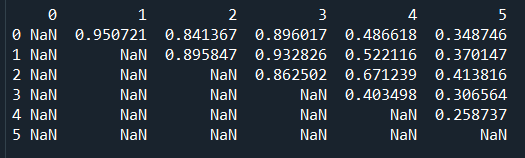
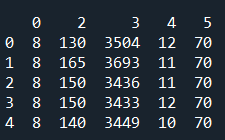
ML HW4 109070032廖品睿

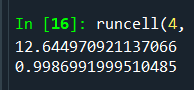
1.

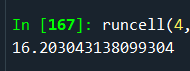
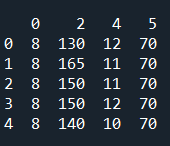
(a) **#High correlation filter**

This is the covariance matrix of each feature and we get rid of the highest correlated feature first, which is feature 1 (displacement)



 the loss is 12.64

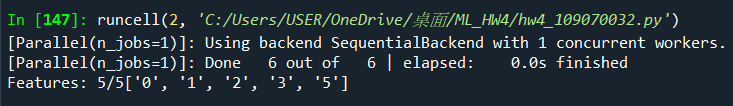
And the R-square is 

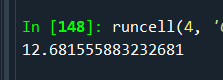
 once we get rid of feature 3

(weight), the loss increases to 16.2, therefore we choose feature cylinders, horsepower, weight, acceleration, model year to predict results.

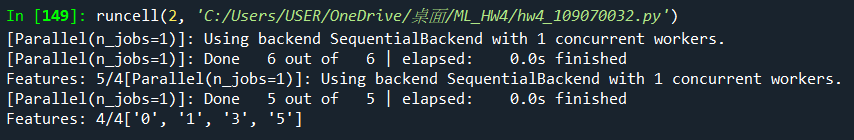
(b) **#Backward selection**

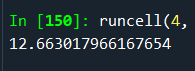
If we choose feature 0,1,2,3,5



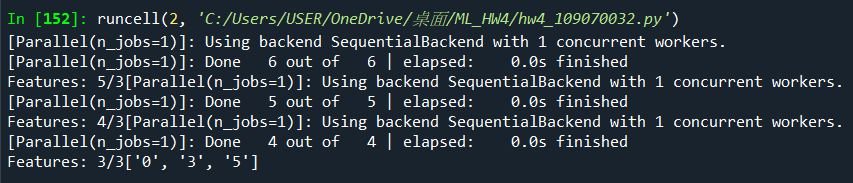
We get this much loss 

If we choose feature 0,1,3,5



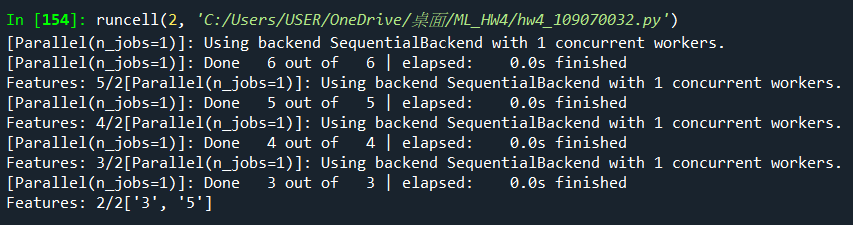
We get this much loss 

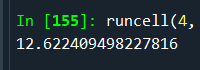
If we choose feature 0,3,5



We get this much loss 

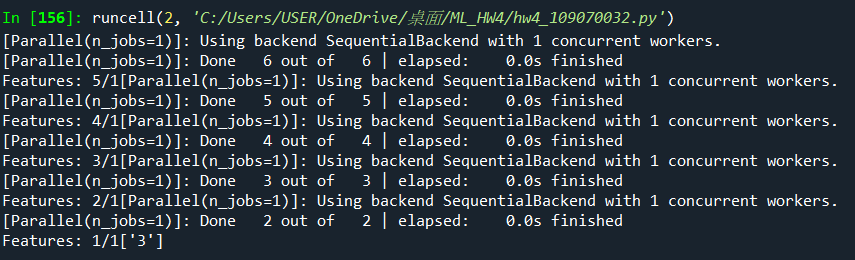
If we choose feature 3,5

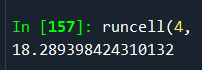


We get this much loss 

And we get R-square close to 

If we choose feature 3

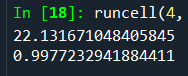


We get this much loss 

The loss gets bigger, therefore we choose the feature 3 (weight) and feature 5 (model year) to build the model.

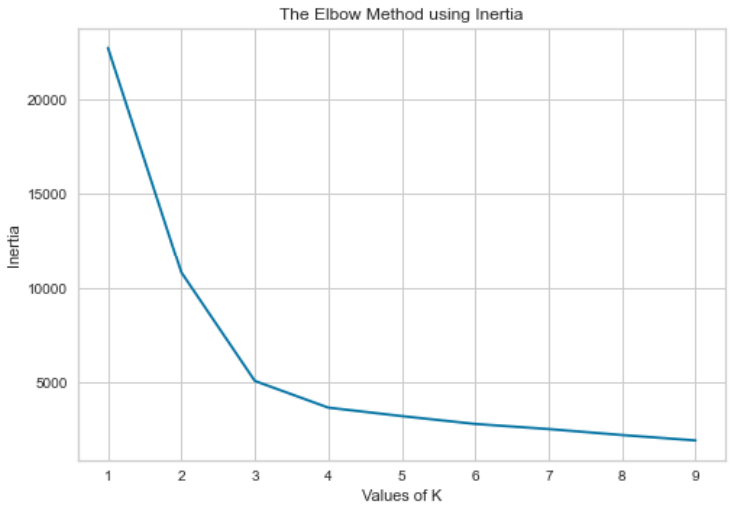
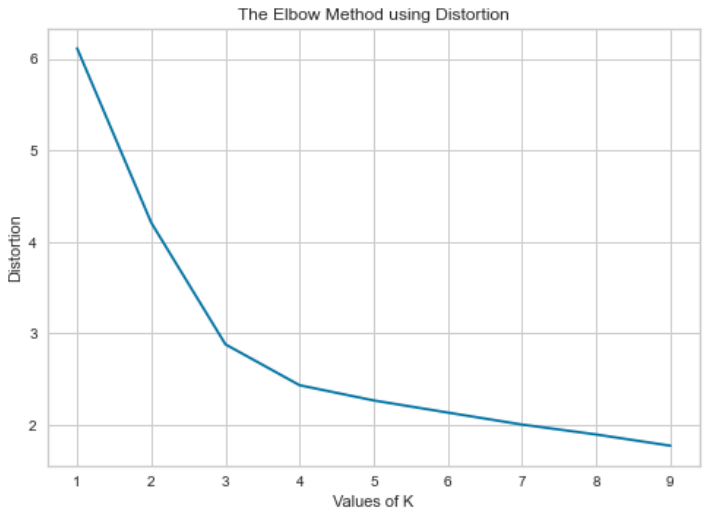
(c)**#PCA**

Using PCA method using only feature 1, we get 24.39 loss, and R-square equal to 0.9977, which is a bit lower than the first two methods because we only choose the features that > 95% variance. Apparently, keeping only one feature for prediction is somehow a bit inaccurate.



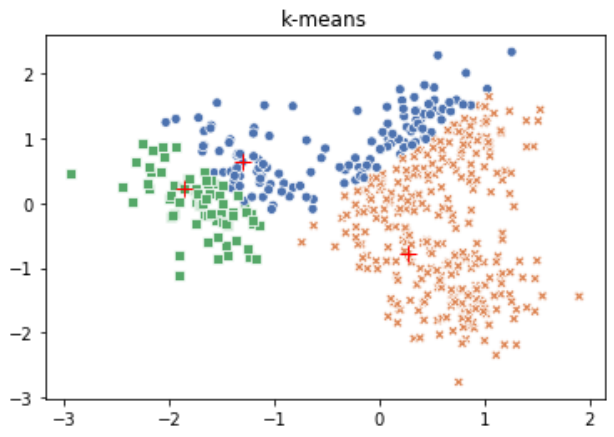
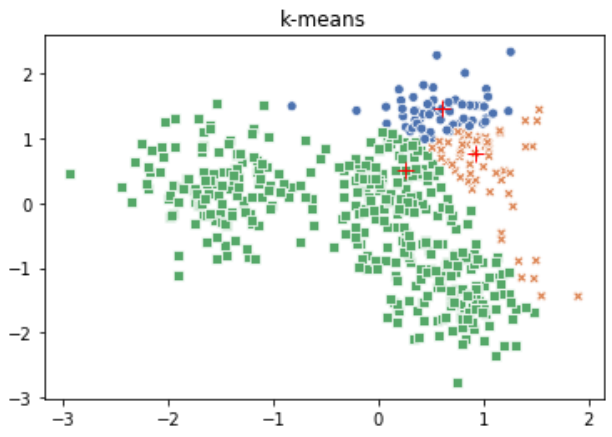
2.

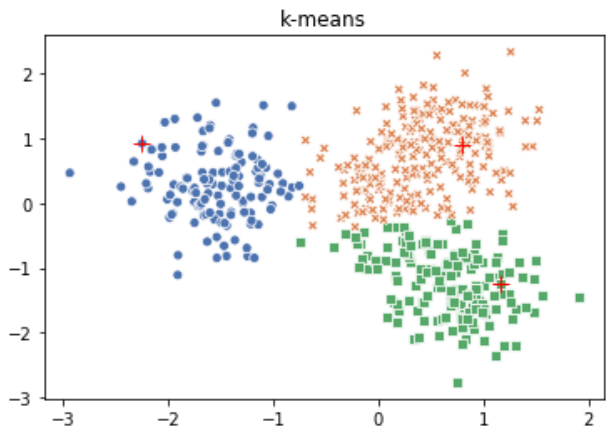
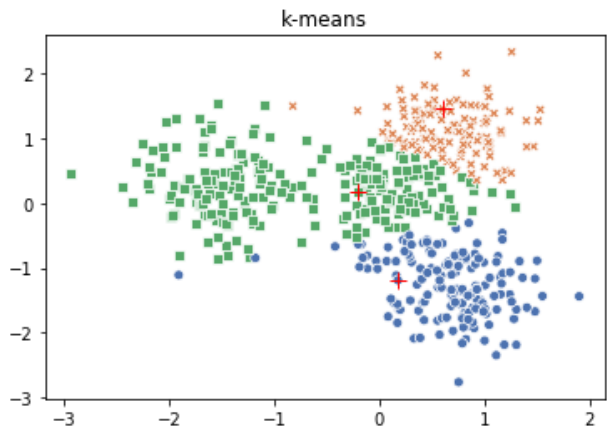
After fitting the dataset into the code below we can find the elbow point under the relatively less distortion score, and because distortion score and inertia score using elbow method is too large in n=2 and fit time is too long in n=4, therefore we choose the optimal n=3 to be the elbow point.



Using handcrafted k-means model to cluster data will have multiple kinds of outcomes because we randomize the center data to plot , therefore the size and the positions of clusters are relatively random, also we assume that we can converge the center in 300 iterations, unlike python k-means model, we get a relatively imprecise result because this method is unsupervised learning.

Note: the red cross represents the center of data, we approximately get 4 kinds of data like the 4 plots below.





The plot below is created by sklearn model in python, which is only similar to the 4th plot above, and this model has the most accuracy and least distortion.

