Visualization and Analysis Report

Analytics and Systems of Big Data

Thallapally Nimisha
CS22B1082
B.Tech in Computer Science and Engineering
IIITDM Kancheepuram

Contents

1 Question 1: Histogram

Problem Statement:
On New Year's Eve, Tina walked into a random shop and was surprised to see a huge crowd there. She is interested to find what kind of products they sell the most, for which she needs the age distribution of customers. Help her to find out the same using histogram. The age details of the customers are given. Identify the type of histogram (e.g. Bimodal, Multimodal, Skewed..etc). Use different bin sizes.

1.1 Solution:

The dataset of ages is:

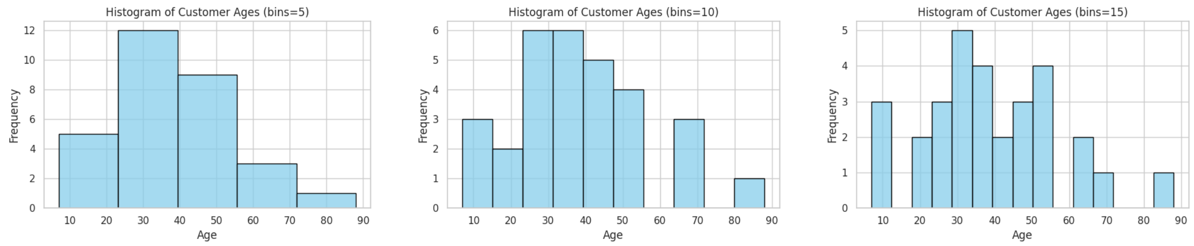
$$7, 9, 27, 28, 55, 45, 34, 65, 54, 67, 34, 23, 24, 66, 53, 45, 44, 88, 22, 33, 55, 35, 33, 37, 47, 41, 31, 30, 29, 12$$

Approach:

- Constructed histograms with different bin sizes.
- Observed distribution shape.

Observation: The distribution shows multiple peaks → **Multimodal**. There is slight right skew due to the outlier at 88.

1.2 Output:



(a) Bin size 5

(b) Bin size 10

(c) Bin size 15

Figure 1: Histograms of customer ages with varying bin sizes

2 Question 2: Stem-Leaf Plot and Outliers

Problem Statement:
A Coach tracked the number of points scored by 30 players. Visualize the data using ordered stem-leaf plot and also detect the outliers and shape of the distribution.

2.1 Solution:

Dataset:

$$22, 21, 24, 19, 27, 28, 24, 25, 29, 28, 26, 31, 28, 27, 22, 39, 20, 10, 26, 24, 27, 28, 26, 28, 18, 32, 29, 25, 31, 27$$

Approach:

- Construct ordered stem-leaf plot.
- Detect outliers using IQR method.

Observation: - Outlier detected at score = 39. - Distribution is slightly right-skewed.

2.2 Output:

Stem-and-Leaf Plot (Ordered):

```
Stem | Leaves
--------------
 1   | 0 8 9
 2   | 0 1 2 2 4 4 4 5 5 6 6 6 7 7 7 7 8 8 8 8 8 9 9
 3   | 1 1 2 9
```

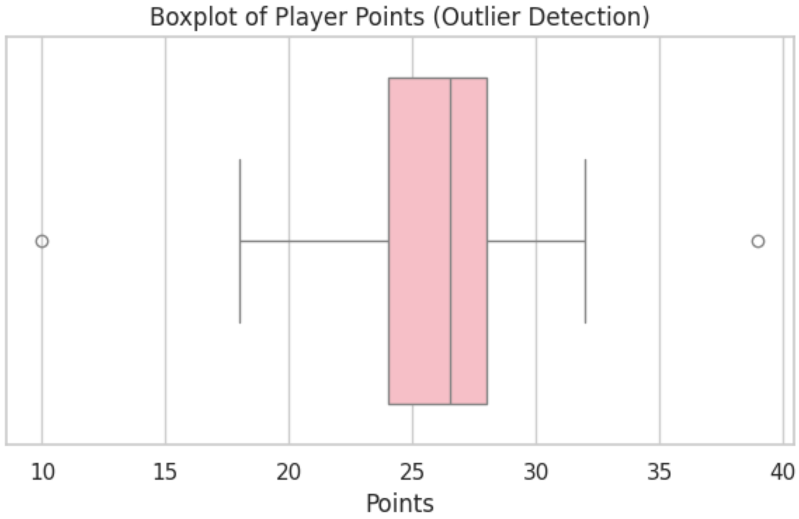


Figure 2: Box Plot

3 Question 3: Density and Rug Plot

Problem Statement:
Visualize water and beverage consumption of 15 people using density plot, rug plot and identify mean, median, mode and skewness.

3.1 Solution:

Water consumption (L): 3.2, 3.5, 3.6, 2.5, 2.8, 5.9, 2.9, 3.9, 4.9, 6.9, 7.9, 8.0, 3.3, 6.6, 4.4

Beverages (L): 2.2, 2.5, 2.6, 1.5, 3.8, 1.9, 0.9, 3.9, 4.9, 6.9, 0.1, 8.0, 0.3, 2.6, 1.4

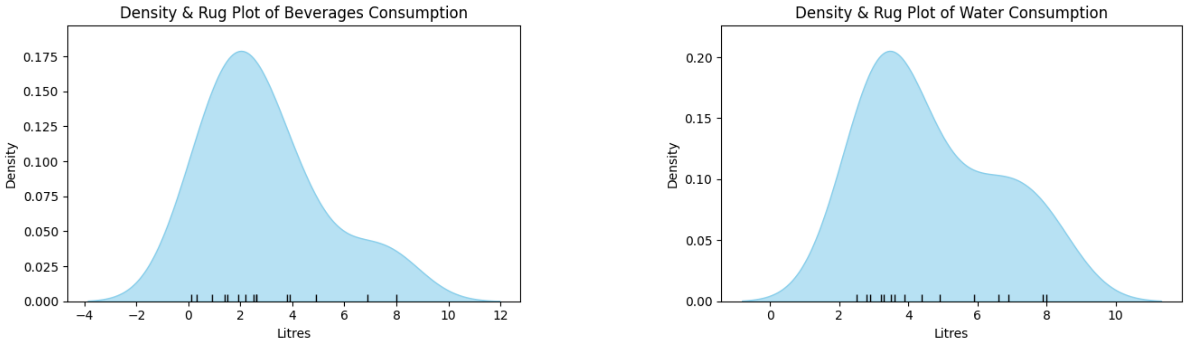
Observation: - Mean and median are close → approximately symmetric distribution. - Beverage distribution shows right skew (due to very low values like 0.1, 0.3).

Summary Statistics:

Statistic	Water (L)	Beverages (L)
Mean	4.69	2.90
Median	3.90	2.50
Mode	None	2.6
Skewness	0.62 (Right-skewed)	0.95 (Right-skewed)

Table 1: Descriptive statistics for water and beverage consumption

3.2 Output:



(a) Density and rug plot for beverages

(b) Density and rug plot for water

Figure 3: Density and rug plots for water and beverages

4 Question 4: Scatter Plot and Correlation

Problem Statement:
A car company wants to predict fuel consumption from car masses.

4.1 Solution:

Dataset:

$$\text{Fuel Used (L)} = \{3.6, 6.7, 9.8, 11.2, 14.7\}$$

$$\text{Mass (tons)} = \{0.45, 0.91, 1.36, 1.81, 2.27\}$$

Correlation: Pearson correlation coefficient = **0.9939** (strong positive linear correlation).

Correlation Analysis between Mass and Fuel Consumption:

- **Correlation Coefficient:** 0.9939
- **Direction:** Positive correlation
- **Relationship Type:** Linear relationship

4.2 Output:

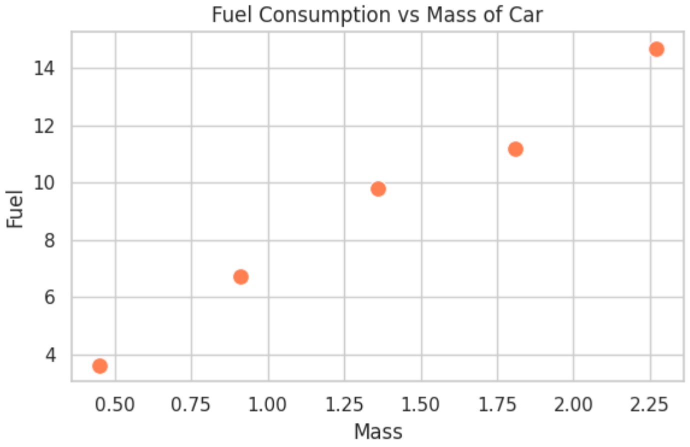


Figure 4: Scatter plot of Mass vs Fuel consumption

5 Question 5: Box Plot and Swarm Plot

Problem Statement

Number of chairs in each class:

$$35, 54, 60, 65, 66, 67, 69, 70, 72, 73, 75, 76, 54, 25, 15, 60, 65, 66, 67, 69, 70, 72, 130, 73, 75, 76$$

Create box plot and swarm plot (with jitter) and find the number of outliers.

5.1 Approach

- Create separate box plot and swarm plot (with jitter) to visualize the distribution and individual data points.
- Detect outliers using the Interquartile Range (IQR) method.

5.2 Computed Results

- Sorted data (for reference): 15, 25, 35, 54, 54, 60, 60, 65, 65, 66, 66, 67, 67, 69, 69, 70, 70, 72, 72, 73, 73, 75, 75, 76, 76, 130.
- Quartiles: $Q_1 = 61.25$, $Q_3 = 72.75$, $\text{IQR} = 11.50$.
- Lower Bound $= Q_1 - 1.5 \times \text{IQR} = 44.0$
- Upper Bound $= Q_3 + 1.5 \times \text{IQR} = 90.0$

Outlier Detection (IQR Method):

- **Q1:** 61.25

- **Q3:** 72.75
- **IQR:** 11.5
- **Lower Bound:** 44.0
- **Upper Bound:** 90.0
- **Outliers:** [35, 25, 15, 130]
- **Number of Outliers:** 4

5.3 Output

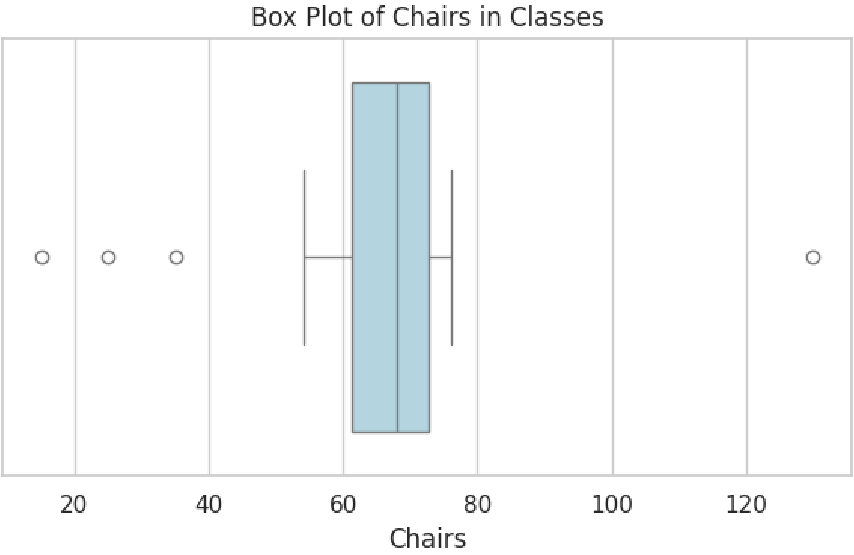


Figure 5: Box plot for number of chairs per class.

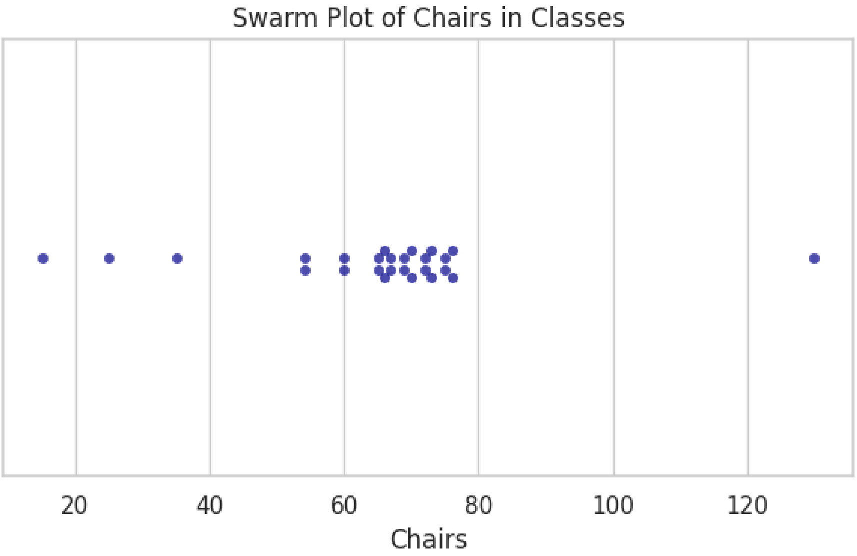


Figure 6: Swarm plot with jitter for number of chairs per class.

6 Question 6: Violin Plots for Random Distributions

6.1 Problem Statement

Generate random numbers f

- (i) Standard Normal distribution
- (ii) Log-Normal distribution

Visualize the data using violin plots.

6.2 Approach

- Use NumPy to sample random numbers:
  - Standard normal: `np.random.randn(N)`
  - Log-normal: `np.random.lognormal(mean, sigma, N)`
- Plot violin plots to show distributions and spread.

6.3 Output

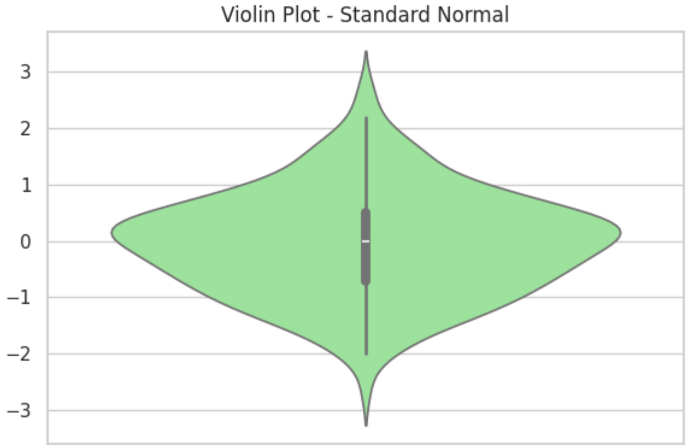


Figure 7: Violin plot for Standard Normal (example: N=200).

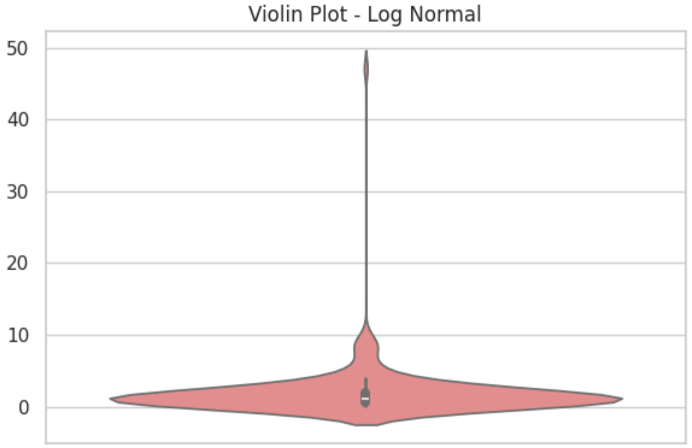


Figure 8: Violin plot for Log-Normal (example: N=200).

7 Question 7: Radar

Problem Statement

The agency wants number of ads per quarter for categories:

Category	Quarter 1	Quarter 2	Quarter 3	Quarter 4
Textile	10	6	8	13
Jewellery	5	5	2	4
Cleaning Essentials	15	20	16	15
Cosmetics	14	10	21	11

Visualize using radar/spider charts.

7.1 Approach

- Prepare categories as axes.

- Plot each quarter as a polygon overlay; fill with pastel transparency.

7.2 Output

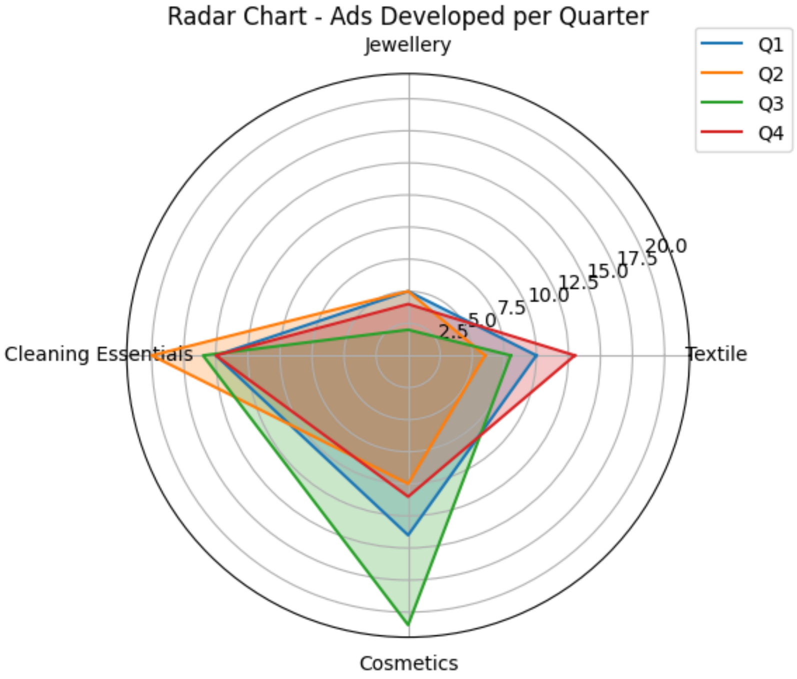


Figure 9: Radar chart showing ads developed per quarter (quarters overlaid).

8 Question 8: Funnel Chart (Time Spent in Product Development)

Problem Statement

Time spent (hours) on product development steps:

Step	Hours
Requirement Elicitation	50
Requirement Analysis	110
Software Development	250
Debugging & Testing	180
Others	70

Visualize using a funnel chart (or horizontal bar chart stacked to look like funnel).

8.1 Approach

- Plot descending bar heights or trapezoids to give a funnel feel.

- Compute total and percent time per step if needed:

$$\%\text{time} = \frac{\text{Hours for step}}{\text{Total hours}} \times 100$$

8.2 Computed Totals

- Total hours = 50 + 110 + 250 + 180 + 70 = **660**
- Example: Software Development = 250 hours = $\frac{250}{660} \times 100 \approx 37.88\%$

8.3 Output

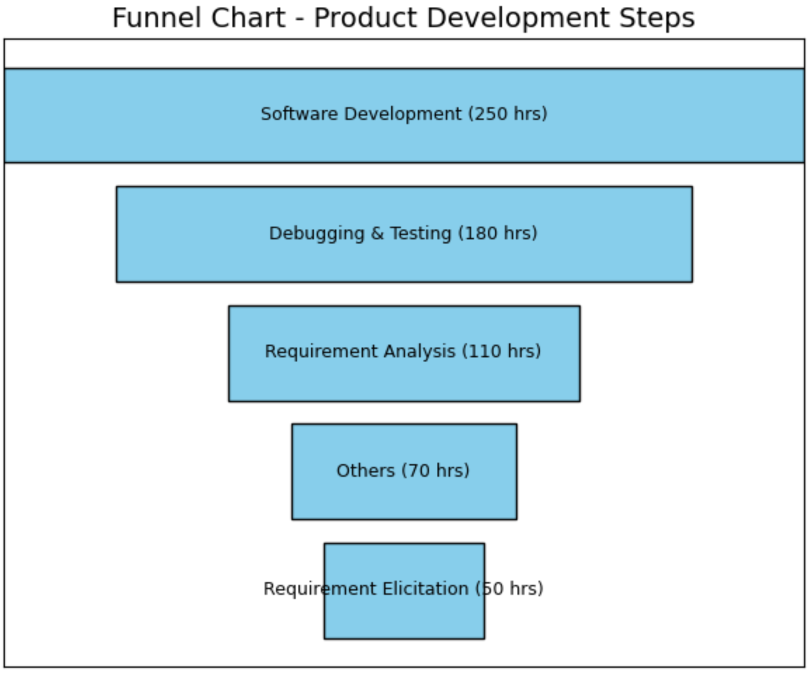


Figure 10: Funnel-style representation of time spent per step.

9 Question 9: Correlation (Ice-Cream Shop)

Problem Statement

Temperature vs Number of Customers:

Temperature	Customers
98	15
87	12
90	10
85	10
95	16
75	7

Find correlation and comment.

9.1 Approach

- Compute Pearson correlation coefficient as in Q4.
- Plot scatter

9.2 Result

- Pearson correlation ≈ **0.9118** — strong positive correlation: as temperature increases, customers increase.

9.3 Output

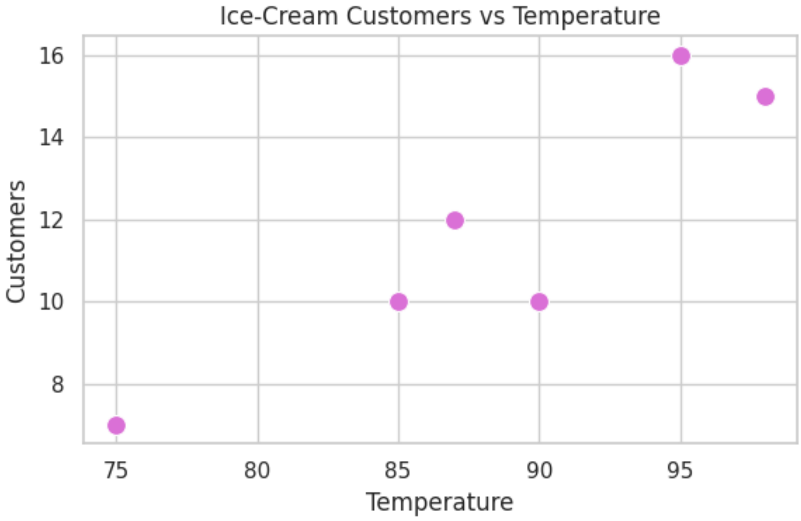


Figure 11: Scatter plot: Temperature vs Number of Customers.