MIS637 - Data Analytics and Machine Learning
Assignment 2

Komal Wavhal (20034443)

Problem Statement: Make up a data set consisting of eight scores on an exam in which one of the scores is an outlier.

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.
c.) Verify that the outlier is indeed an outlier, using the IQR method.

Solution

a.) Find the mean score and the median score, with and without the outlier.

Ans: Dataset: {12, 17, 20, 22, 25, 28, 30, 100}

Mean = {Sum of all Numbers in a dataset} / Total numbers

Median (for odd numbers) = Middle number of the dataset after arranging the numbers in ascending order

Median (for even numbers) = Average of the two middlemost numbers after arranging the numbers in ascending order.

Method	**With Outlier**	**Without Outlier**
1. MEAN	{12+17+20+22+25+28+30+100} / 8 = 34.5	{17+20+22+25+28+30} / 6 = 23.67
2. MEDIAN	{22+25} / 2 = 23.5	22.5

b.) State which measure, the mean or the median, the presence of the outlier affects more, and why.

Ans: From the above observation, the presence of the outlier affects the mean more than the median because the mean is the average of all the numbers in a dataset and if a single number is removed, it affects the average. Whereas, median only focuses on the middlemost number and wouldn't affect much if a number is removed. Hence, the presence of outlier affects the mean more after the outlier is removed.

c.) Verify that the outlier is indeed an outlier, using the IQR method.

Ans: Consider a dataset = {12, 17, 20, 22, 25, 28, 30, 100}

Total numbers in dataset are 8.

Q2 = (22 + 25) / 2 = 23.5

Two halves are (12, 17, 20, 22) and (25, 28, 30, 100).

Q1 = (17 + 20) / 2 = 18.5
Q3 = (28 + 30) / 2 = 29

To find outliers:
1.) IQR = Q3 - Q1 = 29 - 18.5 = 10.5
2.) 1.5 * IQR = 1.5 * 10.5 = 15.75
3.) Outliers will be any points below Q1 - 1.5 * IQR = 18.5 - 15.75 = 2.75 and above Q3 + 1.5 * IQR = 29 + 15.75 = 44.75

4.) Hence, the outlier in the dataset is 100, as it is greater than 44.75.