Course 4 - Process Data from Dirty to Clean

What does data integrity imply?

- accuracy
- completeness
- consistency
- trustworthiness

In what ways data integrity can be compromised?

Data can be compromised every time during:

- **replication** — a process of storing data in multiple locations that can cause those data to be out of sync and inconsistent,
- **transfer** — a process of copying data from one storage device to memory or from one computer to another. If this transfer is interrupted, this can compromise the integrity and load an incomplete dataset.
- **manipulation** — changing the data to make it more organized and easier to read.
- human error, malware, hacking and system failures.

What are data constraints?

Data constraints are criteria that determine the validity.

- **datatype** — date, number, bool…
- **data range** — predefined min and max values
- **mandatory** — can’t be left blank or empty
- **unique** — can’t have duplicates
- **regex** — values must match a prescribed pattern
- **cross-field validation** — condition for multiple fields
- **primary key** — value unique per column
- **foreign key** — values for a column must be unique from a column in another table
- **set-membership** — values for a column must come from a set of discrete values
- **accuracy** — the degree to which the data conforms to the actual entity being measured or described
- **completeness** — the degree to which the data contains all desired components or measures

- **consistency** — the degree to which the data is repeatable from different points of entry or collection.

What to do when there is no data?

● gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after collecting more data

● if there isn't time to collect data, perform analysis on proxy data from another dataset (the most common case)

What to do when there is too little data?

● do the analysis with the proxy data along with actual data

● adjust the analysis to align with the data you already have

What to do with wrong data/ data errors?

○ if the reason for wrong data is that requirements were misunderstood, communicate the requirements again.

○ identify errors and if possible, correct them at the source by looking for a pattern in the errors.

○ if you can't correct errors, ignore the wrong data and go ahead with the analysis if the sample size is large enough, and ignoring errors won't cause systematic bias.

Calculating sample size — terminology

- **population** — the entire group
- **sample** — a subset of a population
- **margin of error** — sample's results are expected to differ from what the result would have been for the entire population. This difference is a margin of error.
- **confidence level** — how confident are you in a survey result. The confidence level is targeted before we start our study because it will affect how big the margin of error is.
- **confidence interval** — the range of possible values that the population's result would be at the confidence level of the study. This range is a simple result +- the margin of error.
- **statistical-significance** — the determination of whether your result could be due to random chance or not.

What to remember when determining the size of the sample?

- don't use a sample size of less than 30
- the confidence level most commonly used is 95%, but 90% can work in some cases
- increase the sample size for:

— a larger confidence level

— decrease the margin of error,

— for greater statistical significance

What task should be completed before analyzing data?

● **determine data integrity** by assessing the accuracy, consistency, and completeness of the data.

● **connect objectivities** to the data to understand how business objectives can be served by an investigation into the data.

● know **when to stop to collect data.**

What makes data insufficient?

- coming from only one source
- continuously updated and is incomplete
- is outdated
- is geographically limited

How to deal with insufficient data?

- identify trends within the available data
- wait for more data if time allows
- discuss with stakeholders and adjust their objectives
- search for the new dataset

What is statistical power?

● probability of getting meaningful results from a test

● the larger the sample size the greater the statistically significant results — that's statistical power.

● SP is usually shown as a value out of one. We need a statistical power of at least 0.8 (80%) to consider the results statistically significant.

● statistically significant means that the results of the test are real and not an error caused by a random chance.

How to determine the best sample size?

Margin of Error Calculator-

The sample size is a part of the population that is representative.

sample size calculators require input on:

1. **confidence level** — the probability that the sample accurately reflects the greater population (90–95% are minimum)
2. **the margin of error** — how close the sample size results are to what our results would be if we used the entire population
3. and **population size**.

What to do with the results?

The calculated sample size is the minimum number to achieve what you input for confidence level and margin of error.

What is the most common cause of dirty data?

Human error.

- typing in a piece of data incorrectly
- inconsistent formatting
- blank fields
- duplicates

Types of dirty data and consequences

- **duplicated** data — skewed metrics or analysis
- **outdated** data — inaccurate insights, decision-making and analytics
- **incomplete** data — decreased productivity, inaccurate insight
- **incorrect/inaccurate** data — inaccurate insight and decision making
- **inconsistent** data — contradictory data points lead to inability to clasify or segment data

What is a data validation?

A data validation is **a tool for checking the accuracy and quality** of data before adding or importing it.

What are the principles of data integrity?

- **Validity** — the concept of using data integrity principles to ensure measures conform to defined business rules or constraints
- **Accuracy** — the degree of conformity of a measure to a standard or a true value
- **Completeness** — the degree to which all required measures are known

Data cleaning tools and techniques

Always make a copy of the dataset first!

remove unwanting data:

- remove duplicates
- remove irrelevant data (that doesn't fit a problem we're trying to solve)
- remove extra spaces and blanks
- fix misspellings, inconsistent capitalization, incorrect punctuation, typos

- use spellcheck, autocorrect, and conditional formatting, and convert text to lowercase, uppercase, or proper case.

What is a merger?

A merger is **an agreement that unites two organizations** into a single new one. All the data from each organization would need to be combined using data merging.

What is data merging?

A data merging is a **process of combining two or more datasets into a single dataset**. Those datasets need to be compatible.

What questions do we need to ask while checking compatibility?

- do we have all the data we need? -do datasets give me the info to answer business questions/solve a business problem?
- does the data we need exist within these datasets?
- do the datasets need to be cleaned?
- are datasets cleaned to the same standard?
- how are missing values handled?
- how recently was the data updated?

What are common data cleaning mistakes?

- no checking for spelling errors
- forgetting to document errors
- no checking for misfield values (when values are entered into a wrong field)
- overlooking missing values
- looking at a subset of data and not the whole picture
- losing track of the business objectives
- not fixing the source of an error
- not analyzing the system prior to data cleaning (to figure out where errors come from — data entry, lack of formats, duplicates…)
- not backing up the data before data cleaning
- not accounting for data cleaning in deadlines/process

What are efficiency tools that data analysts use?

- conditional formatting

- removing duplicates
- formatting dates
- fixing text strings and substrings
- splitting text to columns (data — split text to columns)

Excel basic functions for cleaning data

A function is a set of instructions that performs a specific calculation using the data in a spreadsheet

- COUNTIF (range, condition) — returns the number of cells that match specific criteria
- LEN (range) — returns the length of a text
- LEFT/RIGHT — gives us the set number of characters from the left/right side of the string
- MID — returns a segment from the middle of the text
- CONCATENATE — combines 2 or more text strings
- TRIM — removes leading, trailing, and repeating spaces

What workflow can be automated?

workflow automation is the process of automating parts of your work.

- **modeling the data** — creating DB structure from diagrams, creating business-specific infographics, diagrams, data visualizations, and flowcharts

we can **partially automate:**

- **preparing and cleaning data** — some tasks like detecting missing values
- **data exploration** — some tasks like visualization

What are different data perspectives we can apply to our dataset?

- sorting — we can easily find duplicates
- filtering — for showing only the data that meet specific criteria
- pivot table — data summarization tool for sorting, grouping, counting, total and average data
- VLOOKUP — searches for a certain value in a column
- plotting — putting data in a graph, chart, or other visual

Data cleaning verification checklist:

● **Sources of errors**: Did you use the right tools and functions to find the source of the errors in your dataset?

● **Null data**: Did you search for NULLs using conditional formatting and filters?

● **Misspelled words**: Did you locate all misspellings?

● **Mistyped numbers**: Did you double-check that your numeric data has been entered correctly?

● **Extra spaces and characters**: Did you remove any extra spaces or characters using the TRIM function?

● **Duplicates**: Did you remove duplicates in spreadsheets using the Remove Duplicates function or DISTINCT in SQL?

● **Mismatched data types**: Did you check that numeric, date, and string data are typecast correctly?

● **Messy (inconsistent) strings**: Did you make sure that all of your strings are consistent and meaningful?

● **Messy (inconsistent) date formats**: Did you format the dates consistently throughout your dataset?

● **Misleading variable labels (columns)**: Did you name your columns meaningfully?

● **Truncated data**: Did you check for truncated or missing data that needs correction?

● **Business Logic**: Did you check that the data makes sense given your knowledge of the business?

What are the steps to review the goal of the project?

1. confirm the **business problem**
2. confirm the **goal** of the project
3. verify that **data can solve the problem and is aligned with the goal**

What is the documentation?

The documentation is a process of tracking changes, additions, deletions, and errors during data cleaning. It’s staged chronologically and provides a real-time account of every modification.

What are the advantages of documentation?

- it lets us discover data cleaning errors,
- it is a way to inform other users of changes that have been made,
- and it helps us to determine the quality of the data

What is a changelog?

A changelog is a file containing a chronologically ordered list of modifications made to a project

- in spreadsheets — File / Version history
- in SQL — specify exactly what and when you commit a query, or just add comments as you go while cleaning data in SQL

Difference between changelogs and version history

A changelog can build on the automated version history. Version histories record what was done in a data change, but don't tell us why. Changelogs help us understand the reasons changes have been made.

What type of information a changelog should record?

- data, file, formula, query, or any other **component that changed**
- **description** of that change
- **date** of the change
- a **person who made** that change
- a **person who approved** that change
- **version** number
- reason for the change

Changelog best practices:

- changelogs are for humans, so write clear
- every version should have its own entry
- each change should have its own line
- group the same changes
- versions should be ordered chronologically, latest to newest
- the release date of each version should be noted

How to group categories in changelogs?

All changes usually fall into one of the following categories and should be grouped together:

- **Added** — new features introduced
- **changed** — changes in existing functionality
- **deprecated** — features about to be removed

- **removed** — features that have been removed
- **fixed** — bug fixed
- **security** — lowering vulnerability

What changes should be captured in the changelog while cleaning the dataset?

- treated missing data
- changed formatting
- changed values or cases for data

Most common errors in data

- human mistakes — mistyping or misspelling
- flawed processes — poor design or survey form
- system issues — older systems integrate data incorrectly

How do we import data from one sheet to another?

1. with =**IMPORTRANGE(spreadsheet_url, range_string)** function

IMPORTRANGE —

- data are automatically updated,
- more efficient than copying and pasting on a large set of data
- reduce the chance of errors,
- helpful for data cleaning because we can pick the data relevant to the project
- if we want to share data, we need to allow access first time

2. with =**QUERY(data, "query", [headers])** function

QUERY function-

- extract specific data within a spreadsheet
- faster than filtering manually
- can be combined with other functions for more complex calculations
- we can use a simple SQL statement to extract specific data

Filtering data with the FILTER function

=FILTER(range, condition1, [condition2, …])

FILTER function -

- FILTER function is fully internal to a spreadsheet and doesn’t require the use of query language.
- it lets us view only data that meet our criteria
- faster than the QUERY function