In order to improve performances of classification or clustering models, I have removed outliers from “Age”. The use of bucketizer in this assignment is defined briefly as below:

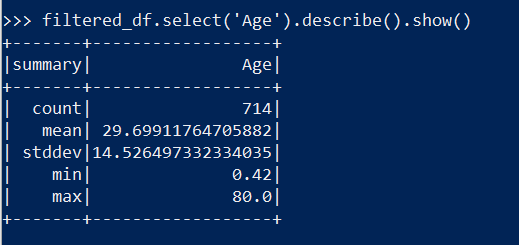
Bucketizer transforms a column of continuous features to a column of feature buckets, where the

buckets are specified by users. It takes a parameter:

Splits: Parameter for mapping continuous features into buckets. With n+1 splits, there are

n buckets. A bucket defined by splits x,y holds values in the range [x,y) except the last

bucket, which also includes y.



There is a huge variation in ‘Age’, which is indicated by the range of to 0.42 to 80.0, and it’s mean being 30. Bucketizing the records to measure ditribution:

