

Course 4:

Process Data from Dirty to Clean

Week 1: The importance of integrity

11-02-2022

Why data integrity is important

A strong analysis depends on the integrity of the data.

Data integrity is the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

Data replication is the process of storing data in multiple locations.

Data transfer is the process of copying data from a storage device to memory, or from one computer to another.

Data manipulation is 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

Data constraint	Definition	Examples
Data type	Values must be of a certain type: date, number, percentage, Boolean, etc.	If the data type is a date, a single number like 30 would fail the constraint and be invalid

Data range	Values must fall between predefined maximum and minimum values	If the data range is 10-20, a value of 30 would fail the constraint and be invalid
Mandatory	Values can't be left blank or empty	If age is mandatory, that value must be filled in
Unique	Values can't have a duplicate	Two people can't have the same mobile phone number within the same service area
Regular expression (regex) patterns	Values must match a prescribed pattern	A phone number must match ###-###-#### (no other characters allowed)
Cross-field validation	Certain conditions for multiple fields must be satisfied	Values are percentages and values from multiple fields must add up to 100%
Primary-key	(Databases only) value must be unique per column	A database table can't have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program.

Set-membership	(Databases only) values for a column must come from a set of discrete values	Value for a column must be set to Yes, No, or Not Applicable
Foreign-key	(Databases only) values for a column must be unique values coming from a column in another table	In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table
Accuracy	The degree to which the data conforms to the actual entity being measured or described	If values for zip codes are validated by street location, the accuracy of the data goes up.
Completeness	The degree to which the data contains all desired components or measures	If data for personal profiles required hair and eye color, and both are collected, the data is complete.
Consistency	The degree to which the data is repeatable from different points of entry or collection	If a customer has the same address in the sales and repair databases, the data is consistent.

Balancing objectives with data integrity

It's important to check that the data you use aligns with the business objective. Good alignment means that the data is relevant and can help you solve a business problem or determine a course of action to achieve a given business objective.

- When there is clean data and good alignment, you can get accurate insights and make conclusions the data supports.
- If there is good alignment but the data needs to be cleaned, clean the data before you perform your analysis.
- If the data only partially aligns with an objective, think about how you could modify the objective, or use data constraints to make sure that the subset of data better aligns with the business objective.

Dealing with insufficient data

Types of insufficient data

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

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 new dataset

Data issue 1: no data

Possible Solutions	Examples of solutions in real life
Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the	If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3

analysis after you have collected more data.	weeks to collect the data from all employees.
If there isn't time to collect data, perform the analysis using proxy data from other datasets. <i>This is the most common workaround.</i>	If you are analyzing peak travel times for commuters but don't have the data for a particular city, use the data from another city with a similar size and demographic.

Data issue 2: too little data

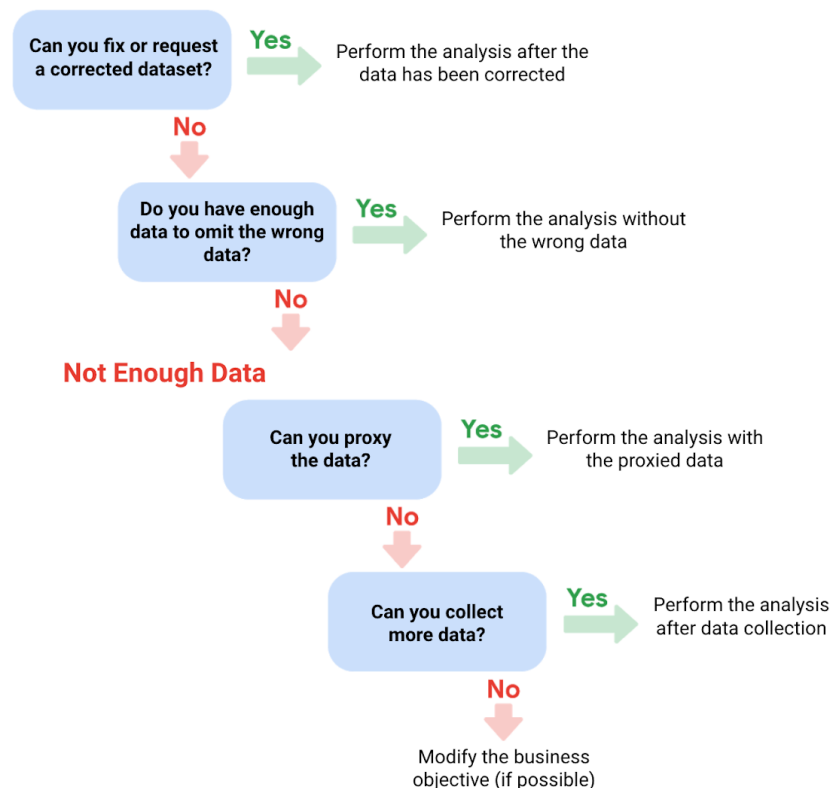
Possible Solutions	Examples of solutions in real life
Do the analysis using proxy data along with actual data.	If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors.
Adjust your analysis to align with the data you already have.	If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: <i>this conclusion applies to adults 25 years and older only.</i>

Data issue 3: wrong data, including data with errors*

Possible Solutions	Examples of solutions in real life
If you have the wrong data because requirements were misunderstood, communicate the requirements again.	If you need the data for female voters and received the data for male voters, restate your needs.

Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors.	If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values.
If you can't correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won't cause systematic bias.	If your dataset was translated from a different language and some of the translations don't make sense, ignore the data with bad translation and go ahead with the analysis of the other data.

Data Errors



The importance of sample size

Population: all possible data values in a certain dataset

Sample size: a part of a population that is representative of the population.

Sampling bias: a sample isn't representative of the population as a whole.

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.

Terminology	Definitions
Population	The entire group that you are interested in for your study. For example, if you are surveying people in your company, the population would be all the employees in your company.
Sample	A subset of your population. Just like a food sample, it is called a sample because it is only a taste. So if your company is too large to survey every individual, you can survey a representative sample of your population.
Margin of error	Since a sample is used to represent a population, the sample's results are expected to differ from what the result would have been if you had surveyed the entire population. This difference is called the margin of error. The smaller the margin of error, the closer the results of the sample are to what the result would have been if you had surveyed the entire population.

Confidence level	How confident you are in the survey results. For example, a 95% confidence level means that if you were to run the same survey 100 times, you would get similar results 95 of those 100 times. Confidence level is targeted before you start your study because it will affect how big your margin of error is at the end of your study.
Confidence interval	The range of possible values that the population's result would be at the confidence level of the study. This range is the sample result +/- the margin of error.
Statistical significance	The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance.

Things to remember when determining the size of your sample:

- Don't use a sample size less than 30. It has been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the average result of a population.
- The confidence level most commonly used is 95%, but 90% can work in some cases.
- For a **higher** confidence level, use a larger sample size
- To **decrease** the margin of error, use a larger sample size
- For **greater** statistical significance, use a larger sample size

12-02-2022

Using Statistical Power


Statistical power is the probability of getting meaningful results from a test.

Hypothesis testing is a way to see if a survey or experiment has meaningful results.

If a test is statistically significant, it means that results of the test are real and not an error caused by random chance. Usually, you need a statistical power of at least 0.8 or 80% to consider your results statistically significant.

Determining the best sample size

Confidence level is the probability that your sample accurately reflects the greater population.

 **Sample Size Calculator**

[Sample Size Calculator](#)


[Sample Size Calculator by Raosoft, Inc.](#)

Evaluate the reliability of your data

Margin of error is the maximum amount that the sample results are expected to differ from those of the actual population.

To calculate margin of error we need:

- population size
- sample size
- confidence level

 **Margin of Error Calculator**

[Margin of Error Calculator](#)

[Sample size calculator - CheckMarket](#)

12-02-2022

Week 2: Sparkling clean data

1 cause of poor quality data = human error

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

Why data cleaning is important

Clean data is data that is complete, correct, and relevant to the problem you're trying to solve.

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 is an indication that a value does not exist in a dataset.

Types of dirty data:



Duplicate data



Outdated data



Incomplete data



Incorrect/inaccurate data



Inconsistent data

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

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

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

14-02-2022

Data-cleaning tools and techniques

- remove duplicate data
- remove irrelevant data
- remove extra spaces and blanks
- fix misspellings
- fix inconsistent capitalization
- fix incorrect punctuation and other typos
- remove formatting

Cleaning data from multiple sources

Merger is an agreement that unites two organizations into a single new one.

Data merging is the process of combining two or more datasets into a single dataset.

Compatibility is 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?
- Does the data need to be cleaned, or are they ready for me to use?
- Are the datasets cleaned to the same standards?

These are some questions to start with before starting analysis on merging data.

Common pitfalls



[10 Google Workspace tips to clean up data](#)

[Top ten ways to clean your data](#)

Data-cleaning features in spreadsheets

Conditional formatting is a spreadsheet tool that changes how cells appear when values meet specific conditions.

Remove duplicates is a tool that automatically searches for and eliminates duplicate entries from a spreadsheet.

Text string is a group of characters within a cell, most often composed of letters.

Split is a tool that divides text around a specified character and puts each fragment into a new, separate cell.

Optimize the data-cleaning process

COUNTIF is a function that returns the number of cells that match a specified value.

`=COUNTIF(range,"value or criteria")`

LEN is a function that tells you the length of a text string by counting the number of characters it contains.

LEFT is a function that gives you a set number of characters from the left side of a text string.

`=LEFT(range,number of characters)`

RIGHT is a function that gives you a set number of characters from the right side of a text string.

`=RIGHT(range,number of characters)`

MID is a function that gives you a segment from the middle of a text string.

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

CONCATENATE is a function that joins together two or more text strings.

`=CONCATENATE(item1, item2)`

TRIM is a function that removes leading, trailing and repeated spaces in data.

=TRIM(range)

Workflow automation

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

Different data perspectives

Sorting can bring duplicates together to make it easy to find them.

Filtering can be used to filter out data using some parameters.

Pivot tables can give a visual for data.

VLOOKUP is 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)

Data mapping is the process of matching fields from one data source to another.

Week3: Cleaning data with SQL

25-02-2022

SQL advantages and disadvantages

Features of Spreadsheets	Features of SQL Databases
Smaller data sets	Larger datasets
Enter data manually	Access tables across a database
Create graphs and visualizations in the same program	Prepare data for further analysis in another of software
Built-in spell check and other useful functions	Fast and powerful functionality
Best when working solo on a project	Great for collaborative work and tracking queries run by all users
Generated with a program	A language used to interact with database programs
Access to the data you input	Can pull information from different sources in the database
Stored locally	Stored across a database
Working independently	Tracks changes across team

Widely used SQL queries

SELECT:

SELECT

name,

city

FROM

customer_data.customer_address

INSERT INTO:

```
INSERT INTO table_name(column_name)
VALUES (values corresponding to column names)
```

```
UPDATE table_name
SET column_name = value
WHERE column_name = value
```

CREATE TABLE IF NOT EXISTS

DROP TABLE IF EXISTS

Running queries does not save the results as a new table so we can either download the results as a csv file or save the data in new table.

Cleaning string variables using SQL

Including **DISTINCT** in SELECT statements.

```
SELECT DISTINCT column_name
FROM table_name
```

Text strings: a group of characters within a cell, commonly composed of letters, numbers or both.

LENGTH / LEN

```
SELECT
    LENGTH(column_name) AS letters_in_column
FROM
    table_name
```

```
SELECT
    column_name
```

FROM

table_name

WHERE

LENGTH(couolumn_name) <condition>

SUBSTR

SUBSTR(column_name,starting_letter_index, no_of_letters)

DISTINCT

TRIM()

CAST() to convert anything from one data type to another

CONCAT() adds strings together to create new text strings that can be used as unique keys

COALESCE() can be used to return non-null values in a list

Week 4: Verify and report on your cleaning results

14-04-2022

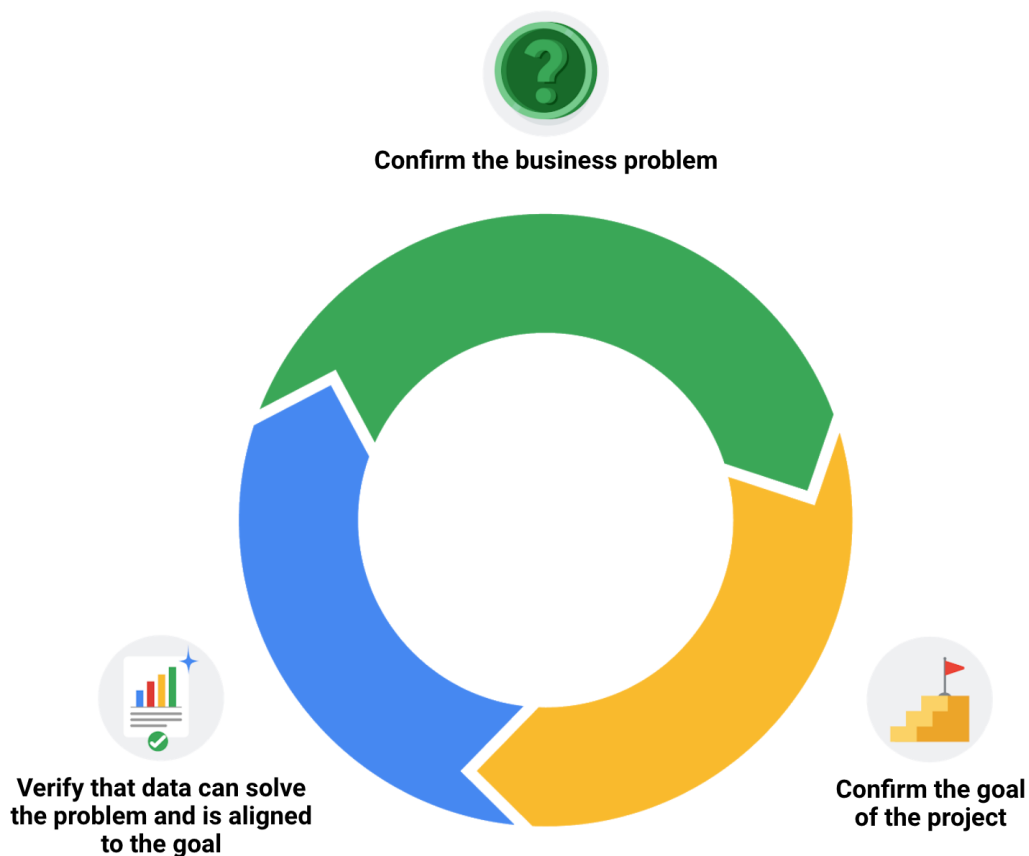
Verifying and reporting results

Verification: A process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable.

Cleaning your data expectations

See the big picture when verifying data-cleaning:

1. consider the business problem
2. consider the goal
3. consider the data



The final step in data cleaning

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

COUNTA: a function that counts the total number of values within a specified range.

CASE statement: The CASE statement goes through one or more conditions and returns a value as soon as a condition is met.

Most common problems in data:

- sources of errors
- null data
- misspelled words
- mistyped numbers
- extra spaces and characters
- duplicates
- mismatched data types
- messy/inconsistent strings
- inconsistent date formats
- misleading variable labels
- truncated data
- business logic

Documenting results and cleaning process

- helps recover data-cleaning errors
- inform other users of changes made
- determine the quality of data

Changelog: A file containing a chronologically ordered list of modifications made to a project.

Function	Syntax (Google Sheets)	Menu Options (Microsoft 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 pseudo SQL (SQL-like) statements or a wizard to import the data.
FILTER	=FILTER(range, condition1, [condition2, ...])	Filter (conditions per column)	Displays only the data that meets the specified conditions.

Week 5: Adding data to your resume

14-04-2022

Creating a Resume

- stick to one page resumes
- keep bullet points limited to 2-3 and keep them brief
- highlight your skills

- Contact information
 - name
 - address
 - phone number
 - email address
 - normally on the top
- Work history
 - start from most recent ones
- Summary (optional)
 - 2-3 sentences highlighting how you can be an asset to the company
- Qualifications
 - mention relevant qualifications for the job
- Skills

Accomplished [x] as measured by [y] by doing [z].

Be direct and coherent.

Problem Action Result statements.

▶ Live Portfolio and Resume Analysis with Data Science Hiring Mana...




[Kaggle CareerCon 2019 | Full Sessions - YouTube](#)

▶ How to Build a Compelling Data Science Portfolio & Resume | Kagg...

▶ Live Breakdown of Common Data Science Interview Questions | Kagg...

Technical skills

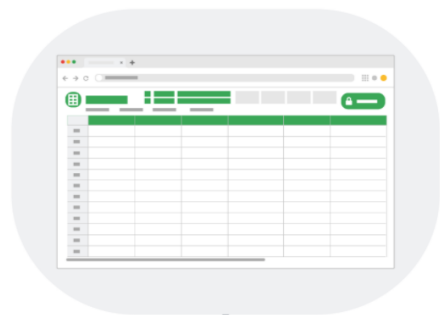
Soft skills

1	Presentation Skills	
2	Collaboration	
3	Communication	
4	Research	
5	Problem-solving skills	
6	Adaptability	
7	Attention to detail	

Professional skills



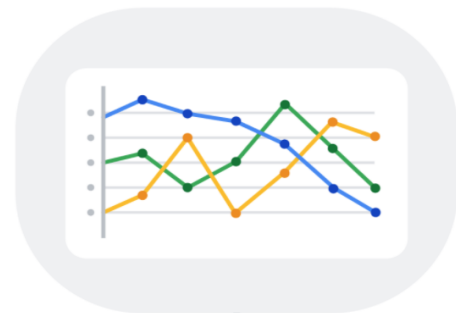
Structured Query
Language (SQL)



Spreadsheets



R or Python-Statistical
Programming



Data Visualization

Best practices for writing about experience

One of the most important functions of a resume is communicating your prior work experience in a favorable light. This can often be challenging, as the one-page format forces job seekers to summarize all of their work experience into a few bullet points.

Resume best practices will help you select the most relevant parts of your work experience and communicate them in the shortest, most impactful way possible.

As you think about how to represent your work experience on your resume effectively, it might be helpful to refer to these best practices:

Focus on your accomplishments first, and explain them using the formula “Accomplished X, as measured by Y, by doing Z.”

- These statements help you communicate the most important things a recruiter or hiring manager is searching for—the impact of your work.
- Whenever possible, use numbers to explain your accomplishments. For example, “Increased manufacturing productivity by 15% by improving shop floor employee engagement,” is better than “Increased manufacturing productivity.”

Phrase your work experience and duties using Problem-Action-Result (PAR) statements.

- For example, instead of saying “was responsible for two blogs a month,” phrase it as “earned little-known website over 2,000 new clicks through strategic blogging.”

Describe jobs that highlight transferable skills (those skills that can transfer from one job or industry to another).

- This is especially important if you are transitioning from another industry into data analytics.
- For example, communication is a skill often used in job descriptions for data analysts, so highlight examples from your work experience that demonstrate your ability to communicate effectively.

Describe jobs that highlight your soft skills.

- These are non-technical traits and behaviors that relate to how you work.
- Are you detail-oriented? Do you have grit and perseverance? Are you a strong critical thinker? Do you have leadership skills?
- For instance, you could give an example of when you demonstrated leadership on the job.
- Showing is always more effective than telling.

This is almost always the hardest part of crafting a resume, especially if you are transitioning from a different career field. However, if you take a moment to think deeply about your previous work experience, you’ll likely discover that you can find ways to represent your work experiences in a way that highlights your ability to do things important to data analyst roles, such as thinking critically or making data-driven decisions.