1.5em 0pt

Assignment 1

Course: EE708, Fundamentals of Data Science and Machine Intelligence

Nikhil Jain

Dept: Electrical Engineering

Roll No: 220709

1. We have two possibilities: scanning and e-mailing the image or using an optical character reader (OCR) to extract text and send a text file. Below is a comparative discussion of both approaches:

Scanning and E-mailing the Image:

- **Advantages:**
  - Maintains the exact formatting, layout, and visual details.
  - Can capture handwritten text or signatures accurately.
  - Works well for complex documents with diagrams, equations, or artistic elements.
- **Disadvantages:**
  - Larger file sizes compared to plain text.
  - Not searchable or editable without OCR.
  - Requires a good image resolution for readability.

Using OCR and Sending a Text File:

- **Advantages:**
  - Produces a searchable and editable document.
  - Smaller file size compared to image-based formats.
  - Can be easily copied, modified, and used in other documents.
- **Disadvantages:**
  - May introduce errors in text recognition, especially with poor-quality scans or complex formatting.
  - May lose original formatting, images, and non-text elements.
  - OCR technology may fail to recognize certain text accurately, potentially generating incorrect content instead of the original message.

When is Each Preferable?

- Scanning and e-mailing an image is preferable when maintaining visual integrity is crucial, such as for legal documents, signatures, or complex layouts.
- OCR and text file transmission are preferable when text needs to be searchable, editable, or efficiently stored and transmitted.

. . . .

2.

a. What is in a junk e-mail that lets us know it is junk?

Junk e-mails, also known as spam, often exhibit certain characteristics that make them identifiable. Some common traits include:

- **Unsolicited content:** Junk e-mails are typically unsolicited and sent to a large number of recipients. These e-mails may advertise products or services not requested by the recipient.
- **Suspicious subject lines:** Subject lines in junk e-mails often contain misleading or sensational phrases like ”You have won a prize!” or ”Exclusive offer just for you!”
- **Irrelevant or random content:** The body of the message may be filled with random, irrelevant, or repetitive content, often with multiple hyperlinks to suspicious websites.
- **Excessive use of capitalization and special characters:** Junk e-mails may use excessive capitalization (e.g., ”BUY NOW!!!”) or special characters to grab attention.
- **Obfuscation techniques:** Spammers often use techniques to obfuscate certain words or links, such as using misspellings or replacing characters with similar-looking ones (e.g., ”fr33” instead of ”free”).
- **Unusual attachments or links:** Junk e-mails may contain strange attachments or links that lead to phishing sites or malware.

By analyzing these patterns, a system can be trained to distinguish junk e-mails from legitimate ones.

b. How can the computer detect junk through a syntactic analysis?

Syntactic analysis, or parsing, involves analyzing the structure and syntax of the e-mail message to detect patterns indicative of junk content. Here’s how this can be applied:

- **Keyword matching:** A common approach in syntactic analysis is to check for specific keywords or phrases commonly found in junk e-mails. For example, words like ”free,” ”buy now,” or ”prize” might be flagged as suspicious.
- **Sentence structure patterns:** Junk e-mails may use specific sentence structures designed to grab attention, such as incomplete sentences or excessive use of punctuation marks (e.g., ”Hurry!! Limited time offer!!!”).
- **Frequency analysis:** High-frequency occurrences of certain phrases or symbols (e.g., repeated use of the word ”FREE”) can be a sign of junk.
- **Abnormal word sequences:** Syntactic analysis can help identify unnatural word sequences that are typical in spam messages, such as disjointed phrases or content that doesn’t follow proper grammatical rules.
- **URL analysis:** Detecting unusual or suspicious domain names in URLs and links within the e-mail can help identify junk e-mails. For example, URLs with random strings or strange domains are often indicative of phishing attempts.

Syntactic analysis is essential because it helps isolate the structural features of an e-mail that may not be captured by simple keyword matching, making the detection system more robust.

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?

The desired action when junk e-mails are detected depends on the preferences of the user or the system being used. Each approach has its own merits:

- **Delete it automatically:** This approach is ideal for users who receive a high volume of junk e-mails and prefer to have them removed from their inboxes without any manual intervention. However, there is a risk of false positives, where a legitimate e-mail may be mistakenly classified as junk.
- **Move it to a different file (e.g., spam folder):** Moving junk e-mails to a designated spam or junk folder is a more cautious approach. This allows users to review suspected junk e-mails later and ensures that no important e-mail is lost. Many e-mail clients already have this feature, which helps organize the inbox without permanently deleting messages.
- **Highlight it on the screen:** This method involves marking suspicious e-mails as junk without taking any action on them automatically. The e-mail remains in the inbox, but it is highlighted or flagged for the user's attention. This method is useful for users who prefer to manually decide whether an e-mail is junk or not.

Ultimately, the choice depends on the user's preferences and the tolerance for false positives or false negatives. For a more efficient system, a combination of these strategies (e.g., move to a spam folder with an option to automatically delete after a certain period) can be implemented.

. . . .

3.

Let us assume we are tasked with building an automated taxi system.

a. Define the constraints: The constraints for an automated taxi system can be divided into several categories:

- **Safety:** The taxi must ensure the safety of the passenger and the people on the road. This includes collision avoidance, adherence to traffic rules, and handling emergency situations.
- **Cost-efficiency:** The system must be cost-effective, minimizing fuel consumption, maintenance costs, and operational costs.
- **Time-efficiency:** The system should minimize travel time by selecting the most optimal route.
- **Comfort:** The ride must be smooth and comfortable for the passengers, including proper suspension and minimal jolts during the ride.
- **Reliability:** The taxi system must function in various weather conditions and environments, including rain, snow, and at night.
- **Security:** There must be mechanisms in place to ensure the safety of the passengers, such as real-time location tracking and emergency communication.

- **Regulatory compliance:** The taxi system must comply with local traffic laws, licensing requirements, and insurance policies.

b. What are the inputs? The inputs to the automated taxi system include:

- **Passenger request:** The location of the passenger and their destination.
- **Road conditions:** Information about road closures, traffic, and other obstructions.
- **Sensor data:** Real-time data from sensors such as cameras, LiDAR, radar, GPS, and accelerometers to understand the surrounding environment.
- **Maps:** Digital maps and routing information for navigation and traffic updates.
- **Weather conditions:** Information about the weather to make decisions related to speed and safety.
- **Fuel/battery level:** Monitoring of fuel or battery status to ensure the taxi does not run out of power.
- **Passenger preferences:** If available, passenger preferences like temperature, music, and route type (scenic, fastest, etc.).

c. What is the output? The output of the automated taxi system includes:

- **Passenger delivery:** Safe and timely delivery of the passenger to the desired destination.
- **Navigation instructions:** Detailed and optimized route instructions for the automated system.
- **Vehicle status:** Status updates such as fuel/battery level, maintenance alerts, and operational status.
- **Passenger interaction:** Confirmation messages regarding the pickup, drop-off, and estimated arrival time.

d. How can we communicate with the passenger? Communication with the passenger can be accomplished through the following methods:

- **In-vehicle display screen:** A screen in the vehicle could display information such as the current route, estimated time of arrival, and vehicle status.
- **Voice interface:** The taxi can use a voice assistant to communicate with the passenger, providing real-time updates on the ride and answering questions.
- **Mobile app:** The passenger can track the taxi's progress and receive notifications regarding the ride through a mobile app.
- **Text notifications:** Automated messages can be sent via SMS or the app, confirming ride requests, pick-up, and drop-off times.
- **Emergency contact:** In case of emergency, the passenger can contact support or an emergency service via a built-in communication system.

e. Do we need to communicate with the other automated taxis; that is, do we need a "language"? Yes, communication between automated taxis can be beneficial. This is where a common communication protocol or "language" would be required for:

- **Coordinating traffic:** Automated taxis could share their locations and routes to avoid congestion, optimize routes, and prevent accidents.
- **Fleet management:** A central system can communicate with multiple taxis to optimize the assignment of rides and allocate taxis to areas of high demand.
- **Emergency situations:** In case of accidents or road obstructions, taxis can communicate to reroute and share information.
- **Safety:** Taxis can exchange information about road conditions, hazards, or traffic updates, improving overall safety.
- **Data exchange:** Sharing operational data such as battery levels, fuel, and maintenance needs could help in better fleet management.

Thus, the system would require a shared protocol or language that all automated taxis understand to communicate effectively in real time.

. . . .

4.

a. What are the parameters?

For a circle, the parameters typically include:

- The radius of the circle, denoted as r.
- The center of the circle, which can be described by the coordinates (x_0, y_0) in a 2D Cartesian coordinate system.

Therefore, the parameters of a circle are:

$$\{r, (x_0, y_0)\}$$

where r is the radius, and (x_0, y_0) are the coordinates of the center.

b. How can the parameters of a circle hypothesis be calculated in such a case?

To calculate the parameters of a circle hypothesis (i.e., to find the radius r and the center (x_0, y_0)), we can use the general equation of a circle in the 2D plane:

$$(x - x_0)^2 + (y - y_0)^2 = r^2$$

Given a set of data points that we believe lie on the circumference of the circle, we can perform the following steps:

1. **Estimating the center (x_0, y_0):** - Use techniques like least squares fitting or optimization algorithms (e.g., gradient descent) to minimize the sum of squared distances from the data points to the center of the circle.

2. **Estimating the radius r:** - Once the center (x_0, y_0) is estimated, the radius can be calculated as the average distance from the center to all the data points:

$$r = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$$

where (x_i, y_i) are the coordinates of the data points, and N is the total number of points.

c. What if it is an ellipse?

For an ellipse, the hypothesis class involves more parameters. The general equation of an ellipse in a 2D Cartesian coordinate system is:

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} = 1$$

where:

- (x_0, y_0) is the center of the ellipse.
- a is the length of the semi-major axis (the longest radius).
- b is the length of the semi-minor axis (the shortest radius).

Therefore, the parameters of an ellipse are:

$$\{a, b, (x_0, y_0)\}$$

where a and b are the lengths of the semi-major and semi-minor axes, and (x_0, y_0) is the center of the ellipse.

To estimate these parameters from data points: 1. Use a method such as least squares fitting or an optimization algorithm to minimize the difference between the observed data points and the ellipse equation. 2. Alternatively, algebraic methods like direct least squares fitting can be employed to solve for a, b, and (x_0, y_0).

d. Why does it make more sense to use an ellipse instead of a circle?

An ellipse often provides a better fit than a circle in many real-world scenarios because:

- **Non-circular distributions:** Data points that approximate a circle in some cases might not lie perfectly on a circle, but they could still fit well on an ellipse with different axis lengths.
- **More flexibility:** A circle assumes symmetry in all directions (i.e., equal radii), which may not be true for real-world data. An ellipse, with different lengths for the semi-major and semi-minor axes, allows for more flexibility in modeling data that does not exhibit perfect circular symmetry.
- **Better approximation of ellipsoidal shapes:** Many objects and patterns in real life (such as planetary orbits, biological shapes, etc.) are more naturally represented as ellipses rather than circles.

Thus, an ellipse can accommodate a wider variety of data shapes, making it a more general and useful model in many applications.

....

5. We are given that the Z-Phone has a mean consumer life of 42 months with a standard deviation of 8 months, and that the lifespan follows a normal distribution. We are tasked with finding the probability that a random Z-Phone will last between 20 and 30 months.

Let the random variable X represent the lifespan of the Z-Phone, which follows a normal distribution with mean $\mu = 42$ months and standard deviation $\sigma = 8$ months. Therefore, $X \sim N(42, 8^2)$.

We need to find $P(20 \leq X \leq 30)$, which is the probability that the lifespan of the Z-Phone lies between 20 and 30 months.

To solve this, we first standardize the values using the Z-score formula:

$$Z = \frac{X - \mu}{\sigma}$$

For $X = 20$:

$$Z_1 = \frac{20 - 42}{8} = \frac{-22}{8} = -2.75$$

For $X = 30$:

$$Z_2 = \frac{30 - 42}{8} = \frac{-12}{8} = -1.5$$

Now, we need to find the probability that the Z-score lies between $Z_1 = -2.75$ and $Z_2 = -1.5$. This can be written as:

$$P(20 \leq X \leq 30) = P(-2.75 \leq Z \leq -1.5)$$

Using standard normal distribution tables or a calculator, we find the following values:

$$P(Z \leq -1.5) \approx 0.0668$$
$$P(Z \leq -2.75) \approx 0.0030$$

Thus, the probability is:

$$P(-2.75 \leq Z \leq -1.5) = P(Z \leq -1.5) - P(Z \leq -2.75) \approx 0.0668 - 0.0030 = 0.0638$$

Therefore, the probability that a given random Z-Phone will last between 20 and 30 months is approximately 0.0638, or 6.38%.

....

6.

We are given the failure times of eight components tested under accelerated conditions. The failure times are as follows:

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

The observation "100" indicates that the unit still functioned at 100 hours, implying it is a censored observation (i.e., the unit survived at least 100 hours but we do not know the exact time of failure).

To calculate a meaningful measure of location for these data, we consider the following:

1. **Censored Data Consideration**: Since the observation 100 indicates that the unit did not fail within the 100-hour period, we treat it as a censored observation, i.e., we know the failure time is at least 100 hours.

2. **Mean**: The mean is calculated as the sum of the failure times divided by the number of components. However, since we have a censored observation, the mean would not fully represent the data. We can compute the mean of the uncensored data first and then consider the censored nature of the data if needed.

The sum of the uncensored failure times is:

$$75 + 63 + 36 + 51 + 45 + 80 + 90 = 440$$

The number of uncensored data points is 7 (since the 100-hour data point is censored). Therefore, the mean of the uncensored data is:

$$\text{Mean} = \frac{440}{7} \approx 62.86$$

3. **Median**: The median is the middle value when the data is ordered. Ordering the data:

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

Since we have 8 data points (including the censored one), the median is the average of the 4th and 5th values in the ordered list:

$$\text{Median} = \frac{63 + 75}{2} = \frac{138}{2} = 69$$

4. **Interpretation**: The median, which is 69 hours, is a meaningful measure of location for these data. This is because the median is less sensitive to extreme values and censored data points, making it a better measure of central tendency when data includes censored observations.

Thus, the meaningful measure of location for these data is the median, and its numerical value is 69.

....

6.

The following cold start times (in seconds) were recorded for two formulations of gasoline.

For the first formulation:

$$1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91$$

For the second formulation:

$$1.83, 1.99, 3.13, 3.29, 2.65, 2.87, 3.40, 2.46, 1.89, 3.35$$

The sample mean, variance, and standard deviation for the two formulations are as follows:

For **Formulation 1**:

- Sample Mean: 2.415 seconds
- Sample Variance: 0.2854 seconds^2
- Sample Standard Deviation: 0.5342 seconds

For **Formulation 2**:

- Sample Mean: 2.686 seconds
- Sample Variance: 0.3833 seconds^2
- Sample Standard Deviation: 0.6191 seconds

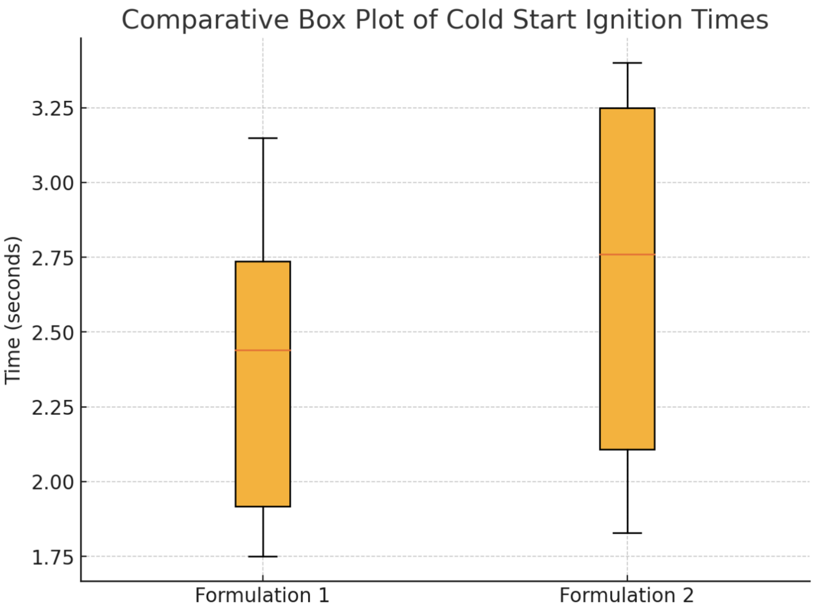


Figure 1: Comparative Box Plot of Cold Start Times for Two Gasoline Formulations

The comparative box plot of the cold start times for both formulations is shown below: **Interpretation:**

- **Formulation 1** shows a smaller spread of values and lower variability compared to **Formulation 2**. The interquartile range (IQR) for Formulation 1 is narrower, indicating less variability in the ignition times.
- **Formulation 2** has a wider spread, with more variation between the values, as seen from the larger IQR and higher standard deviation.
- The median for **Formulation 1** is slightly lower than for **Formulation 2**, suggesting that Formulation 1 has a generally quicker cold start compared to Formulation 2.
- Both formulations show no extreme outliers, but **Formulation 2** has a few more values towards the higher end, indicating that some cold start times may be notably slower than others.

In summary, **Formulation 1** appears to provide more consistent and slightly faster cold start times, while **Formulation 2** has more variability and slightly slower average cold start times.

. . . .

7.

We are given a table of patient records with the following columns: Name, Weight (kg), Height (m), Systolic Blood Pressure (mm Hg), Diastolic Blood Pressure (mm Hg), and Diabetes. The goal is to perform some data transformations on the Weight (kg) variable.

a. Min-Max Normalization of Weight (kg)

The min-max normalization transforms the values of a variable to a specific range, typically [0, 1]. The formula for min-max normalization is given by:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where: - X is the original value, - $X_{\min}$ is the minimum value in the dataset, - $X_{\max}$ is the maximum value in the dataset.

For the Weight (kg) variable, the minimum and maximum values are: - $X_{\min} = 41$ (for A. Patel), - $X_{\max} = 136$ (for F. Marsh).

The normalized Weight (kg) values for each patient are calculated using the above formula. The following table shows the transformed values:

$$\text{Normalized Weight (kg)} = \frac{\text{Weight (kg)} - 41}{136 - 41}$$

Name	**Weight (kg)**	**Normalized Weight (kg)**
P. Lee	50	$\frac{50-41}{136-41} = 0.066$
R. Jones	115	$\frac{115-41}{136-41} = 0.548$
J. Smith	96	$\frac{96-41}{136-41} = 0.408$
A. Patel	41	$\frac{41-41}{136-41} = 0.000$
M. Owen	79	$\frac{79-41}{136-41} = 0.283$
S. Green	109	$\frac{109-41}{136-41} = 0.501$
N. Cook	73	$\frac{73-41}{136-41} = 0.236$
W. Hands	104	$\frac{104-41}{136-41} = 0.464$
P. Rice	64	$\frac{64-41}{136-41} = 0.179$
F. Marsh	136	$\frac{136-41}{136-41} = 1.000$

b. Binning the Weight (kg) variable

The Weight (kg) variable can be divided into three categories: low (less than 60 kg), medium (60-100 kg), and high (greater than 100 kg). The binning can be done using the following ranges:

- Low: Weight $< 60\,\text{kg}$, - Medium: $60 \leq$ Weight $\leq 100\,\text{kg}$, - High: Weight $> 100\,\text{kg}$.

The binning results for each patient are as follows:

Name	**Weight (kg)**	**Weight Category**
P. Lee	50	Low
R. Jones	115	High
J. Smith	96	Medium
A. Patel	41	Low
M. Owen	79	Medium
S. Green	109	High
N. Cook	73	Medium
W. Hands	104	High
P. Rice	64	Medium
F. Marsh	136	High

c. Aggregated Column: Body Mass Index (BMI)

The Body Mass Index (BMI) is defined by the formula:

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

We will calculate the BMI for each patient using their Weight (kg) and Height (m) values.

$$\text{BMI for P. Lee} = \frac{50}{(1.52)^2} = 21.65$$

$$\text{BMI for R. Jones} = \frac{115}{(1.77)^2} = 36.71$$

$$\text{BMI for J. Smith} = \frac{96}{(1.83)^2} = 28.57$$

$$\text{BMI for A. Patel} = \frac{41}{(1.55)^2} = 17.02$$

$$\text{BMI for M. Owen} = \frac{79}{(1.82)^2} = 23.84$$

$$\text{BMI for S. Green} = \frac{109}{(1.89)^2} = 30.47$$

$$\text{BMI for N. Cook} = \frac{73}{(1.76)^2} = 23.59$$

$$\text{BMI for W. Hands} = \frac{104}{(1.71)^2} = 35.61$$

$$\text{BMI for P. Rice} = \frac{64}{(1.74)^2} = 21.13$$

$$\text{BMI for F. Marsh} = \frac{136}{(1.78)^2} = 42.94$$

The BMI values for each patient are shown below:

Name	**BMI**
P. Lee	21.65
R. Jones	36.71
J. Smith	28.57
A. Patel	17.02
M. Owen	23.84
S. Green	30.47
N. Cook	23.59
W. Hands	35.61
P. Rice	21.13
F. Marsh	42.94

. . . .

8.

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

Table 1: Contingency Table of Store vs Product Category

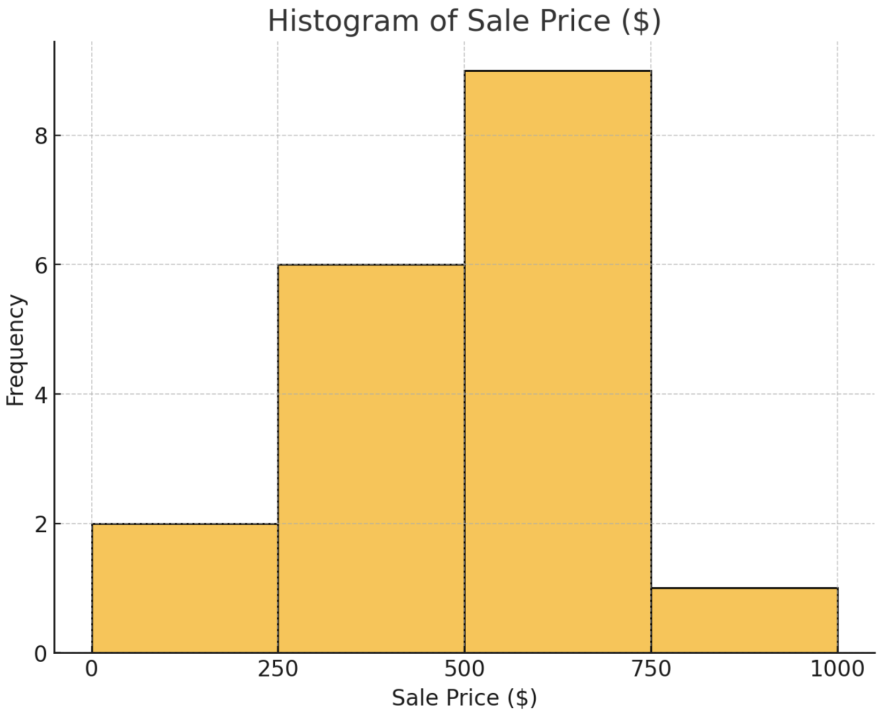


Figure 2: Histogram of Sale Price ($)

9.

Customer	Transaction Count	Total Sale Price ($)
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

Table 2: Summary by Customer

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

Table 3: Summary by Store

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

Table 4: Summary by Product Category

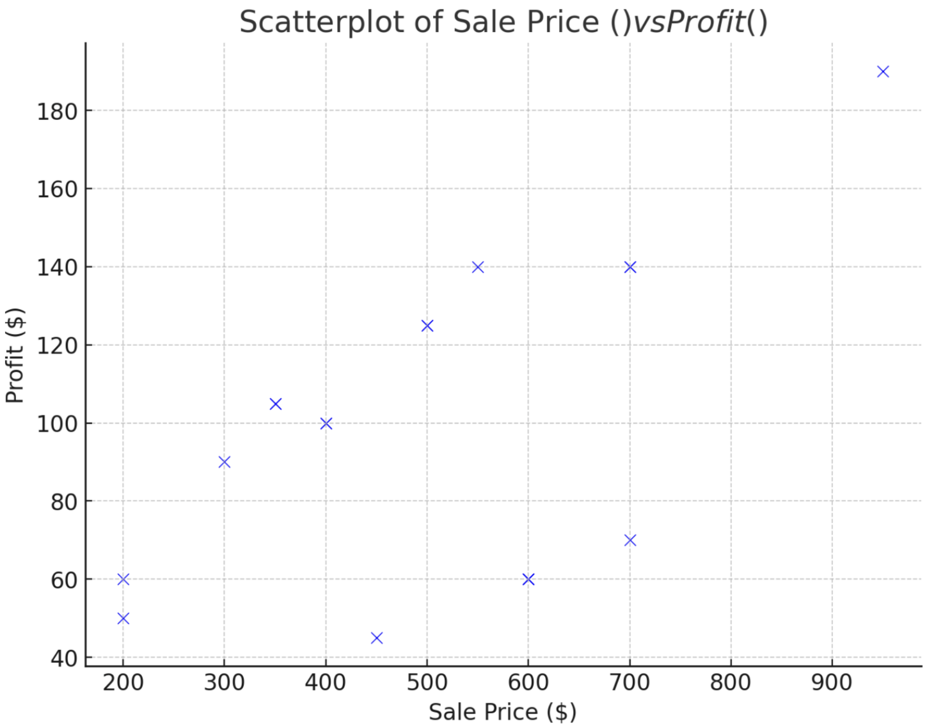


Figure 3: Scatterplot of Sale Price ($) vs Profit ($)

Q9. Classes	Frequency
A	151
B	123
C	68

	Sample Number	Feature 1	Feature 2	Feature 3	Feature 4
count	342.000	342.000	342.000	342.000	342.000
mean	171.500	43.922	17.151	200.915	4201.754
std	98.871	5.460	1.975	14.062	801.955
min	1.000	32.100	13.100	172.000	2700.000
25%	86.250	39.225	15.600	190.000	3550.000
50%	171.500	44.450	17.300	197.000	4050.000
75%	256.750	48.500	18.700	213.000	4750.000
max	342.000	59.600	21.500	231.000	6300.000

Feature	IQR
Feature 1	9.275
Feature 2	3.100
Feature 3	23.000
Feature 4	1200.000

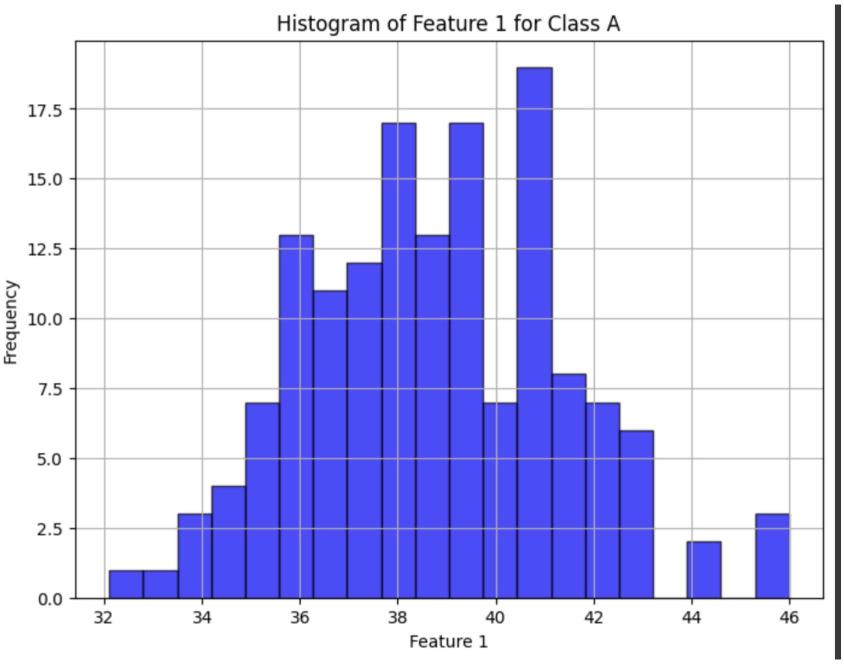


Figure 4: Histogram of Feature 1 for Class A

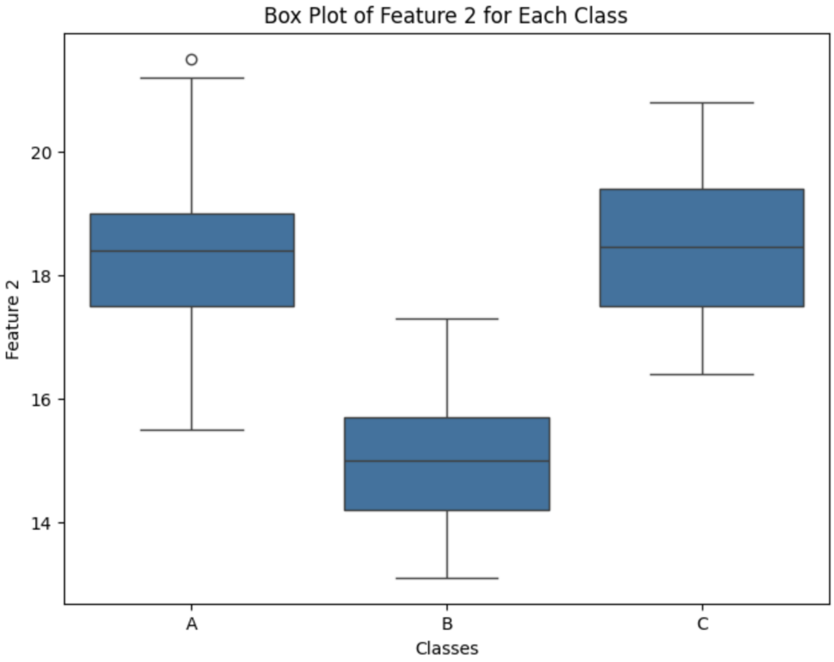


Figure 5: Box Plot of Feature 2 for Each Class

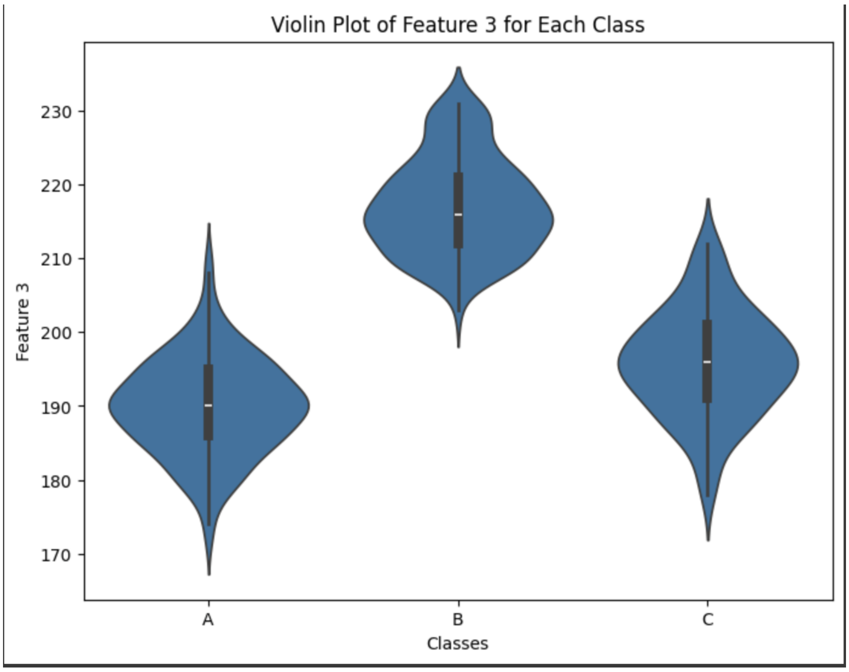


Figure 6: Violin Plot of Feature 3 for Each Class

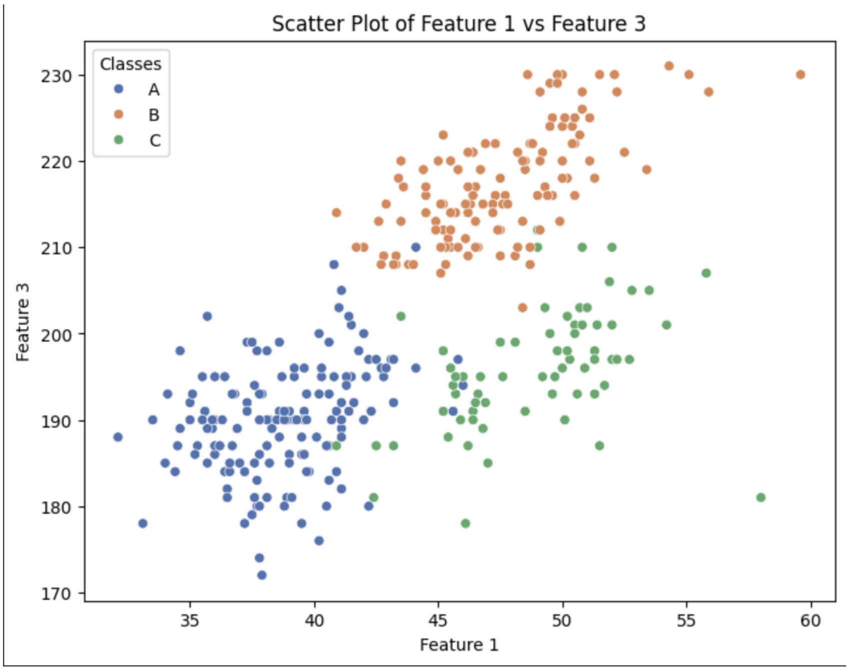


Figure 7: Scatter Plot of Feature 1 vs Feature 3

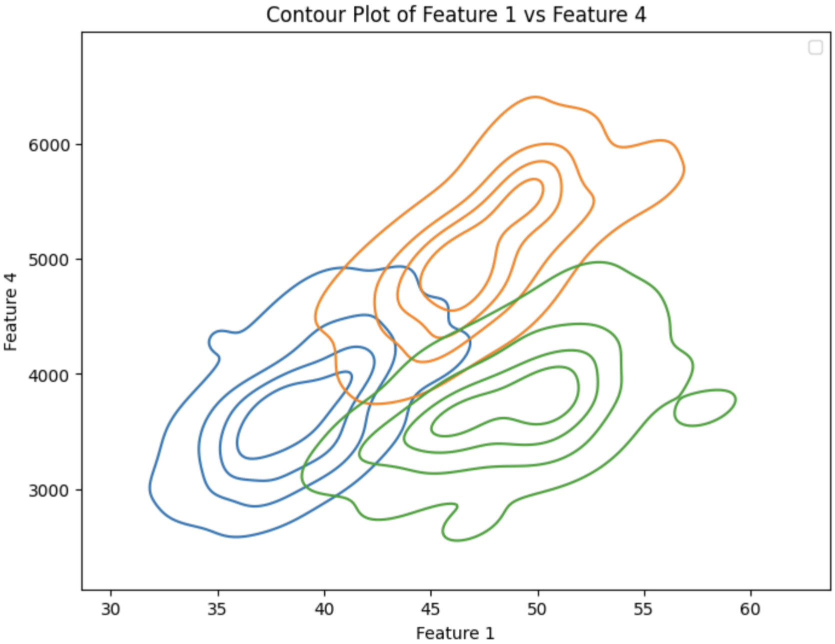


Figure 8: Contour Plot of Feature 1 vs Feature 4

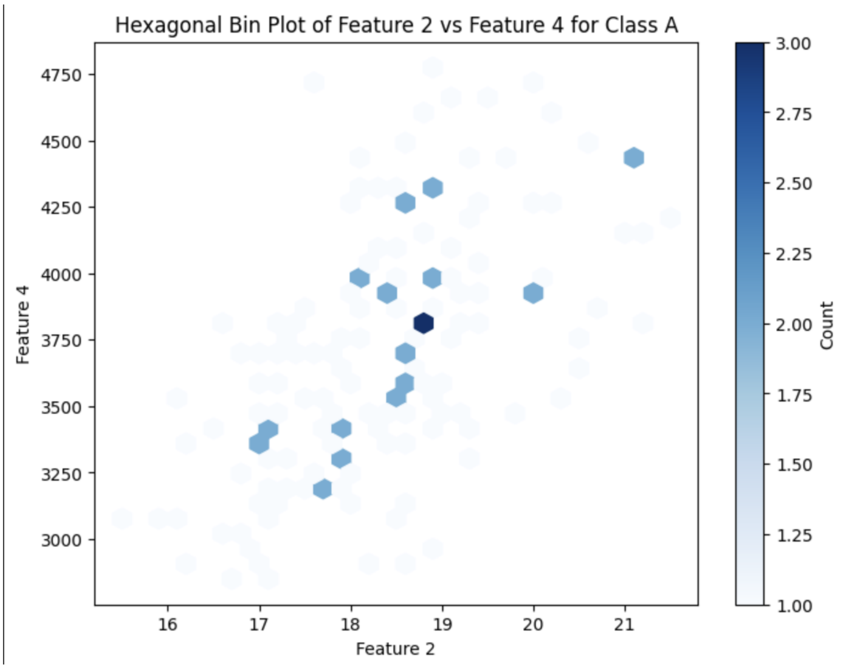
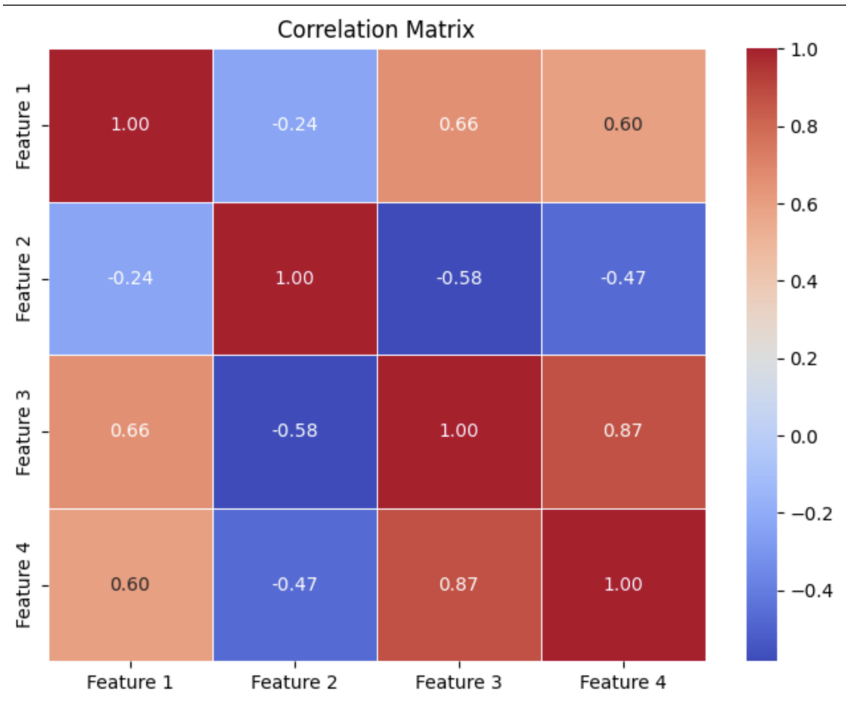


Figure 9: Hexagonal Bin Plot of Feature 2 vs Feature 4 for Class A



	Feature 1	Feature 2	Feature 3	Feature 4
Feature 1	1.00	-0.24	0.66	0.60
Feature 2	-0.24	1.00	-0.58	-0.47
Feature 3	0.66	-0.58	1.00	0.87
Feature 4	0.60	-0.47	0.87	1.00

Figure 10: Correlation Matrix

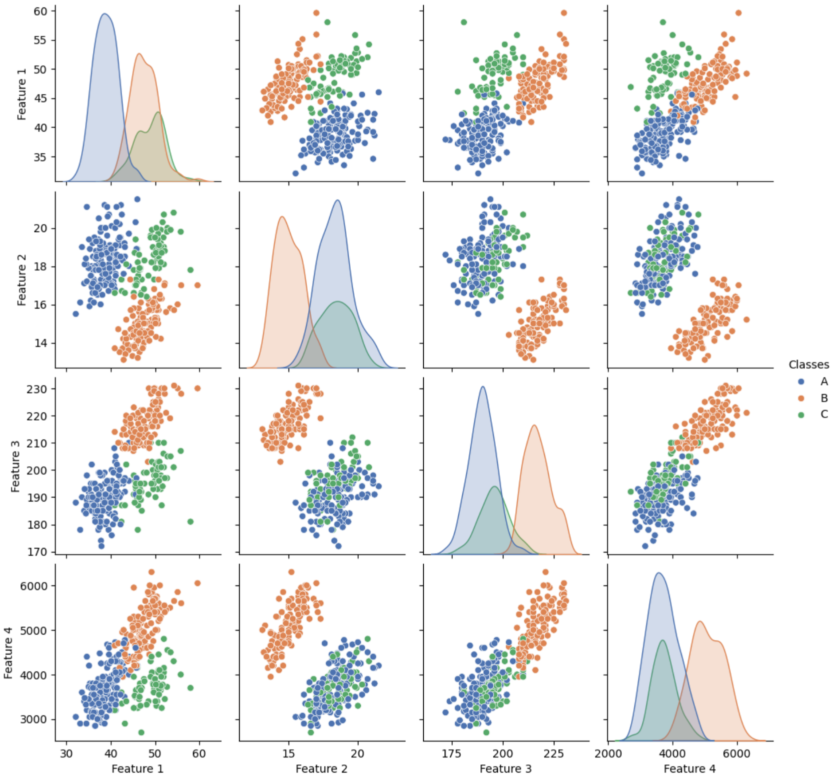


Figure 11: Pair Plot of Features Showing Classes