EE708, Assignment 1 Submission

Dhruv Gupta, Roll no: 240354

Problem 1

Question

Imagine we have two possibilities: We can scan and e-mail the image, or we can use an optical character reader (OCR) and send the text file. Discuss the advantages and disadvantages of the two approaches in a comparative manner.
When would one be preferable over the other?

Output

Here is a comparative analysis of the use cases of an OCR vs scanning an image. I have made a table to understand this:

Feature	**OCR text file**	**Scanned Image**
File Size	Easy to store	Harder to store images
Searchability	Searchable	Not searchable
Editability	Easily editable	Not editable
Accessibility	Highly Accessible	Requires software
Data Loss	Not that accurate	High accuracy

Overall, keeping all these aspects in mind, I think it is advisable to use scanned images when the accuracy of the data matters more (eg: contracts, signatures, forms, etc). and OCR encoded text can be used when we aren't concerned about the original layout being intact to a large extent. (eg: basic classifier models).

Problem 2

Question

Assume we are tasked with building a system to distinguish junk e-mail.
a. What is in a junk e-mail that lets us know it is junk?
b. How can the computer detect junk through a syntactic analysis?
c. What would we like the computer to do if it detects a junk e-mail: delete it automatically, move it to a different file, or just highlight it on the screen?

Output

(a) How to identify a junk email?

- **Irrelevant Content:** Offering products or services that the user would never use, or be very un-inclined towards.
- **Huge Claims:** Things like "You won a prize!", or "Get rich by investing in this!" and other claims that seem scammy.
- Suspicious links, no personalized greeting with a name, etc.
- **Grammar and Keywords:** Bad grammar, words like "rich", "urgent","prize","lottery," etc.
- **Others:** No unsubscribe option, high frequency of domains, fake sender ids, etc.

(b) Detecting a junk email using parsing

- **Using ML Models:** things like Bayesian Filtering to calculate the probability of an email being spam.
- **Using simple rules:** using a rule based system or simple decision tree to classify emails.
- **Analysis of the header:** looking for inconsistencies in the header, senders, IP addresses, etc
- **Frequency analysis:** To notice any odd sorts of words being used excessively, or notice the frequency of similar emails.
- **Misc:** looking for basic things like no unsubscribe option, all caps text, punctuation etc.

(c) What should the computer do?

- When should it delete automatically: when there is high accuracy in the system and minimal false positives (very less chance of accidentally deleting important mails classified as spam).
- When should it move it to a different file: this is a safe approach, as the user can manually check for emails, can be used when the model has decent performance but still misses things here and there.
- Highlighting spam to the screen: can be used in early stages if there are still a lot of false positives or during the testing phase.

Problem 3

Question

Let us say we are tasked with building an automated taxi.

(a) Define the constraints.
(b) What are the inputs?
(c) What is the output?
(d) How can we communicate with the passenger?
(e) Do we need to communicate with the other automated taxis; that is, do we need a 'language'?

Output

(a) Contraints of the problem

- **safety:** the model has to accurately detect obstacles and blockages in real time, and follow traffic rules about signals and speed limit.
- **logistics:** has to have good location detection for pickup/drop, has to deal with conditions like rain or limited visibility, has to monitor its fuel and have a mechanism to refuel when necessary.
- most importantly, the system should have **very low latency** to be able to function in real time.

(b) Inputs There can be a large range of inputs, feature selection and analysis play a huge role here, but I will try to list some of the main ones:

- **Local real-time signals:** inputs like data from the cameras, sensors, traffic signals, road signs, GPS data for location etc.
- **Passenger instructions:** passengers telling the car to manually stop, or to drive slower, control the A/C etc.
- **Vehicle Status:** the fuel level, speed, acceleration, brake health, engine health, etc.

(c) Outputs:

- **Vehicle Controls:** steer, accelerate, brake, change lane, etc. for maneuvering the vehicle.
- **Other Outputs:** info given to the passenger about ETA, location, etc. and signals like indicator lights or signals to local pedestrians.
- **Data collection:** to improve the model further potentially, information about cases where the response wasnt optimal, delayed response, accidents etc.

(d) How to communicate with the passenger? We can communicate with the passenger in many ways, like audio-visually with displays inside the car, or a mobile app integration, we could also make interactive touchscreens so that the passenger can communicate.

(e) Communication with other automated taxis. To communicate with other taxis, a language could be made for taxi to taxi communication, moreover, even manual cars could have some module integrated such that they are able to communicate with the taxis.

But it might be simpler to have each taxi communicate with a centralized system than for them to communicate with each other. The latter option can only be used when necessary, for example, imagine something similar to flight traffic control, which is done for all airplanes in the local area with the ATC, a centralized body.

Problem 4

Question

Let us say our hypothesis class is a circle instead of a rectangle.

(a) What are the parameters?
(b) How can the parameters of a circle hypothesis be calculated in such a case?
(c) What if it is an ellipse?
(d) Why does it make more sense to use an ellipse instead of a circle?

Output

(a) parameters of the problem: for a circle, the parameters are the center of the circle (x,y) and the radius (r), hence we have 3 parameters.
for a standard ellipse, the parameters are the center of the ellipse (x,y), and the semi major and semi minor axes (a,b). hence we have 4 parameters. (note: we could also include a 5th parameter, ie, the angle of rotation for the ellipse in case we are also considering tilted ellipses).

(b) calculating parameters of a circle hypothesis: for calculating the parameters of a circle, we can assume a center for every point and then use a euclidean distance comparison to see the relative distances of every point from that point, if we find a group of points all spaced at about a distance r from the assumed center, we can classify that group as a circle of radius r at the assumed center.

(c) case of an ellipse: to calculate the parameters in case of an ellipse, in case orientation is involved, we can use techniques like PCA to identify the major/minor axes and then check using a similar approach to the circle if a group of points satisfy the equation of the ellipse.

(d) why to use an ellipse instead of a circle? a circle imposes a uniform distance constraint; this is not practical, as most feature distributions that we use don't follow this properly. an ellipse on the other hand, is flexible and takes into account skewness/kurtosis and other factors of distributions, hence it is a better choice, and can reduce misclassification to a great extent.

Problem 5

Question

A manufacturer says the Z-Phone smartphone has a mean consumer life of 42 months with a standard deviation of 8 months. Assuming a normal distribution, what is the probability that a given random Z-Phone will last between 20 and 30 months?

Output

First, we use the PDF for normal dist., given by the formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Then, we substitute the known values into the PDF, so the function for any x becomes:

$$f(x) = \frac{1}{\sqrt{2\pi \cdot 64}} \exp\left(-\frac{(x-42)^2}{2 \cdot 64}\right)$$

Now we convert this to the CDF. We need to calculate the probability between $x = 20$ and $x = 30$, so we can directly integrate the PDF from 20 to 30. This gives us the expression:

$$P(20 \leq X \leq 30) = \int_{20}^{30} f(x)\, dx$$

After solving, assuming the CDF to be $\Phi(x)$, we calculate $\Phi(30) - \Phi(20)$, which is:

$$P(20 \leq X \leq 30) = \Phi(30) - \Phi(20) \approx 0.0668 - 0.0030 = 0.0638$$

Hence the probability that a given Z-Phone will last between 20 and 30 months is approximately **6.38 percent**.

Problem 6

Question

An experiment to investigate the survival time in hours of an electronic component consists of placing the parts in a test cell and running them for 100 hours under elevated temperature conditions. (This is called an "accelerated" life test.) Eight components were tested with the following resulting failure times:

75, 63, 100+, 36, 51, 45, 80, 90

The observation 100+ indicates that the unit still functioned at 100 hours. Is there any meaningful measure of location that can be calculated for these data? What is its numerical value?

Output

This data has an outlier, 100+, which could take any value greater than 100. So, it isnt advisable to use measures of location that are sensitive to outliers.
We cannot use the mean of the data as a good measure of location in this case because outliers can severely affect the value of the mean.
But the measure of median is not affected by outliers in this case, and hence can be used as a measure of location. **The median of this data, which is 63, is an appropriate measure of location.**

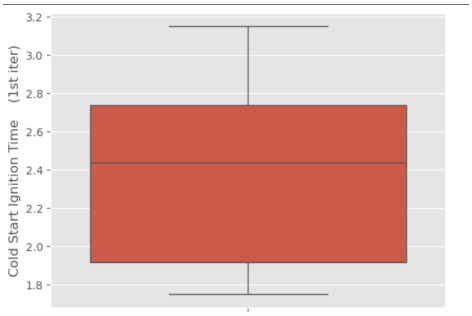
Problem 7

Question

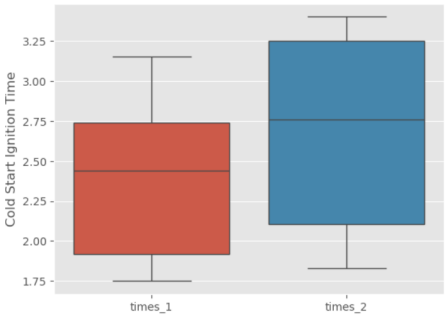
A gasoline manufacturer is investigating the "cold start ignition time" of an automobile engine. The following times (in seconds) were obtained for a test vehicle: 1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91. Calculate the sample mean, sample variance, and sample standard deviation. Construct a box plot of the data. A second formulation of the gasoline was tested in the same vehicle, with the following times (in seconds): 1.83, 1.99, 3.13, 3.29, 2.65, 2.87, 3.40, 2.46, 1.89, and 3.35. Use these new data, along with the cold start times reported in the previous exercise, to construct comparative box plots. Write an interpretation of the information that you see in these plots.

Output

Sample mean for first sample: 2.415
Sample variance for first sample: 0.28537142857142855
Sample standard deviation for first sample: 0.534201674062735



box plot of first data



comparitive box plots

Interpretation of plots

The first plot clearly shows that the engine usually ignites between 2 and 2.7 seconds with the first gasoline formulation, the mean time being around 2.5 seconds.
The second plot is a comparison between the formulations of gasoline, showing that the first one was better as it has a lower mean start time.

Problem 8

Question

A set of 10 hypothetical patient records from a large database are presented in Table 1. Patients with a diabetes value of 1 have type-II diabetes, and patients with a diabetes value of 0 do not have type-II diabetes.
a. Create a new column by normalizing the Weight (kg) variable into the range 0-1 using the minmax normalization.
b. Create a new column by binning the Weight (kg) variable into three categories: low (less than 60 kg), medium (60-100 kg), and high (greater than 100 kg).
c. Create an aggregated column, body mass index (BMI, which is defined by the formula:

$$BMI = \frac{Weight(Kg)}{(Height(m))^2}$$

Output

	Name	Weight	Height	Systolic BP	Diastolic BP	Diabetes	norm_weight	wtcategory	BMI
3	A. Patel	41	1.55	76	125	0	0.000	low	17.066
0	P. Lee	50	1.52	68	112	0	0.095	low	21.641
8	P. Rice	64	1.74	101	132	0	0.242	medium	21.139
6	N. Cook	73	1.76	108	136	0	0.337	medium	23.567
4	M. Owen	79	1.82	65	105	0	0.400	medium	23.850
2	J. Smith	96	1.83	88	136	0	0.579	medium	28.666
7	W. Hands	104	1.71	107	145	1	0.663	high	35.566
5	S. Green	109	1.89	114	159	1	0.716	high	30.514
1	R. Jones	115	1.77	110	154	1	0.779	high	36.707
9	F. Marsh	136	1.78	121	165	1	1.000	high	42.924

Updated table, sorted in the order of normalised weight

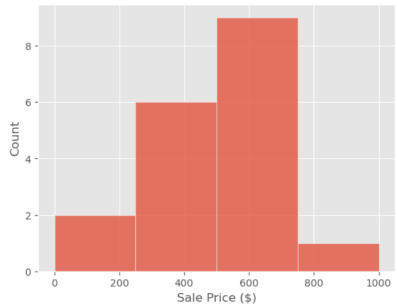
Answer

Added the normalised weight column, using minmax normalisation.
Classified all the inputs into different classes based on weight.
Created an aggregated BMI column, added at the end of the dataframe.

Problem 9

Question

Table 2 shows a series of retail transactions monitored by the main office of a computer store.
a. Create a histogram of Sale Price () using the following intervals: 0 to less than 250, 250 to less than 500, 500 to less than 750, and 750 to less than 1000.
b. Generate a contingency table summarizing the variables Store and Product category.
c. Generate the following summary tables:
i. Grouping by Customer with a count of the number of observations and the sum of the Sale price ($) for each row.
ii. Grouping by Store with a count of the number of observations and the mean Sale price ($) for each row.
iii. Grouping by Product category with a count of the number of observations and the sum of the Profit ($) for each row.
d. Create a scatterplot showing Sales price ($) against Profit ($).



histogram

Product Category	Desktop	Laptop	Printer	Scanner
Store				
New York, NY	3	1	2	4
Washington, DC	2	2	2	2

contingency table

Customer	Count	Total Sale
B. March	3	1700
E. Sims	1	700
G. Hinton	4	2150
H. Fu	1	450
H. Taylor	1	400
J. Bain	1	500
L. Nye	2	900
P. Judd	2	900
S. Cann	1	600
T. Goss	2	750

grouping by customer

part 2 output

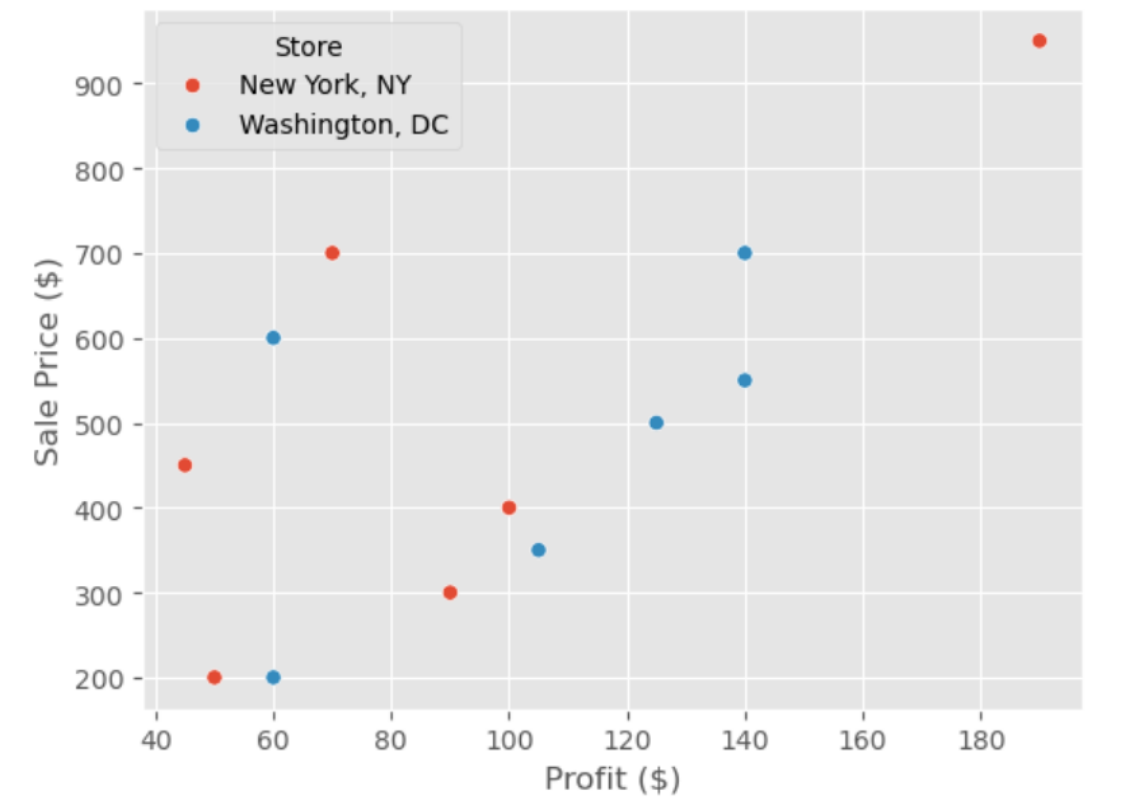
Store	Count	Mean Sales ($)
New York, NY	10	485.0
Washington, DC	8	525.0

part 3 output

Product Category	Count	Net Profit ($)
Desktop	5	295
Laptop	3	470
Printer	4	360
Scanner	6	640

grouping by store and product category

Output



Scatterplot showing Sales Price vs Profit

Problem 10

Question

Write a code to implement the following exploratory data analysis: (use dataset A1.csv)
a. Find the frequency of samples for each class.
b. Generate data description and calculate the interquartile range for all four features.
c. Plot a histogram of feature 1 for class A.
d. Make the box plot for feature 2 for each class separately.
e. Violin plot for feature 3 for each class separately.
f. Scatter plots between feature 1 and feature 3 showing classes separately.
g. Contour plot between feature 1 and feature 4 showing classes separately.
h. Hexagonal bin plot for class A between feature 2 and 4.
i. Correlation matrix for the four features.
j. Pair plot for the four features showing classes separately.

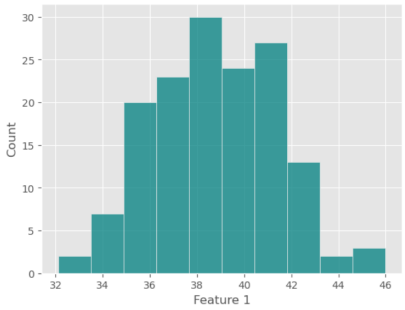
Output

a) part output: {'A': 151, 'B':123, 'C':68}

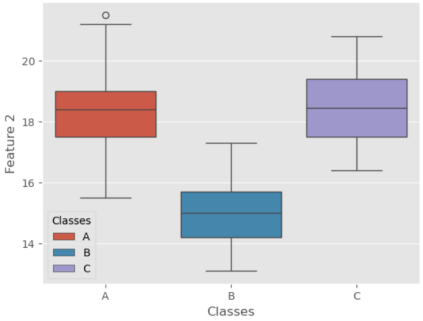
```
        Feature 1   Feature 2   Feature 3    Feature 4
count  342.000000  342.000000  342.000000   342.000000
mean    43.921930   17.151170  200.915205  4201.754386
std      5.459584    1.974793   14.061714   801.954536
min     32.100000   13.100000  172.000000  2700.000000
25%     39.225000   15.600000  190.000000  3550.000000
50%     44.450000   17.300000  197.000000  4050.000000
75%     48.500000   18.700000  213.000000  4750.000000
max     59.600000   21.500000  231.000000  6300.000000

Interquartile Range for all features:
Feature 1       9.275
Feature 2       3.100
Feature 3      23.000
Feature 4    1200.000
dtype: float64
```

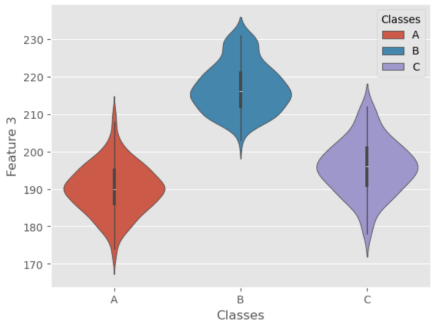
b) data description, interquartile range



c) histogram for feature 1, class A



d) box plot for classwise feature 2

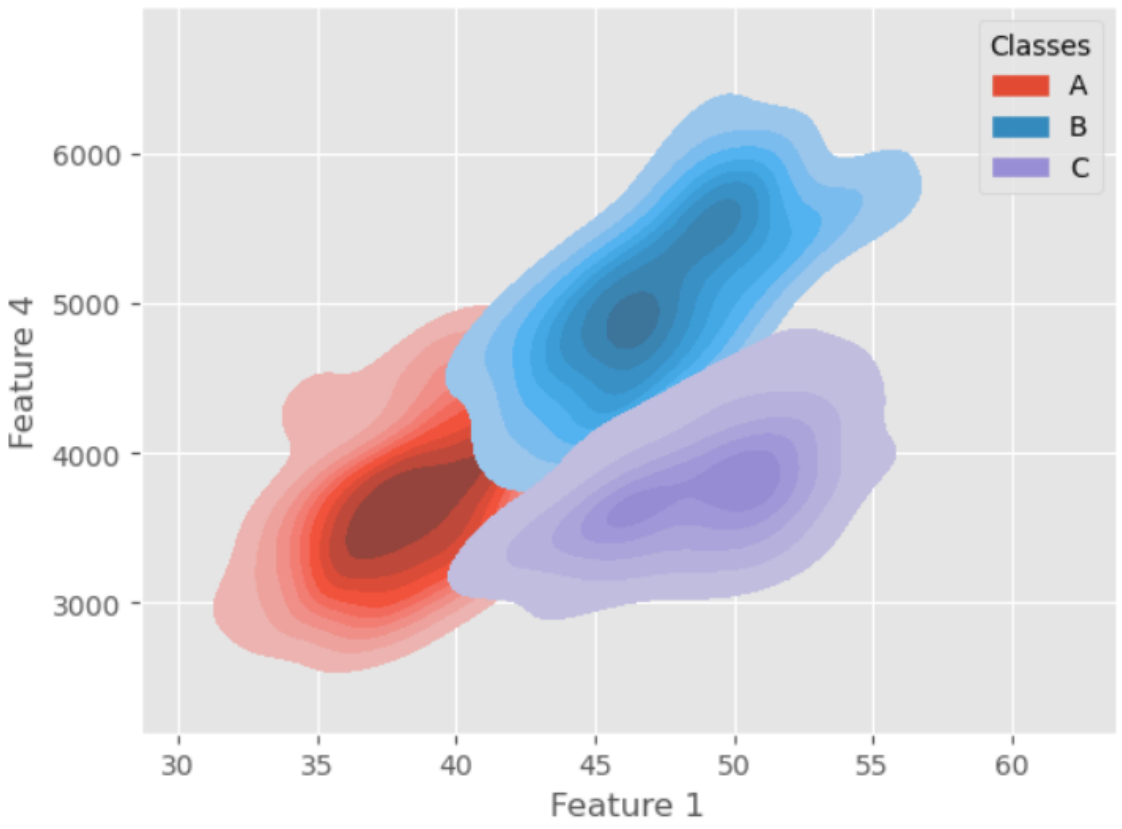


e) violin plot for classwise feature 3

Output

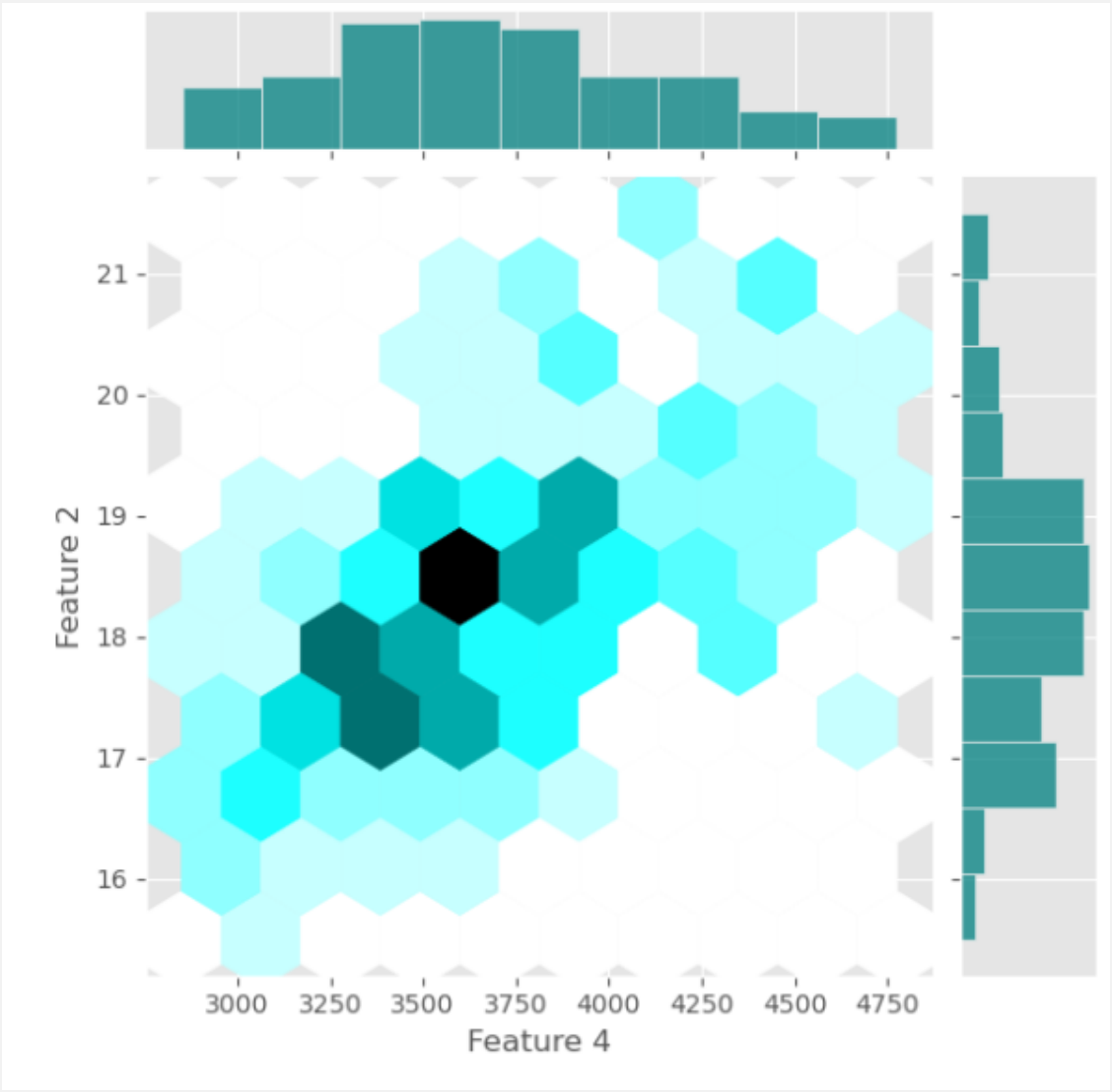
f) scatterplot for classwise comparision of feature 1 and 2

Output



g) contour plot for classwise comparision of feature 1 and 4

Output



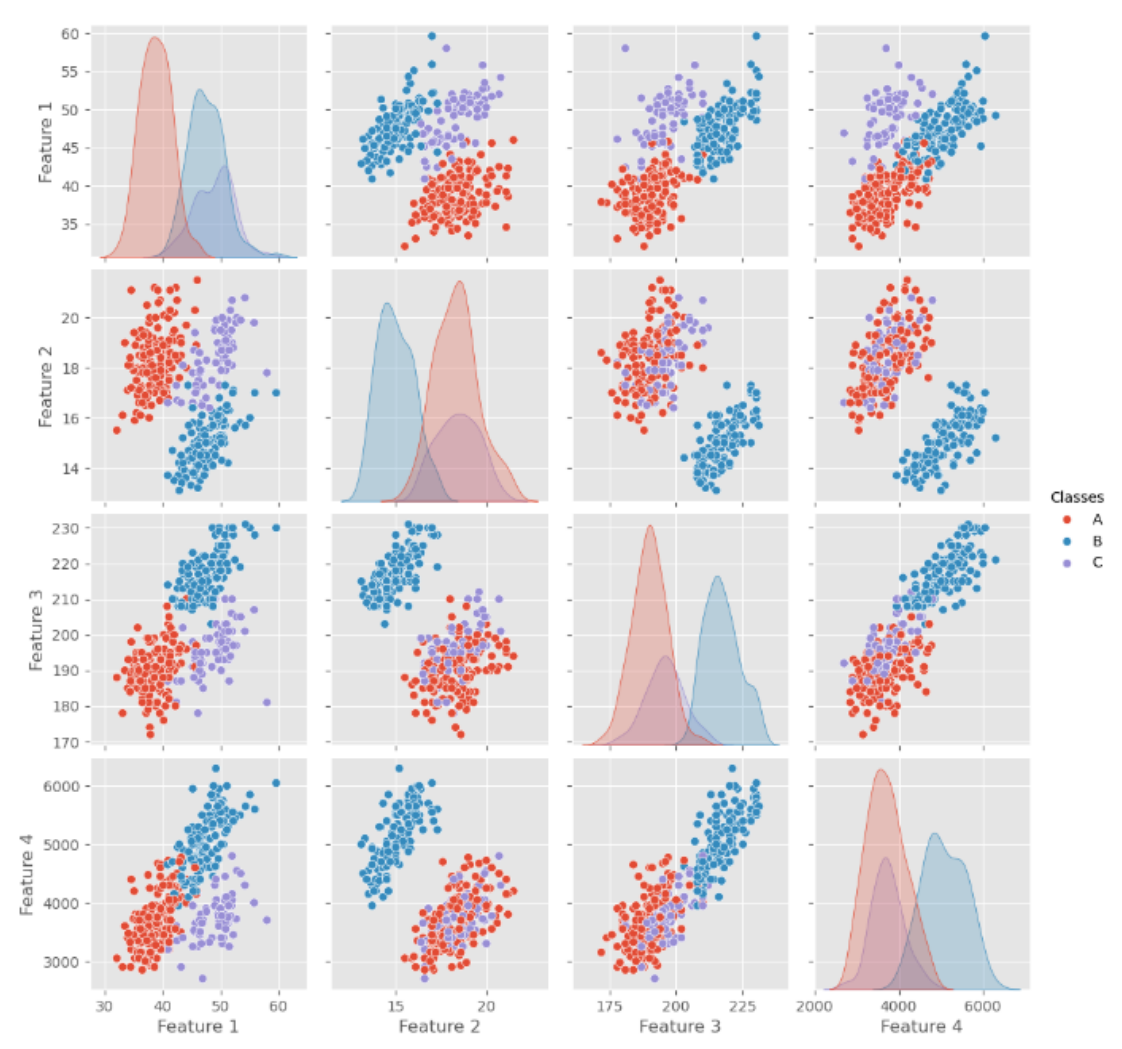
h) hexagonal bin plot for class A's comparision of feature 2 and 4

Output

	Feature 1	Feature 2	Feature 3	Feature 4
Feature 1	1.000000	-0.235053	0.656181	0.595110
Feature 2	-0.235053	1.000000	-0.583851	-0.471916
Feature 3	0.656181	-0.583851	1.000000	0.871202
Feature 4	0.595110	-0.471916	0.871202	1.000000

i) correlation matrix and heatmap for the four features

Output



j) pair plot for four features, classwise