Exercise 0

AUTHOR

Collin Real (yhi267)

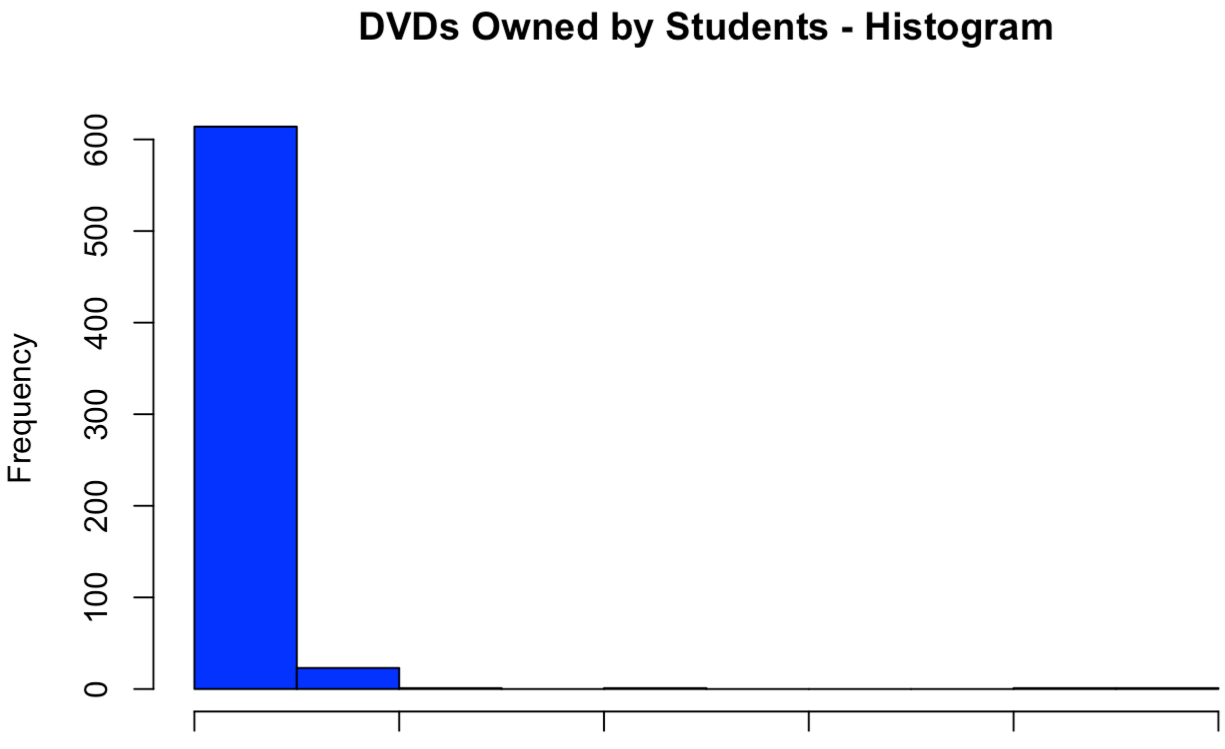
Import package, access the studentdata from package, and show part of the data.

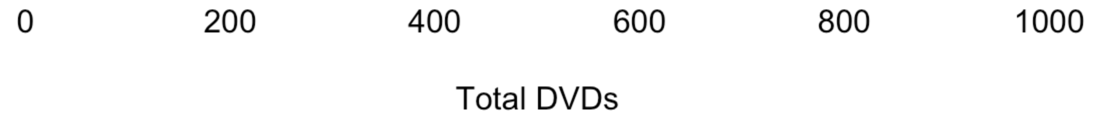
```r
library(LearnBayes)
data(studentdata)
attach(studentdata)
head(studentdata)
```

```
  Student Height Gender Shoes Number Dvds ToSleep WakeUp Haircut  Job Drink
1       1     67 female    10      5   10    -2.5    5.5      60 30.0 water
2       2     64 female    20      7    5     1.5    8.0       0 20.0   pop
3       3     61 female    12      2    6    -1.5    7.5      48  0.0  milk
4       4     61 female     3      6   40     2.0    8.5      10  0.0 water
5       5     70   male     4      5    6     0.0    9.0      15 17.5   pop
6       6     63 female    NA      3    5     1.0    8.5      25  0.0 water
```

1a) Construct a histogram of this variable using the hist command in R.

```r
hist(studentdata$Dvds,
     main="DVDs Owned by Students - Histogram",
     xlab="Total DVDs",
     ylab="Frequency",
     col="blue",
     border="black")
```





1b) Summarize this variable using the summary command in R.

```
summary(studentdata$Dvds)
```

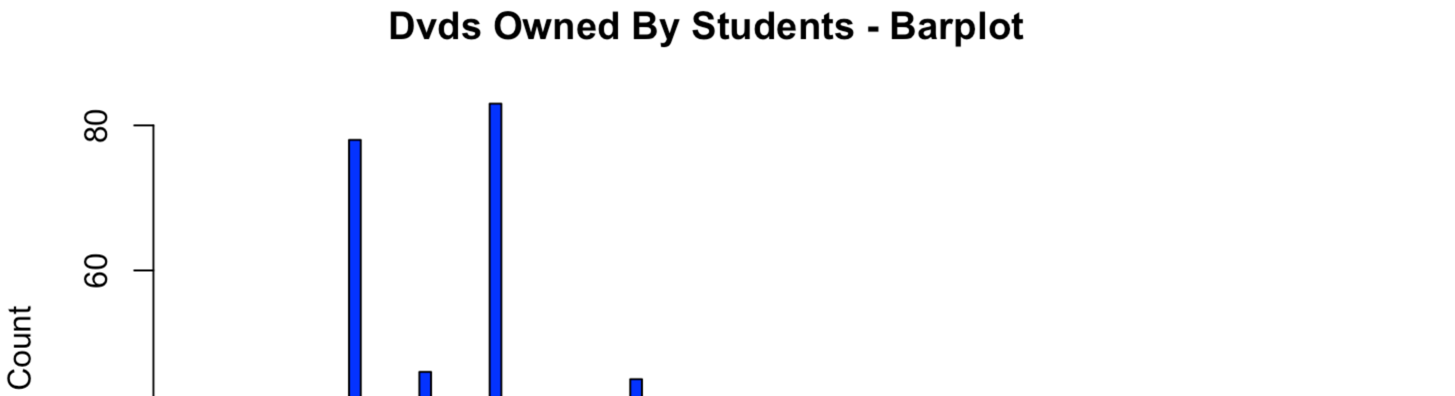
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   10.00   20.00   30.93   30.00 1000.00      16 
```

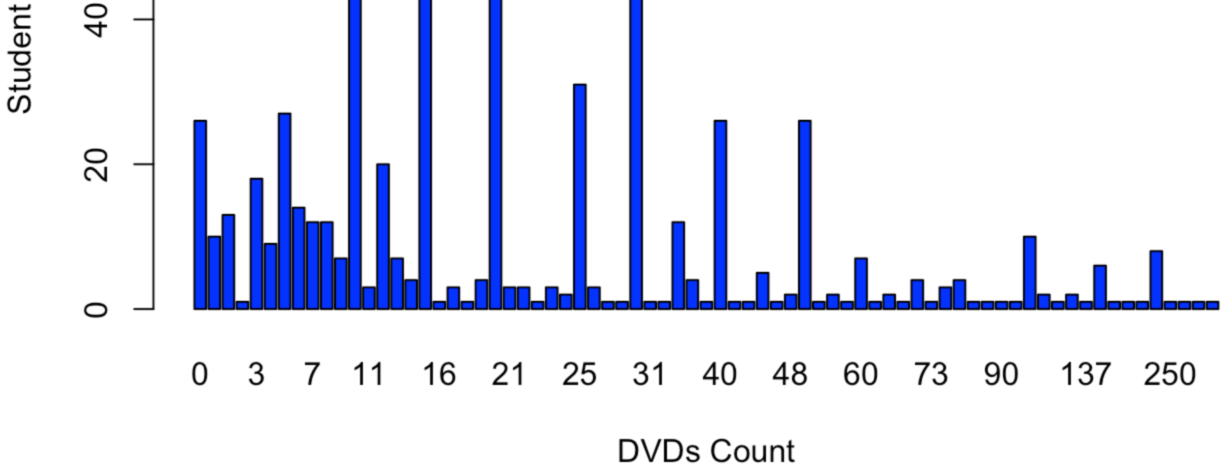
1c) Use the table command in R to construct a frequency table of the individual values of Dvds that were observed. If one constructs a barplot of these tabled values using the command `barplot(table(Dvds),col='red')` one will see that particular response values are very popular. Is there any explanation for these popular values for the number of DVDs owned?

```
table(studentdata$Dvds)
```

```
    0     1     2   2.5     3     4     5     6     7     8     9    10    11    12    13    14 
   26    10    13     1    18     9    27    14    12    12     7    78     3    20     7     4 
   15    16    17  17.5    18    20    21    22  22.5    23    24    25  27.5    28    29    30 
   46     1     3     1     4    83     3     3     1     3     2    31     3     1     1    45 
   31    33    35    36    37    40    41    42    45    46    48    50    52    53    55    60 
    1     1    12     4     1    26     1     1     5     1     2    26     1     2     1     7 
   62    65    67    70    73    75    80    83    85    90    97   100   120   122   130   137 
    1     2     1     4     1     3     4     1     1     1     1    10     2     1     2     1 
  150   152   157   175   200   250   500   900  1000 
    6     1     1     1     8     1     1     1     1 
```

```
barplot(table(studentdata$Dvds),
        main = 'Dvds Owned By Students - Barplot',
        xlab ='DVDs Count',
        ylab = 'Student Count',
        col = 'blue',
        border = 'black')
```

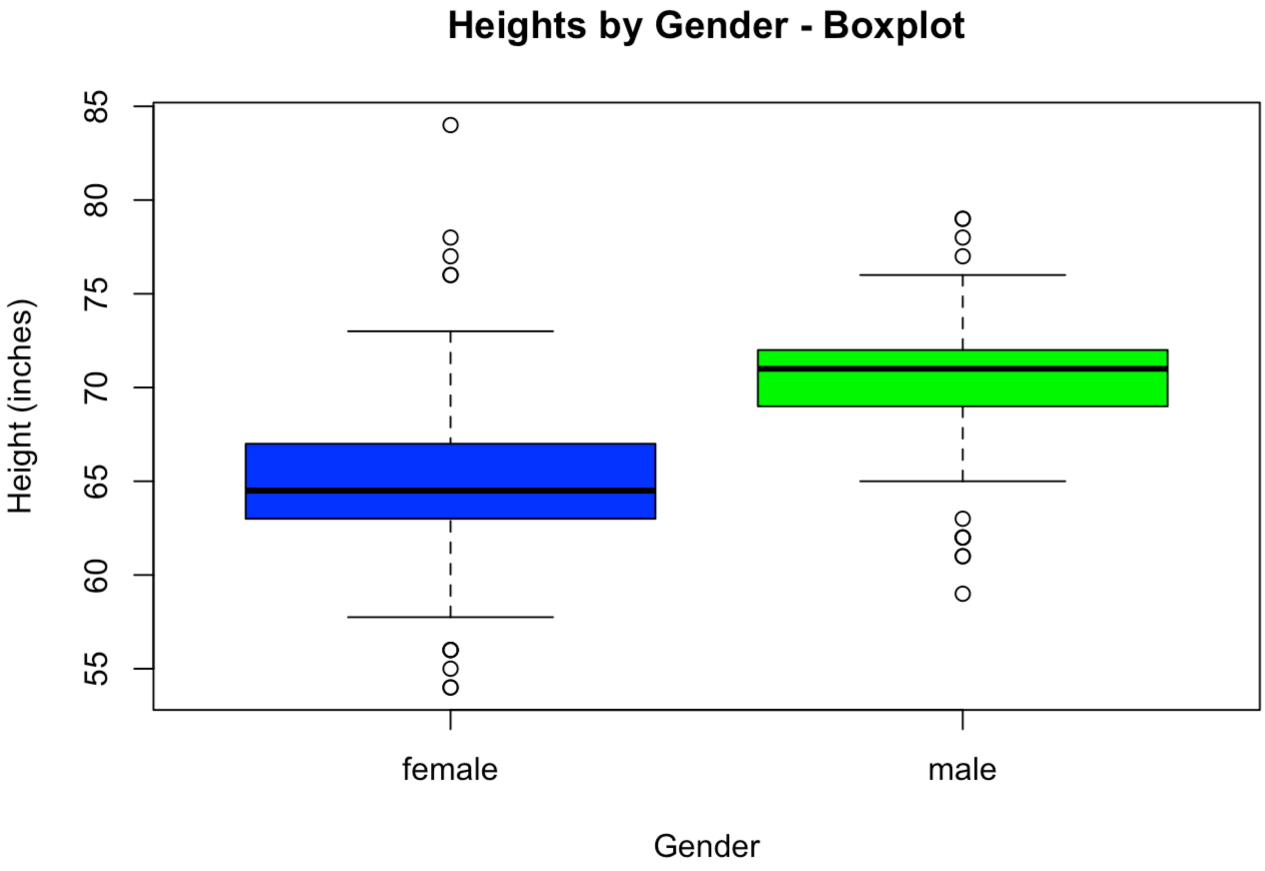




1c) Explanation: There are significant spikes in student count at intervals of 5 & 10, indicating that many students in the survey might not have inputted their exact DVD counts, but rather an estimate.

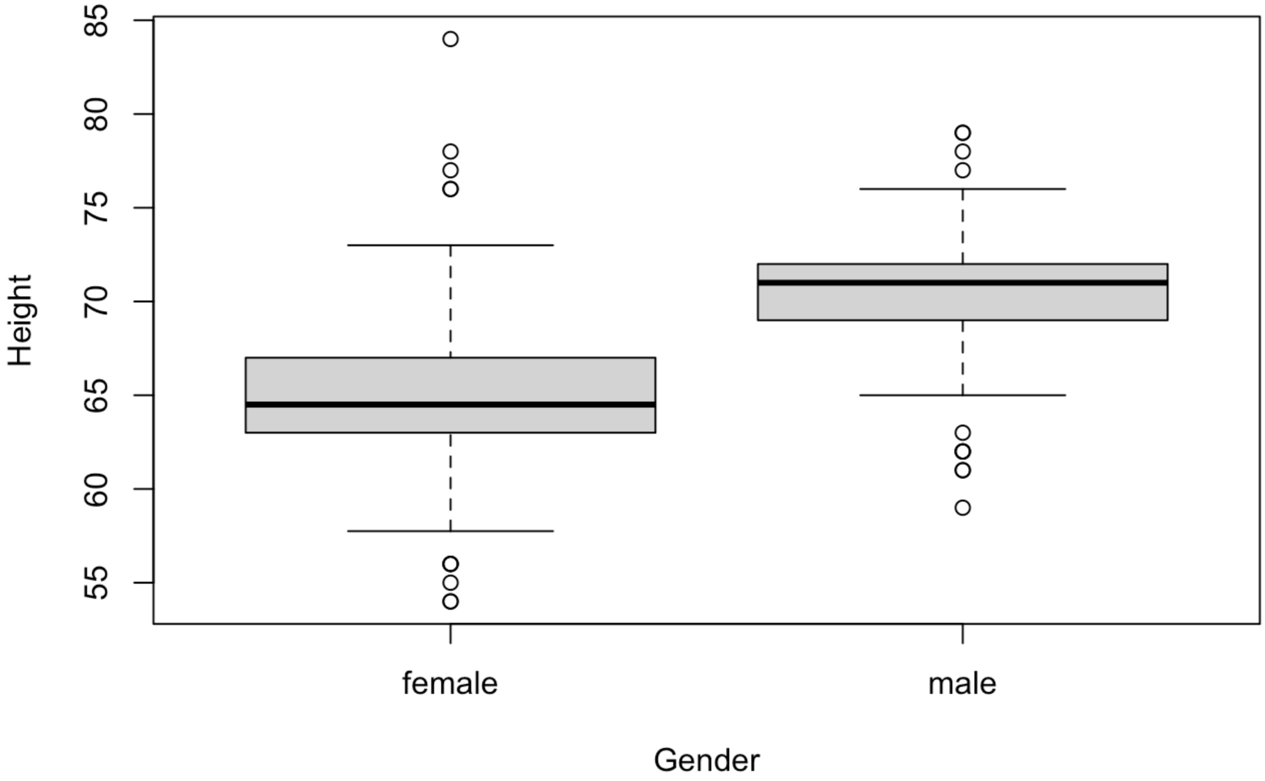
Question 2a) Construct parallel boxplots of the heights using the Gender variable.

```
boxplot(Height ~ Gender,
        data = studentdata,
        main = "Heights by Gender - Boxplot",
        xlab = "Gender",
        ylab = "Height (inches)",
        col = c("blue", "green"))
```



Question 2b) If one assigns the boxplot output to a variable `output=boxplot(Height~Gender)` then output is a list that contains statistics used in constructing the boxplots. Print output to see the statistics that are stored.

```
output = boxplot(Height ~ Gender)
```



```
print(output)
```

```
$stats
      [,1] [,2]
[1,] 57.75   65
[2,] 63.00   69
[3,] 64.50   71
[4,] 67.00   72
[5,] 73.00   76

$n
[1] 428 219

$conf
         [,1]    [,2]
[1,] 64.19451 70.6797
[2,] 64.80549 71.3203

$out
```

```
 [1] 56 76 55 56 76 54 54 84 78 77 56 63 77 79 62 62 61 79 59 61 78 62

$group
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2

$names
[1] "female" "male"
```

Question 2c) On average, how much taller are male students than female students?

```
avg_male_height = mean(Height[Gender == "male"],
                       na.rm = TRUE)
avg_female_height = mean(Height[Gender == "female"],
                         na.rm = TRUE)
height_diff = avg_male_height - avg_female_height

print(paste0("Male height: ",
             round(avg_male_height,
                   digits = 2),
             " inches."))
```

```
[1] "Male height: 70.51 inches."
```

```
print(paste0("Female height: ",
             round(avg_female_height,
                   digits = 2),
             " inches."))
```

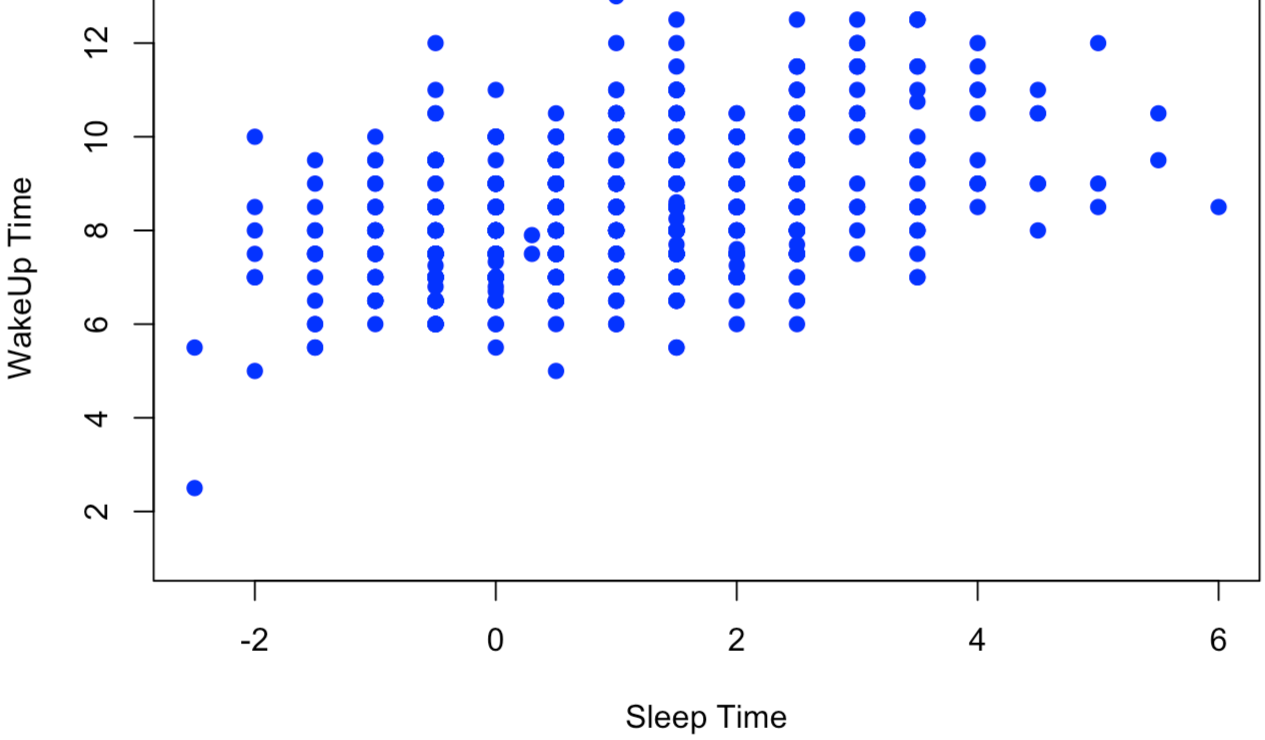
```
[1] "Female height: 64.76 inches."
```

```
print(paste0("On average, male students are ",
             round(height_diff,
                   digits = 2),
             " inches taller than females."))
```

```
[1] "On average, male students are 5.75 inches taller than females."
```

Question 3a) Construct a scatterplot of ToSleep and WakeUp.

```
plot(ToSleep, WakeUp,
     xlab = "Sleep Time",
     ylab = "WakeUp Time",
     main = "ToSleep and WakeUp - Scatterplot",
     pch = 19,
     col = "blue")
```



Question 3b) Find a least-squares fit to these data using the lm command and then place the least-squares fit on the scatterplot using the abline command.

```
# Plot ToSleep and WakeUp on a scatterplot
plot(ToSleep, WakeUp,
     xlab = "Sleep Time",
     ylab = "WakeUp Time",
     main = "ToSleep and WakeUp - Scatterplot",
     pch = 19,
     col = "blue")

# Find a least-squares fit
fit = lm(WakeUp ~ ToSleep)

# Display the summary
summary(fit)
```

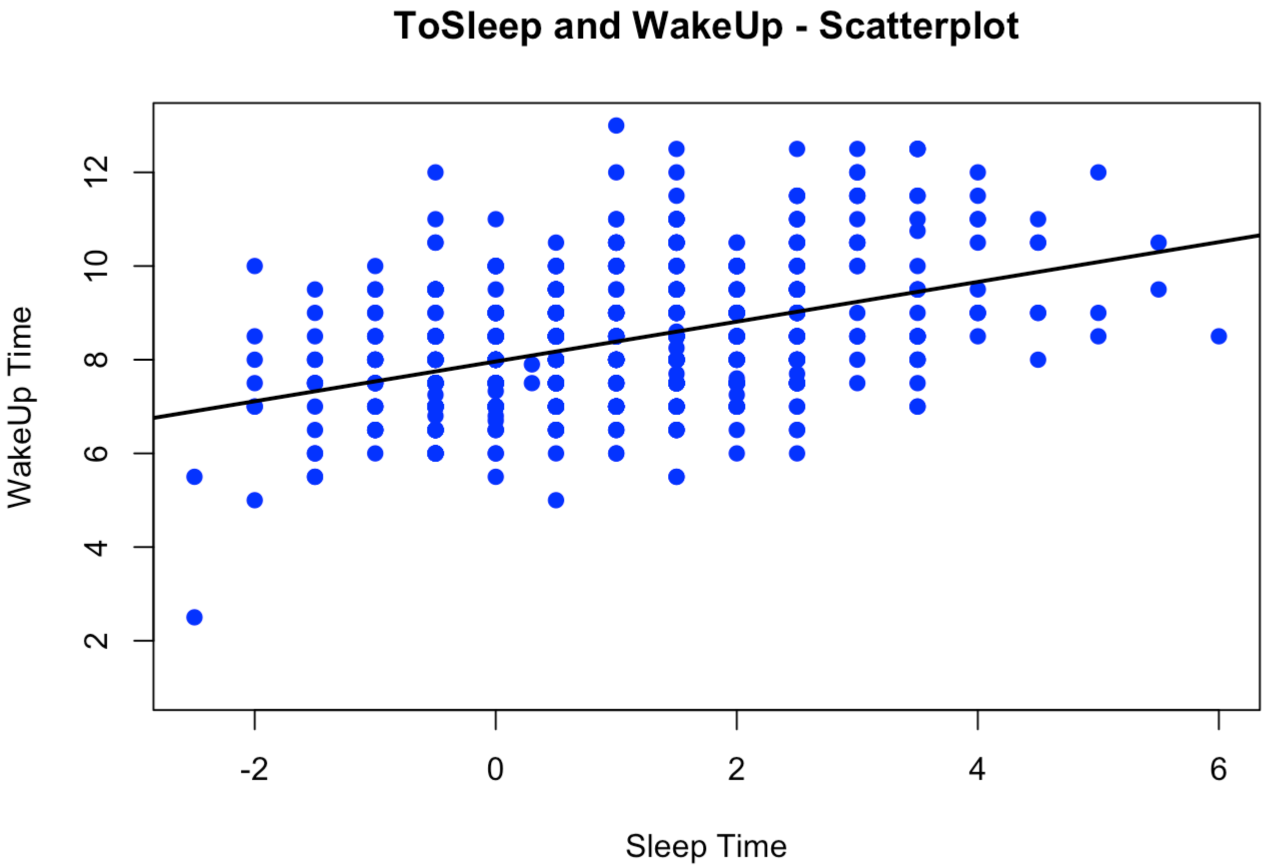
```
Call:
lm(formula = WakeUp ~ ToSleep)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4010 -0.9628 -0.0998  0.8249  4.6125 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.96276    0.06180  128.85   <2e-16 ***
ToSleep      0.42472    0.03595   11.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.282 on 651 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.1765,	Adjusted R-squared:  0.1753 
F-statistic: 139.5 on 1 and 651 DF,  p-value: < 2.2e-16
```

```r
# Add the least-squares fit line to the scatterplot
abline(fit, col = "black", lwd = 2)
```


ToSleep and WakeUp - Scatterplot
WakeUp Time
Sleep Time