

# Data Analysis Course 4: Process Data from Dirty to Clean

## Module 1

A strong analysis depends on the integrity of the data.

data integrity: the accuracy, completeness, consistence, and trustworthiness of the data throughout its lifecycle.

Data is at risk of corruption every time it is:

- replicated
- transferred
- manipulated

data replication: process of storing data in multiple locations.

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

data manipulation: 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

<b>Data constraint</b>	<b>Definition</b>	<b>Examples</b>
<b>Data type</b>	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
<b>Data range</b>	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
<b>Mandatory</b>	Values can't be left blank or empty	If age is mandatory, that value must be filled in
<b>Unique</b>	Values can't have a duplicate	Two people can't have the same mobile phone number within the same service area
<b>Regular expression (regex) patterns</b>	Values must match a prescribed pattern	A phone number must match ###-###-#### (no other characters allowed)
<b>Cross-field validation</b>	Certain conditions for multiple fields must be satisfied	Values are percentages and values from multiple fields must add up to 100%
<b>Primary-key</b>	(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.

<b>Set-membership</b>	(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
<b>Foreign-key</b>	(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
<b>Accuracy</b>	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.
<b>Completeness</b>	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.
<b>Consistency</b>	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.

duplicate data  
not enough data

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.

VLOOKUP is a spreadsheet function that searches for a certain value in a column to return a related piece of information.

Deal with insufficient data

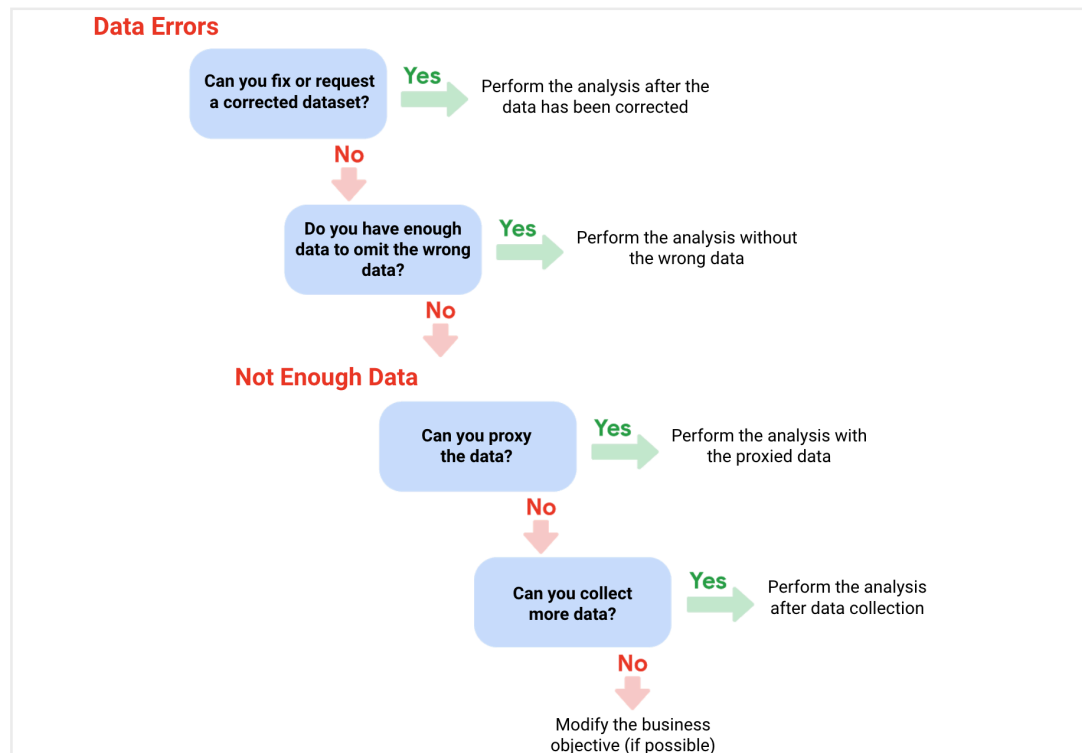
Types:

- 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



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

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.

This recommendation is based on the Central Limit Theorem (CLT) in the field of probability and statistics. As sample size increases, the results more closely resemble the normal (bell-shaped) distribution from a large number of samples. A sample of 30 is the smallest sample size for which the CLT is still valid. Researchers who rely on regression analysis – statistical methods to determine the relationships between controlled and dependent variables – also prefer a minimum sample of 30.

statistical power: probability of getting meaningful results from a test.

- usually represented as a value out of 1.
  - .6=60%: likelihood of getting a statistically significant result.
- the larger the sample size, the greater the statistical power.

hypothesis testing: a way to see if a survey or experiment has meaningful results.

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.

contraindication is a condition that may cause a patient not to take a vaccine due to the harm it would cause them if taken

confidence level: the probability that your sample size accurately reflects the greater population.

- 99% is ideal, but most industries hope for a least 90-95%.

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.

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

- helps you understand how reliable the data from your hypothesis testing is.
- want it to be closer to 0.

To calculate margin of error you need:

- population size
- sample size
- confidence level

A/B testing (or split testing) tests two variations of the same web page to determine which page is more successful in attracting user traffic and generating revenue.

conversion rate: user traffic that gets monetized

confidence interval: determined by the conversion rate and the margin of error

### **Terms and definitions for Course 4, Module 1**

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

Confidence interval: A range of values that conveys how likely a statistical

estimate reflects the population

Confidence level: The probability that a sample size accurately reflects the greater population

Consistency: The degree to which data is repeatable from different points of entry or collection

Cross-field validation: A process that ensures certain conditions for multiple data fields are satisfied

Data constraints: The criteria that determine whether a piece of a data is clean and valid

Data integrity: The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

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

Data range: Numerical values that fall between predefined maximum and minimum values

Data replication: The process of storing data in multiple locations

DATEDIF: A spreadsheet function that calculates the number of days, months, or years between two dates

Estimated response rate: The average number of people who typically complete a survey

Hypothesis testing: A process to determine if a survey or experiment has meaningful results

Mandatory: A data value that cannot be left blank or empty

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

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

Regular expression (Regex): A rule that says the values in a table must match a prescribed pattern

## **Module 2**

Human error is most common cause of poor quality data

Bad data costs companies \$3.1 trillion/year in U.S. alone according to IBM

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 to the problem you're trying to solve.

data engineers: transform data into a useful format for analysis and give it a reliable infrastructure.

- develop, maintain, and test databases, data processors and related systems.

data warehousing specialists: develop processes and procedures to effectively store and organize data.

- make sure that data is available, secure, and backed up to prevent loss.

data cleaning even more important with external data

Null: indication that a value does not exist in the dataset.

- empty fields

Types of dirty data:

- **duplicate data:** Any data record that shows up more than once
- **outdated data:** Any data that is old which should be replaced with newer and more accurate information
- **incomplete data:** Any data that is missing important fields
- **incorrect/inaccurate data:** Any data that is complete but inaccurate
- **inconsistent data:** Any data that uses different formats to represent the same thing

data integrity guidelines

- need rules in place to maintain uniformity

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

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

Principles of data integrity:

- **validity:** The concept of using data integrity principles to ensure measures conform to defined business rules or constraints
- **completeness:** The degree to which all required measures are known
- **consistency:** The degree to which a set of measures is equivalent across systems
- **accuracy:** The degree of conformity of a measure to a standard or a true value

Data cleaning tools & techniques:

- remove duplicates
- remove irrelevant data
- remove extra spaces and blanks
- fix misspellings

- fix inconsistent capitalization
- fix incorrect punctuation and other typos
- removing formatting

Make copy before removing unwanted data

merger: agreement that unites two organizations into single new one.

data merging: 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?
- Does the data need to be cleaned or are they ready for use?
  - Are they cleaned to the same standard?

Common data cleaning pitfalls

- Not checking for spelling errors
- Forgetting to document errors
- Not checking for misfielded values
  - When the values are entered into the wrong field
- Overlooking missing values
- Only looking at a subset of the data
- Losing track of 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 cleaning
- Not accounting for data cleaning in your deadlines/process

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

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

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

- delimiter: specified text separating each item

## **SPREADSHEET FUNCTIONS**

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

Concatenate: a function that joins multiple text strings into a single string

=CONCATENATE(item 1,item 2)



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

syntax: a predetermined structure that includes all required information and its proper placement.

LEN: a function that tells you the length of a text string by counting the number of characters it contains.  
=LEN(range)

LEFT: a function that gives you a set number of characters from the left side of text string  
=LEFT(range,number of characters)

RIGHT: a function that gives you a set number of characters from the right side of text string  
=RIGHT(range,number of characters)

MID: a function that gives you a segment from the middle of a text string.  
=MID(range,reference starting point,number of middle characters)

TRIM: a function that removes leading, trailing, and repeated spaces in data.  
=TRIM(range)

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

Sorting: arranging data in to a meaningful order to make it easier to understand, analyze, and visualize.

Filtering: showing only the data the meets a specific criteria while filtering the rest.

Pivot table: a data summarization tool that is used in data processing

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)  
– stands for "vertical lookup"

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

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

Schema: a way of describing how something is organized.

### Create a 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.
- **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.

### 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

### **Module 3: SQL**

Use it for big data sets

- Process faster than spreadsheets

SQL been around since 1986

Spreadsheets vs. SQL

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 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

```

SUBSTRING
SELECT
    customer_id
FROM
    `silent-complex-430821-f3.customer_data.customer_address`
WHERE
    SUBSTR(country, 1, 2) = 'US';

```

```

SELECT DISTINCT
    customer_id
FROM
    `silent-complex-430821-f3.customer_data.customer_address`
WHERE
    SUBSTR(country, 1, 2) = 'US';

```

```

TRIM: removes any spaces
SELECT
    state
FROM
    `silent-complex-430821-f3.customer_data.customer_address`
WHERE
    LENGTH(state) > 2;

```

```

SELECT
    DISTINCT customer_id
FROM
    `silent-complex-430821-f3.customer_data.customer_address`
WHERE
    TRIM(state) = 'OH';

```

## **HANDS ON ACTIVITY: CLEANING W/ SQL**

### **FILL IN MISSING DATA**

```
SELECT
    *
FROM
    silent-complex-430821-f3.cars.car_info
WHERE
    num_of_doors IS NULL;
```

```
UPDATE
    silent-complex-430821-f3.cars.car_info
SET
    num_of_doors = "four"
WHERE
    make = "dodge"
    AND fuel_type = "gas"
    AND body_style = "sedan";
```

```
UPDATE
    silent-complex-430821-f3.cars.car_info
SET
    num_of_doors = "four"
WHERE
    make = "mazda"
    AND fuel_type = "diesel"
    AND body_style = "sedan";
```

### **ID POTENTIAL ERRORS**

```
SELECT
    DISTINCT num_of_cylinders
FROM
    silent-complex-430821-f3.cars.car_info;
```

```
UPDATE
    silent-complex-430821-f3.cars.car_info
SET
    num_of_cylinders = "two"
WHERE
    num_of_cylinders = "tow";
```

```
SELECT
    MIN(compression_ratio) AS min_compression_ratio,
```

```
MAX(compression_ratio) AS max_compression_ratio
FROM
    silent-complex-430821-f3.cars.car_info;
```

```
SELECT
    MIN(compression_ratio) AS min_compression_ratio,
    MAX(compression_ratio) AS max_compression_ratio
FROM
    silent-complex-430821-f3.cars.car_info;
WHERE
    compression_ratio <> 70;
```

```
SELECT
    COUNT(*) AS num_of_rows_to_delete
FROM
    silent-complex-430821-f3.cars.car_info;
WHERE
    compression_ratio = 70;
```

```
DELETE silent-complex-430821-f3.cars.car_info;
WHERE compression_ratio = 70;
```

### **ENSURE CONSISTENCY**

```
SELECT
    DISTINCT drive_wheels,
    LENGTH(drive_wheels) AS string_length
FROM
    silent-complex-430821-f3.cars.car_info;
```

```
UPDATE
    silent-complex-430821-f3.cars.car_info
SET
    drive_wheels = TRIM(drive_wheels)
WHERE TRUE;
```

```
SELECT
    DISTINCT drive_wheels
FROM
    silent-complex-430821-f3.cars.car_info;
```

### **ADVANCED DATA CLEANING FUNCTIONS**

CAST(): convert data from one datatype to another.

```
SELECT
    CAST(purchase_price AS FLOAT64)
```

```
FROM
  silent-complex-430821-f3.customer_data2.furniture_transactions
ORDER BY
  CAST(purchase_price AS FLOAT64) DESC;
```

—this corrected the column from being read as STRING to FLOAT so that we could properly order the purchase prices

Typecasting: converting data from one type to another

Float: a number that contains a decimal

```
CAST (again)
SELECT
  date,
  purchase_price
FROM
  silent-complex-430821-f3.customer_data2.furniture_transactions
WHERE
  date BETWEEN '2020-12-01' AND '2020-12-31';
```

—looks kinda weird because it's showing as date time, not just date. want to convert.

```
SELECT
  CAST(date AS date) AS date_only,
  purchase_price
FROM
  silent-complex-430821-f3.customer_data2.furniture_transactions
WHERE
  date BETWEEN '2020-12-01' AND '2020-12-31';
```

—now it just shows date data type

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

```
SELECT
  CONCAT(product_code, product_color) AS new_product_code,
FROM
  silent-complex-430821-f3.customer_data2.furniture_transactions
WHERE
  product = 'couch'
ORDER BY
  purchase_size DESC;
```

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

SELECT

COALESCE(product, product\_code) AS product\_info

FROM

silent-complex-430821-f3.customer\_data2.furniture\_transactions;

### **Terms and definitions for Course 4, Module 3**

CAST: A SQL function that converts data from one datatype to another

COALESCE: A SQL function that returns non-null values in a list

CONCAT: A SQL function that adds strings together to create new text strings that can be used as unique keys

DISTINCT: A keyword that is added to a SQL SELECT statement to retrieve only non-duplicate entries

Float: A number that contains a decimal

Substring: A subset of a text string

Typecasting: Converting data from one type to another

### **Module 4**

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

- a stamp of approval
- manually cleaning data

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

Step 1: go back to original dataset and compare it to what you have now.

Step 2: take a big picture view of the project

- Three Steps:
  - consider the business problem you are trying to solve
  - consider the goal of the project
  - consider whether your data is capable of solving the problem and meeting the project objectives

Pivot table: a data summarization tool that is used in data processing

- they can sort, reorganize, group, count, total or average data stored in a database.

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

- SQL function

SELECT

customer\_id

CASE

WHEN first\_name = 'Tnoy' THEN 'Tony'

WHEN first\_name = 'Tmo' THEN 'Tom'

WHEN first\_name = 'Rachle' THEN 'Rachel'

ELSE first\_name

END AS cleaned\_name

FROM

customer\_date.customer\_name;

### **Data-Cleaning Verification Checklist**

Correct the most common problems in data

- 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?

Review the goal of your project

- Confirm the business problem
- Confirm the goal of the project
- Verify that data can solve the problem and is aligned to the goal

Documentation: the process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort.

- errors: crime
- cleaning: gathering evidence
- documentation: detailing exactly what happened for peer review/court

Record of how dataset evolved:

- let's us recover data cleaning errors
  - create clean table rather than overwriting existing table
- gives you a way of informing other users of changes you've made
- helps you determine the quality of the data to be used for analysis

Automated Version Control

Sheets

File > See Version History in Sheets

Right click > show edit history in particular cell

Excel

Use "Track Changes"

SQL

specify what you did and why

add comments

Query history in BigQuery

Changelogs

give an even more detailed record than automated version control

Version histories record what was done in a data change for a project, but don't tell us why. Changelogs help understand reasons for changes.

- No set formal, unless your company prescribes one.

Typically, a changelog records:

- Data, file, formula, query, or 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

#### Version control system

- syncing: ensuring that the most up-to-date version of the query that is being changed.
- code review: asking someone to review a coding change.
- code commit: submitting the updated version of the query to a repository in a company's version control system.

A changelog for a personal project may take any form desired. However, in a professional setting and while collaborating with others, readability is important. These guiding principles help to make a changelog accessible to others:

- Changelogs are for humans, not machines, so write legibly.
- Every version should have its own entry.
- Each change should have its own line.
- Group the same types of changes. For example, Fixed should be grouped separately from Added.
- Versions should be ordered chronologically starting with the latest.
- The release date of each version should be noted.

All the changes for each category should be grouped together. Types of changes usually fall into one of the following categories:

- 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

Markdown is a style of writing common to keep changelogs as a readme file in code repository. (<https://docs.github.com/en/get-started/writing-on-github/getting-started-with-writing-and-formatting-on-github/basic-writing-and-formatting-syntax>)

#### Common data errors

- human error in data entry
- flawed processes
- system issues

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.

### Terms and definitions for Course 4, Module 4

CASE: A SQL statement that returns records that meet conditions by including an if/then statement in a query

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

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

Find and replace: A tool that finds a specified search term and replaces it with something else

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

### Module 5

talk about analyzing previous spending to develop budgets

Communication skills

Include quantitative data

PAR







Problem

Action

Result

Teamwork

Problem-solving

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

Junior or associate data analyst most applicable to my certificate

Healthcare analyst

Marketing analyst

Business intelligence analyst

Financial analyst