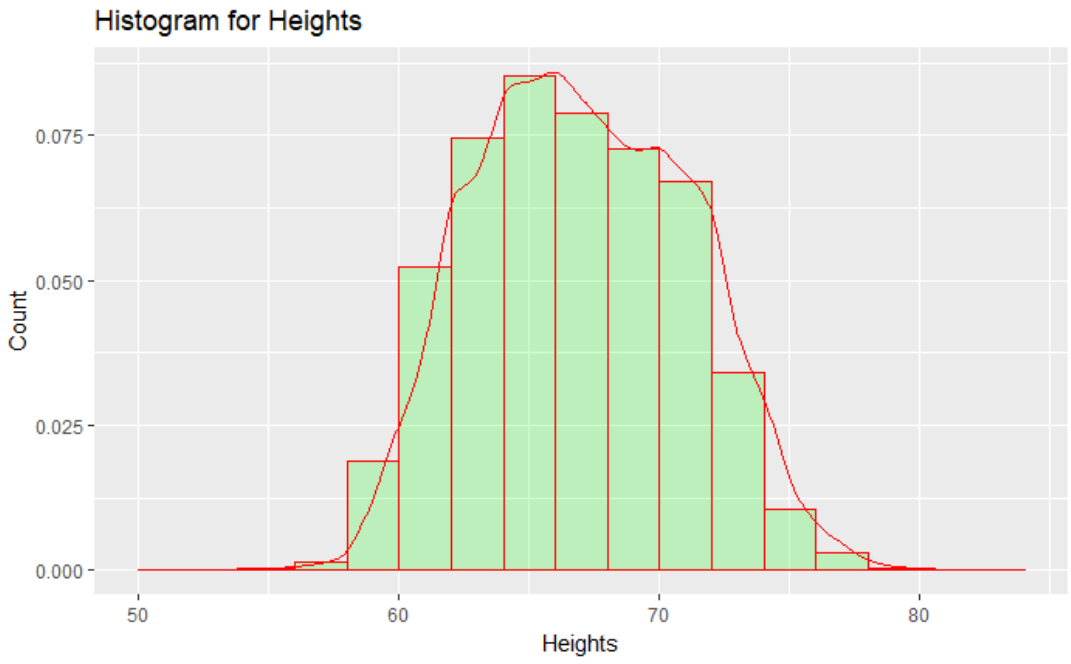
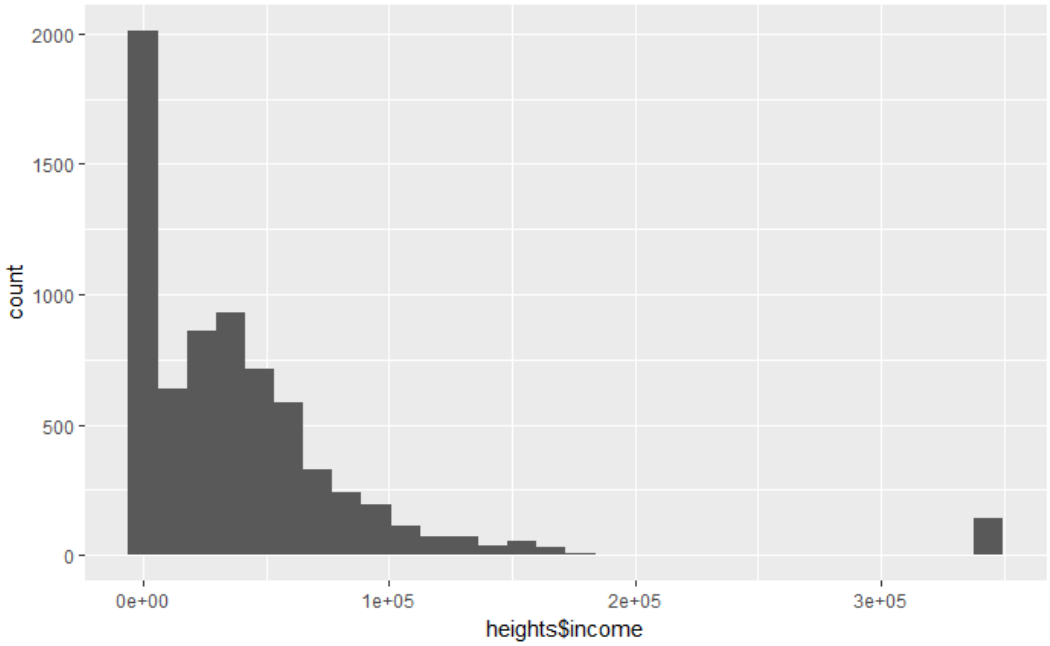
The factors that influence incomes

Peng Tian

Everyone must know that the income is related with the years of schooling, races and sex. But have you ever thought about the relationship between your height and your annual income. The BLS National Longitudinal Surveys (NLS) tracked the income, education, and life circumstances of a large cohort of Americans across several decades. A small sample of the whole data is included in the package modelr. I will use this sample data to try to develop the model that detect the factors influencing incomes.

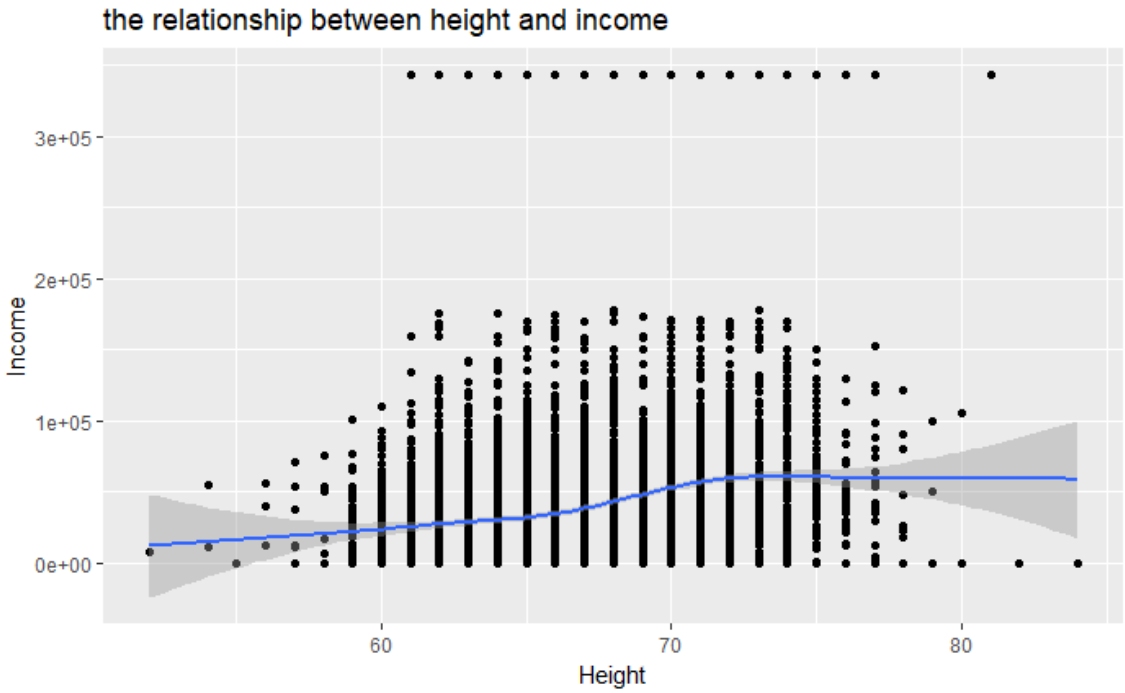
Firstly, we can plot the histogram of height and income. Histogram plot is an estimate of the probability distribution of a continuous variable, so we can get the basic distribution information from our plots about height and income.

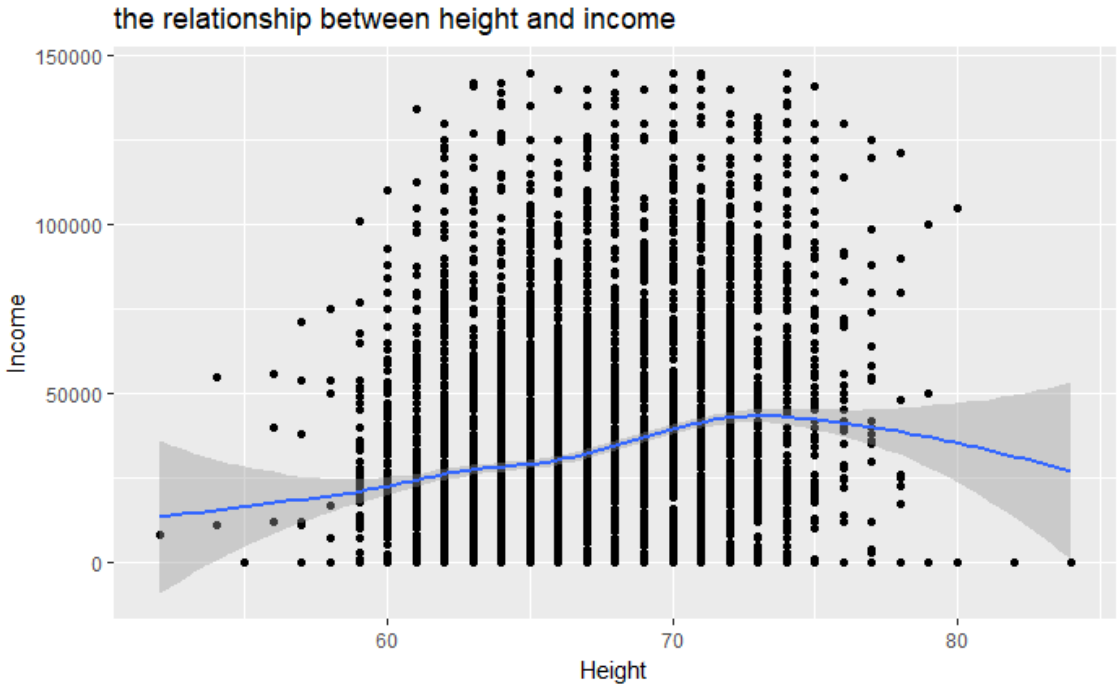




The income axis has a big span because very few people have very high income. That makes the histogram of income weird.

Then we can visualize the relationship between height and income. I added the geom_smooth function, which can aid the eye in seeing patterns between these two variables. Can you detect any relationship between these two variables? How do you describe the relationship? There is a little slop in the plot. That means the higher value of the heights, the more income you can earn. We can inspect the od straight line one the top of the plot. This is because the Bureau of Labor Statistics removed the top 2% of income values and replaced them with the mean value of the top 2% of values. So I am going to throw out these data since it is not the original income and plot the data again.

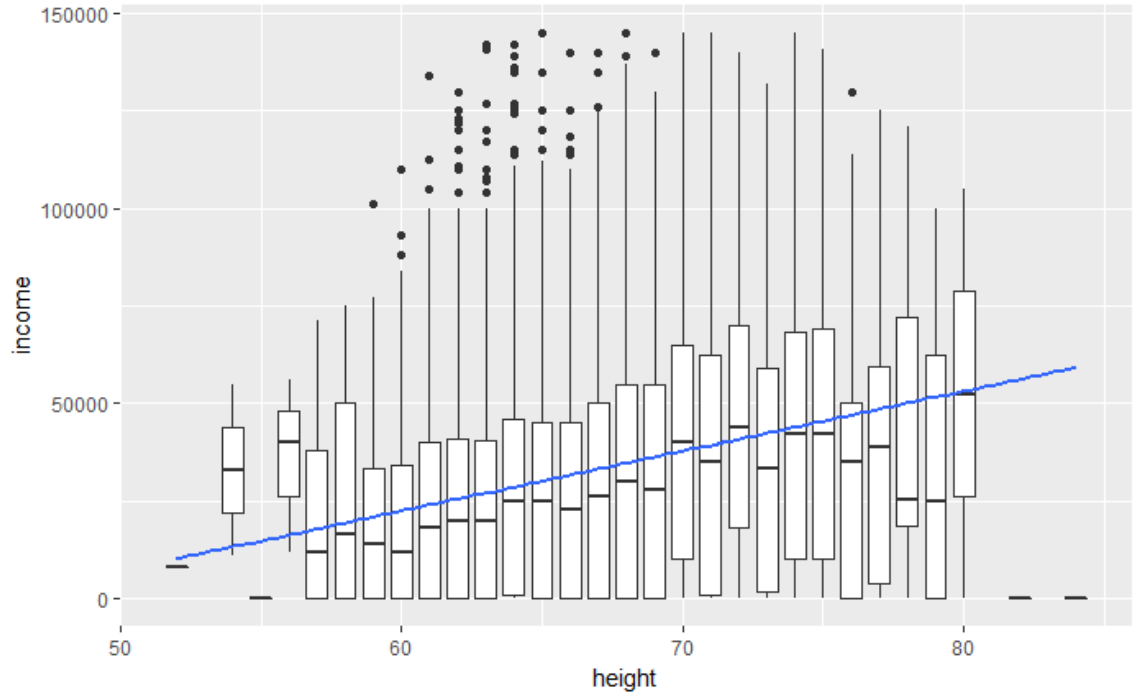


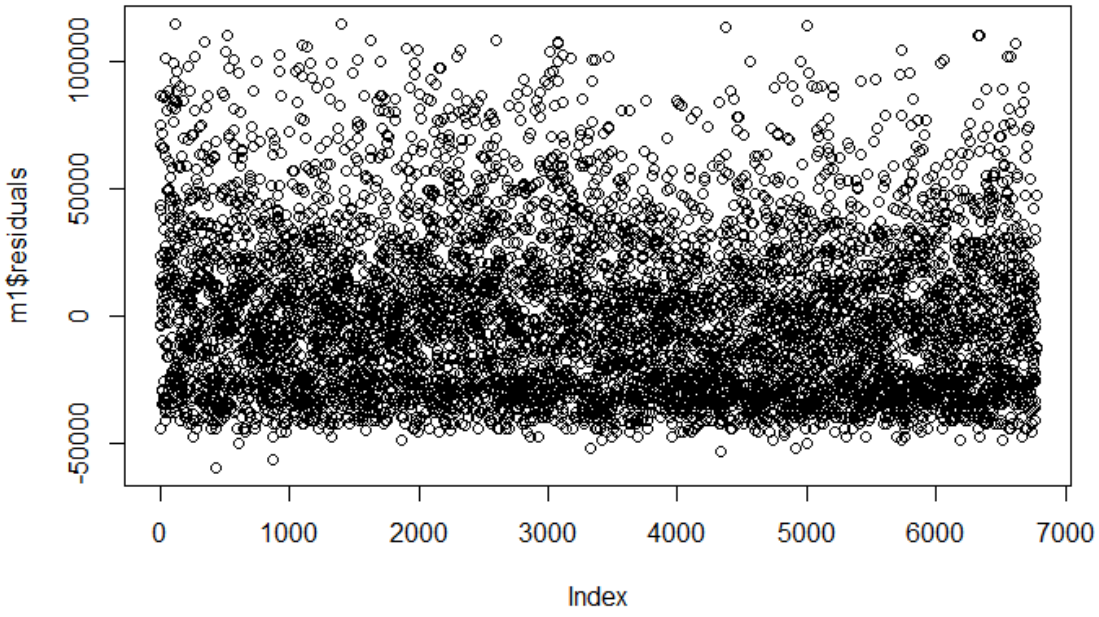


Next step I will try to develop the model. Firstly, I am going to start with the simple linear model.

```
m1 <- lm(income ~ height, data = heights)
```

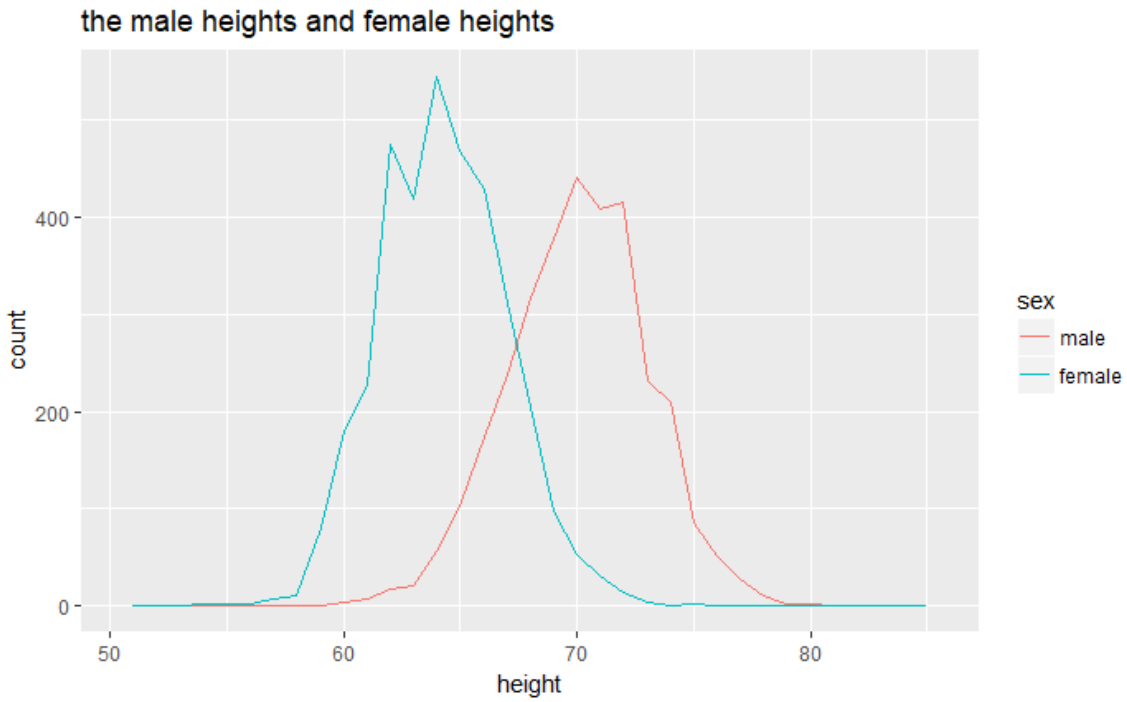
So far my model is Income = -69597 + 1535*Height. We can get the coeffients from the summary of the model. We can visualize our result as well. The model is quite simple so far and the residuals don't behave like white noise.

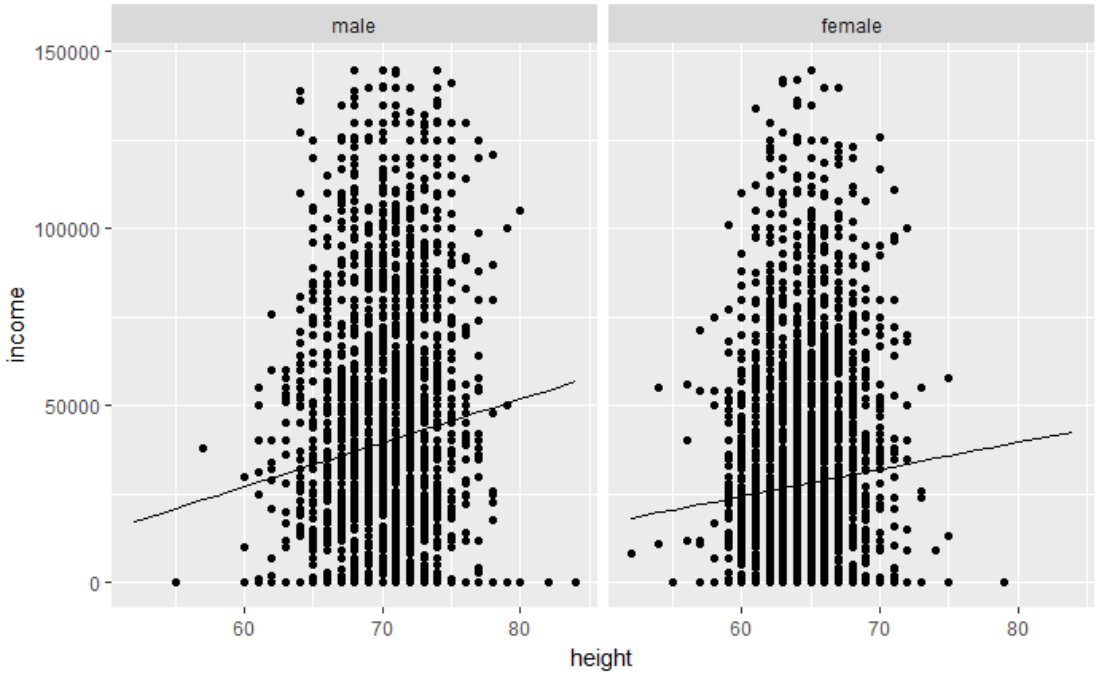




We can add more variables into the model. Firslty, we check the difference between the male heights and femal heights. We try to add the sex and education into the model.

```
m3 <- lm(income ~ height * sex * education, data = heights)
```





We can check the model adequate later. So far the model is getting more complicated. The data we used is just the sample from package modelr. Now, I am going to download another data online.

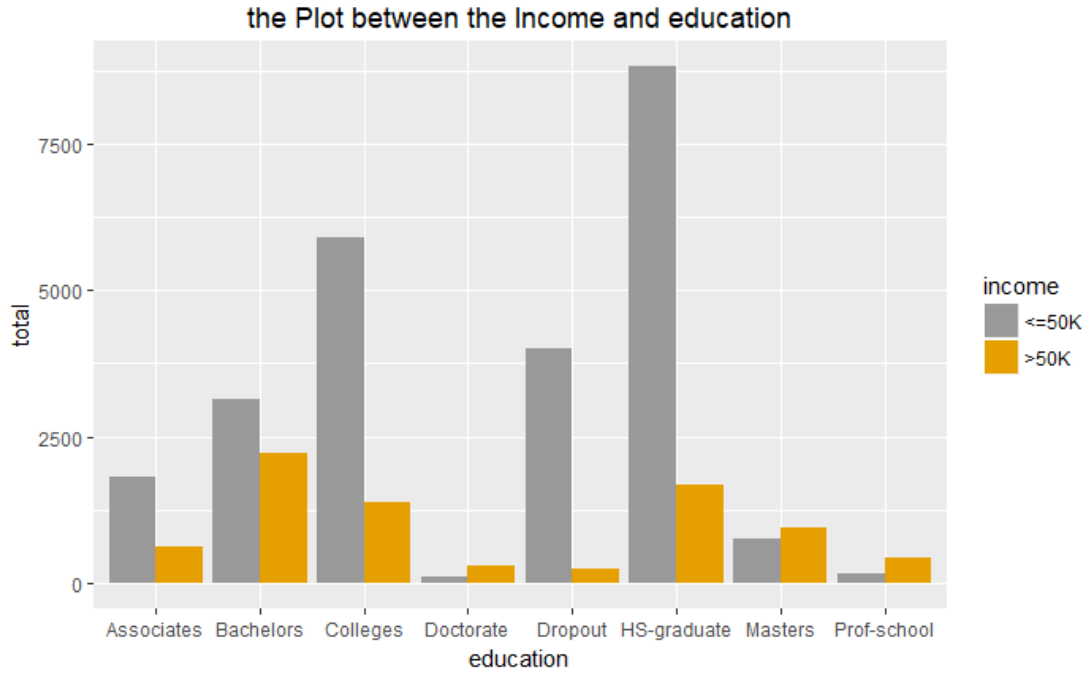
I downloaded the data from http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data, which can give us more information about the incomes.

Next step, I do some data-cleanup. We combine “Preschool”, “1st-4th”, “5th-6th”, “7th-8th”, “9th”, “10th”, “11th” and “12th” groups to “Dropout” group, “Assoc-acdm” and “Assoc-voc” groups to

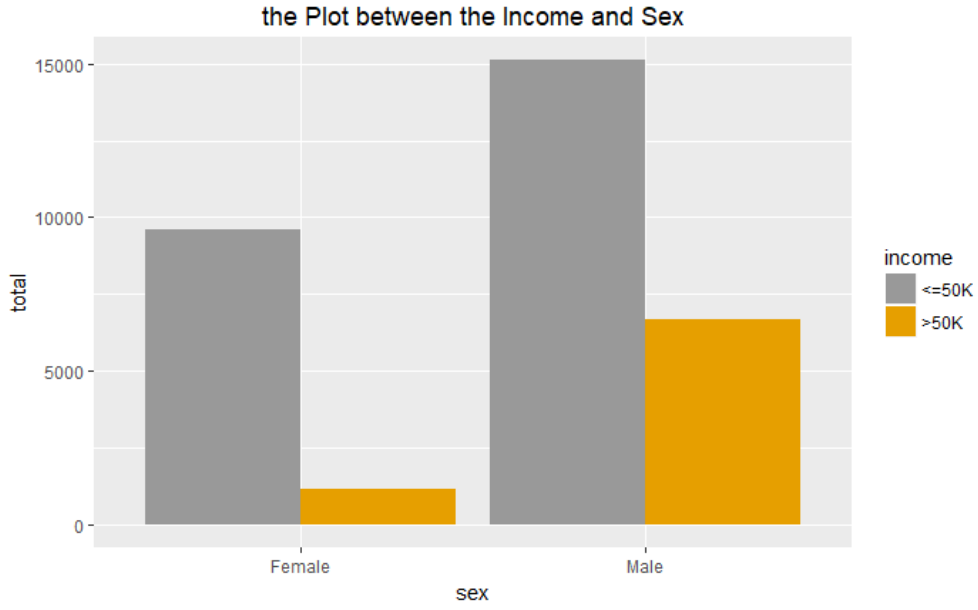
“Associates” group, “HS-grad” and “Some-college” groups to “HS-Graduate” group. Here is the summary after clean-up:

```
Associates    Bachelors     Colleges    Doctorate      Dropout HS-graduate
Masters
      2449         5355         7291          413         4253        10501
1723
Prof-school
       576
```

We can visualize the relationship between income and education



Similarly, we can group the data by sex, and plot the income by sex.



The income by age is more complicated because the age is numerical. So the plot will give the income distribution in different ages.

There are so many factors of influencing income level. I used the two different data source to reveal a part of the relations in this project. These plot can give us some idea about the factors influencing the income level.