# Process Data from Dirty to Clean

Saturday, January 28, 2023 22:22

# Data Integrity

Tuesday, October 4, 2022 18:17

**Data Integrity**: The accuracy, completeness, consistency and trustworthiness of data throughout its lifecycle

Data integrity can be compromised when it's replicated, transferred, manipulated.

**Data replication**: The process of storing data in multiple locations

**Data transfer**: The process of copying data from a storage device to memory, or from one computer to another

**Data manipulation**: The process of changing data to make it more organized and easier to read.

other threats to data integrity:

- human error
- viruses
- malware
- hacking
- system failures

possible flaws in the data

- duplicates > clean data
- not enough / incomplete data > adjust objectives

clean data + alignment to business objectives = accurate conclusions

Types of insufficient data

- data from only one source
- data that keeps updating
- outdated data
- geographically limited data

Ways to address insufficient data:

- Identify trends with the available data
- Wait for more data if time allows

- talk with stakeholders and adjust your objective
- look for a new dataset

## DATA ERRORS

can you fix or request a corrected dataset? → YES → perform analysis after the data has been corrected

↓ NO

do you have enough data to omit the wrong data? → YES → perform the analysis without the wrong data

↓ NO

## NOT ENOUGH DATA

can you proxy the data? → YES → perform the analysis with the proxied data

↓ NO

can you collect more data? → YES → perform the analysis after data collection

↓ NO

modify the business objective (if possible)

## The importance of sample size

**Random sampling:** a way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

**margin of error:** difference between results obtained from surveying a sample vs. the entire population

**confidence level:** how confident you are in the survey results,

population

**confidence level**: how confident you are in the survey results,
95% CL » run same survey 100 times,
get similar results 95 out of 100 times

**confidence interval**: sample result +/- the margin of error,
range of possible values that the population's
result would be at the confidence level of the
study

**statistical significance**: determination of whether or not the
results could be due to random chance

use a minimum sample size of 30 because of the Central Limit Theorem

**What to do when there is no data**

- use proxy data
  new car model sales numbers —proxy→ clicks on car specs on website
  demand for new plant-based meat product → sales data for tofu turkey substitute
- open (public) datasets
  ↳ CSV, JSON, SQLite, and BigQuery datasets

# Statistical Power and margin of error

Tuesday, October 4, 2022 22:10

The probability of getting meaningful results from a test, out of 1. e.g. if statistical power = 0.6 >> 60% chance of getting a statistically significant result from a test, usually need at least 0.8 or 80%.

**Hypothesis testing**: a way to see if a survey or experiment has meaningful results

**Margin of Error**: the maximum amount that the sample results are expected to differ from those of the actual population

To calculate margin of error you need:

- Population size
- Sample size
- Confidence level

# Data Cleaning

Thursday, October 6, 2022 21:20

the #1 cause of poor data quality is human error

- someone entering data incorrectly
- inconsistent formatting
- blank fields
- duplicates

**Dirty Data**: data that is incomplete incorrect, or irrelevant to the problem you're trying to solve

**Clean Data**: data that is complete, correct, and relevant
data is more likely to be clean if it is internal to your company and has been verified and is cared for by the data engineers

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

**Data warehousing specialists** develop processes and procedures to effectively store and organize data

**Null**: an indication that a value does not exist in a dataset
↳ need to clean data that contains Nulls, could remove those datapoints or learn from the fact that they are Null, e.g. question was phrased confusingly

## Types of dirty data:

- Duplicate data: any data record that shows up more than once
- outdated data: old data that should be replaced with newer information
- incomplete data: any data that is missing important fields
- incorrect/inaccurate data
- inconsistent data: uses different formats to represent the same thing

## Recognize & Remedy dirty data

- misspelling, spelling variations, mixed up letters, inconsistent punctuation,

- misspelling, spelling variations, mixed up letters, inconsistent punctuation, typos
- different currencies, units
- inconsistent formatting
- nulls > might need to dig into the data to fill empty fields
- duplicates > delete one of the duplicates
- incorrect labelling
- inconsistent field length > assign certain length to avoid errors
  ↳ tool for determining how many characters can be keyed into a field

Data validation: a tool for checking the accuracy and quality of data before adding or importing it

# Cleaning data from multiple sources

Monday, October 10, 2022 18:38

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

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

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

- Do I have all the data I need?
- Does the data I need exist within these datasets?
- Do the datasets need to be cleaned, or are they ready for use?
- Are the datasets cleaned to the same standard?

## Common data cleaning pitfalls:

- not checking for spelling errors
- forgetting to document errors
- not checking for misfielded values
- overlooking missing values
- looking at a subset of data and not the whole picture
- loosing track of the business objectives
- not fixing the source of the error
- not analyzing the system prior to data cleaning
- not backing up your data prior to data cleansing
- not accounting for data cleaning in your deadlines/process

# In spreadsheets

Monday, October 10, 2022 19:32

**conditional formatting** · a spreadsheet tool that changes how cells appear when values meet specific conditions.

> select range > format > conditional formatting > choose format rule & style

**remove duplicates**

> Data > remove duplicates > ☑ data has header row > select all > remove duplicates

**make format consistent**

> select range > format > number > date (e.g.)

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

**Split**: a tool that divides text around a specific character and puts each fragment into a new, separate cell

**Concatenate**: a function that joins multiple text strings into a single string = CONCATENATE(item 1, item 2)

## Optimize the data-cleaning process

**COUNTIF**: a function that returns the number of cells that match a specific value = COUNTIF(range, "value")

**Syntax**: a predetermined structure that includes all required information and its proper placement

**LEN**: tells the length of a text string by counting the number of characters it contains = LEN(range)

**LEFT**: returns a set number of characters from the left side of a text string

↳ **RIGHT**

↳ **MID** = MID(range, reference starting point, number of middle characters)

**TRIM**: removes leading, trailing, and repeated spaces in data = TRIM(range)

# More data cleaning

Sunday, October 16, 2022 21:23

- Pivot tables
- **VLOOKUP**: a function that searches for a certain value in a column to return a corresponding piece of information = VLOOKUP(data to look up, 'where to look'! Range, column, false)
- plotting to visualize data and quickly identify outliers & errors

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

**Schema**: a way of describing how something is organized

data mapping: ① determine the content of each section/column to ensure it ends up in the right place
② transform data into a consistent format
③ transfer data to its destination (use queries, import wizards, drag&drop)
④ testing phase
- inspect sample piece of data to confirm that it's clean and properly formatted
- spot checks (e.g. check number of nulls)

# Using SQL to clean data

Monday, October 17, 2022 20:24

## SQL capabilities

- handle huge amounts of data

## Spreadsheets & SQL

↳ both: arithmetic, use formulas, join data

COUNIF() >> COUNT + WHERE

| Spreadsheets | SQL |
|---|---|
| • generated with a program (Spreadsheets, Excel) | • a language used to interact with database programs (MySQL, Microsoft SQL Server) |
| • access to the data you input | • can pull information from different sources in the database |
| • stored locally | • stored across a database |
| • small datasets | • larger datasets |
| • working independently | • tracks changes across team |
| • built-in functionalities (e.g. spellcheck) | • useful across multiple programs |

## Widely used SQL queries

- SELECT
- FROM
- INSERT INTO customer_data.customer_address
  (customer_id, address, city, state, zipcode, country)
  VALUES
  (2645, '333 SQL Road', 'Jackson', 'MI', 49202, 'US')
- UPDATE customer data.customer_address
  SET address = '123 New Address'
  WHERE customer_id = 2645
- CREATE TABLE IF NOT EXISTS
- DROP TABLE IF EXISTS

## Cleaning string variables using SQL

- Including DISTINCT in your SELECT statement

- Including DISTINCT in your SELECT statement removes duplicates

```
SELECT
    DISTINCT customer_id
FROM
    customer_data.customer_address
```

- LENGTH / LEN

```
SELECT
    LENGTH(country) AS letters_in_country
FROM
    customer_data.customer_address
```

---

```
SELECT
    country
FROM
    table_name
WHERE
    LENGTH(country) > 2
```

- SUBSTR()

```
SELECT DISTINCT customer_id
FROM table_name
WHERE SUBSTR(country, 1, 2) = 'US'
```

- TRIM()

```
SELECT DISTINCT customer_id
FROM table_name
WHERE TRIM(state) = 'OH'
```

# Advanced data cleaning functions

Wednesday, October 19, 2022 21:23

- CAST() can be used to convert anything from one data type to another

```
CAST(field AS FLOAT64)
```

- ORDER BY

```
SELECT
    purchase_price
FROM
    tablename
ORDER BY
    purchase price DESC
```

typecasting: converting data from one type to another

- 
```
SELECT
    purchase_price, date
FROM tablename
WHERE
    date BETWEEN '2020-12-01' AND '2020-12-31'
```

- CONCAT()

```
SELECT
    CONCAT(product_code, product_color) AS new_product_code
```

- COALESCE() - replaces null values in first column with values from second column

```
SELECT
    COALESCE(product, product_code) AS product_data
FROM tablename
```

# Verifying and reporting results

Thursday, October 20, 2022 21:27

**verification**: a process to confirm that a data cleaning effort was well executed and the resulting data is accurate and reliable

**changelog**: a file containing a chronologically ordered list of modifications made to a project

**Verifying the clean data**

> check for nulls

> use conditional formatting, search manually, use filters

> check for common missspellings noticed during cleanup

> take a **big-picture view** of your project to confirm you are focusing on the business problem at hand

1. consider the business problem
2. consider the goal
3. consider the data : can it solve the problem? Is it aligned to the goal?

> Do the numbers make sense?

**Pivot table**: a data summarization tool that is used in data processing (spreadsheet)

**Find and replace**: a tool that looks for a specified search term in a spreadsheet and allows you to replace it with something else (spreadsheet)

**COUNTA**: a function that counts the total number of values within a specified range (spreadsheet)

**CASE statement** (SQL): The CASE statement goes through one or more conditions and returns a value as soon as a condition is met

```
SELECT
   customer_id,
   CASE
      WHEN first_name = 'Tnoy' THEN 'Tony'
      ELSE first_name
```

```
CASE
    WHEN first_name = 'Tnoy' THEN 'Tony'
    ELSE first_name
    END AS cleaned_name
FROM ...
```

## CHECKLIST

- Sources of error: did you use the right tools and functions to find the source of the errors in your dataset?
- Null data > conditional formatting & filters
- Missspelled words
- Mistyped numbers
- Extra spaces and characters > TRIM, LENGTH
- Duplicates > Remove duplicates, DISTINCT
- Mismatched data types: numeric, date string typecast correctly?
- Messy/inconsistent strings
- Messy/inconsistent date formats
- Misleading variable labels (columns) > give meaningful names
- Truncated data
- Business Logic: does the data make sense given your knowledge of the business?

# Documenting results and the cleaning process

Sunday, October 23, 2022 13:54

**Documentation:** the process of tracking changes, additions, deletions, and errors involved in your data cleaning effort
- recover data cleaning errors
  ↳ create clean table rather than overwriting
- inform other users of changes
- determine quality of data

**Changelog:** file containing a chronologically ordered list of modifications made to a project
→ version history in Google sheets, 'Track changes' in Excel
→ BigQuery's Query History
≙ Engineering Chang Orders, document revision histories

changelog content:
- data, file, formula, query, any other component that changed
- description of what changed
- date of the change
- person who made the change
- person who approved the change
- version number
- reason for the change

## Best practices for changelogs
- write legibly
- every version should have its own entry
- each change should have its own line
- group the same types of changes
- versions should be ordered chronologically starting with the latest
- the release date of each version should be noted

## types of changes
- Added: new features introduced
- Changed: changes in existing functionality
- Deprecated: features about to be removed
- Removed: features that have been removed
- Fixed: bug fixes
- Security: lowering vulnerabilities

## Common data errors
- human error in data entry
- flawed processes
- system issues

## Advanced Functions for Speedy data cleaning (spreadsheets)

| | Syntax (Google Sheets) | MENU OPTIONS (MS Excel) | PRIMARY USE |
|---|---|---|---|
| IMPORTRANGE | =IMPORTRANGE (spreadsheet_url, range_string) | Paste Link (copy the data first) | Imports (pastes) data from one sheet to another and keeps it automatically updated |

| | | | |
|---|---|---|---|
| QUERY | = QUERY (Sheet and Range, "SELECT*" | Data > From other sources > from Microsoft Query | Enables psudo SQL (SQL-like) statements or a wizard to import the data. |
| FILTER | = FILTER (range, condition 1, [condition 2, ...]) | Filter (conditions per column) | Displays only the data that meets the specified conditions |