478 *  
bare_nuclei        0.37600    0.11391   3.301 0.000964 ***
band_chromatin     0.50240    0.18188   2.762 0.005741 ** 
normal_nucleoli    0.17390    0.15117   1.150 0.249997    
mitoses            0.62038    0.49932   1.242 0.214069    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 507.343  on 399  degrees of freedom
Residual deviance:  74.866  on 391  degrees of freedom
AIC: 92.866

Number of Fisher Scoring iterations: 8
```

Step (3) MAE & MSE Again



```
yield_glm_backward <- predict(glm.fit.backward, newdata = TestSet, type='response')
MAE(y_pred = yield_glm_backward, y_true = as.numeric(TestSet$class))
```

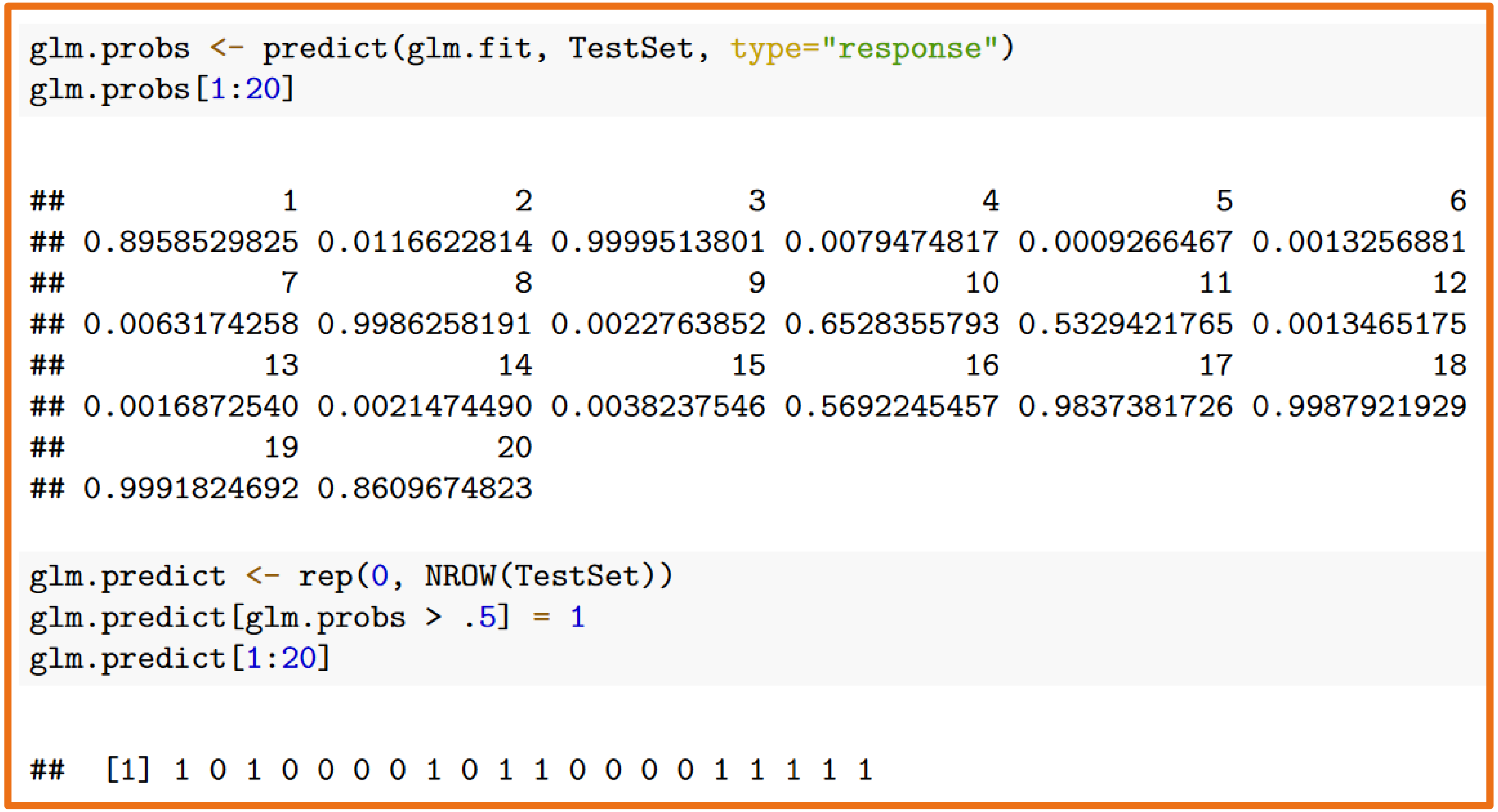
```
## [1] 1.009169
```

```
MSE(y_pred = yield_glm_backward, y_true = as.numeric(TestSet$class))
```

```
## [1] 1.03568
```

- MAE and MSE have the lower values.

Step (4) Accuracy



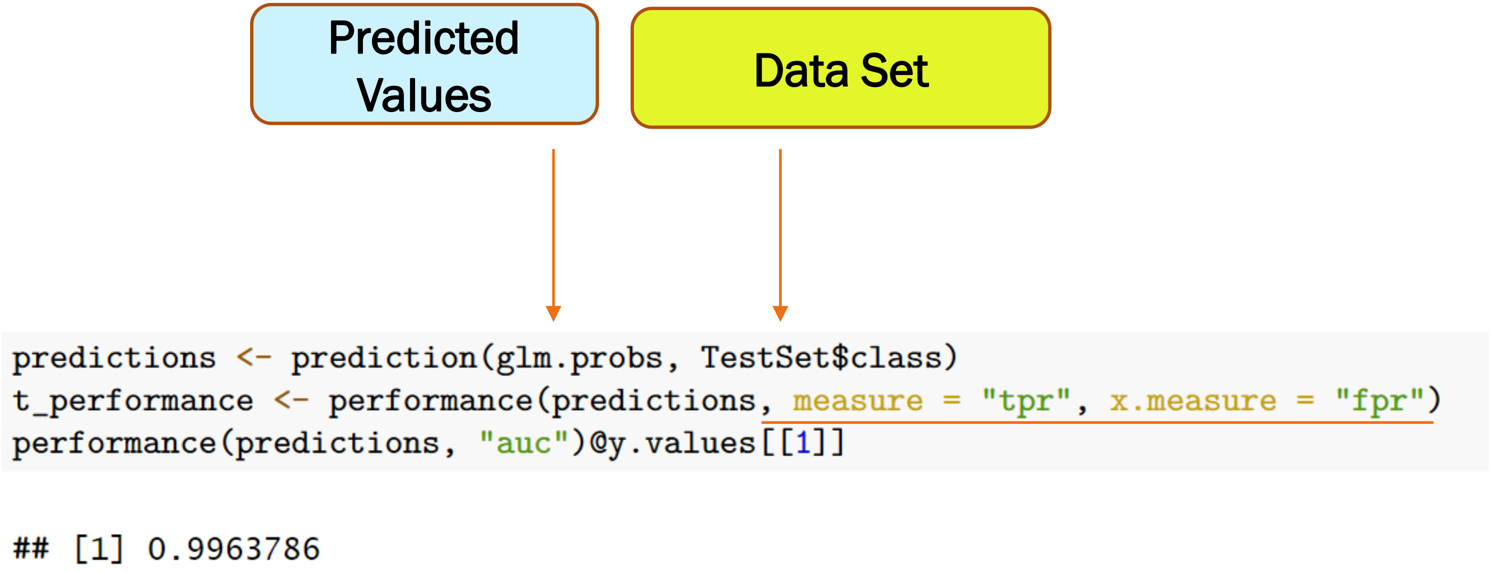
```r
glm.probs <- predict(glm.fit, TestSet, type="response")
glm.probs[1:20]
```

```
##            1            2            3            4            5            6
## 0.8958529825 0.0116622814 0.9999513801 0.0079474817 0.0009266467 0.0013256881
##            7            8            9           10           11           12
## 0.0063174258 0.9986258191 0.0022763852 0.6528355793 0.5329421765 0.0013465175
##           13           14           15           16           17           18
## 0.0016872540 0.0021474490 0.0038237546 0.5692245457 0.9837381726 0.9987921929
##           19           20
## 0.9991824692 0.8609674823
```

```r
glm.predict <- rep(0, NROW(TestSet))
glm.predict[glm.probs > .5] = 1
glm.predict[1:20]
```

```
##  [1] 1 0 1 0 0 0 0 1 0 1 1 0 0 0 0 1 1 1 1 1
```

Step (4) Accuracy



Predicted Values

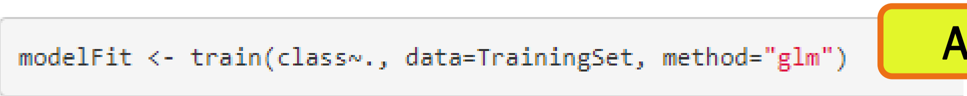
Data Set

```
predictions <- prediction(glm.probs, TestSet$class)
t_performance <- performance(predictions, measure = "tpr", x.measure = "fpr")
performance(predictions, "auc")@y.values[[1]]
```

```
## [1] 0.9963786
```

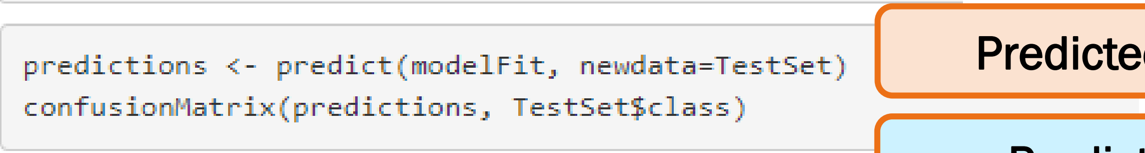
Accuracy 99.63%

(Bonus) Confusion Matrix



```
modelFit <- train(class~., data=TrainingSet, method="glm")
```

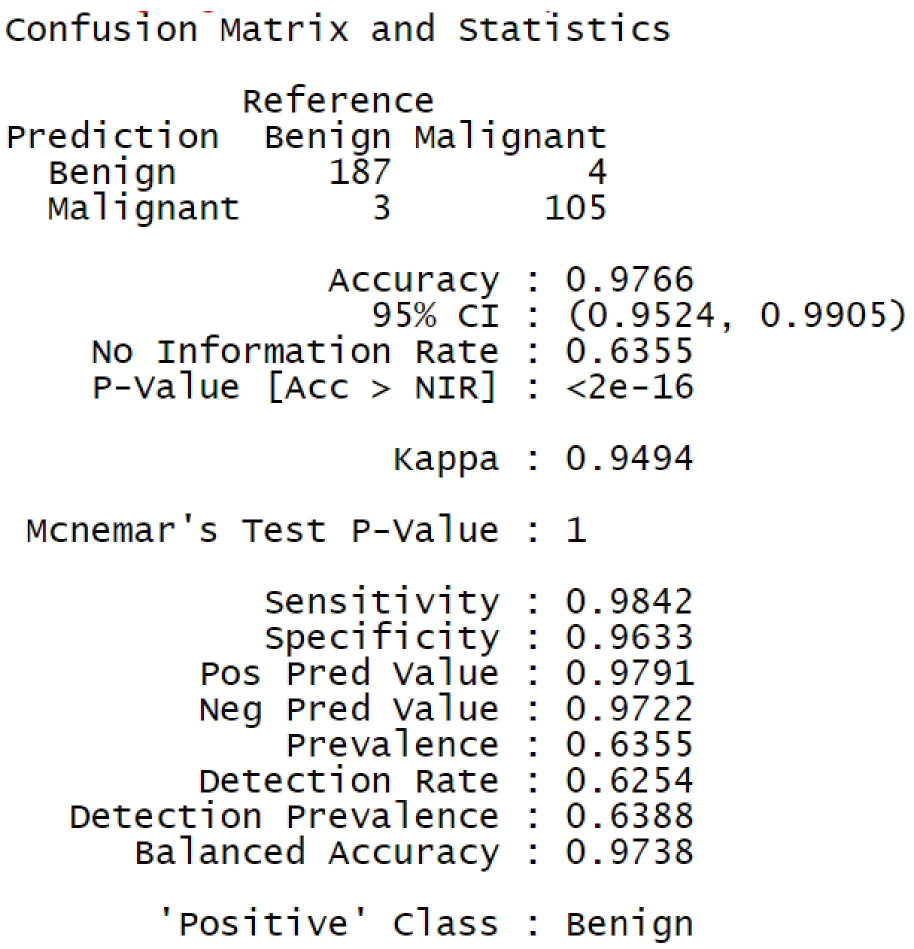
Applying glm to TrainingSet to get model as an input



```
predictions <- predict(modelFit, newdata=TestSet)
confusionMatrix(predictions, TestSet$class)
```

Predicted values in the frame new data, TestSet.

Predictions is categorized by confusionMatrix



```
Confusion Matrix and Statistics

           Reference
Prediction  Benign Malignant
  Benign       187         4
  Malignant      3       105

               Accuracy : 0.9766
                 95% CI : (0.9524, 0.9905)
    No Information Rate : 0.6355
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9494

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9842
            Specificity : 0.9633
         Pos Pred Value : 0.9791
         Neg Pred Value : 0.9722
             Prevalence : 0.6355
         Detection Rate : 0.6254
   Detection Prevalence : 0.6388
      Balanced Accuracy : 0.9738

       'Positive' Class : Benign
```

Step (3) Conclusion

THE MULTIPLE LINEAR REGRESSION MODEL WORKS WELL TO PREDICT THE RESPONSE.

Q & A

Citations

- https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
- https://visualsonline.cancer.gov/details.cfm?imageid=2512

Thank You!