

COURSE 4: PROCESS

WEEK 1

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

There's a chance data can be compromised (**decrease of data integrity**) every time it's **replicated**, **transferred**, or **manipulated** in any way.

Data replication is 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 (**changes at one place or device does not change accordingly and simultaneously on other devices**). This data lacks integrity because different people might not be using the same data for their findings, which can cause inconsistencies.

Data replication, data transfer, data manipulation, human error, viruses, malware, hacking, and system-failures can cause loss of data integrity.

With incomplete data, it's hard to see the whole picture to get a real sense of what is going on.

Learning how to deal with data issues while staying focused on your objective will help set you up for success in your career as a data analyst.

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 colour, 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.

Clean data + alignment to business objective = accurate conclusions

(Alignment refers to the related, relevant data to the 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.

INSUFFICIENT DATA

Types of insufficient data

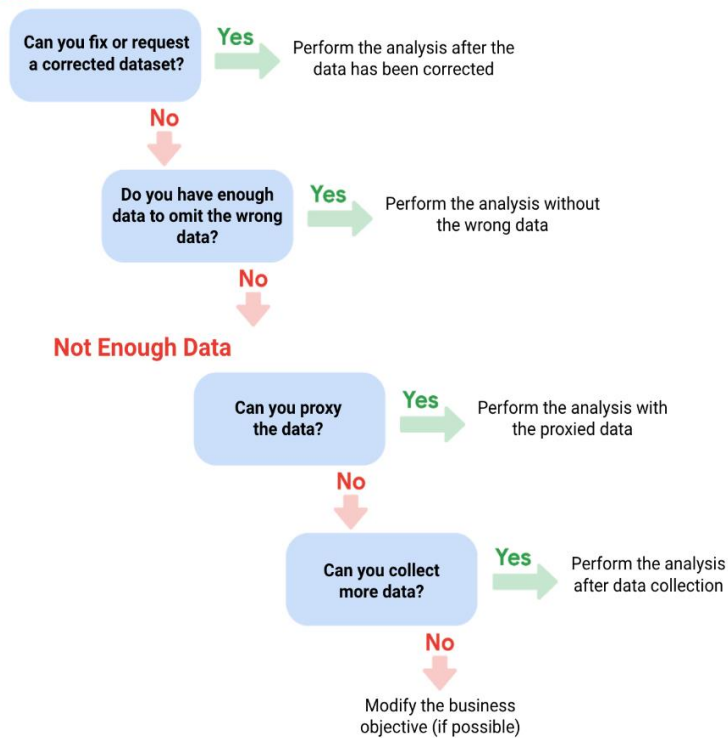
- Data from only one source
- Data that keeps updating
- Outdated data
- Geographically-limited data

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

Decision tree as a reminder of how to deal with data errors or not enough data:

Data Errors



Population and samples.

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

The goal is to get enough information from a small group within a population to make predictions or conclusions about the whole population. The sample size helps ensure the degree to which you can be confident that your conclusions accurately represent the population.

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.

- 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

Sample sizes vary by business problem.

Larger sample sizes have a higher cost.

You should complete the following tasks before **analysing data**:

1. Determine data integrity by assessing the overall accuracy, consistency, and completeness of the data.
2. Connect objectives to data by understanding how your business objectives can be served by an investigation into the data.
3. Know when to stop collecting data.

Data analysts perform pre-cleaning activities to complete these steps. Pre-cleaning activities help you determine and maintain **data integrity**, which is essential to the role of a junior data analyst.

Some focus on statistics, sample size, population, confidence level and more.

Statistical power

The probability of getting meaningful results from a test

If a test is statistically significant, it means the results of the test are real and not an error caused by random chance

Usually, the larger the sample size, the greater the chance you'll have statistically significant results with your test.

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.

Sample size calculator

You need to input the **confidence level, population size, and margin of error** to get the sample size through the calculator.

The confidence level is 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. Industries like pharmaceuticals usually want a confidence level that's as high as possible when they are using a sample size. This makes sense because they're testing medicines and need to be sure they work and are safe for everyone to use.

The confidence level and margin of error don't have to add up to 100 percent. They're independent of each other.

- **Confidence level:** The probability that your sample size accurately reflects the greater population.
- **Margin of error:** The maximum amount that the sample results are expected to differ from those of the actual population.
- **Population:** This is the total number you hope to pull your sample from.
- **Sample:** A part of a population that is representative of the population.
- **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.

Sample size calculator

https://docs.google.com/spreadsheets/d/1kBTvnpH2qOLJx4XWjUG1v-GF4LPmOhequy_9VRyslJ8/template/preview

<https://www.surveymonkey.com/mp/sample-size-calculator/>

Margin of error

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

Based on the sample size, the resulting margin of error will tell us how different the results might be compared to the results if we had surveyed the entire population.

Margin of error helps you understand how reliable the data from your hypothesis testing is. The closer to zero the margin of error, the closer your results from your sample would match results from the overall population.

More technically, the margin of error defines a range of values below and above the average result for the sample. The average result for the entire population is expected to be within that range. We can better understand margin of error by using some examples below.

Calculating the margin of error is particularly helpful when you are given the data to analyse.

Margin of error calculator

https://www.google.com/url?q=https://docs.google.com/spreadsheets/d/1gdhfyA3_vMnQ1cDaGSCshXd5ezLtVPfLhxc9STGq6B8/template/preview&sa=D&source=editors&ust=1625603636025000&usg=AOvVaw0BLkCfYvEulQnuysANCgzO

<https://goodcalculators.com/margin-of-error-calculator/>

week 2

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. 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. **Note that it's not the same as a zero.**

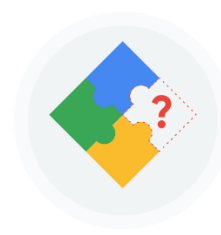
TYPES OF DIRTY DATA



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

Clean data depends largely on the data integrity rules that an organization follows, such as spelling and punctuation guidelines. Data validation is a tool for checking the accuracy and quality of data before adding or importing it.

These include spelling and other text errors, inconsistent labels, formats and field length, missing data, and duplicates.



clean data by removing duplicates, inconsistent data, extra spaces, fixing misspellings, inconsistent capitalization, incorrect punctuation, and other typos.

Data merging is always done in data analysis but the schemas of the different datasets is different and it comes with its own challenges.

Data merging

The process of combining two or more datasets into a single dataset

Compatibility

How well two or more datasets are able to work together

Always check for these errors:



reference link for description

<https://www.coursera.org/learn/process-data/supplement/m3iWu/common-data-cleaning-pitfalls>

Top 10 ways to clean your data my Microsoft.

<https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19>

Cleaning data in spreadsheet by

1. removing the blank spaces by filter method and selecting the remove blank spaces.
2. **transpose** the data from long format to wide format and wide format to long format.
3. remove the extra blank spaces in the texts and the numbers using the **TRIM** command in MS excel.
4. changing the case of the string data using the **UPPER, LOWER** etc. commands.
5. clear formatting using the clear option.

DETAILED CLEANING OF DATA IN MS EXCEL.

1. CONDITIONAL FORMATTING

Conditional formatting

A spreadsheet tool that changes how cells appear when values meet specific conditions

if blank cells or other go to conditional formatting and add new rules accordingly.

2. removing duplicates.
3. splitting text strings using the SPLIT TEXT TO COLUMNS.
4. using the **LEN, LEFT, RIGHT, MID, COUNTIF, TRIM, SPLIT, CONCATINATE** functions in spreadsheets.

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

Using **sorting filtering, pivot tables, VLOOKUP, plotting data** to clean the data.

Sorting involves arranging data into a meaningful order to make it easier to understand, analyse, and visualize. Sorting can also bring duplicate entries closer together for faster identification.

Filtering means showing only the data that meets a specific criterion while hiding the rest.

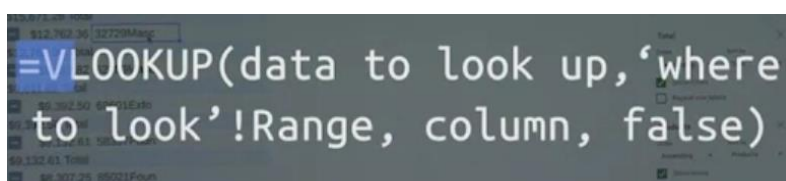
Pivot table is a data summarization tool that is used in data processing.

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

A function that searches for a certain value in a column to return a corresponding piece of information



Data mapping

Data mapping

The process of matching fields from one data source to another

This is very important to the success of data migration, data integration, and lots of other data management activities.

TO transpose the data from the long format to wide format first select the data and copy the data and now in a blank cell right click and select the special paste option and choose the transpose option.

Week 3

Using SQL to clean data

Relational database

A database that contains a series of tables that can be connected to form relationships

SQL can process large amounts of data much more quickly than spreadsheets. This is one of the reasons why data analysts use SQL when working with vast, complex datasets.

Spreadsheet 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

When it comes down to it, **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.

Spreadsheets	SQL
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
Small datasets	Larger datasets
Working independently	Tracks changes across team
Built-in functionalities	Useful across multiple programs

SQL also records changes in queries, which makes it easy to track changes across your team if you're working collaboratively.

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.

HOW DATA IS MEASURED

Data is measured by the number of **bits** it takes to represent it.

Unit	Equivalent to	Abbreviation	Real-World Example
Byte	8 bits	B	1 character in a string
Kilobyte	1024 bytes	KB	A page of text (~4 kilobytes)
Megabyte	1024 Kilobytes	MB	1 song in MP3 format (~2-3 megabytes)
Gigabyte	1024 Megabytes	GB	~300 songs in MP3 format
Terabyte	1024 Gigabytes	TB	~500 hours of HD video
Petabyte	1024 Terabytes	PB	10 billion Facebook photos
Exabyte	1024 Petabytes	EB	~500 million hours of HD video
Zettabyte	1024 Exabytes	ZB	All the data on the internet in 2019 (~4.5 ZB)

USING SQL QUERIES TO CLEAN DATA

DISTINCT for getting distinct values rather than duplicate values in the select query if duplicate exists.

LENGTH (column_name) for getting the length of the strings.

SUBSTRING(column_name,from_where_to_start,how_many_characters) to get a substring from a string value.

TRIM (column_name) for removing extra spaces in the string values .

MAX (column_name) for finding the maximum numeric value in the given column.

MIN (column_name) for finding the minimum numeric value in the given column.

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.

WHERE (column_name) IS NULL for finding the rows containing NULL values or empty values.

CAST(column_name, to_which_datatype_u_want_to_convert)

this function converts the datatype of the selected field to the given datatype.

```
select cast(date as date) as date_only, purchase_price
from my-project-0242-362611.prac_data.leet_code
where date BETWEEN '2020-12-1' AND '2020-12-31'
```

the **between** clause gives the dates between 1 December 2020 to 31 December 2020.

CONCAT(column1_name, column2_name, column_n_name)

to concatenate two or more strings and create new unique values.

```
select concat(product_code, product_color) as combo_code, product_color
from my-project-0242-362611.prac_data.leet_code
where product = 'couch'
```

COALESCE()

Can be used to return non-null values in a list

COALESCE (first_column_to_check,second_column_to_check..)

this function checks for values in the first column provided then if there are NULL values in the first column it will give the values from the second column provided.

CASE statement

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

```
1 SELECT
2     customer_id,
3     CASE
4         WHEN first_name = 'Tnoy' THEN 'Tony'
5         WHEN first_name = 'Tmo' THEN 'Tom'
6         WHEN first_name = 'Rachle' THEN 'Rachel'
7         ELSE first_name
8     END AS cleaned_name
9 FROM
10    customer_data.customer_name
```

generally the case function will check for SPELLING ERRORS in the column specified for the GIVEN CONDITION with the correct spelling and return the corrected data if incorrect is found.

Definition

Query

Return a limited number of characters to create substrings from longer strings of text



SUBSTR()



Return the length of a string of text by counting the number of characters it contains



LENGTH()/LEN()



Pull data from any table in a database



SELECT FROM



Pull data from a specific place in a table, typically a table column



SELECT FROM WHERE



Remove leading, trailing, and repeated spaces in data



TRIM()



Change existing data in a database



UPDATE



Remove data from a database



DELETE



Add strings together to create new text strings that can be used as unique keys



CONCAT()



Add new data into a database



INSERT INTO



Return non-null values in a list



COALESCE()



Convert data from one datatype to another



CAST()



WEEK 4

Verifying the data after cleaning it.

Verification

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

Changelog

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

Different steps to verify your clean data.

1. Comparing your old unclean data with the cleaned data.
2. 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.

It includes

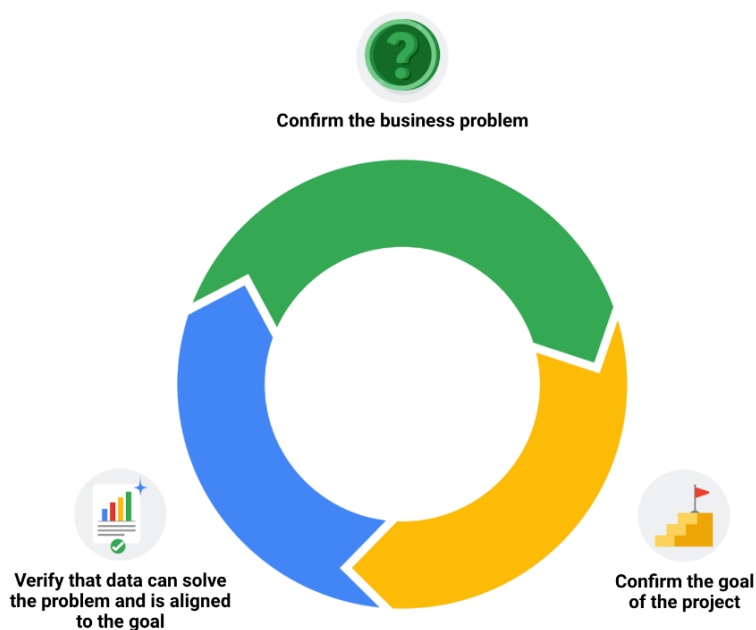
First, consider the business problem you're trying to solve with the data.

Second, you need to consider the goal of the project.

And third, you need to consider whether your data is capable of solving the problem and meeting the project objectives.

Correct the most common problems

- **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?



Documenting results and the cleaning process

Documentation

The process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort

First two benefits of documentation –

- 1) recalling the errors that were cleaned and
- 2) informing others of the changes -- assume that the data errors *aren't* fixable. She then added that when the data errors are fixable, the documentation needs to record how the data was fixed. Data-cleaning documentation is important in both cases.

CHANGE LOGS IN SPREADSHEETS

use the feature of version history in spreadsheets to check the changes made in entire spreadsheet as well as the individual cells.

CHANGE LOGS IN DATABASE

the change logs in database changes as per the software program we are using , usually there are **version history** available to restore to the previous version.

also add comments to check for the changes done regularly.

Typically, a changelog records this type of information:

- 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.

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

Advanced functions for speedy data cleaning

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.

using the copy link feature in MS excel to automatically update the values in the other sheet as the values changes in the present sheet.

ALL THE ABOVE FUCNTION THAT WE HAVE USED IN THE MS EXCEL CAN BE PERFORMED AUTOMATICALLY USING THE EXCEL **POWER QUERY OPTION** . to use select the table or range and go data tab and select the from table option and do the necessary things.

lastly using the filter option to for different filtering.

The **FILTER** function might run faster than the **QUERY** function. But keep in mind, the **QUERY** function can be combined with other functions for more complex calculations. For example, the **QUERY** function can be used with other functions like **SUM** and **COUNT** to summarize data, but the **FILTER** function can't.

WEEK 5

job search resume building and more....

<https://www.coursera.org/learn/process-data/supplement/DQgYG/careercon-resources-on-youtube>

(a link for Kaggle's job assistance program)

some of the skills from this course will be..

- Strong analytical skills
- Pattern recognition
- Relational databases and SQL
- Strong data visualization skills
- Proficiency with spreadsheets, SQL, R, and Tableau

The most important skills for data analyst









1. Structured Query Language (SQL): SQL is considered a basic skill that is pivotal to any entry-level data analyst position. SQL helps you communicate with databases, and more specifically, it is designed to help you retrieve information from databases. Every month, thousands of data analyst jobs posted require SQL, and knowing how to use SQL remains one of the most common job functions of a data analyst.

2. Spreadsheets: Although **SQL is popular, 62% of companies still prefer to use** spreadsheets for their data insights. When getting your first job as a data analyst, the first version of your database might be in spreadsheet form, which is still a powerful tool for reporting or even presenting data sets. So, it is important for you to be familiar with using spreadsheets for your data insights.

3. Data visualization tools: Data visualization tools help to simplify complex data and enable the data to be visually understood. After gathering and analysing data, data analysts are tasked with presenting their findings and making that information simple to grasp. Common tools that are used in data analysis include Tableau, MicroStrategy, Data Studio, Looker, Datarama, Microsoft Power BI, and many more. Among these, Tableau is best known for its ease of use, so it is a must-have for beginner data analysts. Also, studies show that data analysis jobs requiring Tableau are expected to grow **about 34.9% over the next decade.**

4. R or Python programming: Since only less than a third of entry-level data analyst positions require knowledge of Python or R, you don't need to be proficient in programming languages as an entry-level data analyst. But, R or Python are great additions to have as you become more advanced in your career.

SOME OF THE SOFT SKILLS INCLUDE

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