Name: Yeshaswini Venkatesha Gupta

Assignment 1

1. (Total: 15 points)This exercise involves the Auto data set. Make sure that the missing values have been removed from the data.

(a) (2 points) Which of the predictors are quantitative, and which are qualitative?

Ans: quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year
qualitative: origin, name

(b) What is the range (e.g., minimum and maximum) of each quantitative predictor?

Ans:

```
data[['mpg','cylinders','displacement','horsepower','weight','acceleration','year']].describe()
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	year
	9.0	3	68	46	1613	8.0	70
	46.6	8	455	230	5140	24.8	82

(c) What is the mean and standard deviation of each quantitative predictor?

Ans:

```
[ ] data[['mpg','cylinders','displacement','horsepower','weight','acceleration','year']].mean()
```

```
mpg              23.445918
cylinders         5.471939
displacement    194.411990
horsepower      104.469388
weight         2977.584184
acceleration     15.541327
year             75.979592
dtype: float64
```

```
data[['mpg','cylinders','displacement','horsepower','weight','acceleration','year']].std()
```

```
mpg              7.805007
cylinders        1.705783
displacement   104.644004
horsepower      38.491160
weight         849.402560
acceleration     2.758864
year             3.683737
dtype: float64
```

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset that remains?

Ans:

```
[8] data = data.drop(labels=range(10, 85), axis=0)
```

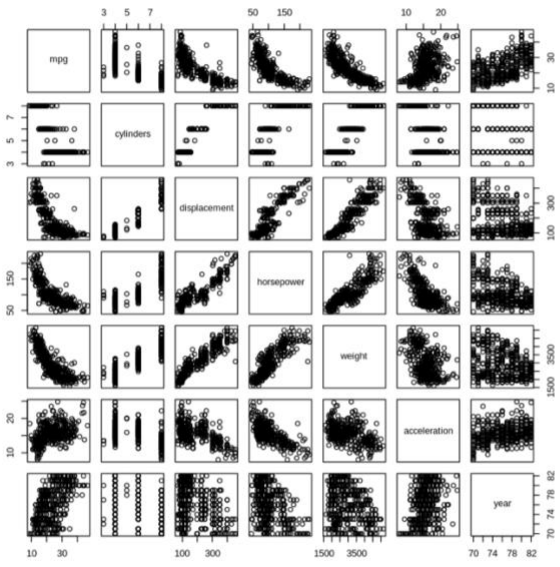
```
data[['mpg','cylinders','displacement','horsepower','weight','acceleration','year']].describe()
```

mean	24.404954	5.374613	187.566563	NaN	2937.111455	15.715480	77.145511	1.5
std	7.901449	1.656600	99.986169	NaN	809.745689	2.715394	3.138699	0.8
min	11.000000	3.000000	68.000000	NaN	1649.000000	8.500000	70.000000	1.0
25%	18.000000	4.000000	100.500000	NaN	2217.000000	14.000000	75.000000	1.0
50%	23.900000	4.000000	146.000000	NaN	2800.000000	15.500000	77.000000	1.0
75%	30.600000	6.000000	250.000000	NaN	3512.000000	17.300000	80.000000	2.0
max	46.600000	8.000000	455.000000	NaN	4997.000000	24.800000	82.000000	3.0

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

Ans: It has both linear and non-linear relationships. For example, year and mpg tend to have a linear relationship whereas horsepower and mpg tend to have a non-linear relationship.

```
sns.pairplot(data)
```



(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Ans: Yes, we see variables with both positive and negative relationships to the mpg result. For instance: year and mpg tend to have a positive relationship whereas horsepower and mpg tend to have a negative relationship.

2. (Total: 20 points) This exercise involves the Boston housing data set.

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

Ans: No. of rows: 506

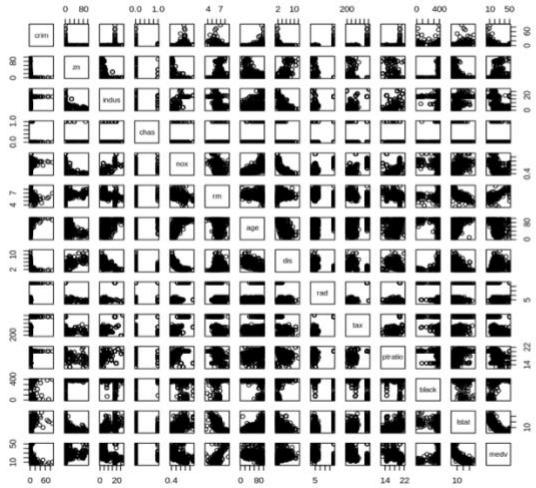
No. of columns: 14

Each row represent the set of predictor observations for a given Neighborhood in Boston. Each column represent each predictor variable for which an observation was made in 506 neighborhoods of Boston.

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings:

Ans:

```
sns.pairplot(data)
```



Not much can be discerned other than the fact that some variables appear to be correlated.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Ans: Based on the correlation coefficients and their corresponding p-values, there is indeed an association between the crime rate (crim) and the other predictors.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Ans:

```
data[['crim']].describe()
```

	crim
count	506.000000
mean	3.613524
std	8.601545
min	0.006320
25%	0.082045
50%	0.256510
75%	3.677082
max	88.976200

```
[18] data[['tax']].describe()
```

	tax
count	506.000000
mean	408.237154
std	168.537116
min	187.000000
25%	279.000000
50%	330.000000
75%	666.000000
max	711.000000

```
data[['ptratio']].describe()
```

	ptratio
count	506.000000
mean	18.455534
std	2.164946
min	12.600000
25%	17.400000
50%	19.050000
75%	20.200000
max	22.000000

```
[22] selection = data[data.crim>10].shape[0]
     selection/506
```

```
0.1067193675889328
```

11% of the neighborhood's crime rates above 10%

```
[23] selection = data[data.crim>50].shape[0]
     selection/506
```

```
0.007905138339920948
```

0.8% of the neighborhoods have crime rates above 10%

```
[24] selection = data[data.tax<600].shape[0]
     selection/506
```

```
0.7292490118577075
```

73% of the neighborhood pay under 600$

```
[25] selection = data[data.tax>600].shape[0]
     selection/506
```

```
0.2707509881422925
```

```
[ ]  26% of the neighborhood pay over 600$
```

- ⇨ Considering the median and maximum crime rate value which are 0.26% and 89% . There are few neighborhoods where crime rates are very high.
- ⇨ Based on the scatter plot there are few neighborhood where the tax rates are higher. The median and average tax amount are $330 and $408.20

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

```
selection = data[data.chas==1].shape[0]
selection
```

```
35
```

Ans:

35 datasets

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

```
[28] data[['ptratio']].describe()
```

	ptratio
count	506.000000
mean	18.455534
std	2.164946
min	12.600000
25%	17.400000
50%	19.050000
75%	20.200000
max	22.000000

Ans:

Median = 19.05

(g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors?

Ans:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.9	30.59	5

```
      crim                zn              indus            chas        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
      nox               rm             age              dis        
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax           ptratio          black       
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
     lstat            medv      
 Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 6.95   1st Qu.:17.02  
 Median :11.36   Median :21.20  
 Mean   :12.65   Mean   :22.53  
 3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :37.97   Max.   :50.00  
```

Suburb number **399**

- Crime is high when compared to other neighborhoods.
- Non-retail business acres is high when compared to most suburbs.
- This is one of the suburb that does not bound the Charles river.
- Nitrogen concentration is one of the highest.
- Average number of rooms per dwelling is lowest.
- Lowest weighted mean of distances to five Boston employment centers.
- Highest full value property tax.
- Highest pupil teacher ratio by town.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Ans:

```
[31] selection = data[data.rm>7].shape[0]
     selection

     64
```

There are 64 subrubs with more than 7 rooms per dwelling

```
[32] selection = data[data.rm>8].shape[0]
     selection

     13
```

```
There are 13 subrubswith more than 7 rooms per dwelling
```

3. (Total: 24 points)In this question, you should use the Carseats data set to predict the sales in a new store with Price=$120,Advertising=$10000, ShelveLoc = Good, 'Urban=Yes, US=Yes.

(a) Fit a multiple regression model to predict Sales using Price, Advertising Urban, and US. Write out the model in equation form, being careful to handle the qualitative variables properly.

Ans: Model in Equation Form Since we know general equation is described as:
$h\beta(x) = \beta 0 + \beta 1x$

In this case, Price=120, Advertising =10000, ShelveLoc = Good, Urban=Yes, US=Yes. The model can be written as follows: Sales = 13.0113 + -0.0546 * Price + 0.1203 * Advertising + -0.0388 * Urban_Yes+ 0.0585 * US_Yes the corresponding values:
Sales = 13.0113 + (-0.0546 * 120) + (0.1203 * 1) + (-0.0388 * 1) + (0.0585 * 1)Sales(in thousands of units) = 6.5993

(b) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!

```
[ ]  import statsmodels.api as sm
     from scipy import stats
     x = data[['Price', 'Advertising', 'Urban', 'US']]
     x2 = sm.add_constant(x_train)
```

Ans:

```
import statsmodels.api as sm
from scipy import stats
x2 = sm.add_constant(x)
est = sm.OLS(y.astype(float), x2.astype(float))
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.282
Model:                            OLS   Adj. R-squared:                  0.275
Method:                 Least Squares   F-statistic:                     38.77
Date:                Tue, 12 Oct 2021   Prob (F-statistic):           2.21e-27
Time:                        01:35:13   Log-Likelihood:                -916.11
No. Observations:                 400   AIC:                             1842.
Df Residuals:                     395   BIC:                             1862.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.0113      0.633     20.544      0.000      11.766      14.256
Price         -0.0546      0.005    -10.710      0.000      -0.065      -0.045
Advertising    0.1203      0.025      4.845      0.000       0.072       0.169
Urban         -0.0388      0.264     -0.147      0.883      -0.558       0.481
US             0.0585      0.345      0.170      0.865      -0.620       0.737
==============================================================================
Omnibus:                        1.094   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.579   Jarque-Bera (JB):                0.983
Skew:                           0.120   Prob(JB):                        0.612
Kurtosis:                       3.036   Cond. No.                         629.
==============================================================================
```

Observations:

- The coefficient related to Price is negative and increase of Price by 1 dollar will have an effect of decrease in average sales by 54.6 units.
- The coefficient related to advertising is positive and increase of Advertising budget by 1 dollar will have an effect of increase in sales by 120 units.
- The coefficient related to Urban is negative and if the store is present in urban location, it will have an effect in average sales decreasing by 38 units or we can say that the sales in urban stores are 38 units lesser than in the rural stores as an interpretation of this data.
- The coefficient related to US is positive and if the store is present in the US, it will have an effect in average by increase in 58 units.

(c) Using the model from (a), predict sales in the new store and calculate 68% and 95%confidence intervals.

Ans: We know that:
68% confidence interval for $\beta 1$: $\hat{\beta}1 \pm SE(\hat{\beta}1)$

95% confidence interval for β1: β^1 ± 2*SE(β^1)
From (a) sales = 6.5993
Standard error = 2.4
Therefore, 68% confidence interval: 6.5993 ± 2.4 = (8.9993, 4.1993)
95% confidence interval: 6.5993 ± 2*2.4 = (11.3993, 1.7993)

(d) Using the model from (a), what is the probability that sales will be greater than 12000units in the new store?

Ans: Probability (sales>12)=P(sales-7.682)/2.82
=P(12-7.682)/2.82
=P(sales> 2.379) = 1-P(sales<=2.379) = 0.057

Therefore there is 5.7% chance that sales will be more than 12000

(e) Using the model from (a), what is the probability that sales will be between 6000 and10000 units in the new store?

Ans: Probability(6 <= sales <= 10)
=P(6-7.682)/2.82<= P(sales-7.682)/2.82 <= P(10-7.682)/2.82
= 0.4325 <= P <= 0.2061
=0.3614
Therefore 36% chance that the data will fall in between 6000 and 10000

(f) For which of the predictors can you reject the null hypothesisH_0:β_j= 0?

Ans: If the p-value(p>|t|) is less than 0.05, we reject the null hypothesis. In this case, Urban and US have p-value greater than 0.05 and the null hypothesis cannot be rejected for these predictors.
For Advertising, the p-value is zero and hence the null hypothesis can be rejected.

(g) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. Using this model, predict sales in the new store and calculate 68% and 95% confidence intervals.

Ans:
68% confidence interval for β1: β^1 ± SE(β^1)
95% confidence interval for β1: β^1 ± 2*SE(β^1)
From (a) sales = 6.5993
Standard error = 2.4
Therefore, 68% confidence interval: 6.5993 ± 2.4 = (8.9993, 4.1993)
95% confidence interval: 6.5993 ± 2*2.4 = (11.3993, 1.7993)

```
[ ]  import statsmodels.api as sm
     from scipy import stats
     x = data[['Price', 'Advertising', 'Urban', 'US']]
     x2 = sm.add_constant(x_train)
```

```
import statsmodels.api as sm
from scipy import stats
x2 = sm.add_constant(x)
est = sm.OLS(y.astype(float), x2.astype(float))
est2 = est.fit()
print(est2.summary())
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.282
Model:	OLS	Adj. R-squared:	0.278
Method:	Least Squares	F-statistic:	77.91
Date:	Tue, 18 Feb 2020	Prob (F-statistic):	2.87e-29
Time:	21:52:57	Log-Likelihood:	-916.14
No. Observations:	400	AIC:	1838.
Df Residuals:	397	BIC:	1850.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.0034	0.607	21.428	0.000	11.810	14.196
Price	-0.0546	0.005	-10.755	0.000	-0.065	-0.045
Advertising	0.1231	0.018	6.809	0.000	0.088	0.159

Omnibus:	1.120	Durbin-Watson:	1.964
Prob(Omnibus):	0.571	Jarque-Bera (JB):	1.006
Skew:	0.121	Prob(JB):	0.605
Kurtosis:	3.037	Cond. No.	599.

(h) How well do the models in (a) and (g) fit the data?

Ans: Adjusted R-squared value compares the goodness of fit for a model and accounts for the factors that are not significant for the model. As per comparison of the two models the adjusted R-squared value of model(g) [0.275]> adjusted R-squared value of model (a) [0.278]. It seems that the first model is facing overfitting, hence the second model(g) fits the data better than (a)

4. (Total: 27 points) This problem involves the sales data set for Toyota Corolla, which can be found in the file ToyotaCorolla.csv. The data set contains 1436 observations on the following 10 variables.

   (a) Which of the predictors are quantitative, and which are qualitative?

   Ans: **Quantitative**: Price, Age, Milage, Displacement, Doors, Weight, Horsepower
   Qualitative: FuelType, MetColor, Automatic

   (b) What is the range (i.e., min and max) of each quantitative predictor?

   Ans:

```
data.describe()
```

	Price	Age	Mileage	Horsepower	MetColor	Automatic	Displacement	Doors	Weight
count	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000
mean	13199.266017	55.947075	42584.589136	101.502089	0.674791	0.055710	95.720752	4.033426	2364.442897
std	4461.162081	18.599988	23305.402480	14.981080	0.468616	0.229441	11.587120	0.952677	115.945453
min	5351.000000	1.000000	1.000000	69.000000	0.000000	0.000000	79.000000	2.000000	2205.000000
25%	10394.000000	44.000000	26719.000000	90.000000	0.000000	0.000000	85.000000	3.000000	2293.000000
50%	12177.000000	61.000000	39388.500000	110.000000	1.000000	0.000000	98.000000	4.000000	2359.000000
75%	14699.000000	70.000000	54072.000000	110.000000	1.000000	0.000000	98.000000	5.000000	2392.000000
max	39975.000000	80.000000	150993.000000	192.000000	1.000000	1.000000	122.000000	5.000000	3560.000000

(c) What is the mean and standard deviation of each quantitative predictor?

Ans:

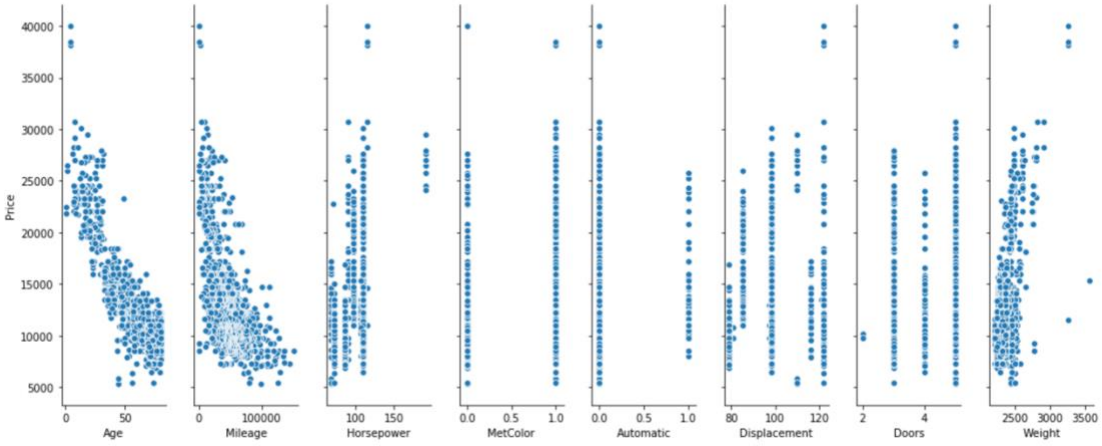
```
data.describe()
```

	Price	Age	Mileage	Horsepower	MetColor	Automatic	Displacement	Doors	Weight
count	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000	1436.000000
mean	13199.266017	55.947075	42584.589136	101.502089	0.674791	0.055710	95.720752	4.033426	2364.442897
std	4461.162081	18.599988	23305.402480	14.981080	0.468616	0.229441	11.587120	0.952677	115.945453
min	5351.000000	1.000000	1.000000	69.000000	0.000000	0.000000	79.000000	2.000000	2205.000000
25%	10394.000000	44.000000	26719.000000	90.000000	0.000000	0.000000	85.000000	3.000000	2293.000000
50%	12177.000000	61.000000	39388.500000	110.000000	1.000000	0.000000	98.000000	4.000000	2359.000000
75%	14699.000000	70.000000	54072.000000	110.000000	1.000000	0.000000	98.000000	5.000000	2392.000000
max	39975.000000	80.000000	150993.000000	192.000000	1.000000	1.000000	122.000000	5.000000	3560.000000

(d) Investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

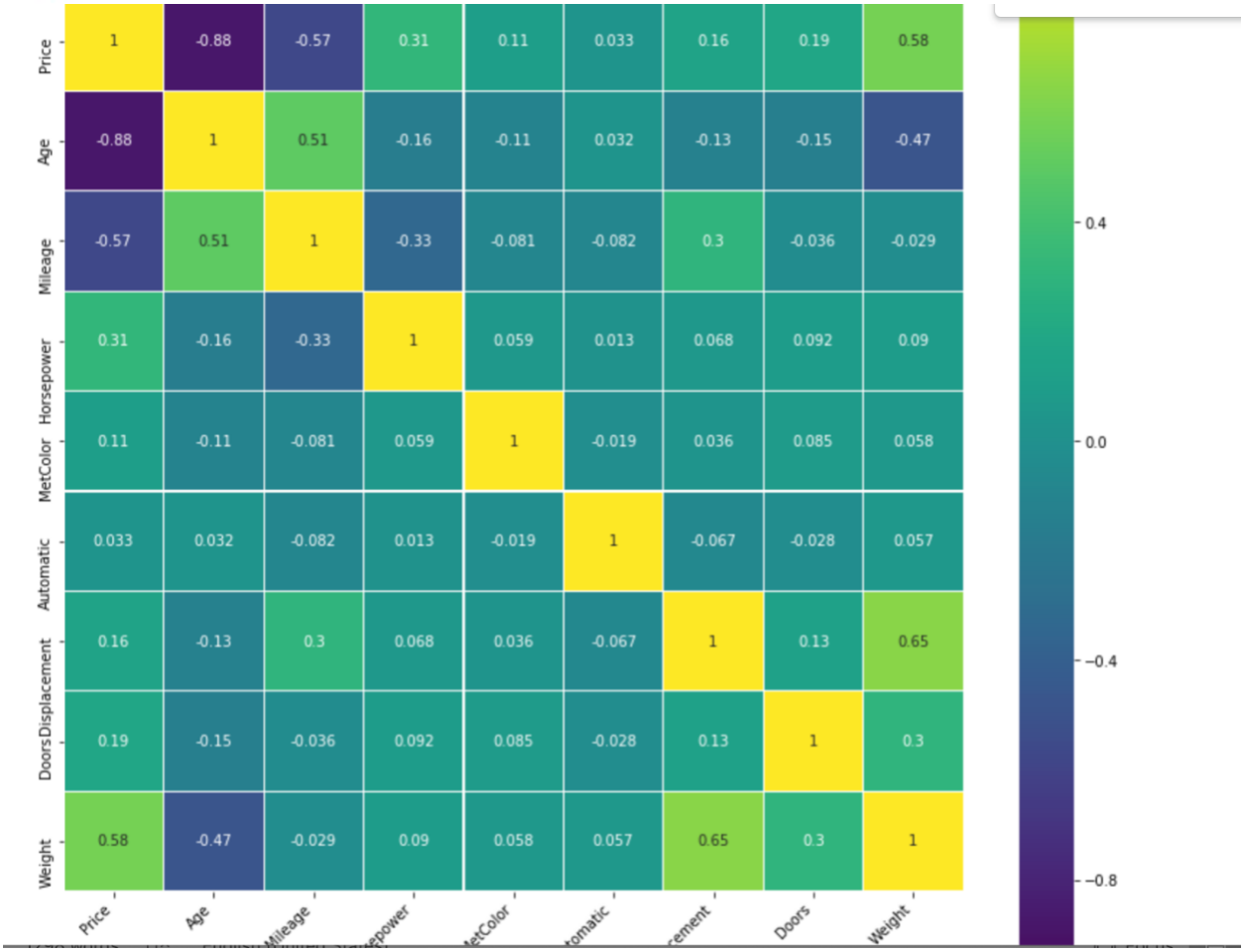
Ans:

```
In [46]: sb.pairplot(data, x_vars=['Age', 'Mileage', 'Horsepower', 'MetColor', 'Automatic','Displacement', 'Doors', 'Weight'], y_vars
         ='Price', height=6, aspect=0.3, diag_kind="reg")
Out[46]: <seaborn.axisgrid.PairGrid at 0x1e0981beb08>
```



The figure above shows the pair plot with predictors such as ‘Age’, ‘Milage’, ‘Horsepower’, ‘MetColor’, ‘Automatic’, ‘Displacement’, ‘Doors’, ‘Weight’ against ‘Price’.

```
In [45]: plt.figure(figsize=(14,14))
ax = sb.heatmap(df.corr(), vmin=-1, vmax=1, center=0, annot=True, linewidths=0.1, cmap="viridis", square="True")
ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
ax.set_ylim(9, 0)

Out[45]: (9, 0)
```



The darker shade of blue shows less correlation and the light-yellow shows high correlation, as per the heatmap light green shows slight correlation between predictors.

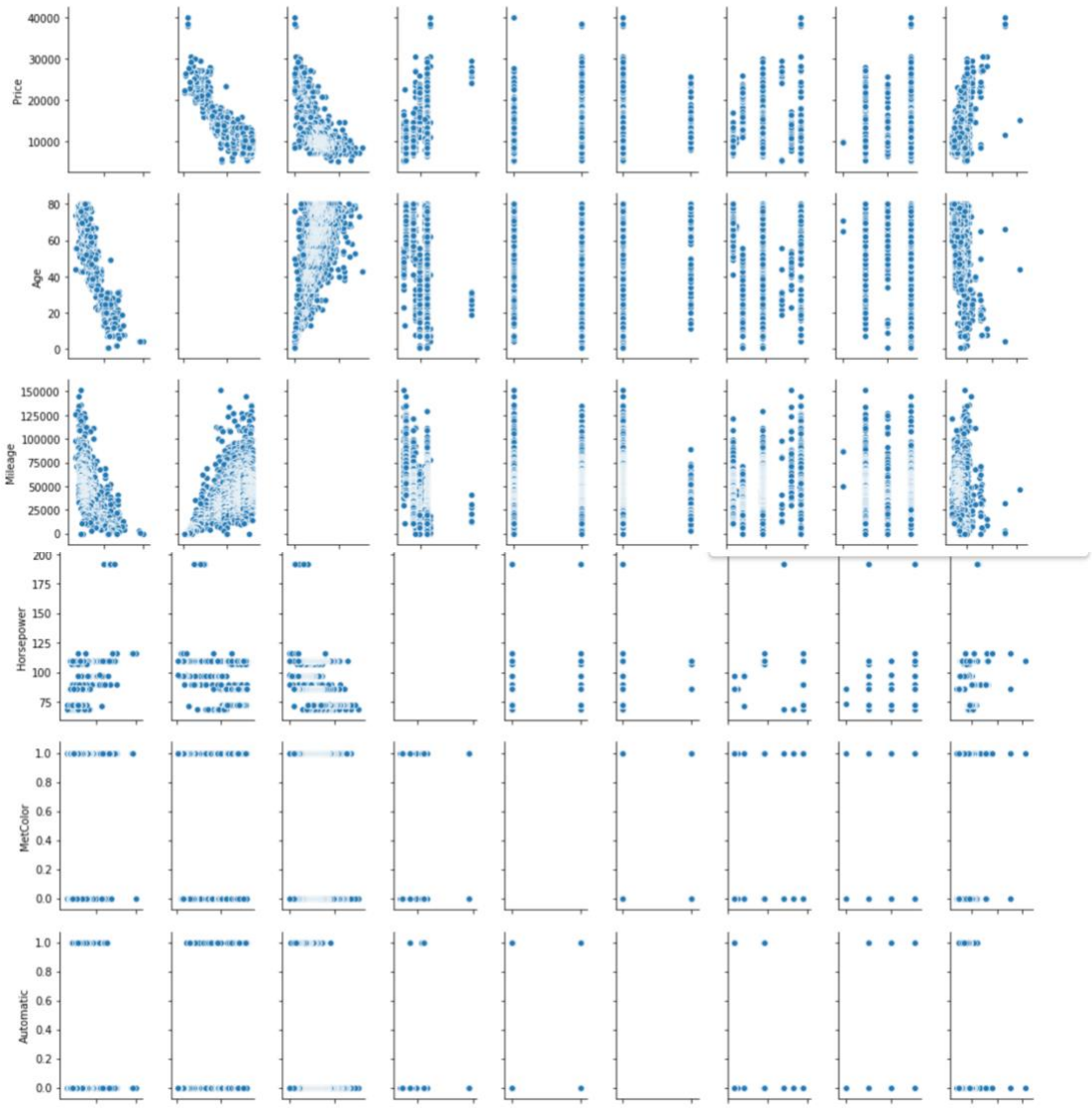
The correlation between factors such as age and mileage is quite significant, similar is the correlation between weight and price.

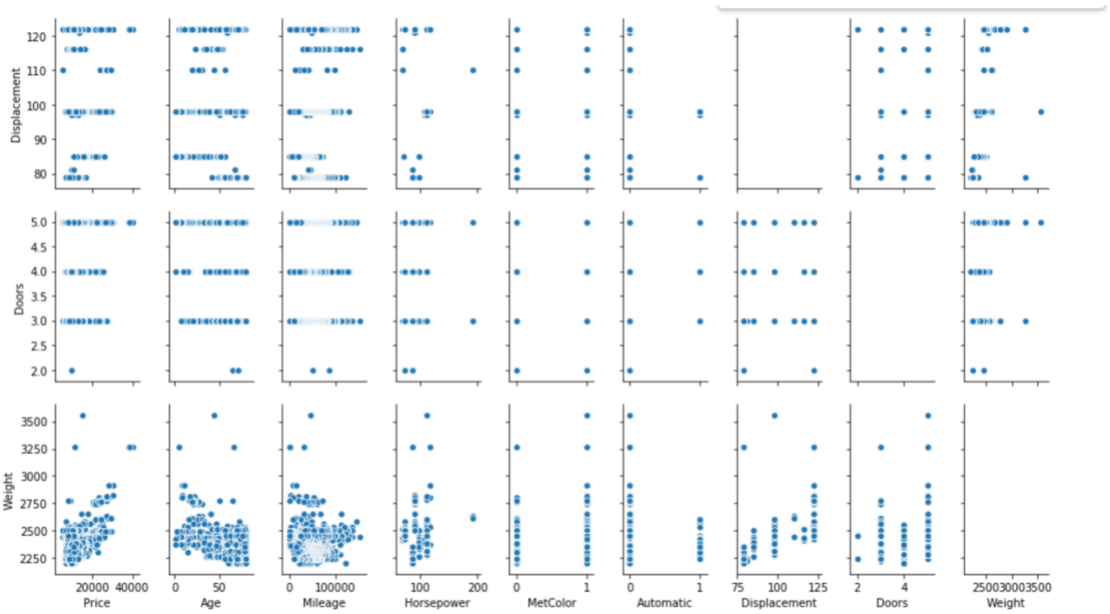
There is negative correlation between predictors such as age and price from heatmap.

The figure below shows the pair plot for the predictors with all predictors included.

In [47]:
```
sb.pairplot(data, diag_kind="scatter", aspect=0.6)
```

Out[47]: <seaborn.axisgrid.PairGrid at 0x1e09a2c75c8>





(e) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

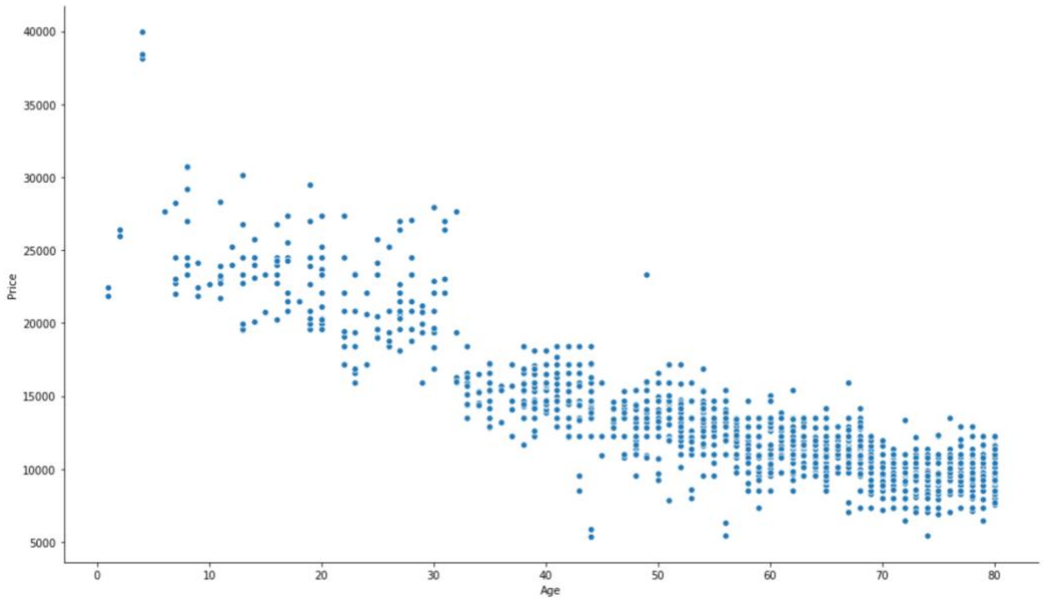
Ans: Based on the heatmap and the pair plot above it is evident that milage is directly proportional to age, higher the age, higher the milage. Horsepower or price increase result in decrease in milage.

(f) Fit a simple linear regression with Price as the response and Age as the predictor.

1. Is there a relationship between the predictor and the response?

Ans:

```
In [53]: sb.pairplot(data, x_vars=['Age'], y_vars='Price', height=8, aspect=1.7, diag_kind="reg")
Out[53]: <seaborn.axisgrid.PairGrid at 0x1e09d2b5f08>
```



Model equation: Price = 24962 + (-210.2481*Age)

In [86]:
```
import statsmodels.api as sm
from scipy import stats
x2 = sm.add_constant(x)
est = sm.OLS(y, x2)
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.768
Model:                            OLS   Adj. R-squared:                  0.768
Method:                 Least Squares   F-statistic:                     4758.
Date:                Wed, 19 Feb 2020   Prob (F-statistic):               0.00
Time:                        22:44:35   Log-Likelihood:                -13054.
No. Observations:                1436   AIC:                         2.611e+04
Df Residuals:                    1434   BIC:                         2.612e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.496e+04    179.699    138.910      0.000    2.46e+04    2.53e+04
Age         -210.2481      3.048    -68.978      0.000    -216.227    -204.269
==============================================================================
Omnibus:                      359.248   Durbin-Watson:                   1.214
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2773.781
Skew:                           0.946   Prob(JB):                         0.00
Kurtosis:                       9.540   Cond. No.                         187.
==============================================================================
```

Since the p-value(p>|t|) is zero for the predictor age, the null hypothesis can be rejected, and we can observe a relationship between predictor and response.

2. How strong is the relationship between the predictor and the response?

Ans: Since R-squared is the measure of how closely the data is fitted to the regression line, in this case, the R-squared value can help reveal the strength of relationship between the predictor and response. The R-squared value is 0.768, which signifies that 76.8% of the change in price can be described by the age which is strong in comparison to other predictors.

3. What is the predicted price associated for a car with an age of 48 months? What are the associated 95% confidence intervals?

Ans: Price = 24962 + (-210.2481*Age)
Price = 24962 + (-210.2481*48)
Price = 14870.0912 in dollars
Standard Deviation: 2147
95% confidence interval: 24962 $\pm$ 2*2147

(g) Fit a multiple linear regression with Price as the response and all other variables the predictors.

1. Is there a relationship between the predictors and the response?

Ans:

```
In [72]: import statsmodels.api as sm
         from scipy import stats
         x2 = sm.add_constant(x_encoded)
         est = sm.OLS(y, x2)
         est2 = est.fit()
         print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     935.7
Date:                Wed, 19 Feb 2020   Prob (F-statistic):               0.00
Time:                        22:43:15   Log-Likelihood:                -12635.
No. Observations:                1435   AIC:                         2.529e+04
Df Residuals:                    1424   BIC:                         2.535e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -5007.2497   1592.543     -3.144      0.002   -8131.231   -1883.268
Age             -150.0655      3.204    -46.840      0.000    -156.350    -143.781
Mileage           -0.0320      0.003    -12.305      0.000      -0.037      -0.027
Horsepower        73.0285      7.048     10.362      0.000      59.203      86.854
MetColor          68.8163     92.215      0.746      0.456    -112.076     249.709
Automatic        409.5908    193.300      2.119      0.034      30.408     788.774
Displacement     -78.0478     10.504     -7.430      0.000     -98.653     -57.443
Doors             -9.5930     49.306     -0.195      0.846    -106.312      87.127
Weight            11.1450      0.672     16.577      0.000       9.826      12.464
FuelType_Diesel 3937.1244    627.355      6.276      0.000    2706.485    5167.764
FuelType_Petrol 1380.3624    408.972      3.375      0.001     578.110    2182.615
==============================================================================
Omnibus:                      269.752   Durbin-Watson:                   1.619
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2763.003
Skew:                          -0.556   Prob(JB):                         0.00
Kurtosis:                       9.706   Cond. No.                     1.83e+06
==============================================================================
```

The p-value(p>|t|) is zero for all predictors except Doors and MetColor, so we can't reject the null hypothesis for these two predictors, but all the other factors exhibit significant relationship to the response(price).

2. How strong is the relationship between the predictors and the response?

Ans: Since R-squared is the measure of how closely the data is fitted to the regression line, in this case, the R-squared value can help reveal the strength of relationship between the predictor and response. The R-squared value is 0.868, which signifies that 86.8% of the change in price can be described by the age which is strong in comparison to other predictors.

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

Ans: The predictors which show significant relationship to the response are:
Age, Mileage, Horsepower, Automatic, Displacement, Weight, FuelType

4. What does the coefficient for the age variable suggest? How accurate can you estimate the effect of age on price?

Ans:

```
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -5007.2497   1592.543     -3.144      0.002   -8131.231   -1883.268
Age             -150.0655      3.204    -46.840      0.000    -156.350    -143.781
```

We observe that with an increase in Age by 1 month, there is a decrease in price by 150 dollars. The accuracy of the estimation can be deducted by the R-squared value, which is 0.768.

5. What is the predicted price associated for a car with a mileage of 45000 miles, 48 months, diesel, automatic transmission, 4 doors, 2568 pounds, a displacement of 122 cu. inches, a horsepowerof 90, and non-metallic color? What are the associated 95% confidence intervals?

Ans: Price =-5007.2497 + (-150.0655* Age) + (-0.0320 * Mileage) + 73.0285 * Horsepower + 68.8163 * MetColor + 409.5908 * Automatic + (-78.0478 * Displacement) + (-9.5930 * Doors) + (11.1450 * Weight) + (3937.1244 * FuelType_Diesel)
Therefore, Price = 17625.0429 dollars
Standard Deviation = 1618.8560
95% confidence interval: 17625.0429 +-2*1618.856

6. Which predictors matter most for predicting the price for a car? (Find the first and the second most important variables)

Ans: The first most important value is Age as we saw a high R-squared value for a single predictor, also observed the relation from heatmap and pair plot.

The second most important variable is Weight with a positive correlation as observed from the heatmap.

5. This problem involves the Boston data set. We want to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

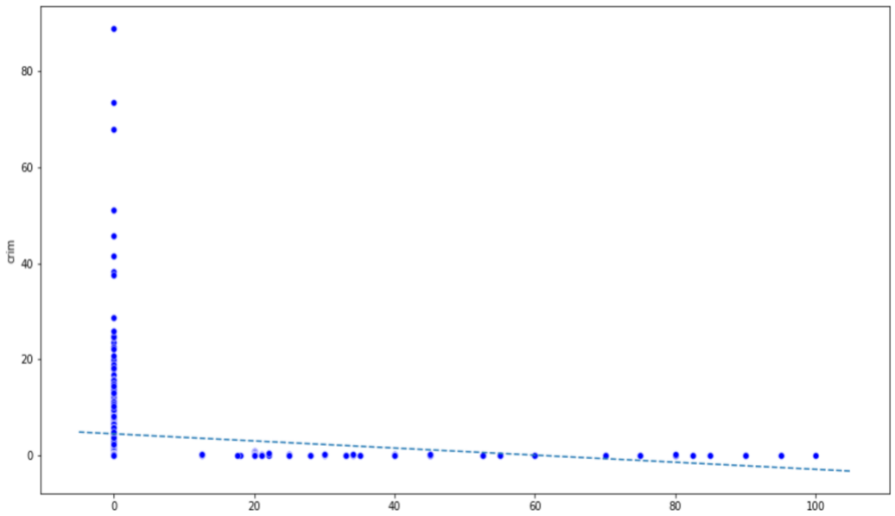
(a) (3 points) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

Ans:

```
[12] import statsmodels.api as sm

     X = data[['zn']]
     X = sm.add_constant(X, prepend=True)
     y = data['crim']

     model = sm.OLS(y, X)
     result = model.fit()
     print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.429  -4.222  -2.620   1.250  84.523

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="zn", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = 4.4537 -0.0739 * x_vals
plt.plot(x_vals, y_vals, '--')
```

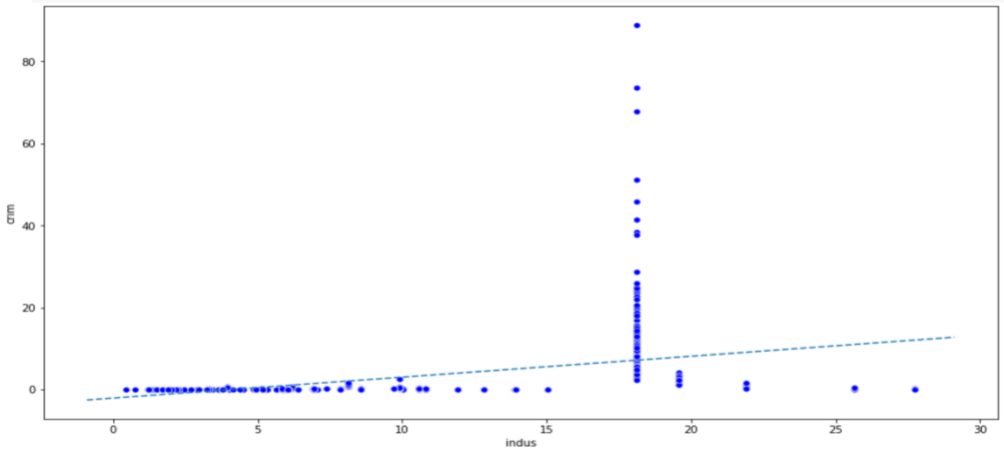


- From the above analysis, we can say that there is a low negative relationship between “crime” and “zn”

```
X = data[['indus']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-11.972  -2.698  -0.736   0.712  81.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06374    0.66723  -3.093  0.00209 **
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="indus", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -2.063 +0.5098 * x_vals
plt.plot(x_vals, y_vals, '--')
```

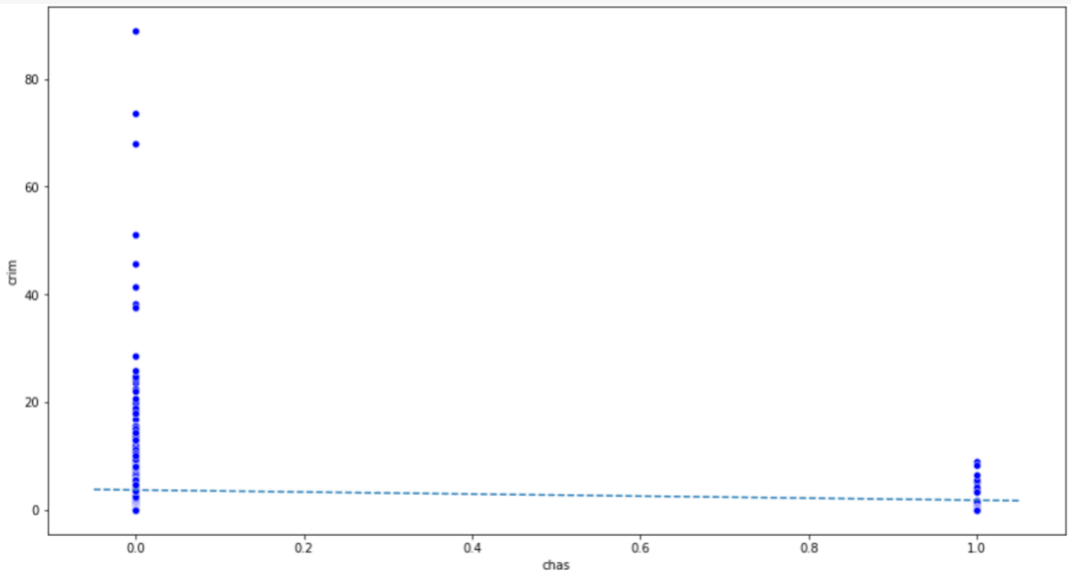


Its seen from the graph that there is a slightly positive relationship between the “crime” and “indus”

```
X = data[['chas']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.738  -3.661  -3.435   0.018  85.232

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.7444     0.3961   9.453   <2e-16 ***
chas          -1.8928     1.5061  -1.257    0.209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="chas", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = 3.7444  -1.8928 * x_vals
plt.plot(x_vals, y_vals, '--')
```

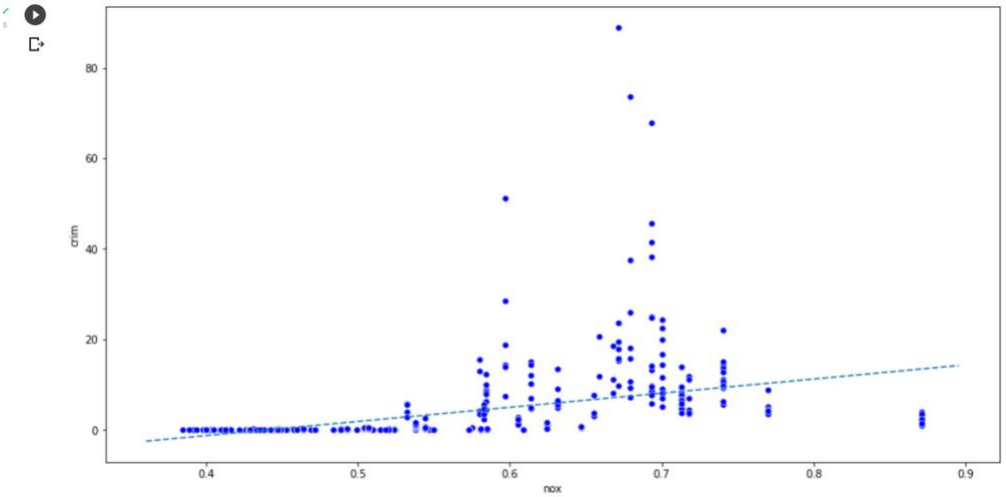


From the above graph we can say that changes in “chas” is not accompanied by an increase in the crime rate. Therefore there is no relationship between them

```
X = data[['nox']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.371   -2.738   -0.974    0.559   81.728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
nox           31.249      2.999  10.419  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared:  0.1772, Adjusted R-squared:  0.1756
F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="nox", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -13.7199 + 31.2485 * x_vals
plt.plot(x_vals, y_vals, '--')
```



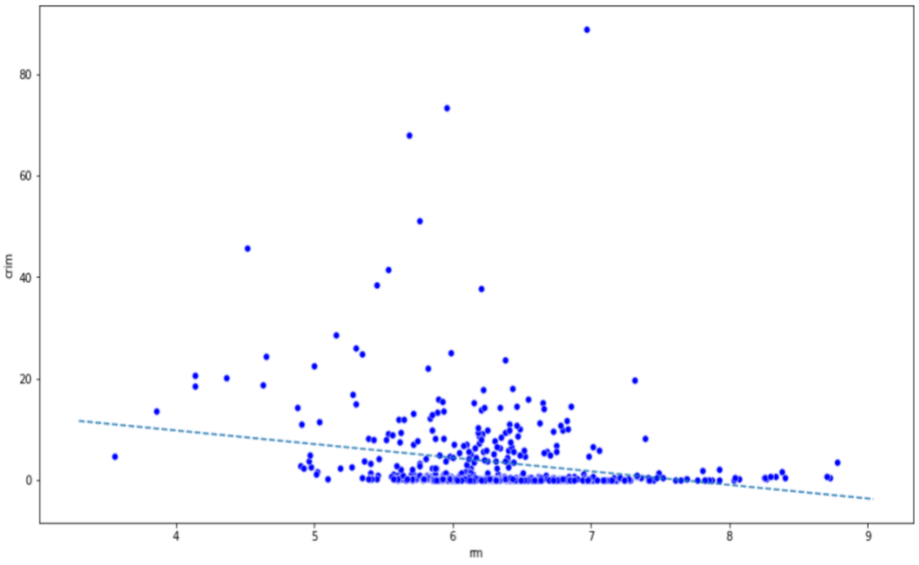
The above graph confirms that there exists a slightly lower positive correlation between crime and nitrogen oxides concentration

```
X = data[['rm']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
   Min     1Q Median     3Q    Max
-6.604 -3.952 -2.654  0.989 87.197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.482      3.365   6.088 2.27e-09 ***
rm            -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="rm", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = 20.4818 + -2.6841 * x_vals
plt.plot(x_vals, y_vals, '--')
```

```
[<matplotlib.lines.Line2D at 0x7f6a4e784910>]
```

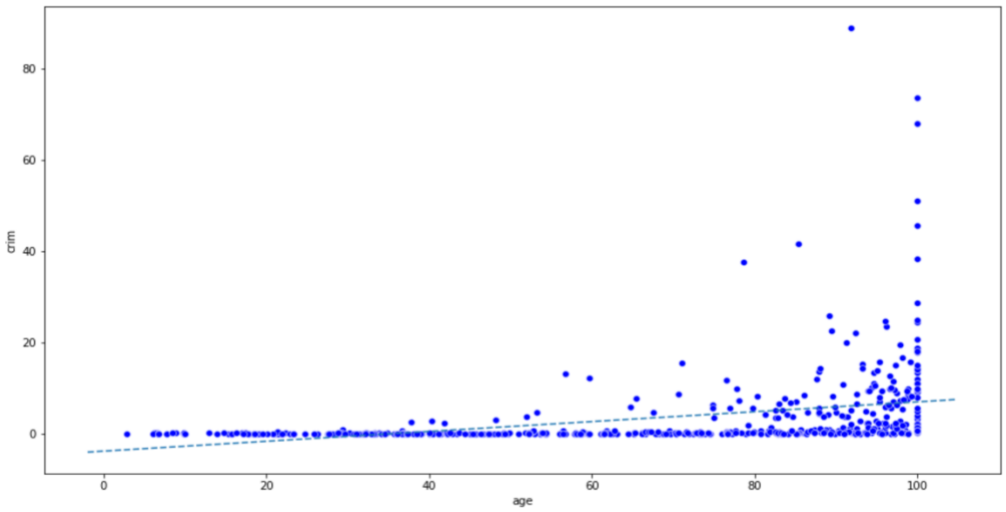


The graph shows the relationship between the crime rate and rm is defined by low sloping value.

```
X = data[['age']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
   Min      1Q  Median      3Q     Max
-6.789  -4.257  -1.230   1.527  82.849

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
age          0.10779    0.01274   8.463 2.85e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.057 on 504 degrees of freedom
Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="age", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -3.7779 + 0.1078 * x_vals
plt.plot(x_vals, y_vals, '--')
```

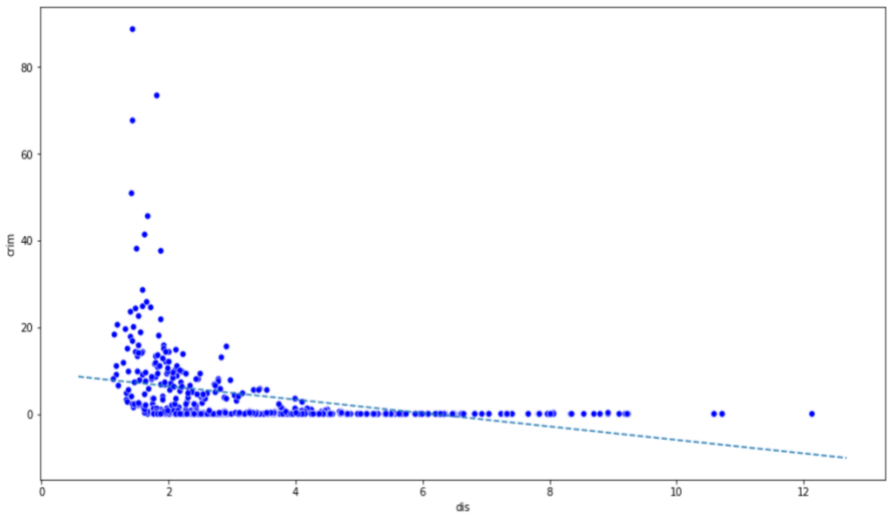


From the figure we can say there exists a low positive relationship between crime and age

```
X = data[['dis']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.708  -4.134  -1.527   1.516  81.674

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.4993     0.7304  13.006   <2e-16 ***
dis           -1.5509     0.1683  -9.213   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,	Adjusted R-squared:  0.1425
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="dis", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = 9.4993 -1.5509 * x_vals
plt.plot(x_vals, y_vals, '--')
```

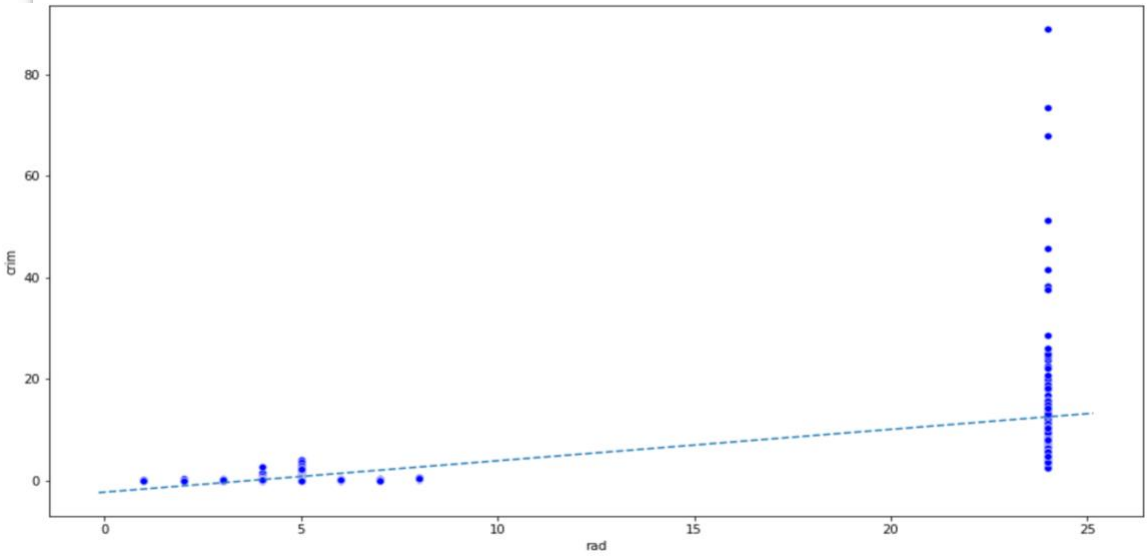


There is a low sloping relation between crime and dis

```
X = data[['rad']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.164   -1.381   -0.141    0.660   76.433

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
rad          0.61791    0.03433  17.998  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared:  0.3913, Adjusted R-squared:  0.39
F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="rad", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -2.2872 + 0.6179 * x_vals
plt.plot(x_vals, y_vals, '--')
```

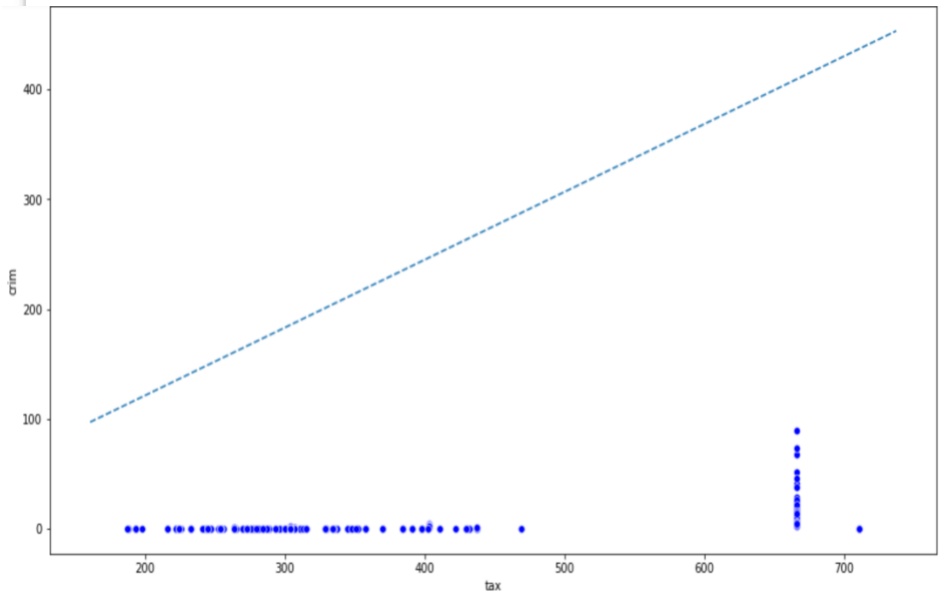


The graph confirms the existence of statistically significant relationship between crime and rad

```
X = data[['tax']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.513   -2.738   -0.194    1.065   77.696

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
tax          0.029742   0.001847   16.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="tax", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -2.2872 + 0.6179 * x_vals
plt.plot(x_vals, y_vals, '--')
```

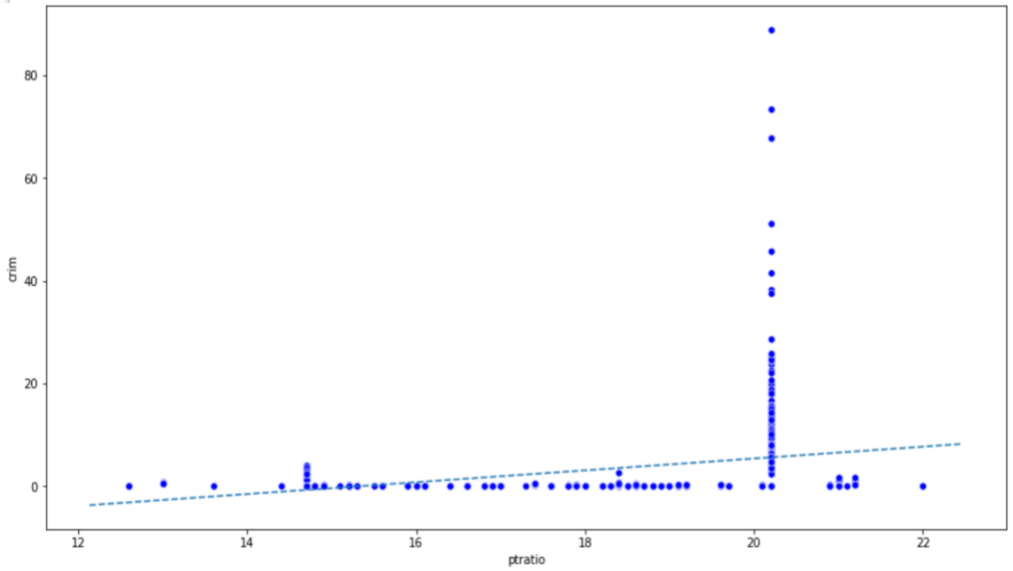


The graph above shows that there is a positive relationship between crime and tax

```
X = data[['ptratio']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
   Min      1Q  Median      3Q     Max
-7.654  -3.985  -1.912   1.825  83.353

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="ptratio", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -17.6469  + 1.1520 * x_vals
plt.plot(x_vals, y_vals, '--')
```

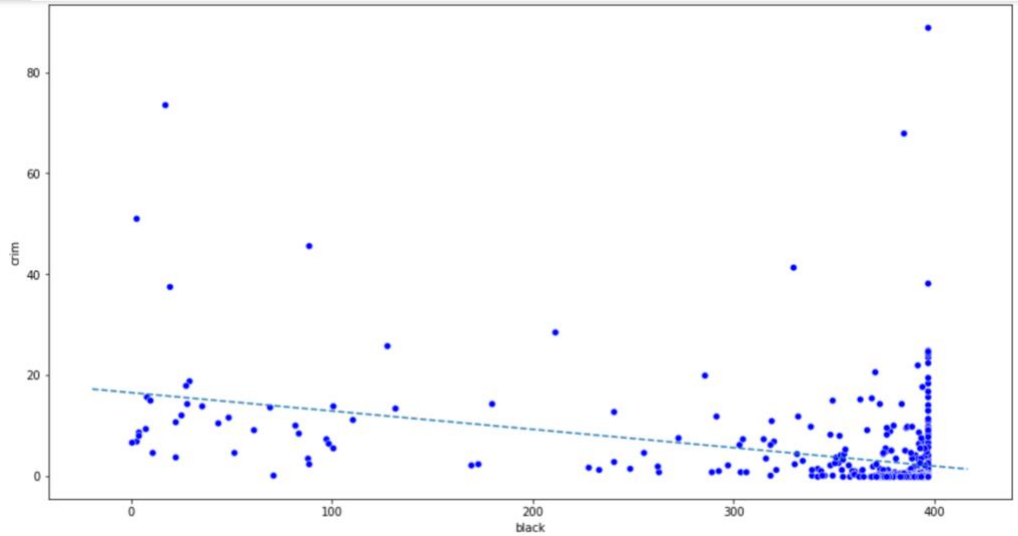


There is low positive correlation between ptratio and crime

```
X = data[['black']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.756  -2.299  -2.095  -1.296  86.822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.553529   1.425903  11.609   <2e-16 ***
black       -0.036280   0.003873  -9.367   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483, Adjusted R-squared:  0.1466
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="black", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = 16.5535  -0.0363 * x_vals
plt.plot(x_vals, y_vals, '--')
```

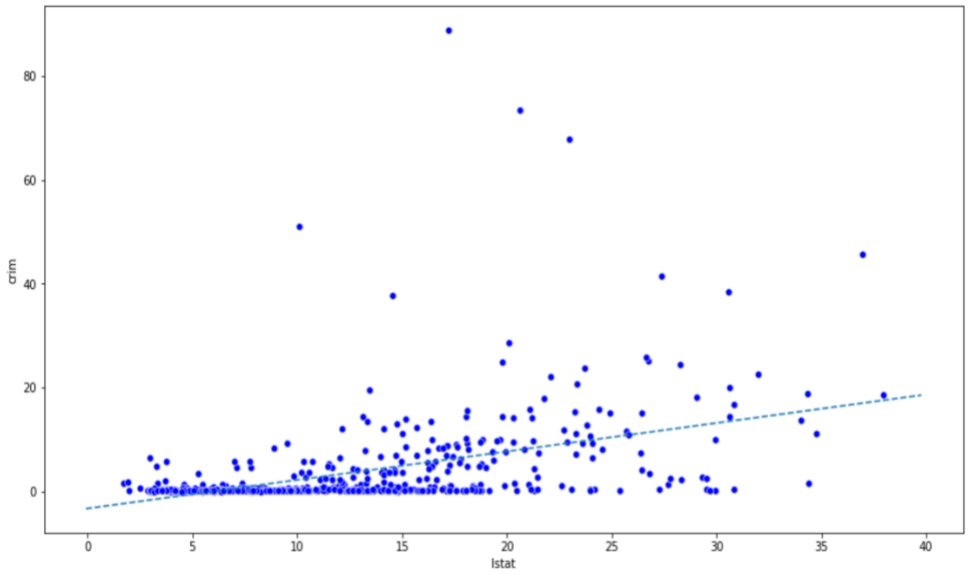


The graph confirms the statistical significant relationship between crime and black

```
X = data[['lstat']]
X = sm.add_constant(X, prepend=True)
y = data['crim']
model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.925  -2.822  -0.664   1.079  82.862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
lstat        0.54880    0.04776  11.491  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="lstat", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = -3.3305  + 0.5488 * x_vals
plt.plot(x_vals, y_vals, '--')
```

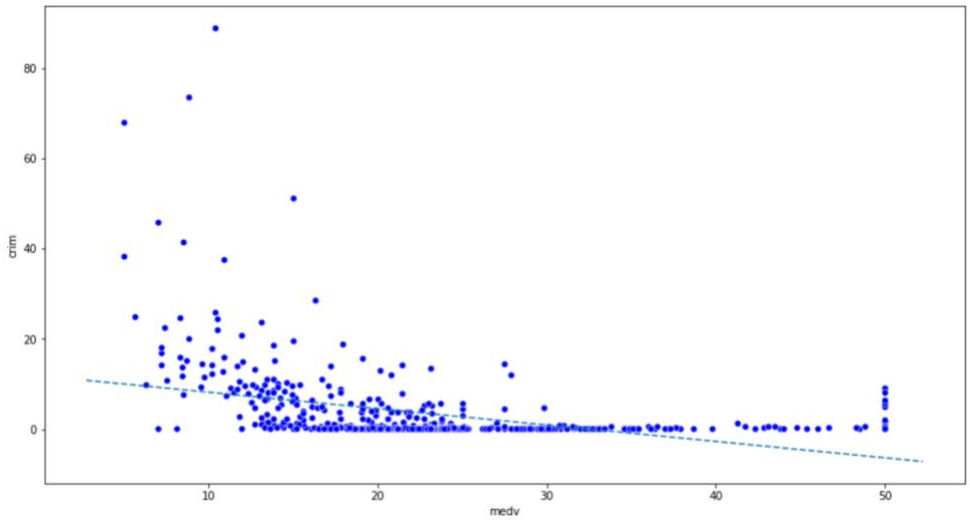


The graph shows that there is a statistical significant positive relation between crime and lstat

```
X = data[['medv']]
X = sm.add_constant(X, prepend=True)
y = data['crim']
model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.071 -4.022 -2.343   1.298 80.957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79654    0.93419   12.63   <2e-16 ***
medv        -0.36316    0.03839   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
fig = plt.figure(figsize=(15,8))
ax = fig.add_subplot(111)

ax = sns.scatterplot(x="medv", y="crim", color='b', data=data)

x_vals = np.array(ax.get_xlim())
y_vals = 11.7965  -0.3632 * x_vals
plt.plot(x_vals, y_vals, '--')
```



The relationship between medv and crime is slightly sloping as shown

In conclusion there is a statistically significant relationship between predictors and the response for every variable except chas The R squared and adjusted R squared values are very low for all the above models, and therefore these predictors describe a small amount of variation in crime response.

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0: \beta_j = 0$?

Ans:

```
X = data[['zn', 'indus', 'chas', 'nox','rm','age','dis','rad','tax','ptratio','black','lstat','medv']]
X = sm.add_constant(X, prepend=True)
y = data['crim']

model = sm.OLS(y, X)
result = model.fit()
print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.924 -2.120  -0.353   1.019  75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

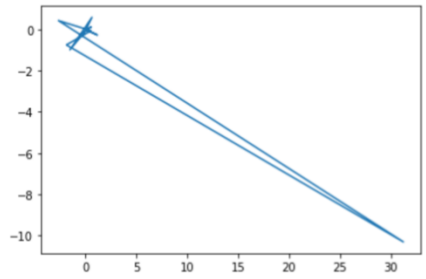
From the result obtained we can reject the null hypothesis of “zn”, “dis”, “rad”, “black”, “medv”. The multiple regression model generally does not fit Boston dataset very well because the R squared value is 0.45 and the adjusted R squared value is 0.43

(c) (4 points) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on they-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on they-axis.

Ans:

```
[58] simple_reg = [-0.07393, 0.50978, -1.8928, 31.249, -2.684, 0.10779, -1.5509, 0.61791, 0.0297, 1.1520, -0.0362, 0.55, -0.36]
     multiple_reg = [0.044855, -0.0638, -0.75, -10.313, 0.430, 0.0014, -0.987, 0.588, -0.003, -0.271, -0.007, 0.126, -0.198]
     plt.plot(simple_reg, multiple_reg)
```

```
[<matplotlib.lines.Line2D at 0x7f6a4d55a210>]
```



Univariate coefficients and multiple coefficients have distinct difference. This is because the slope of simple regression model represents the average effect of an increase in predictor ignoring the other predictors in the dataset. But the multiple regression holds other predictors fixed and its slope represents the average effect of an increase in the predictor.

(d) 4 points) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor x, fit a model of the form $y=\beta_0+\beta_1x+\beta_2x_2+\beta_3x_3+$

```
[60] def poly(x, p):
         x = np.array(x)
         X = np.transpose(np.vstack((x**k for k in range(p+1))))
         return np.linalg.qr(X)[0][:,1:]
     X = sm.add_constant(poly(data['zn'],3), prepend=True)


     Y = data['crim']

     model = sm.OLS(Y, X)
     result = model.fit()
     print(result.summary())
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.821  -4.614  -1.294   0.473  84.130

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.846e+00  4.330e-01  11.192  < 2e-16 ***
zn          -3.322e-01  1.098e-01  -3.025  0.00261 **
I(zn^2)      6.483e-03  3.861e-03   1.679  0.09375 .
I(zn^3)     -3.776e-05  3.139e-05  -1.203  0.22954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.372 on 502 degrees of freedom
Multiple R-squared:  0.05824,    Adjusted R-squared:  0.05261
F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
```

```
Residuals:
   Min     1Q Median     3Q    Max
-8.278 -2.514  0.054  0.764 79.713

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.6625683  1.5739833   2.327   0.0204 *
indus       -1.9652129  0.4819901  -4.077 5.30e-05 ***
I(indus^2)   0.2519373  0.0393221   6.407 3.42e-10 ***
I(indus^3)  -0.0069760  0.0009567  -7.292 1.20e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.423 on 502 degrees of freedom
Multiple R-squared:  0.2597, Adjusted R-squared:  0.2552
F-statistic: 58.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
   Min     1Q Median     3Q    Max
-9.110 -2.068 -0.255  0.739 78.302

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   233.09      33.64   6.928 1.31e-11 ***
nox         -1279.37     170.40  -7.508 2.76e-13 ***
I(nox^2)     2248.54     279.90   8.033 6.81e-15 ***
I(nox^3)    -1245.70     149.28  -8.345 6.96e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.234 on 502 degrees of freedom
Multiple R-squared:  0.297,  Adjusted R-squared:  0.2928
F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
   Min     1Q Median     3Q    Max
-9.762 -2.673 -0.516  0.019 82.842

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.549e+00  2.769e+00  -0.920  0.35780
age          2.737e-01  1.864e-01   1.468  0.14266
I(age^2)    -7.230e-03  3.637e-03  -1.988  0.04738 *
I(age^3)     5.745e-05  2.109e-05   2.724  0.00668 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.84 on 502 degrees of freedom
Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
F-statistic: 35.31 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.757  -2.588   0.031   1.267  76.378

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.0476     2.4459  12.285  < 2e-16 ***
dis         -15.5543     1.7360  -8.960  < 2e-16 ***
I(dis^2)      2.4521     0.3464   7.078 4.94e-12 ***
I(dis^3)     -0.1186     0.0204  -5.814 1.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.331 on 502 degrees of freedom
Multiple R-squared:  0.2778, Adjusted R-squared:  0.2735
F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.381  -0.412  -0.269   0.179  76.217

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.605545   2.050108  -0.295    0.768
rad          0.512736   1.043597   0.491    0.623
I(rad^2)    -0.075177   0.148543  -0.506    0.613
I(rad^3)     0.003209   0.004564   0.703    0.482

Residual standard error: 6.682 on 502 degrees of freedom
Multiple R-squared:    0.4,  Adjusted R-squared:  0.3965
F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.273  -1.389   0.046   0.536  76.950

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.918e+01  1.180e+01   1.626    0.105
tax         -1.533e-01  9.568e-02  -1.602    0.110
I(tax^2)     3.608e-04  2.425e-04   1.488    0.137
I(tax^3)    -2.204e-07  1.889e-07  -1.167    0.244

Residual standard error: 6.854 on 502 degrees of freedom
Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
F-statistic:  97.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
   Min     1Q Median     3Q    Max
-6.833 -4.146 -1.655  1.408 82.697

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  477.18405  156.79498   3.043  0.00246 **
ptratio      -82.36054   27.64394  -2.979  0.00303 **
I(ptratio^2)   4.63535    1.60832   2.882  0.00412 **
I(ptratio^3)  -0.08476    0.03090  -2.743  0.00630 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.122 on 502 degrees of freedom
Multiple R-squared:  0.1138, Adjusted R-squared:  0.1085
F-statistic: 21.48 on 3 and 502 DF,  p-value: 4.171e-13
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.096  -2.343  -2.128  -1.439  86.790

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.826e+01  2.305e+00   7.924  1.5e-14 ***
black       -8.356e-02  5.633e-02  -1.483    0.139
I(black^2)   2.137e-04  2.984e-04   0.716    0.474
I(black^3)  -2.652e-07  4.364e-07  -0.608    0.544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.955 on 502 degrees of freedom
Multiple R-squared:  0.1498, Adjusted R-squared:  0.1448
F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16
Residuals:
    Min      1Q  Median      3Q     Max
-15.234  -2.151  -0.486   0.066  83.353

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2009656  2.0286452   0.592   0.5541
lstat       -0.4490656  0.4648911  -0.966   0.3345
I(lstat^2)   0.0557794  0.0301156   1.852   0.0646 .
I(lstat^3)  -0.0008574  0.0005652  -1.517   0.1299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.629 on 502 degrees of freedom
Multiple R-squared:  0.2179, Adjusted R-squared:  0.2133
F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-24.427  -1.976  -0.437   0.439  73.655

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.1655381  3.3563105  15.840  < 2e-16 ***
medv        -5.0948305  0.4338321 -11.744  < 2e-16 ***
I(medv^2)    0.1554965  0.0171904   9.046  < 2e-16 ***
I(medv^3)   -0.0014901  0.0002038  -7.312 1.05e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.569 on 502 degrees of freedom
Multiple R-squared:  0.4202, Adjusted R-squared:  0.4167
F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16

Residuals:
   Min     1Q Median     3Q    Max
-3.738 -3.661 -3.435  0.018 85.232

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7444     0.3961   9.453   <2e-16 ***
chas         -1.8928     1.5061  -1.257    0.209
I(chas^2)         NA         NA      NA       NA
I(chas^3)         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124,   Adjusted R-squared:  0.001146
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

Predictor	Statistically Significant	
	I(predictor)^2	I(predictor)^3
zn	✗	✗
Indus	✓	✓
nox	✓	✓
rm	✗	✗
age	✓	✓
dis	✓	✓
rad	✗	✗
tax	✗	✗
ptratio	✓	✓
black	✗	✗
lstat	✗	✗
medv	✓	✓
chas	NA	NA

The table shows whether there is evidence of non-linear relationship between crime and predictors.