Week 1: Get to know about DA

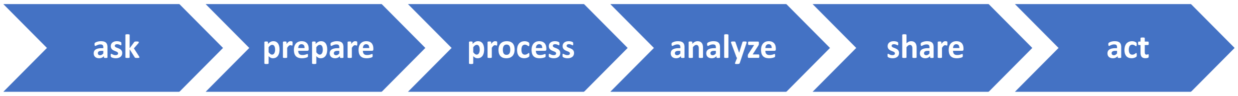
1. Definition:

Data Analysis: The collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making.

Data Analytics: the science of data

DA helps organizations completely rethink sth they do or point out another direction.

2. Six steps of the data analysis process



ask → prepare → process → analyze → share → act

(One of the ways of data analysis)

Theory	*Case study:* An organization was experiencing **a high turnover rate** among new hires. Many employees left the company before the end of their first year on the job. The analysts used the data analysis process to answer the following question: **how can the organization improve the retention rate for new employees?**
Step 1: Ask • ***Define the problem*** you're trying to solve • Make sure you fully understand the stakeholder's expectations • Focus on the actual problem and avoid any distractions • Collaborate with stakeholders and keep an open line of communication • Take a step back and see the whole situation in context *Questions to ask yourself in this step:* • What topic are you exploring?	• **Asked** effective questions and collaborated with leaders and managers (who were interested in the outcome of their people analysis) ○ What do you think new employees need to learn to succeed in their first job year? ○ Have you gathered data from new employees before? If so, may we have access to the **historical data**? ○ Do you believe managers with higher retention rates offer new employees something **extra or unique**? ○ What do you suspect is a **leading cause** of dissatisfaction among new employees? ○ **By what percentage** would you like employee retention to **increase** in the next fiscal year?

• What is the problem you are trying to solve? • What metrics will you use to measure your data to achieve your objective? Who are the stakeholders? • Who is your audience for this analysis and how does this affect your analysis process and presentation? • How will this data help your stakeholders make decisions? *Key tasks* • Choose a case study • Identify the problem • Determine key stakeholders • Explore the data and establish metrics	
Step 2: Preparation • Collect data and store it appropriately • Identify how it’s organized • Sort and filter the data • Determine the credibility of the data *Questions to ask yourself in this step:* • Where is your data located? • How is the data organized? • Are there issues with bias or credibility in this data? Does your data ROCCC? • How are you addressing licensing, privacy, security, and accessibility? • How did you verify the data’s integrity? • How does it help you answer your question? • Are there any problems with the data?	• **Built a timeline** • Identified data needed to achieve the successful result (in the previous step) In this case, the analysts chose to **gather the data** from an online survey of new employees. • Prepare: ○ They *developed specific questions* ○ To ask about employee satisfaction with different business processes, such as hiring and onboarding, and their overall compensation. ○ They *established rules* for who would have *access* to the data collected ○ In this case, anyone outside the group wouldn't have access to the raw data, but could view summarized or aggregated data. ○ For example, an individual's compensation wouldn't be available, but salary ranges for groups of individuals would be viewable. ○ They *finalized* what specific information would be gathered, and how best to *present the data visually*.

	The analysts *brainstormed possible project- and data-related issues* and how to avoid them.
Step 3: Process • Check the data for errors • Transform the data into the right type • Document the cleaning process • Choose your tools *Questions to ask yourself in this step:* What **data errors** or **inaccuracies might get in** my way of getting the best possible answer to the problem I am trying to solve? How can I clean my data so the information I have is more consistent? *What tools are you choosing and why?* *Have you ensured your data's integrity?* *What steps have you taken to ensure that your data is clean?* *How can you verify that your data is clean and ready to analyze?* *Have you documented your cleaning process so you can review and share those results?*	• They *restricted access* to the data to a *limited number of analysts.* • They *cleaned the data* to make sure it was *complete, correct, and relevant*. Certain data was aggregated and summarized without revealing individual responses. • They *uploaded raw data to an internal data warehouse* for an additional layer of security. ⇨ Data is complete, correct, relevant, and free of errors and outliers.
Step 4: Analyse At this stage, you might ***sort and format your data*** to make it easier to: Combine data from multiple sources Create tables with your results Organize and format your data Perform calculations Identify trends and relationships *Questions to ask yourself in this step:* What story is my data telling me? How will my data help me solve this problem? Who needs my company's product or service? What type of person is most likely to use it?	From the completed survey, they find: • Employees who experienced a long and complicated hiring process were most likely to leave the company. • Employees who experienced an efficient and transparent evaluation and feedback process were most likely to remain with the company. ⇨ **Document** exactly what they found in the analysis, no matter what the results.

How should you organize your data to perform analysis on it? Has your data been properly formatted? What surprises did you discover in the data? What trends or relationships have you found in the data? How do these insights answer your question or solve the problem?	
Step 5: Share • Determine the best way to share your findings • Create effective data visualizations • Present your findings • Ensure your work is accessible to your audience *Questions to ask yourself in this step:* How can I make what I present to the stakeholders engaging and easy to understand? What would help me understand this if I were the listener? What story does your data tell? How do your findings relate to your original question? Who is your audience? What is the best way to communicate with them? Can data visualization help you share your findings? Is your presentation accessible to your audience?	**Sharing the report carefully**: • They shared the report with managers who met or exceeded the minimum number of direct reports with submitted responses to the survey. • They presented the results to the managers to make sure they had the full picture. • They asked the managers to personally deliver the results to their teams. This process gave managers an opportunity to **communicate the results** with the right context. As a result, they could have productive team conversations about next steps to improve employee engagement.
Step 6: Act	**Step 6: Act**

• Share next steps with your stakeholders • Determine if more data could give you new insights • Upload to your portfolio *Questions to ask yourself in this step:* How can I use the feedback I received during the share phase (step 5) to actually meet the stakeholder's needs and expectations? What is your final conclusion based on your analysis? How can you apply your insights? Are there any next steps you or your stakeholders can take based on your findings? Is there additional data you could use to expand on your findings? How can you feature your case study in your portfolio?	work with leaders within their company and decide how best to **implement changes and take actions** based on the findings. These were their recommendations: • Standardize the hiring and evaluation process for employees based on the most efficient and transparent practices. • Conduct the same survey annually and compare results with those from the previous year. A year later, the same survey was distributed to employees. Analysts anticipated that a comparison between the two sets of results would indicate that the action plan worked. Turns out, the changes improved the retention rate for new employees and the actions taken by leaders were successful!

Additional Resource

To learn more about some recent applications of data analytics in the business world, check out the article "4 Examples of Business Analytics in Action" from Harvard Business School. The article reveals how corporations use data insights to optimize their decision-making process. Please note that the first example in the article contains a minor error in the second paragraph, but the example is still a valid one.

Correction to article in bold below: Microsoft's Workplace Analytics team hypothesized that moving the 1,200-person group from five buildings to four could improve collaboration by **increasing** the number of employees per building and by reducing the distance that staff needed to travel for meetings.

3. Data ecosystems

3.1. Definition

- Data ecosystems: The various elements that interact with one another -> produce, manage, store, organize, analyze, and share data.

Include: hardware and software tools

- Cloud: a place to keep data online
- **Dataset:** A collection of data that can be manipulated or analyzed as one unit

3.2. How data informs better decisions

- Data-driven decision-making: Using facts to guide business strategy
- Data alone will never be as powerful as data combined with human experience, observation, and sometimes even intuition
- a data analyst finds data, analyzes it, and uses it to uncover trends, patterns and relationships. (the past, future)
- It's important to include insights from people who are familiar with the business problem. These people are called **subject matter experts**, and they have the **ability to look at the results of data analysis** and identify any inconsistencies, make sense of gray areas, and eventually validate choices being made

3.3. Data and guts distinct

- Gut instinct: an intuitive understanding of something with little or no explanation. (giác quan thứ 6)
- It's essential that data analysts focus on the data to ensure they make informed decisions. Should not ignore data (biased). Guts distinct can make mistakes.
  => to figure out – use past experience and business knowledge (maybe just a touch of gut instinct)
- To find the perfect balance (way -to-do), try asking yourself:
  - What kind of results are needed?
  - Who will be informed?
  - Am I answering the question being asked?
  - How quickly does a decision need to be made?

  For instance: If rush -> rely more on knowledge and experience. If not, more data-driven.

Week 2

1. Data skills

- Analytical skills: qualities and characteristics associated with solving problems using facts
  - Curiosity
  - Understanding contexts: understanding where information fits into the "big picture"
    Context: The condition in which something exists or happens

    - Having a technical mindset: breaking big things into smaller steps
    - Data design: organize information
    - Data strategy: management of the people, process, and tools udes in DA

2. About Analytical thinking

2.1. Def:

Analytical thinking: Identifying + defining a problem -> solving it by using data in an organized, step-by-step manner

2.2. 5 key aspects:

- Visualization: the graphical representation of info
- Strategy
- Problem-orientation
- Correlation: understand the relationship btw factors.
  Correlation # causation
- Big-picture and detail-oriented thinking

2.3. Core analytical skills

Think analytically, critically (knowing the right question to ask), and creatively (create a new path of solution)

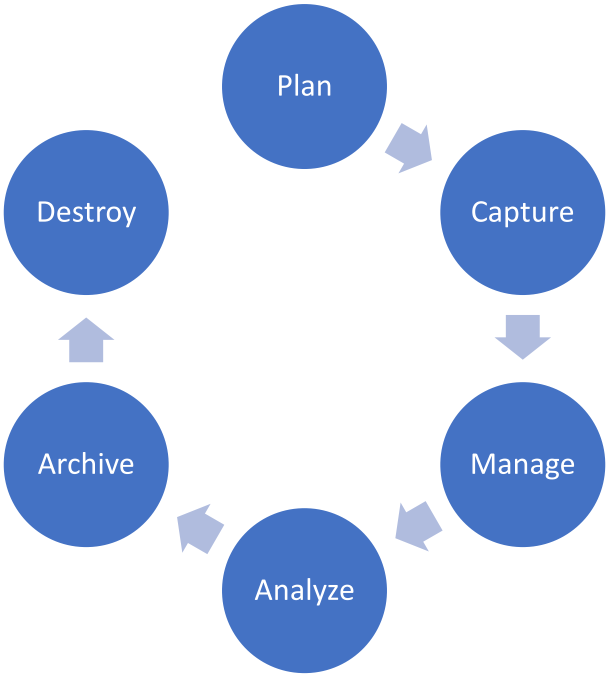
- Firstly, think about the root cause (the reason why a problem occurs).
    - Ask “why?” 5 times.
- Ask “where are the gaps in our process?”
    - Gap analysis: a method for examining and evaluating how a process works currently in order to get where you want to be in the future.
    - Compare where u are now and where u want to be -> identify gaps -> determine how to bridge them
- Ask “What did we not consider before?”

3. Thinking about outcomes

Week 3

1. The Data Cycle

1.1. Stages of the data cycle



- Planning: a business decides
  - what kind of data it needs
  - how it will be managed throughout its life cycle
  - who will be responsible for it
  - the optimal outcomes
- Capture: (common method) getting data from outside resources (publicly dataset, company’s own documents and files inside a database)
- Manage:
  - How we care for our data
  - How and where it's stored
  - The tools used to keep it safe and secure
  - The actions taken to make sure that it's maintained properly
- Analyze: the data is used to solve problems, make great decisions, and support business goals
- Archive: storing relevant data in an available place, but may not be used again
- Destroy: Remove data from storage and delete any shared copies of the data

WARNING: Data life cycle # data analysis life cycle

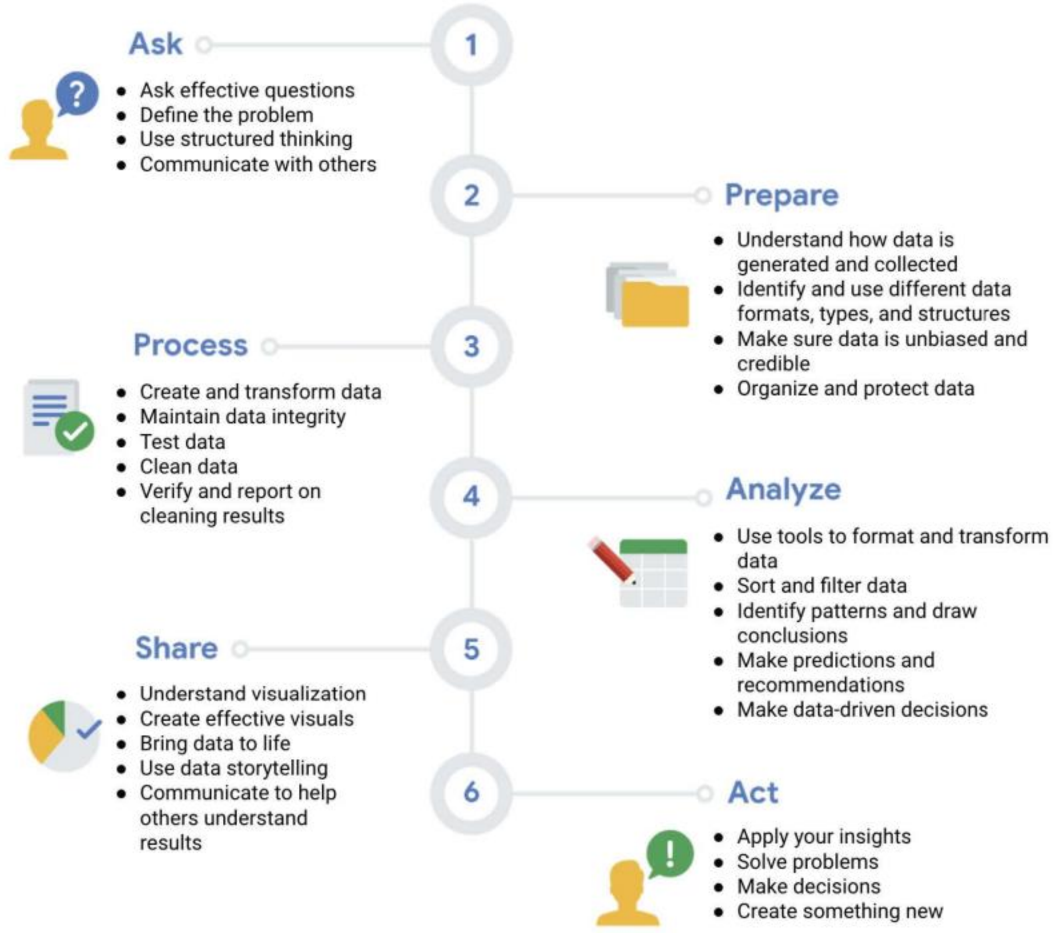
Manage deals with data that is still useful, archive deals with relevant data, but may not be used again.

1.2. Variations of the data life cycle
Individual stages in the data life cycle will vary from company to company or by industry or sector

⇨ Govern how data is handled so that it is accurate, secure, and available to meet your organization's needs.

2. Outlining the data analysis process

2.1. The 6 phases



1 Ask
- Ask effective questions
- Define the problem
- Use structured thinking
- Communicate with others

2 Prepare
- Understand how data is generated and collected
- Identify and use different data formats, types, and structures
- Make sure data is unbiased and credible
- Organize and protect data

3 Process
- Create and transform data
- Maintain data integrity
- Test data
- Clean data
- Verify and report on cleaning results

4 Analyze
- Use tools to format and transform data
- Sort and filter data
- Identify patterns and draw conclusions
- Make predictions and recommendations
- Make data-driven decisions

5 Share
- Understand visualization
- Create effective visuals
- Bring data to life
- Use data storytelling
- Communicate to help others understand results

6 Act
- Apply your insights
- Solve problems
- Make decisions
- Create something new

2.2. Key data analyst tools

Spreadsheets: Microsoft Excel and Google Sheets.

Purpose: collect and organize data.

Spreadsheets structure data in a meaningful way by letting you

- Collect, store, organize, and sort information
- Identify patterns and piece the data together in a way that works for each specific data project
- Create excellent data visualizations, like graphs and charts.

Databases and query languages: Structured Query Language (SQL) programs include MySQL, Microsoft SQL Server, and BigQuery.

A database is a collection of structured data stored in a computer system.

A query language is a computer programming language that allows you to retrieve and manipulate data from a database.

Query languages

- enables data analysts to request, retrieve, and update information from a database.

Spreadsheets	Databases
Software applications	Data stores - accessed using a query language (e.g. SQL)
Structure data in a row and column format	Structure data using rules and relationships
Organize information in cells	Organize information in complex collections
Provide access to a limited amount of data	Provide access to huge amounts of data
Manual data entry	Strict and consistent data entry
Generally one user at a time	Multiple users
Controlled by the user	Controlled by a database management system
Use both tools depending on the purpose and situation.	

Visualization tools: Tableau and Looker.

Data analysts use a number of visualization tools, like graphs, maps, tables, charts, and more.

These tools

- Turn complex numbers into a story that people can understand

- Help stakeholders come up with conclusions that lead to informed decisions and effective business strategies
- Have multiple features

- **Tableau**'s simple drag-and-drop feature lets users create interactive graphs in dashboards and worksheets

- **Looker** communicates directly with a database, allowing you to connect your data right to the visual tool you choose

A career as a data analyst also involves using programming languages, like R and Python, which are used a lot for statistical analysis, visualization, and other data analysis.

3. Difference between data analysis process and data life cycle.

	Data analysis process	Data life cycle
Similarities	Involve planning and asking questions	
Differences	The ask phase: focuses on big-picture strategic thinking about business goals	The Plan phase: focuses on the fundamentals of the project, such as what data you have access to, what data you need, and where you're going to get it.

Week 4

1. Spread sheet

The spreadsheet contains cells (ô), rows (hàng), and columns (cột)

Attribute: A characteristic/ quality of data used to label a column in a table.

-> Column name, header, column labels, etc.

Row is also called an observation.

Resources to learn about Spreadsheet and Excel

Google Sheets Training and Help

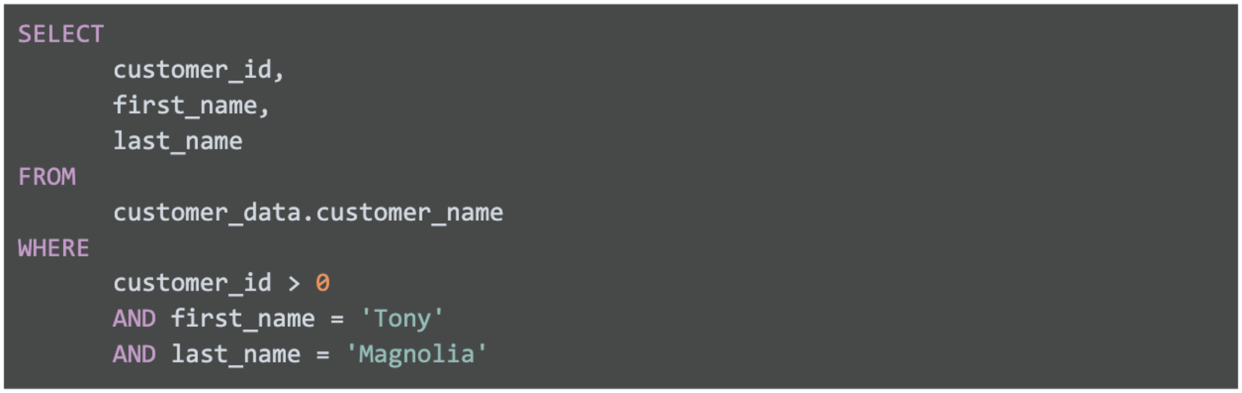
Google Sheets Cheat Sheet

2. SQL

Query: A request for data/ information from a data base

Syntax: the predetermined structure of a language that includes all required words, symbols, and punctuation, as well as their proper placement.

- **SELECT** ***columns***
- **FROM** *table name*.
- **WHERE** *certain conditions* (to filter for certain information)
  Eg:



```
SELECT
       customer_id,
       first_name,
       last_name
FROM
       customer_data.customer_name
WHERE
       customer_id > 0
       AND first_name = 'Tony'
       AND last_name = 'Magnolia'
```

1. **SELECT** the columns named **customer_id**, **first_name**, and **last_name**
2. **FROM** a table named **customer_name** (in a dataset named **customer_data**) (The dataset name is always followed by a dot, and then the table name.)
3. But only return the data **WHERE** customer_id is greater than **0**, first_name is **Tony**, and last_name is **Magnolia**.

Results:

customer_id	first_name	last_name
1967	Tony	Magnolia
7689	Tony	Magnolia

Overall:

```
SELECT
        ColumnA,
        ColumnB,
        ColumnC
FROM
        Table where the data lives
WHERE
        Condition 1
        AND Condition 2
        AND Condition 3
```

Note:

- Use **Capitalization, indentation, and semicolons** flexibly to review/trouble shoot it **easier**
- The **percent sign (%)** is used as a wildcard to match >=1 characters. In some databases an asterisk (*) is used as the wildcard instead of a percent sign (%).

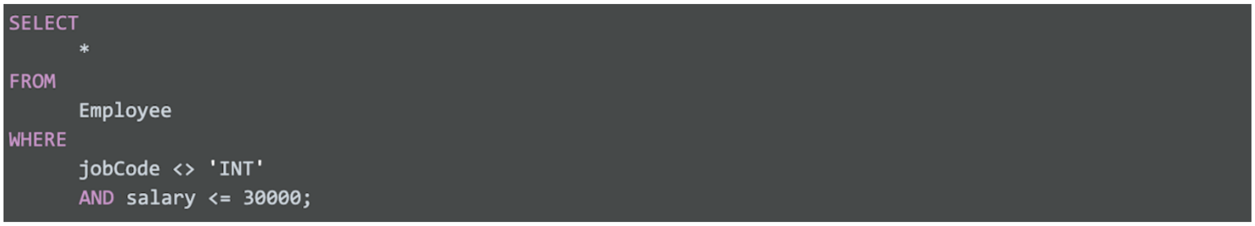
  Eg: to find a specific customer with the last name Chavez:

  WHERE field1 = 'Chavez'

  Find all customers with the last name that begins with the letters “Ch,"

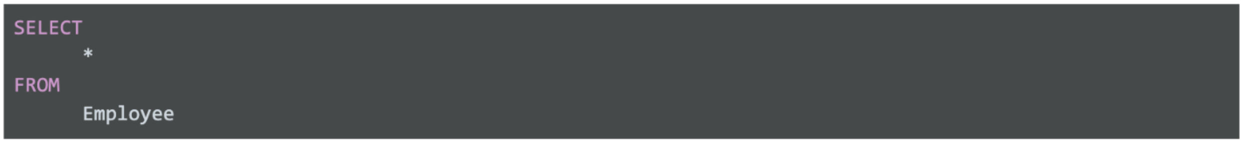
  WHERE field1 LIKE 'Ch%'

- Exclude conditions: choose rows except abc -> columns <> ‘abc’
  <> : "does not equal"



```
SELECT
      *
FROM
      Employee
WHERE
      jobCode <> 'INT'
      AND salary <= 30000;
```

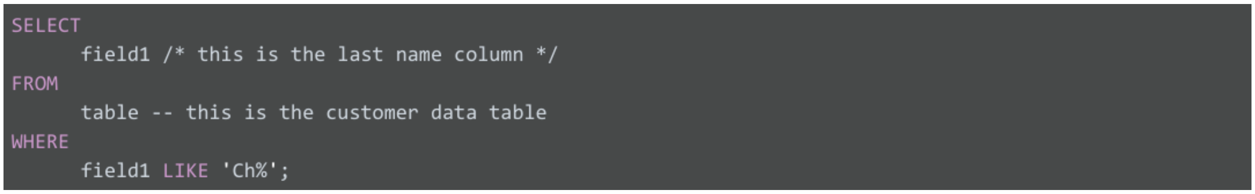
- **SELECT *:** selecting all of the columns in the table (use sparingly)



```
SELECT
      *
FROM
      Employee
```

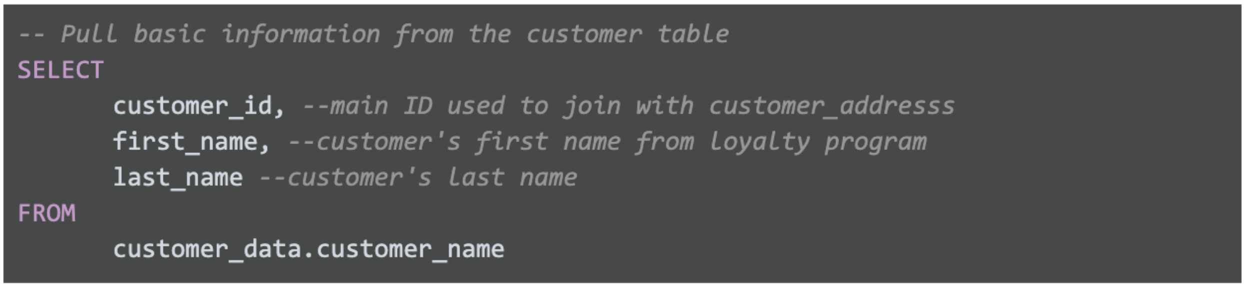
- **Comments:** you can place comments alongside your SQL to help you remember what the name represents
  How: **/*** Comments ***/**, or -- Comments

Eg:



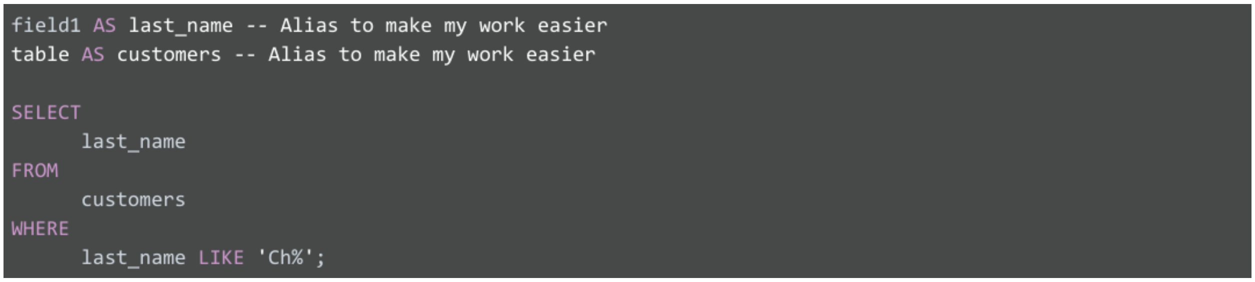
```
SELECT
        field1 /* this is the last name column */
FROM
        table -- this is the customer data table
WHERE
        field1 LIKE 'Ch%';
```

Comments can be added outside and within a statement to provide an overall description of what you are going to do, step-by-step notes about how you achieve it, and why you set different parameters/conditions.



```
-- Pull basic information from the customer table
SELECT
        customer_id, --main ID used to join with customer_addresss
        first_name, --customer's first name from loyalty program
        last_name --customer's last name
FROM
        customer_data.customer_name
```

- **Aliases:** assigning a new name or **alias** to the column/table names
  => make them easier to work with (and avoid the need for comments)



```
field1 AS last_name -- Alias to make my work easier
table AS customers -- Alias to make my work easier

SELECT
        last_name
FROM
        customers
WHERE
        last_name LIKE 'Ch%';
```

Note: An alias doesn't change the actual name of a column or table in the database.

Resources for SQL

W3Schools SQL Tutorial: Click the green **Start learning SQL now** button or the **Next** button to begin the tutorial.

SQL Cheat Sheet

3. Data Visualization

Purpose: to make data easier to read, to create interesting graphs, and to reinforce data analysis

Steps to plan a data visualization

Step 1: Explore the data for patterns
Ask for the information
Step 2: Plan your visuals
Know your target audience: whom will I present my data to?

EG: your audience is *sales oriented* -> the data visualization you use should:

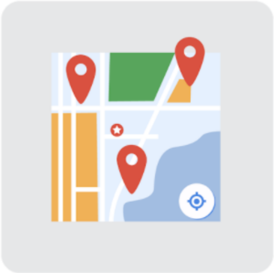
- Show *sales numbers* over time
- Connect *sales* to *location*
- Show *the relationship between sales* and *website use*
- Show *which customers* fuel growth

Step 3: Create your visuals

It is the process of trying different visualization formats and making adjustments

Eg:

Line charts can track sales over time



Maps can connect sales to locations

Donut charts can show customer segments

Bar charts can compare total visitors and visitors that make a purchase

Data visualization toolkit

1. **Spreadsheets (Microsoft Excel or Google Sheets):** for creating **simple visualizations** (bar graphs, pie charts, some advanced visualizations like maps, and waterfall and funnel diagrams (shown in the following figures))

2. Visualization software (Tableau): For a **wide variety of data** and includes an **interactive dashboard** that lets you and your stakeholders click to explore the data interactively.

Exploring Tabelau: How-to Video

Viz of the Day: beautiful visuals ranging from the Hunt for (Habitable) Planets to Who's Talking in Popular Films.

3. Programming language (R with RStudio): When using Rstudio

Exploring RStudio, RStudio Cheatsheets and the RStudio Visualize Data Primer.

Week 5

1. Data in business

Issue: a topic/subject to investigate

Question: design to discover information

Problem: An ostacle/ complication that needs to be worked out

Business task: Question/problem data analysis ansers for a business

If can't find the insight -> try to find a different patterns: ask other question, come at in a different way, etc.

2. Data and fairness

Fairness: ensuring that your data doesn't create/reinforce bias

Note:

consider all of the available data on the field

think about the other surrounding factors that impact the data

collected self reported data in a separate system

oversampled non-dominant groups to ensure the model was including them

If it is a volunteer workshop -> hard to conclude.

Asking, "Why?" when reviewing the results of data analysis.

3. Jobs in data fields

Interview tips:

- Think about a time where you've used data to solve a problem, whether it's in your professional or personal projects
- increase your professional network (increase your online footprint, reach out to other analysts on LinkedIn, join local meet-ups with other data scientists)
- have your LinkedIn updated along with websites like GitHub, where you can showcase a lot of the data analysts projects you've done.
- prepare questions for the interviewer (not broad questions). They should be questions to understand the team and the job better

- given a case study in an interview, you should expect to be given a business problem along with the sample data set: ensure you are analyzing the data and coming up with a solution that relates back to that data
- Look for the recruiter.