Accelerated Stats Chapter 2 Problem Set

Tianwei Liu

9/9/2019

First of all, we load the .Rdata file into the current project Load Data File

```
load("Ch2_Exercise2_Olympics.Rdata")
```

2. (a) Summarize the medals, athletes and GDP data.

```
summary(dta$medals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.751   0.000  37.000
```

```
summary(dta$athletes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    2.00   18.17   13.00  230.00
```

```
summary(dta$GDP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0110  0.1212  0.3849  1.1691  1.5127 14.5230     156
```

(b) List the first five observations for the country, year, medals, athletes, and GDP data

```
head (dta$country, n=5)
```

```
## [1] "Albania" "Albania" "Albania" "Albania" "Albania"
```

```
head (dta$year, n=5)
```

```
## [1] 1980 1984 1988 1992 1994
```

```
head (dta$medals, n=5)
```

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

```
head (dta$athletes, n=5)
```

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

```
head (dta$GDP, n=5)
```

```
## [1]     NA 0.0641 0.0637 0.0206 0.0587
```

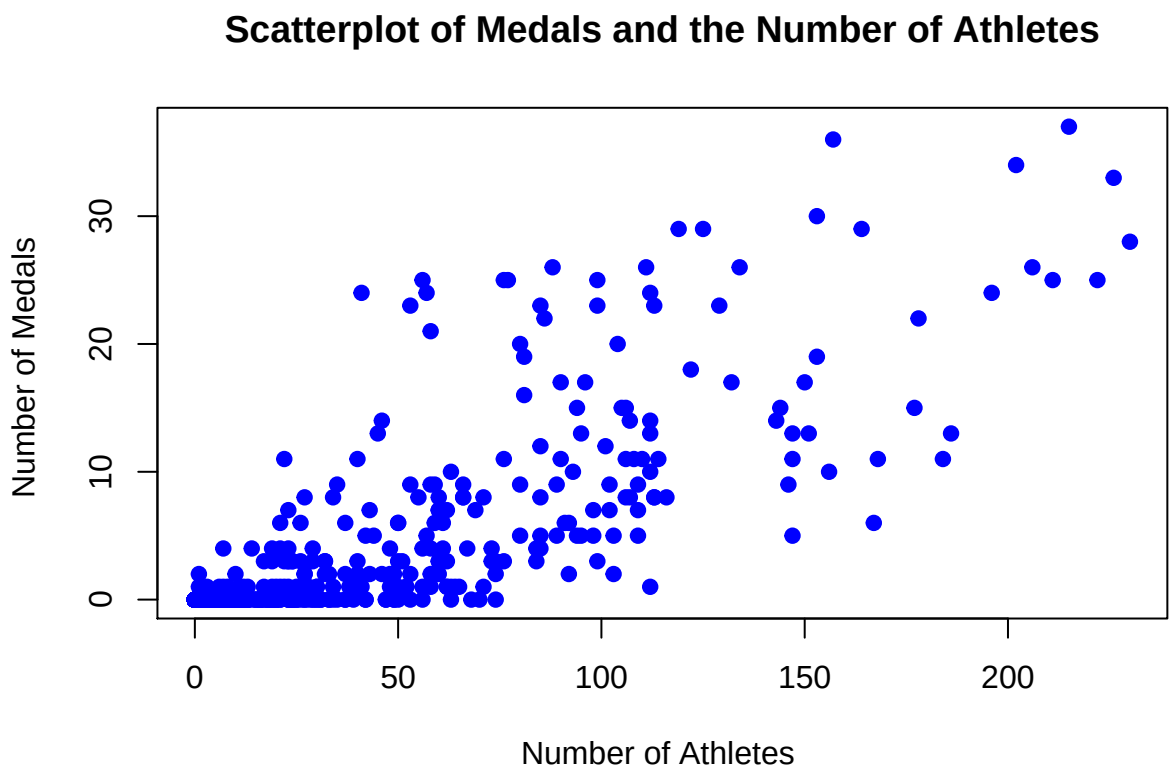
(c) How many observations are there for each year?

```
table (dta$year)
```

```
##
## 1980 1984 1988 1992 1994 1998 2002 2006 2010 2014
##  117  117  117  113  110  110  110  110  109  109
```

(d) Produce a scatterplot of medals and the number of athletes. Describe the relationship depicted.

```
plot(dta$athletes, dta$medals, main="Scatterplot of Medals and the Number of Athletes",
   xlab="Number of Athletes ", ylab="Number of Medals ", pch=19, col = c("blue"))
```



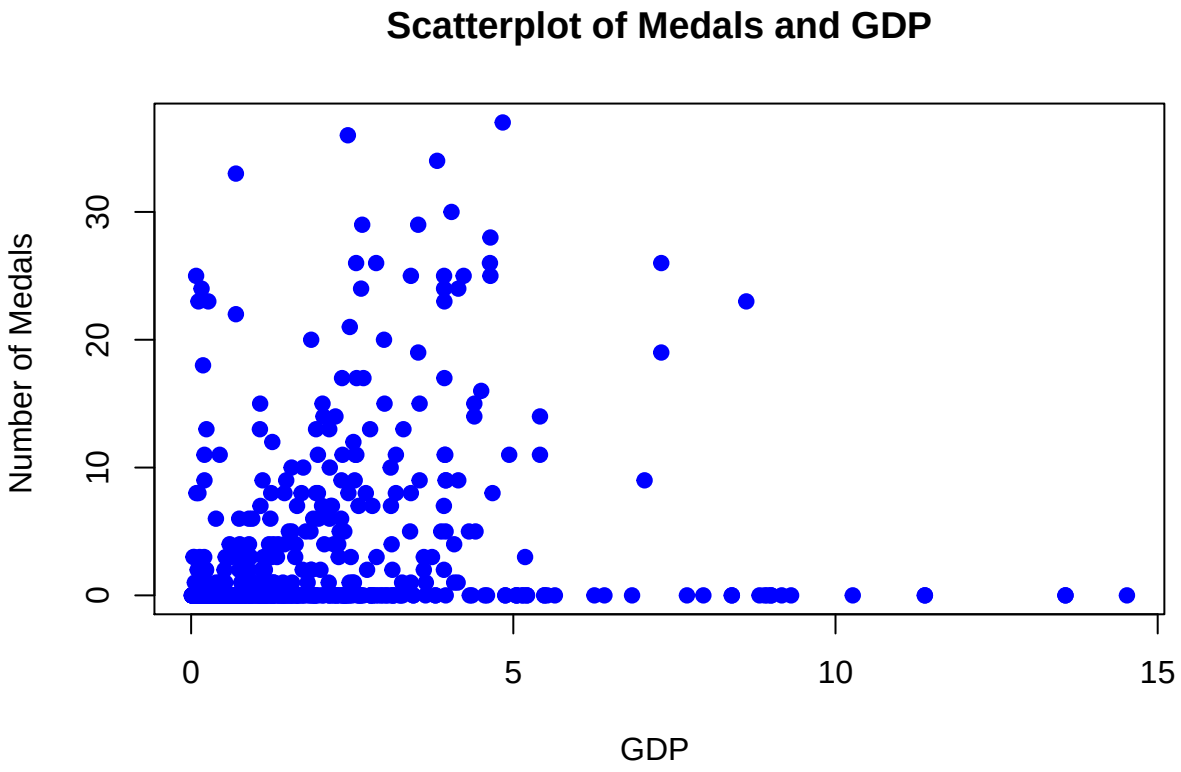
As we can see from the scatterplot above, as the number of athletes increases, the number of medals won also increases.

(e) Explain any suspicion you might have that other factors could explain the observed relationship between the number of athletes and medals.

Let's think of the core model where Y is the number of medals and X is the number of athletes. There might be endogeneity embedded in the independent variable. For example, a richer country is able to recruit more athletes and invests more heavily into the training; so it is more likely to have a larger group of athletes and these athletes win more medals.

(f) Create a scatterplot of medals and GDP. Briefly describe any clear patterns.

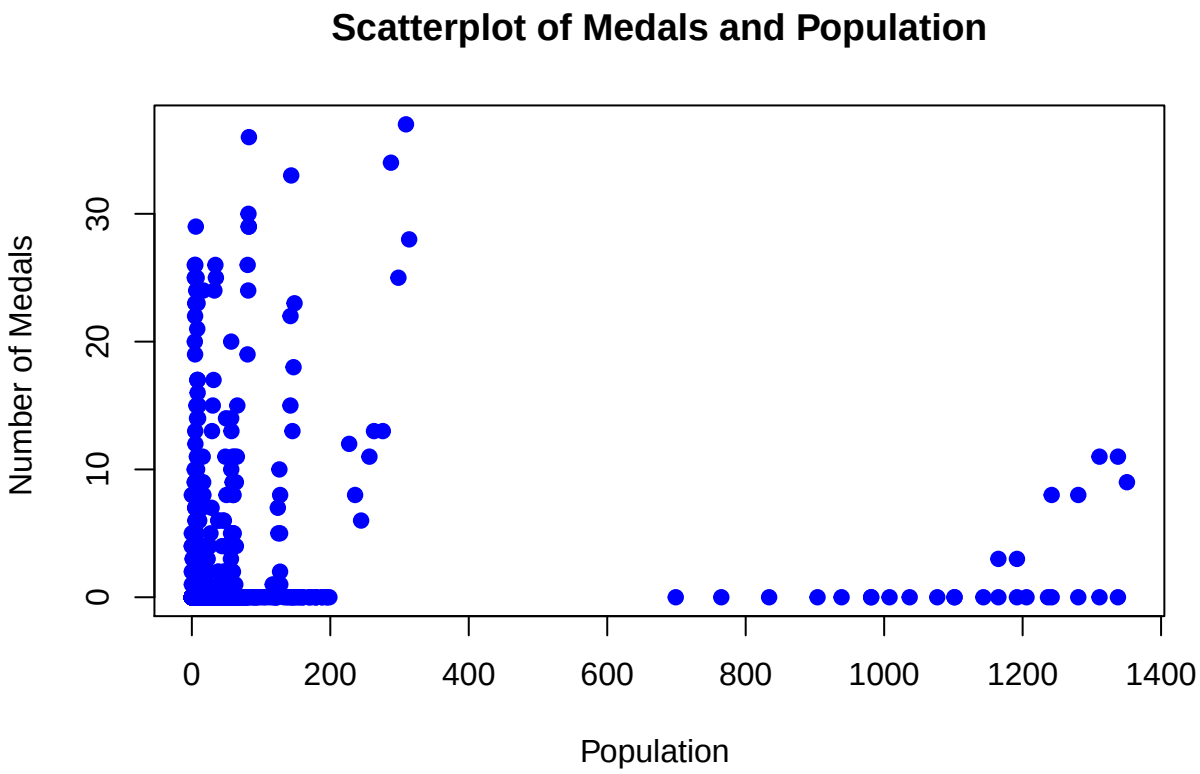
```
plot(dta$GDP, dta$medals, main="Scatterplot of Medals and GDP",
   xlab="GDP ", ylab="Number of Medals ", pch=19, col = c("blue"))
```



Pattern: the higher the country's GDP is, the more models it is likely to win. However, there are also many countries with a relatively high GDP but have not won any medals.

(g) Create a scatterplot of medals and population. Briefly describe any clear patterns.

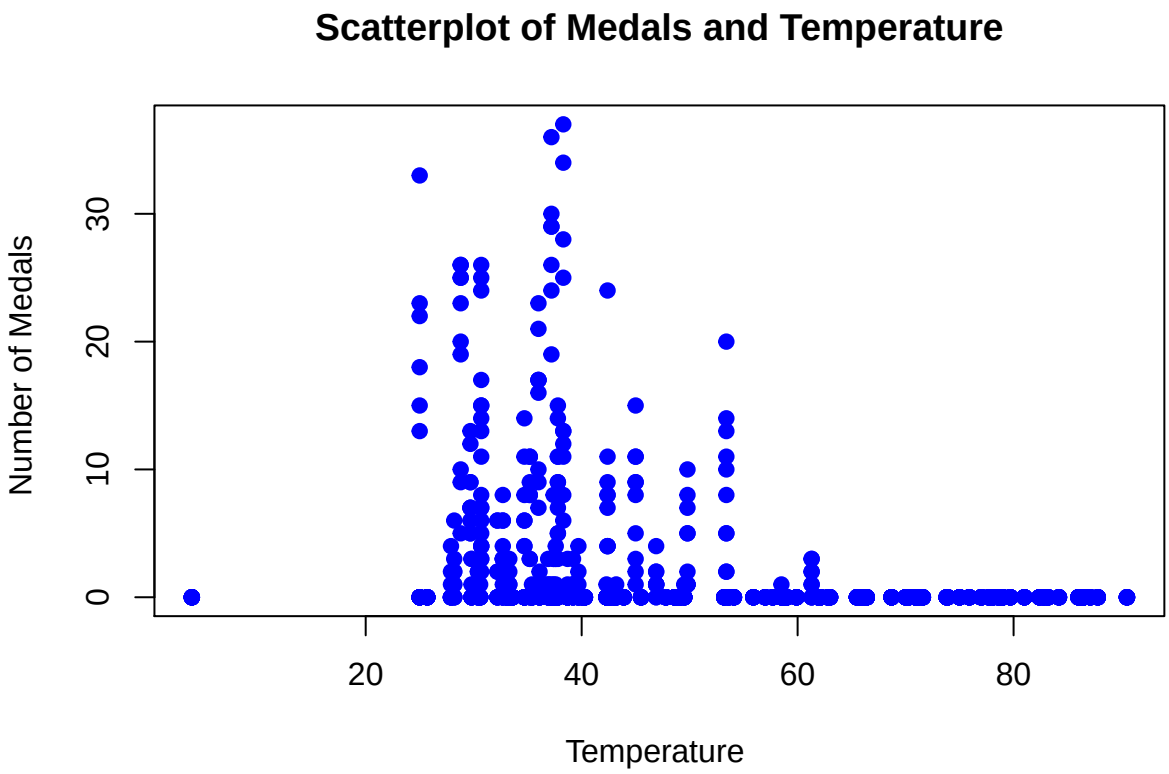
```
plot(dta$population, dta$medals, main="Scatterplot of Medals and Population",
   xlab="Population ", ylab="Number of Medals ", pch=19, col = c("blue"))
```



Pattern: countries with a small population (0-200, unit 100,000) spread over the number of medals the win; larger countries, espeically with population over about 700, win few medals or even 0 medal.

(h) Create a catterplot of medals and temperature. Briefly describe any clear patterns.

```
plot(dta$temp, dta$medals, main="Scatterplot of Medals and Temperature",
   xlab="Temperature ", ylab="Number of Medals ", pch=19, col = c("blue"))
```



Pattern: countries with a lower average high temperature (in a winter month) tend to win more medals than countries having a higher average high temperature in a winter month.

3. (a) Summarize the wage, height (both height85 and height81), and sibling variables. Discuss briefly.

```
load("Ch2_Exercise3_Height_and_Wages_US.Rdata")
summary (dta$wage96)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000    6.743   10.783   14.177   16.213 1533.333     5756
```

```
summary (dta$height81)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   48.00   64.00   67.00   67.01   70.00   83.00     543
```

```
summary (dta$height85)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   48.00   64.00   67.00   67.08   70.00   81.00    1823
```

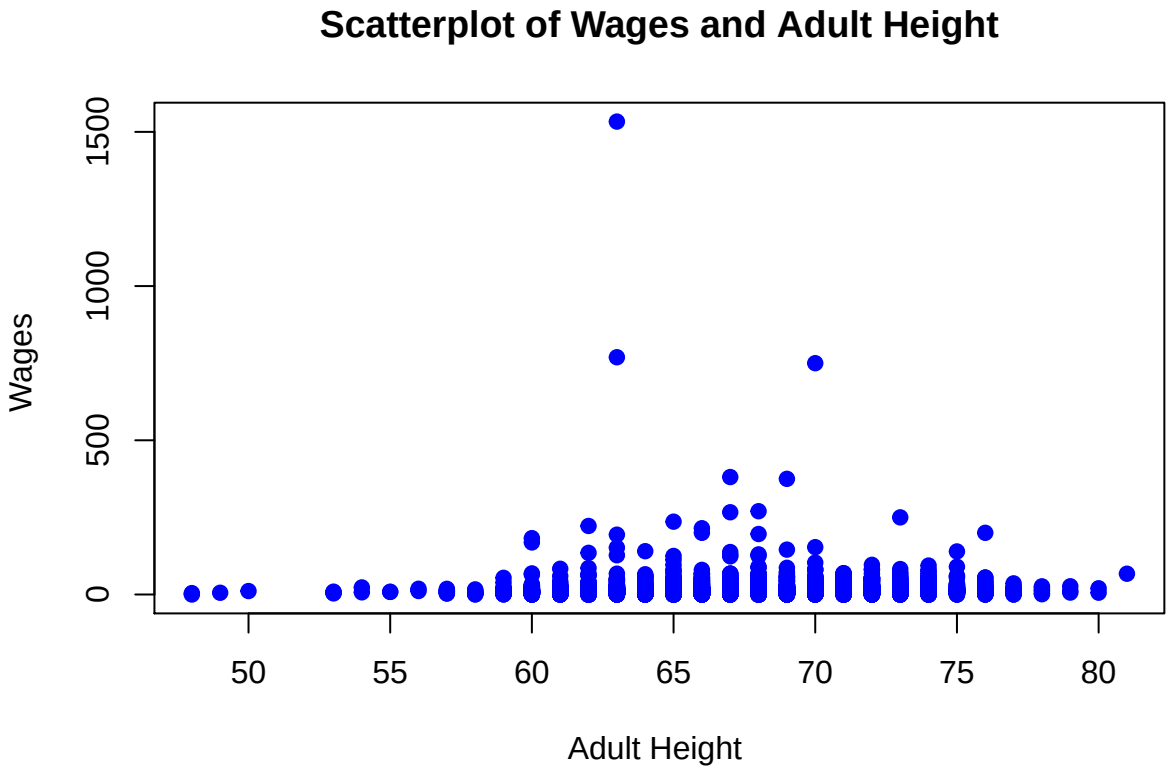
```
summary (dta$siblings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.000   2.000   3.000   3.844   5.000  29.000
```

Brief discussion: wage has a unbelievably high maximum, so we should look into the extreme values of wages; adolescent height and adult height are pretty similar, despite minor difference in the maximums; the min of number of siblings is negative, which is impossible, so there are errors in the siblings variable.

(b) Create a scatterplot of wages and adult height (height85). Discuss any distinctive observations.

```
plot(dta$height85, dta$wage96, main="Scatterplot of Wages and Adult Height",
   xlab="Adult Height ", ylab="Wages ", pch=19, col = c("blue"))
```



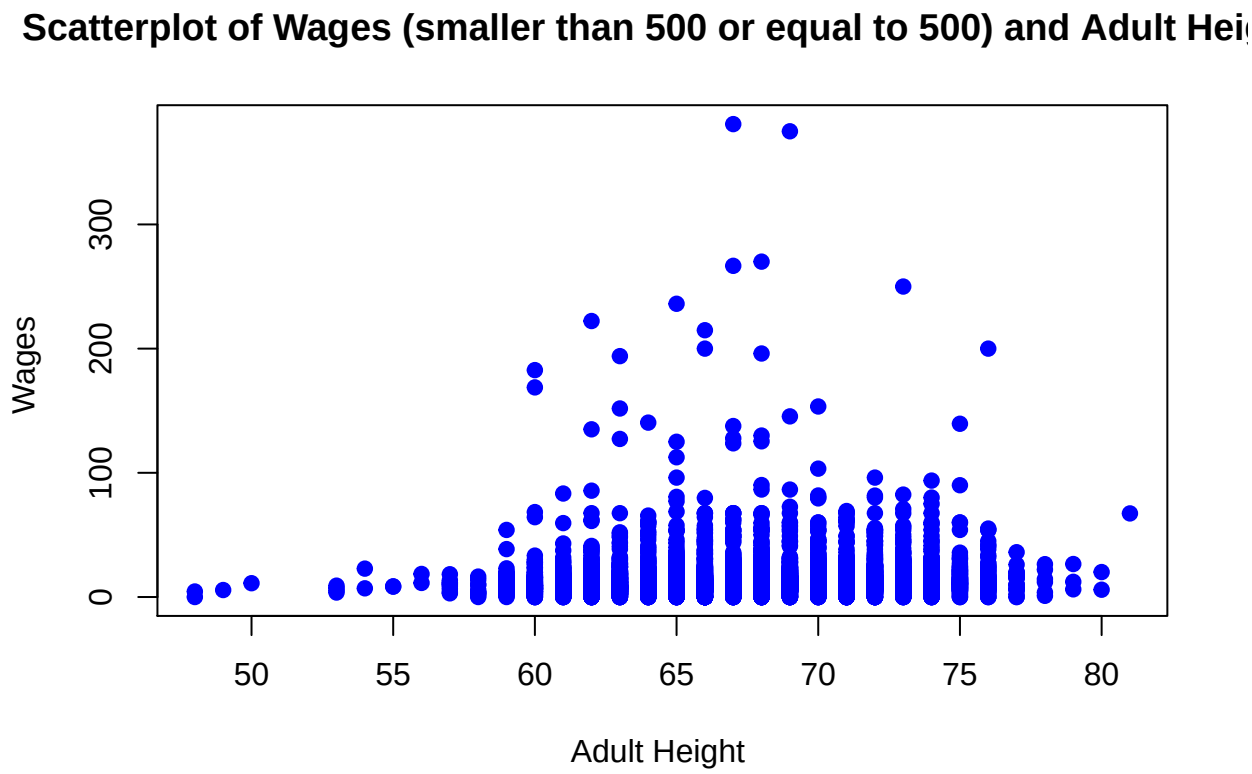
Pattern: adults with a height between 60-70 are more likely to gain a higher wage than other height groups. Note that there are three distinctive observations whose wages are significantly above 500. The existence of these three observations significantly enlarged the range of the Y-axis and thereby made variations among the 0-500 wage group insignificant. Therefore, we should remove the three observations and focus on observations with wages under 500 dollars per hour.

(c) Create a scatterplot of wages and adult height that excludes the observations with wages above 500 Dollars per hour.

```
dta$wage96[dta$wage96>=500] <- NA
summary(dta$wage96)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   6.736  10.772  13.743  16.186 380.769    5759
```

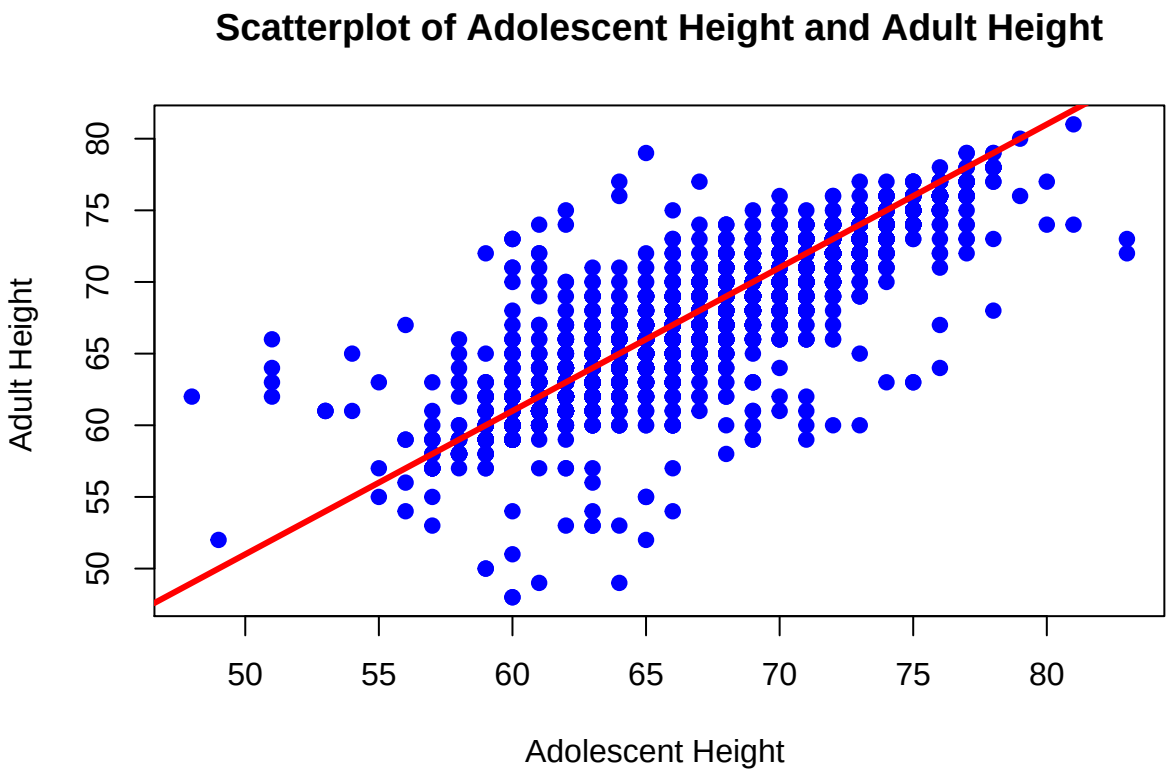
```
plot(dta$height85, dta$wage96, main="Scatterplot of Wages (smaller than 500 or equal to 500) and Adult H
```



Pattern: This looks like a normal distribution to me.

(d) Create a scatterplot of adult height against adolescent height. Identify the set of observations where people's adolescent height is more than their adult height. Do you think we should use these observations in any future analysis we conduct with this data? why or why not?

```
plot(dta$height81, dta$height85, main="Scatterplot of Adolescent Height and Adult Height", xlab="Adoles
abline (1,1, col= "red", lwd =3)
```



The set of people whose Adolescent Height is more than their Adult Height is under the red line (y=x). I don't think we should use this dataset for future analysis because it does not really make sense that half of the sample are shorter when they become adults.