Applied Regression With R

Bruce Campbell

July 17, 2017

Wed Jul 19 14:40:28 2017

Chapter 1 examples

```
if (!require(faraway)) {
    install.packages("faraway")
    library(faraway)
}

if (!require(HistData)) {
    install.packages("HistData")
    library(HistData)
}
```

```
## Loading required package: HistData
```

Diabetes survey on Pima Indians

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix.

```
# Loads or lists available datasets
data(pima, package = "faraway")
head(pima)
```

```
##   pregnant glucose diastolic triceps insulin  bmi diabetes age test
## 1        6     148        72      35       0 33.6    0.627  50    1
## 2        1      85        66      29       0 26.6    0.351  31    0
## 3        8     183        64       0       0 23.3    0.672  32    1
## 4        1      89        66      23      94 28.1    0.167  21    0
## 5        0     137        40      35     168 43.1    2.288  33    1
## 6        5     116        74       0       0 25.6    0.201  30    0
```

```
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           bmi          diabetes           age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
```

```
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##       test
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

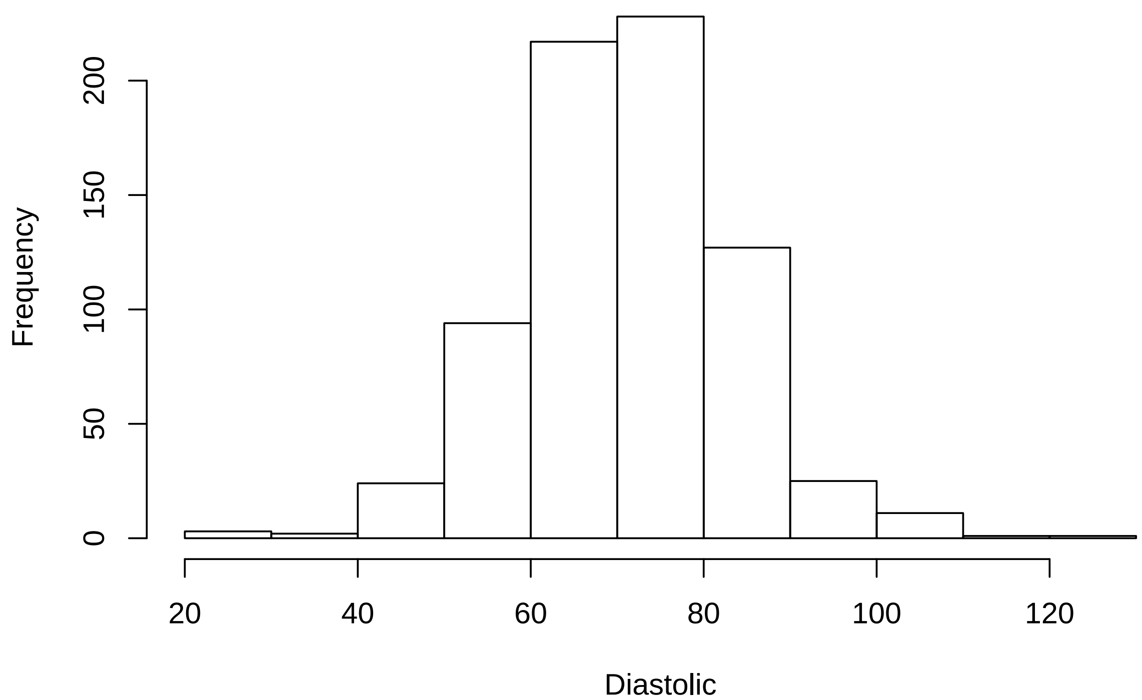
```r
# From the summary - we see that we have zero's for physical variable We set
# them to NA - this is an important part ot due diligence in statistics
# Check that the values make sense.  sort(pima$diastolic)
pima$diastolic[pima$diastolic == 0] <- NA
pima$glucose[pima$glucose == 0] <- NA
pima$triceps[pima$triceps == 0] <- NA
pima$insulin[pima$insulin == 0] <- NA
pima$bmi[pima$bmi == 0] <- NA
pima$test <- factor(pima$test)
summary(pima$test)
```

```
##   0   1
## 500 268
```

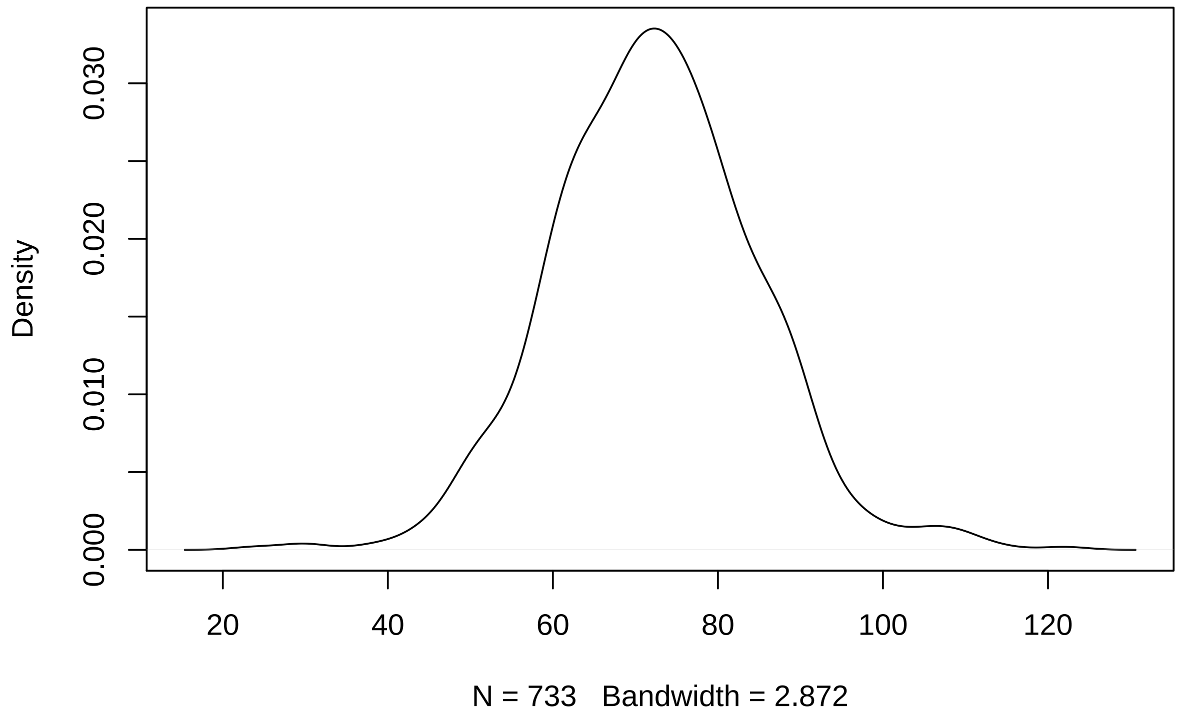
```r
levels(pima$test) <- c("negative", "positive")
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   : 44.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7   Mean   : 72.41   Mean   :29.15
##  3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##                   NA's   :5       NA's   :35       NA's   :227
##     insulin            bmi           diabetes           age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780   Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437   1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725   Median :29.00
##  Mean   :155.55   Mean   :32.46   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##  NA's   :374      NA's   :11
##        test
##  negative:500
##  positive:268
##
##
##
##
##
```

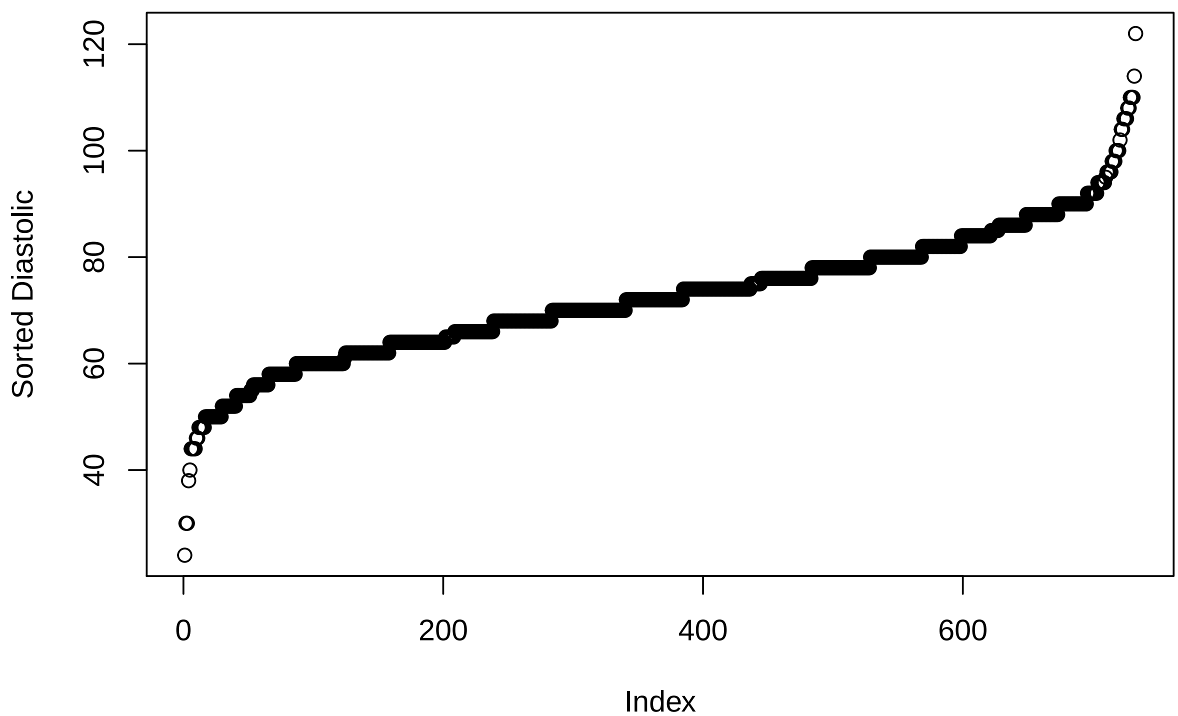
```r
hist(pima$diastolic, xlab = "Diastolic", main = "")
```



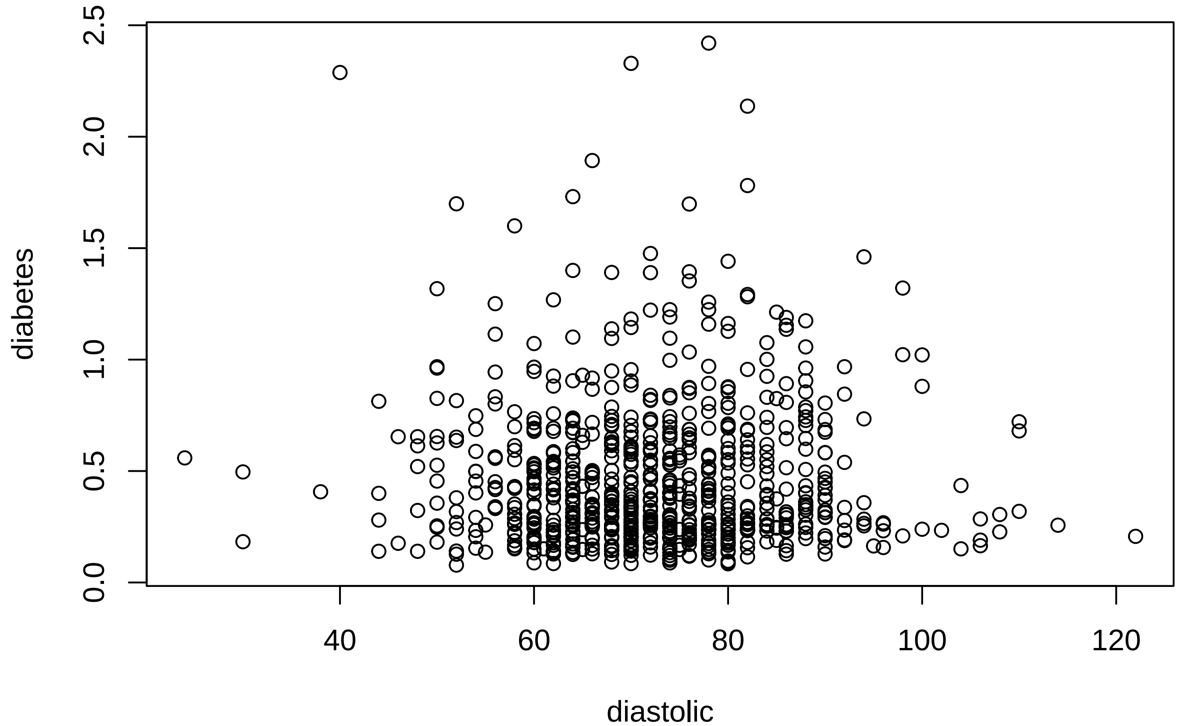

```r
plot(density(pima$diastolic, na.rm = TRUE), main = "")
```



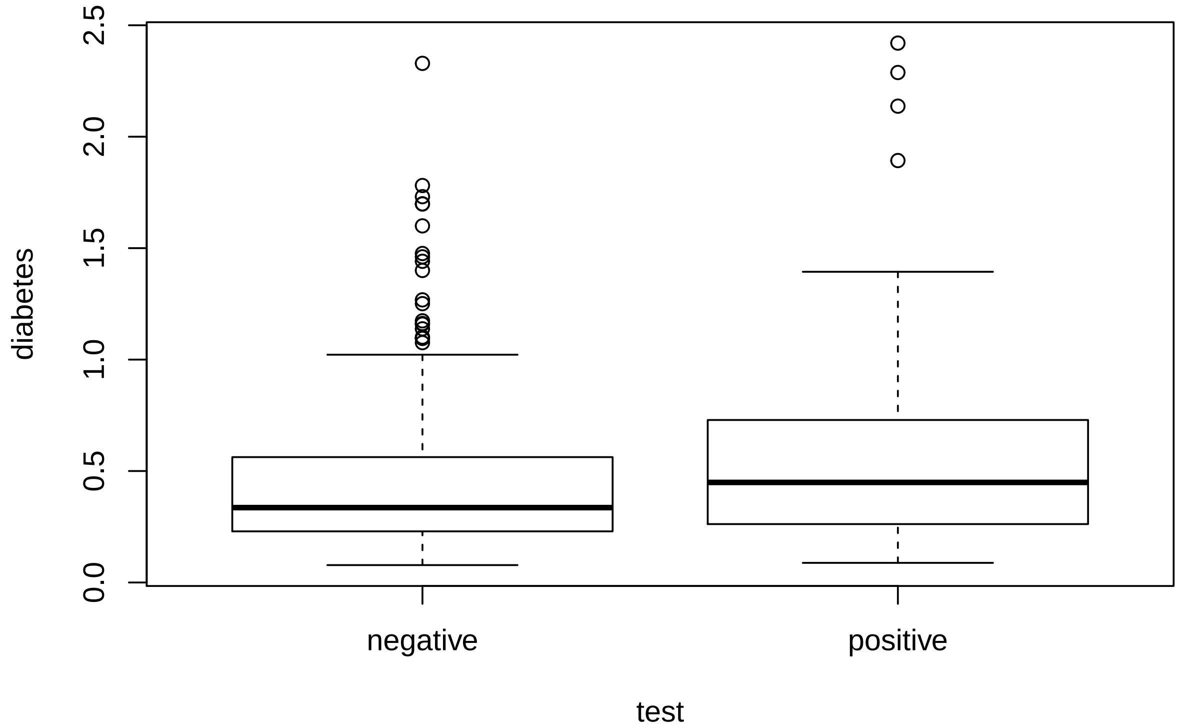

```
plot(sort(pima$diastolic), ylab = "Sorted Diastolic")
```




```
plot(diabetes ~ diastolic, pima)
```

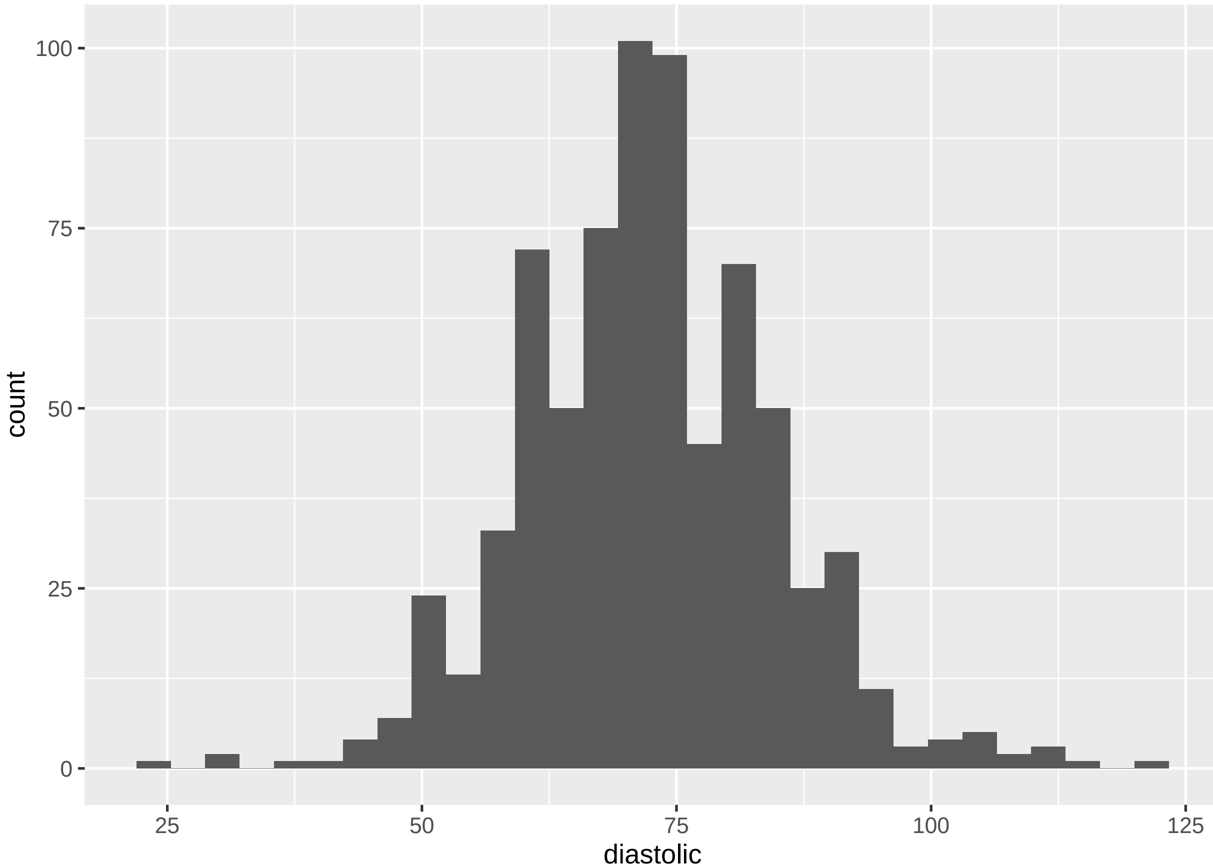



```
plot(diabetes ~ test, pima)
```

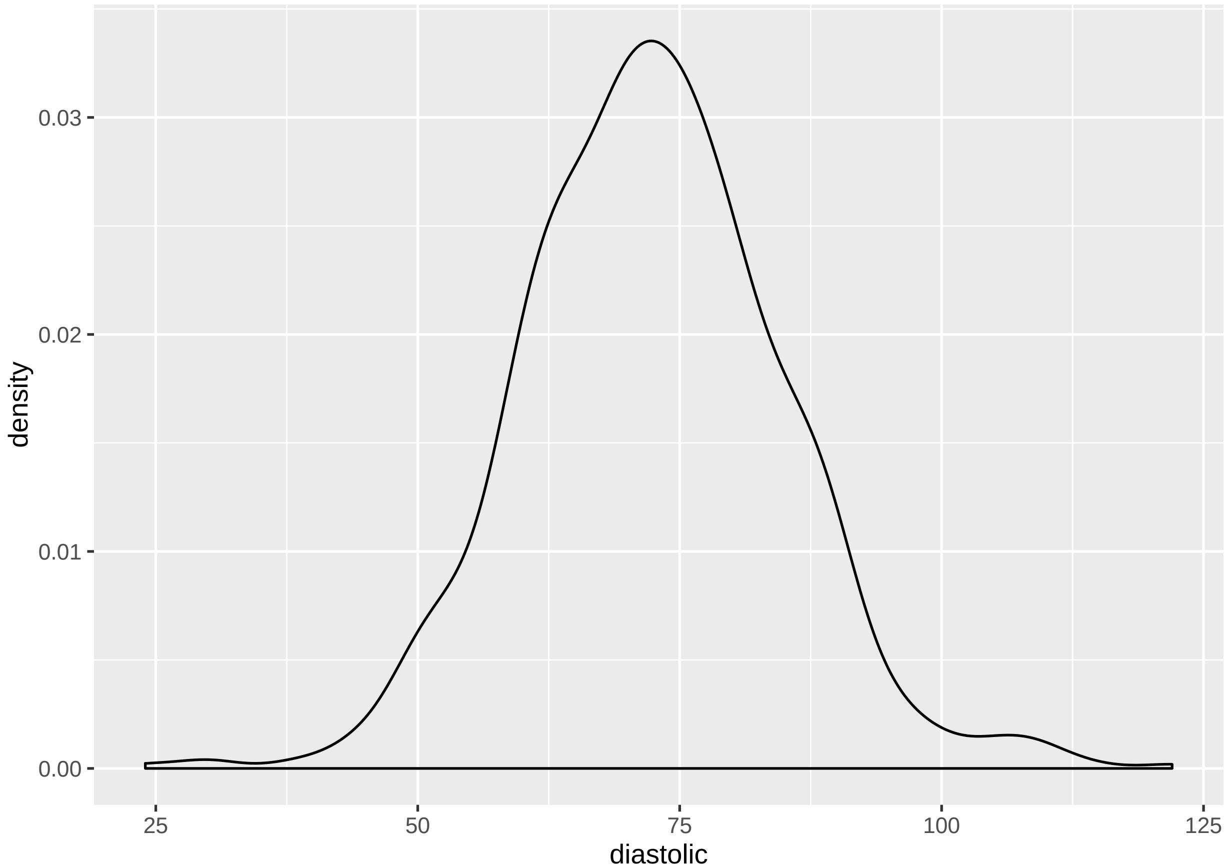



```
require(ggplot2)
ggplot(pima, aes(x = diastolic)) + geom_histogram()
```

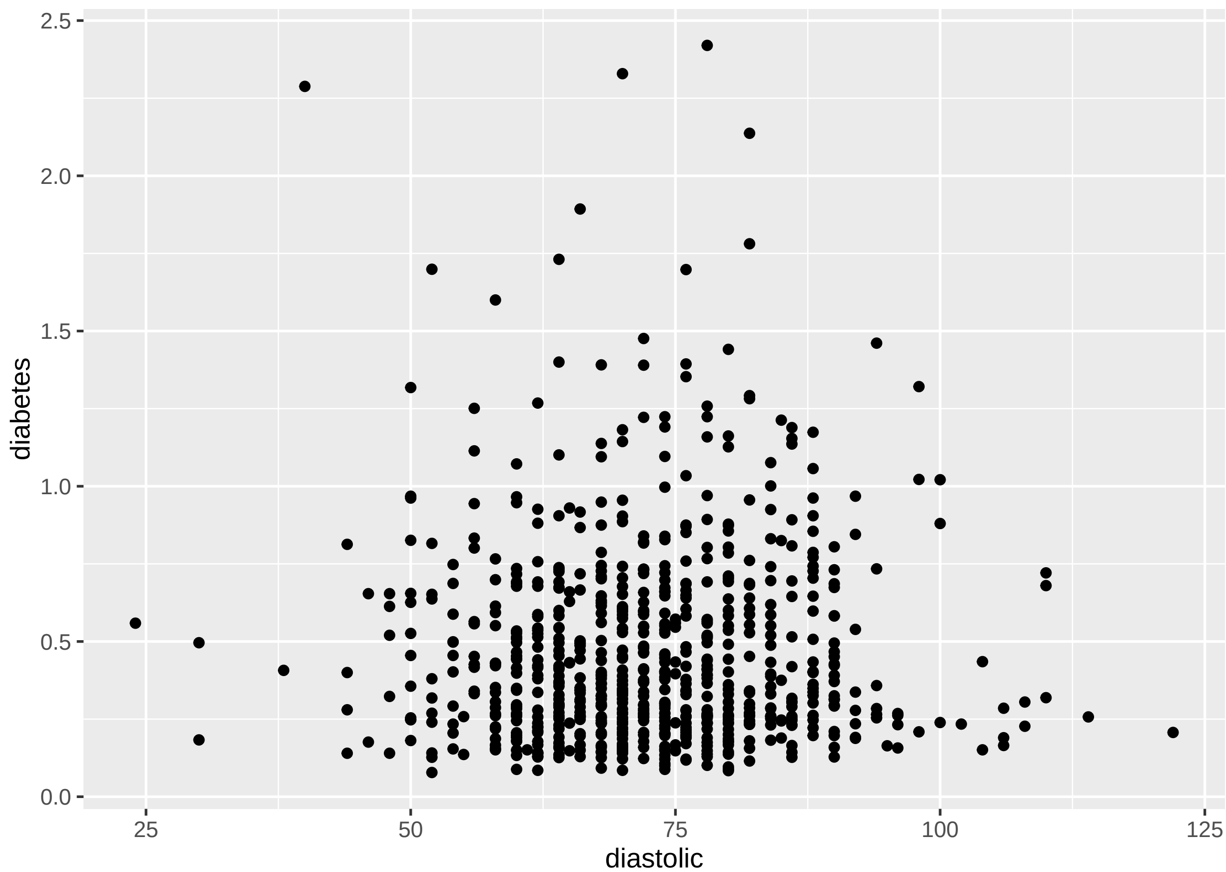
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



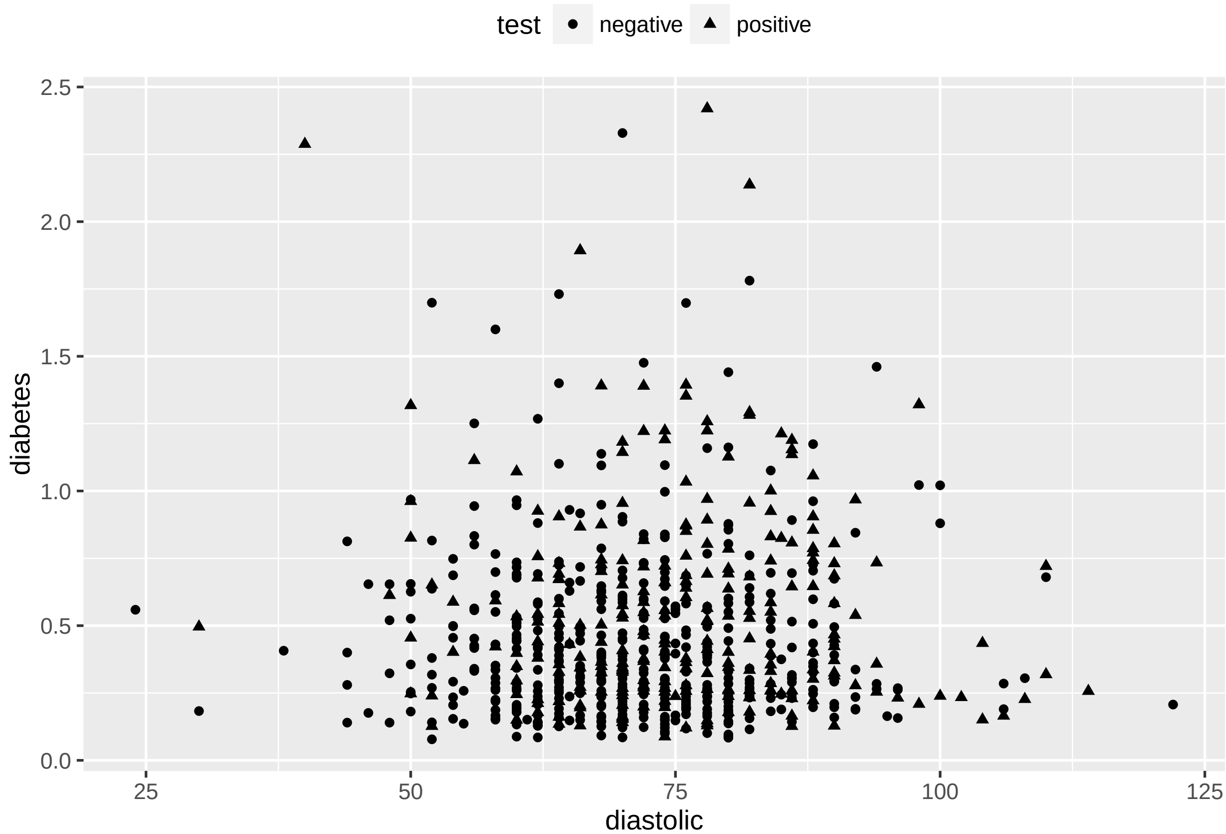

```r
ggplot(pima, aes(x = diastolic)) + geom_density()
```



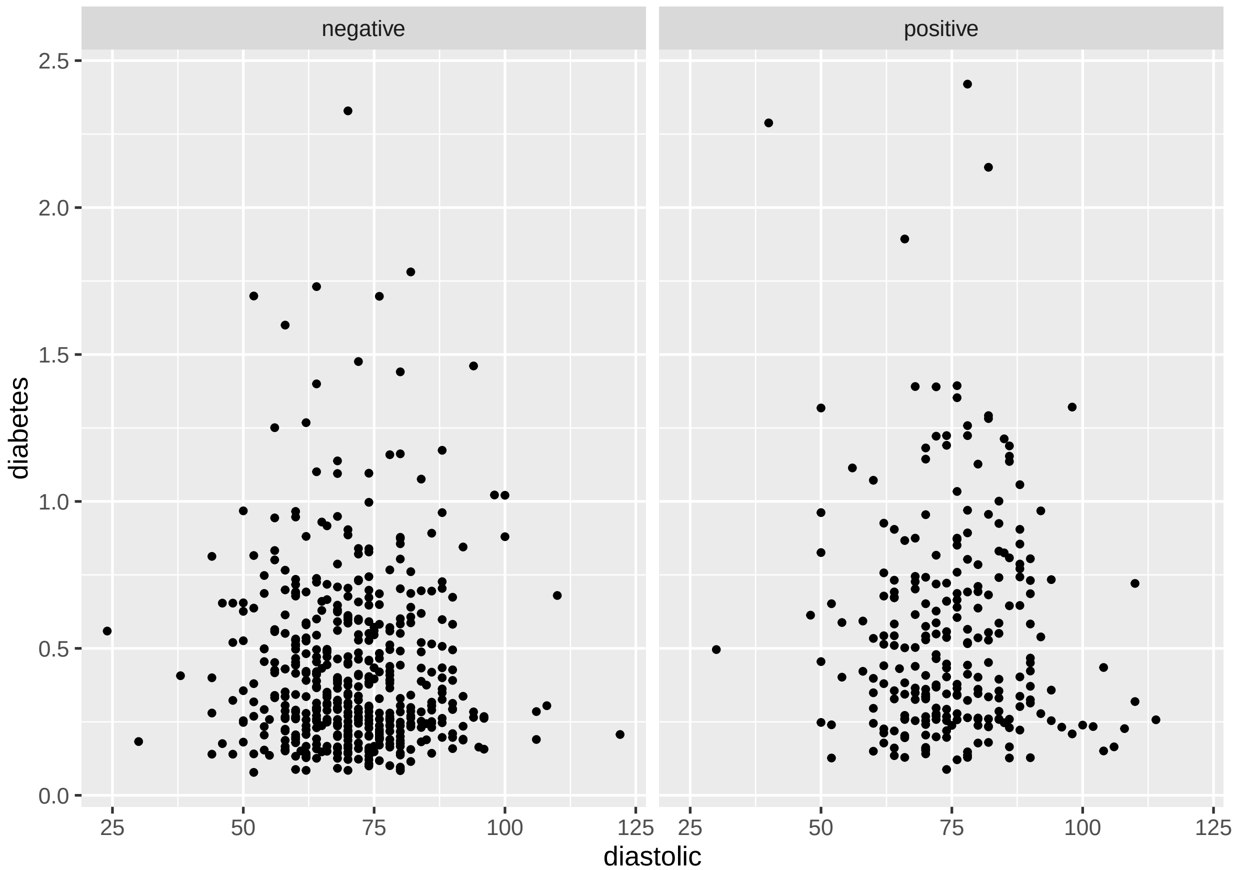

```
ggplot(pima, aes(x = diastolic, y = diabetes)) + geom_point()
```




```
ggplot(pima, aes(x = diastolic, y = diabetes, shape = test)) + geom_point() +
    theme(legend.position = "top", legend.direction = "horizontal")
```



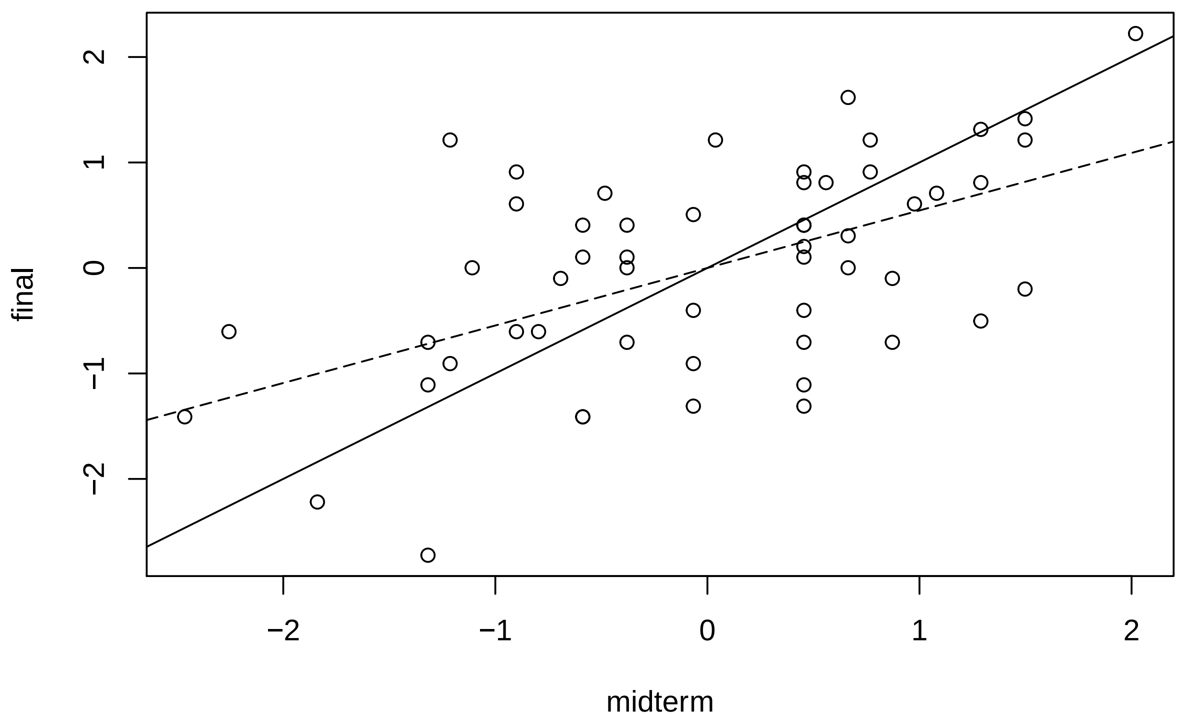

```
ggplot(pima, aes(x = diastolic, y = diabetes)) + geom_point(size = 1) + facet_grid(~test)
```



Marks in a statistics class

Marks from Statistics 500 one year at the University of Michigan

```
data(stat500)
stat500 <- data.frame(scale(stat500))
plot(final ~ midterm, stat500)
abline(0, 1)
g <- lm(final ~ midterm, stat500)
abline(coef(g), lty = 5)
```




```
cor(stat500)
```

```
##            midterm      final         hw     total
## midterm 1.0000000 0.54522775 0.27205756 0.8444568
## final   0.5452277 1.00000000 0.08733764 0.7788629
## hw      0.2720576 0.08733764 1.00000000 0.5644286
## total   0.8444568 0.77886293 0.56442864 1.0000000
```

Mayer's 1750 data on the Manilius crater on the moon

In 1750, Tobias Mayer collected data on various landmarks on the moon in order to determine its orbit. The data involving the position of the Manilius crater resulted in a least squares like problem. The example is discussed in Steven Stigler's History of Statistics

```
data(manilius, package = "faraway")
head(manilius)
```

```
##        arc sinang  cosang group
## 1 13.16667 0.8836 -0.4682     1
## 2 13.13333 0.9996 -0.0282     1
## 3 13.20000 0.9899  0.1421     1
## 4 14.25000 0.2221  0.9750     3
## 5 14.70000 0.0006  1.0000     3
## 6 13.01667 0.9308 -0.3654     1
```

```r
(moon3 <- aggregate(manilius[, 1:3], list(manilius$group), sum))
```

```
##   Group.1      arc  sinang  cosang
## 1       1 118.1333  8.4987 -0.7932
## 2       2 140.2833 -6.1404  1.7443
## 3       3 127.5333  2.9777  7.9649
```

```r
solve(cbind(9, moon3$sinang, moon3$cosang), moon3$arc)
```

```
## [1] 14.5445859 -1.4898221  0.1341264
```

```r
lmod <- lm(arc ~ sinang + cosang, manilius)
coef(lmod)
```

```
## (Intercept)      sinang      cosang
## 14.56162351 -1.50458123  0.09136504
```

```r
data(GaltonFamilies, package = "HistData")
plot(childHeight ~ midparentHeight, GaltonFamilies)
lmod <- lm(childHeight ~ midparentHeight, GaltonFamilies)
coef(lmod)
```

```
##     (Intercept) midparentHeight
##      22.6362405       0.6373609
```

```r
abline(lmod)
(beta <- with(GaltonFamilies, cor(midparentHeight, childHeight) * sd(childHeight)/sd(midparentHeight)))
```

```
## [1] 0.6373609
```

```r
(alpha <- with(GaltonFamilies, mean(childHeight) - beta * mean(midparentHeight)))
```

```
## [1] 22.63624
```

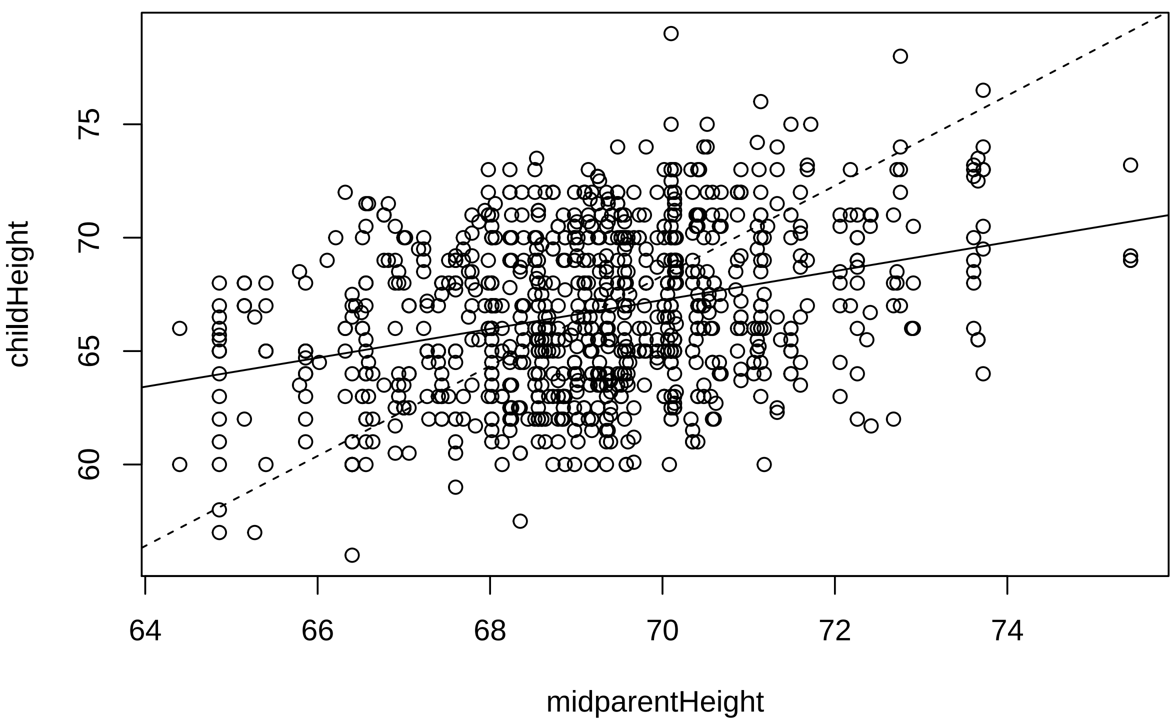
```r
(beta1 <- with(GaltonFamilies, sd(childHeight)/sd(midparentHeight)))
```

```
## [1] 1.985858
```

```r
(alpha1 <- with(GaltonFamilies, mean(childHeight) - beta1 * mean(midparentHeight)))
```

```
## [1] -70.68889
```

```r
abline(alpha1, beta1, lty = 2)
```



Homework Chapter 1

We're asked to make numerical and graphical summaries of a variety of datasets. We are instructed to limit the output to a quantity that abusy reader would find sufficient to get a basic understanding of the data.

- teengamb
- uswages
- prostate
- sat
- divusa

Study of teenage gambling in Britain

The teengamb data frame has 47 rows and 5 columns. A survey was conducted to study teenage gambling in Britain. This frame contains the following columns:

sex 0=male, 1=female

status Socioeconomic status score based on parents' occupation

income in pounds per week

verbal verbal score in words out of 12 correctly defined

gamble expenditure on gambling in pounds per year

```
data(teengamb, package = "faraway")

head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

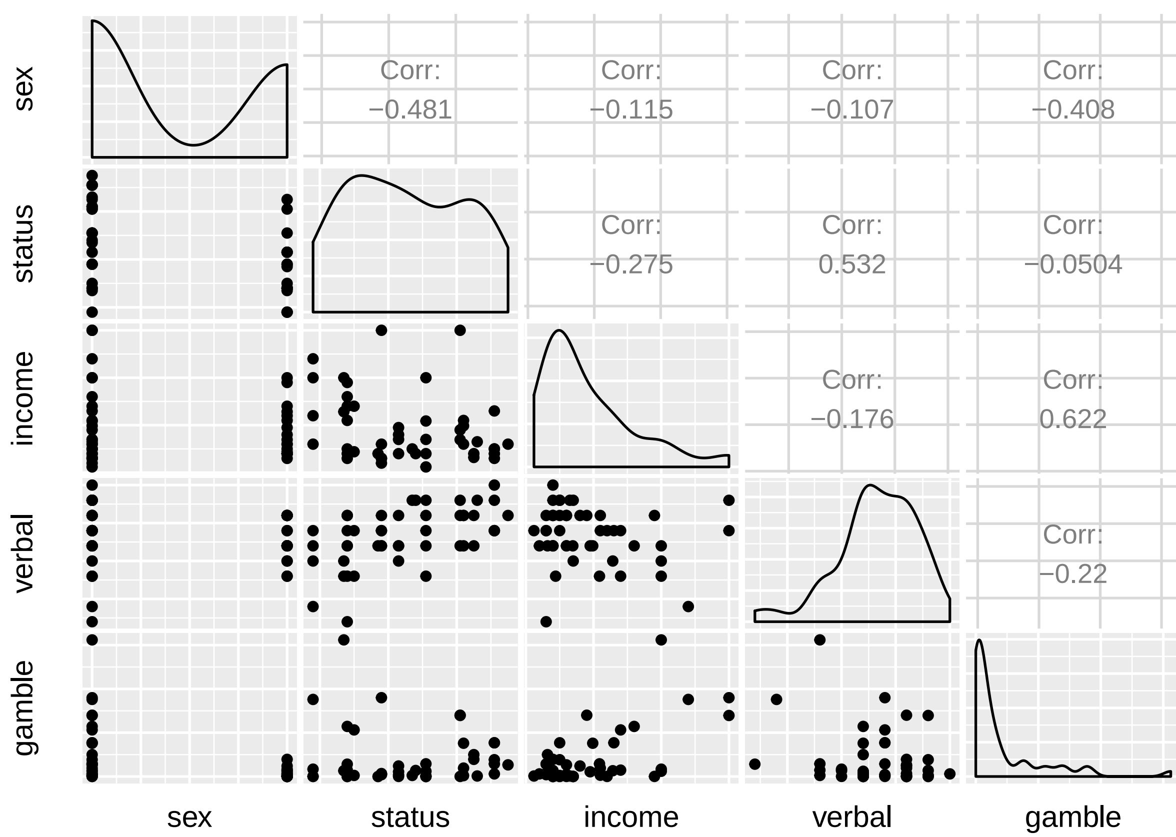
```
require(GGally)
```

```
## Loading required package: GGally

##
## Attaching package: 'GGally'

## The following object is masked from 'package:faraway':
##
##     happy
```

```
library(ggplot2)
require(GGally)
ggpairs(teengamb) + theme(axis.line = element_blank(), axis.text = element_blank(),
    axis.ticks = element_blank())
```



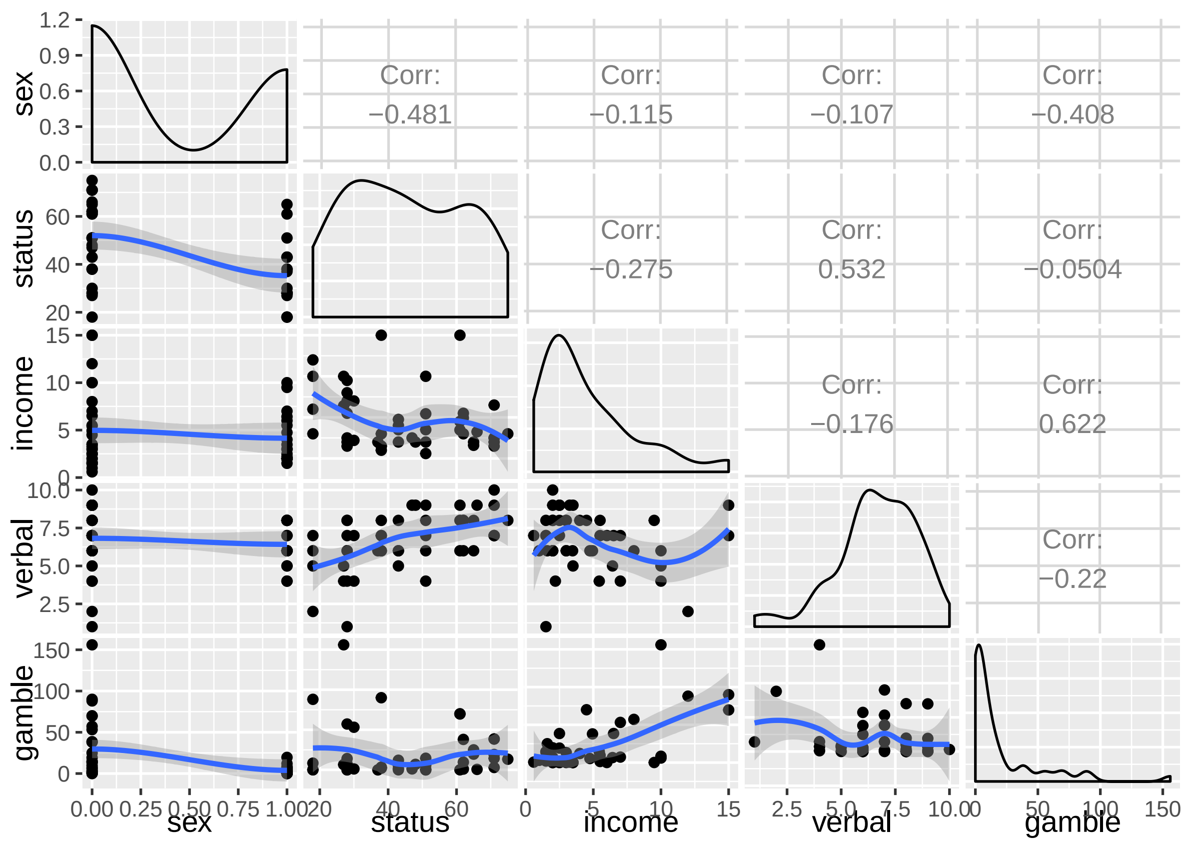
This data set is not well matched by sex so we'll be cautious in making inference on how sex influences gambling status. At first glance we may be tempted to note that gambling values are higher for males, but this may be due to variability in the population of gamblers.

We do note that there is an association between income and gamble. Gamble and Income appear to be right skewed fat tailed distributions.

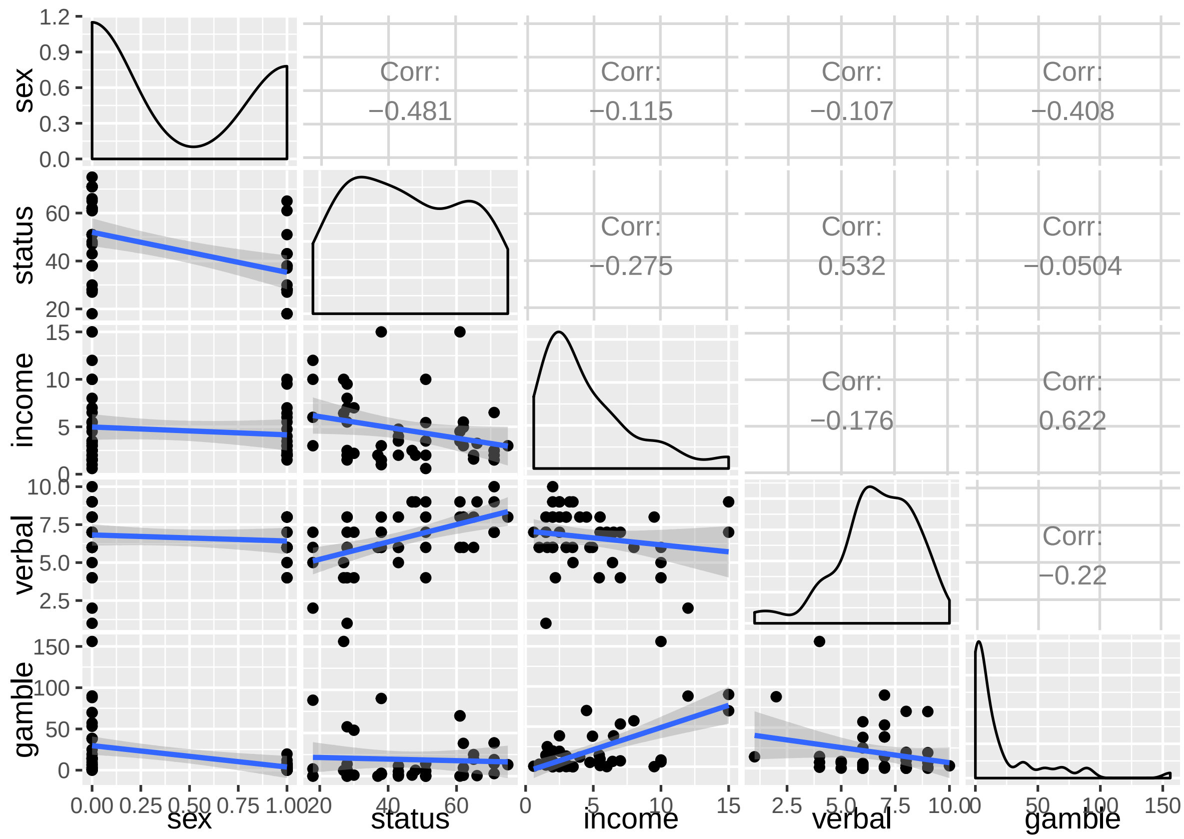
Here we add LOESS and LM models to the pairs plots. LOESS is fitting by local polynomial regression.

```
my_fn <- function(data, mapping, method = "loess", ...) {
    p <- ggplot(data = data, mapping = mapping) + geom_point() + geom_smooth(method = method,
        ...)
    p
}

# Default loess curve
ggpairs(teengamb, lower = list(continuous = my_fn))
```




```
# Use wrap to add further arguments; change method to lm
ggpairs(teengamb, lower = list(continuous = wrap(my_fn, method = "lm")))
```


sex
status
income
verbal
gamble
Corr:
-0.481
Corr:
-0.115
Corr:
-0.107
Corr:
-0.408
Corr:
-0.275
Corr:
0.532
Corr:
-0.0504
Corr:
-0.176
Corr:
0.622
Corr:
-0.22