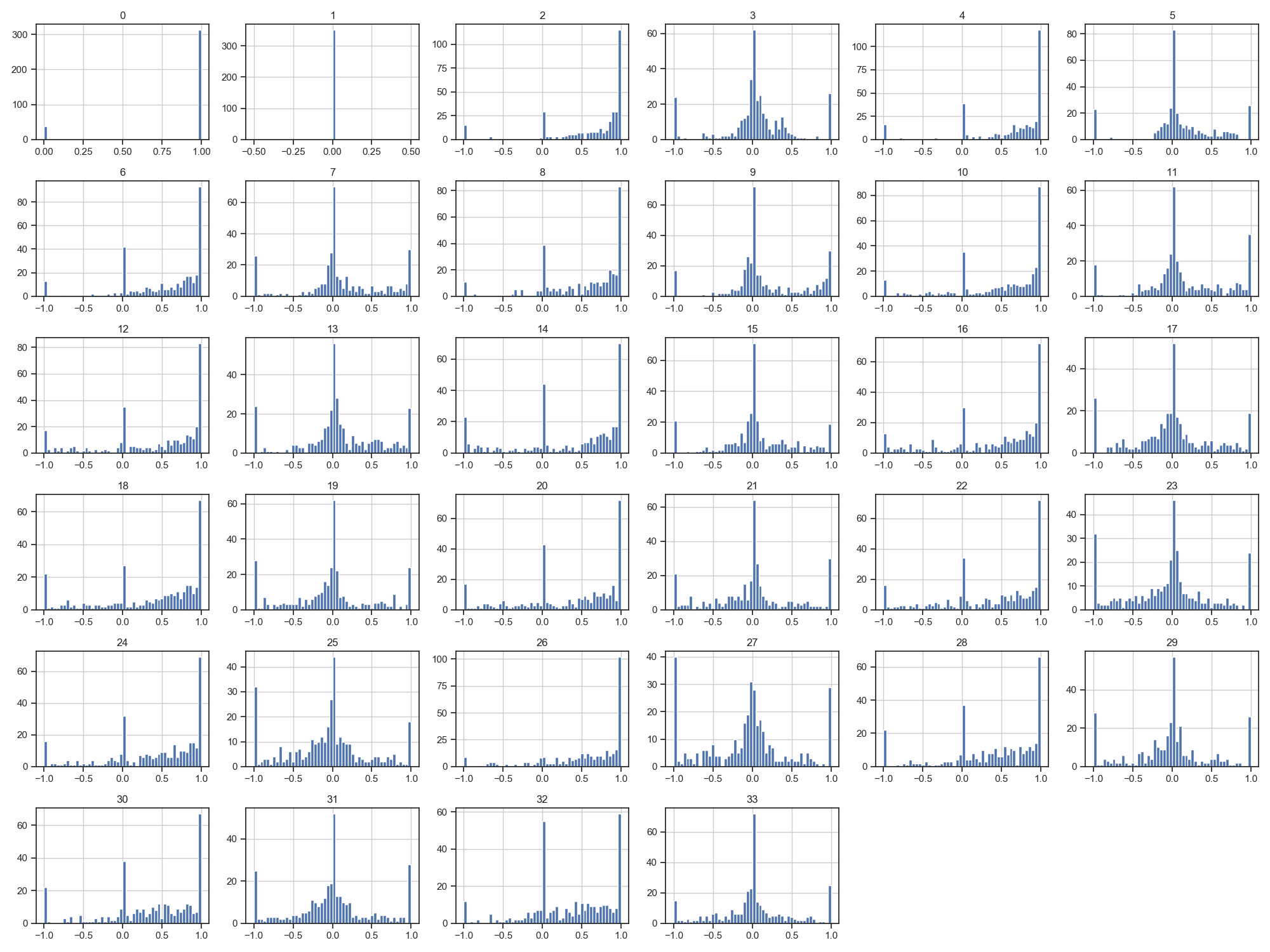
How did we process the data?

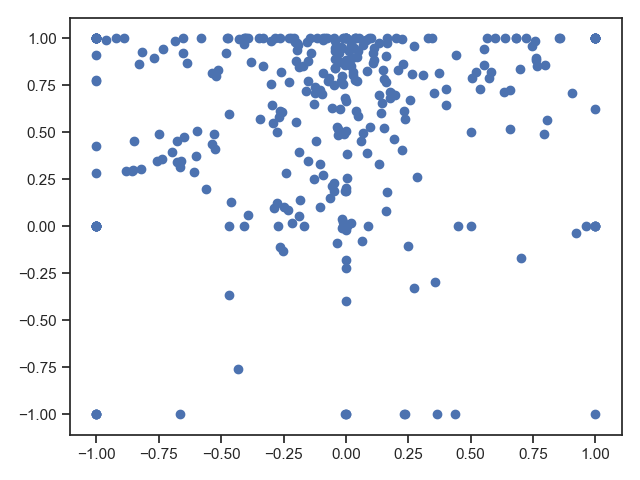
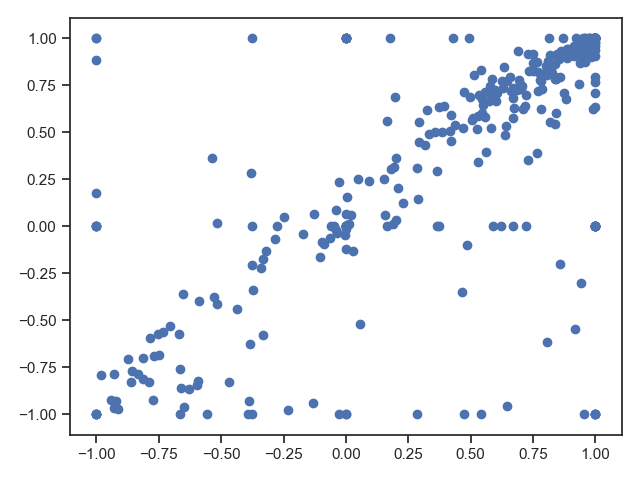
Check mean, standard deviation, min, and max to get a brief picture of the data distribution of each feature and filter out the invalid values such as ‘99999’. Split the X matrix into continuous and categorical to ease the training for both Logistic regression and Naïve Bayes. Categorical features are further spitted into binary arrays.

Dataset1:

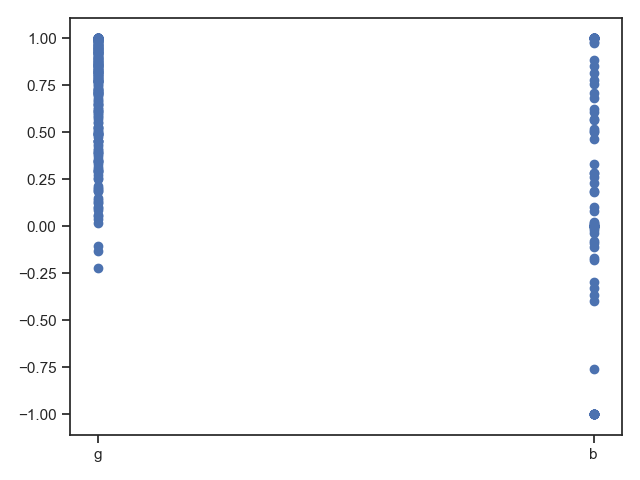
Looking at histograms we can see most of features are following gaussian-toward pattern where most of data points are centered in the middle, however still quite some points fall under either -1 or 1. Which may be due to the characteristic of feature itself.



Looking at scatter plots, some pair-features are showing linear correlations, and some are not.

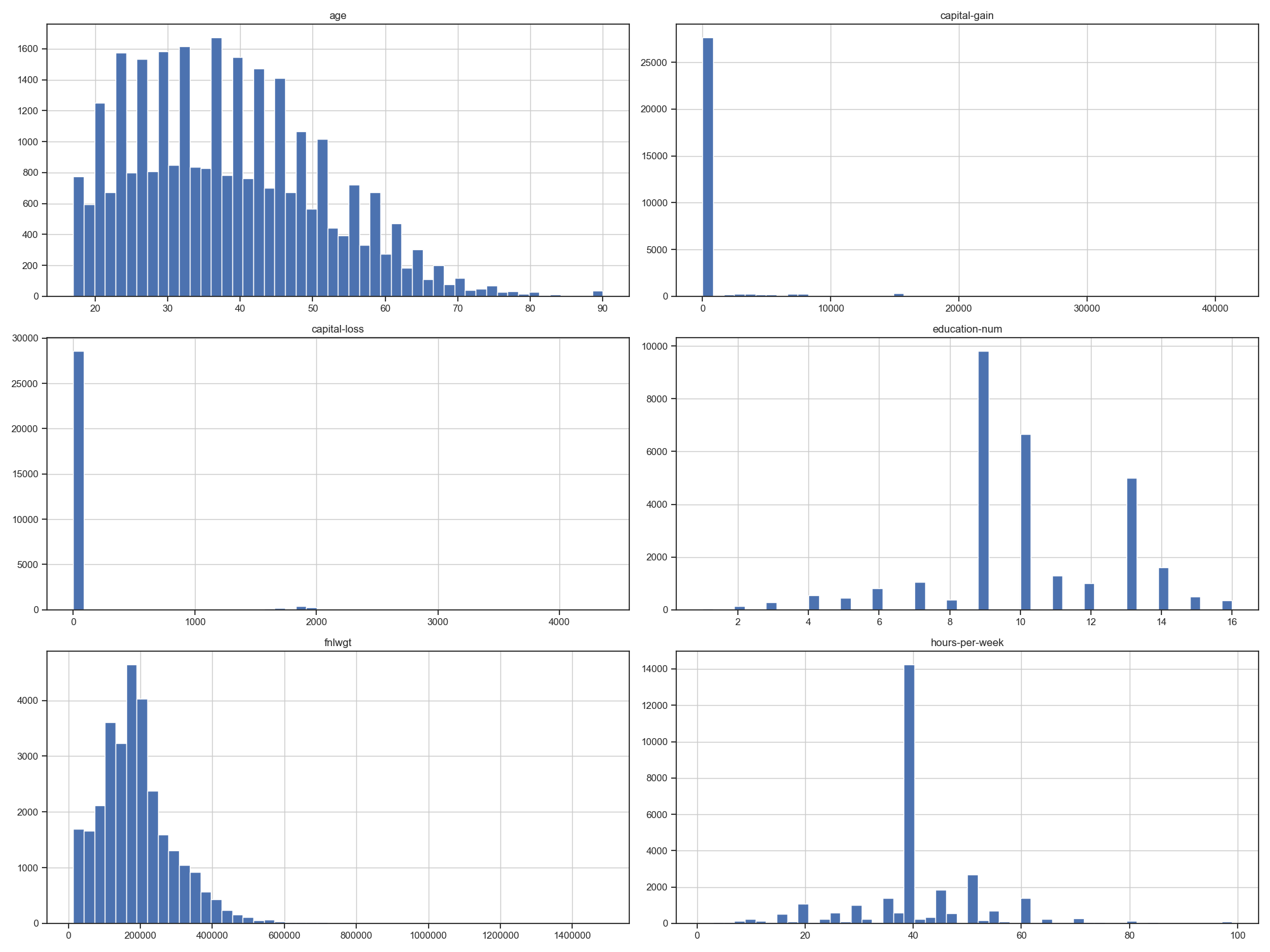


Some features are distributed differently in different class, which means these features do have predicting power.

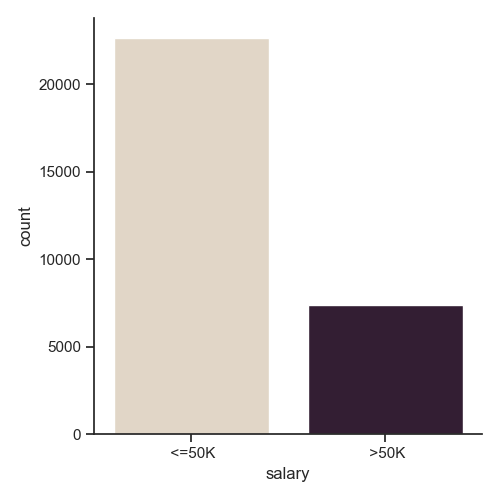
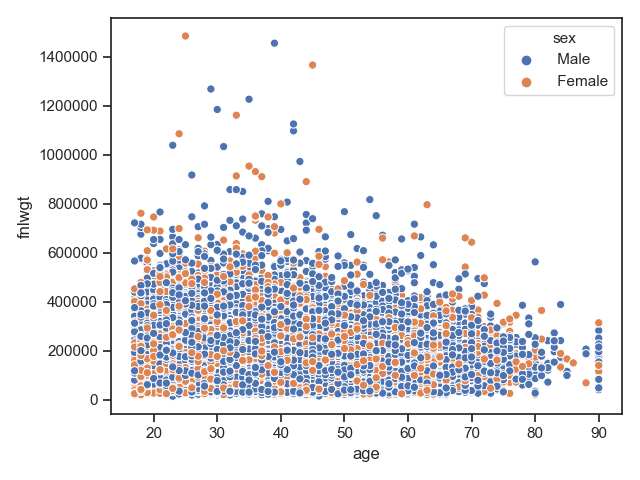


Dataset2

In our sample, age distribution is more centralize between 20- 40; most of people with 9+ years of education; fnlwgt are centralized under 30000; majority work hours is 37.5.

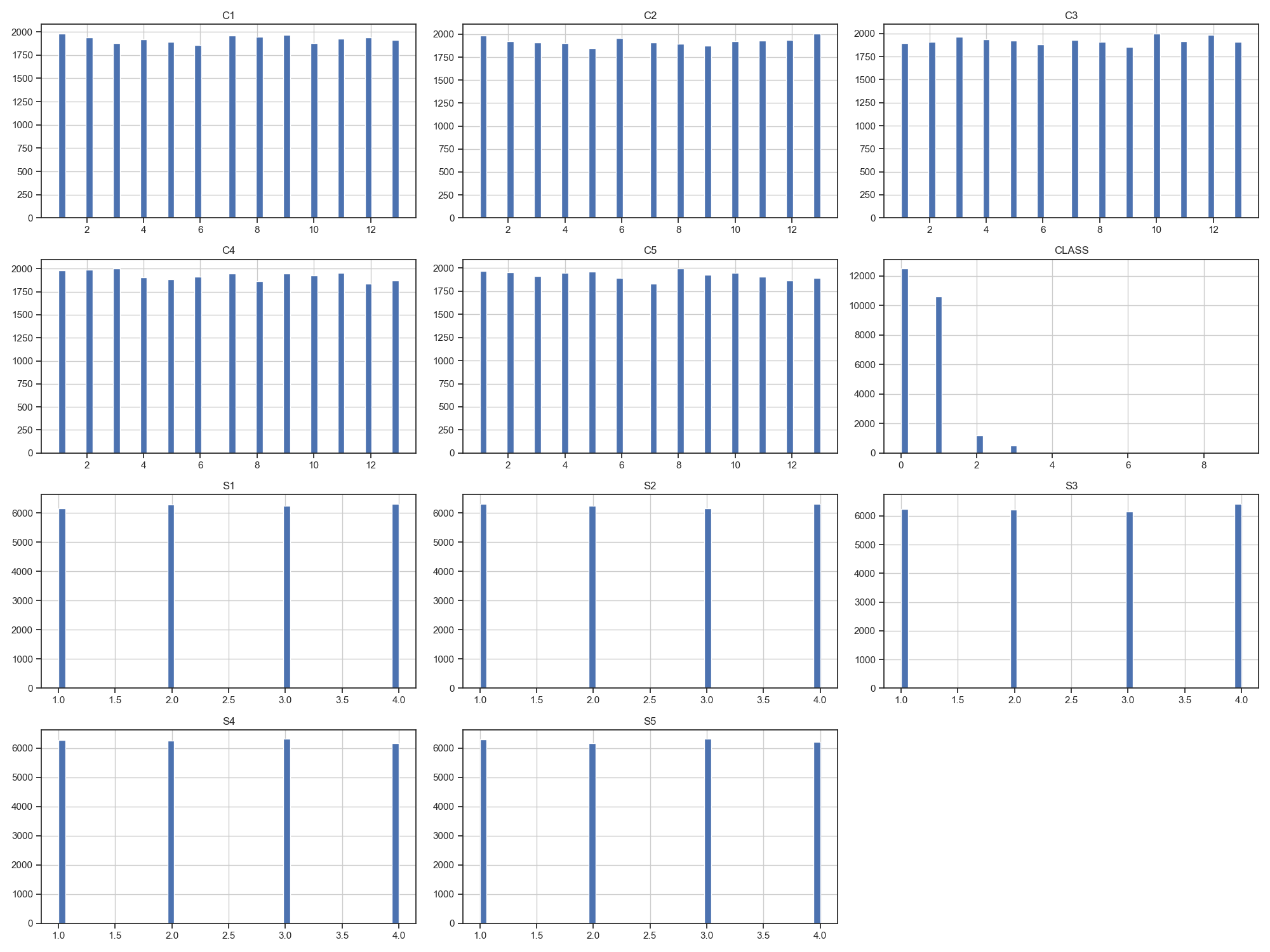


Correlation is seen between ages and fnlwgt, the younger the age the higher the fnlwgt.



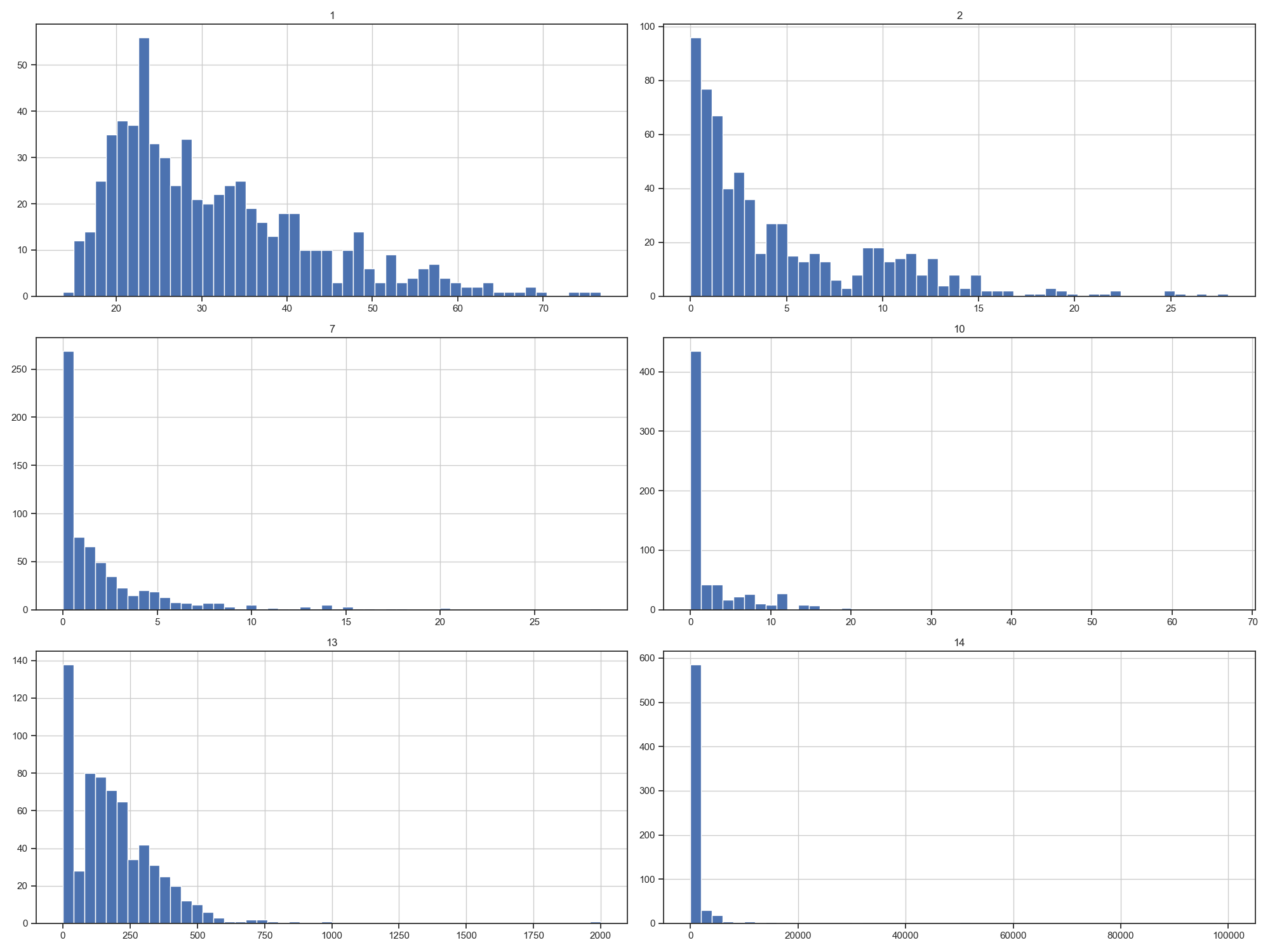
Dataset3

Both Suit and Rank are distributed equally of each card which means our poker sample is unbiased. As the level of class goes higher the probability is shrinking dramatically. (extremely rate to see royal flush)



Dataset4

We can see our continuous features are skewed to left (smaller values) with fat tail on the right.



Linear Correlation can be seen between continuous features. Also features are showing predicting power (the greater of value of feature 7 the higher the chance being classified as +).

