JSC270 Assignment 2 Answers Xiang Gao

Link to the GitHub repo with the python notebook:
https://github.com/xiangtgao/JSC270_Assg2/blob/xiang_branch/JSC270_Assignment2.ipynb

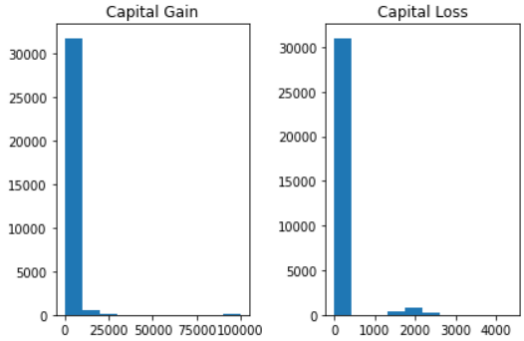
Initial Data Exploration.

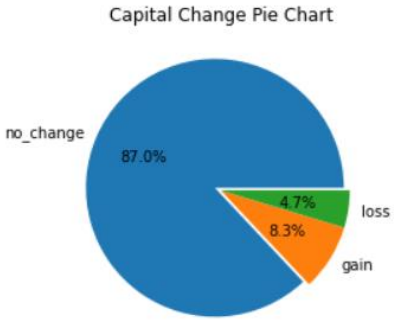
1. The data types match with the descriptions. The numerical data described in the file has int64 as dtype, and the other categorical data has object as dtype.

2. The missing values in the dataset was originally represented with '?'. Since there are no misspecified data types, I do not need to cast any types. Also, the missing data is now represented as np.nan; the number of missing values is shown on the right. workclass has 1836 missing values, occupation has 1843 missing values, and native_country has 583.

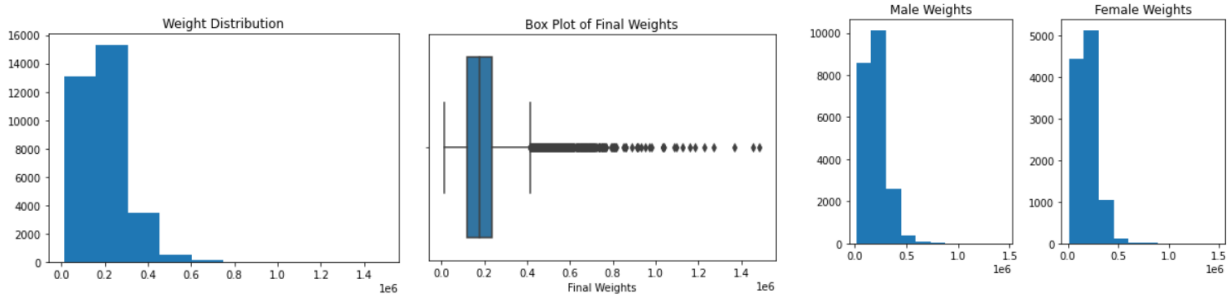
```
age                    0
workclass           1836
fnlwgt                 0
education              0
education_num          0
marital_status         0
occupation          1843
relationship           0
race                   0
sex                    0
capital_gain           0
capital_loss           0
hours_per_week         0
native_country       583
gross_income_group     0
```

3. The plots for distributions of capital gain/loss are shown below. These variables should be transformed since they are essentially telling one thing: the net change in capital for person, and their distributions are extremely left skewed. I combined these two variables into one categorical variable that consists of three categories: no change, gain, and loss. In this way, I created a pie chart that shows percentage of people in each category. This pie chart is also shown below.



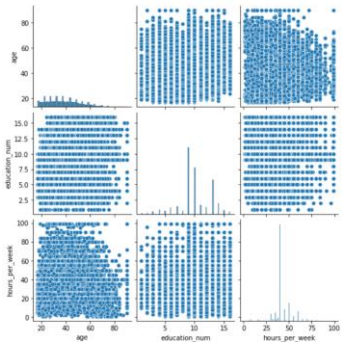


4. The distribution of fnlwgt is not symmetrically distributed bur rather left skewed. The distribution for men and women overall looks identical by eyeballing the shape; however, we need to notice that the number of men in the dataset double the number of women if we look at the frequency label on the left. On the other hand, the final weights are distributed similarly, ranging from 0 to below 0.5 mainly. I do not think we need to exclude these outliers. The weighting is meant to reflect how this individual represent the population based on demographics, and there is quite a bit of people that are considered as outliers in the above boxplot. I think this information should be reflected in our analysis rather than discarded since there are people differs from others to some extent.



Correlation.

a) From the pair plot, none of them seem to have a linear relationship whatsoever. The matrix shows a 0.148 for cor(education_num, hours_per_week) which can be considered as a weak positive correlation.

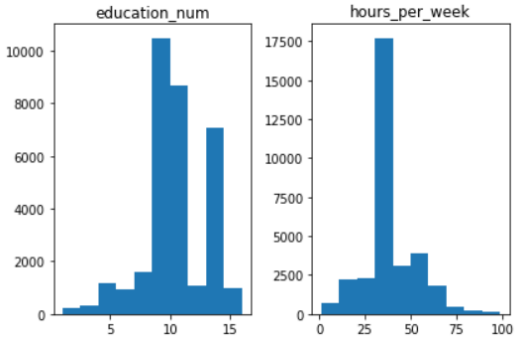


	age	education_num	hours_per_week
age	1.000000	0.036527	0.068756
education_num	0.036527	1.000000	0.148123
hours_per_week	0.068756	0.148123	1.000000

b) education_num and hours_per_week are the only pair with correlation greater than |0.1|. We performed the hypothesis test under the null hypothesis that these two variables are not correlated. The p-value of 4.24e-159 indicates that there is strong evidence to reject the null hypothesis. The correlation of 0.148 suggests a weak positive linear relationship, and the findings are not surprising as well-educated population will have stable job and more work hours.

c) The correlation for women (0.1787) is slightly stronger than men's (0.1368) but not by a large margin; both have a weak positive linear relationship. Moreover, the two p-values (e-91 and e-78) both show strong evidence to reject the null hypothesis and accept that the two variables are correlated. I think this result is expected since gender does not really play a role in the relationship between education and work hours. The correlation is stronger for women because well-educated women are probably more involved in professional development; besides, the sample size of women in this dataset is only half of men which can also be a reason.

d) First, notice the covariance only slightly decreases when going from unweighted 4.705 to 4.634 weighted. The variance of education_num went from 6.62 (unweighted) to 6.83 (weighted). This means that the data points on the "edge" (the ones that can increase dispersion) are weighted more; hence, these samples are underrepresented, and from the histograms, we know that these underrepresented samples are mainly people with less education. The samples from the center and right (one mode) of the distribution are well-represented. The opposite goes with hours_per_work, the unweighted variance 152 decreases to weighted 146. There is less dispersion after weighting which means that samples from center of the distribution are underrepresented. And the people with less hours and people with more hours has small weighting and well represented.



Regression.

a) Yes, men do tend to work more than women in our model. The coefficient for males is 6.0177 (men work 6 hours more), and this result is supported by the p-value given.

b) The trend of men working more hours than women remain the same, but the coefficient does slightly decrease a bit to 5.97. education_num also has a positive coefficient which aligns with our previous result, and it is significant based on the p-value given 0.000.

c) The trend from previous question is still the same with a little decrease in coefficient. education_num's coefficient also decreased a bit. The indicator of gross_income_group (income more than 50K) has a positive coefficient of 4.5175, which means people with more income also work more. All these results are significant by the corresponding p-values. We can use the Akaike Information Criterion to assess the models. The last model with 3 parameters has the lowest AIC which should be preferred. Notice AIC considers not just the goodness of the fit but also of overfitting. We could perform a forward selection during the model fitting procedure: pick the best one-variable model, and then add one more to compare all two-variable model; finally, compare the best two-variable model with the three-variable model.

Bonus Question.

After pushing symbols, we can obtain the following: