### Project

**Is there a gender income imbalance? Research on the phenomenon and causes of gender income gap.**

Social and Cultural Analytics

28 January 2019

[I. Introduction 12](#_Toc534668805)

[II. Data screening and acquisition 14](#_Toc534668806)

[III. Aim and Methodologies of the Study 16](#_Toc534668807)

[IV. Data processing and visualization 17](#_Toc534668808)

[1. Data Preparation 17](#_Toc534668809)

[2. Exploration of Gender Income Gap 20](#_Toc534668810)

[3. Predicting Modeling 26](#_Toc534668811)

[A. Decision Tree 26](#_Toc534668812)

[B. Neural Networks 30](#_Toc534668813)

[4. Binomial Logistic Regression 33](#_Toc534668814)

[V. Conclusion 37](#_Toc534668815)

[Ⅵ. Bibliography 38](#_Toc534668816)

# I. Introduction

The issue of wage gap between different genders has been a long-lasting controversial topic, and this problem has plagued women for decades. Income disparity is likely to reinforce stereotypes about women and subconsciously deepen gender discrimination.

In reality, organizations and individuals from different fields have made great efforts to reduce the inequality. As early as 1996, the National Committee on Pay Equity had set up a symbolic day, called Equal Pay Day, aiming to eliminate the gender income discrimination (Gibbon, 2018). Recently, the PayMeToo, a campaign encourages employers to be accountable to women, which was also in full swing in Britain (Topping, 2018).

However, these efforts have not achieved great success yet. A census published by the U.S. Department of Commerce this year shows that full-time female income accounts for only 80.5% of men's income under the same working conditions in 2017, and there are no obvious improvements compared with the data in 2016 (Fontenot et al., 2018). If annual income ratios continue to change at the same rate as they have since 1960, it will take another 41 years to reach parity (Institute for Women's Policy Research, 2018)

To eliminate gender pay gap, it is important to analyze the causes of this problem. An idea given by Stanley and Jarrell (1998) suggests that women usually focus more on family production and parenting, which leads them to lose opportunities to develop their careers for a long time and fail to accumulate the corresponding human capital in their work. From a historical perspective, men and women have assumed different roles in their families and economies (ibid.). However, these opinions, which are based on artificial statistics or experience, are difficult to be accurate and credible.

In the era of digital society, an increasing amount of digitized data and statistical tools provide researchers with a more scientific and cost-effective way to discover the essence of cultural and social phenomena. Manovich (2005, in Manovich, 2016) develops a concept of cultural analytics, which refers to the analysis of large amounts of cultural data through computer and visualization technology. This concept has been widely applied to the study of digital humanities today. Researchers could get deeper insights from culture and society by collecting and processing massive datasets from the target population.

In this study, R scrips, predicting modeling and binomial regression analysis would be used to study the relevant variables related to income in the "Census Income" dataset and analyzed the data structure under different variables more specifically. I aim to identify the phenomenon of gender wage gap inequity caused by different factors in the context of digital social analysis and to provide more specific suggestions for further narrowing the income gap.

# II. Data screening and acquisition

To pursue this analysis, it is important to obtain the data including all the factors that may affect income. But this is unrealistic, because it is difficult to pinpoint all the revenue-related factors without doing a lot of surveys and research, and organizations that count those data sets often refuse to disclose them because of privacy and other issues. However, not all factors are equally important to income levels. We can learn from some of the commonly accepted factors or inferences from previous studies to find out the reasons that affect the income imbalance by controlling and comparing these variables.

Researchers generally believe that education background, career choice and working-hours play an important role in personal salary (Salary.com, 2018; Discover your money, 2017; Mehdikarimi et al., 2015). In addition, a study did by Huber, Bookstein and Fieder (2010) indicates that the physical characteristics and socio-economic status of women determine that they have to spend more time on fertility and family. Therefore, employers are more willing to pay male candidates higher wages when they are hiring.

Based on these factors that affect income, I have found that a data frame in which each row corresponds to one person and the columns cover income, gender, and these factors are needed. Moreover, in order to ensure the accuracy of the prediction of the impact of different attributes on income, the instance should be as much as possible.

Many attempts have been made but failed in the process of acquiring data. I first tried to get the data from data.gov, an official website created by U.S. government to provide machine readable data. But through keyword searches such as "income", "census" and "pay gap", I did not find data suitable for this project. Then, I attempted to get useful dataset from some existing studies. However, the data in those studies has been processed by the researchers and they did not disclose the raw data of their researches.

Finally, I get access to a suitable data set from the UC Irvine Machine Learning Repository. This data set, also known as "Census Income" data set, was extracted by Kohavi and Becker (1996) from the 1994 census database, which was originally used to predict whether an individual make over $50K per year. Although the data set has been around for 24 years, it still has great research value for this project. In this data frame including 32,560 instances and 14 different attributes, each row represents a respondent, and these attributes contain not only basic information such as gender, age, race and income, but also factors such as education, occupation, working hours, relationship and marital status that may cause income gap. Moreover, the names of the respondents in this data set have been removed, which ensures their privacy to the greatest extent.

# III. Aim and Methodologies of the Study

The purpose of this study is to investigate and analyze the income gaps between different genders and visualize them in a clear way. Therefore, this project will be analyzed from three aspects with different methodologies.

In the first part, I analyzed whether the gender income gap exists by comparing the proportion of men and women earning more than $50K per year under the same variables, like occupation and education. To implement it, I made use of **dplyr** package and **ggplot2** package to aggregate data and to visualize it, following Blanke’s lecture notes (2018).

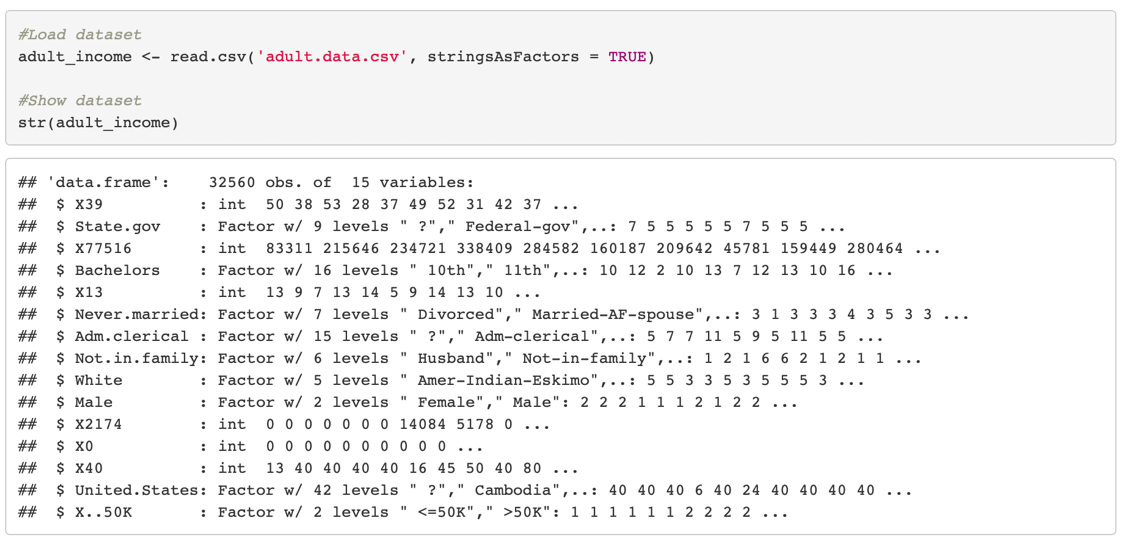
In the second part, a decision tree and a neural network were built respectively to analyze the influence of different variables on the income level. Through the representation of the decision tree and neural network, the most important variables affecting the income level in the data set could be identified (Pandey, 2018; Blanke, 2018).

Because the income in the data set is a binary variable (whether each individual’s income is over $50,000 or not), I used binomial logistic regression to further analyze the positive and negative effects of different factors on women's income by running the function **glm()** in the last section. After finishing it, it could be possible to draw a relatively comprehensive conclusion about the factors that cause lower incomes of women and to make some suggestions for narrowing the gender income gap.

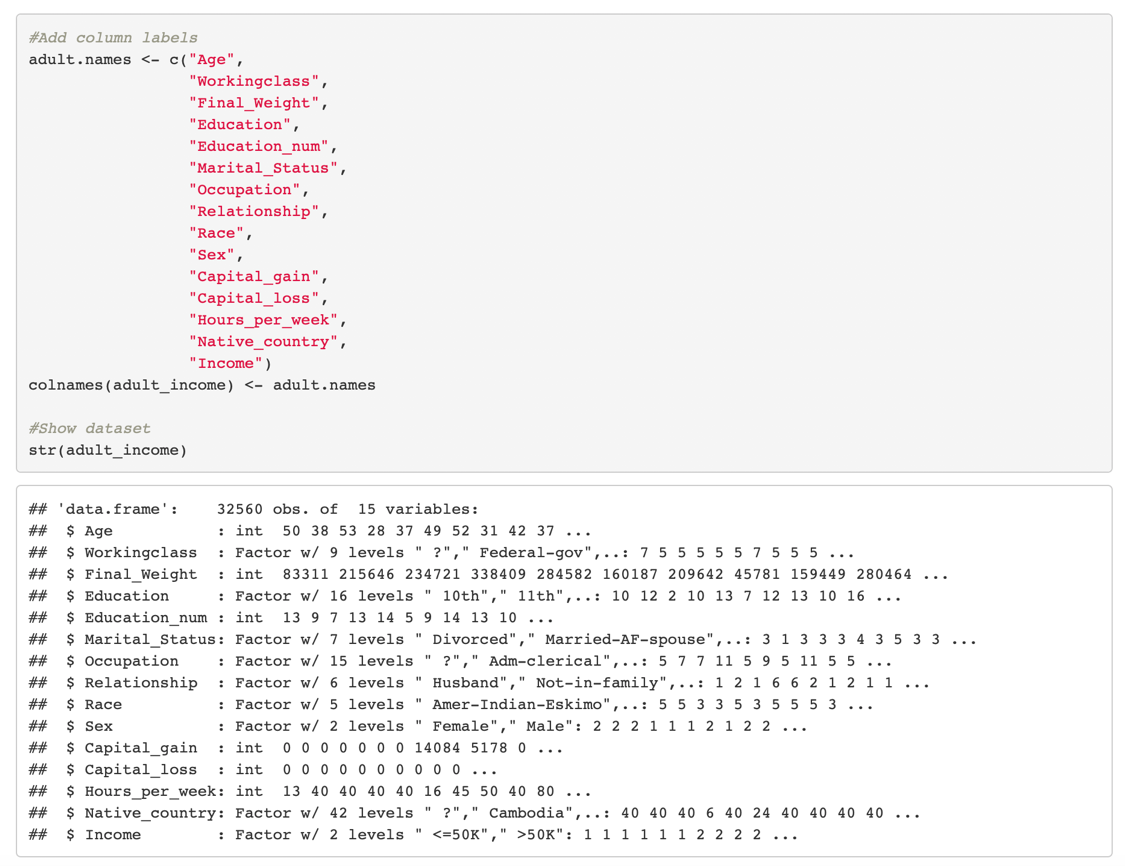
# IV. Data processing and visualization

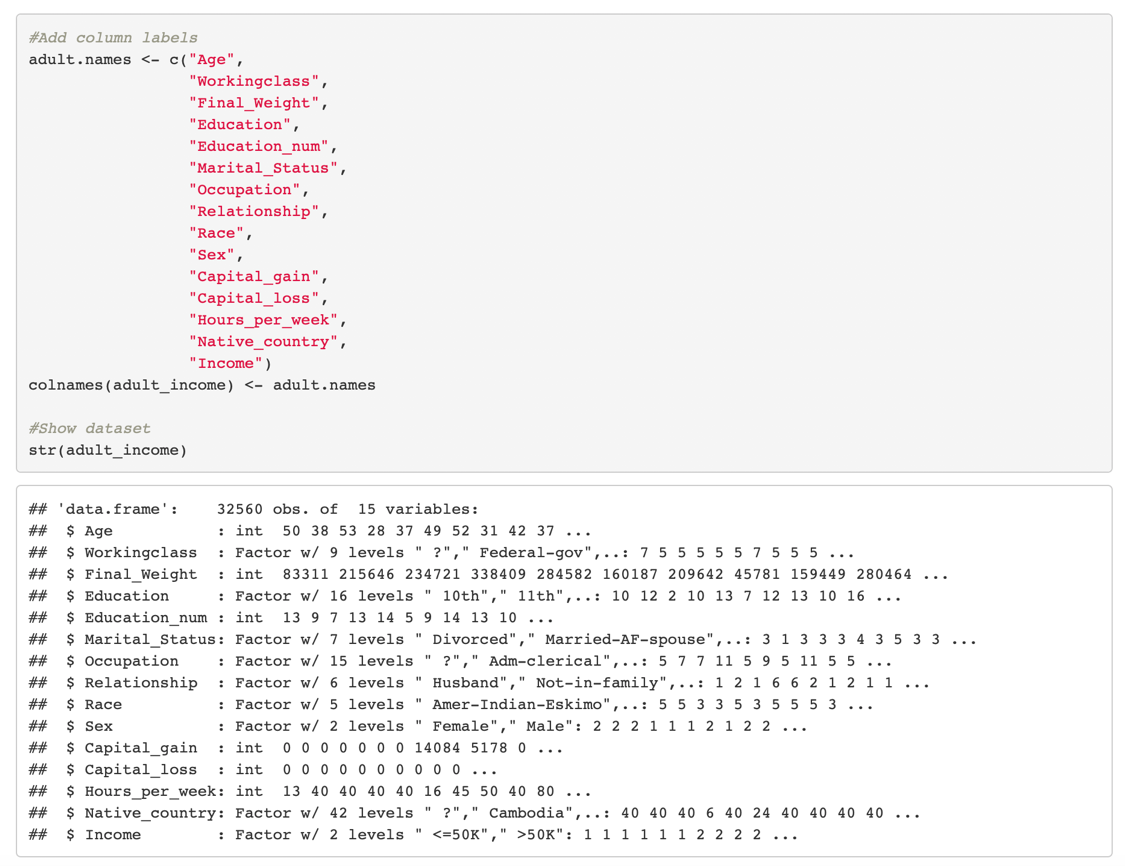
## 1. Data Preparation

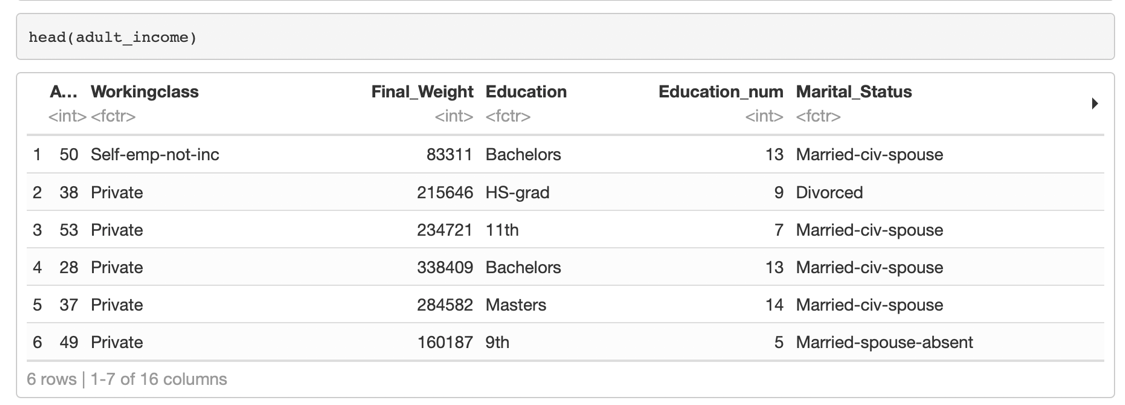
In order to achieve subsequent analysis, the data is loaded into R Studio by using the function **read.csv()** and naming it **adult\_income**. Then running the **str()** function to display the internal structure of this data frame.



It could be seen that the **adult\_income** data frame lacks column labels, so adding the names of the different columns and checking the structure of it again.

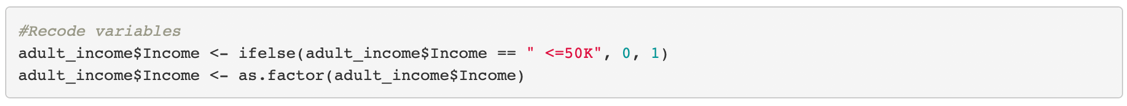




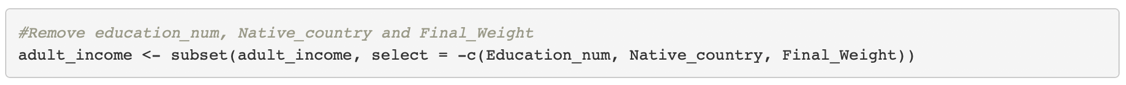


This data frame contains 32,560 observations and 15 variables. Excluding response variable “Income”, there are 5 continuous variables and 9 categorical variables left.

In order to make the following data processing more convenient, I first changed “Income” as a factor vector of 1 or 0 (1 means “>50K”, and 0 means “<=50K”).



I found that column “Education\_num” matched the same information described by “Education”, so I removed it. I also deleted column “Native\_country”, since more than 90% of the respondents were "United States". In addition, the column Final\_Weight was deleted because it cannot be reasonably interpreted.



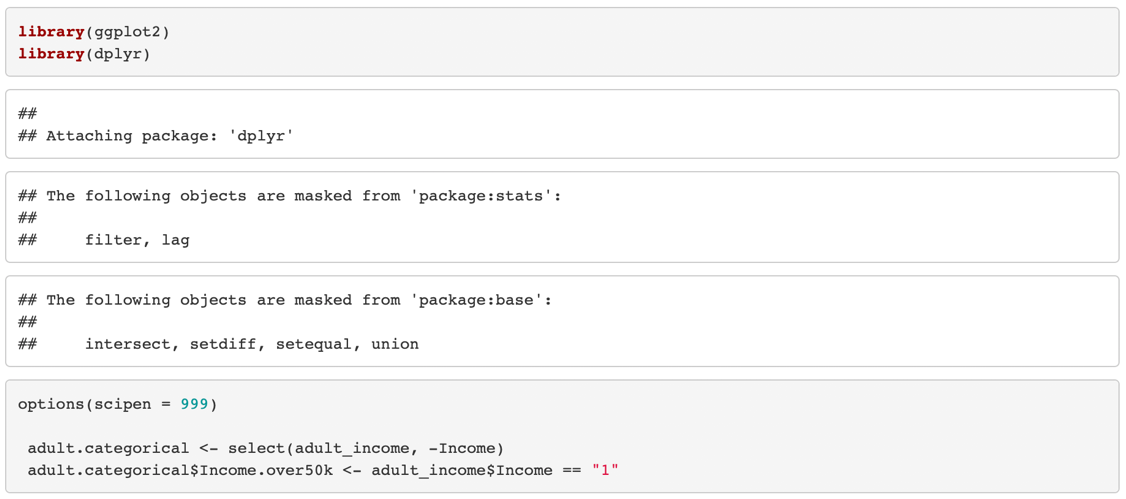
I also noticed that there were some values appear as "?" in this dataset, so I assigned these values to NA and removed the observations containing NA from the data frame through the function **na.omit()**.



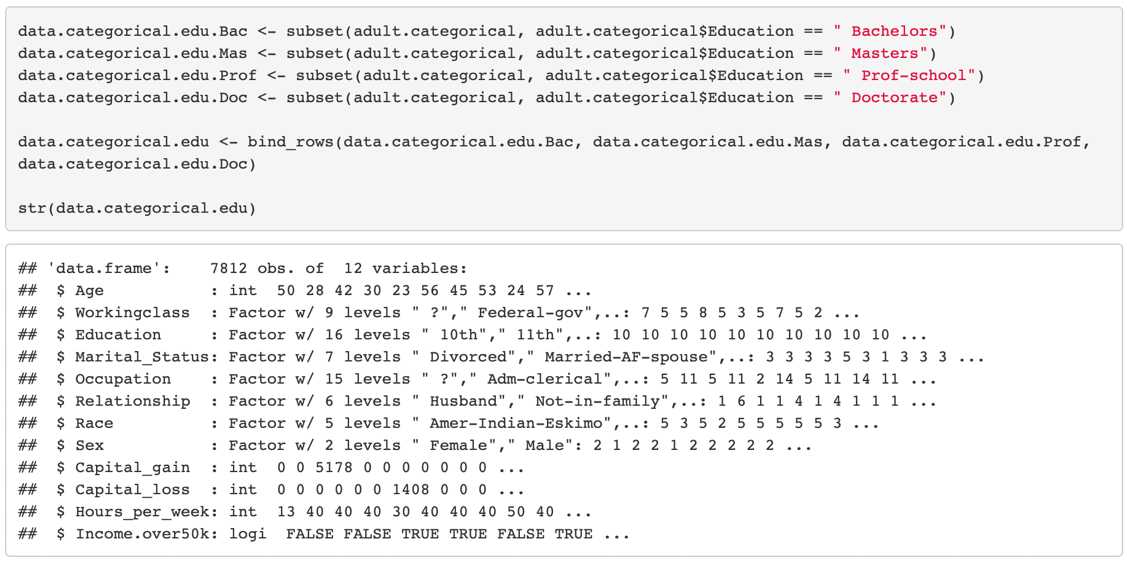
## 2. Exploration of Gender Income Gap

After cleaning data, I compared the proportion of male and female incomes over $50K from four widely discussed aspects (“Education”, “Race”, “Occupation”, and “Age”) to verify whether the gender income gap exists.

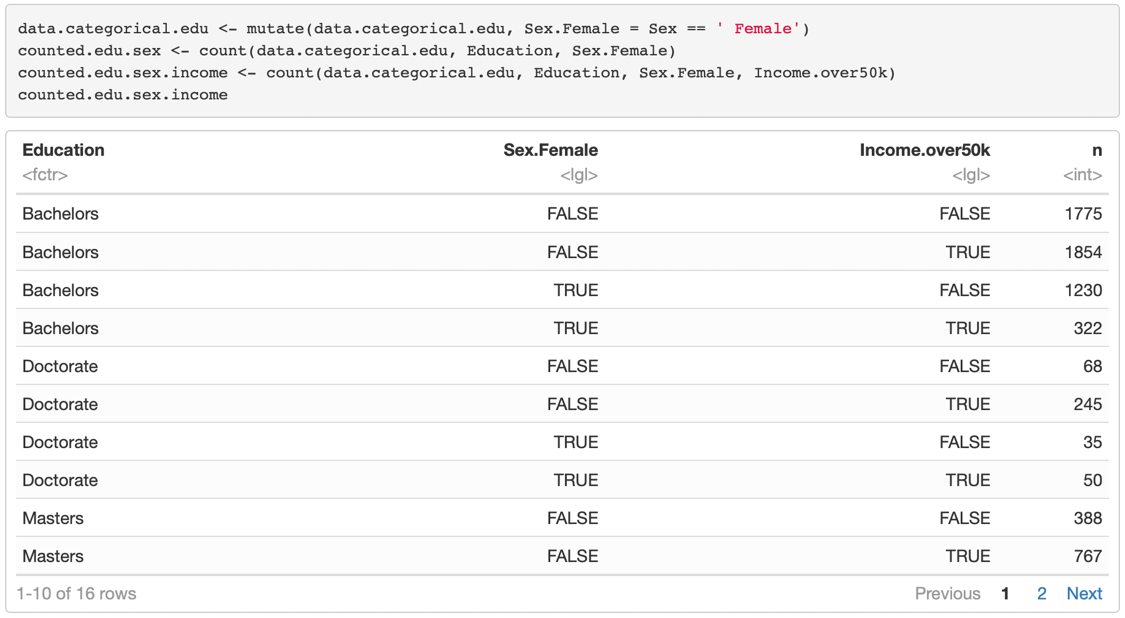
**dplyr** package and **ggplot2** package were used to aggregate and visualize the dataset. In addition, I created a new dataset called **adult.categorical**, in which the column containing the logical (TRUE or FALSE) variable named "Income.over50k" replacing the previous "Income" column.



When analyzing whether men and women have equal incomes under different education backgrounds, I chose respondents from "Bachelors", "Masters", "Profess" and "Doctorate", and combined them into a new dataset, because it is widely believed that highly educated men and women have no income gap (Salary.com, 2018).

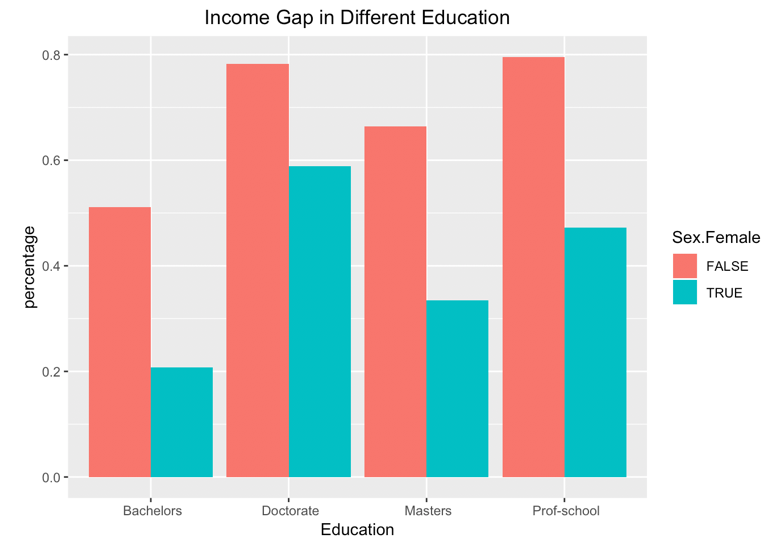


To compare the proportion, I used the function **mutate()** from **dplyr** package to compute and append one new column called “Sex.Female”, which contained logical variables for women that were TRUE and men that were FALSE. Then, I used **count()** function to find out the number of female and male who earned more than $50K under the same education background.



After this step, I used the functions **left\_join()**, **filter()** and **mutate()** to create another dataset called **percent.edu.sex.income**, which contained the proportion of men and women earning more than $50K at different educational levels. Then, I plot it by **ggplot2**.

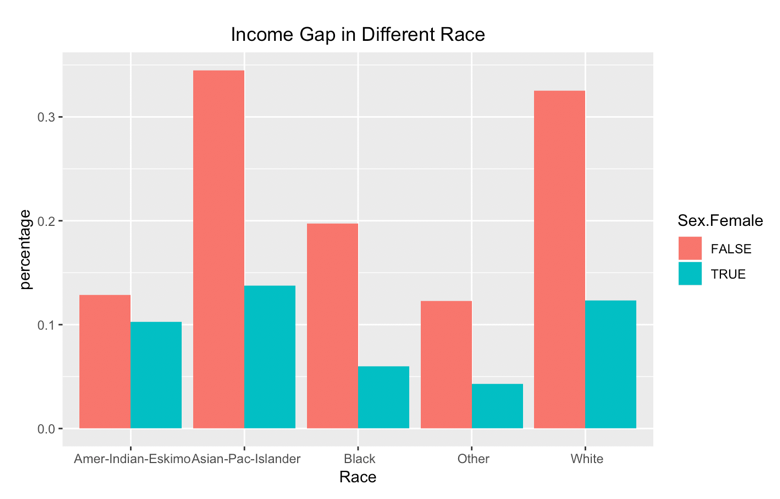
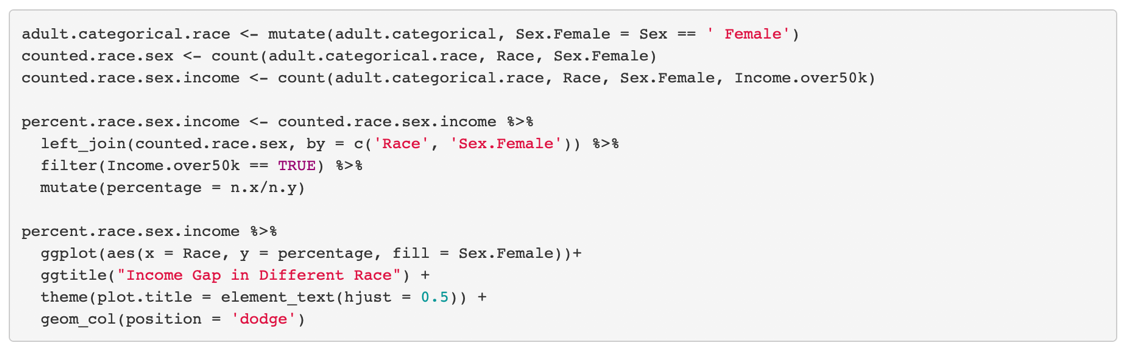




*Figure1 Income Gap in Different Education*

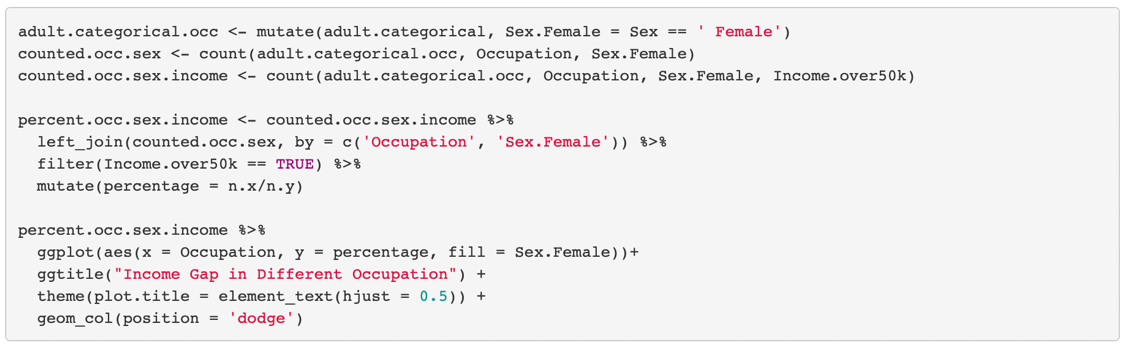
Figure 1 shows that the income inequality between men and women is always there no matter what the degree is.

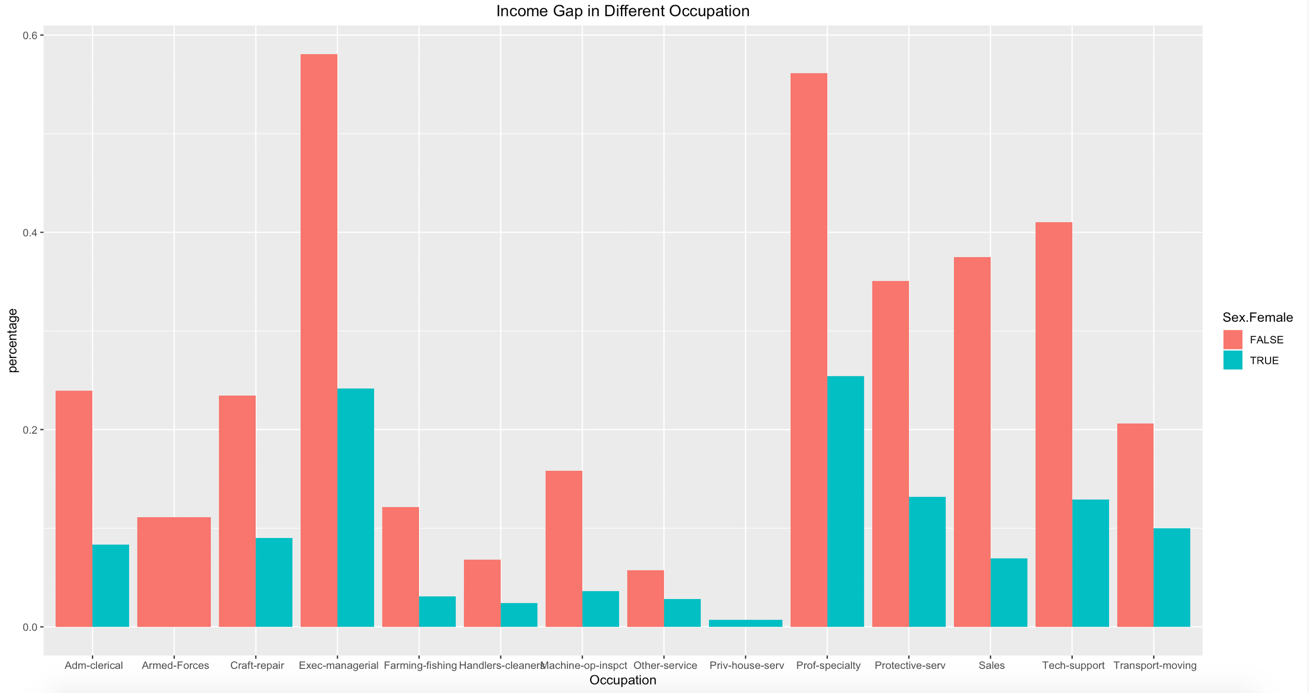
Repeating the same operation, we can see from Figure 2 that even in the same race, men are more likely to get a higher salary than women.



*Figure2 Income Gap in Different Race*

Also, figure 3 shows that it is difficult for female workers to get high salaries in both high - and low-income industries. The reason why "Amed-Forces" and "Priv-house-serv" appear the same color is that the data lackes male or female samples.

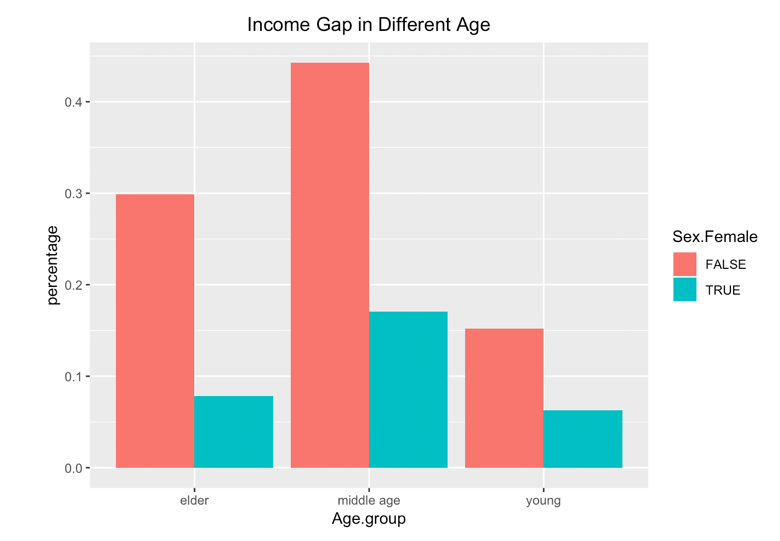




*Figure3 Income Gap in Different Occupation*

In terms of analyzing whether there are income inequality at different ages, I noticed that "Age" is a numeric vector with a large span of ages by using the function **summary()**. Therefore, I divided them into three age groups: elder (66-90), middle age (36-65) and young (17-35). As can be seen from figure 4, the phenomenon that women have lower incomes is still common in different age groups.





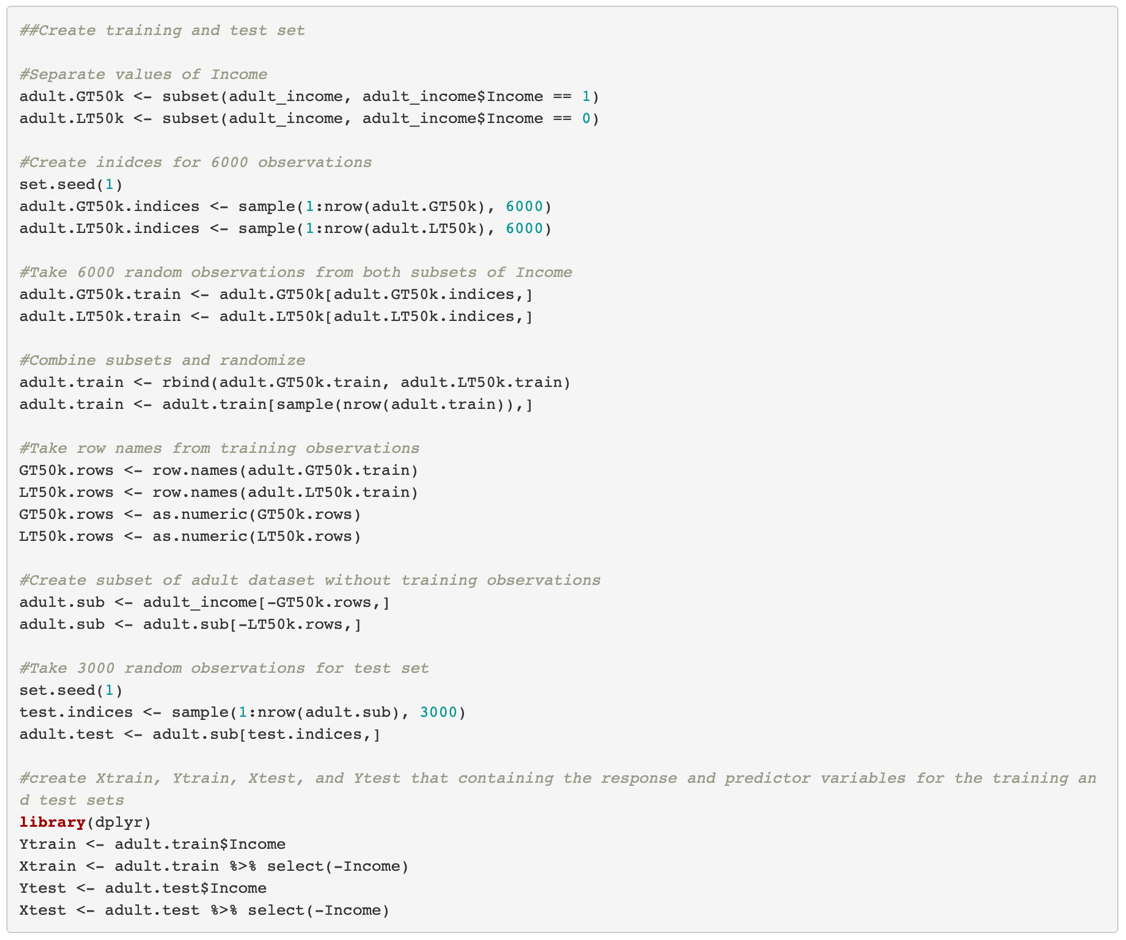
*Figure4 Income Gap in Different Age*

## 3. Predicting Modeling

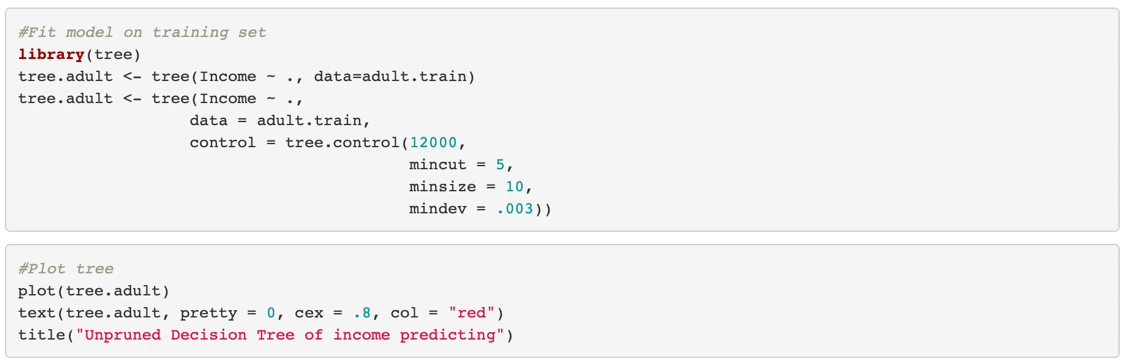
To compare the influence of different attributes on the income level, I used the decision tree model and the neural network model for analysis.

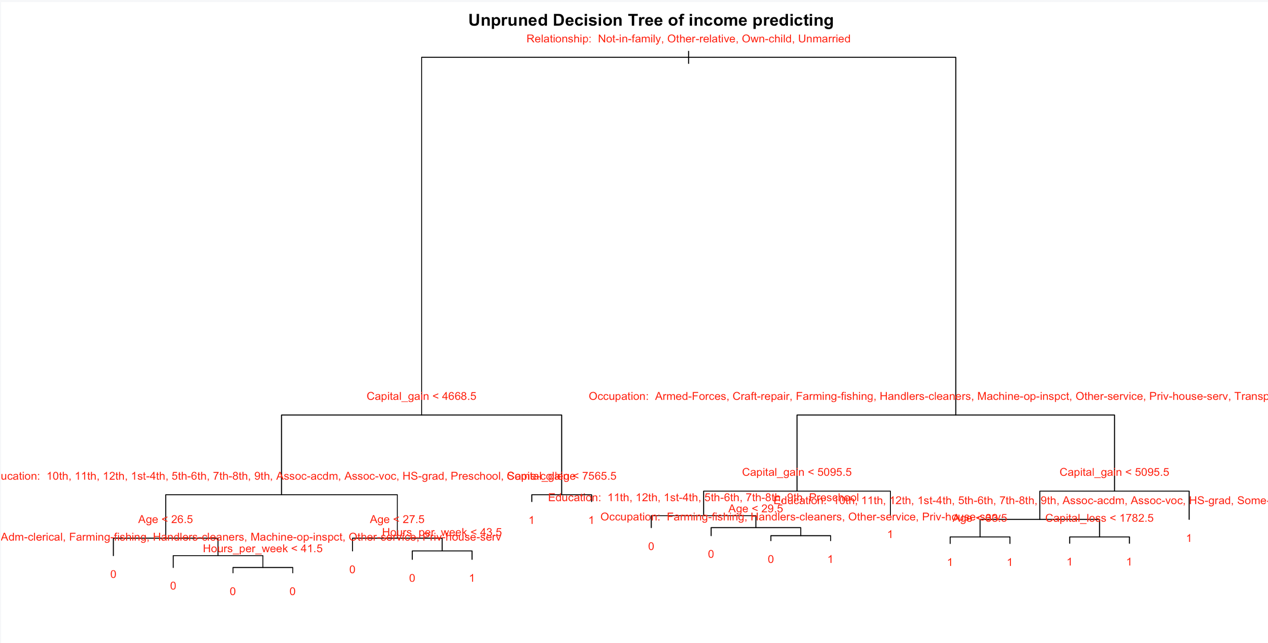
### A. Decision Tree

Since the number of samples with income higher than $50K in the data set is far lower than those that are less than or equal to $50K, I under-sampled the data and randomly selected 12,000 observations for training data set and 3,000 observations for testing data set.

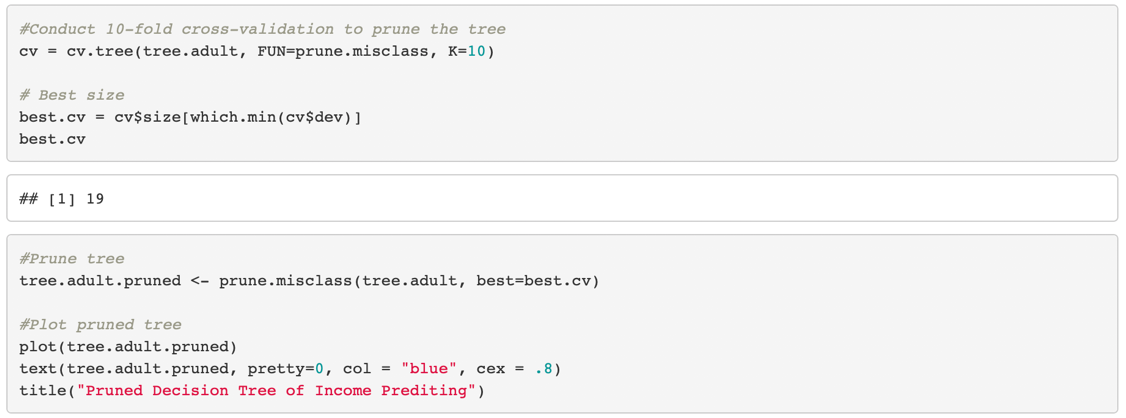


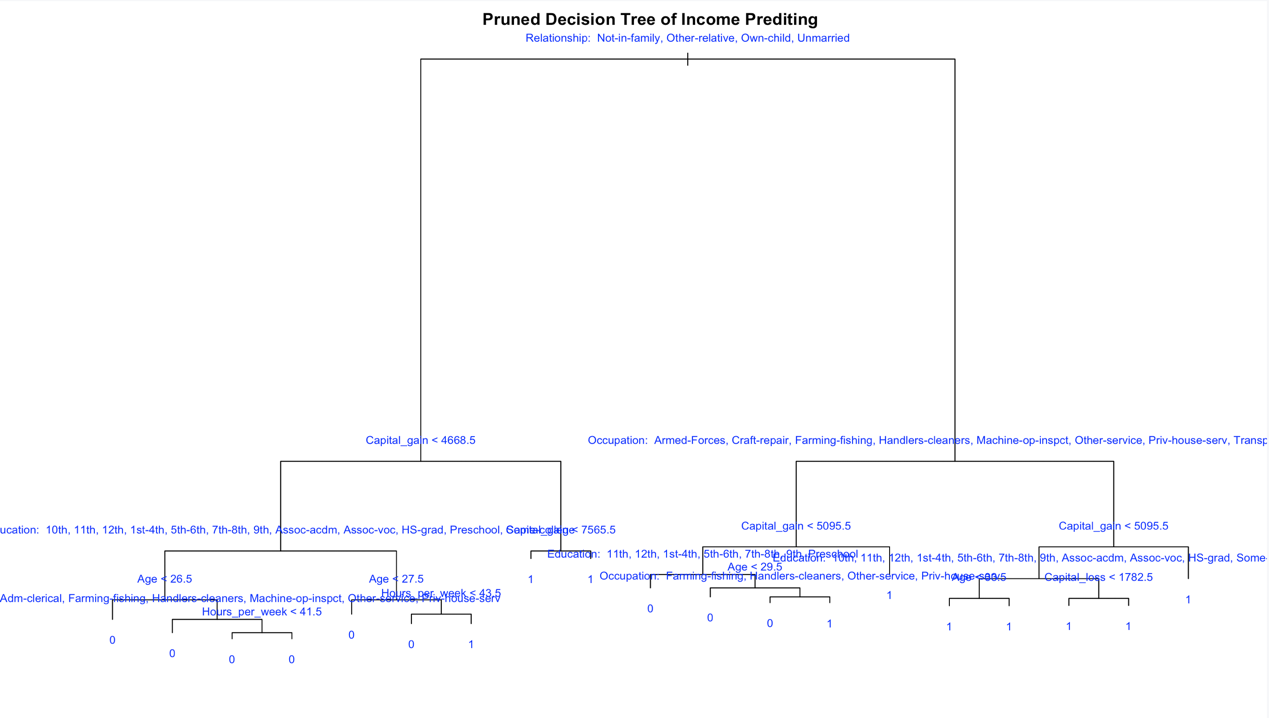
Through using the **tree()** function from **tree** package, I created a decision tree that took all the other variables as independent variables and income as a response variable.





To reduce the error rate, I performed a 10-fold cross-validation to prune it.

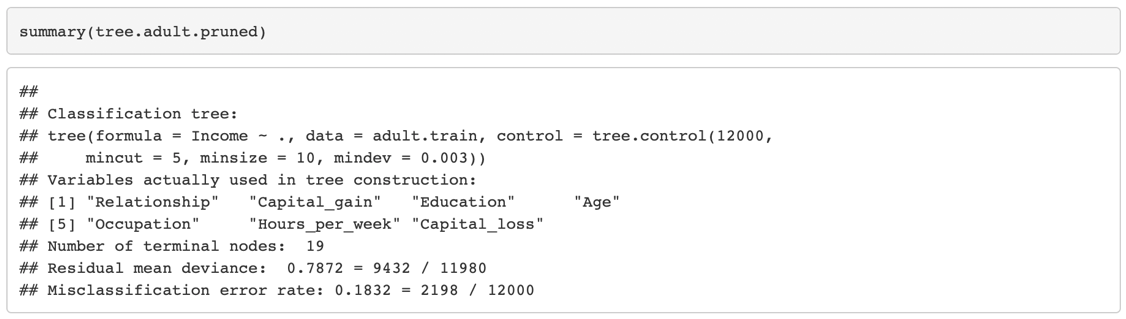
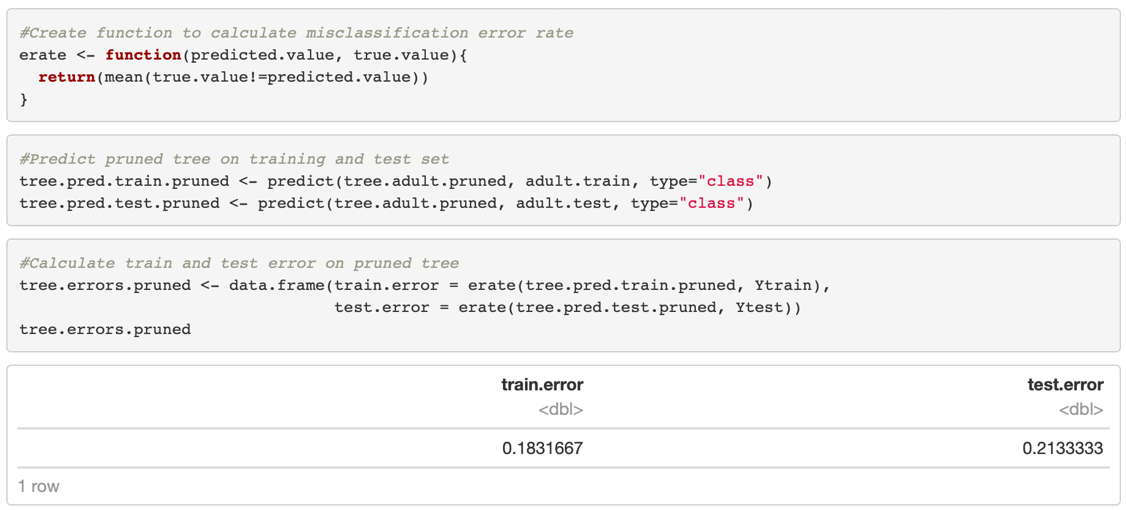




*Figure5 Pruned Decision Tree of Income Predicting*

The text at each split represents a prediction of high or low income, and if a sample contains these values, it will move to the left (“<=$50K”). We could find from the first split that an individual will move to the right of the tree when he or she is marriedwhich means this individual is more likely to earn a high income. We can also see that if a person's education level is lower than "some-college", he or she can hardly get a high salary.

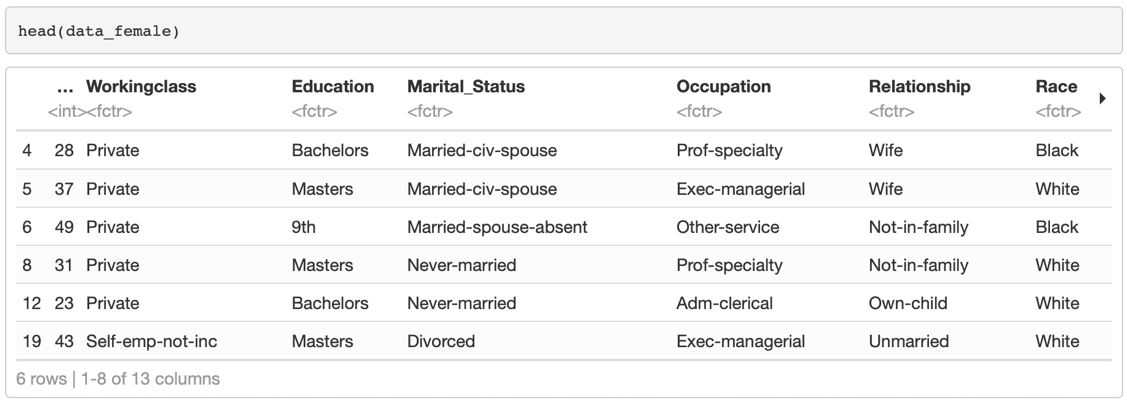
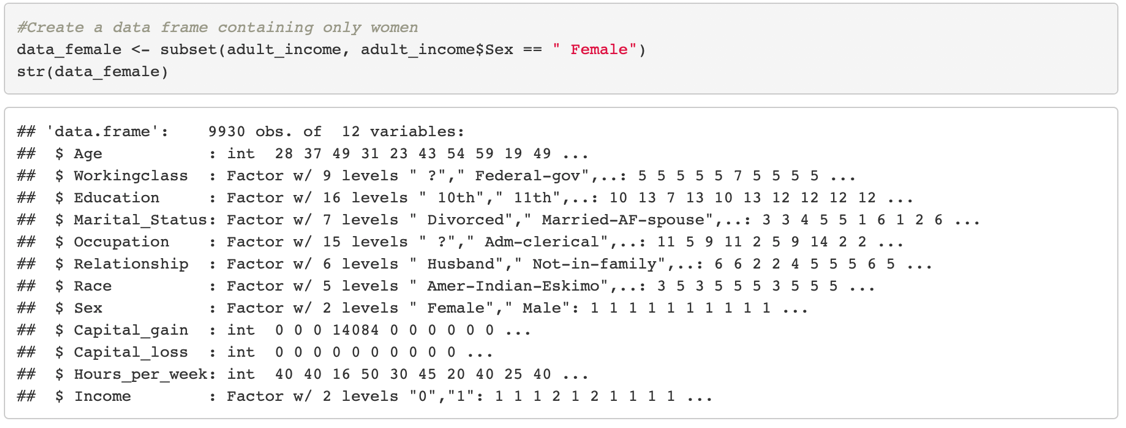
However, it is still difficult to analyze the impact of different variables from this plot. As a result, after calculating the training and test error rates, I used the **summary()** function to observe the variables in this tree that would affect the predicted results.



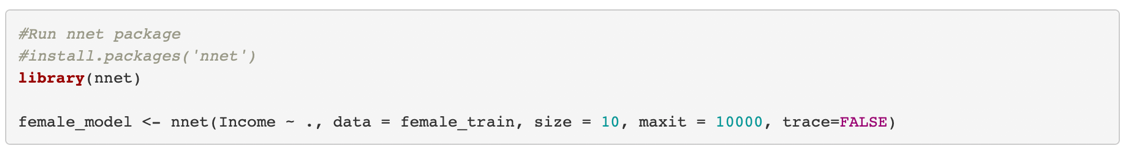
The training error rate and test error rate were 18.3% and 21.3% respectively, and “Relationship”, “Capital\_gain”, “Education”, “Age”, “Occupation”, “Hours\_per\_week” and “Capital\_loss” were the most important predictors of income.

### B. Neural Networks

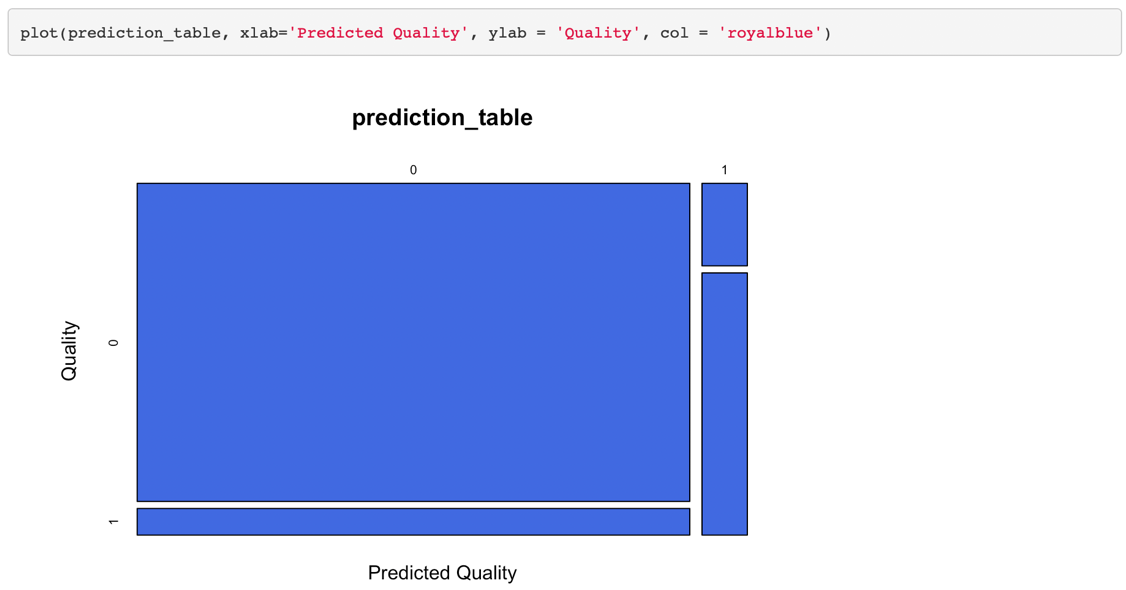
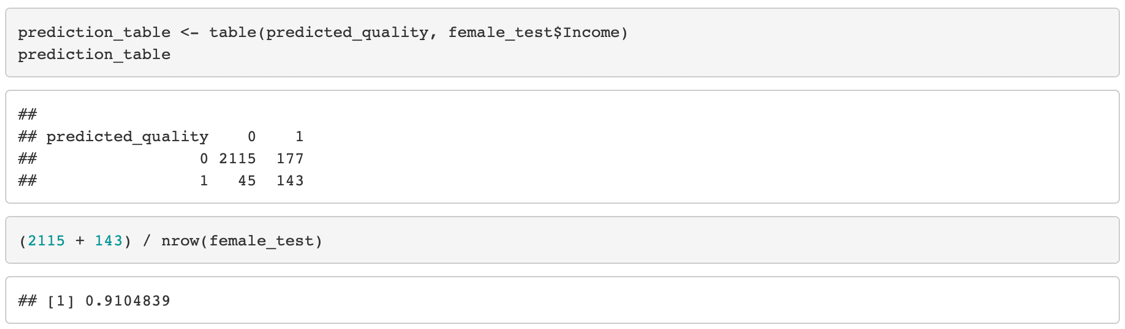
To further study the impact of different variables on women’s income, I extracted a data frame called **data\_female** containing only “Female” value and created random training data set and testing data set from it.



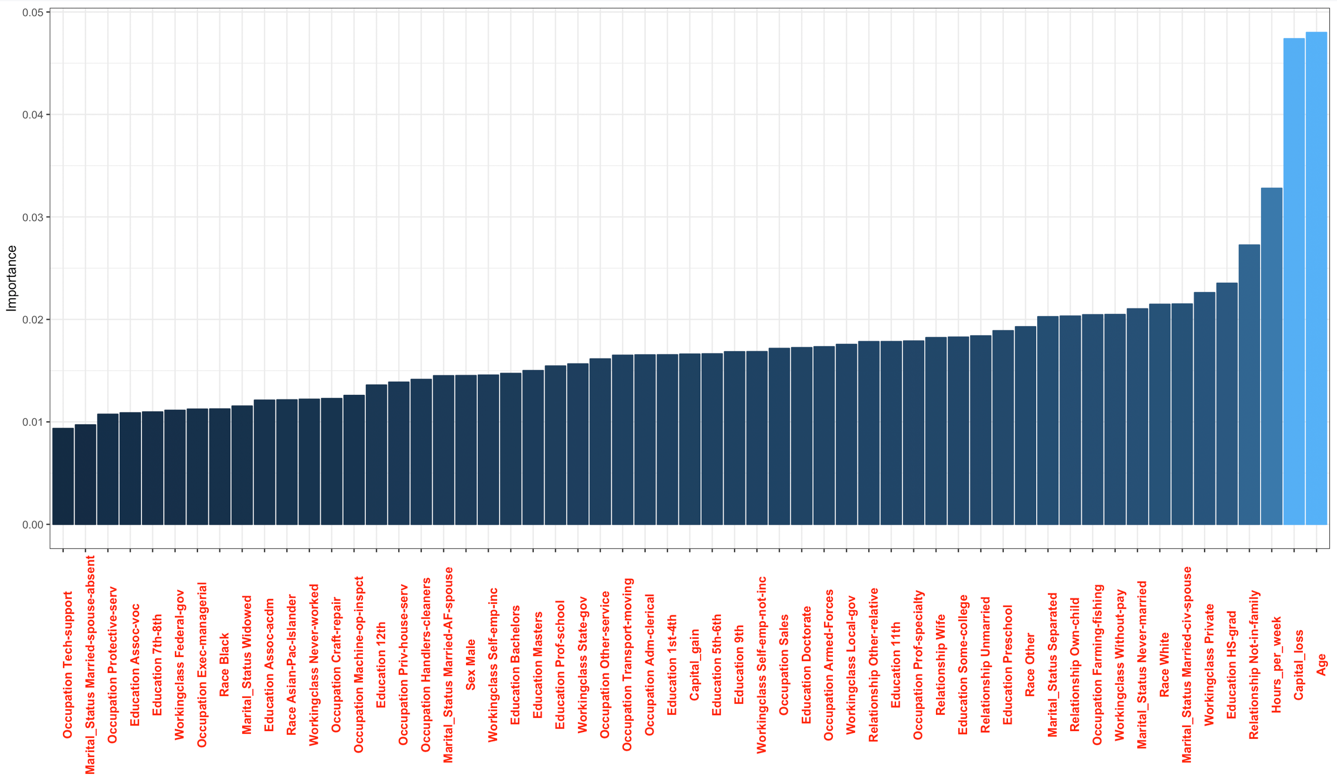
Next, I ran the **nnet** package to train the neural network based on female\_train data and simulated it using the female\_test data with function **predict()**.



After calculating the accuracy of model is 91%, I used the **plot()** function to present it as a mosaic plot.



To understand which features have the greatest impact on female income, I used the **garson()** function from **NeuralNetTools** package with **ggplot**.



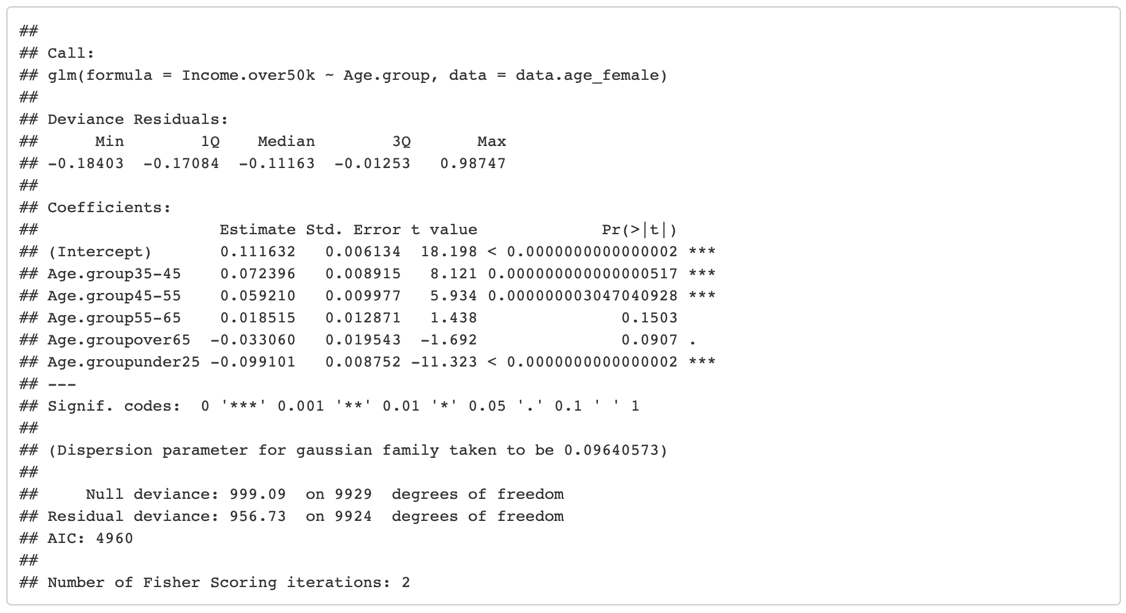
*Figure6 The Importance of Different Factors to Female Income*

From this chart, we can see that the prediction of neural network is basically consistent with that of decision tree. "Age", "Relationship", "Hours\_per\_week" and "Capital\_loss" all have a great impact on female income.

## 4. Binomial Logistic Regression

Binomial logistic regression is used to further explain whether the factors that have a significant impact on women's earnings were positive or negative.

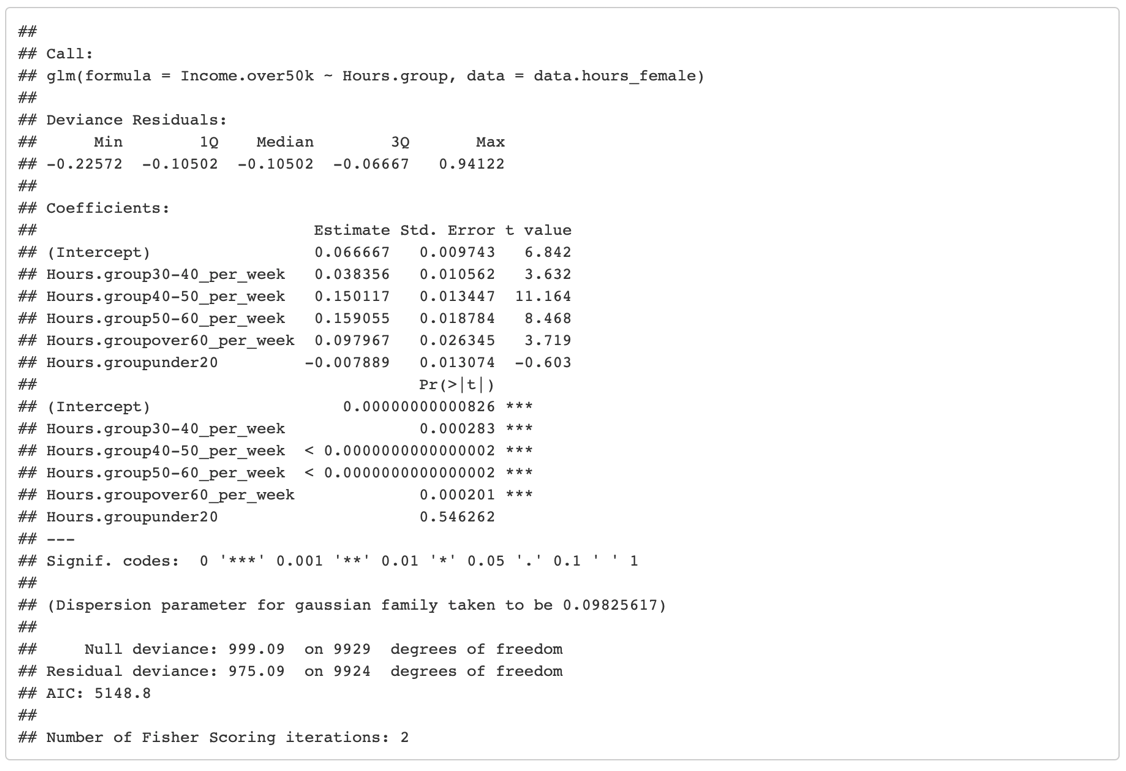
In the process of analyzing age, I divided age into six groups to make the results more accurate. Then, I used **glm()** function to analyze the relationship between “Age.group” and “Income.over50k”.



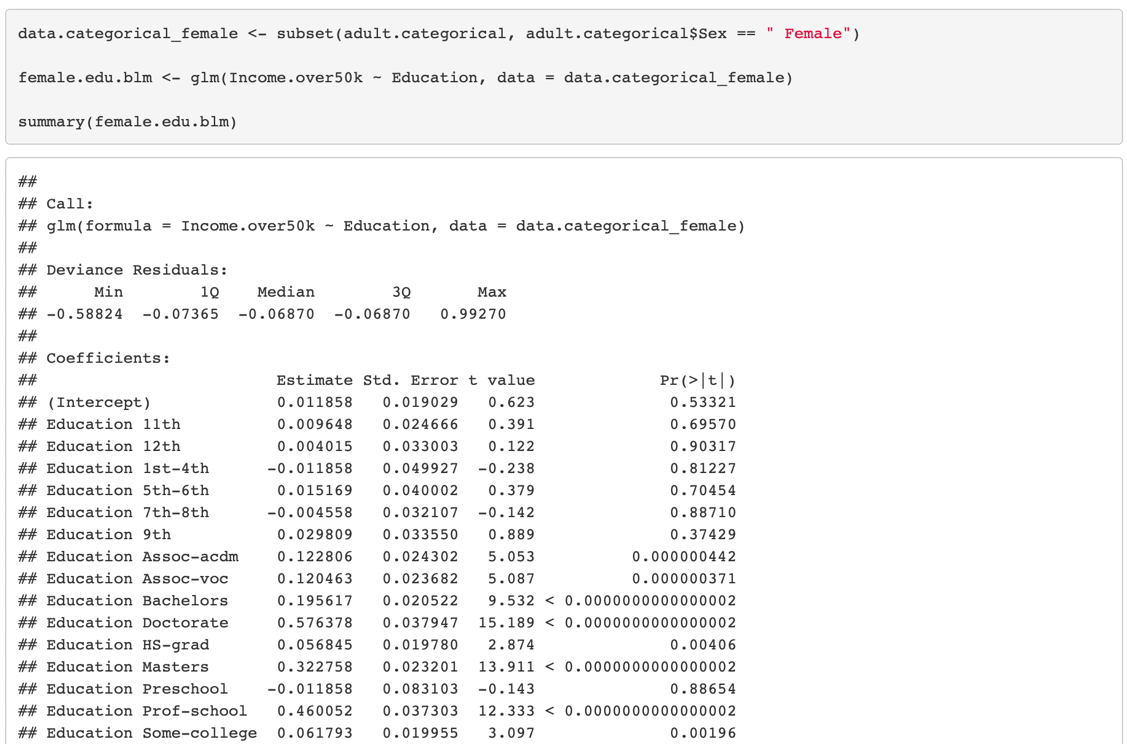
I noticed that women between the 35 and 45 were most likely to earn high salaries. On the contrary, it is difficult for women under 25.

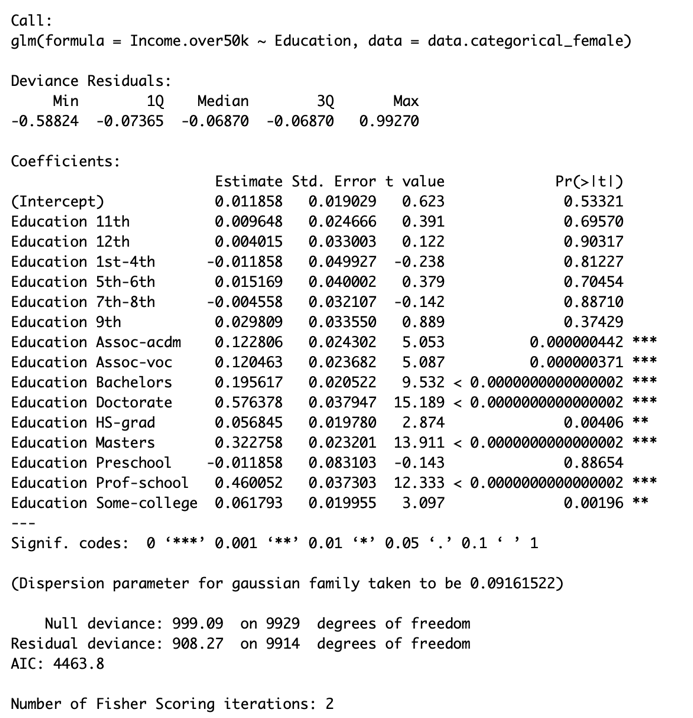
I applied the same method to verify “Hours\_per\_week”, and it shows that working 40 to 60 hours a week plays a positive role in getting high wages, while working less than 20 hours is negative.



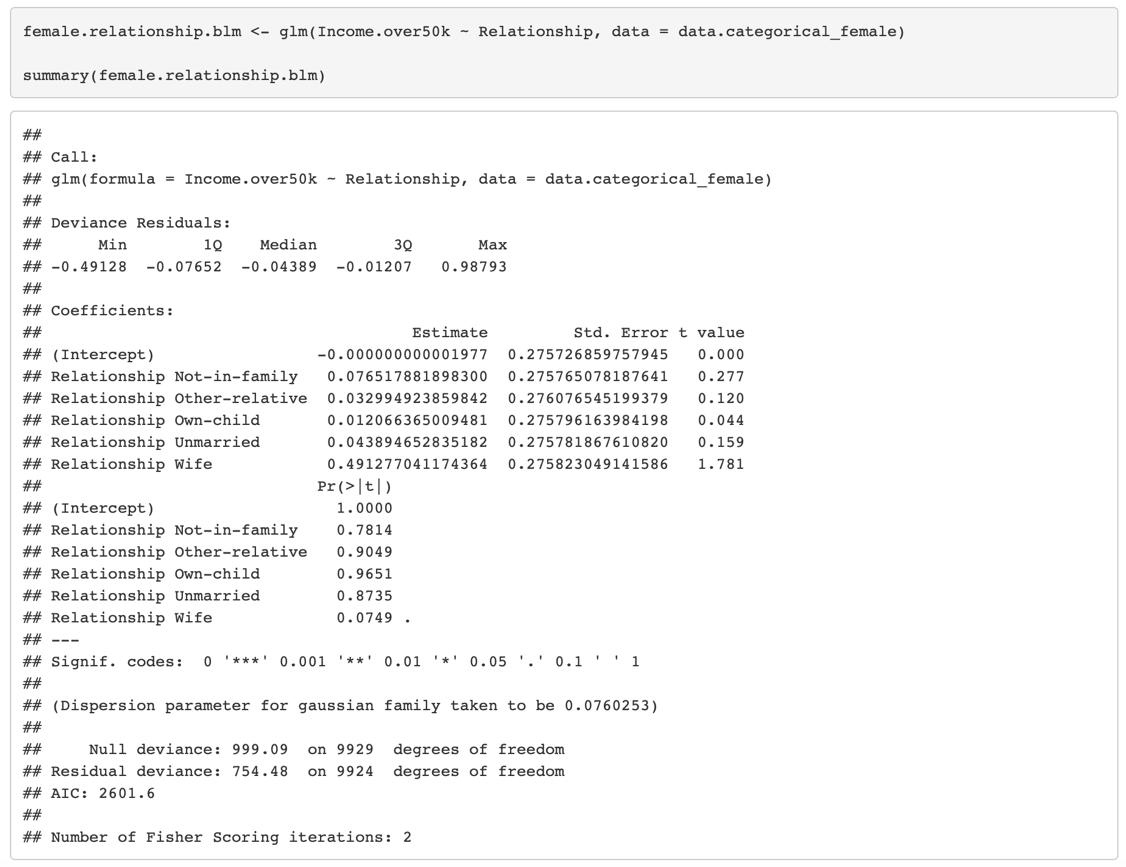


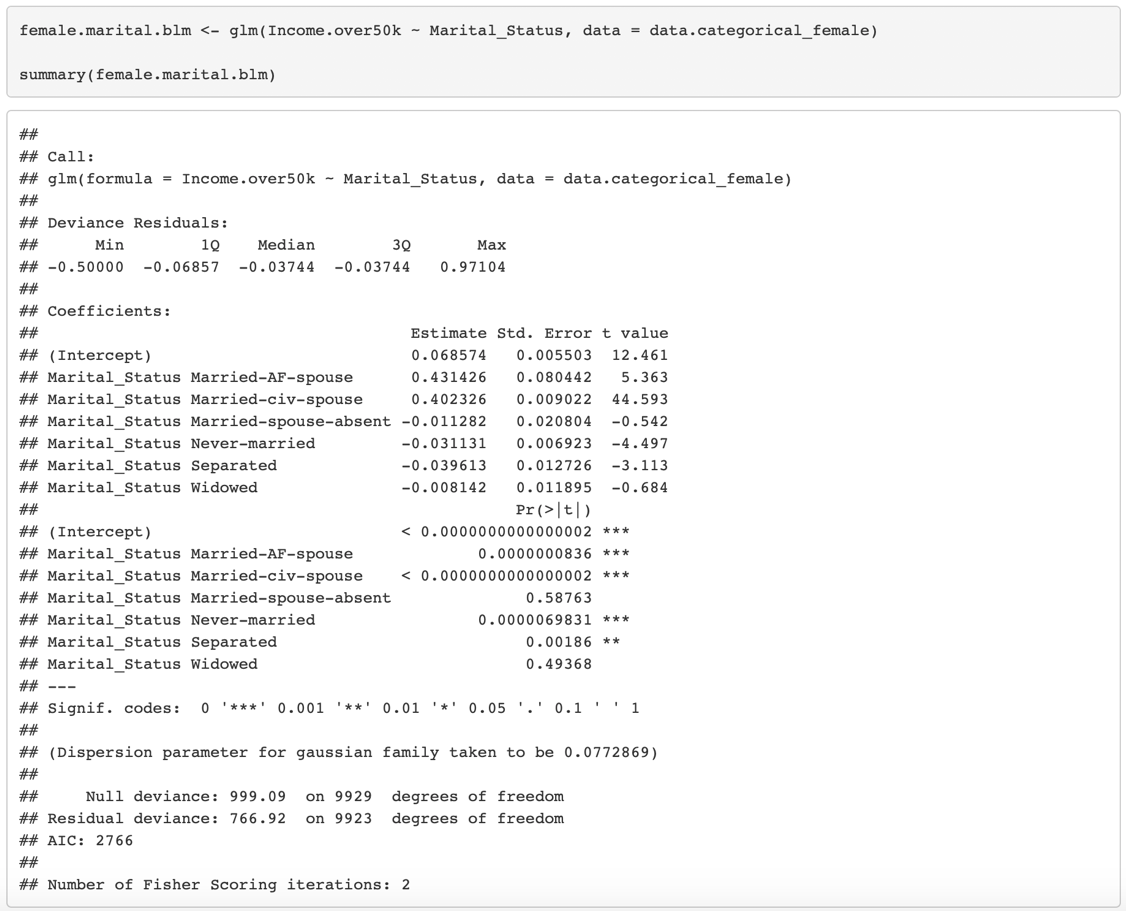
Through the analysis of "Education", I found the results, as predicted by Salary.com (2018), that education is positively related to income.



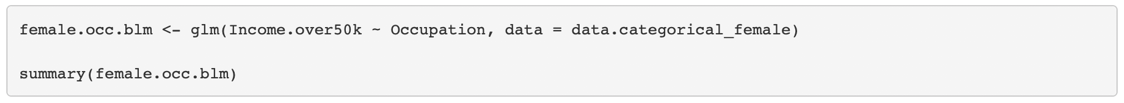


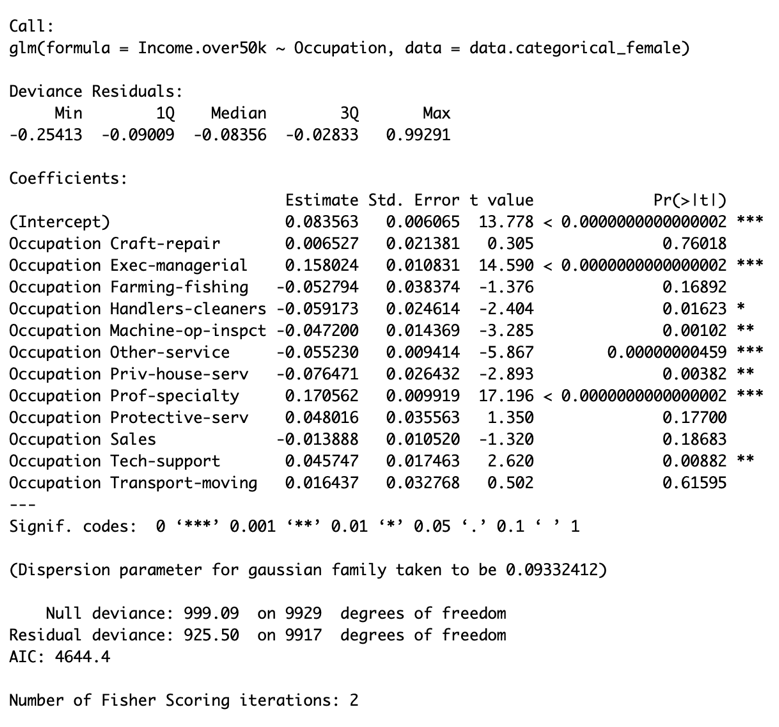
Similarly, according to the analysis of “Relationship” and “Marital\_Status”, married women are more likely to receive higher wages than unmarried or divorced women, which is basically consistent with predictions of “Age”.





Finally, the analysis of “Occupation” shows that managers or technical professionals tend to earn high salaries, while those in service or sales often earn less.





# V. Conclusion

After data comparison, predictive modeling, and binomial regression analysis, we have a more comprehensive understanding of the income gap between different genders and the causes of these problems. It is common that women are less likely to be treated equally by employers in the job search process. Based on the analysis of age and marital status, we speculate that most women will focus on family and child rearing before the age of 35, and this stereotype has largely caused the reality that women's income lower than that of men.

In order to change the disadvantage position of women in the workplace, except calling on the society to pay attention to this issue, improving their education and giving them appropriate employment guidance are also feasible.

However, there are still some limitations in this project. Since the data I used only included 32,560 samples, of which only 9,930 were women, the predicted results may not be accurate enough. In addition, the selection of factors affecting women's income is not sufficient, which may lead to the omission of other important variables in the analysis process. By improving these limitations and applying the results of the research, I believe that the problem of low female income will be greatly improved.

# Ⅵ. Bibliography

Blanke, T., 2018. communities of Digital Culture and Society 2, *Lecture Notes,* [online] Available at: <https://keats.kcl.ac.uk/pluginfile.php/3511745/mod\_resource/content/0/Communities\_of\_Digital\_Culture\_and\_Society\_2.html> [Accessed 22 Dec. 2018].

Blanke, T., 2018. Predicting and Modelling - A Neural Network Experience, *Lecture Notes*, [online] Available at: <https://keats.kcl.ac.uk/pluginfile.php/2998527/mod\_resource/content/0/Predicting Modelling.html> [Accessed 22 Dec. 2018].

Blanke, T., 2018. Visuals and Arts, *Lecture Notes*, [online] Available at: https://keats.kcl.ac.uk/ pluginfile.php/2998518/mod\_resource/content/0/SCA-Week8-handouts.pdf [Accessed 22 Dec. 2018]

CALFAS, J., 2018. Women Have Pushed for Equal Pay for Decades. It's Sad How Little Progress We've Made, *Money*. [online] Available at: <http://time.com/money/5225986/equal-pay-day-2018-gender-wage-gap/> [Accessed 23 Dec. 2018].

ESTONIANCREDITUNION., 2017. Income Essentials – What factors affect your income?, *Discover your money*, [online] Available at: <https://blog.estoniancu.com/2017/01/06/income-essentials-what-factors-effect-your-income/> [Accessed 24 Dec. 2018].

Fontenot, K., Semega, J. and Kollar, M., 2018. Income and Poverty in the United States: 2017, *census.gov,* [online] Available at: <https://www.census.gov/content/dam/Census/library/ publications/2018/demo/p60-263.pdf> [Accessed 24 Dec. 2018].

Gibbon, K., 2018. Equal Pay Day 2018: More Than a Wage Gap, *Royal Society for the encouragement of Arts, Manufactures and Commerce*, [online] Available at: <https://www.thersa.org/discover/ publications-and-articles/rsa-blogs/2018/11/equal-pay-day> [Accessed 23 Dec. 2018].

Huber, S., Bookstein, F. and Fieder, M., 2010. Socioeconomic status, education, and reproduction in modern women: An evolutionary perspective. *American Journal of Human Biology*, 22(5), pp.578-587.

Kohavi, R. and Becker, B., 1996. delve adult dataset detail, *Cs.toronto.edu*, [online] Available at: <http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html> [Accessed 23 Dec. 2018].

Institute for Women's Policy Research., 2018. Women’s Median Earnings as a Percent of Men’s Median Earnings, 1960-2017 (Full-time, Year-round Workers) with Projection for Pay Equity in 2059, *Institute for Women's Policy Research*, [online] Available at: <https://iwpr.org/publications/pay-equity-projection-1960-2017/> [Accessed 24 Dec. 2018].

Manovich, L., 2016. The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. *Journal of Cultural Analytics*, pp.1-15.

Mehdikarimi, S., Norris, S. and Stalzer, C., 2015. Regression Analysis of the Relationship between Income and Work Hours, *Georgia Institute of Technology*, [online] Available at: <https://smartech.gatech.edu/handle/1853/53299> [Accessed 21 Dec. 2018].

Salary.com., 2018. Eight Factors That Can Affect Your Pay, Salary.com, [online] Available at: <https://www.salary.com/articles/eight-factors-that-can-affect-your-pay/> [Accessed 20 Dec. 2018].

Stanley, T. and Jarrell, S., 1998. Gender Wage Discrimination Bias? A Meta-Regression Analysis. *Human Resources*, 33(4), pp.947-973.

Pandey, P., 2018. A Guide to Machine Learning in R for Beginners: Decision Trees, *Medium*, [online] Available at: <https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb> [Accessed 23 Dec. 2018].

Topping, A., 2018. Exclusive: #PayMeToo campaign launched to close gender pay gap, *the Guardian*, [online] Available at: <https://www.theguardian.com/news/2018/apr/02/exclusive-paymetoo-campaign-launched-close-gender-pay-gap> [Accessed 23 Dec. 2018].