# Linear Regression in Python

Use markdown to indicate

- objectives
- data exploration
- model building
- making predictions.

Also comment on each of these

## Q1. Simple Linear Regression Model

```
x = np.array([1, 2, 3, 4, 5, 6]]).reshape(-1, 1)
y = np.array([6,1,9,5,17,12])
```

- Explore the data (scatter plot)
- Build a linear regression model for y in terms of x.
  - Print out the slope and intercept
  - Print $R^{2,}$ (model.score(x,y)
  - Comment on the model.
- Find the predicted values of y for the values of x.
  - Draw a scatter plot for both actual and predicted values.
  - Draw the line of best fit.

## Q2. Calculate Linear Regression Parameters

```
x = np.array([1, 2, 3, 4, 5, 6]])
y = np.array([6,1,9,5,17,12])
```

For the regression line y = a x + b, and x and y given above, calculate the values of a and b using the following equations.

$$a = (n\Sigma xy - (\Sigma x)(\Sigma y)) / (n\Sigma x^2 – (\Sigma x)^2)$$

$$b = (\Sigma y - a(\Sigma x)) / n$$

a is the slope and b is the intercept in the linear regression model. The values of a and b should be the same as found by LinearRegression fit() function.

Use x*y to multiply the corresponding elements of x and y.
Use sum (or np.sum) to sum the elements of an array.

## Q4. Linear Regression – Stackloss

- Read the stackloss dataset from stackloss.csv.
- Data Exploration
  - Get a summary of numerical features
  - Get a matrix of correlation coefficients between the variables
  - Draw a matrix of scatterplots.
  - Comment - which input variables correlate the most strongly with stackloss?
- Build a linear regression model for stackloss in terms of the other (input) variables.
  - Split the data into training and test data.
  - Build the model using the training data.
  - Print out intercept, coefficients and $R^2$.
- For the new input data make predictions for the stackloss
  newData= [[72, 20, 85], [75, 25, 80]]
  - comment on the values obtained.

## Q5a. Linear Regression – Fish

- Read the fish dataset from fish.csv.
- Explore the data
  - Get a summary of numerical features
  - Get a matrix of correlation coefficients between the variables
  - Draw a matrix of scatterplots.
  - Which input variables correlate the most strongly with weight?
- Build a Linear Regression Model **for weight in terms of the other numeric input variables.**
  - Define X and y
  - Split data into training and test data.
  - Print out the intercept and the coefficients of the input variables.
  - Print out the value of $R^2$
  - Print out the value of the RMSE for the test data.

## Q5b Linear Regression – Fish

- Add to the the script above.
- Build a model where the weight depends on all the other variables including species.
- Onehot encode the data using  X = pd.get_dummies(X).
- Compare the $R^2$ and RMSE with the values obtained without Species.