Assume data is collected and is ready for analysis

1 Statistical analysis

1. Count, mean, median, mix, min, STD, Quartile

*display(data.describe())*

1. Correlation analysis

Create scatter matrix to observer correlation and distribution

*pd.plotting.scatter\_matrix(data, alpha = 0.3, figsize = (14,8), diagonal = 'kde');*

1. Data Preparation
2. If data is not normally distributed or skewed , it is necessary to perform non-linear transform: natural logarithm or Box-Cox transform

*np.log()*

1. Outlier removal

Interquartile range\*1.5

Remove outlier if data point is considered to be an outlier in multiple features

1. Feature transform /Variable reduction– PCA
2. Clustering and scoring

*sklearn.mixture.GausianMixture()*

*sklearn.metrics.silhouette\_score*

1. Data remapping – center is representation of each cluster

*pca.inverse\_transform(Centers)*

*np.exp(Centers)*

5 Use cluster as Engineering feature to classify new data points