Table Analysis Report

Firstly, I decided to create a division between the continuous and categorical data, in which I opted to divide these features by having the numerical data in the continuous table and the non-numerical data inside the categorical table. I deliberately not added ‘id’ and ‘target’ because ‘id’ is not needed in any of the data tables and ‘target’ is the result of the data therefore, I did not need to do any analysis or replacement on it.

Continuous features contain: age, fnlwgt, education number, capital gain, capital loss, hours per week.

Categorical features contain: work class, education, marital status, occupation, relationship, race, sex, native country.

After executing my program without doing any transformations to the data I found out that there were outliers in the original data set, at first, I tried using the clamp transformation in the continuous features using the first quartile value and the third quartile value, a method that did not work for all the data, therefore I had to use the mean and the standard deviation method.

I then created a range with a lower value (that would turn into the minimum value in the column if the number was lower than zero) and a higher value, using the method mentioned above for both values once I had my threshold, I created a random number between that threshold and use a mask instead of replacing the data to not damage the original data.

in the continuous features I use that threshold in every category except for age as that category should not be limited by my threshold. For capital gain and capital loss in the continuous features I did not want to remove or mask the zeros as they were almost 30k people with that value this would make sense as not many people have a profit or loss from selling property.

As there were no missing values in the continuous features, I decided to have my missing percentage to be the Sum of the values I masked using the threshold and divided them by the total count in the data. And as age was not being masked, I had to add it manually to the table.

For the categorical features I replaced the ‘?’ in the data with ‘NaN’ values and find out that only work class, occupation, and native country had missing data. To find the missing percentage, I used the sum of the ‘NaN‘ values and divided it by the total count of the data.

in the continuous features there were very high cardinality values in comparison to the categorical cardinality but I decided to not do anything with the age feature as it seemed normal that there can be multiple ages, for the fnlwgt I was not sure what this feature meant so I preferred to keep it as it was, for capital gain I found it normal to have many different values as there can be multiple profits amount from the sale of a property and something that is contradictory is the capital loss which should have about the same cardinality as capital gain does and instead it is very low in comparison.

In my opinion the cardinality in the continuous and categorical features looked regular therefore I decided to leave it as it was for both cases