Data Cleaning

Freitag, 11. Juli 2025 10:29

Dirty Data
L> incomplete, incorrect or irrelevant data
L> to the problem we're trying to solve

Clean Data
L> is complete, correct and relevant
L> to the problem we're trying to solve

Data Engineers
L> transform data into a useful format for analysis
L> and give it a reliable infrastructure

Data Warehousing Specialists
L> develop processes and procedures
L> to effectively store and organise data

Null
L> indication that a value does not exist in a dataset
L> it means that the customer skipped that question
L> it is not a 0 - a zero can be a zero as a response

Field
L> a single piece of information from a row or a column of a spreadsheet

Field Length
L> a tool for determining how many characters can be keyed into a field

Data Validation
L> a tool for checking the accuracy and quality of data
L> before adding or importing it
L> a form of data cleansing

Duplicate Data

Description	Possible causes	Potential harm to businesses
Any data record that shows up more than once	Manual data entry, batch data imports, or data migration	Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval

Outdated Data

Description	Possible causes	Potential harm to businesses
Any data that is old which should be replaced with newer and more accurate information	People changing roles or companies, or software and systems becoming obsolete	Inaccurate insights, decision-making, and analytics

Incomplete Data

Description	Possible causes	Potential harm to businesses
Any data that is missing important fields	Improper data collection or incorrect data entry	Decreased productivity, inaccurate insights, or inability to complete essential services

Datei auswählen Keine ausgewählt

Incorrect/Inaccurate Data

Description	Possible causes	Potential harm to businesses
Any data that is complete but inaccurate	Human error inserted during data input, fake information, or mock data	Inaccurate insights or decision-making based on bad information resulting in revenue loss

Inconsistent Data

Description	Possible causes	Potential harm to businesses
Any data that uses different formats to represent the same thing	Data stored incorrectly or errors inserted during data transfer	Contradictory data points leading to confusion or inability to classify or segment customers

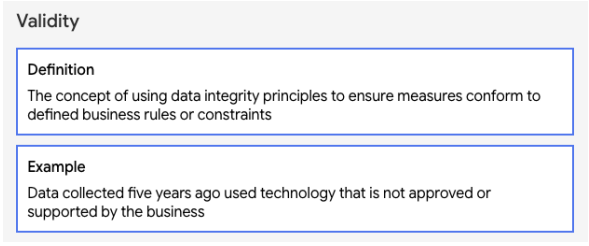
Dirty Data has huge impact on companies across various fields!

Data Integrity [Rep]

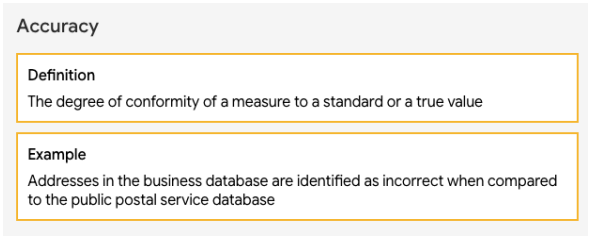
Freitag, 11. Juli 2025 10:29

Data Integrity
L> ensuring that data is correct
L> Repetition - 4 parts

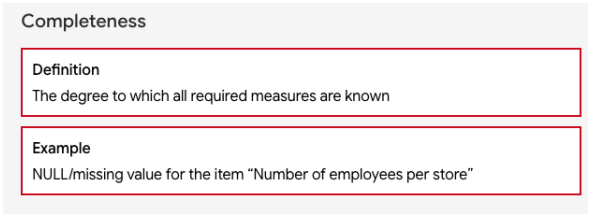
Validity
L> measures conform to defined business rules or constraints



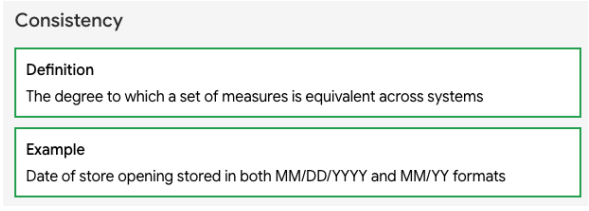
Accuracy
L> compared to reliable outside sources



Completeness
L> ensuring that all important data is included



Consistency
L> measures are formatted or structured
L> the same way across systems



Datei auswählen Keine ausgewählt

Data Merging / Errors

Freitag, 11. Juli 2025 10:29

Merger
L> an agreement that unites two organisations into a single new one
L> in case you have data from two sources

Data Merging
L> process of combining two or more datasets into a single dataset

Compatibility
L> how well two or more datasets are able to work together

How To Access Compatibility
L> Do I have all the data I need?
L> Does the data I need exist within these datasets?
L> Do the data need to be cleaned, or are they ready for me to use?
L> Are the datasets cleaned to the same standard?

Errors While Cleaning Data

Common mistakes to avoid
L> Not checking for spelling errors
L> Forgetting to document errors *(can be a big time safer)*
L> Not checking for misfielded values
L> Overlooking missing values
L> Only looking at a subset of the data *(every dataset equal attention)*
L> Losing track of business objectives *(new discoveries while cleaning)*
L> Not fixing the source of the error *(fixing same error again and again)*
L> Not analysing the system prior to data cleaning
  L> learn the source of dirty data
L> Not backing up your data prior to data cleaning *(create backups)*
L> Not accounting for data cleaning in your deadlines/process
  L> invest enough time into data cleaning! *(but effectively)*

10 Google Workspace Tips to clean up Data
https://support.google.com/a/users/answer/9604139?hl=en#zippy=

Datei auswählen Keine ausgewählt

Cleaning Spreadsheets 1

Freitag, 11. Juli 2025 10:29

Cleaning Data With Blank Fields
L> apply a filter on every column
L> where it only shows empty cells
L> go column after column
L> delete all empty cells

How To Transpose Data
L> from long format to wide format
L> copy the whole table
L> right-click and *"Paste-Special"*
L> and then choose *"Transpose"*

Get Rid of Extra Spaces in cells
L> select the whole sheet
L> select *"Data"*
L> select *"Data clean-up"*
L> select *"Trim Whitespace"*
L> optional: =TRIM(A1)
L> and then drag auto-fill square

Add-On: Change lower/uppercase/Proper case text
L> process string data
L> Add-On is the easiest way to clean up string data
L> All Uppercase *(makes all letters BIG)*

Delete All Formatting
L> highlight all Rows 1 - 8
L> select the tab *"Format"*
L> select *"Clear Formatting"*

Datei auswählen Keine ausgewählt

Cleaning Spreadsheets 2

Freitag, 11. Juli 2025 10:29

Conditional Formatting
L> changes how cells appear when values meet specific conditions

Remove Duplicates
L> tool that automatically searches for duplicates
L> and deletes duplicate entries from the spreadsheet

Text String
L> a group of characters within a cell
L> mostly composed of letters

Split
L> tool that divides text around a specific character
L> puts each fragment into a new, separate cell
L> e.g. commas recognised as delimiters
L> delimiter = specified text separator

Concatenate
L> function joining multiple text strings into a single string

Auto Data Format Setting
L° mark the column
L> select *"Format"*
L> select *"Number"*
L> select "Date"

Splitting Data into Columns
L> when column has commas or similar
L> you can automatically split into columns
L> mark column
L> select the tab *"Data"*
L> select *"split into columns"*
L> there is a **Function!**
L> =SPLIT(Cell, "Split-Sign")
L> Split-Sign is what determines the place to split

Fixing Errors?! Using "Splitting Data into Columns"
L> there was a #VALUE! Error
L> a number in a field hat this "717" format
L> and the splitting function just removed the " "
L> and solved the error

Datei auswählen Keine ausgewählt

Cleaning Spreadsheets 3

Freitag, 11. Juli 2025 10:29

Function
L> set of instructions that performs a specific calculation
L> using the data in a spreadsheet

Syntax
L> predetermined structure
L> including all required information
L> and it's proper placement
L> what I always write about function before e.g.

The COUNTIF Function
L> write into an empty cell
L> =COUNTIF (Range:Range; "Value")
L> e.g. =COUNTIF(I2:I72, "<100")
L> count how many cells have value under 100
L> can be used to clean data
L> can be used to check data integrity

The LEN Function
L> if certain piece of information has certain length
L> =LEN(Cell) , e.g. =LEN(A2)
L> tells me how many numbers/letters cell is made of
L> spreading it all over a column can help to find mistakes

The LEFT/RIGHT/MID Function
L> can return a segment of a cell
L> starting from the left, right or from the middle
L> =LEFT(Cell, Number Of Letters)
L> e.g. =LEFT(A2, 5) ; =RIGHT(A2, 4)
L> =MID(Cell, Starting Point Letter Number, Number Of Letters)
L> e.g. =MID(A2, 3, 2)

The CONCATENATE Function
L> joins together two or more text strings
L> put info from two cells into one cell
L> =CONCATENATE(Cell 1, Cell 2)
L> e.g. =CONCATENATE(H2, I2)

The TRIM Function
L> removing leading, trailing and repeated spaces in data
L> if there are too many "spaces" or "-" or ".."
L> =TRIM(Cell), e.g. =TRIM(A2)

Datei auswählen Keine ausgewählt

Cleaning Spreadsheets 4

Freitag, 11. Juli 2025 10:29

Sorting
L> arranging data into a meaningful order
L> to make it easier to understand, analyse and visualise

Filtering
L> showing only the data that meets a specific criteria
L> while hiding the rest

Pivot Table
L> data summarisation tool used in data processing

VLOOKUP
L> means virtual look up

Data Mapping
L> process of matching fields from one databases to another
L> critical for data migration, integration and more

Schema
L> way of describing how something is organised

Primary Key
L> references a column in which each value is unique

Foreign Key
L> a field within a table that is a primary key in another table

Pivot Tables
L> data summarisation tool
L> used in data processing and data cleaning
L> we used it to copy 2 rows into a table
L> go to the tab *"Insert"*
L> click on *"Pivot Table"*
L> there is an self-explaining editor

The VLOOKUP Function
L> vertically searches for a certain value in a column
L> return a corresponding piece of information
L> but usually we're searching across multiple sheets/databases
L> that's what we will do now
L> =VLOOKUP(data to look up, 'where to look'!Range, column, false)
L> =VLOOKUP(A2, 'Sheet 2'!A1:B31, 2, false)
L> **take value in A2** of Sheet 1
L> **checking value in Sheet 2** from A1 to B31
L> we **want** the **value in the 2nd column**
L> *"false"* means, it will **only search for exact matches**
L> Optional: Locking the column! A2 or A$2

Creating a Chart
L> we learned that already
L> mark the column *"Prizes"*
L> open the tab *"Insert"*
L> click on *"Chart"*

The CONCATENATE Function
L> can put to strings of to cells into one together
L> we will use it to put two datasets together
L> First: We can add Spaces between the united strings
L> =CONCATENATE(Cell , " ", Cell)
L> the " " can make us put sth in between
L> it could also be "_"

"Data Cleaning Checklist"

Determine the size of the dataset:

Large datasets may have more data quality issues and take longer to process. This may impact your choice of data cleaning techniques and how much time to allocate to the project.

Determine the number of categories or labels:

By understanding the number and nature of categories and labels in a dataset, you can better understand the diversity of the dataset. This understanding also helps inform data merging and migration strategies.

Datei a**Identify missing data:**

Recognizing missing data helps you understand data quality so you can take appropriate steps to remediate the problem. Data integrity is important for accurate and unbiased analysis.

Identify unformatted data:

Identifying improperly or inconsistently formatted data helps analysts ensure data uniformity. This is essential for accurate analysis and visualization.

Explore the different data types:

Understanding the types of data in your dataset (for instance, numerical, categorical, text) helps you select appropriate cleaning methods and apply relevant data analysis techniques.

Glossary

Freitag, 11. Juli 2025 10:29

Terms and definitions for Course 4, Module 2

Clean data: Data that is complete, correct, and relevant to the problem being solved

Compatibility: How well two or more datasets are able to work together

CONCATENATE: A spreadsheet function that joins together two or more text strings

Conditional formatting: A spreadsheet tool that changes how cells appear when values meet specific conditions

Data engineer: A professional who transforms data into a useful format for analysis and gives it a reliable infrastructure

Data mapping: The process of matching fields from one data source to another

Data merging: The process of combining two or more datasets into a single dataset

Data validation: A tool for checking the accuracy and quality of data

Data warehousing specialist: A professional who develops processes and procedures to effectively store and organize data

Delimiter: A character that indicates the beginning or end of a data item

Dirty data: Data that is incomplete, incorrect, or irrelevant to the problem to be solved

Duplicate data: Any record that inadvertently shares data with another record

Field length: A tool for determining how many characters can be keyed into a spreadsheet field

Incomplete data: Data that is missing important fields

Inconsistent data: Data that uses different formats to represent the same thing

Incorrect/inaccurate data: Data that is complete but inaccurate

LEFT: A function that returns a set number of characters from the left side of a text string

LEN: A function that returns the length of a text string by counting the number of characters it contains

Length: The number of characters in a text string

Merger: An agreement that unites two organizations into a single new one

MID: A function that returns a segment from the middle of a text string

Null: An indication that a value does not exist in a dataset

Outdated data: Any data that has been superseded by newer and more accurate information

Remove duplicates: A spreadsheet tool that automatically searches for and eliminates duplicate entries from a spreadsheet

Split: A function that divides text around a specified character and puts each fragment into a new, separate cell

Substring: A smaller subset of a text string

Text string: A group of characters within a cell, most often composed of letters

TRIM: A function that removes leading, trailing, and repeated spaces in data

Unique: A value that can’t have a duplicate

Datei auswählen Keine ausgewählt