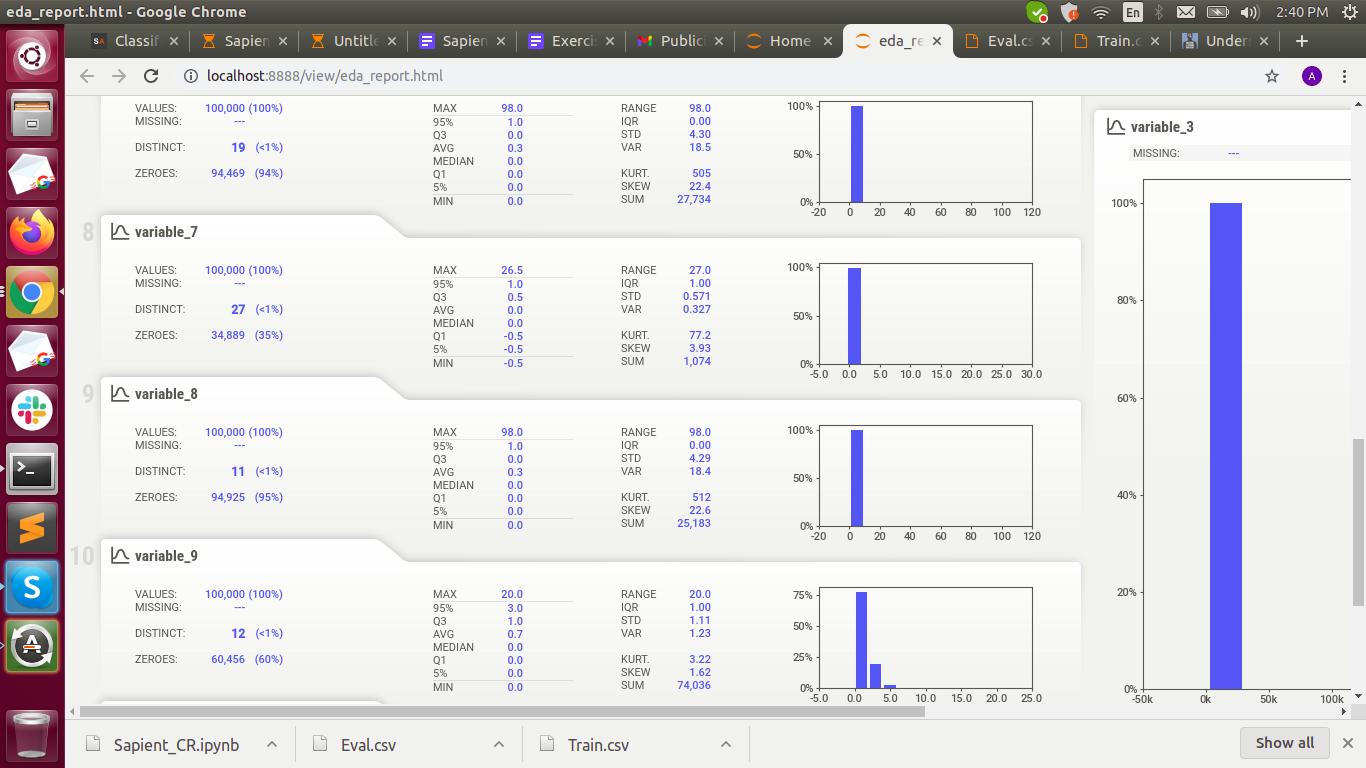
**Introduction:**

The task was to solve a binary-classification problem. The given dataset is highly skewed towards class 0 with lots of outliers on almost every feature .

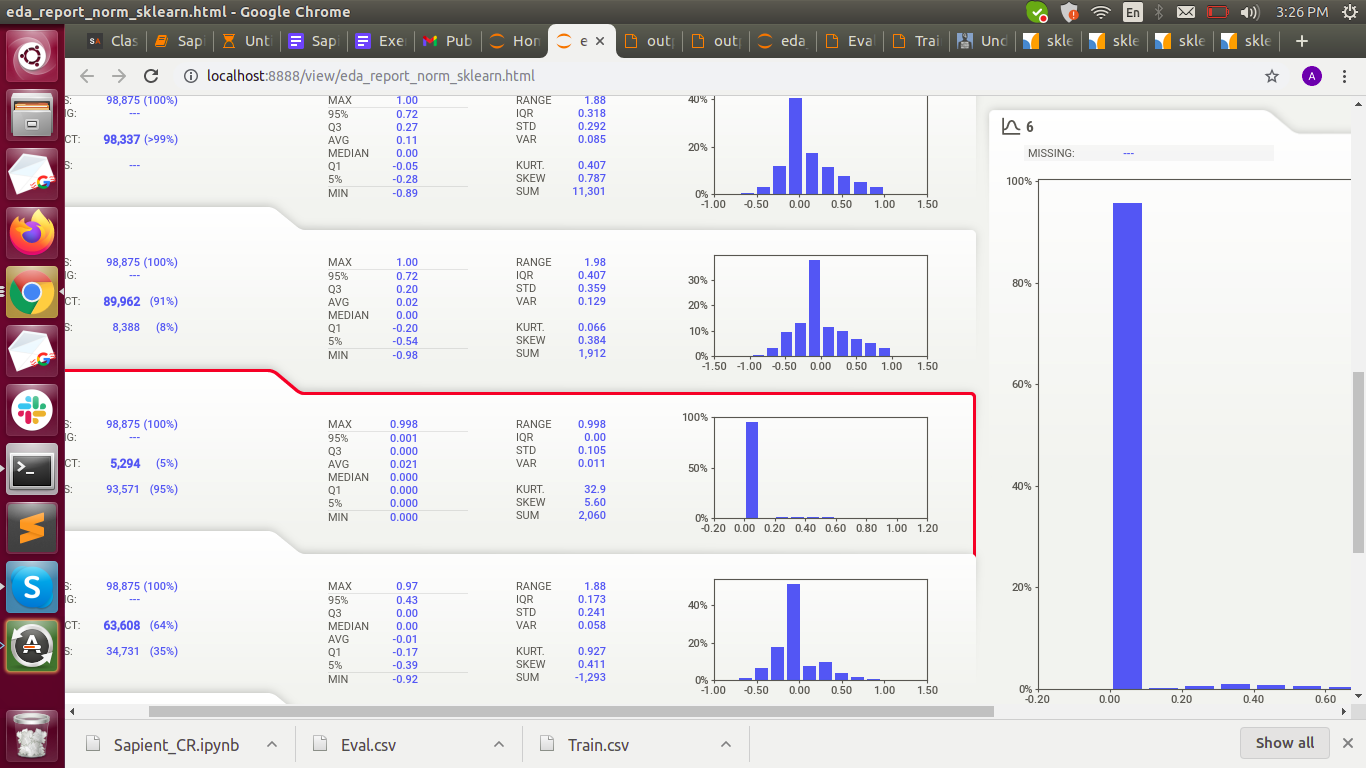
EDA:

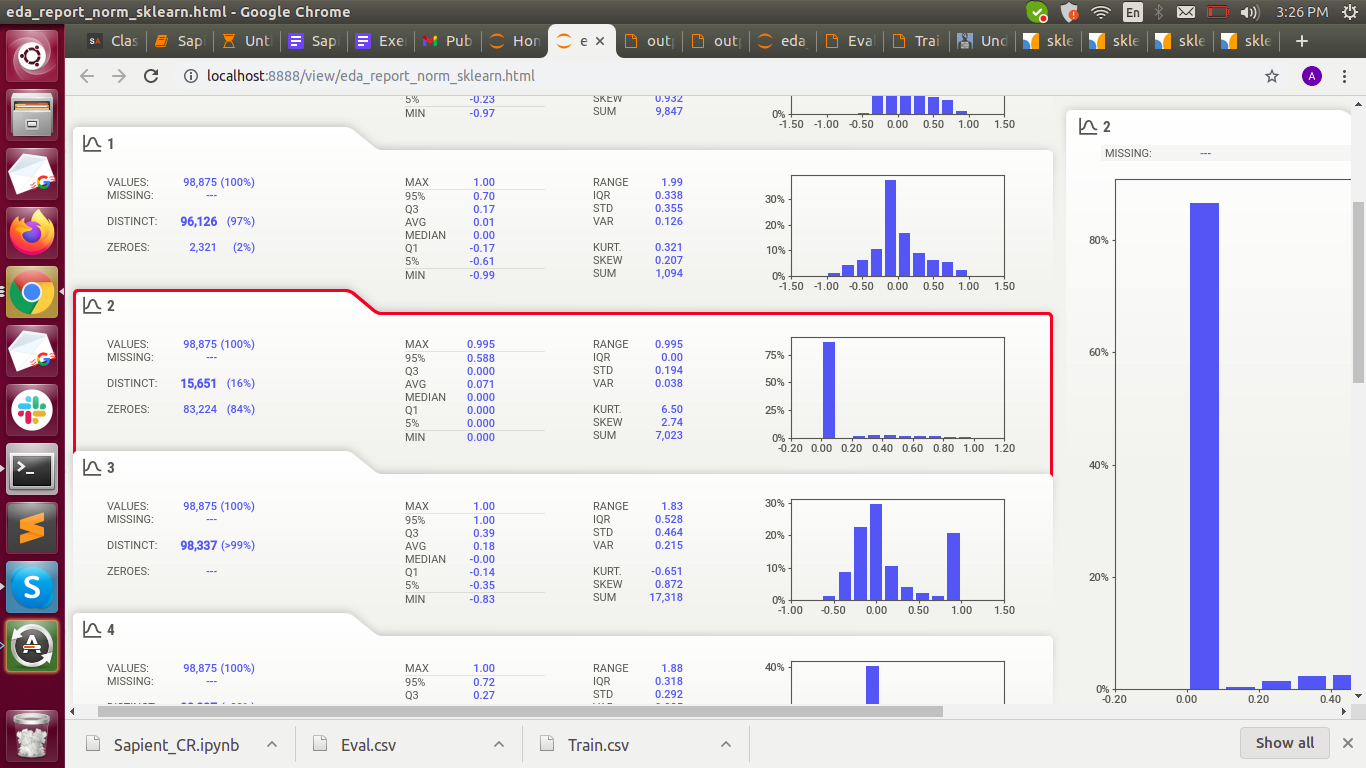
A detailed HTML report on the raw data is available in file *‘eda\_report.html’* .





After performing pre-processing such as removing Nan, removing outliers, Normalization and Undersampling. The EDA report on the resultant dataset is *‘eda\_report\_norm\_sklearn.html*’.





**Approach:**

* The first step was to do EDA, which showcased as discussed above that there are many outliers in the data and the sample was highly biased towards class0 with fewer samples for class 1.
* The second challenge was to address outliers, two steps were taken in this regard.

- Clipping some outliers from training where z score was higher than 5.

- Normalising the sample to create a more balanced distribution for the features.

* After doing these changes I ran a number of samples models for doing this classification task but the accuracy was stuck at 93% (roughly the number of samples for majority class)
* The next step was to do under-sampling to balance the training data to have better representation of both the classes, the technique I am using is condensed nearest neighbours, unfortunately due to the lack of computing power I was not able to complete the test for this run.
* We can also set up more biased thresholds for doing the classification ultimately for getting better results.