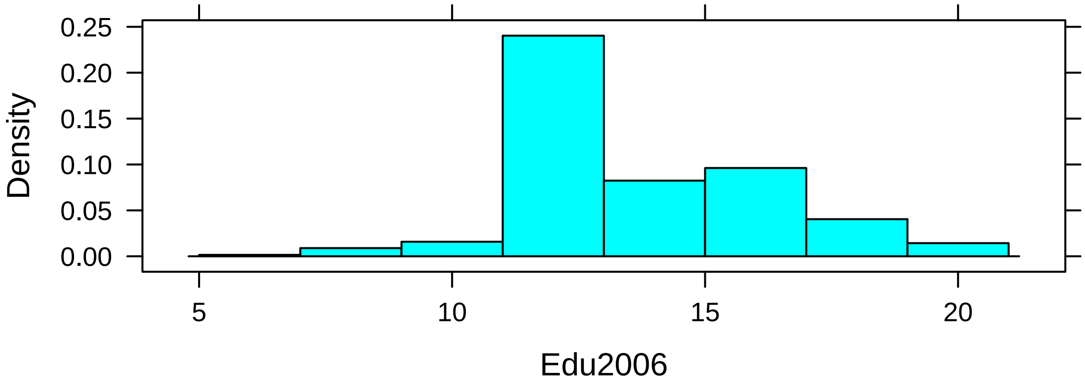
STAT 220: Homework 2

Zixi Li

(a)

Here I plot a histogram for the variable Edu2006 in density scale.

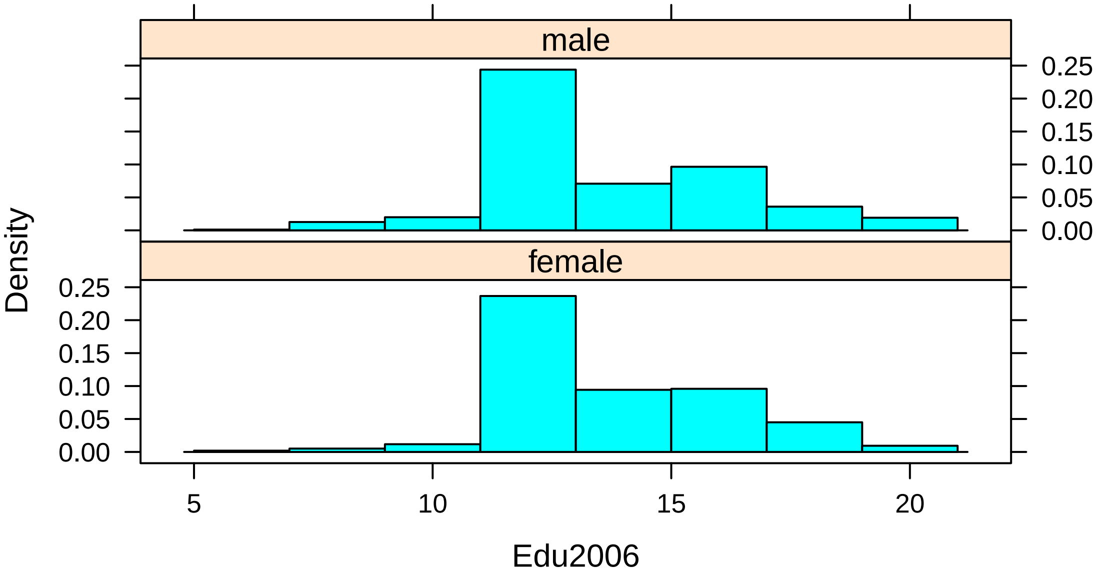
```r
histogram(~ Edu2006, data=NLSY, width=2, xlab="Edu2006")
```



(b)

Here I plot two histograms for the variable Edu2006 of females and males.

```r
histogram(~Edu2006 | Gender, data=NLSY, width=2, xlab="Edu2006",layout=c(1,2))
```



Here I export the favorite summaries (mean, SD, five-number summary) for Edu2006 of males and females.

```
favstats(Edu2006 ~ Gender, data=NLSY)
```

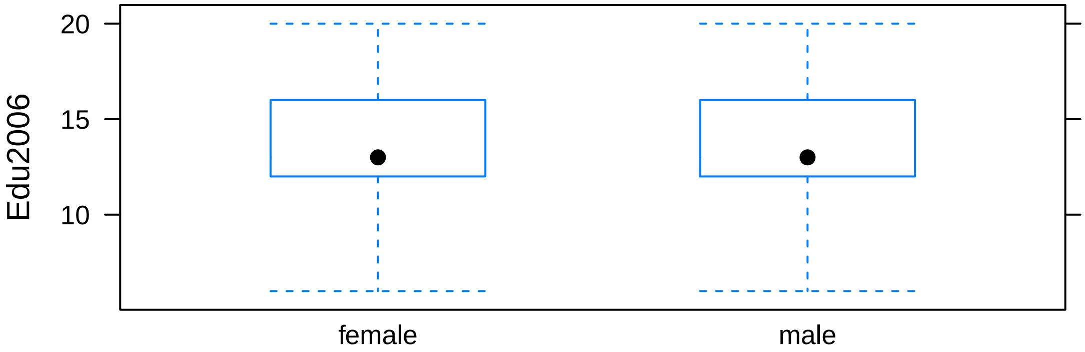
```
##   Gender min Q1 median Q3 max    mean      sd    n missing
## 1 female   6 12     13 16  20 13.9703 2.41226 1278       0
## 2   male   6 12     13 16  20 13.8132 2.58827 1306       0
```

From the histogram and the summary, males and females in the data set have the same distributions in their education levels.

(c)

Here I make a side-by-side boxplot comparing the education levels of males and females.

```
bwplot(Edu2006 ~ Gender, data=NLSY)
```

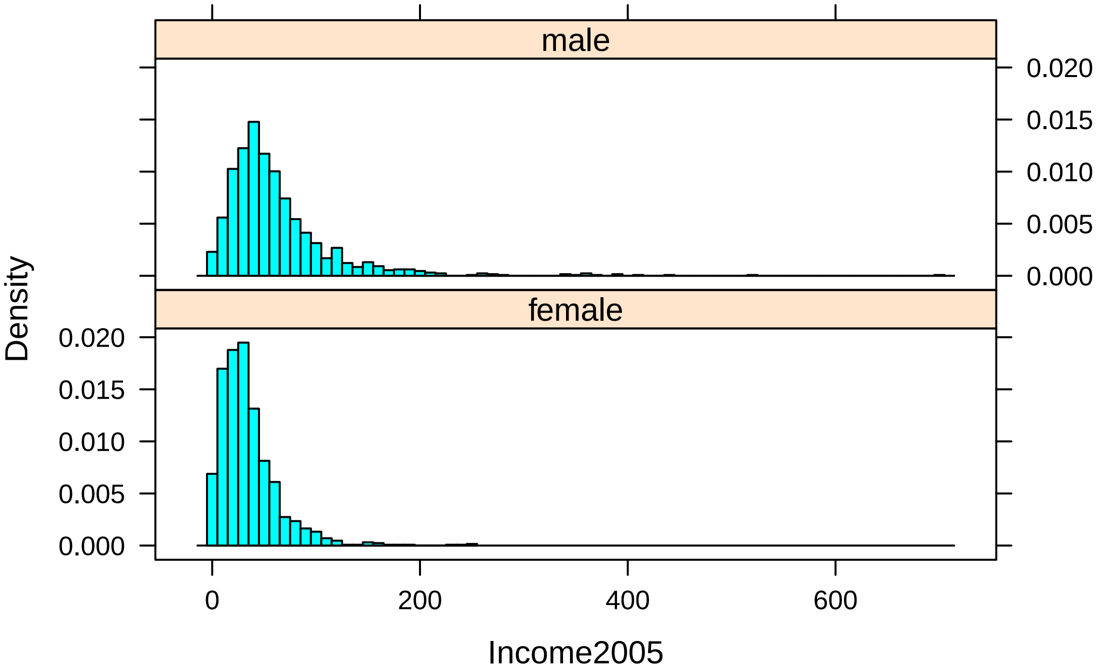


The histogram gives a good sense of the distribution (the overall shape) of the variable Edu2006 but the boxplot does not.

(d)

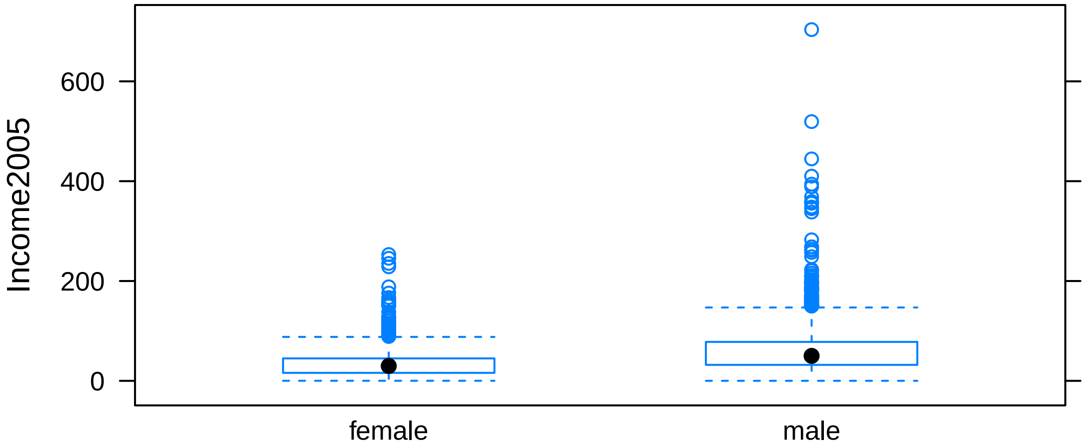
Here I make two histograms comparing the Income2005 of males and females.

```
histogram(~Income2005 | Gender, data=NLSY, width=10, xlab="Income2005", layout=c(1,2))
```



Here I make a side-by-side boxplot comparing the Income2005 of males and females.

```
bwplot(Income2005 ~ Gender, data=NLSY)
```



Here I export the favorite summaries comparing the Income2005 of males and females.

```
favstats(Income2005 ~ Gender, data=NLSY)
```

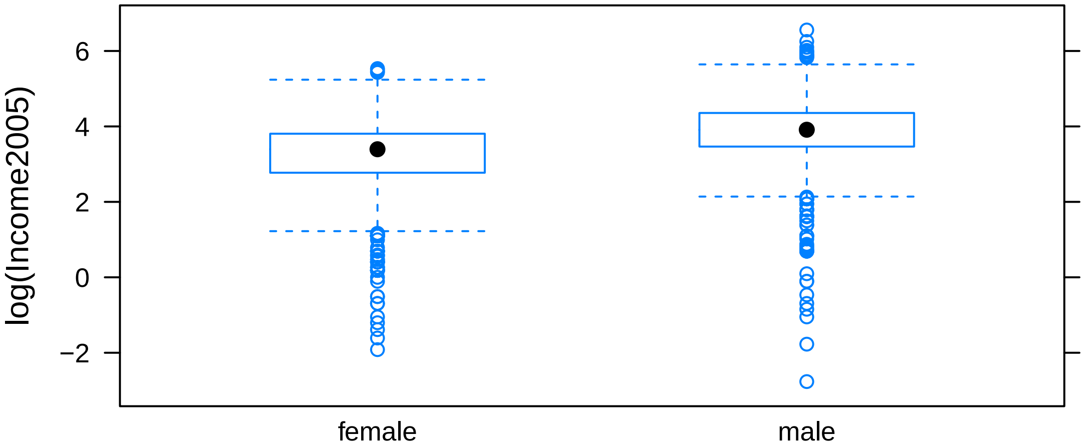
```
##   Gender   min Q1  median Q3     max    mean      sd    n missing
## 1 female 0.147 16 29.8105 45 253.043 35.2107 28.7764 1278       0
## 2   male 0.063 32 50.0000 78 703.637 63.3187 55.8611 1306       0
```

From the two histograms we know that the shape is right skewed. From the boxplot we know that males had a higher income. From the standard deviation we know that males had a higher variability in their distributions of incomes.

(e)

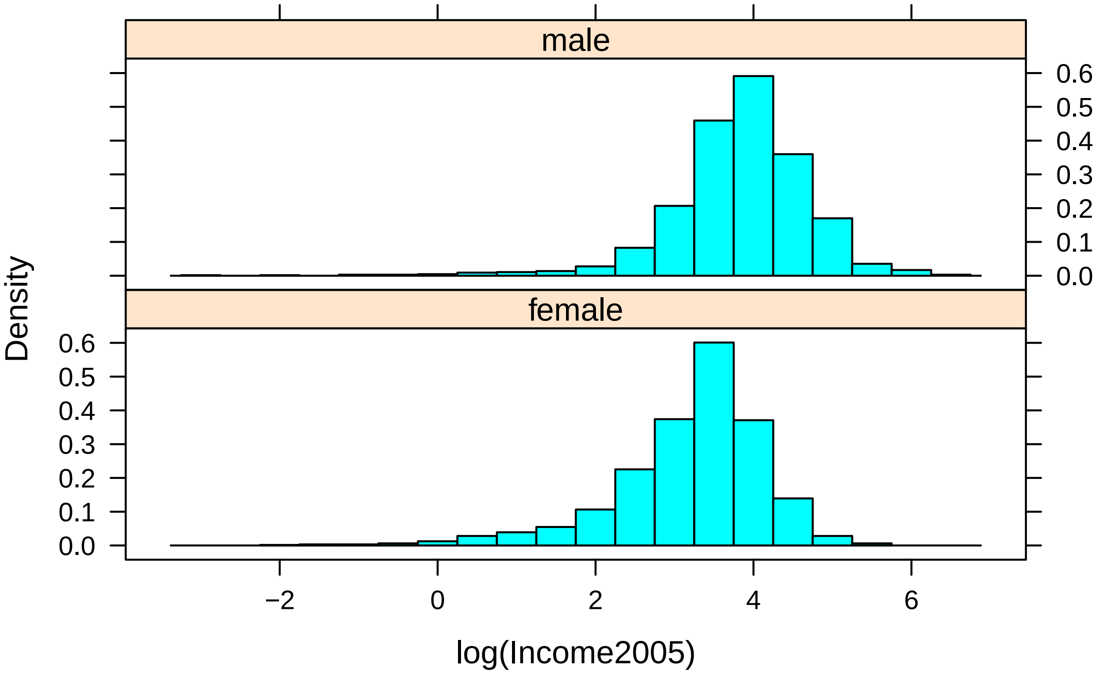
Here I make a side-by-side boxplot comparing the logarithm of Income2005 of males and females.

```
bwplot(log(Income2005) ~ Gender, data=NLSY)
```



Here I make two histograms comparing the logarithm of Income2005 of males and females.

```
histogram(~log(Income2005) | Gender, data=NLSY, width=0.5, xlab="log(Income2005)", layout=c(1,2))
```



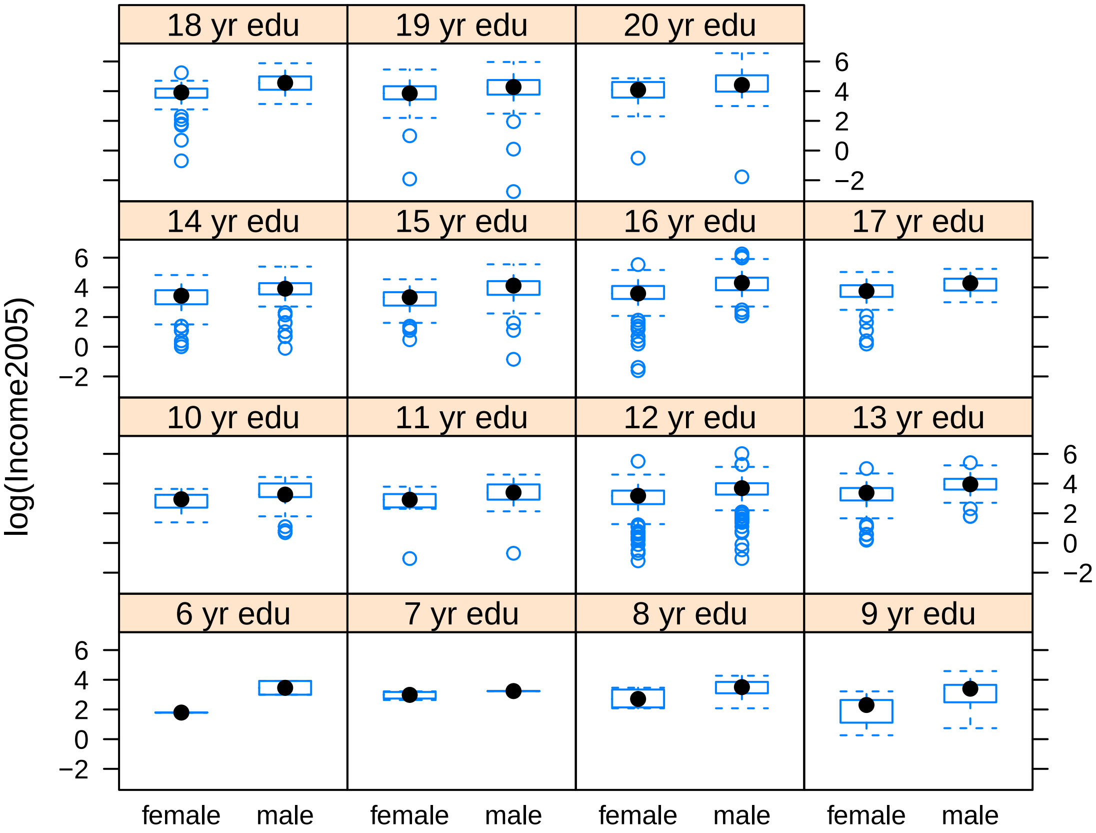
The distributions after the log transformation is left skewed.

(f)

Here I split the data by the levels of Edu2006, and make a side-by-side boxplot comparing the logarithm of Income2005 of males and females.

```
NLSY$Edu2006.fac = factor(NLSY$Edu2006, labels = paste(6:20,"yr edu"))
```

```
bwplot(log(Income2005) ~ Gender | Edu2006.fac, data=NLSY)
```

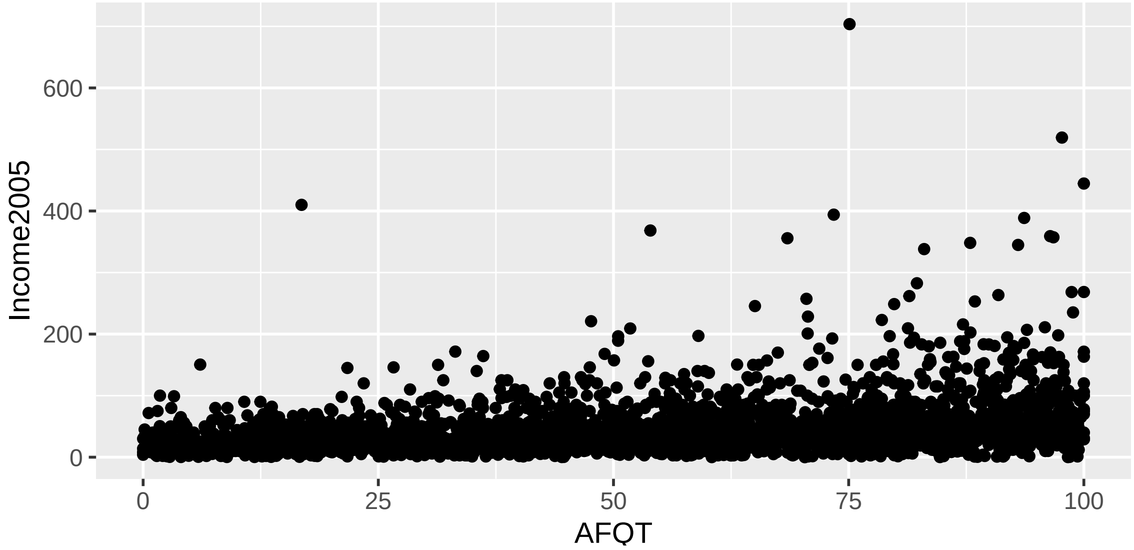


From the above boxplot we know men earn more than women, even after adjusted for their education level.

(g)

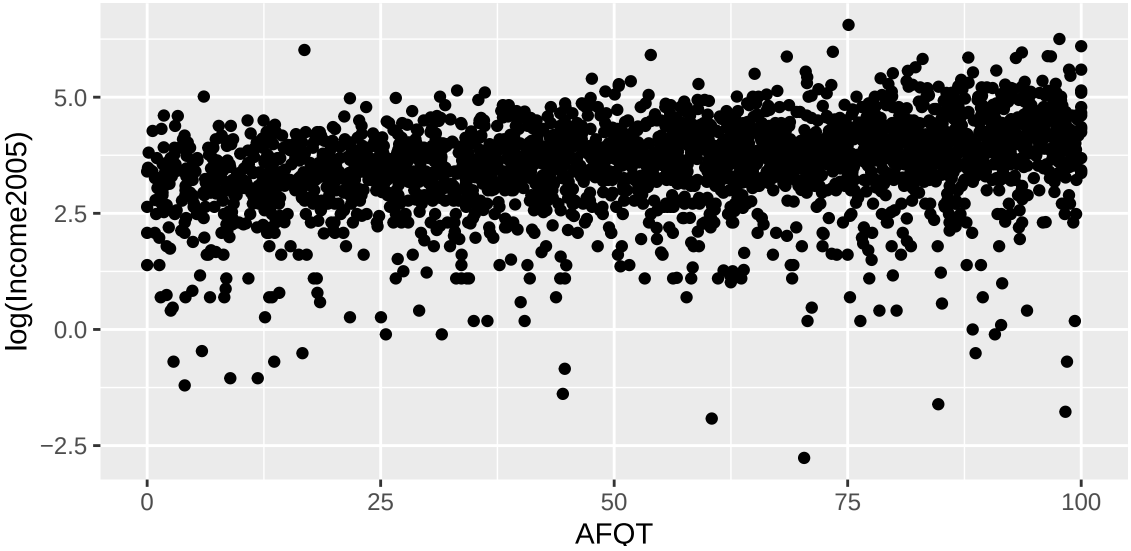
Here I make a scatter plot between AFQT and Income2005.

```
qplot(AFQT, Income2005, data=NLSY, xlab="AFQT", ylab="Income2005")
```



Here I make a scatter plot between AFQT and logarithm of Income2005.

```r
qplot(AFQT, log(Income2005), data=NLSY, xlab="AFQT", ylab="log(Income2005)")
```

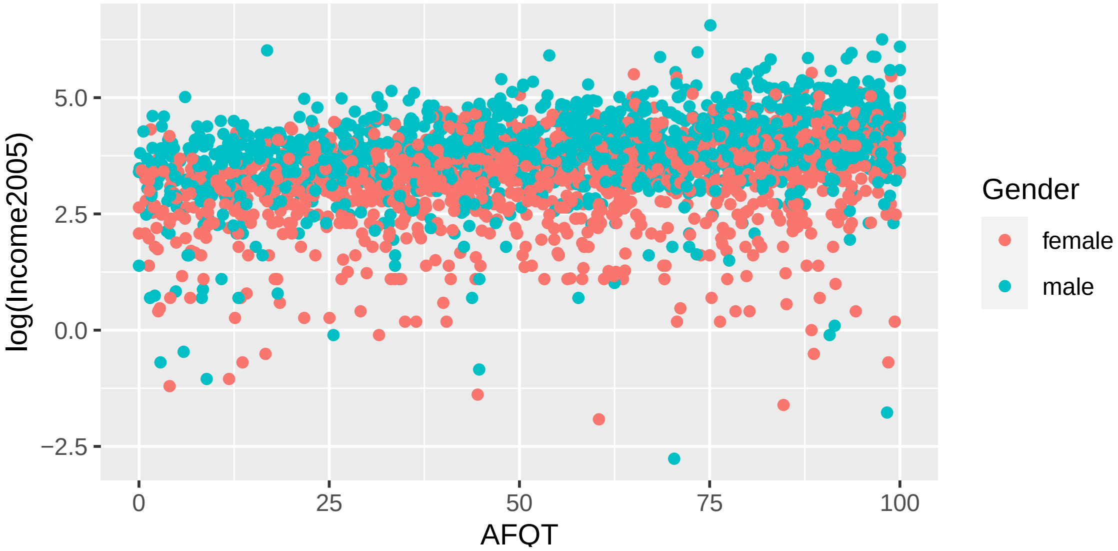


1. The Income2005 had a weak positive association with AFQT.
2. The variability of Income2005 had a positive association with AFQT.
3. The variability of the logarithm of Income2005 had a positive association with AFQT.

(h)

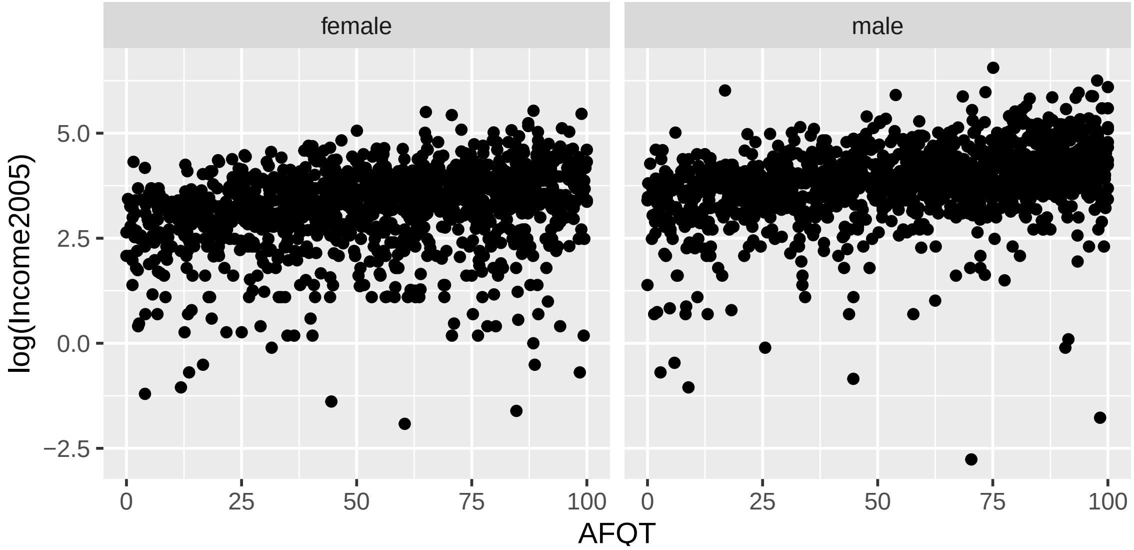
Here I make a color-coded scatterplot between AFQT and Income2005, and the color of points represents the Gender of the subject.

```r
qplot(AFQT, log(Income2005), data=NLSY, xlab="AFQT", ylab="log(Income2005)",color=Gender)
```



Here I make two separate scatterplots between AFQT and Income2005 for men and women.

```r
qplot(AFQT, log(Income2005), data=NLSY, xlab="AFQT", ylab="log(Income2005)",facets = ~Gender)
```

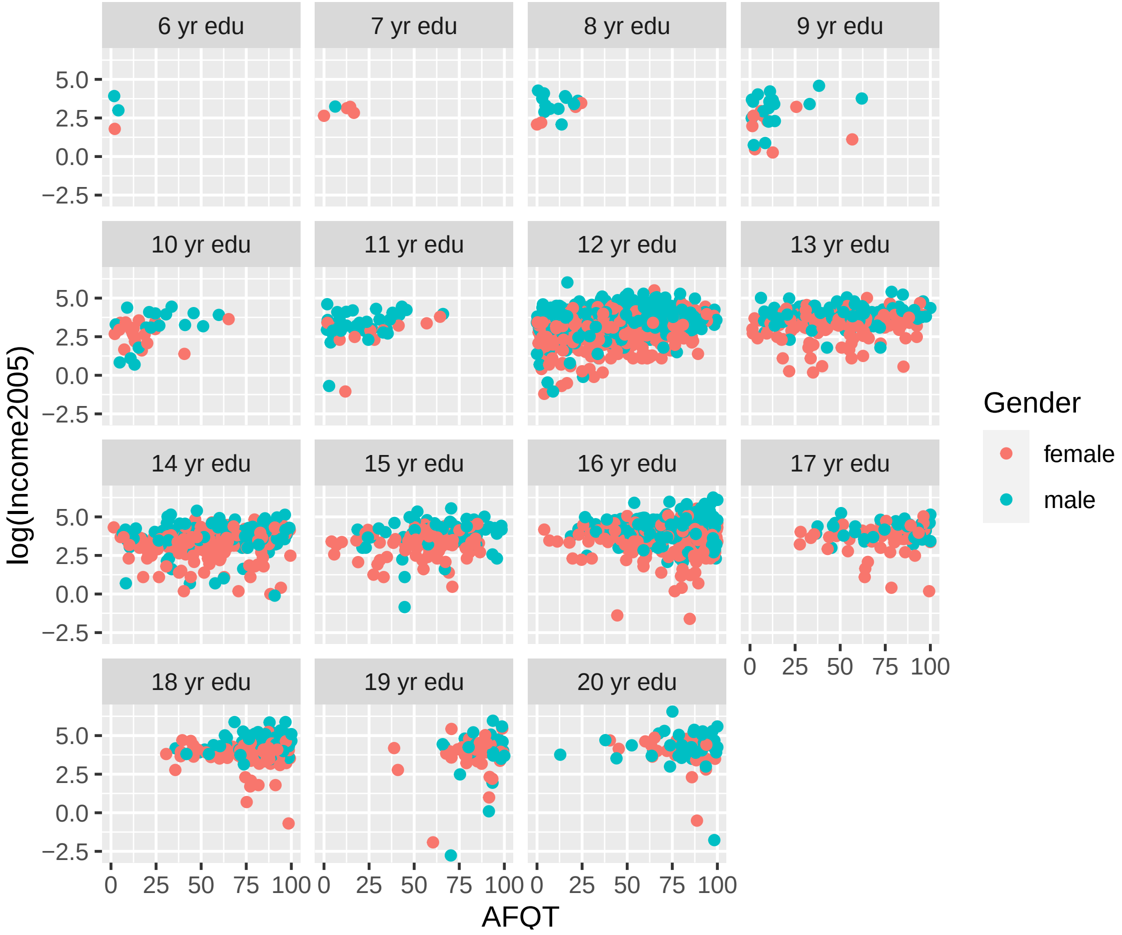


1. For females, the logarithm of Income2005 had a positive association with AFQT, and the variability of the logarithm of Income2005 also had a positive association with AFQT. For males, the results are the same: the logarithm of Income2005 had a positive association with AFQT, and the variability of the logarithm of Income2005 also had a positive association with AFQT.
2. Comparing men and women with similar intelligence test score percentiles, men earn more than women in general.

(i)

Here I make color-coded scatterplots between AFQT and Income2005 for each level of years of education, using the color of points to represent the Gender of the subjects.

```
qplot(AFQT, log(Income2005), data=NLSY, xlab="AFQT", ylab="log(Income2005)",
      color=Gender, facets=~Edu2006.fac)
```



Comparing men and women with the same years of education and with similar intelligence test score percentiles, men earn more than women in general.