An Introduction to Statistical Learning (with python application) Chapter 2 Exercise: (2), (3), (5)

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

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.

Answer: As CEO salary is a numerical value so quantitative in nature that's why it is related to Regression problem. As per problem statement we have to find out current factors affecting CEO salary so it is inference. Here, sample size will be n = 500 as we collect data of top 500 firms and recorded variables or predictors are p = 3.

b) (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 20 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 ten other variables.

Answer: As we have to detect its success or failure then it will be Classification problem.
As we are interested in its future status such as success or failure so it is prediction.
Here, n=20 as we collect data of 20 similar products.
p = 13 as there are total 13 variables or predictors which need to study for the required prediction.

c) (c) We are interested in predicting the % 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 2012. 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.

Answer: Its regression problem as we are interested in finding % change of exchange rate. As stated in the first line we have to find out future value then it is prediction.
Here, n = 52 number of weeks in year 2012 as we are collecting weekly data.
P = 3 as there are 3 variables or predictors which need to study for the required prediction.

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

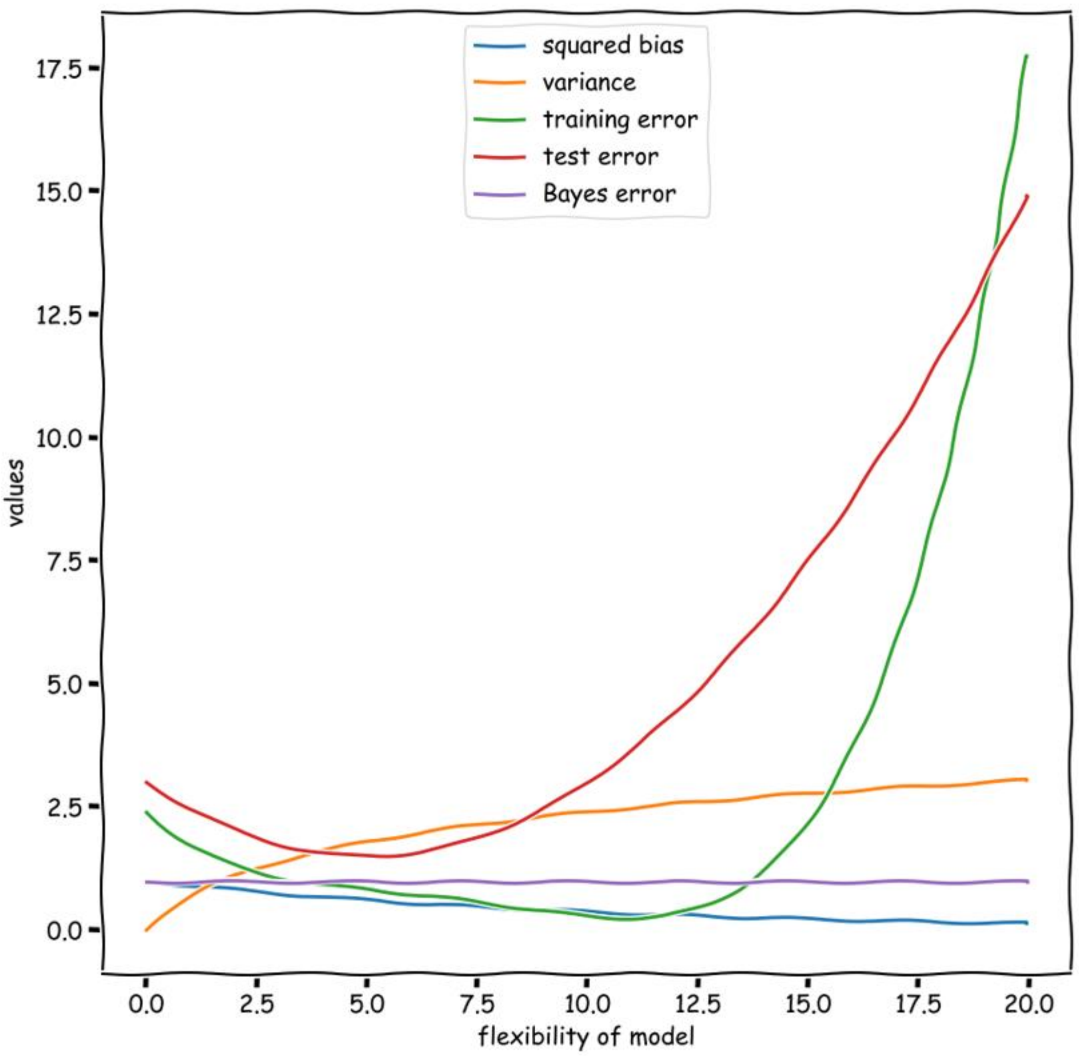
Code:

```
import numpy as np
import matplotlib.pyplot as plt

x = np.arange(0.0, 20.0, 0.05)
x

def typical_squared_bias(x):
   return np.exp(-0.1 * x)
def variance(x):
   return np.log(1+x)
def training_error(x):
    return 2.38936 - 0.825077*x + 0.176655*x**2 - 0.0182319*x**3 + 0.00067091*x**4
def test_error(x):
    return 3 - 0.6*x + .06*x**2
def bayes_error(x):
    return x + 1 – x
plt.xkcd()

plt.figure(figsize=(10, 10))
plt.plot(x,typical_squared_bias(x), label='squared bias')
plt.plot(x, variance(x), label='variance')
plt.plot(x, training_error(x), label='training error')
plt.plot(x, test_error(x), label='test error')
plt.plot(x, bayes_error(x), label='Bayes error')
plt.legend(loc='upper center')
plt.xlabel('flexibility of model')
plt.ylabel('values')
plt.savefig('exercise3.png')
plt.show()
```



(b) Explain why each of the five curves has the shape displayed in part (a).

Answer: Squared Bias value will be decreased as model flexibility increases.

Variance will be around zero with no flexibility and it start increasing with the model flexibility.

Training Error: Training error typically decreases as the model becomes more flexible, but when flexibility increases certain level training error is also increase exponentially due to noise.

Test error follows more of U-shaped curve, with an optimal point representing the best trade-off between bias and variance

Bayes error seems constant through any level of flexibility.

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Answer: Advantages of very flexible (versus a less flexible) approach are it may be less bias and provide better results if input data is sufficient.
Where as analysis take more time and model training will be more computational. It also has more risk or overfitting and less clear interpretability.

More flexible approach is best suits when sample size is large and variables are limited or less.
Where as less flexible approach produces better result with small sample size and large number of predictors.

Open Intro Statistics (4th Edition) Chapter 1/2 Exercise: 1.4 (Page 21) 2.12 (Page 67)

1.4 a) Identify the main research question to study?
Answer: To find out the effectiveness of Buteyko method on reduction of Asthma symptoms and improvement quality of life?

1,4 b) Who are the subjects in this study and how many are included?
Answer: Asthma patients aged 18-69 are subjects and 600 are included in the study.

1.4 c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

Answer: Variables are:
Treatment group: categorical as divided in 2 groups.
Quality of life: numerical as measure in scale of 0-10
Activity: numerical as measure in scale of 0-10
Asthma symptoms: numerical as measure in scale of 0-10
Medication reduction: numerical as measure in scale of 0-10
Usually when we rate on scale 0-10 we won't consider decimal value and recorded as whole value if we do the same here then all numerical scale values are discrete.
And treatment group is not categorize order or rank wise so it is nominal.

2.12) Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the media.
Answer: Median seems near to 80 and mean will be lower than median for the given histogram.

Answer for R part:

Answer a): data frame has 506 rows and 14 columns,

This data frame contains the following columns:

Crim – per capita crime rate by town.

Zn – proportion of residential land zoned for lots over 25,000 sq.ft.

Indus – proportion of non-retail business acres per town.

Chas – Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

Nox – nitrogen oxides concentration (parts per 10 million).

Rm – average number of rooms per dwelling.

Age – proportion of owner-occupied units built prior to 1940.

Dis – weighted mean of distances to five Boston employment centres.

Rad – index of accessibility to radial highways.

Tax – full-value property-tax rate per $10,000.

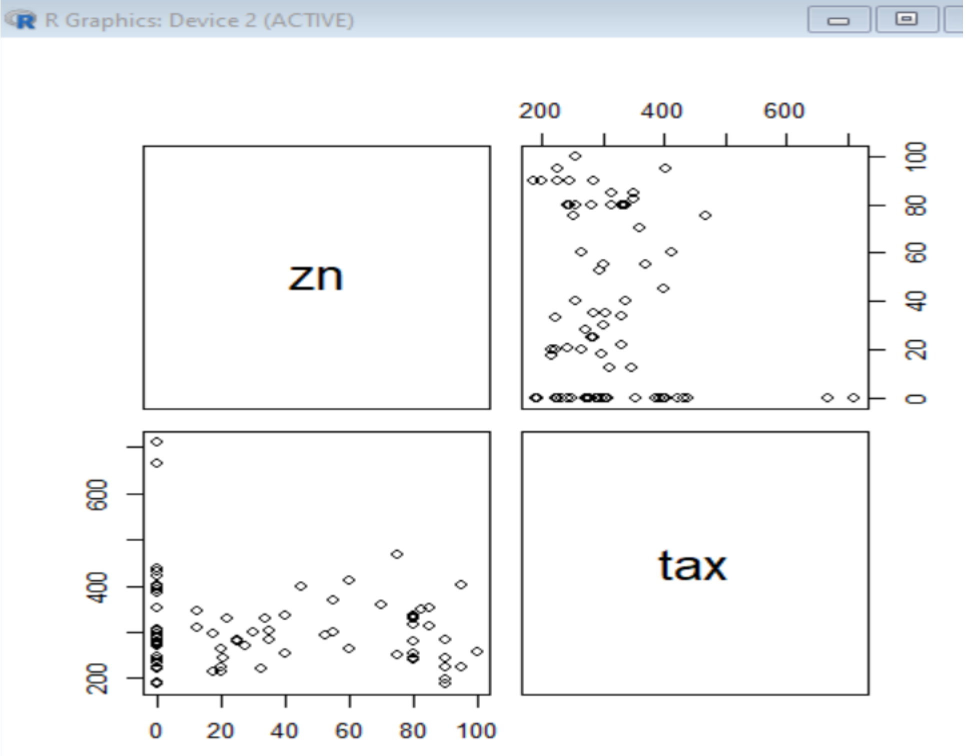
Ptratio – pupil-teacher ratio by town.

Black – $1000(Bk-0.63)$2 where Bk is the proportion of blacks by town.

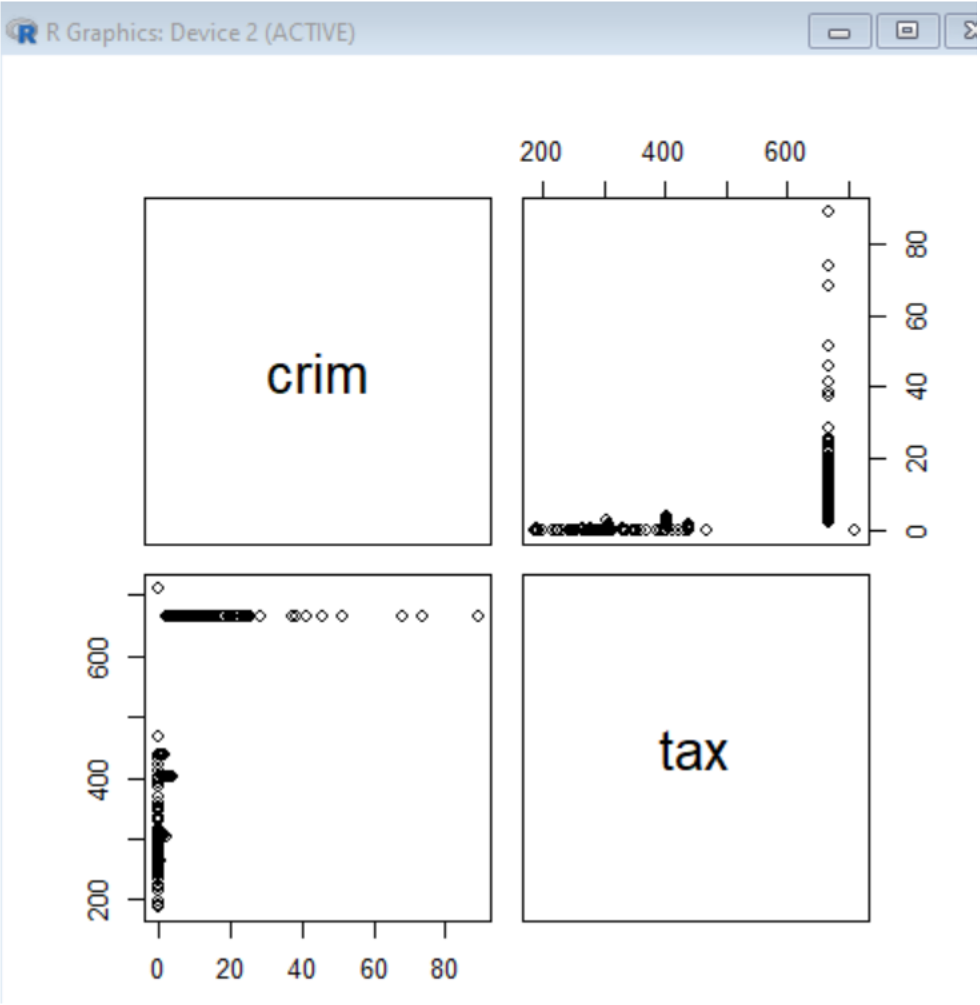
Lstat – lower status of the population (percent).

Medv – median value of owner-occupied homes in $1000s.

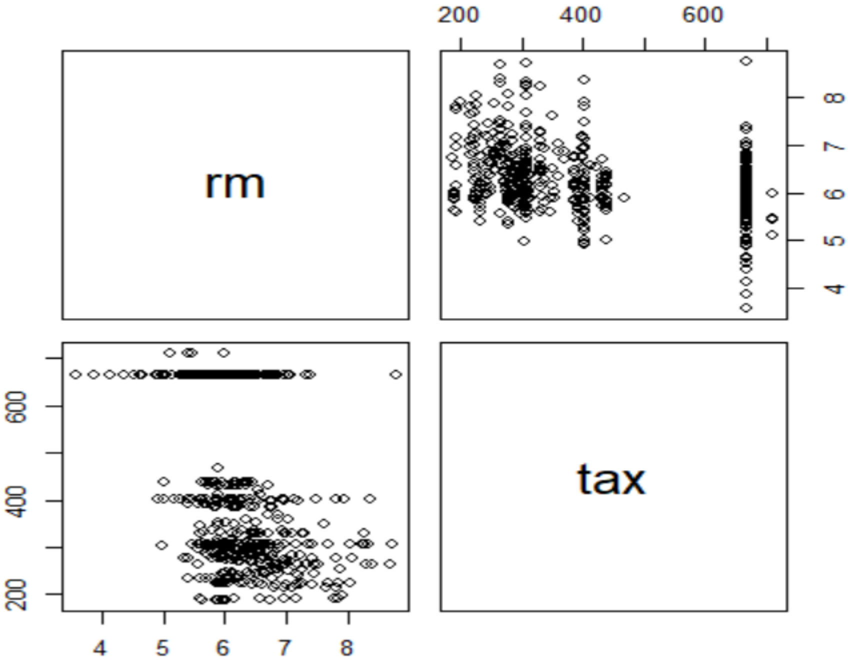
Answer b): some pairwise graph
Graph 1) Places with lower taxes have higher land zoned for lots over 25,000 sq ft.



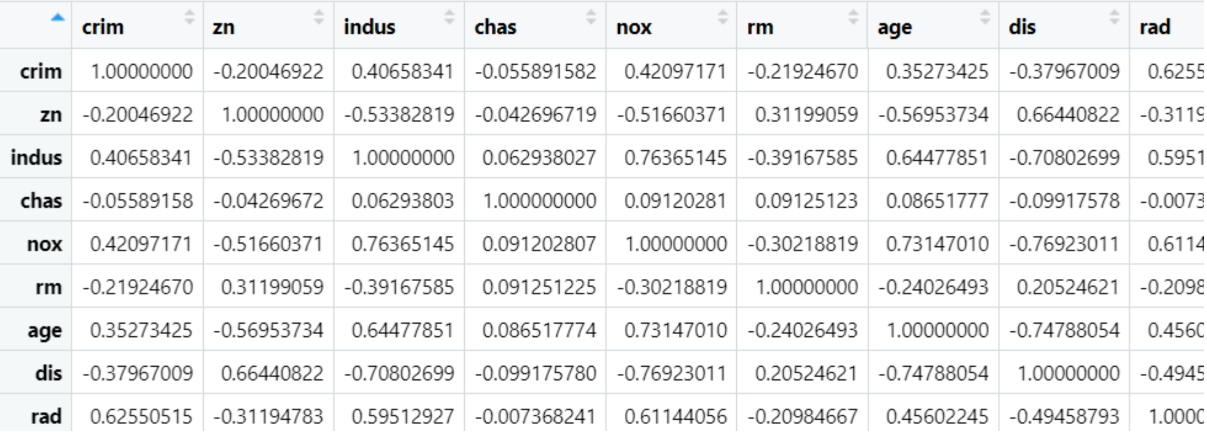
Graph 2) Higher taxes places have higher crime rates



3) High tax more average room.



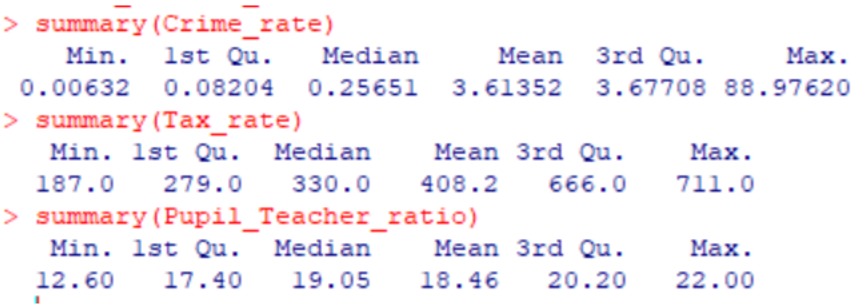
Answer c) correlation matrix is used to find associated predictors with per capita crime rate
Yes, there are predictors associated with per capita crime rate. Variables such as percentage of lower-status population (lstat), nitrogen oxide concentration (nox), and property tax rate (tax), proportion of non-retail business acres (INDUS), accessibility to radial highways(rad) are positively associated.



	crim	zn	indus	chas	nox	rm	age	dis	rad
crim	1.00000000	-0.20046922	0.40658341	-0.055891582	0.42097171	-0.21924670	0.35273425	-0.37967009	0.6255
zn	-0.20046922	1.00000000	-0.53382819	-0.042696719	-0.51660371	0.31199059	-0.56953734	0.66440822	-0.3119
indus	0.40658341	-0.53382819	1.00000000	0.062938027	0.76365145	-0.39167585	0.64477851	-0.70802699	0.5951
chas	-0.05589158	-0.04269672	0.06293803	1.000000000	0.09120281	0.09125123	0.08651777	-0.09917578	-0.0073
nox	0.42097171	-0.51660371	0.76365145	0.091202807	1.00000000	-0.30218819	0.73147010	-0.76923011	0.6114
rm	-0.21924670	0.31199059	-0.39167585	0.091251225	-0.30218819	1.00000000	-0.24026493	0.20524621	-0.2098
age	0.35273425	-0.56953734	0.64477851	0.086517774	0.73147010	-0.24026493	1.00000000	-0.74788054	0.4560
dis	-0.37967009	0.66440822	-0.70802699	-0.099175780	-0.76923011	0.20524621	-0.74788054	1.00000000	-0.4945
rad	0.62550515	-0.31194783	0.59512927	-0.007368241	0.61144056	-0.20984667	0.45602245	-0.49458793	1.0000

Answer d)

Some Boston suburbs have higher per capita crime rate.
Some Boston suburbs have higher Tax rates.
Some some Boston suburbs have higher Pupil Teacher ratio.



```
> summary(Crime_rate)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620 
> summary(Tax_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  187.0   279.0   330.0   408.2   666.0   711.0 
> summary(Pupil_Teacher_ratio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.60   17.40   19.05   18.46   20.20   22.00 
```

Answer E) 35 suburbs bound to Charles river.

Answer F) 19.05 is the pupil-teacher ratio among towns.

Answer G) Median is 399,
This suburb has the highest number of owner-occupied units built prior 1940(age), highest index of accessibility to radial highways(rad) and highest proportion African American descents(black),

Answer H) 64 Suburb have more than seven room per dwelling and 13 have more than 8 room per dwelling. House with more rooms may reflects high income.