Question 1

In a standard boxplot, the whiskers extend to the smallest and largest data points within 1.5 times the IQR from the Q1 and Q3. Any points beyond this range are considered outliers.

It may not work well in skewed distributions, as extreme values may be classified as outliers even if they are part of a long-tail distribution.The same thing would happen with small datasets, the IQR rule may label too many points as outliers, even if they are legitimate observations.

Question 2

In highly skewed data, the IQR-based whiskers may be asymmetric, making it difficult to interpret whether extreme values are true outliers or just part of the skew.

Alternative methods are using Log transformations, which will help to normalize skewed data or density plots, which provide smoother visualization of distribution without predefined whiskers.

Question 3

Mean is the arithmetic average. Median is the middle value.

Boxplots prioritize the median, because It gives a better representation of the central tendency in non-symmetric distributions. It is also less affected by extreme values than the mean.

It can be misleading if the data has a long-tailed right-skew, the median may under-represent the actual data range.

Question 4

A right-skewed boxplot suggests that most data points are at lower values, with a few extreme high values.

Effects on statistical measures: Variance increases due to large extreme values. Skewness coefficient is positive, indicating asymmetry. Model assumptions may be violated if normality is required, affecting statistical tests and regression models.

Question 5

Boxplots are particularly useful for comparing multiple groups in high-dimensional Data because they allow side-by-side comparison of medians, spreads, and outliers across multiple categories.They are also effective for high-dimensional data because they summarize distributions in a compact way.

The limitations are that overlapping distributions may be difficult to distinguish. Boxplots do not show multimodality and with small sample sizes, the IQR and outlier thresholds may be unreliable.

Question 6

selecting an inappropriate number of bins in a histogram can lead to this effects:

Too few bins: Oversmooths data, hiding meaningful variations.
Too many bins: Introduces excessive noise, making patterns difficult to interpret.
KDE estimates a smooth probability density function. Choosing an inappropriate bandwidth leads to overfitting if we have chosen too small bandwidth and over smoothing if we have chosen too large bandwidth.

Question 7

A histogram is used for continuous data, where the x-axis represents intervals or bins, and the height of the bars indicates the frequency of values within each bin. A bar chart is used for categorical data, where each bar corresponds to a distinct category, and the height represents the frequency or proportion of that category.
In a histogram different bin widths can lead to different conclusions about the distribution. In a bar chart, the width of the bars is arbitrary and does not affect the meaning of the visualization since the categories are discrete and do not have a natural ordering. This makes bin choice irrelevant for bar charts but very important for histograms.

Question 8

Histograms can sometimes misrepresent data distributions if we choose the wrong bin width. If the bins are too wide, important details in the data may be lost. If the bins are too narrow, the histogram can become too noisy, making it difficult to identify meaningful patterns.
For example, If a histogram of house prices uses wide bins, it might group $200,000 and $500,000 houses together, making prices look more similar than they are. If the bins are too narrow, small price differences can create too many peaks, making the data look more scattered than it really is.
A density plot (KDE) or a violin plot can give a clearer picture without this problem.To solve the issue we can use kernel density estimation (KDE). KDE provides a smooth estimate of the probability density function, avoiding false binning effects.

Question 9

A density plot (KDE) is different from a histogram because it smooths the data instead of using fixed bins. A histogram shows counts within each bin, while a density plot estimates the overall shape of the data. The choice of bandwidth in a density plot is important—if it's too small, the plot will be too bumpy, and if it's too large, important details may disappear.

Question 10

The area under a density plot is always equal to 1 because it represents a probability density function (PDF), which describes the likelihood of different values occurring within a dataset. In probability theory, a PDF must integrate to 1 over its entire range to ensure that it accounts for all possible outcomes. This property allows density plots to be used for comparing distributions, even when datasets have different sample sizes, because the total area remains 1 for both distributions, they can be meaningfully compared without being biased by sample size differences.