

Big Data Analysis and Investigation on Influential Factors on United States Public Incomes under Information Digitization

Bohua Cheng^{1,*}, Bowei Cheng¹

¹ Xi'an Tie Yi High School, Xi'an, China

*Corresponding author: 921680665@qq.com

Abstract. There are multiple factors influencing the income level of the public, such as education, gender, race. To investigate this issue, an adult census income dataset was analyzed. The data was collected by the United States Census Bureau [1]. The research primarily focuses on the influence of years of education, age, race, and sex on the public income. The result from statistical analysis shows strong relationships between income and age, race, sex, and years of education. Age affects income; senior people tend to get higher income than junior people. Sex affects income; men tend to get higher income than women. Race also affects income; white and Asian tend to get higher income than other people in United States.

Keywords: data science, statistics, R language programming.

1. Introduction

1.1 Motivation

There are various factors that affect income, and we analyzed four essentials of them: age, sex, race, and education. For age, senior people have more experience than junior and therefore tend to earn more. In the article “Average Income Tends to Rise with Age” [2], author Erica York mentions incomes tend to rise as taxpayers grow older and more experience. In the company, bosses are more willing to pay competent people than newcomers.

For race, because of unique history in US, African American and other minority do not have the same right to compete with another person in the public. Like Asian could get higher income. For the race, the difference has been weakened during the modern life. However, the problem behind the situation was originated from the last century [3]. Because the native American and black people were regarded as slaves, they weren't allowed to have the education. From that time, the unbalance among races exaggerated. Until the emancipation proclamation, some slaves could have their rights to enable their children to have education. Nevertheless, there were still unbalance for black and native American people [4]. For example, they were not allowed to take the bus at the same carriage with white people; black people were given less than the white people, Like Rosa Parks event. Although they had the right to have education, they didn't have the money to support them to have education because of these reasons. For a family, if they didn't have money, they would make their children work directly. What's more, although they would enter the school of black people, the education resources made them have less competitive than others. Also, this will decrease the chance for them to receive higher education. For education, if a person has a high degree, it means that the person known a lot about professional knowledge.

1.2 Dataset Description

This data was collected from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics) [5]. This data was used to compare people's income. There are 48842 instances in the dataset, and the data was collected on 1996-05-01.

When collecting the data, researcher used 3 controls:

- 1.single cell estimates of the 16 + population in each state.
- 2.Hispanic descent was controlled by age and sex.
3. Control by race, age and gender.

In the original dataset, there are 15 columns: “age”, “work class”, “final weight”, “education”, “education years”, “marital status”, “occupation”, “relationship”, “race”, “sex”, “capital gain”, “capital loss”, “hours per week”, “native country” and “income”. “final weight” represents specific socio-economic characteristics of the population. People who have the similar demographic characteristics have the similar weights. However, in the modified dataset, we just keep age, education, education years, occupation, race, sex, hours per week, native country, generation and income. We have deleted some columns from the dataset like capital gain, capital loss, work class, final weight and marital status. Since we just interested in four factors in the research, we assume that other factors won’t affect the result. We also add some columns in the Table 1.

Table 1. Dataset description

Generation	Under 30, 31 to 50, 51 to 70, Above 70
Income-Dummy	1 if Income is greater than 50K, 0 if income is less than 50K
Sex-Dummy	1 if Male, 0 if Female
US Born-Dummy	1 if born in the US, 0 if not
College-Dummy	1 if the person has a bachelors degree or higher, 0 if not
White-Dummy	1 if a person is white, 0 if otherwise
Private-Dummy	1 if the person’s workclass is private, 0 if otherwise
Gov-Dummy	1 if someone is a federal government worker, 0 if otherwise
Under30-Dummy	1 if person is under 30, 0 if not
30to50-Dummy	1 if person is between 30 and 50, 0 if not
50to70-Dummy	1 if person is between 50 and 70, 0 if not
Over70-Dummy	1 if person is over 70, 0 if not

These columns help us build pictures more easily in R.

2. Research methods

We used R Studio to analyze the dataset. We plotted box plot, bar chart, etc. We showed the relationships between each factor and yearly income. In order to show the result directly, the level of income is divided into more than 50k and less than or equal to 50k. The statistical methods to analyze the dataset are listed below.

2.1 Box Plot

Box plot is a method for graphically depicting groups of numerical data through their quartiles which are the values that divide a list of number into quarters. To make a box plot, a central box is drawn from the first quartile to the third quartile. A line in the box marks the median. Lines (called whiskers) extend from the box out to the smallest and largest observations that are not outliers. Then, outliers are marked with a special symbol such as an asterisk. Here is an example displaying the 51 observations for gross state product (in billions of dollars) from textbook Introduction to Statistics and Data Analysis [6]:

Gross Product: 110 24 134 62 1119 142 142 34 54 419 254 40 31 426 174 85 77 107 129 32 165 239 295 161 62 163 20 52 63 41 319 48 707 236 17 341 82 105 364 30 100 21 160 646 60 16 231 193 40 158 18.

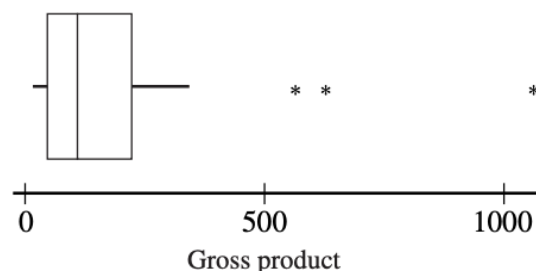


Fig. 1 boxplot of gross state product

2.2 Histogram

A histogram is an approximate representation of the distribution of numerical data. Histogram determines the dispersion and bias of the dataset. Histogram allows people to see the size of the dataset immediately. Also, it displays the data distribution so people can identify whether the data is distributed normally or skewed. Frequency in histogram is represented by a rectangle centered over the corresponding value and the area of the rectangle is proportional to the corresponding frequency. Here is an example displaying the Math SAT score in 2005 from textbook Introduction to Statistics and Data Analysis [6]:

Table 2. Relative Frequency Distribution of Math SAT Score

Score	Frequency	Relative Frequency
200–249	33,841	.02 $\leftarrow \frac{33,841}{1,475,623}$
250–299	67,117	.05
300–349	124,032	.08
350–399	166,446	.11
400–449	210,449	.14
450–499	255,904	.17
500–549	225,335	.15
550–599	188,780	.13
600–649	111,350	.08
650–699	58,762	.04
700–749	22,270	.02
750–800	11,337	.01
Total	1,475,623	1.00

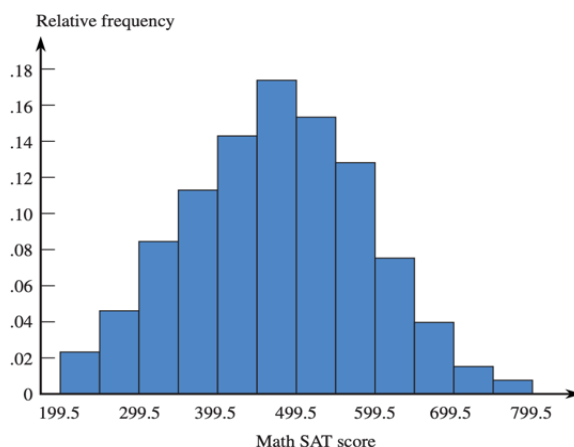


Fig. 2 Histogram of Relative Frequency Distribution of Math SAT Score

2.3 Barchart

Bar graphs represent each category as a bar. The bar heights show the category counts or percentages. Bar graphs are easier to make than pie charts and are also easier to read. Categorical data could be summarized in a frequency distribution and displayed graphically using a bar chart. Here is an example displaying identity theft rates provided in article “Fraud, Identity Theft Afflict Consumers” from textbook Introduction to Statistics and Data Analysis [6]:

Table 3. Identity theft rates

State	Identity Theft (Cases per 100,000 Population)
California	122.1
Arizona	142.5
Oregon	87.8

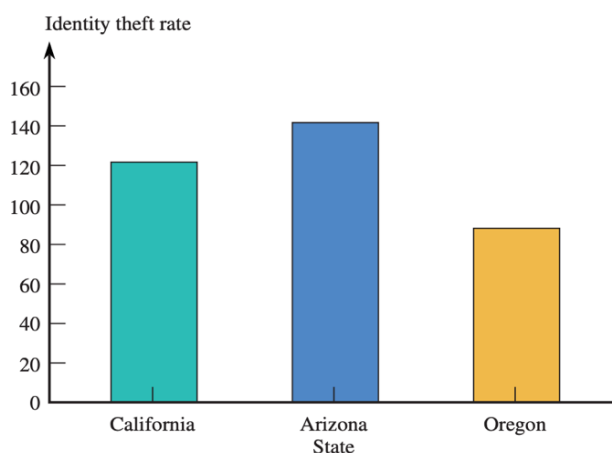


Fig. 3 Bar chart of identity theft rates

2.4 Mean

The mean is the usual numerical average, calculated as the sum of the data values divided by the number of values. It is nearly universal to represent the mean of a sample with the symbol $\bar{x}$, read as “x-bar.” If the n observations are $X_1, X_2, \dots, X_n$, their mean is

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{X_1 + X_2 + \dots + X_n}{n} \quad (1)$$

The median of a sample is the midpoint of a distribution, the number such that about half observations are smaller and about half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median is the center observations in the ordered list.

2.5 Quartile

The lower quartile (Q_1) is the median of the data values that are located below the median.

The upper quartile (Q_3) is the median of the data values that are located above the median.

These values are called quartiles because, along with the median and the extremes, they divide the ordered data approximately into quarters.

FORMULA Finding Quartiles

Q_1 = lower quartile = median of lower half of the ordered data values

Q_3 = upper quartile = median of upper half of the ordered data values

3. Analysis

3.1 The relationship between income and age

Firstly, the relationship between different generations and income level was analyzed. The generations are categorized into three groups: from 30 to 50, from 50 to 70, and over 70.

From figure 4, the bar chart shows the income level distribution from age 30 to 50. The chart shows that only 30% people tend to have higher income level.

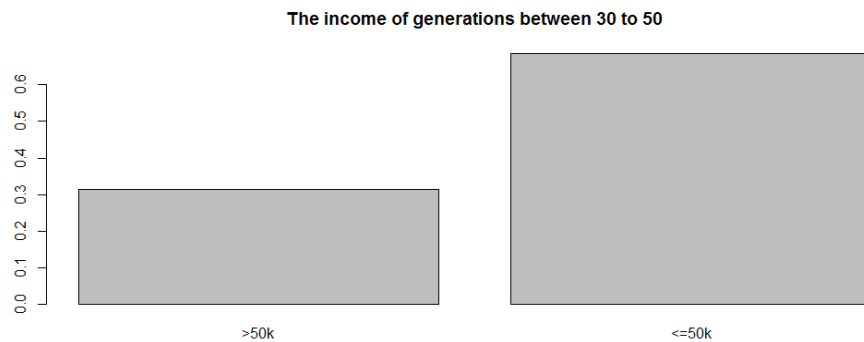


Fig. 4 The proportion of people having income over or less than 50k between 30 to 50

In figure 5, the bar chart shows the income level distribution form 50 to 70. The distribution is similar to the first one: number of people making more than 50k yearly income is less than number of people making less than 50k. However, the ratio of making over 50k income in this generation is higher than the generation with age 30 to 50.

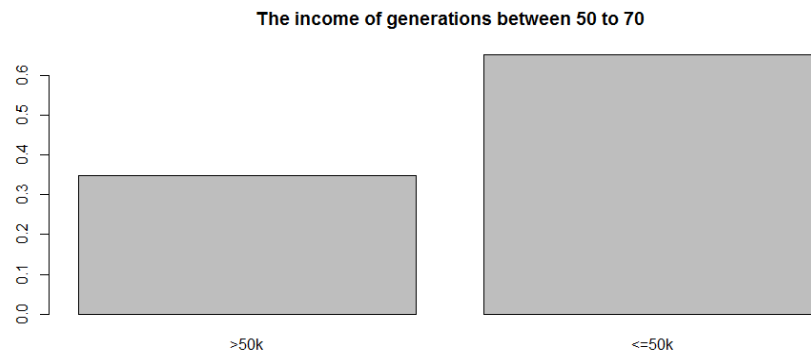


Fig. 5 The proportion of people having income over or less than 50k between 50 to 70

In figure 6, the bar chart shows the income level distribution from age over 70. The percentage of people who have higher income level is only 20%, which is much less than other two generations.

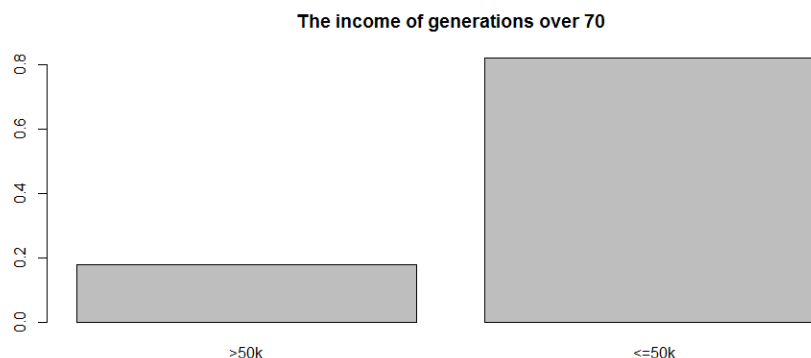


Fig. 6 The proportion of people having income over or less than 50k over 70

The results above roughly show people in age 30 to 50 to have higher income. In order to search for more concrete relationship, we investigate the relationship between age and income.

The figure 7 shows the overall relationship between age and income. It is obvious that people who have higher income tend to be order than people who have lower income level.

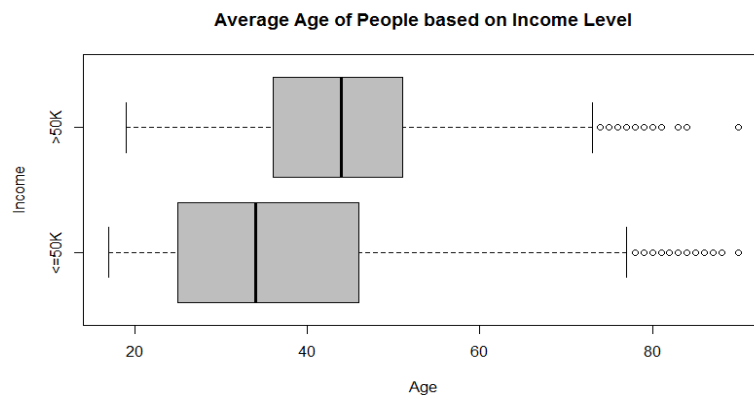


Fig. 7 Average Age of People based on Income Level

3.2 The Relationship between Race and Income

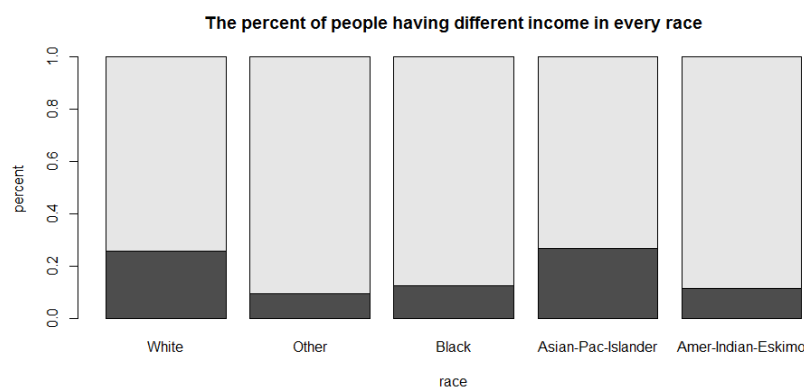


Fig. 8 Income Level Among All Races

For figure 8 the black part in each bar represents the proportion of people getting yearly income over 50k, while the white part represents the proportion of people getting yearly income less than 50k. The result shows that Asian-Pac Islander and white people can get more money than others

3.3 Sex and Income

Firstly, the percentage of higher income level in both genders is investigated. From Figures 9, the bar chart shows that the percentage of male in making more than 50k annually is more than 30%, which is almost the triple of the percentage of female in making more than 50k annually.

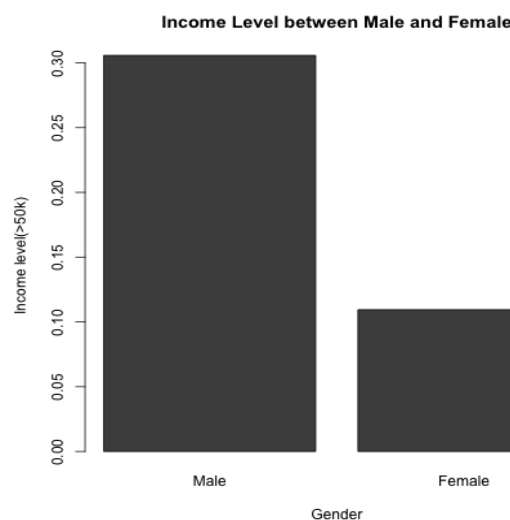


Fig. 9 Income Level between Male and Female

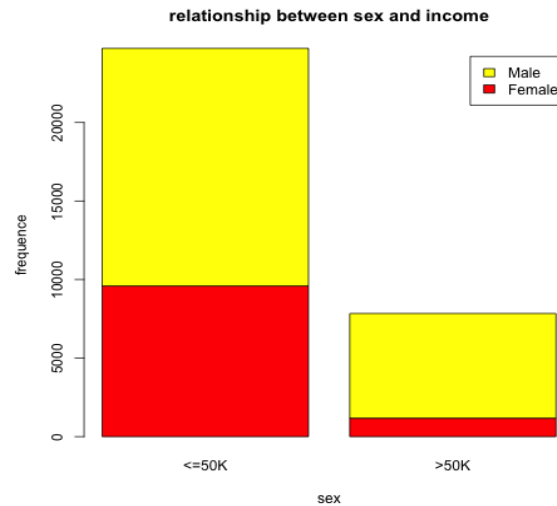


Fig. 10 Frequencies of each income level between Male and Female

The income levels in each gender are evaluated separately. From figure 10 the stacked bar chart shows the discrepancy between people making more than 50k and making equal or less than 50k. The discrepancy in Male is much less than female, which further proves higher percentage of more annual income in male than female.

3.4 Education and Income

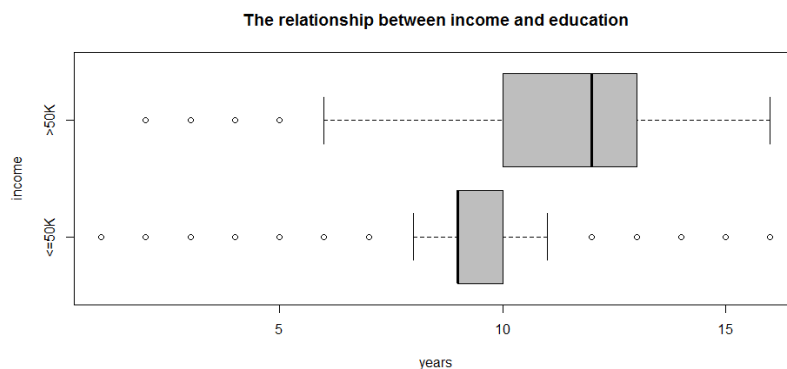


Fig. 11 The relationship between income and Education

The overall relationship between years of education is analyzed. From the figure 11, the box plot shows that people receive more education would generally have higher income.

3.5 Race and education

From figure 12, the distribution of years of education is displayed. It is shown that the Asian-Pac-Islander people tend to have average education years at 8-9 or 12.

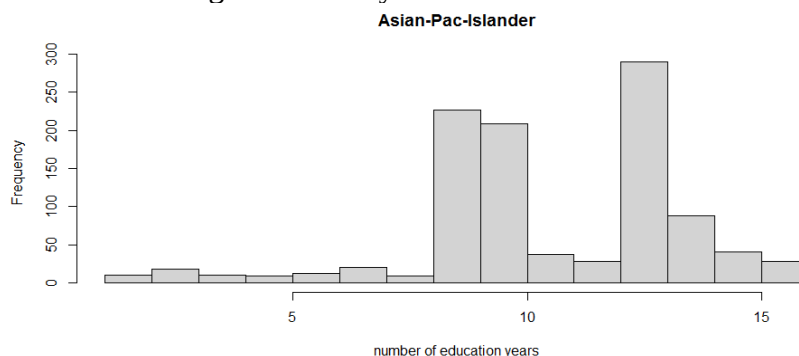


Fig. 12 Education Years Distribution for Asian

Figure 13 shows the distribution of number of education years of the African American. For black people, they are more likely to have 8-9 years of education. Fewer people have the chance to pursue a higher degree.

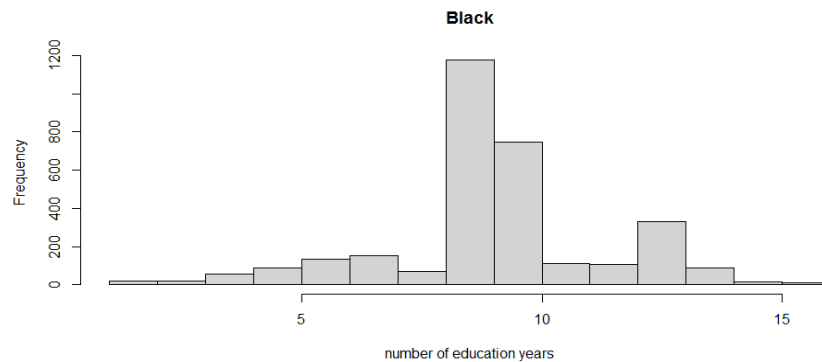


Fig. 13 Education Years Distribution for Black

Figure 14 shows the distribution of number of education years of the Whites. White people are more likely to have 8-9 or 12 years of education. Some people have the chance to pursue college-level degree or higher degree.

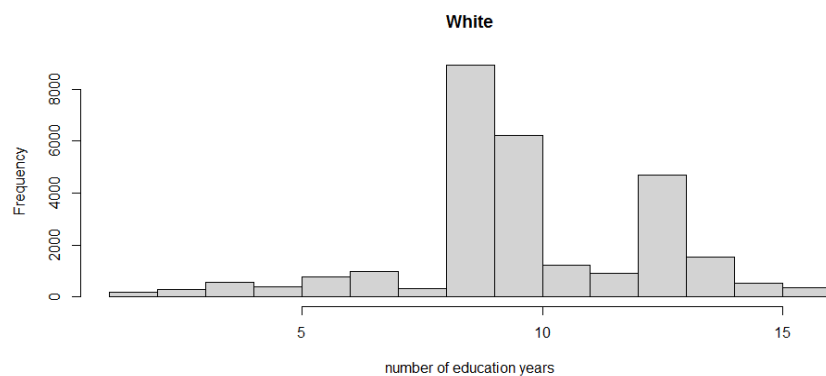


Fig. 14 Education Years Distribution for White

Figure 15 shows the distribution of number of education years of the Indian. The distribution for Amer-Indian-Eskimo people is similar as the distribution of the black. They are more likely to have 8-9 years of education.

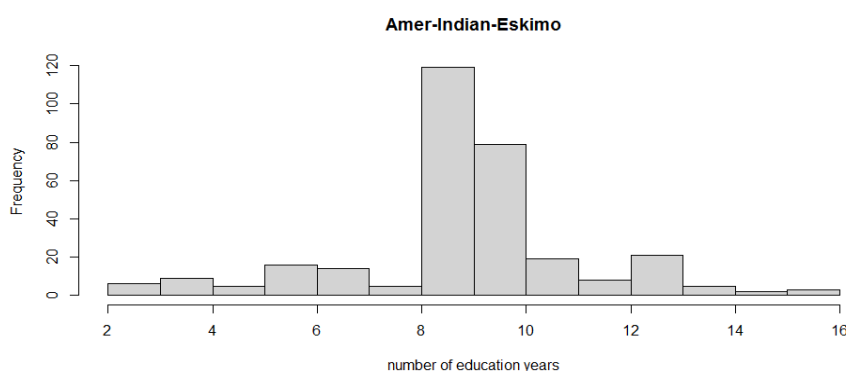


Fig. 15 Education Years Distribution for Amer-Indian-Eskimo

3.6 Sex and Education

Figure 16 shows the distribution of number of education years of female, and figure 17 shows the distribution of number of education years of male. From these two figures, it is reasonable to tell there is little discrepancy of the average education level between male and female since the distribution in two figures are very similar.

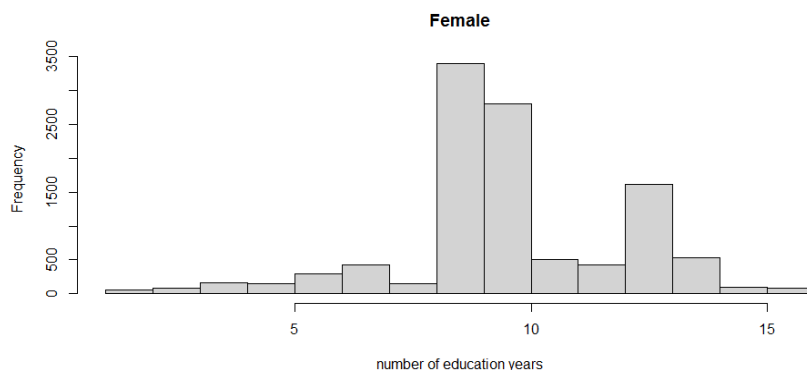


Fig. 16 Education Years Distribution for Female

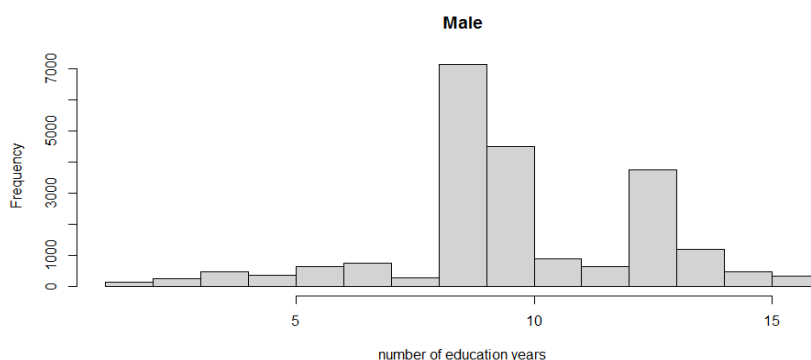


Fig. 17 Education Years Distribution for Male

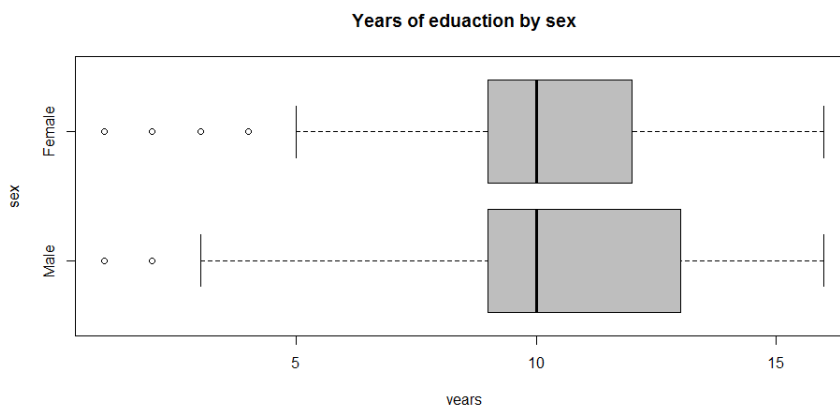


Fig. 18 Overall Education Years Distribution vs Gender

From the plots, we could see that compared with the Amer-Indian-Eskimo people, the Asian-Pac-Islander people would have higher education years; compared with the black people, the white people would have a longer education; compared with the males, more females would have education for 5 to 10 years and fewer would have a education for 10 more years. Also, males and females generally have similar education level.

4. Conclusion

People's age, sex, year of education, and race will affect people's income. When people become older like 40 years old, their income will increase because their skills would also increase. However, when they are too old like over 70 years old, people's income would like to decrease because the company would like to find new workers who are younger than them and more active than them.

It's easy to see that men's income is higher than that of women because most of the men are stronger than women, and they could do more job and harder things than women. Also, women would like to

choose part-time job than men, because most of the women would like to go home to take care of their children, like Laura D' Andrea Tyson explained [7].

First, the male-dominated society has lasted for thousands of years. Men generally hold high positions, but they also bear much greater social responsibilities than women. Second, although on average, women's psychological endurance is higher than that of men, but women in high stress state of the body will rapidly decline, third, female fertility and physiology are unique to women.

According to the result, the reason of the difference should be found to solve the problem. The unbalance between female and male has been existed for a long time. According to history of many countries, many females just were allowed to have the ability to know some basic things. From the chart we could know that male would have an longer education years average than female. In the history, female was just considered as housewife to take care of children. For race, there are many reasons about that, like race discrimination in America, African American only could get lower education than others. As a result, they just could get lower income than others.

Acknowledgments

Thanks to Dr. Sara Kališnik from ETH Zurich (Public University in Zurich) and Jiamu Bai from NYU for guiding and conducting the relevant researches and paper drafting.

References

- [1] Bureau, US Census. "Census.gov Homepage." Census.gov, 23 Dec. 2021, <http://www.census.gov/en.html>.
- [2] York, Erica. "Average Income Tends to Rise with Age: Measuring Income Inequality." Tax Foundation, Tax Foundation, 31 July 2020, <https://taxfoundation.org/average-income-age/>
- [3] Financial Samurai, et al. "Income by Race: Why Is Asian Income so High?" Financial Samurai, 5 Oct. 2021, <https://www.financialsamurai.com/income-by-race-why-is-asian-income-so-high/>.
- [4] McWhorter, John. "Black People Should Stop Expecting White America to 'Wake up' to Racism." The Daily Beast, The Daily Beast Company, 14 Apr. 2017, <https://www.thedailybeast.com/black-people-should-stop-expecting-white-america-to-wake-up-to-racism>.
- [5] Learning, UCI Machine. "Adult Census Income." Kaggle, 7 Oct. 2016, <https://www.kaggle.com/uciml/adult-census-income>.
- [6] Peck, Roxy, et al. Introduction to Statistics and Data Analysis. Cengage, 2020.
- [7] Weissmann, Jordan. "Why Are Women Paid Less?" The Atlantic, Atlantic Media Company, 13 Jan. 2014, <https://www.theatlantic.com/business/archive/2012/10/why-are-women-paid-less/263776/>