Part II - Data Analysis

Initial Data Exploration

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

The data types for this data frame are generally correct as per the column description in the text file. More specifically, categorical variables have type `object` and continuous variables have type `int64`. The only thing I can argue is that we might convert some of the numerical variables into type `float64` since they are continuous, but this is not necessary because we have integer values for all those variables in all rows.

2) How are missing values represented in this data? 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.

From the text file description, missing values are marked with `"?"`. There are 1836 missing values in the column `workclass`, 1843 in `occupation`, and 583 in `native_country`. Overall, there are 2399 rows with missing values. This is written in the text file description (i.e. 32651 before removing missing values and 30162 after). The use of the `len` function confirms with this description.

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).

Since `capital_gain` and `capital_loss` are numerical variables, we can use a histogram to plot the distribution.

It's very interesting that `capital_gain` and `capital_loss` have the majority of the data at 0. This is probably an indication of missing data in these numerical variables. Hence, I think it is helpful to convert these numerical variables into categorical variables by classifying them into a specific range. The following tables show the categories of the newly created variables and the counts, obtained from calling `value_counts()`.

`capital_gain_group`	count
[0, 5000)	30913
[5000, 10000)	878
[10000, 15000)	157
[15000, 20000)	360
[20000, 25000)	38
[25000, 30000)	49
[30000, 35000)	5

`capital_gain_group`	count
[40000, 45000)	2
[95000, 100000)	159

`capital_loss_group`	count
[0, 500)	31053
[500, 1000)	25
[1000, 1500)	100
[1500, 2000)	1058
[2000, 2500)	281
[2500, 3000)	33
[3000, 3500)	2
[3500, 4000)	6
[4000, 4500)	3

The tables confirm with the preprocessed data that there are a lot of rows with values less than 5000 for `capital_gain` and 500 for `capital_loss`. However, by abandoning specific values and adopting groups, the analysis might become more suggestive because all groups are more meaningful and do not have a value of 0 like before.

4) 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.

From the histogram and the boxplot, it is clear that the data for `fnlwgt` is right-skewed and not symmetrically distributed, with some observations having large `fnlwgt` values. The boxplot also identified many outliers using the $1.5\times IQR$ rule.

From the graphs that differentiate male and female, we can see no obvious distinction between the distribution of fnlwgt for male and that for female. Both distributions are right-skewed with a peak centered at around 200000. Both exhibit some outliers as per the $1.5\times IQR$ rule according to the boxplots. Overall, I would not exclude the outliers identified by the boxplots from my analysis. The main reason is that the outliers are not individual. In other words, there are many points outside the maximum whisker, not just one. Thus, it is a systematic pattern that `fnlwgt` is outside the *normal* range, so we shouldn't exclude some values just because we want the analysis to look good.

Correlation

1) Find the correlation between `age`, `education_num`, and `hours_per_week`.

	age	education_num	hours_per_week
age	1	0.036527	0.068756
education_num	0.036527	1	0.148123
hours_per_week	0.068756	0.148123	1

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

No, these 3 variables do not appear to be correlated. I have constructed a correlation table and a pairplot above to assess this. From the correlation table, we see that the correlation between each pair of the variables is smaller than 0.15, which indicates a linear correlation is very weak among the variables. Additionally, the above pairplot confirms with this result. We can see that the data points are randomly scattered on the plots, and no discernable pattern can be detected. This is an indication of no correlation.

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

From the correlation table, the only pair of variables with a correlation coefficient $>|0.1|$ is the pair of `education_num` and `hours_per_week`. We use the `pearsonr` function in the `scipy.stats` library to conduct the test. We set $\alpha=0.05$. From the test, the r-value is 0.148 (rounded to 3 decimals), confirming with the value in the correlation table, and has a corresponding p-value of 4.237×10^{-159} (rounded to 3 decimals). This p-value is very small, indicate that there is a correlation between `education_num` and `hours_per_week`. This result is not expected. As per the scatterplot of `education_num` vs. `hours_per_week`, the points are very scattered and show no discernable pattern. A possible explanation for this small p-value is that the large sample size magnified the significance by reducing the standard error. Nevertheless, even though there might be a correlation, the correlation is weak.

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

From the result, the r-values for male and female are not similar. Specifically, the correlation for male is positive and the correlation for female is negative. At the significance level $\alpha=0.05$, the p-value of 4.023×10^{-19} for male shows that the correlation is statistically significant, whereas the p-value for female is 0.063, indicating that the existing correlation is probably due to chance. Nevertheless, as mentioned above, the large sample size magnified the significance by reducing the standard error. Since the absolute values of the correlations are very small, or smaller than $|0.1|$, I would say that this difference is tolerable and expected, and that there is probably no correlation between `education_num` and `age` for both male and female.

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

	education_num	hours_per_week
education_num	6.61888991	4.70533794

	`education_num`	`hours_per_week`
`hours_per_week`	4.70533794	152.45899505

From the covariance matrix above, the variance for `education_num` is 6.619, the variance for `hours_per_week` is 152.459, and the covariance between the two variables is 4.705. This covariance is rather small, from which we can conclude that the correlation between the two variables is small, confirming with our conclusion in parts (a) and (b). Nevertheless, we also see a rather large variance from `hours_per_week`, which might need special attention if we were to fit a regression model using this variable.

Regression

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

a) Do men tend to work more hours?

Yes, men tend to work more hours. From the summary of the linear regression model, men work 6.0177 hours more than women on average. This result has a t-value of 42.510 and a very small p-value correspondingly. In other words, we reject H_0 and conclude that the difference in working hours between men and women is significant.

b) 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?

From the summary of `model2`, the trend in hours worked by men vs. women remain the same. After adding `education_num`, the regression coefficient for `sex` is still 5.9709, with a t-value of 42.653 and a very small p-value. This is an indication that `sex` is still a significant predictor for `hours_per_week`. The coefficient for `education_num` is 0.6975, and it is also statistically significant, with a t-value of 27.244 and a very small p-value. The 95% confidence interval for `sex` is $[5.697, 6.245]$ and the 95% confidence interval for `education_num` is $[0.647, 0.748]$. The confidence interval for `education_num` does not include 0, which confirms that the coefficient is statistically significant.

c) 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?

First of all, all three models have very similar coefficients of existing variables. For the variable `sex`, the coefficients for all three models are about 5 to 6. The coefficients of `education_num` for `model2` and `model3` are also very similar. All three models have very small p-values for the coefficients, indicating the coefficients are all statistically significant in predicting `hours_per_week`.

To interpret `sex` in `model1`, on average, men work 6.0177 hours more than women. This value can also be seen as the difference between the mean of working hours for men and the mean of working hours for women in this sample. In `model2`, the coefficient for `sex` is 5.9709 and can be interpreted as that men typically work 5.9709 hours more than women, given that they have the same value of `education_num`. The `sex` coefficient in `model3` can be interpreted in a similar way: Keeping `education_num` and `gross_income_group` the same, it is predicted that men work 5.1010 hours more than women on average.

Several factors in the summary tables can help us decide which model is the best, such as R^2, adjusted R^2, AIC, and BIC. By adding more predictors, the value of R^2 increases from 0.053 to 0.074 to 0.094. The values of adjusted R^2 in all three models are also the same as the respective R^2 values up to 3 decimals. From the value of R^2, we see that 9.4% of the variability in working hours can be explained by `model3`, which is the highest among the three models. For AIC and BIC, we look for smaller values, in which case `model3` is still the best by having 2.529×10^5 as the values for both AIC and BIC. Since AIC and BIC account for some penalty in the number of predictors, and `model3` still has the smallest values, we conclude that `model3` is the best model among the three.

Nevertheless, a R^2 value of 0.094 is still very low, and 2.529×10^5 as the AIC and BIC is a huge value. From these, we conclude that all three models poorly predict the number of working hours. In other words, `sex`, `education_num`, and `gross_income_group` might not be the best predictors in this situation.

Running a partial F-test between `model2` and `model3`, we get an F-value of 741.409 and a corresponding p-value of 1.914×10^{-161}. This result also confirms with the conclusion above that `model3` is the best because removing the variable `gross_income_group` makes the linear model less useful. Similarly, the partial F-test between `model1` and `model2` shows a statistically significant result as well. This means that `education_num` is also a significant variable in the linear model.