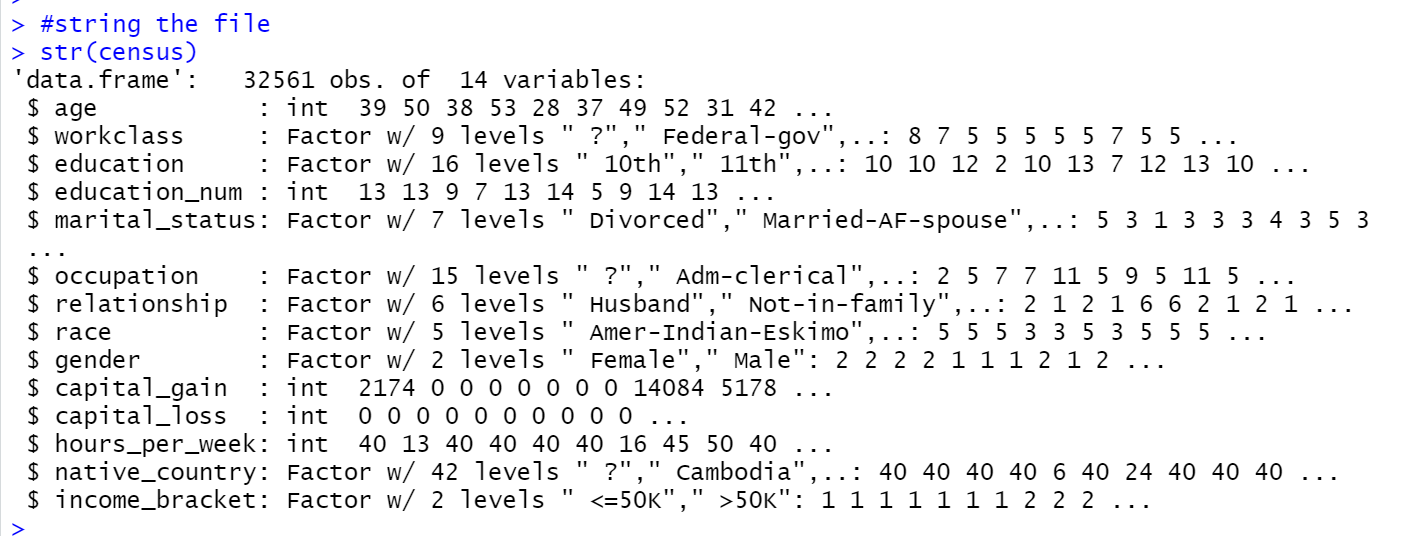
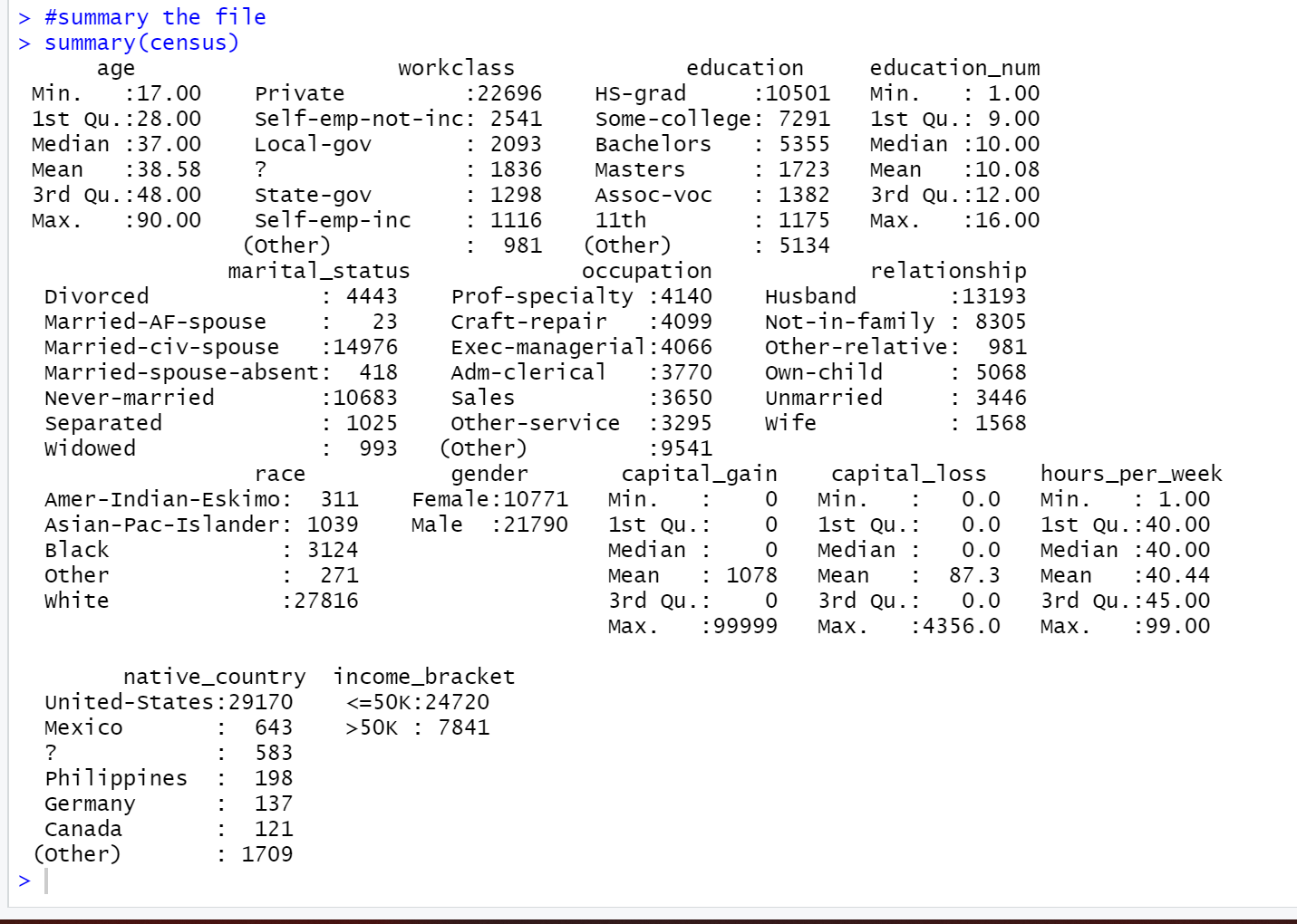
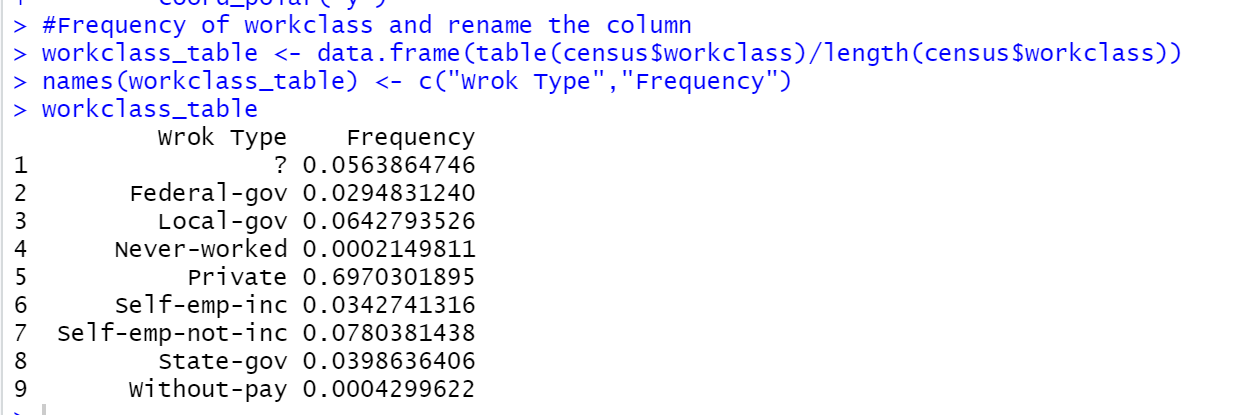
Yijun Wang

ALY 6010

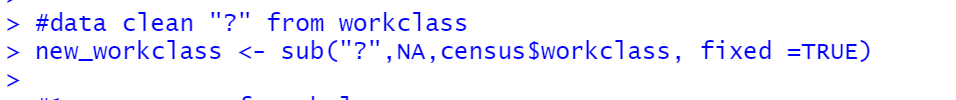


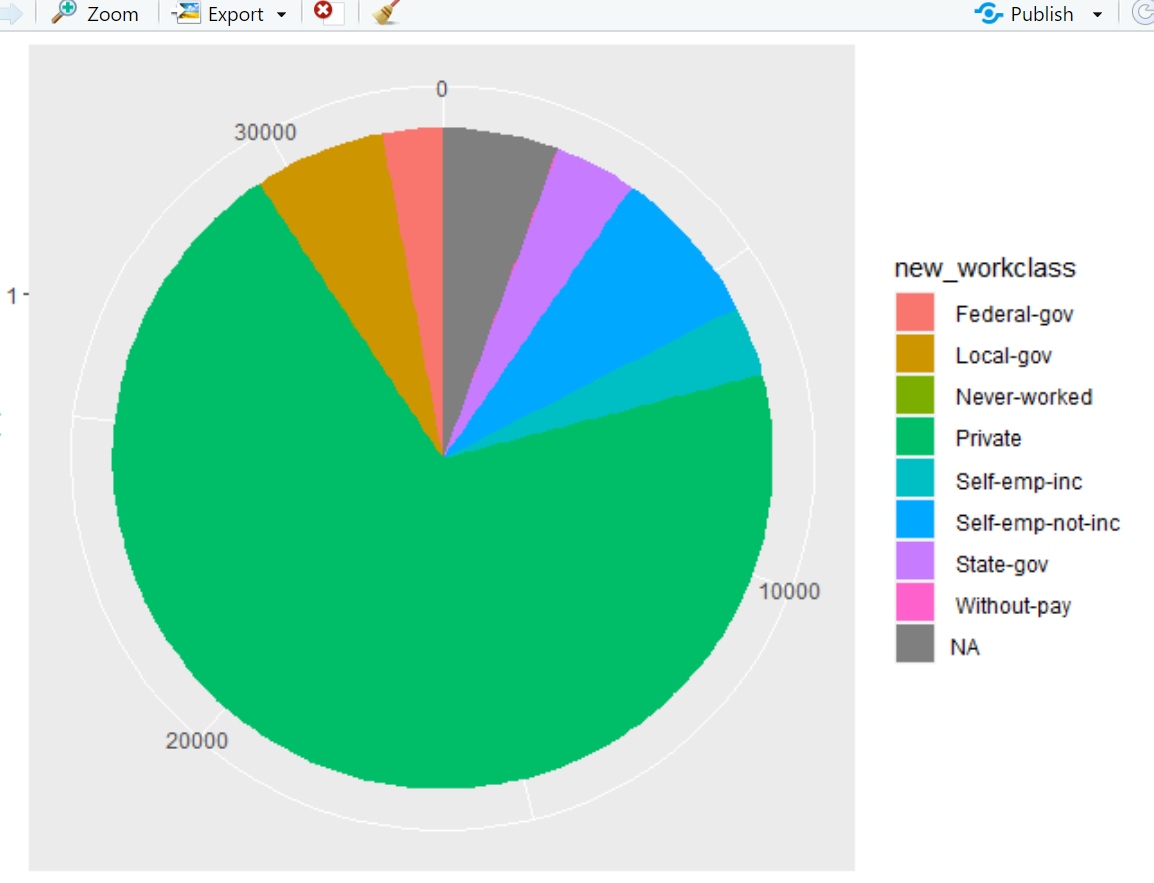
From the string of dataset, we know that there is 32561 obs. and 14 variables in this dataset. The goal of this study is to examine the dataset statistically and provide relevant, in-depth findings about the dataset, including data levels, descriptive statistics, and data distribution.



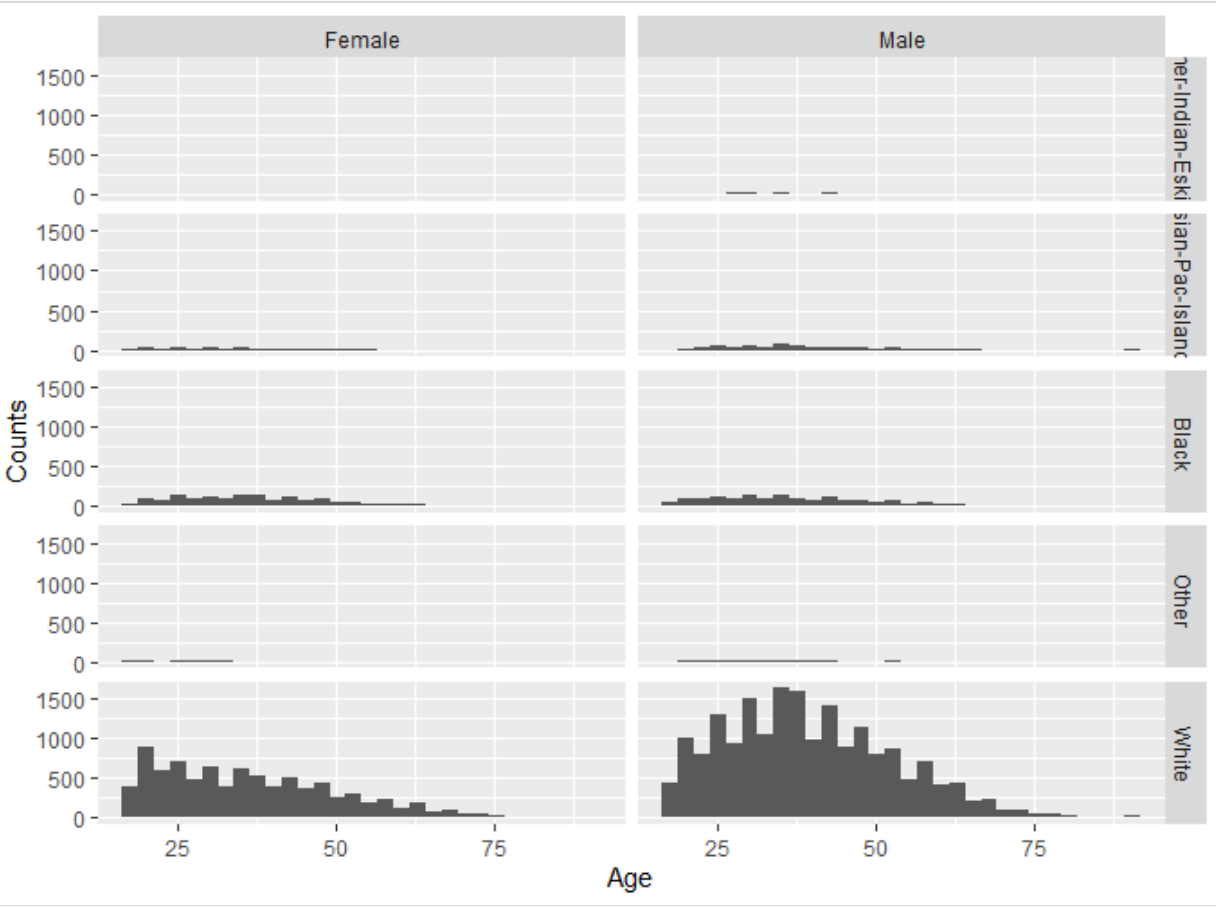


We can see there are 583 data are missing from the column of native country in summary, and the frequency of “?”has 5.6%. When different sets of data are compared, the data clearly shows the difference distribution of each group, which is highly useful for generating conclusions, so I will remove 583 data from the column of native country.

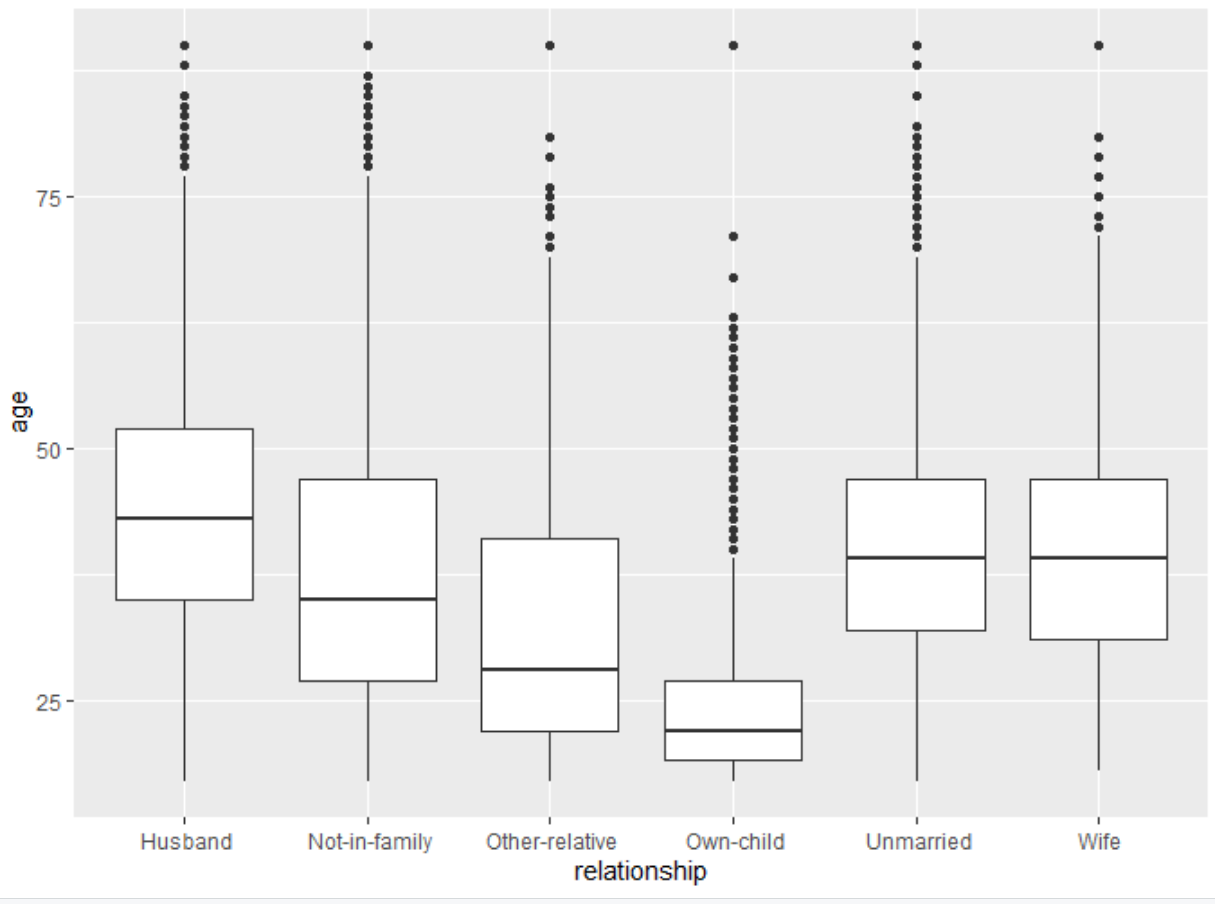




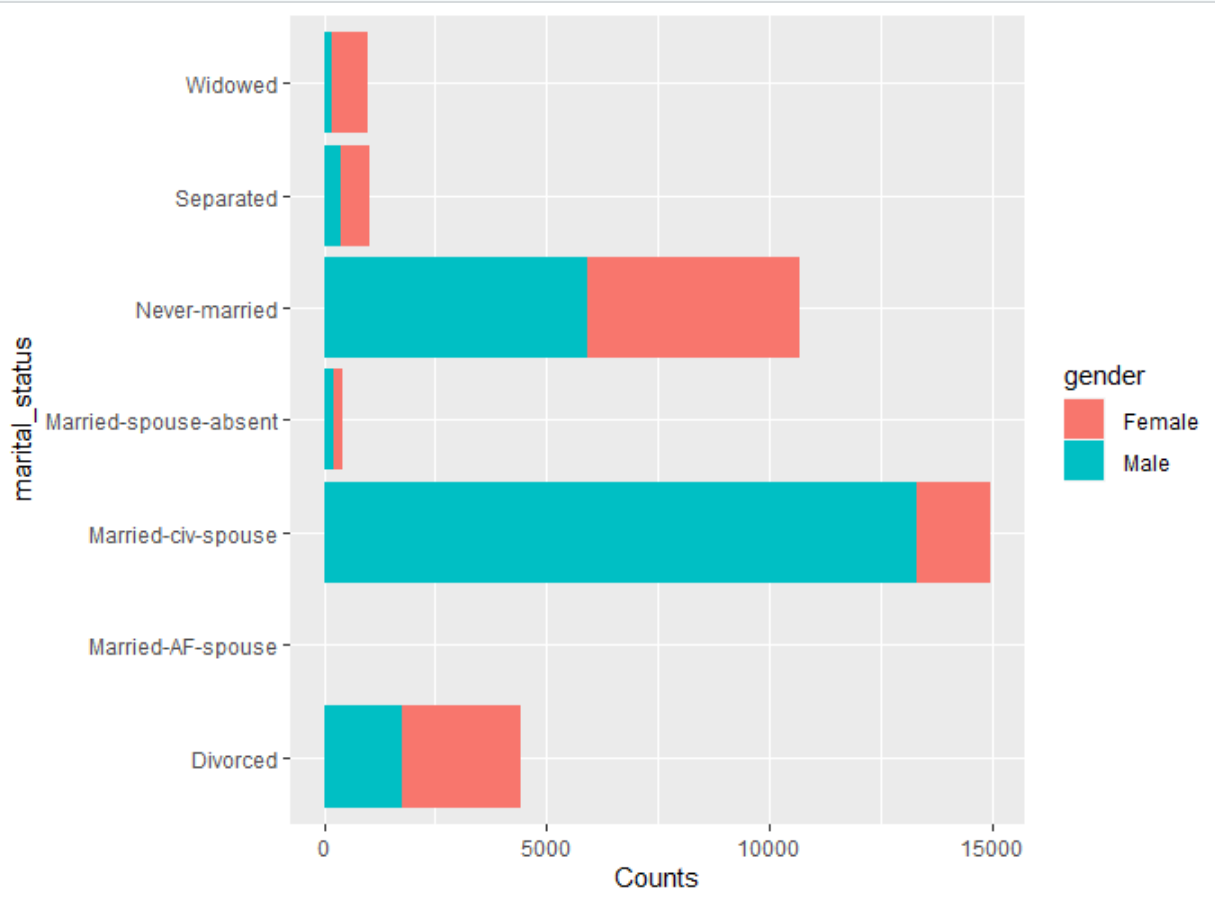
This is Workclass’s pie chart. Through this pie chart, we can demonstrate that the number of “Private” and “Self-emp-not-inc” is, in fact, the bulk of the dataset.



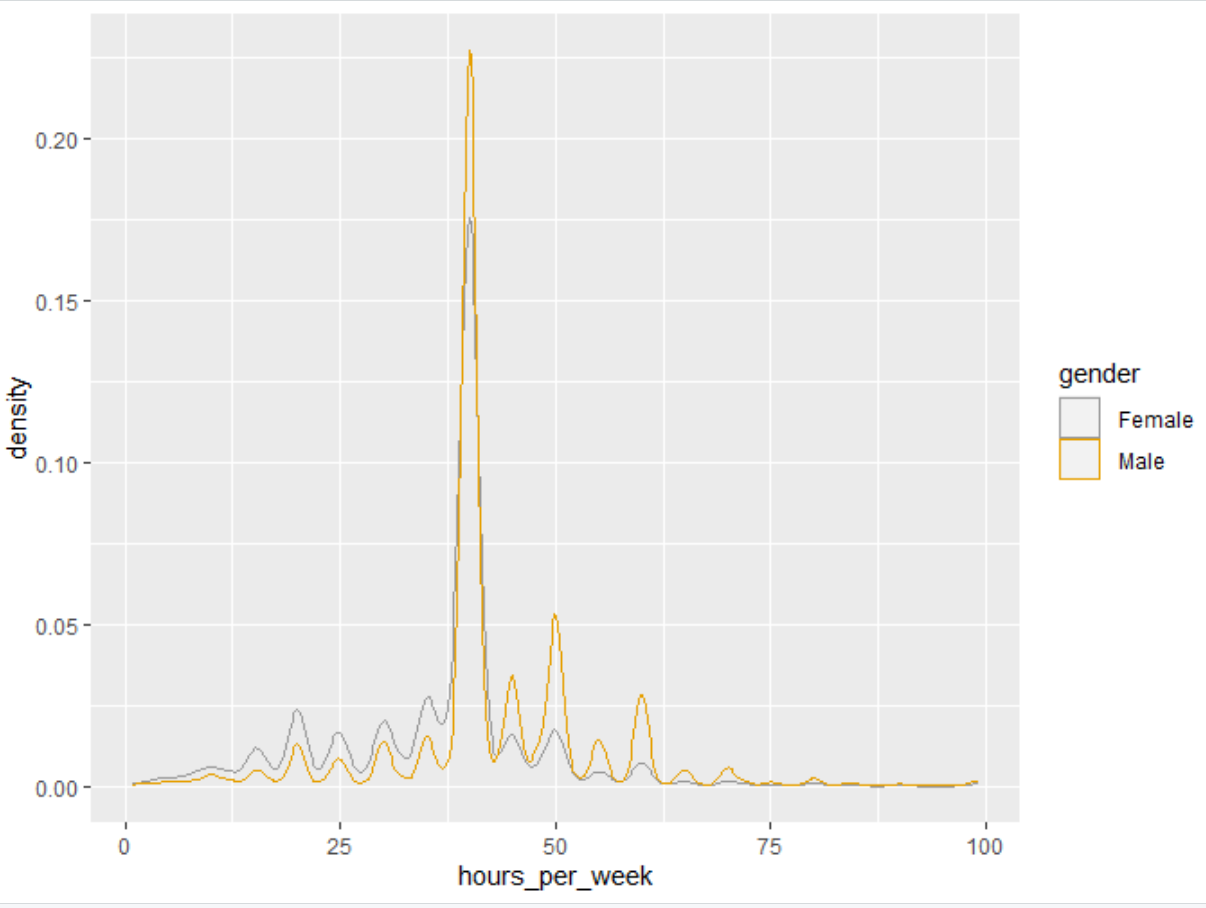
This chart is gender Distribution between age and Race. According to the histogram, the number of whites in this census data is the highest, followed by blacks, and Asian-Pacific-Islanders come in third. There are more white males than white women among them, especially those over the age of 40, with over 1,500. And Amer-Indian-Eskimo has the fewest women. Also, I can see that this is a normal distribution from Male of White, and White Female is right-skewed.



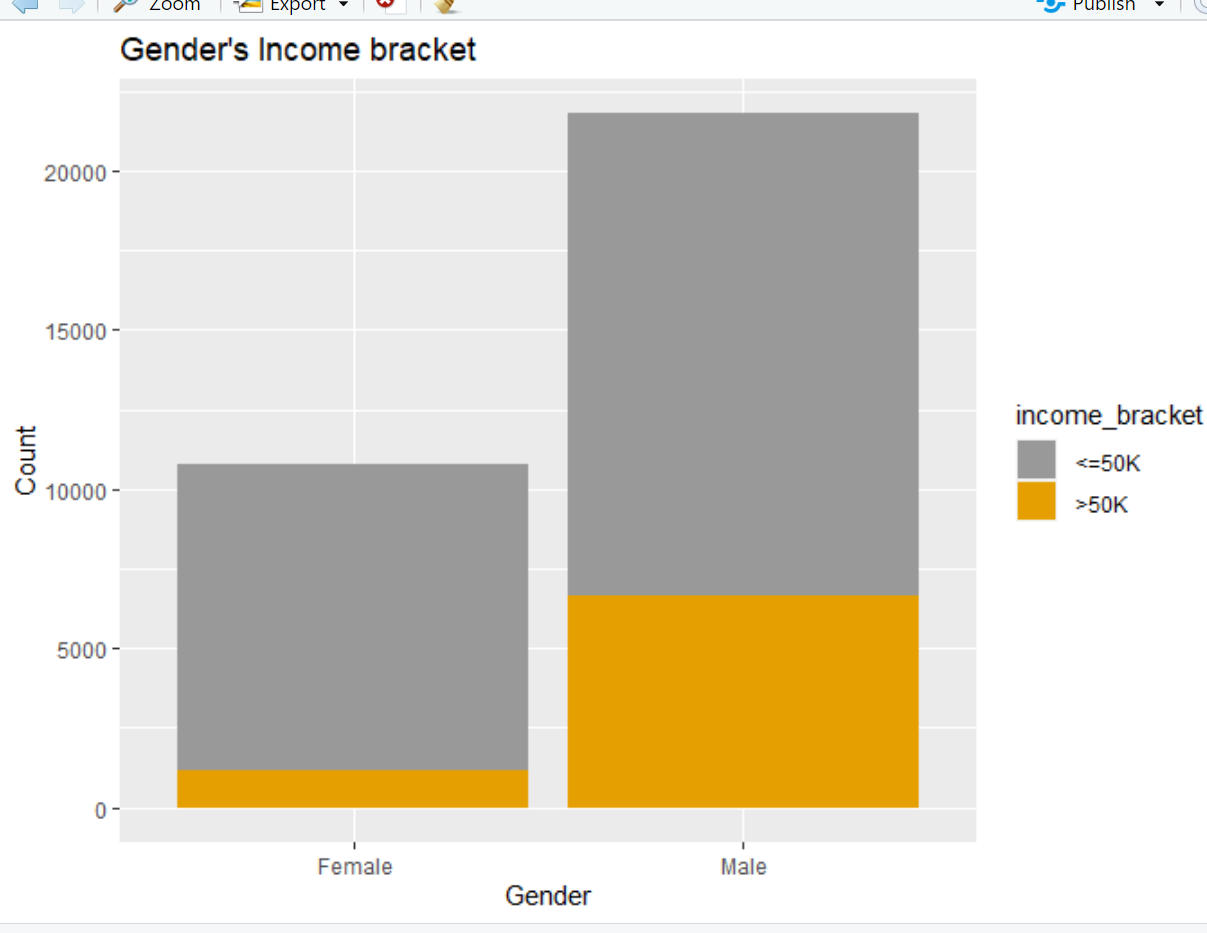
This visualization is a boxplot of age and connection. We can observe that there are some points outside the minimum and maximum on the age's boxplot, implying that there are points outside the range of the dataset, indicating that some outliers exist. The unique feature is that there is no outlier in these six groupings of associations that are less than boxplot min. Also, different relationship has different mean. “Own- child” has the lowest mean, this signifies that the average age of this group is low, and the average data of "Husband" is greater than that of other groups.



This visualization is Female and male’s Marital status. According to this graph, males have the largest proportion of "Married-civ-spouse", while females have the highest proportion of "Never-married." Regardless of the ratio of men to women, "Married-AF-Spouse" is the least.



This is density of working hours. Overall, regardless of gender, the density of 30-40 hours of labor each week is the greatest. More specifically, during this time period, men's working hours are longer than women's. In the 25-hour work week, however, females work more hours than males.



Gender's income level is in this range. Obviously, whether it is "=50k" or ">50k," males outnumber women. Males with "<=50K" have the greatest percentage, Female with “>50K” has the least number.

In short, males work more hours than women in general, hence men have greater ">50K" wages than women. White males, as well as "married-civ-spouse" men, made up the lion's share of the population at this period. Because the age of the "Wife" is largely steady, there aren't as many outliers in other categories.