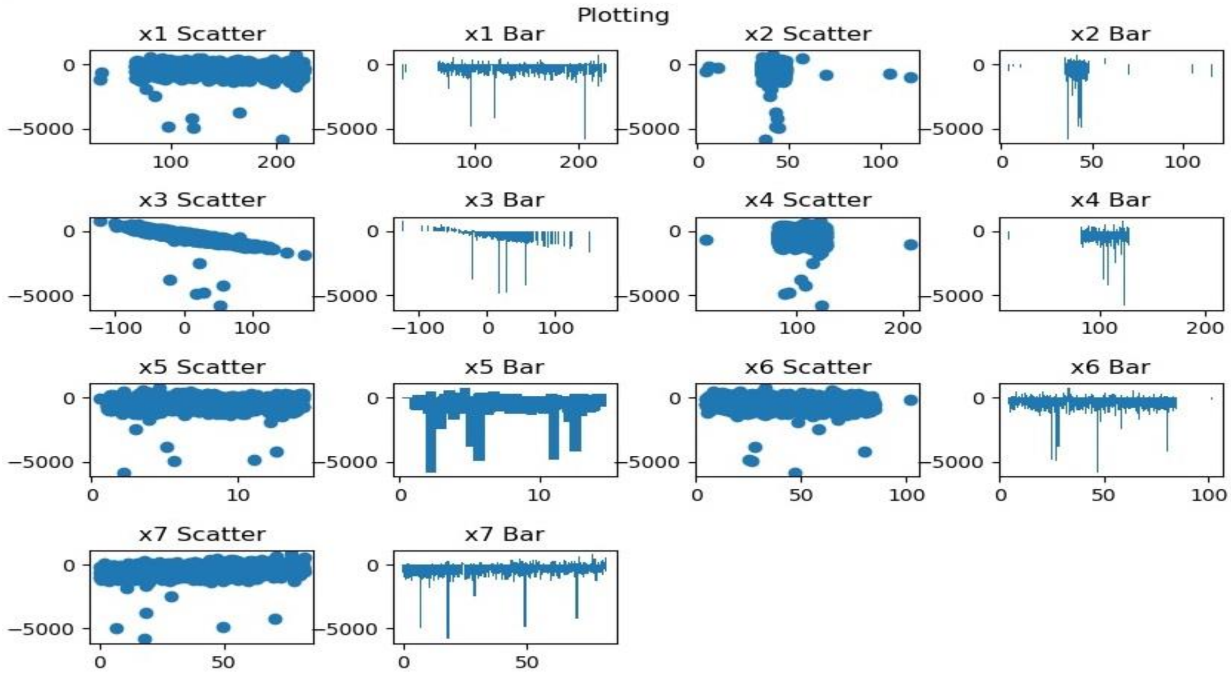
Mohanad Alhayek

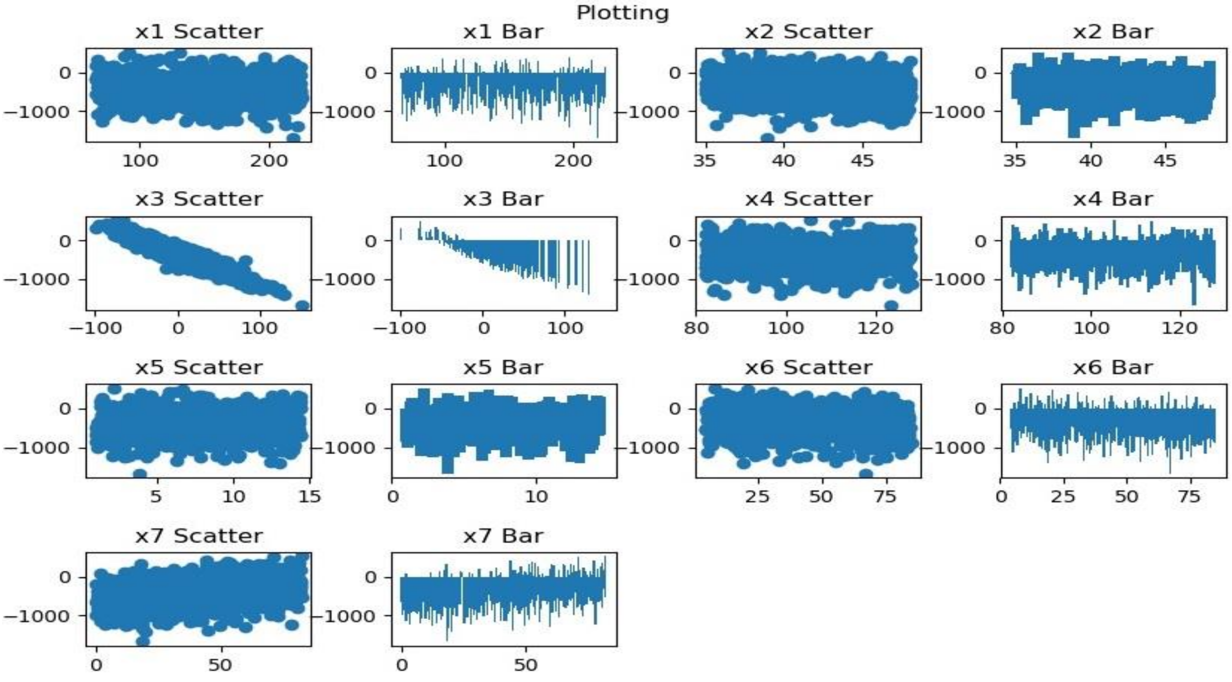
Explore Data

Before & after data:

Before Data Clean up



After Data Clean up



Cleaning process:

After I was able to plot all the data and see them, it was easier to use the scatter plots to identify the points that are much far away from the huge chunks of data. Since labels are shared between all the features (x1,x2,x3…etc) I decided to carefully cut the labels that have outliners in all the features.

Then, cutting the x1 for each feature slowly from one side of the graph (the left side first) and then work on the right side until the presented data looks as clean as possible.

It did also help me to see if my cutting code worked by running the <code>fc mo.csv cut_mo.csv</code> in the windows command terminal.

Attribute/features	Heavily or lightly correlated to the label	Uncorrelated	Positively correlated	Negatively correlated?
X1	Lightly correlated	No	False	True
X2	Lightly	Yes	N/A	N/A
X3	Heavily	No	False	True
X4	Lightly	No	False	True
X5	Lightly	Yes	N/A	N/A
X6	Lightly	No	False	True
X7	Lightly	No	True	False

Key definitions*

- Heavily correlated: correlation coefficient range (0.5 < = abs(coefficient) <= 1.00)
- Lightly correlated: correlation coefficient range (0 < abs(coefficient) < 0.5)
- Uncorrelated: the correlations coefficient is zero or linear (zero slope)
- Positively correlated: correlation coefficient is positive (coefficient > 0)
- Negatively correlated: the correlation coefficient is negative (coefficient < 0)

Coefficient is m in the linear equation of f(x)=mx+c

* Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools and techniques to build intelligent system 1st edition by Aurelien Geron (pg 83 figure 2-14)