Project 1

Anscombe Quartet

Imports

In [ ]:
```
# The Anscombe Quartet (Python)

# demonstration data from
# Anscombe, F. J. 1973, February. Graphs in statistical analysis.
#  The American Statistician 27: 17-21.

# prepare for Python version 3x features and functions
from __future__ import division, print_function

# import packages for Anscombe Quartet demonstration
import pandas as pd  # data frame operations
import numpy as np  # arrays and math functions
import statsmodels.api as sm  # statistical models (including regression)
import matplotlib.pyplot as plt  # 2D plotting
```

Set, shift, and analyze data

In [ ]:
```
# define the anscombe data frame using dictionary of equal-length lists
anscombe = pd.DataFrame({'x1' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x2' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x3' : [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5],
    'x4' : [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8],
    'y1' : [8.04, 6.95,  7.58, 8.81, 8.33, 9.96, 7.24, 4.26,10.84, 4.82, 5.68],
    'y2' : [9.14, 8.14,  8.74, 8.77, 9.26, 8.1, 6.13, 3.1,  9.13, 7.26, 4.74],
    'y3' : [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73],
    'y4' : [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89]})

# shifting the data by 6
anscombe = anscombe + 6
anscombe
```

Out[ ]:

	x1	x2	x3	x4	y1	y2	y3	y4
0	16	16	16	14	14.04	15.14	13.46	12.58
1	14	14	14	14	12.95	14.14	12.77	11.76
2	19	19	19	14	13.58	14.74	18.74	13.71
3	15	15	15	14	14.81	14.77	13.11	14.84
4	17	17	17	14	14.33	15.26	13.81	14.47
5	20	20	20	14	15.96	14.10	14.84	13.04
6	12	12	12	14	13.24	12.13	12.08	11.25
7	10	10	10	25	10.26	9.10	11.39	18.50
8	18	18	18	14	16.84	15.13	14.15	11.56
9	13	13	13	14	10.82	13.26	12.42	13.91
10	11	11	11	14	11.68	10.74	11.73	12.89

In [ ]:

```
anscombe.describe()
```

Out[ ]:

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	15.000000	15.000000	15.000000	15.000000	13.500909	13.500909	13.500000	13.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	10.000000	10.000000	10.000000	14.000000	10.260000	9.100000	11.390000	11.250000
25%	12.500000	12.500000	12.500000	14.000000	12.315000	12.695000	12.250000	12.170000
50%	15.000000	15.000000	15.000000	14.000000	13.580000	14.140000	13.110000	13.040000
75%	17.500000	17.500000	17.500000	14.000000	14.570000	14.950000	13.980000	14.190000
max	20.000000	20.000000	20.000000	25.000000	16.840000	15.260000	18.740000	18.500000

Create linear regression models

In [ ]:

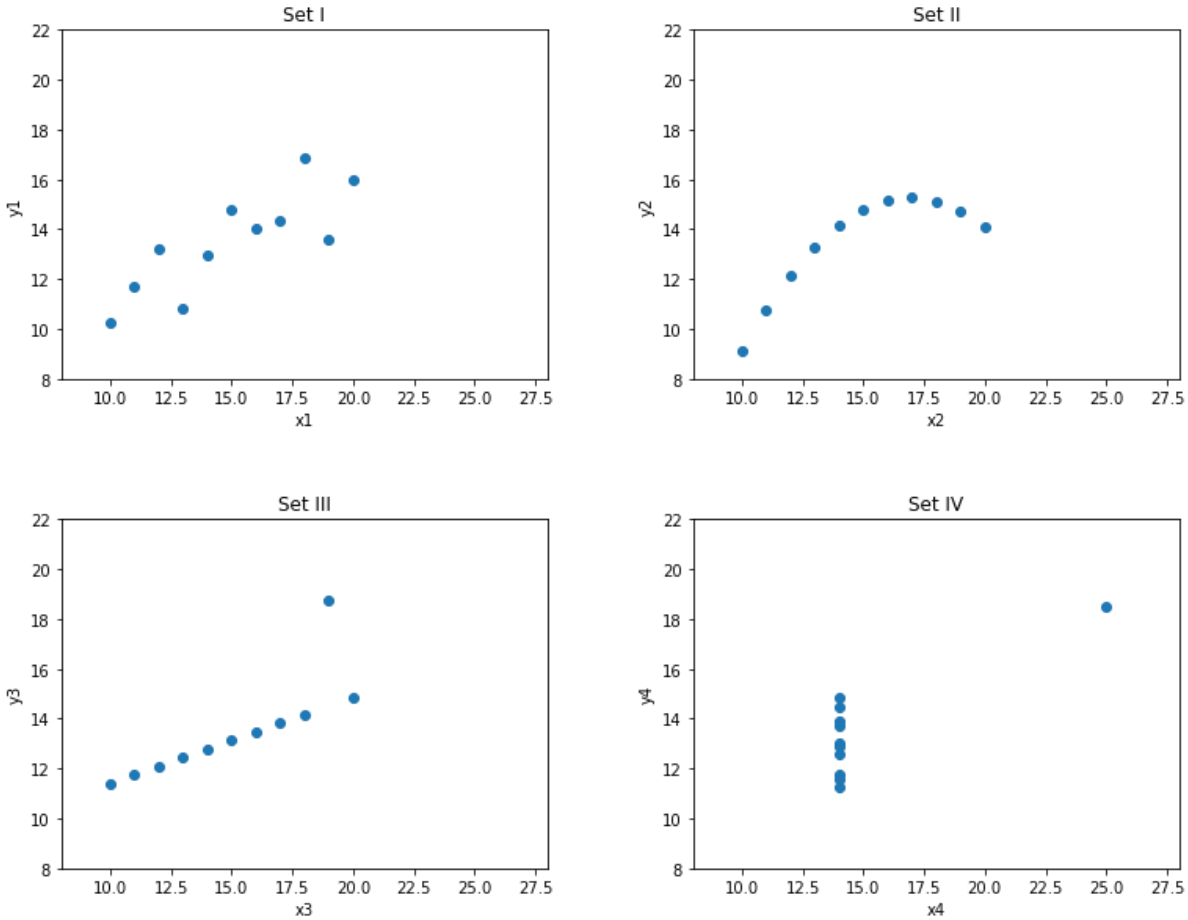
```
# fit linear regression models by ordinary least squares
set_I_design_matrix = sm.add_constant(anscombe['x1'])
set_I_model = sm.OLS(anscombe['y1'], set_I_design_matrix)
print(set_I_model.fit().summary())

set_II_design_matrix = sm.add_constant(anscombe['x2'])
set_II_model = sm.OLS(anscombe['y2'], set_II_design_matrix)
print(set_II_model.fit().summary())

set_III_design_matrix = sm.add_constant(anscombe['x3'])
set_III_model = sm.OLS(anscombe['y3'], set_III_design_matrix)
print(set_III_model.fit().summary())

set_IV_design_matrix = sm.add_constant(anscombe['x4'])
set_IV_model = sm.OLS(anscombe['y4'], set_IV_design_matrix)
print(set_IV_model.fit().summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     y1   R-squared:                       0.667
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     17.99
Date:                Fri, 06 Jan 2023   Prob (F-statistic):            0.00217
Time:                        19:38:59   Log-Likelihood:                -16.841
No. Observations:                  11   AIC:                             37.68
Df Residuals:                       9   BIC:                             38.48
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.9995      1.807      3.319      0.009       1.911      10.088
x1             0.5001      0.118      4.241      0.002       0.233       0.767
==============================================================================
Omnibus:                        0.082   Durbin-Watson:                   3.212
Prob(Omnibus):                  0.960   Jarque-Bera (JB):                0.289
Skew:                          -0.122   Prob(JB):                        0.865
Kurtosis:                       2.244   Cond. No.                         74.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly sp
ecified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                     y2   R-squared:                       0.666
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     17.97
Date:                Fri, 06 Jan 2023   Prob (F-statistic):            0.00218
Time:                        19:38:59   Log-Likelihood:                -16.846
No. Observations:                  11   AIC:                             37.69
Df Residuals:                       9   BIC:                             38.49
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.0009      1.808      3.318      0.009       1.910      10.092
x2             0.5000      0.118      4.239      0.002       0.233       0.767
==============================================================================
Omnibus:                        1.594   Durbin-Watson:                   2.188
Prob(Omnibus):                  0.451   Jarque-Bera (JB):                1.108
Skew:                          -0.567   Prob(JB):                        0.575
Kurtosis:                       1.936   Cond. No.                         74.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly sp
ecified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                     y3   R-squared:                       0.666
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     17.97
Date:                Fri, 06 Jan 2023   Prob (F-statistic):            0.00218
Time:                        19:38:59   Log-Likelihood:                -16.838
```

```
No. Observations:                  11   AIC:                             37.68
Df Residuals:                       9   BIC:                             38.47
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.0041      1.807      3.323      0.009       1.916      10.092
x3             0.4997      0.118      4.239      0.002       0.233       0.766
==============================================================================
Omnibus:                       19.540   Durbin-Watson:                   2.144
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               13.478
Skew:                           2.041   Prob(JB):                      0.00118
Kurtosis:                       6.571   Cond. No.                         74.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly sp
ecified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                     y4   R-squared:                       0.667
Model:                            OLS   Adj. R-squared:                  0.630
Method:                 Least Squares   F-statistic:                     18.00
Date:                Fri, 06 Jan 2023   Prob (F-statistic):            0.00216
Time:                        19:38:59   Log-Likelihood:                -16.833
No. Observations:                  11   AIC:                             37.67
Df Residuals:                       9   BIC:                             38.46
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.0023      1.806      3.323      0.009       1.917      10.088
x4             0.4999      0.118      4.243      0.002       0.233       0.766
==============================================================================
Omnibus:                        0.555   Durbin-Watson:                   1.662
Prob(Omnibus):                  0.758   Jarque-Bera (JB):                0.524
Skew:                           0.010   Prob(JB):                        0.769
Kurtosis:                       1.931   Cond. No.                         74.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly sp
ecified.
```

```
c:\Python310\lib\site-packages\scipy\stats\_stats_py.py:1477: UserWarning: kurtosis
test only valid for n>=20 ... continuing anyway, n=11
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
c:\Python310\lib\site-packages\scipy\stats\_stats_py.py:1477: UserWarning: kurtosis
test only valid for n>=20 ... continuing anyway, n=11
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
c:\Python310\lib\site-packages\scipy\stats\_stats_py.py:1477: UserWarning: kurtosis
test only valid for n>=20 ... continuing anyway, n=11
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
c:\Python310\lib\site-packages\scipy\stats\_stats_py.py:1477: UserWarning: kurtosis
test only valid for n>=20 ... continuing anyway, n=11
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

In [ ]:
```
# create scatter plots
fig = plt.figure(figsize=(12,9))

set_I = fig.add_subplot(2, 2, 1)
set_I.scatter(anscombe['x1'],anscombe['y1'])
set_I.set_title('Set I')
set_I.set_xlabel('x1')
set_I.set_ylabel('y1')
set_I.set_xlim(8, 28)
set_I.set_ylim(8, 22)

set_II = fig.add_subplot(2, 2, 2)
set_II.scatter(anscombe['x2'],anscombe['y2'])
set_II.set_title('Set II')
set_II.set_xlabel('x2')
set_II.set_ylabel('y2')
set_II.set_xlim(8, 28)
set_II.set_ylim(8, 22)

set_III = fig.add_subplot(2, 2, 3)
set_III.scatter(anscombe['x3'],anscombe['y3'])
set_III.set_title('Set III')
set_III.set_xlabel('x3')
set_III.set_ylabel('y3')
set_III.set_xlim(8, 28)
set_III.set_ylim(8, 22)

set_IV = fig.add_subplot(2, 2, 4)
set_IV.scatter(anscombe['x4'],anscombe['y4'])
set_IV.set_title('Set IV')
set_IV.set_xlabel('x4')
set_IV.set_ylabel('y4')
set_IV.set_xlim(8, 28)
set_IV.set_ylim(8, 22)

plt.subplots_adjust(left=0.1, right=0.925, top=0.925, bottom=0.1,
    wspace = 0.3, hspace = 0.4)
```


Set I
Set II
Set III
Set IV
x1
y1
x2
y2
x3
y3
x4
y4