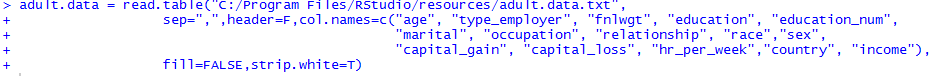
**Census Income Data Set**

The goal is to test your understanding of data summarization, exploration, and modeling.

* Census Income Data Set in the UCI Machine Learning Repository<https://archive.ics.uci.edu/ml/datasets/Census+Income>  
  adult.data has the actual data and the remaining files include descriptions of the data.

Using the read.table() function, I read “adult.data” and changed the variables’ name to make more readable.

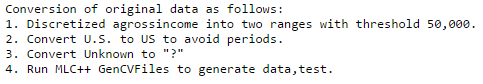


I could got 32,561 rows of data like the following picture.

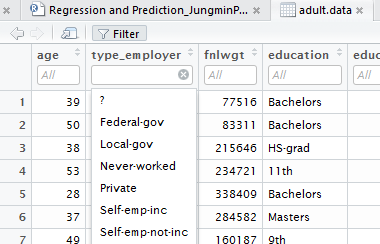


* Data exploration
  + Missing and/or invalid values

In data analysis, it is often necessary to clean a dataset before analyzing it. Data that contain missing and invalid values must be removed or replaced by some estimate so as to not bias any comparisons we make between the methods.  According to the UCI Machine Learning Repository, they converted unknown value to “?”. In other words, these values are invalid values, and I removed these data.

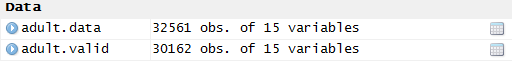


By “Filter” function, we can quickly check which column has “?”. As a result, type\_employer, occupation, and country have “?” which is invalid values.



Using the na.omit() function, I [subset](http://datascienceplus.com/subset-which-ifelse-functions/) the original dataset selecting the relevant data only. Therefore, only 30,162 rows of data left.





* + Summary statistics

It has a combination of continuous and categorical variables. The continuous variables range over the entire space permitted by the nature of each particular variable, and the categorical variable have 2 to more than 30 categories.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type** | **Permitted Value** | **Include in the Analysis** |
| Age | Continuous | / | Y |
| Education(Number of Years) | Continuous | / | N |
| Capital Gain | Continuous | / | Y |
| Capital Loss | Continuous | / | Y |
| Working Hours per Week | Continuous | / | Y |
| Final Weight | Continuous | / | N |
| Work Class | Categorical | 8 Categorical | Y |
| Education | Categorical | 16 Categorical | Y |
| Marital Status | Categorical | 7 Categorical | Y |
| Occupation | Categorical | 14 Categorical | Y |
| Relationship | Categorical | 6 Categorical | Y |
| Race | Categorical | 5 Categorical | Y |
| Sex | Categorical | 2 Categorical | Y |
| Native Country | Categorical | 41 Categorical | Y |

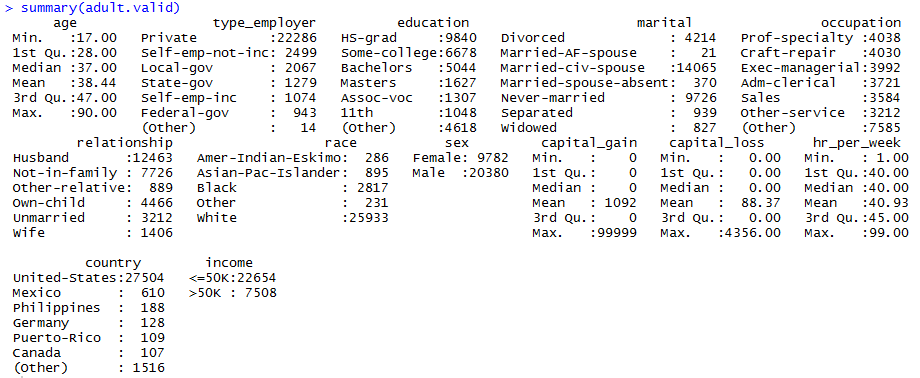
Final Weight is assumed to be used in the sampling process to construct the dataset and not to be related to income classification, so I excluded this variable. Also, I excluded the variable education (Number of Years) to avoid duplication with the categorical variable education.



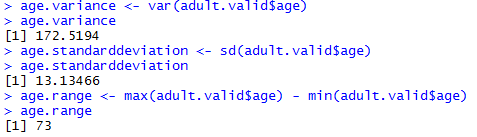
If you see the table, you can see that there are not education\_num and fnlwgt variable.

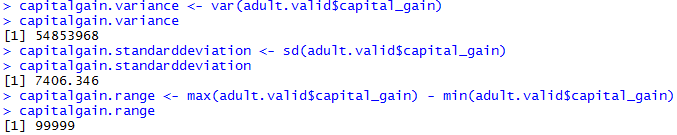


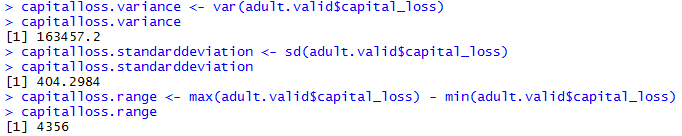
To see summary data such as mean, median, and quartile, I used the summary() function. As you can see from the following chart, if the variable is categorical variable, it is impossible to calculate mean, median, and quartile. Instead, we can know how many data each category has.

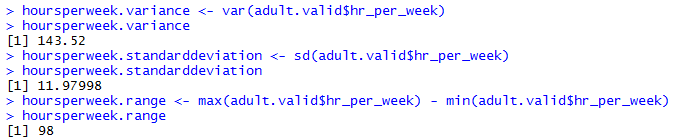


For variance, standard deviation, and range, I needed to calculate separately.

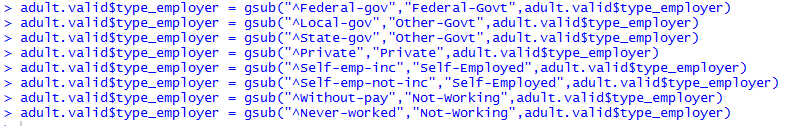


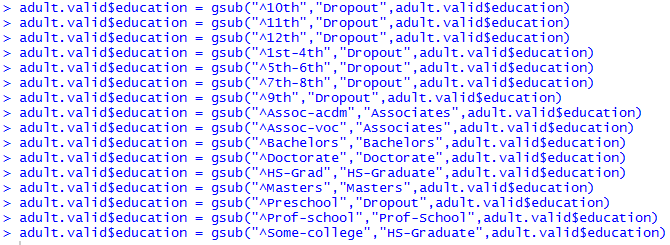


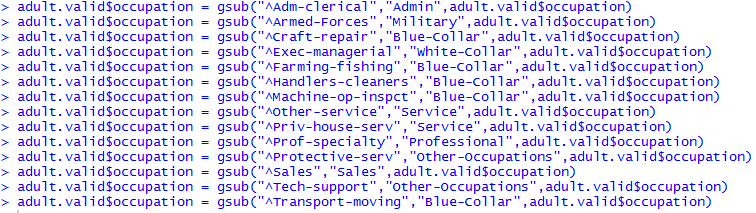




However, as you saw the chart before, some variables have many categories. So, by combining similar group, I tried to reduce the number of categories. This method helps to simplify the calculation, as well as to make the output more readable. I changed variables; type\_employeer, education, occupation, country.





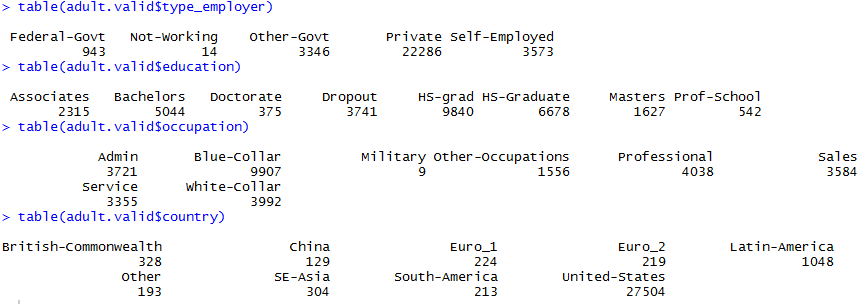




You can see how many variables changed by the following chart.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Type** | **Previous Permitted Value** | **Current Permitted Value** |
| Work Class | Categorical | 8 Categorical | 5 Categorical |
| Education | Categorical | 16 Categorical | 8 Categorical |
| Occupation | Categorical | 14 Categorical | 8 Categorical |
| Native Country | Categorical | 41 Categorical | 9 Categorical |

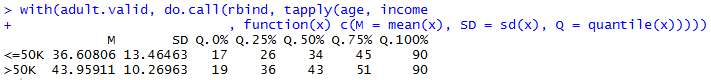
After I reduced the number of category, I checked again how many data each category has.



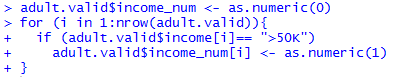
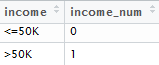
* Conditional summaries -- multivariate analysis & Distributions & Correlations

Our goal is to identify the characteristics that may distinguish individuals whose income is greater than $50K from those whose income is less than 50K. So, this part is important to see the pattern. I compared every factors with income.

As you see from the chart, on average people who earn over $50,000 is older than people who earn under $50,000.



To calculate correlation, I added new column called income\_num to change income data to numeric data.

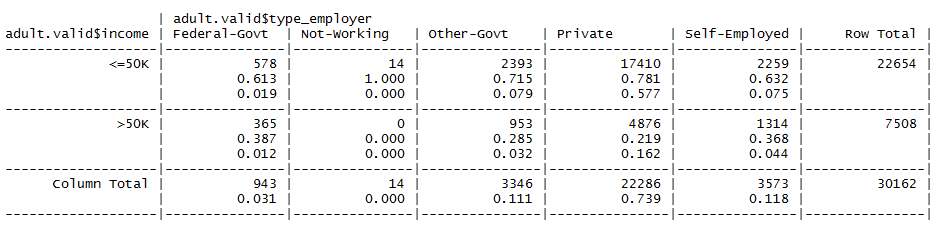
 

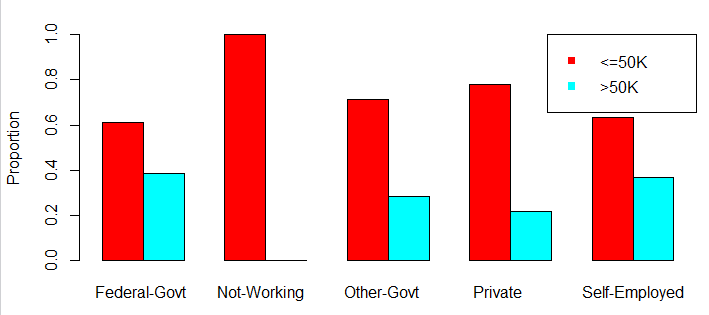
Finally, I computed correlation, and the outcome is 0.24 which means there is a weak positive relationship between age and income.

If independent variable is categorical, speaking of correlation is somehow abusive, since correlation is defined by means and categorical variables do not have mean. Speaking of association is better. To analyze the relationship between categorical variables, I used chi-square test instead of calculating correlation.

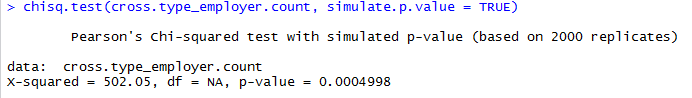
As you can see from the chart and graph, compared with other people, there is a strong possibility that self-employed people can earn money over $50,000 per year.





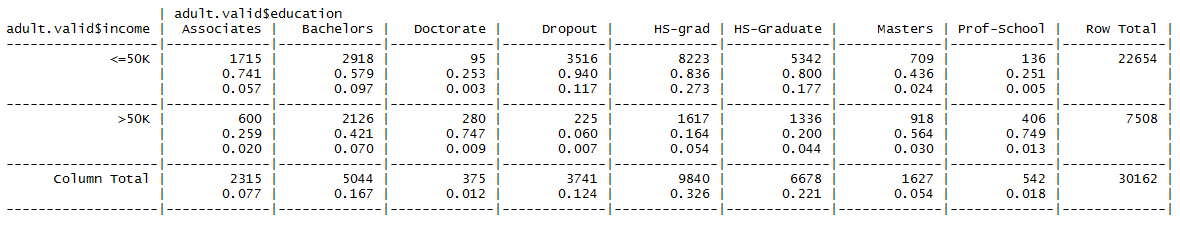


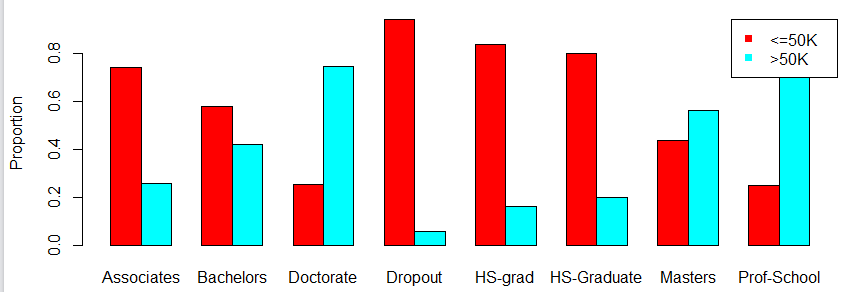
Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between work class and income.



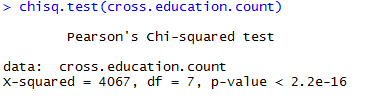
As you can see from the chart and graph, compared with other people, there is a strong possibility that someone who earned the doctorate and professional school degree can earn money over $50,000 per year.





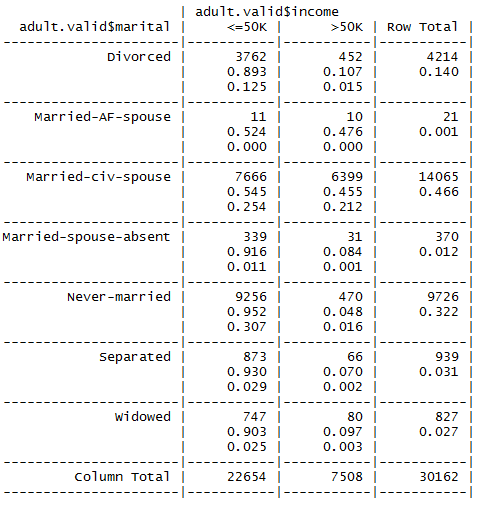


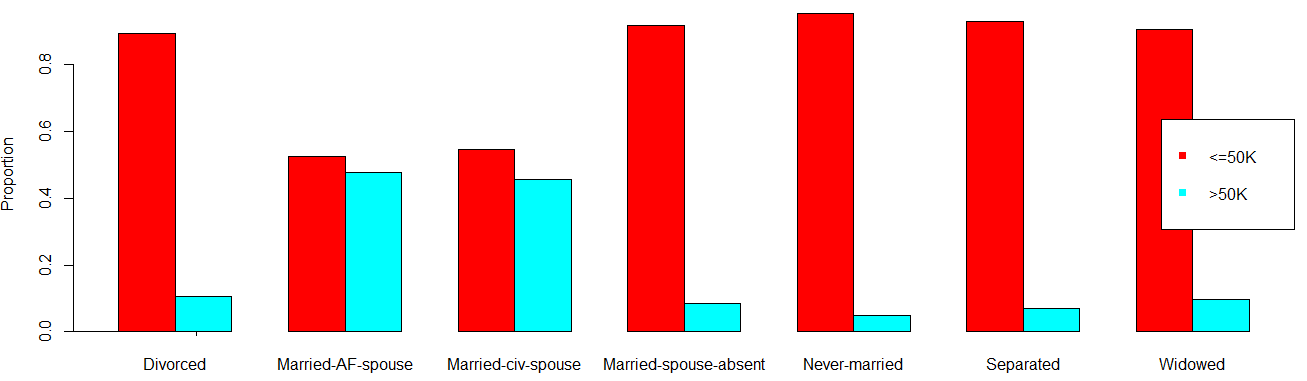
Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between education and income.



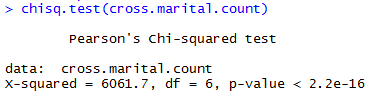
As you can see from the chart and graph, compared with other people, there is a strong possibility that someone who is married and is living together can earn money over $50,000 per year.





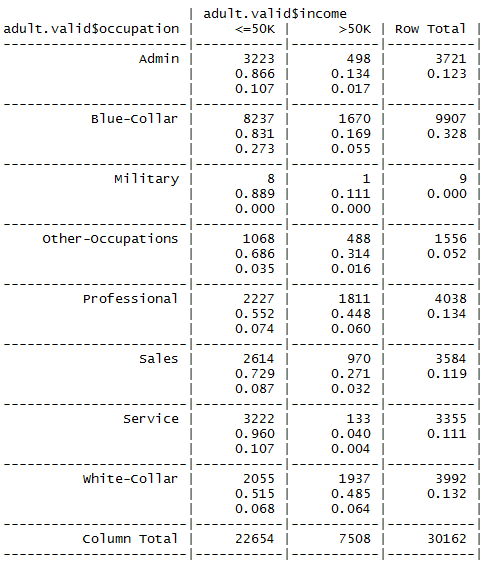


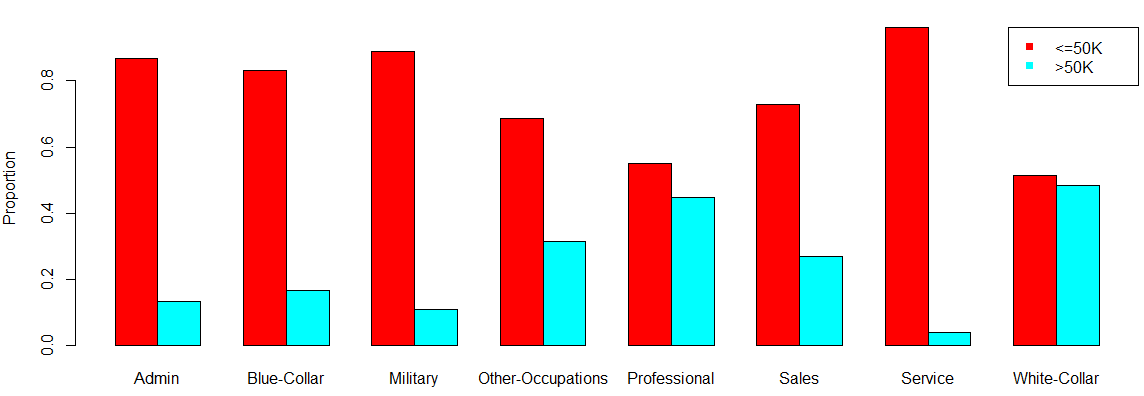
Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between marital and income.



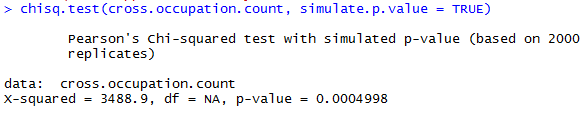
As you can see from the chart and graph, compared with other people, there is a strong possibility that someone who has white collar or professional occupation can earn money over $50,000 per year.





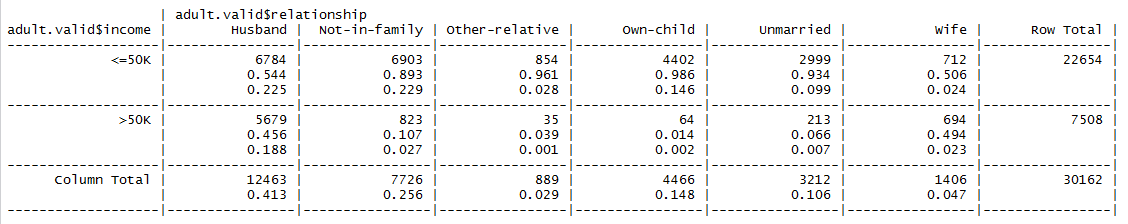


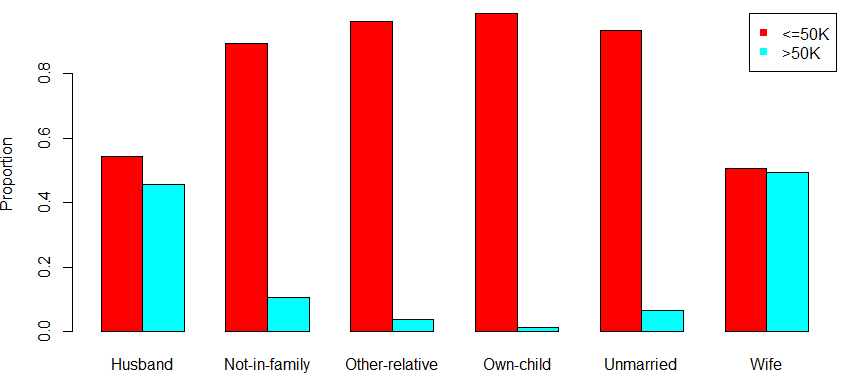
Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between occupation and income.



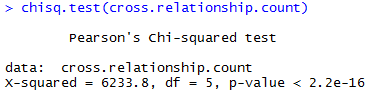
As you can see from the chart and graph, compared with other people, there is a strong possibility that someone who has their spouse can earn money over $50,000 per year.





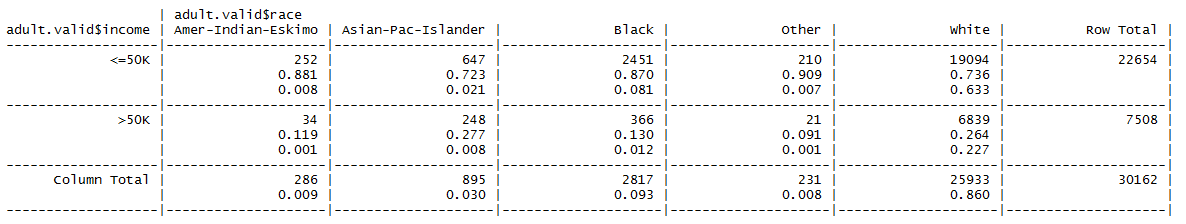


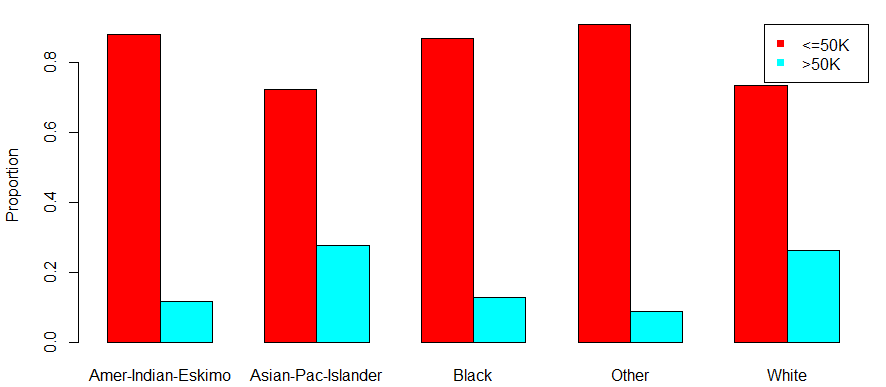
Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between relationship and income.



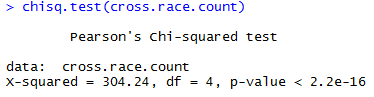
As you can see from the chart and graph, compared with other people, there is a strong possibility that someone whose race is Asian or white can earn money over $50,000 per year.





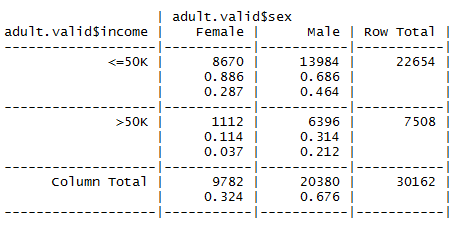


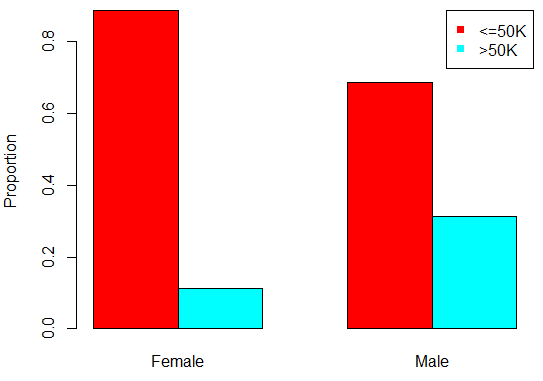
Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between race and income.



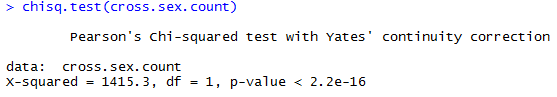
As you can see from the chart and graph, compared with female, there is a strong possibility that male can earn money over $50,000 per year.

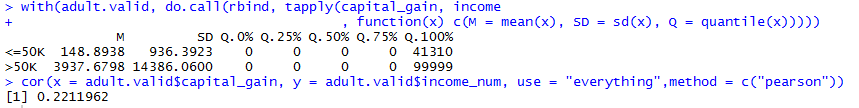


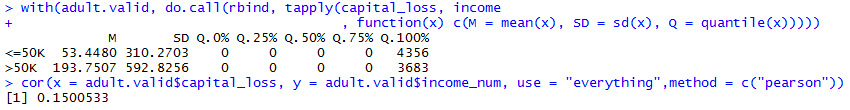


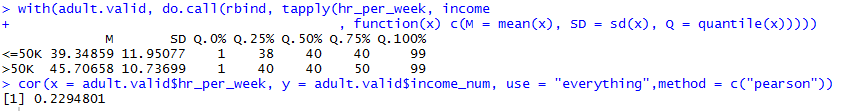


Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between sex and income.

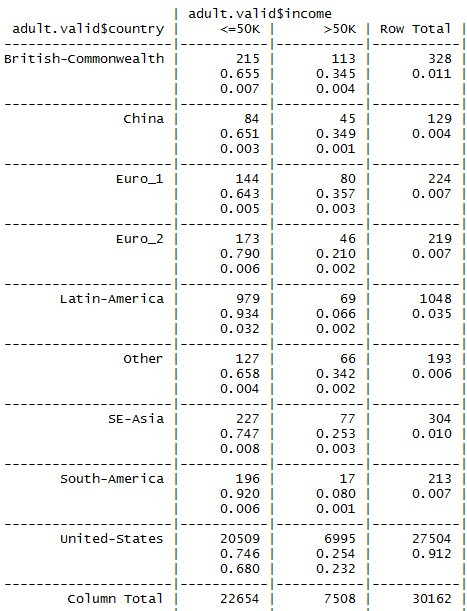


As you can see from the chart, on average people who earn over $50,000 has more capital gain than people who earn under $50,000. I computed correlation, and the outcome is 0.22 which means there is a weak positive relationship between capital gain and income.

As you can see from the chart, on average people who earn over $50,000 has more capital loss than people who earn under $50,000. I computed correlation, and the outcome is 0.15 which means there is a weak positive relationship between capital loss and income.

As you can see from the chart, on average people who earn over $50,000 has more capital loss than people who earn under $50,000. I computed correlation, and the outcome is 0.15 which means there is a weak positive relationship between capital loss and income.

As you can see from the chart and graph, compared with other country, there is a strong possibility that people from China, Euro1, or British common wealth can earn money over $50,000 per year.



Finally, by Chi-square test, p-value is smaller than 0.05, so I can tell that there is association between country and income.

