Exploration Analysis of Economic Dataset

```
library(pacman)
```

```
## Warning: package 'pacman' was built under R version 4.1.3
```

```
p_load(tidyverse,rvest,tseries,lmtest)

read.csv("D:\\Econometria\\earning_data.csv",sep=",")%>%as_tibble()
```

```
## # A tibble: 50,742 x 13
##        X   age female  hisp educa~1 earni~2 hours  week union uncov region  race
##    <int> <int>  <int> <int>   <int>   <dbl> <int> <int> <int> <int>  <int> <int>
##  1     1    52      0     0      12  146000    45    52     0     0      1     1
##  2     2    38      0     0      18   50000    45    52     0     0      1     1
##  3     3    38      0     0      14   32000    40    51     0     0      1     1
##  4     4    41      1     0      13   47000    40    52     0     0      1     1
##  5     5    42      0     0      13  161525    50    52     1     0      1     1
##  6     6    66      1     0      13   33000    40    52     0     0      1     1
##  7     7    51      0     0      16   37000    44    52     0     0      1     1
##  8     8    49      1     0      16   37000    44    52     0     0      1     1
##  9     9    33      0     0      16   80000    40    52     0     0      1     1
## 10    10    52      1     0      14   32000    40    52     0     0      1     1
## # ... with 50,732 more rows, 1 more variable: marital <int>, and abbreviated
## #   variable names 1: education, 2: earnings
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
economics<-read.csv("D:\\Econometria\\earning_data.csv",sep=",")%>%as_tibble()
```

Columns of the dataset.

```
head(economics)
```

```
## # A tibble: 6 x 13
##       X   age female  hisp educat~1 earni~2 hours  week union uncov region  race
##   <int> <int>  <int> <int>    <int>   <dbl> <int> <int> <int> <int>  <int> <int>
## 1     1    52      0     0       12  146000    45    52     0     0      1     1
## 2     2    38      0     0       18   50000    45    52     0     0      1     1
## 3     3    38      0     0       14   32000    40    51     0     0      1     1
## 4     4    41      1     0       13   47000    40    52     0     0      1     1
## 5     5    42      0     0       13  161525    50    52     1     0      1     1
## 6     6    66      1     0       13   33000    40    52     0     0      1     1
## # ... with 1 more variable: marital <int>, and abbreviated variable names
## #   1: education, 2: earnings
## # i Use `colnames()` to see all variable names
```

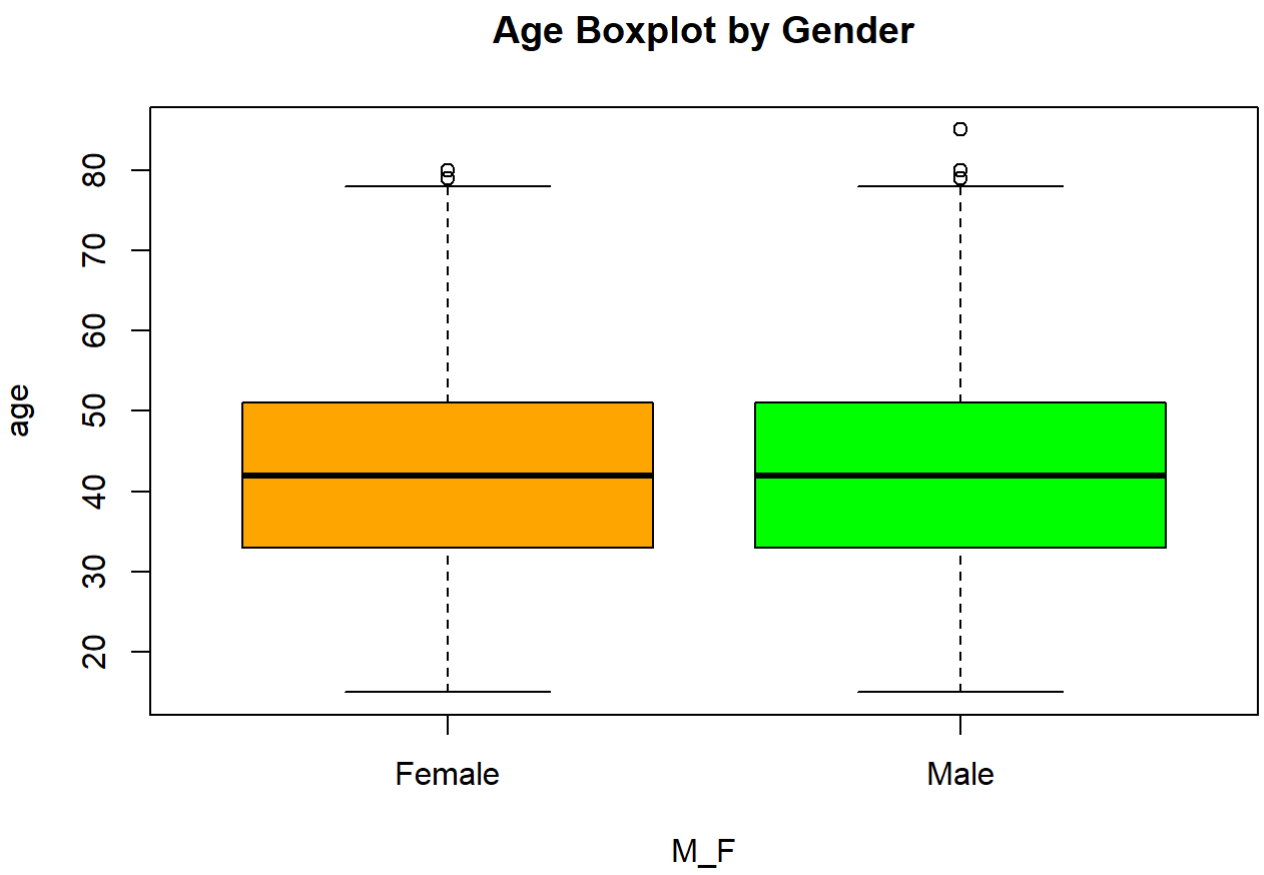
```
tail(economics)
```

```
## # A tibble: 6 x 13
##       X   age female  hisp educat~1 earni~2 hours  week union uncov region  race
##   <int> <int>  <int> <int>    <int>   <dbl> <int> <int> <int> <int>  <int> <int>
## 1 50737    25      1     0       11   20000    37    52     0     0      4     4
## 2 50738    58      1     0       11   30000    40    52     0     0      4     4
## 3 50739    62      1     0       16   35000    40    52     0     0      4     4
## 4 50740    58      0     0       12   75000    50    52     0     0      4     1
## 5 50741    45      1     0       12   40000    60    52     0     0      4     1
## 6 50742    40      0     0       11   60000    40    52     0     0      4     9
## # ... with 1 more variable: marital <int>, and abbreviated variable names
## #   1: education, 2: earnings
## # i Use `colnames()` to see all variable names
```

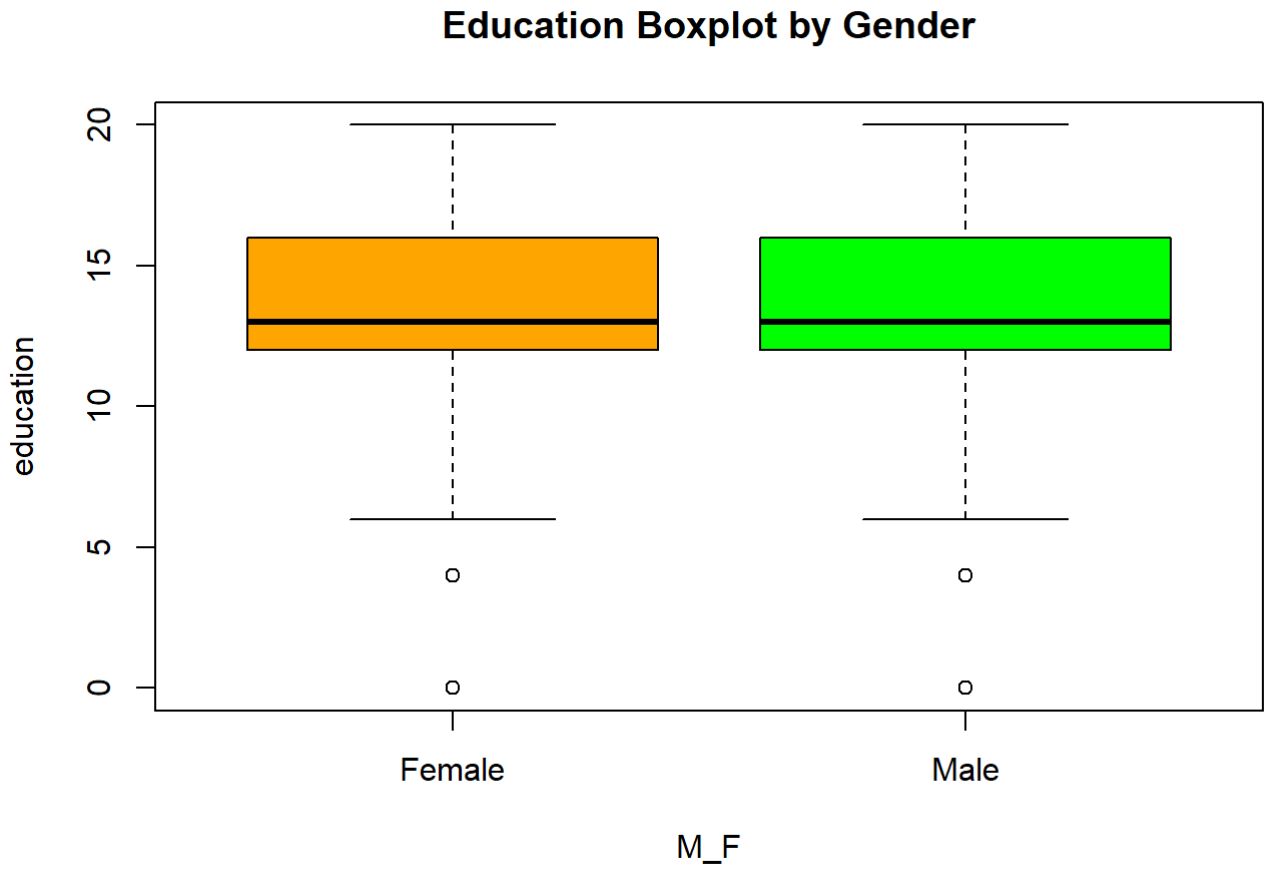
```
economics%>%mutate(M_F=if_else(female==0,"Male","Female"),Marital_status=if_else(marital==1,
"Married","Not Married"))->economics
```

Reading data pattern by boxplots:

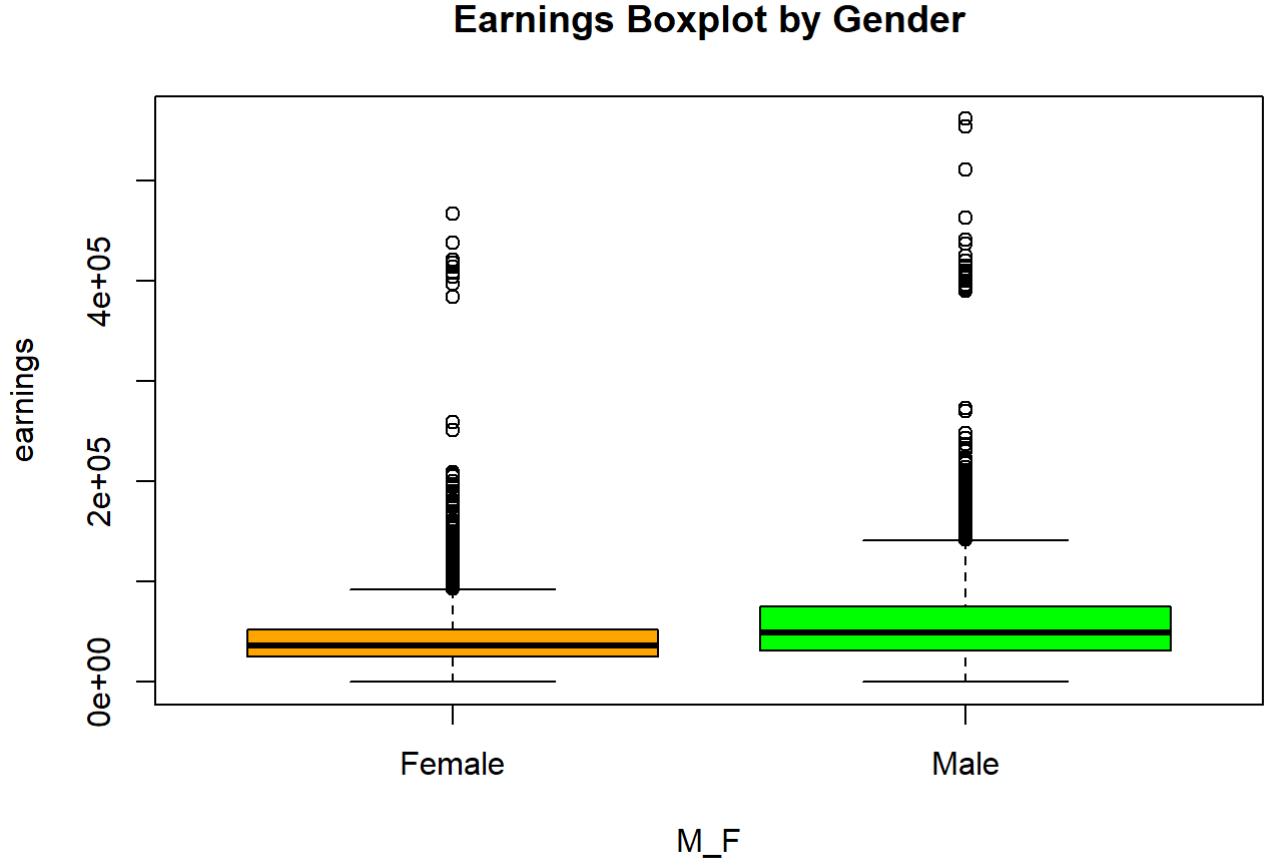
```
boxplot(age~M_F,col=c("orange","green"),main="Age Boxplot by Gender",data=economics)
```




```
boxplot(education~M_F,col=c("orange","green"),main="Education Boxplot by Gender",data=economi
cs)
```

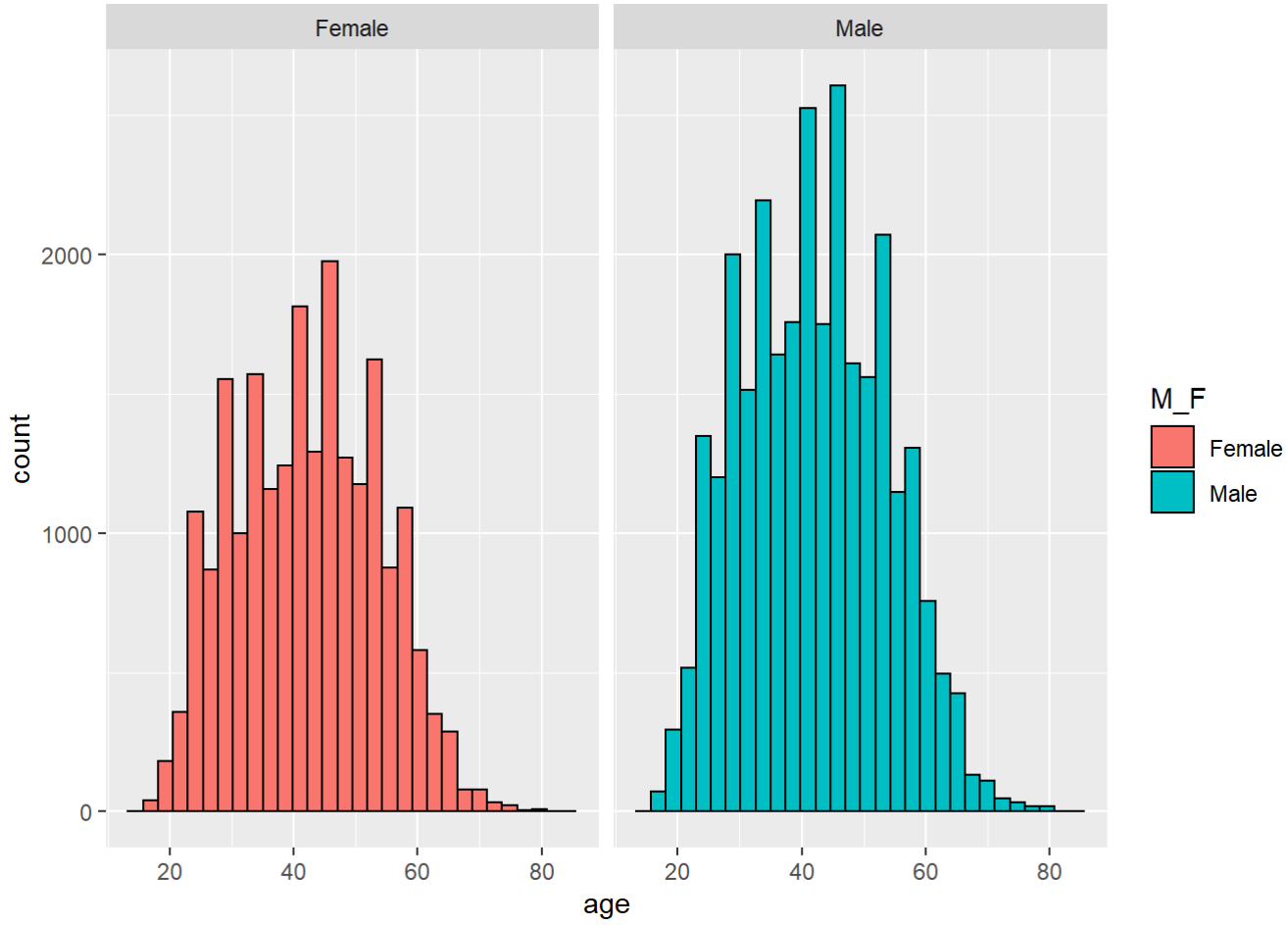



```
boxplot(earnings~M_F,col=c("orange","green"),main="Earnings Boxplot by Gender",data=economic
s)
```



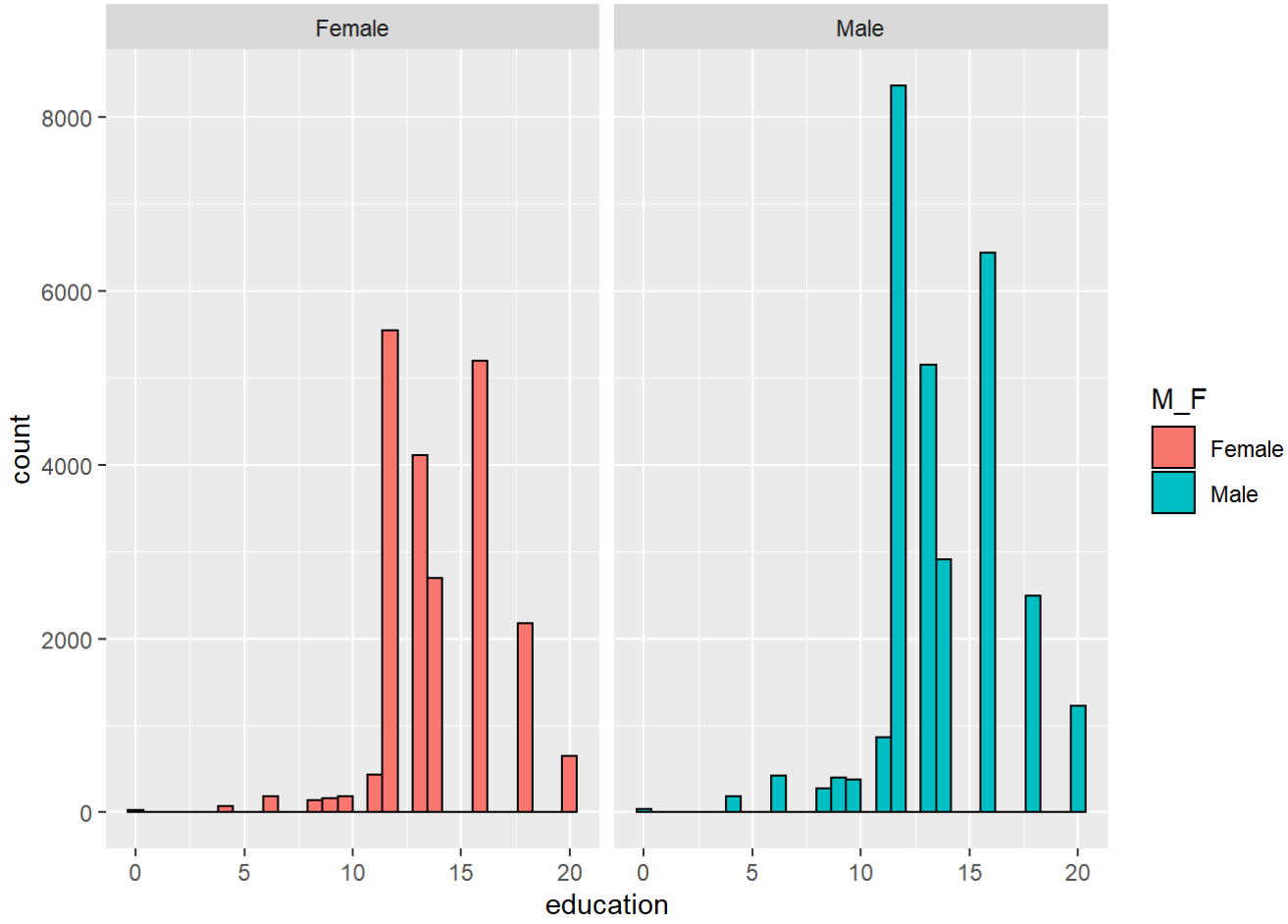
```
ggplot(data=economics,aes(x=age,fill=M_F))+geom_histogram(color="black")+facet_wrap(~M_F)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



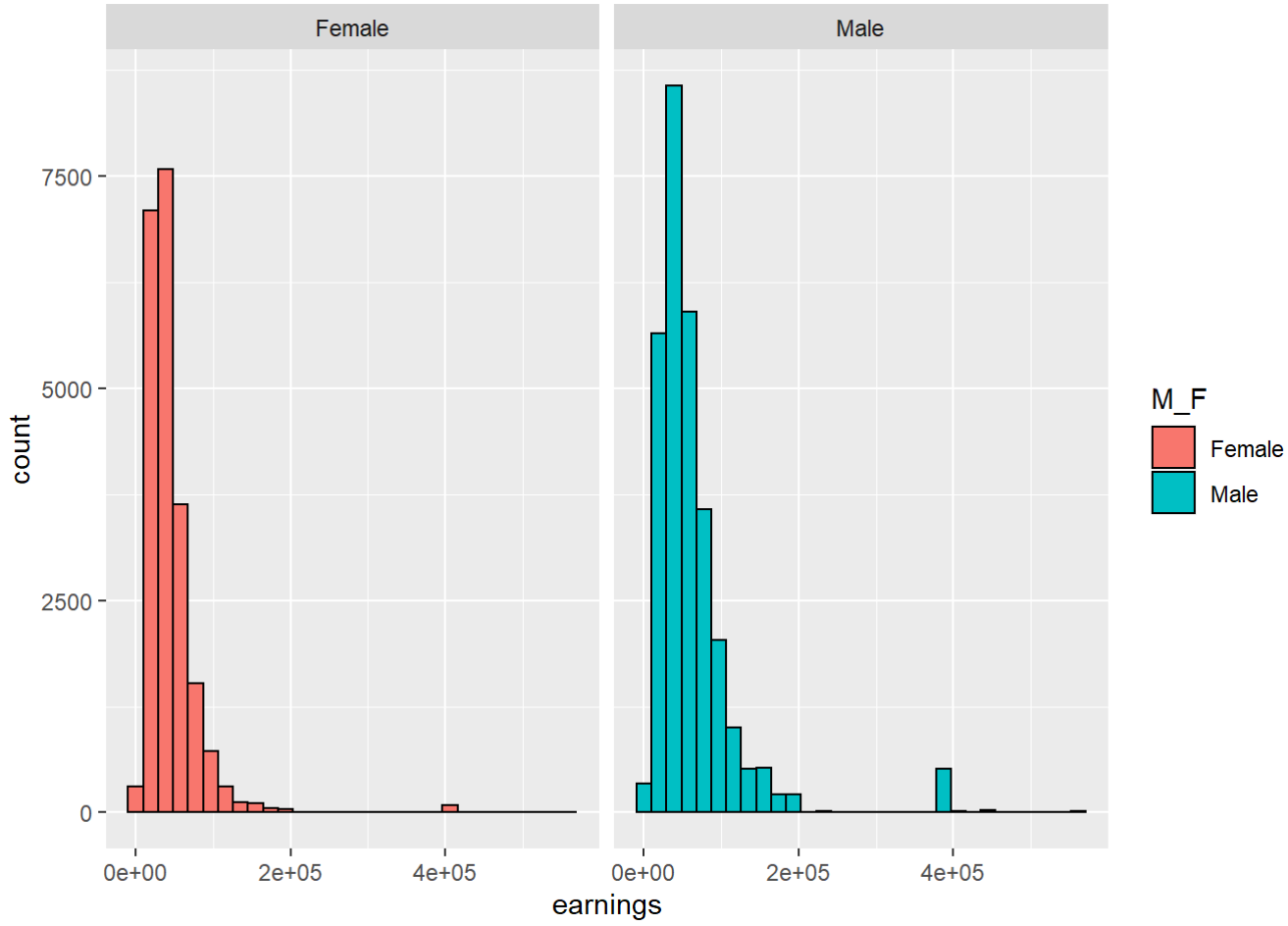

```
ggplot(data=economics,aes(x=education,fill=M_F))+geom_histogram(color="black")+facet_wrap(~M_F)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data=economics,aes(x=earnings,fill=M_F))+geom_histogram(color="black")+facet_wrap(~M_
F)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
model<-lm(formula=earnings~age+education+marital+female,data=economics)

anova(model)
```

```
## Analysis of Variance Table
##
## Response: earnings
##              Df     Sum Sq    Mean Sq F value    Pr(>F)
## age           1 3.1052e+12 3.1052e+12 1440.31 < 2.2e-16 ***
## education     1 1.9655e+13 1.9655e+13 9117.03 < 2.2e-16 ***
## marital       1 1.4177e+12 1.4177e+12  657.59 < 2.2e-16 ***
## female        1 4.8153e+12 4.8153e+12 2233.54 < 2.2e-16 ***
## Residuals 50737 1.0938e+14 2.1559e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lmtest::bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1557.3, df = 4, p-value < 2.2e-16
```