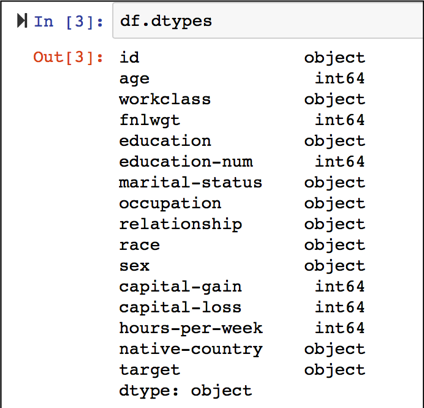
# Machine Learning Assignment 1 DATA QUALITY REPORT

For this assignment, I will be using the ‘pandas’ toolkit to analyse this data set. By studying this data set describes the income of an adult, which has features of capital gain, capital loss and target.

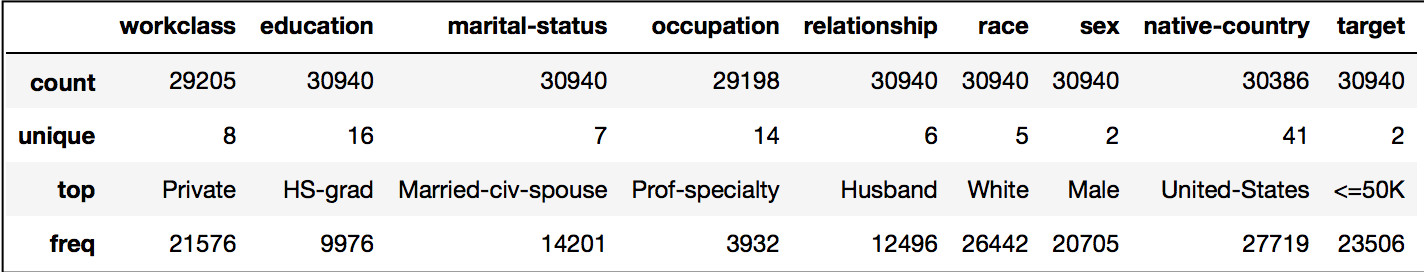
There’s two txt files:’feature\_names.txt’ and ‘dataset.txt’. The file ‘feature\_names.txt’ contains the headers that corresponds to each attribute of a row in ‘dataset.txt’. Many of the missing values were represented as ‘?’. Pandas has a built in value to replace these values that does not fall under categorical or continuous as ‘NaN’ / not a number.

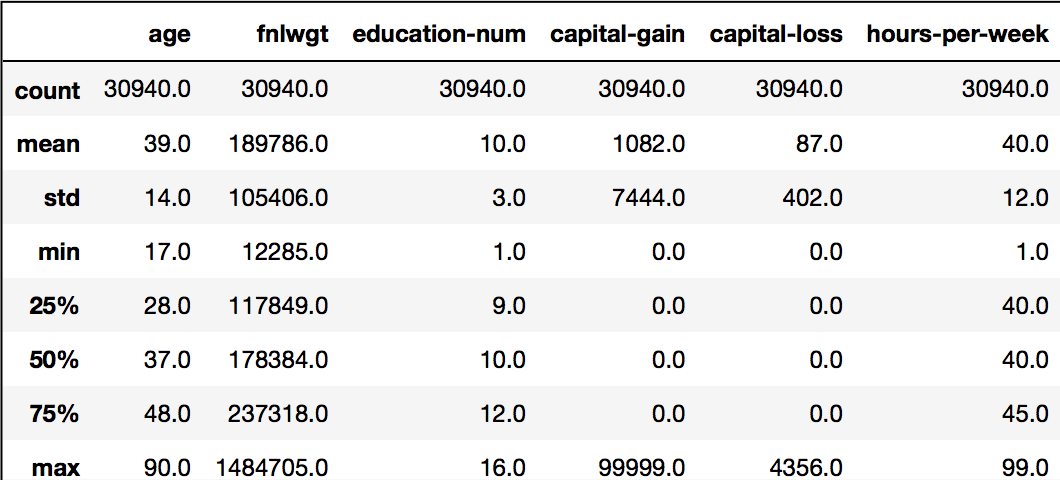
## Continuous vs Categorical

To see whether one feature is categorical or continuous, I would use the ‘dtypes’ command. Each feature corresponds to either ‘object’ for categorical, as categorical features tends to contain string characters, and ‘int64’ for continuous, as continuous features contains numbers.



I used ‘.describe()’ function to refer to for error checking in both tables.





Continuous deals with the numbers. For this dataset we find the number of occurrences, percentage of missing values, cardinality, minimum value, first quartile, mean, median, third quartile, maximum value and the standard deviation.

Categorical features include string characters of the data set. We find the number of occurrences, percentage of missing values, cardinality, mode, mode frequency and percentage of the mode. We also fine the mode of the second value of the data set, index[1].

Both of these data frames refer to the continuous and categorical tables of the the data quality report tables.

To find the cardinality of a data frame, I used the ‘.nunique()’ function to return the unique value for each feature. This shows the data that is unique, as in no reoccurring values, to the data frame.

I find one issue on the quality of the data are the missing values. To calculate the percentage of the missing values, I gather the values that are null and then I multiply by 100 and divide by the amount.

With handling the quartiles and modes of the data, reading each data in a for loop is required. To find the first and third quartile, they both transpose to ‘.quantile(0.25)’ and ‘.quantile(0.75), respectively. The data is written at an increasing order, to find the lowest and highest value, and could be represented as a bell curve.

To find the frequency of the mode, the ‘loc[]’ function is used to locate the data in the frame.

In my opinion, one way to improve the quality issues would be to normalise the data. This way, it would prevent repetition to occur. Since the data set is so large, by eliminating some units of measure it’ll minimise the size of the data set and it allows comparing categorical and continuous data easier.