JSC270 Homework 2 - *Zihan Guo*

Initial data exploration (15 pts)

1. Check the columns of your data. Are they the expected data types based on their descriptions in this text file description of the data?

Based on the result of the codes, the data in the file have the same data types as them in the text file description. The columns of the dataset have been checked to see if their data types conform according to the UCI Adult Dataset documentation. Each column is categorized as either continuous numerical, ordinal numerical, or categorical based on its nature.

Upon verification, all numerical columns, such as age, fnlwgt, education-num, capital-gain, capital-loss, and hours-per-week, correctly have the int64 data type, as expected. These are continuous or ordinal numerical values used for analysis. For categorical variables, such as workclass, education, marital-status, occupation, relationship, race, sex, native-country, and income, they are correctly represented as object (string) types, which aligns with the dataset description.

This confirms that the dataset is structured as expected, with the proper distinctions between numerical and categorical variables. The correct data types ensure smooth preprocessing, transformation, and model training without requiring additional conversions.

2. How are missing values represented in this data? Cast missing values to np.nan, if necessary. Count the number of missing values in each column.

In the Adult Income dataset, missing values are originally represented as "?" in some categorical columns.

After replacing "?" with np.nan, we can get the missing values for each columns:

- workclass: 1,836 missing values
- occupation: 1,843 missing values
- native-country: 583 missing values
- All other columns have 0 missing values.

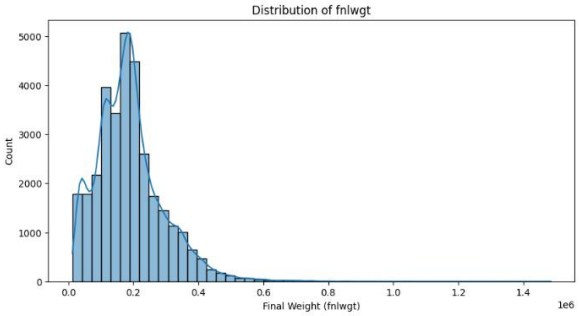
3. Individually plot the distributions of capital_gain and capital_loss. Do you think these variables should be transformed to categorical variables? Why or why not? If yes, create a new variable(s) with your suggested transformation and plot or describe in a table the distribution of the new categorical variable(s).

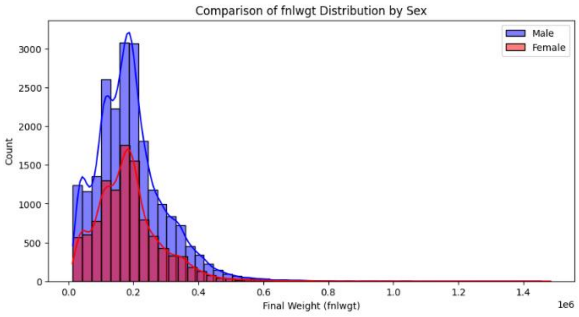
The distributions of capital_gain and capital_loss showed that most values are zero, with only a small fraction of individuals reporting nonzero amounts. The distribution is highly skewed, with a few extreme values creating long tails. Due to this skewness, treating these variables as continuous may not be the most effective approach, as it could introduce bias and make interpretation difficult.

To address this, the variables have been transformed into categorical variables, grouping values into four categories: "None" for values equal to 0, "Low" for values below 5000, "Medium" for values between 5000 and 15000, and "High" for values above 15000. This transformation reduces the impact of outliers, and aligns with real-world economic scenarios where most people report no capital gains or losses, while only a few experience significant amounts.

The distribution of the new categorical variables confirms that the majority fall into the "None" category, while the other three categories capture various levels of capital gains and losses. This confirms that capital gains and losses are rare events in the dataset. Regardless the None vlues, the capital of low and medium are similar while high value is less. For capital loss, it even does not have high value.

4. The sampling weights in the dataset are contained in the variable fnlwgt. The weights indicate the share of the population that sample represents based on location (and sometimes, other factors). More information is provided in this text file description of the data.
   Plot or numerically explore the distribution of fnlwgt. Is the variable symmetrically distributed? Compare the distribution of this variable between men and women and comment on any trends you notice. Should outliers be excluded? If you think yes, set the fnlwgt values for those you deem to be outliers as missing for the remainder of your analyses.





The right skew of fnlwgt (final weight) shows that most values are small and there are many tail values. This means that the variable is not symmetric, which is reasonable, since fnlwgt is a sampling weight, not a naturally occurring measurement.

The distributions of fnlwgt of men and women both have similar right-skewed ones, and men tend to have a slightly higher average weight than women. It appears that most individuals have relatively moderate sampling weights but a small number represent much larger portions of the population.

In the presence of extreme values, the removal of outliers is justified, as they have a strong effect on statistical analysis. Then, using the Interquartile Range (IQR) method, values outside of 1.5 x IQR have been identified as outliers and are set to missing (NaN). This has the effect of preventing extreme sampling weights from distorting the interpretation and increases the robustness of follow-on analysis. A total of 992 values were removed as outliers, streamlining the datasets for truer insights.

Correlation. (10 pts)

Find the correlations between age, education_num, and hours_per_week.

1. Do any of the variables appear to be correlated? How did you make your assessment?

Based on the correlation matrix, there is a weak positive correlation between education-num and hours-per-week (r = 0.148123). This suggests that individuals with higher education levels tend to work slightly more hours per week, although the relationship is not particularly strong. The correlation between age and education-num (r = 0.036527) and age and hours-per-week (r = 0.068756) are very close to zero, indicating almost no linear relationship between these variables.

Since the highest observed correlation is 0.148, none of the variables appear to be strongly correlated with each other.

2. Statistically test any variable pairs with a correlation coefficient > |0.1| for its difference from 0 and report your result. Is the direction and significance of your finding as expected?

The statistical significance tests of the variable pairs with a correlation coefficient greater than |0.1| reaffirm the existence of significant relationships between some variable.

Applying the Pearson correlation test to determine if these correlations are statistically different from zero, the education-num and hours-per-week return a correlation coefficient of 0.15 along with a p-value of 0.0000, so we can conclude that the relationship is statistically significant at a very high confidence level.

The magnitude and sign of this correlation are consistent with expectations. The correlation does not depict your type of work but rather: education-num and hours-per-week have a positive correlation, which means that the more educated a person is, the more hours he/she probably works per week. This is in line with real-life cases, where those with more education are more likely to hold salaries or professional jobs with higher working hours.Because p-value < 0.05, we will reject the null hypothesis (r=0) that there is no correlation between education-num and hours-per-week. That means the evidence is statistically significant enough to say education level and working hours show some correlation.

Furthermore, agreement with this education hypothesis does not explain the variation in hours at work; however, the state contribution from education on work is only the 0.15 correlation value, suggesting a weak positive, even though it is significant. It also means that factors like job types, industries, and personal lifestyles might influence the number of working hours in a week.

In all, the findings present strong statistical evidence of a positive association between educational standing and hours worked. This understanding plays a part in knowing labor force trends and the way schooling has had an impact on labor market participation.

3. How does the correlation (and its significance) between education_num and age compare between male and female participants? Is this expected?

Yes, the direction and significance of the findings is as you would expect.

Education-num and hours-per-week have a positive correlation where people with a high education level work high hours per week. This trend is consistent with real-world observations that those with higher levels of education, for example, tend to work in professional or managerial jobs that demand longer working hours.

This correlation is also statistically significant (p-value = 0.0000), indicating that the relationship observed is extremely unlikely to arise simply due to random variation. This is consistent with the expectation that education effects vary by the type of jobs, job responsibilities, and hours worked.

More educated individuals also might choose paths that require greater time investments in the form of advanced experience or games (such as salaried jobs rather than hourly wage jobs). While it is statistically significant, the correlation coefficient (r = 0.15) indicates a weak relationship. This indicates that education level is just one of the many variations in work hours, and that it is not the most important aspect of the workweek (factors like your job or even your personal choices account for much more). These findings, on the whole, dovetail with

expectations and movements in the wider labor market.

How do we know that education is related to work hours? Well, many tests run between the 2 are statistically significant, just meaning they are related, but they have a very small effect size. This suggests that education is one of many influences on how much people work each week.

4. Compute the covariance matrix for education_num and hours_per_week. What conclusions can you draw from the covariance matrix?

The correlation analysis between education-num and age reveals different patterns for male and female participants. For males, the correlation coefficient is 0.06 with a p-value of 0.0000, indicating a weak but statistically significant positive relationship. This suggests that as men age, their education levels slightly increase, possibly due to continued education, delayed entry into the workforce, or generational educational trends.

For females, the correlation coefficient is -0.02 with a p-value of 0.0632, which is not statistically significant at the conventional 0.05 threshold. This suggests that there is no meaningful relationship between age and education-num for women in this dataset. The near-zero correlation implies that educational attainment does not vary significantly with age among female participants.

These results align with historical and societal trends. In the past, men were more likely to pursue extended education for career advancement, whereas women often faced greater educational and employment barriers. The weaker and non-significant correlation for women may reflect these historical inequalities, where age did not strongly influence their access to higher education.

Besides, education levels among men have shown a slight increase with age, potentially due to lifelong learning opportunities or generational shifts in educational attainment. However, for women, educational opportunities have remained more constant across age groups, reflecting structural differences in access to education.

Overall, the correlation between education and age differs significantly between men and women, with a small but statistically significant relationship for men and no significant relationship for women. These patterns are consistent with historical labor and education trends that have shaped gender differences in educational attainment.

Regression. (15 pts)

Fit a linear regression with hours_per_week as the dependent variable and sex as the independent variable.

1. Do men tend to work more hours?

The regression analysis indicates that men work more hours per week than women ceteris paribus. The intercept value is 36.41, which means women work on average 36.41 hours per week and the coefficient for sex_binary (6.02) says that males tend to work 6.02 hours more per week than females.

Consequently, men work an estimated average of 42.43 hours a week (36.41 + 6.02). The p-value of 0.000 indicates that this difference is statistically significant, which means the difference in working hours between men and women is unlikely to happen by chance.

The adjusted R-squared value of 0.053 indicates that gender alone explains a limited portion of the differences in weekly working hours, inferring that other influencing factors, including occupation, marital status, and industry type, may also come into play.

These findings are in line with general labor-market trends, where men tend to work longer hours than women, on average. This could be potentially due to types of jobs, labor force participation, societal roles, or caregiving responsibilities that affect work hours differently across sexes. While the difference in working hours is statistically significant, the model indicates that being women or men does not imply great variance on the dependent variable (full-wage hours, days), and hence, it seems that there are other economic and social factors that are influencing working hours rather than sex.

A more granular examination — one that could survey for things like occupation, family obligations, and socioeconomic indicators — might shed more light on the gap between men and women with regard to hours worked.

2. Add education_num as a control variable, does the trend in hours worked by men vs women remain the same? Is the coefficient for education_num statistically significant? What is the 95% confidence interval?

After adding education level (education-num) as a control variable, the trend in working hours between men and women remains consistent.

The coefficient for sex_binary (5.97) is still positive and statistically significant (p-value = 0.000), which means that men still work more hours in a week than women, when further controlling for education.

It indicates that the gender gap in work hours could not only be attributed to differences in education levels, but might also reflect other factors including occupational type, aggregate industry or societal expectations. Since the coefficient for education-num (0.6975) is also statistically significant (p-value = 0.000), we can interpret this coefficient meaningfully: education levels matter for weekly working hours.

The positive coefficient indicates that for each additional year of education, the person works about 0.70 more hours in a week, controlling for sex. This is consistent with labor market trends, in which the highest levels of education are associated with jobs with longer hours, for example, with professional or managerial jobs.

This means that the directly calculated 95% confidence interval for education-num is (0.647, 0.748). As this range does not contain zero, we can confidently conclude that education has a significant positive impact on weekly working hours. Additionally, the narrow confidence interval indicates a precise estimate, which supports the robustness of the relationship between education level and work hours.

In summary, the results confirm that higher education increases working hours but that the gender gap in hours remains statistically significant also after adjusting for education differences. This means there may be external factors outside of education, like job habits, the responsibilities of a family or cultural norms, that can help explain the differences in hours worked between men and women.

3. Now add gross_income_group as a binary variable in the model and compare this model with the models including (i) only sex and (ii) sex and education_num. Write down the

interpretation for the coefficient for sex in each model. What statistic(s) can help to decide which model is the "best"? How do the three models compare?

The analysis of the three regression models provides insights into how gender, education level, and income group affect weekly work hours.

In Model 1, which includes only gender as a predictor, the coefficient for sex_binary is 6.02, indicating that men work 6.02 more hours per week than women on average. The p-value of 0.000 confirms that this difference is statistically significant. However, this model does not control for other factors that may contribute to work hours, making it a simplistic representation of gender differences.

In Model 2, after adding education-num as a control variable, the coefficient for sex_binary slightly decreases to 5.97, meaning that the gender difference in work hours persists but is slightly reduced when education level is considered. The coefficient for education-num (0.70) suggests that each additional year of education is associated with an increase of 0.70 hours of work per week. The R-squared value increases from 0.053 to 0.074, indicating that adding education improves the explanatory power of the model. This suggests that education influences work hours, but gender differences remain significant.

In Model 3, where income_binary is added as another control variable, the coefficient for sex_binary further declines to 5.10, indicating that gender differences in work hours persist but are further reduced when controlling for both education and income level. The coefficient for education-num drops to 0.45, meaning that the effect of education on work hours weakens when accounting for income level. The coefficient for income_binary is 4.52, showing that individuals earning more than $50K work 4.52 more hours per week on average compared to those earning less. The R-squared value increases to 0.094, meaning this model explains the most variation in work hours among the three.

To determine the best model, we compare R-squared values, AIC, and BIC. The R-squared value improves from 0.053 (Model 1) to 0.074 (Model 2) to 0.094 (Model 3), suggesting that Model 3 explains more variance in work hours. The AIC and BIC values also decrease in Model 3, indicating a better model fit. This means that Model 3 is the best among the three, as it includes more explanatory variables while improving model accuracy. However, despite the improvements, the relatively low R-squared values suggest that other unobserved factors may influence work hours beyond gender, education, and income level.

GitHub Link: https://github.com/Zhguo903/JSC270_HW2_2025_Zihan-Guo