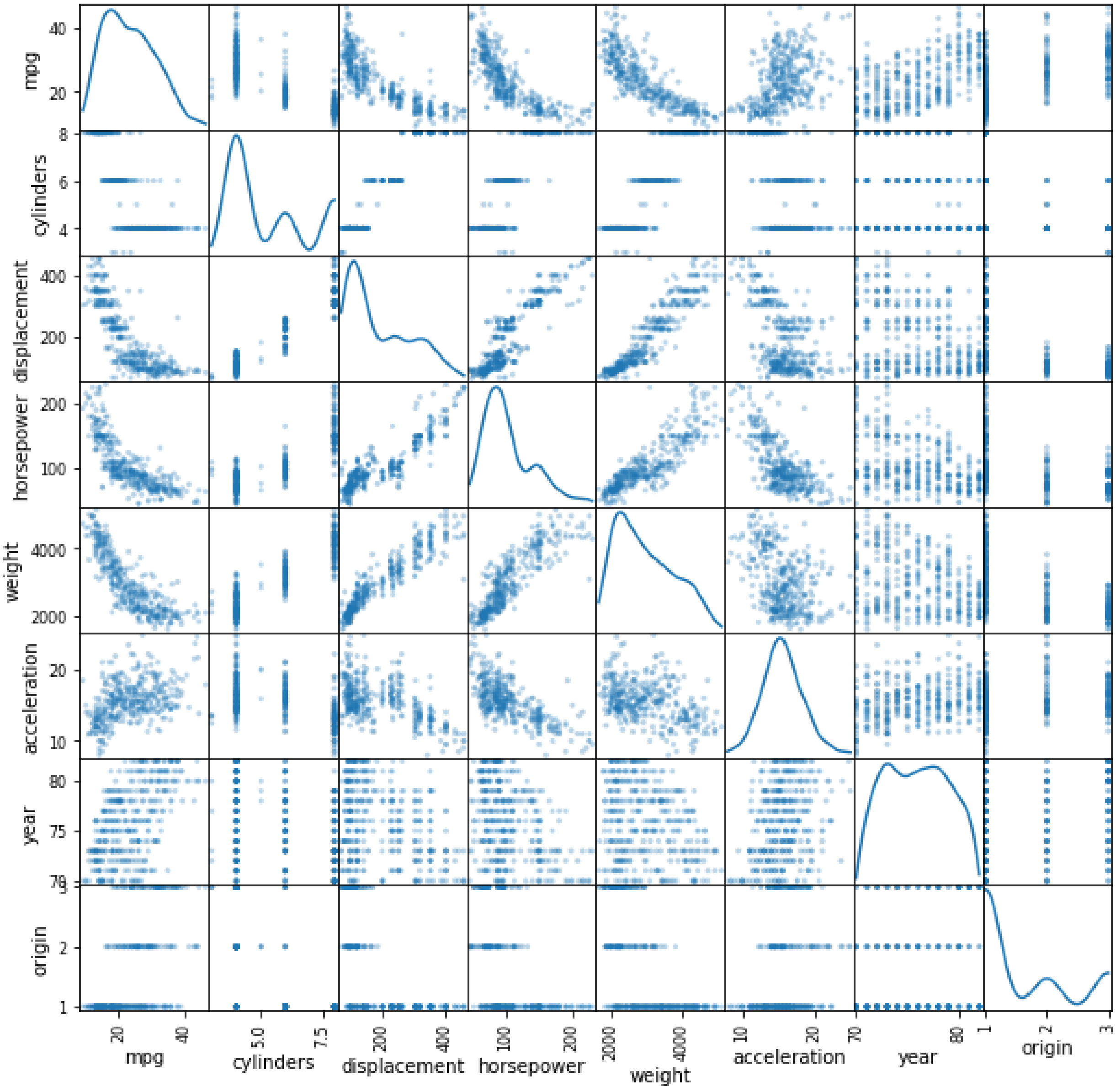
Problem Set #4

MACS 30100, Dr. Evans

1. Multiple linear regression

a. See the codes

b.



c.

d.

i. displacement, year, weight, and origin

ii. cylinders, horsepower, acceleration

iii. Miles per gallon of a car improves by 7.5 % as the car becomes an year older

e)

i. Based on the scatterplot matrix variables that seem to have a non-linear relationship with mpg are: displacement, horsepower, weight, and acceleration

ii. Adjusted R-squared for the new regression is 0.866 which is better than the older regression adjusted R squared of 0.818

iii. The coefficient of displacement term has become negative from positive and is no longer statistically significant. The coefficient of its square term is very small and is also not significant.

iv. It is still not significant

(f) Predicted value 7.506582e+06.

2. Classification problem: KKN by hand and in Python

a)

Obs.	**x1**	**x2**	**x3**	**Y**	distance	**Ecul. Dist.**	**Rank**
5	-1	0	1	Green	2	1.414213562	1
6	1	1	1	Red	3	1.732050808	2
2	2	0	0	Red	4	2	3
4	0	1	2	Green	5	2.236067977	4
1	0	3	0	Red	9	3	5
3	0	1	3	Red	10	3.16227766	6

b) Green, because for K= 1, we will take the shortest distance

c) Red, for K=3 the p(red) = 2/3, while p(green) = 1/3

d) If the decision boundary is highly non-linear, then the K value should be small.

e) Green

3. Multivariable logistic (logit) regression

a) weight and year are statistically significant at 5 % level of significance

b) see the codes

c)

B1	-1.8304
B2	0.0281
B3	0.0153
B4	-0.0075
B5	0.1094
B6	0.6432
B7	0.2444
B0	-26.7996

d) Predicts lower mpg better 86 times out of 98 (88% precision).