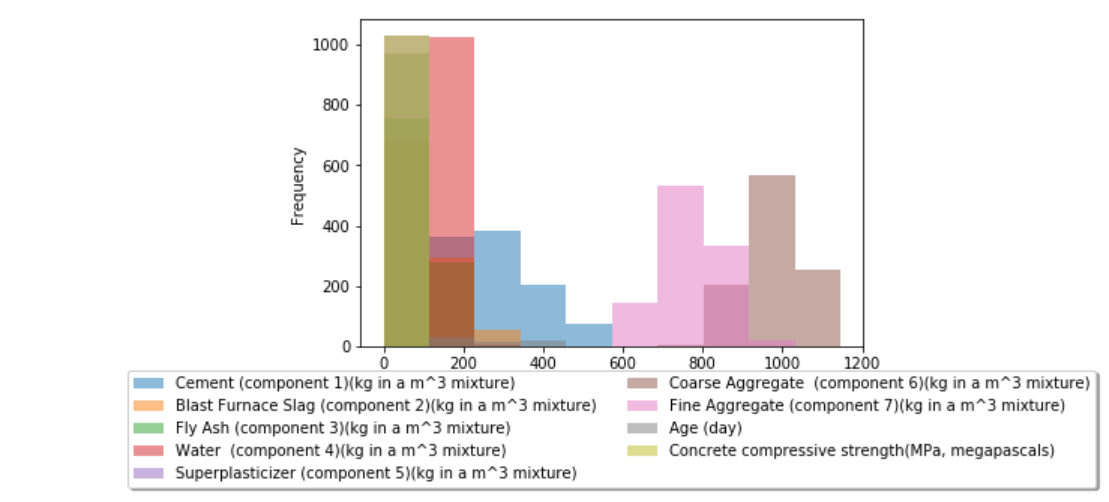
27.11.2017

**Mahalete Haile**

**Justin Uchechukwu Nwachukwu**

**ASSIGNMENT I**

Histogram showing the general distribution of the Concrete compressive strength dataset, components .

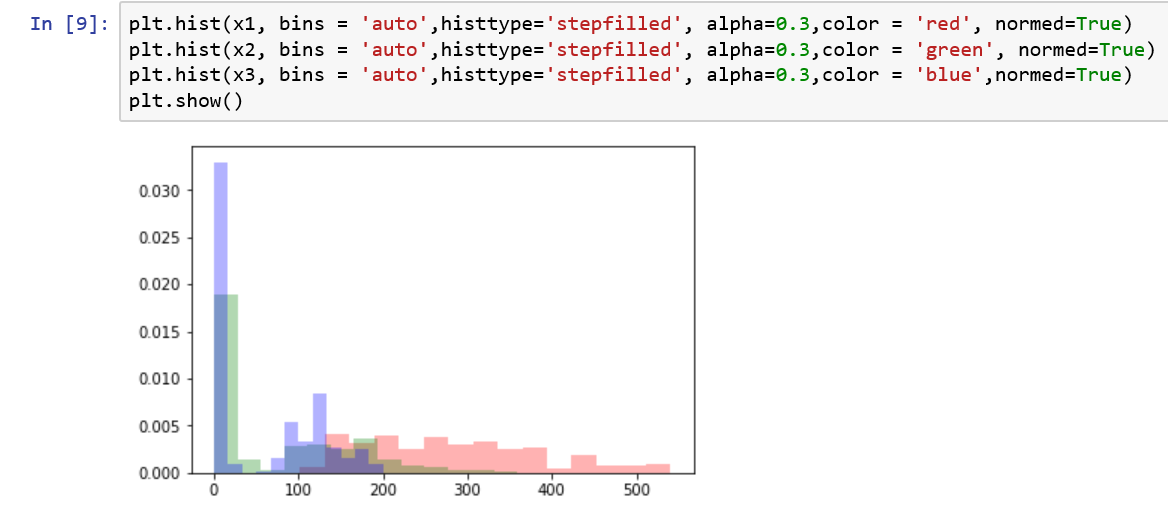


1. Plot histograms of the attributes. To determine the number of bins use at least three different methods introduced in the lectures (Sturges’ rule, Scott’s rule, square-choice, FreedmanDiaconis rule).

**Sturges’ rule, Scott’s rule, square-choice rule is used for selecting the optimal bin width for a histogram**

In Comparing the results–  there’s a high level of similarity in the histograms., between the Struges rule and Scott rule. The size of the bins is quite similar with each component . In using the square choice rule we see that the bin size has increased making the bins slightly slimmer in size.

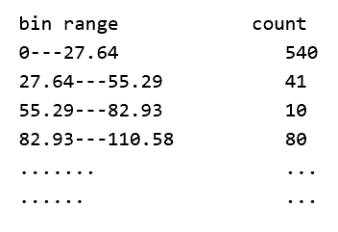
**Using bins= 'auto ' 3 component in same histogram**



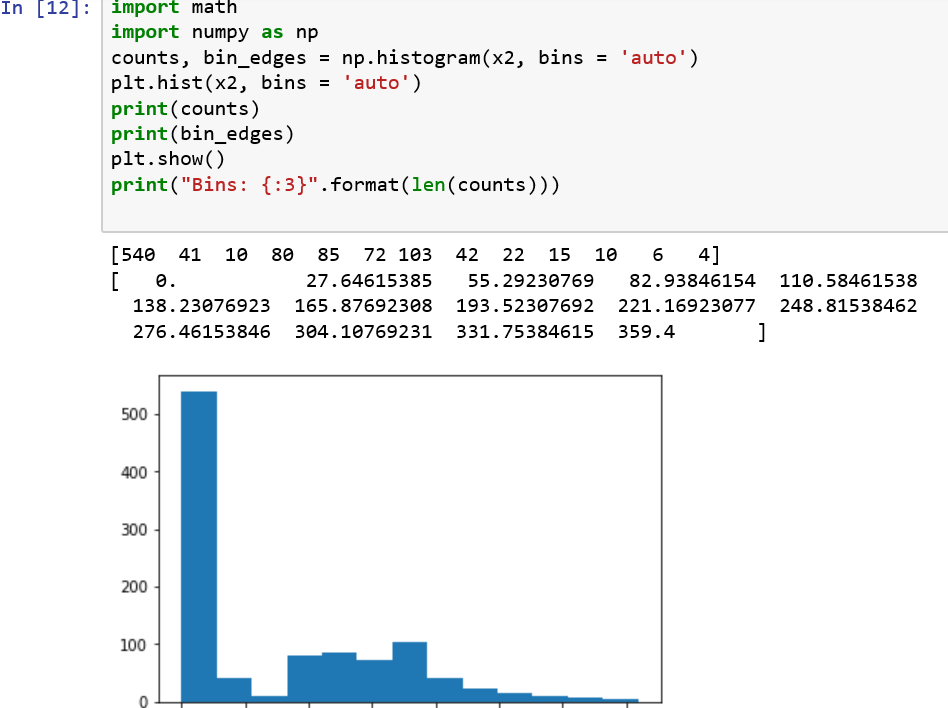
**On the other hand using Component2 = auto the bins=13**

Histogram: it is similar to barcharts ,but a histogram groups the numbers into range.(from below you can see x2 values rage from 0 to 359.5)counts: number of elements in each bin (ex:540 elements in 1st bin ,41 in second bin ...) The bin\_edges on the other hand:edges of bars(ex: in first bar 0 is one edge 27.64 is other edge of the bar(this means from 0-27.64 thereare 540 numbers in that range) , in second bar it started at 27.64 so that is one edge and other is 55.29...)

bin range count

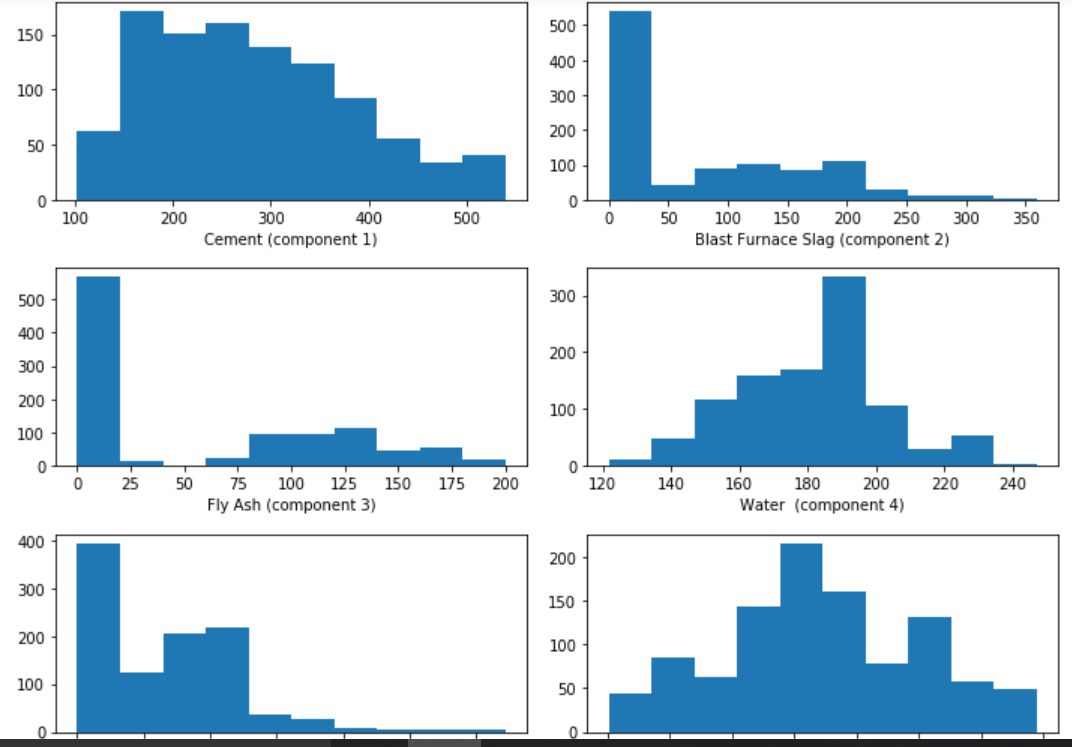


Using single component2 and setting the bins ='auto' we can see that the number of bins =13

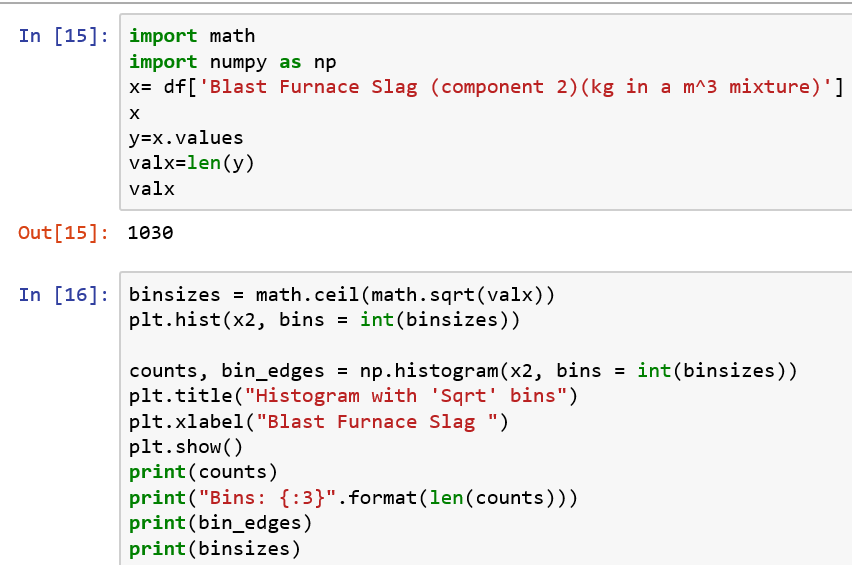


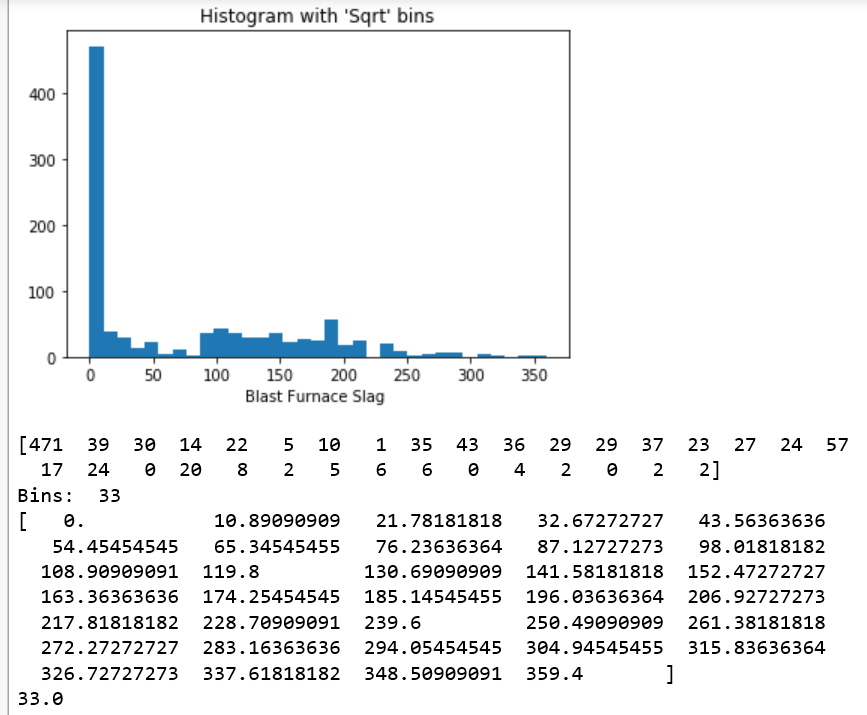
Bin:13

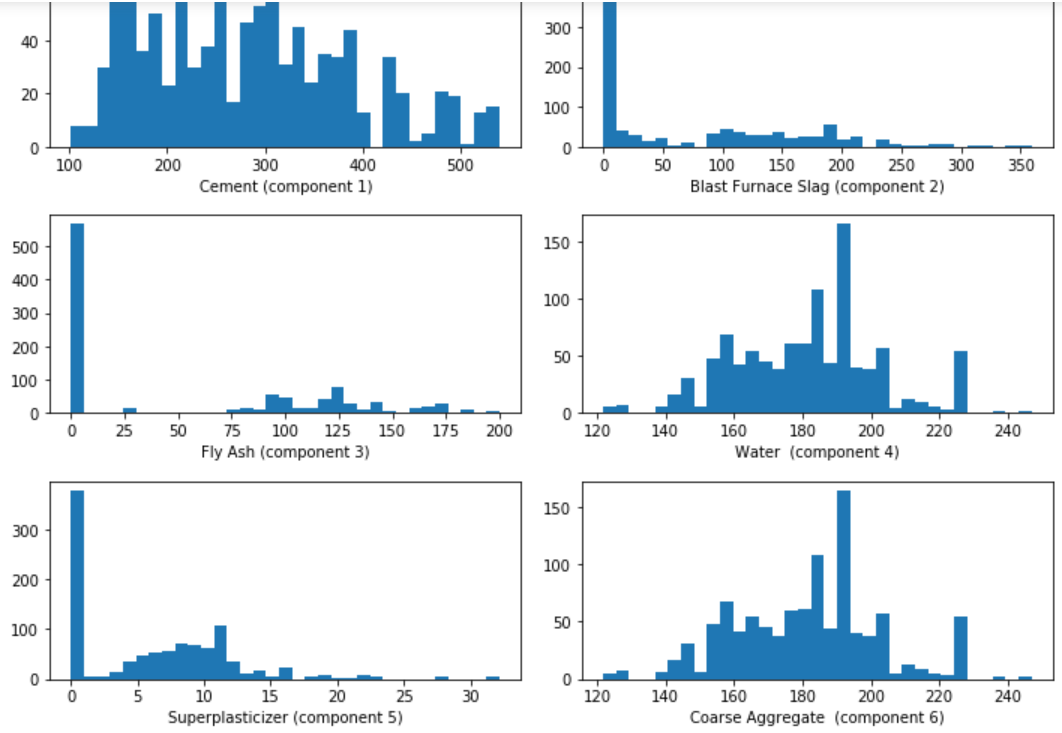
Here we can see the changes in bins for every component.



Using square-choice rule (bins =33) for component 2. Again, for the single component 2 of Blast Furnace Slag code below, we can see bins width has changed (slimed down) . We can see the same effect on each component as well.

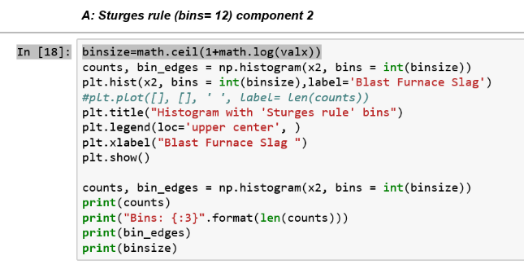


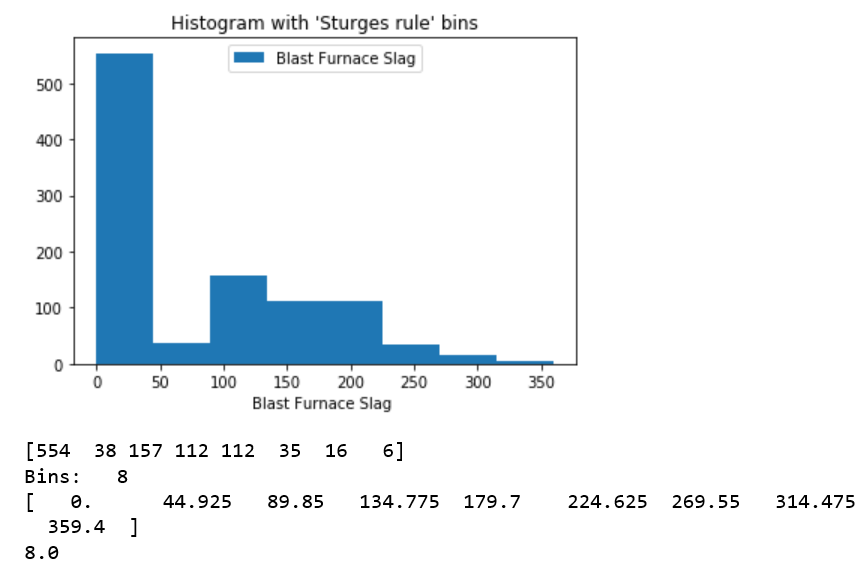


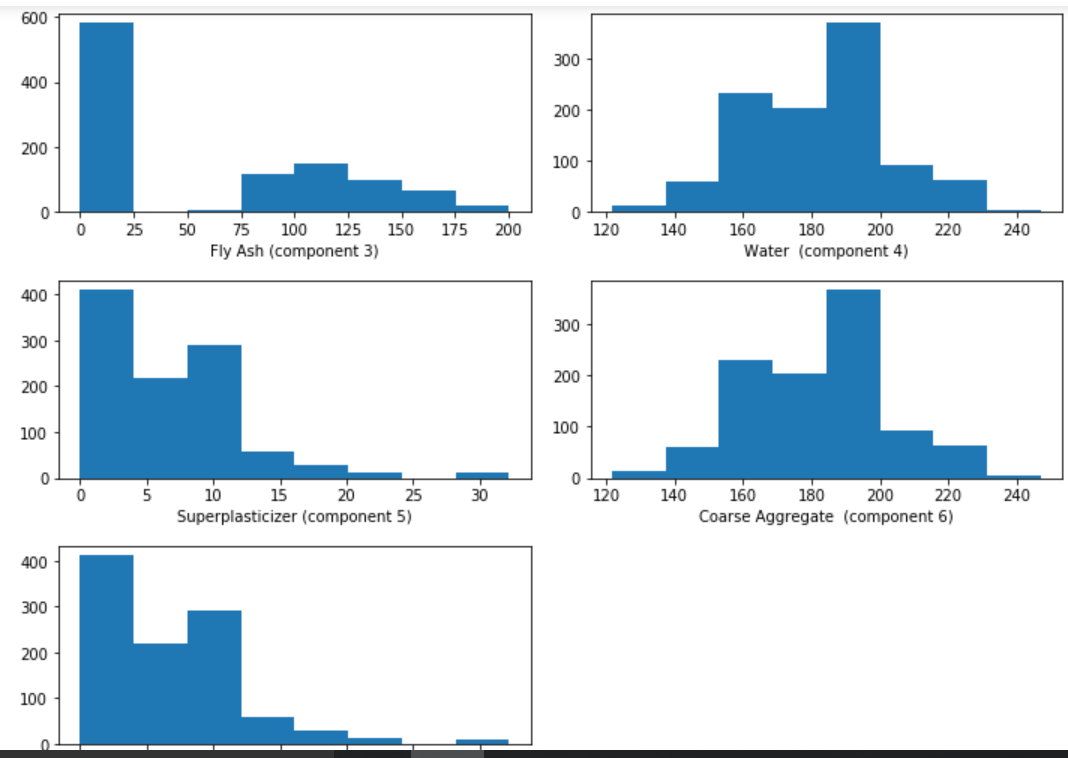


***Sturges rule***

*When plotting the histogram using Sturges rule, for component 2, we can see the number of bins is reduced to 12 . And we can see that the bins have become larger.*



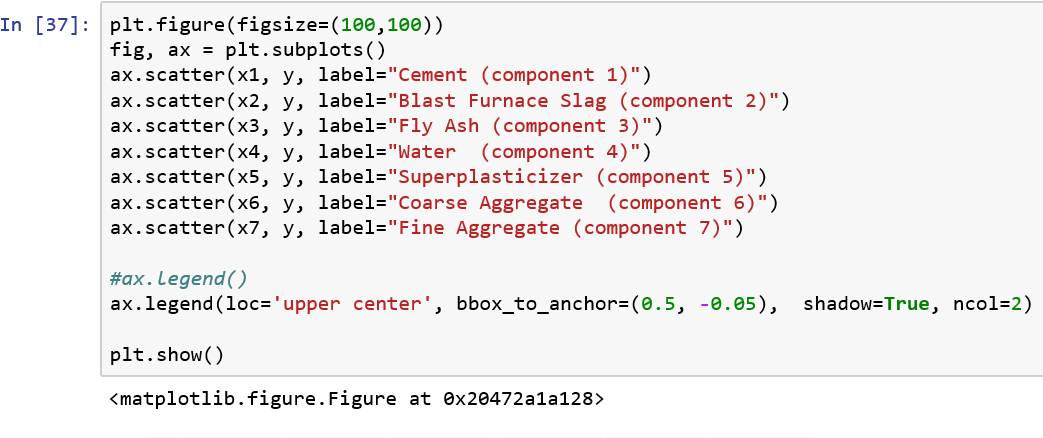


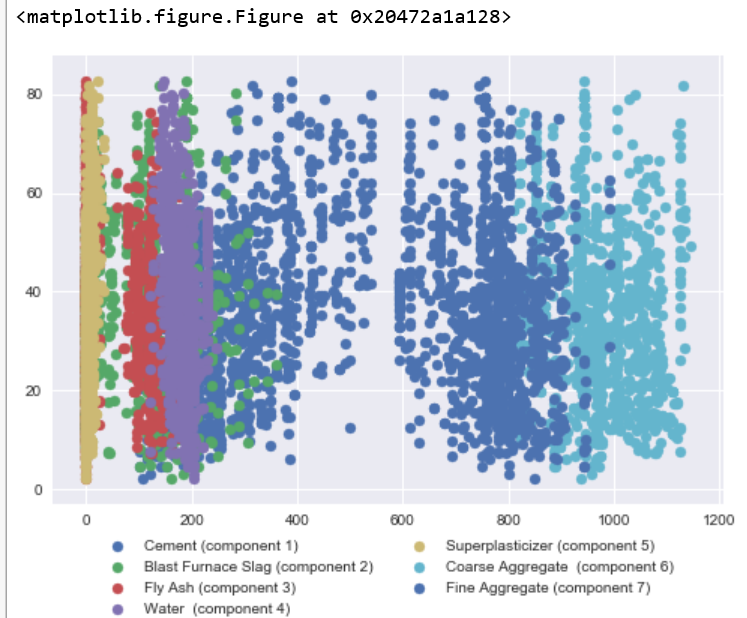


**1.2) Produce scatter plot of the data and the parallel coordinates representation (if you don’t find an implementation, it is easy to code yourself too).**

The plot below shows the scattered plot of the data. The Y axsis has theConcrete compressive strength(MPa, megapascals) and the X axsis contains all the other components.

y=df['Concrete compressive strength(MPa, megapascals) ']





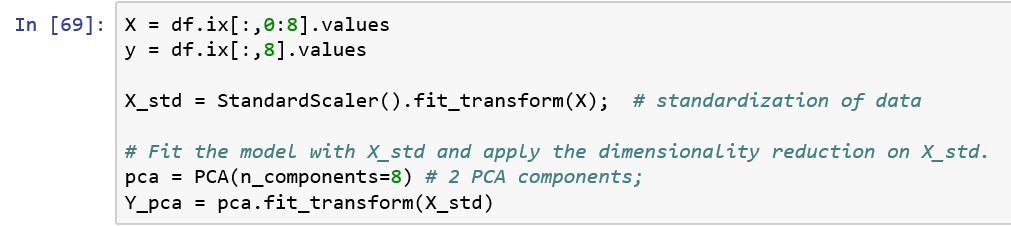
**Parallel coordinates representation of the data**

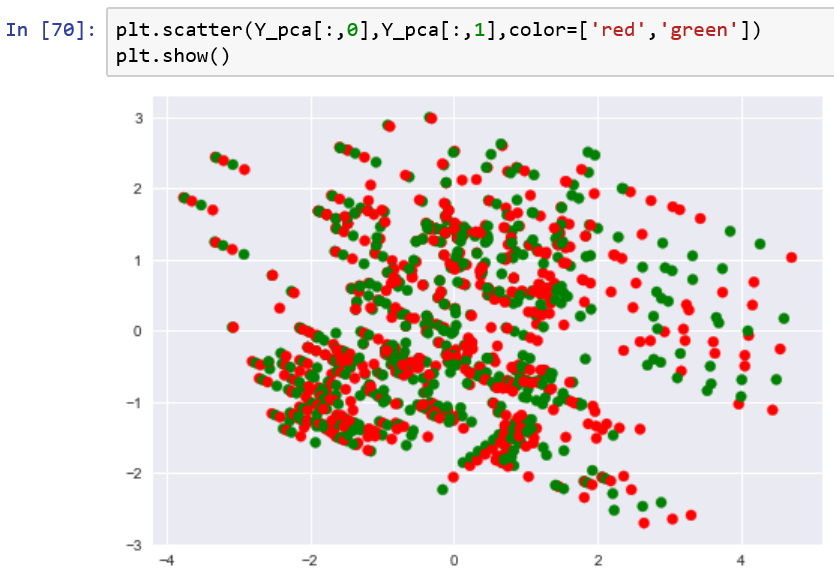


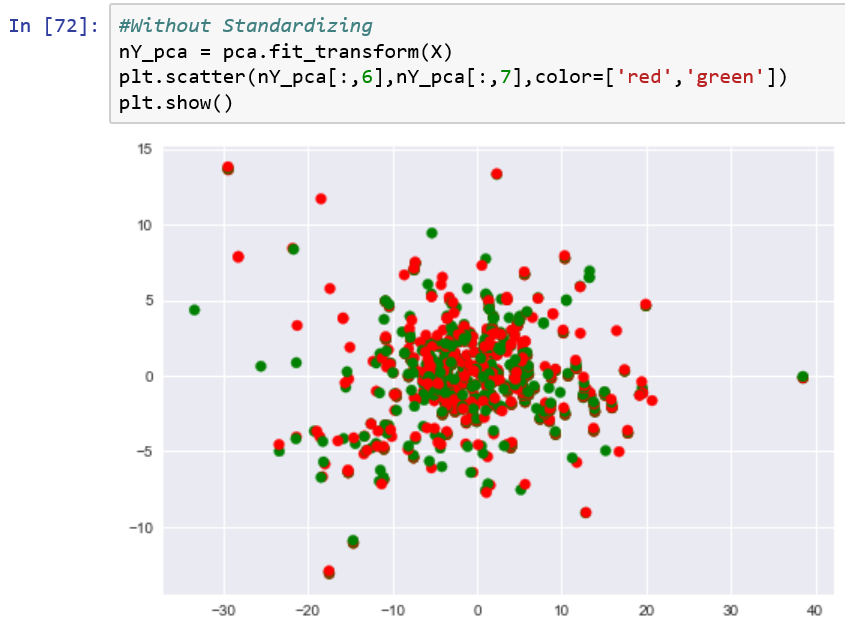
* 1. **Principal component analysis (PCA) with and without normalization: project the data to the first two principal components. Compare the results.**

We can see that the PCA in the normalized version (first picture) the dataset are more sparse while the one which is not normalized shows a more clustered dataset .The

Principle Component Analysis (PCA) is a method of dimensionality reduction. It has applications far beyond visualization. It finds new axes on which the data is most spread out. From these new axes, we can choose those with the most extreme spreading and project onto this plane







**1.5) Calculate the Pearson and Kendall’s tau correlation tables for the attributes.**



