softtech®
group

# DATA CLEANING AND PROCESSING

Data Sheet

# Data Cleaning and Processing

## Data Cleaning

Data cleaning, also called data cleansing, is the process of preparing data for the analysis, ensuring your data is correct, consistent, and usable. Incorrect, incomplete, irrelevant, duplicated, or improperly formatted data being removed or modified before the analysis. This data is usually not necessary or helpful when it comes to analyzing data because it may hinder the process or provide inaccurate results. The process is mainly used in databases where improper or irrelevant part of the data is identified and then altered or deleted.

### Better data beats fancier algorithms

Garbage in gets you garbage out. If you have a properly cleaned data set, even simple algorithms can learn impressive insights from the data!

## Fundamentals of Data Cleaning

### Eliminate Unsolicited Observations

First step to data cleaning is removing unwanted observations from the dataset.

- Duplicate Observations
- Irrelevant Observations

### Filter Unwanted Anomalies

Outliers/Anomalies can cause issues with certain types of models. However, there must be a good reason to remove an out lieras some can be informative. Outlier removal can help your

### Fix Structural Errors

Structural errors are those that arise during measurement, data transfer, or other types of "poor housekeeping". For instance, typos or inconsistent capitalization, etc.

### Handle Missing Data

Missing data is a tricky issue and can't be ignore in a data set. Missing categorical data Missing numeric data It must be handled in a practical way as most algorithms do not accept missing values. Dropping or imputing missing values can be sub-optimal and leads to loss in information.

### Some of the Best Data Cleaning Tools of 2020

SISENSE

ANSWER ROCKET

TIBCO

DOMO

Dundas

trenData

oxcyon

Phocas
Got data. Get results.

LUMENORE
A NETLINK PLATFORM

CXO Software

## Data Processing

Data processing is the conversion of raw data to useable information. Data is worked on to generate results for an improvement to an existing solution or resolving a problem. It follows a cycle of inputs (raw data) being fed to a process to produce outputs (information & insights). This processing gives the data a form and context essential to be read by systems and utilized by organization.

## Stages of Data Processing

### Data Collection

The first step and very crucial is the collection of data. Data is gathered from available sources, including data lakes and data warehouses. This stage provides the baseline from which to measure and a target on what to improve.

### Data Preparation

Also referred to as "Pre-processing" stage, is where the data is cleaned up and organized and is checked for accuracy and errors making it suitable for further analysis, processing, and exploration.

### Data Input

It is the stage where cleaned and verified data is entered into its destination and is coded or converted into usable information for the systems to read. It follows a formal and strict syntax to breakdown the complex data.

### Data Processing

The data inputted in the previous stage is subjected to different means and methods for interpretation and to generate outputs. It is done through machine learning algorithms, depending on the data type and the data source.

### Data Output

Here, the data is transmitted and displayed to the user and becomes finally usable to non-data scientists. It is presented to user in the form of graphical reports, audio, video, images, plain text, etc.

### Data Storage

The final stage of data processing cycle where data and metadata are then stored for future use. Some of this information may be put to use immediately. Properly stored data is easily retrievable to be used

## Some of the Best Data Processing Tools of 2020

tableau public

rapidminer

Power BI

OpenRefine

Open for Innovation KNIME

HubSpot

Soft Tech Group can help your organization overcome the data challenges you are facing. The Data Engineers at Soft Tech Group will clean and transform your data so that it “plays nicely with data from other sources”. Our team makes sure that your data is used in the most productive and meaningful manner that can increase the fundamental value of your organization.