MSDS 6372 Project 2

Contents

Data Set 1: Osteoporosis in Women

From Hosmer, Lemeshow, and Sturdivant (2013), Applied Logistic Regression, 3rd Edition. The Global Longitudinal Study of Osteoporosis in Women (GLOW) is an international study of osteoporosis in women aged 55 years and over. The major goals of the study are to examine prevention and treatment of fractures and distribution of risk factors among older women. Complete details on the study as well as a list of GLOW publications may be found at the Center for Outcomes Research web site, http://www.outcomes-umassmed.org/glow. There are over 60K observations in the original data set. This data set contains a sample of 500 of them. The link below is to a website with the data set and description of the variables. The data set in question is called "glow500".

https://www.umass.edu/statdata/statdata/data/glow/index.html Note: If you choose this data set, you MAY NOT use the Hosmer, Lemeshow, and Sturdivant text to help you in your analysis. You may only use Chapter 1 in order to obtain a description of the data.

Of course if you dont have the book

https://www.umass.edu/statdata/statdata/data/glow/glow.pdf provides definitions to the variables.

The Global Longitudinal Study of Osteoporosis in Women (GLOW) (2005-2014) was a prospective cohort study of physician practices in the provision of prophylaxis and treatment against osteoporotic fractures. The goal of this research was to improve understanding of the risk and prevention of osteoporosis-related fractures among female residents of 10 countries who were 55 years of age and older. GLOW enrolled over 60,000 women through over 700 physicians in 10 countries, and conducted annual follow-up for up to 5 years through annual patient questionnaires.

Setup:

Data Import and Cleaning

Missing values were not detected in dataset. Special characters were removed from column headings. What we know/don't know about the sample (500): 1. We do not know if the subjects are distributed equally around the world. We will assume that the same percentage from each region was selected for the sample in this dataset. 2. Based on the Sub_ID(Subject ID), we can assume that the datat is independent sample of participants. 3.

```
library(here)
```

```
## here() starts at D:/2018-stats2-project/stats2proj2
```

```
##
## Attaching package: 'here'
```

```
## The following object is masked from 'package:plyr':
##
##     here
```

```
glow_data_file <- here("data", "glow500.csv")
dataset <- read.csv(glow_data_file, sep=",", stringsAsFactors = TRUE, header=TRUE,na.strings=c(""))
#dataset <- read.csv("C:/Users/carol/OneDrive/Documents/MSDS6372/Proj2/glow500.csv", sep=",", stringsAs

# List rows of data that have missing values
Missing_values <- dataset[!complete.cases(dataset),]

# Create new dataset without missing data
dataset <- na.omit(dataset)

#remove FRACSCORE feature per professor Turner
drops <- c("FRACSCORE")
dataset <- dataset[ , !(names(dataset) %in% drops)]

#Cleanup column names
colnames(dataset)[colnames(dataset)=="ï..SUB_ID"] <- "SUB_ID"
```

Grouping Variables as Continuous, Categorical, and ID

```
numericVar <- dataset[,5:8]
ID_var <- dataset[,c(1:3)]
set_noID <- dataset[4:14]
categoricalVar <- set_noID[,-c(2:5)]
```

Create a vector of all categorical variables and run frequency 2X2s with Mosaic plots.

Chi-Square Test For the 2-way tables the chisq test independence will show if 2 categorical variables are related in some population. Null Hypothesis: The two categorical variables are independent. Alternative Hypothesis: The two categorical variables are dependent

Variable: PRIORFRAC 41% of subjects with Prior Franctures also had current Fractures but only make up 25% of the overall subjects in the sample that had prior fractures. The Chi-squared p-value favors overwhemingly the alternative hypothesis that the PRIORFRAC variable is dependent on Fracture variable.

Variable: PREMENO 80% of the sample subjects are not in Pre-Menopausehad of which 24% had fractures. The same frequency of 25% Premenopausal women had fractures. The Chi-squared p-value favors the null hypothesis that the PREMENO variable is independent on Fracture variable.

Variable: MOMFRAC 13% of subjects have Mothers with a history of fractures. Out of those 13%, 36% of subjects also had fractures. The Chi-squared p-value favors the alternative hypothesis that the MOMFRAC variable is probably dependent on Fracture variable.

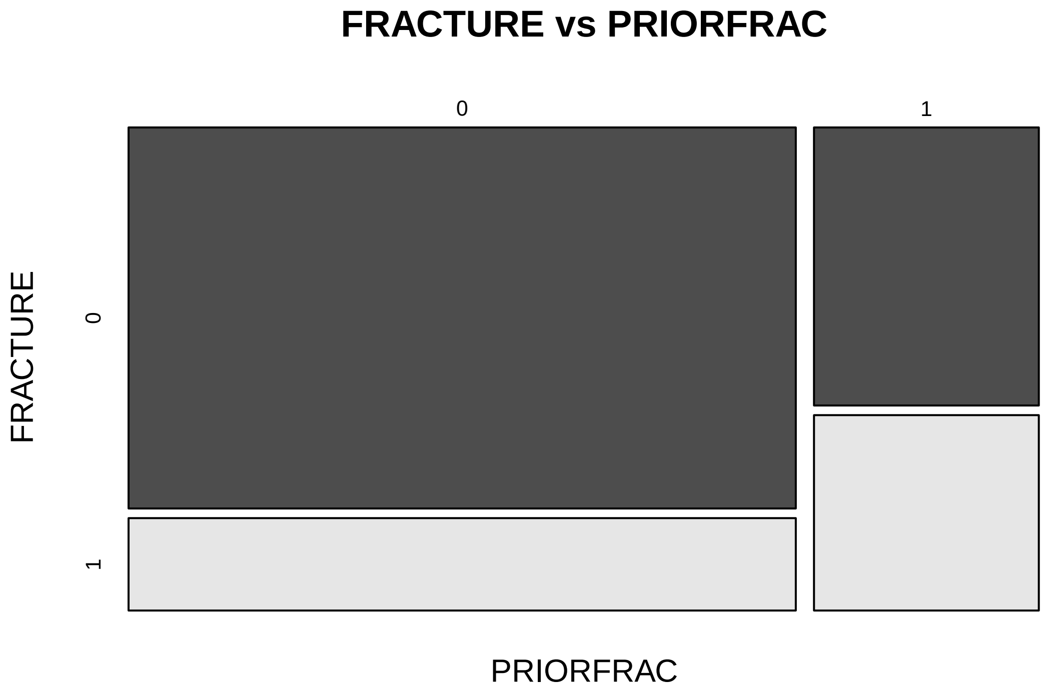
Variable: ARMASSIST 62% (312/500) subjects do not have Armassist of which 20% had fractures. Of those with Armassist, 33% had fractures. The Chi-squared p-value favors the alternative hypothesis that the ARMASSIST variable is most likely dependent on Fracture variable.

Variable: SMOKE In the dataset, 93% of subjects are non-smokers of which 26% had fractures. 7% of the subjects who were smokers of which 26% had no fractures. Although the subjects are not balance in smoker vs non-smoker category, the p-value for Chi-squared test shows .47 we favor the alternative hypothesis that the Smoke variable is dependent on the Fracture.

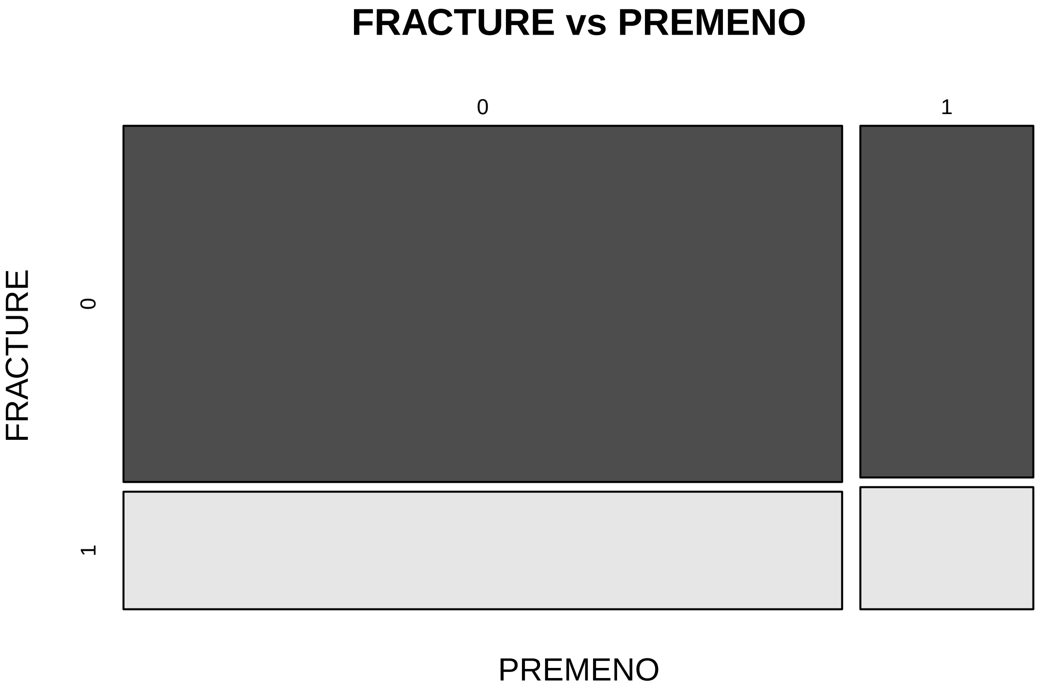
Variable: RATERISK Raterisk shows the frequency of subjects in each Raterisk level is between 29%-33%. This is pretty even in terms of how many subjects are within each Raterisk. For those that did have Fractures, their probability of a fracture increased with the level of Raterisk. This makes sense.

```
categoricalVarVec  <- c("PRIORFRAC","PREMENO","MOMFRAC","ARMASSIST","SMOKE","RATERISK")
for(categoricalVar in categoricalVarVec){
  CrossTable(dataset[,categoricalVar], dataset$FRACTURE, chisq = TRUE , expected = TRUE, dnn=c(categori
  mosaicplot(CrossTable(dataset[ ,categoricalVar], dataset$FRACTURE)$t, main=paste("FRACTURE vs",catego
}
```

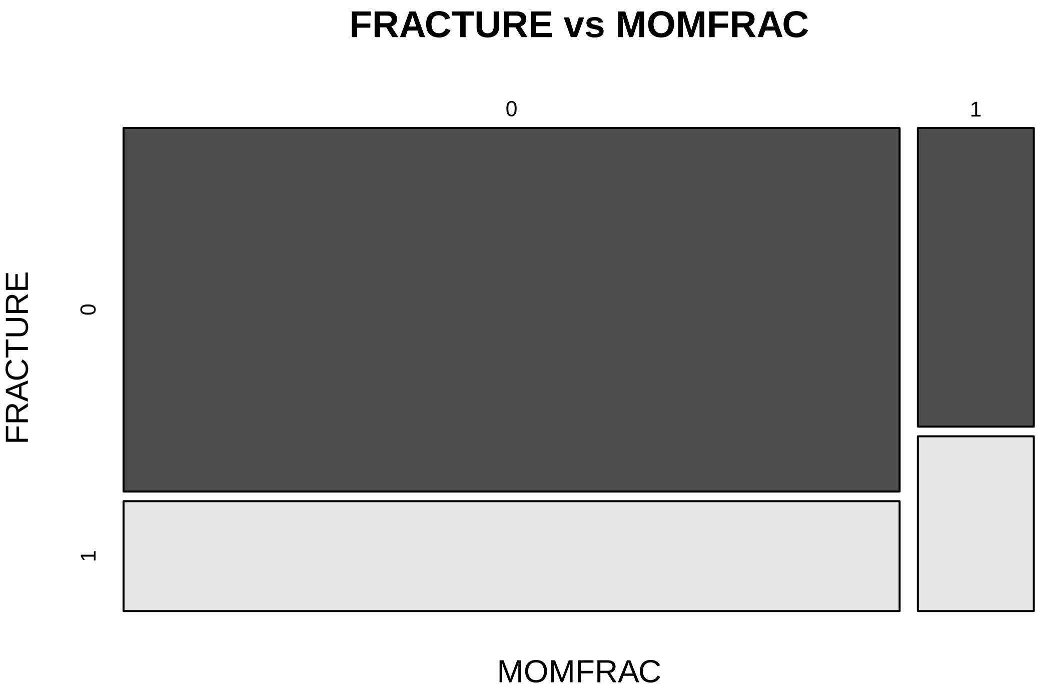
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##              | FRACTURE
##    PRIORFRAC |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       301 |        73 |       374 |
##              |   280.500 |    93.500 |           |
##              |     1.498 |     4.495 |           |
##              |     0.805 |     0.195 |     0.748 |
##              |     0.803 |     0.584 |           |
##              |     0.602 |     0.146 |           |
## -------------|-----------|-----------|-----------|
##            1 |        74 |        52 |       126 |
##              |    94.500 |    31.500 |           |
##              |     4.447 |    13.341 |           |
##              |     0.587 |     0.413 |     0.252 |
##              |     0.197 |     0.416 |           |
##              |     0.148 |     0.104 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       375 |       125 |       500 |
##              |     0.750 |     0.250 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  23.78123     d.f. =  1     p =  1.079299e-06
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  22.63532     d.f. =  1     p =  1.958512e-06
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##                           | dataset$FRACTURE
## dataset[, categoricalVar] |         0 |         1 | Row Total |
## --------------------------|-----------|-----------|-----------|
##                         0 |       301 |        73 |       374 |
##                           |     1.498 |     4.495 |           |
##                           |     0.805 |     0.195 |     0.748 |
##                           |     0.803 |     0.584 |           |
##                           |     0.602 |     0.146 |           |
## --------------------------|-----------|-----------|-----------|
##                         1 |        74 |        52 |       126 |
##                           |     4.447 |    13.341 |           |
##                           |     0.587 |     0.413 |     0.252 |
##                           |     0.197 |     0.416 |           |
##                           |     0.148 |     0.104 |           |
## --------------------------|-----------|-----------|-----------|
##              Column Total |       375 |       125 |       500 |
##                           |     0.750 |     0.250 |           |
## --------------------------|-----------|-----------|-----------|
##
##
```



```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##              | FRACTURE
##      PREMENO |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       303 |       100 |       403 |
##              |   302.250 |   100.750 |           |
##              |     0.002 |     0.006 |           |
##              |     0.752 |     0.248 |     0.806 |
##              |     0.808 |     0.800 |           |
##              |     0.606 |     0.200 |           |
## -------------|-----------|-----------|-----------|
##            1 |        72 |        25 |        97 |
##              |    72.750 |    24.250 |           |
##              |     0.008 |     0.023 |           |
##              |     0.742 |     0.258 |     0.194 |
##              |     0.192 |     0.200 |           |
##              |     0.144 |     0.050 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       375 |       125 |       500 |
##              |     0.750 |     0.250 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  0.038372     d.f. =  1     p =  0.844698
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  0.004263556     d.f. =  1     p =  0.9479384
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##                           | dataset$FRACTURE
## dataset[, categoricalVar] |         0 |         1 | Row Total |
## --------------------------|-----------|-----------|-----------|
##                         0 |       303 |       100 |       403 |
##                           |     0.002 |     0.006 |           |
##                           |     0.752 |     0.248 |     0.806 |
##                           |     0.808 |     0.800 |           |
##                           |     0.606 |     0.200 |           |
## --------------------------|-----------|-----------|-----------|
##                         1 |        72 |        25 |        97 |
##                           |     0.008 |     0.023 |           |
##                           |     0.742 |     0.258 |     0.194 |
##                           |     0.192 |     0.200 |           |
##                           |     0.144 |     0.050 |           |
## --------------------------|-----------|-----------|-----------|
##              Column Total |       375 |       125 |       500 |
##                           |     0.750 |     0.250 |           |
## --------------------------|-----------|-----------|-----------|
##
##
```

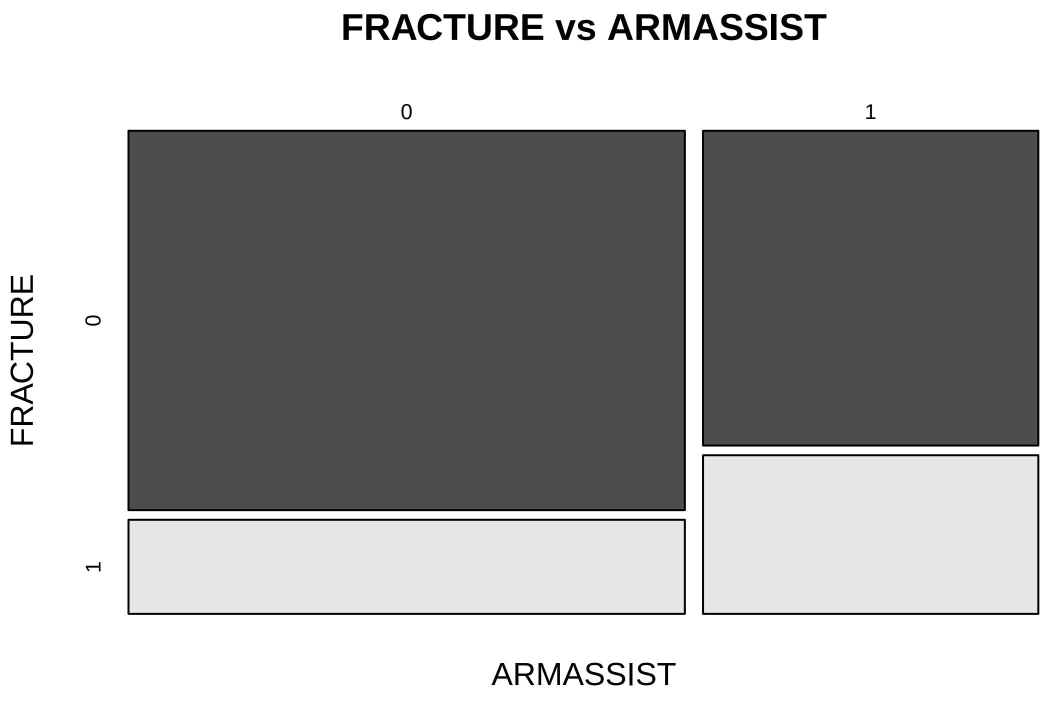


```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##              | FRACTURE
##      MOMFRAC |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       334 |       101 |       435 |
##              |   326.250 |   108.750 |           |
##              |     0.184 |     0.552 |           |
##              |     0.768 |     0.232 |     0.870 |
##              |     0.891 |     0.808 |           |
##              |     0.668 |     0.202 |           |
## -------------|-----------|-----------|-----------|
##            1 |        41 |        24 |        65 |
##              |    48.750 |    16.250 |           |
##              |     1.232 |     3.696 |           |
##              |     0.631 |     0.369 |     0.130 |
##              |     0.109 |     0.192 |           |
##              |     0.082 |     0.048 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       375 |       125 |       500 |
##              |     0.750 |     0.250 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  5.664604     d.f. =  1     p =  0.01731063
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  4.957265     d.f. =  1     p =  0.02598127
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##                           | dataset$FRACTURE
## dataset[, categoricalVar] |         0 |         1 | Row Total |
## --------------------------|-----------|-----------|-----------|
##                         0 |       334 |       101 |       435 |
##                           |     0.184 |     0.552 |           |
##                           |     0.768 |     0.232 |     0.870 |
##                           |     0.891 |     0.808 |           |
##                           |     0.668 |     0.202 |           |
## --------------------------|-----------|-----------|-----------|
##                         1 |        41 |        24 |        65 |
##                           |     1.232 |     3.696 |           |
##                           |     0.631 |     0.369 |     0.130 |
##                           |     0.109 |     0.192 |           |
##                           |     0.082 |     0.048 |           |
## --------------------------|-----------|-----------|-----------|
##              Column Total |       375 |       125 |       500 |
##                           |     0.750 |     0.250 |           |
## --------------------------|-----------|-----------|-----------|
##
##
```

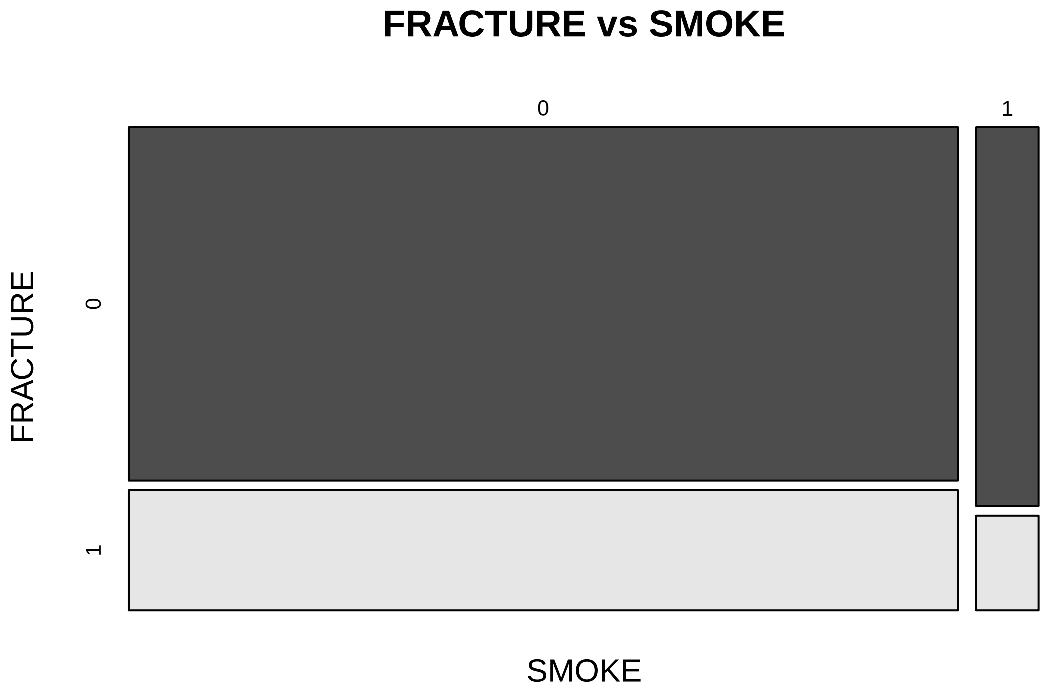


```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##              | FRACTURE
##    ARMASSIST |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       250 |        62 |       312 |
##              |   234.000 |    78.000 |           |
##              |     1.094 |     3.282 |           |
##              |     0.801 |     0.199 |     0.624 |
##              |     0.667 |     0.496 |           |
##              |     0.500 |     0.124 |           |
## -------------|-----------|-----------|-----------|
##            1 |       125 |        63 |       188 |
##              |   141.000 |    47.000 |           |
##              |     1.816 |     5.447 |           |
##              |     0.665 |     0.335 |     0.376 |
##              |     0.333 |     0.504 |           |
##              |     0.250 |     0.126 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       375 |       125 |       500 |
##              |     0.750 |     0.250 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  11.63848     d.f. =  1     p =  0.0006460138
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  10.92244     d.f. =  1     p =  0.0009500637
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
```

```
##                          | dataset$FRACTURE 
## dataset[, categoricalVar] |         0 |         1 | Row Total | 
## --------------------------|-----------|-----------|-----------|
##                         0 |       250 |        62 |       312 | 
##                           |     1.094 |     3.282 |           | 
##                           |     0.801 |     0.199 |     0.624 | 
##                           |     0.667 |     0.496 |           | 
##                           |     0.500 |     0.124 |           | 
## --------------------------|-----------|-----------|-----------|
##                         1 |       125 |        63 |       188 | 
##                           |     1.816 |     5.447 |           | 
##                           |     0.665 |     0.335 |     0.376 | 
##                           |     0.333 |     0.504 |           | 
##                           |     0.250 |     0.126 |           | 
## --------------------------|-----------|-----------|-----------|
##              Column Total |       375 |       125 |       500 | 
##                           |     0.750 |     0.250 |           | 
## --------------------------|-----------|-----------|-----------|
## 
## 
```

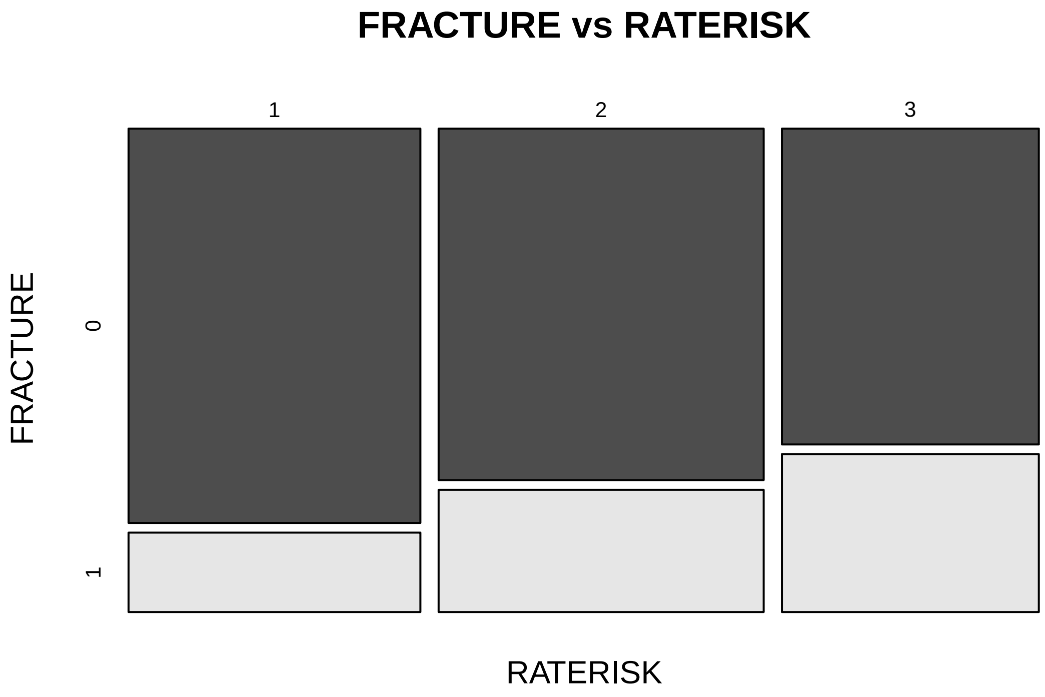


```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##              | FRACTURE
##        SMOKE |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       347 |       118 |       465 |
##              |   348.750 |   116.250 |           |
##              |     0.009 |     0.026 |           |
##              |     0.746 |     0.254 |     0.930 |
##              |     0.925 |     0.944 |           |
##              |     0.694 |     0.236 |           |
## -------------|-----------|-----------|-----------|
##            1 |        28 |         7 |        35 |
##              |    26.250 |     8.750 |           |
##              |     0.117 |     0.350 |           |
##              |     0.800 |     0.200 |     0.070 |
##              |     0.075 |     0.056 |           |
##              |     0.056 |     0.014 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       375 |       125 |       500 |
##              |     0.750 |     0.250 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  0.5017921     d.f. =  1     p =  0.4787137
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  0.2560164     d.f. =  1     p =  0.6128703
##
##
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##                          | dataset$FRACTURE
## dataset[, categoricalVar] |         0 |         1 | Row Total |
## -------------------------|-----------|-----------|-----------|
##                        0 |       347 |       118 |       465 |
##                          |     0.009 |     0.026 |           |
##                          |     0.746 |     0.254 |     0.930 |
##                          |     0.925 |     0.944 |           |
##                          |     0.694 |     0.236 |           |
## -------------------------|-----------|-----------|-----------|
##                        1 |        28 |         7 |        35 |
##                          |     0.117 |     0.350 |           |
##                          |     0.800 |     0.200 |     0.070 |
##                          |     0.075 |     0.056 |           |
##                          |     0.056 |     0.014 |           |
## -------------------------|-----------|-----------|-----------|
##             Column Total |       375 |       125 |       500 |
##                          |     0.750 |     0.250 |           |
## -------------------------|-----------|-----------|-----------|
##
##
```



```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##              | FRACTURE
##     RATERISK |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |       139 |        28 |       167 |
##              |   125.250 |    41.750 |           |
##              |     1.509 |     4.528 |           |
##              |     0.832 |     0.168 |     0.334 |
##              |     0.371 |     0.224 |           |
##              |     0.278 |     0.056 |           |
## -------------|-----------|-----------|-----------|
##            2 |       138 |        48 |       186 |
##              |   139.500 |    46.500 |           |
##              |     0.016 |     0.048 |           |
##              |     0.742 |     0.258 |     0.372 |
##              |     0.368 |     0.384 |           |
##              |     0.276 |     0.096 |           |
## -------------|-----------|-----------|-----------|
##            3 |        98 |        49 |       147 |
##              |   110.250 |    36.750 |           |
##              |     1.361 |     4.083 |           |
##              |     0.667 |     0.333 |     0.294 |
##              |     0.261 |     0.392 |           |
##              |     0.196 |     0.098 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       375 |       125 |       500 |
##              |     0.750 |     0.250 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  11.54688     d.f. =  2     p =  0.003109037
##
##
##
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  500
##
##
##                           | dataset$FRACTURE
## dataset[, categoricalVar] |         0 |         1 | Row Total |
## --------------------------|-----------|-----------|-----------|
##                         1 |       139 |        28 |       167 |
##                           |     1.509 |     4.528 |           |
##                           |     0.832 |     0.168 |     0.334 |
##                           |     0.371 |     0.224 |           |
##                           |     0.278 |     0.056 |           |
## --------------------------|-----------|-----------|-----------|
##                         2 |       138 |        48 |       186 |
##                           |     0.016 |     0.048 |           |
##                           |     0.742 |     0.258 |     0.372 |
##                           |     0.368 |     0.384 |           |
##                           |     0.276 |     0.096 |           |
## --------------------------|-----------|-----------|-----------|
##                         3 |        98 |        49 |       147 |
##                           |     1.361 |     4.083 |           |
##                           |     0.667 |     0.333 |     0.294 |
##                           |     0.261 |     0.392 |           |
##                           |     0.196 |     0.098 |           |
## --------------------------|-----------|-----------|-----------|
##              Column Total |       375 |       125 |       500 |
##                           |     0.750 |     0.250 |           |
## --------------------------|-----------|-----------|-----------|
##
##
```



Exploratory Data Analysis

Summary Tables

Assumptions This is a prospective study which means its a study over time of a group of similar individuals who differ with respect to certain factors under a study and how these factors affect rates of a certain outcome (Fracture vs No-Fracture) Linearity -

Independence of errors - Based on SUB_ID(Subject ID) we confirm each record is an independent sample. Multicollinearity - Weight and BMI are highly correlated.

```
# display the first 20 rows
print(head(dataset, n=20))
```

```
##    SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT      BMI PREMENO
## 1       1       1     14         0  62   70.3    158 28.16055       0
## 2       2       4    284         0  65   87.1    160 34.02344       0
## 3       3       6    305         1  88   50.8    157 20.60936       0
## 4       4       6    309         0  82   62.1    160 24.25781       0
## 5       5       1     37         0  61   68.0    152 29.43213       0
## 6       6       5    299         1  67   68.0    161 26.23356       0
## 7       7       5    302         0  84   50.8    150 22.57778       0
## 8       8       1     36         1  82   40.8    153 17.42919       0
## 9       9       1      8         1  86   62.6    156 25.72321       0
## 10     10       4    282         0  58   63.5    166 23.04398       0
## 11     11       6    315         0  67   67.6    153 28.87778       0
## 12     12       1     34         0  56  117.9    167 42.27473       0
## 13     13       6    315         0  59   67.1    162 25.56775       0
## 14     14       1     33         0  72   57.6    165 21.15702       0
## 15     15       1     23         0  64   61.2    160 23.90625       1
## 16     16       3    179         0  68   78.0    161 30.09143       0
## 17     17       4    284         0  67  105.7    165 38.82461       0
## 18     18       4    283         0  69   65.8    162 25.07240       0
## 19     19       3    179         1  78   81.6    162 31.09282       0
## 20     20       6    313         0  60   56.7    157 23.00296       0
##    MOMFRAC ARMASSIST SMOKE RATERISK FRACTURE
## 1        0         0     0        2        0
## 2        0         0     0        2        0
## 3        1         1     0        1        0
## 4        0         0     0        1        0
## 5        0         0     0        2        0
## 6        0         0     1        2        0
## 7        0         0     0        1        0
## 8        0         0     0        2        0
## 9        0         0     0        2        0
## 10       0         0     0        1        0
## 11       1         0     1        1        0
## 12       0         1     1        2        0
## 13       0         0     1        1        0
## 14       0         1     0        1        0
## 15       0         0     0        2        0
## 16       0         1     0        1        0
## 17       0         0     0        1        0
## 18       0         0     0        2        0
## 19       0         1     0        3        0
## 20       0         0     0        2        0
```

```
# display the dimensions of the dataset
print(dim(dataset))
```

```
## [1] 500  14
```

```
# list types for each attribute
print(sapply(dataset,class))
```

```
##    SUB_ID   SITE_ID    PHY_ID PRIORFRAC       AGE    WEIGHT    HEIGHT 
## "integer" "integer" "integer" "integer" "integer" "numeric" "integer" 
##       BMI   PREMENO   MOMFRAC ARMASSIST     SMOKE  RATERISK  FRACTURE 
## "numeric" "integer" "integer" "integer" "integer" "integer" "integer"
```

```
# summarize the dataset
print(summary(set_noID))
```

```
##    PRIORFRAC          AGE            WEIGHT           HEIGHT     
##  Min.   :0.000   Min.   :55.00   Min.   : 39.90   Min.   :134.0  
##  1st Qu.:0.000   1st Qu.:61.00   1st Qu.: 59.90   1st Qu.:157.0
##  Median :0.000   Median :67.00   Median : 68.00   Median :161.5
##  Mean   :0.252   Mean   :68.56   Mean   : 71.82   Mean   :161.4
##  3rd Qu.:1.000   3rd Qu.:76.00   3rd Qu.: 81.30   3rd Qu.:165.0
##  Max.   :1.000   Max.   :90.00   Max.   :127.00   Max.   :199.0
##       BMI           PREMENO          MOMFRAC       ARMASSIST
##  Min.   :14.88   Min.   :0.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:23.27   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000
##  Median :26.42   Median :0.000   Median :0.00   Median :0.000
##  Mean   :27.55   Mean   :0.194   Mean   :0.13   Mean   :0.376
##  3rd Qu.:30.79   3rd Qu.:0.000   3rd Qu.:0.00   3rd Qu.:1.000
##  Max.   :49.08   Max.   :1.000   Max.   :1.00   Max.   :1.000
##      SMOKE          RATERISK       FRACTURE
##  Min.   :0.00   Min.   :1.00   Min.   :0.00
##  1st Qu.:0.00   1st Qu.:1.00   1st Qu.:0.00
##  Median :0.00   Median :2.00   Median :0.00
##  Mean   :0.07   Mean   :1.96   Mean   :0.25
##  3rd Qu.:0.00   3rd Qu.:3.00   3rd Qu.:0.25
##  Max.   :1.00   Max.   :3.00   Max.   :1.00
```

```r
# Standard Deviations for the non-categorical columns
std=sapply(set_noID,sd)
print('The standard deviations are:')
```

```
## [1] "The standard deviations are:"
```

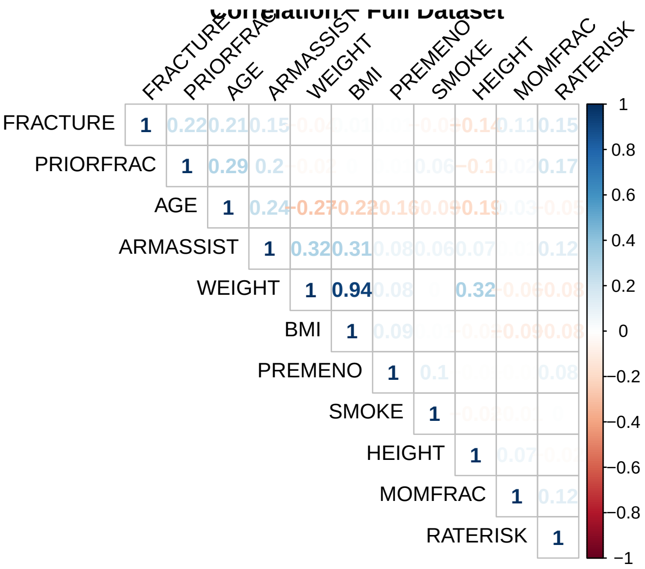
```r
print(std)
```

```
##  PRIORFRAC        AGE     WEIGHT     HEIGHT        BMI    PREMENO
##  0.4345961  8.9895372 16.4359918  6.3554928  5.9739583  0.3958249
##    MOMFRAC  ARMASSIST      SMOKE   RATERISK   FRACTURE
##  0.3366402  0.4848651  0.2554025  0.7922470  0.4334464
```

```r
# Skewness
#The further the distribution of the skew value from zero,
# the larger the skew to the left (negative skew value) or right (positive skew value).
#library(e1071) # the library for skewness
library(e1071)

skew=apply(set_noID[,c(1:11)], 2, skewness)
print(skew)
```

```
##  PRIORFRAC        AGE     WEIGHT     HEIGHT        BMI    PREMENO
## 1.13900707 0.42737676 0.80951443 0.25181286 0.85717095 1.54304591
##    MOMFRAC  ARMASSIST      SMOKE   RATERISK   FRACTURE
## 2.19379594 0.51045948 3.36049899 0.07085255 1.15123817
```

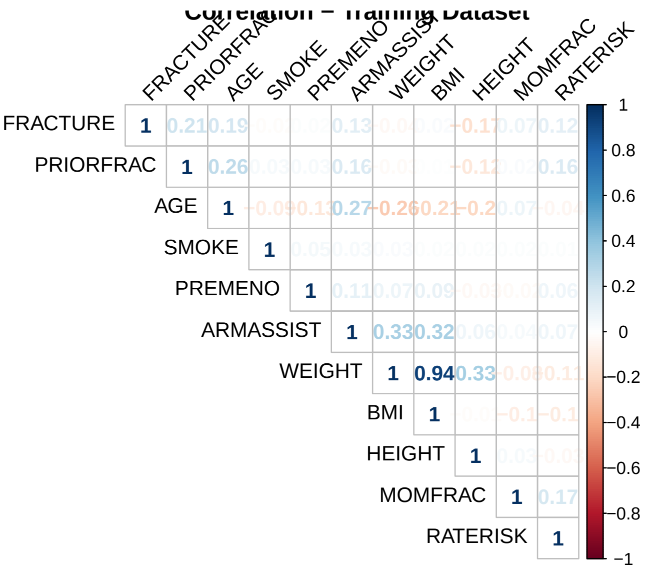
```r
# Correlations
library(corrplot)

#Full dataset without ID columns
corrplot(cor(set_noID), method = "number", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45,main="Correlation - Full Dataset")
```



```
#Training data set without ID columns
#split the data into training and validation sets
library(caret)
```

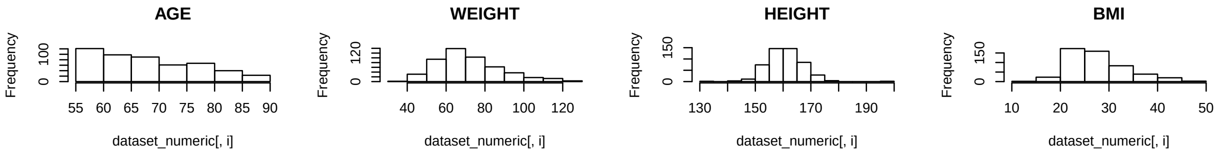
```
## Loading required package: lattice
```

```
set.seed(84)
validation_index = createDataPartition(dataset$FRACTURE, p=0.75, list=FALSE)
validationData = set_noID[-validation_index,]
trainingData = set_noID[validation_index,]
corrplot(cor(trainingData), method = "number", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, main="Correlation - Training Dataset")
```

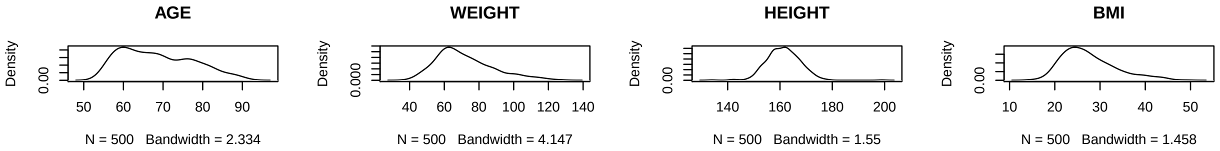


```
# Data visualizations
dataset_numeric = numericVar
```

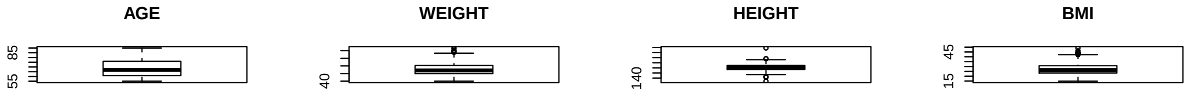
```
#Histograms
par(mfrow=c(3,4)) # put four figures in a row (2*4)
for (i in 1:4) {
  hist(dataset_numeric[,i],main=names(dataset_numeric)[i])
}

#Density Plots
par(mfrow=c(3,4))
```



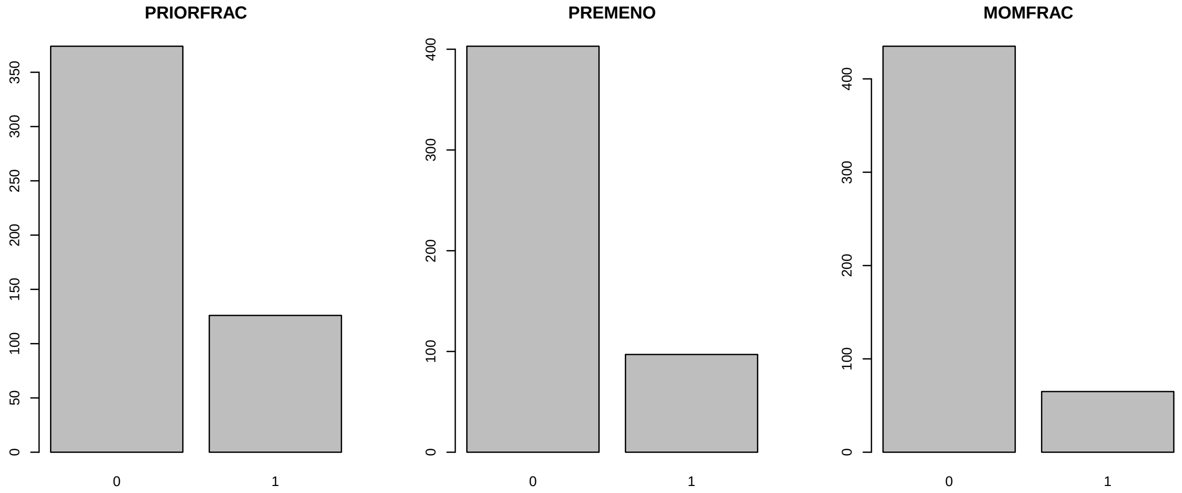
```
for(i in 1:4) {
  plot(density(dataset_numeric[,i]), main=names(dataset_numeric)[i])
}

#Box And Whisker Plots
par(mfrow=c(3,4))
```

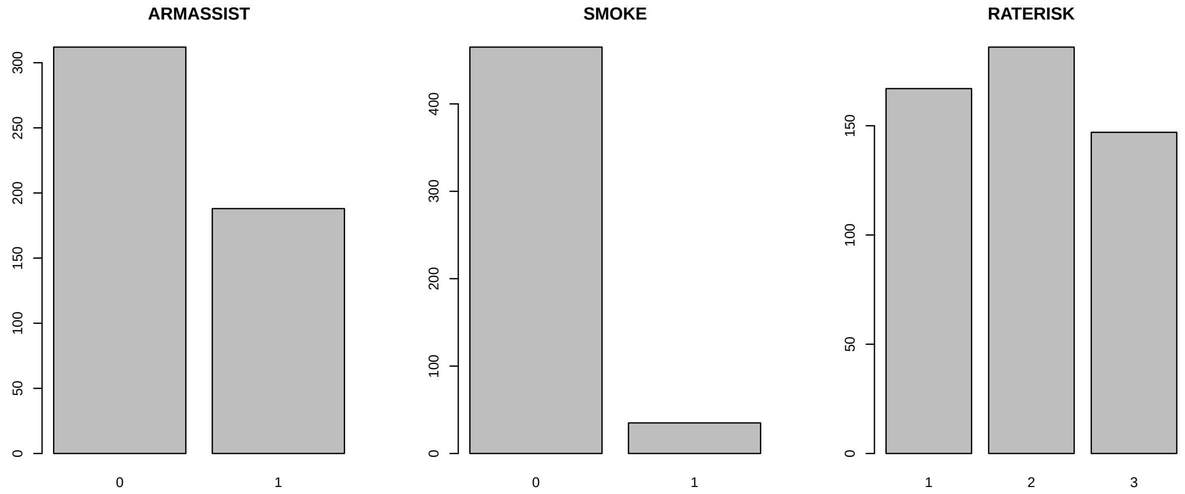


```r
for(i in 1:4) {
  boxplot(dataset_numeric[,i], main=names(dataset_numeric)[i])
}

#Barplots, which is used to count the accurances for categorical attributes
dataset_categorical = set_noID[,-c(2:5)]
par(mfrow=c(1,3))
```



```r
for(i in 1:7) {
  counts <- table(dataset_categorical[,i]) # get the count for each categorical value
  name <- names(dataset_categorical)[i]
  barplot(counts, main=name)
}
```

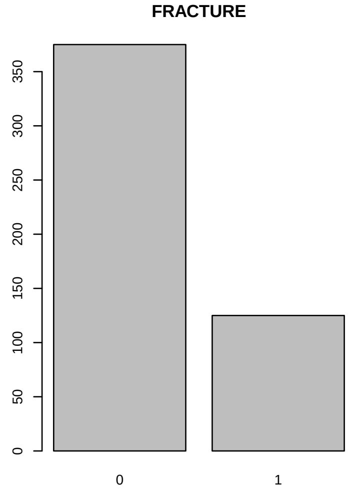




```
#Multivariate Visualization
library(corrplot) # for function corrplot()
correlations1=cor(dataset_numeric)
print(correlations1)
```

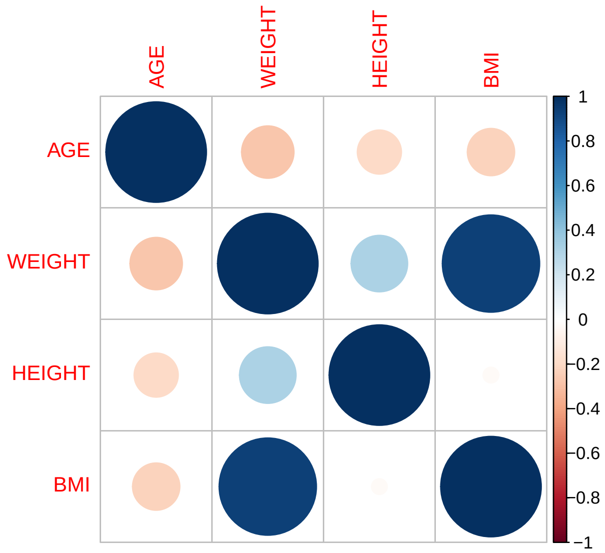
```
##                AGE     WEIGHT      HEIGHT         BMI
## AGE     1.0000000 -0.2715964 -0.19264861 -0.22125651
## WEIGHT -0.2715964  1.0000000  0.31596915  0.93733603
## HEIGHT -0.1926486  0.3159691  1.00000000 -0.02437689
## BMI    -0.2212565  0.9373360 -0.02437689  1.00000000
```

```
par(mfrow=c(1,1))
```

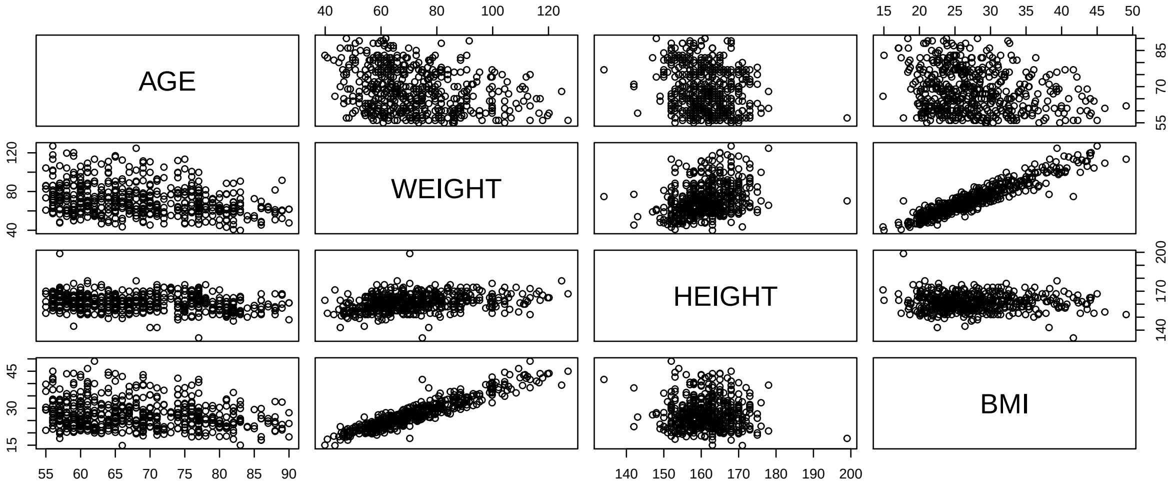


```
corrplot(correlations1, methods="circle")
```

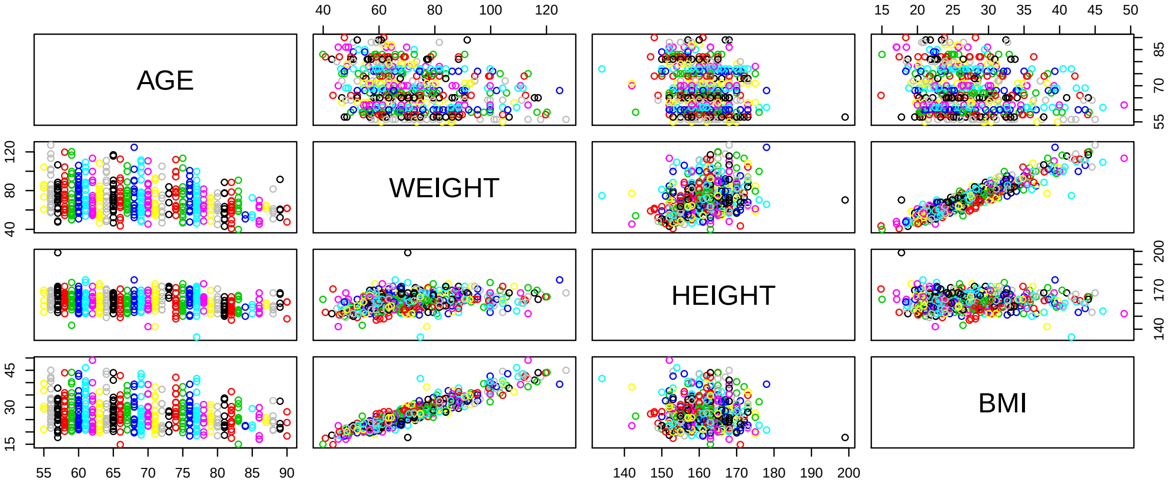
```
## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt
## = tl.srt, : "methods" is not a graphical parameter

## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col
## = tl.col, : "methods" is not a graphical parameter

## Warning in title(title, ...): "methods" is not a graphical parameter
```



```
# pair-wise scatterplots of the numeric attributes
par(mfrow=c(1,1))
pairs(dataset_numeric)
```



```
#Scatterplot Matrix By Class (use different color to distinguish different class)
par(mfrow=c(1,1))
pairs(dataset_numeric, col=dataset[,5])
```



```
#Density By Class
library(caret)

# load the data
data(iris)

# density plots for each attribute by class value
x <- dataset_numeric
y <- dataset[,5]
scales <- list(x=list(relation="free"), y=list(relation="free"))
par(mfrow=c(1,1))
featurePlot(x=dataset_numeric, y=dataset[,5], plot="density", scales=scales)
```

```
## NULL
```

```
#Box And Whisker Plots By Class
featurePlot(x=dataset_numeric, y=dataset[,5], plot="box")
```

```
## NULL
```

Logistic Regression

Train / Test

Training set will be 70% of dataset and Test set will be remaining 30%

```
#smp_size <- floor(0.70 * nrow(dataset))
#set.seed(1234)
#train_ind <-sample(seq_len(nrow(dataset)), size=smp_size)
#test <-dataset[-train_ind,]
#train <-dataset[train_ind,]
```

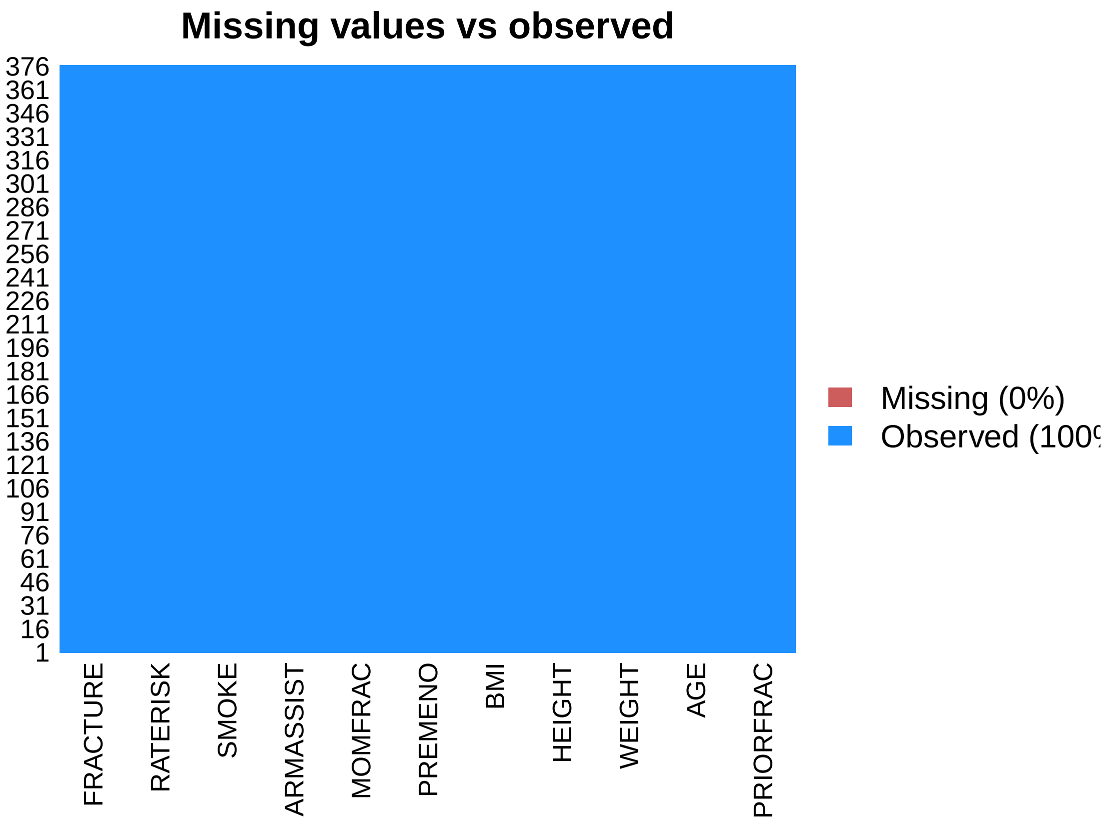
```r
# split the data into training and validation sets
library(caret)
set.seed(84)
validation_index = createDataPartition(dataset$FRACTURE, p=0.75, list=FALSE)
validationData = set_noID[-validation_index,]
trainingData = set_noID[validation_index,]

#check for Missing Data
sapply(trainingData,function(x) sum(is.na(x)))
```

```
## PRIORFRAC       AGE    WEIGHT    HEIGHT       BMI   PREMENO   MOMFRAC
##         0         0         0         0         0         0         0
## ARMASSIST     SMOKE  RATERISK  FRACTURE
##         0         0         0         0
```

```r
sapply(trainingData, function(x) length(unique(x)))
```

```
## PRIORFRAC       AGE    WEIGHT    HEIGHT       BMI   PREMENO   MOMFRAC
##         2        36       120        33       318         2         2
## ARMASSIST     SMOKE  RATERISK  FRACTURE
##         2         2         3         2
```

```r
missmap(trainingData, main = "Missing values vs observed") #library(Amelia)
```



##Build Model

Question of Interest? What are the odds of getting a fracture, given certain conditions?

```
set.seed(84)
model <- glm(FRACTURE~.,family = "binomial" (link='logit'), data=set_noID)
model
```

```
##
## Call:  glm(formula = FRACTURE ~ ., family = binomial(link = "logit"),
##     data = set_noID)
##
## Coefficients:
## (Intercept)    PRIORFRAC          AGE       WEIGHT       HEIGHT
##   -16.03863      0.67285      0.03915     -0.12152      0.06564
##         BMI      PREMENO      MOMFRAC    ARMASSIST        SMOKE
##     0.33126      0.10438      0.63679      0.35875     -0.31360
##    RATERISK
##     0.37666
##
## Degrees of Freedom: 499 Total (i.e. Null);  489 Residual
## Null Deviance:       562.3
## Residual Deviance: 503.9     AIC: 525.9
```

```
summary(model)
```

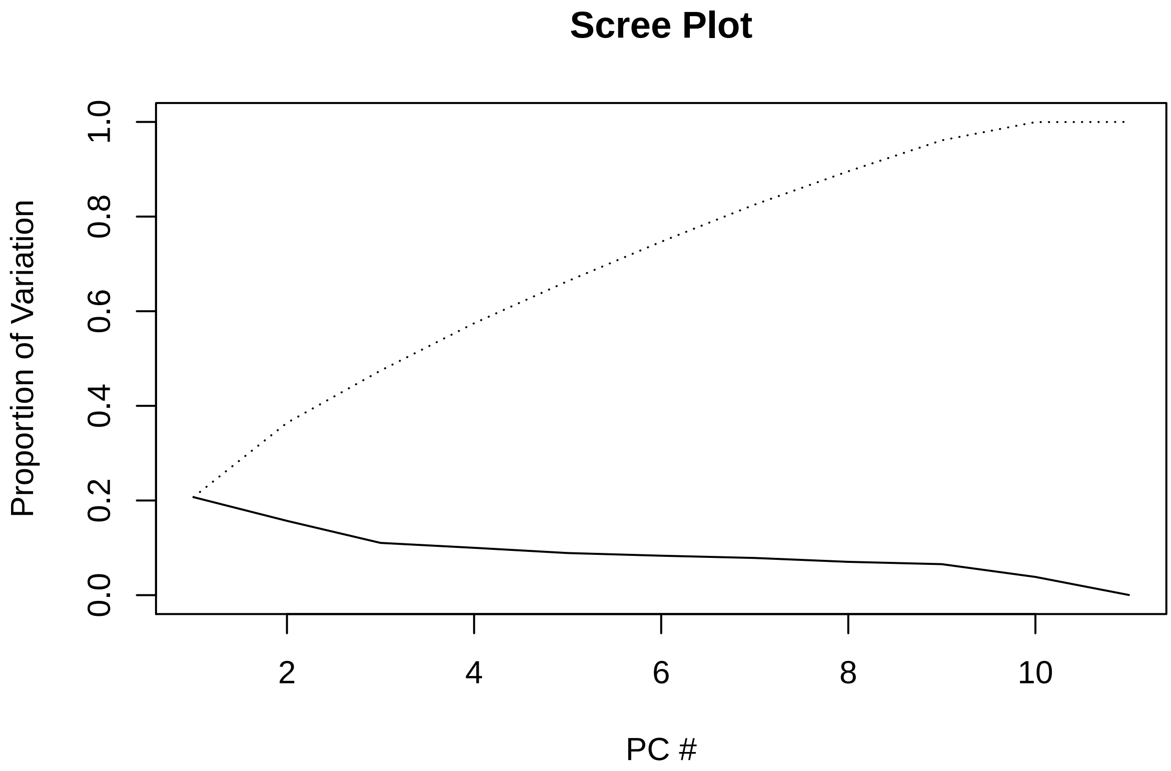
```
##
## Call:
## glm(formula = FRACTURE ~ ., family = binomial(link = "logit"),
##     data = set_noID)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6882  -0.7254  -0.5654  -0.0960   2.2111
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.03863   12.66515  -1.266  0.20538
## PRIORFRAC     0.67285    0.24967   2.695  0.00704 **
## AGE           0.03915    0.01471   2.662  0.00778 **
## WEIGHT       -0.12152    0.08658  -1.404  0.16044
## HEIGHT        0.06564    0.07815   0.840  0.40098
## BMI           0.33126    0.22326   1.484  0.13787
## PREMENO       0.10438    0.28486   0.366  0.71406
## MOMFRAC       0.63679    0.30767   2.070  0.03848 *
## ARMASSIST     0.35875    0.25615   1.401  0.16134
## SMOKE        -0.31360    0.46222  -0.678  0.49747
## RATERISK      0.37666    0.14896   2.529  0.01145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 562.34  on 499  degrees of freedom
## Residual deviance: 503.87  on 489  degrees of freedom
## AIC: 525.87
##
## Number of Fisher Scoring iterations: 4
```

Interpretation of logistic regression model: Weight, height, BMI, Premeno, Armassist, and Smoke are not statistically significant variables. Priorfrac and Age are statistically significant variables and have the lowest p-value indicating a strong association with having a Fracture.

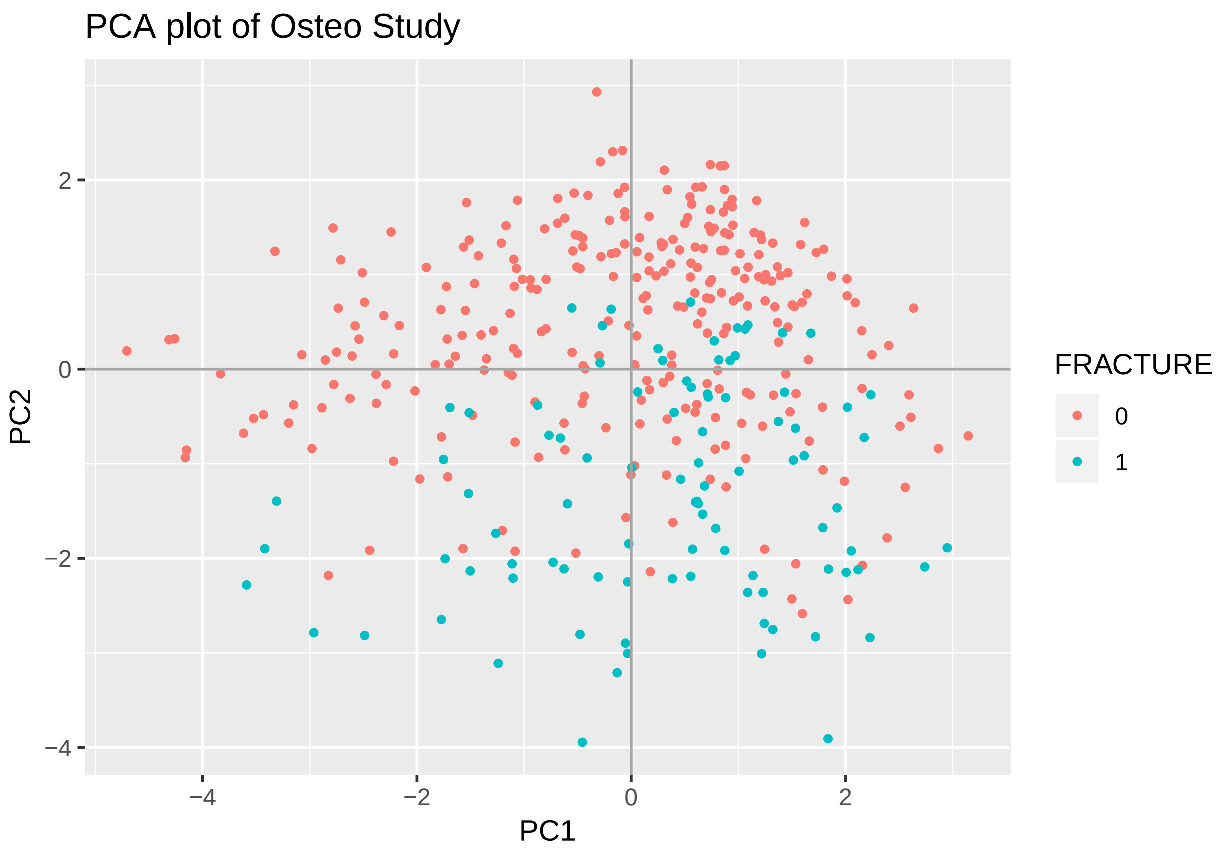
```
##glmnet
dat <- categoricalVar

#Get Training Set
dat.train <- trainingData

dat.train.x <- dat.train[,1:ncol(dat.train)]
dat.train.y <- dat.train$FRACTURE

dat.train.y <- as.factor(as.character(dat.train.y))

#PCA
pc.result<-prcomp(dat.train.x,scale.=TRUE)
pc.scores<-pc.result$x
pc.scores<-data.frame(pc.scores)
pc.scores$FRACTURE<-dat.train.y

PCA <- pc.result$rotation
PCA
```

```
##                   PC1         PC2         PC3         PC4         PC5
## PRIORFRAC  0.07405715 -0.47973677  0.04196542  0.09391798 -0.13590836
## AGE        0.25108797 -0.44246880 -0.36503870 -0.23606187 -0.19052036
## WEIGHT    -0.64071593 -0.07386956 -0.01492425 -0.09380628  0.04511049
## HEIGHT    -0.22890271  0.24187107  0.22588400 -0.52730271 -0.16245665
## BMI       -0.59713909 -0.16251927 -0.09572198  0.09472088  0.11445051
## PREMENO   -0.11505234 -0.07693268  0.32458960  0.55307931  0.16142556
## MOMFRAC    0.11007196 -0.13173669  0.46323427 -0.46128883  0.06458940
## ARMASSIST -0.25734750 -0.47047462 -0.02257374 -0.19863859 -0.16959440
## SMOKE     -0.04972241  0.01391670  0.31483340  0.26442893 -0.86124378
## RATERISK   0.10918438 -0.20973926  0.61992312 -0.03166000  0.26572927
## FRACTURE   0.07750976 -0.44147408  0.02641809  0.10280064  0.17988806
##                    PC6         PC7         PC8         PC9        PC10
## PRIORFRAC  0.009110347 -0.51653703  0.12726745 -0.65076852 -0.16932571
## AGE        0.245357588  0.13034942  0.07530809  0.08177107  0.65295027
## WEIGHT    -0.120183672 -0.03235101  0.02116421 -0.08190948  0.22942735
## HEIGHT     0.327444296 -0.34498645 -0.46918526 -0.08143823  0.14936479
## BMI       -0.245140543  0.09823153  0.18404370 -0.05154601  0.18874838
## PREMENO    0.622965173  0.25696963 -0.17636604 -0.16498653  0.17626188
## MOMFRAC   -0.148794845  0.59700737  0.10567492 -0.38334831 -0.04799385
## ARMASSIST  0.368115390  0.11545494  0.07410041  0.40014322 -0.57380527
## SMOKE     -0.230599843  0.09342985 -0.02981247  0.11155875  0.10632225
## RATERISK  -0.084062927 -0.37403086  0.29943887  0.43734477  0.24549790
## FRACTURE  -0.389848139  0.04558067 -0.76508915  0.12573228 -0.01463757
##                    PC11
## PRIORFRAC -0.0073515954
## AGE       -0.0025546320
## WEIGHT     0.7054674906
## HEIGHT    -0.2415073104
## BMI       -0.6662434928
## PREMENO    0.0042646791
## MOMFRAC    0.0004639497
## ARMASSIST -0.0001841958
## SMOKE     -0.0025392898
## RATERISK  -0.0004029694
## FRACTURE   0.0048813762
```

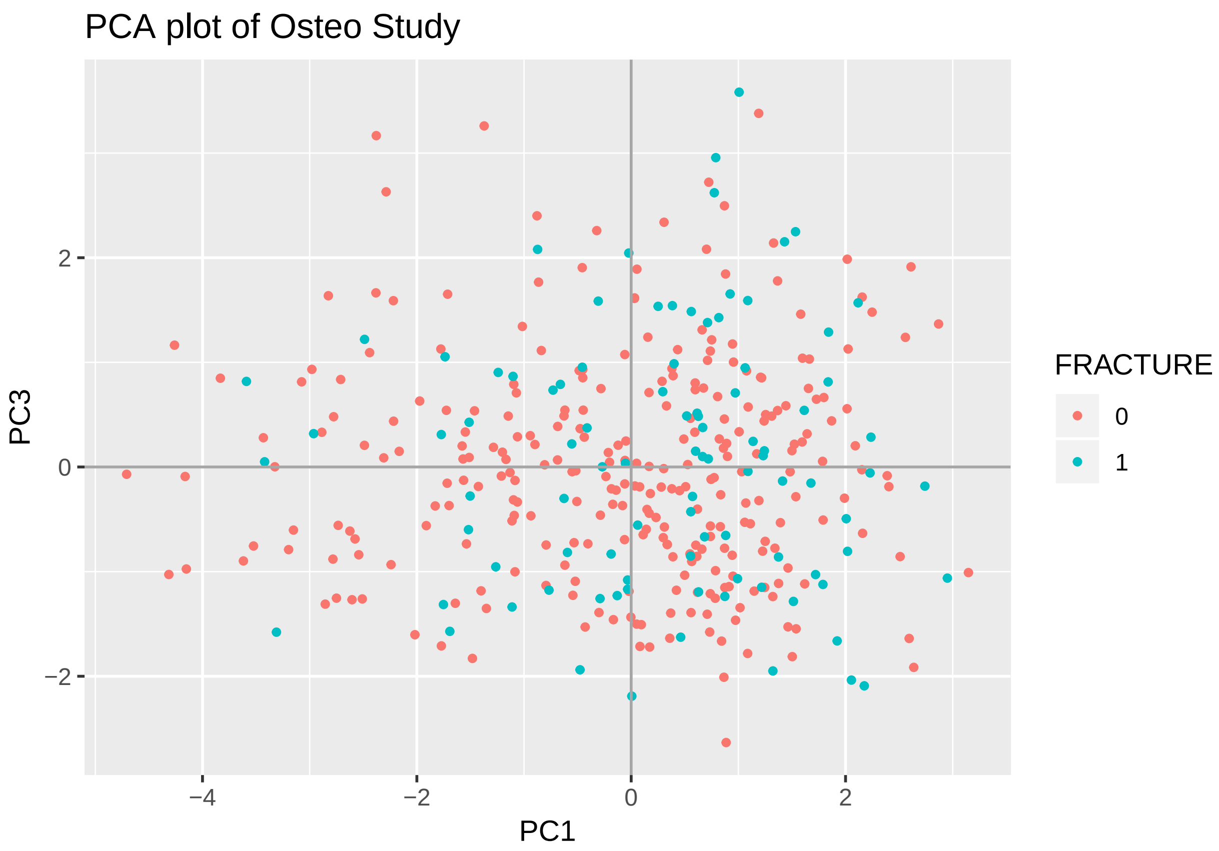
```
#Scree plot
pc.eigen<-(pc.result$sdev)^2
pc.prop<-pc.eigen/sum(pc.eigen)
pc.cumprop<-cumsum(pc.prop)
plot(1:11,pc.prop,type="l",main="Scree Plot",ylim=c(0,1),xlab="PC #",ylab="Proportion of Variation")
 lines(1:11,pc.cumprop,lty=3)
```



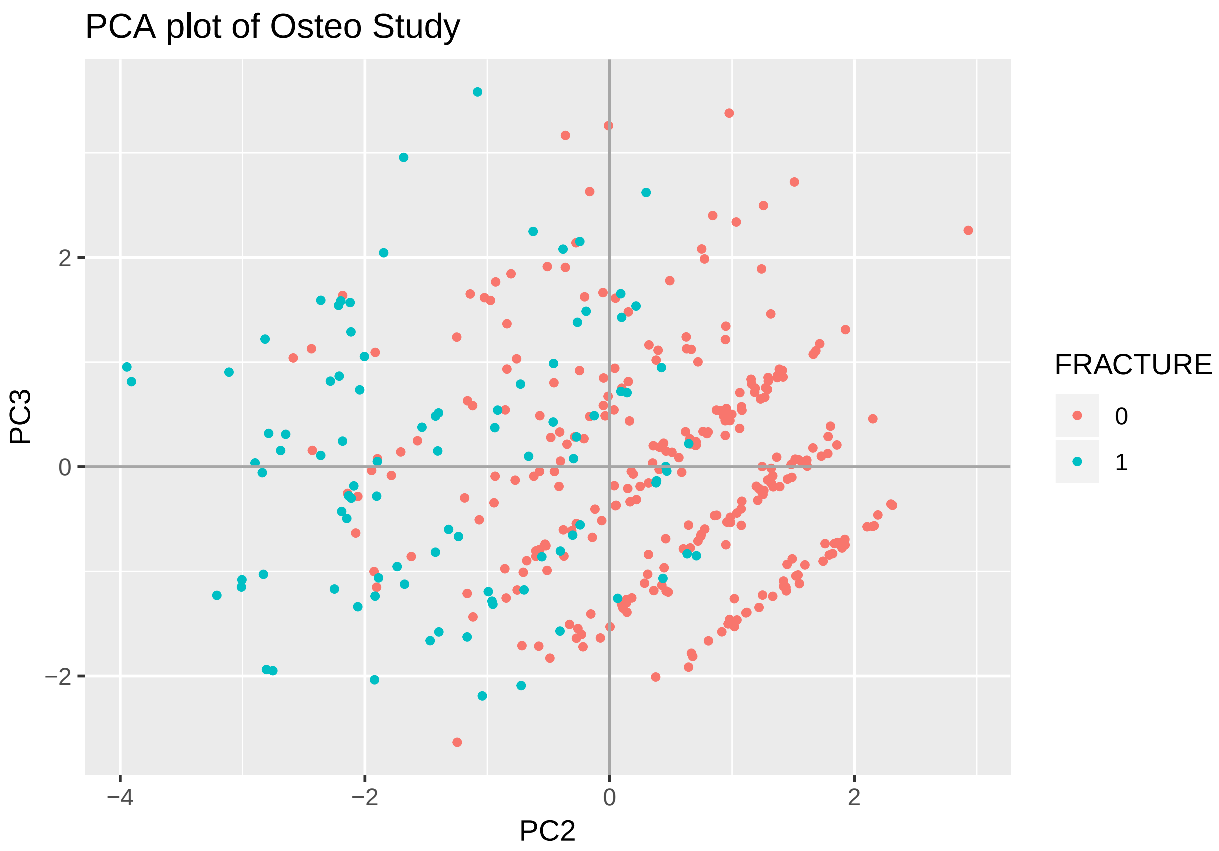
```
#Use ggplot2 to plot the first few pc's
ggplot(data = pc.scores, aes(x = PC1, y = PC2)) +  geom_point(aes(col=FRACTURE), size=1)+ geom_hline(yi
```



```
ggplot(data = pc.scores, aes(x = PC1, y = PC3)) +  geom_point(aes(col=FRACTURE), size=1)+ geom_hline(yi
```

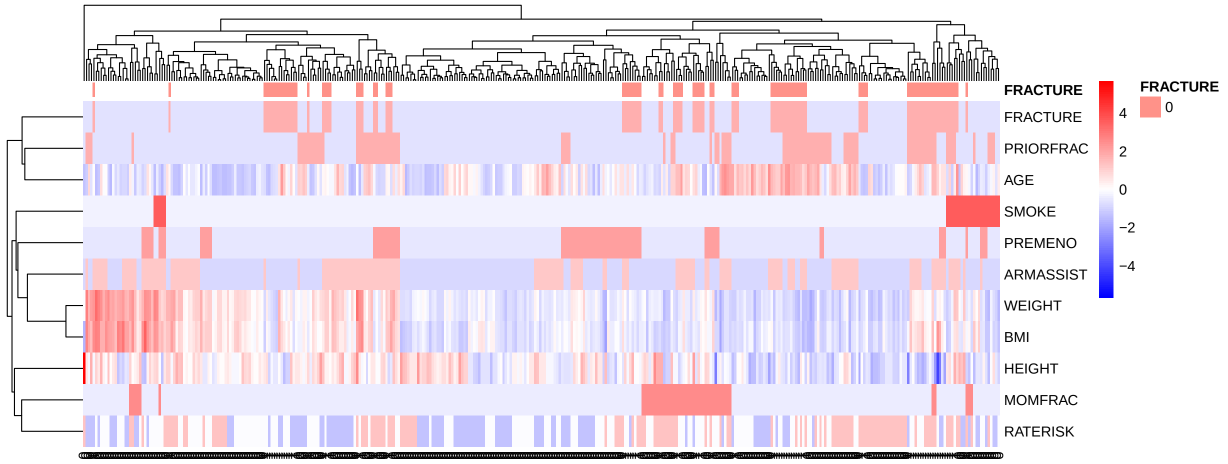


```
ggplot(data = pc.scores, aes(x = PC2, y = PC3)) +  geom_point(aes(col=FRACTURE), size=1)+ geom_hline(yi
```

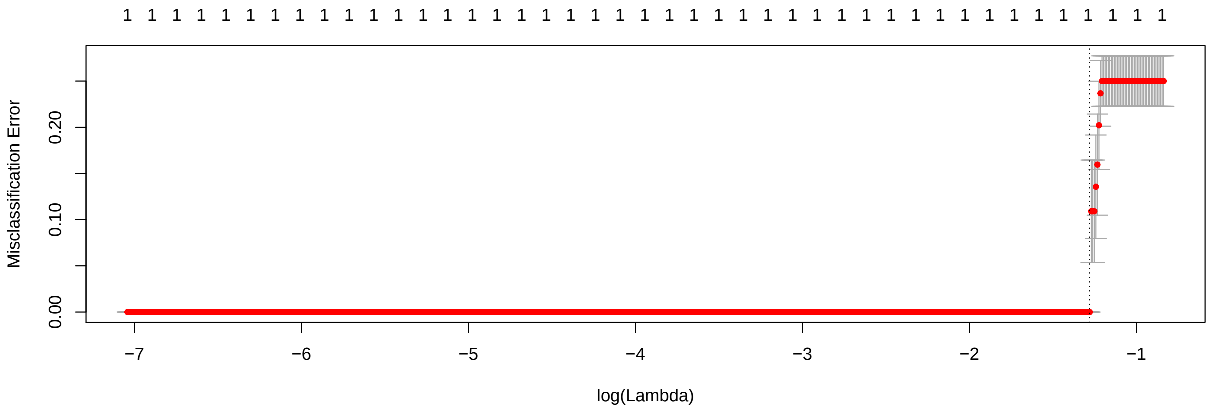


##Clustering

```
#Lets look at a heatmap using hierarchical clustering to see if the
#response naturually clusters out using the predictors

#Transposting the predictor matrix and giving the response categories its
#row names.
library(RColorBrewer)
x<-t(dat.train.x)
colnames(x)<-dat.train.y
pheatmap(x,annotation_col=data.frame(FRACTURE=dat.train.y),scale="row",legend=T,color=colorRampPalette(
```



```
##logistic regression
dat.train.x <- as.matrix(dat.train.x)
library(glmnet)
cvfit <- cv.glmnet(dat.train.x, dat.train.y, family = "binomial", type.measure = "class", nlambda = 100
plot(cvfit)
```



```
coef(cvfit, s = "lambda.min")
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                     1
## (Intercept) -1.653758
## PRIORFRAC    .
## AGE          .
## WEIGHT       .
## HEIGHT       .
## BMI          .
## PREMENO      .
## MOMFRAC      .
## ARMASSIST    .
## SMOKE        .
## RATERISK     .
## FRACTURE     1.726571
```