Cleaning

1, Data Quality Research: Good quality data has three attributes: uniqueness, continuity, and consistency. We can test the bias of the data from these three levels.

1. Verify that the data is duplicated.
2. There are no missing values between the lowest and highest values of the check attribute, and some values need to be unique. In general, I will pay special attention to these points when testing continuity:

Use of NA value

Use of Null values

Use of null values

Use of special symbols (such as ?, %)

Make use of null values (such as "Unknown", "empty", "unknown")

1. Check whether the data write format conforms to the unified specification, and whether there will be behavior such as field overload. In the field, there are mainly three types:

Character type (eg CHAR, VARCHAR)

Numerical type (eg INTEGER, DECIMAL)

Date type (eg DATE, TIMESTAMP)

2, If we find that there is a missing value, we will automatically add the value to those missing values or delete it directly.

3, Smoothing-out noisy data:

Binning:

Smooth ordered data by values ​​around the data (near values). First, sort the data, and then divide the data into equal-frequency boxes. Replace the values ​​in the box with the same indicator values. Therefore, in the binning operation, the larger the width of the box, the more obvious the smoothing effect. The box can be of equal width or can be defined by itself.

Here, through the different indicators, the bin can be divided into:

The box average is smooth: each value in the box is replaced with the average of the data in each box.

The median box is smooth: each value in the box is replaced with the median value of the data in each bin.

The box boundary is smooth: the maximum and minimum values ​​in the defined box are the boundary values, and each value in the box is replaced with the nearest one.

Regression:

Fit a noisy variable to a straight line (linear regression) or a curve (curve regression). The general approach here is to use a linear regression or multiple linear regression to fit a line or multidimensional plane. In turn, the noise data can be replaced by smooth lines or faces.

Outlier Analysis:

Outliers are detected by clustering methods. The values ​​that fall outside the cluster are outliers that can be replaced or directly removed by the cluster's centralized trend indicator.

Data cleaning tool

Data Scrubbing Tool: Potter's Wheel

Data Auditing Tool: ACL

Data Migration Tool: ETL