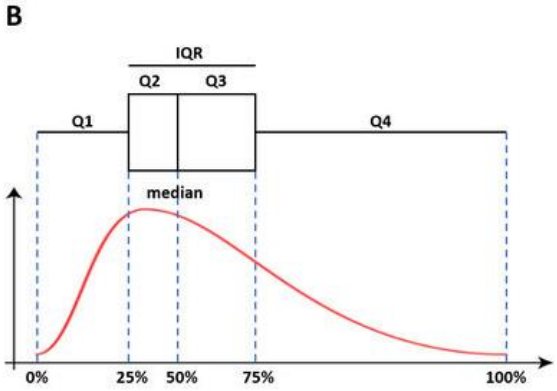
Homework N2

Part 1: Theoretical Questions

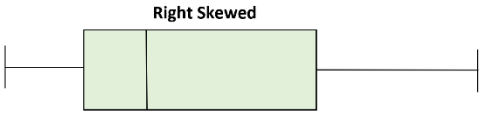
1. In a boxplot, whiskers are typically set at Q1−1.5×IQR (minimum) and Q3+1.5×IQR (maximum), where IQR is the interquartile range. However, this rule can does not identify values as outliers in skewed distributions or fail to account for extreme values.

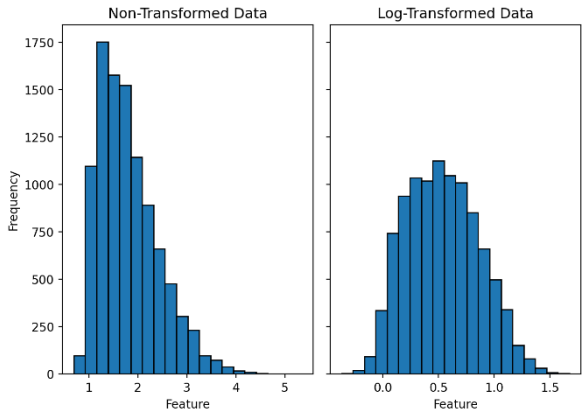
Example: In a dataset of incomes, most values are between 150,000 – 400,000 but some exceeding 500,000 the 1.5 × IQR rule might incorrectly label legitimate higher salaries as outliers due to the dataset's right skewness.(as shown in the graph)



2. In a heavily skewed dataset, a boxplot can misrepresent outliers by overestimating them in long ends. Alternative methods like the median absolute deviation (MAD) or strong statistical approaches tailored to skewness, such as log transformation can provide more accurate outlier detection.

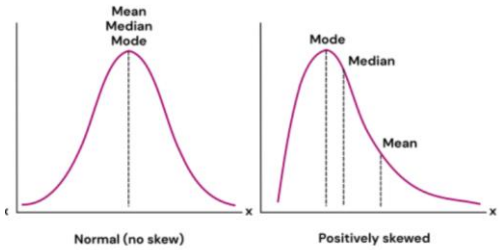
Example: In a dataset of house prices with a peak around $200,000 and another around $1,000,000, the 1.5 × IQR rule may wrongly flag many high-end homes as outliers, while a log transformation could better handle for the dataset's skewed distribution. (as shown in the graphs)





3. The mean is sensitive to extreme values, while the median is less affected by outliers. Boxplots use the median for this reason, but this can hide important patterns like groups of values at different levels.

   Example: In a dataset of house prices, with most homes at $200,000 and some over $1,000,000, the mean is pulled higher by the expensive homes, while the median and boxplot focus on the $200,000 range and might miss the pattern of luxury homes.



4. Right skewness means that the data has a long tail on the right, which means most values are at the lower end and a few of them have extreme higher values. This increases the variance and results in a positive skewness coefficient, which can affect models assuming normality.

   Example: In income data, right skewness shows that while most people earn average wages, a few very high earners increase the variance, making normality assumptions in models less accurate.

5. Boxplots are helpful for comparing multiple groups because they show clear visualization of the distribution (e.g. central tendency). However, it can be difficult if in the groups have a too much overlap or if there are only a few data points in a group, making it hard to see the differences.

   Example: In a study comparing test scores from different study methods, boxplots might not show clear differences if the scores overlap a lot, or if one method has only a few students.

6. Choosing the wrong number of bins in a histogram can either hide important patterns or make the data look too messy, especially in datasets with varying densities or multimodal distributions. For kernel density estimation (KDE), the bin width affects how smooth or detailed the estimated distribution is, with wider bins blending data and narrower ones potentially over fitting it.

   Example: In a dataset of exam scores, using too few bins might make low, average, and high scores appear similar, while using too many bins could make the data look overly spread out, missing the clear groups.

7. Histograms show distribution of continuous data, with bin choice affecting frequency interpretation, while bar charts display categories with fixed, non-variable width bars. In histograms, bins group data points, while bar charts simply count distinct categories.

Example: In a histogram of test scores, bins group ranges of scores, while in a bar chart of study methods, each bar represents a separate method, with no binning needed.

8. A histogram can distort the perception of a dataset's distribution if the bin size or number of bins is poorly chosen. Too few bins can hide important variations, while too many bins can create an overly jagged appearance.

Example: In a dataset of exam scores, using very wide bins might group together high and low scores, misleadingly suggesting a uniform distribution, while very narrow bins could exaggerate minor differences, making the distribution appear more spread out than it really is. Alternative visualizations like KDE or violin plots can provide smoother, more accurate representations by adjusting for these issues.

9. A density plot estimates the probability density function using continuous smoothing, while a histogram shows the distribution by dividing data into discrete bins. Density plots are more interpretable as they show a smooth curve, but the choice of kernel function and bandwidth can affect their shape, especially in sparse datasets.

Example: In a sparse dataset with few data points, a poor choice of bandwidth in a KDE can either smooth out important features or create too many bumps, obscuring the actual distribution.

10. The area under a density plot is always 1 because it represents the total probability of all outcomes, which, according to probability theory, must sum to 1. This ensures that the plot accurately reflects the likelihood of the data.

Example: When comparing two distributions with different sample sizes, the area under each density plot will still be 1, meaning the plots are normalized. This allows for fair comparisons of shapes, even if one distribution has more data points than the other.