**Artificial Intelligence 2 – Assignment 1**

Student: Ronan Dillon Student Number: C12355251

The dataset used for this assignment appears to attempt to determine the features which may cause a person to earn above or below a fifty thousand salary. Using features like marital status, occupation and education which can be used to view a difference in salary, while other features like sex, race and native country can be used to determine whether there is a bias in a particular feature.

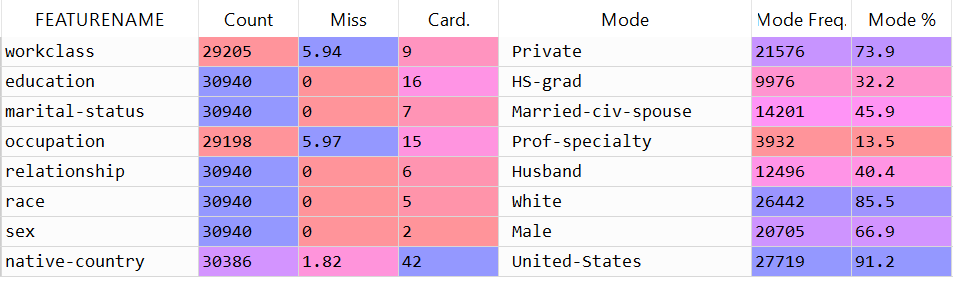
There are some issues with the data, although many of the features contain completely clean data, others do not. This can be a problem when performing analysis. Some records are missing data which could be due to human error when entering. Dirty data can also occur when converting data. Missing data only affects three features, and the percentage of missing data is less than six percent, so with ninety four percent of the data being ok it is not a big issue but it can be prevented by making data easier for users to enter and making sure the result can be easily converted to CSV. There could also be a no answer button so that rather than a user leaving the question blank and skipping it they can indicate that they did not have an actual answer for this question.

There are some things I would change in the data set provided, in my opinion there are too many different job titles in the occupation feature. I feel this could be categorized into smaller groups. This would make identifying occupations which receive the highest salaries easier while not making the data too vague and not providing any concrete information. The problem with having too many occupations is that outliers can become very influential when processing the data. Another feature which contains too many answers is the education field. I feel this is another feature which can be grouped together a bit more while not becoming too vague and useless.

Some features have a high cardinality which can cause the resulting data to be too vague. A way to combat high cardinality is to make a question more specific so not as many people fall under the same category. The more specific the data is the more likely analysis will return a stronger result.

**Data Quality Report**

**Category Features**



**Continuous Features**