HW3 答案

1. 從 UCI 下載 Concrete Compressive Strength Data Set

下載檔案

In [24]:
```
import pandas as pd
df=pd.read_excel("/Users/chuanyang/Downloads/Concrete_Data.xls")
df.head()
```

Out[24]:

	Cement (component 1) (kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3) (kg in a m^3 mixture)	Water (component 4) (kg in a m^3 mixture)	Superplasticizer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.986111
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.887366
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.269535
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.052780
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.296075

更改column name

In [27]:
```
df.columns =['Cement', 'Blast Furnace Slag', 'Fly Ash', 'Water', 'Superplasticizer', 'Coarse Aggregate',
'Fine Aggregate', 'Age', 'Concrete compressive strength']
df.head()
```

Out[27]:

	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Concrete compressive strength
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.986111
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.887366
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.269535
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.052780
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.296075

Name -- Data Type -- Measurement -- Description

Cement (component 1) -- quantitative -- kg in a m3 mixture -- Input Variable
Blast Furnace Slag (component 2) -- quantitative -- kg in a m3 mixture -- Input Variable
Fly Ash (component 3) -- quantitative -- kg in a m3 mixture -- Input Variable
Water (component 4) -- quantitative -- kg in a m3 mixture -- Input Variable
Superplasticizer (component 5) -- quantitative -- kg in a m3 mixture -- Input Variable
Coarse Aggregate (component 6) -- quantitative -- kg in a m3 mixture -- Input Variable
Fine Aggregate (component 7) -- quantitative -- kg in a m3 mixture -- Input Variable
Age -- quantitative -- Day (1~365) -- Input Variable
Concrete compressive strength -- quantitative -- MPa -- Output Variable

2. 請算出9個變數間的相關係數

*** 方法1 :直接的出相關係數

In [28]:
```
df.corr()
```

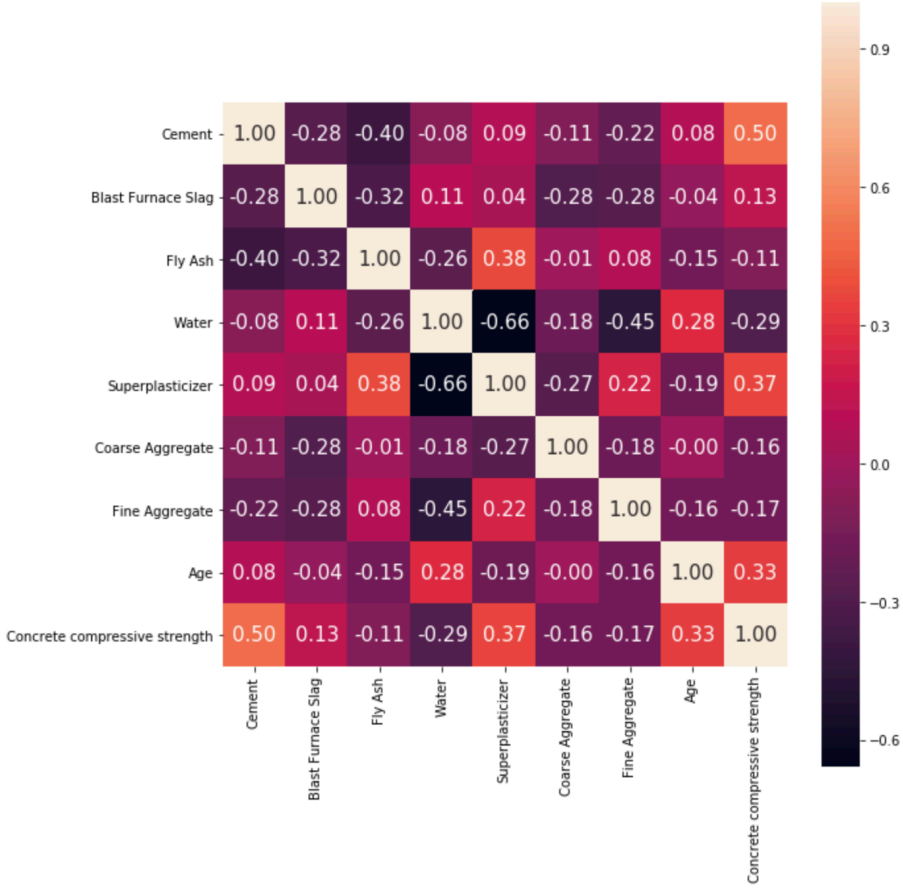
Out[28]:

	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Concrete compressive strength
Cement	1.000000	-0.275193	-0.397475	-0.081544	0.092771	-0.109356	-0.222720	0.081947	0.497833
Blast Furnace Slag	-0.275193	1.000000	-0.323569	0.107286	0.043376	-0.283998	-0.281593	-0.044246	0.134824
Fly Ash	-0.397475	-0.323569	1.000000	-0.257044	0.377340	-0.009977	0.079076	-0.154370	-0.105753
Water	-0.081544	0.107286	-0.257044	1.000000	-0.657464	-0.182312	-0.450635	0.277604	-0.289613
Superplasticizer	0.092771	0.043376	0.377340	-0.657464	1.000000	-0.266303	0.222501	-0.192717	0.366102
Coarse Aggregate	-0.109356	-0.283998	-0.009977	-0.182312	-0.266303	1.000000	-0.178506	-0.003016	-0.164928
Fine Aggregate	-0.222720	-0.281593	0.079076	-0.450635	0.222501	-0.178506	1.000000	-0.156094	-0.167249
Age	0.081947	-0.044246	-0.154370	0.277604	-0.192717	-0.003016	-0.156094	1.000000	0.328877
Concrete compressive strength	0.497833	0.134824	-0.105753	-0.289613	0.366102	-0.164928	-0.167249	0.328877	1.000000

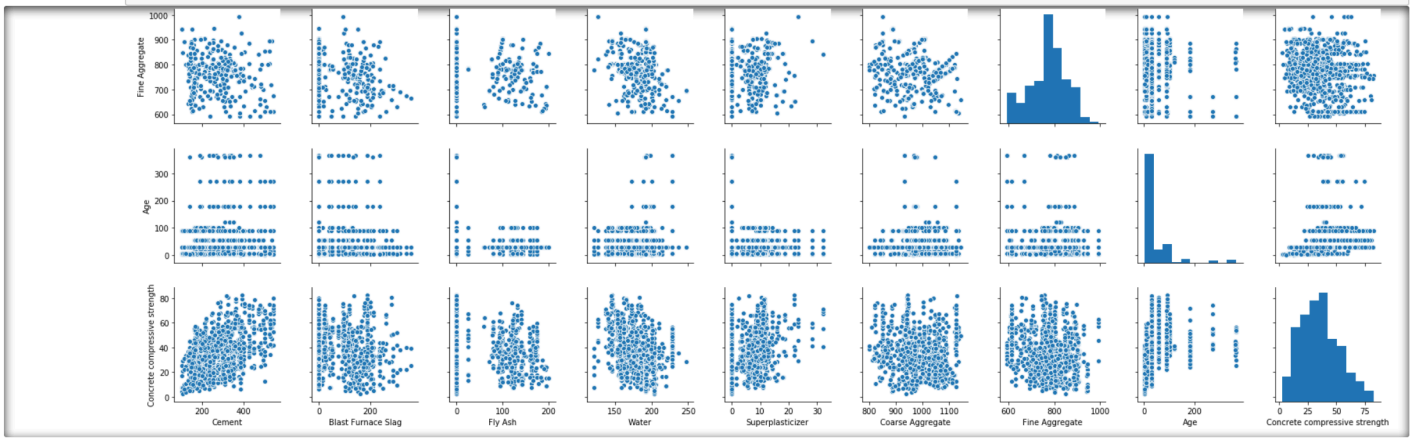
*** 方法2 : 算出相關係數並劃出熱力圖

In [32]:
```
import numpy as np
cm = np.corrcoef(df.values.T)
plt.figure(figsize=(10,10))
hm = sns.heatmap(cm,
                 cbar=True,
                 annot=True,
                 square=True,
                 fmt='.2f',
                 annot_kws={'size': 15},
                 yticklabels=cols,
                 xticklabels=cols)

plt.tight_layout()
plt.show()
```



In [33]:
```
import matplotlib.pyplot as plt
import seaborn as sns
sns.pairplot(df)
plt.tight_layout()
#plt.savefig('scatterplot.png', dpi=300)
plt.show()
```



3. 得出使用用8個特徵預測 Concrete compressive strength的線性迴歸模型

In [115]:
```
from sklearn.model_selection import train_test_split
#取得x跟y
X = df.iloc[:,:-1].values
y = df.iloc[:, -1].values
#training data 與 testing data的切割
X_train, X_test, y_train, y_test = train_test_split(
    X, y, random_state=8)
```

訓練模型

In [116]:
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
slr = LinearRegression()
slr.fit(X_train, y_train)
#print(slr.coef_)
for x in range(0,8):
    print(df.columns[x],",",'coef: %.8f' % slr.coef_[x])
```

```
Cement , coef: 0.12497617
Blast Furnace Slag , coef: 0.10756450
Fly Ash , coef: 0.09257633
Water , coef: -0.14994099
Superplasticizer , coef: 0.19312049
Coarse Aggregate , coef: 0.01664178
Fine Aggregate , coef: 0.02078749
Age , coef: 0.12085950
```

預測模型

結果：發現MSE比較高，而且train的R^2比test的R^2高

In [117]:
```
#預測模型
y_train_pred = slr.predict(X_train)
y_test_pred = slr.predict(X_test)
print('MSE train: %.3f, test: %.3f' % (
        mean_squared_error(y_train, y_train_pred),
        mean_squared_error(y_test, y_test_pred)))
print('R^2 train: %.3f, test: %.3f' % (
        r2_score(y_train, y_train_pred),
        r2_score(y_test, y_test_pred)))
```

```
MSE train: 103.685, test: 120.356
R^2 train: 0.633, test: 0.547
```

4. 觀察是否有迴歸係數與相關係數異異號的情況

比較各X與y

發現有異號 : Fly Ash, Coarse Aggregate, Fine Aggregate

In [118]:
```
#相關係數結果
df.corr()["Concrete compressive strength"]
```

Out[118]:
```
Cement                          0.497833
Blast Furnace Slag              0.134824
Fly Ash                        -0.105753
Water                          -0.289613
Superplasticizer                0.366102
Coarse Aggregate               -0.164928
Fine Aggregate                 -0.167249
Age                             0.328877
Concrete compressive strength   1.000000
Name: Concrete compressive strength, dtype: float64
```

In [119]:
```
#回歸結果
for x in range(0,8):
    print(df.columns[x],",','coef: %.8f' % slr.coef_[x])
```

```
Cement , coef: 0.12497617
Blast Furnace Slag , coef: 0.10756450
Fly Ash , coef: 0.09257633
Water , coef: -0.14994099
Superplasticizer , coef: 0.19312049
Coarse Aggregate , coef: 0.01664178
Fine Aggregate , coef: 0.02078749
Age , coef: 0.12085950
```

5. 進行行行資料預處理理（ex. 刪除部分特徵），以求得迴歸係數與相關係數均同號的線性迴歸模型

In [120]:
```
df.columns
```

Out[120]:
```
Index(['Cement', 'Blast Furnace Slag', 'Fly Ash', 'Water', 'Superplasticizer',
       'Coarse Aggregate', 'Fine Aggregate', 'Age',
       'Concrete compressive strength'],
      dtype='object')
```

發現有異號 : Fly Ash, Coarse Aggregate, Fine Aggregate

In [121]:
```
index=[0,1,3,4,7]
X1 = X[:,index ]
X1_train, X1_test, y_train, y_test = train_test_split(
    X1, y, random_state=8)
slr_new = LinearRegression()
slr_new.fit(X1_train, y_train)
y1_train_pred = slr_new.predict(X1_train)
y1_test_pred = slr_new.predict(X1_test)
```

新的回歸結果

In [122]:
```
for x in range(0,len(index)):
    print(index[x],":",df.columns[(index[x])],'coef: %.8f' % slr_new.coef_[x])
```

```
0 : Cement coef: 0.08442900
1 : Blast Furnace Slag coef: 0.06168285
3 : Water coef: -0.19841103
4 : Superplasticizer coef: 0.54630191
7 : Age coef: 0.11736351
```

原始回歸結果

In [123]:
```
for x in range(0,8):
    print(x,":",df.columns[x],",','coef: %.8f' % slr.coef_[x])
```

```
0 : Cement , coef: 0.12497617
1 : Blast Furnace Slag , coef: 0.10756450
2 : Fly Ash , coef: 0.09257633
3 : Water , coef: -0.14994099
4 : Superplasticizer , coef: 0.19312049
5 : Coarse Aggregate , coef: 0.01664178
6 : Fine Aggregate , coef: 0.02078749
7 : Age , coef: 0.12085950
```

新的回歸結果的MSE &R^2

In [124]:
```
print('MSE train: %.3f, test: %.3f' % (
        mean_squared_error(y_train, y1_train_pred),
        mean_squared_error(y_test, y1_test_pred)))
print('R^2 train: %.3f, test: %.3f' % (
        r2_score(y_train, y1_train_pred),
        r2_score(y_test, y1_test_pred)))
```

```
MSE train: 113.221, test: 126.465
R^2 train: 0.600, test: 0.524
```