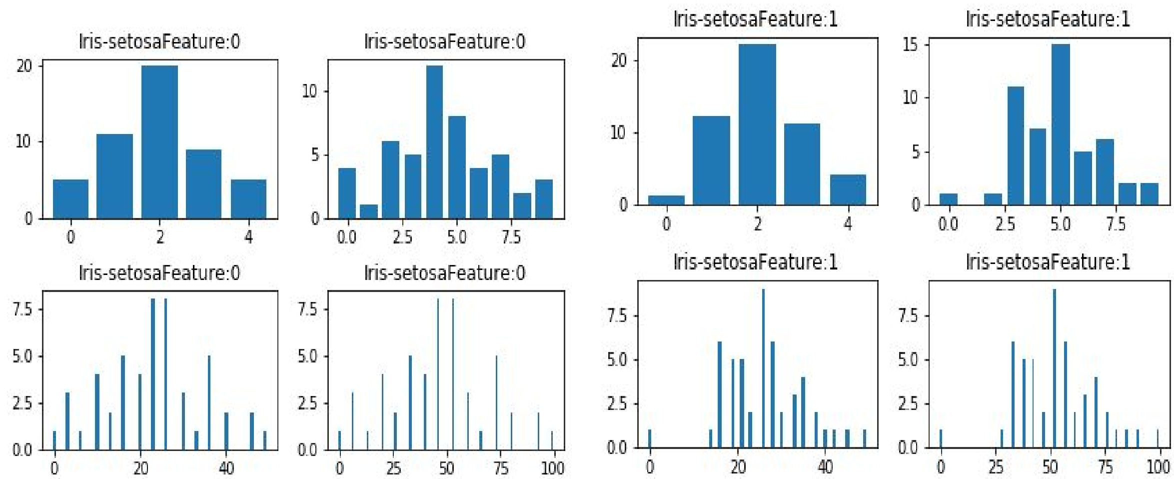
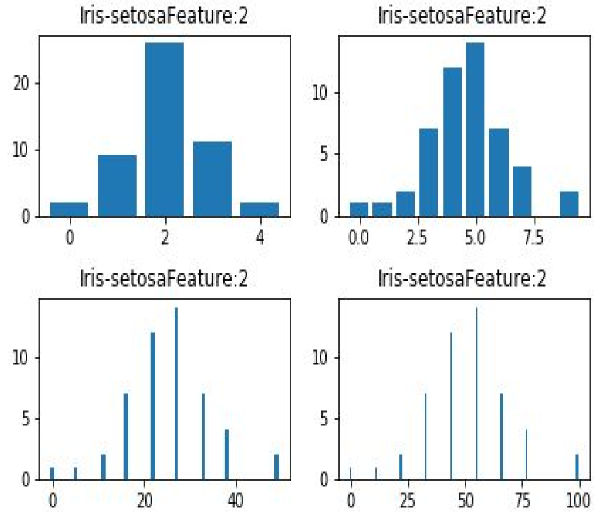
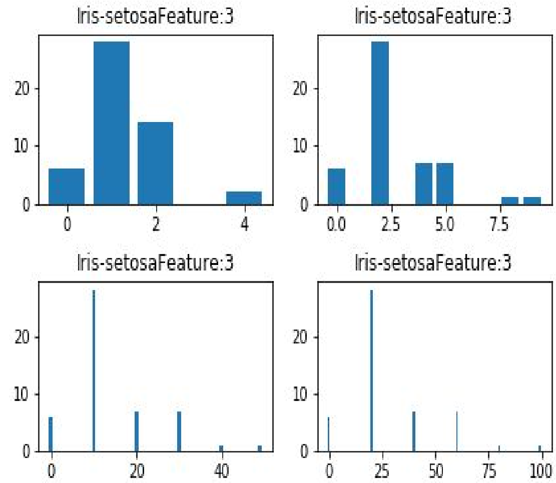
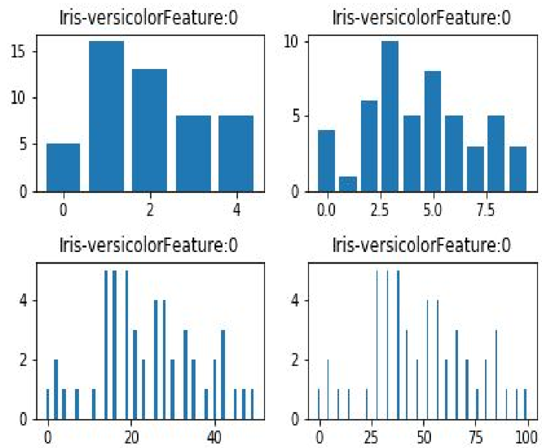
Assignment1
Name: I-Hsien Huang
SID: 862057578
Question 1: Feature distribution [35%]
result: Based on the class , the features and the bins and the data set
[IRIS DATA]
I know the professor ask to put the range of x_label, but it will make the picture become very huge and not easy to view the labels on it. Therefore, I put my original pictures here, and the version with X_label in the compressed files with my code.

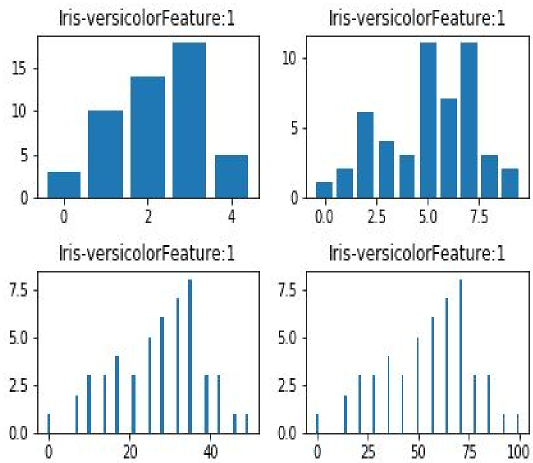
	iris-setosa	versicolor	virginical
Feature0	symmetric/bimodel	negaive skewed/ multi-model	symmetric/Multi-model
Feature1	symmetric/Multi-model	negaive skewed/ multi-model	symmetric/uni-imodel
Feature2	symmetric/bimodel	negaive skewed/ multi-model	skewed/bi-model
Feature3	postivie skewed/uni-imodel	symmetric/ bi-model	symmetric/bi-imodel

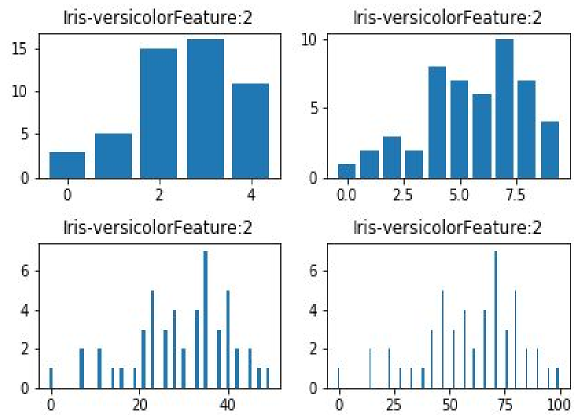


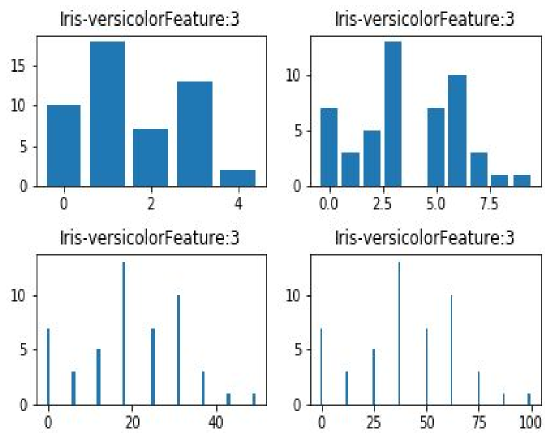

Iris-setosaFeature:2
Iris-setosaFeature:2
Iris-setosaFeature:2
Iris-setosaFeature:2

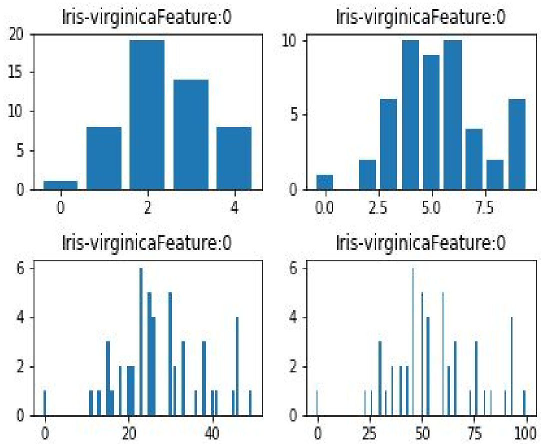

Iris-setosaFeature:3
Iris-setosaFeature:3
Iris-setosaFeature:3
Iris-setosaFeature:3

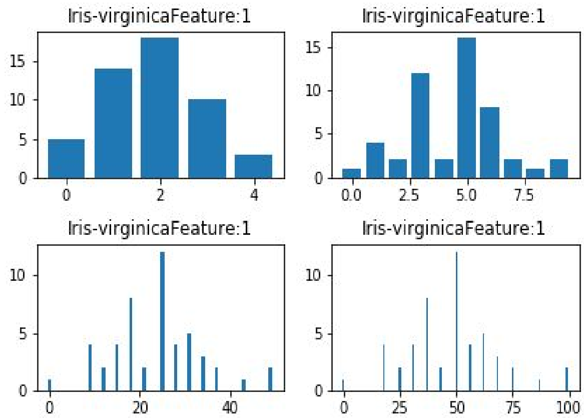

Iris-versicolorFeature:0
Iris-versicolorFeature:0
Iris-versicolorFeature:0
Iris-versicolorFeature:0

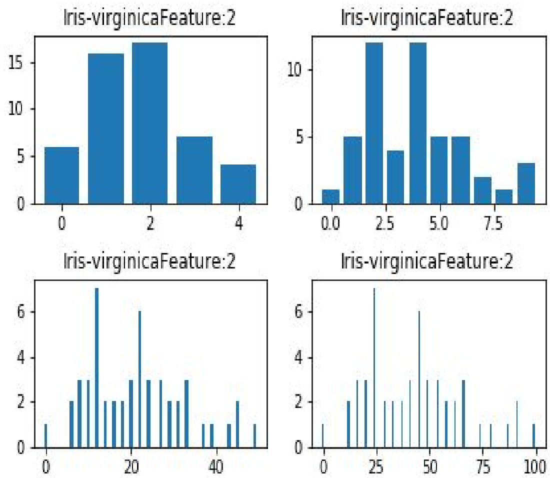

Iris-versicolorFeature:1
Iris-versicolorFeature:1
Iris-versicolorFeature:1
Iris-versicolorFeature:1

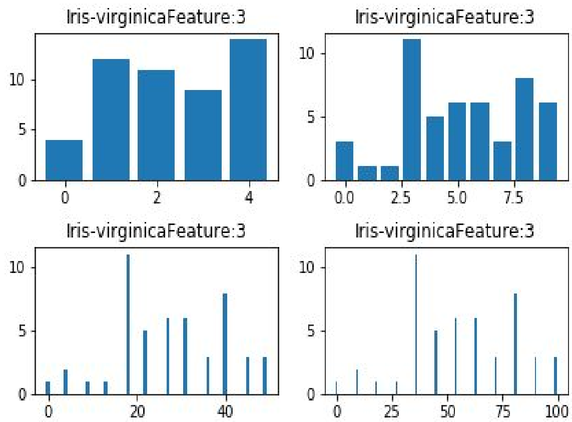

Iris-versicolorFeature:2
Iris-versicolorFeature:2
Iris-versicolorFeature:2
Iris-versicolorFeature:2


Iris-versicolorFeature:3
Iris-versicolorFeature:3
Iris-versicolorFeature:3
Iris-versicolorFeature:3



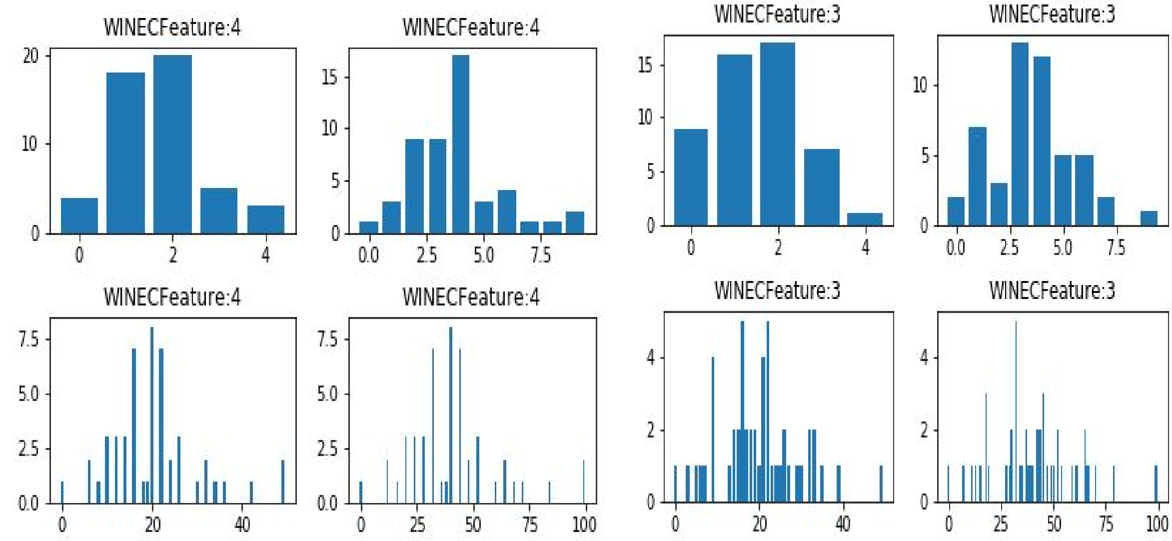


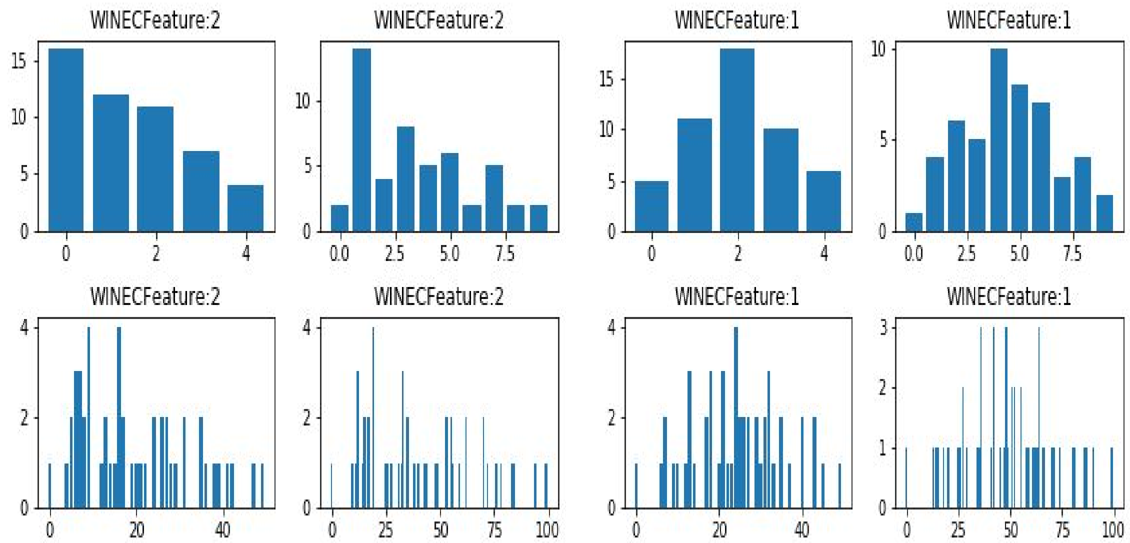


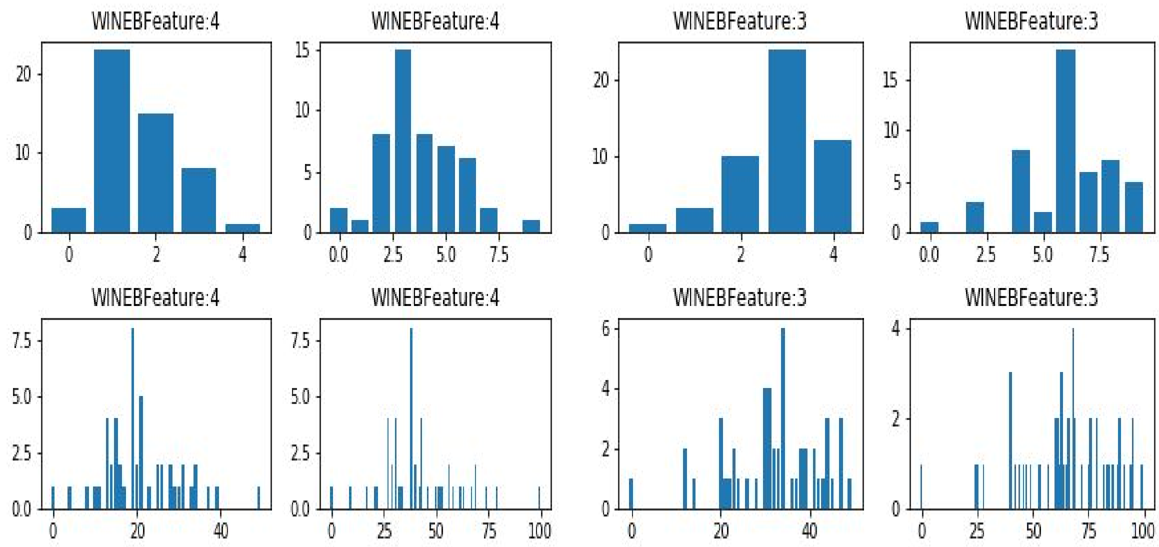


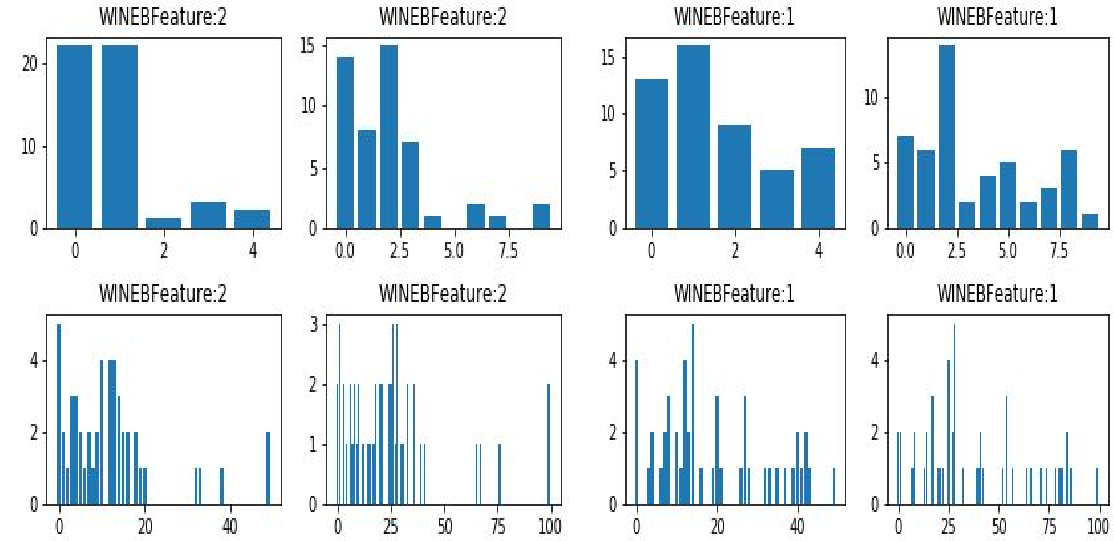
[WINE DATA]

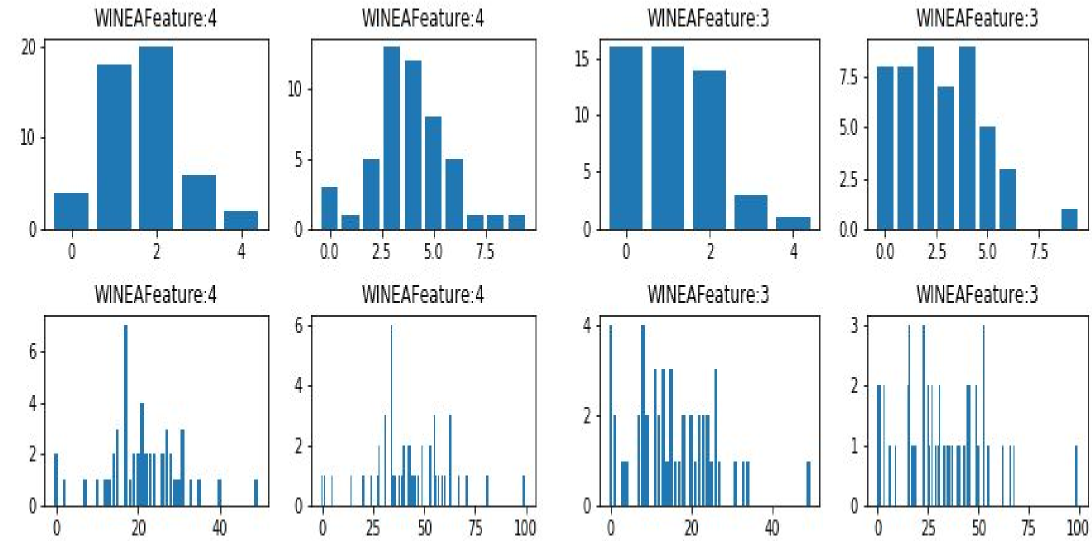
	wine A	wine B	Wine C
Feature1	symmetric/multi-model	skewed/ multi-model	symmetric/Multi-model
Feature2	skewed/ bi-model	skewed/ multi-model	skewed/bi-imodel
Feature3	skewed/ multi-model	skewed/ multi-model	skewed/uni-model
Feature4	skewed/ multi-model	skewed/ uni-model	skewed/multi-imodel

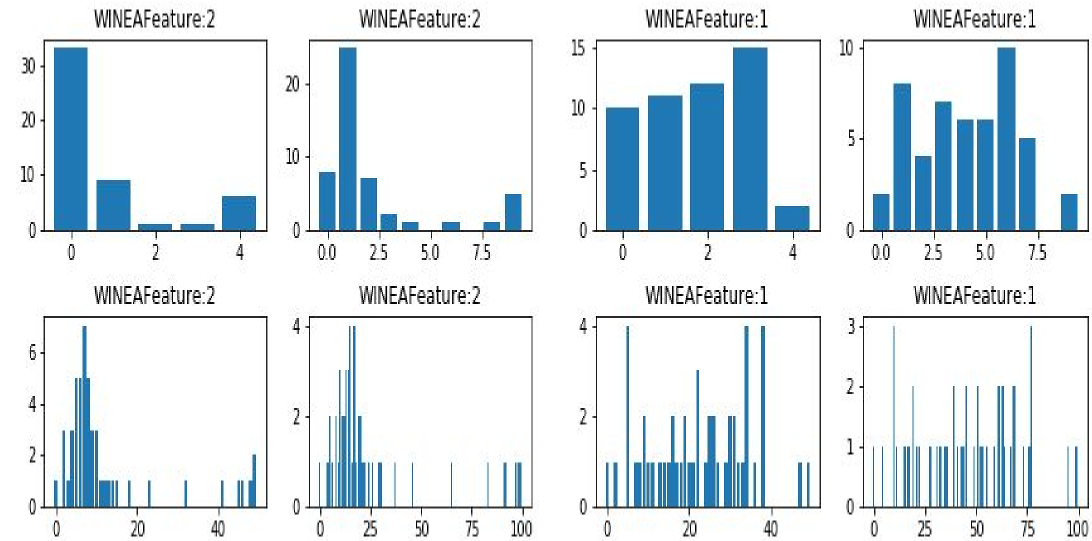

WINECFeature:4
WINECFeature:4
WINECFeature:3
WINECFeature:3
WINECFeature:4
WINECFeature:4
WINECFeature:3
WINECFeature:3


WINECFeature:2
WINECFeature:2
WINECFeature:1
WINECFeature:1
WINECFeature:2
WINECFeature:2
WINECFeature:1
WINECFeature:1


WINEBFeature:4
WINEBFeature:4
WINEBFeature:3
WINEBFeature:3
WINEBFeature:4
WINEBFeature:4
WINEBFeature:3
WINEBFeature:3

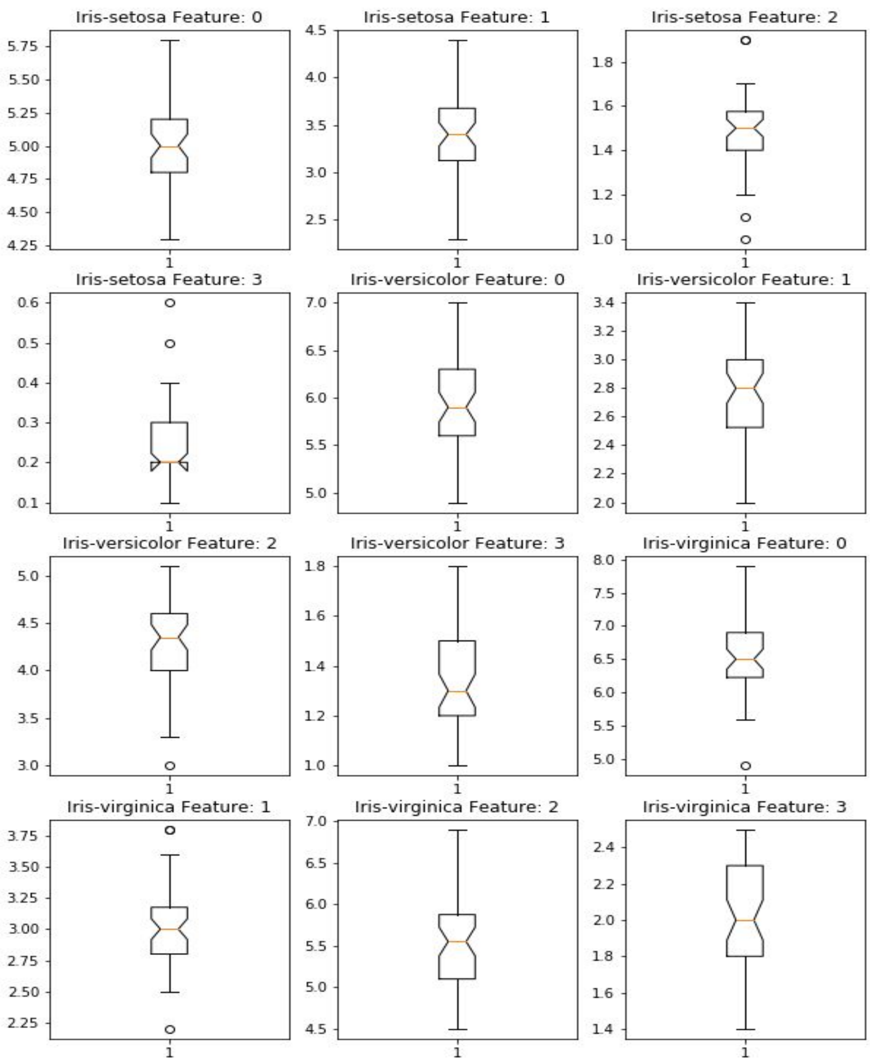




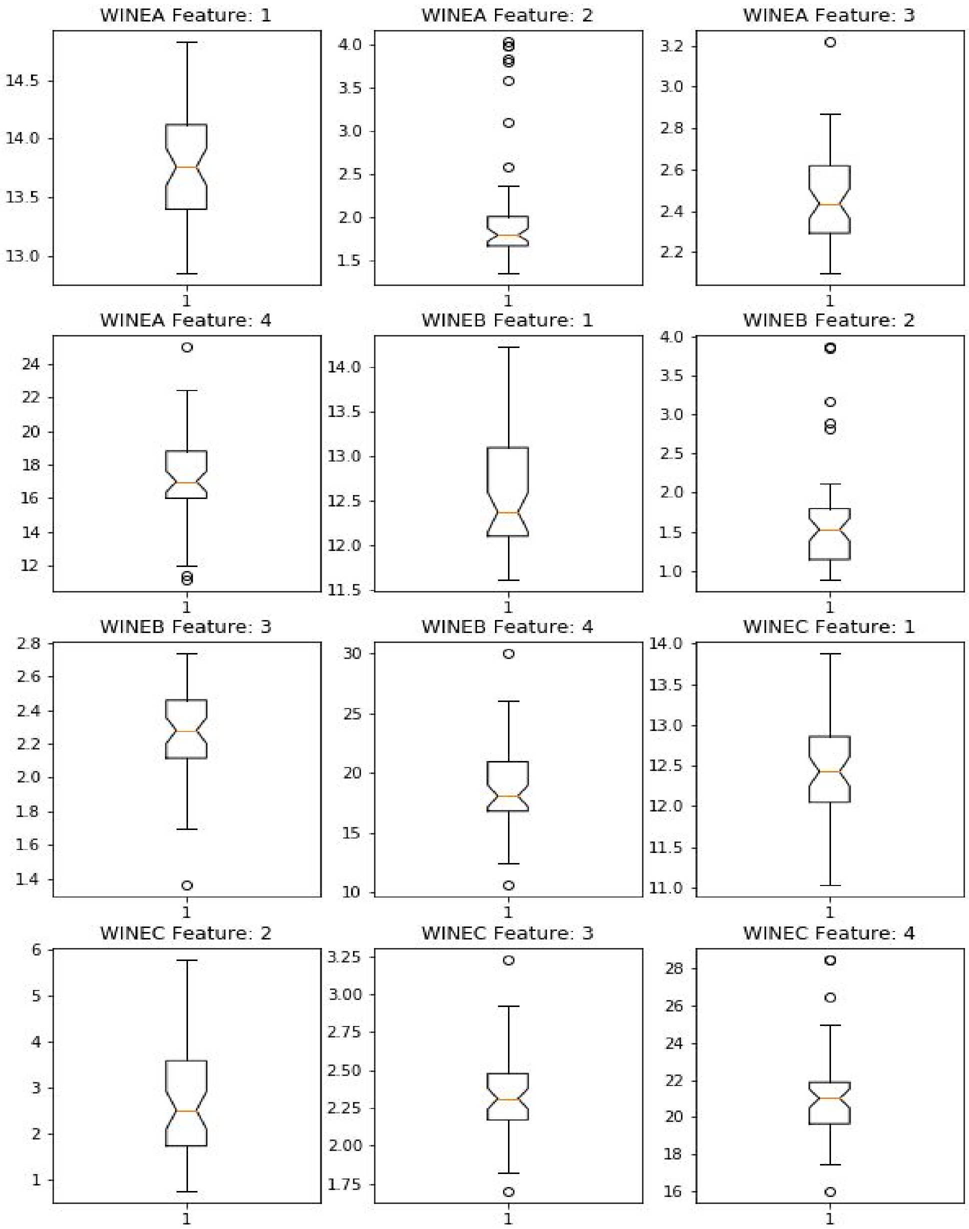


[SUM UP] Basically, the answer might be affected by the subjective opinion, for the reason that there are so many model between two kinds of model.

2) [20%] For the same data (organized in the same way as above), plot their Box-plots. You may use a library function for this
[IRIS]



[WINE]


WINEA Feature: 1
WINEA Feature: 2
WINEA Feature: 3
WINEA Feature: 4
WINEB Feature: 1
WINEB Feature: 2
WINEB Feature: 3
WINEB Feature: 4
WINEC Feature: 1
WINEC Feature: 2
WINEC Feature: 3
WINEC Feature: 4

Question 2: Relations between features and data points [60%]

1) [20%] Correlation Plots :

1a)calculate in my code

1b)

[IRIS]

[1.0, -0.1094, 0.8718, 0.818]

[-0.1094, 1.0, -0.4205, -0.3565]

[0.8718, -0.4205, 1.0, 0.9628]

[0.818, -0.3565, 0.9628, 1.0]]

[WINE]

[[1.0, 0.0944, 0.2115, -0.3102, 0.2708, 0.2891, 0.2368, -0.1559, 0.1367, 0.5464, -0.0717, 0.0723, 0.6437],

[0.0944, 1.0, 0.164, 0.2885, -0.0546, -0.3352, -0.411, 0.293, -0.2207, 0.249, -0.5613, -0.3687, -0.192],

[0.2115, 0.164, 1.0, 0.4434, 0.2866, 0.129, 0.1151, 0.1862, 0.0097, 0.2589, -0.0747, 0.0039, 0.2236],

[-0.3102, 0.2885, 0.4434, 1.0, -0.0833, -0.3211, -0.3514, 0.3619, -0.1973, 0.0187, -0.274, -0.2768, -0.4406],

[0.2708, -0.0546, 0.2866, -0.0833, 1.0, 0.2144, 0.1958, -0.2563, 0.2364, 0.2, 0.0554, 0.066, 0.3934],

[0.2891, -0.3352, 0.129, -0.3211, 0.2144, 1.0, 0.8646, -0.4499, 0.6124, -0.0551, 0.4337, 0.6999, 0.4981],

[0.2368, -0.411, 0.1151, -0.3514, 0.1958, 0.8646, 1.0, -0.5379, 0.6527, -0.1724, 0.5435, 0.7872, 0.4942],

[-0.1559, 0.293, 0.1862, 0.3619, -0.2563, -0.4499, -0.5379, 1.0, -0.3658, 0.1391, -0.2626, -0.5033, -0.3114],

[0.1367, -0.2207, 0.0097, -0.1973, 0.2364, 0.6124, 0.6527, -0.3658, 1.0, -0.0252, 0.2955, 0.5191, 0.3304],

[0.5464, 0.249, 0.2589, 0.0187, 0.2, -0.0551, -0.1724, 0.1391, -0.0252, 1.0, -0.5218, -0.4288, 0.3161],

[-0.0717, -0.5613, -0.0747, -0.274, 0.0554, 0.4337, 0.5435, -0.2626, 0.2955, -0.5218, 1.0, 0.5655, 0.2362],

[0.0723, -0.3687, 0.0039, -0.2768, 0.066, 0.6999, 0.7872, -0.5033, 0.5191, -0.4288, 0.5655, 1.0, 0.3128],

[0.6437, -0.192, 0.2236, -0.4406, 0.3934, 0.4981, 0.4942, -0.3114, 0.3304, 0.3161, 0.2362, 0.3128, 1.0]]

1c)

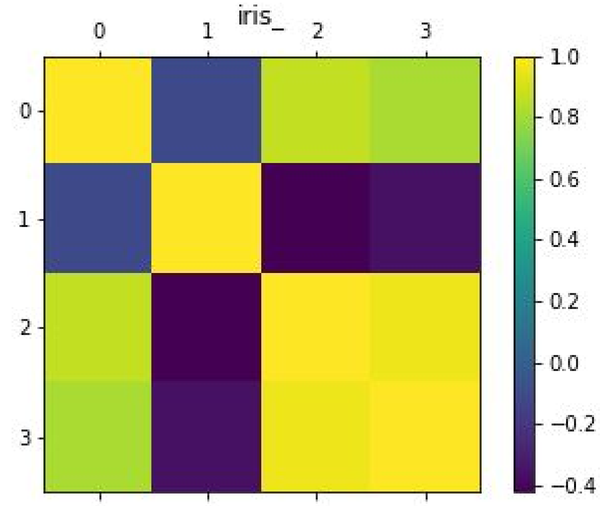
[iris] minumum (4*4-4)/2 = 6 (upper triangle and the diagonal is equal to 1)

[wine] (13*13-13)/2 =78

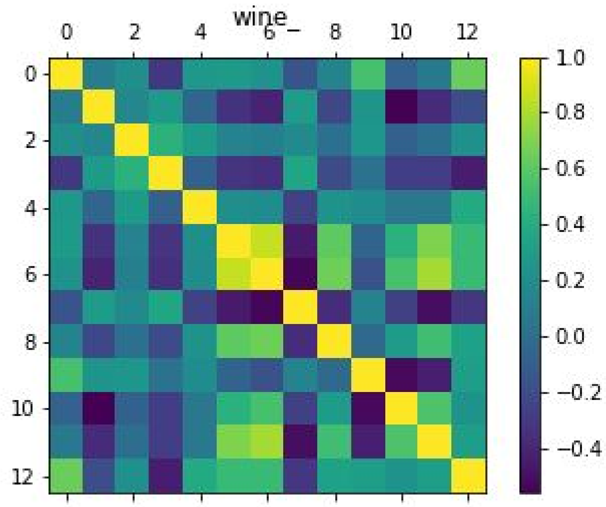
1d)

[iris]

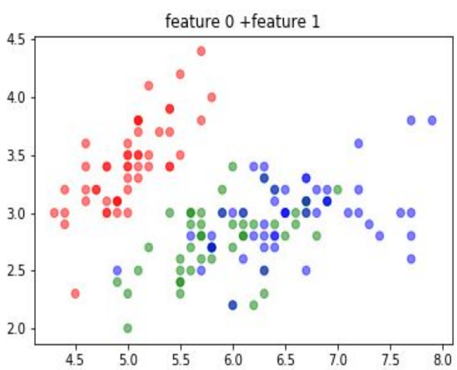
the diagonal the exact correlated and features 3 and 4 are really correlated. This information is very useful, for example, if I would like to drop out some features to save space, I can choose either 3 or 4, because each of them can represent both of them.

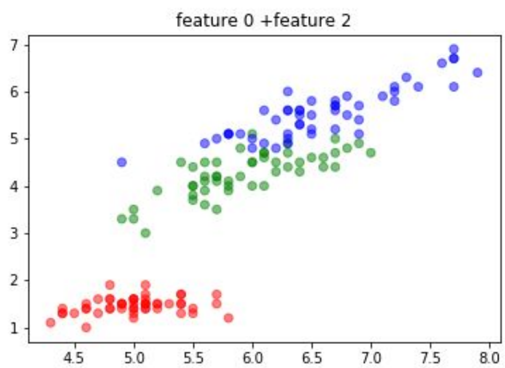


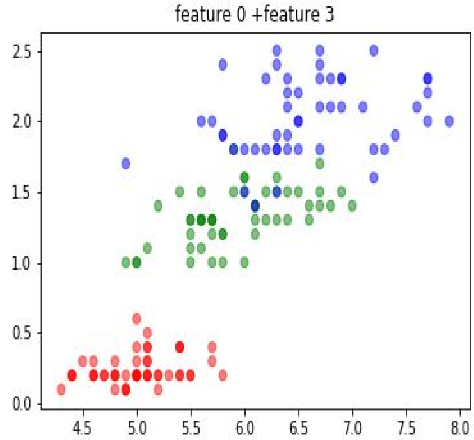
[WINE]
ex features 5,6 | featurs 10 11 are correalted
when we have these stuff, it will help us preprocess our data and analyze our data.
Preprocessing is a really significant process in ML.

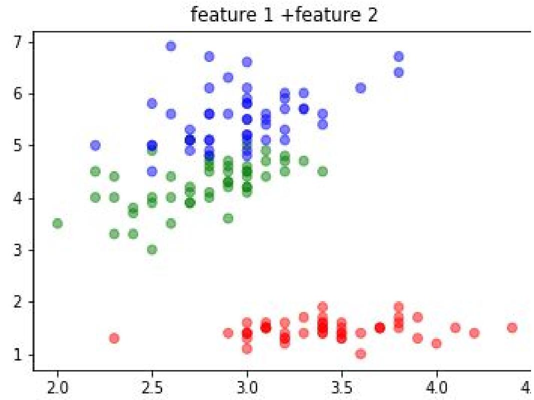


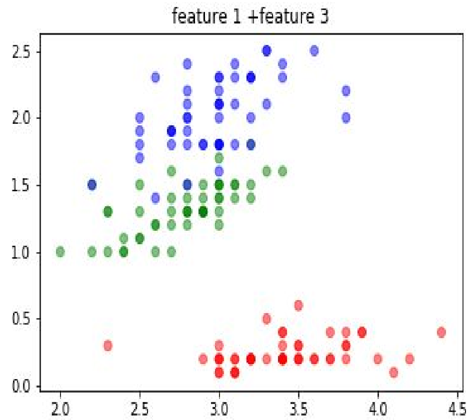
2) [20%] Scatterplots [only for the “Iris” dataet] :
a)

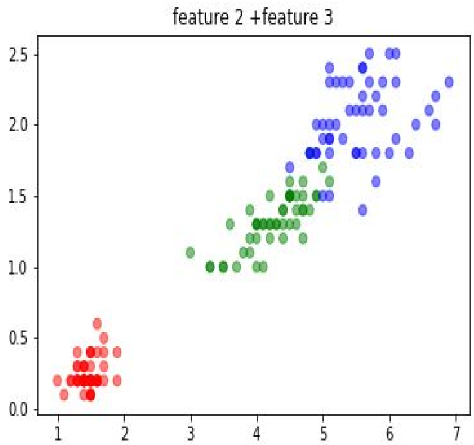












bC)

setosa:red | versicolor:green | virginical:blue

FEATURE 0+1:

setosa and others : discriminate

versicolor and setosa non-discriminate

FEATURE 0+2:

setosa and others : discriminate

versicolor and setosa non-discriminate(some parts are overlaped)

FEATURE 0+2:

setosa and others : discriminate

versicolor and setosa are discriminate(but some parts are still overlaped)

FEATURE 1+2:

setosa and others : discriminate

versicolor and setosa non-discriminate(some parts are overlaped)

FEATURE 1+3:

setosa and others : discriminate

versicolor and setosa are discriminate(but some parts are still overlaped)

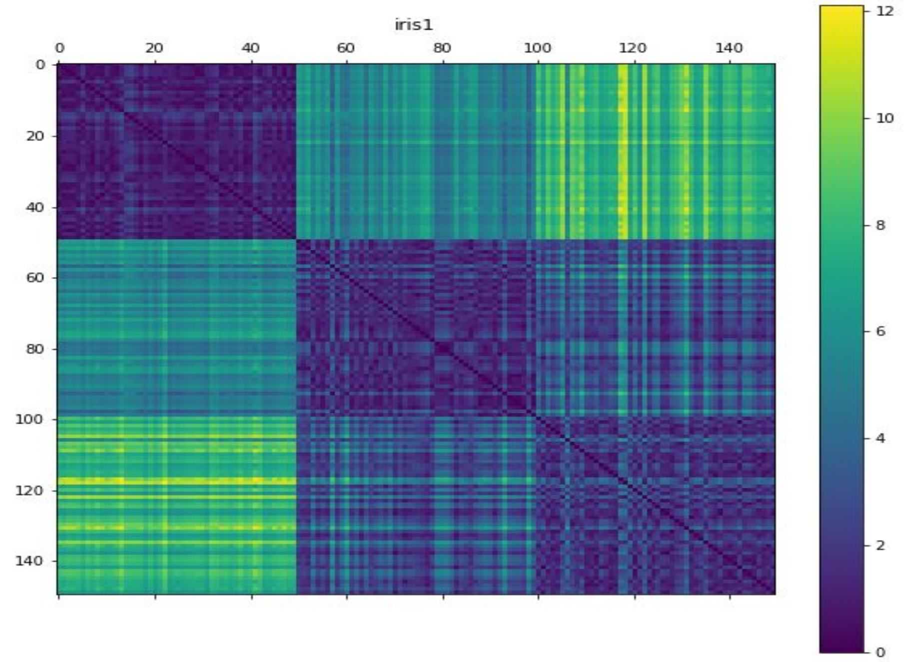
FEATURE 2+3:

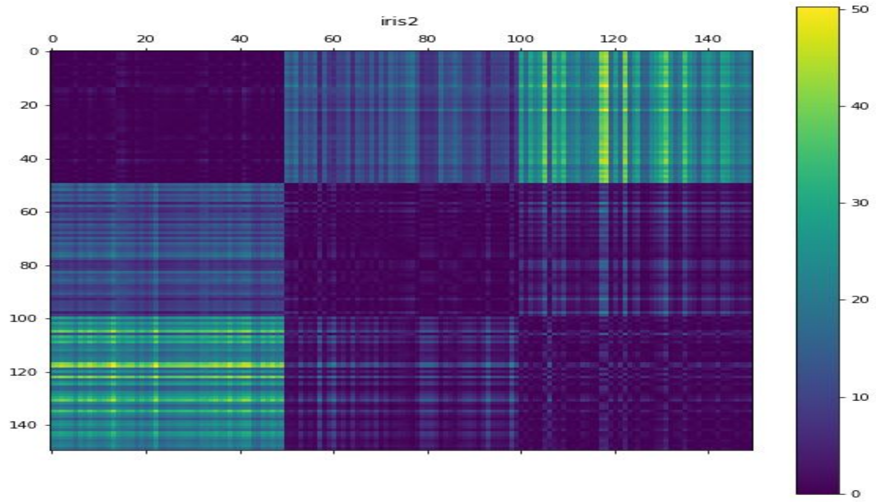
setosa and others : discriminate

versicolor and setosa are discriminate(but some parts are still overlaped)

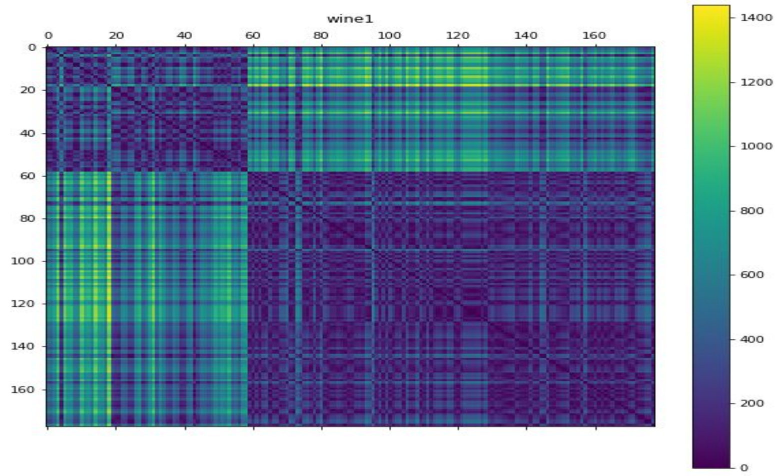
3.[60%] Distances :

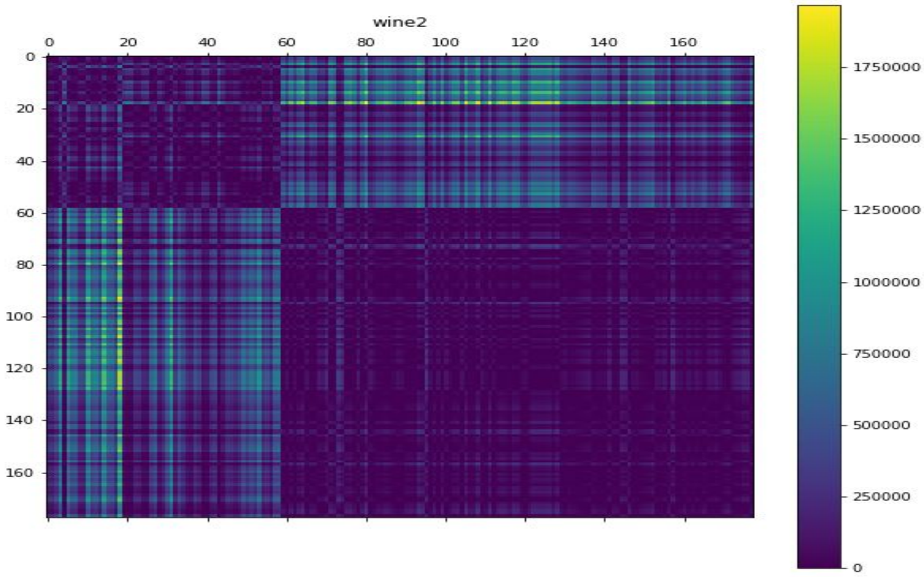
a) implement in my code and followe the fomula
b) implement in my code and followe the fomula
c) iris: (150*150-150)/2 = 11175(upper bound without diagonal)
   wine: (178*178-178)/2 =15753
d) iris





[wine]





Basically, p1,p2 will affect the intensity of the distance. It means that the larger distance between x,y will amplify when p equal to 2. Sometimes, we should normalized our dataset to make it fair. There are several ways to do that. But normalization is an important part in calculating the distance. EX x,y: feature 1 10000:10010 feature 2 : 1 and 11. In p = 1 it looks like the same distance in features 1 and 2. But Actually the proportion is not the same.

e)
iris: p =1 143/150 p=2 144/150
wine: p=1 150/178, p=2 142/150

The result might be a little different when we decide the situation that the distance is the same.