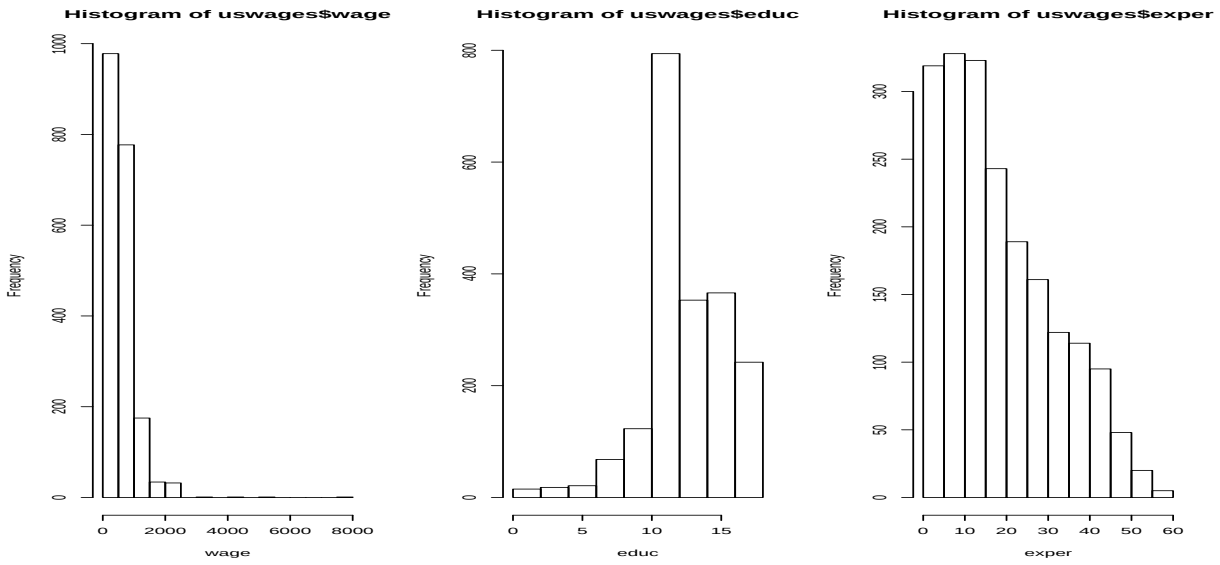
Stat 500 – Homework 1 (Solutions)

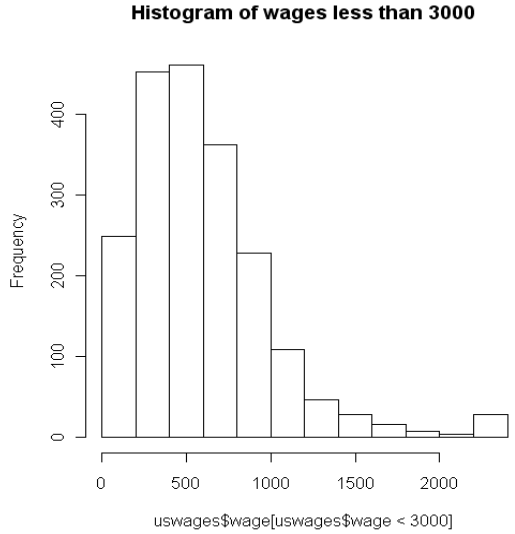
After loading the data, fix the categorical variables **race, smsa, ne, mw, so, we,** and **pt** so that they are treated as factors (e.g. for **race,** use the following command – do similarly for others). Also treat the negative values of **exper** as missing data. Then do the summary:

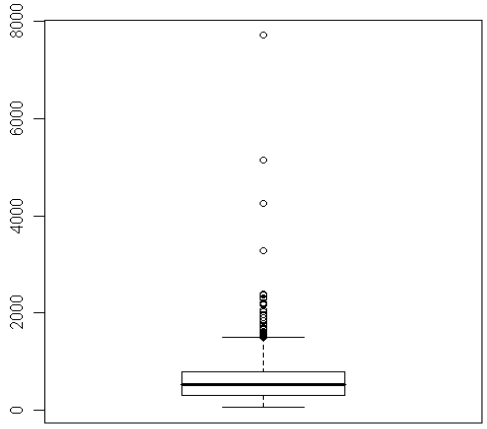
```
> uswages$race<-as.factor(uswages$race)
> uswages$exper[uswages$exper<0]<-NA   # changing negative values to NA
> summary(uswages)
      wage              educ           exper       race     smsa     ne
 Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   0:1844   0: 488   0:1542
 1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1: 156   1:1512   1: 458
 Median : 522.32   Median :12.00   Median :16.00
 Mean   : 608.12   Mean   :13.11   Mean   :18.74
 3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00
 Max.   :7716.05   Max.   :18.00   Max.   :59.00
                                   NA's   :33.00
 mw       so       we       pt
 0:1503   0:1375   0:1580   0:1815
 1: 497   1: 625   1: 420   1: 185
```

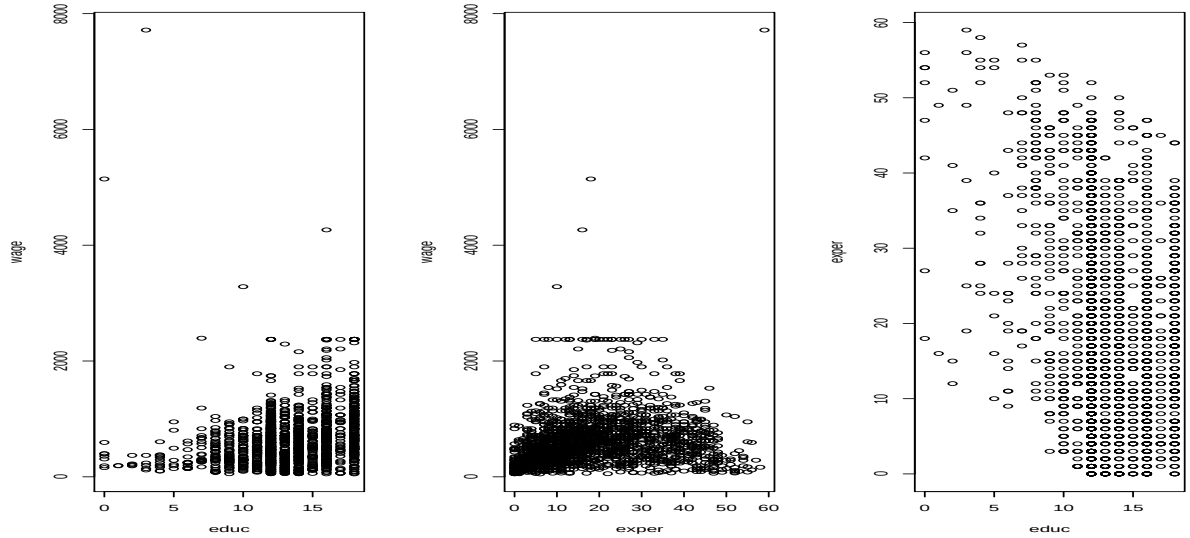


In this dataset, the mean of **wage** is much larger than the median, suggesting the distribution is right skewed and may have large outliers, as confirmed by the histogram and boxplot. It suggests us make a histogram for wages less than 3000. From the other two histograms, we can conclude that the distributions of **educ** and **exper** are moderately left skewed and moderately right skewed respectively. Following are the pairwise scatterplots between the 3 numerical variables. We do not see any strong pattern, which is justified by the correlation matrix shown below:

```
> round(cor(uswages[,1:3]),2)
```








```
      wage  educ exper
wage  1.00  0.25  0.18
educ  0.25  1.00 -0.30
exper 0.18 -0.30  1.00
```

Finally we examine the distribution of **wage** across the levels of the different factors. From the side-by-side boxplots below, we see that on average, the wage of a **black** worker is slightly less than a **white** worker, the wage of a **part-time** worker is a bit less than a **full-time** worker, and the wage is slightly higher if the worker is from a **Standard Metropolitan Statistical Area.** Other boxplots tell that there is no real difference in wages among workers from different regions of the country (North East, Midwest, West, and South). The command is the following:

```
>boxplot(uswages[uswages$ne==1,1],uswages[uswages$mw==1,1],uswages[uswages$so==1,1]
```

```
+,uswages[uswages$we==1,1])
```