# Support Vector Machine for salary data

Problem statement:

Prepare a classification model using SVM for salary data

Data Description:

age -- age of a person

workclass -- A work class is a grouping of work

education -- Education of an individuals

maritalstatus -- Marital status of an individulas

occupation -- occupation of an individuals

relationship --

race -- Race of an Individual

sex -- Gender of an Individual

capitalgain -- profit received from the sale of an investment

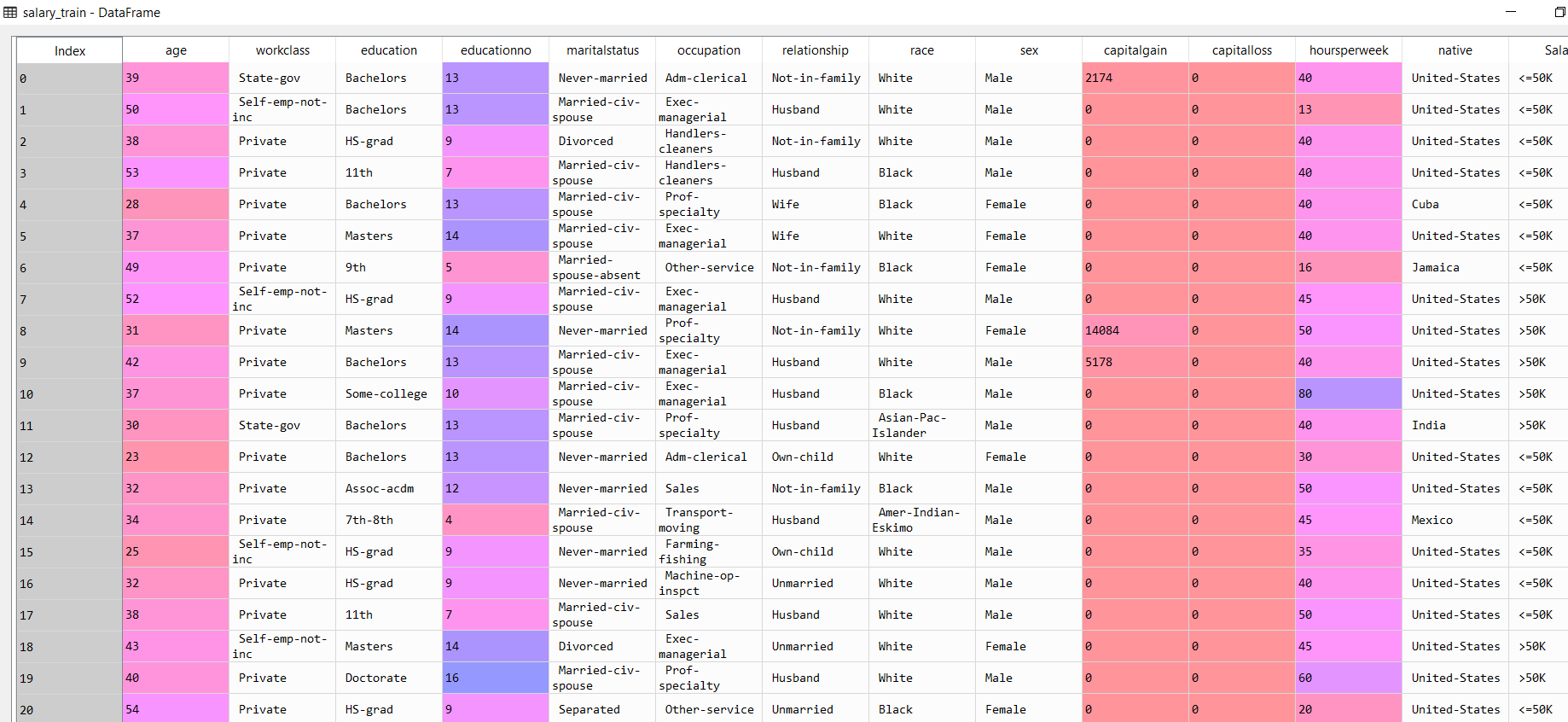
capitalloss -- A decrease in the value of a capital asset

hoursperweek -- number of hours work per week

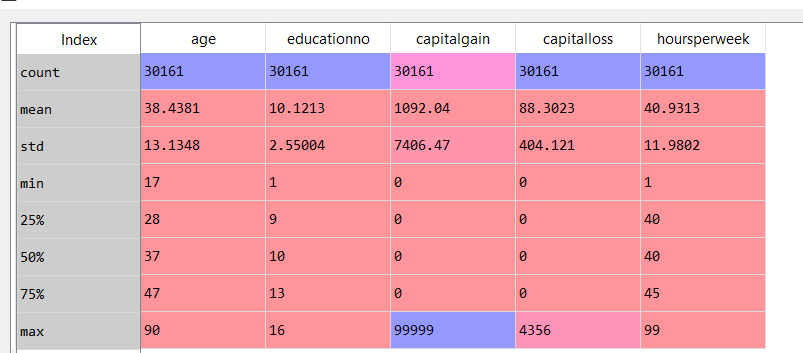
native -- Native of an individual

Salary -- salary of an individual

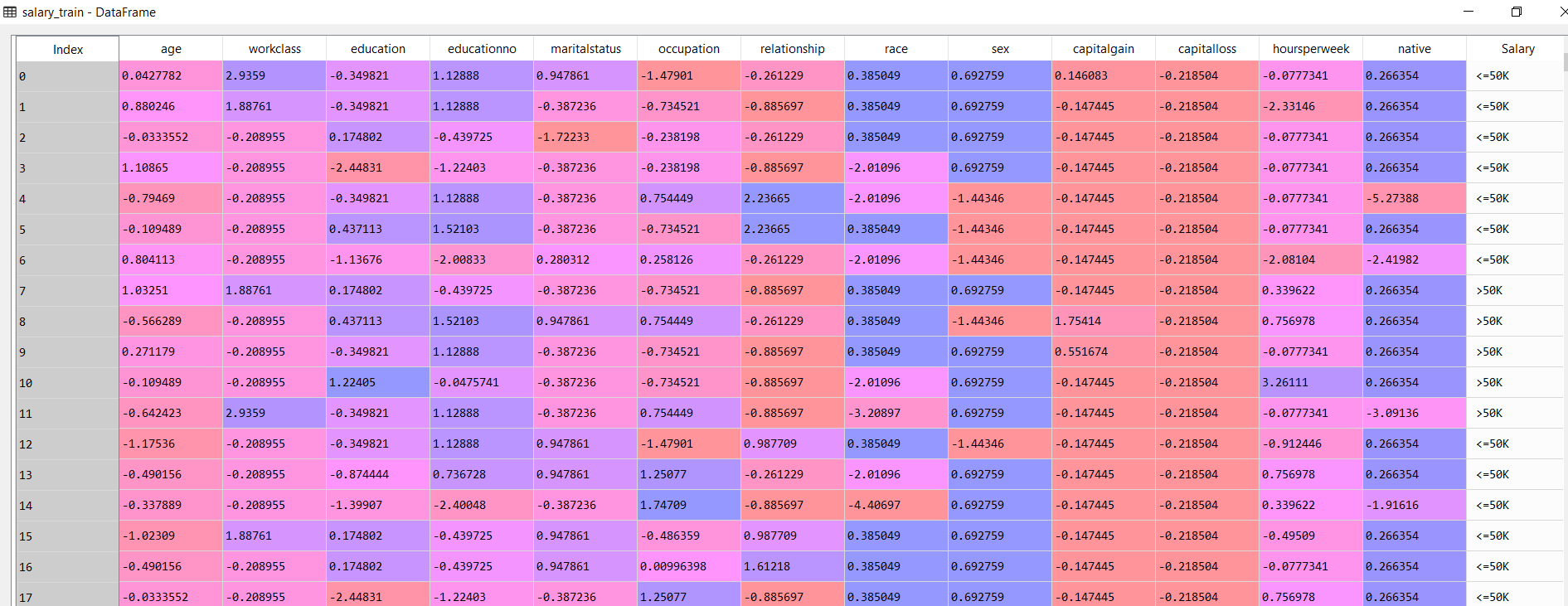
The following is the dataset of salary training data



The following is the descriptive statistics of only the continuous data



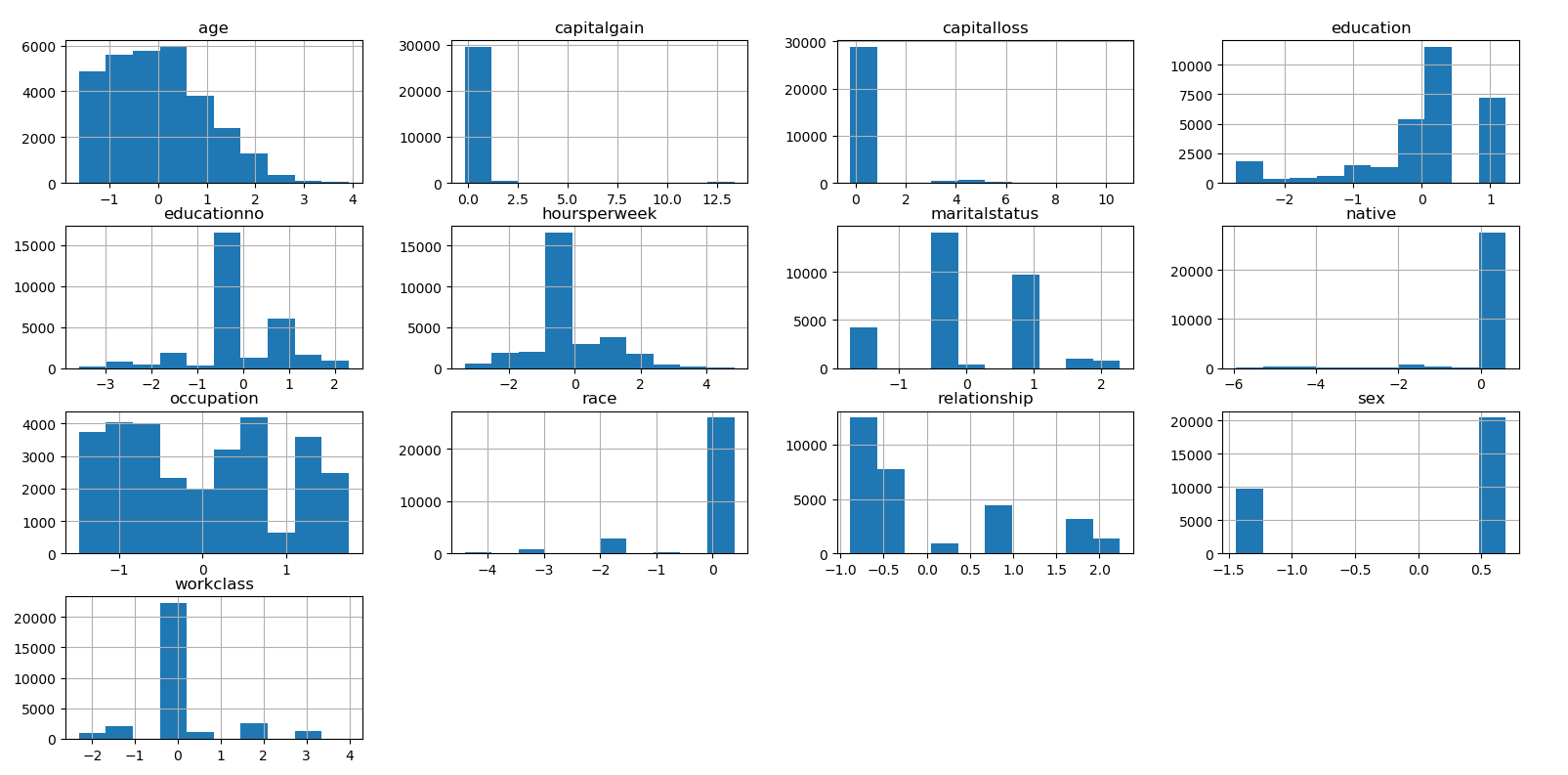
The range of the variables ‘captialgain’ and capitalloss’ have ranges far greater than the other variables. Hence we first convert all the categorical data into continuous data and then we normalize the entire data using the normalize function as follows:



The above is the data that is normalized data

The following are the exploratory data analysis

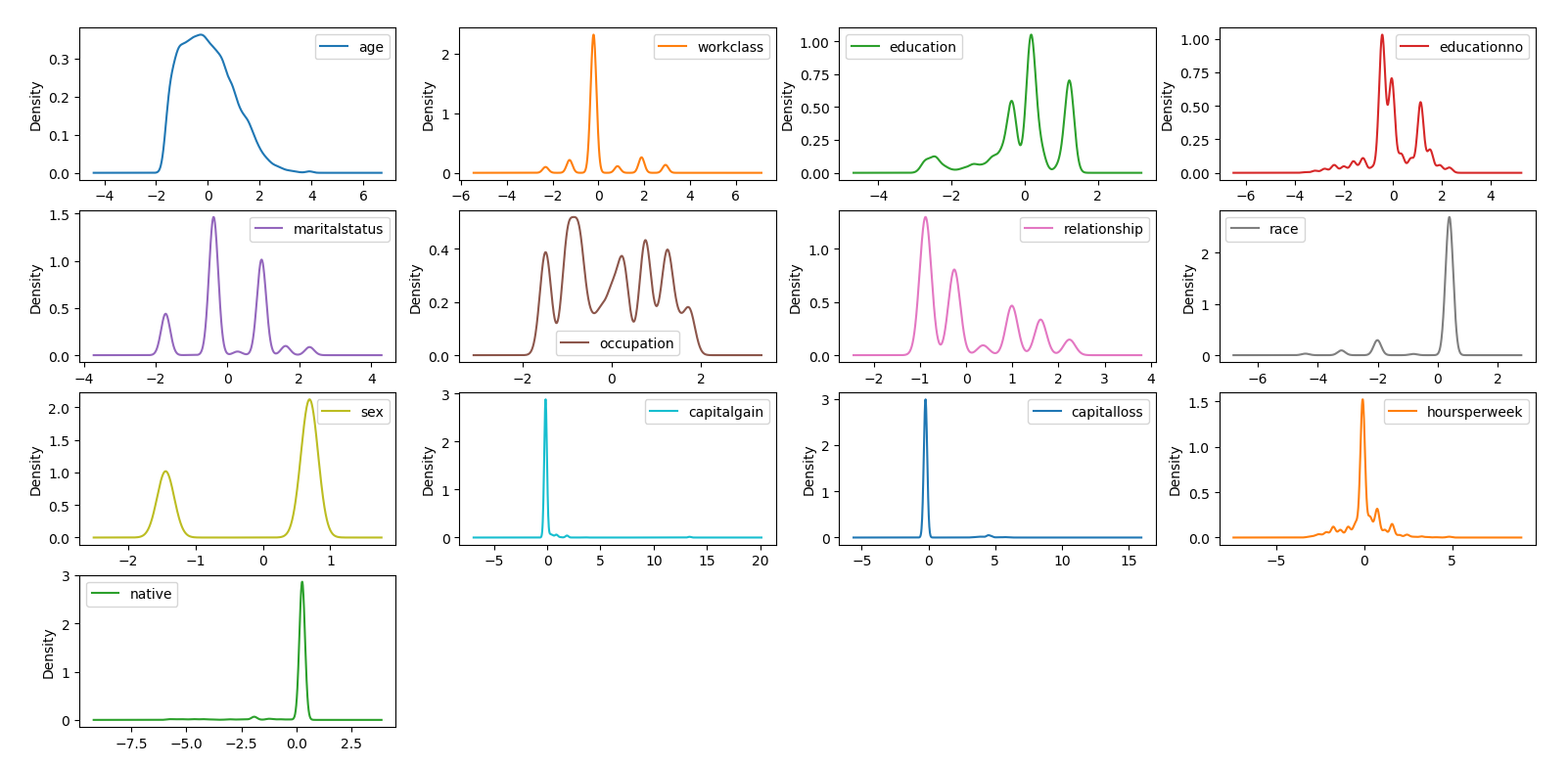
Histogram



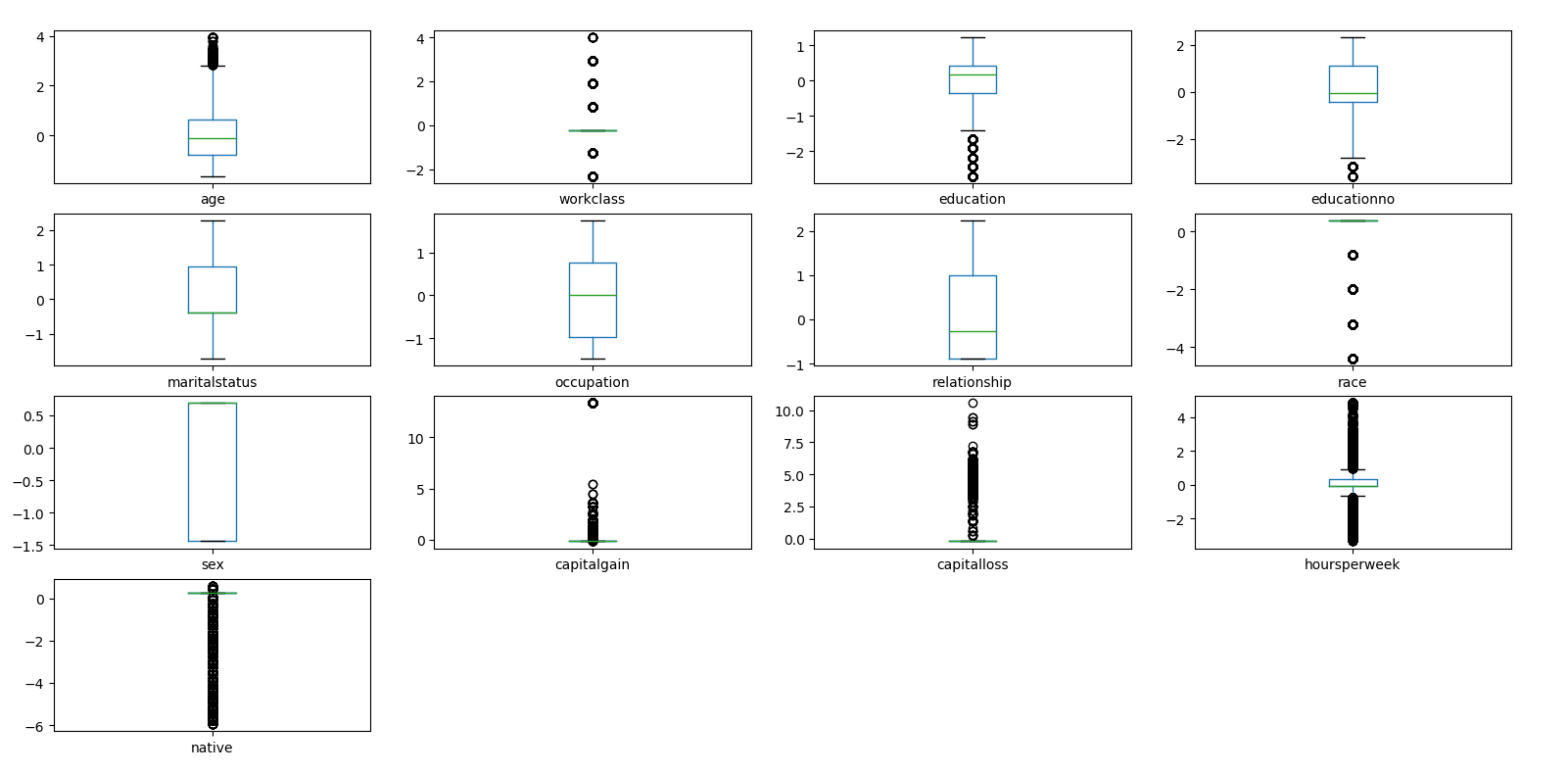
From the above data, the variable age is having near Gaussian distribution but with slightly positively skewed or right skewed.

The data with variable ‘hourseperweek’ has very high kurtosis

The following is the plot for continuous density function

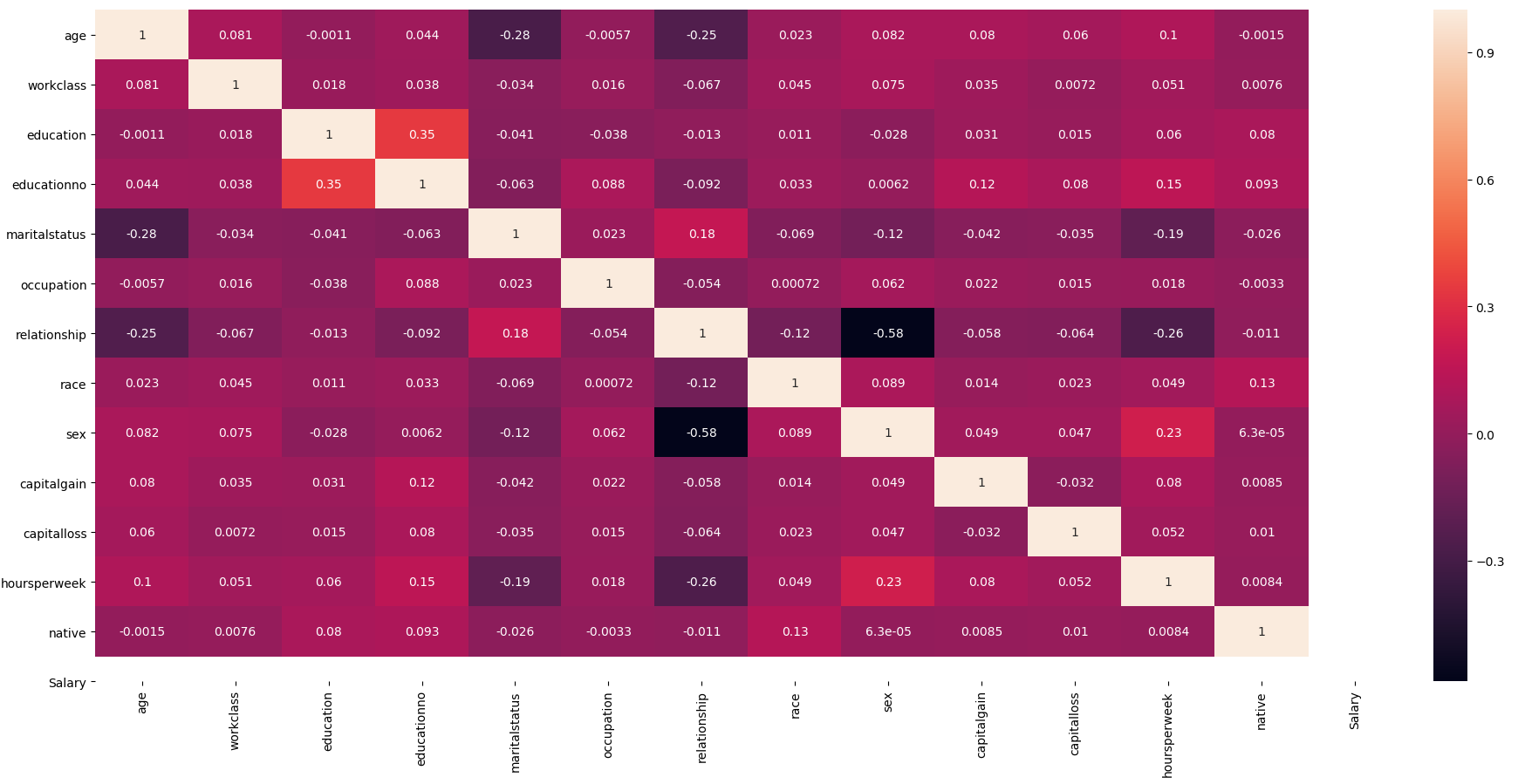


Boxplots



Except the variables maritalstatus, occupation, relationship and sex, the rest of the variables have many outliers.

The following is the heat map and their corresponding co-relation values:



The highest co-relation factor is between the variables relationship and sex which is a negative 0.58. This will not constitute for colinearity issues between the variables. The co-relation between other variables are not significant enough to cause any issue.

We feed the data through the Support Vector Classifier SVC imported from the sklearn package

With the help of different kernels we will estimate the accuracy and select the best model

**Result of Kernel = linear**



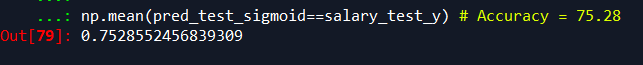
**Result of Kernel = poly**



**Result of Kernel = rbf**



**Result of Kernel = sigmoid**



Of the above model the model with the kernel rbf provides the highest accuracy with 84.65%