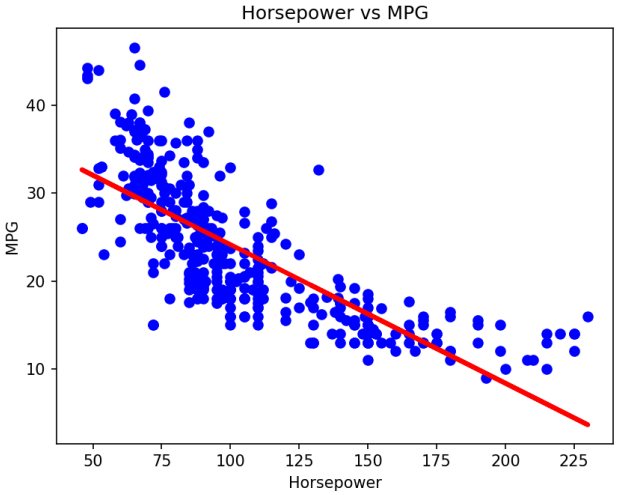
Conceptual and Theoretical Questions

1. The Null Hypothesis would be that newspapers, tv, and radio advertising have no effect on sales of a company. Based on the p-values we can say that tv and radio have a significant impact on the number of sales where we can accept the Null hypothesis for the newspaper due to the large p-value.
2. KNN classification uses the local majority of the surrounding data points to predict the value of the output whereas KNN regression attempts to predict the value based on the mean of the local data points.
3.
    a. The equation for a males salary is y = 50 + 20 gpa + 0.07 iq + 0.01 (gpa * iq) and for a female is y = 85 + 10 gpa + 0.07 iq + 0.01 (gpa * iq) and by comparing these 2 lines with a fixed gpa and iq a male will earn more as long as his gpa is high enough meaning statement iii is correct.
    b. Using the equation from above
    Y = 85 + 10 * 4 + 0.07 * 110 + 0.01 * 440 = 137 which results in $137,000
    c. False: The interaction term is small due to the size of the values that it represents to not over effect the data
4.
    a. I would expect the linear regression to fit the data better because as mentioned in the question the true fit is linear and the cubic regression might overfit the data causing a higher RSS due to the overfitting model
    b. Once again due to the nature of the data we can predict that the cubic regression will overfit the data and create more error in the training.
    c. There is not enough information to make a confident statement about which regression model will have a lower RSS. In general the cubic model will be much better at fitting non-linear data as there is much more flexibility but without knowing the data it is difficult to say with much confidence.
    d. Due to the vagueness of the question we can be even less sure about which model will have less error with the test data. Without know how far the data is away from linear there is no way to try and formulate a prediction of which model will be more accurate without actually completing it.

Applied Questions with Owen Rago

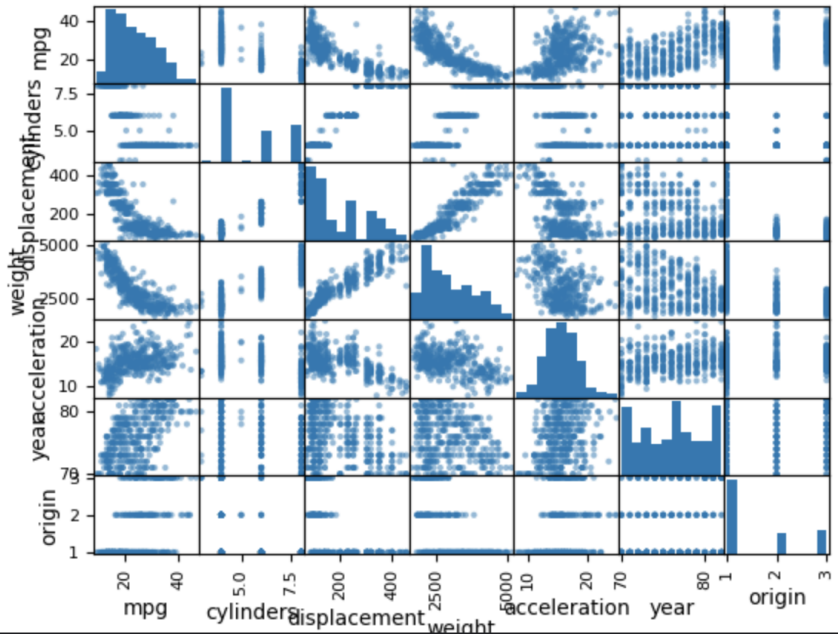
1.
    a.
        i. There is a relationship, the coefficient of determination is 0.606
        ii. Moderately strong, had a value of 0.606
        iii. negative
        iv. 24.94 mpg

b.

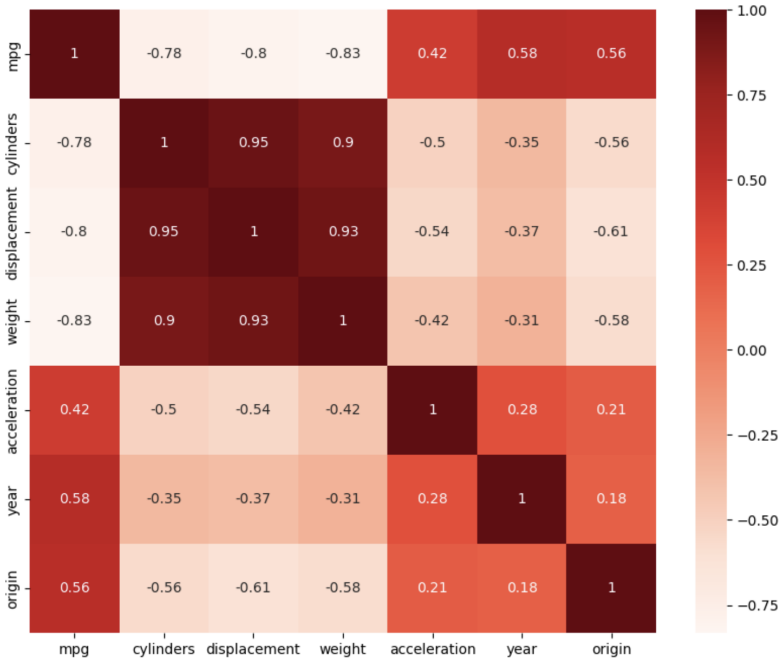

Horsepower vs MPG
MPG
Horsepower
40
30
20
10
50
75
100
125
150
175
200
225

2.

a.



b.



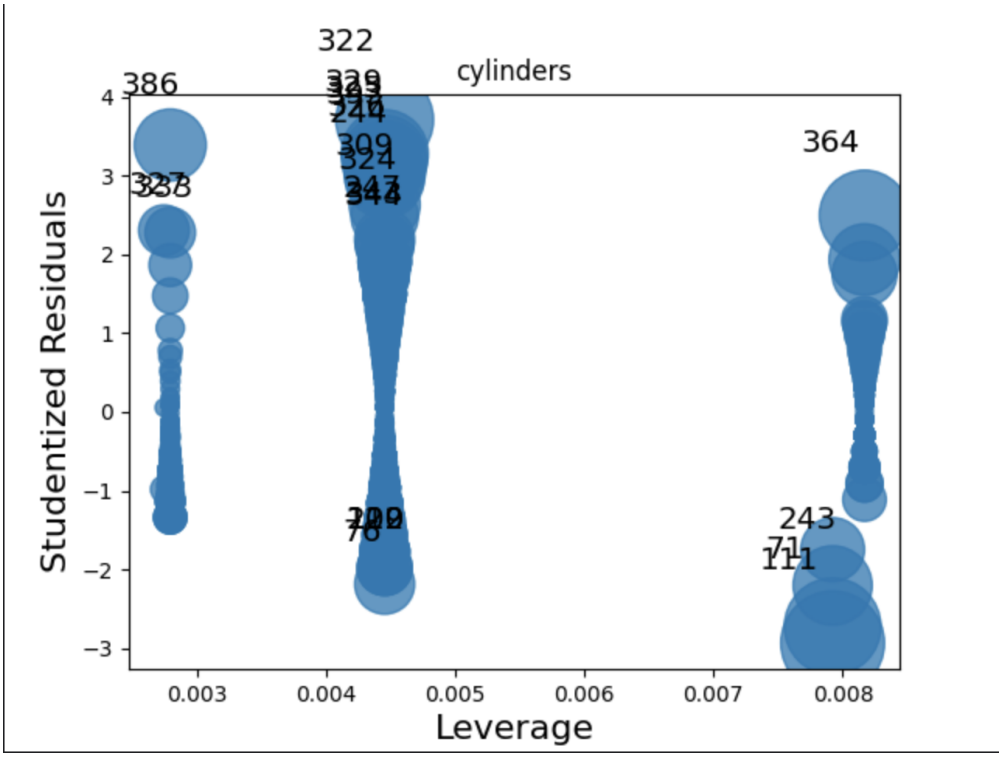
c.

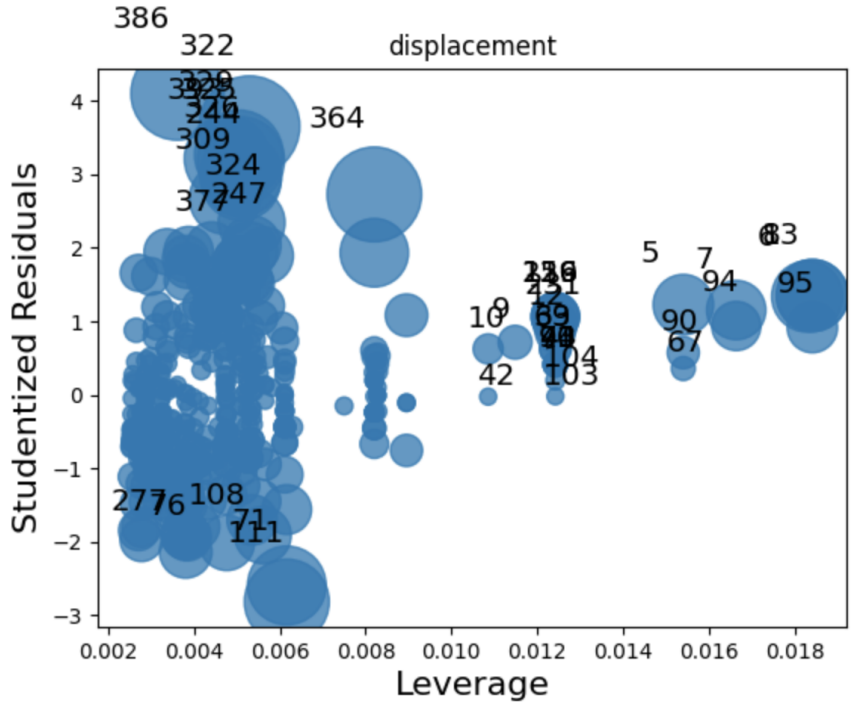
i. There is a moderately strong relationship between mpg and cylinders, displacement, horsepower and weight and a weaker one between mpg and acceleration, year and origin.

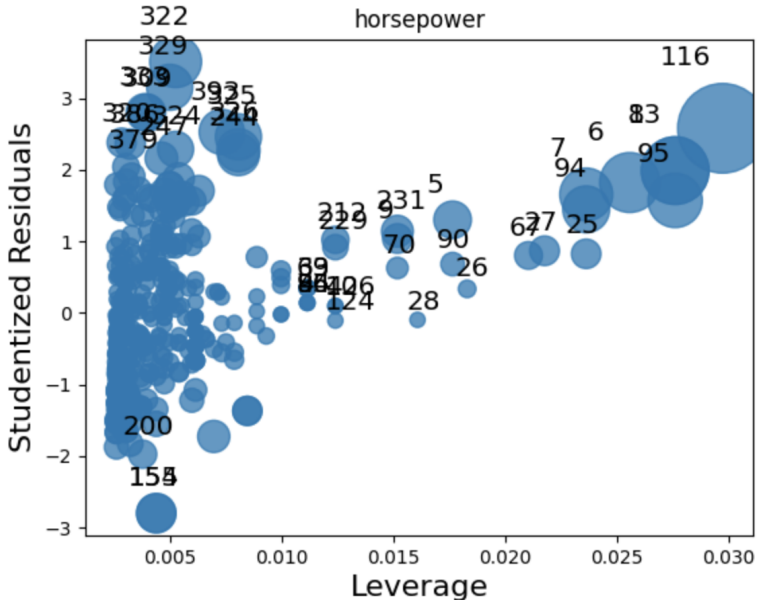
ii. cylinders, displacement, horsepower and weight seem to have a statistically significant relationship

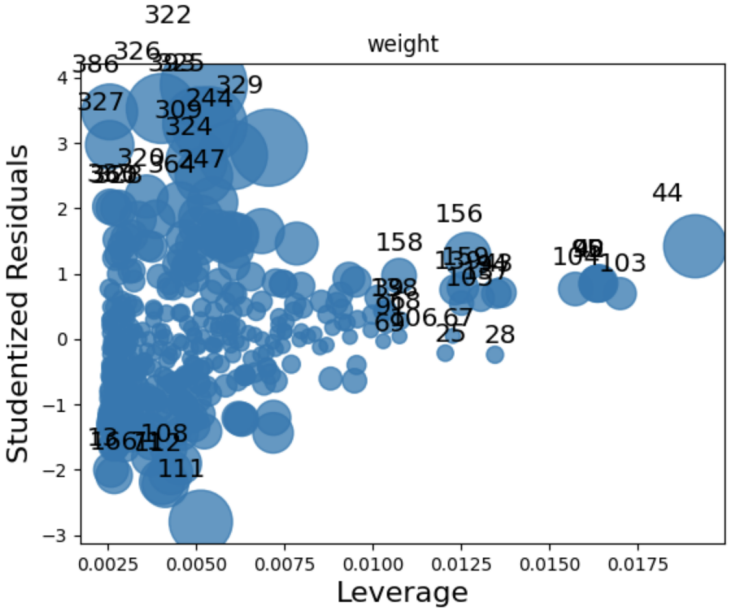
iii. The coefficient of year suggests that there is a small positive correlation.

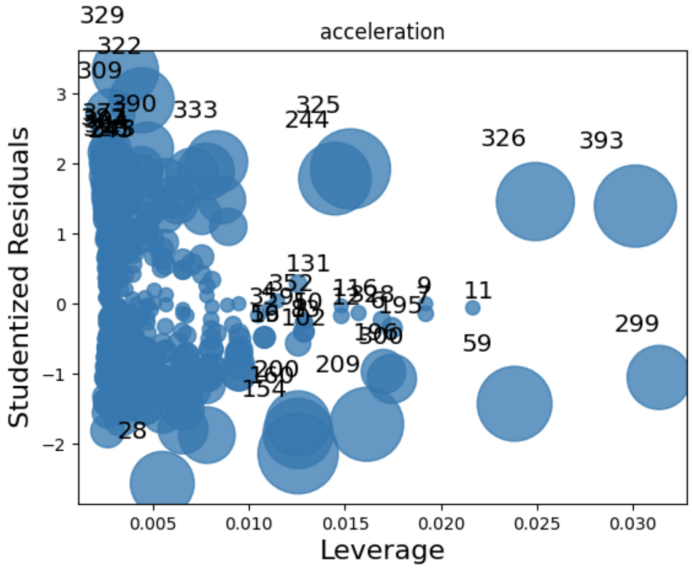
d.

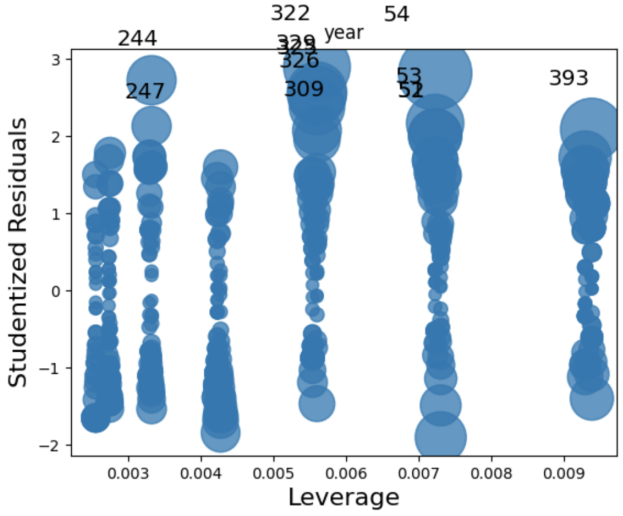


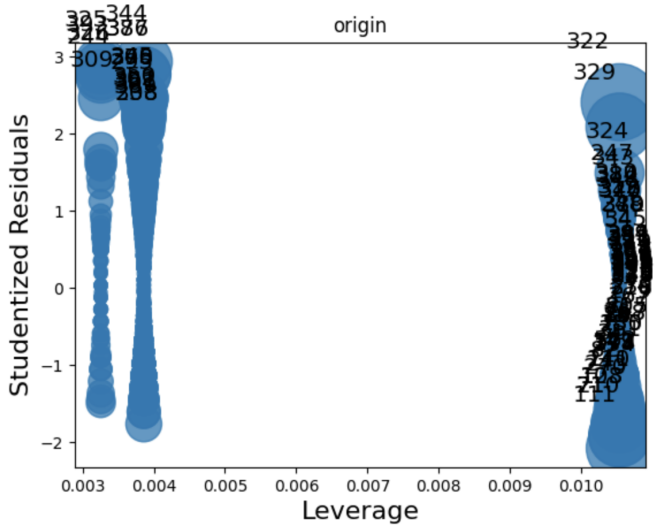



horsepower
Studentized Residuals
Leverage
322
329
116
200
0.005
0.010
0.015
0.020
0.025
0.030


weight
Studentized Residuals
Leverage
322
326
329
44
156
158
111
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175


acceleration
Studentized Residuals
Leverage
329
322
309
390
333
325
244
326
393
131
11
299
59
209
154
28


year
Studentized Residuals
Leverage
322
54
244
326
309
247
53
52
393


origin
Studentized Residuals
Leverage
322
329
324

e. Yes, a few do, most notably displacement and year as well as weight and year
f. Yes, some do, acceleration and horsepower as well as weight and horsepower seem statistically significant but are weaker than using all predictors with no interaction terms.
g. X^2 generally made it worst while log and square root helped a lot of the predictors