Assignment 1 – Data Quality

Written Analysis

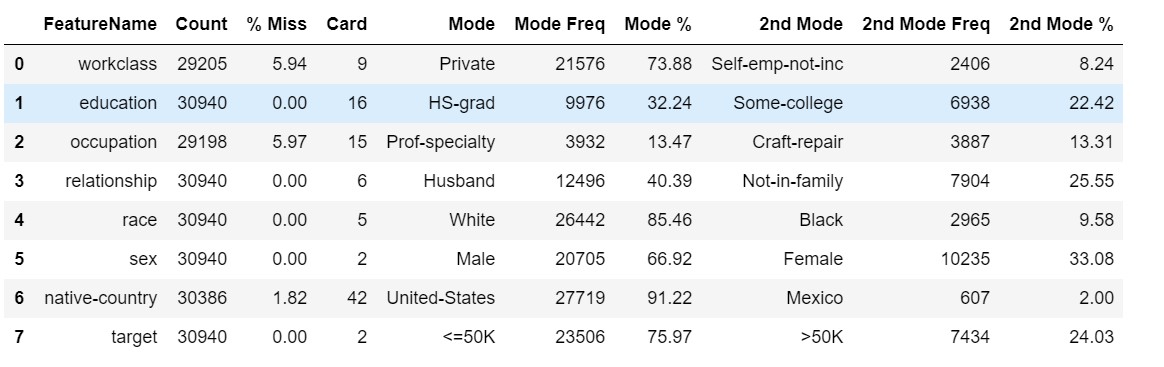
The dataset has in total 15 different features. Those with categorical sense are WorkClass, education, occupation, relationship, race, sex, native-country and target. The other features, which are of the continuous type, are age, fnlwgt, education-num, capital-gain, capital-loss and hours-per-week.

After analysing the data, if the collection were to be grouped by certain criteria, the categorical features are more ideal and easier due to the lower cardinality of such features. The features that have a cardinality of less than ten are “workclass, relationship, race, sex and target”. As part of the system that would use the dataset, “sex and target” can be set as Boolean flags due to their cardinality rate of 2. It is advised to use the other features that have higher cardinality rate as the main filters of the dataset.

However, these features are more likely to be impacted by outliers that dominate the set. This is true for “workclass”, “race” and “native-country”, as one part of each feature hold more than 70% of their respective areas. To work with this so the predictive model can be trained more effectively, a clamp would be put onto these areas to handle these fields.

Missing data exist in the set in the features “workclass”, “occupation” and “native-country”. However, since with native-country it is only 2% and with the other two fields it is 6%, we can train the model to use imputation to make an informed “guess” based off the other inputs associated with the missing features. Assuming this decision is being made, documentation about how it is implemented should be enforced.

A table of the section can be displayed below

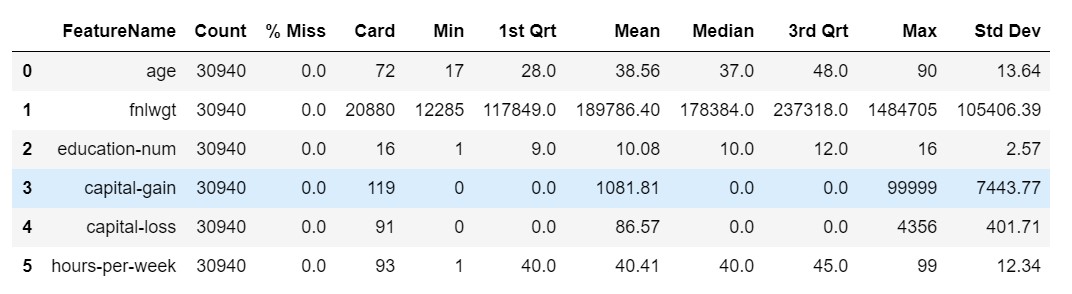


The continuous features, while not the optimal choice for filter reasoning due to higher cardinality rate, as just as important data as they are less likely to be influenced by an outlier. The only cases exception are hours-per-week, capital-gain and capital-loss. The outliers in these cases do not affect the calculation aspect of the data but the result is changed. Like the previous outliers, a clamp would be suitable here to assist train the model more effectively.

There seems to be a correlation between age increase with fewer workers. The older groups might have an impact of the rest of the dataset. More analysis needs to be covered to test that idea.

The features education-num and fnlwgt seem to be the ideal type, where there is no missing data nor outliers that impact the data, which means no alteration is necessary to ensure the model is trained to handle these fields.

A table of the section can be displayed below



In short, the categorical features seem to be the better option for filter of the data through their low cardinality rate. The continuous features are as important as the higher cardinality rate as it is less susceptible to outliers. For the outliers in the set, a clamp is ideal to filter them out and train the model more effectively. Once this part is completed, a imputation can be used on the missing data as they are very low in percentage, allowing the model to make an informed decision on these elements.