1. The null hypothesis that the given advertisement method does not affect sales.

Based on the P values, we can see that this is untrue for intercept, TV, and radio because they all have a very low p value, meaning they aff>

We can see that this is likely true for newspaper because it has a high p value, meaning it does not likely affect sales.

2. KNN regression is for quantitative, continuous (takes an average of the neighbors value)

KNN classification is for discreet classification problems (assigns the classifier with the highest representation in the pool of neighbors)

3. Assumed equation:

$$Salary = 50 + 20GPA + 0.07IQ + 35\,Gender + 0.01GPA * IQ - 10GPA * Gender$$

   a. Part iii is correct. This is because when substituting the known values for the ‘gender’ attribute, we get that males average:
   $50 + 20GPA + 0.07IQ + 0.01GPA * IQ$, while females average:
   $85 + 10GPA + 0.07IQ + 0.01GPA \times IQ$. We can combine this to say that males make more if:
   $50 + 20GPA + 0.07IQ + 0.01GPA * IQ > 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ$
   which simplifies to say that males make more if $GPA > 3.5$
   b. $Salary = 50 + 20GPA + 0.07IQ + 35\,Gender + 0.01GPA * IQ - 10GPA * Gender$
   $Salary = 50 + 20(4) + 0.07(110) + 35\,(1) + 0.01(4) * (110) - 10(4) * (1)$
   $Salary = 137.1$
   c. False. The coefficient by itself does not give enough information to make this claim. The way to determine this would be to use the p test as alluded to in problem 1.

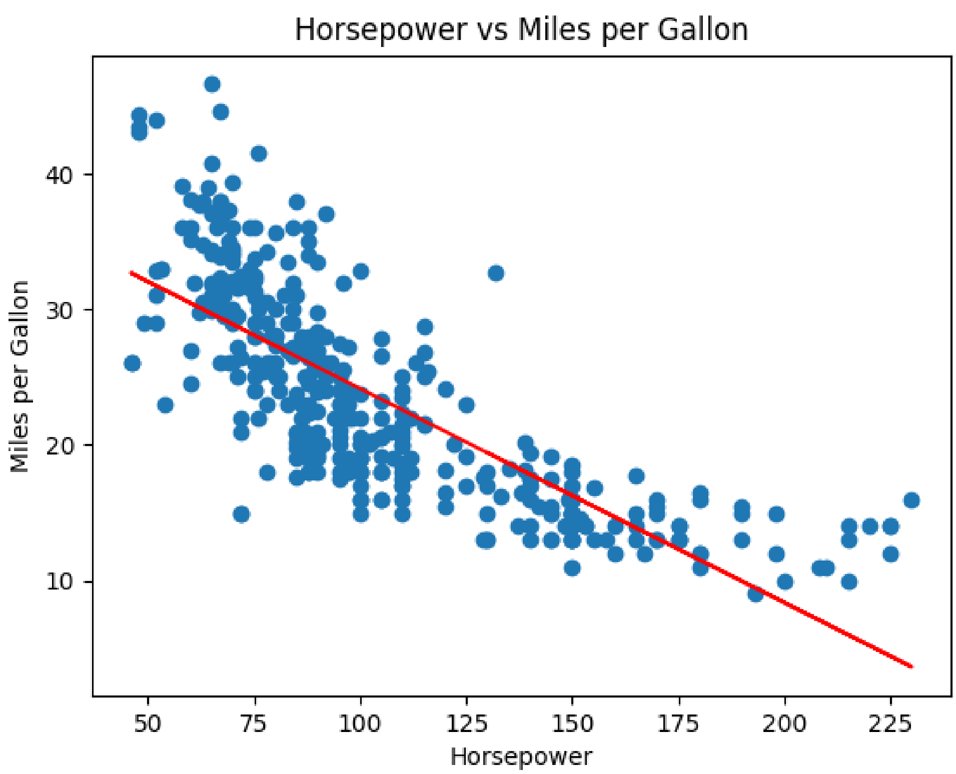
4.

   a. We would expect the TRAINING rss to be less for the cubic model on average. This is because the higher complexity would be able to overfit to the data, creating variance, but lowering the TRAINING error. Despite the fact that the original data is generated by a linear relationship, the cubic model would likely be able to match the noise a bit better than the linear model. At worst, the cubic model would find the β_2 and β_3 to be 0, making it effectively just an inefficient linear model. Therefore, the cubic model could have a lower RSS, but should never have a higher one, so we can expect it to be lower. Note that this is only true of the TRAINING rss, and test RSS should be expected to be higher in the cubic.
   b. As mentioned above, since the data is generated by a linear function, the cubic model would create bias, and overfit to the training data, causing a higher test rss. We can expect a lower rss on the linear model as it is the same order as the generator function.
   c. As above, with training rss, a higher order model will always have a lower training rss because it can fit the noise in the data better. More complexity = lower training rss.
   d. There is not enough information to tell. If the data is almost linear, then we can expect a linear function to perform better in general, because the cubic will

overfit. However, if the generator function is, for example, on the order of 3 or greater, we can expect a lower test rss from the cubic model.

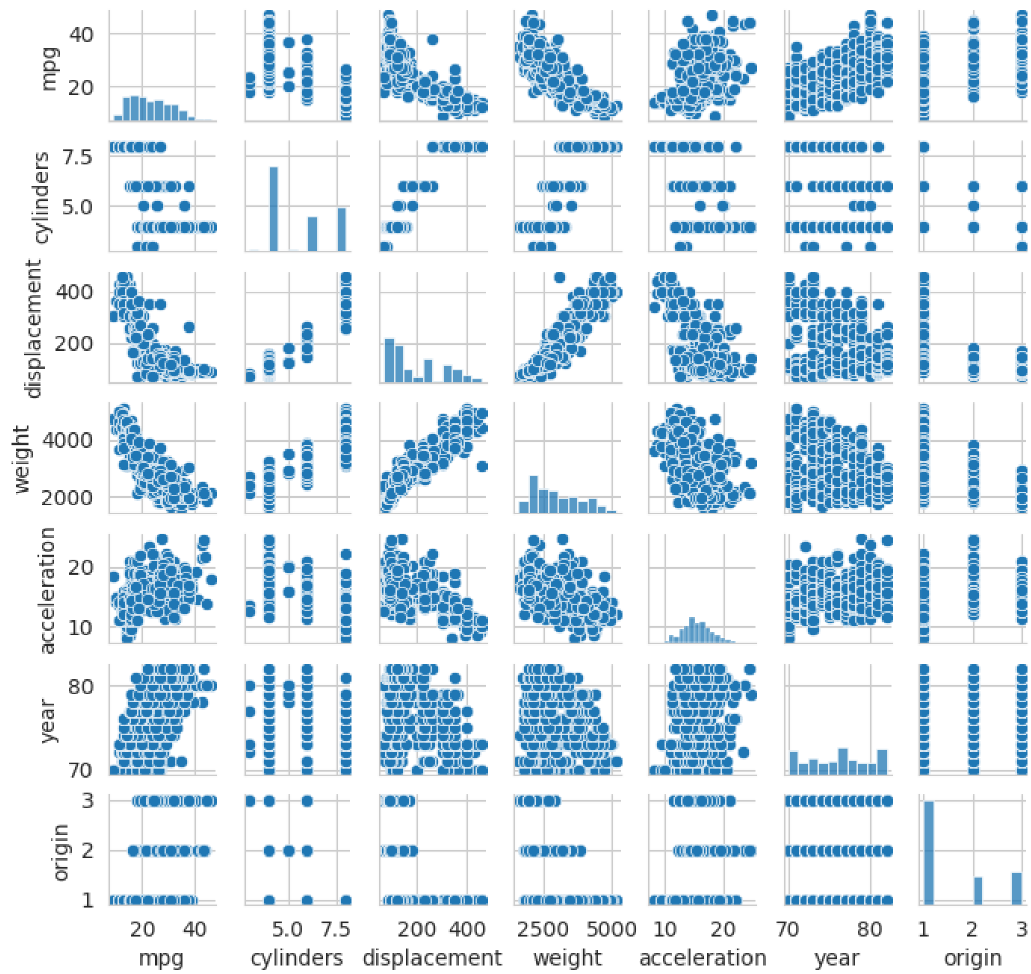
5.
6. The formula for simple linear regression is $\widehat{Y} = \widehat{\beta 0} + \widehat{\beta} 1x$ (can't get subscript to work correctly). If we substitute $\widehat{\beta 0}$ for the formula in the description, we get $\widehat{Y} = \bar{Y} - \widehat{\beta}1\bar{x} + \widehat{\beta}1x$. Because this is the equation of a line, plugging in a value for x will give us the corresponding y value. If we plug in $\bar{x}$ for x : $\widehat{Y} = \bar{Y} - \widehat{\beta}1\bar{x} + \widehat{\beta}1\bar{x} = \bar{Y}$. Therefore, the point $(\bar{x}, \bar{y})$.

Applied Questions

1.
    a.
        i. Yes because the R^2 value is 0.60594825788943
        ii. Somewhat strong, because about 60% of the variation in mpg is predicted by horsepower
        iii. Negative, slope of -0.15784473
        iv. 24.94061135
        v. 

2.

a. Scatterplot Matrix:



b. Correlation Matrix:

	mpg	cylinders	displacement	weight	acceleration	year	origin
mpg	1	-0.7776175081	-0.8051269467	-0.8322442148	0.4233285369	0.5805409661	0.5652087567
cylinders	-0.7776175081	1	0.9508233008	0.8975273403	-0.5046833793	-0.3456474403	-0.5689315895
displacement	-0.8051269467	0.9508233008	1	0.9329944041	-0.5438004967	-0.3698552067	-0.6145351146
weight	-0.8322442148	0.8975273403	0.9329944041	1	-0.416839202	-0.3091198808	-0.5850053547
acceleration	0.4233285369	-0.5046833793	-0.5438004967	-0.416839202	1	0.2903161133	0.212745808
year	0.5805409661	-0.3456474403	-0.3698552067	-0.3091198808	0.2903161133	1	0.1815277184
origin	0.5652087567	-0.5689315895	-0.6145351146	-0.5850053547	0.212745808	0.1815277184	1

c.

Predicting mpg using cylinders
Intercept: 42.91550535343909
Coefficient: [-3.55807837]
R2 Score: 0.6046889889441245

Predicting mpg using displacement
Intercept: 35.12063593840391
Coefficient: [-0.06005143]

R2 Score: 0.6482294003193044

Predicting mpg using horsepower
Intercept: 39.93586102117047
Coefficient: [-0.15784473]
R2 Score: 0.6059482578894348

Predicting mpg using weight
Intercept: 46.21652454901758
Coefficient: [-0.00764734]
R2 Score: 0.6926304331206254

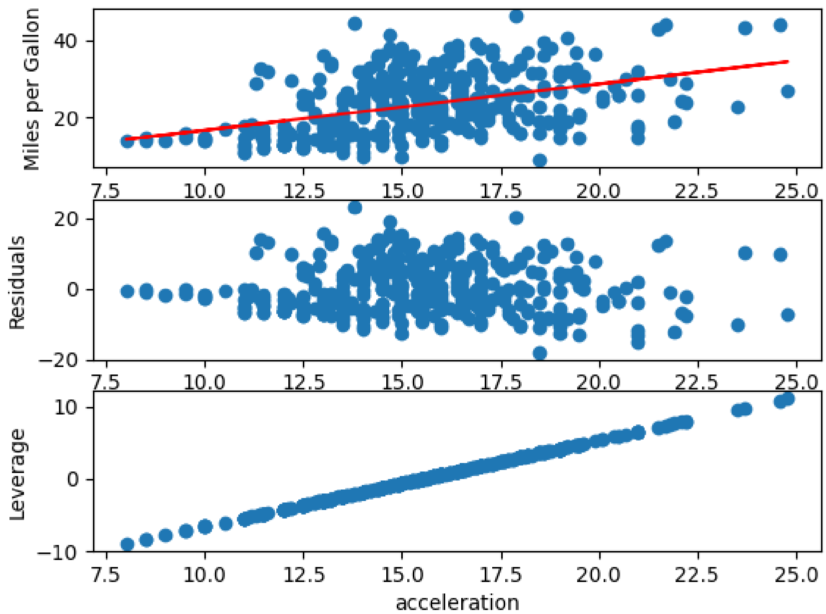
Predicting mpg using acceleration
Intercept: 4.833249804843799
Coefficient: [1.19762419]
R2 Score: 0.1792070501562546

Predicting mpg using year
Intercept: -70.01167409014336
Coefficient: [1.23003546]
R2 Score: 0.33702781330962295

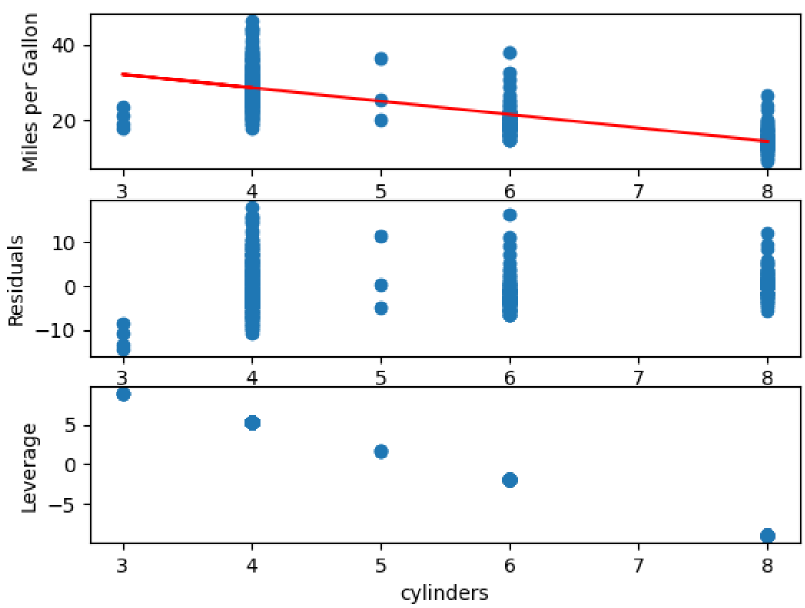
Predicting mpg using origin
Intercept: 14.81197361541246
Coefficient: [5.47654748]
R2 Score: 0.3194609386689675

i. For all values there is a relationship, just some are not very descriptive (such as acceleration)
ii. Weight, horsepower, displacement, and cylinders are all decent predictors, but weight is the best with a r^2 of 0.69
iii. The coefficient for year suggests that from 1970 to 1982, the average mpg of the data generally increased. This makes sense as new advancements would allow for more efficient engines and lighter vehicles.

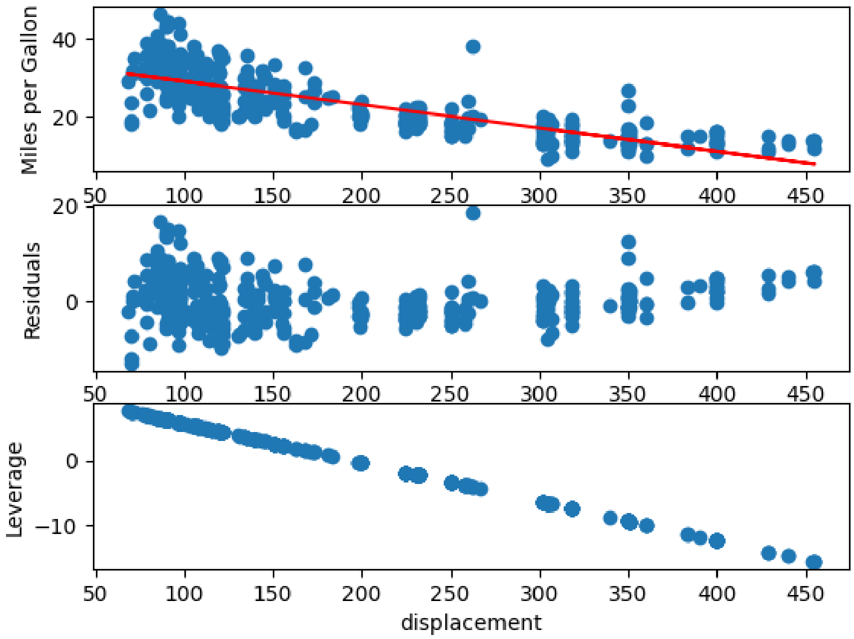
d.



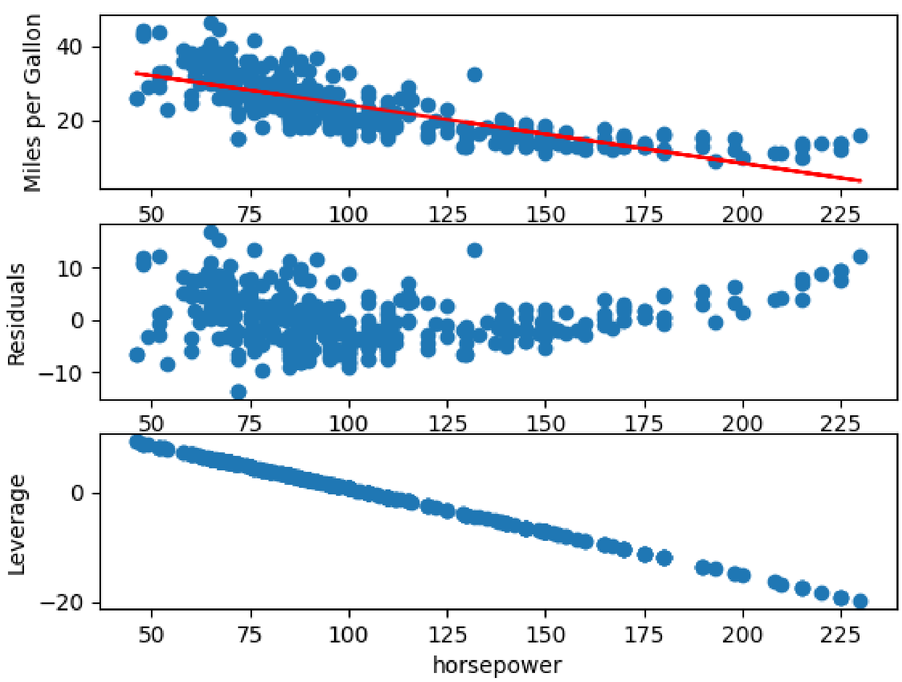
Residuals show that there is far less variation for lower values of acceleration.



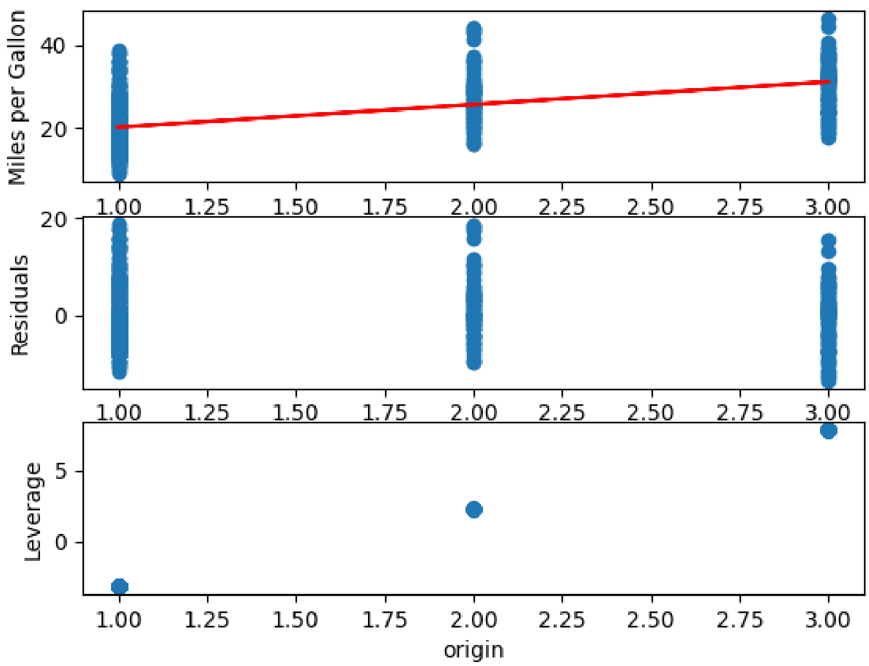
The values at 3 cylinders have a surprising leverage considering that there seem to be few of them. This is not very descriptive due to the large grained value of cylinders not allowing for a lot of variation.



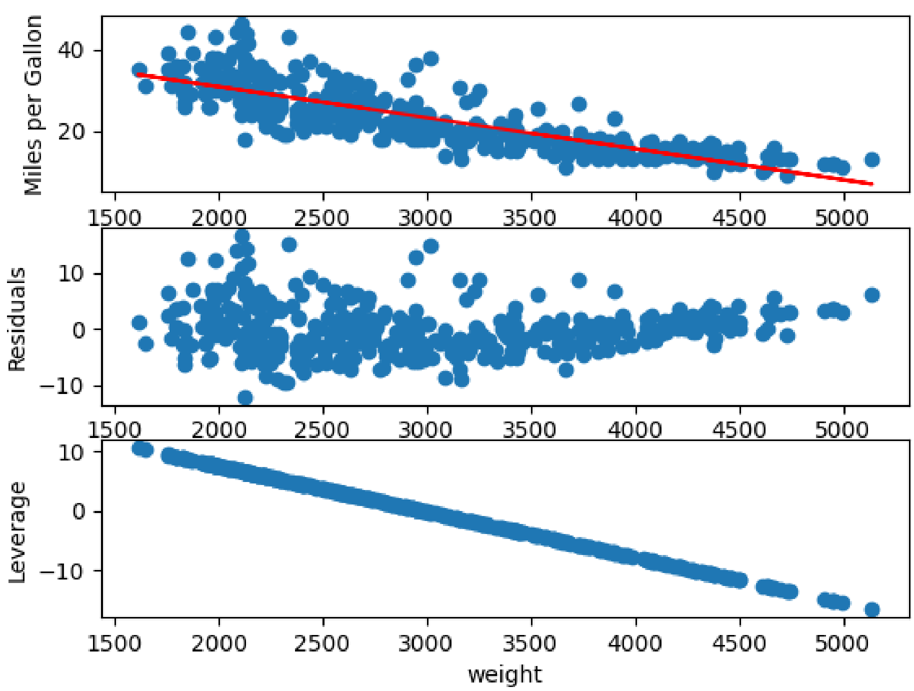
There is a clear outlier at just over 250 displacement that comes in much higher than normal. Variance decreases as displacement increases, but there is also far more data for lower displacement vehicles (as they are much more common).



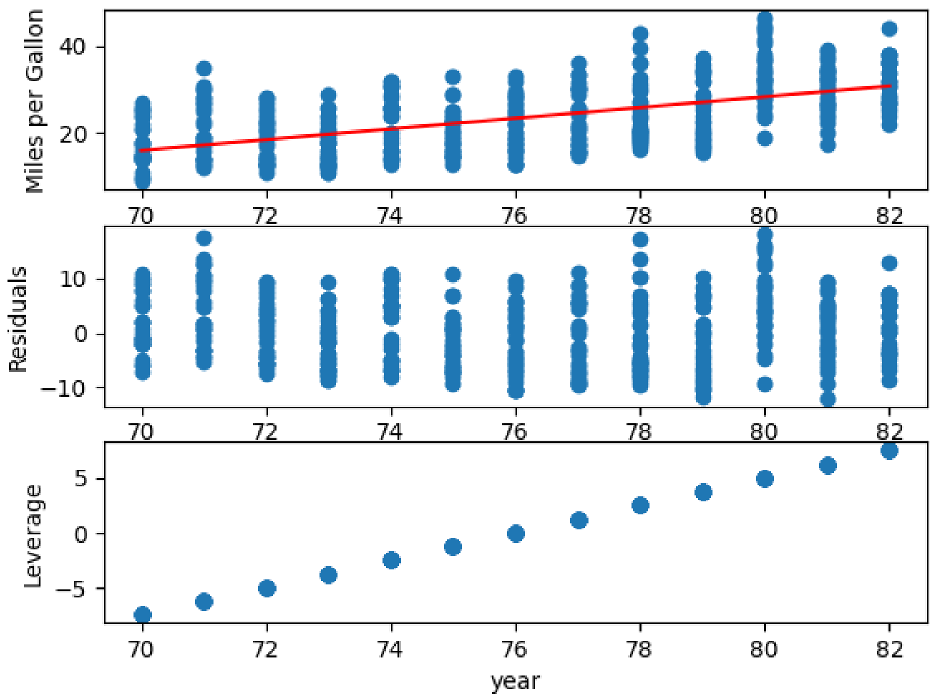
Clearly another decent predictor of mpg. Clearly left skewed like the previous. The residuals show a slight parabolic shape which could imply that introducing a non-linear term could provide better results.



This doesn't mean anything at all, since origin is just an enum for countries. The scatterplot may have some meaning but putting a line to it is not helpful.



Decent fit. Variation decreases as weight increases.



Shows little more than a general trend toward better mpg as time increases. Interesting to note that each year has a similar variance in mpg values.

e.

Predicting mpg using cylinders x cylinders
Intercept: 33.216581988812166
Coefficient: [-0.29748351]
R2 Score: 0.5934472096925252

Predicting mpg using cylinders x displacement
Intercept: 30.989620265007336
Coefficient: [-0.00611767]
R2 Score: 0.6128468289184077

Predicting mpg using cylinders x horsepower
Intercept: 32.49813405356838
Coefficient: [-0.01444064]
R2 Score: 0.6031409605327316

Predicting mpg using cylinders x weight
Intercept: 34.294799210406666
Coefficient: [-0.00061675]

R2 Score: 0.6519639183732271

Predicting mpg using cylinders x acceleration
Intercept: 40.92789254500349
Coefficient: [-0.21146136]
R2 Score: 0.3515194612646172

Predicting mpg using cylinders x year
Intercept: 42.480533737757355
Coefficient: [-0.04602299]
R2 Score: 0.5235641708125532

Predicting mpg using cylinders x origin
Intercept: 19.663640342408446
Coefficient: [0.48200682]
R2 Score: 0.030928157699728898

Predicting mpg using displacement x cylinders
Intercept: 30.989620265007336
Coefficient: [-0.00611767]
R2 Score: 0.6128468289184077

Predicting mpg using displacement x displacement
Intercept: 29.257667934664347
Coefficient: [-0.00011929]
R2 Score: 0.5660372846661386

Predicting mpg using displacement x horsepower
Intercept: 29.889010319034202
Coefficient: [-0.00026942]
R2 Score: 0.5600156104929326

Predicting mpg using displacement x weight
Intercept: 31.263398468891392
Coefficient: [-1.18161003e-05]
R2 Score: 0.6285078407559446

Predicting mpg using displacement x acceleration
Intercept: 36.93353104785503
Coefficient: [-0.004708]
R2 Score: 0.6025228861921829

Predicting mpg using displacement x year
Intercept: 35.29576522247327

Coefficient: [-0.00081002]
R2 Score: 0.617733761954898

Predicting mpg using displacement x origin
Intercept: 33.44015515235274
Coefficient: [-0.03921957]
R2 Score: 0.2065745127912666

Predicting mpg using horsepower x cylinders
Intercept: 32.49813405356838
Coefficient: [-0.01444064]
R2 Score: 0.6031409605327316

Predicting mpg using horsepower x displacement
Intercept: 29.889010319034202
Coefficient: [-0.00026942]
R2 Score: 0.5600156104929326

Predicting mpg using horsepower x horsepower
Intercept: 30.465772857016926
Coefficient: [-0.0005665]
R2 Score: 0.5073670089832611

Predicting mpg using horsepower x weight
Intercept: 32.91603270088104
Coefficient: [-2.79140389e-05]
R2 Score: 0.6203603485412151

Predicting mpg using horsepower x acceleration
Intercept: 47.72989725518964
Coefficient: [-0.0156611]
R2 Score: 0.6271509270660596

Predicting mpg using horsepower x year
Intercept: 40.451416248529924
Coefficient: [-0.00215843]
R2 Score: 0.5634336794996258

Predicting mpg using horsepower x origin
Intercept: 23.232932177439157
Coefficient: [0.00141404]
R2 Score: 0.0001339830518214402

Predicting mpg using weight x cylinders

Intercept: 34.294799210406666
Coefficient: [-0.00061675]
R2 Score: 0.6519639183732271

Predicting mpg using weight x displacement
Intercept: 31.263398468891392
Coefficient: [-1.18161003e-05]
R2 Score: 0.6285078407559446

Predicting mpg using weight x horsepower
Intercept: 32.91603270088104
Coefficient: [-2.79140389e-05]
R2 Score: 0.6203603485412151

Predicting mpg using weight x weight
Intercept: 34.46926951055215
Coefficient: [-1.14998452e-06]
R2 Score: 0.650735168904341

Predicting mpg using weight x acceleration
Intercept: 40.53176874235595
Coefficient: [-0.00037716]
R2 Score: 0.3407467995920773

Predicting mpg using weight x year
Intercept: 45.70696643040454
Coefficient: [-9.88190265e-05]
R2 Score: 0.597694315492433

Predicting mpg using weight x origin
Intercept: 20.30400637108656
Coefficient: [0.00073153]
R2 Score: 0.021686269517247836

Predicting mpg using acceleration x cylinders
Intercept: 40.92789254500349
Coefficient: [-0.21146136]
R2 Score: 0.3515194612646172

Predicting mpg using acceleration x displacement
Intercept: 36.93353104785503
Coefficient: [-0.004708]
R2 Score: 0.6025228861921829

Predicting mpg using acceleration x horsepower
Intercept: 47.72989725518964
Coefficient: [-0.0156611]
R2 Score: 0.6271509270660596

Predicting mpg using acceleration x weight
Intercept: 40.53176874235595
Coefficient: [-0.00037716]
R2 Score: 0.3407467995920773

Predicting mpg using acceleration x acceleration
Intercept: 14.597504984386502
Coefficient: [0.035518]
R2 Score: 0.16302352249011032

Predicting mpg using acceleration x year
Intercept: 2.561770244902391
Coefficient: [0.01764212]
R2 Score: 0.2730276186569306

Predicting mpg using acceleration x origin
Intercept: 14.938653635570311
Coefficient: [0.34065906]
R2 Score: 0.3913118973914146

Predicting mpg using year x cylinders
Intercept: 42.480533737757355
Coefficient: [-0.04602299]
R2 Score: 0.5235641708125532

Predicting mpg using year x displacement
Intercept: 35.29576522247327
Coefficient: [-0.00081002]
R2 Score: 0.617733761954898

Predicting mpg using year x horsepower
Intercept: 40.451416248529924
Coefficient: [-0.00215843]
R2 Score: 0.5634336794996258

Predicting mpg using year x weight
Intercept: 45.70696643040454
Coefficient: [-9.88190265e-05]
R2 Score: 0.597694315492433

Predicting mpg using year x acceleration
Intercept: 2.561770244902391
Coefficient: [0.01764212]
R2 Score: 0.2730276186569306

Predicting mpg using year x year
Intercept: -23.65807503688161
Coefficient: [0.00814042]
R2 Score: 0.34135140899954497

Predicting mpg using year x origin
Intercept: 14.485654992443445
Coefficient: [0.07446939]
R2 Score: 0.36379995000204024

Predicting mpg using origin x cylinders
Intercept: 19.663640342408446
Coefficient: [0.48200682]
R2 Score: 0.030928157699728898

Predicting mpg using origin x displacement
Intercept: 33.44015515235274
Coefficient: [-0.03921957]
R2 Score: 0.2065745127912666

Predicting mpg using origin x horsepower
Intercept: 23.232932177439157
Coefficient: [0.00141404]
R2 Score: 0.0001339830518214402

Predicting mpg using origin x weight
Intercept: 20.30400637108656
Coefficient: [0.00073153]
R2 Score: 0.021686269517247836

Predicting mpg using origin x acceleration
Intercept: 14.938653635570311
Coefficient: [0.34065906]
R2 Score: 0.3913118973914146

Predicting mpg using origin x year
Intercept: 14.485654992443445
Coefficient: [0.07446939]

R2 Score: 0.36379995000204024

Predicting mpg using origin x origin
Intercept: 19.1925466197405
Coefficient: [1.35775385]
R2 Score: 0.3006914705169893

As you can see, there are some that are significant, but none are better than the best simple model, so this approach is not worth it.

f.

Predicting mpg using log of cylinders
Intercept: 56.605930177297864
Coefficient: [-20.05995044]
R2 Score: 0.6034457474746462

Predicting mpg using log of displacement
Intercept: 85.69058349210749
Coefficient: [-12.1384539]
R2 Score: 0.6863348898210174

Predicting mpg using log of horsepower
Intercept: 108.69970699574486
Coefficient: [-18.58218476]
R2 Score: 0.6683347641192137

Predicting mpg using log of weight
Intercept: 209.9433404696741
Coefficient: [-23.43173838]
R2 Score: 0.7126631343895841

Predicting mpg using log of acceleration
Intercept: -27.834291263755443
Coefficient: [18.80125132]
R2 Score: 0.19000944491763905

Predicting mpg using log of year
Intercept: -377.87302742877563
Coefficient: [92.6985447]
R2 Score: 0.3323743535175887

Predicting mpg using log of origin
Intercept: 20.107449427745856
Coefficient: [9.77178191]
R2 Score: 0.3297926714189251

Predicting mpg using square root of cylinders
Intercept: 62.81122386010685
Coefficient: [-17.02711648]
R2 Score: 0.6058312164839386

Predicting mpg using square root of displacement
Intercept: 47.118389122432674
Coefficient: [-1.75877931]
R2 Score: 0.6745852005048071

Predicting mpg using square root of horsepower
Intercept: 58.705172037217494
Coefficient: [-3.50352375]
R2 Score: 0.6437035832706475

Predicting mpg using square root of weight
Intercept: 69.67217695049604
Coefficient: [-0.85560124]
R2 Score: 0.7057597690908815

Predicting mpg using square root of acceleration
Intercept: -14.177285156676295
Coefficient: [9.58150742]
R2 Score: 0.18548306674505965

Predicting mpg using square root of year
Intercept: -162.7122169137047
Coefficient: [21.3629367]
R2 Score: 0.33474111552905306

Predicting mpg using square root of origin
Intercept: 5.3237428792775745
Coefficient: [14.86174592]
R2 Score: 0.3258151973984553
As you can see, both the log and sqrt of weight have marginally better r^2 values than weight (the previous best). NOTE that x^2 was already done in the previous section and produced no improvement.

3.

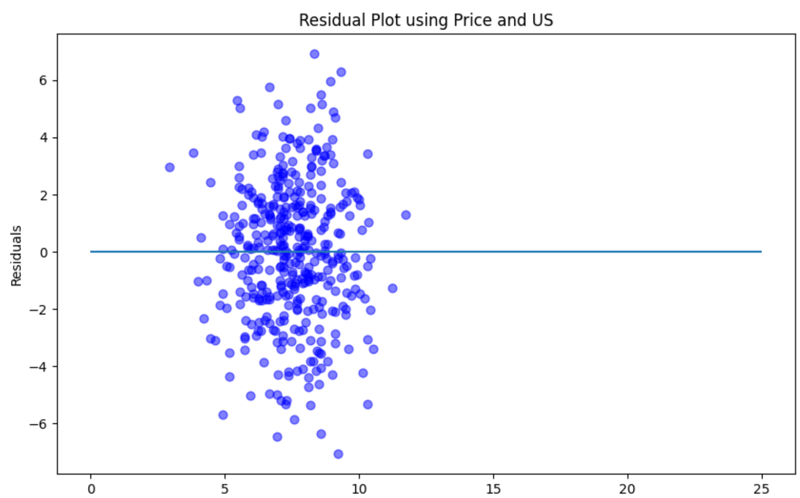
a.

Coefficients: Price: -0.0544588491775822, Urban: -0.021916150814141, US: 1.200572697794116
Intercept: 13.043468936764896

b. With no other data, we can expect the seat to have 13 thousand sales. Price negatively impacts sales, more expensive = less sales at a 1:0.05 ratio. Being in an urban area reduces the number of sales by .02 thousand sales. Being in the US increases sales by 1.2 thousand sales.
c. Sales = -0.0544588491775822Price + -0.021916150814141Urban + 1.200572697794116US. Where Urban is 1 if urban or 0 otherwise, and US is 1 if US and 0 otherwise
d. Urban has a high p value, we cannot reject the null hypothesis.

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.0435	0.651	20.036	0.000	11.764	14.323
Price	-0.0545	0.005	-10.389	0.000	-0.065	-0.044
Urban	-0.0219	0.272	-0.081	0.936	-0.556	0.512
US	1.2006	0.259	4.635	0.000	0.691	1.710

e.
Coefficients: Price: -0.05447763247978729, US: 1.1996429432266782
Intercept: 13.030792754615764
f. a) has a R^2 value of 0.23927539218405547, while e) has 0.23926288842678567. They fit almost exactly the same, but both are decent.
g.
95% confidence interval for Price: [-0.05447763 -0.05447763]
95% confidence interval for US: [1.19964294 1.19964294]



h.
The distribution of the residuals is still quite normal, no major outliers. This would be an expected outcome from a normal distribution.

4.

a.
Predicting crim using zn
Intercept: 4.453693755086385
Coefficient: [-0.07393498]

R2 Score: 0.04018790803211081

Predicting crim using indus
Intercept: -2.0637426063278133
Coefficient: [0.50977633]
R2 Score: 0.16531007043075152

Predicting crim using chas
Intercept: 3.7444468365180477
Coefficient: [-1.89277655]
R2 Score: 0.0031238689633057426

Predicting crim using nox
Intercept: -13.719882309974336
Coefficient: [31.2485312]
R2 Score: 0.17721718179269375

Predicting crim using rm
Intercept: 20.481804177792405
Coefficient: [-2.68405122]
R2 Score: 0.048069116716083604

Predicting crim using age
Intercept: -3.7779063179682684
Coefficient: [0.10778623]
R2 Score: 0.12442145175894637

Predicting crim using dis
Intercept: 9.49926164655728
Coefficient: [-1.55090168]
R2 Score: 0.1441493749253987

Predicting crim using rad
Intercept: -2.2871594483103497
Coefficient: [0.61791093]
R2 Score: 0.39125668674998915

Predicting crim using tax
Intercept: -8.528369093069161
Coefficient: [0.02974225]
R2 Score: 0.3396142433788122

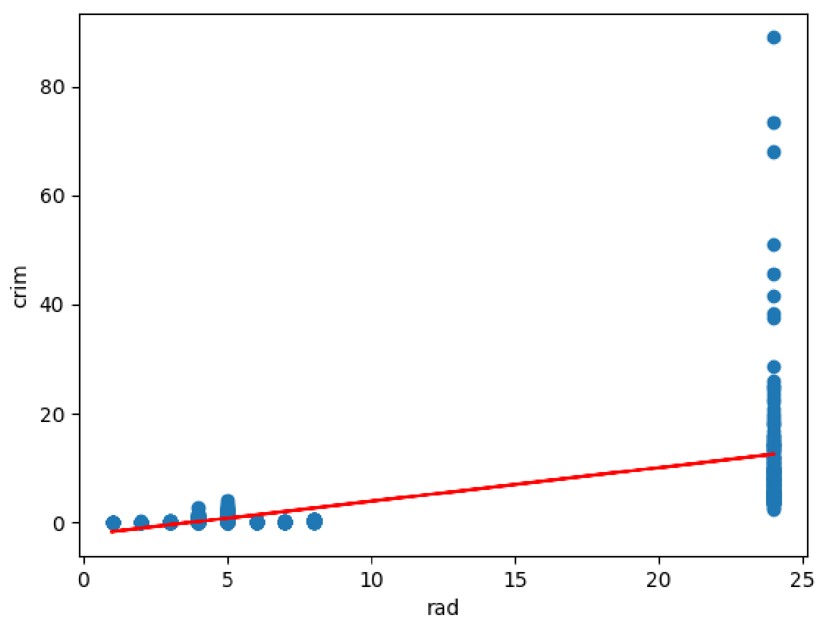
Predicting crim using ptratio
Intercept: -17.64693347244794

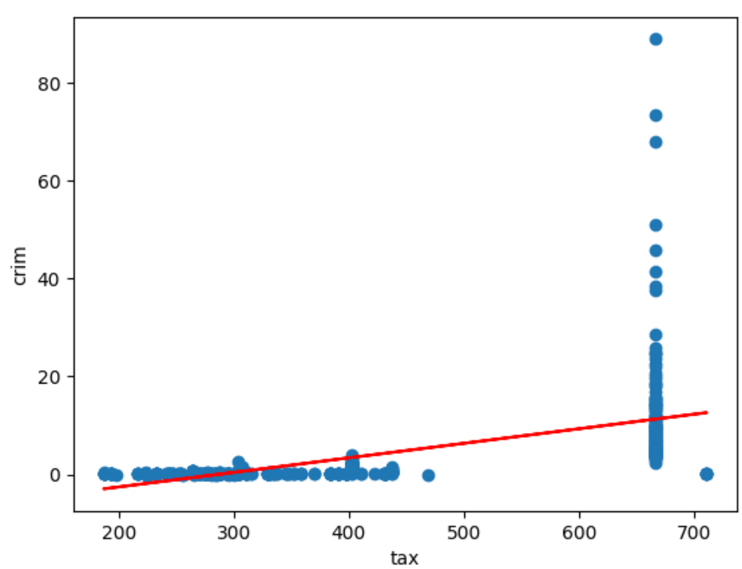
Coefficient: [1.15198279]
R2 Score: 0.0840684389437365

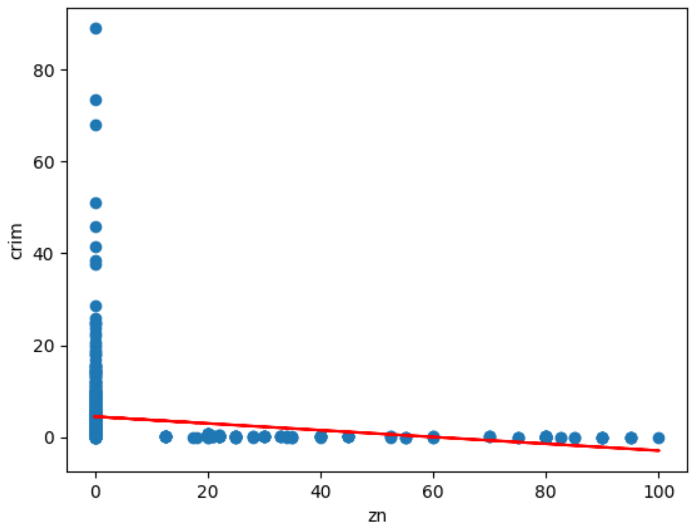
Predicting crim using lstat
Intercept: -3.330538057145062
Coefficient: [0.54880478]
R2 Score: 0.2075909325343357

Predicting crim using medv
Intercept: 11.796535750221913
Coefficient: [-0.36315992]
R2 Score: 0.15078046904975706

The predictors that best capture the variance of the data are tax, rad, and zn (corresponding R^2 values above). Below are their graphs to show why this is the case.


crim
rad


crim
tax

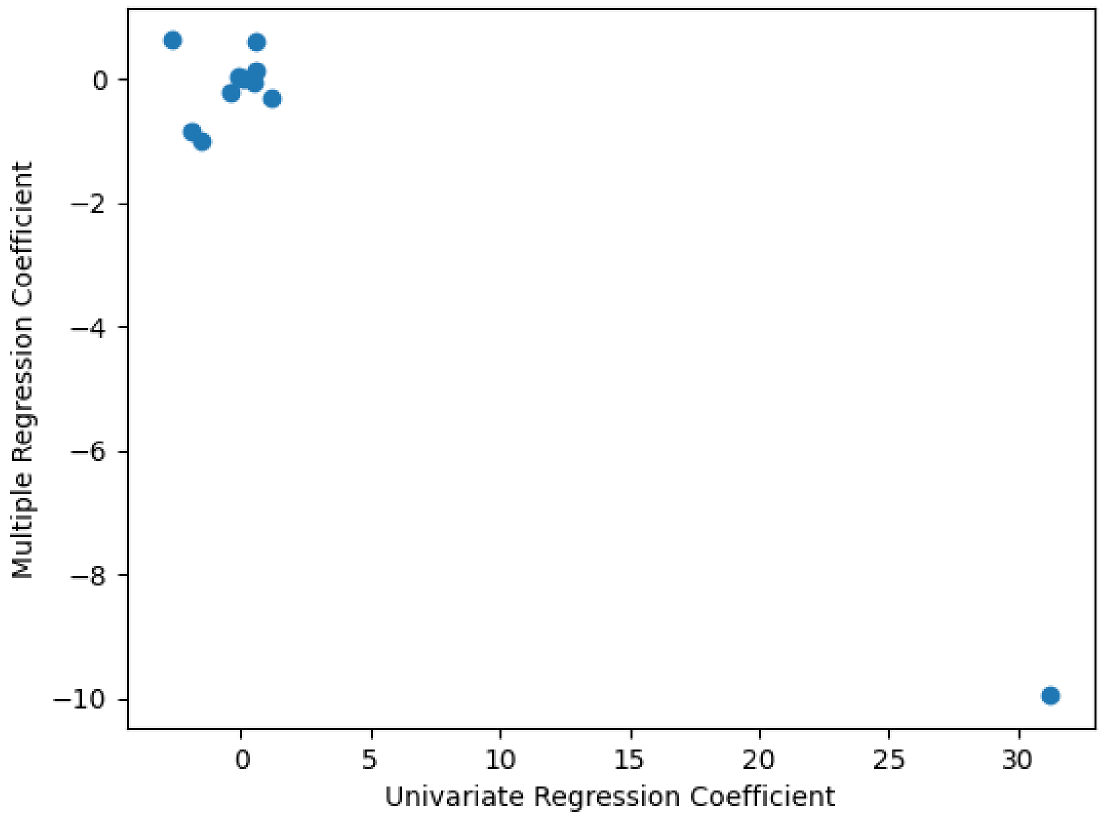


b.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   crim   R-squared:                       0.449
Model:                            OLS   Adj. R-squared:                  0.436
Method:                 Least Squares   F-statistic:                     33.52
Date:                Tue, 15 Nov 2022   Prob (F-statistic):           2.03e-56
Time:                        16:19:17   Log-Likelihood:                -1655.4
No. Observations:                 506   AIC:                             3337.
Df Residuals:                     493   BIC:                             3392.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.7784      7.082      1.946      0.052      -0.136      27.693
zn             0.0457      0.019      2.433      0.015       0.009       0.083
indus         -0.0584      0.084     -0.698      0.486      -0.223       0.106
chas          -0.8254      1.183     -0.697      0.486      -3.150       1.500
nox           -9.9576      5.290     -1.882      0.060     -20.351       0.436
rm             0.6289      0.607      1.036      0.301      -0.564       1.822
age           -0.0008      0.018     -0.047      0.962      -0.036       0.034
dis           -1.0122      0.282     -3.584      0.000      -1.567      -0.457
rad            0.6125      0.088      6.997      0.000       0.440       0.784
tax           -0.0038      0.005     -0.730      0.466      -0.014       0.006
ptratio       -0.3041      0.186     -1.632      0.103      -0.670       0.062
lstat          0.1388      0.076      1.833      0.067      -0.010       0.288
medv          -0.2201      0.060     -3.678      0.000      -0.338      -0.103
==============================================================================
Omnibus:                      663.436   Durbin-Watson:                   1.516
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            80856.852
Skew:                           6.579   Prob(JB):                         0.00
Kurtosis:                      63.514   Cond. No.                     1.24e+04
==============================================================================
```

With threshold of P= 0.01, the only predictors that we can reject the null hypothesis on are dis, rad, and medv.

c.



d.

Predicting crim using zn, zn^2, and zn^3
Intercept: 4.846050076279062
Coefficient: [-3.32188415e-01 6.48263365e-03 -3.77579253e-05]
R2 Score: 0.05824197422258326

Predicting crim using indus, indus^2, and indus^3
Intercept: 3.6625682786867575
Coefficient: [-1.96521293 0.2519373 -0.00697601]
R2 Score: 0.2596578579195665

Predicting crim using chas, chas^2, and chas^3
Intercept: 3.790769651062254
Coefficient: [1.23257144e+14 -6.16285720e+13 -6.16285720e+13]
R2 Score: 0.002741838752170711

Predicting crim using nox, nox^2, and nox^3

Intercept: 233.08659066305339
Coefficient: [-1279.37125166 2248.54405256 -1245.70287375]
R2 Score: 0.2969778956287379

Predicting crim using rm, rm^2, and rm^3
Intercept: 112.62459631863406
Coefficient: [-39.15013634   4.55089591 -0.17447695]
R2 Score: 0.06778606116878627

Predicting crim using age, age^2, and age^3
Intercept: -2.548763403675663
Coefficient: [2.73653131e-01 -7.22959558e-03 5.74530704e-05]
R2 Score: 0.17423099358657324

Predicting crim using dis, dis^2, and dis^3
Intercept: 30.047611563456307
Coefficient: [-15.55435349   2.45207217 -0.11859864]
R2 Score: 0.27782477308673637

Predicting crim using rad, rad^2, and rad^3
Intercept: -0.6055447455773111
Coefficient: [0.51273604 -0.07517736 0.003209]
R2 Score: 0.40003687202422356

Predicting crim using tax, tax^2, and tax^3
Intercept: 19.183581466940225
Coefficient: [-1.53309613e-01 3.60826646e-04 -2.20371513e-07]
R2 Score: 0.36888207966295994

Predicting crim using ptratio, ptratio^2, and ptratio^3
Intercept: 477.1840461034031
Coefficient: [-82.36053772   4.63534723 -0.08476032]
R2 Score: 0.11378157744698159

Predicting crim using lstat, lstat^2, and lstat^3
Intercept: 1.2009655811881563
Coefficient: [-0.44906559 0.05577942 -0.00085737]
R2 Score: 0.21793243242225602

Predicting crim using medv, medv^2, and medv^3
Intercept: 53.165538094375584
Coefficient: [-5.09483054e+00 1.55496490e-01 -1.49010277e-03]
R2 Score: 0.4202002565634151

As can be seen, medv performed the best when used in a higher order linear regression. However, it still did not outperform the multivariate regression, and only marginally better than the univariate in terms of the R^2 value. There is not evidence to support that this data should be represented with a higher order function.