Github Repo: https://github.com/PatuanPurba/JSC270_HW2_2025_PPurba

Data Exploration

Number 1:
By using .dtypes, we can see that columns that are described as continuous on the website appear as int64. Further, almost all of these columns seem to get rounded into the whole numbers, except for the 'education-num' column. It appears that 'education-num' is a type of discrete. Apart from continuous type, the data types are also correct for columns that are described as categorical type on the website. However, there is an additional type of value, ' ?', in columns 'workclass', 'occupation', and 'native_country'.

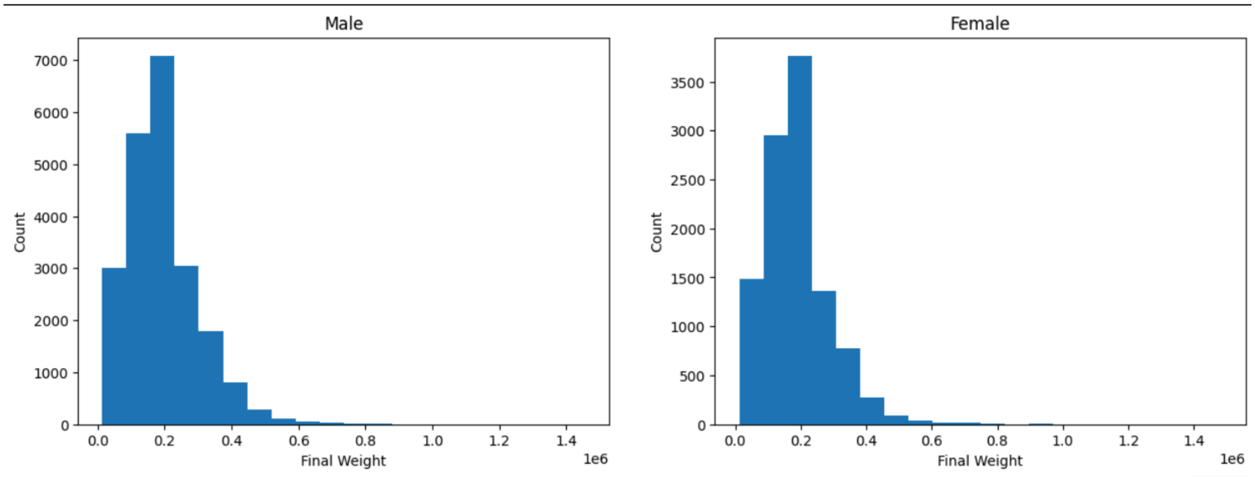
Number 2:
By using .isna().sum(), we can see that there is no np.nan on the table, so we need to determine what's classified as missing values in the tables. It appears that not all column has missing values, but some columns have ' ?' as the values. We will classify this value as a missing value in our table. The total number of missing values in each column:

- 'workclass': 1836
- 'occupation': 1843
- 'native-country': 583

To make future analysis easier, we will replace this value with np.nan.

Number 3:
By plotting the 'capital_gain' and 'capital_loss'columns individually, we can see an interesting pattern. For each column, the dataset is divided into 2, which are zero value or non-zero value. More than around 90% of data points have zero value in these columns. In this scenario, we may transform the column into a categorical value. Further, we can see on the plot, especially on the "capital_loss", there are 3 regions. The 3 regions are zero, middle, and high. We can use this finding as a baseline for our transformation.

Number 4:



From the graph, we can see that the data is heavily right-skewed, with most of the data points having a value of 0.2. Further, we can see huge similarities in distribution between “Men” and “Women”. Both of the graphs start at 0, peak at 0.2, and start to vanish at 0.6 until 0.8. Further, there is a slightly extreme outlier for women where there are some data points of 0.9.

Heavy right skewness makes many data points lie outside the 1.5 IQR range, leading to many data points that can be considered outliers. By using 1.5 IQR on the boxplot, we can see that every value higher than 0.4 can be classified as an outlier. However, we know from the definition that this column represents the population representation. A high value may represent an underrepresents group of people during the survey. An extremely high value may tell that the group is extremely underrepresented. Therefore, we shouldn’t exclude the outliers in this case.

Correlation

For this analysis, we only use columns “education_num”, “age” and “hours_per_week”

We find the correlation between these columns on our datasets by using the pre-built function pd.corr() and graph it by using sns.heatmap(). From the graph, we can see that these 3 columns have a positive correlation with each other. However, the correlation coefficient tends to be small with 2 of them only around 0.06. The correlation that is quite big is only between “education_num” and “hours_per_week”.

We do hypothesis testing to analyze these 2 columns further. By computing the t_value from the correlation, we get the p_value is 0.0. Therefore, we have very strong evidence to reject the null hypothesis. Therefore, we can fairly say a positive correlation exists between “education_num” and “hours_per_week”. However, the direction of the correlation is not really as expected. We usually expect that the more educated people, the less they will spend on work hours. However, this finding suggests that more educated people will have more hours of work.

We also usually suspect that older males will have a higher education than females due to job requirements. We can analyze this by comparing the correlation between both of them. From the analysis, we know the correlation for “Male” is 0.06 (p-value: 0) while for “Female” is -0.017 (p-value: 0). This finding aligns with our expectations beforehand. Additionally, we can also derive that younger females tend to have more education than the “older” generation, becoming a great sign of gender equality.

Lastly, we will try to find the covariance matrix between “education_num” and “hours_per_week”. We can see that the covariance is positive means a positive relationship between these columns. However, the value is relatively smaller than the variances, indicating a weak relationship.

Regression

We use linear regression to make a prediction on “hours_per_week” based on other columns

In the first model, we use “sex” as our dependent variable. From the summary, we can see that the coefficient for males is higher than for females with a difference of around 6. It suggests that males, on average have 6 hours more of work than females.

In the second model, we use “education_num” as a control variable to increase our model capacity. From the summary, we can see that the previous trend still happens in this model. From the coefficient, we can say that Males still have 6 hours more work than females. Further, we can also see from the model that “education_num” has a p-value of 0, which means that this column is significant in the model. The 95% Confidence Interval of the coefficient of this column is (0.647, 0.748)

In the last model, we add “gross_income_group” to the independent variables. First, we can write down the interpretation of “sex”’s coefficient on each model.

- **First Model:** Every data point will have at least 26.2977 hours per week (const in the model. If Male, it will get an additional 16.15 hours per week while Female only 10.13 hours.
- **Second Model:** Every datapoint will have at least 21.6 + 0.7 * value of its “education_num” hours. Then, If Male, it will get an additional 13.7841 hours. while Females only 7.81 hours
- **Third Model:** Every data point will have at least 18.1155 + 0.45 * value of its “education_num” hours. Then, if its “gross_income_group” is bigger than 50K, it will get an additional 4.5175 hours. Finally, if Male, it will get an additional 13.8746 hours. while Female only8.7736 hours

To determine the best model, we can see the value of R squared, which tells how much percentage does our model can explain the variance of data and the RMSE. Looking at the summary, the 3rd model is better with R^2 is 0.094 and RMSE of 11.75 while the 1st model only has an R^2 value of 0.053 and RMSE of 12 and the 2nd model only has an R^2 value of 0.074 and RMSE of 11.88.

Bonus Question: We know on simple linear regression: $\hat{y} = \beta_0 + \beta_1 X + \varepsilon$

We know, Corr(X, y) = $\frac{Cov(X, y)}{Sd(X)Sd(y)}$

From the Lecture Slide, we know $\beta_1 = \frac{Cov(X, y)}{Var(X)} = \frac{Cov(X, y)}{Sd(X)^2} = \frac{Cov(X, y)}{Sd(X)^2} \cdot \frac{Sd(y)}{Sd(y)} = \frac{Cov(X, y)}{Sd(X)Sd(y)} \cdot \frac{Sd(y)}{Sd(X)}$

Therefore, we know that :

β_1 = Corr(X, y) * $\frac{Sd(y)}{Sd(X)}$