Final Project Report

I. Data description

1. General information

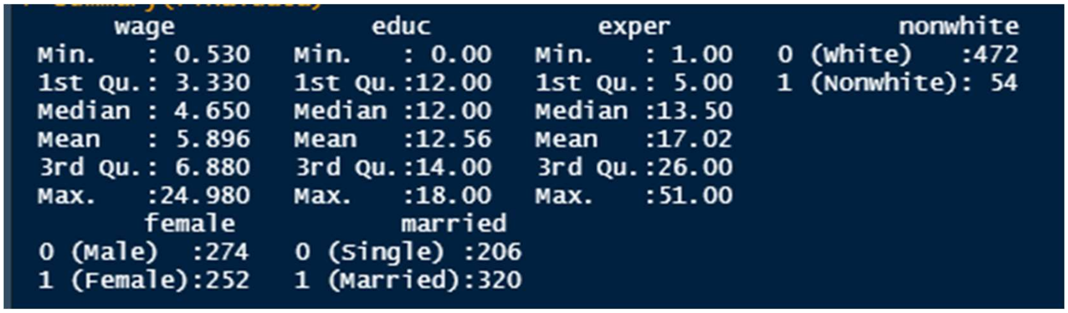
   Dataframe is taken from wage1.csv (data 1 folder). It contains 526 rows and 6 column, which are: wage, educ, exper, nonwhite, female, married.

2. Dataframe structure

No.	Atrribute	Description
1	wage	Worker's average hourly earnings
2	educ	Worker years of education
3	exper	Worker's years potential experience
4	nonwhite	Worker's ace (=1 if nonwhite, =0 if white)
5	female	Worker's gender(=1 if female, =0 if male)
6	married	Worker's marriage status (=1 if married, =0 if single

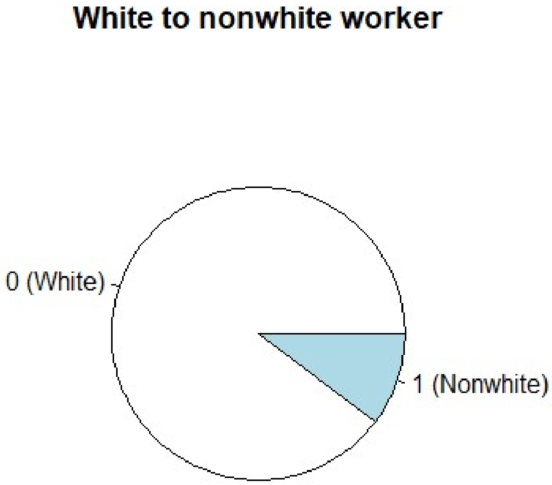
II. Descriptive statistics

1. Data overview



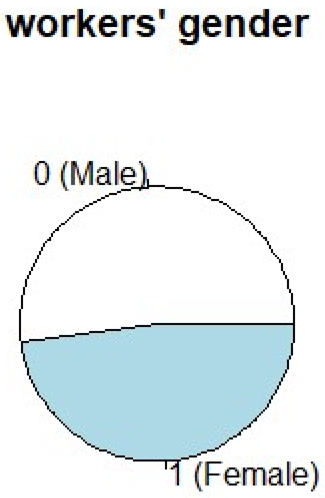
```
      wage             educ            exper               nonwhite
 Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   0 (White)   :472
 1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1 (Nonwhite): 54
 Median : 4.650   Median :12.00   Median :13.50
 Mean   : 5.896   Mean   :12.56   Mean   :17.02
 3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00
 Max.   :24.980   Max.   :18.00   Max.   :51.00
         female           married
 0 (Male)  :274   0 (Single) :206
 1 (Female):252   1 (Married):320
```

2. "nonwhite" attribute



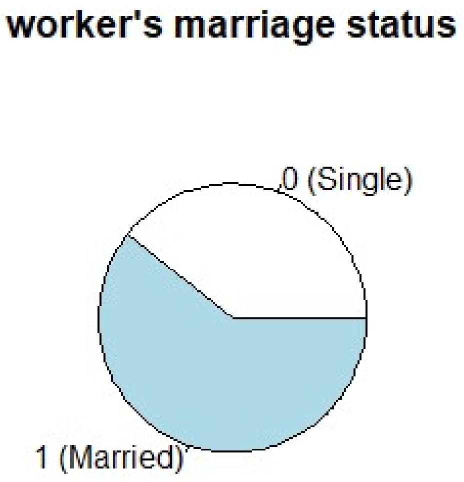
Overrall, white workers outnumber the nonwhite workers.

3. “female’ attribute



Overrall, the amount of female workers is almost the same to the amount of male workers

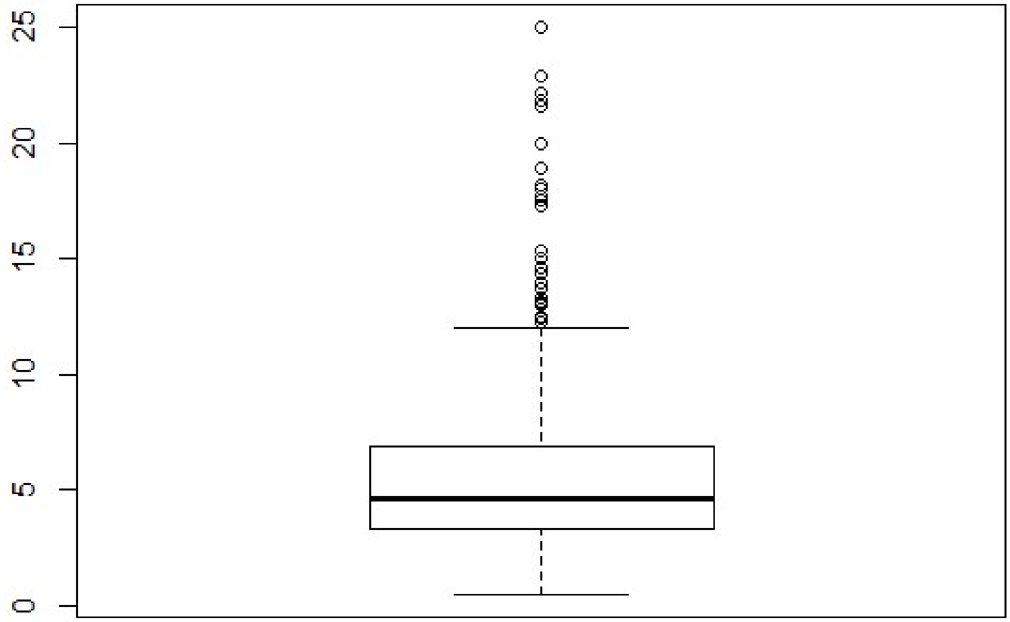
4. “married” attribute

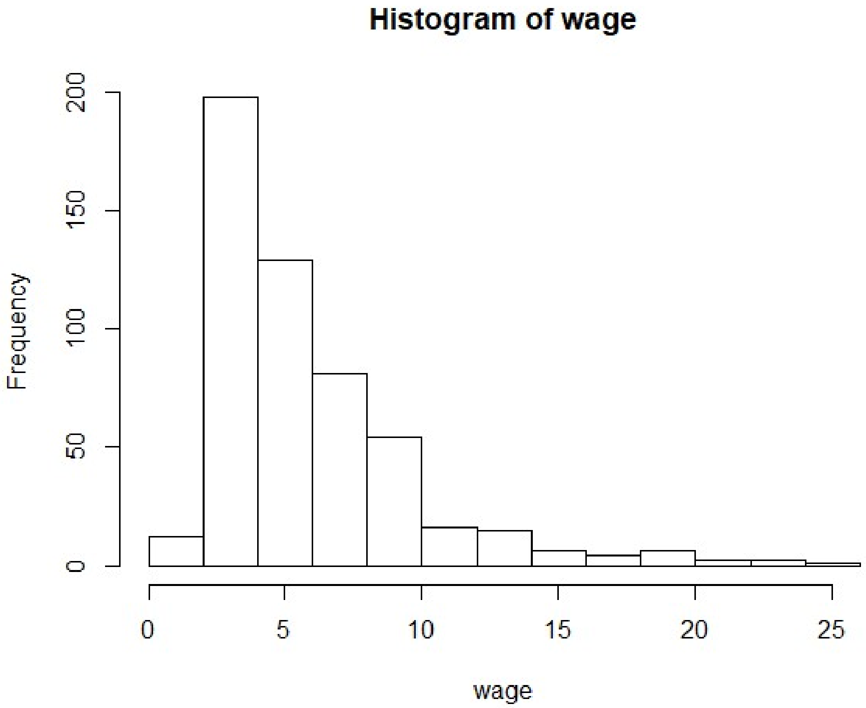


Overrall, there are more married workers than single workers.

5. "wage" attribute

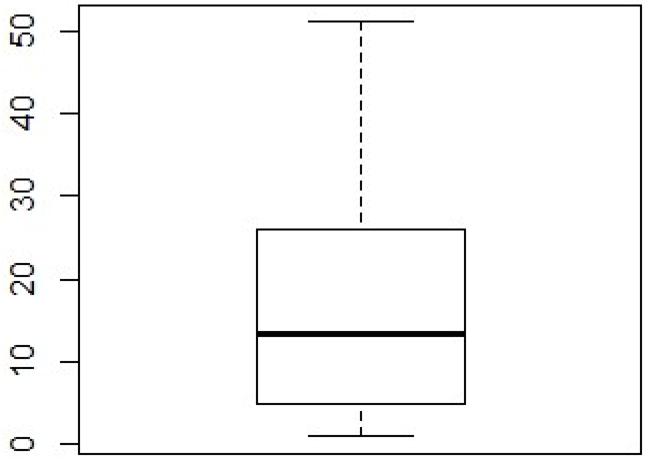
   Boxplot of "wage"

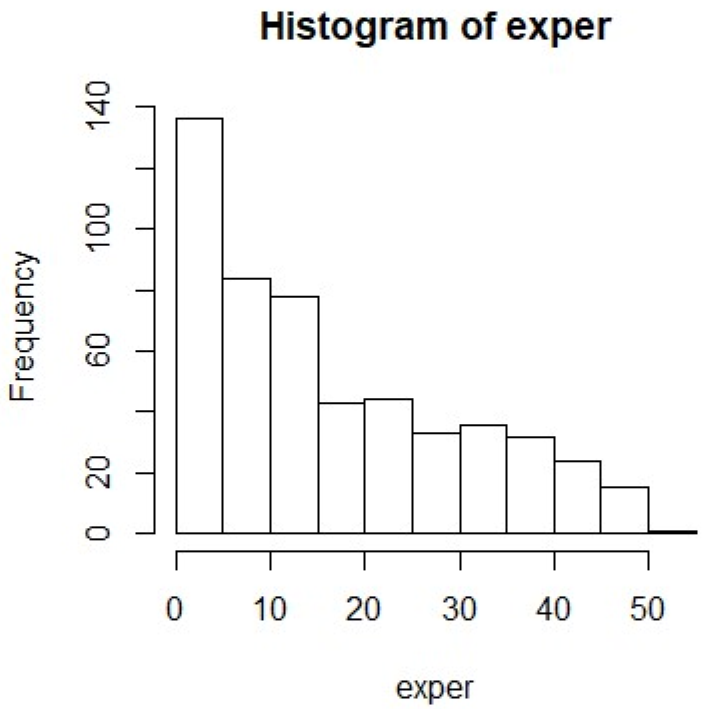




As can be seen, the median of the workers' hourly earning is about 5

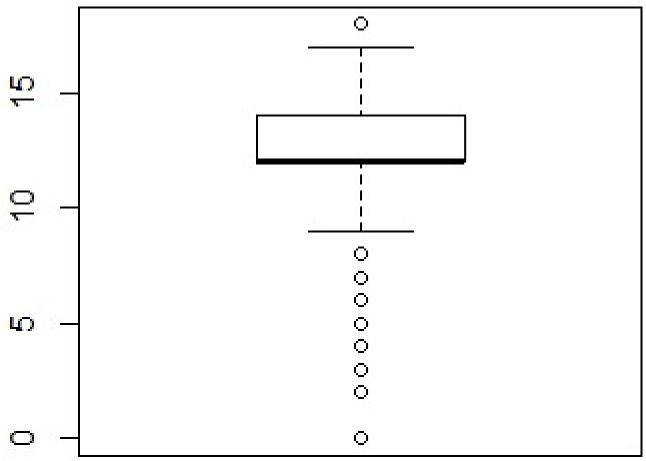
6. ”exper” attribute
   Boxplot of “exper”

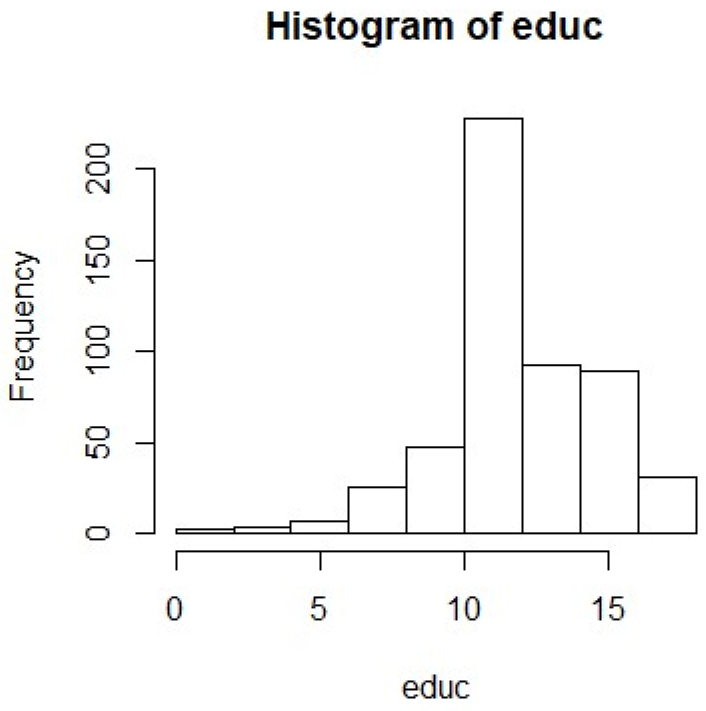




As can be seen, the median of the workers’ years of experience is about 12.

7. "educ attribute"

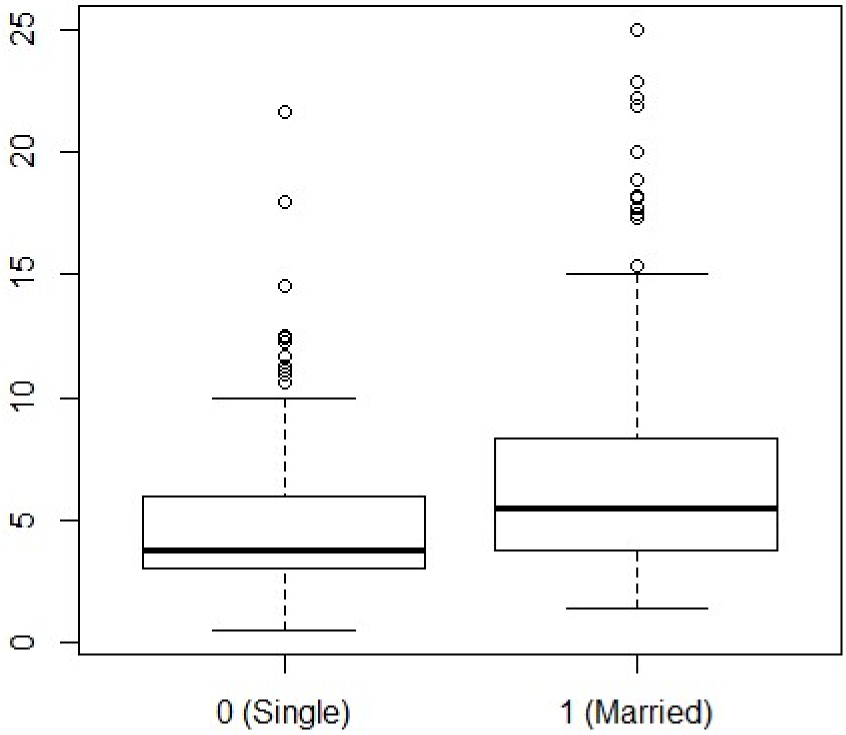




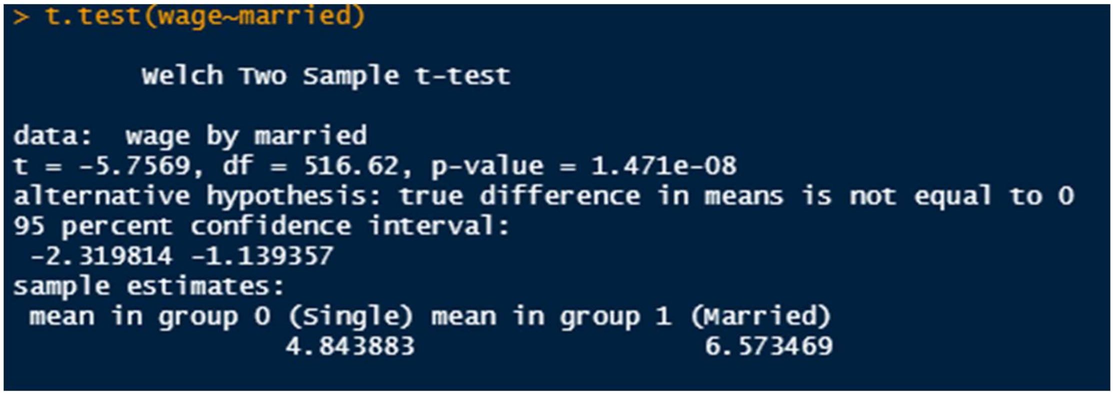
As can be seen, the median of the workers' years of eduction is about 12.

III. Inference statistics

1. Wage and marriage status



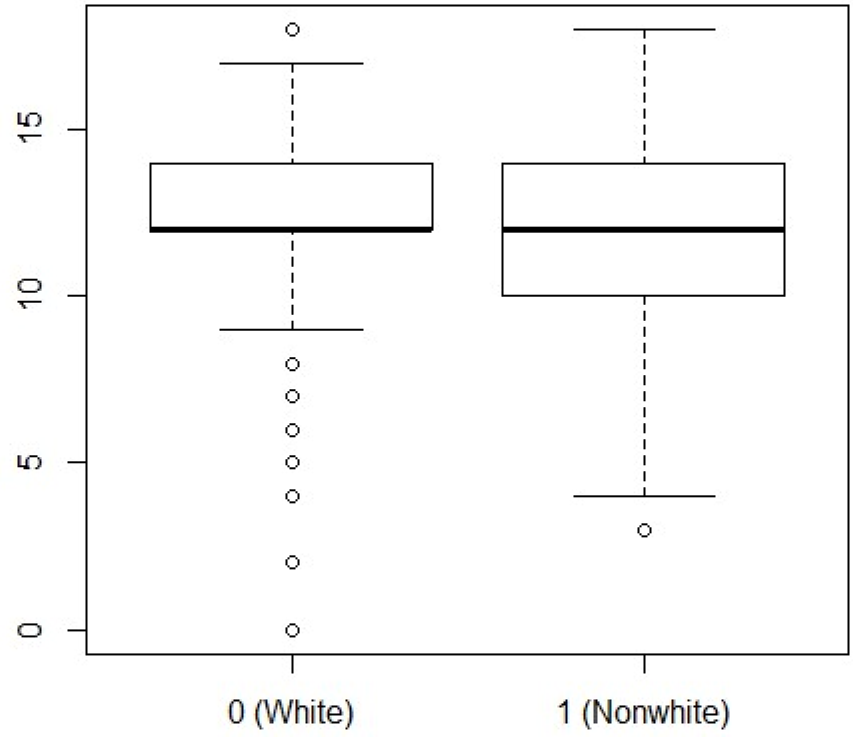
We will test if the single workers have the same wage than the married workers.



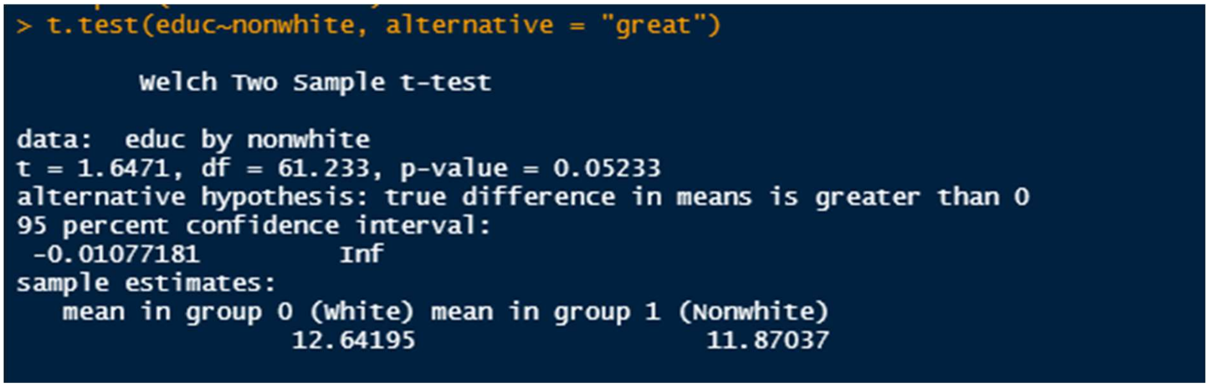
```
> t.test(wage~married)

        Welch Two Sample t-test

data:  wage by married
t = -5.7569, df = 516.62, p-value = 1.471e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.319814 -1.139357
sample estimates:
 mean in group 0 (Single) mean in group 1 (Married)
                 4.843883                  6.573469
```

Since the p-value is almost unnoticeable, we reject the null hypothesis at 90%, 95%, 99% test and conclude that the average wage of single worker is not equal to th the wage of married worker

2. Year of education and ethnicity



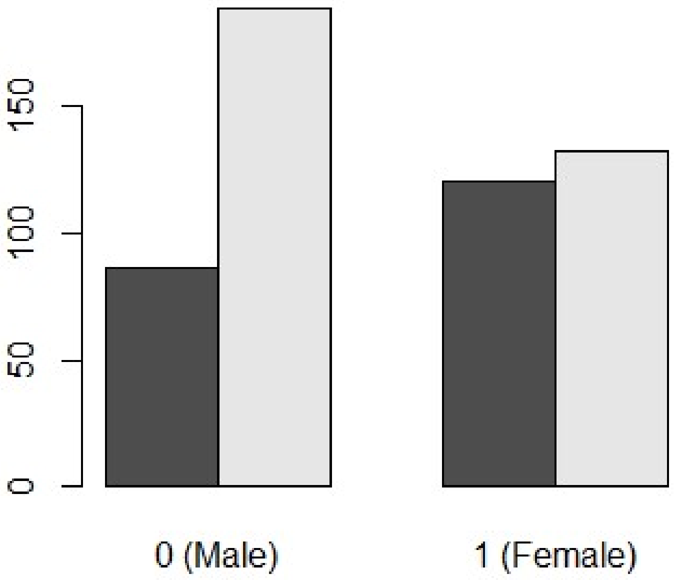
We will test if the white workers have greater average years of eduction than the married workers.



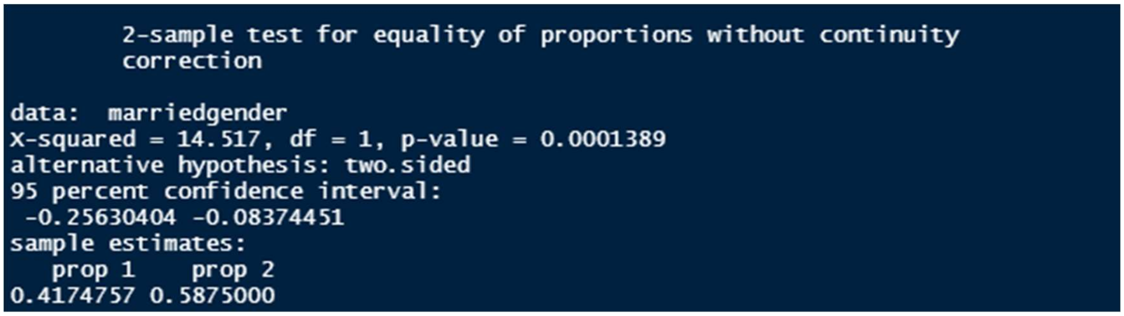
```
> t.test(educ~nonwhite, alternative = "great")

        Welch Two Sample t-test

data:  educ by nonwhite
t = 1.6471, df = 61.233, p-value = 0.05233
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.01077181         Inf
sample estimates:
   mean in group 0 (White) mean in group 1 (Nonwhite)
                  12.64195                   11.87037
```

We accept the null hypothesis at 95% test and conclude that the average years of eductions of white worker is not greater to that of nonwhite worker

3. Marriage status and gender



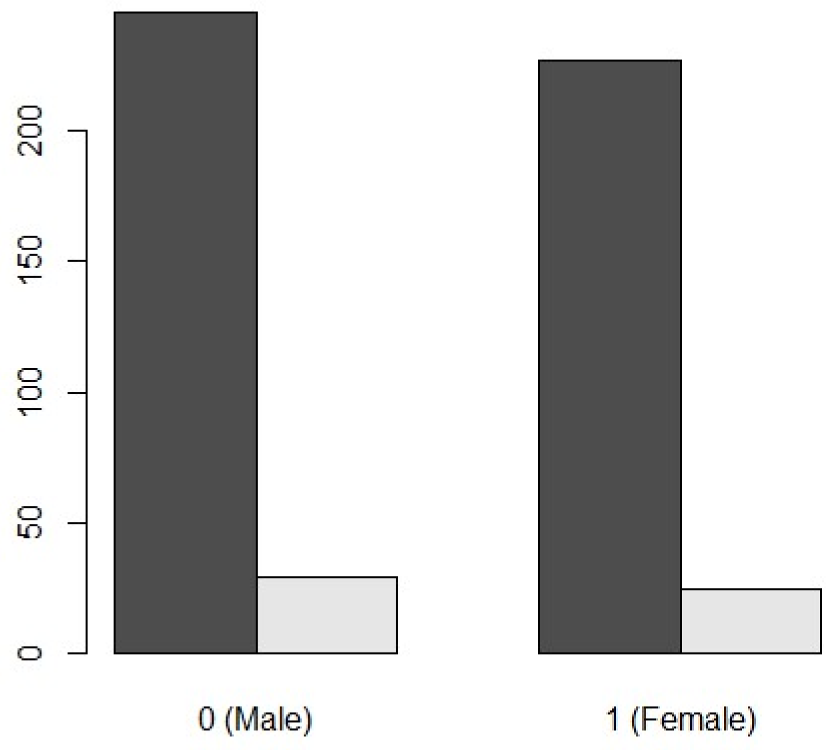
We will test if the proportion of male that is married is equal to the proportion of married women that is married.



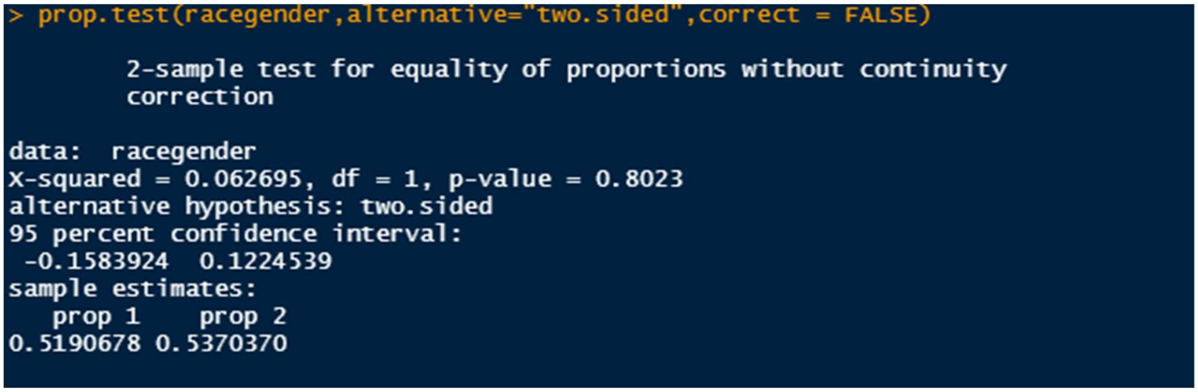
```
	2-sample test for equality of proportions without continuity
	correction

data:  marriedgender
X-squared = 14.517, df = 1, p-value = 0.0001389
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.25630404 -0.08374451
sample estimates:
   prop 1    prop 2 
0.4174757 0.5875000 
```

Since p-value is small(< 0.001) we reject the null hypothesis and conclude that the proportion of male that is married is not equal to the proportion of women that is married.

4. Race and gender



We will test if the proportion of male that is white is equal to the proportion of married women that is white.



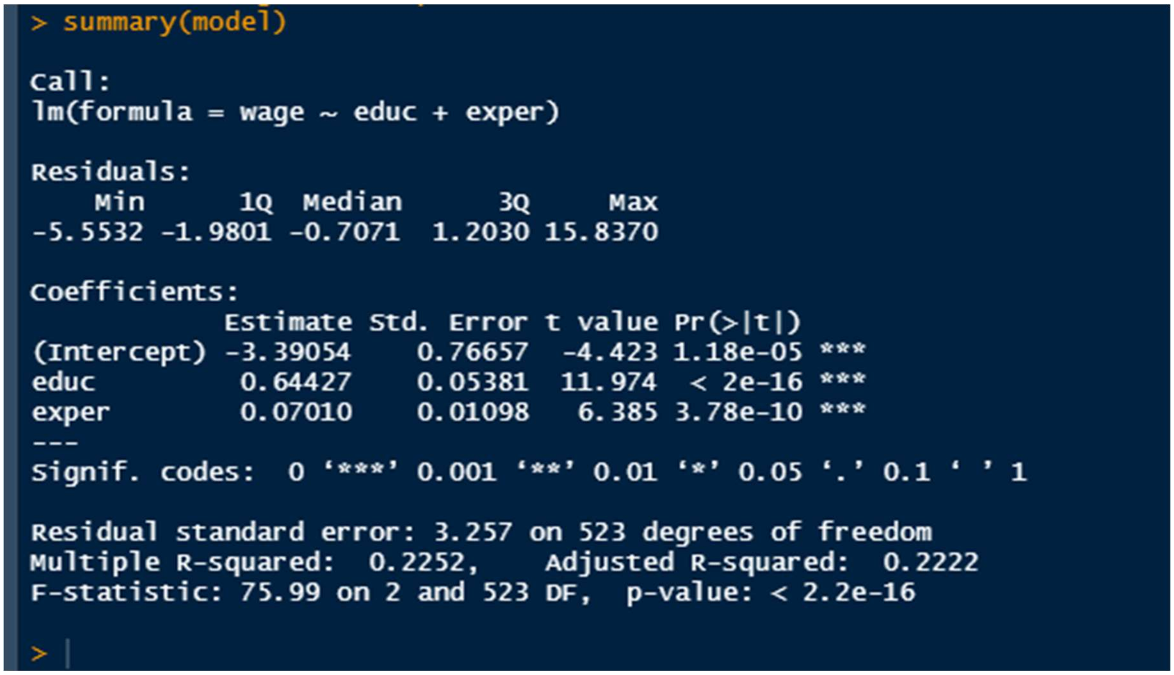
```
> prop.test(racegender,alternative="two.sided",correct = FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  racegender
X-squared = 0.062695, df = 1, p-value = 0.8023
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1583924  0.1224539
sample estimates:
   prop 1    prop 2 
0.5190678 0.5370370 
```

Since p-value is large, we accpect the null hypothesis at 90%, 95% , and 99% test and conclude that the proportion of male that is white is equal to the proportion of women that is white.

IV. Regression models

We will construct a multiple regession model for “wage” with “educ” and “exper” with the formula :

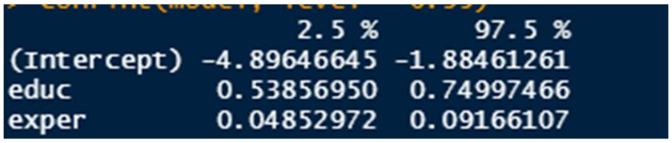
$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \varepsilon$$



```
> summary(model)

Call:
lm(formula = wage ~ educ + exper)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5532 -1.9801 -0.7071  1.2030 15.8370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
educ         0.64427    0.05381  11.974  < 2e-16 ***
exper        0.07010    0.01098   6.385 3.78e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.257 on 523 degrees of freedom
Multiple R-squared:  0.2252,	Adjusted R-squared:  0.2222 
F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16

> |
```

- The estimate for the intercept β_0 is -3.39054, which means when the average educ is 0 and exper is 0 , average wage is -3.39054.
- The estimate for educ coeffiecient β_1 is 0.64427 ,which means whenever the average educ increase by 1, the average wage increase by 0.64427
- The estimate for educ coeffiecient β_2 is 0.07010, which means whenever the average educ increase by 1, the average wage increase by 0.07010

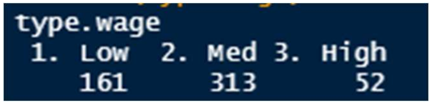
The 95% confident intervals of three cofficients:



```
                  2.5 %      97.5 %
(Intercept) -4.89646645 -1.88461261
educ         0.53856950  0.74997466
exper        0.04852972  0.09166107
```

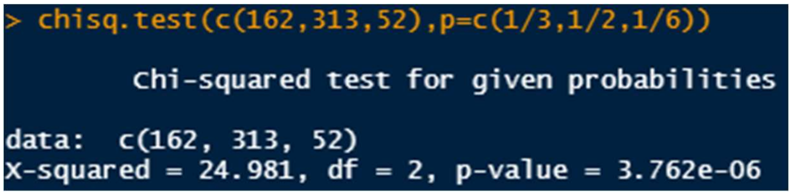
V. Goodness of fit test

We will categorize the “wage” attribute into 3 categories: Low (below 3.5), Med(from 3.5 to 10), High(above 10):



```
type.wage
 1. Low  2. Med 3. High 
    161     313      52 
```

We will test if the distribution of those 3 categories are equal to 1/3, 1/2 and 1/6 respectively.
Null hypothesis: the distribution of 3 categories are equal to 1/3, 1/2 and 1/6 respectively
Alternative hypothesis: : the distribution of 3 categories are not equal to 1/3, 1/2 and 1/6 respectively



```
> chisq.test(c(162,313,52),p=c(1/3,1/2,1/6))

        Chi-squared test for given probabilities

data:  c(162, 313, 52)
X-squared = 24.981, df = 2, p-value = 3.762e-06
```

Since p-value is insignificant, we reject the null hypothesis and conclude that the distribution of 3 categories are not equal to 1/3, 1/2 and 1/6 respectively