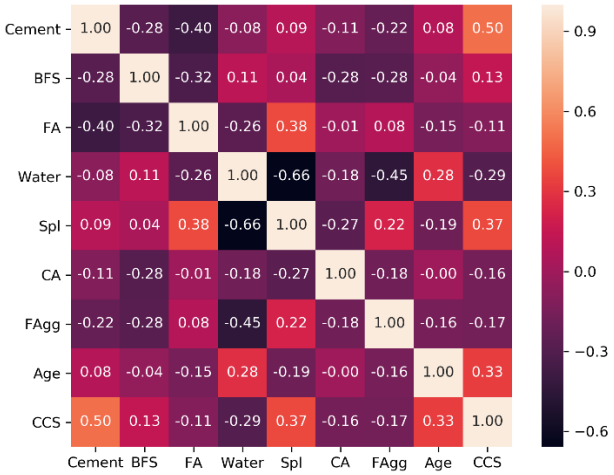
一. 從 UCI 下載 Concrete Compressive Strength Data Set

```
cols = ['Cement','BFS','FA','Water','Spl','CA','FAgg','Age','CCS']
data = pd.read_excel('Concrete_Data.xls')
data.columns = cols
```

1. 下載下來後，因 Columns name 過於長，故以上圖 List 順序更換 Columns name。

二. 請算出 9 個變數間的相關係數



1. 計算出相關係數後，並將其繪製成 heatmap 如上圖，CCS 代表著 Target。根據由上圖可見得此資料集大部分的資料彼此之間都無達到高度相關(>0.7)。唯獨變數 Cement 與 Concrete compressive strength(Target)之間的關係數來到了 0.5。
2. 值得慶幸的，在相關係數中看起來並未有某兩個獨立變相相關係數過高的問題(>0.7)，因此於此圖目前可以暫時排除共線性問題。當然實際上仍需觀察迴歸係數。
3. 除了正相關之外，於其中還可以發現某些較高的負相關例如 Spl(Superplasticizer)跟 Water 之間的相關係數就來到了-0.66。

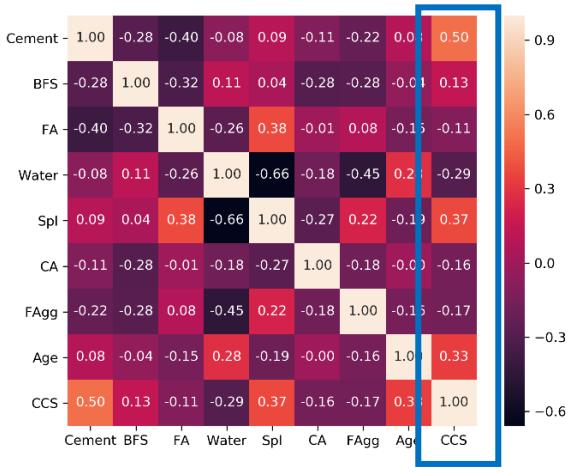
三. 得出使用用 8 個特徵預測 Concrete compressive strength 的線性迴歸模型

```
MSE train : 102.685918 | MSE test: 126.366075
R^2 train : 0.626947 | R^2 test: 0.567500
```

1. 經過初步簡單的線性回歸後結果如上圖，可見得在 MSE、R 平方兩種評估指標中，該模型的結果差強人意。

四. 觀察是否有迴歸係數與相關係數異號的情況

```
訓練完畢，得到迴歸係數:
 [13.1087892   9.37243504  6.06930049 -2.64527913  1.9713868   1.6393208
  1.86081386  7.27644206]
```



1. 從迴歸係數中我們須再與前面的相關係數一起看，可以見得其中第三(FA)、第六(CA)、第七個(FAgg)係數與相關係數不同號，可見得其存在共線性關係，因此我們必須對其做必要的處理。

五. 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

1. 將上述提到的三個變相簡單刪除，分別是 FA、CA、FAgg。

六. 結論

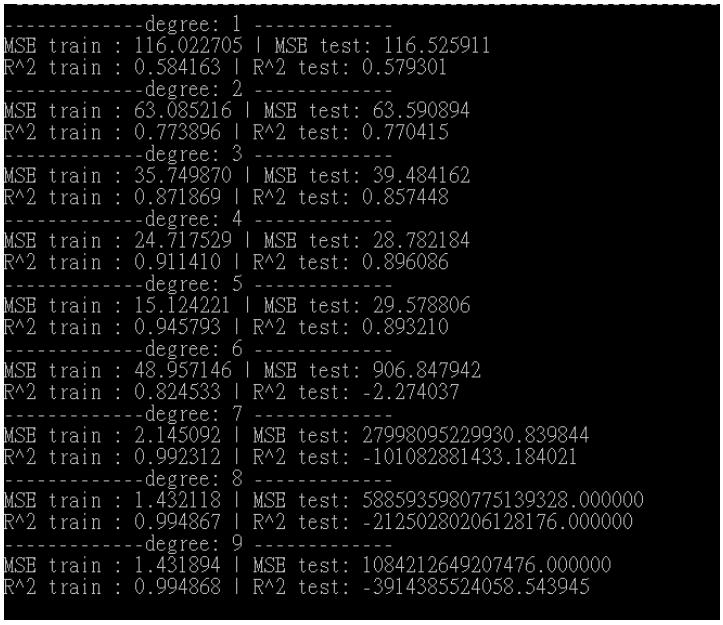
```
訓練完畢，得到迴歸係數:
 [13.1087892   9.37243504  6.06930049 -2.64527913  1.9713868   1.6393208
  1.86081386  7.27644206]
MSE train : 102.685918 | MSE test: 126.366075
R^2 train : 0.626947 | R^2 test: 0.567500
```

未處理

```
訓練完畢，得到迴歸係數:
 [ 8.58930649  5.31894647 -4.4720828   3.46017832  6.67257231]
MSE train : 114.732027 | MSE test: 121.310204
R^2 train : 0.582552 | R^2 test: 0.587531
```

已處理

1. 在經過簡單將三個存在問題的變相去除後，可以發現處理完後無論是在測試集的 MSE 與 R 平方皆有改進，但目測起來結果並不是很好，因此接下來將使用 Polynomial Features 的方式增進結果。



2. 為了測試參數對於模型的好壞，使用 degree=1~10 來測試，可見得大約在 degree=4 的時候模型表現最佳，從 5 之後就逐漸有 Overfitting 的問題產生。

3. 最終可以得到於測試集中 R 平方達到 0.88 的水準。