MIS 637-A

KNOWLEDGE DISCOVERY IN DATABASES

ASSIGNEMNT 2

DATA SET:

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

STUDENT	SCORE (/100)
S101	87
S102	83
S103	30
S104	80
S105	92
S106	90
S107	85
S108	93

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

Mean Score (with outlier): (Sum of all 8 scores)/8

(87+83+30+80+92+90+85+93)/8

640/8

⇨ **80**

Mean Score (without outier): (Sum of 7 scores)/7

(87+83+80+92+90+85+93)/7

610/7

⇨ **87.14**

Median Score (with outlier): 8 values
Sorting data set: 87,83,30,80,92,90,85,93
=>30,80,83,85,87,90,92,93
Median: Sum of Middle 2 Values
(85+87)/2
⇨ 172/2
⇨ **86**

Median Score (without outlier): 7 values

Sorted data set: 80,83,85,87,90,92,93

Median: ((7+1)/2)th element

⇨ 4th element

⇨ **87**

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

	With Outlier	**Without Outlier**
Mean	80	87.14
Median	86	87

As seen from the above table (generated from part a), we can see that the Mean is affected by 7.14 points once the median is removed, whereas the Median is affected by 1point.
The Mean is more easily affected by outliers as compared to the Median because the mean incorporates the actual numerical value of the outlier, whereas the Median doesn't. Hence the Median can be a better measure of Central Tendency.

c. Verify that the outlier is indeed an outlier, using the IQR method.
IQR: Interquartile Range= 75th Percentile-25th Percentile;
Sorting Data Set: 30,80,83,85,87,90,92,93
Median (85+87)/2

⇨ 86

25th percentile: (80+83)/2

⇨ 163/2

⇨ 81.5

75th percentile: (90+92)/2

⇨ 182/2

⇨ 91

Hence IQR: 75th Percentile-25th Percentile

⇨ 91-81.5

⇨ 9.5

Outlier Verification: Outlier< 25th Percentile -1.5*IQR, Q3+1.5*IQR<Outlier

⇨ Q1-1.5*IQR: 81.5-1.5*9.5

⇨ 67.25

⇨ Q3+1.58*IQR: 91+1.5*9.5

⇨ 105.25

The outliers are the scores below 67.25 and above 105.25. Hence the score 30 in the data set is a verified outlier.