Wages, Gender and other factors

In this notebook we make use of a small data set("CPS1985.xlsx") concerning employees and examine:

- the composition of the work force with respect to various employee characteristics(variables).
- whether there is any sign of of possible wage difference between men and women and
- if there is any profound correlation among the variables

We first import the relevant data which are available as an Excel file.

```
setwd("D:/data/Econometrics and Applied Statistics")
library(readxl)
cps <- read_excel("CPS1985.xlsx")
```

We first take a look at the data

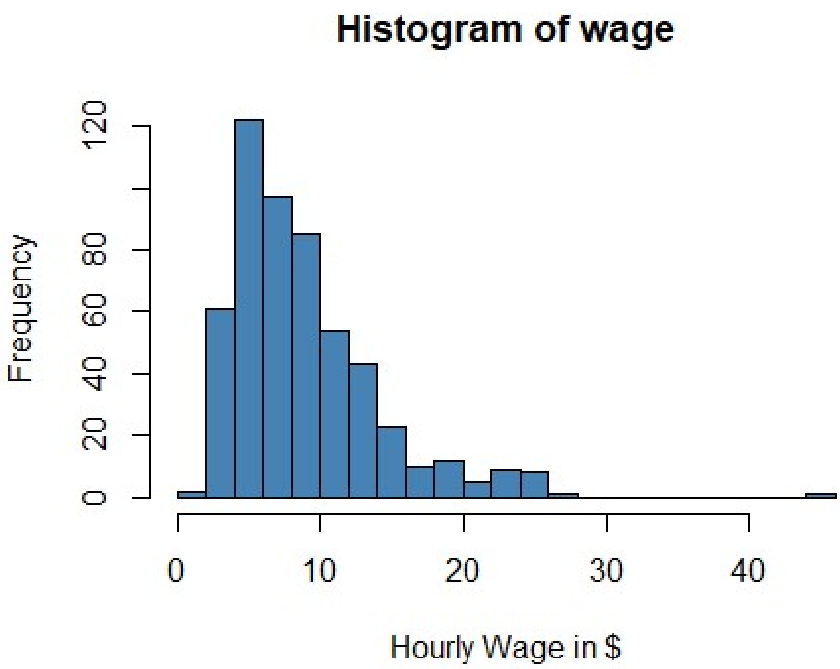
```
summary(cps)
```

```
##     rownames         wage          education       experience          age
##  Min.   :  1   Min.   : 1.000   Min.   : 2.00   Min.   : 0.00   Min.   :18.0
##  1st Qu.:134   1st Qu.: 5.250   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:28.0
##  Median :267   Median : 7.780   Median :12.00   Median :15.00   Median :35.0
##  Mean   :267   Mean   : 9.032   Mean   :13.03   Mean   :17.78   Mean   :36.8
##  3rd Qu.:400   3rd Qu.:11.250   3rd Qu.:15.00   3rd Qu.:26.00   3rd Qu.:44.0
##  Max.   :533   Max.   :44.500   Max.   :18.00   Max.   :55.00   Max.   :64.0
##   ethnicity            region              gender            occupation
##  Length:533         Length:533         Length:533         Length:533
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     sector              union              married
##  Length:533         Length:533         Length:533
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

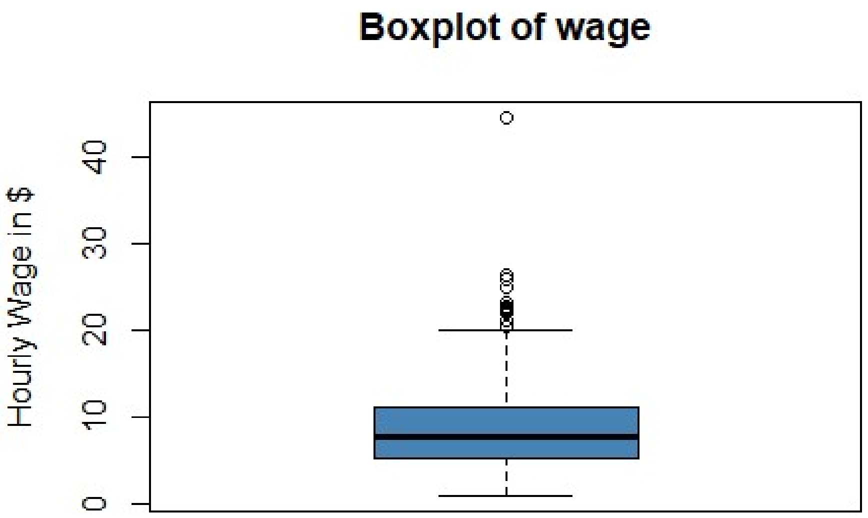
from which we get the basic descriptive statistics for each numerical-variable and the length for each character-variable.

Additionally, let us draw a Histogram and a Boxplot for each variable. From the diagrams we can check for skewness and look for outliers (that we may or may not decide to get rid off).

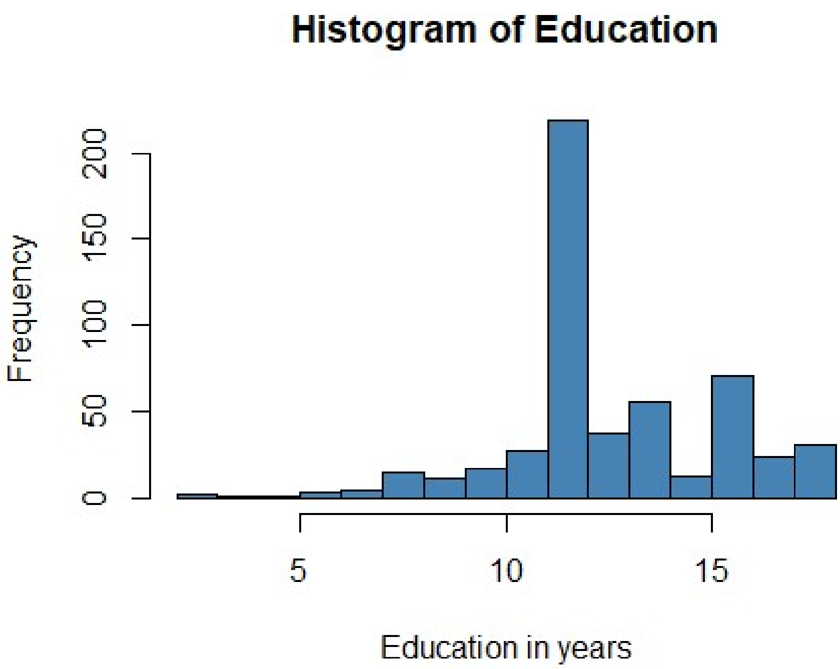
```
hist(cps$wage,
xlab = "Hourly Wage in $",
main = "Histogram of wage",
col = "steelblue", breaks = 20)
```



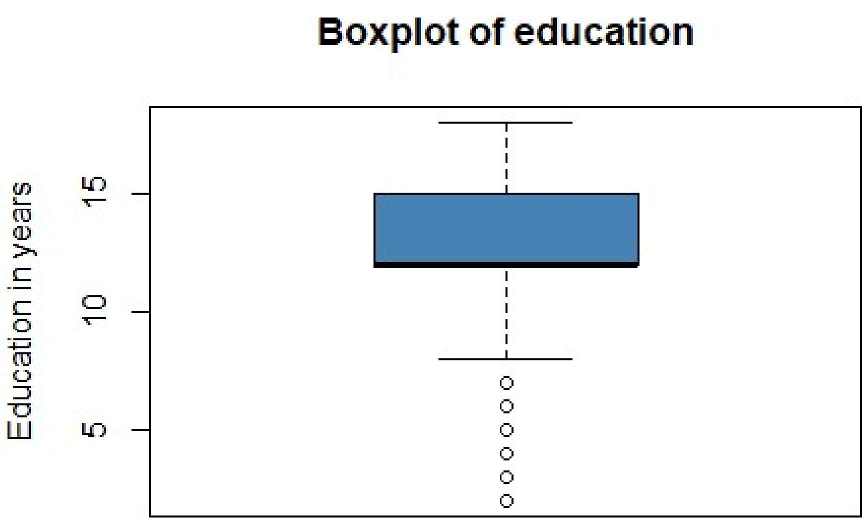

```r
boxplot(cps$wage,
ylab = "Hourly Wage in $",
main = "Boxplot of wage",
col = "steelblue")
```



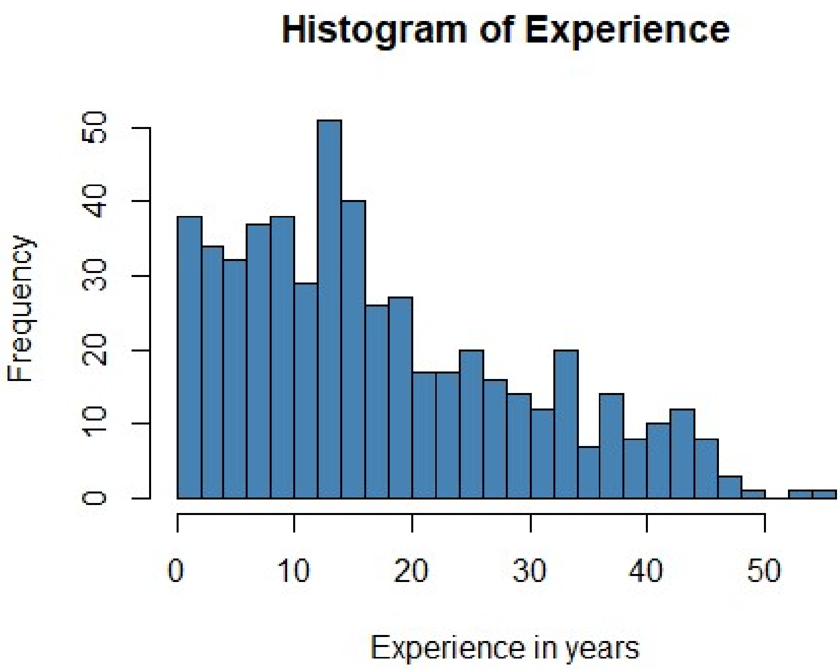

```r
hist(cps$education,
xlab = "Education in years",
main = "Histogram of Education",
col = "steelblue", breaks = 20)
```



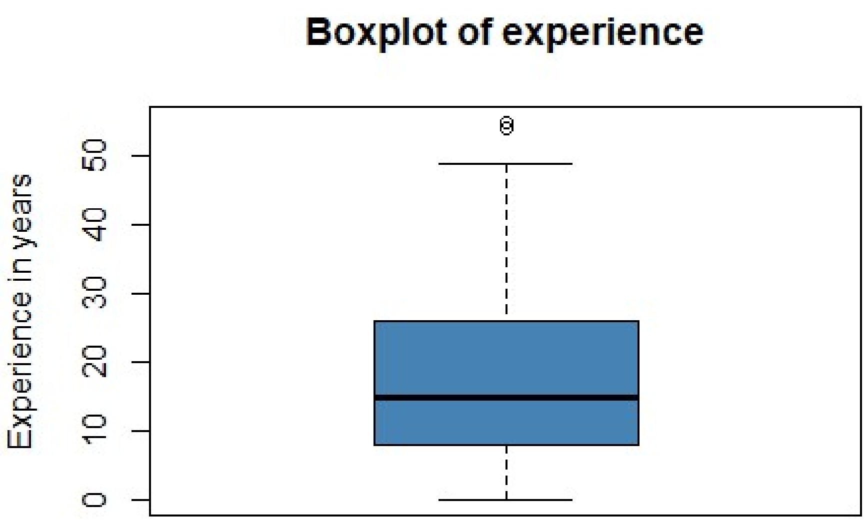

```
boxplot(cps$education,
ylab = "Education in years",
main = "Boxplot of education",
col = "steelblue")
```



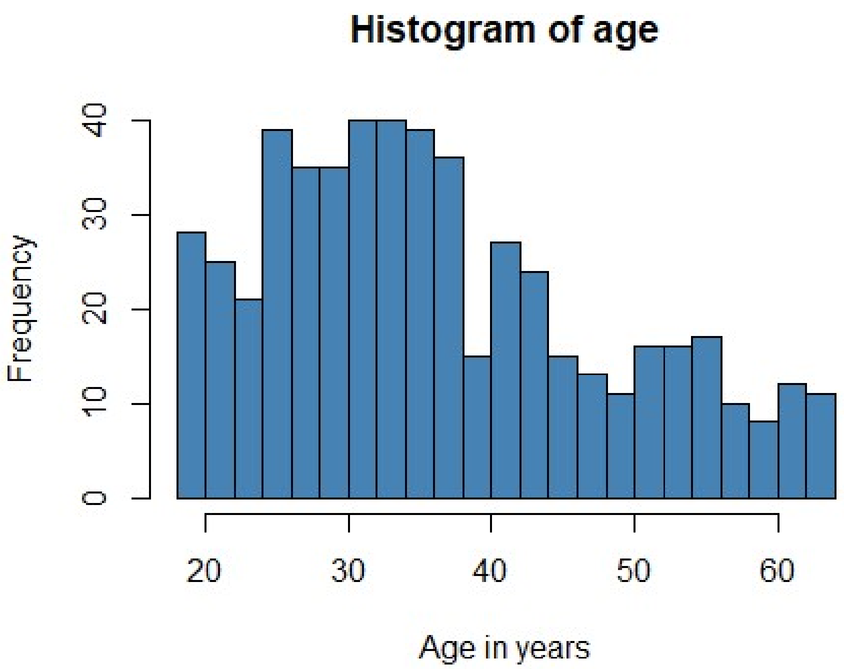

```
hist(cps$experience,
xlab = "Experience in years",
main = "Histogram of Experience",
col = "steelblue", breaks = 20)
```



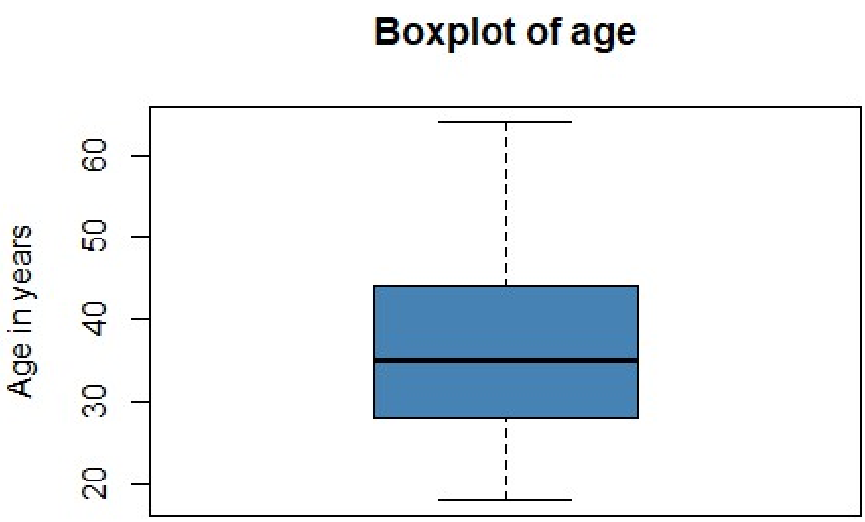

```r
boxplot(cps$experience,
ylab = "Experience in years",
main = "Boxplot of experience",
col = "steelblue")
```




```r
hist(cps$age,
xlab = "Age in years",
main = "Histogram of age",
col = "steelblue", breaks = 20)
```




```
boxplot(cps$age,
ylab = "Age in years",
main = "Boxplot of age",
col = "steelblue")
```



We calculate also the distribution of each categorical variable

```
table(cps$ethnicity)
```

```
##
##      cauc hispanic    other
##       439       27       67
```

```r
table(cps$region)
```

```
##
## other south
##   377   156
```

```r
table(cps$gender)
```

```
##
## female   male
##    244    289
```

```r
table(cps$occupation)
```

```
##
## management     office      sales   services  technical     worker
##         55         97         38         83        105        155
```

```r
table(cps$sector)
```

```
##
##  construction manufacturing         other
##            24            98           411
```

```r
table(cps$union)
```

```
##
##  no yes
## 437  96
```

```r
table(cps$married)
```

```
##
##  no yes
## 184 349
```

We can also calculate each distribution across the possible values of a specific categorical variable. For example, a statistic of interest would be the distribution of working sector, marital status or wage across gender

```r
table(cps$gender, cps$sector)
```

```
##
##          construction manufacturing other
##   female            2            38   204
##   male             22            60   207
```

```r
table(cps$gender, cps$married)
```

```
##
##           no yes
##   female  83 161
##   male   101 188
```

```r
tapply(cps$wage, cps$gender, summary)
```

```
## $female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.750   4.718   6.840   7.891  10.000  44.500 
## 
## $male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   6.000   8.930   9.995  13.000  26.290
```

Even more specifically, we are interested in the mean wage, standard deviation and total number of observations of each distinct state of gender(male or female). To retrieve this information we group the initial data by gender and proceed the aforementioned calculations

```
library(dplyr)
avgs <- cps %>%
  group_by(gender) %>%
  summarise(mean(wage),
            sd(wage),
            n())
print(avgs)
```

```
## # A tibble: 2 × 4
##   gender `mean(wage)` `sd(wage)` `n()`
##   <chr>         <dbl>      <dbl> <int>
## 1 female         7.89       4.73   244
## 2 male           9.99       5.29   289
```

A first glance naive conclusion from the above table would be that the average wage for women is about 2$ less than the average wage for men. But is this really true?

To compare the mean wage for women and men and test statistical significance of the difference between them we should split the initial data in 2 appropriate subgroups (men, women) and perform a t-test applied on the variable “wage”

```
male_obs <- cps %>% dplyr::filter(gender == "male")

female_obs <- cps %>% dplyr::filter(gender == "female")
t.test(male_obs$wage, female_obs$wage)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  male_obs$wage and female_obs$wage
## t = 4.8498, df = 529.24, p-value = 1.627e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.251789 2.956317
## sample estimates:
## mean of x mean of y 
##  9.994913  7.890861
```

The above result confirms that the the difference in means is not equal to 0.

To illuminate the procedure we perform the above calculation also manually. To do so we return to the table “avgs” getting access to the estimated E(wage), se(wage) and number of observation for each gender.

```r
# split the dataset by gender
male <- avgs %>% dplyr::filter(gender == "male")

female <- avgs %>% dplyr::filter(gender == "female")

# rename columns of both splits
colnames(male)   <- c("Gender", "Y_bar_m", "s_m", "n_m")
colnames(female) <- c("Gender", "Y_bar_f", "s_f", "n_f")
male

## # A tibble: 1 × 4
##   Gender Y_bar_m   s_m   n_m
##   <chr>    <dbl> <dbl> <int>
## 1 male      9.99  5.29   289

female

## # A tibble: 1 × 4
##   Gender Y_bar_f   s_f   n_f
##   <chr>    <dbl> <dbl> <int>
## 1 female    7.89  4.73   244
```

Now, considering the wage_male and wage_female as independent variables, we then have that for the variable gap = wage_male - wage_female follows asymptotically a t_stastic with:

E(gap) = E(wage_male) - E(wage_female) , var(gap) = var(wage_male) + var(wage_female) , thus

<u>estimated E(gap)</u> is gap_bar Y_bar_m - Y_bar_f

<u>estimated se(gap)</u> is (s_m2 /n_m +s_f2 /n_f)$^{1/2}$

```r
gap <- male$Y_bar_m - female$Y_bar_f

gap_se <- sqrt(male$s_m^2 / male$n_m + female$s_f^2 / female$n_f)
```

So, we finally calculate the 95% confidence interval as follows

```r
gap_ci_l <- gap - 1.96 * gap_se

gap_ci_u <- gap + 1.96 * gap_se

result <- cbind(gap, gap_se, gap_ci_l, gap_ci_u)


print(result, digits = 3)

##      gap gap_se gap_ci_l gap_ci_u
## [1,] 2.1  0.434     1.25     2.95
```

Our result coincides with the result of the automated t-test we have alternatively performed.

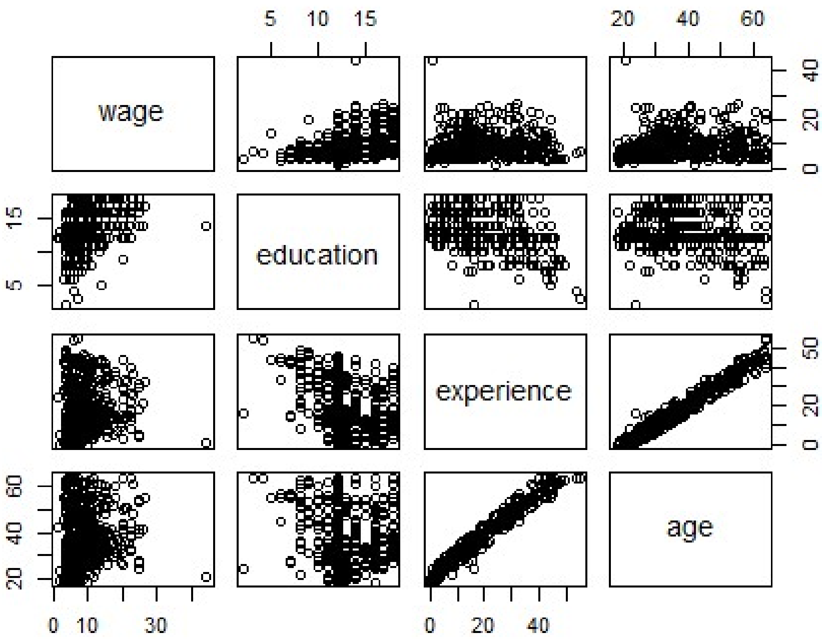
As a final task we examine correlation between the continuous numerical variables of the dataset. To that end we calculate the correlation and create the corresponding scatterplot for each pair

```r
library(corrplot)

## corrplot 0.92 loaded
```

```r
subset=cps[,2:5]
cor1=cor(subset)
corrplot.mixed (cor1, lower.col='black', number.cex=.7)
```


```r
pairs(subset)
```



Among others we observe a very high positive correlation between experience and age, a relatively high positive correlation between wage and education and a relatively high negative correlation between education and experience.