**PART II - Data Analysis**

**Link to my Github Repo**: <https://github.com/ChloeH88/JSC270_HW2_2022_CHuang.git>

Note: All answers in this pdf can be found in the colab file.

**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](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names)?

Answer: all columns are expected data types,

as all continuous attributes are int64 and the categorical ones are object.

1. 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.

Answer: Missing values are in categorical variables, and are represented by ' ?'

Number of missing values for workclass: 1836

Number of missing values for occupation: 1843

No other missing values.

1. 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).

Answer: I think we can convert these two variables into categorical ones,

because approximately 50% of the values are 0, so I decided to have 2 categories: 0 and not 0.

Also, after dividing them up into categories/labels,

we can see more clearly how many counts are in each group.

And making conclusions based on categories(has capital gain/ no capital gain group)

makes the data more understandable to audiences.

But, there are definitely some drawbacks, like

we lose some precision by ignoring the distinction between different numeric values (e.g. 100000 and 10000 capital\_gain), and cannot apply numerical statistical methods on them.

1. 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](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names).

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.

Answer: fnlwgt is not symmetrically distributed; it's right skewed

**Compare distribution:**

The shape of distribution between men and women are quite similar, and the median and mode are also similar.

However, noticing the y-axis, men's weight is approximately 1.5 times higher than women's.

**Outliers:**

I do think outliers needed to be included, because weight is artificially assigned to help us decide if two people have similar demographic characteristics

**Correlation. (10 pts)**

1. 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?

From the pairwise scatter plot and correlation matrix, since **\*\*education\_num and hours\_per\_week\*\*** have correlation coefficient as 0.148, they seem to have a **\*\*very weak positive correlation.\*\***

* 1. 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?

From the Pearson correlation test, we know the correlation coefficient (0.148) shows there is a very weak positive correlation. While the extremely low p-value confirms the fact that there is indeed a linear correlation between education\_num and hours\_per\_week.

* 1. How does the correlation (and its significance) between *education\_num* and *age* compare between male and female participants? Is this expected?

From the coefficients: for man, as age goes up, education\_num increases, whereas for women, education\_num decreases as age goes up.

However, if we set the significance level at 5%, the 0.063 p-value for women leads to the conclusion that there may be no correlation between education\_num and age for women. While the very low p-value for men shows there may be some association between education\_num and age for them.

The result is sort of expected. For men, they may have more money to pursue their academic path as their age increases. Yet, for women, it really depends on their socioclass, financial situation and many other social/cultural factors. Those factors may push women to take care of their home/child instead of persuing education. Because of diverse social/cultural difference, we can hardly conclude association between women's age and education\_num.

* 1. Compute the covariance matrix for *education\_num* and *hours\_per\_week*. What conclusions can you draw from the covariance matrix?

Positive covariance indicates the direction of relationship between varialbes, thus education\_num and hours\_per\_week are positively correlated. But covariance doesn't give us the strength of the correlation, and it doesn't account for difference in unit between variables.

**Regression. (15 pts)**

1. 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?

Yes, because the coefficient for sex (6.02) represents the estimated average difference in hours\_per\_week between men and women. And since 'Male' is the non-baseline variable, Male = 1 and women = 0. Thus, hours\_per\_week for men \*\*minus\*\* estimated hours\_per\_week for women = 6.02 > 0, meaning men tend to work more hours.

And since p-value is very low (approaching 0), this coefficient is statistically significant.

* 1. 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?

(1) Trend remains the same (men works more hours than women), since the coefficient for sex is 5.97 > 0.

(2) The coefficient for education\_num is statistically significant as its p-value is extremely small.

(3) The 95% confidence interval for the coefficient for education\_num is (0.647, 0.748)

* 1. 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?

**(1) Interpretation of coefficient for sex:**

* **(i) only sex**: the estimated average difference in hours\_per\_week between men and women, i.e. men's hours\_per\_week **-**women's hours\_per\_week = 6.02.
* **(ii) sex and education\_num:** the estimated average difference between men's hours\_per\_week and women's (actually men's - women's) is 5.97, when education\_num = 0
* **(ii)** **sex + education\_num + gross\_income\_group:** the estimated average difference between men's hours\_per\_week and women's (actually men's - women's) is 5.1, when education\_num = 0 and gross\_income\_group <= 50k.

**(2) What statistics help understsand which model is the best?**

R squared, RSE, and RMSE can all help determine which model is the best.

**(3) How do models compare?**

As we see that RSE and RMSE for model 3(model with sex + education\_num + gross\_income\_group) is the smallest, meaning model 3 is the best model interms of accurately predicting the response variable. And its R-squared is the largest from the OLS regression results above, indicating this linear regression model fits the observed data the best. Hence, I argue that model 3 is the best.