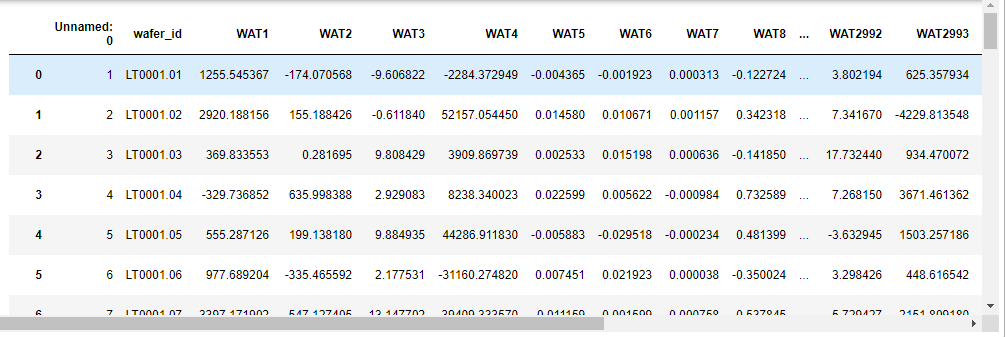
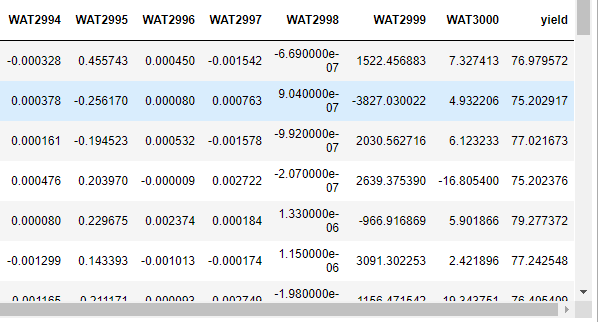
HW9 B063012054

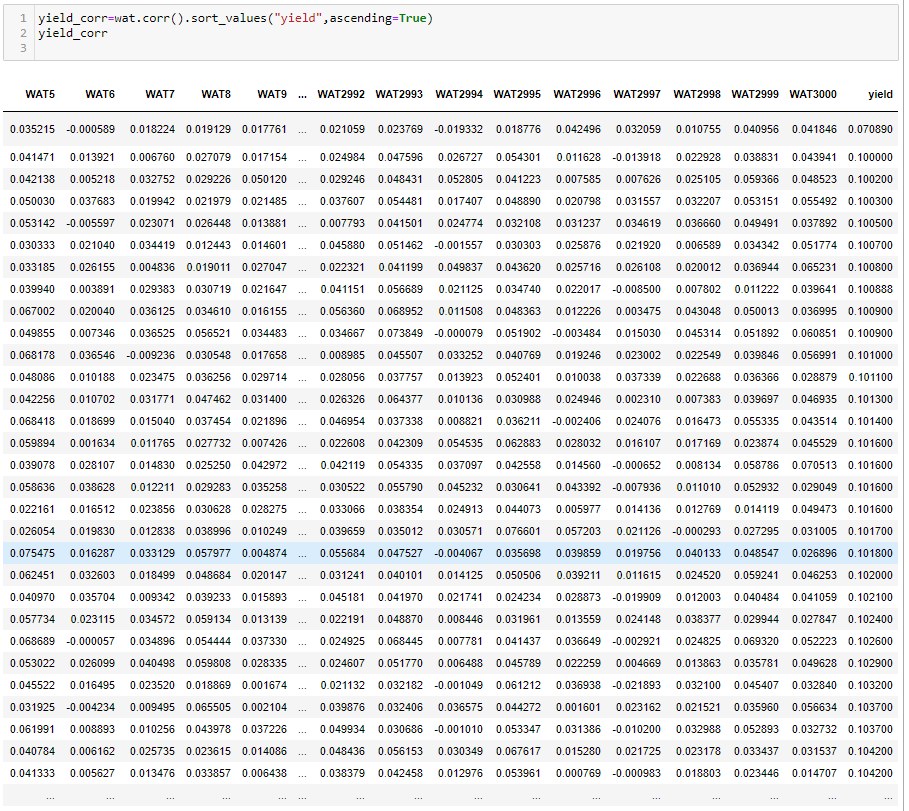
use pandas to analysis data and discuss the correlation between WAT and Yield

1.Merge the “yield data” and “WAT data” and print them

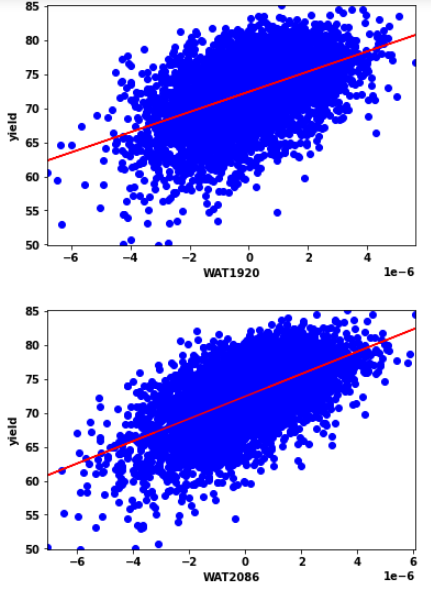
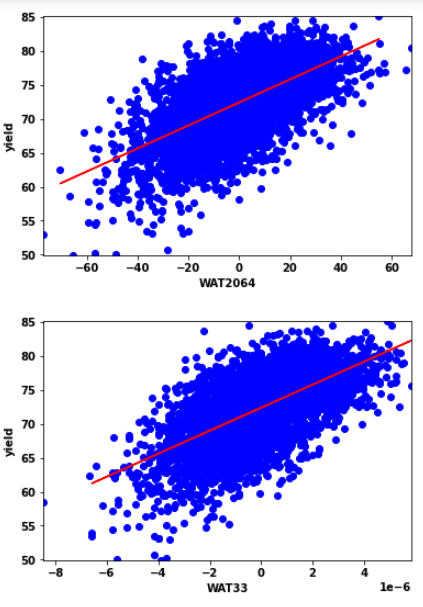


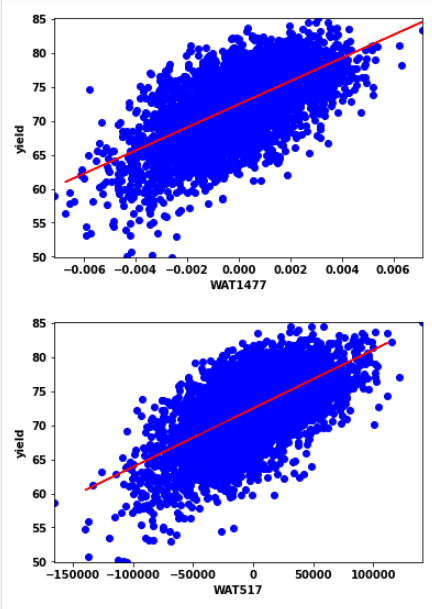
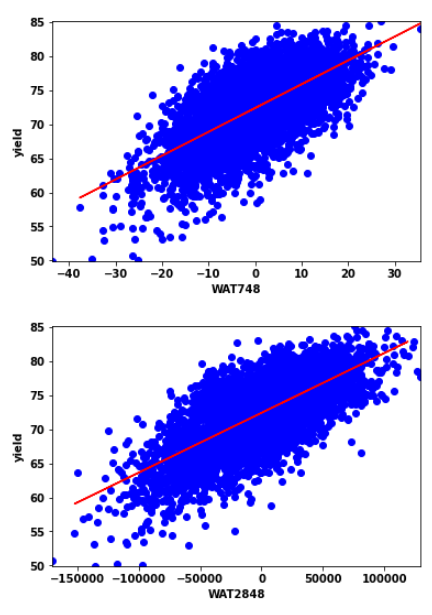


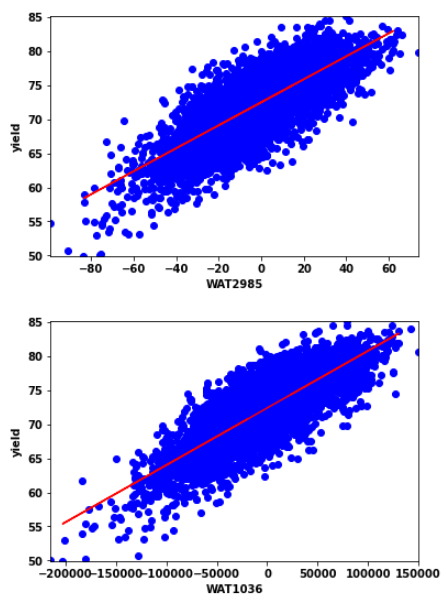
2.Calculate the pearson correlation coefficient value of each WAT to Yield, and sort them



3.According to their correlation, and draw the first 10 scatter plots with high correlation

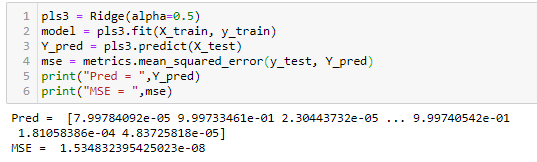
 

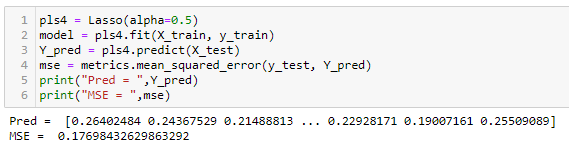
 



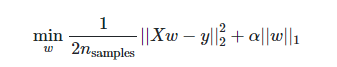
use pandas to predict dismission rate and know how feature data affects the target data

1.use other regression model to predict dismission rate



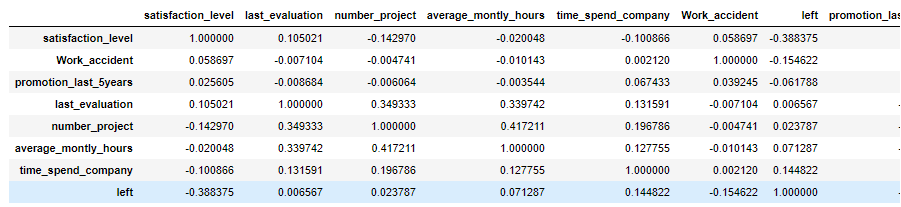


明顯看到Ridge預測出的值非常小， 根據Lasso定義的公式，與Ridge僅差別於前面的1/2n項，該項壓抑了異常多與異常少的數值，因此其結果值不會因為數量過多或過少出現極大或極小的情況。



2.and discuss which features affects accuracy

從pearson correlation看



最大值出現在time spend company，因此可以推測這個數值表現最接近離職率，由於我們是用回歸線做預測，進一步猜測由於表現相近，其預測的結果主要受time spend company影響。

3.discuss which model is better

我認為Lasso模型最好，因為不僅考慮了衰減量還考慮了資料量，對於不同尺度的資料能有更好的適應力