ECS171 HW1 Report

Jiahui Li : 915544392

Instruction to Run the Code

I split the code file for one file per question (problem1.py, problem2.py,etc) and put some shared code, eg.read the data, in one file (basicfunction.py).

The following report shows the results, plots and tables for each problem. Since problem3 is to write a solver, it has nothing to report except the code, this report doesn't have the answer for problem3.

Problem 1

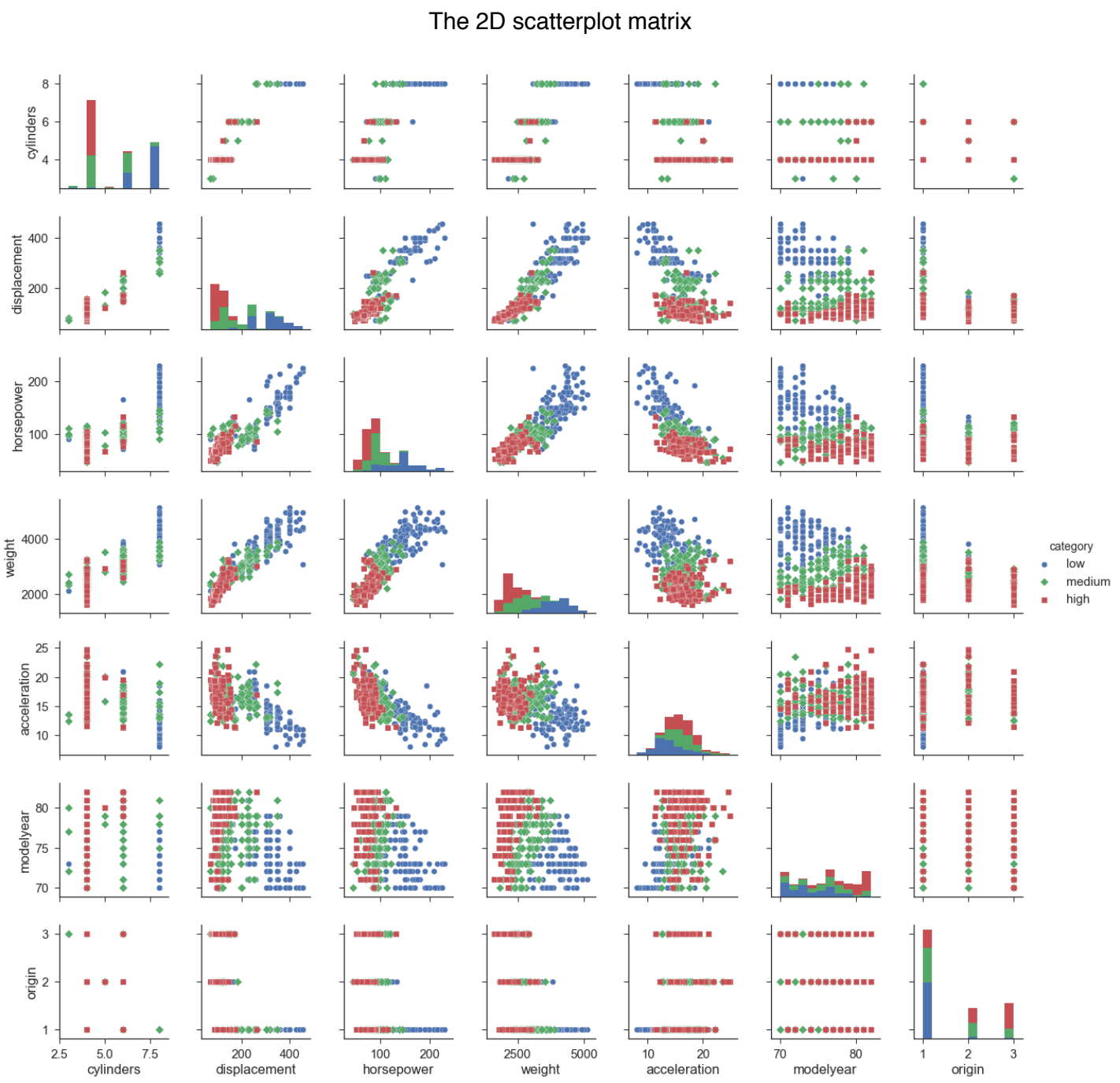
The thresholds are 18.73 and 26.93.

low : mpg <= 18.73, sample size = 131.

medium : 18.73 < mpg <= 26.93, sample size = 130.

high : mpg > 26.93, sample size = 131.

problem 2



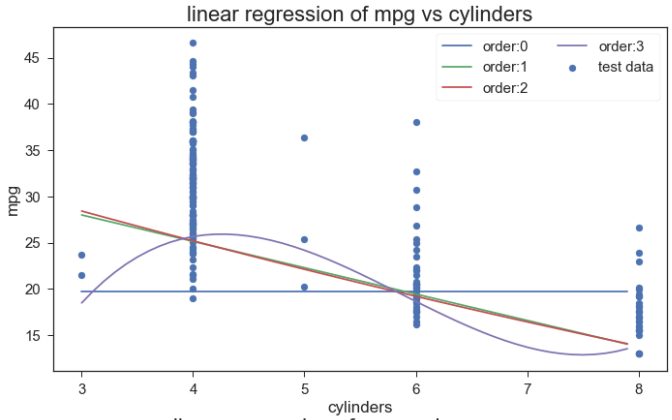
From the plot, we can find that the most informative pair-wise feature combination is horsepower and weight since it is relatively easy to distinguish the three mpg category by checking if horsepower and weight are over or below some thresholds.

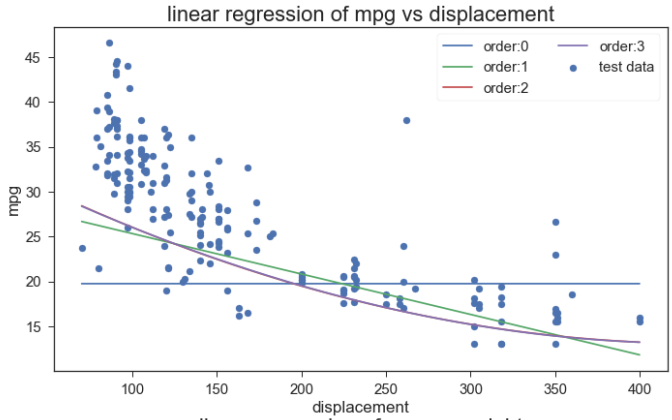
problem 4

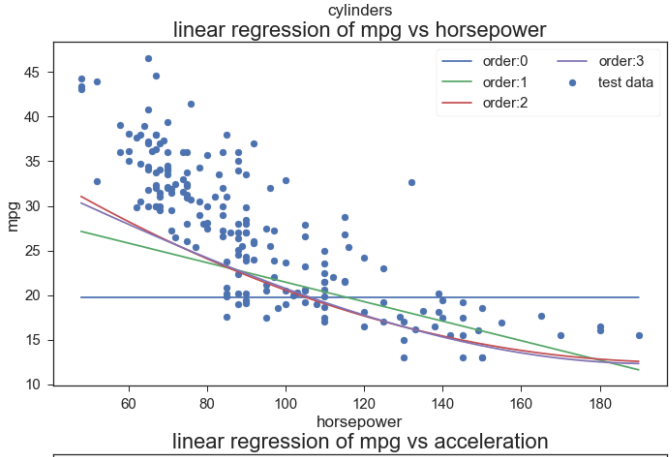
The mse of the training and the testing with linear regression on each single variable are showed in the table.

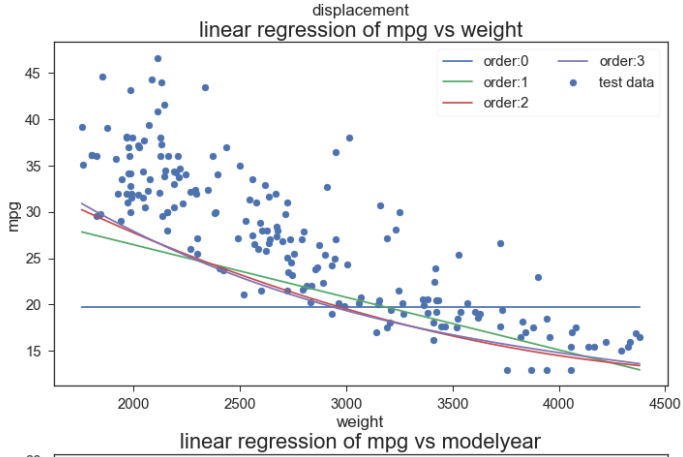
The name of each column means this: eg. "train_0th" means the mse of 0th order on training set.

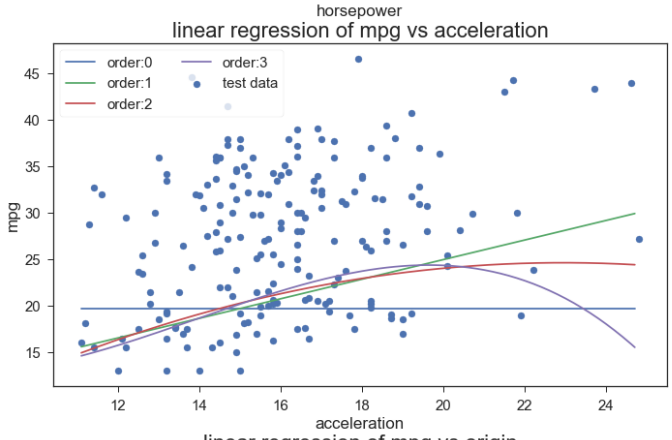
	train_0th	test_0th	train_1st	test_1st	train_2nd	test_2nd	train_3rd	test_3rd
cylinders	34.0549	116.623308	8.595830	55.707917	8.578095	55.469676	7.434722	51.851011
displacement	34.0549	116.623308	7.479179	52.419933	6.000853	48.676818	6.000330	48.675018
horsepower	34.0549	116.623308	11.511310	56.582734	8.906969	46.934573	8.847085	47.458707
weight	34.0549	116.623308	6.266344	50.477596	4.870518	48.335228	4.788272	48.673278
acceleration	34.0549	116.623308	24.838396	96.831856	24.202522	95.563763	23.636051	99.362201
modelyear	34.0549	116.623308	32.771386	66.237911	32.475902	52.314284	32.368586	641.357355
origin	34.0549	116.623308	18.693183	83.108650	17.008752	85.573273	17.008752	85.573273

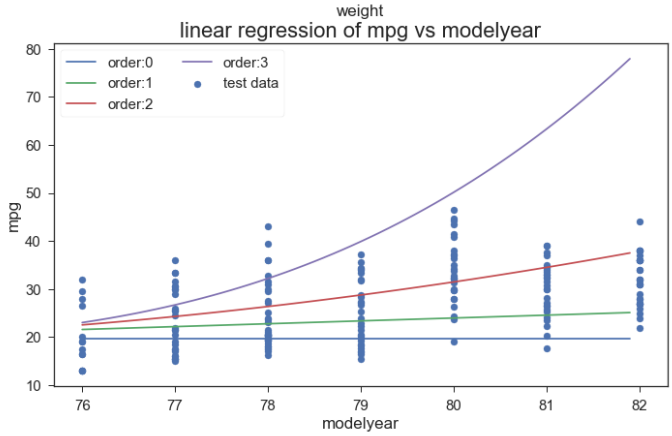


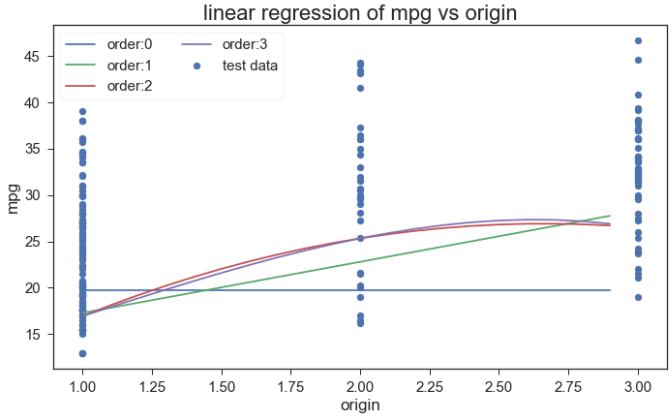












The above plot is regression line for the testing set. From the mse table and the plot, we can see the best polynomial order in the test set is 2nd order since except origin, 2nd order of all the variables is better then 0th and 1st. And, for some variables, the 3rd order become worse, especially modelyear.The most informative feature for mpg consumption is horsepower since when it is 2nd order and 3rd order, it has the least mse, and although for 1st order, it has second least mse, but just a little bit larger than the least.

problem 5

The mse of the training and the testing with linear regression on all 7 variables are showed in the table.

Each row show different order: eg. "0th" means 0th order polynomial.

	train	test
0th	34.054900	116.623308
1st	5.099396	30.521729
2nd	3.177745	42.319377

problem 6

The precision for training and the testing with logistic regression are showed in the table.

I calculate the precision for different category separately. For each, precision is TP/(TP + FP).

	train	test
low	0.905660	0.731707
medium	0.813333	0.433962
high	0.842105	1.000000

problem 7

For better predicting the new data, I train the models with whole dataset(392 data).

From the second-order, multi-variate polynomial regression, we can predict the MPG value is 20.71(round to 2 decimal place), so we expect the MPG rating is medium since the medium threshold is (18.73,26.93] .

From logistic regression, we can predict the MPG category it belongs is **low**.

problem 8

Since it's a car driven by the horse, so I don't think it should use mobile oil to move. There is no a concept like MPG for it, that means no model can be used to predict the MPG of it.