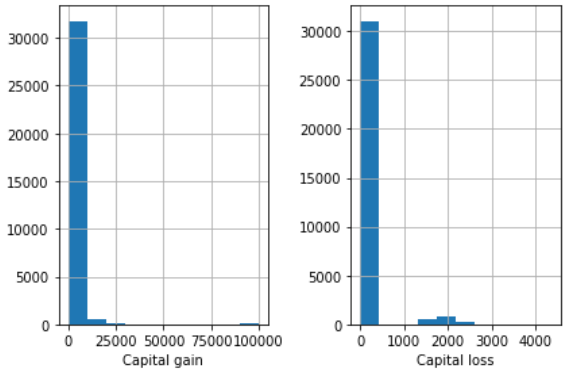
JSC270H1 - Assignment 2 Part II

Initial Data Exploration:

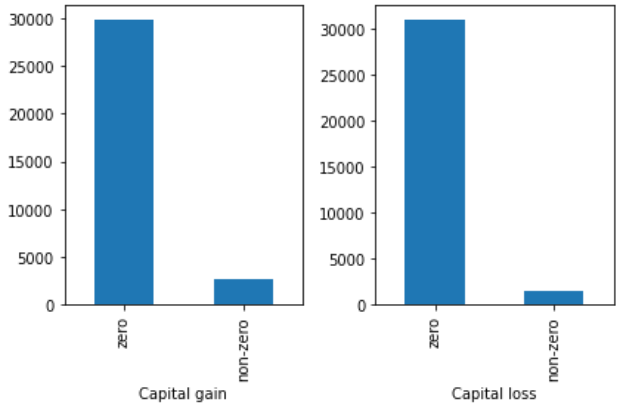
1. The columns that are numerical data (int64 in the above table) are 'age', 'fnlwgt', 'education_num', 'capital_gain', 'capital_loss', and 'hours_per_week'. In the text file, these columns are denoted as continuous data.

   All the other columns are categorical data. This is expected since variables like age, capital gain and loss, and hours per week are all described by numbers, whereas education, occupation, and race are better described by categories. Note that 'fnlwgt', the final weight, is indicated as continuous data in the text file description of the data.

2. Missing values are represented with ' ?'.

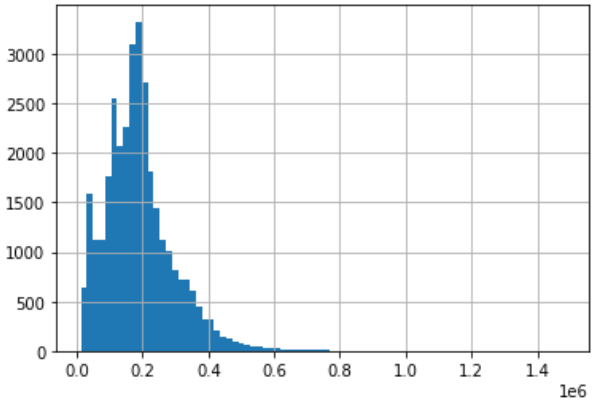
3. 

   I think these variables should be transformed to categorical variables. In the plots of capital gain and loss shown above, there seems to be many zero values in both capital gain and loss. As such, I will separate the zero values from the non-zero values in new categorical variables called “capital_gain_category” and “capital_loss_category”, which will have the values "zero" and "non-zero".

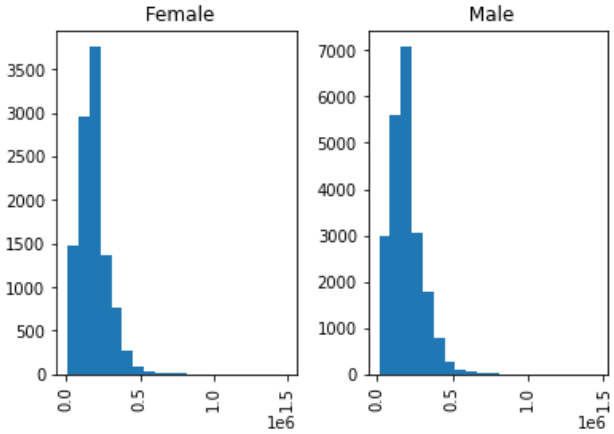


From the plots of “capital_gain_category” and “capital_loss_category” above, there appears to be significantly more zero values than non-zero values for the variables “capital_gain” and “capital_loss”.

4. Below is the histogram of the distribution of fnlwgt:



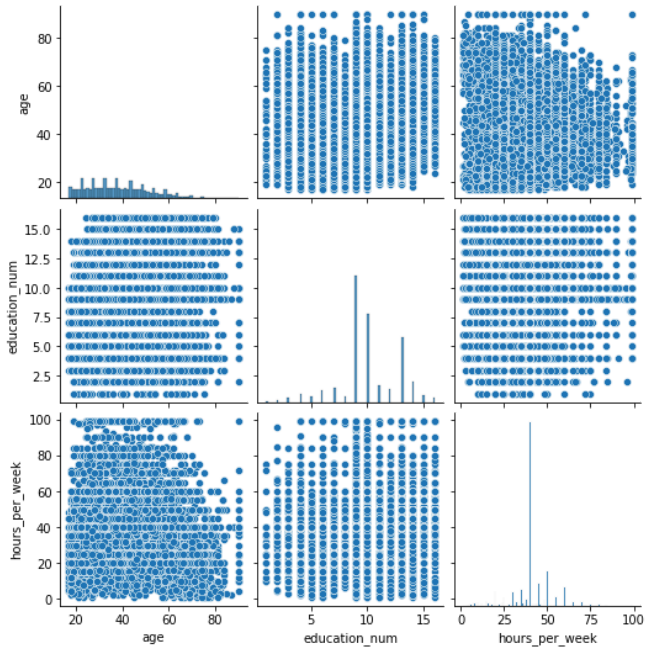
The variable fnlwgt is not symmetrically distributed; it appears to be right-skewed in the above graph. On the left side of the mode, there appears to be smaller modes, so the distribution is multimodal. On the right side, the distribution of values roughly follows the negative exponential graph.



The distribution of fnlwgt is identical between men and women, as shown in the above plots of fnlwgt for males and females. For instance, the individual distributions are both right-skewed. However, there is more data for fnlwgt for males than females.
There does not appear to be any significant outliers outside of the tails, and so they will not be excluded.

Correlation:

1. a) Below are the correlation plots for 'age', 'education_num', 'hours_per_week':



None of the variables appear to be correlated. Correlation between two variables is when adjusting the value of one variable leads to a change in the value of the other, and vice versa. As shown in the non-diagonal subplots, this does not occur since in each subplot, there is a variable whose value remains constant as the other variable varies. (This is shown by the vertical or horizontal lines in the subplots.)

b) The variable pairs with correlation coefficient > |0.1| are “capital_gain” and “education_num” (coefficient 0.122630), and “hours_per_week” and “education_num” (coefficient 0.148123).

The p-values for the Pearson test on both pairs are 0.000. This means we are very confident about the alternative hypothesis that the coefficients are different from 0. In other words, we are confident that the variable pairs are positively correlated. This is expected since it makes sense that the higher a person’s level of education is, the more likely they are to earn more and work hard (e.g., as determined by the number of hours they work). It could also be that someone who works more also earns more, like if they are paid by hourly wage, and this itself is correlated with their level of education.

c) For males, the coefficient is 0.060486 and the p-value is 0.000. For females, the coefficient is -0.017899 and the p-value is 0.063. Thus, the correlation between education_num and age is positive for males and negative for females. The coefficient for females has a higher absolute value, while the p-value for females (0.063) is also higher (above 0.05 but below 0.1), indicating less certainty of this correlation (the null hypothesis is that they are not correlated, while the alternative is that they are).

This means that we are more certain that education_num and age positively correlate for males than we are certain that they negatively correlate for females. This is somewhat expected, since older males are likely to have spent more years in education, such as by attaining master's or PhD degrees. Furthermore, notice that the dataset was created in 1994, a time when more women entered higher education compared to earlier generations. It might be possible that back then, there were many older women who were less educated than younger ones, thus creating the observed negative correlation.

d) The covariance matrix gives us the covariance between education_num and hours_per_week and the variance for each variable. The covariance between education_num and hours_per_week is 4.705338, which is positive and thus indicates that as the variables grow in the same direction (they either increase together or decrease together). The variance of education_num is 6.618890, while the variance of hours_per_week is 152.458995, which is significantly higher. This is likely because different people in different positions and jobs have different working environments, meaning they are likely to spend time differently, and also that some people would want to work longer for promotions or bonuses, for instance.

Regression:

1. a) Yes. The intercept, or the value for the estimated expected hours_per_week for females, is 36.4104. The coefficient for males is 6.0177, which means that compared to females (the intercept), the expected hours_per_week for males is 6.0177 higher than that for females. Thus, with a higher expected hours_per_week, males tend to work more hours.

   b) The general trend remains the same in that males still tend to work more hours than females, but the difference between the expected value for males and females is smaller (5.9709 for this model, compared to 6.0177 for the previous). The p-value for the coefficient of education_num is 0.000, indicating that it is statistically significant. The 95% confidence interval for the coefficient of education_num is [0.647, 0.748], while the coefficient itself is 0.6975.

   c) For the first model ('hours_per_week' ~ 'sex'), the coefficient of 'sex' is 6.0177. The interpretation for this is the same as in 1a).

   For the second model ('hours_per_week' ~ 'sex' and 'education_num' is the control), the coefficient of 'sex' for males (with education_num of 0) is 5.9709, with the data on females (with education_num of 0) as the intercept. This means that as we move from data on females to males who have not spent time in education, the expected number of hours worked in a week increases by 5.9709, indicating that such males tend to work 5.9709 more hours on average compared with such females.

For the third model with ‘sex’, ‘gross_income_group’, and ‘education_num’, the coefficient of ‘sex’ is 5.1010 for males (with <= 50K for gross_income_group and education_num of 0). As before, the data on females (with <= 50K for gross_income_group and education_num of 0) is the intercept. This means that as we move from data on females to males whose gross income is below 50K and who have not spent time in education, the expected number of hours worked in a week increases by 5.1010, indicating that such males tend to work 5.1010 more hours on average compared with such females.

The statistics that can help to determine the ‘best’ model are the adjusted R-squared and RMSE values. The closer the R-squared value is to 1, the better the fit is, while RMSE values should ideally be as low as possible (i.e., near 0).

The adjusted R-squared for the first model with only ‘sex’ and ‘gross_income_group’ is 0.053, for the second model (‘sex’ + ‘education_num') is 0.074, and for the last model (‘sex’ + ‘education_num’ + ‘gross_income_group’) is 0.094. They are similar, but since the value for the last model is the closest to 1, it is the best fit. Similarly, the RMSE value for the first model is 144, for the second is 141, and for the last is 138. Again, they are similar, but since the RMSE value for the last is the lowest, it is the best fit.

Overall, the model with ‘sex’ + ‘education_num’ + ‘gross_income_group’ is the best fit.