Multiple Linear Regression

Data Exploration

(Multiple) Linear Regression

- In practise there is normally more than one independent variable
- The plane of best fit is given by
- $y = a_1 * x_1 + a_2 * x_2 \ldots + a_n * x_n + b$
- Again the parameters are obtained by minimising the sum of the squares of the errors.

Multiple Linear regression

$$y = \sum_{k} a_k x_k + b$$

- y is the dependent variable
- x_k are independent variables
- a_k, b are parameters
- For Example
- StackLoss = a_1*AirFlow + a_2*WaterTemp + a_3*AcidTemp + b

Stackloss Dataset

- https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/stackloss.html
- 21 observations of 4 variables
- Obtained from 21 days of operation of a plant for the oxidation of ammonia (NH3) to nitric acid (HNO3)
- Stackloss (the dependent variable) is 10 times the percentage of the ingoing ammonia to the plant that escapes

Data Exploration

```
import pandas as pd
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt


stacklossDF = pd.read_csv("data/stackloss.csv")
# print(stacklossData)

print(stacklossDF.describe())
print(stacklossDF.corr())

scatter_matrix(stacklossDF)
# plt.show()
# plt.savefig('plots/p3stacklossScatter.png')
```

Data Exploration

- pd.DataFrame.describe()
  - Summary of numeric features
  - Mean, Min, Max, std etc.
- pd.DataFrame.corr()
  - Matrix of correlation coefficients between the variables
- pd.scatter_matrix()
  - A matrix of scatterplots

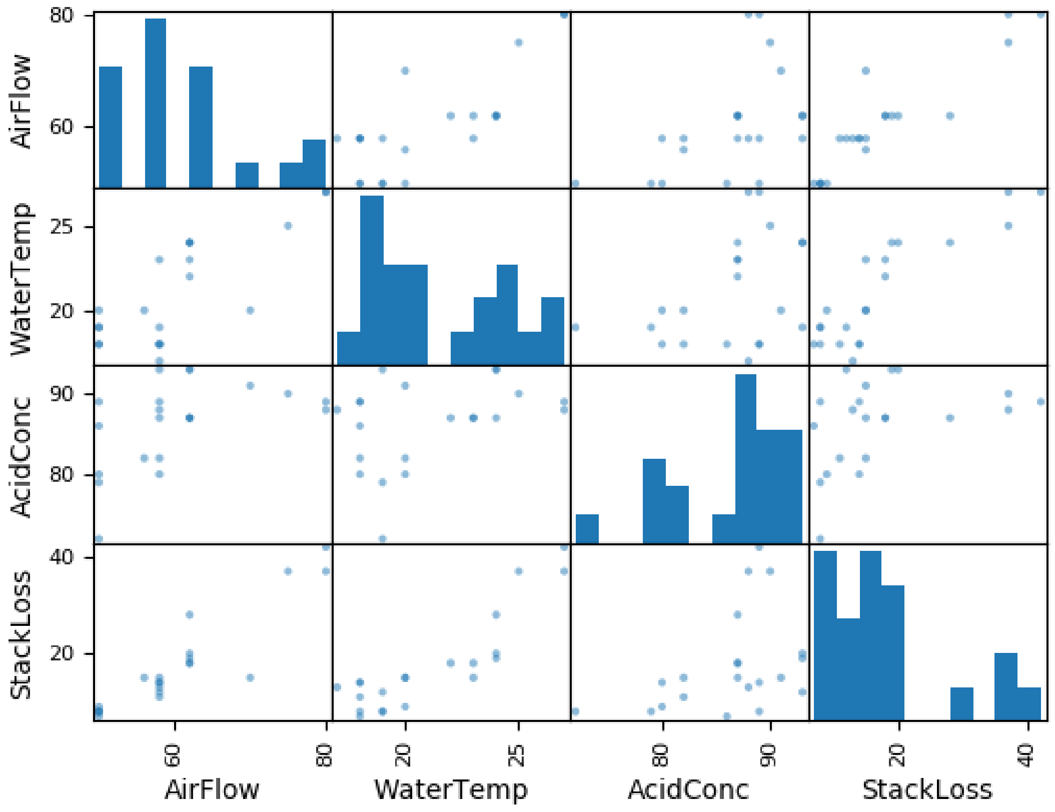
DataFrame.describe()

```
         AirFlow    WaterTemp  AcidConc   StackLoss
count  21.000000  21.000000   21.000000  21.000000
mean   60.428571  21.095238   86.285714  17.523810
std     9.168268   3.160771    5.358571  10.171623
min    50.000000  17.000000   72.000000   7.000000
25%    56.000000  18.000000   82.000000  11.000000
50%    58.000000  20.000000   87.000000  15.000000
75%    62.000000  24.000000   89.000000  19.000000
max    80.000000  27.000000   93.000000  42.000000
```

DataFrtame.corr()

```
            AirFlow   WaterTemp  AcidConc  StackLoss
AirFlow     1.000000  0.781852   0.500143  0.919663
WaterTemp   0.781852  1.000000   0.390940  0.875504
AcidConc    0.500143  0.390940   1.000000  0.399830
StackLoss   0.919663  0.875504   0.399830  1.000000
```

pandas.plotting.scatter_matrix()



Scatter Plot Matrix - Analysis

- There seems to be a fairly strong linear correlation between StackLoss and AirFlow
- Also between StackLoss and WaterTemp
- Not so strong between StackLoss and AcidConc
- Maybe even a non-linear relationship between StackLoss and AcidConc.
- Even so, the normal thing to do is proceed using all three variables.
- We would not be surprised if AcidConc is not as strong a predictor as the other variables.

Marginal Relationships

- The scatter plots show the marginal relationships between variables without regard to other variables.
- Note that the absence of a correlation between an independent variable (predictor) and the dependent variable (output) does not mean that the dependent variable is not useful as a predictor.
- (For example, points that should be near a line of best fit could be being moved away from it by the values of other variables.)

Marginal Relationships

- With multiple linear regression we are primarily concerned with how the dependent or output variable relates to the independent (predictor) variable simultaneously.

Feature Reduction

- It is nearly always safest to use all the features unless there is a good reason not to.
- If a predictor variable has close to zero variance (it does not vary much) then it can have very little predictor value.
- If two predictor variables are very strongly correlated, then one is redundant and possibly can be omitted.
- For performance reasons a model with a reduced set of features with similar accuracy to the full model can be used.