**Part 1: Theoretical Questions**

1. **In a dataset with a non-normal distribution and potential extreme values, how are the whiskers in a boxplot determined, and what are the limitations of the standard IQR-based rule in such cases?**

Boxplot whiskers go up to 1.5 × IQR past the quartiles, but they stop at the last actual data point within that range. If the data has extreme values or is not normal, this rule might miss real extreme points or label normal values as outliers.

1. **Given a dataset with heavy skewness and multiple peaks, how can a boxplot misrepresent outliers, and what alternative methods exist for identifying them more accurately?**

If the data has skewness or multiple peaks, a boxplot might not show outliers correctly. It assumes a balanced spread, so it could mark too many or too few outliers. Instead, violin plots, KDE, or log transformations can show outliers better.

1. **Explain the conceptual difference between median and mean in the context of nonsymmetric distributions. Why does a boxplot prioritize the median, and in what cases could this choice obscure important data characteristics?**

The median is the middle value, while the mean is the average. Mean gets affected by extreme values, while median stays stable. A boxplot prefers median because it’s more reliable for skewed data, but this can hide shifts if the mean and median are far apart.

1. **If a boxplot exhibits strong right skewness, what can you infer about the underlying probability distribution? How would this skewness affect statistical measures such as variance, skewness coefficient, and potential model assumptions?**

If a boxplot shows strong right skewness, it means the data has many small values and a few very large ones. This makes variance bigger, the skewness number positive, and can break assumptions in models that expect symmetry.

1. **Why are boxplots particularly useful for comparing multiple groups in high-dimensional data? What are the limitations of boxplots when dealing with overlapping distributions or categorical variables with small sample sizes?**

Boxplots are great for comparing multiple groups because they quickly show medians, spread, and outliers in one view. This makes them useful in high-dimensional data where you need to compare distributions across different categories. However, boxplots hide details about the shape of the data. If two groups have a lot of overlap, the boxplots may look the same even if the distributions are actually different.

1. **What are the theoretical consequences of selecting an inappropriate number of bins in a histogram, particularly in datasets with varying density regions or multimodal distributions? How does bin width selection affect kernel density estimation (KDE)?**

Choosing the wrong number of bins in a histogram can make the data look misleading. If there are too few bins, the histogram might look too smooth, hiding important details like peaks or gaps in the distribution. On the other hand, if there are too many bins, the histogram can become too noisy, showing random fluctuations that aren’t meaningful. For example, a multimodal distribution may look unimodal when too wide and few bins are selected.

1. **Histograms and bar charts both use rectangular bars to display data. How does the interpretation of frequency differ in these two visualizations, and why is bin choice irrelevant in bar charts but crucial in histograms?**

Histograms group data into bins to see distribution, while bar charts compare categories. In histograms, bin size changes how patterns look, but in bar charts, bins don’t exist—just fixed categories.

1. **Under what conditions might a histogram distort the perception of a dataset’s distribution? Provide an example where binning choices lead to misleading conclusions, and explain how alternative visualizations (e.g., KDE or violin plots) could address these distortions.**

Histograms can mislead by making data look too smooth. Example: A dataset with two peaks may look like one big peak if bins are too wide. KDE or violin plots fix this by smoothing data and show the true shape of the data.

1. **How does a density plot differ from a histogram in terms of its mathematical foundation and interpretability? What challenges arise when choosing a kernel function and bandwidth for density estimation, particularly in sparse datasets?**

Histograms count data in bins, **grouping data into intervals** (bins) and **counting how many values** fall into each bin, while density plots estimate probabilities over a continuous range. Density plots provide a smoother view of the data but require choosing a kernel function and bandwidth. Too much smoothing hides details, while too little creates false patterns, especially in small datasets.

1. **Explain why the area under a density plot is always equal to 1. How does this property relate to probability theory, and what implications does it have for comparing distributions with different sample sizes?**

Density plot area is always 1 because it shows probabilities, and the sum of probabilities is 1. Because of this, density plots allow fair comparisons between datasets of different sizes, since they focus on relative distribution instead of raw numbers. However, too much or too little smoothing in KDE can distort the data.