CS 4342 Assignment #1

Ivan Martinovic

Questions

Conceptual and Theoretical Questions

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size *n* is extremely large, and the number of predictors *p* is small.

A more flexible model would yield greater performance for this case. Since large data set more accurately reflects relationship, then a more flexible model can better fit the data, than a less flexible one.

(b) The number of predictors *p* is extremely large, and the number of observations *n* is small.

In the case where we have a small number of observations, a less flexible model generally yields better performance. This is due to the fact that less flexible models tend to require less data points to for a more accurate estimate than more flexible models.

(c) The relationship between the predictors and response is highly non-linear.

In the case where the relationship between the predictors and the response is highly non-linear, a more flexible model would yield a better performance. This is due to the fact that less flexible models, because they are linear or slightly non-linear, will have a stronger bias when estimating highly non-linear relationships; compared to their more flexible (and more non-linear) counterparts. And since the test MSE is proportional to the variance and bias squared of our model function, the error due to higher biases of the less flexible models (because it is a quadratic relationship) easily overshadows the error due to higher variances of the more flexible models.

(d) The variance of the error terms, i.e. σ2=Var(ϵ), is extremely high.

In the case where the variance of the error terms is extremely high, a less flexible model would yield a better performance. This is due to the fact that more flexible models tend to imitate the measurement errors of observations. When the variance in the error terms is high, the danger of overfitting rises. Therefore, less flexible models, which tend to capture less errors, will yield a greater performance in the case where the variance of the error terms is high.

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.

This is a regression problem since we are trying to model a numeric value, namely the CEO salary. This is an inference problem, since we want to understand the relationship, rather than just predict an outcome. The sample size n is 500, the number of predictors p is 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 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.

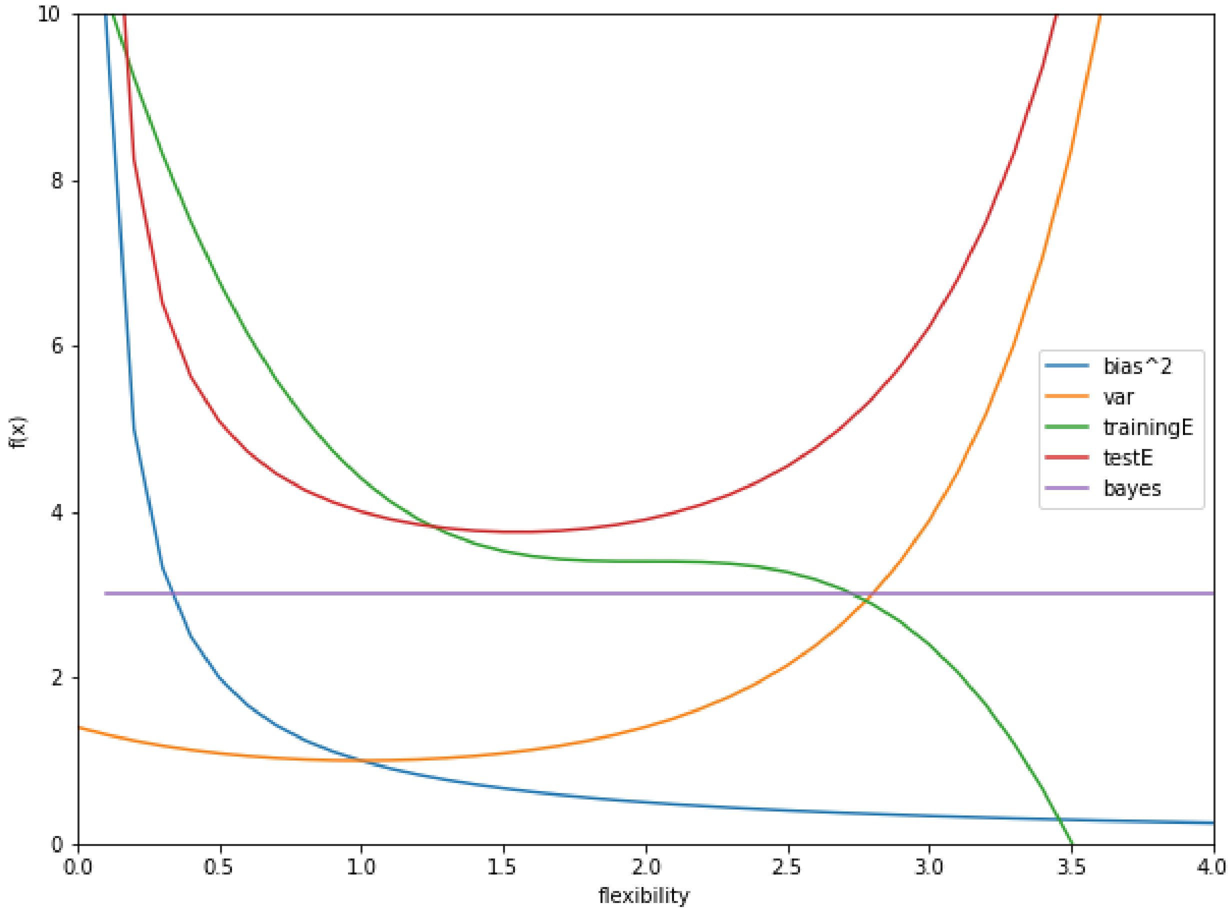
The sample size n is 20, and the number of predictors p is 13. This is a classification problem; we are modeling a qualitative variable, i.e. whether a product is a success or a failure. This is a prediction problem, we do not need to understand the relationship between the predictors and the response variables, we just want to predict whether a product will succeed or not.

(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.

This is a regression problem, since we are estimating a numerical variable (the % change in the USD/Euro exchange rate). As the question suggests, this is a prediction problem, i.e. we are interested in predicting the % change, not really in understanding the relationship between the USD/Euro exchange rate and the weekly changes in the world stock markets. There are 52 weeks in 2012, hence n is 52. Number of predictors p is 3.

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.



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

The bias of the statistical learning method decreases as the statistical method becomes more flexible. This is by definition since more flexible methods, make less assumptions about the relationships between the predictors and observed variable.

The variance of the statistical learning method increases as the statistical method becomes more flexible. This is due to the fact that, as a statistical method becomes more flexible, a small change in one of the observations would lead to a much larger change in the shape of the function.

The Bayes error curve is a horizontal line. It is the "ideal" error rate, i.e. the error rate below which no learning method can produce a lower test MSE for. It does not depend on the learning method, but rather on the actual data itself.

The training error decreases as the flexibility increases. As the model becomes more flexible, it starts approaching the data points, and the difference between the predicted values and the response values in the training data set becomes smaller.

The test error is the sum of the Bayes error rate, the bias squared and variance of the model function. Initially, the test error decreases as the flexibility of the learning method increases, since more flexible methods are less biased. But eventually we reach a point of diminishing returns, and we start overfitting our data. At this point our learning method has become too flexible and it started to model the errors in the observations, not the relationship itself. This is also exactly the point where the increase in the variance of the model function becomes greater than the decrease in bias squared. Therefore the test error starts increasing.

4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
Example 1: We want to determine whether a vehicle is a car, a truck, or a bus at the tollbooth station to determine how much to charge. The goal here is prediction, we do not care to understand how the system determines to what category a vehicle belongs to, as long as it is the correct one.

Example 2: We want to determine what letter is written on an image of a piece of paper. The goal here is prediction, we do not care to understand how the system determines which letter is written, as long as it identifies the correct letter.

Example 3: We want to submit a resume to our dream job for a company, and we want to determine which sections of our resume should be longest. We (somehow) collect resumes of all applicants to the company, and whether they got accepted or not. The goal here is inference, we want to determine the relationship between the lengths of the sections and whether the applicant got accepted, so that we may alter our resume to give us better chances of getting hired.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Example 1: We wish to predict the changes in the price of a stock. We gather some predictors: suck as current stock value, state of the economy, expectations for the future etc. The response variable is the price of the stock. In such a case prediction is the answer … we just care to know if a stock is going to increase or decrease in value in the future.

Example 2: We developed a hypothesis for a physics experiment which we want to test. Suppose it is Hooke's law for springs. The predictor variable is the force exerted on a spring. The response variable is the change of length of the spring. The goal of our physics model is both inference and prediction. Using the model we want to know what is the underlying relationship between the force exerted on a spring and its elongation. However, we would also like to use this model to make accurate predictions about the future.

Example 3: Suppose we are a coach for a strength sport athlete. We want to understand what is fatiguing for our athlete. The predictor variables are exercise selection, weight lifted, total volume and intensity. Our response variable is the athletes perceived fatigue level. This is an inference problem, because we want to understand through our model the relationship between the predictor variables and the response variables in order for us as a coach to write better training programs for our athlete in the future.

(c) Describe three real-life applications in which cluster analysis might be useful.

Example 1: In medicine, more specifically during PET scans, cluster analysis can be used to differentiate between different types of tissue in a three-dimensional image for many different purposes.

Example 2: Clustering can be used to divide a digital image into distinct regions for border detection or object recognition.

Example 3: Cluster analysis can be used to identify areas where there are greater incidences of particular types of crime. Using such information it is possible to manage law enforcement resources more effectively.

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?

We choose a more or less flexible model depending on how complex our problem is. We should choose less flexible models for less complex problems, since less flexible models can better capture less complex relationships and avoid overfitting. We should choose more flexible models for more complex problems, since more flexible models can better capture complexities and avoid biases.

A less flexible model is also more easily interpretable. Less flexible models are usually simpler, and they let us form more concise inferences between the predictor variables and the response variable. Hence less flexible models are preferred when inference is the goal.

Real life phenomena are, however, rarely linear or simple. Therefore more flexible approaches usually give a more accurate relationship between the predictor variables and the response variable. Hence they are usually preferred when accuracy of prediction is the goal.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

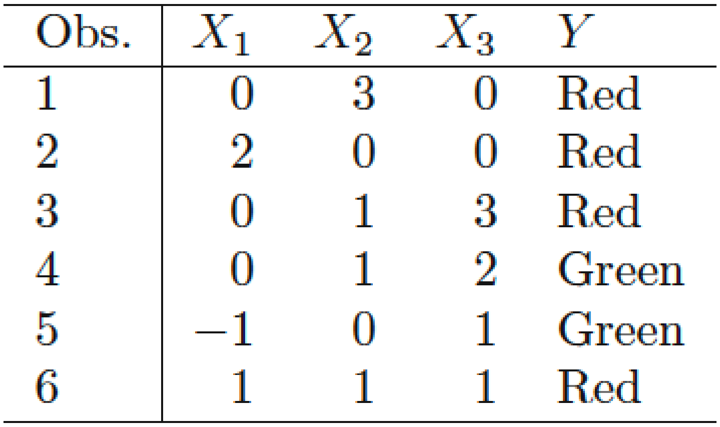
In a parametric approach to statistical learning we first assume the functional form or shape of relationship we are trying to model. Then we proceed to finding the parameters using our training data, which makes our model best fit the observer phenomena.

In a non-parametric approach to statistical learning we do not make explicit assumptions about the functional form of the relationship we are trying to model. Instead we seek to create an estimate of the relationship that gets as close to the data points as possible without being too rough or wiggly.

The main advantage of parametric methods is that now instead of trying to estimate which functional form best matches the relationship, we have reduced the problem to just estimating the parameters of a single functional form based on some best-fit criteria (which is generally easier). The main disadvantage of such methods is that they have a possibility that the functional form used to model the relationship is very different than the relationship itself, in which case the resulting model will not fit the model well.

The main advantage of non-parametric methods is that, by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a larger range of possible shapes for the relationship. The main disadvantage for non-parametric methods is that, since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations is required in order to obtain an accurate estimate for f.

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.
Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.



Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.
Dist(Obs1) = 3 = sqrt(9)
Dist(Obs2) = 2 = sqrt(4)
Dist(Obs3) = sqrt(1 + 9) = sqrt(10)
Dist(Obs4) = sqrt(1 + 4) = sqrt(5)
Dist(Obs5) = sqrt(1 + 1) = sqrt(2)
Dist(Obs6) = sqrt(1 + 1 + 1) = sqrt(3)

(b) What is our prediction with K = 1? Why?
The prediction is that Y is Green. We take one nearest neighbor to X1=X2=X3=0, which is Obs5, and since it is Green, we assign Green to our Y.

(c) What is our prediction with K = 3? Why?
The prediction is that Y is Red. We take 3 nearest neighbors to X1=X2=X3=0, which are Obs2, Obs5 and Obs6. Since most neighbors (Obs2 and Obs6) are green, then we assign Green to our Y.

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

If the Bayes decision boundary is highly non-linear, then we would expect the best value for K to be small. This is due to the fact that as K grows, the KNN method becomes less flexible i.e. more linear. Hence a smaller K, will lead to a more flexible model, which will in-turn better capture the high non-linearity of the Bayes decision boundary.

Applied Questions (*Note: see end of document for code***)**
1. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?
Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year
Qualitative: origin, name

(b) What is the range of each quantitative predictor?
Mpg range: 37.6
cylinders range: 5
displacement range: 387.0
horsepower range: 184
weight range: 3527
acceleration range:16.8
year range:12

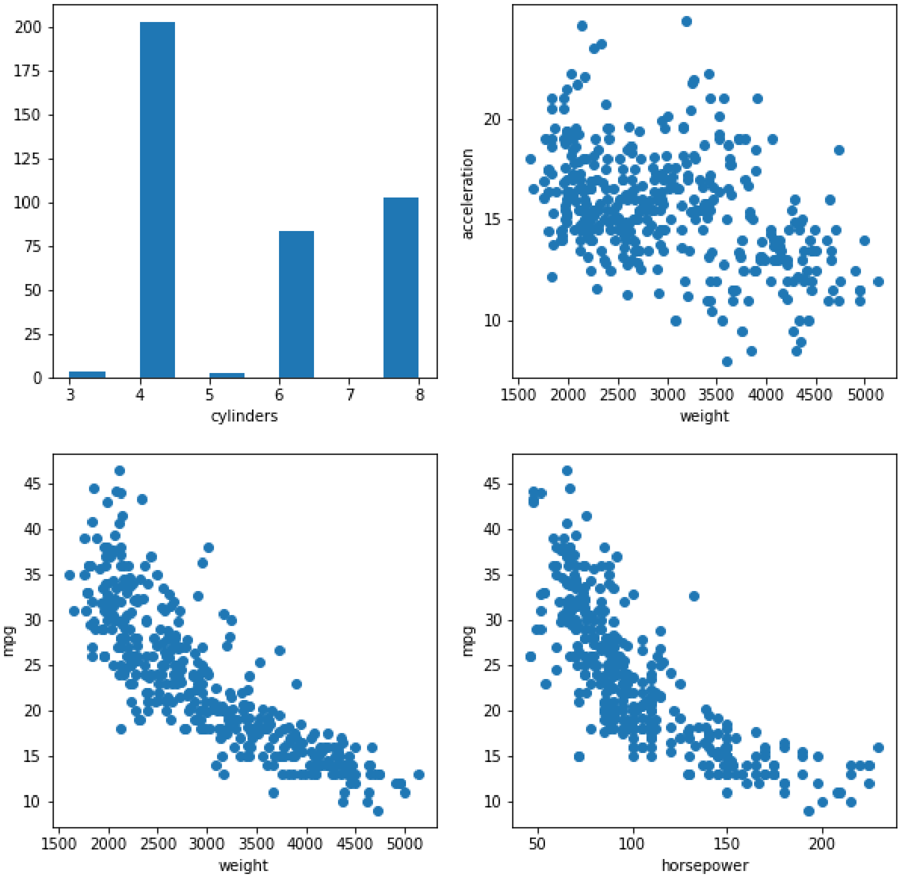
(c) What is the mean and standard deviation of each quantitative predictor?

Mpg mean:23.51587	Mpg variance 61.08894
cylinders mean:5.45844	cylinders variance 2.88807
displacement mean:193.53275	displacement variance 10867.65384
horsepower mean:104.46939	horsepower variance 1477.78988
weight mean:2970.26196	weight variance 717130.46034
acceleration mean:15.55567	acceleration variance 7.54343
year mean:75.99496	year variance 13.58184

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

```
Mpg range: 37.6              Mpg mean:23.51587            Mpg variance 61.08894
cylinders range: 5           cylinders mean:5.45844       cylinders variance 2.88807
displacement range: 387.0    displacement mean:193.53275  displacement variance 10867.65384
horsepower range: 184        horsepower mean:104.46939    horsepower variance 1477.78988
weight range: 3527           weight mean:2970.26196       weight variance 717130.46034
acceleration range:16.8      acceleration mean:15.55567   acceleration variance 7.54343
year range:12                year mean:75.99496           year variance 13.58184
```

(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.



Top left graph shows that most cars are 4-cylinder, followed by 8-cylincer and 6 cylinder cars. 3 and 5-cylinder cars are least common.
Top right graph shows that, although there is a lot of variance in the data, acceleration generally decreases with weight of the vehicle.
Bottom left graph shows that, miles per gallon decreases with weight of the vehicle.
Bottom right graph shows that, miles per gallon decreases with horsepower of the vehicle.

(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.

The bottom two graphs suggest that mpg depends on the weight of the vehicle and its horsepower. Miles per gallon seems to be inversely proportional to the weight and the horsepower of the vehicle.

Hints:

- Range: The **range of a set** of data is the difference between the highest and lowest values in the **set**. To find the **range**, first order the data from least to greatest. Then subtract the smallest value from the largest value in the **set**.

- You can use NumPy package. It provides functions for the mean, min, and max of arrays.

- Load Data: You can use Pandas package. You might use functions in Pandas like read_csv.

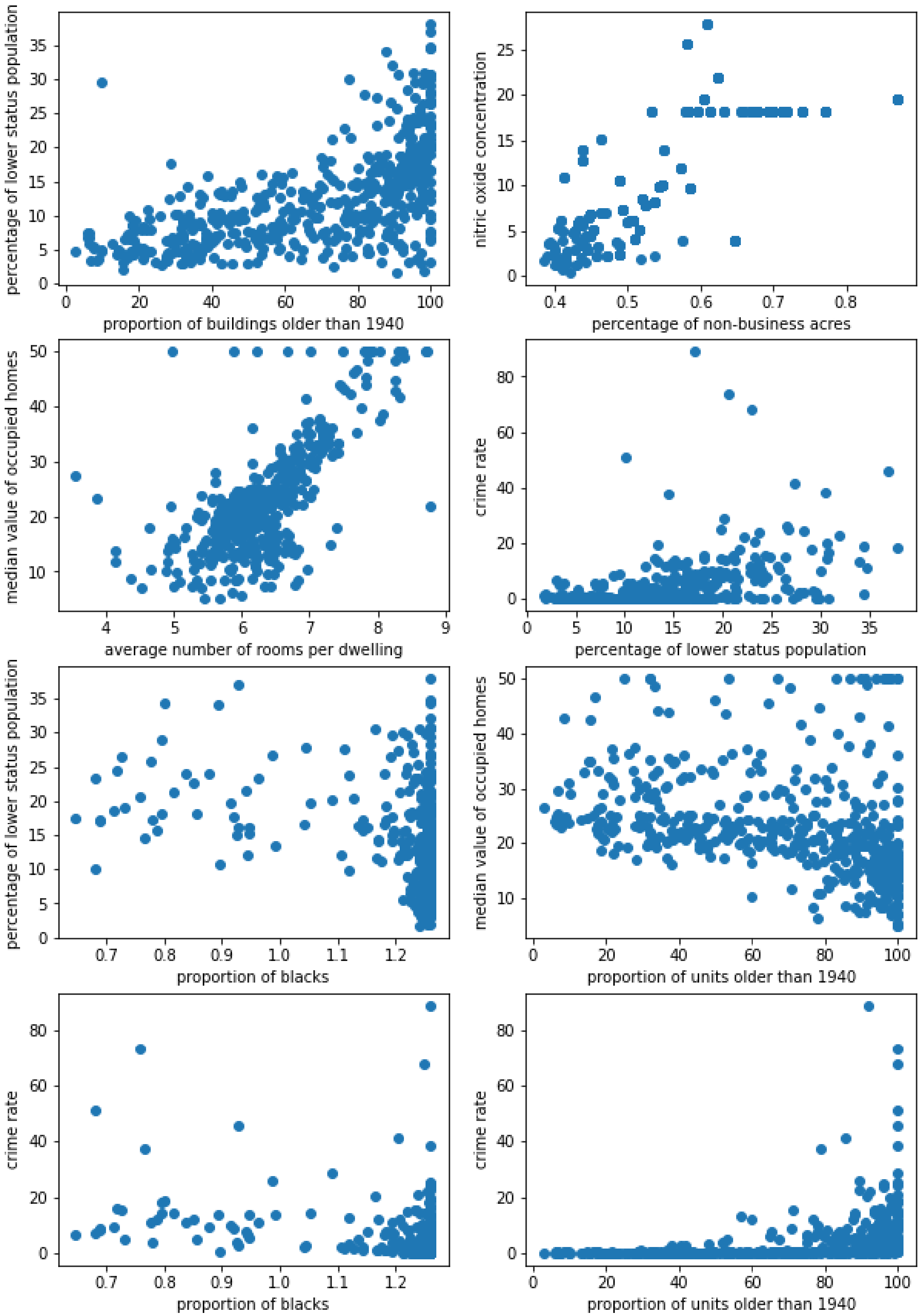
- Scatter plots: check this
https://jakevdp.github.io/PythonDataScienceHandbook/04.02-simple-scatter-plots.html

Points: 15

2. 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?
There are 506 rows; these represent our observed data set.
There are 14 columns; these represent our predictor variables.

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


percentage of lower status population
0
5
10
15
20
25
30
35
0
20
40
60
80
100
proportion of buildings older than 1940
nitric oxide concentration
0
5
10
15
20
25
0.4
0.5
0.6
0.7
0.8
percentage of non-business acres
median value of occupied homes
10
20
30
40
50
4
5
6
7
8
9
average number of rooms per dwelling
crime rate
0
20
40
60
80
0
5
10
15
20
25
30
35
percentage of lower status population
percentage of lower status population
0
5
10
15
20
25
30
35
0.7
0.8
0.9
1.0
1.1
1.2
proportion of blacks
median value of occupied homes
10
20
30
40
50
0
20
40
60
80
100
proportion of units older than 1940
crime rate
0
20
40
60
80
0.7
0.8
0.9
1.0
1.1
1.2
proportion of blacks
crime rate
0
20
40
60
80
0
20
40
60
80
100
proportion of units older than 1940

Starting at the top left plot, we see that the percentage of the lower status population increases as the proportion of old buildings increases, indicating that lower status population tends to live in older buildings.

Top right plot shows that nitric oxide concentrations seem to be higher in areas where there is a higher proportion of non-retail business (aka industry) acres.

Left plot in the second row shows that the median value of occupied homes tends to rise as the average number of rooms per dwelling rises. In other words, houses with more rooms are generally more expensive.

Right plot in the second row shows that crime rate is generally higher in areas where the population is of lower status.

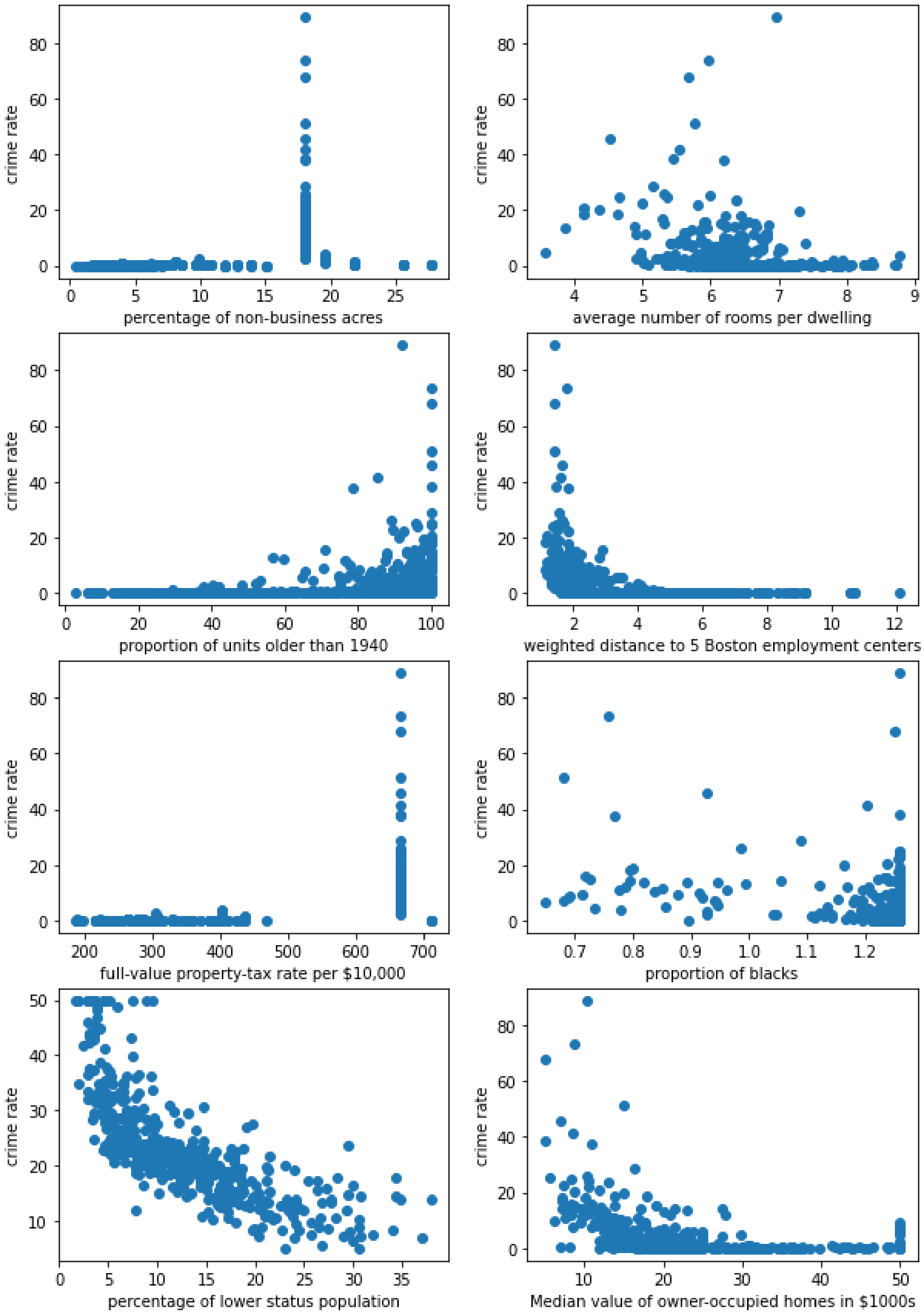
Left plot in the third row shows that percentage of lower status population stays relatively constant regardless of the proportion of blacks per area.

Right plot in the third row shows that the median values of homes stays relatively constant for most proportions of units older than 1940. Only when the proportions get considerably high (80% - 100% range) does the median value start to decline.

Left plot in the last row shows that the crime rate stays relatively constant regardless of the proportion of blacks per area.

Right plot in the last row shows that crime rate is relatively constant where the proportion of units older than 1940 is between (0% and 80%). However as the proportion of old units increases past 80% the crime rate also goes up.

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


crime rate
0
20
40
60
80
0
5
10
15
20
25
percentage of non-business acres
crime rate
0
20
40
60
80
4
5
6
7
8
9
average number of rooms per dwelling
crime rate
0
20
40
60
80
0
20
40
60
80
100
proportion of units older than 1940
crime rate
0
20
40
60
80
2
4
6
8
10
12
weighted distance to 5 Boston employment centers
crime rate
0
20
40
60
80
200
300
400
500
600
700
full-value property-tax rate per $10,000
crime rate
0
20
40
60
80
0.7
0.8
0.9
1.0
1.1
1.2
proportion of blacks
crime rate
10
20
30
40
50
0
5
10
15
20
25
30
35
percentage of lower status population
crime rate
0
20
40
60
80
10
20
30
40
50
Median value of owner-occupied homes in $1000s

Starting at the top left scatter plot, we see that except for one outlier, crime rate does not depend on the percentage of non-business acres. However when the percentage of non-business acres is around 18%, the crime rate seems to sky-rocket.

Top right scatter plot shows an interesting Bell-curve relationship between the crime rate and the average number of rooms per dwelling centered at around 5.5 rooms. The crime rate increases as the average number of rooms increases from 3 to 5.5, and then drops as the average number of rooms increases past 5.5.

Left plot in the second row was discussed in the previous question.

Right plot in the second row indicates that there is an inverse relationship between the crime rate and the weighted distance to 5 Boston employment centers. As the distance increases, the crime rate drops.

Left plot in the third row shows a similar relationship of crime rate and full-value property tax as one seen in the top left graph (the relationship between the crime rate and percentage of non-business acres). The crime rate does not seem to depend on the property-tax rate, except for one outlier. When the tax rate is around 675, the crime rate seems to sky-rocket.

Right plot in the third row was discussed in the previous question.

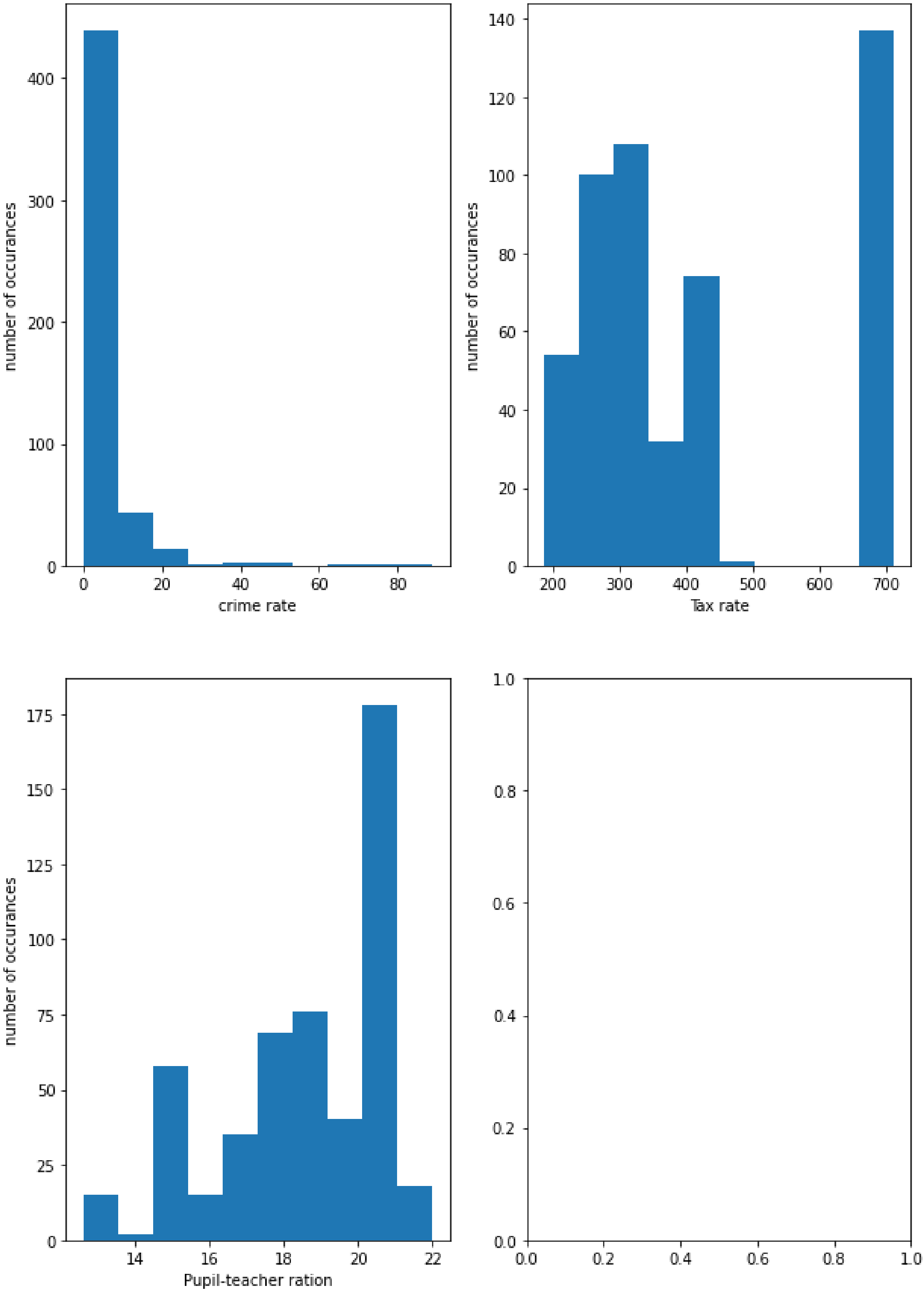
Bottom left graph shows an inverse relationship between the crime rate and the percentage of lower status population The crime rate seems to drop as the percentage of lower status population increases.

Bottom right graph seems to indicate an inverse relationship between the crime rate and the median value of owner-occupied homes. As the median value of the homes increases, the crime rate drops.

(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.
There seems to be a few suburbs with crime rates of over 60%. Crime rate range is 88.96988 which is huge. Since most suburbs have a crime rate of almost 0, this range indicates that the few suburbs with extremely high rates of crime are massive outliers.

Most full-value property-tax rate falls below 500 per 10 000$. However, the mode is actually around 700. There is a gap of tax rates between 500 and 700 per 10 000$ and the range of the tax rate is 524. This leads to the mean of the tax rate to be in the 500-700 range, even though none of the property tax rates is in that specified range.

The pupil-teacher ratio peaks at 22, although the mode is 21 and the mean is around 19. The range of the pupil-teacher ratio is 9.4, which indicates that some schools (outliers) have a pupil-teacher ratio which is about half the mean value.



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

There are 35 suburbs in this data which bound the Charles river.

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

The median pupil-teacher ratio among the towns in this data set is 19.1.

(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? Comment on your findings.

There are actually 2 suburbs with the lowest median value of owner occupied homes: suburb 398 and suburb 405 both with a value of median value of owner occupied homes in 1000$'s of 5.

Suburb 398 has a crime rate of 38.3518 and Suburb 405 has a crime rate of 67.9208. The range of the crime rate is between 0.00632 and 88.9762, implying that suburb 398 is in the mid-to-high range of crime rates, whereas suburb 405 is in the high range of crime rates.

Both suburbs have a proportion 0 of residential land zoned for lots over 25,000 sq.ft.

Both suburbs have a proportion of non-retail business acres per town of 18.1. The range of the industry proportion is from 0.46 to 27.74, implying that both suburbs are in the middle of the pack.

Both suburbs do not bound the Charles river.

Both suburbs have nitric oxides concentration of 0.693 (parts per 10 million). The range for nitric oxides concentrations is from (0.385 to 0.871), implying that the nitric oxide concentrations in the two suburbs are on the higher end.

The average number of rooms per dwelling in Suburb 398 is 5.453 and in Suburb 405 it is 5.683. The range for the average number of rooms per dwelling is from 3.561 to 8.78, implying that the two suburbs are in the "middle of the pack" when it comes to average number of rooms per dwelling.

The proportion of owner-occupied units built prior to 1940 is 100 in both suburbs. With a range of 2.9 to 100, we can say that both suburbs are on the high end when it comes to old buildings.

The weighted distances to five Boston employment centres for Suburb 398 is 1.4896 and for Suburb 405 it is 1.4254. With a range of 1.1296 to 12.1265, we see that both suburbs are on the low end of the spectrum.

Both suburbs have an index of accessibility to radial highways of 24.

Both suburbs have a full-value property-tax rate per $10,000 of 666. With a tax rate range from 187 to 711, we see that both of the suburbs are on the high-end of the spectrum.

Both suburbs have a pupil-teacher ratio of 20.2. With a range from 12.6 to 22, we see that both suburbs are on the higher-end of the spectrum.

The "black" coefficient of Suburb 398 is 396.90, and that of Suburb 405 is 384.97. With a range from 0.32 to 396.9, we see that both suburbs are on the high end of the spectrum.

The percentage of lower status of the population in Suburb 398 is 30.59, and in Suburb 405 it is 22.98. With a range from 1.73 to 37.97, we see that Suburb 405 is in the "middle of pack", whereas Suburb 398 is on the high end.

(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.

There are 64 suburbs with an average of more than seven rooms per dwelling.

There are 13 suburbs with an average of more than eight rooms per dwelling (these are Suburbs 13, 97, 163, 204, 224, 225, 226, 232, 233, 253, 257, 262, 267 and 364).

All of the suburbs have a very low crime rate (below 1%). There are only 2 small exceptions: Suburb 163 with a crime rate of 1.51902% and Suburb 364 with a crime rate of 3.47428%.

The proportion of residential land zones for lots over 25,000 sq ft. are most commonly 0 or 20. There are again two exceptions: Suburb 204 with a 95 zone proportion, and Suburb 253 with a 22 zone proportion.

The proportion of non-retail business acres per town is generally below 7, and with a range from 0.46 to 27.74 it is generally low. There are two outliers: Suburb 163 and Suburb 364 with a proportion of 19.58 and 18.10 respectively.

Interestingly, only 2 Suburbs bound the Charles river: Suburb 163 and Suburb 364.

The nitric oxides concentration (parts per 10 million), ranges from 0.4161 to 0.7180, with Suburb 204 on the low end and Suburb 364 on the high end. Most other suburbs sit around 0.5040, and with a general range between 0.385 and 0.871, we can say that they are in the "middle of the pack".

No suburb has an average number of rooms per dwelling over 9. The suburb with the highest number of rooms per dwelling is Suburb 364 with 8.780 rooms per dwelling.

The proportion of owner-occupied units built prior to 1940 is generally above 70. There are two exceptions: Suburb 204 with a proportion of 31.9 and Suburb 253 with a very low proportion of 8.4. Considering that the general range is from 2.9 to 100, we can say that most suburbs are on the high end when it comes to the proportion of old homes.

The weighted distances to five Boston employment centres take on a wide variety of values ranging from 1.8010 to 8.9067, with Suburb 257 on the low end and Suburb 253 on the high end. With a general range from 1.1296 to 12.1265, we can say that most Suburbs are in the "middle of the pack".

The index of accessibility to radial highways is most commonly 8 or 4.

Most values for full-value property-tax rate per $10,000 range between 265 and 330. There are two outliers: on the low end we have Suburb 204 with a tax rate of 224; and on the high end we have Suburb 364 with a tax rate of 666. With a general tax rate range between 187 and 711, we can say that most Suburbs are in the "middle of the pack"

The pupil-teacher ratio is generally lower than the overall mean. Most are 17.4. The lowest ones are in suburbs 257, 262 and 267, with a value of 13. This is very close to the overall minimum of 12.6. The highest ratio is that of Suburb 364, with a value of 20.2. This is very close to the overall maximum of 22.0.

The "black" index is over 350 for all Suburbs. Considering the overall range is between 0.32 and 396.9, we can say that all Suburbs are on the high-end of the spectrum.

The percentage of lower status of the population is below 7.5 for all Suburbs. With the overall range between 1.73 and 37.97, we can say that all Suburbs are on the low end of the spectrum.

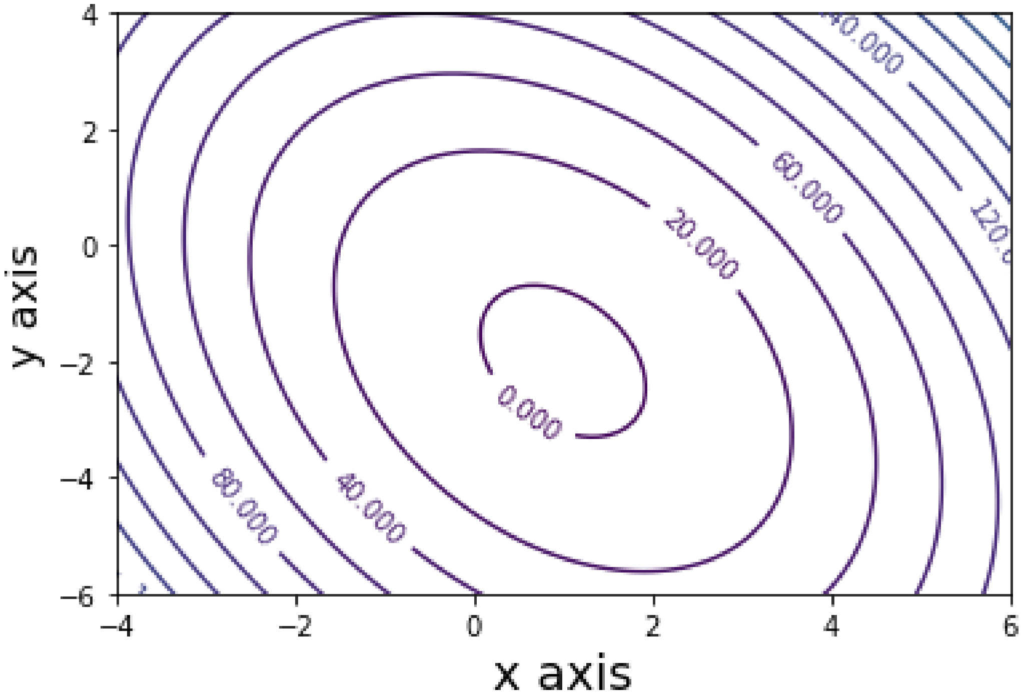
The median value of owner-occupied homes in $1000's is generally between 37.6 and 50.0. There is a single outlier, namely Suburb 364 with a median value of 21.9. With a general range from 5.0 to 50.0, we can say that almost all suburbs are on the high-end of the spectrum.

Points: 15

Mathematics and Probability Questions
1. Minimum and Maximum of a function

For the following function $(x,)=4\ x2+2\ y2-4\ x+6\ y+2\ x\ y+5$

a. Show the contour plot of the function



b. Find the partial derivative with respect to x and y

$$\frac{\delta f}{\delta x} = 8x - 4 + 2y$$
$$\frac{\delta f}{\delta y} = 4y + 6 + 2x$$

c. Find the minimum point of the function
Minimum point is where both partial derivatives are zero:

$$\frac{\delta f}{\delta x} = 8x - 4 + 2y = 0$$
$$\frac{\delta f}{\delta y} = 4y + 6 + 2x = 0$$

Solving the above linear system of equations gives:

$$x = 1 \; and \; y = \; -2$$

Therefore function reaches minimum at point (1, -2).

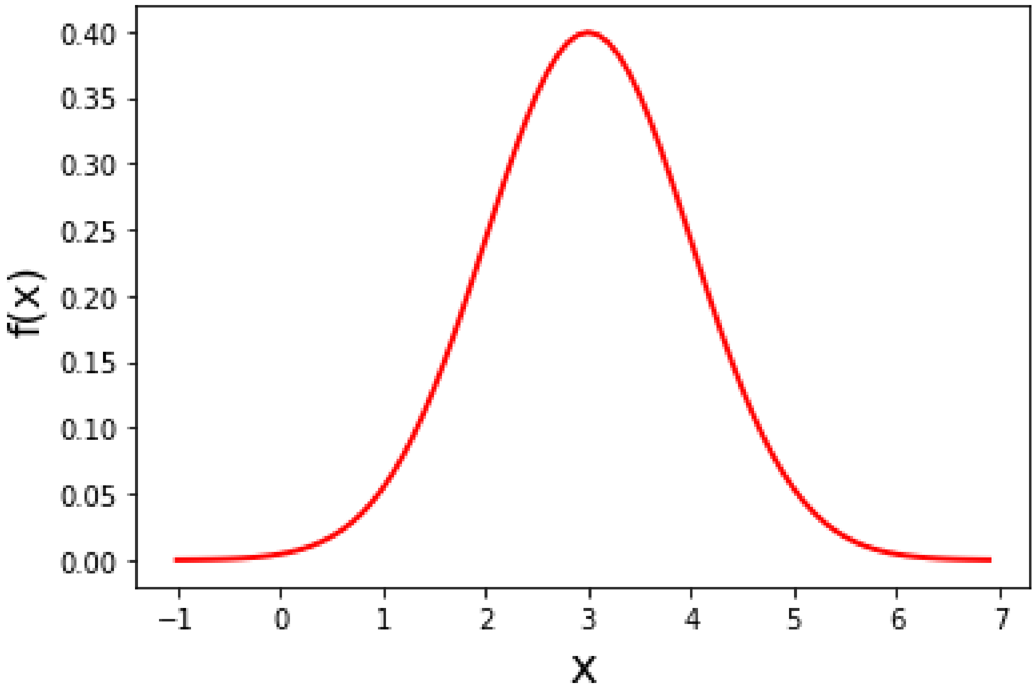
Hints:
- Check https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.contour.html for contour plot

Points: 10
2. Maximum Likelihood Estimation (MLE) (10)

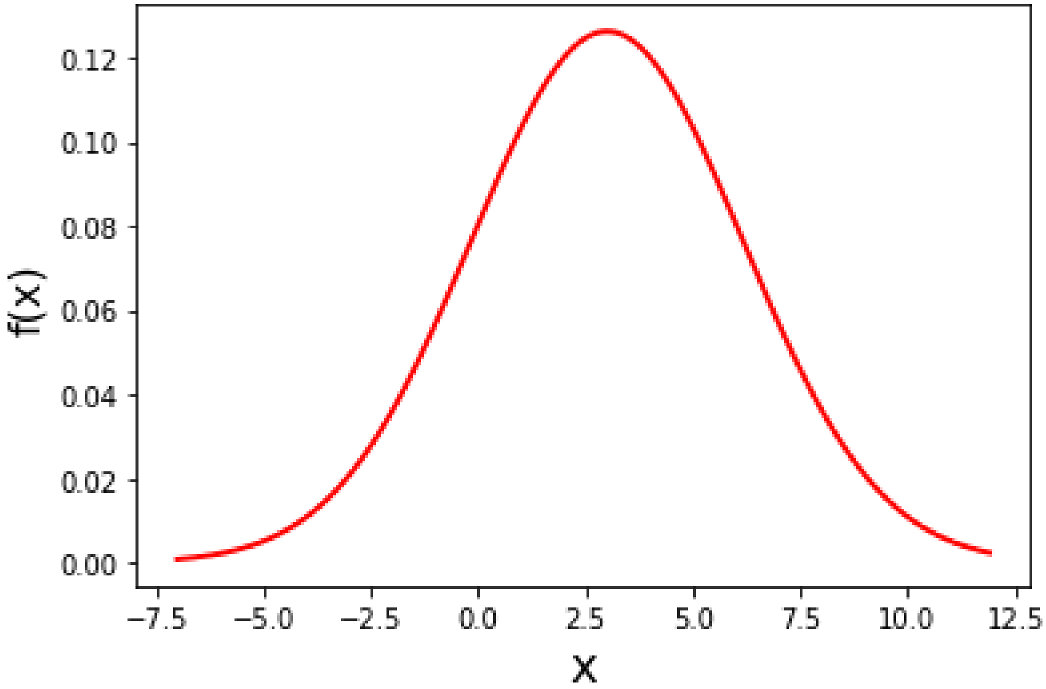
For the normal distribution with mean m and variance σ^2; its pdf is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

a. Plot the curve for m=3 and σ=1



b. Plot the curve for $m=3$ and $\sigma=\sqrt{10}$



c. Let's assume, we have N samples - $\{x1,2,....,xN\}$. The likelihood function is defined by

$$L(x_1, x_2, \dots, x_N; m, \sigma) = \prod_{i=1}^{N} f(x_i) = f(x_1) * f(x_2) * \dots * f(x_N)$$

find the MLE for m and σ.

To find MLE we need to differentiate the above equation with respect to m and with respect to δ. To do so more easily, we use the log trick (since log is a function that always increases as x increases, minimizing log(x) always minimizes x):

$$\mathrm{G} = \log(L(x_1, x_2, \dots, x_N; m, \sigma)) = \sum_{i}^{N} \ln\big(f(x_i)\big) = \ln(f(x_1)) + \cdots + \ln(f(x_N))$$

Substituting for f(x) gives:

$$\mathrm{G} = \sum_{i}^{N} ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - m)^2}{2\sigma^2}}\right) = \sum_{i}^{N}\left[ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x_i - m)^2}{2\sigma^2}\right]$$

$$= \frac{-N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i}^{N}(x_i - m)^2$$

Now take the partial derivative of G with respect to m, and set it to zero:

$$\frac{\delta G}{\delta m} = \frac{\delta}{\delta m}\left\{\frac{-N}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i}^{N}(x_i - m)^2\right\} = -\frac{1}{2\sigma^2}\sum_{i}^{N}\frac{\delta}{\delta m}\left[(x_i - m)^2\right] = 0$$

$$\sum_{i}^{N}\frac{\delta}{\delta m}[(x_i - m)^2] = 0$$

$$\sum_{i}^{N} 2m - 2x_i = 0$$

$$\sum_{i}^{N} m - x_i = 0$$

$$Nm - \sum_{i}^{N} x_i = 0$$

$$m = \frac{1}{N}\sum_{i}^{N} x_i$$

Or put in words, the value of m which gives the maximum likelihood estimation is precisely the arithmetic average of the values of X.

Similarly we take the partial derivative of G with respect to δ, and set it to zero:

$$\frac{\delta G}{\delta \sigma} = \frac{\delta}{\delta \sigma}\left\{\frac{-N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i}^{N}(x_i - m)^2\right\} = 0$$

$$\frac{\delta}{\delta \sigma}\left\{\frac{-N}{2}\ln(2\pi\sigma^2)\right\} - \frac{\delta}{\delta \sigma}\left\{\frac{1}{2\sigma^2}\sum_{i}^{N}(x_i - m)^2\right\} = 0$$

$$\left\{\frac{-N}{2}4\pi\sigma\frac{1}{2\pi\sigma^2}\right\} - \left\{\frac{-1}{\sigma^3}\sum_{i}^{N}(x_i - m)^2\right\} = 0$$

$$\left\{\frac{1}{\sigma^3}\sum_{i}^{N}(x_i - m)^2\right\} - \left\{\frac{N}{\sigma}\right\} = 0$$

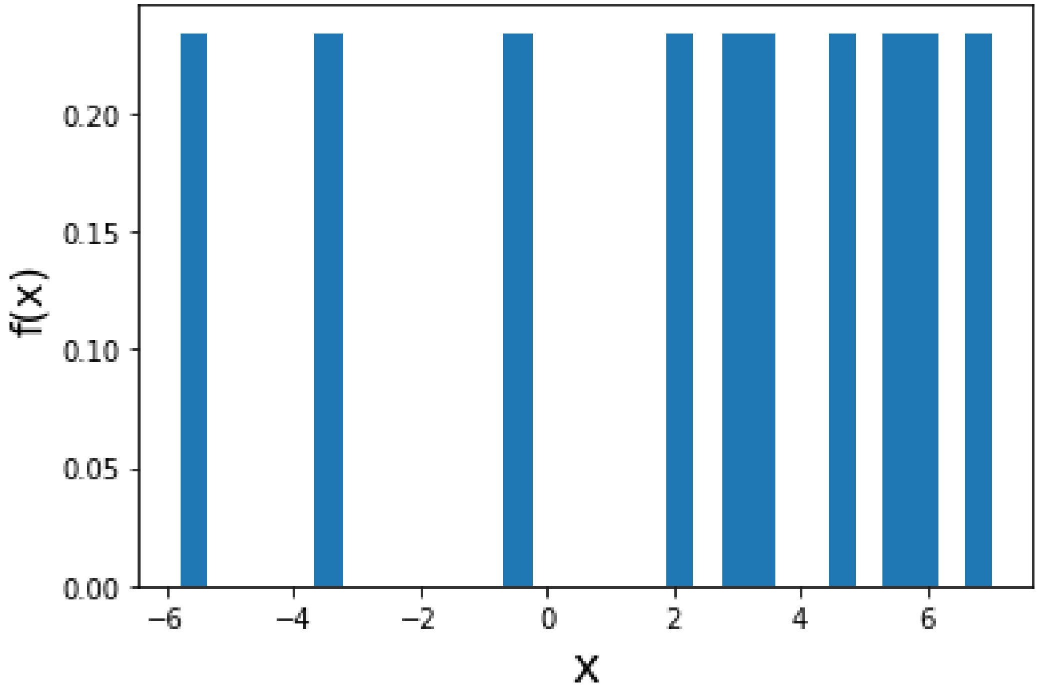
$$\frac{1}{\sigma^2}\sum_{i}^{N}(x_i - m)^2 = N$$

$$\sigma^2 = \frac{1}{N}\sum_{i}^{N}(x_i - m)^2$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{N}(x_i - m)^2}$$

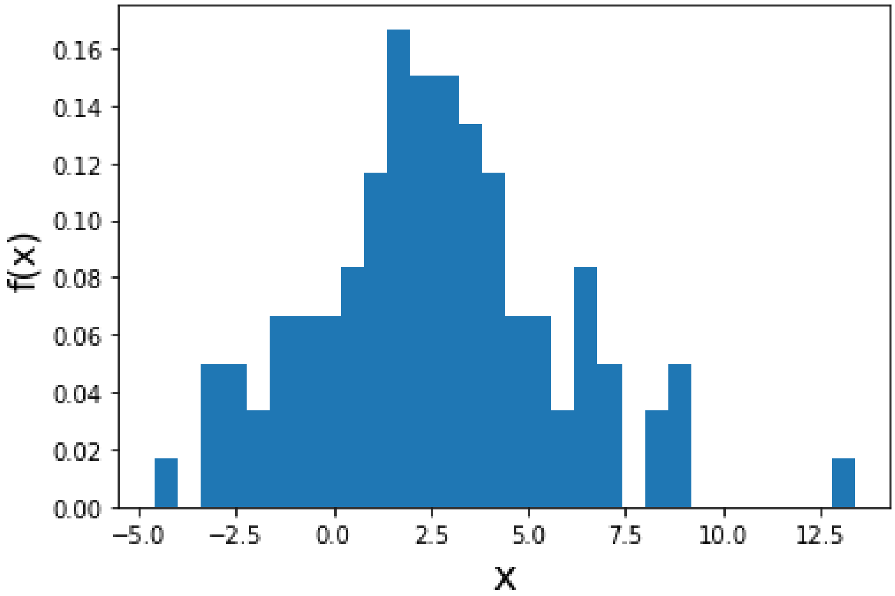
In other words our standard deviation should equal the standard deviation of the sample.

d. Draw 10 samples from a normal distribution with $m=3$ and $\sigma=\sqrt{10}$. Show its histogram, and calculate it's mean and variance



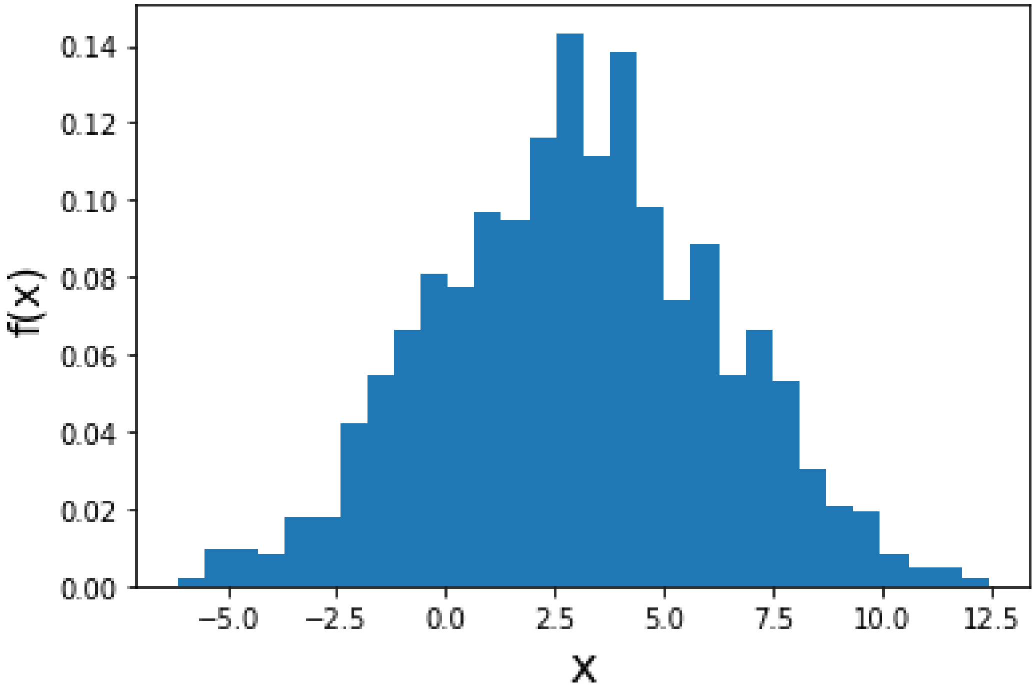
mean = 3.9068563140076558
variance = 7.40771644105278

e. Draw 100 samples from a normal distribution with $m=3$ and $\sigma=\sqrt{10}$. Show its histogram, and calculate it's mean and variance



mean = 2.585994853
var = 9.79893128473

f. Draw 1000 samples from a normal distribution with $m{=}3$ and $\sigma{=}\sqrt{10}$. Show its histogram, and calculate it's mean and variance



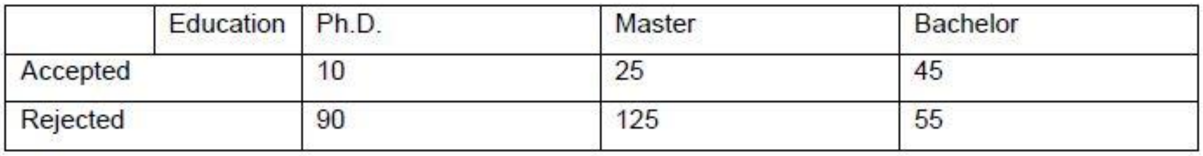
mean = 3.0292491051592516
variance = 10.388716183540502

Hints:

- Check
https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html
for random normal number generator

3. Bayes Rule and Conditional Distribution (15)

For a company, we have collected the following information for their hiring process over the last 10 years.



	Education	Ph.D.	Master	Bachelor
Accepted		10	25	45
Rejected		90	125	55

a. What is the probability of an applicant to have PhD?

$$p(PhD) = \frac{100}{100 + 150 + 100} = \frac{100}{350} = 0.2857$$

b. What is probability of being accepted if you have at least a Master Degree?

$$p(Accepted \mid Master) = \frac{25}{150} = 0.16667$$

c. What is probability of being accepted?

$$p(Accepted) = \frac{10 + 25 + 45}{350} = \frac{80}{350} = 0.22857$$

d. What is probability of having Ph.D. if the candidate being accepted?

$$p(PhD\ |Accepted) = \frac{10}{10 + 25 + 45} = \frac{10}{80} = 0.125$$

Code Section:

Theoretical Question 3a)

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

X = np.arange(0, 20, .1)

bias = 1 / X
variance = 2**((((X-1)*0.7)**2))
trainingE = -(X-2)**3 + 3.4
testE = bias + variance + 2
bayes = 3*X/X

plt.xlabel('flexibility')
plt.ylabel('f(x)')

plt.xlim(0, 4)
plt.ylim(0,10)

plt.plot(X, bias , label='bias^2')
plt.plot(X, variance, label='var')
plt.plot(X, trainingE, label='trainingE')
plt.plot(X, testE, label='testE')
plt.plot(X, bayes, label='bayes')
plt.legend()

fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 7.5)
```

Applied Question 1)

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

```
!pip install pandas
import pandas as pd

data = pd.read_csv("Auto.csv")
mpg = sorted(data['mpg'])
cylinders = sorted(data['cylinders'])
displacement = sorted(data['displacement'])
horsepower = sorted(pd.to_numeric(data[data['horsepower'] != '?']['horsepower']))
#print(horsepower)
weight = sorted(data['weight'])
acceleration = sorted(data['acceleration'])
year = sorted(data['year'])

#part b)
print("Mpg range: " + str(mpg[-1]- mpg[0]))
print("cylinders range: "+ str(cylinders[-1] - cylinders[0]))
print("displacement range: "+ str(displacement[-1] - displacement[0]))
print("horsepower range: "+ str(horsepower[-1] - horsepower[0]))
print("weight range: "+ str(weight[-1] - weight[0]))
print("acceleration range:" +str(acceleration[-1] - acceleration[0]))
print("year range:" + str(year[-1] - year[0]))

#part c)

print("Mpg mean:" + format(np.mean(mpg), ".5f") + "\t\t Mpg variance "+ (format(np.var(mpg), ".5f")))
print("cylinders mean:" + format(np.mean(cylinders), ".5f") + "\t\t cylinders variance "+ (format(np.var(cylinders), ".5f")))
print("displacement mean:" + format(np.mean(displacement), ".5f") + "\t displacement variance "+ (format(np.var(displacement), ".5f")))
print("horsepower mean:" + format(np.mean(horsepower), ".5f") + "\t horsepower variance "+ (format(np.var(horsepower), ".5f")))
print("weight mean:" + format(np.mean(weight), ".5f") + "\t\t weight variance "+ (format(np.var(weight), ".5f")))
print("acceleration mean:" + format(np.mean(acceleration), ".5f") + "\t acceleration variance "+ (format(np.var(acceleration), ".5f")))
print("year mean:" + format(np.mean(year), ".5f") + "\t\t year variance "+ (format(np.var(year), ".5f")))


#part d)
dataPrime = pd.concat([data[1:10], data[85:]])
#print(dataPrime)
data = pd.read_csv("Auto.csv")
```

```
mpg = sorted(data['mpg'])
cylinders = sorted(data['cylinders'])
displacement = sorted(data['displacement'])
horsepower = sorted(pd.to_numeric(data[data['horsepower'] != '?']['horsepo
wer']))
#print(horsepower)
weight = sorted(data['weight'])
acceleration = sorted(data['acceleration'])
year = sorted(data['year'])

print("Mpg range: " + str(mpg[-1]-
 mpg[0]) + "\t\t\t Mpg mean:" + format(np.mean(mpg), ".5f") + "\t\t Mpg va
riance "+ (format(np.var(mpg), ".5f")))
print("cylinders range: "+ str(cylinders[-1] -
 cylinders[0]) + "\t\t cylinders mean:" + format(np.mean(cylinders), ".5f"
) + "\t\t cylinders variance "+ (format(np.var(cylinders), ".5f")))
print("displacement range: "+ str(displacement[-1] -
 displacement[0]) + "\t displacement mean:" + format(np.mean(displacement)
, ".5f") + "\t displacement variance "+ (format(np.var(displacement), ".5f
")))
print("horsepower range: "+ str(horsepower[-1] -
 horsepower[0]) + "\t\t horsepower mean:" + format(np.mean(horsepower), ".
5f") + "\t horsepower variance "+ (format(np.var(horsepower), ".5f")))
print("weight range: "+ str(weight[-1] -
 weight[0]) + "\t\t weight mean:" + format(np.mean(weight), ".5f") + "\t\t
 weight variance "+ (format(np.var(weight), ".5f")))
print("acceleration range:" +str(acceleration[-1] -
 acceleration[0]) + "\t\t acceleration mean:" + format(np.mean(acceleratio
n), ".5f") + "\t acceleration variance "+ (format(np.var(acceleration), ".
5f")))
print("year range:" + str(year[-1] -
 year[0]) + "\t\t\t year mean:" + format(np.mean(year), ".5f") + "\t\t yea
r variance "+ (format(np.var(year), ".5f")))

#part e)
fig, ax = plt.subplots(nrows= 2, ncols= 2)

#figure(figsize=(8, 6), dpi=80))

plt.sca(ax[0,0])
plt.xlabel('cylinders')
plt.hist(data['cylinders'])  # density=False would make counts

plt.sca(ax[0,1])
plt.yticks(np.arange(0, acceleration[-1], 5.0))
plt.xlabel('weight')
plt.ylabel('acceleration')
plt.scatter(data['weight'],data['acceleration'])

plt.sca(ax[1,0])
plt.yticks(np.arange(0, mpg[-1], 5.0))
plt.xlabel('weight')
plt.ylabel('mpg')
plt.scatter(data['weight'],data['mpg'])

plt.sca(ax[1,1])
plt.xticks(np.arange(0, horsepower[-1], 50))
plt.yticks(np.arange(0, mpg[-1], 5.0))
plt.xlabel('horsepower')
plt.ylabel('mpg')
modifiedHP = pd.to_numeric(data[data['horsepower'] != '?']['horsepower'])
modifiedMPG = pd.to_numeric(data[data['horsepower'] != '?']['mpg'])
plt.scatter(modifiedHP, modifiedMPG)


fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 10)
```

Applied Question 2

```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
import math

!pip install pandas
import pandas as pd

data = pd.read_csv("Boston.csv")

#part b)
"""
fig, ax = plt.subplots(nrows= 4, ncols= 2)

plt.sca(ax[0,0])
plt.xlabel('proportion of buildings older than 1940')
plt.ylabel('percentage of lower status population')
plt.scatter(data['age'], data['lstat'])
plt.sca(ax[0,1])
plt.xlabel('percentage of non-business acres')
plt.ylabel('nitric oxide concentration')
plt.scatter(data['nox'],data['indus'])

plt.sca(ax[1,0])
plt.xlabel('average number of rooms per dwelling')
plt.ylabel('median value of occupied homes')
plt.scatter(data['rm'],data['medv'])

plt.sca(ax[1,1])
plt.xlabel('percentage of lower status population')
plt.ylabel('crime rate')
plt.scatter(data['lstat'],data['crim'])

plt.sca(ax[2,0])
plt.xlabel('proportion of blacks')
plt.ylabel('percentage of lower status population')
plt.scatter(0.63 + (data['black'] / 1000) ** 0.5 , data['lstat'])

plt.sca(ax[2,1])
plt.xlabel('proportion of units older than 1940')
plt.ylabel('median value of occupied homes')
plt.scatter(data['age'],data['medv'])

plt.sca(ax[3,0])
plt.xlabel('proportion of blacks')
plt.ylabel('crime rate')
plt.scatter(0.63 + (data['black'] / 1000)**0.5 , data['crim'])

plt.sca(ax[3,1])
plt.xlabel('proportion of units older than 1940')
plt.ylabel('crime rate')
plt.scatter(data['age'],data['crim'])

fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 15)
"""

"""
#part c)
fig, ax = plt.subplots(nrows= 4, ncols= 2)

plt.sca(ax[0,0])
plt.xlabel('percentage of non-business acres')
plt.ylabel('crime rate')
plt.scatter(data['indus'], data['crim'])

plt.sca(ax[0,1])
plt.xlabel('average number of rooms per dwelling')
plt.ylabel('crime rate')
plt.scatter(data['rm'],data['crim'])

plt.sca(ax[1,0])
plt.xlabel('proportion of units older than 1940')
plt.ylabel('crime rate')
plt.scatter(data['age'],data['crim'])

plt.sca(ax[1,1])
plt.xlabel('weighted distance to 5 Boston employment centers')
plt.ylabel('crime rate')
plt.scatter(data['dis'],data['crim'])

plt.sca(ax[2,0])
plt.xlabel('full-value property-tax rate per $10,000')
plt.ylabel('crime rate')
plt.scatter(data['tax'],data['crim'])

plt.sca(ax[2,1])
plt.xlabel('proportion of blacks')
plt.ylabel('crime rate')
plt.scatter(0.63 + (data['black'] / 1000) ** 0.5 , data['crim'])

plt.sca(ax[3,0])
plt.xlabel('percentage of lower status population')
plt.ylabel('crime rate')
plt.scatter(data['lstat'],data['medv'])

plt.sca(ax[3,1])
plt.xlabel('Median value of owner-occupied homes in $1000s')
plt.ylabel('crime rate')
plt.scatter(data['medv'] , data['crim'])

#plt.hist(data['tax'], bins = len(set(data['tax'])))


fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 15)
"""
```

```
#part d)
'''
crim = sorted(data['crim'])
tax = sorted(data['tax'])
ptratio = sorted(data['ptratio'])

print("Crime rate range: " + str(crim[-1]- crim[0]))
print("Tax rate range: "+ str(tax[-1] - tax[0]))
print("Pupil-teacher ratio range: "+ str(ptratio[-1] - ptratio[0]))

fig, ax = plt.subplots(nrows= 2, ncols= 2)
plt.sca(ax[0,0])
plt.xlabel('crime rate')
plt.ylabel('number of occurances')
plt.hist(crim)

plt.sca(ax[0,1])
plt.xlabel('Tax rate')
plt.ylabel('number of occurances')
plt.hist(tax)

plt.sca(ax[1,0])
plt.xlabel('Pupil-teacher ration')
plt.ylabel('number of occurances')
plt.hist(ptratio)

fig = matplotlib.pyplot.gcf()
fig.set_size_inches(10, 15)
'''

#part e)
'''
charles = ((data[data['chas'] != 0]['chas']))
print(len(charles))
'''

#part f)
'''
ptratio = sorted(data['ptratio'])
print(ptratio[(math.floor(len(ptratio)/2))])
'''

#part g)
'''
lowestIncome = data[data['medv'] == sorted(data['medv'])[0]]
```

```
print(lowestIncome)
crim = sorted(data['crim'])
zn = sorted(data['zn'])
indus = sorted(data['indus'])
chas = sorted(data['chas'])
nox = sorted(data['nox'])
rm = sorted(data['rm'])
age = sorted(data['age'])
dis = sorted(data['dis'])
rad = sorted(data['rad'])
tax = sorted(data['tax'])
ptratio = sorted(data['ptratio'])
black =  sorted(data['black'])
lstat = sorted(data['lstat'])
medv = sorted(data['medv'])

print(lowestIncome['crim'])
print("Crime range " + str(crim[0]) + " - " + str(crim[-1]))
print(lowestIncome['zn'])
print("zn range " + str(zn[0]) + " - " + str(zn[-1]))
print(lowestIncome['indus'])
print("indus range " + str(indus[0]) + " - " + str(indus[-1]))
print(lowestIncome['chas'])
print("chas range " + str(chas[0]) + " - " + str(chas[-1]))
print(lowestIncome['nox'])
print("nox range " + str(nox[0]) + " - " + str(nox[-1]))
print(lowestIncome['rm'])
print("rm range " + str(rm[0]) + " - " + str(rm[-1]))
print(lowestIncome['age'])
print("age range " + str(age[0]) + " - " + str(age[-1]))
print(lowestIncome['dis'])
print("dis range " + str(dis[0]) + " - " + str(dis[-1]))
print(lowestIncome['rad'])
print("rad range " + str(rad[0]) + " - " + str(rad[-1]))
print(lowestIncome['tax'])
print("tax range " + str(tax[0]) + " - " + str(tax[-1]))
print(lowestIncome['ptratio'])
print("ptratio range " + str(ptratio[0]) + " - " + str(ptratio[-1]))
print(lowestIncome['black'])
print("black range " + str(black[0]) + " - " + str(black[-1]))
print(lowestIncome['lstat'])
print("lstat range " + str(lstat[0]) + " - " + str(lstat[-1]))
print(lowestIncome['medv'])
print("medv range " + str(medv[0]) + " - " + str(medv[-1]))
```

```
'''
#part h)
print(len(data[data['rm'] > 7.0]))
modifiedData = data[data['rm'] > 8.0]
print(len(modifiedData))
crim = sorted(data['crim'])
zn = sorted(data['zn'])
indus = sorted(data['indus'])
chas = sorted(data['chas'])
nox = sorted(data['nox'])
rm = sorted(data['rm'])
age = sorted(data['age'])
dis = sorted(data['dis'])
rad = sorted(data['rad'])
tax = sorted(data['tax'])
ptratio = sorted(data['ptratio'])
black =  sorted(data['black'])
lstat = sorted(data['lstat'])
medv = sorted(data['medv'])

print(modifiedData['crim'])
print("Crime range " + str(crim[0]) + " - " + str(crim[-1]))
print(modifiedData['zn'])
print("zn range " + str(zn[0]) + " - " + str(zn[-1]))
print(modifiedData['indus'])
print("indus range " + str(indus[0]) + " - " + str(indus[-1]))
print(modifiedData['chas'])
print("chas range " + str(chas[0]) + " - " + str(chas[-1]))
print(modifiedData['nox'])
print("nox range " + str(nox[0]) + " - " + str(nox[-1]))
print(modifiedData['rm'])
print("rm range " + str(rm[0]) + " - " + str(rm[-1]))
print(modifiedData['age'])
print("age range " + str(age[0]) + " - " + str(age[-1]))
print(modifiedData['dis'])
print("dis range " + str(dis[0]) + " - " + str(dis[-1]))
print(modifiedData['rad'])
print("rad range " + str(rad[0]) + " - " + str(rad[-1]))
print(modifiedData['tax'])
print("tax range " + str(tax[0]) + " - " + str(tax[-1]))
print(modifiedData['ptratio'])
print("ptratio range " + str(ptratio[0]) + " - " + str(ptratio[-1]))
print(modifiedData['black'])
print("black range " + str(black[0]) + " - " + str(black[-1]))
print(modifiedData['lstat'])
print("lstat range " + str(lstat[0]) + " - " + str(lstat[-1]))
print(modifiedData['medv'])
print("medv range " + str(medv[0]) + " - " + str(medv[-1]))
```

Mathematical Question 1

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

X = np.arange(-10, 10, .1)
Y = np.arange(-10, 10, .1)
Z = np.zeros((X.size,Y.size))
XX, YY = np.meshgrid(X, Y)


for i, x in enumerate(X):
  for j, y in enumerate(Y):
    Z[j,i] = 4*x**2+2*y**2-4*x+6*y+2*x*y+5

# Plot the points using matplotlib
print(Z)
plt.xlim([-4, 6])
plt.ylim([-6, 4])
CS = plt.contour(XX, YY, Z, 50)

plt.xlabel('x axis', fontsize=18)
plt.ylabel('y axis', fontsize=16)

plt.clabel(CS, inline=True, fontsize=10)
plt.show()
```

Mathematical Question 2

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

mu, sigma = 3, math.sqrt(10)
s = np.random.normal(mu, sigma, 1000)
print(np.mean(s))
print(np.var(s))

bins = np.arange(-7, 12, .1)
count, bins, ignored = plt.hist(s, 30, density=True)
```

```
#plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *
#               np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
#         linewidth=2, color='r')

plt.xlabel('x', fontsize=18)
plt.ylabel('f(x)', fontsize=16)

plt.show()
```