Handling Outliers

You have learnt what missing values are and how to treat them. Now, let's move to the next concept of data cleaning, which is outliers.

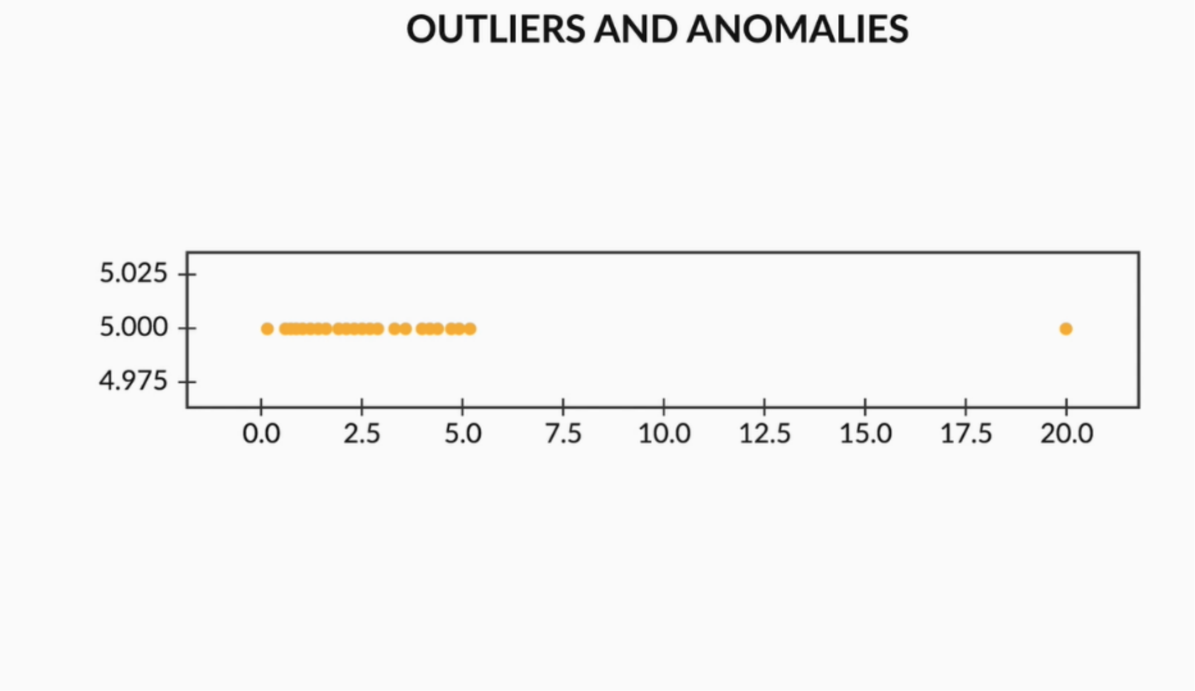
The definition of outliers is as follows:

Outliers are values that are much beyond or far from the next nearest data points.

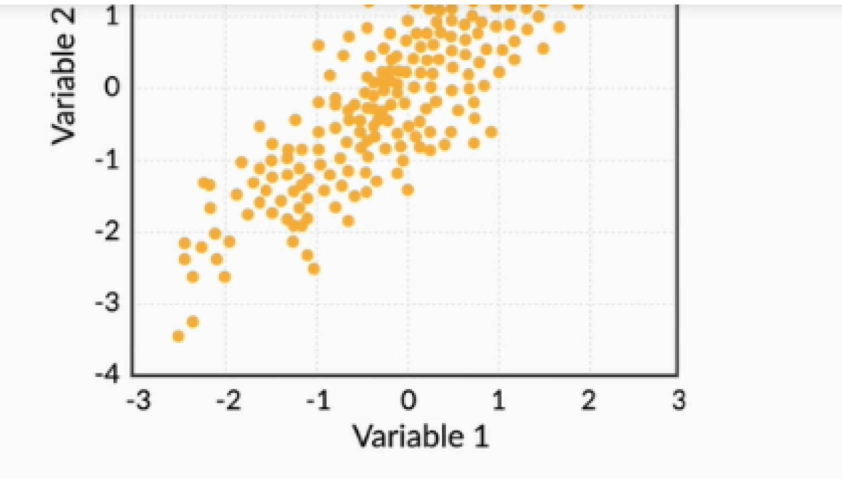
In this video, Rahim will help you understand the concept of outliers.

✓

lie beyond the range of expected values. You can get a better understanding of univariate outliers from the image below. Here, almost all the points lie between 0 and 5.0, and one point is extremely far away (at 20.0) from the normal norms of this data set.



- **Multivariate outliers:** While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value. These are called multivariate outliers. You can refer to the image below to get a better understanding of multivariate outliers.



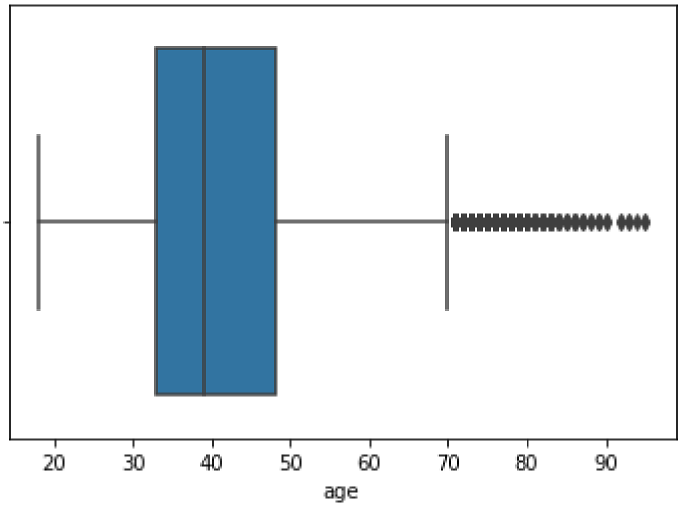
Now, let's proceed to the next video, where you will learn about the reasons behind the appearance of outliers in data and how to treat them.

Now, the major approaches to the treatment of outliers can include:

- Imputation
- Deletion of outliers
- Binning of values
- Capping the outliers

In the process of handling missing values and outliers of different columns, you are already performing univariate analysis. You will learn more about it in further sessions. In this video, you will learn how to implement all your learning on the bank marketing dataset.

Also, the 70-90 age group is sparsely populated and participate in opening the term deposit account, which is why these set of people fall out of the box plot but they are not outliers and can be considered as normal values.



Let's listen to Rahim as he explains the variable 'balance'.

An important aspect that has been covered in this video is **quantiles**. Sometimes, it is beneficial if you look into the quantiles instead of the box plot, mean or median. Quantile may give you a fair idea about the outliers. If there is a huge difference between the maximum value and the 95th or 99th quantiles, then there are outliers in the data set.

In the next segment, you will learn about the standardisation process in EDA.

< > **Question 1/4** Mandatory ✓

Outliers

Consider the following two statements:

1. The difference between the maximum value of the balance variable and the 99th percentile is too high.
2. The difference between the 99th percentile value and the 95th percentile value of the balance variable is in the normal range, meaning it is not too high.

Based on the above two statements, choose the correct option which concludes that the balance variable has outliers in it.

○ Statement 1 is alone sufficient to conclude that balance variables have outliers.

○ Both the statements are insufficient to conclude that the balance variable has outliers.

- Feedback:

  Both the statements are simultaneously required for you to infer from statement 1 that if the maximum value of the variable is far from the 99th percentile, it gives a clear idea that there are outliers in the data set. In addition, from the 2nd statement, you can see that there is no huge difference between the quantile of 95th and 99th.

○ Any of the statement is alone sufficient to conclude that the balance variable has outliers.

✓ Your answer is Correct. Attempt 1 of 2 Continue

Report an error

PREVIOUS

Impute/Remove Missing Values

NEXT

Standardising Values

An important aspect that has been covered in this video is **quantiles**. Sometimes, it is beneficial if you look into the quantiles instead of the box plot, mean or median. Quantile may give you a fair idea about the outliers. If there is a huge difference between the maximum value and the 95th or 99th quantiles, then there are outliers in the data set.

In the next segment, you will learn about the standardisation process in EDA.

< > **Question 2/4** Mandatory ✓

Outliers

Which of the following methods can be used to identify the outliers (univariate/multivariate) in the dataset?

○ Box plot can be used to plot the single variable and find its interquartile range and quantiles. ✓ Correct

Feedback:

Box plot gives a clear picture of all the points and visualises the quantiles to infer knowledge about the outliers.

○ The difference of each point from the mean/median value in the dataset is alone sufficient to identify whether a point is an outlier or not.

An important aspect that has been covered in this video is **quantiles**. Sometimes, it is beneficial if you look into the quantiles instead of the box plot, mean or median. Quantile may give you a fair idea about the outliers. If there is a huge difference between the maximum value and the 95th or 99th quantiles, then there are outliers in the data set.

In the next segment, you will learn about the standardisation process in EDA.

< > **Question 3/4** Mandatory ✓

Outliers

What is the mean and 75th percentile of the salary variable in bank marketing data set, respectively?

○ 57004, 60000

○ 57004, 70000 ✓ Correct

Feedback:

Just write the following code to describe the salary variable. You will find the mean and 75th

percentile.

```
inp1.salary.describe()
```