Process Data from Dirty to Clean
Google

Course 4 : Process Data from Dirty to Clean

Week1-The importance of integrity

Data integrity

The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

Data integrity can be compromised in lots of different ways. There's a chance data can be compromised every time it's replicated, transferred, or manipulated in any way.

- **Data replication**: The process of storing data in multiple locations. If you're replicating data at different times in different places, there's a chance your data will be out of sync. This data lacks integrity because different people might not be using the same data for their findings, which can cause inconsistencies.
- **Data transfer**: The process of copying data from a storage device to memory, or from one computer to another. If your data transfer is interrupted, you might end up with an incomplete data set, which might not be useful for your needs.
- **Data manipulation**: The process of changing the data to make it more organized and easier to read. Data manipulation is meant to make the data analysis process more efficient, but an error during the process can compromise the efficiency.
- **Other threats**: Data can also be compromised through human error, viruses, malware, hacking, and system failures.

Clean data + alignment to business objective = accurate conclusions
Alignment to business objective + newly discovered variables + constraints = accurate conclusions

Maintaining data integrity helps ensure a close alignment of data and business objectives because the data is likely to be accurate, complete, consistent, and trustworthy.

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- Outdated data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

Consider the following data issues and suggestions on how to work around them.
When you are getting ready for data analysis, you might realize you don't have the data you need or you don't have enough of it. In some cases, you can use what is known as proxy data in place of the real data. Think of it like substituting oil for butter in a recipe when you don't have butter. In other cases, there is no reasonable substitute and your only option is to collect more data.

Data issue 1: no data

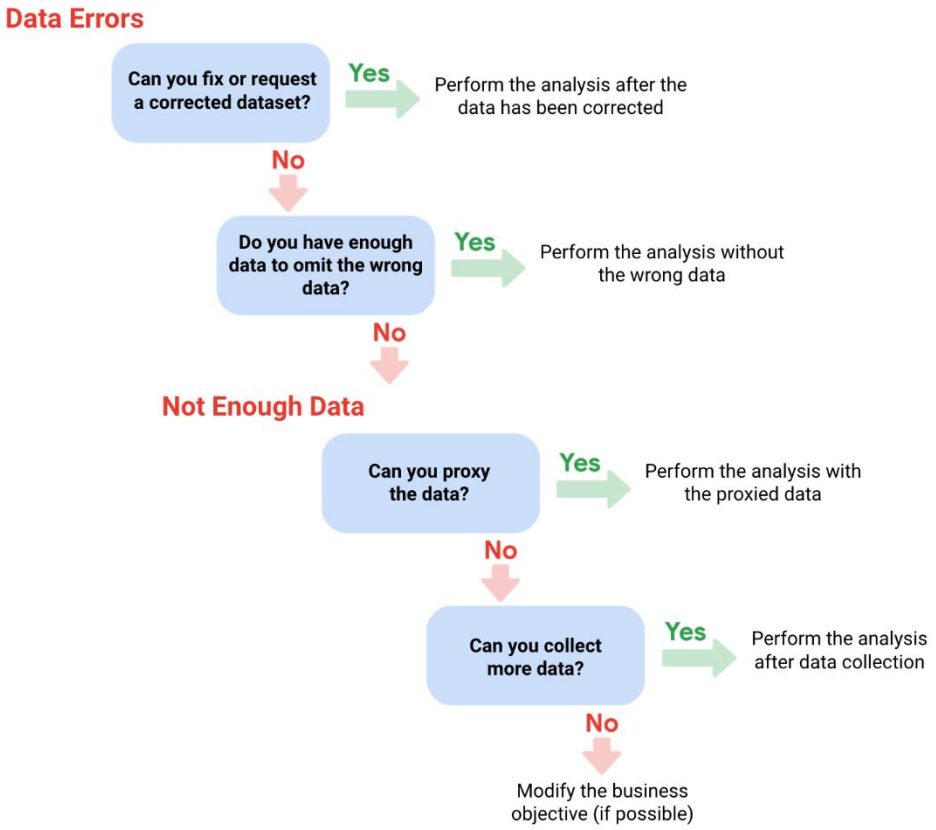
- Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data.
- If there isn't time to collect data, perform the analysis using **proxy data** from other datasets. This is the most common workaround.

Data issue 2: too little data

- Do the analysis using proxy data along with actual data.
- Adjust your analysis to align with the data you already have.

Data issue 3: wrong data, including data with errors

- If you have the wrong data because requirements were misunderstood, communicate the requirements again.
- Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors.
- 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.



Proxy data examples

Sometimes the data to support a business objective isn't readily available. This is when proxy data is useful. Take a look at the following scenarios and where proxy data comes in for each example:

Business scenario	How proxy data can be used
A new car model was just launched a few days ago and the auto dealership can't wait until the end of the month for sales data to come in. They want sales projections now.	The analyst proxies the number of clicks to the car specifications on the dealership's website as an estimate of potential sales at the dealership.
A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years.	The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years.
The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren't publicly available yet.	The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier.

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.

[Caluculating sample size]

Pre-cleaning activities help you determine and maintain data integrity and are important because they increase the efficiency and success of your data analysis tasks. One of the objectives of pre-cleaning activities is to address insufficient data. If you know that your data is accurate, consistent, and complete, you can be confident that your results will be valid. Stakeholders will be pleased if you connect the data to business objectives. And, knowing when to stop collecting data will allow you to finish your tasks in a timely manner without sacrificing data integrity. Data analysts perform pre-cleaning activities to complete these steps.

Statistical power

The probability of getting meaningful results from a test. Statistical power can be calculated and reported for a completed experiment to comment on the confidence one might have in the conclusions drawn from the results of the study. It can also be used as a tool to estimate the number of observations or sample size required in order to detect an effect in an experiment.

Hypothesis testing

A way to see if a survey or experiment has meaningful results.

Statistical power is usually shown as a value out of one. If a test is statistically significant, it means the 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.

Confidence level

The probability that your sample accurately reflects the greater population. Having a 99 percent confidence level is ideal but most industries hope for at least a 90 or 95 percent confidence level.

Estimated response rate: If you are running a survey of individuals, this is the percentage of people you expect will complete your survey out of those who received the survey.

A sample size calculator tells you how many people you need to interview (or things you need to test) to get results that represent the target population. To calculate sample size using an online calculator, it's necessary to input the confidence level, margin of error, and population size. Calculator. The calculated sample size is the minimum number to achieve what you input for confidence level and margin of

error. If you are working with a survey, you will also need to think about the estimated response rate to figure out how many surveys you will need to send out.

Margin of error

The maximum amount that the sample results are expected to differ from those of the actual population. The closer to zero the margin of error, the closer your results from your sample would match results from the overall population. The more people you include in your survey, the more likely your sample is representative of the entire population. Decreasing the confidence level would also have the same effect, but that would also make it less likely that your survey is accurate. To calculate margin of error, you need three things: population size, sample size, and confidence level. Calculator.

In order for an experiment to be statistically significant, the results should be real and not caused by random chance.

Week2-Sparkling-clean data

Dirty data and Clean data

Dirty data is data that is incomplete, incorrect, or irrelevant to the problem you're trying to solve. 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. This means they develop, maintain, and test databases, data processors and related systems.
Data warehousing specialists develop processes and procedures to effectively store and organize data. They make sure that data is available, secure, and backed up to prevent loss.

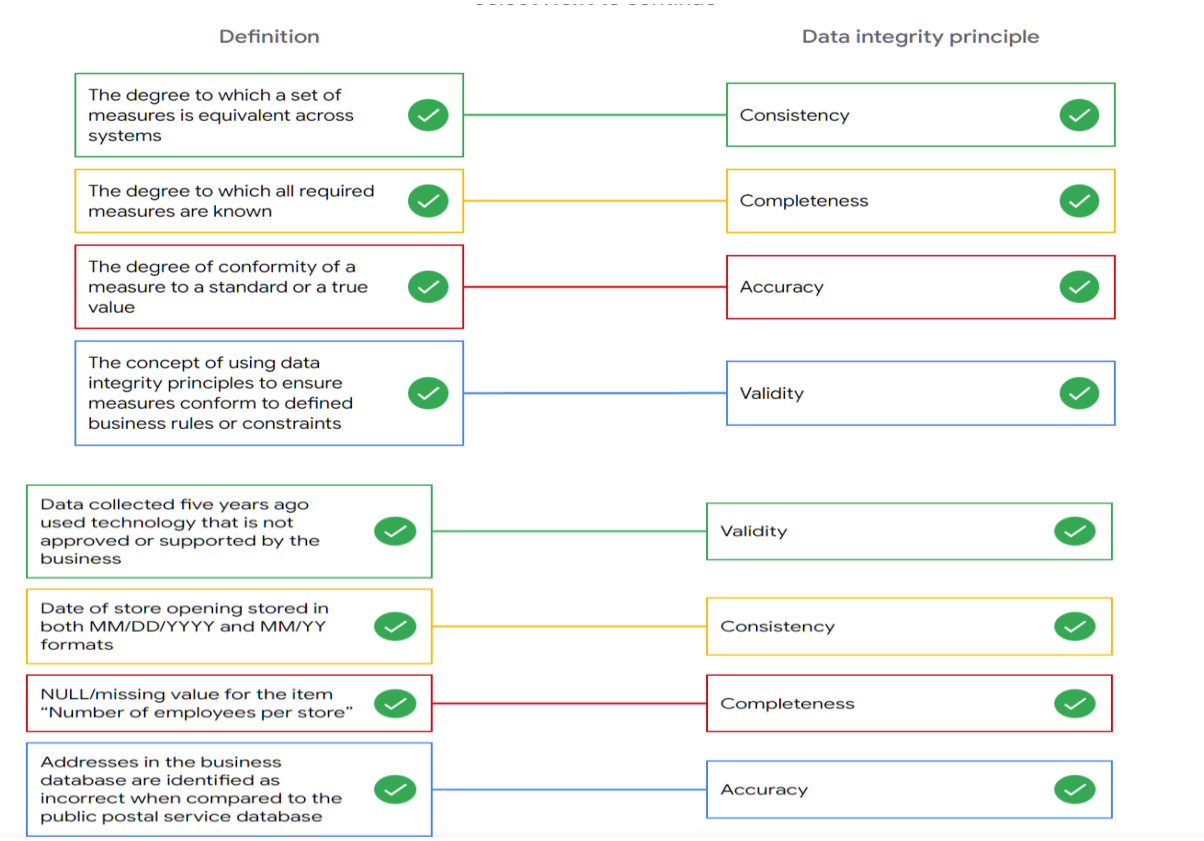
A **null** is an indication that a value does not exist in a data set.

Types of dirty data

Description	Possible Causes	Potential harm to businesses
Duplicate data: 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: 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: 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: 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: 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

A **field** is a single piece of information from a row or column of a spreadsheet. **Field length** is a tool for determining how many characters can be keyed into a field. Using the field length tool to specify the number of characters in each cell in the column could be part of data validation.

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



Data merging

The process of combining two or more datasets into a single dataset. This presents a unique challenge because when two totally different datasets are combined, the information is almost guaranteed to be inconsistent and misaligned. In data analytics, **compatibility** describes how well two or more datasets are able to work together.

Key questions to think about to avoid redundancy and to confirm that the datasets are compatible:

- 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 me to use?
- Are the datasets cleaned to the same standard?

Common data-cleaning pitfalls

- **Not checking for spelling errors**: Misspellings can be as simple as typing or input errors. Most of the time the wrong spelling or common grammatical errors can be detected, but it gets harder with things like names or addresses. For example, if you are working with a spreadsheet table of customer data, you might come across a customer named "John" whose name has been input incorrectly as "Jon" in some places. The spreadsheet's spellcheck probably won't flag this, so if you don't double-check for spelling errors and catch this, your analysis will have mistakes in it.
- **Forgetting to document errors**: Documenting your errors can be a big time saver, as it helps you avoid those errors in the future by showing you how you resolved them. For example, you might find an error in a formula in your spreadsheet. You discover that some of the dates in one of your columns haven't been formatted correctly. If you make a note of this fix, you can reference it the next time your formula is broken, and get a head start on troubleshooting. Documenting your errors also helps you keep track of changes in your work, so that you can backtrack if a fix didn't work.
- **Not checking for misfielded values**: A misfielded value happens when the values are entered into the wrong field. These values might still be formatted correctly, which makes them harder to catch if you aren't careful. For example, you might have a dataset with columns for cities and countries. These are the same type of data, so they are easy to mix up. But if you were trying to find all of the instances of Spain in the country column, and Spain had mistakenly been entered into the city column, you would miss key data points. Making

sure your data has been entered correctly is key to accurate, complete analysis.

- **Overlooking missing values**: Missing values in your dataset can create errors and give you inaccurate conclusions. For example, if you were trying to get the total number of sales from the last three months, but a week of transactions were missing, your calculations would be inaccurate. As a best practice, try to keep your data as clean as possible by maintaining completeness and consistency.
- **Only looking at a subset of the data**: It is important to think about all of the relevant data when you are cleaning. This helps make sure you understand the whole story the data is telling, and that you are paying attention to all possible errors. For example, if you are working with data about bird migration patterns from different sources, but you only clean one source, you might not realize that some of the data is being repeated. This will cause problems in your analysis later on. If you want to avoid common errors like duplicates, each field of your data requires equal attention.
- **Losing track of business objectives**: When you are cleaning data, you might make new and interesting discoveries about your dataset-- but you don't want those discoveries to distract you from the task at hand. For example, if you were working with weather data to find the average number of rainy days in your city, you might notice some interesting patterns about snowfall, too. That is really interesting, but it isn't related to the question you are trying to answer right now. Being curious is great! But try not to let it distract you from the task at hand.
- **Not fixing the source of the error**: Fixing the error itself is important. But if that error is actually part of a bigger problem, you need to find the source of the issue. Otherwise, you will have to keep fixing that same error over and over again. For example, imagine you have a team spreadsheet that tracks everyone's progress. The table keeps breaking because different people are entering different values. You can keep fixing all of these problems one by one, or you can set up your table to streamline data entry so everyone is on the same page. Addressing the source of the errors in your data will save you a lot of time in the long run.
- **Not analyzing the system prior to data cleaning**: If we want to clean our data and avoid future errors, we need to understand the root cause of your dirty data. Imagine you are an auto mechanic. You would find the cause of the problem before you started fixing the car, right? The same goes for data. First, you figure out where the errors come from. Maybe it is from a data entry error, not setting up a spell check, lack of formats, or from duplicates. Then, once you understand where bad data comes from, you can control it and keep your data clean.

- **Not backing up your data prior to data cleaning**: It is always good to be proactive and create your data backup before you start your data clean-up. If your program crashes, or if your changes cause a problem in your dataset, you can always go back to the saved version and restore it. The simple procedure of backing up your data can save you hours of work-- and most importantly, a headache.
- **Not accounting for data cleaning in your deadlines/process**: All good things take time, and that includes data cleaning. It is important to keep that in mind when going through your process and looking at your deadlines. When you set aside time for data cleaning, it helps you get a more accurate estimate for ETAs for stakeholders, and can help you know when to request an adjusted ETA.

Refer to these "top ten" lists for data cleaning in Microsoft Excel and Google Sheets to help you avoid the most common mistakes:

- Top ten ways to clean your data: Review an orderly guide to data cleaning in Microsoft Excel.
- 10 Google Workspace tips to clean up data: Learn best practices for data cleaning in Google Sheets.

Cleaning is a fundamental step in data science as it greatly increases the integrity of the data. If data analysis is based on bad or “dirty” data, it may be biased, erroneous, and uninformed. Good data science results rely heavily on the reliability of the data. Data analysts clean data to make it more accurate and reliable. This is important for making sure that the projects you will work on as a data analyst are completed properly.

- **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.
- In data analytics, a **text string** is a group of characters within a cell, commonly composed of letters, numbers or both. An important characteristic of a text string is its length, which is the number of characters in it. A **substring** is a smaller subset of a text string.
- **Split** is a tool that divides a text string around the specified character and puts each fragment into a new and separate cell.

- **COUNTIF** is a function that returns the number of cells that match a specified value. =COUNTIF(range, "value")
- **LEN** is a function that tells you the length of the text string by counting the number of characters it contains. =LEN(range)
- **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 multiple text strings into a single string. =CONCATENATE(item 1, item 2)
- **TRIM** is a function that removes leading, trailing, and repeated spaces in data. =TRIM(range)

Different methods that data analysts use to look at data differently and how that leads to more efficient and effective data cleaning: Some of these methods include sorting and filtering, pivot tables, a function called VLOOKUP, and plotting to find outliers.

- For data cleaning, you can use sorting to put things in alphabetical or numerical order, so you can easily find a piece of data. Sorting can also bring duplicate entries closer together for faster identification. Filters, on the other hand, are very useful in data cleaning when you want to find a particular piece of information.
- **Pivot tables** sort, reorganize, group, count, total or average data stored in the database. In data cleaning, pivot tables are used to give you a quick, clutter-free view of your data. You can choose to look at the specific parts of the data set that you need to get a visual in the form of a pivot table.
- **VLOOKUP** stands for vertical lookup. It's 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 up' !Range, column, false)
- When you plot data, you put it in a graph chart, table, or other visual to help you quickly find what it looks like. Plotting is very useful when trying to identify any skewed data or outliers.

Data mapping

Data mapping is the process of matching fields from one data source to another. Different systems store data in different ways. Data mapping helps us note these kinds of differences so we know when data is moved and combined it will be compatible.

1. The first step to data mapping is identifying what data needs to be moved. This includes the tables and the fields within them. We also need to define the desired format for the data once it reaches its destination.
2. Next comes mapping the data. Depending on the schema and number of primary and foreign keys in a data source, data mapping can be simple or very complex. A **schema** is a way of describing how something is organized. For more challenging projects there's all kinds of data mapping software programs you can use. These data mapping tools will analyze field by field how to move data from one place to another then they automatically clean, match, inspect, and validate the data. They also create consistent naming conventions, ensuring compatibility when the data is transferred from one source to another. When selecting a software program to map your data, you want to be sure that it supports the file types you're working with, such as Excel, SQL, Tableau, and others.
3. The next step is to transform the data into a consistent format.
4. Now that everything's compatible, it's time to transfer the data to it's destination. There's a lot of different ways to move data from one place to another, including querying, import wizards, and even simple drag and drop.
5. We would still want to make sure everything was transferred properly. We'll go into the testing phase of data mapping. For this, you inspect a sample piece of data to confirm that it's clean and properly formatted. It's also a smart practice to do spot checks on things such as the number of nulls. For the test, you can use a lot of the data cleaning tools such as data validation, conditional formatting, COUNTIF, sorting, and filtering.
6. Once you've determined that the data is clean and compatible, you can start using it for analysis.

Week3-Cleaning data with SQL

SQL can process large amounts of data much more quickly than spreadsheets.

Where the data lives will decide which tool you use. If you are working with data that is already in a spreadsheet, that is most likely where you will perform your analysis. And if you are working with data stored in a database, SQL will be the best tool for you to use for your analysis. SQL can handle huge amounts of data, can be adapted

and used with multiple database programs, and offers powerful tools for cleaning data. SQL is also a well-known standard in the professional community.

Data stored in a SQL database is useful to a project with multiple team members because they can access the data at the same time, use SQL to interact with the database program, and track changes to SQL queries across the team.

Structured Query Language, or SQL, is a language used to talk to databases. Learning SQL can be a lot like learning a new language — including the fact that languages usually have different dialects within them. Some database products have their own variant of SQL, and these different varieties of SQL dialects are what help you communicate with each database product. These dialects will be different from company to company and might change over time if the company moves to another database system. So, a lot of analysts start with Standard SQL and then adjust the dialect they use based on what database they are working with. Standard SQL works with a majority of databases and requires a small number of syntax changes to adapt to other dialects.

- We can use `SELECT` to specify exactly what data we want to interact with in a table. If we combine `SELECT` with `FROM`, we can pull data from any table in this database as long as they know what the columns and rows are named.
- We can also insert new data into a database or update existing data. We can use the `INSERT INTO` query to put that information in. We also want to specify which columns we're adding this data to by typing their names in the parentheses.
- If we want to create a new table for an updated database, we can use the `CREATE TABLE IF NOT EXISTS` statement. Just running a SQL query doesn't actually create a table for the data we extract. It just stores it in our local memory.
- If you're creating lots of tables within a database, you'll want to use the `DROP TABLE IF EXISTS` statement to clean up.
- `DELETE` removes data from a database.
- `UPDATE` changes existing data in a database.

- Including `DISTINCT` in your `SELECT` statement removes duplicates.
- In a query, if you use the `LENGTH()`, `SUBSTR()`, or `TRIM()` function in a WHERE clause, you can select data based on a string condition. `SUBSTR()` and `TRIM()` functions can be used to clean string variables. `LENGTH()` can be used in the general cleaning process to check if the data is as expected, but it does not actually clean strings.
  - If we already know the length our string variables are supposed to be, we can use `LENGTH(column)` to double-check that our string variables are consistent. For some databases, this query is written as `LEN(column)`, but it does the same thing.
  - `SUBSTR(column, starting position, number of letters including starting position)` is the substring function.
  - `TRIM(column)` function is really useful if you find entries with extra spaces and need to eliminate those extra spaces for consistency.

- `MIN(column)` and `MAX(column)` returns the minimum and maximum numerical values respectively in the specified column.
- `COUNT(*)` returns the number of rows.

- `CAST(column AS data_type)` can be used to convert anything from one data type to another.
- `ORDER BY` statement allows us to order rows in the specified column in descending or ascending order as specified in the statement.
- `CONCAT(column1, column2)` lets you add strings together to create new text strings that can be used as unique keys.
- `COALESCE(column to check first, column to check second if the first column is null)` can be used to return non-null values in a list. Null values are missing values.

Week4-Verify and report on your cleaning results

Verification

Verification is a process to confirm that a data cleaning effort was well- executed and the resulting data is accurate and reliable. It involves rechecking your clean dataset, doing some manual clean ups if needed, and taking a moment to sit back and really think about the original purpose of the project. That way, you can be confident that the data you collected is credible and appropriate for your purposes.

Reporting is a great opportunity to show stakeholders that you're accountable, build trust with your team, and make sure you're all on the same page of important project details. Different strategies for reporting include creating **data- cleaning reports**, **documenting your cleaning process**, and using **changelog**.

Changelog

A changelog is a file containing a chronologically ordered list of modifications made to a project. It's usually organized by version and includes the date followed by a list of added, improved, and removed features. Changelogs are very useful for keeping track of how a dataset evolved over the course of a project. They're also another great way to communicate and report on data to others. They can be referred to during the verification period if there are errors or questions.

- The first step in the verification process is going back to your original unclean data set and comparing it to what you have now, review the dirty data and try to identify any common problems.
- Another key part of verification involves taking a big-picture view of your project. This is an opportunity to confirm you're actually focusing on the business problem, that you need to solve, and the overall project goals; and to make sure that your data is actually capable of solving that problem and achieving those goals. Taking a big picture view of your project involves doing three things:
    i. Consider the business problem you're trying to solve with the data. Taking a problem-first approach to analytics is essential at all stages of any project.
    ii. Consider the goal of the project. On top of that, you also need to know whether the data you've collected and cleaned will actually help your company achieve that goal.
    iii. Consider whether your data is capable of solving the problem and meeting the project objectives.

- **COUNTA** counts the total number of values within a specified range. Note that there's also function called **COUNT**, which only counts the numerical values within a specified range.
- If you're working in SQL, you can address misspellings using a `CASE` statement. The `CASE` statement goes through one or more conditions and returns a value as soon as a condition is met. You should add a CASE statement as a `SELECT` clause. The typo would be a condition and the correction would be the returned value for the condition.

Documentation

The process of tracking changes, additions, deletions and errors involved in your data cleaning effort. Changelogs are good example of this, since it's staged chronologically, it provides a real-time account of every modification. Documenting data-cleaning makes it possible to be transparent about your process, keep team members on the same page, and demonstrate to project stakeholders that you are accountable.

Having a record of how a data set evolved does three very important things:

- Lets us recover data-cleaning errors (recalling the errors that were cleaned).
- Documentation gives you a way to inform other users of changes you've made.
- Documentation helps you to determine the quality of the data to be used in analysis.

Most software applications have a kind of history tracking built in. You can use and view a changelog in spreadsheets and SQL to achieve similar results.

- In the spreadsheet, we can use *Sheets Version History*, which provides a real-time tracker of all the changes and who made them from individual cells to the entire worksheet. If you want to check out changes in a specific cell, we can right-click and select *Show Edit History*.
- The way you create and view a changelog with SQL depends on the software program you're using.
    - In BigQuery, *Query History* tracks all the queries you've run. You can click on any of them to revert back to a previous version of your query or to bring up an older version to find what you've changed.

While your team can view changelogs directly, stakeholders can't and have to rely on your report to know what you did. There're plenty of ways we could go about documenting what we did, one common way is to just create a doc listing out the steps we took and the impact they had. If we were working with SQL, we could include a comment in the statement describing the reason for a change without affecting the execution of the statement.

Clean data is important to the task at hand. But the data-cleaning process itself can reveal insights that are helpful to a business. The feedback we get when we report on our cleaning can transform data collection processes, and ultimately business development. With consistent documentation and reporting, we can uncover error patterns in data collection and entry procedures and use the feedback we get to make sure common errors aren't repeated. Maybe we need to reprogram the way the data is collected or change specific questions on the survey form. In more extreme cases, the feedback we get can even send us back to the drawing board to rethink expectations and possibly update quality control procedures. For example, sometimes it's useful to schedule a meeting with a data engineer or data owner to make sure the data is brought in properly and doesn't require constant cleaning.

Some advanced functions that can help you speed up the data cleaning process in spreadsheets. Below is a table summarizing three functions and what they do:

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.