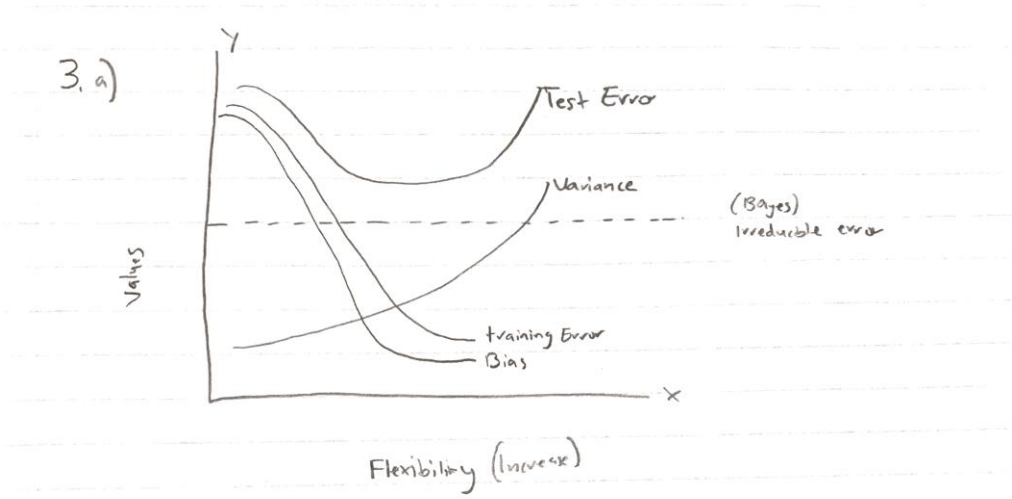
Siddhant Rai

Chapter 2

3.

a)



b) Bias and variance are inversely related in this meaning as flexibility increases , these is an increase in variance, and bias decreases. This is because Bias is used to simplify a complex regression while Variance estimates how close our training data set is to the model how our estimated f^(x) of we change a single point on the data set. For training vs test error, the training data decreases, and test initially decrease before increases again as flexibility increases. When we are given a small training & large test dataset, we to the overfit the model because the patterns in the train may not be in the test. And with the irreducible error(Bayes), it is horizontal and constant, and the test error never crosses it

c) The advantages of a flexible approach for regression or classification are that we can reduce bias and estimate a larger number of parameters in a nonlinear model. However, the disadvantages are if the parameters are too close to the error this can cause overfitting, in the case of flexible approach, it is preferrable if we want to the interpret the model's parameters, but in the case of less flexible, we only want the prediction.

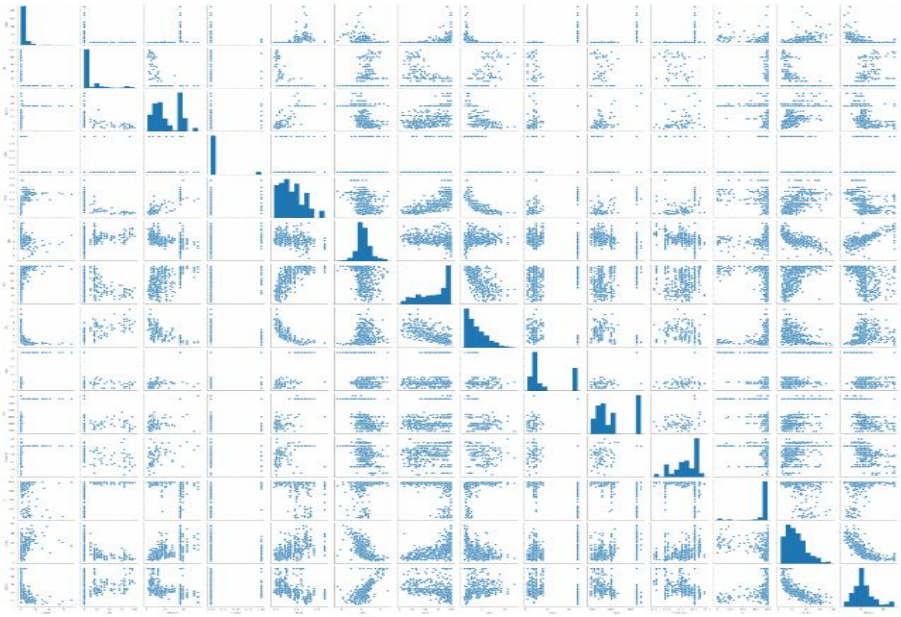
10.

a) There are **506 rows & 14** columns in the dataset. The rows represent the value, and the columns represent the labels for each value.

```
Boston_Housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
```

b)



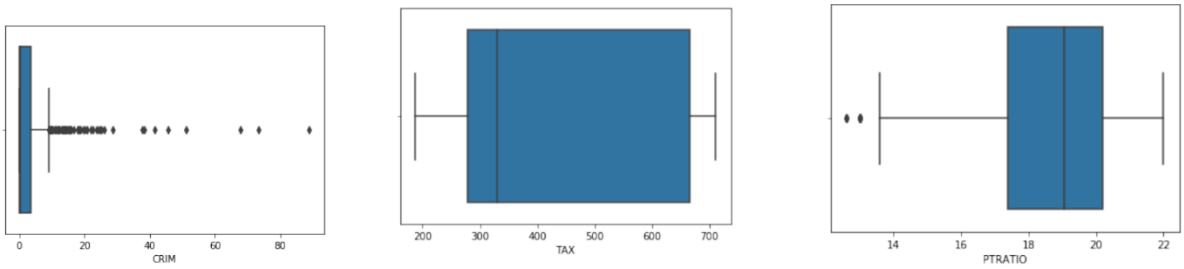
The findings based on my pair plot among all columns. There is a **positive correlation** between RM & MDEV. **Negative correlation** between LSTAT & RM, NOX & DIS, and LSTAT & MDEV.

c)

The is a between crime rate per capital and the following predictors

i. **Relatively and significantly positive correlation** NOX (0.42), RAD (0.62), TAX (0.58), and LSTAT (0.45).
ii. **Slightly positive correlation** AGE (0.35), PTRATIO (0.29), INDUS (0.4)
iii. **Marginally negative** CHAS (-0.055)
iv. **Relatively Negative correlation** Zin (-0.2), RM (-0.22), DIS (-0.38), B (-0.38),MDEV (-0.39)

d)



For this problem I did three boxplots

i. The Crime rate per capita boxplot is skewed mostly right and shows there is a **low rate** of crime between 0-10 range on the boxplot. There is however a large rate of outliers close to the Largest non-outlier, while there are approximately 8 suburbs who have a crime rate in the outlier range above 40.
ii. Tax(full-value property-tax rate per $10,000) boxplot is normally distributed, slightly left skewed, but overall, no outliers, median of 350 and range between 200-700 so no extreme high or low tax rates.
iii. Ptratio (pupil-teacher ratio by town) boxplot is skewed toward the left and has a median of 19, median 19, range 12-22, there are two outliers below the range, but not a large amount, and overall, no extreme high or low ratios.

e) **35** suburbs bound the Charles River.

```
Chas = Boston_Housing[Boston_Housing['CHAS']==1]
Chas['CHAS'].count()
```

```
35
```

f) The median pupil-teacher ratio among the towns is **21.2**.

```
: Median = Boston_Housing['MDEV']
  Median.describe()
```

```
: count    506.000000
  mean      22.532806
  std        9.197104
  min        5.000000
  25%       17.025000
  50%       21.200000
  75%       25.000000
  max       50.000000
  Name: MDEV, dtype: float64
```

g) The suburb with the lowest median value of the owner-occupied homes has a **MDEV value of 5.**

```
from statistics import median
Lowest = Boston_Housing[Boston_Housing['MDEV'] < Boston_Housing['MDEV'].
 ↪median()]
Lowest.min()
```

```
CRIM         0.0136
ZN           0.0000
INDUS        1.6900
CHAS         0.0000
NOX          0.3850
RM           4.1380
AGE         18.5000
DIS          1.1370
RAD          1.0000
TAX        188.0000
PTRATIO     14.7000
B            0.3200
LSTAT        5.5700
MDEV         5.0000
dtype: float64
```

The other predictors compared to MDEV are:		Range of the predictors:
CRIM	**0.0136**	88.96988
ZN	**0.0000**	100.00000
INDUS	**1.6900**	27.28000
CHAS	**0.0000**	1.00000
NOX	**0.3850**	0.48600
RM	**4.1380**	5.21900
AGE	**18.5000**	97.10000
DIS	**1.1370**	10.99690
RAD	**1.0000**	23.00000
TAX	**188.0000**	524.00000
PTRATIO	**14.7000**	9.40000
B	**0.3200**	396.58000
LSTAT	**5.5700**	36.24000

Compare to the Range , all these values **are way below the range** in each predictor except for R M & NOX which are the closest to the range. There are also values of 0 for ZN and CHAS, 0.32 f or B, 1 for RAD, the min for the predictors in both columns.

h) There are **64 suburbs** which have 7 or more rooms per dwelling & **13 suburbs** which have 8 or more rooms per dwelling. For the 13 suburbs, the people who live in these homes have a mean AGE of 71.5, and low crime rate meaning these are more than likely nursing homes for older individuals.

```
Seven =  Boston_Housing[Boston_Housing['RM'] >= 7]
Seven['RM'].count()
```

```
64
```

```
Eight =  Boston_Housing[Boston_Housing['RM'] >= 8]
print(Eight['RM'].count())
print(Eight['AGE'].mean())
Eight
```

```
13
71.53846153846153
```

Chapter 3

3.

a) y = 50 + 20 GPA +0.07 IQ +35 Gender +0.01 GPA*IQ – 10 GPA*Gender

Gender: {Male:0, Female:1}

Males: 50 +20 GPA +0.07 IQ +0.01 GPA*IQ

Females: 50 +20 GPA +0.07 IQ +35 + 0.01 GPA*IQ – 10 GPA

Subtract: Females = 35 -10 GPA -> GPA =3.5

iii) is correct because in the equation above when we subtract at a GPA average of 3.5 or more, women earn less than men.

a) Female: 50 +20 GPA +0.07 IQ +35 + 0.01 GPA*IQ – 10 GPA
50 +20(4) +0.07(110) +35 + 0.01 (4*110) – 10(4) **= 137.1 -> $137,100**

b) **False**, because even if the interaction is small, if the p value is below 0.05 the coefficient is statistically significant.

15.

a) **Every predictor except Charles River is statistically significant**. When looking at the graphs, the CHAS dummy variables is completely separated and has no normal distribution or correlation. But overall, the predictors are far away from fitted regression line when looking at other predictor's correlation with crime.

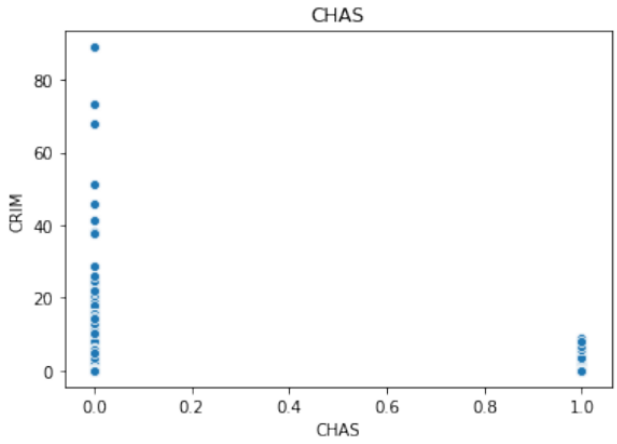
```
CHAS = smf.ols('CRIM ~ CHAS', data = Boston_Housing).fit()
CHAS.summary()
```

OLS Regression Results

Dep. Variable:	CRIM	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	1.546
Date:	Sat, 23 Jan 2021	Prob (F-statistic):	0.214
Time:	21:44:21	Log-Likelihood:	-1805.3
No. Observations:	506	AIC:	3615.
Df Residuals:	504	BIC:	3623.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.7232	0.396	9.404	0.000	2.945	4.501
CHAS	-1.8715	1.505	-1.243	0.214	-4.829	1.086

Omnibus:	562.698	Durbin-Watson:	0.822
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30864.755
Skew:	5.205	Prob(JB):	0.00
Kurtosis:	39.818	Cond. No.	3.96



b) We can reject the null hypothesis for **RAD, DIS,ZN, NOX and MDEV** because they have a p-value below 0.005.

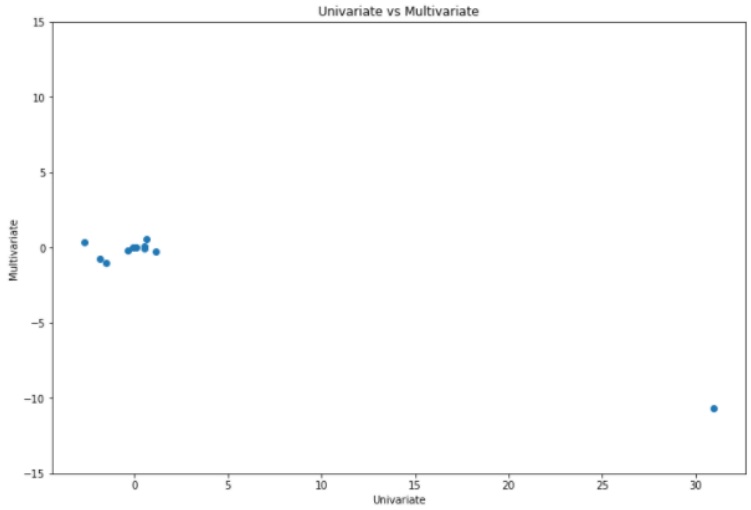
OLS Regression Results

Dep. Variable:	CRIM	R-squared:	0.448
Model:	OLS	Adj. R-squared:	0.434
Method:	Least Squares	F-statistic:	30.73
Date:	Sat, 23 Jan 2021	Prob (F-statistic):	2.04e-55
Time:	21:44:23	Log-Likelihood:	-1655.7
No. Observations:	506	AIC:	3339.
Df Residuals:	492	BIC:	3399.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	17.4184	7.270	2.396	0.017	3.135	31.702
AGE	0.0020	0.018	0.112	0.911	-0.033	0.037
B	-0.0069	0.004	-1.857	0.064	-0.014	0.000
CHAS	-0.7414	1.186	-0.625	0.532	-3.071	1.588
DIS	-0.9950	0.283	-3.514	0.000	-1.551	-0.439
INDUS	-0.0616	0.084	-0.735	0.463	-0.226	0.103
LSTAT	0.1213	0.076	1.594	0.112	-0.028	0.271
MDEV	-0.1992	0.061	-3.276	0.001	-0.319	-0.080
NOX	-10.6455	5.301	-2.008	0.045	-21.061	-0.230
PTRATIO	-0.2787	0.187	-1.488	0.137	-0.647	0.089
RAD	0.5888	0.088	6.656	0.000	0.415	0.763
RM	0.3811	0.616	0.619	0.536	-0.829	1.591
TAX	-0.0037	0.005	-0.723	0.470	-0.014	0.006
ZN	0.0449	0.019	2.386	0.017	0.008	0.082

Omnibus:	662.271	Durbin-Watson:	1.515
Prob(Omnibus):	0.000	Jarque-Bera (JB):	82701.666
Skew:	6.544	Prob(JB):	0.00
Kurtosis:	64.248	Cond. No.	1.58e+04

c)



d) There is nonlinearity for the following variables **INDUS, NOX, DIS, PTRATIO, MDEV**

INDUS

Dep. Variable:	CRIM	**R-squared:**	0.257
Model:	OLS	**Adj. R-squared:**	0.252
Method:	Least Squares	**F-statistic:**	57.86
Date:	Sat, 23 Jan 2021	**Prob (F-statistic):**	3.88e-32
Time:	21:46:17	**Log-Likelihood:**	-1731.0
No. Observations:	506	**AIC:**	3470.
Df Residuals:	502	**BIC:**	3487.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.6410	1.576	2.310	0.021	0.545	6.737
INDUS	-1.9533	0.483	-4.047	0.000	-2.901	-1.005
np.power(INDUS, 2)	0.2504	0.039	6.361	0.000	0.173	0.328
np.power(INDUS, 3)	-0.0069	0.001	-7.239	0.000	-0.009	-0.005

Omnibus:	611.416	**Durbin-Watson:**	1.118
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	51547.097
Skew:	5.815	**Prob(JB):**	0.00
Kurtosis:	51.059	**Cond. No.**	2.47e+04

NOX

Dep. Variable:	CRIM	**R-squared:**	0.292
Model:	OLS	**Adj. R-squared:**	0.288
Method:	Least Squares	**F-statistic:**	69.14
Date:	Sat, 23 Jan 2021	**Prob (F-statistic):**	1.94e-37
Time:	21:46:36	**Log-Likelihood:**	-1718.6
No. Observations:	506	**AIC:**	3445.
Df Residuals:	502	**BIC:**	3462.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	230.1421	33.734	6.822	0.000	163.864	296.420
NOX	-1264.1021	170.860	-7.398	0.000	-1599.791	-928.414
np.power(NOX, 2)	2223.2265	280.659	7.921	0.000	1671.816	2774.637
np.power(NOX, 3)	-1232.3894	149.687	-8.233	0.000	-1526.479	-938.300

Omnibus:	612.604	**Durbin-Watson:**	1.159
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	52872.508
Skew:	5.824	**Prob(JB):**	0.00
Kurtosis:	51.705	**Cond. No.**	1.36e+03

DIS

Dep. Variable:	CRIM	**R-squared:**	0.276
Model:	OLS	**Adj. R-squared:**	0.272
Method:	Least Squares	**F-statistic:**	63.74
Date:	Sat, 23 Jan 2021	**Prob (F-statistic):**	6.20e-35
Time:	21:46:51	**Log-Likelihood:**	-1724.4
No. Observations:	506	**AIC:**	3457.
Df Residuals:	502	**BIC:**	3474.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	29.9496	2.448	12.235	0.000	25.140	34.759
DIS	-15.5172	1.737	-8.931	0.000	-18.931	-12.104
np.power(DIS, 2)	2.4479	0.347	7.061	0.000	1.767	3.129
np.power(DIS, 3)	-0.1185	0.020	-5.802	0.000	-0.159	-0.078

Omnibus:	577.986	**Durbin-Watson:**	1.133
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	42441.952
Skew:	5.310	**Prob(JB):**	0.00
Kurtosis:	46.592	**Cond. No.**	2.10e+03

PTRATIO

Dep. Variable:	CRIM	**R-squared:**	0.112
Model:	OLS	**Adj. R-squared:**	0.107
Method:	Least Squares	**F-statistic:**	21.21
Date:	Sat, 23 Jan 2021	**Prob (F-statistic):**	5.99e-13
Time:	21:47:07	**Log-Likelihood:**	-1775.9
No. Observations:	506	**AIC:**	3560.
Df Residuals:	502	**BIC:**	3577.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	474.0255	156.823	3.023	0.003	165.915	782.135
PTRATIO	-81.8089	27.649	-2.959	0.003	-136.131	-27.487
np.power(PTRATIO, 2)	4.6039	1.609	2.862	0.004	1.444	7.764
np.power(PTRATIO, 3)	-0.0842	0.031	-2.724	0.007	-0.145	-0.023

Omnibus:	572.978	**Durbin-Watson:**	0.949
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36189.609
Skew:	5.303	**Prob(JB):**	0.00
Kurtosis:	43.050	**Cond. No.**	3.02e+06

MDEV

Dep. Variable:	CRIM	**R-squared:**	0.416
Model:	OLS	**Adj. R-squared:**	0.413
Method:	Least Squares	**F-statistic:**	119.2
Date:	Sat, 23 Jan 2021	**Prob (F-statistic):**	2.65e-58
Time:	21:47:15	**Log-Likelihood:**	-1670.0
No. Observations:	506	**AIC:**	3348.
Df Residuals:	502	**BIC:**	3365.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	52.9386	3.366	15.725	0.000	46.325	59.553
MDEV	-5.0774	0.435	-11.668	0.000	-5.932	-4.222
np.power(MDEV, 2)	0.1551	0.017	8.995	0.000	0.121	0.189
np.power(MDEV, 3)	-0.0015	0.000	-7.277	0.000	-0.002	-0.001

Omnibus:	568.100	**Durbin-Watson:**	1.360
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	47296.533
Skew:	5.084	**Prob(JB):**	0.00
Kurtosis:	49.259	**Cond. No.**	3.67e+05