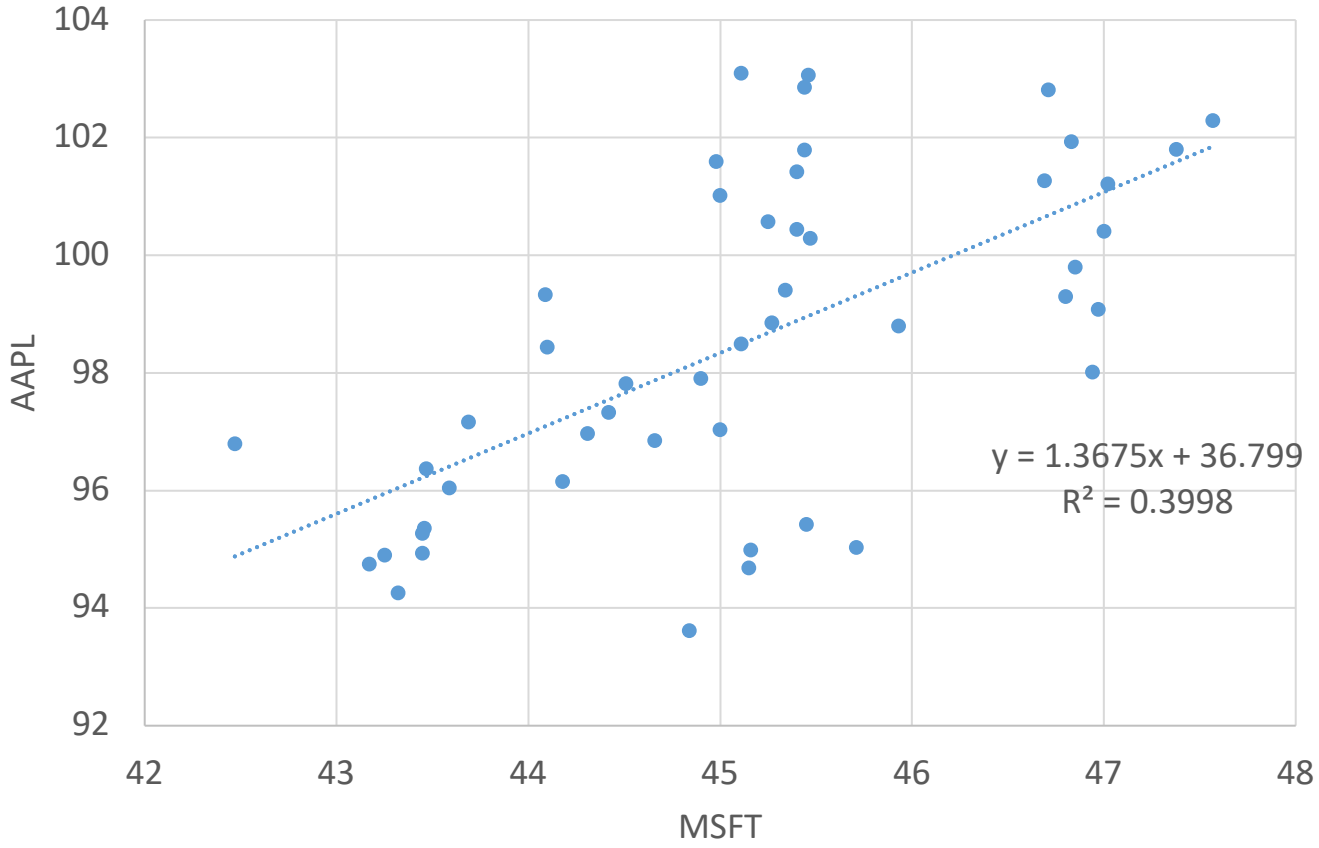
Understanding statistics and Experimental design

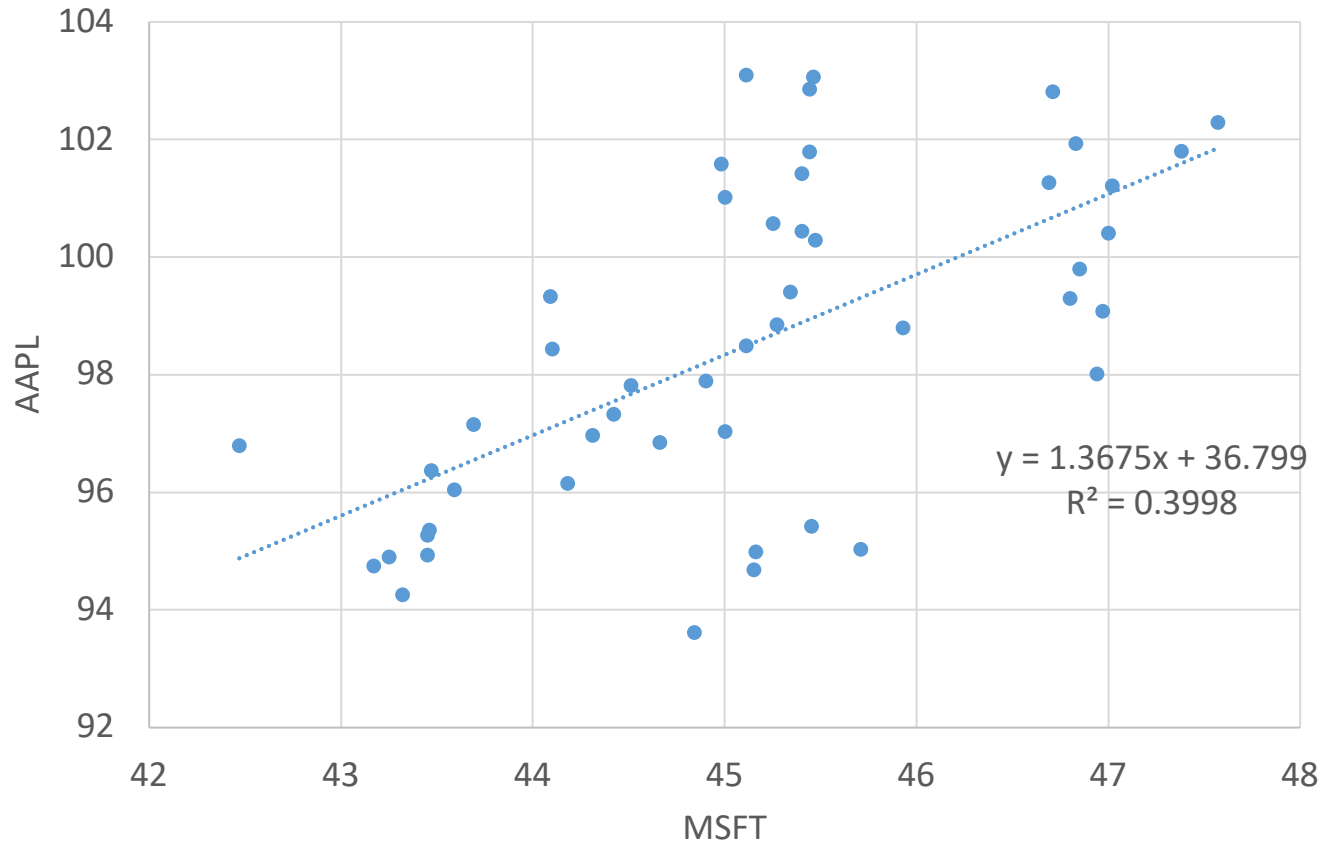
Exercise Week 10

Correlation between MSFT and AAPL

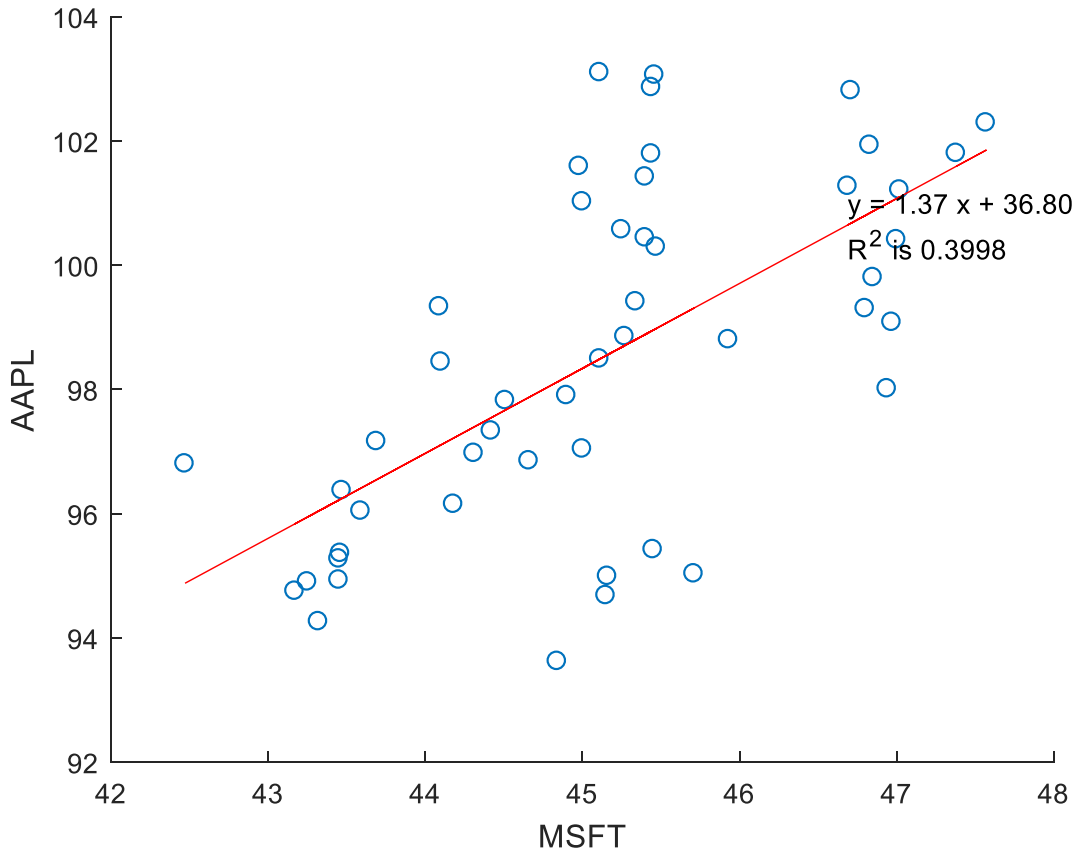


Excel

Select dataset -> Insert -> Charts -> insert scatter plot
Add (+ sign) trend line -> double click trend line -> check display Equation on chart & display R-squared value on chart

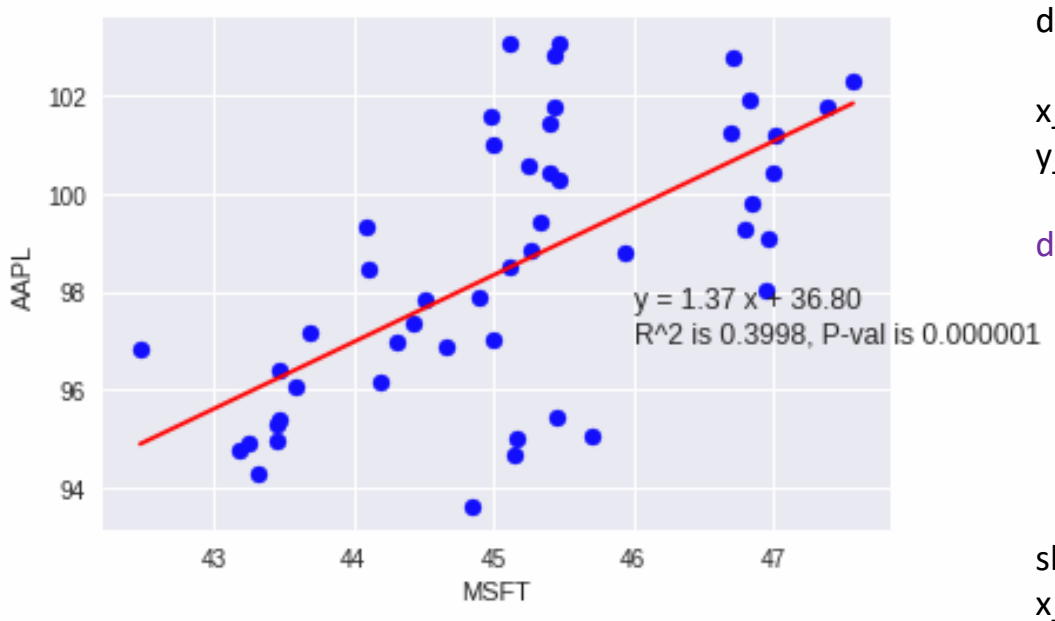


Matlab




```
data = readtable('CorrelationRegression-data.csv');

x_data = data.MSFT;
y_data = data.AAPL;

p_fit = polyfit(x_data,y_data,1);

y_fit = polyval(p_fit,x_data);

y_resid = y_data - y_fit;

SSresid = sum(y_resid.^2);
SStotal = (length(y_data)-1) * var(y_data);

rsq = 1 - SSresid/SStotal;

figure
hold on
% plot(data.MSFT, data.AAPL, 'ob')
scatter(x_data,y_data)
plot(x_data,y_fit,'r')
txt = sprintf('y = %.2f x + %.2f', p_fit(1), p_fit(2));
text(46, 96, txt, 'HorizontalAlignment','left')
xlabel('MSFT')
ylabel('AAPL')
```

Python




```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats

data = pd.read_csv('CorrelationRegression-data.csv')

x_data = np.array(data["MSFT"])
y_data = np.array(data["AAPL"])

def plot_points(data):
    x_data = np.array(data["MSFT"])
    y_data = np.array(data["AAPL"])
    plt.scatter(x_data, y_data, color = 'blue')
    plt.xlabel('MSFT')
    plt.ylabel('AAPL')

slope, intercept, r_value, p_value, std_err = stats.linregress(x_data,y_data)
x_fit = np.linspace(min(x_data), max(x_data), 1000)
y_fit = slope*x_fit+intercept

plot_points(data)
plt.plot(x_fit,y_fit, color='red')
txt = 'y = %.2f x + %.2f \nR^2 is %.2f, P-val is %f' %(slope, intercept, r_value, p_value)
plt.text(46, 97, txt, fontsize=12)
plt.show()
```