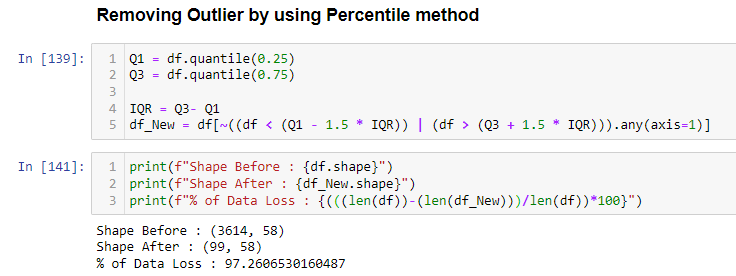


Note:

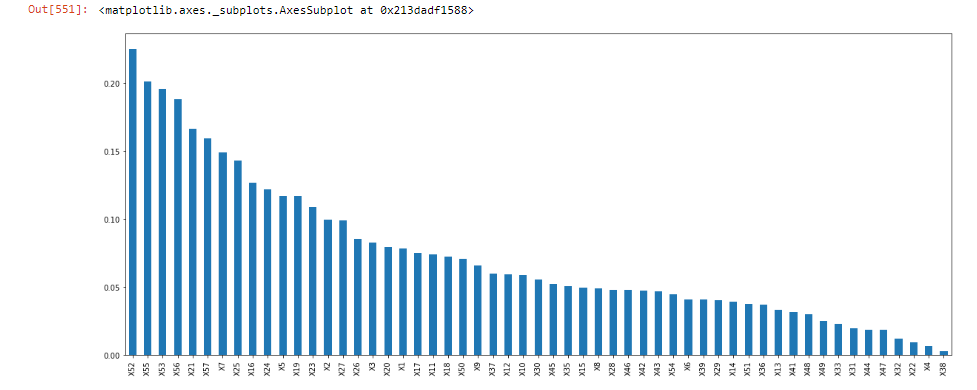
By using Z-Score method for remove skewness in which we are losing 52%of our Data.



**Note:**

By using Percentile method for remove skewness in which we are losing 97%of our Data.

Using a technique like Z-Score, Percentile led to huge amount of data loss mainly in high skewed variables.



Including all

Set 2

Set 1

1) We can observe the 3 sets of feature which help us to predict dependent variable as per mark above.

2) The less features we use (which are influence the dependent variable) the better for use, because it leads to consume less time, memory to predict the dependent variable.

**Logistic regression:-**

Algorithms like regression which are mainly consider the dataset is normally distributed, which follows the distance based concepts.

Using Power transformation with default technique “Yeo-Johnson” to convert independent variable in Gaussian like distribution (because it accept value <=0 which is not accepted by ‘Box-Cox’).

By using “Set 1” feature with intention of time & space complexity logistic regression help to give us 91% accuracy with less bias and less variance.

**RandomForest:-**

Algorithms like Neural Network, SVM, Tree based are not make any assumption over distribution of independent variable. Because this algorithms are work on decision base concept.SO RF approach is used without data transformation.

Without using any transformation technique by using “Set 1” RF able to predict 94% of accuracy. With balanced dataset.

**Sampling:**

Down sampling of majority equal to minority class balance our dataset which helped to avoid misbalancing of target variable.

Correlation:

1st remove the feature which have correlation >70% the selecting features by using mutual information gain.

**Performance Analysis:**

**Linear regression:** Train Accuracy: - 94.75% \_\_\_ Test Accuracy: - 90.37%

**RandomForest** : Train Accuracy: - 99.87% \_\_\_ Test Accuracy: - 92.40%

By applying cross validation on both models we may see that the CV is bit higher that test data. By using different feature selection, extraction technique we able to find best features that help to reduce overfitting by cross validation prospective.