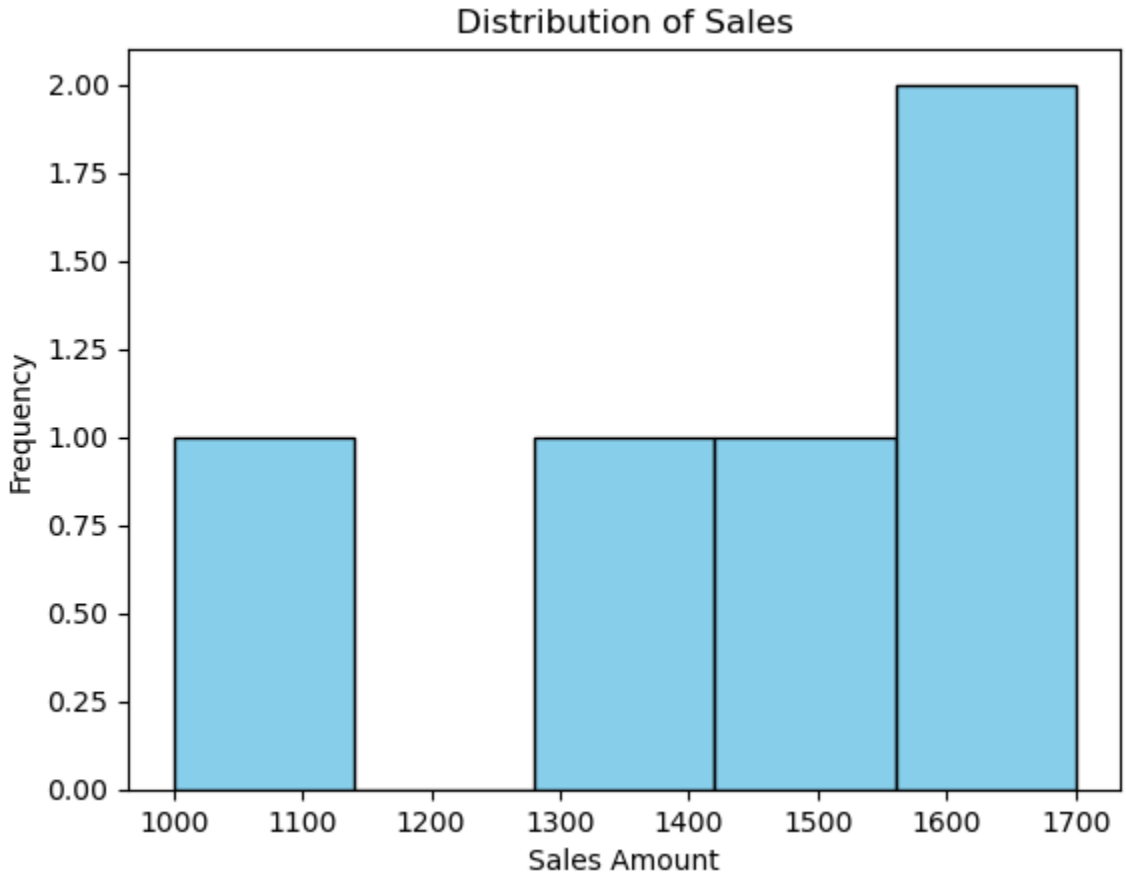
```
import pandas as pd

import matplotlib.pyplot as plt


# 1. Load your CSV file

#    On Windows, either use a raw string (prefix with r) or escape backslashes.

file_path = r'D:\csvfiles\sales_data.csv'

df = pd.read_csv(file_path)

# 2. Inspect the data

print(df.head())

print(df.info())

print(df.describe())
```

```
plt.hist(df['Sales'], bins=5, color='skyblue', edgecolor='black')

plt.title("Distribution of Sales")

plt.xlabel("Sales Amount")

plt.ylabel("Frequency")

plt.show()
```

The histogram is a visualization of how the five monthly “Sales” values in the dataset are spread out across different ranges (bins).



1. Identify the minimum and maximum

- Minimum sale = 1000
- Maximum sale = 1700

So the full range is from 1000 up to 1700, which is a span of

$$1700 - 1000 = 700.$$

2. Compute the bin width

We want 5 bins of equal width, so each bin's width is

$$\text{bin width} = \frac{(\max - \min)}{\#\text{ of bins}} = \frac{700}{5} = 140.$$

---

3. Determine all bin edges

Starting at the minimum (1000), each edge is +140 from the previous:

1. **Edge 0 (start)**: 1000
2. **Edge 1**: 1000 + 140 = 1140
3. **Edge 2**: 1140 + 140 = 1280
4. **Edge 3**: 1280 + 140 = 1420
5. **Edge 4**: 1420 + 140 = 1560
6. **Edge 5 (end)**: 1560 + 140 = 1700

So the five bins cover these intervals (we'll label them Bin 1 through Bin 5):

- **Bin 1:** [1000, 1140)
- **Bin 2:** [1140, 1280)

- **Bin 3:** [1280, 1420)
- **Bin 4:** [1420, 1560)
- **Bin 5:** [1560, 1700]

(Note: By convention, the left edge is inclusive "[" and the right edge is exclusive ")", except that the very last bin can include the max.)

---

4. Locate which bin 1300 falls into

Check 1300 against each interval:

- **Is 1300 ≥ 1000 and < 1140?**
  • No, because 1300 ≥ 1140.
- **Is 1300 ≥ 1140 and < 1280?**
  • No, because 1300 ≥ 1280.
- **Is 1300 ≥ 1280 and < 1420?**
  • Yes—1300 lies between 1280 and 1420.

Hence 1300 sits in Bin 3 (the [1280, 1420) interval).

---

5. Confirm the "1 month" count

Since the only sales value that lands in the [1280, 1420) interval is 1300, that makes the frequency (count) for Bin 3 equal to **1**.

To summarize:

Sales Value	Falls Into Which Bin Interval?	Bin #
1000	[1000, 1140)	1
1300	[1280, 1420)	3
1500	[1420, 1560)	4
1600	[1560, 1700]	5
1700	[1560, 1700]	5

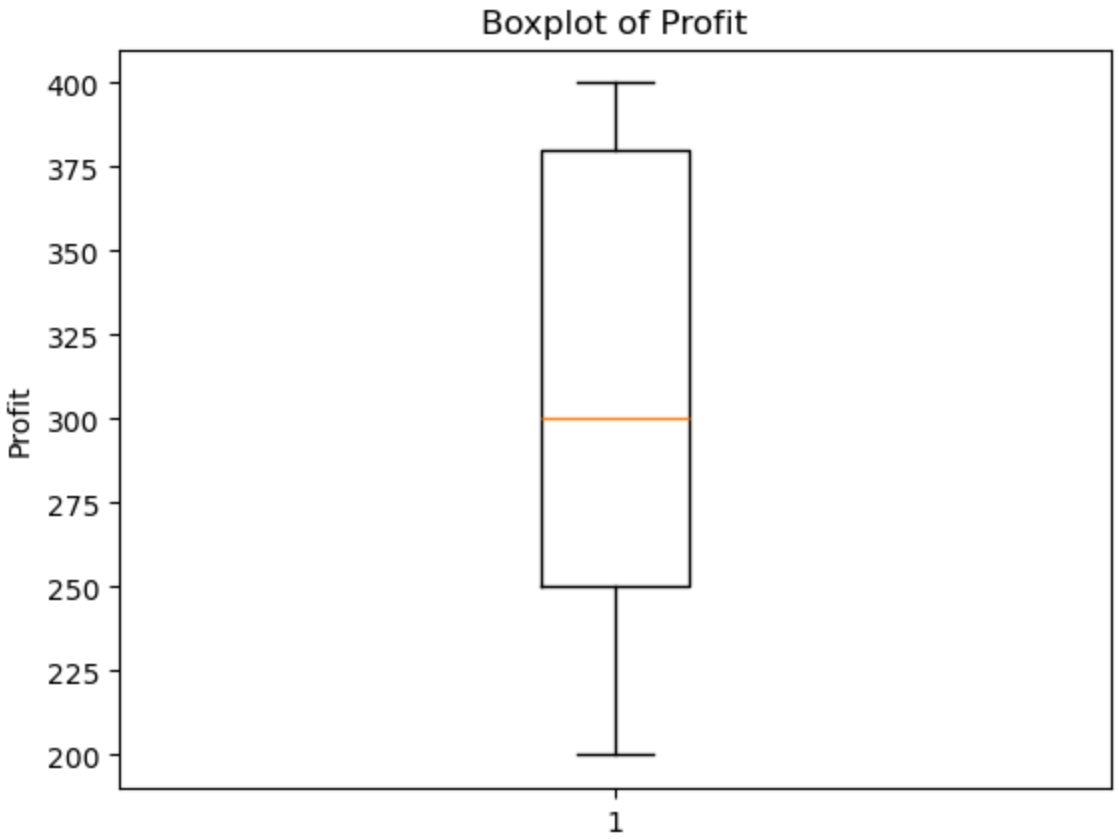
Sales values: [1000, 1500, 1300, 1700, 1600]

Bins (intervals):

Bin 1: 1000–1140 → contains [1000] → frequency = 1
Bin 2: 1140–1280 → contains [] → frequency = 0
Bin 3: 1280–1420 → contains [1300] → frequency = 1
Bin 4: 1420–1560 → contains [1500] → frequency = 1
Bin 5: 1560–1700 → contains [1600, 1700] → frequency = 2

In general, the **height** of each bar in a histogram is its **frequency**—how many observations (data points) fall into that bin.

```
# 4. Plot: Boxplot of Profit (Identify outliers, spread)
plt.boxplot(df['Profit'])
plt.title("Boxplot of Profit")
plt.ylabel("Profit")
plt.show()
```



A **boxplot (or "box-and-whisker" plot)** gives you a quick visual summary of how your **Profit values are distributed**. Let's walk through exactly what each part of that boxplot means for five Profit numbers:

Profit values: [200, 300, 250, 400, 380]

---

1. Sort the data

First, put the numbers in order:

[200, 250, 300, 380, 400]

To find the **median** of the list $200, 250, 300, 380, 400$, follow these steps:

1. **Sort the data (if it isn't already sorted)**

   In this case, the list is already in ascending order:

[200, 250, 300, 380, 400]

Count how many values there are

There are 5 numbers in total.

3. **For an odd number of values, the median is the middle element**
   - When $n = 5$, the "middle" position is at index $\frac{n+1}{2} = \frac{5+1}{2} = 3$ (counting 1st, 2nd, 3rd, etc.).
   - In zero-based indexing (like Python), that corresponds to index 2 (because indices run 0,1,2,3,4).
4. **Pick the 3rd element in the sorted list**

1st element: 200
2nd element: 250
3rd element: 300 ← this is the median
4th element: 380
5th element: 400

---

2. Key statistics

1. **Minimum (min)**
   - The smallest value is **200**.
2. **First Quartile (Q1)**
   - Q1 is the "middle" of the lower half.
   - With five points, the overall median is 300, so the lower half is [**200, 250].**
   - The median of **[200, 250] is (200 + 250)/2 = 225.**
3. **Median (Q2)**

- The overall middle value is **300** (the third number in the sorted list).

4. **Third Quartile (Q3)**
   - Q3 is the "middle" of the upper half, which is [380, 400].
   - The median of [380, 400] is (380 + 400)/2 = **390**.
5. **Maximum (max)**
   - The largest value is **400**.

---

3. How those numbers map to the boxplot

1. **The Box (the "interquartile range" or IQR)**
   - Bottom of box = Q1 = **225**
   - Top of box = Q3 = **390**
   - **So the box spans from 225 up to 390.**
   - Everything inside that box represents the middle **50% of your data (i.e., from the 25th percentile up to the 75th percentile).**
2. **The Line inside the Box (the Median)**
   - **Drawn at Profit = 300.**
   - This line shows that half of your five profits are below 300 and half are above 300.
3. **Whiskers (lines extending from the box)**
   - Lower whisker runs from Q1 down to the minimum (200).
   - Upper whisker runs from Q3 up to the maximum (400).
   - In this simple case, none of the points are far enough from Q1 or Q3 to be considered "outliers," so the whiskers simply touch the min and max.
4. **Outliers (dots beyond the whiskers)**

   a. What Is an "Outlier" in a Boxplot?

   - In a boxplot, any data point that lies "far outside" the bulk of the values is marked with a dot (or a small circle).

- The standard rule is:
  Anything more than 1.5 × IQR below the first quartile (Q1) or above the third quartile (Q3) is called an outlier.

---

2. Compute the IQR (Interquartile Range)

profit values sorted:

[200, 250, 300, 380, 400]

- **Q1 (first quartile)** = 225
- **Q3 (third quartile)** = 390

So

IQR = Q3 – Q1 = 390 – 225 = 165.

---

3. Find the "1.5 × IQR" Distance

Multiply that IQR by 1.5:

1.5 × IQR = 1.5 × 165 = 247.5.

This number (247.5) tells us how far from the "middle 50% box" a point has to be to count as an outlier.

---

4. Calculate the Cutoff Values

1. **Lower cutoff** = Q1 – (1.5 × IQR)
   = 225 – 247.5
   = –22.5

   In other words, any profit below –22.5 would be considered an outlier on the low side.

2. **Upper cutoff** = Q3 + (1.5 × IQR)
   = 390 + 247.5
   = 637.5

   Any profit above 637.5 would be an outlier on the high side.

---

5. Compare Your Actual Profit Values

profits are:

[200, 250, 300, 380, 400]

- Is any profit **< −22.5**? No (the smallest is 200).
- Is any profit **> 637.5**? No (the largest is 400).

Since **all five numbers lie between −22.5 and 637.5**, none of them fall outside the cutoff range. **Therefore, there are no outliers**, and you won't see any extra dots beyond the whiskers in the boxplot.

In Plain English

1. **Find the middle 50% box** (225 up to 390).
2. **Stretch that box** out by 1.5 times its height (165×1.5 = 247.5) in both directions.
3. **Any point outside that stretched range** would be "too far away" and get a dot.
4. **All your profits (200–400)** stay inside those stretched limits, so **no dots appear**—no outliers.

---

4. Reading the plot Box from 225 to 390:

50% of your profits fall between $225 and $390.

- **Median line at 300**:
  Half of the months had Profit below $300, half above.
- **Whiskers down to 200 and up to 400**:
  Your full range of profits is from $200 (the lowest month) to $400 (the highest month).
- **No outlier dots**:
  None of your profit values are unusually far from the rest.

In plain terms

- The **lowest profit** was $200, and the **highest profit** was $400.
- The **"middle" profit** (median) is $300.
- Most months' profits (the central 50%) lie between about $225 and $390.
- There are no extremely unusual profit values.