Data Mining - EX2

Deadline: Friday, Mehr 27, 1403 - October 16, 2024

1. Which approach among the four for managing missing data is most likely to result in an underestimation of the variable's dispersion (e.g., standard deviation)? What advantages does this approach offer?

Consider the following data and answer exercises 2-4:

15	9	35	19	145	25	13	31	3	19	11	23

2. Do the following.

- **a.** Identify the outlier.
- **b.** Verify that this value is an outlier, using the Z-score method.
- **c.** Verify that this value is an outlier, using the IQR method.

3. Investigate how the outlier affects the mean and median by doing the following.

- **a.** Find the mean score and the median score, with and without the outlier.
- **b.** State which measure, the mean or the median, the presence of the outlier affects more,
and why.

4. Complete the following tasks:

- **a.** Divide the data into four bins with equal widths. (First method)
- **b.** Divide the data into three bins, each containing four records. (Second method)
- **c.** Explain why each of the binning methods above may not be the most effective solution.