JSC270 – AS2

Shih-Ting (Cindy) Huang

**Data Exploration**

1. Check the columns of your data. Are they the expected data types based on their descriptions?
   1. Yes, they are the expected data types. The continuous variables have type int64 while the categorical variables have type object in the read data.
2. Missing values are represented by the character “?”.
   1. Workclass: 1836
   2. Occupation: 1843
   3. Native\_country: 583
3. Yes, these variables should be transformed because they are heavily left-skewed.
4. Looking at both the boxplot and histogram for women and men’s final weight, we see that they share approximately the same mean at around 0.2. There seems to be more variance for men’s final weight as there are more data points lying in and beyond the third quartile. I don’t think outliers should be excluded because there seem to be a significant amount of them, suggesting that perhaps there is another category of the population unaccounted for.

**Correlation**

1. S
   1. There doesn’t seem to be a strong correlation between these variables. The strongest correlation occurs between education\_num and hours\_per\_week with a correlation coefficient of 0.148123.
   2. D
   3. The correlation between age and education\_num is -0.0179 for female and 0.06049 for males.