Data Mining Sections

Nagwa El-Araby - Mohammed El-Barbeer - Aml Magdy

Information System Department

2019 - 2020

Outlier:

- What is an outlier?

An outlier is a data point in a dataset that is distinct from all other observation

A data point that lies outside the overall distribution of the dataset

- What is the reason for outlier to exist in data set ?

1-variability in the data

2-an experimental measurement error

- What are the impacts of having outlier in a dataset ?

1-It causes various problems during our statistical analysis

2-it may cause a significant impact on the mean and standard diviations

- Various ways of finding the outlier :

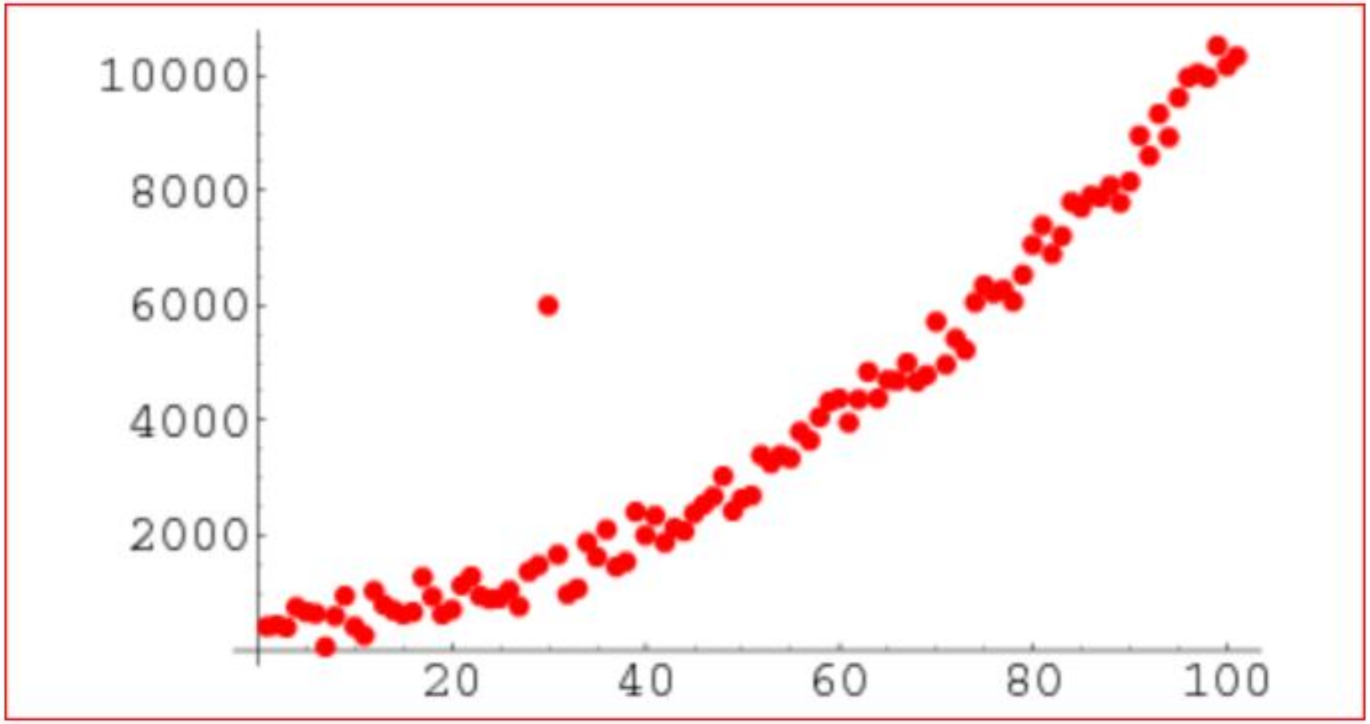
1-Scatter Plot

2-Z-score

3-Box plot

- Scatter Plot:

We can see the scatter plot and it shows us if a data point lies outside the overall distribution of the dataset

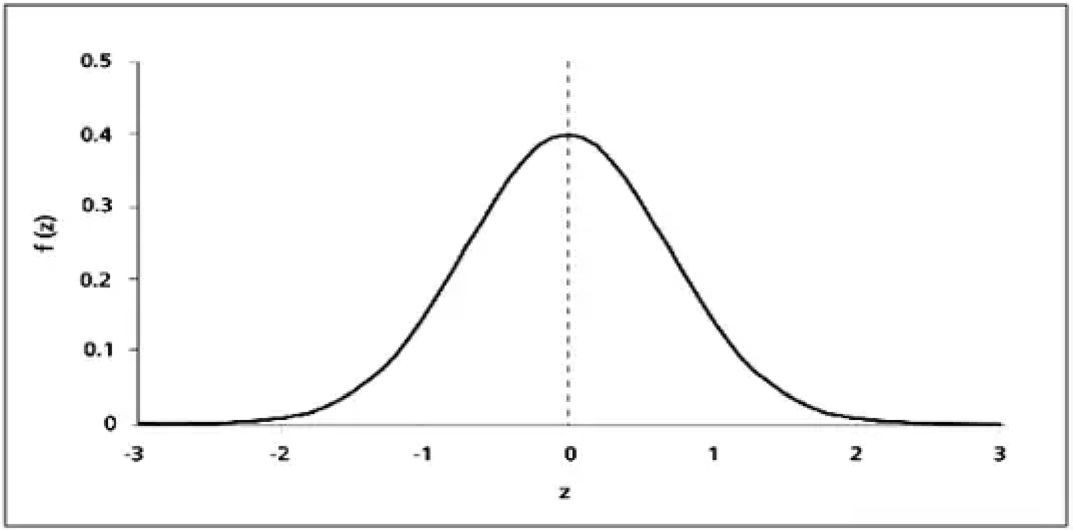


Scatter plot to identify an outlier

- Z-Score:

A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units

- The z-score is positive if the value lies above the mean, and negative if it lies below the mean.



- How do you calculate the z-score?

The formula for calculating a z-score is is $z = (x-\mu)/\sigma$, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

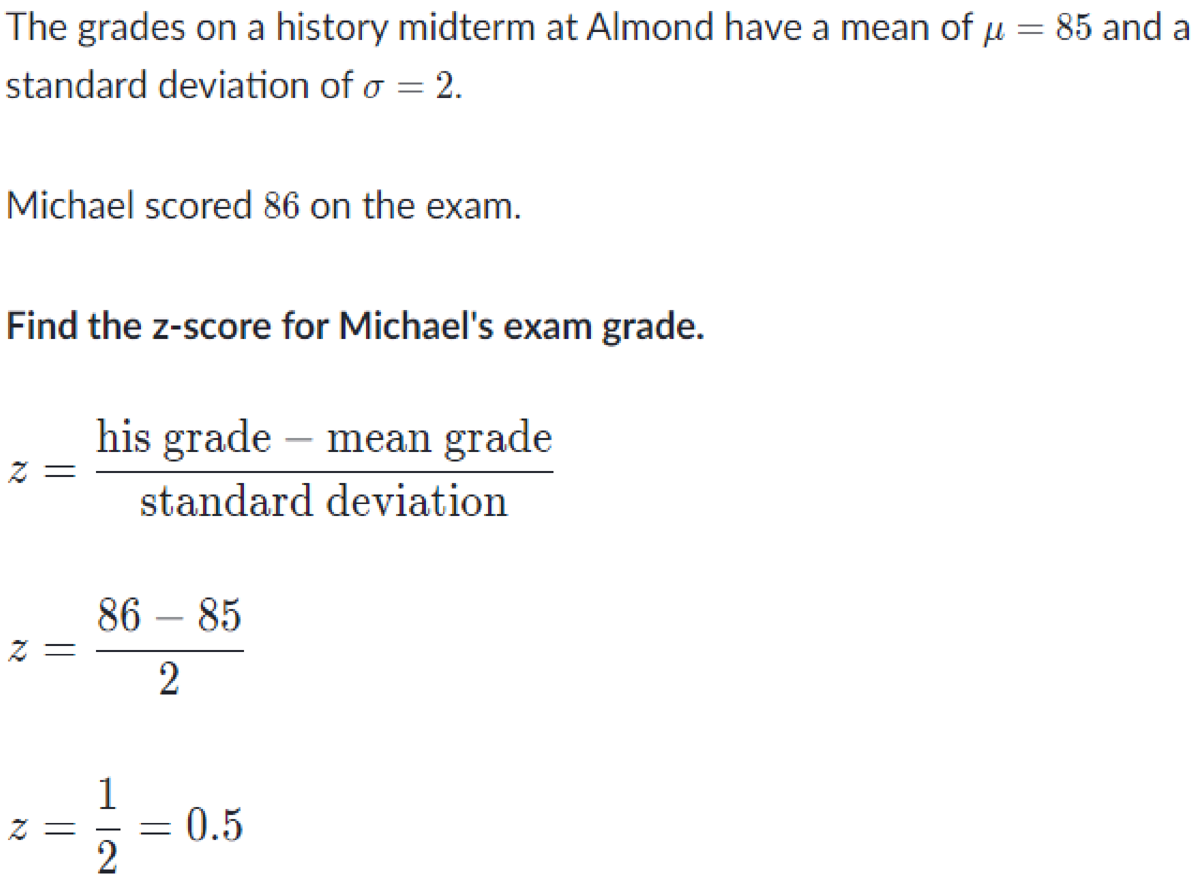
- How do you interpret a z-score?
- The value of the z-score tells you how many standard deviations you are away from the mean.

1-If a z-score is equal to 0, it is on the mean.

2-A positive z-score indicates the raw score is higher than the mean average. For example.

3-A negative z-score reveals the raw score is below the mean average.

Exercise:



The grades on a history midterm at Almond have a mean of $\mu = 85$ and a standard deviation of $\sigma = 2$.

Michael scored 86 on the exam.

Find the z-score for Michael's exam grade.

$$z = \frac{\text{his grade} - \text{mean grade}}{\text{standard deviation}}$$

$$z = \frac{86 - 85}{2}$$

$$z = \frac{1}{2} = 0.5$$

- Using IQR:

IQR tells how spread the middle values are. It can be used to tell when a value is too far from the middle.

Arrange the data in increasing order

Calculate first(q1) and third quartile(q3)

Find interquartile range (q3-q1)

Find lower bound q1*1.5

Find upper bound q3*1.5

Anything that lies outside of lower and upper bound is an outlier

Exercise:

Find the outliers, if any, for the following data set:

10.2, 14.1, 14.4. 14.4, 14.4, 14.5,
14.5, 14.6, 14.7, 14.7, 14.7, 14.9,
15.1, 15.9, 16.4

10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

To find out if there are any outliers, I first have to find the IQR. There are fifteen data points, so the median will be at the eighth position:

$$(15 + 1) \div 2 = 8$$

Then $Q_2 = 14.6$.

There are seven data points on either side of the median. The two halves are:

10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5

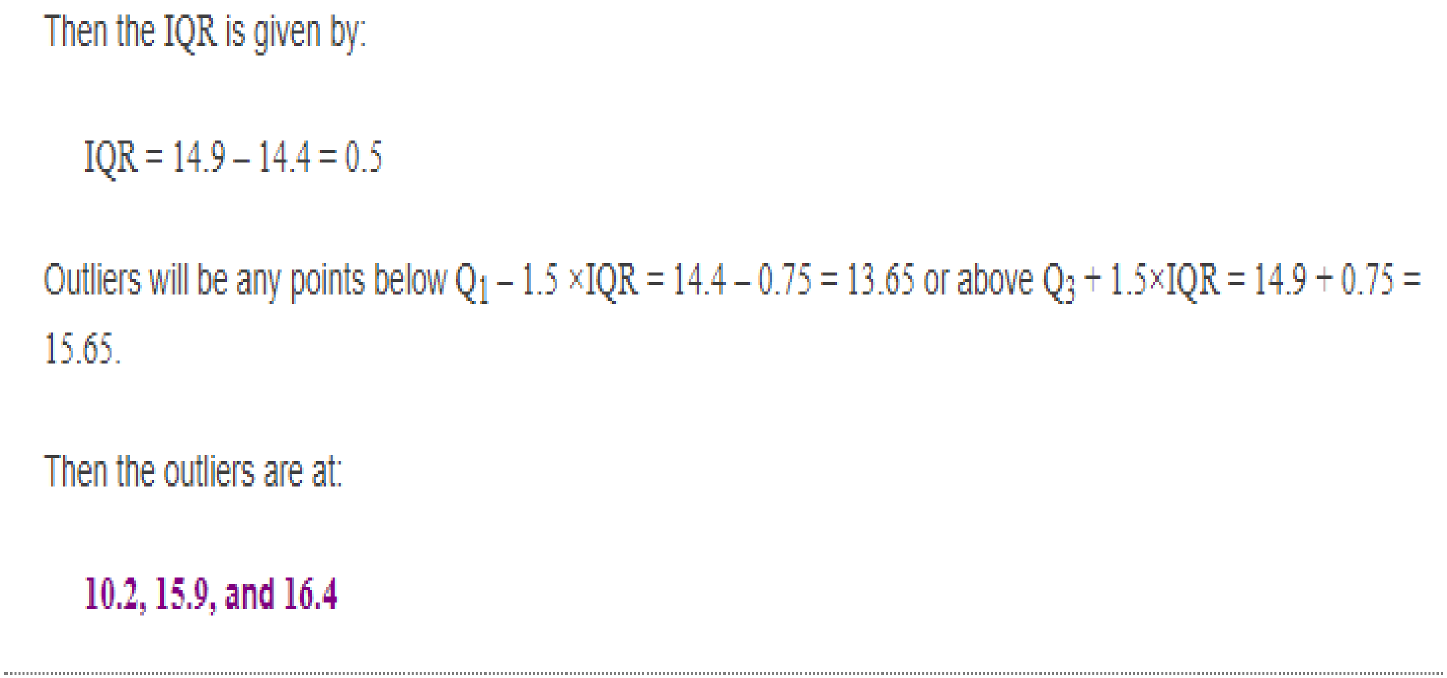
...and:

14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

Q_1 is the fourth value in the list, being the middle value of the first half of the list; and Q_3 is the twelfth value, being th middle value of the second half of the list:

$Q_1 = 14.4$

$Q_3 = 14.9$



Then the IQR is given by:

IQR = 14.9 – 14.4 = 0.5

Outliers will be any points below $Q_1 - 1.5 \times IQR = 14.4 - 0.75 = 13.65$ or above $Q_3 + 1.5 \times IQR = 14.9 + 0.75 = 15.65$.

Then the outliers are at:

10.2, 15.9, and 16.4

Thank You