Introduction to Data Engineering: Handling data issues

Ivan Marin
Daitan Group
imarin@daitangroup.com

Data issues

Data issues are data instances that are different from what is expected.

These issues could be intrinsic to the phenomena being studied or extrinsic.

Most common data issues

The most common data issues are:

- Missing values
- Outliers or Anomalies
- Duplicate data
- “Dirty” data

Missing Values

Values that are missing is a quite common issue.

Beware of the representation of what constitutes a missing value!

Null, NULL, None, ‘’, NA, nan

can all be representing a missing value!

Anomalies or Outliers

Anomalies and outliers are general terms for data that was not expected in the dataset and that *is interesting from the analyst perspective.*

How to detect anomalies and treat them is a large subject.

Duplicate data

Duplicate data is usually inserted on a failure in the ingestion or transformation process. Common cases:

- duplicate columns
- duplicate fields
- duplicate entries

“Dirty” data

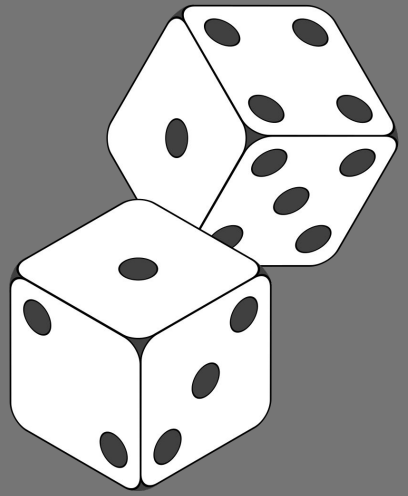
Dirty data is usually inserted because of a failure in the ingestion or transformation process. Common cases:

- empty columns
- encoding problems
-

No data set is perfect

It's important to understand the *cause of the error*. The correction to be applied on the dataset depends on it.

Causes for Data issues



Causes for Data Issues

Sources for data issues:

- the phenomena or process itself
- collection process or equipment
- transmission or storage
- transformation process

Causes for Data Issues

Data issues can be caused by the phenomena itself.

The *phenomena itself* may have errors or oscillations that cause missing data, etc.

These issues may not be considered errors.

Causes for Data Issues

Oklahoma Climatological Survey, Tipton, OK Mesonet station

The most common source of error is the collection process or equipment.

- Writing errors from human operators
- Overloaded sensor
- Cabling or operation errors
- OSMAR errors
- “External Forces”

Causes for Data Issues

Arizona, USA

Transmission or storage errors are also frequent

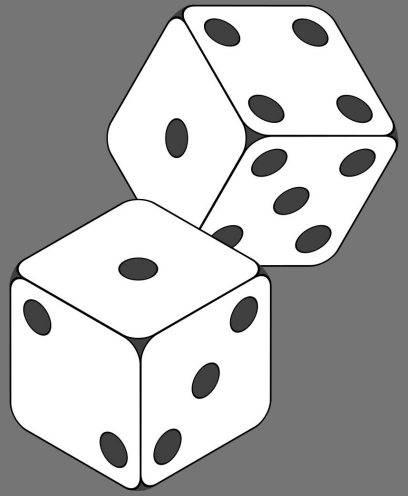
- Connection error (several possible reasons)
- Wrong format
- Difference between source and destination encoding
- Failure of part of the storage

Causes for Data Issues

Transformation errors usually happen on the data pipeline

- Data changed, transformation rules are the same
- Data is the same, transformation rules changed

Managing Data issues



How to handle Data Issues

The following techniques can be applied to handle data issues:

- Partial Deletion
- Imputation
- Interpolation

The application of each technique will depend on the nature of the data

Partial Deletion

Partial deletion is the removal of data instances to correct the data issue. The applicability depends on the frequency of the data issue.

- Missing values small compared to valid data: removal can reduce noise or bias
- Missing values large compared to valid data: removal can improve model performance

Partial Deletion

- Removal can be done by entire row (listwise deletion): can reduce sample size or introduce bias
- Removal can also be done on entire column

Imputation

Imputation is the replacement of the data issue with other value.

- Can introduce bias
- Depends on data type
  - Categorical values can be replaced with the mode, but it can skew the dataset
  - Same for ordinal values, with the addition of the median
  - Real values can use statistics like mean, min and max
- Highly dependent on data distribution

Interpolation

Interpolation consists in constructing new data instances based on existing data.

- Only for ratio values
- Several methods: polynomial, splines
- Linear regression can also be used
- Can lower the variance and overestimate the correlation

Anomalies and Outliers

The first step to handling an anomaly or outlier is *detecting* a data point as an anomaly or outlier.

- There is no general rule to say if a data point is an anomaly or not
- Graphs can help
- Understanding of the business and domain can definitely help

Anomalies and Outliers

After determining that a data point is an anomaly, the approach is similar to data issues:

- Retaining the outlier as valid data
- Deletion
- Imputation

Anomalies and Outliers

Even if the data point is an outlier, if it was generated by the phenomena being studied, it should be kept!

- Even if the data is very distinct from the rest of the data
- There are methods that can handle large outliers
- Understanding the cause can help handling it

Anomalies and Outliers

The detection and handling outliers is a large area of research. For now, it's important to think about how the changes caused by keeping, changing our removing an outlier will impact the data.

Data Duplication

Duplicate records are records that exist more than once in a dataset. Duplication can be:

- Identical duplicates
- Partial duplicates

Identical duplicates are usually caused by the ingestion process.

Partial duplicates are records have the same information in some fields but diverge in others.

Data Duplication

Data duplication can be handled by:

- Deterministic methods
- Probabilistic methods
- Machine Learning methods

Data Duplication

Deterministic methods identify duplicates using established rules that guarantee uniqueness.

- Unique identifier comparison (single or dependent)
- Bit level comparison

Text fields are hard to be deduplicated using this method because of possible variations

Data Duplication

Probabilistic methods use weights and measures by identifier to evaluate the probability of a record being a duplicate.

- If the probability is above a threshold, a record is considered duplicate
- Effective for partial duplicates
- Text fields can be handled using distance metrics (Jaro-Winkler e.g.)
- Normalization is relevant in this method

Data Duplication

Machine Learning can be applied on data deduplication:

- Training a supervised classifier on identified duplicated records
- Also effective for partial duplication

The fix can make the problem worse

Depending on the source of error in the data and the data characteristics, the correction can make the data issue worse

- Introduction of bias and reduction of variance
- Change of data distribution
- Masking real effects and phenomena

Example time

- https://colab.research.google.com/drive/1ytd4-0UJRA3J7qh23O7QKLRQ4kTZYD9x
-

How to handle Missing Values

There are three types of missing values:

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

How to handle Missing Values

Missing Completely At Random is defined as the chance that a data instance is missing is not related to the value of the variable is supposed to have.

- External causes independent of the phenomena
- Lab worker dropping a test tube

How to handle Missing Values

Missing At Random is defined as dependent on the phenomena being studied, but not related to the expected variable value.

- Hard to detect, only if you include the cause it can be detected

How to handle Missing Values

Missing Not At Random is directly related to the phenomena being studied and the expected variable value.

- Why this data instance is missing cannot be ignored