# Assignment 1

Durbasmriti Saha

FUNDAMENTALS OF DATA SCIENCE AND MACHINE INTELLIGENCE

January 31, 2025

1) "Scanning" means using a scanner to convert physical paper documents into a digital image file, essentially creating a digital copy of the document that can be stored and accessed electronically on a computer. It's the process of digitizing paper documents.
Optical Character Recognition (OCR) is the process that converts an image of text into a machine-readable text format.
Both methods have their advantages and disadvantages:

**Advantages of Scan and email:**

- It is a faster process in the case of simple documents.
- It preserves the original formatting and visuals.
- This method is very useful when there is unusual fonts or complex handwriting.
- It stores images and graphics as it was in the original documents.

**Disadvantages of using scan and email image:**

- The scanned image/document cannot be edited, So if there were mistakes in the original documents then that cannot be corrected in the scanned documents.
- It takes a lot of manhours for complex and large amounts of documents since each has to be scanned individually.
- We cannot search for specific words or phrases within the image, making it difficult to find relevant information quickly.
- For data extraction, it needs manpower to do so from scanned documents which can be hectic for large data.

**Advantages of using OCR:**

- It makes the digitized documents completely text searchable. This helps readers to quickly lookup numbers, addresses, names, and various other parameters that would have otherwise taken hours.

- Readers can edit the digitized documents hence highly beneficial since mistakes in the original documents are getting corrected.
- With the usage of machine learning and AI with OCR, digitized documents can even predict words when it becomes difficult to read handwritten characters hence making it more readable.
- When there is a large amount of data, OCR is much faster than traditional scanning, as it extracts data by its own hence saving manual labor.

**Disadvantages of using OCR:**

- If there are imperfections in original documents like complex handwriting, or unusual fonts, it becomes harder for OCR to read and extract data from them making lower quality of digitized documents.
- Sometimes OCR can be slow since it has to analyze each image and convert it into text, which can take some time.
- OCR systems are expensive.
- Since OCRs are not perfectly accurate, digitized documents have to be manually checked and corrected by humans.

2) a) **There are various ways to know if a particular email is a junk or not:**

- Most of the junk emails have subject lines related to sales, offers, investment opportunities, new treatments, requests for money, information on packages we never ordered, etc.
- These emails contain links that are different from their destination links. We can hover over those link to see if that match with the one showing up on the bottom-left corner of the screen.
- Generally, legitimate emails address the recipient by name in the greeting to get their attention. Companies and institutions can use email marketing tools that create personalized greeting for each person on their email lists. If an email has a generic greeting such as “Hello” or “Dear Valued Customer,” there is a good chance that it is spam.
- Sometimes spam emails contain typos to avoid getting picked up, While legitimate institutions have editors to check for typos.

b) The computer can detect junk emails by examining the grammatical structure of the emails, recognizing the frequent usage of certain characters/phrase like ’$’, ’!’, ”sale”, ”offers” etc. It could also identify the pattern of writing in the spam mails which would help it segregating spam emails from the genuine ones.

c) I would like the computer to move spam emails to a seperate files instead of completely deleting or removing from devices. This is because many times computers misclassify the good emails as spam which is worse than classifying spams as good. Moving spams to different files helps in both cases and hence always a preferable idea.

3) a) There are a lot of constraints for an automated taxi:

First of all is obviously **Safety**:

- The taxi must obey all the traffic rules.
- It must Avoid collisions with pedestrians, vehicles or any other obstacles.
- It must have an inbuilt system to account for unexpected situations like sudden brakes by other vehicles which otherwise will cause massive collisions.

The second is about **Ethics**. There may be situations of ethical dilemmas and in those cases, it must prioritize its passengers' lives.
Third, it must have the proper **navigation** system so that it reaches its destination correctly.
fourth, in a country like India, where there are problems of proper roads and traffic, its hardware and software must be of best notch so that it can service perfectly.
Next, It must be **efficient** in terms of power consumption and time taken.
Apart from these, it must also have **emergency services** in case its hardware or software malfunction occurs and most importantly the system should be cost-effective to build, maintain, and operate.

b) The inputs are:

- Visual information about traffic signals, obstacles nearby, and road conditions that can be captured by camera sensors installed in the system.
- Pick-up and drop-off locations of passengers obtained via app or in-car interface.
- audio information like siren, horns or other external sounds.
- real-time location through GPS for proper navigation.
- Weather conditions like temperature, humidity, rain, snow which can affect driving.

c) The outputs are:

- optimal route planning considering all the information about traffic and weather to optimize travel time while maintaining passengers' safety.
- advance notice to passengers about estimated time to be taken or any delays also about any emergency such as system failure, fuel amount left, etc.
- It must be able to perform emergency braking.

- storing passengers data which may come in handy later if any misfortune happens.

d) Ways to communicate with the passenger: 1) Mobile App:

- For tasks like booking of the available taxi, payment processing, and destination inputs.
- notification about lift time, delays or any other emergencies.

2) In car interface:

- screen displaying about speed of the taxi, current location map, destination inputs, fuel amount left.
- voice announcement to update about important information like whether destination reached or not, delays, traffic,

e) Yes, communication with the other automated taxis are essential to have an efficient and safe drive. It's like the digital equivalent of eye contact or hand gestures between two human drivers in non-automated taxis.

- Automated taxis need to share real-time data about their position, and speed to avoid collisions, especially in small roads and complex traffic scenarios like intersections.
- There may be situations where multiple taxis are competing for the same space (e.g., at a busy intersection and narrow lanes), they can negotiate and decide who to go first similar to how human drivers negotiate.
- Additionally if one taxi detects an unexpected scenario (e.g., an accident, road construction, or a pedestrian crossing), it can immediately alert nearby taxis, allowing them to adjust their routes or driving behavior in advance.

**4)** a) The parameters for a circle are:

- Centre : (h,k)
- Radius : r

The equation of the circle is:

$$(x-h)^2 + (y-k)^2 = r^2.$$

b) First, initialize $(h, k, r)$ with some reasonable values. Let's say we have data points $(x_i, y_i)$. Then:

$$h = \text{mean of } x_i, \quad k = \text{mean of } y_i, \quad r = \text{average distance from } (h,k) \text{ to the data points.}$$

The error function $E$ is defined as:

$$E = \sum_{i=1}^{n} \left( \sqrt{(x_i - h)^2 + (y_i - k)^2} - r \right)^2.$$

where n = number of data points Next we have to find values of h, k, r for which E is minimum. For this we need to compute the partial derivatives of the error function $E(h, k, r)$ with respect to $h$, $k$, and $r$:

$$\frac{\partial E}{\partial h} = -2\sum_{i=1}^{n}\left(\frac{x_i - h}{\sqrt{(x_i-h)^2+(y_i-k)^2}}\right)\left(\sqrt{(x_i-h)^2+(y_i-k)^2} - r\right),$$

$$\frac{\partial E}{\partial k} = -2\sum_{i=1}^{n}\left(\frac{y_i - k}{\sqrt{(x_i-h)^2+(y_i-k)^2}}\right)\left(\sqrt{(x_i-h)^2+(y_i-k)^2} - r\right),$$

$$\frac{\partial E}{\partial r} = -2\sum_{i=1}^{n}\left(\sqrt{(x_i-h)^2+(y_i-k)^2} - r\right).$$

Then we have to adjust the values of $h$, $k$, $r$, and learning rate $\alpha$:

$$h_{\text{new}} = h_{\text{old}} - \alpha \cdot \frac{\partial E}{\partial h}$$

$$k_{\text{new}} = k_{\text{old}} - \alpha \cdot \frac{\partial E}{\partial k}$$

$$r_{\text{new}} = r_{\text{old}} - \alpha \cdot \frac{\partial E}{\partial r}$$

We keep on updating and repeat this process till E converges to minimum or a maximum number of iteration is reached.
c) If it is an ellipse then the parameters are:
1) Centre (h,k)
2) Semi-major axis: a
3) semi-minor axis: b
4) angle $\theta$, that the major axis makes with x-axis.
The equation of an ellipse is:

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1.$$

d) An ellipse is better than a circle in many ways:

- An ellipse is more general form. Even circle is an special case of ellipse.
- Many a times, data points are not symmetrically distributed around a central point, in these type of cases, an ellipse would be a better fit.
- The angle $\theta$ can also account for orientation of data points, if data points are not aligned along coordinate axes in case.

**5)** We are given mean($\mu$) = 42, standard deviation($\sigma$) = 8.
To find: $P(20 \leq X \leq 30)$ where X is the lifespan of a Z-Phone.

We first need to convert data to standard normal distribution using the standardization formula $Z = \frac{X-\mu}{\sigma}$. So,

$$P(20 \leq X \leq 30) = P\left(\frac{20-42}{8} \leq Z \leq \frac{30-42}{8}\right) = P(-2.75 \leq Z \leq -1.5)$$

Using standard normal distribution table :

$$P(Z \leq -2.75) \approx 0.0030$$

and

$$P(Z \leq -1.5) \approx 0.0668$$

using

$$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$

We get,

$P(-2.75 \leq Z \leq -1.5) = P(Z \leq -1.5) - P(Z \leq -2.75) = 0.0668 - 0.0030 = 0.0638 \quad (\text{or } 6.38\%)$

**6) Median** would be a meaningful measure. The sorted data in ascending order is 36, 45, 51, 63, 75, 80, 90, $100^+$.
$Median = \frac{63+75}{2} = 69$.
Since total number of data points is 8 so middle terms are 4th and 5th term.

**7)** Data for the first formulation (in seconds): 1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91

$$mean = \mu = \sum_{i=1}^{n} x_i = \frac{19.32}{8} = 2.415 \quad seconds$$

$$standard \quad variance = \sigma^2 = \frac{\sum_{i=1}^{n-1}(x_i - \mu)^2}{n-1} = \frac{1.9976}{7} = 0.286 \quad seconds$$

$$Sample \quad Standard \quad Deviation(\sigma) = \sqrt{0.286} \approx 0.5348 \quad seconds$$

Data for the second formulation (in seconds): 1.83, 1.99, 3.13, 3.29, 2.65, 2.87, 3.40, 2.46, 1.89, 3.35

$$mean = \mu = \sum_{i=1}^{n} x_i = \frac{26.86}{10} = 2.686 \quad seconds$$

$$standard \quad variance = \sigma^2 = \frac{\sum_{i=1}^{n-1}(x_i - \mu)^2}{n-1} = \frac{3.44964}{9} \approx 0.383 \quad seconds$$

$$Sample \quad Standard \quad Deviation(\sigma) = \sqrt{0.383} \approx 0.619 \quad seconds$$

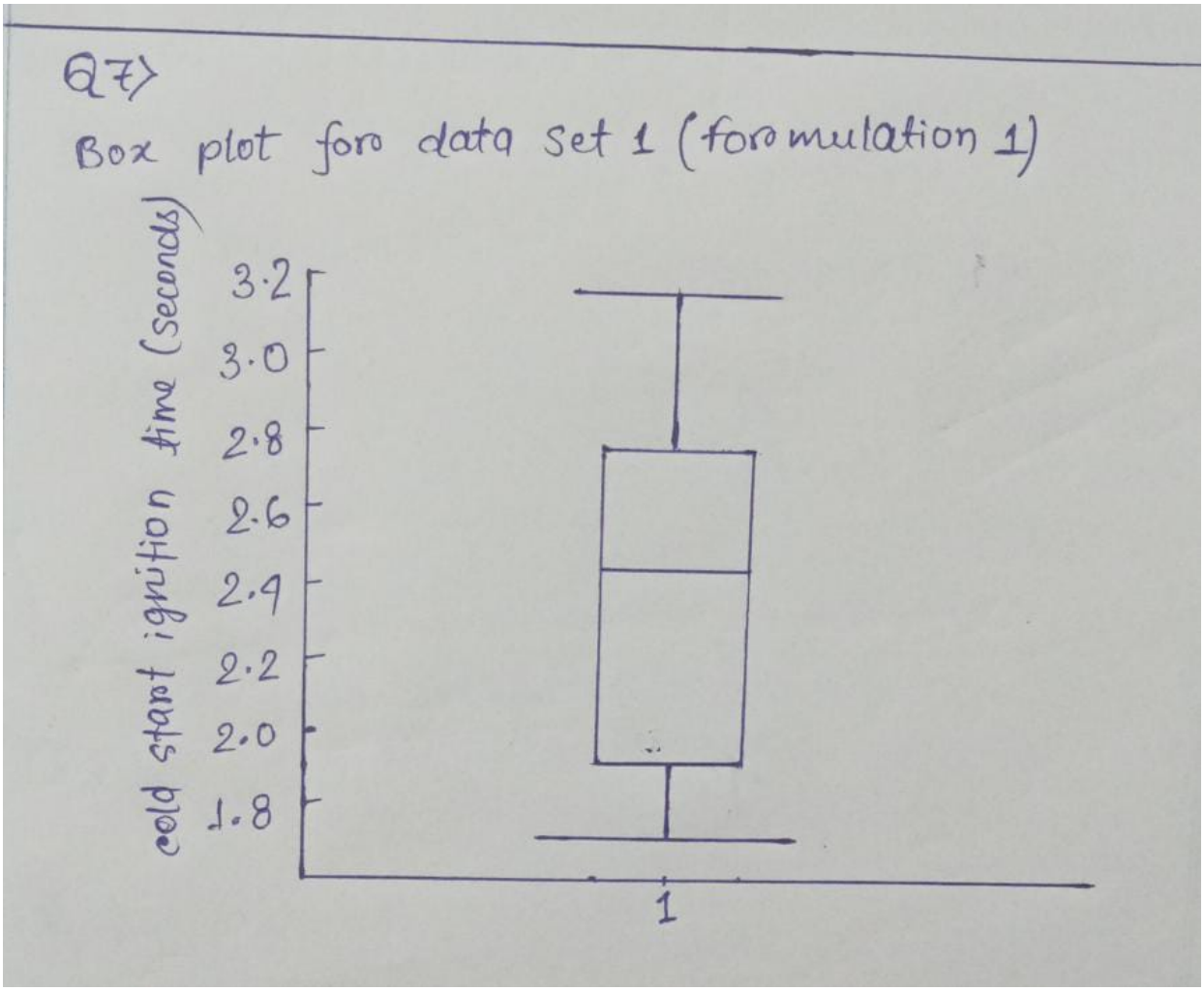
Interpretation of the Box Plots:

**Central Tendency:**
The median cold start ignition time for the first formulation is 2.44 seconds, while for the second formulation, it is 2.686 seconds. This suggests that the second formulation generally takes longer to ignite.

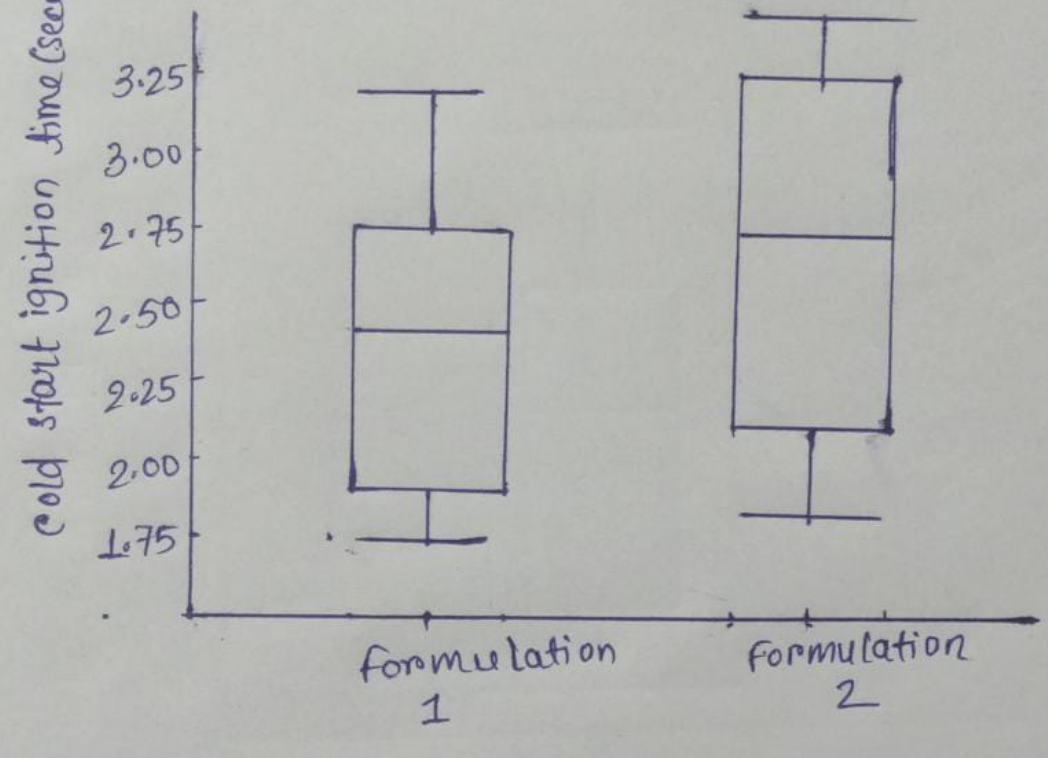
**Variability:**
The interquartile range (IQR) for the first formulation is 1.18 seconds (Q3 - Q1 = 3.09 - 1.91), while for the second formulation, it is 1.34 seconds (Q3 - Q1 = 3.29 - 1.95). The second formulation has slightly more variability in ignition times.

**Outliers:** Neither formulation appears to have significant outliers, as all data points fall within the range of the whiskers in the box plots.

**Comparison:** The second formulation tends to have longer ignition times and greater variability compared to the first formulation. This could indicate that the first formulation is more consistent and efficient for cold starts.



Q7) comparative Box plot



**8)** The updated table:

| | Name | Weight (kg) | Height (m) | Systolic Blood Pressure (mm Hg) | Diastolic Blood Pressure (mm Hg) | Diabetes | Normalized Weight | Weight Category | BMI (kg/m²) |
|---|---|---|---|---|---|---|---|---|---|
| 0 | P. Lee | 50 | 1.52 | 68 | 112 | 0 | 0.094737 | low | 21.641274 |
| 1 | R. Jones | 115 | 1.77 | 110 | 154 | 1 | 0.778947 | high | 36.707204 |
| 2 | J. Smith | 96 | 1.83 | 88 | 136 | 0 | 0.578947 | medium | 28.666129 |
| 3 | A. Patel | 41 | 1.55 | 76 | 125 | 0 | 0.000000 | low | 17.065557 |
| 4 | M. Owen | 79 | 1.82 | 65 | 105 | 0 | 0.400000 | medium | 23.849777 |
| 5 | S. Green | 109 | 1.89 | 114 | 159 | 1 | 0.715789 | high | 30.514263 |
| 6 | N. Cook | 73 | 1.76 | 108 | 136 | 0 | 0.336842 | medium | 23.566632 |
| 7 | W. Hands | 104 | 1.71 | 107 | 145 | 1 | 0.663158 | high | 35.566499 |
| 8 | P. Rice | 64 | 1.74 | 101 | 132 | 0 | 0.242105 | medium | 21.138856 |
| 9 | F. Marsh | 136 | 1.78 | 121 | 165 | 1 | 1.000000 | high | 42.923873 |

Figure 1: Table of Patient Records

9) a) histogram of Sale Price :

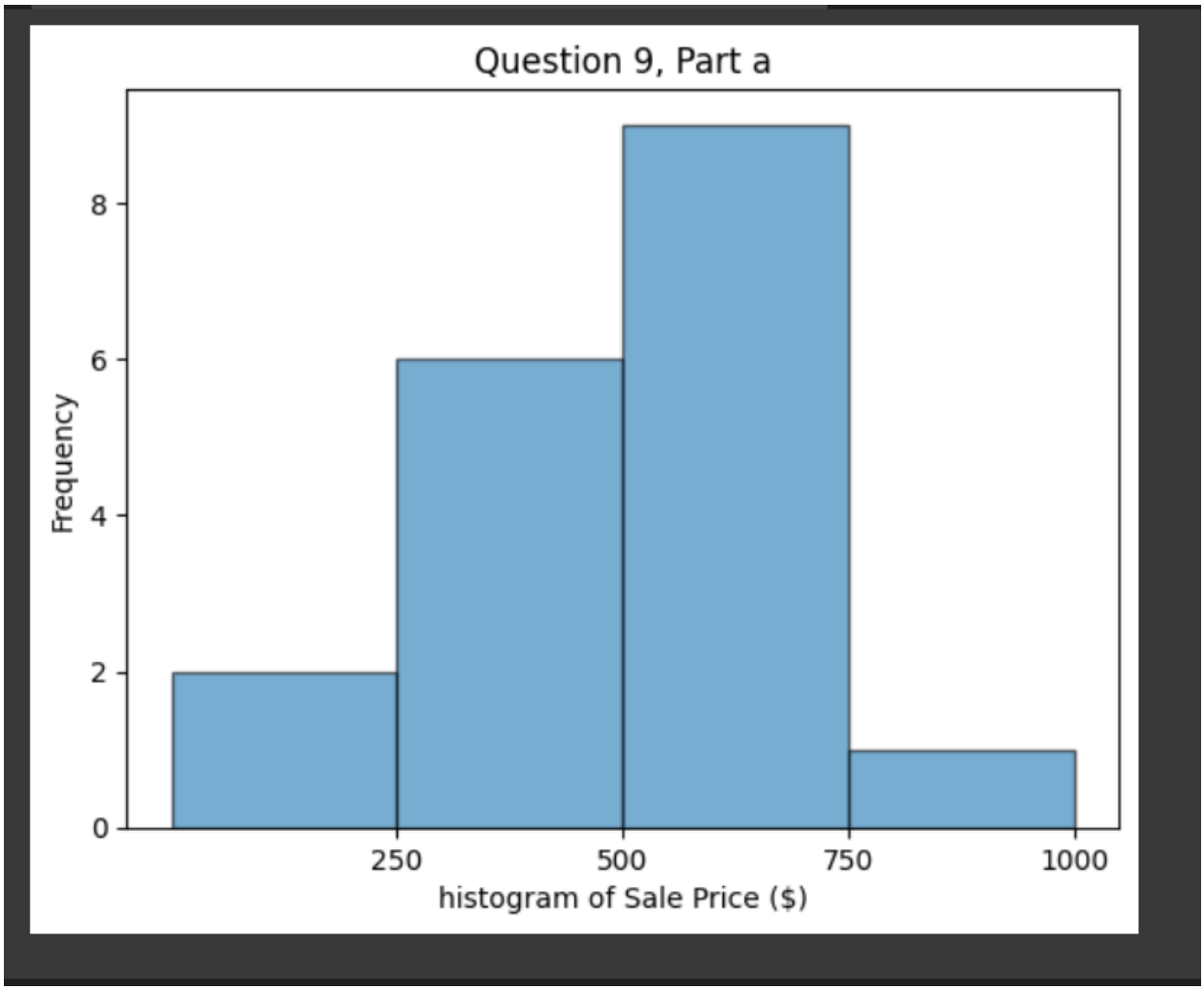



Figure 2: Histogram of Sale Price

9) b) contingency table summarizing the variables Store and Product category:

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

Figure 3: Contingency table

9) c) i) Grouping by Customer:

| | Customer | Number_of_Transactions | Total_Sale_Price |
|---|---|---|---|
| 0 | B. March | 3 | 1700 |
| 1 | E. Sims | 1 | 700 |
| 2 | G. Hinton | 4 | 2150 |
| 3 | H. Fu | 1 | 450 |
| 4 | H. Taylor | 1 | 400 |
| 5 | J. Bain | 1 | 500 |
| 6 | L. Nye | 2 | 900 |
| 7 | P. Judd | 2 | 900 |
| 8 | S. Cann | 1 | 600 |
| 9 | T. Goss | 2 | 750 |

Figure 4: Grouping by customer

c) ii) Grouping by Store:

| | Store | Number_of_Transactions | Mean_Sale_Price |
|---|---|---|---|
| 0 | New York, NY | 10 | 485.0 |
| 1 | Washington, DC | 8 | 525.0 |

Figure 5: Grouping by Store

c) iii) Grouping by Product category:

| | Product Category | Number_of_Transactions | Total_Profit |
|---|---|---|---|
| 0 | Desktop | 5 | 295 |
| 1 | Laptop | 3 | 470 |
| 2 | Printer | 4 | 360 |
| 3 | Scanner | 6 | 640 |

Figure 6: Grouping by Product category

d) A scatterplot showing Sales price against Profit:

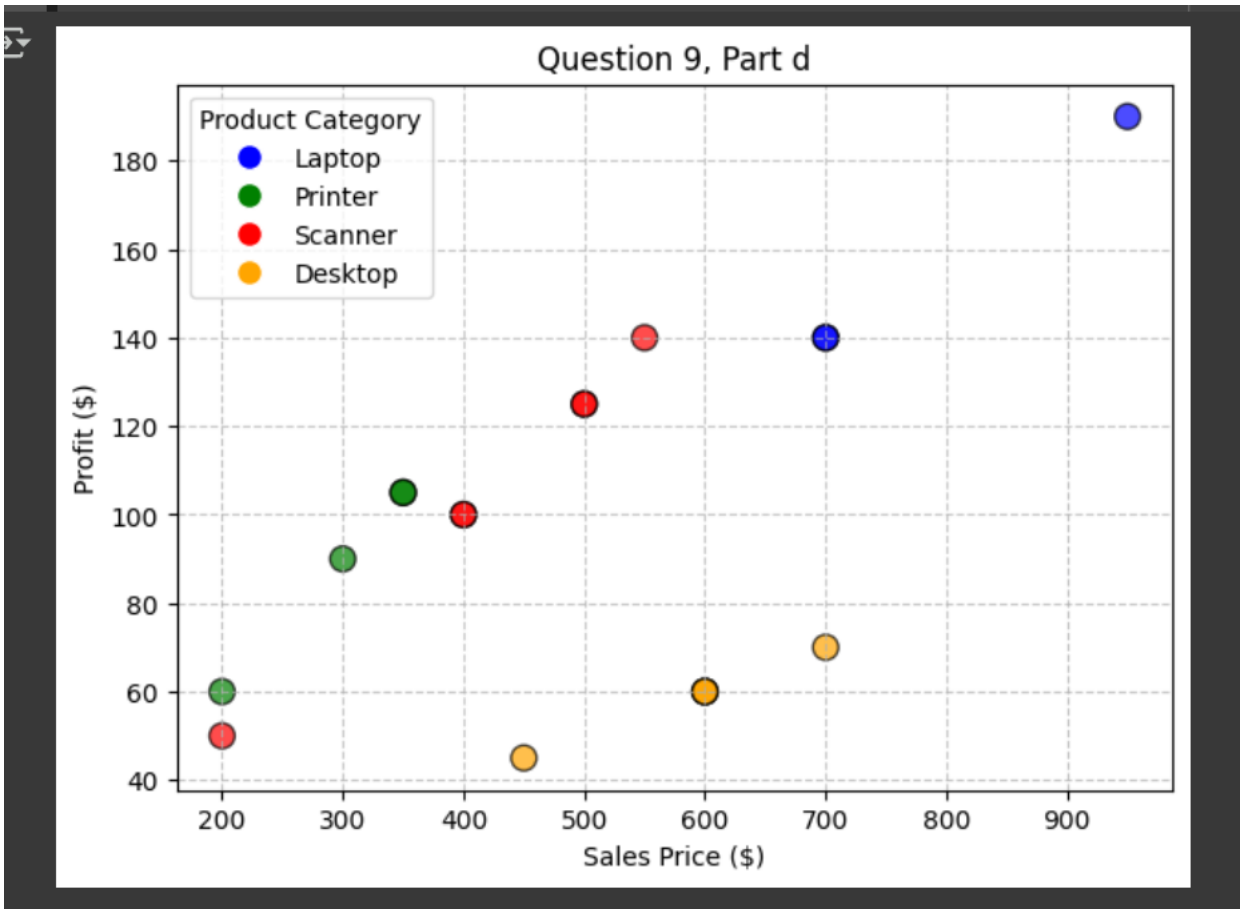



Figure 7: Scatterplot of Sales against Profit

a) frequency of samples for each class:

```
Frequency of each class:
Classes
A    151
B    123
C     68
Name: count, dtype: int64
```

b) data description and calculate the interquartile range for all four features:

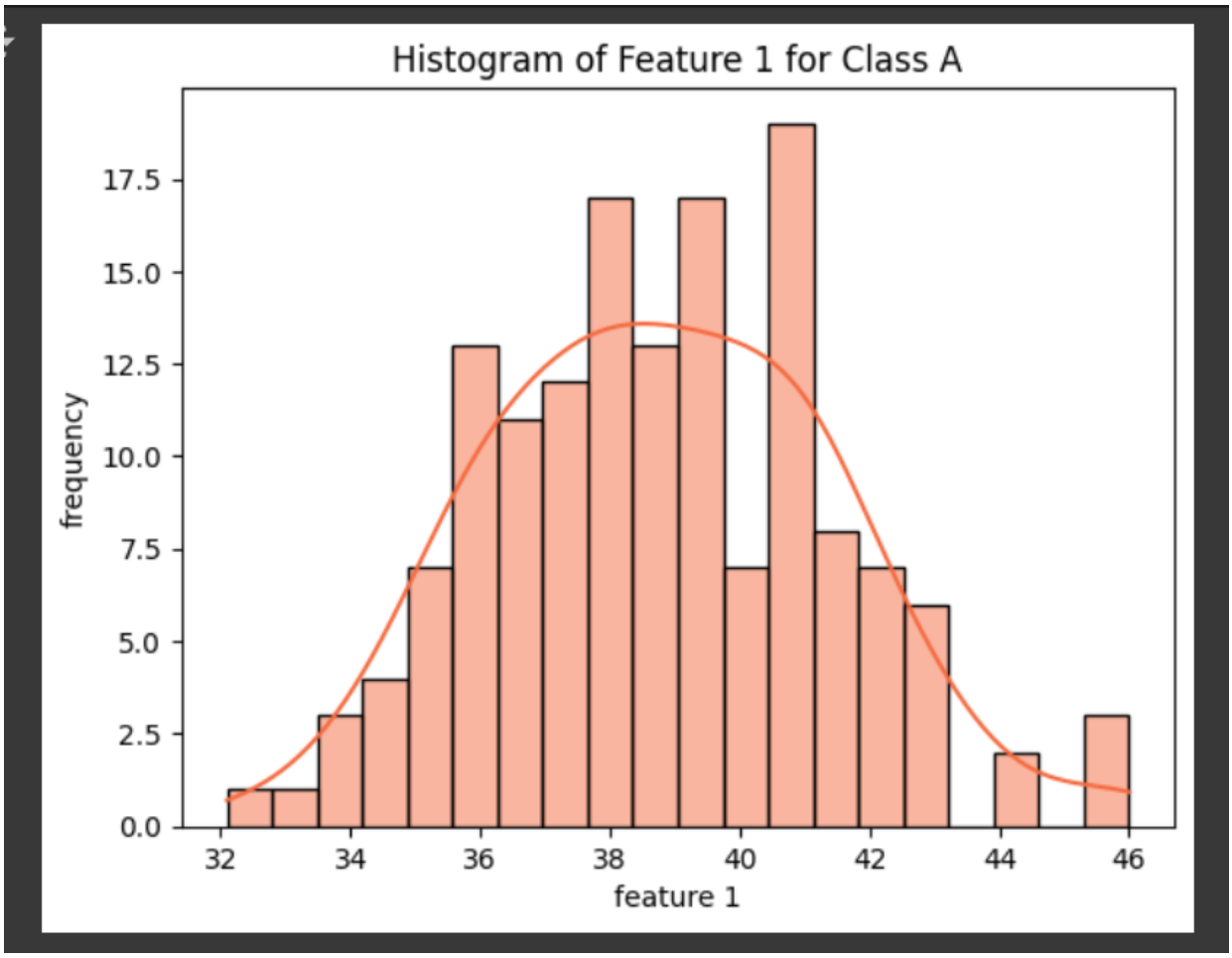
```
Data Description:
       Sample Number   Feature 1   Feature 2   Feature 3    Feature 4
count     342.000000  342.000000  342.000000  342.000000   342.000000
mean      171.500000   43.921930   17.151170  200.915205  4201.754386
std        98.871128    5.459584    1.974793   14.061714   801.954536
min         1.000000   32.100000   13.100000  172.000000  2700.000000
25%        86.250000   39.225000   15.600000  190.000000  3550.000000
50%       171.500000   44.450000   17.300000  197.000000  4050.000000
75%       256.750000   48.500000   18.700000  213.000000  4750.000000
max       342.000000   59.600000   21.500000  231.000000  6300.000000

Interquartile Range (IQR):
```

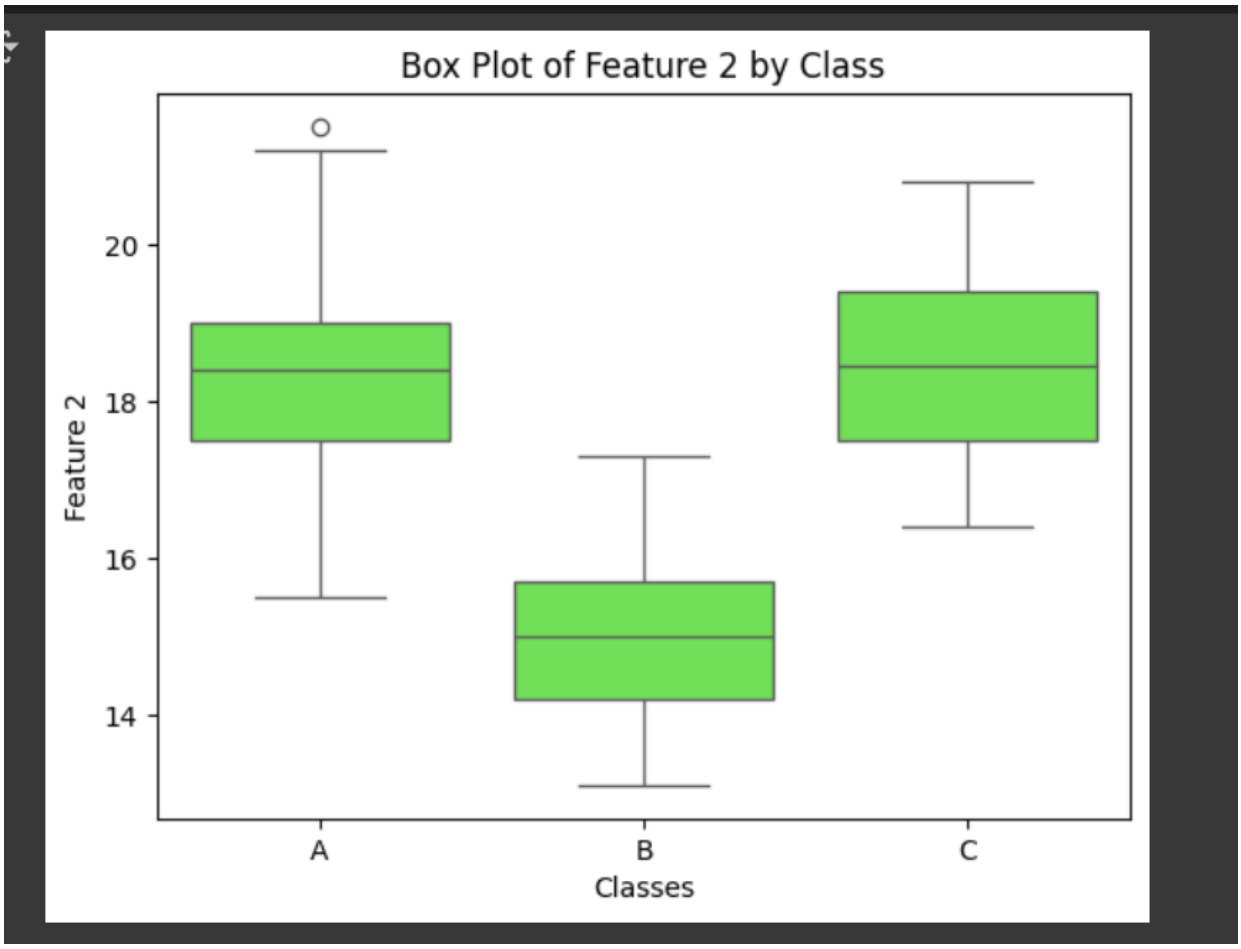
| | 0 |
|---|---|
| Sample Number | 170.500 |
| Feature 1 | 9.275 |
| Feature 2 | 3.100 |
| Feature 3 | 23.000 |
| Feature 4 | 1200.000 |

dtype: float64

c) A histogram of feature 1 for class A:

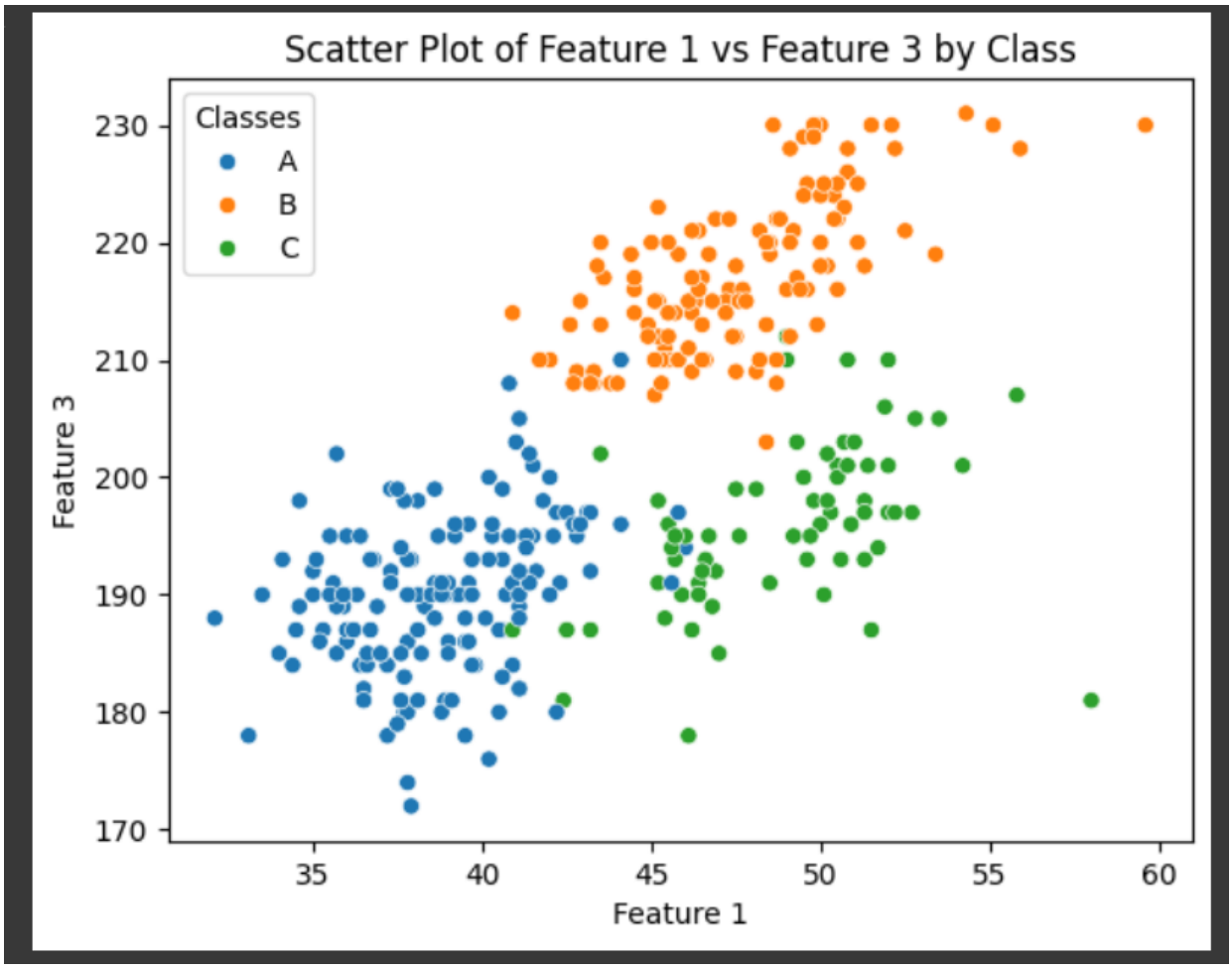


d) The box plot for feature 2 for each class separately:

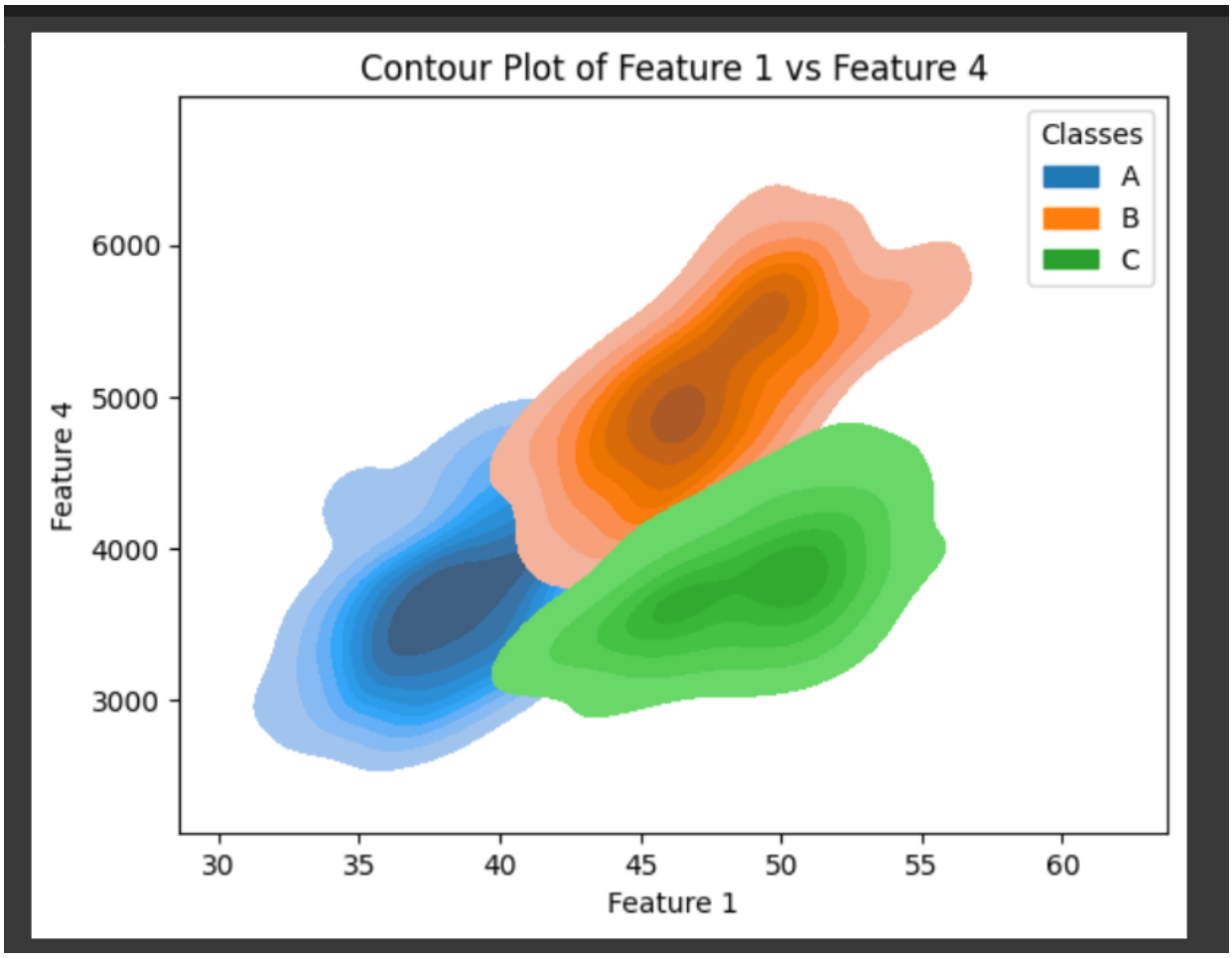


e) Violin plot for feature 3 for each class separately:

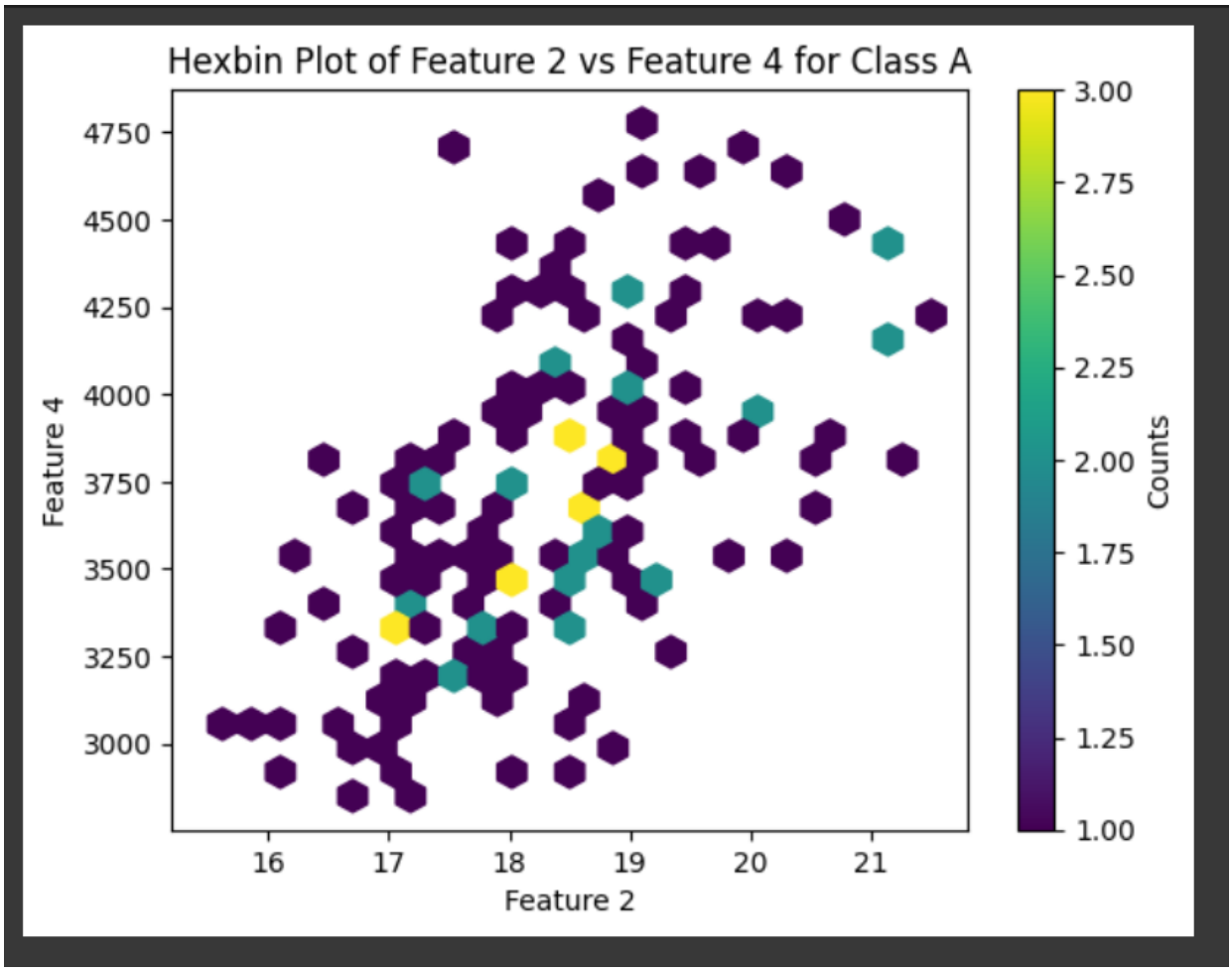
f) Scatter plots between feature 1 and feature 3 showing classes separately:



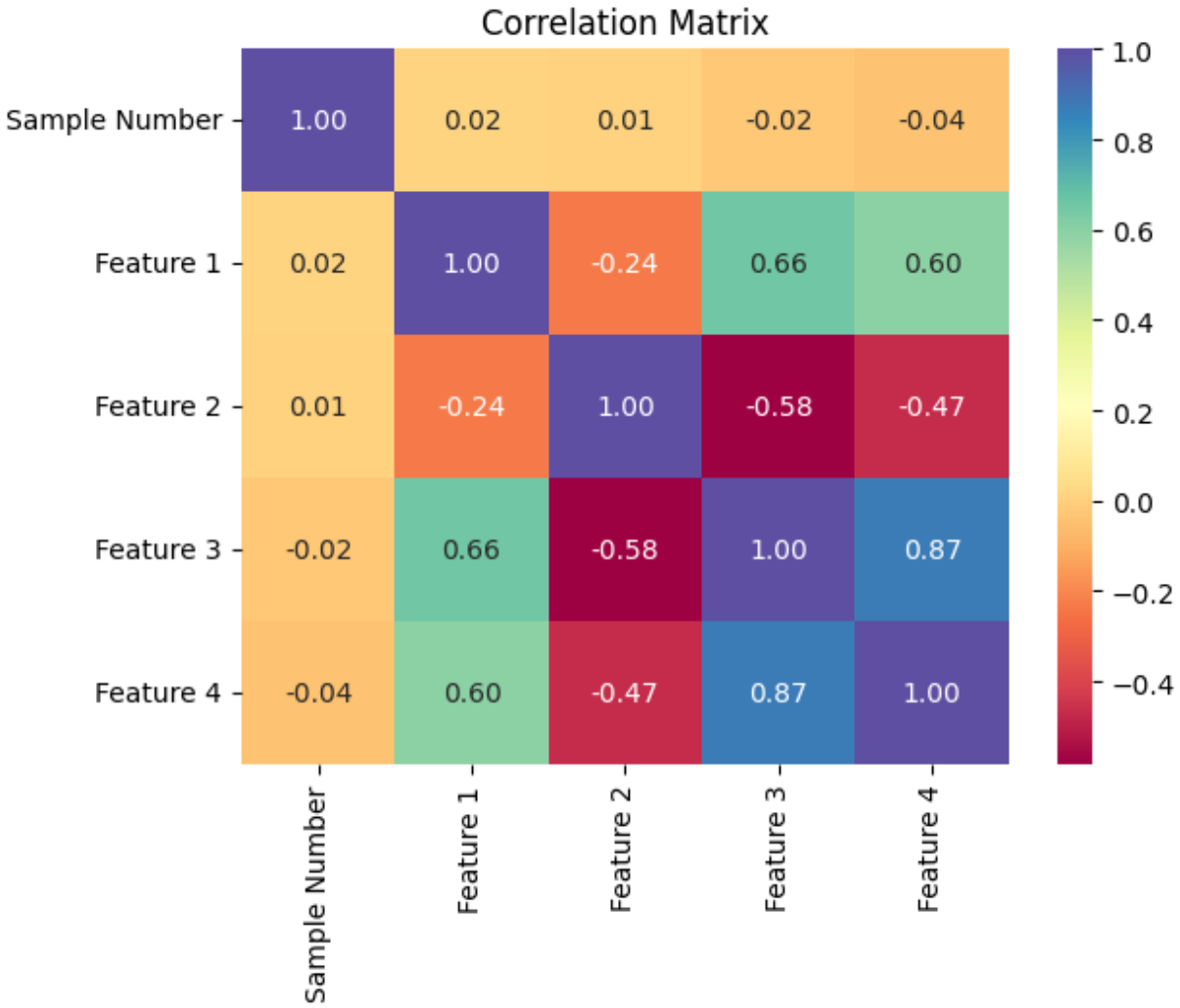
g) Contour plot between feature 1 and feature 4 showing classes separately:



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: