Assignment1 Report

Lan Chen B00809814

1. Analyze the dataset, what problems do you see in the data? In which columns?

There are some bad data and missing data of the dataset. I found them on below columns:

1) Age: there are negative values for age which make no sense
2) Workclass: there are missing values which marked as '?' and also spelling mistake of the value caused by manual input
3) Occupation: there are missing values which marked as '?' and also spelling mistake of the value caused by manual input

2.For every column that you had to work on, explain how you tried to fix the data and justify your decision. If you used any libraries, briefly describe what they do.

1) First, I calculated the number of the instances with missing value. It's about 1000 rows. It's not so many in a dataset with 30,000 rows, so I decide to drop them.
2) For negative value in the column age, I used a mean value to replace those values.
   Used libraries:
   - Numpy: to calculate the mean value
3) For the wrong data in categorial columns workclass and occupation, I matched them to the correct values based on the similarity.
   Used libraries:
   - Levenshtein: to calculate the similarity

3.After preprocessing the data, plot a histogram (use matplotlib) to display the data distribution for one numeric and for a categorical column that you had fixed.

1) Histogram for column age

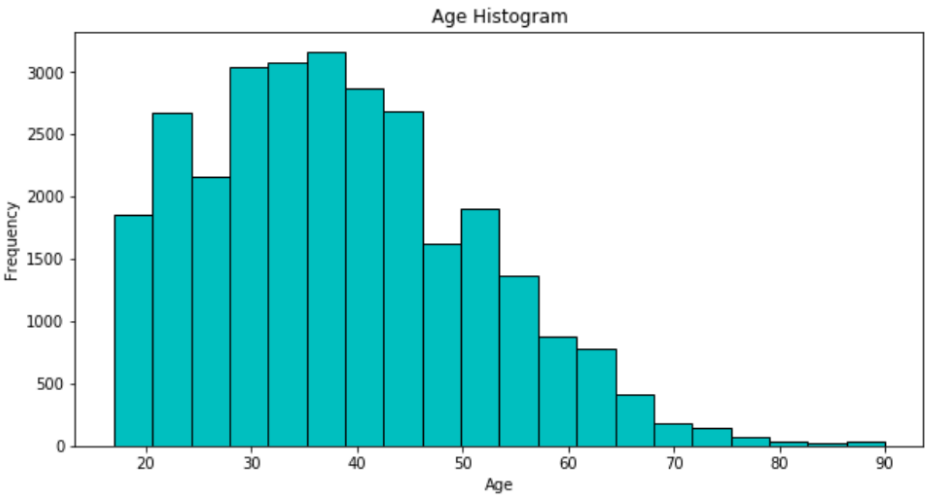

Figure 1 Age histogram

2) Histogram for column workclass

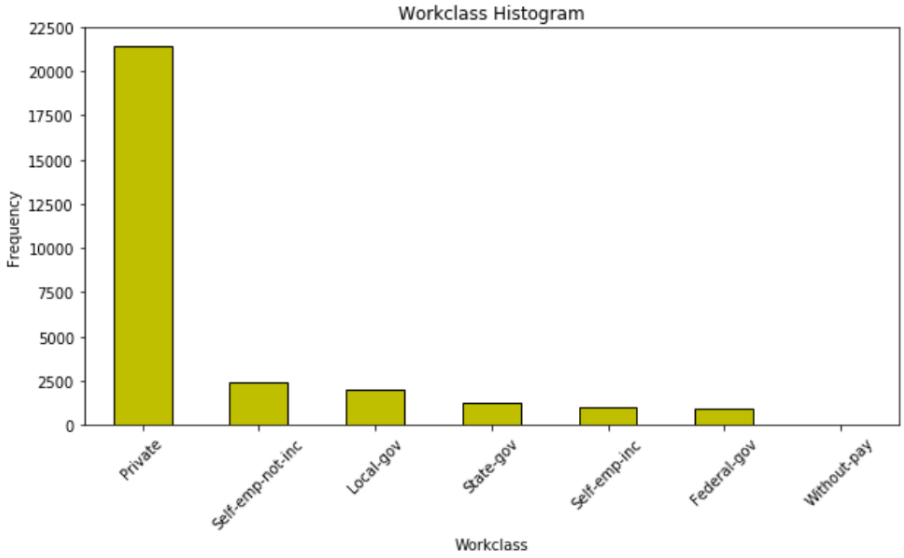

Figure 2 workclass histogram