Yurim Park hw2

1. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n, the number of observations, and p, the number of features.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Regression and inference with n = 500 and p = 3

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 30 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and twelve other variables.

Classification and prediction with n = 20 and p = 13

(c) We are interested in predicting the percent change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2016. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Regression and prediction with n = 52 and p = 3

2. Suppose we with to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point,

X1 = X2 = X3 = 0.

i. What is our prediction with K = 1? Why?

The nearest neighbor to test point (0,0,0) is Obs 5 (-1, 0, 1) with euclidean distance ~1.41. Since Obs 5 was Green, we predict (K=1) that the test point will also be Green.

ii. What is our prediction with K = 3? Why?

The nearest three neighbors to test point (0,0,0) are Obs 5, Obs 6 (with distance ~1.73), and Obs 2 (with distance 2). Since Obs 5 was Green, Obs 6 was Red, and Obs 2 was Red, we predict (K=3) the test point wil be the majority – Red.

(b) Compute the Manhattan distance between each observation and the same test point.

i. What is our prediction with K = 1? Why?

With K=1, we select the nearest neighbor to the test point. The observation with the smallest Manhattan distance is Observation 5 (Manhattan distance = 2), which corresponds to the category "Green". Therefore, our prediction with K=1 would be "Green".

ii. What is our prediction with K = 3? Why?

With K=3, we select the three nearest neighbors to the test point. The three observations with the smallest Manhattan distances are Observations 1, 5, and 6 (Manhattan distances = 2, 2, 3 respectively). Two of them correspond to the category "Green", and one corresponds to the category "Red". Since there are more "Green" observations among the nearest neighbors, our prediction with K=3 would be "Green".

3. Load the Boston data set, which is part of the ISLP library. Details can be found at https://intro-stat-learning.github.io/ISLP/datasets/Boston.html.

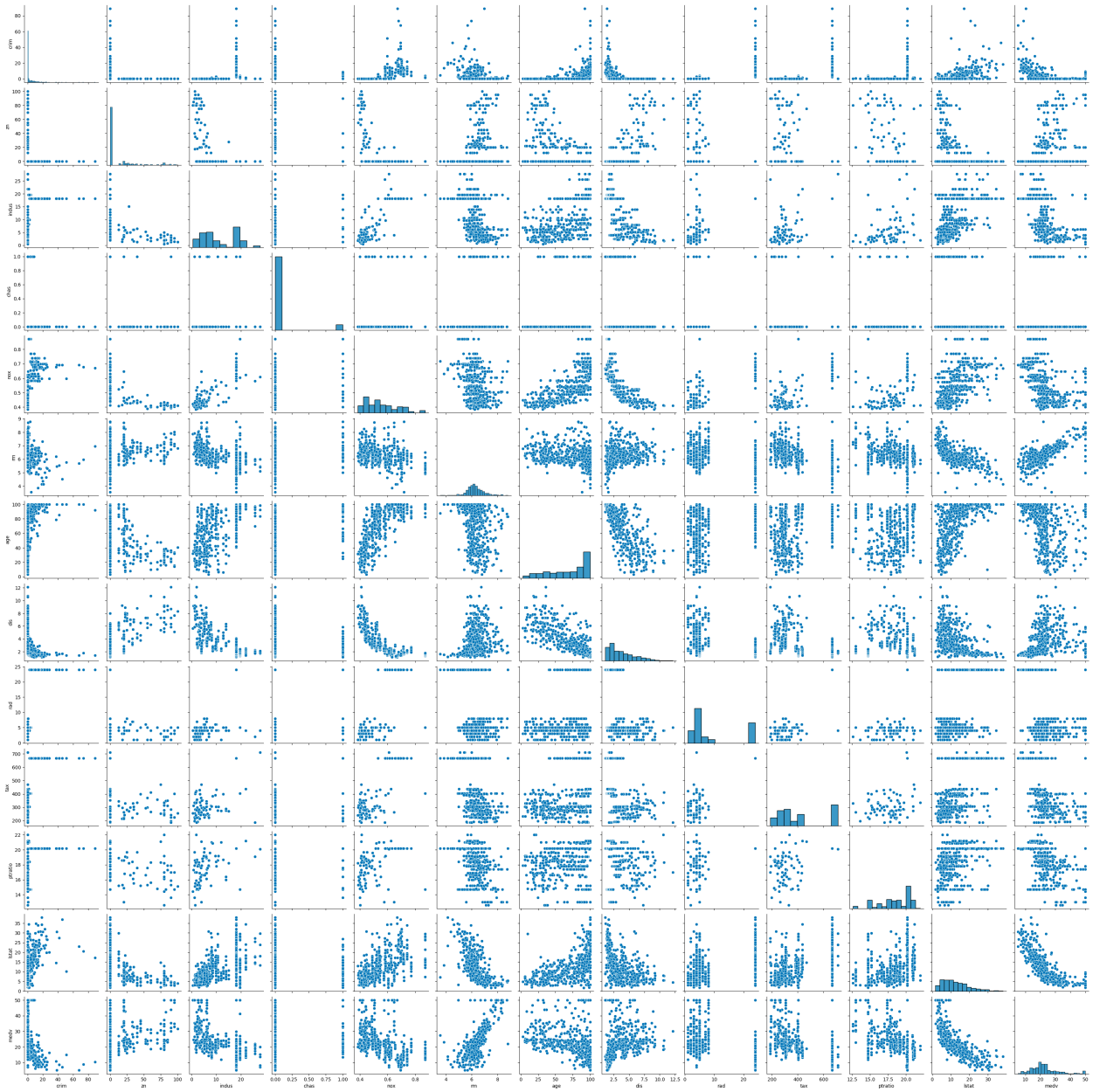
(a) How many rows are in this data set? How many columns? What do the rows represent?

The total number of rows = 506

The total number of columns = 14

The columns are variables that influence the housing values of the suburbs of Boston

(b) Make pairwise scatterplots of the predictors (columns) in this data set with the per capita crime rate.



(i) Property tax rate and index of accessibility to radial highways is strongly correlated. It indicates that town areas are better connected to radial highways where property tax rates are higher. However, from the scatter plot we can observe that the correlation could be impacted due to outliers. It would be ideal to validate the correlation basis partial correlation coefficient.

(ii) Lower status of the population and median value of the owner occupied homes is strongly and negatively correlated. The relationship seems to be non linear. Probably home ownership is less for lower population.

(iii) Average number of rooms per dwelling is positively correlated to Median value of owner occupied homes. It means that higher value areas will have larger dwellings.

*(iv) Median value of owner occupied homes has poor correlation to Charles River Dummy Variable and distance to five **Boston employment centres.*

(v) Nitrogen oxides concentration and age (proportion of owner-occupied units built prior to 1940) are positively correlated and non linear in relationship.

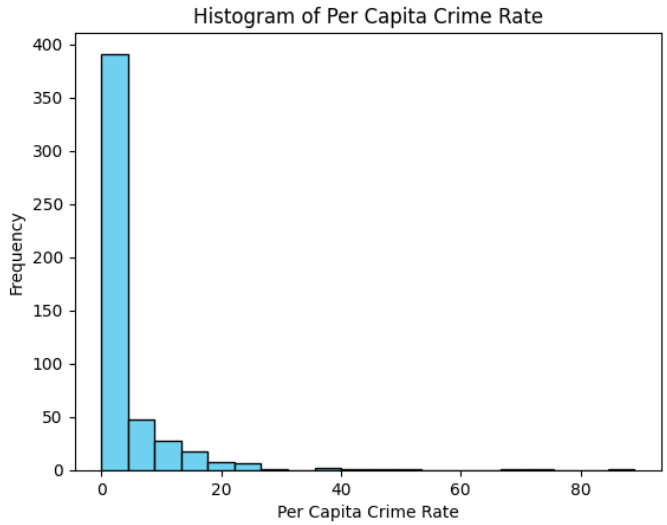
(c) Are any of the predictors correlated with per capita crime rate? If so, explain the relationship. (Assume "correlated" to mean a Pearson correlation coefficient of at least ±0.5.)

Crime rate by town seems to have weak to moderate correlations with other predictors. The scatter plots give the impression that the index of accessibility to radial highways or property tax rate do not affect thecrime rate. Distance from Boston employment centres has a negative effect over crime rate, however seems to be weak.

The number of rooms per dwelling is also negatively correlated with crime rate and is weak. Lower status ofthe population is positively correlated to crime rate. Probably the lower status of the population pushes the crime rate up.Crime rate and age of owner occupied units built prior 1940 is also positively correlated,

however it seems that crime rate is higher for units with higher age. Crime rate seems to be higher for areas where proportion of non-retail business acres is low.

(d) Provide a histogram for the per capita crime rate.



(e) How many suburbs of Boston have a crime rate larger than 30?

Number of suburbs with a crime rate larger than 30: 8

(f) Provide the range of each predictor.

Range of each predictor:

crim 88.96988

zn 100.00000

indus 27.28000

chas 1.00000

nox 0.48600

rm 5.21900

age 97.10000

dis 10.99690

rad 23.00000

tax 524.00000

ptratio 9.40000

lstat 36.24000

medv 45.00000

dtype: float64

(g) How many of the suburbs in this data set bound the Charles river?

Number of suburbs that bound the Charles River: 35

(h) What is the median pupil-teacher ratio among the towns in this data set?

Median pupil-teacher ration among towns is 19.05.

4. (18 points) Adapted from ISLP 3.7.8.

This question involves the use of simple linear regression on the Auto data set. Details can be found at https://intro-stat-learning.github.io/ISLP/datasets/Auto.html.

(a) Use the sm.OLS() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summarize() function to print the results. Note: For simplicity, drop any records that have missing values for horsepower and use all remaining data for training.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.606
Model:                            OLS   Adj. R-squared:                  0.605
Method:                 Least Squares   F-statistic:                     599.7
Date:                Tue, 05 Mar 2024   Prob (F-statistic):           7.03e-81
Time:                        01:16:42   Log-Likelihood:                -1178.7
No. Observations:                 392   AIC:                             2361.
Df Residuals:                     390   BIC:                             2369.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.9359      0.717     55.660      0.000      38.525      41.347
horsepower    -0.1578      0.006    -24.489      0.000      -0.171      -0.145
==============================================================================
Omnibus:                       16.432   Durbin-Watson:                   0.920
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.305
Skew:                           0.492   Prob(JB):                     0.000175
Kurtosis:                       3.299   Cond. No.                         322.
==============================================================================
```

i. Explain why we can deduce that there is a significant association between the predictor and the response.

A significant association between the predictor and the response is indicated by a significant coefficient (with a low p-value), a confidence interval that does not include zero, and possibly a high R-squared value. These statistical measures provide evidence that changes in the predictor variable are associated with changes in the response variable, and that this association is unlikely to be due to random chance.

ii. How strong is the relationship between the predictor and the response? Use the R2 statistic to support your answer.

For a unit increase in horsepower, our model predicts mpg will decrease by -0.1578. So for example, increasing horsepower by 10 is expected to decrease efficiency by -1.578 mpg.

iii. Is the relationship between the predictor and the response positive or negative? How can you tell?

Negative. The coefficient of the predictor variable is negative.

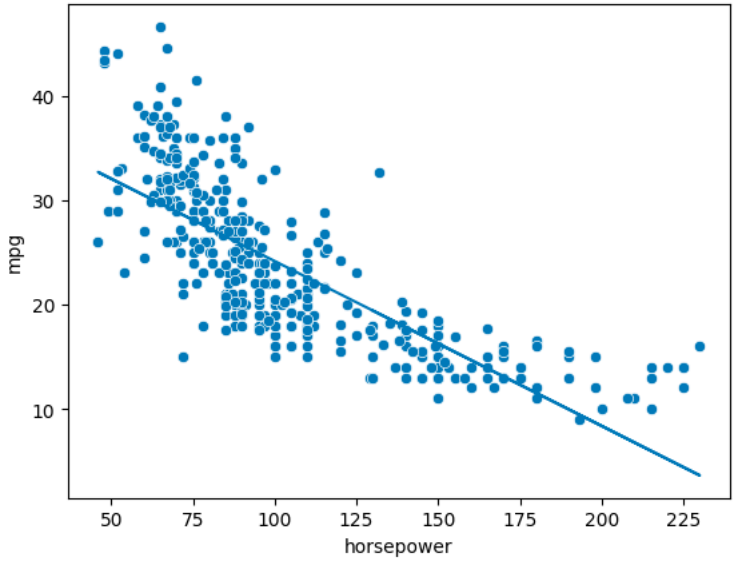
iv. According to this model, what is the predicted mpg for my 2023 Volkswagen Golf, which has a horsepower of 241?

Predicted mpg for a 2023 Volkswagen Golf with 241 horsepower: 1.895 mpg

v. Why might it be inappropriate to use this model to estimate the mpg of my Golf?

The fuel efficiency of a car (mpg) is influenced by a wide range of factors beyond just horsepower. These include the car's weight, aerodynamics, engine efficiency, transmission type, tire resistance, and more. A model that only considers horsepower ignores these other important factors.

(b) Plot the response and the predictor in a new set of axes ax. Use the ax.axline() method or the abline() function defined in the lab to display the least squares regression line.



5. (25 points) Adapted from ISLP 3.7.9. This question involves the use of multiple linear regression on the Auto data set.

(a) Use the sm.OLS() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summarize() function to print the results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.821
Model:                            OLS   Adj. R-squared:                  0.818
Method:                 Least Squares   F-statistic:                     252.4
Date:                Tue, 05 Mar 2024   Prob (F-statistic):          2.04e-139
Time:                        01:59:08   Log-Likelihood:                -1023.5
No. Observations:                 392   AIC:                             2063.
Df Residuals:                     384   BIC:                             2095.
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -17.2184      4.644     -3.707      0.000     -26.350      -8.087
cylinders       -0.4934      0.323     -1.526      0.128      -1.129       0.142
displacement     0.0199      0.008      2.647      0.008       0.005       0.035
horsepower      -0.0170      0.014     -1.230      0.220      -0.044       0.010
weight          -0.0065      0.001     -9.929      0.000      -0.008      -0.005
acceleration     0.0806      0.099      0.815      0.415      -0.114       0.275
year             0.7508      0.051     14.729      0.000       0.651       0.851
origin           1.4261      0.278      5.127      0.000       0.879       1.973
==============================================================================
Omnibus:                       31.906   Durbin-Watson:                   1.309
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.100
...
```

i. Which predictors appear to have a statistically significant relationship to the response?

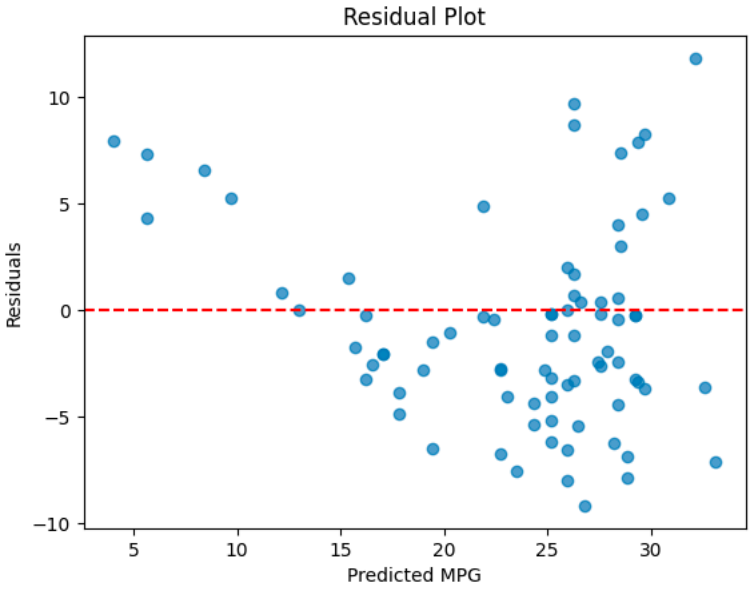
The predictors displacement, weight, year, and origin have statistically significant relationships with the response variable mpg at the 0.05 significance level. These variables significantly affect the miles per gallon that a vehicle can achieve, according to this model's results.

ii. What does the coefficient for the year variable suggest?

Interpreting this coefficient suggests that newer cars tend to have higher fuel efficiency (measured in mpg) compared to older cars. This positive coefficient implies that as the model year of a car increases by one unit, the mpg is expected to increase by approximately 0.7508 units, on average.

In practical terms, this suggests that advancements in automotive technology, improvements in engine efficiency, changes in emission standards, or other factors associated with newer model years contribute to higher fuel efficiency, as reflected in the observed data and captured by the regression model.

iii. Produce a residual plot which compares the observed mpg values to the predicted ones.



iv. Compute the MSE for this model.

Mean Squared Error: 22.153237123863413

(b) Fit a new linear model which only contains the significant variables in part (a) as predictors. Use the summarize() function to print the results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.606
Model:                            OLS   Adj. R-squared:                  0.605
Method:                 Least Squares   F-statistic:                     599.7
Date:                Tue, 05 Mar 2024   Prob (F-statistic):           7.03e-81
Time:                        21:43:12   Log-Likelihood:                -1178.7
No. Observations:                 392   AIC:                             2361.
Df Residuals:                     390   BIC:                             2369.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.9359      0.717     55.660      0.000      38.525      41.347
horsepower    -0.1578      0.006    -24.489      0.000      -0.171      -0.145
==============================================================================
Omnibus:                       16.432   Durbin-Watson:                   0.920
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.305
Skew:                           0.492   Prob(JB):                     0.000175
Kurtosis:                       3.299   Cond. No.                         322.
==============================================================================
```

i. In your new model, do all of the predictors now have a statistically significant relationship to the response?

Horsepower now has a statistically significant relationship to the response variable (mpg).

Because the p-value for the intercept is effectively 0.000 (not explicitly shown but implied by the high t-statistic value), indicating that the intercept is statistically significant. This means the model predicts a significant non-zero mpg value when horsepower is zero.

Also, the coefficient for horsepower is -0.1578 with a p-value effectively at 0.000, as indicated by its t-statistic of -24.489. This clearly shows that horsepower is statistically significant in predicting mpg.

ii. Compute the MSE for this new model.

Mean Squared Error (MSE): 23.943662938603104

iii. If we tested both models (a) and (b) on unseen holdout data, which model would you expect to have a smaller MSE? Why?

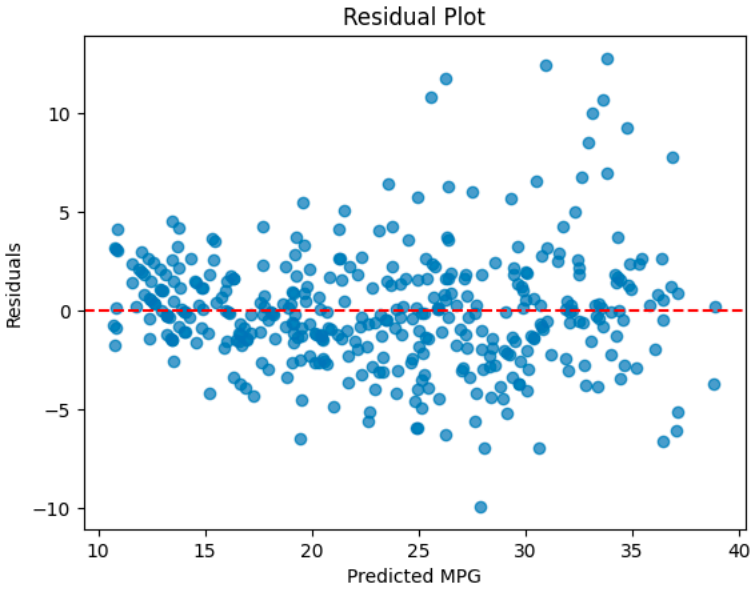
Model (a). Because model (a) includes additional significant predictors or interactions that are relevant and it could improve the model's explanatory power without causing

overfitting. So it could have a smaller MSE and provide more accurate representation of the data.

(c) Fit a new linear model which uses the variables year, origin, weight, and weight^2 as predictors. Use the summarize() function to print the results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.852
Model:                            OLS   Adj. R-squared:                  0.850
Method:                 Least Squares   F-statistic:                     556.5
Date:                Tue, 05 Mar 2024   Prob (F-statistic):          5.30e-159
Time:                        22:38:44   Log-Likelihood:                -986.86
No. Observations:                 392   AIC:                             1984.
Df Residuals:                     387   BIC:                             2004.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.4422      4.058     -0.109      0.913      -8.420       7.536
year           0.8247      0.044     18.675      0.000       0.738       0.912
origin         0.5026      0.243      2.064      0.040       0.024       0.981
weight        -0.0204      0.002    -13.281      0.000      -0.023      -0.017
weight^2    2.213e-06   2.33e-07      9.487      0.000    1.75e-06    2.67e-06
==============================================================================
Omnibus:                       65.205   Durbin-Watson:                   1.366
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              164.775
Skew:                           0.814   Prob(JB):                     1.66e-36
Kurtosis:                       5.727   Cond. No.                     2.94e+08
==============================================================================
```

i. Produce a residual plot which compares the observed mpg values to the predicted ones.



ii. The coefficient for weight^2 is extremely low. Does this suggest that we should remove this variable from the model? Why or why not?

No. The coefficient of weight 2 being extremely low does not necessarily mean that it should be removed from the model. It's because the coefficient of weight^2 is statistically significant, as indicated by the p-value associated with its t-test. A statistically significant coefficient, regardless of its magnitude, suggests that the variable contributes to explaining variation in the dependent variable, in this case, mpg.

6. (12 points) Adapted from ISLP 3.7.10. This question should be answered using the Carseats data set. Details can be found at https://intro-stat-learning.github.io/ISLP/datasets/Carseats.html.

(a) Fit a multiple regression model to predict Sales using Price and US. Use the summarize() function to print the results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.198
Model:                            OLS   Adj. R-squared:                  0.196
Method:                 Least Squares   F-statistic:                     98.25
Date:                Tue, 05 Mar 2024   Prob (F-statistic):           7.62e-21
Time:                        21:31:59   Log-Likelihood:                -938.23
No. Observations:                 400   AIC:                             1880.
Df Residuals:                     398   BIC:                             1888.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.6419      0.633     21.558      0.000      12.398      14.886
Price         -0.0531      0.005     -9.912      0.000      -0.064      -0.043
==============================================================================
Omnibus:                        2.537   Durbin-Watson:                   1.892
Prob(Omnibus):                  0.281   Jarque-Bera (JB):                2.611
Skew:                           0.175   Prob(JB):                        0.271
Kurtosis:                       2.816   Cond. No.                         591.
==============================================================================
```

(b) Provide an interpretation of each coefficient in the model.

When price increases by $1000 and other predictors are held constant, sales decrease by 54.459 unit sales. In otherwords, when price increases by $1000, the number of carseats sold decrease by 54,459.

A store's sale is not affected by whether or not it is in a Urban area.

A store in the US sales 1200 more carseats (in average) than a store that is abroad.

(c) Fit a new regression model that includes the interaction of Price and US as predictors. Use the summarize() function to print the results.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.239
Model:                            OLS   Adj. R-squared:                  0.234
Method:                 Least Squares   F-statistic:                     41.52
Date:                Tue, 05 Mar 2024   Prob (F-statistic):           2.39e-23
Time:                        21:34:05   Log-Likelihood:                -927.66
No. Observations:                 400   AIC:                             1863.
Df Residuals:                     396   BIC:                             1879.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           12.9748      0.953     13.614      0.000      11.101      14.849
Price                           -0.0540      0.008     -6.613      0.000      -0.070      -0.038
US                               1.2958      1.252      1.035      0.301      -1.166       3.757
Price_US_interaction            -0.0008      0.011     -0.078      0.937      -0.022       0.020
==============================================================================
Omnibus:                        0.659   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.719   Jarque-Bera (JB):                0.740
Skew:                           0.092   Prob(JB):                        0.691
Kurtosis:                       2.898   Cond. No.                     1.74e+03
==============================================================================
```

(d) Given the estimated coefficient of the interaction term, describe how the relationship between Price and Sales is different when US is “Yes” or “No.”

When US is "No": The relationship between Price and Sales follows the base coefficient for Price. In our hypothetical case, each unit increase in Price leads to a 0.05 unit decrease in Sales, assuming US="No" translates to the US variable being 0, which negates the interaction term.

When US is "Yes": The impact of Price on Sales is modified by the interaction term. You add the interaction term's coefficient to the Price coefficient when US="Yes". So, for a product sold in the US, the relationship between Price and Sales would be -0.05 (Price coefficient) + 0.02 (interaction coefficient) = -0.03. This means that for products sold in the US, each unit increase in Price leads to a smaller decrease in Sales (0.03 units) compared to products sold outside the US.

The interaction term's positive coefficient suggests that the deterrent effect of Price on Sales is mitigated in the US market. Possible reasons could include higher disposable income, brand preferences, or other market dynamics that make consumers in the US less sensitive to price increases than consumers in other markets.

(e) Should this interaction be included in this model? Why or why not?

No.

Because it lacks theoretical Justification. If there's no theoretical rationale for why the effect of Price on Sales should differ based on whether the product is sold in the US or not, then including the interaction term may not be appropriate. In this case, adding unnecessary complexity to the model can lead to overfitting and reduce interpretability.