Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 6: OUTLIER DETECTION

Lesson 6: Outliers

Outliers are points that are far from others (typically in an abnormal way). In general, they can create analytic challenges by distorting individual measures or relationships and potentially leading to mistaken conclusions. Thus, it is important to understand the presence of outliers in data sets. Some of the most common causes of outliers in a data set include the following.

- data entry errors (human errors)
- measurement errors (instrument errors)
- experimental errors (data extraction or experiment planning/executing errors)
- intentional errors (dummy outliers made to test detection methods)
- data processing errors (data manipulation or data set unintended mutations)
- sampling errors (extracting or mixing data from wrong or various sources)
- natural (not an error, novelties in data)

The most common remedy for outliers involves either fixing errors or removing the outlier. However, there are times when one would leave an outlier in the data.

Help

DLA ▾

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 6: OUTLIER DETECTION

As you look at this, obviously you do not believe one of our customers is 812 years old! Most likely this was typed in incorrectly and possibly this customer is 81 or 82 years old. You should want to investigate this and fix the value. If you cannot find the necessary information to fix this, you might consider deleting this entry so that you do not introduce problems. For example, what if this customer is 23 years old? You should not settle for a value like 81 just because you cannot find information about this customer.

However, not all outliers are errors. There may be real data points that result from a skewed data set. If after investigating, you discover that is a correct data point, then it probably should be retained in the data set. Suppose that the customer ages were the following:

34 45 44 32 46 39 81

In this data, most of the ages are between 30 and 50. The value of 81 seems abnormal compared to the majority of ages, so you could consider this an outlier. Suppose after investigating this value, you determine that this is in fact an 81-year-old customer. Since this is not an error, you should be hesitant to remove this value from the data unless there is a strong justification for doing so.

Since outliers can be problematic for many statistical analyses, you should identify potential outliers before performing the analyses. It is easy to identify it when the observations are just a bunch of numbers and it is one-dimensional. However, when you have thousands of observations that are multidimensional, you will need more clever ways to detect those values.

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 6: OUTLIER DETECTION

ADA Accommodation

Privacy Policy | Terms of Service

Honor Code

Help

DLA ▾