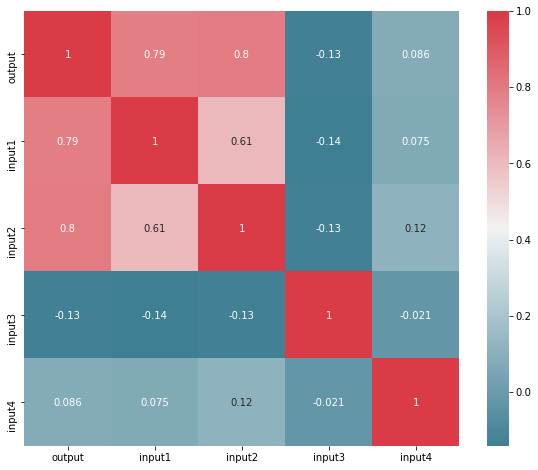
## Feature Selection - with Correlation Matrix

**1. Correlation Matrix**

A ***correlation matrix*** is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient



Implementation:

Python3

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**import** **matplotlib.pyplot** **as** **plt**

**import** **seaborn** **as** **sns**

*# import data*

my\_df = pd.read\_csv("feature\_selection\_sample\_data.csv")

*# run correlation matrix and plot*

f, ax = plt.subplots(figsize=(10, 8))

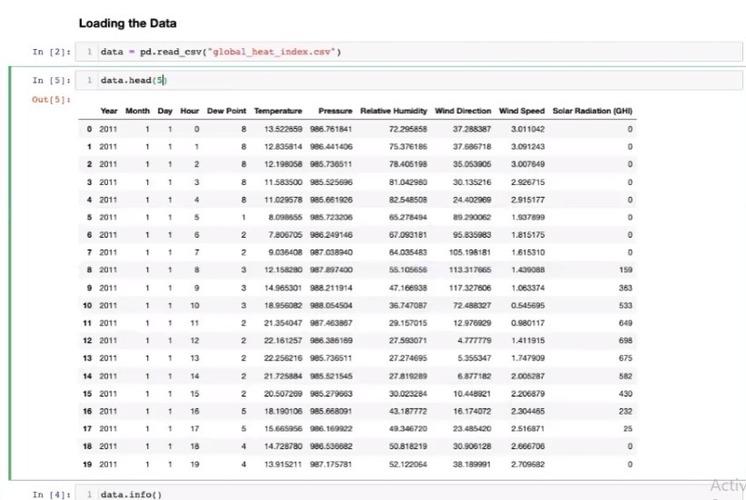
corr = my\_df.corr()

sns.heatmap(corr, mask=np.zeros\_like(corr, dtype=np.bool),

cmap=sns.diverging\_palette(220, 10, as\_cmap=**True**),

square=**True**, ax=ax)

**Let's take an example of dataset**



**MIN-Max scaling:**

**MinMaxScaler** scales all the data features in the range *[0, 1]* or else in the range *[-1, 1]* if there are negative values in the dataset. This scaling compresses all the inliers in the narrow range *[0, 0.005]*.   
In the presence of outliers, StandardScaler does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation. This leads to the shrinkage in the range of the feature values.

