Allen Church Chapter 2 - Exercise 3 Accelerated Statistics for Public Policy 9/16/19

```r
load("/Users/allenchurch/Ch2_Exercise3_Height_and_Wages_US.RData")
ex3 <- dta
```

3a. Summarize wage, height (both height85 and height81)

```r
summary(ex3$wage96)
```

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

```r
summary(ex3$height85)
```

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

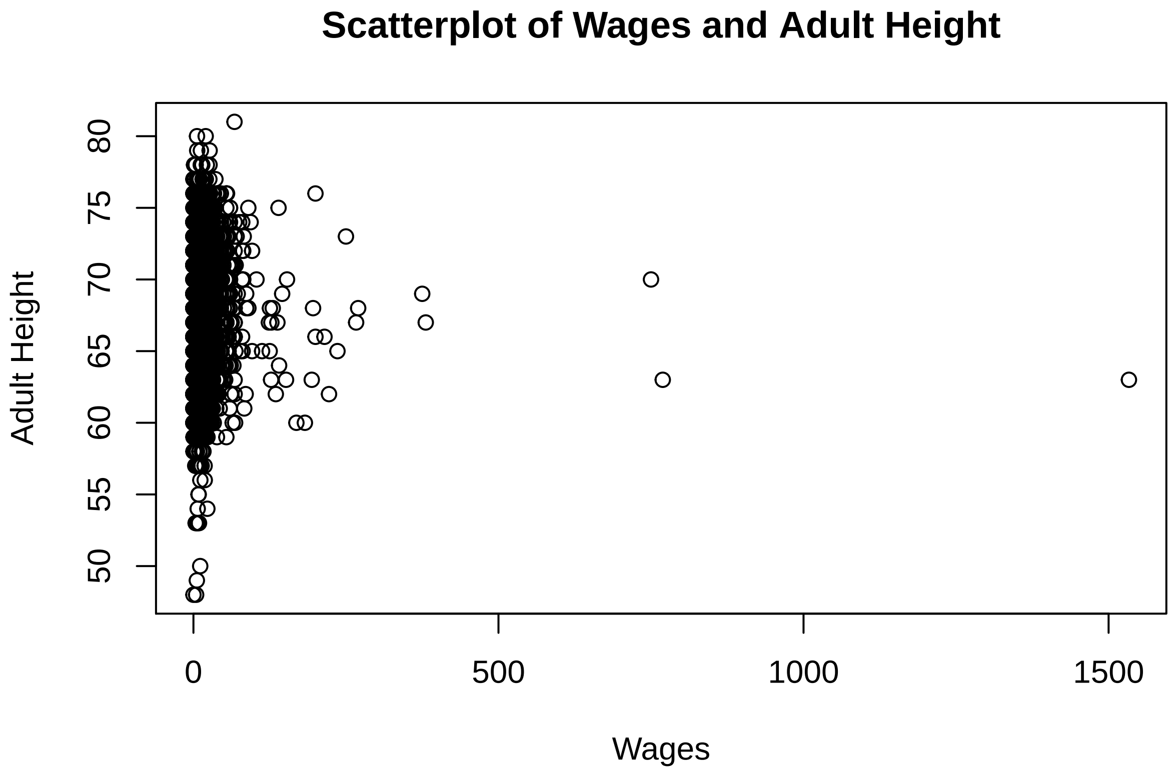
```r
summary(ex3$height81)
```

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

3b. Create scatterplot of wages and adult height (height85). Discuss any distinctive observations.

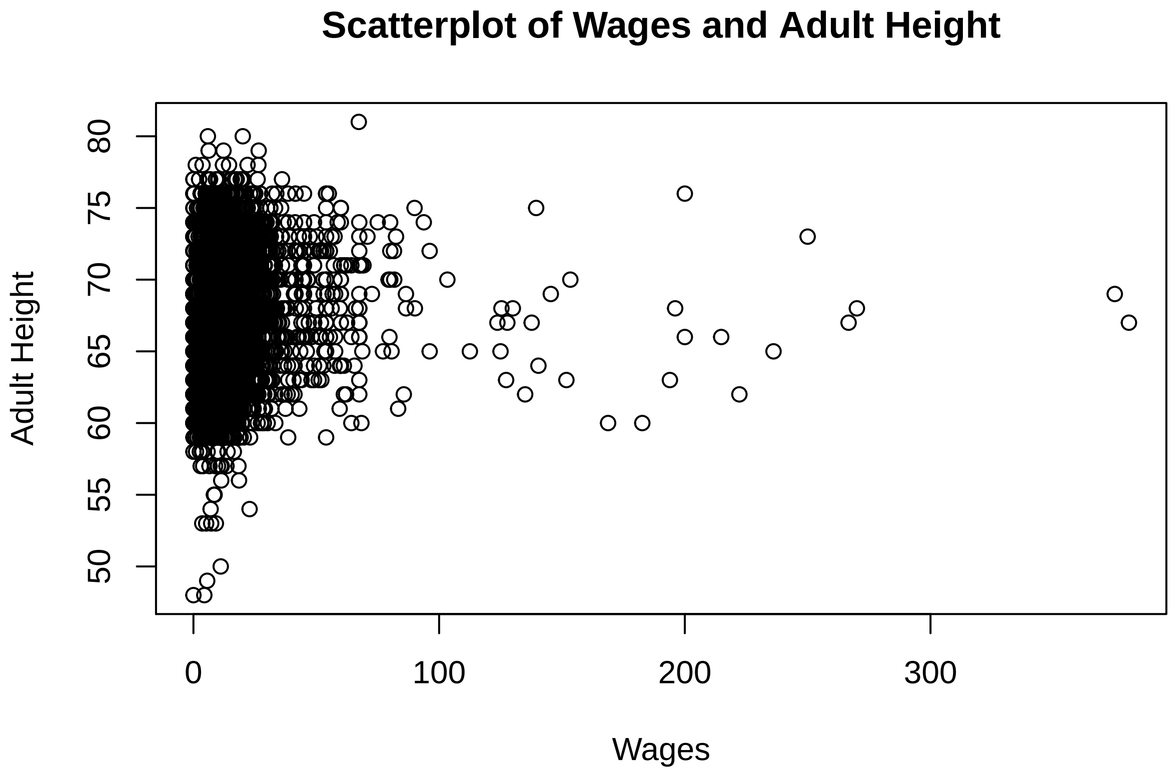
The scatterplot below shows many 0 values - potentially null or error values - for wage. There are also 3 significant outliers with large wage values, which could indicate an error in these observations as well. The scatterplot shows a positive correlation between higher adult height and higher wages.

```r
plot(ex3$wage96, ex3$height85, main="Scatterplot of Wages and Adult Height",
     xlab="Wages", ylab = "Adult Height")
```



3c. Create scatterplot of wages and adult height that exclues observations of wages above $500 per hour.

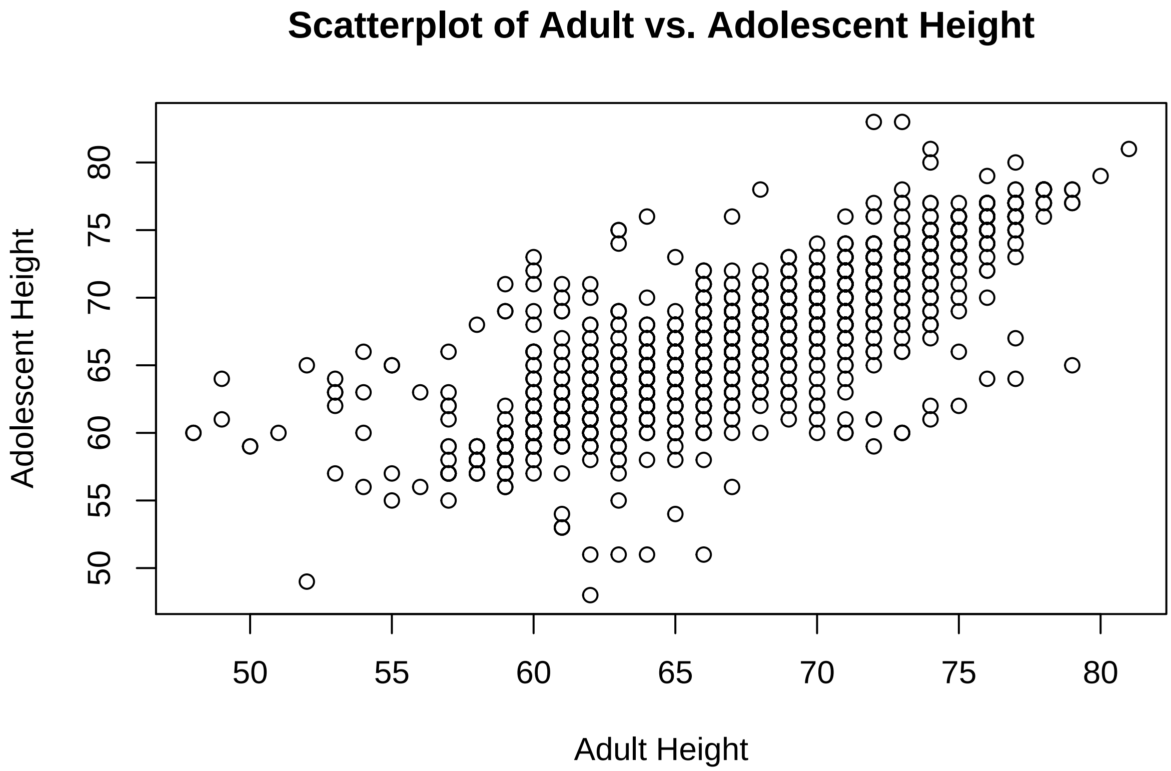
```r
#Subset dataset and specify wage96 column to be above 500
ex4 <- subset(ex3, wage96<500)

plot(ex4$wage96, ex4$height85, main="Scatterplot of Wages and Adult Height",
     xlab="Wages", ylab = "Adult Height")
```



3d. Create 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? Why or why not?

The table below shows that 1805 out of 12,686 observations have higher adolescent height than adult height. I believe we should exclude these observations in future analyses, as these data points contribute to the clutter of the graph. Also, we should drop the incorrect observations as it will help our findings to be internally valid.

```r
plot(ex3$height85, ex3$height81, main="Scatterplot of Adult vs. Adolescent Height",
     xlab="Adult Height", ylab = "Adolescent Height")
```



```
#Create new column 'shrink' that will return TRUE if adolescent height (height81)
#is greater than adult height (height85)
ex3$shrink <- ex3$height81 > ex3$height85

#The table below shows that there are 1805 out of 12,686 observations that have a higher adolescent hei
table(ex3$shrink)
```

```
##
## FALSE  TRUE
##  8776  1805
```

```
#The which function below will return the row IDs for observations that
#have a higher adolescent height than adult height.

#To save paper, I have commented it out and excluded the output
#which(ex3$shrink==TRUE)
```