Question 1

The whiskers in a boxplot are typically set at 1.5 times the interquartile range from the first and third quartiles. Any data points beyond this range are considered outliers. However, in datasets with non-normal distributions or extreme values, this rule may either miss significant outliers (if the data has heavy tails) or flag too many points as outliers (if the data has a high skew)

Question 2

A boxplot assumes a single unimodal distribution, so in a heavily skewed or multimodal dataset, it may incorrectly classify important data points as outliers. This is because extreme values in one mode might appear as outliers when they are actually part of a separate peak. Alternative methods include Violin plots (show distribution shape), Histogram or KDE plots (show multimodal behavior)

Question 3

The mean is sensitive to outliers and shifts towards the tail in skewed distributions

The median is resistant to extreme values and represents the central tendency better in non-symmetric data. A boxplot prioritizes the median because it is more robust and represents the dataset's center accurately in non-normal data.

Question 4

A right skewed boxplot suggests:

Positive skewness - skewness coef. > 0, Higher variance - due to extreme values on the right, mean > median - pulled by large values, potential violation of normality - affecting statistical tests like t-tests or linear regression

Question 5

Boxplots allow easy comparison of central tendency and spread across multiple groups, making them great for high-dimensional datasets.

However: . overlapping distributions make it hard to distinguish groups
. small sample sizes lead to misleading quartiles and outlier detection

Question 6

- Too few bins oversmooth the data and hide structure
- Too many bins make the histogram noisy

In Kernel Density Estimation, bin width choice translates bandwidth selection:

- Small bandwidth - captures too much noise
- Large bandwidth - oversmooths the distribution, hiding details.

Optimal bin width should be chosen based on the data

Question 7

- Histograms show frequency distributions of numerical data, where bin width affects the representation
- Bar charts represent categorical data, where binning is not needed

In histograms, choosing inappropriate bin widths can lead to misleading patterns, while in bar charts, the width of bars is arbitrary and does not affect interpretation

Question 8

A histogram can distort data when:

- Uneven bin widths create artificial patterns
- Too few bins smooth out important features
- Too many bins create excessive noise

For example in a bimodal dataset, poor bin choices may take it appear unimodal. Better alternatives: • Kernel Density Estimation
• Violin plots

Question 9

A density plot smooths the distribution using a Kernel function, unlike a histogram which groups data into bins: Challenges:

- Kernel choice affects smoothing
- Bandwidth selection determines how much the curve smooths:
  - : too large → oversmooths, losing details
  - , too small - retains too much noise

For sparse datasets; KDE can create misleading shapes, so proper bandwith selection is critical

Question 10

Density plot represents a probability density function. A fundamental property of any probability density function is that the total probability over all possible values must sum to 1. This ensures that the density function properly represents probabilities - no matter how the data is distributed, the sum of all probabilities within the dataset must equal to 1.

The area under a density plot always being 1 ensures that distributions with different sample sizes can be fairly compared, as it normalizes frequencies and prevents larger datasets from visually dominating the comparison.

Part 2

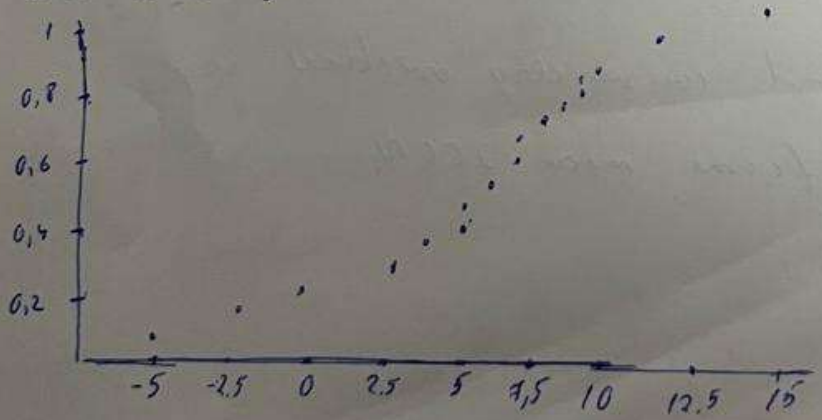
Question 1.

Sorted data: $-5, -2, 0, 3, 4, 5, 5, 6, 7, 7, 8, 9, 9, 10, 12, 15$

$ECDF(x) = \frac{\text{rank of } x}{\text{\# of points}}$ $\quad ECDF(-5) = \frac{1}{16} = 0,0625$ $\quad ECDF(0) = \frac{3}{16} = 0,1875$

$ECDF(-2) = \frac{2}{16} = 0,125$ $\quad ECDF(3) = \frac{4}{16} = 0,25$

and so on for all values we get the following graph



Question 2

Minimum : -5

Q_1 (25th percentile) - 15,5

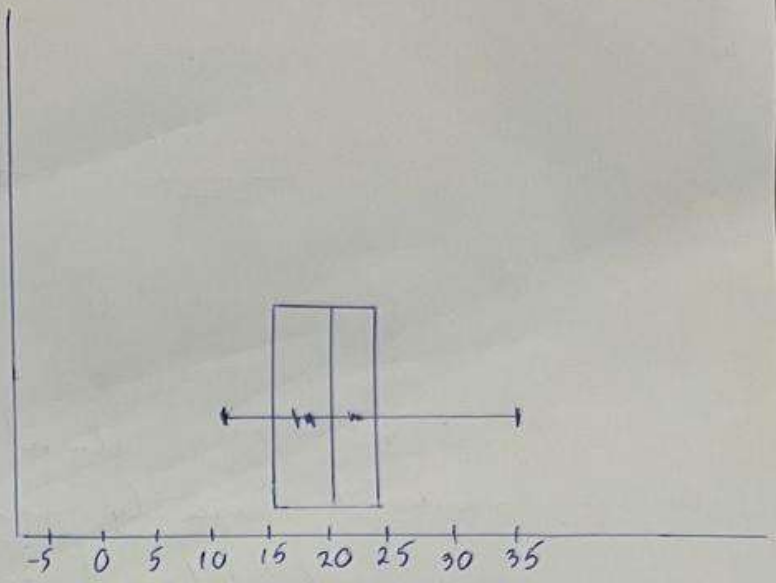
median -20

Q_3 : -24

$IQR = Q_3 - Q_1 = 8,5$

Lower limit = $Q_1 - 1,5 \cdot IQR = 2,75$

Upper limit = $Q_3 + 1,5 \cdot IQR = 36.75$



Question 3

Minimum : -10

Maximum : 105

range = 105 - (-10) = 115

5 bins : $\frac{115}{5} = 25$ (bin width)

bins : [-10, 13], [14, 37], [38, 61], [62, 85], [86, 109]

Bin 1: 1 value

Bin 2 : 0 values

Bin 3: 5 values

Bin 4: 10 values

Bin 5: 8 values