Google Professional Certificate: Data Analytics

Course 1: Data, Data, Everywhere

Introduction to the course

Data is a collection of facts. This can include numbers, pictures, words, etc. Data analysis is the collection, transformation and organization of data in order to draw conclusions, make predictions and drive informed decision making.

Computer + Brain + Skills + Traits = Job Success !

Roadmap of the Learnings through the Modules

Skill sets you will build:

- Using data in everyday life
- Thinking analytically
- Applying tools from the data analytics toolkit
- Showing trends and patterns with data visualizations
- Ensuring your data analysis is fair

Skill sets you will build:

- Asking SMART and effective questions
- Structuring how you think
- Summarizing data
- Putting things into context
- Managing team and stakeholder expectations
- Problem-solving and conflict-resolution

Skill sets you will build:

- Ensuring ethical data analysis practices
- Addressing issues of bias and credibility
- Accessing databases and importing data
- Writing simple queries
- Organizing and protecting data
- Connecting with the data community (optional)

Skill sets you will build:

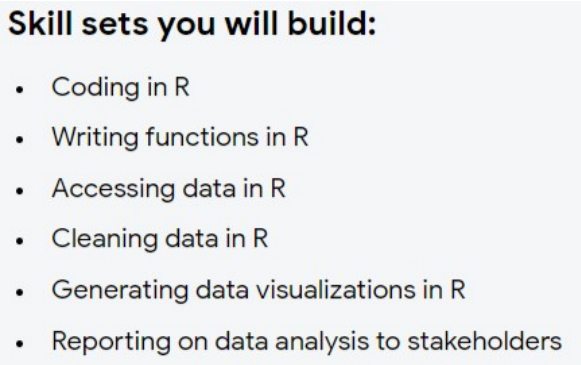
- Connecting business objectives to data analysis
- Identifying clean and dirty data
- Cleaning small datasets using spreadsheet tools
- Cleaning large datasets by writing SQL queries
- Documenting data-cleaning processes

Skill sets you will build:

- Sorting data in spreadsheets and by writing SQL queries
- Filtering data in spreadsheets and by writing SQL queries
- Converting data
- Formatting data
- Substantiating data analysis processes
- Seeking feedback and support from others during data analysis

Skill sets you will build:

- Creating visualizations and dashboards in Tableau
- Addressing accessibility issues when communicating about data
- Understanding the purpose of different business communication tools
- Telling a data-driven story
- Presenting to others about data
- Answering questions about data



Job of a Data Analyst

A company uses data from multiple sources, mainly from customer feedbacks. A company uses this data to know more about the customer such as their buying habits, expense range, etc. A data analyst does the analysis of these data. Data analysis is conducted in healthcare, industries, businesses, etc. A successful data analyst is one which has a balance of personal and technical skills. Observation and intuition are also some powerful tools for data analysts but one must always have a balance between the data and intuition to have the most efficient analysis.

A data analyst must always make a fair data analysis. There shouldn't be any bias about the data analysis. There should be no bias from the step of collection of data until the act phase of data analysis. There are many other jobs which sound the same but are essentially different. Some examples are:

- Business analyst — analyzes data to help businesses improve processes, products, or services
- Data analytics consultant — analyzes the systems and models for using data
- Data engineer — prepares and integrates data from different sources for analytical use
- Data scientist — uses expert skills in technology and social science to find trends through data analysis
- Data specialist — organizes or converts data for use in databases or software systems
- Operations analyst — analyzes data to assess the performance of business operations and workflows

Other industry-specific specialist positions that you might come across in your data analyst job search include:

- Marketing analyst — analyzes market conditions to assess the potential sales of products and services
- HR/payroll analyst — analyzes payroll data for inefficiencies and errors
- Financial analyst — analyzes financial status by collecting, monitoring, and reviewing data
- Risk analyst — analyzes financial documents, economic conditions, and client data to help companies determine the level of risk involved in making a particular business decision
- Healthcare analyst — analyzes medical data to improve the business aspect of hospitals and medical facilities

There is a lot of differences between a Data Analyst, Data Scientist and a Data Specialist. Some of them are given below:

Decoding the job description

	Data Analysts	Data Scientists	Data Specialists
Problem solving	Use existing tools and methods to solve problems with existing types of data	Invent new tools and models, ask open-ended questions, and collect new types of data	Use in-depth knowledge of databases as a tool to solve problems and manage data
Analysis	Analyze collected data to help stakeholders make better decisions	Analyze and interpret complex data to make business predictions	Organize large volumes of data for use in data analytics or business operations
Other relevant skills	• Database queries • Data visualization • Dashboards • Reports • Spreadsheets	• Advanced statistics • Machine learning • Deep learning • Data optimization • Programming	• Data manipulation • Information security • Data models • Scalability of data • Disaster recovery

Important factors to consider when searching for the dream job

1. Industry
    a. Financial Services
    b. Telecom
    c. Tech
2. Tools
3. Location
4. Travel
5. Culture

Steps of Data Analysis

1. Ask (questions to define the problem)
    - Ask effective questions to the stakeholders
    - Ask effective questions to the people interested in the results
    - Business challenge/Objective/Question
2. Prepare (data by collecting and storing the information)
    - Planning a timeline
    - Deciding on an effective communication channel
    - What data needs to be analyzed
    - Choosing the data collection method
    - Collecting the data

- Data generation, collection, storage and data management

3. Process (data by cleaning and checking the information)
    - Make sure that the people the data is based on are made known of how the data is collected, stored, managed and protected.
    - Data cleaning and data integrity
4. Analyze (data to find patterns, relationships and trends)
    - Conduct the analysis of the data and find out trends, predictions, etc.
    - Data exploration, visualization and analysis
5. Share (on the data and use the analysis results)
    - The results were shared in an appropriate, legal and ethical manner with the people who required the results.
    - Communicating and interpreting results
6. Act
    - Implement the changes and compare the results with another round of data analysis to check if the plan worked!
    - Putting your insights together to solve the problem

Disciplines in Data Analytics

There are a total of 3 disciplines in Data analytics which are namely: Statistics, Machine Learning and Analytics. Each separated by how many decisions one wants to make. If one wants to make few important decisions under uncertainty, then it is statistics. If one wants to make multiple decisions or automate the decisions, then it is Artificial Intelligence or Machine Learning. And lastly, if one does not know the number of decisions and want to find the answer based on inspiration, counter the unknown, etc. then, it is Analytics.

Choose the specialization on which type of impact suits your personality.
(Analytics is for me)

Data Ecosystems

An ecosystem is a group of elements which interact with each other. It can be either large or small. Thus, one can say that data ecosystem is nothing but the various elements that interact with each other to produce, manage, store, organize, analyze and share data.

The elements in a data ecosystem include the hardware and the software tools and the people who use them. Data is found to be stored in the cloud, a virtual database.

Data Scientist vs. Data Analyst

Data Science is defined as creating new ways of modelling and understanding the unknown by using raw data. A data scientist will create new questions using data whereas a data analyst finding answers to existing questions through finding insights from existing sources.

Data analysis is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision making. Data analytics on

the other hand is the science of data. It is a broad term which encompasses everything from managing and using data to the tools and methods that data workers use every day.

Data driven decision making

One of the most powerful ways to put data to work is with data driven decision making which is defined as using facts to guide business strategy. Through the data driven decision making approach, one takes the current data, analyzes it, then use the insights gained to suggest/predict further. It is always better to include insights from people familiar with the business problem, these people are called Subject Matter Experts and are useful to identify any data inconsistencies, make sense of grey areas and eventually validate the choices being made.

Data vs. Experience

It is essential that data analysts focus on the data to ensure they make informed decisions. One should not ignore the data based on the experience as it may lead to biased decisions. Decisions based on gut feelings without any data to back up can cause mistakes.

The key to being a successful junior data analyst in blending the data, business knowledge and a little bit of gut instinct in an exact mix required for the project. It depends on the goals of the analysis and hence analysts often should ask the following questions –

- How do I define the success of the project?
- 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?

Different Data Analysis Life Cycles

1. EMC's data analysis life cycle
    - Discovery
    - Preprocessing data
    - Model planning
    - Model building
    - Communicating results
    - Operationalize
2. SaaS's iterative Life Cycle
    - Ask
    - Prepare
    - Explore
    - Model
    - Implement
    - Act
    - Evaluate
3. Project based data analytics life cycle

- Identifying the problem
- Designing data requirements
- Preprocessing data
- Performing data analysis
- Visualizing data

4. Big data analytics Life Cycle
    - Business case evaluation
    - Data identification
    - Data acquisition and filtering
    - Data extraction
    - Data validation and cleaning
    - Data aggregation and representation
    - Data analysis
    - Data visualization
    - Utilization of analysis results

Key skills required by a Data Analyst

- Analytical Skills

    Qualities and characteristics associated with solving problems using facts. There are a total of 5 essential points one should focus on:

    i. Curiosity – All about wanting to learn something.
    ii. Understanding context – Context is the condition in which something exists or happens.
    iii. Having a technical mindset – The ability to break things down into smaller steps and work with them in an orderly or logical way.
    iv. Data design – How you organize information.
    v. Data strategy – The management of the people, processes and tools using data analysis.

Analytical Thinking

Analytical thinking is the identifying and defining a problem and then solving it by using data in an organized, step-by-step manner. There are 5 aspects of analytical thinking, which are:

i. Visualization – the graphical representation of information
ii. Strategy – helps stay focused and on track. It helps to improve quality and usefulness of the data
iii. Problem Orientation – used in order to identify, describe and solve problems
iv. Correlation – A relationship between two or more data. But correlation does not mean causation.
v. Big picture and detail oriented thinking – Being able to see the big picture but at the same time see the details.

One needs to ask a lot of questions when working on a problem. A few given below:

- What is the root cause of the problem? To answer this question, we ask 5 whys.

- Where are the gaps in our process? To answer this question, we use gap analysis. Gap analysis is a method of examining and evaluating how a process works currently in order to get where you want to be in the future. The general approach to gap analysis is:
    - Understanding where you are now compared to where you want to be.
    - Understand the gaps between the current and future state and determine how to bridge them.
- What did we not consider before? A question asked to figure out any question or procedure might be missing from the process to make better decisions.

Data Life Cycle

Data analysts bring data to life by using data analysis tools such as spreadsheets, Databases (collection of data stored in a computer system), Query languages and visualization softwares. The standard Life cycle of data is:

- Plan – This starts well before an analysis project. What kind of data is needed? How it would be managed? Who would be responsible for it? and the possible outcomes are all taken care of in this phase.
- Capture – In this phase the data is collected from different sources and brought into the organization. There are multiple ways to collect data such as collecting it from outside sources, reusing past data, etc.
- Manage – How one cares for the data? How it's stored and the tools used and the actions to keep the data secure. It is very important to data cleansing.
- Analyze – The data is used to solve problems and make decisions to complete business goals.
- Archive – Archiving data means storing data which may not be used again.
- Destroy – Data is destroyed using tools to protect the company's information and privacy.

There are different types of Data life cycles just like how there are different Software process life cycles. Some examples are:

- Plan – Acquire – Maintain – Access – Evaluate – Archive.
- Plan – Acquire – Process – Analyze – Preserve – Publish/Share.
- Capture – Qualify – Transform – Utilize – Report – Archive – Purge.
- Generation – Collection – Processing – Storage – Management – Analysis – Visualization – Interpretation.

Data Analysis Life Cycle

As seen before, there are 6 stages to a data analysis – Ask, Prepare, Process, Analyze, Share and Act. In the Ask phase we define the problem and make sure we understand the stakeholder's expectations. Stakeholders are people who have invested time and resources into a project and are interested in the outcome.

"Look at the current state and identify how it's different from the ideal state."

In the prepare step, the data analysts collect and store data for the upcoming analysis process. In the process step, the data analysts find and eliminate any inaccuracies that can get in the way of the results. This usually involves cleaning data, transforming data, combining it with more useful datasets and removing outliers. It also involves removing or fixing inaccuracies.

The next step is the analyzing phase wherein tools are used to transform and organize the information to draw useful conclusions, predictions and draw insights for decision making. Following the analysis phase, is the share phase. Here data analysts interpret results and share them with others to help stakeholders make data-driven decisions. Visualization of data is the key component of this phase.

Data Analyst Toolkit

- Spreadsheets: Microsoft Excel or Google Sheets
- Databases and Query Languages: MySQL, BigQuery, etc.
- Visualization Tools: Tableau, Looker, etc.

Basics of Spreadsheet

	A	B	C	D
1			This	
2	This is a cell.		is	
3	This	is	a	row.
4			column.	
5				

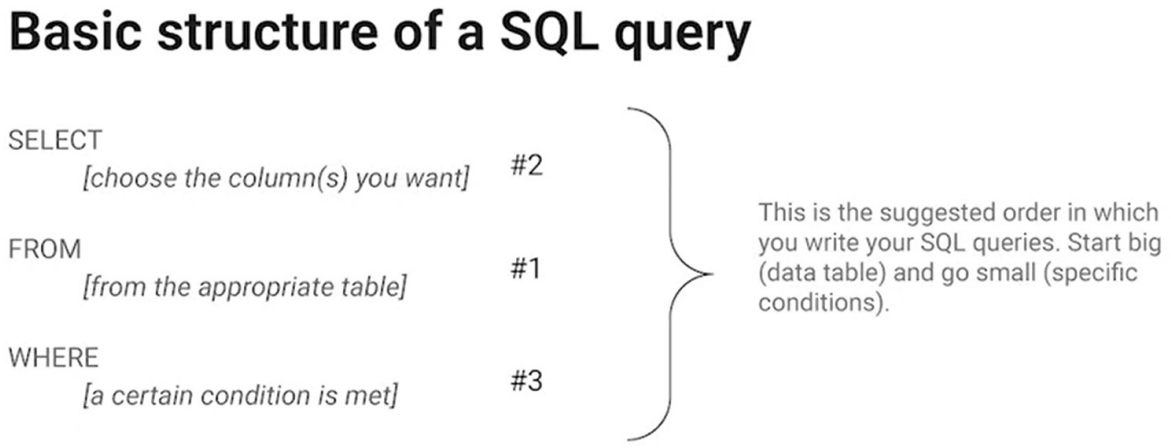
Attribute – A characteristic or quality of data used to label a column in a table.

Observation – All of the attributes for something contained in a row of a data table.

Formula – A set of instructions used to perform a calculation using the data in a spreadsheet.

Basics of SQL

There are a number of databases which require SQL such as Oracle, MySQL, PostgreSQL, Microsoft SQL Server, etc. For whichever database we use, we use SQL in the same format as for others. In SQL, Query is universal. A query is a request for data or information from a database.



"For more details, please visit: https://www.w3schools.com/sql/default.asp"

Basics of Visualization

Data visualization is the graphical representation of data. Certain tools such as Tableau, Looker, Power BI, etc. are used for the purpose of visualizing data and information. There are certain steps a data analyst shall follow when visualizing data. One such process is:

- Exploring the data for patterns
- Planning your visuals
- Creating your visuals

The Analytical Skills Table

Analytical Skill	Strength	Developing	Emerging	Comment/Plan/Goal
Curiosity	✓			I am curious enough to start courses, watch YouTube tutorials, etc. to get to know stuff.
Context			✓	I feel I still can't get a grasp on the big picture and end up taking the smaller bits and pieces

				into account when making decisions.
Technical Mindset	✓			I have developed a knack of noting things down of what I need to do and further break it down into smaller bits for ease of execution or assigning.
Data Design	✓			Check out my laptop, the most organized you would ever find.
Data Strategy		✓		I do think about the people but processes and tools are something I don't know much about. I will learn about them in the short run.

- Curiosity – A desire to know more about something, asking right questions.
- Context – Understanding where information fits in thing picture.
- Technical Mindset – Breaking things into smaller pieces.
- Data Design – Thinking about how to organize data and information.
- Data Strategy – Thinking about the people, processes and tools used in data analytics.

- Strength: The area you feel is a strength.
- Developing: The are you have some experiences in but has significant room for growth.
- Emerging: New to you but will soon gain experience.

Course 2: Ask Questions to make Data Driven Decisions

Introduction

We need to ask appropriate and great question that can lead us to gain insights which would eventually lead to solving business problems. We have to develop something known as Structured Thinking which is the process of recognizing the current problem or situation, organizing available information, revealing gaps and opportunities and identifying the options.

Marketing analysis is the process of measuring, analyzing and managing a company's marketing strategy and budget. Often, this involves identifying the company's target audience which includes the people the company is trying to reach.

Data Analysts work with basically 6 different problem types, such as:

- Making predictions – Using data to make an informed decision about how things may be in the future.
- Categorizing things – Assigning information to different groups or clusters based on common features.
- Spotting something unusual – Identify data which is different from the normal.
- Identifying themes – Grouping categorized information into broader concepts.
- Discovering connections – Finding similar challenges faced by different entities and combining data and insights to address them.
- Finding patterns – Using historical data to understand what happened in the past and is therefore likely to happen again.

The 6 Data Analysis Phases

Step 1: Ask

It's impossible to solve a problem if you don't know what it is. These are some things to consider:

- 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:

1. What are my stakeholders saying their problems are?
2. Now that I've identified the issues, how can I help the stakeholders resolve their questions?

Step 2: Prepare

You will decide what data you need to collect in order to answer your questions and how to organize it so that it is useful. You might use your business task to decide:

- What metrics to measure
- Locate data in your database
- Create security measures to protect that data

Questions to ask yourself in this step:

1. What do I need to figure out how to solve this problem?
2. What research do I need to do?

Step 3: Process

Clean data is the best data and you will need to clean up your data to get rid of any possible errors, inaccuracies, or inconsistencies. This might mean:

- Using spreadsheet functions to find incorrectly entered data
- Using SQL functions to check for extra spaces
- Removing repeated entries
- Checking as much as possible for bias in the data

Questions to ask yourself in this step:

1. What data errors or inaccuracies might get in my way of getting the best possible answer to the problem I am trying to solve?
2. How can I clean my data so the information I have is more consistent?

Step 4: Analyze

You will want to think analytically about your data. At this stage, you might sort and format your data to make it easier to:

- Perform calculations
- Combine data from multiple sources
- Create tables with your results

Questions to ask yourself in this step:

1. What story is my data telling me?
2. How will my data help me solve this problem?
3. Who needs my company's product or service? What type of person is most likely to use it?

Step 5: Share

Everyone shares their results differently so be sure to summarize your results with clear and enticing visuals of your analysis using data via tools like graphs or dashboards. This is your chance to show the stakeholders you have solved their problem and how you got there. Sharing will certainly help your team:

- Make better decisions
- Make more informed decisions
- Lead to stronger outcomes
- Successfully communicate your findings

Questions to ask yourself in this step:

1. How can I make what I present to the stakeholders engaging and easy to understand?
2. What would help me understand this if I were the listener?

Step 6: Act

Now it's time to act on your data. You will take everything you have learned from your data analysis and put it to use. This could mean providing your stakeholders with recommendations based on your findings so they can make data-driven decisions.

Questions to ask yourself in this step:

1. How can I use the feedback I received during the share phase (step 5) to actually meet the stakeholder's needs and expectations?

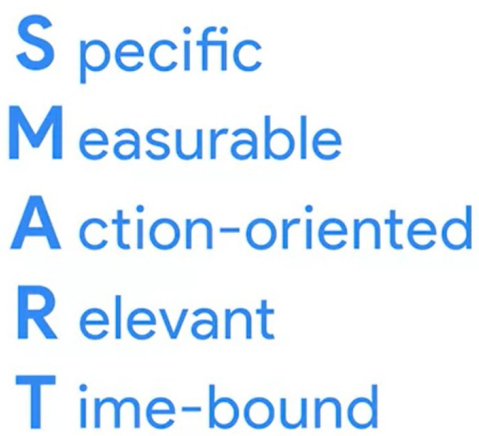
These six steps can help you to break the data analysis process into smaller, manageable parts, which is called **structured thinking**. This process involves four basic activities:

1. Recognizing the current problem or situation
2. Organizing available information
3. Revealing gaps and opportunities
4. Identifying your options

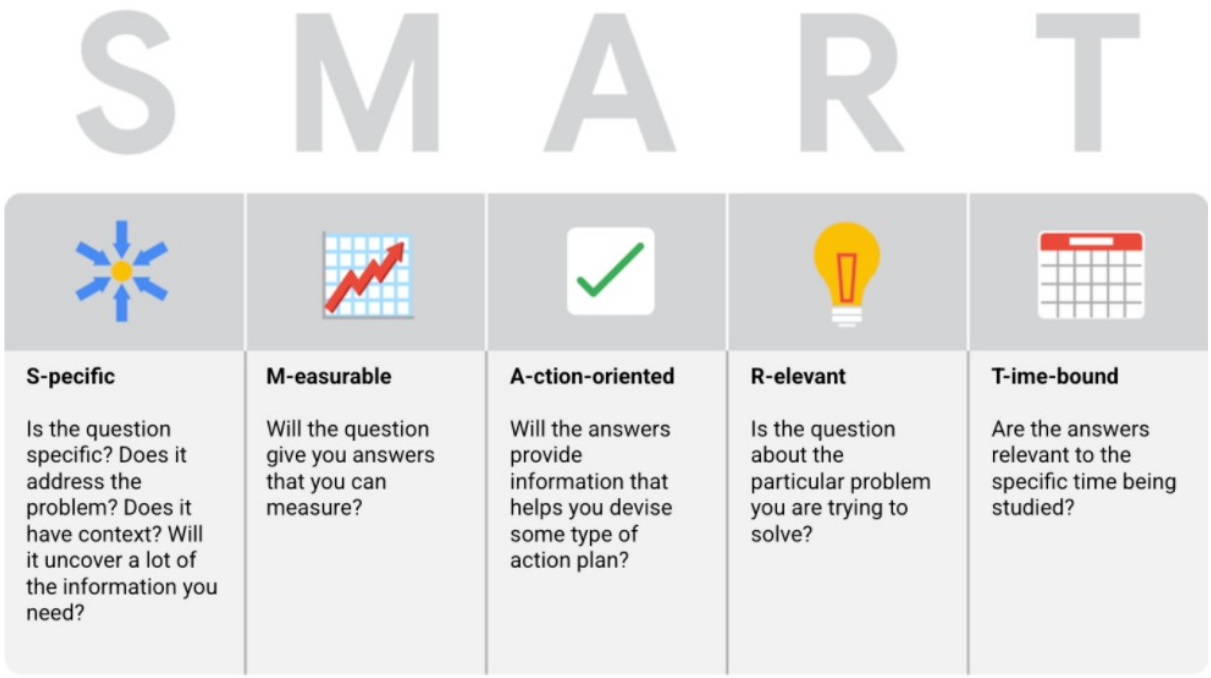
When you are starting out in your career as a data analyst, it is normal to feel pulled in a few different directions with your role and expectations. Following processes like the ones outlined here and using structured thinking skills can help get you back on track, fill in any gaps and let you know exactly what you need.

Asking Effective Questions

The data analysis job starts with the "Ask" phase and thus it is important we ask the right questions which are followed by the S.M.A.R.T Methodology. The questions should also be fair and unbiased.



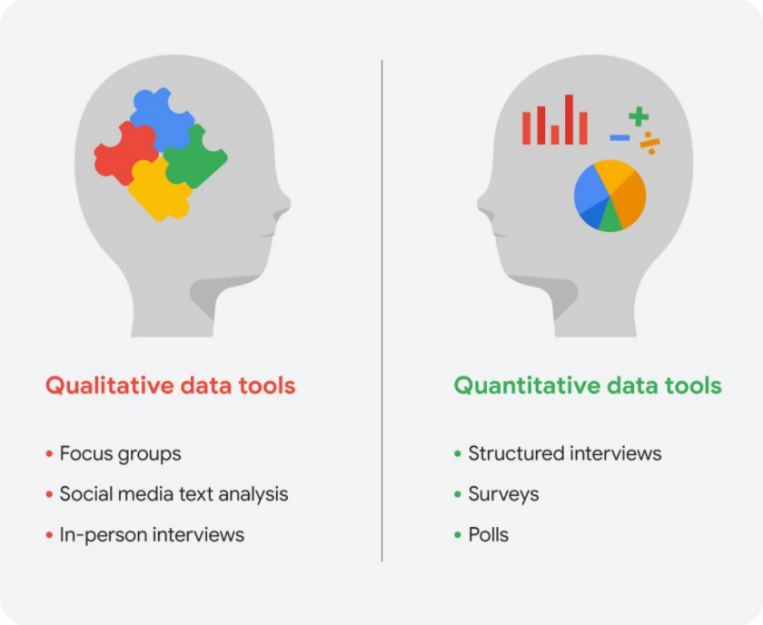
- Specific
    - Specific questions are simple, significant, and focused on a single topic or a few closely related topics.
- Measurable
    - Measurable questions can be quantified and assessed.
- Action Oriented
    - Action Oriented questions are those which encourage change.
- Relevant
    - Relevant questions matter, are important, and have significance to the problem being solved.
- Time Bound
    - Time bound questions are those which specify the time period that has to be studied.



How Data Inspires Decision

Business use data to make decisions. There are 2 ways to do this, one is by Data driven decision making and another by Data inspired decision making. Data inspired decision making explores the different data sources to find out what they have in common. Data is of 2 types:

- Quantitative
    - Specific and Objective measures of numerical facts.
- Qualitative
    - Subjective or Explanatory measures of qualities and characteristics.



Data analysts will generally use both types of data in their work. Usually, qualitative data can help analysts better understand their quantitative data by providing a reason or more thorough explanation.

Data needs to be presented to the shareholders in the share phase. The two common data presentation tools are:

- Reports
  - A report is a static collection of data given to stakeholders periodically.
  - Advantages
    - High level historical data
    - Easy to design
    - Pre-cleaned and sorted data
  - Disadvantages
    - Continual maintenance
    - Less visually appealing
    - Are static
- Dashboards
  - A dashboard monitors live, incoming data.
  - Advantages
    - Dynamic, automatic and interactive
    - More stakeholder access
    - Visually appealing
  - Disadvantages
    - Labor intensive design
    - Can be confusing
    - Have potentially unclean data

Types of dashboards

For a refresher, consider the different types of dashboards a business may use. Often, businesses will tailor a dashboard for a specific purpose. The three most common categories are:

- **Strategic**: focuses on long term goals and strategies at the highest level of metrics
- **Operational:** short-term performance tracking and intermediate goals
- **Analytical:** consists of the datasets and the mathematics used in these sets

Small data	Big data
Describes a data set made up of specific metrics over a short, well-defined time period	Describes large, less-specific data sets that cover a long time period
Usually organized and analyzed in spreadsheets	Usually kept in a database and queried
Likely to be used by small and midsize businesses	Likely to be used by large organizations
Simple to collect, store, manage, sort, and visually represent	Takes a lot of effort to collect, store, manage, sort, and visually represent
Usually already a manageable size for analysis	Usually needs to be broken into smaller pieces in order to be organized and analyzed effectively for decision-making

Volume	Variety	Velocity	Veracity
The amount of data	The different kinds of data	How fast the data can be processed	The quality and reliability of the data

Challenges and benefits

Here are some **challenges** you might face when working with big data:

- A lot of organizations deal with data overload and way too much unimportant or irrelevant information.
- Important data can be hidden deep down with all of the non-important data, which makes it harder to find and use. This can lead to slower and more inefficient decision-making time frames.
- The data you need isn't always easily accessible.
- Current technology tools and solutions still struggle to provide measurable and reportable data. This can lead to unfair algorithmic bias.
- There are gaps in many big data business solutions.

Now for the good news! Here are some **benefits** that come with big data:

- When large amounts of data can be stored and analyzed, it can help companies identify more efficient ways of doing business and save a lot of time and money.
- Big data helps organizations spot the trends of customer buying patterns and satisfaction levels, which can help them create new products and solutions that will make customers happy.
- By analyzing big data, businesses get a much better understanding of current market conditions, which can help them stay ahead of the competition.
- As in our earlier social media example, big data helps companies keep track of their online presence—especially feedback, both good and bad, from customers. This gives them the information they need to improve and protect their brand.

Spreadsheet Errors

Error	Description	Example
#DIV/0!	A formula is trying to divide a value in a cell by 0 (or an empty cell with no value)	=B2/B3, when the cell B3 contains the value 0
#ERROR!	(Google Sheets only) Something can't be interpreted as it has been input. This is also known as a parsing error.	=COUNT(B1:D1 C1:C10) is invalid because the cell ranges aren't separated by a comma
#N/A	A formula can't find the data	The cell being referenced can't be found
#NAME?	The name of a formula or function used isn't recognized	The name of a function is misspelled
#NUM!	The spreadsheet can't perform a formula calculation because a cell has an invalid numeric value	=DATEDIF(A4, B4, "M") is unable to calculate the number of months between two dates because the date in cell A4 falls after the date in cell B4
#REF!	A formula is referencing a cell that isn't valid	A cell used in a formula was in a column that was deleted
#VALUE!	A general error indicating a problem with a formula or with referenced cells	There could be problems with spaces or text, or with referenced cells in a formula; you may have additional work to find the source of the problem.

Spreadsheet Commands

Command	Chromebook	PC	Mac
Create new workbook	Control+N	Control+N	Command+N
Open workbook	Control+O	Control+O	Command+O
Save workbook	Control+S	Control+S	Command+S
Close workbook	Control+W	Control+W	Command+W
Undo	Control+Z	Control+Z	Command+Z
Redo	Control+Y	Control+Y	Command+Y
Copy	Control+C	Control+C	Command+C
Cut	Control+X	Control+X	Command+X
Paste	Control+V	Control+V	Command+V
Paste values only	Control+Shift+V	Control+Shift+V	Command+Shift+V
Find	Control+Shift+F	Control+F	Command+F
Find and replace	Control+H	Control+H	Command+Shift+F
Insert link	Control+K	Control+K	Command+K
Bold	Control+B	Control+B	Command+B
Italicize	Control+I	Control+I	Command+I
Underline	Control+U	Control+U	Command+U
Zoom in	Control+Plus (+)	Control+Plus (+)	Option+Command+Plus (+)
Zoom out	Control+Minus (-)	Control+Minus (-)	Option+Command+Minus (-)
Select column	Control+Spacebar	Control+Spacebar	Command+Spacebar

Select row	Shift+Spacebar	Shift+Spacebar	Up Arrow+Spacebar
Select all cells	Control+A	Control+A	Command+A
Edit the current cell	Enter	F2	F2
Comment on a cell	Ctrl + Alt + M	Alt+I+M	Option+Command+M
Insert column to the left	Ctrl + Alt + = (with existing column selected)	Alt+Shift+I, then C	⌘ + Option + = (with existing column selected)
Insert column to the right	Alt + I, then O	Alt+Shift+I, then O	Ctrl + Option + I, then O
Insert row above	Ctrl + Alt + = (with existing row selected)	Alt+Shift+I, then R	⌘ + Option + = (with existing row selected)
Insert row below	Alt + I, then R, then B	Alt+Shift+I, then B	Ctrl + Option + I, then B

Structured Thinking

A big part of the data analyst job is to develop a structured approach and use critical thinking to find the best solution. Structured thinking is the process of recognizing the current problem or situation, organizing available information, revealing gaps and opportunities and identifying the options.

The starting place for structured thinking is the problem domain. A problem domain is the specific area of analysis that encompasses every activity affecting or affected by the problem. One way of practicing structured thinking is by using the Scope of Work (SOW). An SOW is an agreed upon outline of the work you are going to perform on a project.

A statement of Work is very different from a Scope of work, both abbreviated as SOW. A statement of work is a document that clearly identifies the products and services a vendor or contractor will provide to an organization. It includes objectives, guidelines, deliverables, schedule, and costs. A scope of work is project-based and sets the expectations and boundaries of a project. A scope of work may be included in a statement of work to help define project outcomes.

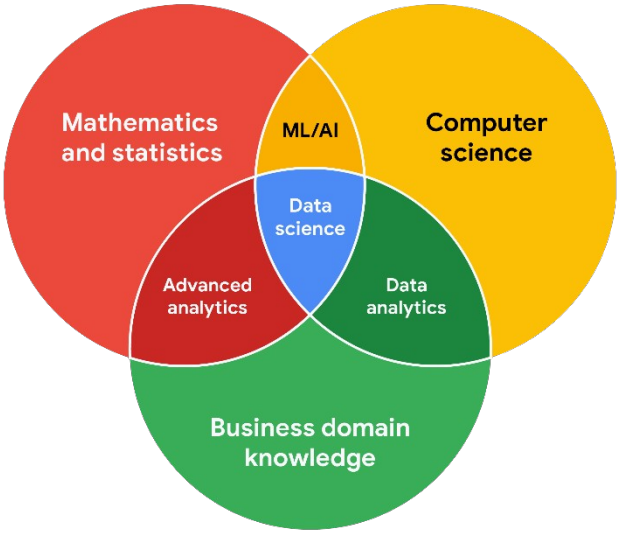
Communication as a key skill

Stakeholders are people that have invested time, interest and resources into the projects one works as a data analyst. A project manager is a person who is in charge of planning and executing the project. A key role of the project manager is to keep the entire project on track and oversee the progress of the entire team. There are different stakeholders, a few of them are:

- Executive team
    - The executive team provides strategic and operational leadership to the company. They set goals, develop strategy, and make sure that strategy is executed effectively. The executive team might include vice presidents, the chief marketing officer, and senior-level professionals who help plan and direct the company's work. These stakeholders think about decisions at a very high level.
- Customer Facing Team
    - The customer-facing team includes anyone in an organization who has some level of interaction with customers and potential customers. Typically, they

compile information, set expectations, and communicate customer feedback to other parts of the internal organization.

- Data Science Team
    - Organizing data within a company takes teamwork.



To work efficiently with the stakeholders, one must follow certain practices:

- Discuss Goals
- Feel empowered to say no
- Plan for the unexpected
- Know your project
- Start with words and visuals
- Communicate often

One must always have 3 questions answered before starting a data analysis project.

1. Who are the primary and secondary stakeholders?
2. Who is managing the data?
3. Where can you go for help?

Before you communicate, think about:

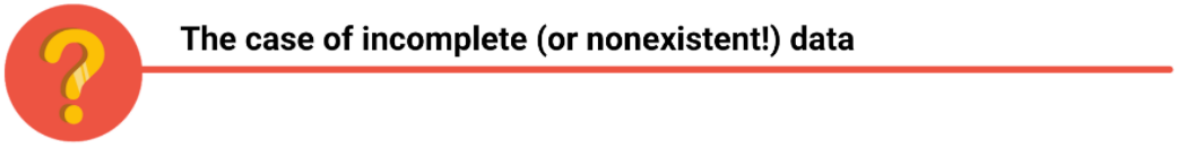
- Who the audience is?
- What they already know?
- What they need to know?
- How you can communicate that effectively to them?

Tips for effective communication in workplace:

- Learn as you go and ask questions.
- Read emails regularly, read out yours.
- Keep emails formal but mold yourself according to how the workplace environment is.
- Answer in a timely manner.
- Flag problems early for the stakeholders.
- Set realistic and objective goals for yourself in the project.

Course 3: Prepare Data for Exploration

Limitations of Data



If you have incomplete or nonexistent data, you might realize during an analysis that you don't have enough data to reach a conclusion. Or, you might even be solving a different problem altogether! For example, suppose you are looking for employees who earned a particular certificate but discover that certification records go back only two years at your company. You can still use the data, but you will need to make the limits of your analysis clear. You might be able to find an alternate source of the data by contacting the company that led the training. But to be safe, you should be up front about the incomplete dataset until that data becomes available.

If you're collecting data from other teams and using existing spreadsheets, it is good to keep in mind that people use different business rules. So one team might define and measure things in a completely different way than another. For example, if a metric is the total number of trainees in a certificate program, you could have one team that counts every person who registered for the training, and another team that counts only the people who completed the program. In cases like these, establishing how to measure things early on standardizes the data across the board for greater reliability and accuracy. This will make sure comparisons between teams are meaningful and insightful.

Dirty data refers to data that contains errors. Dirty data can lead to productivity loss, unnecessary spending, and unwise decision-making. A good data cleaning effort can help you avoid this. As a quick reminder, data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When you find and fix the errors - while tracking the changes you made - you can avoid a data disaster. You will learn how to clean data later in the training.

Tell a clear story

Avinash Kaushik, a Digital Marketing Evangelist for Google, has lots of great tips for data analysts in his blog: Occam's Razor. Below are some of the best practices he recommends for good data storytelling:

- **Compare the same types of data**: Data can get mixed up when you chart it for visualization. Be sure to compare the same types of data and double check that any segments in your chart definitely display different metrics.
- **Visualize with care**: A 0.01% drop in a score can look huge if you zoom in close enough. To make sure your audience sees the full story clearly, it is a good idea to set your Y-axis to 0.
- **Leave out needless graphs:** If a table can show your story at a glance, stick with the table instead of a pie chart or a graph. Your busy audience will appreciate the clarity.
- **Test for statistical significance:** Sometimes two datasets will look different, but you will need a way to test whether the difference is real and important. So remember to run statistical tests to see how much confidence you can place in that difference.
- **Pay attention to sample size**: Gather lots of data. If a sample size is small, a few unusual responses can skew the results. If you find that you have too little data, be careful about using it to form judgments. Look for opportunities to collect more data, then chart those trends over longer periods.

Be the judge

In any organization, a big part of a data analyst's role is making sound judgments. When you know the limitations of your data, you can make judgment calls that help people make better decisions supported by the data. Data is an extremely powerful tool for decision-making, but if it is incomplete, misaligned, or hasn't been cleaned, then it can be misleading. Take the necessary steps to make sure that your data is complete and consistent. Clean the data before you begin your analysis to save yourself and possibly others a great amount of time and effort.

Teamwork

Before the meeting

If you are organizing the meeting, you will probably talk about the data. Before the meeting:

- Identify your objective. Establish the purpose, goals, and desired outcomes of the meeting, including any questions or requests that need to be addressed.
- Acknowledge participants and keep them involved with different points of view and experiences with the data, the project, or the business.
- Organize the data to be presented. You might need to turn raw data into accessible formats or create data visualizations.
- Prepare and distribute an agenda. We will go over this next.

Crafting a compelling agenda

A solid meeting agenda sets your meeting up for success. Here are the basic parts your agenda should include:

- Meeting start and end time
- Meeting location (including information to participate remotely, if that option is available)
- Objectives
- Background material or data the participants should review beforehand

Here's an example of an agenda for an analysis project that is just getting started:

Sample Agenda

Your name	Data Analysis Project
Phone	October 6, 2020 9:30 - 10:30 PST
Email	Group Meeting Room 1

Meeting attendees: Elon, Dae, Olivia, Kiri, Pedro

Reason for meeting: Project orientation. Set goals and draft timelines for the project.

Goals

- Read the meeting agenda
- Review project goals
- Plan project timelines

Questions

- Does anyone have any suggestions for the agenda?
- What sources of data have been identified and which variables will be tracked?
- What is the earliest milestone the team can schedule? What progress would the milestone mark?

Next steps

- What should we address in the next meeting?
-
-

Sharing your agenda ahead of time

After writing your agenda, it's time to share it with the invitees. Sharing the agenda with everyone ahead of time helps them understand the meeting goals and prepare questions, comments, or feedback. You can email the agenda or share it using another collaboration tool.

During the meeting

As the leader of the meeting, it's your job to guide the data discussion. With everyone well informed of the meeting plan and goals, you can follow these steps to avoid any distractions:

- Make introductions (if necessary) and review key messages
- Present the data
- Discuss observations, interpretations, and implications of the data
- Take notes during the meeting
- Determine and summarize next steps for the group

After the meeting

To keep the project and everyone aligned, prepare and distribute a brief recap of the meeting with next steps that were agreed upon in the meeting. You can even take it a step further by asking for feedback from the team.

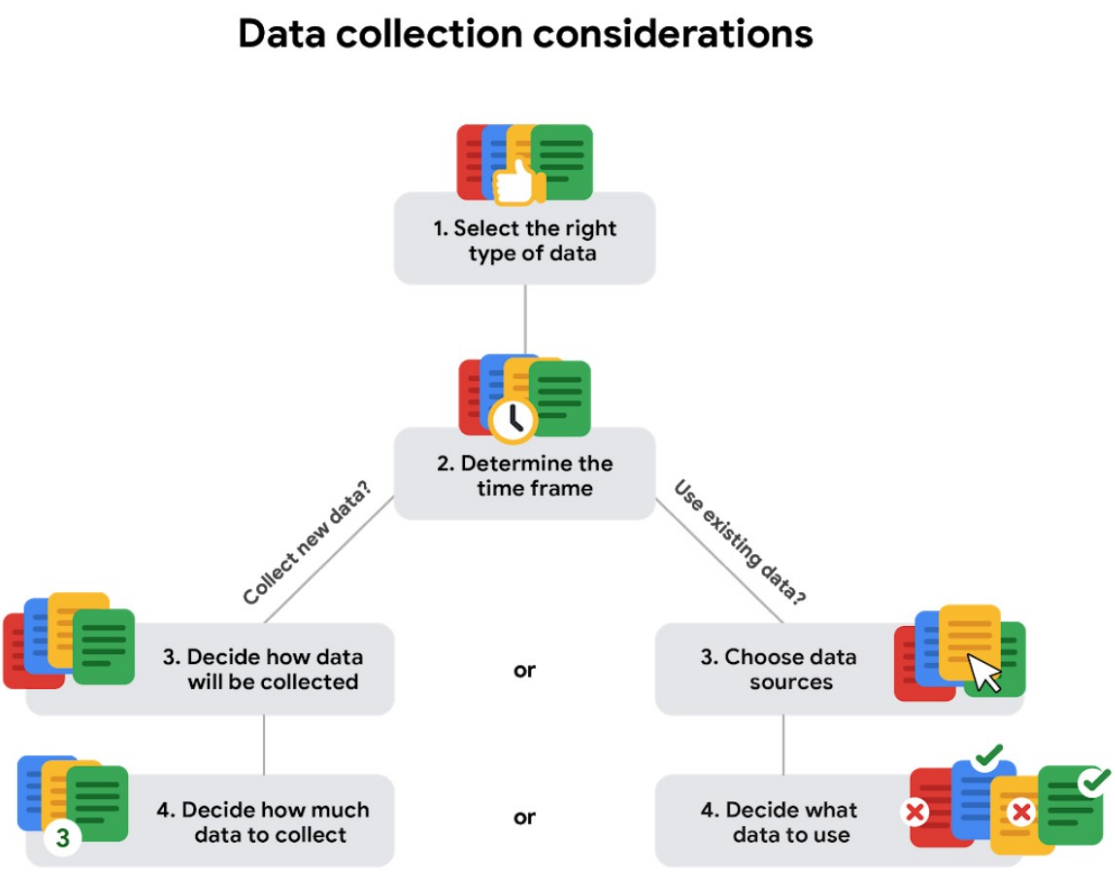
- Distribute any notes or data
- Confirm next steps and timeline for additional actions
- Ask for feedback (this is an effective way to figure out if you missed anything in your recap)

How is data collected

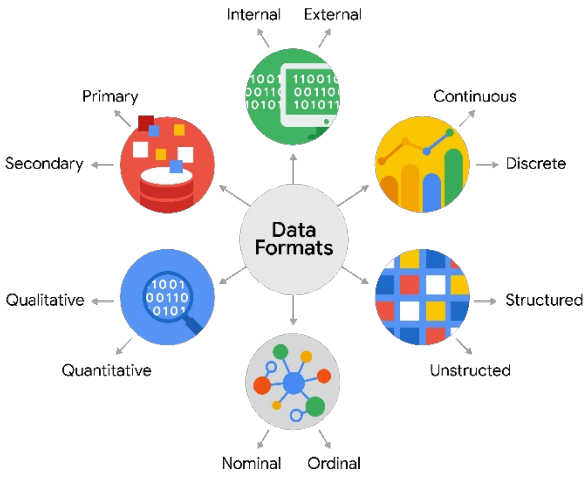
Interviews	Observations	Forms
Questionnaires	Surveys	Cookies

Data Collection Considerations

How the data will be collected	Choose data sources	Decide what data to use
How much data to collect	Select the right data type	Determine the time frame



Types of Data



Primary vs. Secondary

Data Format Classification	Definition	Examples
Primary data	Collected by a researcher from first-hand sources	- Data from an interview you conducted - Data from a survey returned from 20 participants - Data from questionnaires you got back from a group of workers
Secondary data	Gathered by other people or from other research	- Data you bought from a local data analytics firm's customer profiles - Demographic data collected by a university - Census data gathered by the federal government

Internal vs. External

Data Format Classification	Definition	Examples
Internal data	Data that lives inside a company's own systems	- Wages of employees across different business units tracked by HR - Sales data by store location - Product inventory levels across distribution centers
External data	Data that lives outside of a company or organization	- National average wages for the various positions throughout your organization - Credit reports for customers of an auto dealership

Continuous vs Discrete

Data Format Classification	Definition	Examples
Continuous data	Data that is measured and can have almost any numeric value	- Height of kids in third grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature
Discrete data	Data that is counted and has a limited number of values	- Number of people who visit a hospital on a daily basis (10, 20, 200) - Room's maximum capacity allowed - Tickets sold in the current month

Qualitative vs. Quantitative

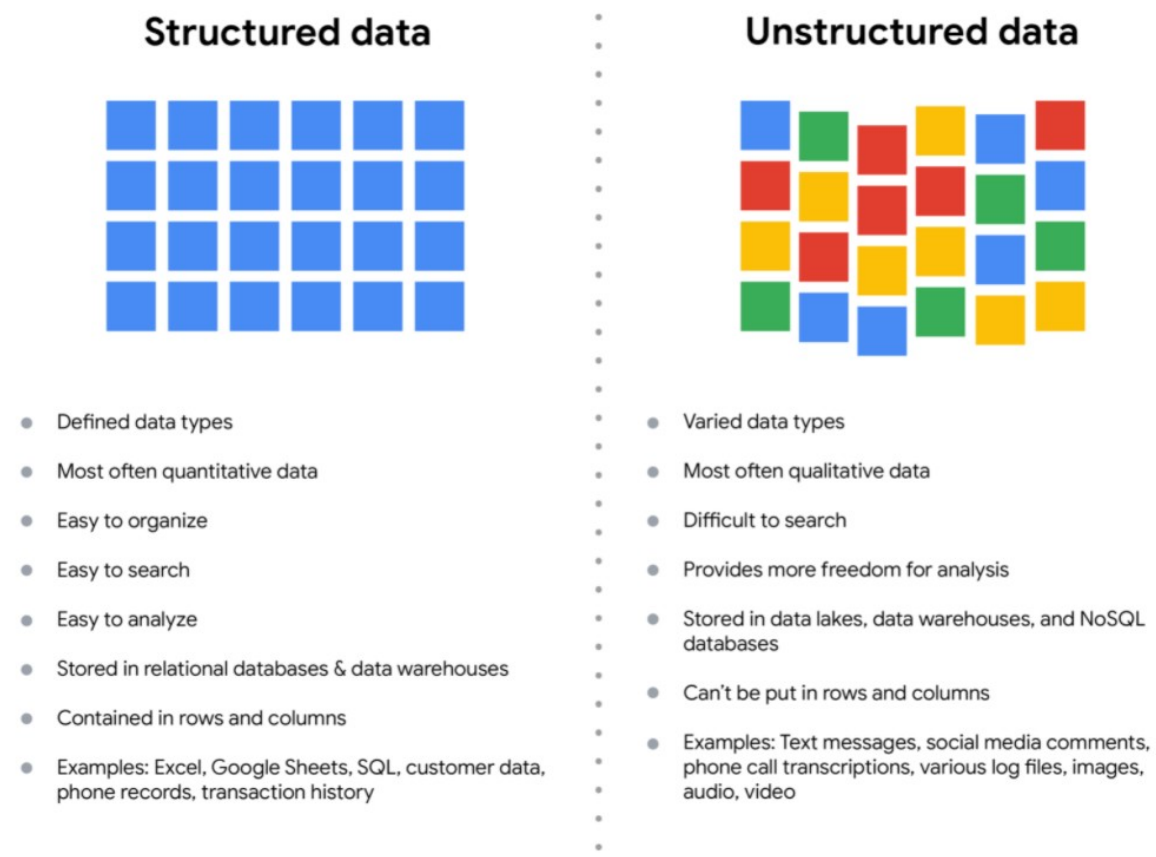
Data Format Classification	Definition	Examples
Qualitative	Subjective and explanatory measures of qualities and characteristics	- Exercise activity most enjoyed - Favorite brands of most loyal customers - Fashion preferences of young adults
Quantitative	Specific and objective measures of numerical facts	- Percentage of board certified doctors who are women - Population of elephants in Africa - Distance from Earth to Mars

Nominal vs. Ordinal

Data Format Classification	Definition	Examples
Nominal	A type of qualitative data that isn't categorized with a set order	- First time customer, returning customer, regular customer - New job applicant, existing applicant, internal applicant - New listing, reduced price listing, foreclosure
Ordinal	A type of qualitative data with a set order or scale	- Movie ratings (number of stars: 1 star, 2 stars, 3 stars) - Ranked-choice voting selections (1st, 2nd, 3rd) - Income level (low income, middle income, high income)

Structured vs. Unstructured

Data Format Classification	Definition	Examples
Structured data	Data organized in a certain format, like rows and columns	- Expense reports - Tax returns - Store inventory
Unstructured data	Data that isn't organized in any easily identifiable manner	- Social media posts - Emails - Videos



Structured data

As we described earlier, **structured data** is organized in a certain format. This makes it easier to store and query for business needs. If the data is exported, the structure goes along with the data.

Unstructured data

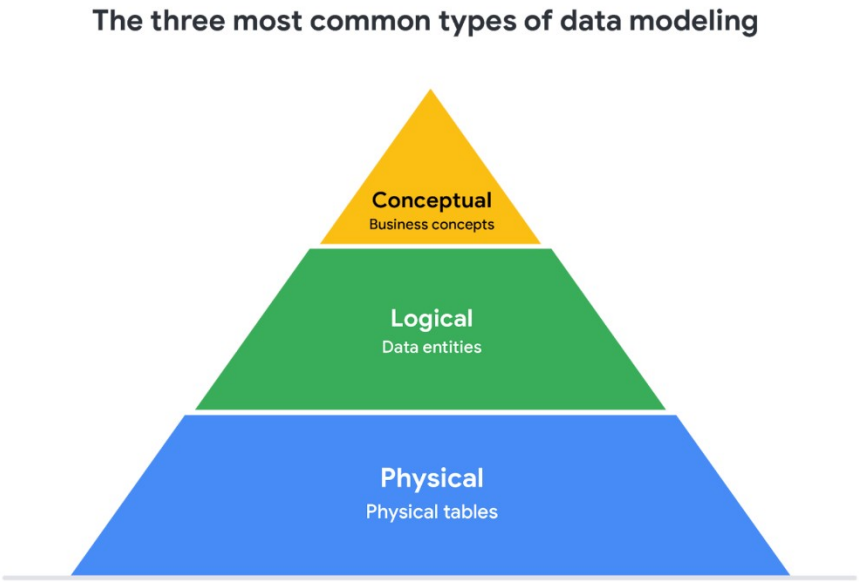
Unstructured data can't be organized in any easily identifiable manner. And there is much more unstructured than structured data in the world. Video and audio files, text files, social media content, satellite imagery, presentations, PDF files, open-ended survey responses, and websites all qualify as types of unstructured data.

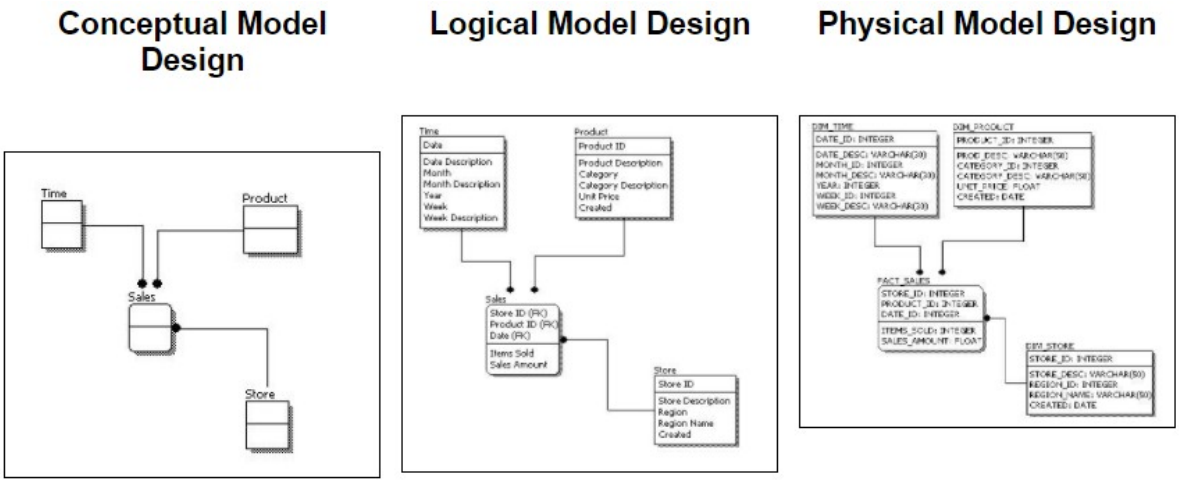
Data Modeling

Data modeling is the process of creating diagrams that visually represent how data is organized and structured. These visual representations are called data models. The 3 most common types of data modelling are:

- Conceptual Data modelling
  - It gives a high-level view of the data structure, such as how data interacts across an organization. A conceptual data model doesn't contain technical details.
- Logical Data modelling
  - It focuses on the technical details of a database such as relationships, attributes, and entities. But it doesn't spell out actual names of database tables.
- Physical Data modelling

- It depicts how a database operates. A physical data model defines all entities and attributes used.

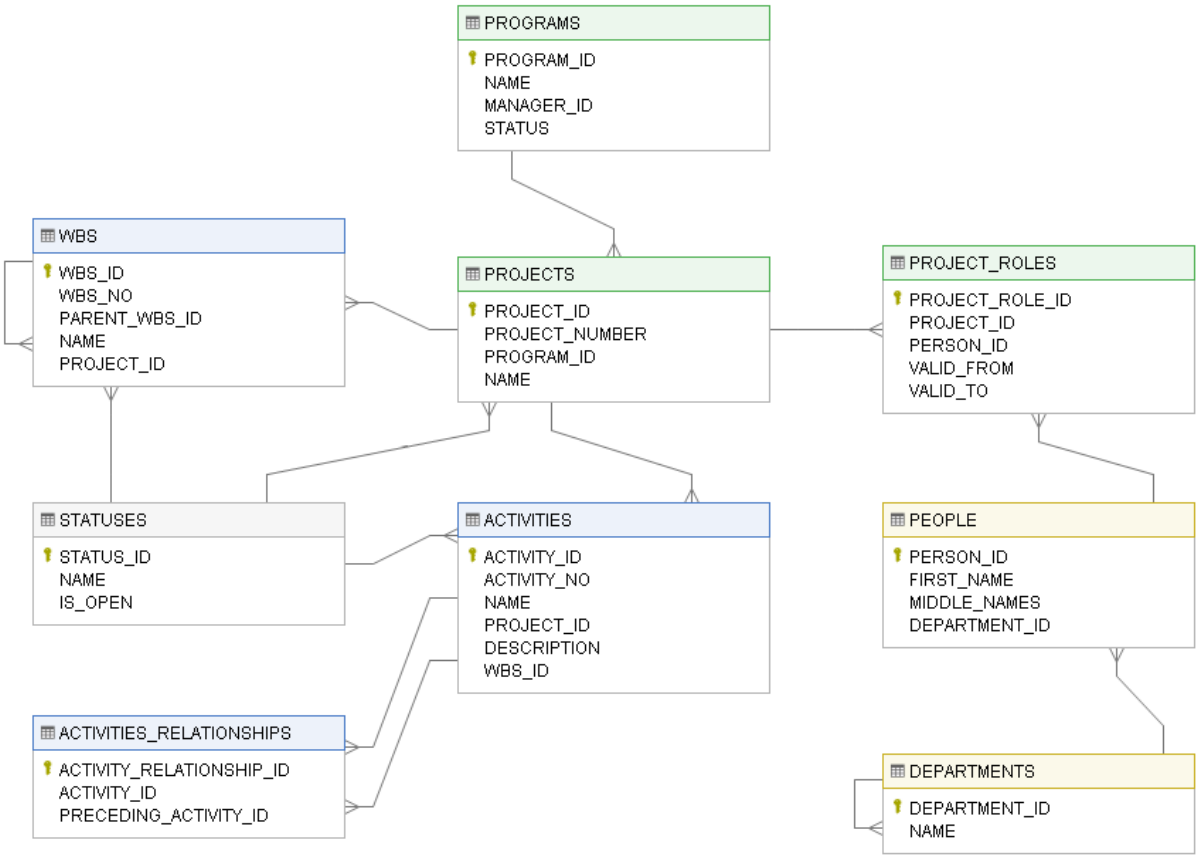




There are a lot of approaches when it comes to developing data models, but two common methods are the Entity Relationship Diagram (ERD) and the Unified Modeling Language (UML) diagram. ERDs are a visual way to understand the relationship between entities in the data model. UML diagrams are very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations, and their relationships.

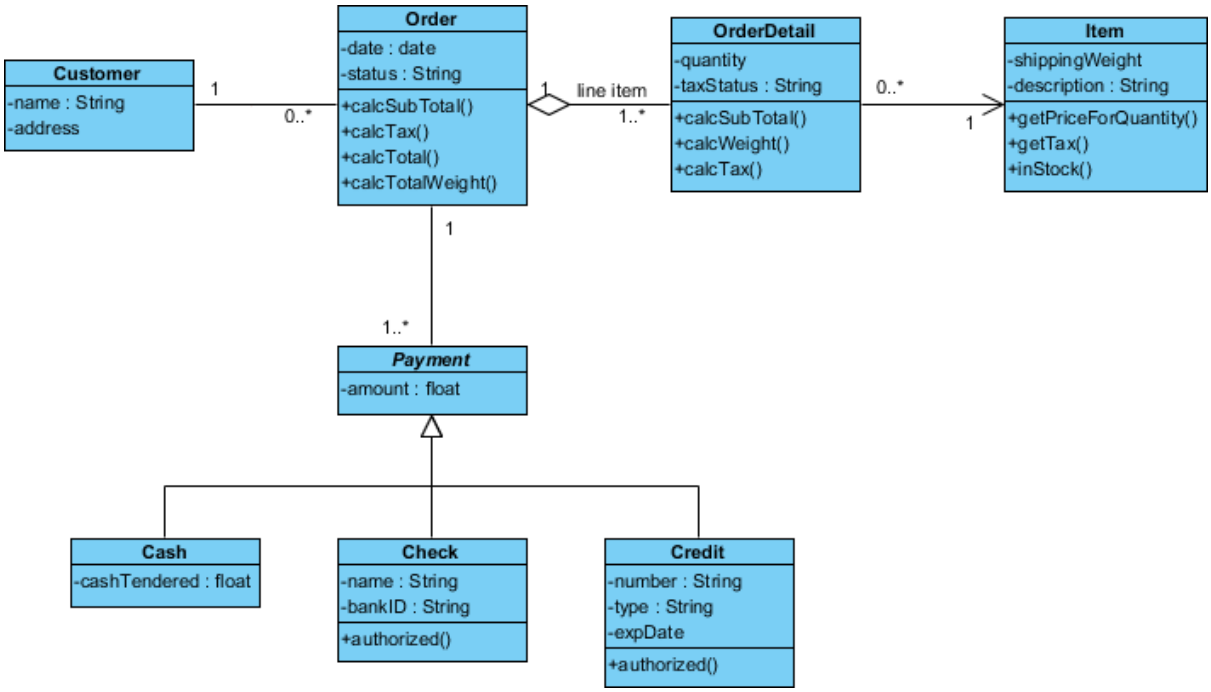
- Entity Relationship Diagrams
  - Also referred to as ER diagrams or ERDs. Entity-Relationship modeling is a default technique for modeling and the design of relational (traditional) databases. It consists of:
    - Entities
      - Representing objects (or tables in relational database),
    - Attributes
      - Of entities including data type,
    - Relationships

- Between entities/objects (or foreign keys in a database).



An example of an Entity Relationship (ER) Diagram

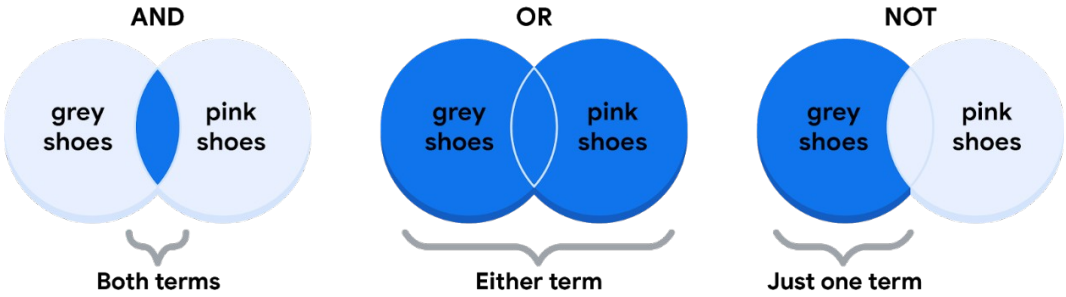
- UML Class Diagrams
  - UML (Unified Modeling Language) is a standardized family of notations for modeling and design of information systems. It was derived from various existing notations to provide a standard for software engineering. A Class diagrams are equivalent of ERDs in relational world. A class diagram consists of:
    - Classes
      - Equivalent of entity in relational world
    - Attributes
      - Of a class (same as in an ERD) including data type,
    - Methods
      - Associated to specific class, representing its behavior
    - Relationships
      - Between Objects
        - (instances of Classes) differentiated into Dependency, Association, Aggregation and Composition (equivalent to relationships in an ERD).
      - Between Classes
        - Of two kinds Generalization/Inheritance and Realization/Implementation (this has no equivalent in relational world).



An example of an UML Class Diagram

Types of Data

- Numerical
- String
- Boolean



- Wide and Long data

Wide data is preferred when	Long data is preferred when
Creating tables and charts with a few variables about each subject	Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank
Comparing straightforward line graphs	Performing advanced statistical analysis or graphing

Data Transformation

Data transformation is the process of changing the data's format, structure, or values. Data transformation usually involves:

- Adding, copying, or replicating data
- Deleting fields or records
- Standardizing the names of variables
- Renaming, moving, or combining columns in a database
- Joining one set of data with another

- Saving a file in a different format. For example, saving a spreadsheet as a comma separated values (CSV) file.

Reason for transforming data:

- Data **organization**: better organized data is easier to use
- Data **compatibility**: different applications or systems can then use the same data
- Data **migration**: data with matching formats can be moved from one system to another
- Data **merging**: data with the same organization can be merged together
- Data **enhancement**: data can be displayed with more detailed fields
- Data **comparison**: apples-to-apples comparisons of the data can then be made

Data Anonymization

Personally identifiable information, or PII, is information that can be used by itself or with other data to track down a person's identity. Data anonymization is the process of protecting people's private or sensitive data by eliminating that kind of information. Typically, data anonymization involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values. Organizations have a responsibility to protect their data and the personal information that data might contain. As a data analyst, you might be expected to understand what data needs to be anonymized, but you generally wouldn't be responsible for the data anonymization itself.

Healthcare and financial data are two of the most sensitive types of data. These industries rely a lot on data anonymization techniques. After all, the stakes are very high. That’s why data in these two industries usually goes through de-identification, which is a process used to wipe data clean of all personally identifying information. Here is a list of data that is often anonymized:

- Telephone numbers
- Names
- License plates and license numbers
- Social security numbers
- IP addresses
- Medical records
- Email addresses
- Photographs
- Account numbers

Open Data

In data analytics, open data is part of data ethics, which has to do with using data ethically. Openness refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

- Be available and accessible to the public as a complete dataset
- Be provided under terms that allow it to be reused and redistributed
- Allow universal participation so that anyone can use, reuse, and redistribute the data

Data can only be considered open when it meets all three of these standards.

Databases in Data Analytics

A relational database is a database that contains a series of tables that can be connected to show relationships. Basically, they allow data analysts to organize and link data based on what the data has in common. Tables in a relational database are connected by the fields they have in common. A primary key is an identifier that references a column in which each value is unique. By contrast, a foreign key is a field within a table that is a primary key in another table. A table can have only one primary key, but it can have multiple foreign keys. These keys are what create the relationships between tables in a relational database, which helps organize and connect data across multiple tables in the database.

Some tables don't require a primary key. A primary key may also be constructed using multiple columns of a table. This type of primary key is called a composite key. Databases use a special language to communicate called a query language. Structured Query Language (SQL) is a type of query language that lets data analysts communicate with a database.

Meta-Data

A meta data is data about the data. There are 3 types of meta data which are common in data analytics:

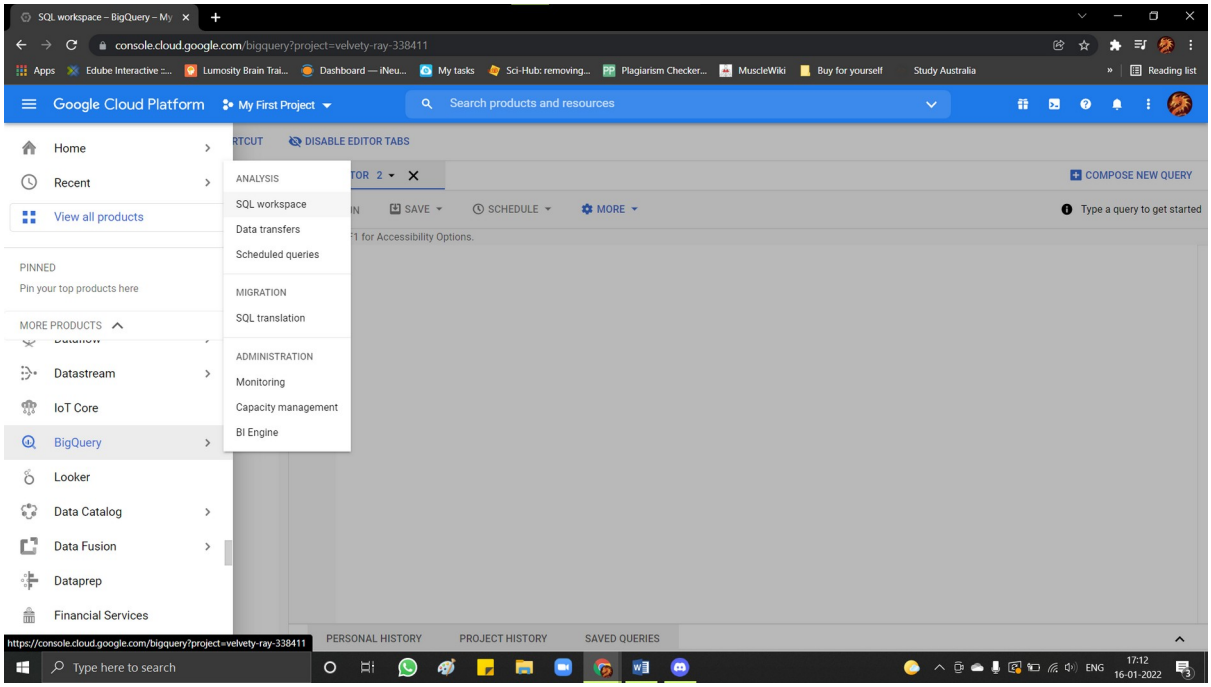
- Descriptive – Describes a piece of data and can be used to identify the piece of data at a later point in time.
- Structural – Indicates how a piece of data is organized and whether it is part of one, or more than one data collection.
- Administrative – Indicates the technical source of a digital asset.

Elements of meta data consists of the following:

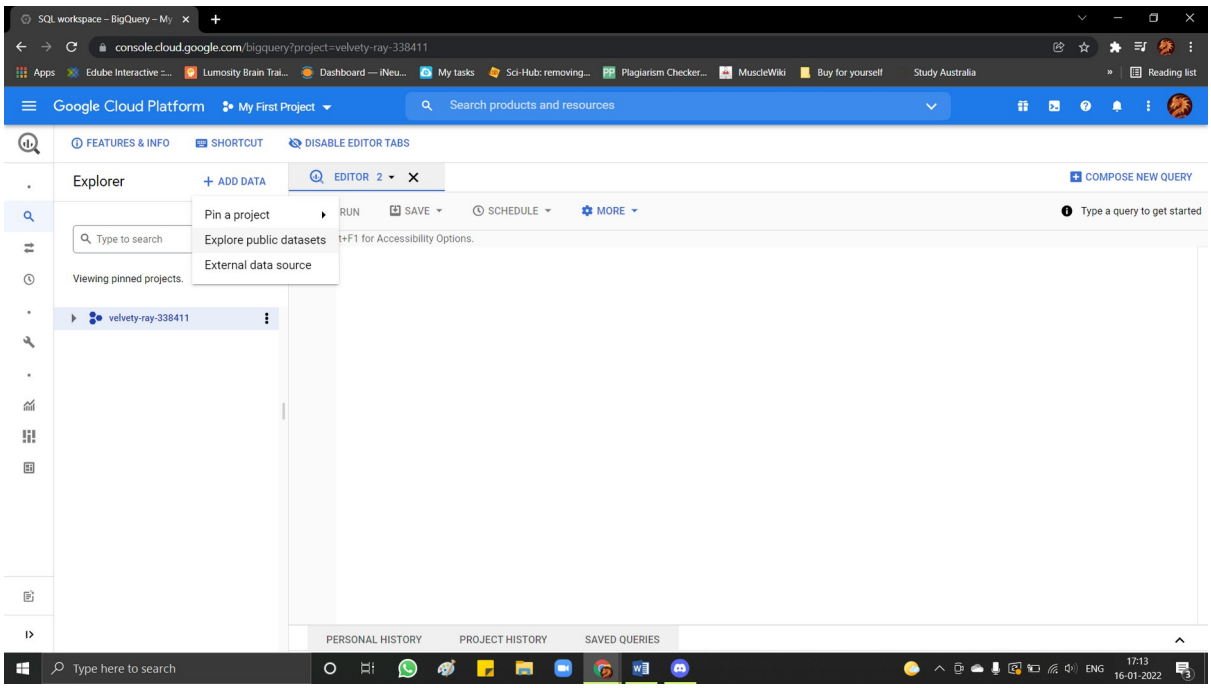
- Title and description
- Tags and categories
- Who created it and when
- Who last modified it and when
- Who can access or update it

BigQuery

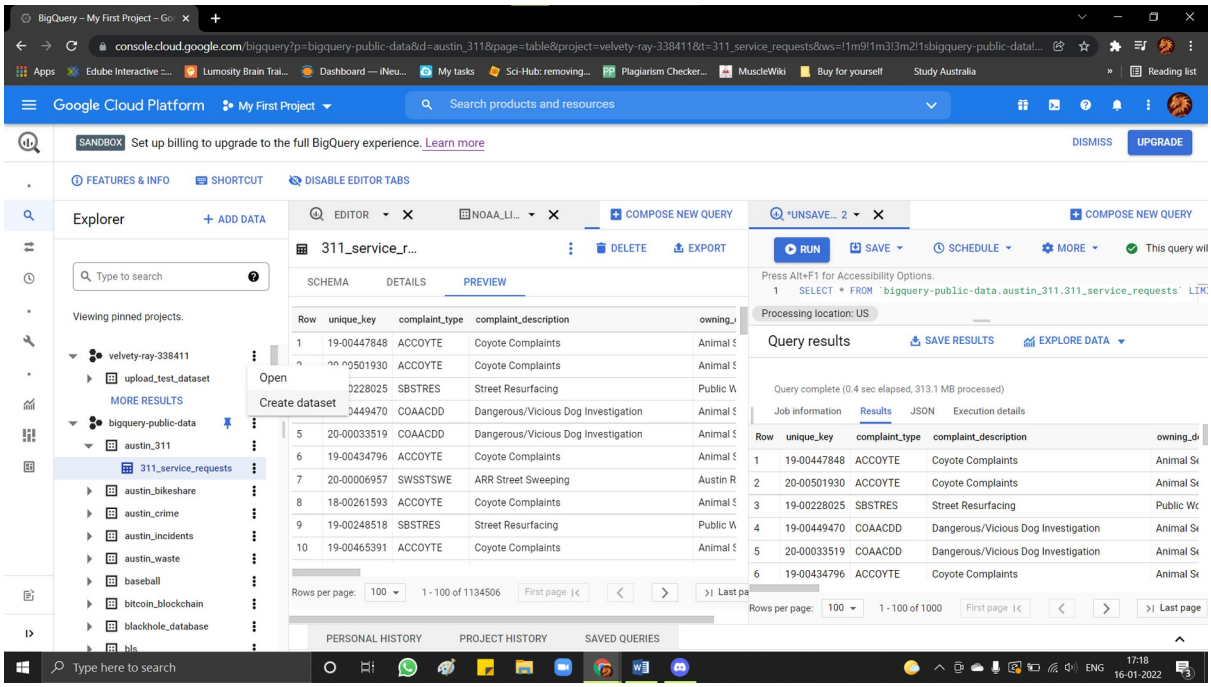
1. To open Query Workspace:



2. To add datasets to Query Workspace



3. To create a dataset



Structure of Syntax of BigQuery

```
Press Alt+F1 for Accessibility Options.
1   SELECT * FROM `bigquery-public-data.sunroof_solar.solar_potential_by_postal_code` LIMIT 1000
```

- Follows the same rules as SQL:
  - FROM
  - SELECT
  - WHERE
- Words before first dot is the Database name.
- Words after first dot represents the table name.
- * represents that we are selecting all columns from the dataset.

Best practices on SQL

- Capitalize the queries.
- Certain databases are Case Sensitive, so keep in mind of case sensitivity when querying.
- Use Single quotes (‘’) When you want strings to be identifiable in *any* SQL dialect
- Use Double quotes (“”) When your string contains an apostrophe or quotation marks
- Use Double dash (--) or Slash Asterisk (/* */) as a comment in queries.
- Use proper indentation.

Effective Organization of Data

Benefits of organizing data:

- Makes it easier to find and use.
- Helps to avoid making mistakes during your analysis
- Helps to protect your data.

How to organize data

- Foldering and Sub-foldering
- Naming it appropriately
- Archiving older folders
- Aligning naming and storage practices with team
- Develop metadata practices
- Make copies appropriately and store in different places

File naming conventions

- Work out conventions early
- Align file naming with your team
- Make sure the file names are meaningful
- Keep file names short and sweet
- Format dates in DDMMYYYY format
- Lead revision numbers with 0
- Use dashes and underscores
- Create a text file having all naming conventions

Data Security

Data security means protecting data from unauthorized access or corruption by putting safety measures in place. Usually the purpose of data security is to keep unauthorized users from accessing or viewing sensitive data.

Encryption uses a unique algorithm to alter data and make it unusable by users and applications that don't know the algorithm. This algorithm is saved as a "key" which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.

Tokenization replaces the data elements you want to protect with randomly generated data referred to as a "token." The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

Integrity of the Data

A good analysis depends on the integrity of the data. Here are some other things to watch out for:

- Data replication compromising data integrity

- Data transfer compromising data integrity
- Data manipulation compromising data integrity

A few Data constraints and their examples.

Data constraint	Definition	Examples
Data type	Values must be of a certain type: date, number, percentage, Boolean, etc.	If the data type is a date, a single number like 30 would fail the constraint and be invalid
Data range	Values must fall between predefined maximum and minimum values	If the data range is 10-20, a value of 30 would fail the constraint and be invalid
Mandatory	Values can't be left blank or empty	If age is mandatory, that value must be filled in
Unique	Values can't have a duplicate	Two people can't have the same mobile phone number within the same service area
Regular expression (regex) patterns	Values must match a prescribed pattern	A phone number must match ###-###-#### (no other characters allowed)
Cross-field validation	Certain conditions for multiple fields must be satisfied	Values are percentages and values from multiple fields must add up to 100%
Primary-key	(Databases only) value must be unique per column	A database table can't have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program.
Set-membership	(Databases only) values for a column must come from a set of discrete values	Value for a column must be set to Yes, No, or Not Applicable
Foreign-key	(Databases only) values for a column must be unique values coming from a column in another table	In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table
Accuracy	The degree to which the data conforms to the actual entity being measured or described	If values for zip codes are validated by street location, the accuracy of the data goes up.
Completeness	The degree to which the data contains all desired components or measures	If data for personal profiles required hair and eye color, and both are collected, the data is complete.
Consistency	The degree to which the data is repeatable from different points of entry or collection	If a customer has the same address in the sales and repair databases, the data is consistent.

You can gain powerful insights and make accurate conclusions when data is well-aligned to business objectives. 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.

Clean data + alignment to business objective = accurate conclusions

Alignment to business objective + additional data cleaning = accurate conclusions

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

What to do when there is an issue with the data?

Data issue 1: no data

Possible Solutions	Examples of solutions in real life
Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data.	If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees.
If there isn't time to collect data, perform the analysis using proxy data from other datasets. *This is the most common workaround.*	If you are analyzing peak travel times for commuters but don't have the data for a particular city, use the data from another city with a similar size and demographic.

Data issue 2: too little data

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

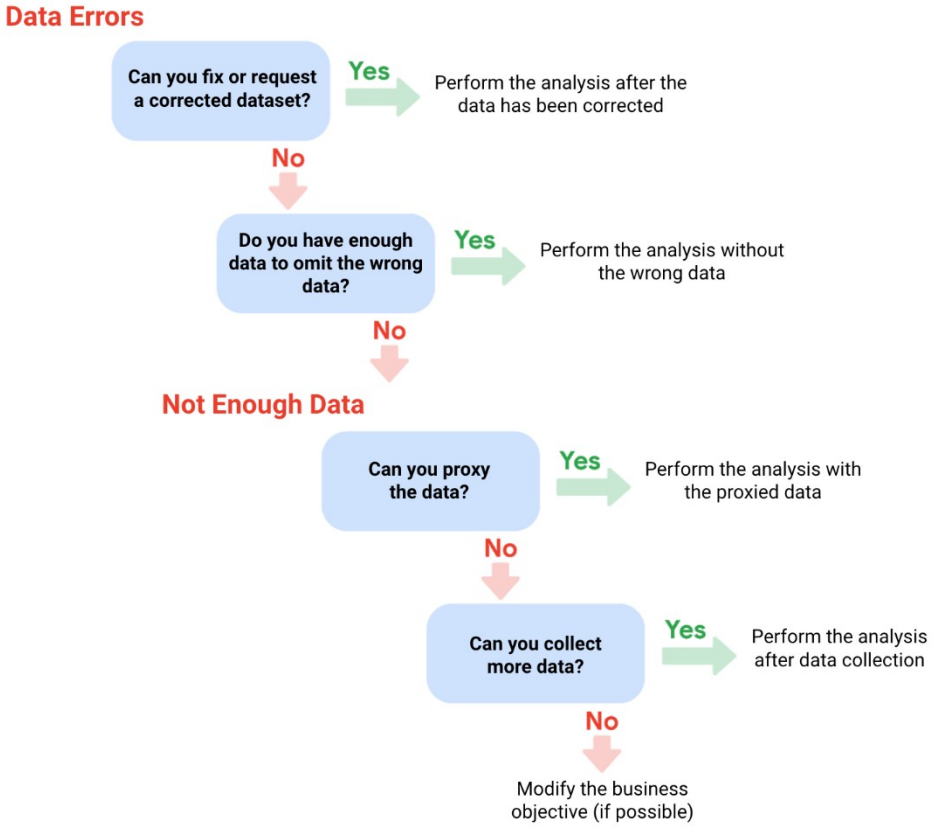
Data issue 3: wrong data, including data with errors*

Possible Solutions	Examples of solutions in real life
If you have the wrong data because requirements were misunderstood, communicate the requirements again.	If you need the data for female voters and received the data for male voters, restate your needs.
Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors.	If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values.
If you can't correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won't cause systematic bias.	If your dataset was translated from a different language and some of the translations don't make sense, ignore the data with bad translation and go ahead with the analysis of the other data.

** Important note: sometimes data with errors can be a warning sign that the data isn't reliable. Use your best judgment.*

Certain terminologies to remember:

Terminology	Definitions
Population	The entire group that you are interested in for your study. For example, if you are surveying people in your company, the population would be all the employees in your company.
Sample	A subset of your population. Just like a food sample, it is called a sample because it is only a taste. So if your company is too large to survey every individual, you can survey a representative sample of your population.
Margin of error	Since a sample is used to represent a population, the sample's results are expected to differ from what the result would have been if you had surveyed the entire population. This difference is called the margin of error. The smaller the margin of error, the closer the results of the sample are to what the result would have been if you had surveyed the entire population.
Confidence level	How confident you are in the survey results. For example, a 95% confidence level means that if you were to run the same survey 100 times, you would get similar results 95 of those 100 times. Confidence level is targeted before you start your study because it will affect how big your margin of error is at the end of your study.
Confidence interval	The range of possible values that the population's result would be at the confidence level of the study. This range is the sample result +/- the margin of error.
Statistical significance	The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance.



Things to remember when determining the size of the sample

When figuring out a sample size, here are things to keep in mind:

- 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. 99% is ideal.

Increase the sample size to meet specific needs of your project:

- 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

The limit of 30 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.

Sample size will vary based on the type of business problem you are trying to solve. Large sample size has a higher cost. Larger the sample size, the more statistically significant would be the result. A statistical power of the data is calculated. It is a unity value, hence if a dataset result has the statistical value of 0.6, it means that it has 60% chances of the result being accurate and 40% chance of wrong. Normally a statistical power of 0.8 or 80% is considered to be statistically significant.

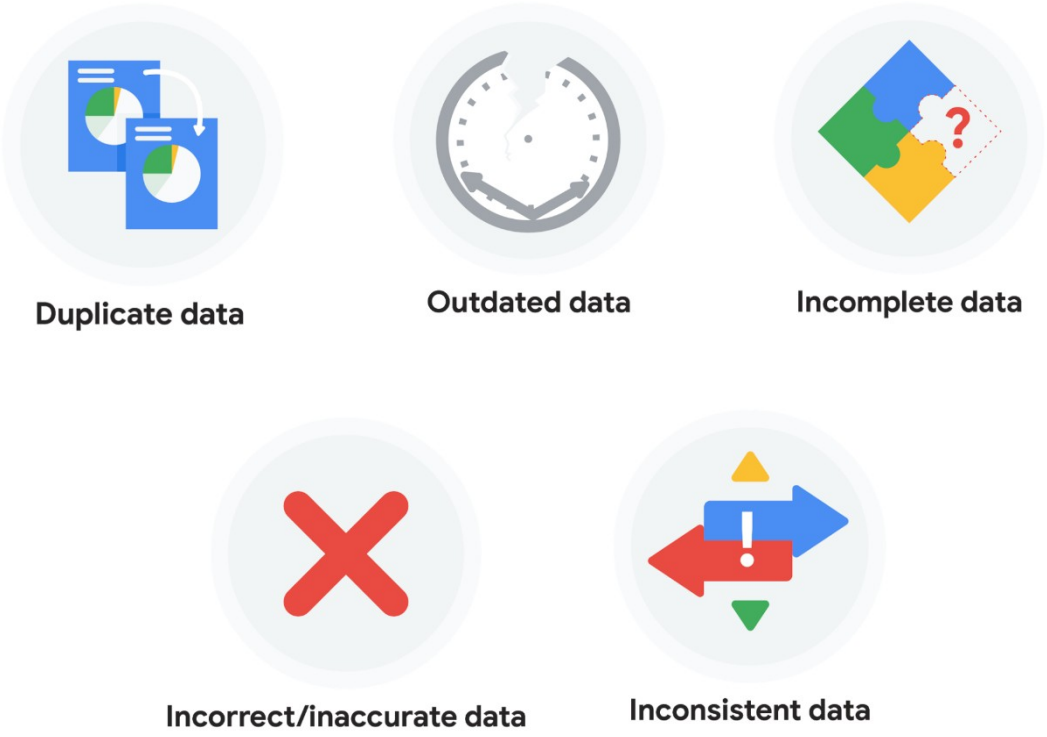
Sample Size is calculated on the basis of Population size, Confidence level and Margin of Error.

Examples of types of datasets that can serve as alternate data sources

- Proxy data samples
- Open or Public datasets
- CSV, JSON, SQLite, and BigQuery datasets

Dirty Data

Dirty data is data that is incomplete, incorrect, or irrelevant to the problem you are trying to solve. Types of dirty data includes:



Data cleaning tools and techniques

- Creating a copy before deleting certain data.
- Have a constant format throughout.
- Remove duplicates.
- Remove irrelevant data.
- Remove extra spaces and blanks.
- Fixing misspellings.
- Fixing inconsistent capitalization.
- Fixing incorrect punctuation and other typos.
- Removing formatting when dataset has data from different sources.

Common Data cleaning pitfalls

Features of SQL and Spreadsheet

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.

Steps to take a big picture of the project

1. Consider the business problem.
2. Consider the goal of the project.
3. Consider the data (whether it is capable of solving the problem).

Data cleaning verification checklist

1. Correct the most common mistakes

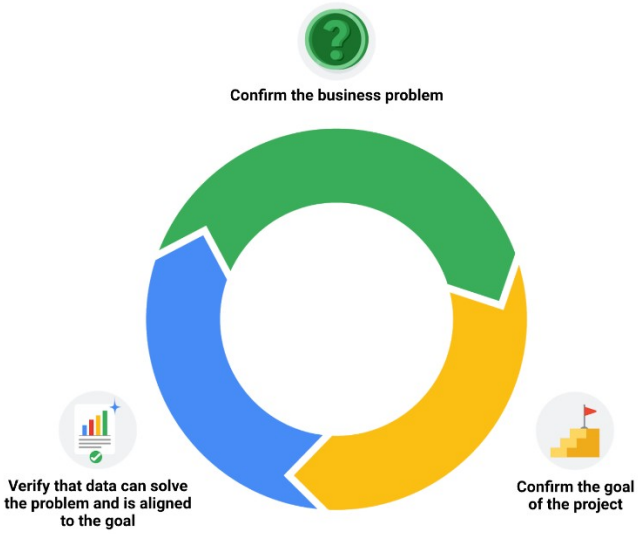
Make sure you identified the most common problems and corrected them, including:

- 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?

2. Review the goal of the project

Once you have finished these data cleaning tasks, it is a good idea to review the goal of your project and confirm that your data is still aligned with that goal. This is a continuous process. Three steps to keep in mind are:

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



Changelogs

A changelog can build on your automated version history by giving you an even more detailed record of your work. This is where data analysts record all the changes they make to the data. 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

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.

Analysis of Data

Analysis is the process used to make sense of the data collected. It means taking the right steps to proceed and think about your data in different ways. The goal of analysis is to identify trends and relationships within the data so that you can accurately answer the question you're asking. There are 4 phases to a proper analysis:

- Organize the data
- Format and adjust the data
- Get input from others
- Transform the data

Best practices for searching online:

- Thinking skills
- Data analytical terms
- Basic knowledge of tools

Sorting vs Filtering

Sorting is when you arrange data into a meaningful order to make it easier to understand, analyze, and visualize. It ranks your data based on a specific metric you choose.

Filtering is used when you are only interested in seeing data that meets a specific criterion, and hiding the rest. Filtering is really useful when you have lots of data. You can save time by zeroing in on the data that is really important or the data that has bugs or errors.

Transforming data in SQL

In this reading, you will go over the conversions that can be done using the CAST function. There are also more specialized functions like COERCION to work with big numbers, and UNIX_DATE to work with dates. UNIX_DATE returns the number of days that have passed since January 1, 1970 and is used to compare and work with dates across multiple time zones. You will likely use CAST most often. Certain common conversions are:

Starting with	CAST function can convert to:
Numeric (number)	- Integer - Numeric (number) - Big number - Floating integer - String
String	- Boolean - Integer - Numeric (number) - Big number - Floating integer - String - Bytes - Date - Date time - Time - Timestamp
Date	- String - Date - Date time - Timestamp

VLOOKUP

Functions can be used to quickly find information and perform calculations using specific values. VLOOKUP, or Vertical Lookup, searches for a certain value in a spreadsheet column and returns a corresponding piece of information from the row in which the searched value is found. We use VLOOKUP when:

- Populating data in a spreadsheet
- Merging data from one spreadsheet with data in another

VLOOKUP syntax

A VLOOKUP function is available in both Microsoft Excel and Google Sheets. You will be introduced to the general syntax in Google Sheets. (You can refer to the resources at the end of this reading for more information about VLOOKUP in Microsoft Excel.)

```
VLOOKUP(10003, A2:B26, 2, FALSE)
```

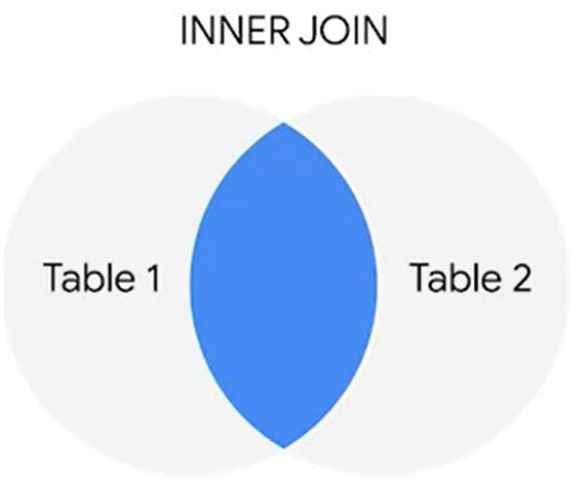
Here is the syntax.

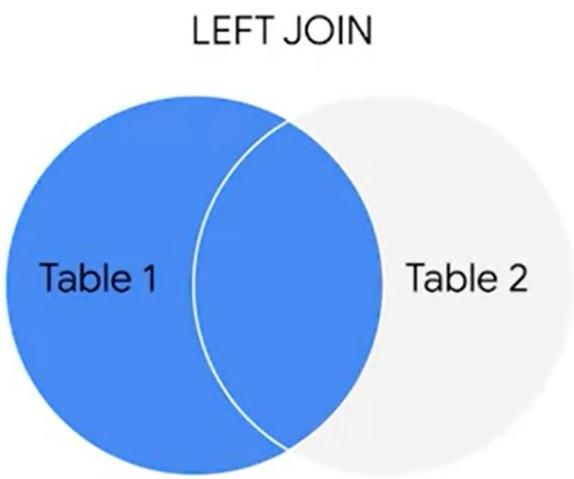
```
VLOOKUP(search_key, range, index, [is_sorted])
```

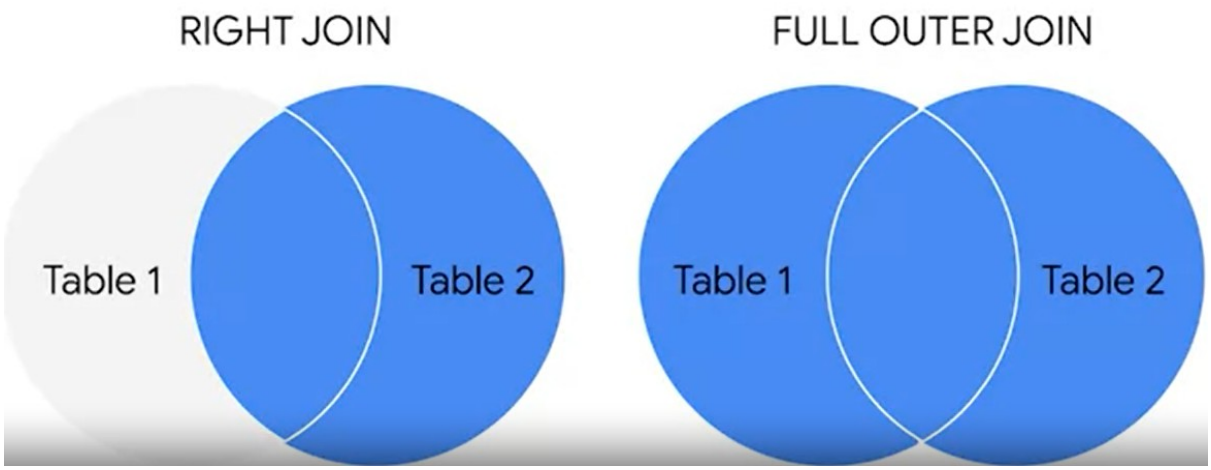
Remember

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)

JOIN IN SQL







The general JOIN syntax

```
SELECT
   -- table columns from tables are inserted here
   table_name1.column_name
   table_name2.column_name
FROM
   table_name1
JOIN
   table_name2
ON table_name1.column_name = table_name2.column_name
```

INNER JOIN

INNER is *optional* in this SQL query because it is the default as well as the most commonly used JOIN operation. You may see this as JOIN only. INNER JOIN returns records if the data lives in both tables. For example, if you use INNER JOIN for the 'customers' and 'orders' tables and match the data using the customer_id key, you would combine the data for each customer_id that exists in both tables. If a customer_id exists in the customers table but not the orders table, data for that customer_id isn't joined or returned by the query.

```
SELECT
   customers.customer_name,
   orders.product_id,
   orders.ship_date
FROM
   customers
INNER JOIN
   orders
ON customers.customer_id = orders.customer_id
```

LEFT JOIN

You may see this as LEFT OUTER JOIN, but most users prefer LEFT JOIN. Both are correct syntax. LEFT JOIN returns all the records from the left table and only the matching records from the right table. Use LEFT JOIN whenever you need the data from the entire first table and values from the second table, if they exist. For example, in the query below, LEFT JOIN will return customer_name with the corresponding sales_rep, if it is available. If there is a customer who did not interact with a sales representative, that customer would still show up in the query results but with a NULL value for sales_rep.

```
SELECT
   customers.customer_name,
   sales.sales_rep
FROM
   customers
LEFT JOIN
   sales
ON customers.customer_id = sales.customer_id
```

RIGHT JOIN

You may see this as RIGHT OUTER JOIN or RIGHT JOIN. RIGHT JOIN returns all records from the right table and the corresponding records from the left table. Practically speaking, RIGHT JOIN is rarely used. Most people simply switch the tables and stick with LEFT JOIN. But using the previous example for LEFT JOIN, the query using RIGHT JOIN would look like the following:

```
SELECT
   sales.sales_rep,
   customers.customer_name
FROM
  sales
RIGHT JOIN
  customers
ON sales.customer_id = customers.customer_id
```

FULL OUTER JOIN

You may sometimes see this as FULL JOIN. FULL OUTER JOIN returns all records from the specified tables. You can combine tables this way, but remember that this can potentially be a large data pull as a result. FULL OUTER JOIN returns all records from *both* tables even if data isn't populated in one of the tables. For example, in the query below, you will get all customers and their products' shipping dates. Because you are using a FULL OUTER JOIN, you may get customers returned without corresponding shipping dates or shipping dates without corresponding customers. A NULL value is returned if corresponding data doesn't exist in either table.

```
SELECT
   customers.customer_name,
   orders.ship_date
FROM
   customers
FULL OUTER JOIN
   orders
ON customers.customer_id = orders.customer_id
```

COUNT and COUNT DISTINCT

Count in SQL is a query that returns the number of rows in a specified range. Count distinct on the other hand is a query in SQL that only return the distinct values in a specified range.

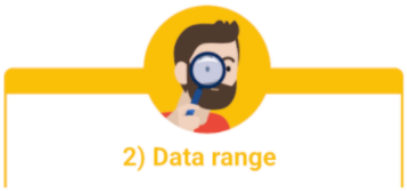
Alias in SQL

Aliases are used in SQL queries to create temporary names for a column or table. Aliases make referencing tables and columns in your SQL queries much simpler when you have table or column names that are too long or complex to make use of in queries. Aliasing is the process of using aliases. In SQL queries, aliases are implemented by making use of the AS command. The basic syntax for the AS command can be seen in the following query for aliasing a table:

```
SELECT column_name(s)
FROM table_name AS alias_name;
```

Types of Data Validation

- **Purpose**: Check that the data matches the data type defined for a field.
- **Example**: Data values for school grades 1-12 must be a numeric data type.
- **Limitations**: The data value 13 would pass the data type validation but would be an unacceptable value. For this case, data range validation is also needed.

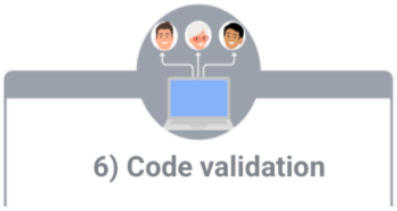


- **Purpose**: Check that the data falls within an acceptable range of values defined for the field.
- **Example**: Data values for school grades should be values between 1 and 12.
- **Limitations**: The data value 11.5 would be in the data range and would also pass as a numeric data type. But, it would be unacceptable because there aren't half grades. For this case, data constraint validation is also needed.

- **Purpose**: Check that the data meets certain conditions or criteria for a field. This includes the type of data entered as well as other attributes of the field, such as number of characters.
- **Example**: Content constraint: Data values for school grades 1-12 must be whole numbers.
- **Limitations**: The data value 13 is a whole number and would pass the content constraint validation. But, it would be unacceptable since 13 isn't a recognized school grade. For this case, data range validation is also needed.

- **Purpose**: Check that the data makes sense in the context of other related data.
- **Example**: Data values for product shipping dates can't be earlier than product production dates.
- **Limitations**: Data might be consistent but still incorrect or inaccurate. A shipping date could be later than a production date and still be

- **Purpose**: Check that the data follows or conforms to a set structure.
- **Example**: Web pages must follow a prescribed structure to be displayed properly.
- **Limitations**: A data structure might be correct with the data still incorrect or inaccurate. Content on a web page could be displayed properly and still contain the wrong information.



- **Purpose:** Check that the application code systematically performs any of the previously mentioned validations during user data input.
- **Example:** Common problems discovered during code validation include: more than one data type allowed, data range checking not done, or ending of text strings not well defined.
- **Limitations:** Code validation might not validate all possible variations with data input.

Temporary Tables

Temporary tables are exactly what they sound like—temporary tables in a SQL database that aren't stored permanently. A few features of temp tables are:

- They are automatically deleted from the database when you end your SQL session.
- They can be used as a holding area for storing values if you are making a series of calculations. This is sometimes referred to as pre-processing of the data.
- They can collect the results of multiple, separate queries. This is sometimes referred to as data staging. Staging is useful if you need to perform a query on the collected data or merge the collected data.
- They can store a filtered subset of the database. You don't need to select and filter the data each time you work with it. In addition, using fewer SQL commands helps to keep your data clean.

Temporary table creation in BigQuery

Temporary tables can be created using different clauses. In BigQuery, the **WITH** clause can be used to create a temporary table. The general syntax for this method is as follows:

```
WITH
new_table_data AS (

SELECT *

FROM
Existing_table

WHERE
Tripduration >=60
)
```

Temporary table creation in other databases (not supported in BigQuery)

The following method isn't supported in BigQuery, but most other versions of SQL databases support it, including SQL Server and mySQL. Using **SELECT** and **INTO**, you can create a temporary table based on conditions defined by a **WHERE** clause to locate the information you need for the temporary table. The general syntax for this method is as follows:

```
SELECT
*
INTO
AfricaSales
FROM
GlobalSales
WHERE
Region = "Africa"
```

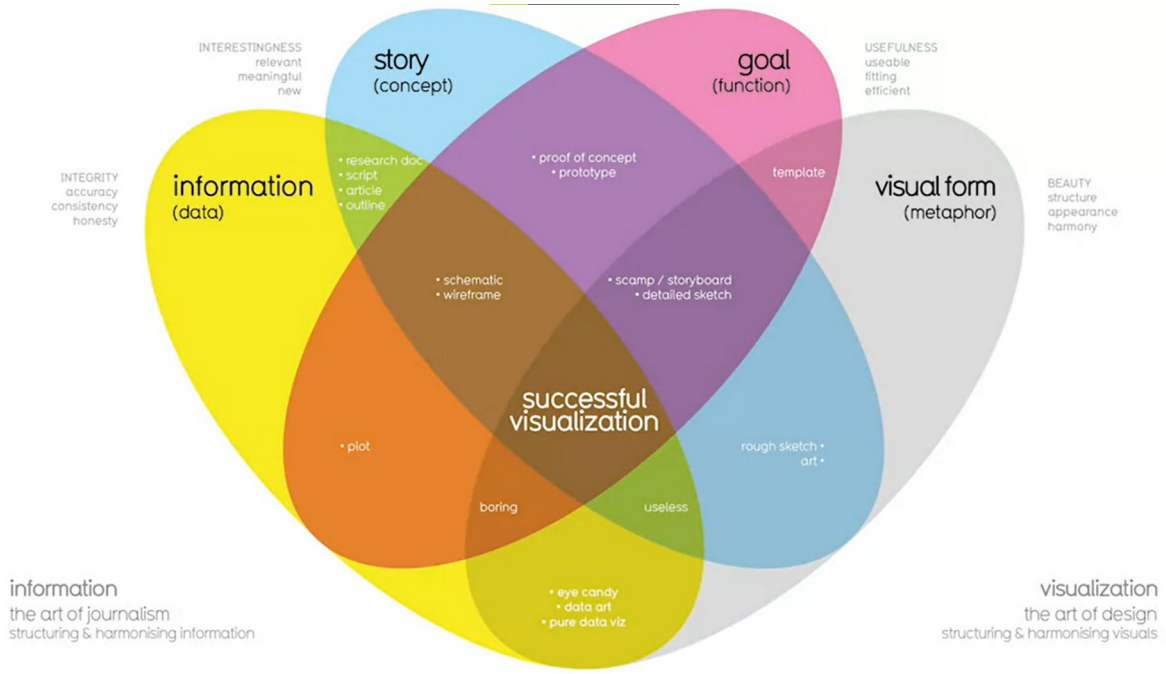
Best practices when working with temporary tables

- **Global vs. local temporary tables:** Global temporary tables are made available to all database users and are deleted when all connections that use them have closed. Local temporary tables are made available only to the user whose query or connection established the temporary table. You will most likely be working with local temporary tables. If you have created a local temporary table and are the only person using it, you can drop the temporary table after you are done using it.
- **Dropping temporary tables after use:** Dropping a temporary table is a little different from deleting a temporary table. Dropping a temporary table not only removes the information contained in the rows of the table, but removes the table variable definitions (columns) themselves. Deleting a temporary table removes the rows of the table but leaves the table definition and columns ready to be used again. Although local temporary tables are dropped after you end your SQL session, it may not happen immediately. If a lot of processing is happening in the database, dropping your temporary tables after using them is a good practice to keep the database running smoothly.

Effective Visualization

There are different frameworks for organizing your thoughts about visualization. A few of them are:

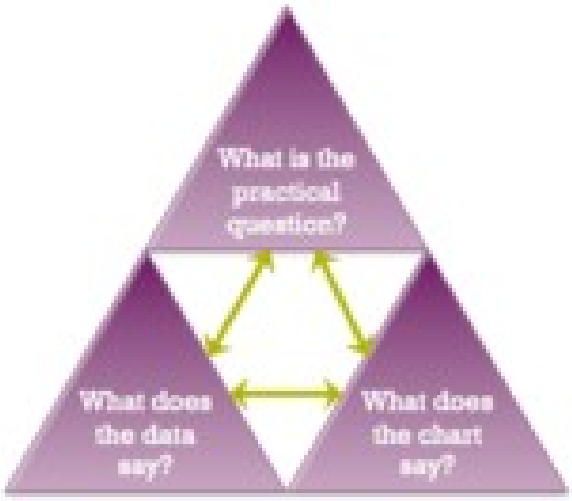
- The McCandless Method



  - Information: the data you are working with
  - Story: a clear and compelling narrative or concept
  - Goal: a specific objective or function for the visual
  - Visual form: an effective use of metaphor or visual expression

- Kaiser Fung's Junk Charts Trifecta Check-up
  - What is the practical question?
  - What does the data say?
  - What does the visual say?

Junk Charts Trifecta Checkup



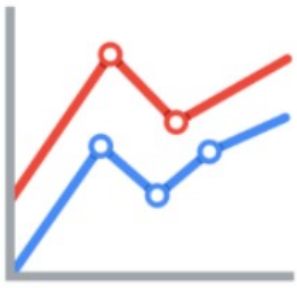
- Pre-attentive attributes

  Pre-attentive attributes are the elements of a data visualization that people recognize automatically without conscious effort. The essential, basic building blocks that make visuals immediately understandable are called marks and channels.

Marks

Marks are basic visual objects like points, lines, and shapes. Every mark can be broken down into four qualities:

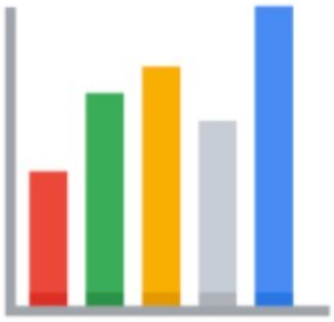
1. **Position** - Where a specific mark is in space in relation to a scale or to other marks



2. **Size** - How big, small, long, or tall a mark is

3. **Shape** - Whether a specific object is given a shape that communicates something about it

4. **Color** - What color the mark is



Channels

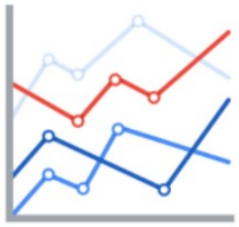
Channels are visual aspects or variables that represent characteristics of the data. Channels are basically marks that have been used to visualize data. Channels will vary in terms of how effective they are at communicating data based on three elements:

1. **Accuracy** - Are the channels helpful in accurately estimating the values being represented?

For example, color is very accurate when communicating categorical differences, like apples and oranges. But it is much less effective when distinguishing quantitative data like 5 from 5.5.

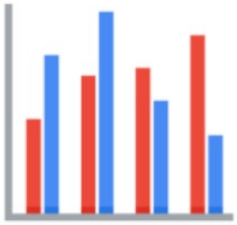
2. **Popout** - How easy is it to distinguish certain values from others?

There are many ways of drawing attention to specific parts of a visual, and many of them leverage pre-attentive attributes like line length, size, line width, shape, enclosure, hue, and intensity.



3. **Grouping** - How good is a channel at communicating groups that exist in the data?

Consider the proximity, similarity, enclosure, connectedness, and continuity of the channel.



Principle	Description
Choose the right visual	One of the first things you have to decide is which visual will be the most effective for your audience. Sometimes, a simple table is the best visualization. Other times, you need a more complex visualization to illustrate your point.
Optimize the data-ink ratio	The data-ink entails focusing on the part of the visual that is essential to understanding the point of the chart. Try to minimize non-data ink like boxes around legends or shadows to optimize the data-ink ratio.
Use orientation effectively	Make sure the written components of the visual, like the labels on a bar chart, are easy to read. You can change the orientation of your visual to make it easier to read and understand.
Color	There are a lot of important considerations when thinking about using color in your visuals. These include using color consciously and meaningfully, staying consistent throughout your visuals, being considerate of what colors mean to different people, and using inclusive color scales that make sense for everyone viewing them.
Numbers of things	Think about how many elements you include in any visual. If your visualization uses lines, try to plot five or fewer. If that isn't possible, use color or hue to emphasize important lines. Also, when using visuals like pie charts, try to keep the number of segments to less than seven since too many elements can be distracting.

What to avoid	Why
Cutting off the y-axis	Changing the scale on the y-axis can make the differences between different groups in your data seem more dramatic, even if the difference is actually quite small.
Misleading use of a dual y-axis	Using a dual y-axis without clearly labeling it in your data visualization can create extremely misleading charts.
Artificially limiting the scope of the data	If you only consider the part of the data that confirms your analysis, your visualizations will be misleading because they don't take all of the data into account.
Problematic choices in how data is binned or grouped	It is important to make sure that the way you are grouping data isn't misleading or misrepresenting your data and disguising important trends and insights.
Using part-to-whole visuals when the totals do not sum up appropriately	If you are using a part-to-whole visual like a pie chart to explain your data, the individual parts should add up to equal 100%. If they don't, your data visualization will be misleading.
Hiding trends in cumulative charts	Creating a cumulative chart can disguise more insightful trends by making the scale of the visualization too large to track any changes over time.
Artificially smoothing trends	Adding smooth trend lines between points in a scatter plot can make it easier to read that plot, but replacing the points with just the line can actually make it appear that the point is more connected over time than it actually was.

Correlation and Causation

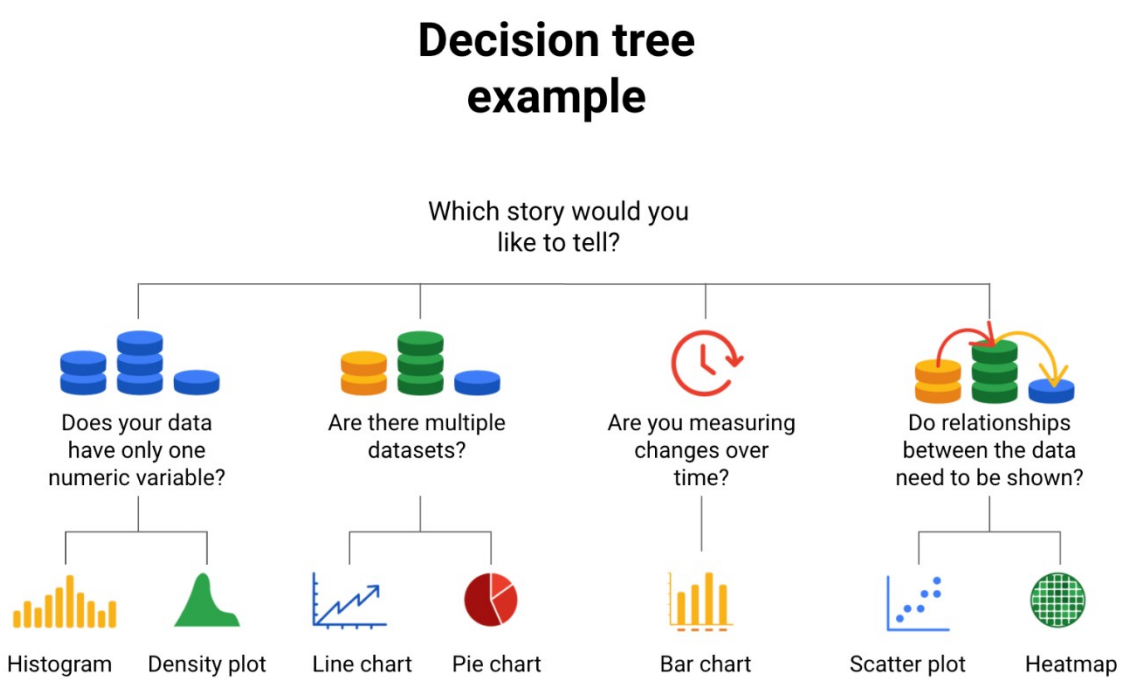
Correlation in statistics is the measure of the degree to which two variables move in relationship to each other. It is important to remember that correlation doesn't mean that one event causes another. But, it does indicate that they have a pattern with or a relationship to each other. If one variable goes up and the other variable also goes up, it is a positive correlation. If one variable goes up and the other variable goes down, it is a negative or inverse correlation. If one variable goes up and the other variable stays about the same, there is no correlation.

Causation refers to the idea that an event leads to a specific outcome. For example, Lightning causes thunder. In your data analysis, remember to:

- Critically analyse any correlations that you find
- Examine the data's context to determine if a causation makes sense (and can be supported by all of the data)
- Understand the limitations of the tools that you use for analysis

Decision Trees

A decision tree is a decision-making tool that allows you, the data analyst, to make decisions based on key questions that you can ask yourself. Each question in the visualization decision tree will help you make a decision about critical features for your visualization. Below is an example of a basic decision tree to guide you towards making a data-driven decision about which visualization is the best way to tell your story.



Elements of Art

- Line
- Shape
  - Good for depicting size contrasts.
- Color
  - Hue (the color), Intensity (brightness or dullness) and Value (How much light is reflected) make up the color
- Space
- Movement

Principles of design

- Balance
  - The design of a data visualization is balanced when the key visual elements, like color and shape, are distributed evenly. This doesn't mean that you need complete symmetry, but your visualization shouldn't have one side distracting from the other.
- Emphasis
  - Your data visualization should have a focal point, so that your audience knows where to concentrate. In other words, your visualizations should emphasize the most important data so that users recognize it first. Using color and value is one effective way to make this happen. By using contrasting colors, you can make certain that graphic elements—and the data shown in those elements—stand out.
- Movement
  - Movement can refer to the path the viewer's eye travels as they look at a data visualization, or literal movement created by animations. Movement in data

visualization should mimic the way people usually read. You can use lines and colors to pull the viewer's attention across the page.

- Pattern
  - You can use similar shapes and colors to create patterns in your data visualization.
- Repetition
  - Repeating chart types, shapes, or colors adds to the effectiveness of your visualization. The repetition of the colors helps the audience understand that there are distinct sets of data.
- Proportion
  - Proportion is another way that you can demonstrate the importance of certain data. Using various colors and sizes helps demonstrate that you are calling attention to a specific visual over others. If you make one chart in a dashboard larger than the others, then you are calling attention to it. It is important to make sure that each chart accurately reflects and visualizes the relationship among the values in it.
- Rhythm
  - This refers to creating a sense of movement or flow in your visualization. Rhythm is closely tied to the movement principle. If your finished design doesn't successfully create a flow, you might want to rearrange some of the elements to improve the rhythm.
- Variety
  - Your visualizations should have some variety in the chart types, lines, shapes, colors, and values you use. Variety keeps the audience engaged. But it is good to find balance since too much variety can confuse people. The variety you include should make your dashboards and other visualizations feel interesting and unified.
- Unity
  - This means that your final data visualization should be cohesive. If the visual is disjointed or not well organized, it will be confusing and overwhelming.

Data visualizations have three essential elements: clear meaning, a sophisticated use of contrast, and refined execution. The 4 elements of a successful visualization are:

- Information (data)
  - The information or data that you are trying to convey is a key building block for your data visualization. Without information or data, you cannot communicate your findings successfully.
- Story (concept)
  - Story allows you to share your data in meaningful and interesting ways. Without a story, your visualization is informative, but not really inspiring.
- Goal (function)
  - The goal of your data visualization makes the data useful and usable. This is what you are trying to achieve with your visualization. Without a goal, your visualization might still be informative, but can't generate actionable insights.
- Visual Form (metaphor)
  - The visual form element is what gives your data visualization structure and makes it beautiful. Without visual form, your data is not visualized yet.

4 Phases of the design process

- Empathize
  - You think about the emotions and needs of the target audience.
- Define
  - You define the audience's needs, problems, and your insights.
- Ideate
  - You generate your database ideas.
- Prototype
  - You put all charts, dashboards and other visualizations together.
- Test
  - You show to team members, get critics and then do all the stuff necessary and then show to stakeholders.

As interactive dashboards become more popular for data visualization, new importance has been placed on efficiency and user-friendliness.

Headlines, subtitles, labels, and annotations help you turn your data visualizations into more meaningful displays. When you present a visualization, they should be able to process and understand the information you are trying to share in the first five seconds. Certain techniques are:

- Headline
  - A **headline** is a line of words printed in large letters at the top of a visualization to communicate what data is being presented. It is the attention grabber that makes your audience want to read more.
- Subtitles
  - A **subtitle** supports the headline by adding more context and description. Adding a subtitle will help the audience better understand the details associated with your chart. Typically, the text for subtitles has a smaller font size than the headline.

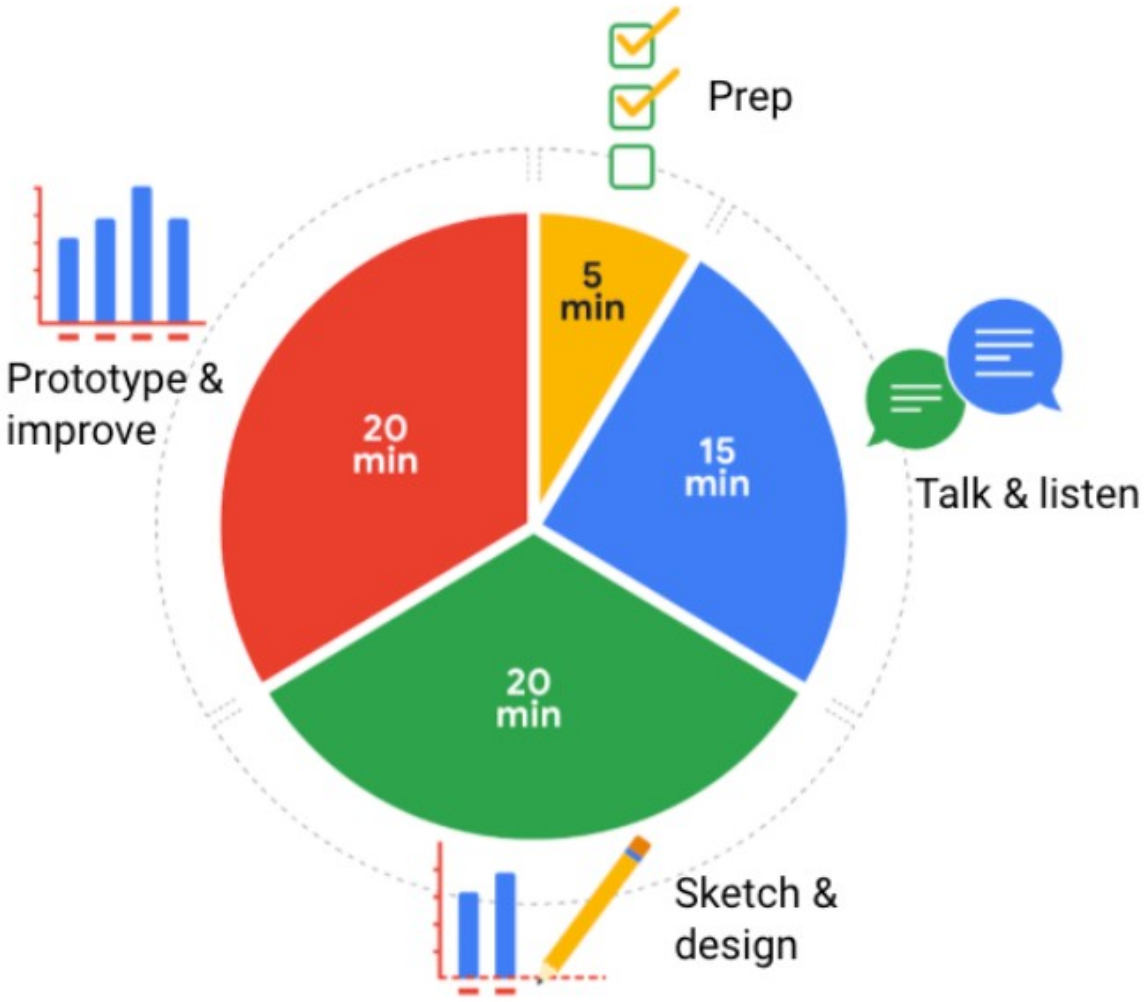
- Labels
  - A **label** in a visualization identifies data in relation to other data. Most commonly, labels in a chart identify what the x-axis and y-axis show. Always make sure you label your axes. Data can also be labeled directly in a chart instead of through a chart legend. This makes it easier for the audience to understand data points without having to look up symbols or interpret the color coding in a legend.
- Annotations
  - An **annotation** briefly explains data or helps focus the audience on a particular aspect of the data in a visualization.

Visualization components	Guidelines	Style checks
Headlines	- **Content**: Briefly describe the data - **Length**: Usually the width of the data frame - **Position**: Above the data	- Use brief language - Don't use all caps - Don't use italic - Don't use acronyms - Don't use abbreviations - Don't use humor or sarcasm
Subtitles	- **Content**: Clarify context for the data - **Length**: Same as or shorter than headline - **Position**: Directly below the headline	- Use smaller font size than headline - Don't use undefined words - Don't use all caps, bold, or italic - Don't use acronyms - Don't use abbreviations
Labels	- **Content**: Replace the need for legends - **Length**: Usually fewer than 30 characters - **Position**: Next to data or below or beside axes	- Use a few words only - Use thoughtful color-coding - Use callouts to point to the data - Don't use all caps, bold, or italic
Annotations	- **Content**: Draw attention to certain data - **Length**: Varies, limited by open space - **Position**: Immediately next to data annotated	- Don't use all caps, bold, or italic - Don't use rotated text - Don't distract viewers from the data

Ways to make Visualizations accessible

- Labeling
- Text alternatives
- Text based format
- Distinguishing
- Simplify

Chart creating time allotment



- Prep (5 min):
    - Create the mental and physical space necessary for an environment of comprehensive thinking. This means allowing yourself room to brainstorm *how* you want your data to appear while considering the amount and type of data that you have.
- Talk and listen (15 min):
    - Identify the object of your work by getting to the “ask behind the ask” and establishing expectations. Ask questions and really concentrate on feedback from stakeholders regarding your projects to help you hone how to lay out your data.
- Sketch and design (20 min):
    - Draft your approach to the problem. Define the timing and output of your work to get a clear and concise idea of what you are crafting.
- Prototype and improve (20 min):
    - Generate a visual solution and gauge its effectiveness at accurately communicating your data. Take your time and repeat the process until a final visual is produced. It is alright if you go through several visuals until you find the perfect fit.

3 Data storytelling steps

- Engage your audience
  - Capturing and holding someone's interest and attention.
- Create compelling visuals
  - One has to show the story and not just tell it and compelling visuals are the way to go.
- Tell the story in an interesting narrative
  - A narrative has a beginning, a middle, and an end. The visualizations should also have an organized beginning, middle and end.

Effective Data Story

In data analytics, data storytelling is communicating the meaning of a dataset with visuals and a narrative that is customized for a particular audience. One needs to ask a few questions when reviewing the visualizations for presentations:

- How does the visualization help set the context?
- How does the visualization help clarify the data?
- Do you notice a data visualization best practice?
- How does the visualization perform against the 5 second rule?
- How does the visualization help make a point?

Live vs Static

Identifying whether data is live or static depends on certain factors:

- How old is the data?
- How long until the insights are stale or no longer valid to make decisions?
- Does this data or analysis need updating on a regular basis to remain valuable?

Static data involves providing screenshots or snapshots in presentations or building dashboards using snapshots of data. There are pros and cons to static data.

Pros:

- Can tightly control a point-in-time narrative of the data and insight
- Allows for complex analysis to be explained in-depth to a larger audience

Cons:

- Insight immediately begins to lose value and continues to do so the longer the data remains in a static state
- Snap-shots can't keep up with the pace of data change

Live data means that you can build dashboards, reports, and views connected to automatically updated data.

Pros:

- Dashboards can be built to be more dynamic and scalable
- Gives the most up-to-date data to the people who need it at the time when they need it
- Allows for up-to-date curated views into data with the ability to build a scalable "single source of truth" for various use cases

- Allows for immediate action to be taken on data that changes frequently
- Alleviates time/resources spent on processes for every analysis

Cons:

- Can take engineering resources to keep pipelines live and scalable, which may be outside the scope of some companies' data resource allocation
- Without the ability to interpret data, you can lose control of the narrative, which can cause data chaos (i.e. teams coming to conflicting conclusions based on the same data)
- Can potentially cause a lack of trust if the data isn't handled properly

Compelling Presentation Tips

The narrative of the presentation requires the following:

- Characters
  - The people affected by your story
- Settings
  - What's going on?
  - How often its happening?
  - What tasks are involved?
  - Other background information about the data.
- Plot
  - Also called the conflict.
  - It is what creates the tension in the current situation.
  - For example: A challenge from the competitor, etc.
  - Should reveal the problem your analysis is solving.
  - Compel the characters to act.
- Big Reveal
  - Show how the characters can solve the problem they are facing by, for example, becoming more competitive, inventing a new system, or the ultimate goal of the project.
- Aha moment
  - Share recommendations and why you think the recommendation will help the company.

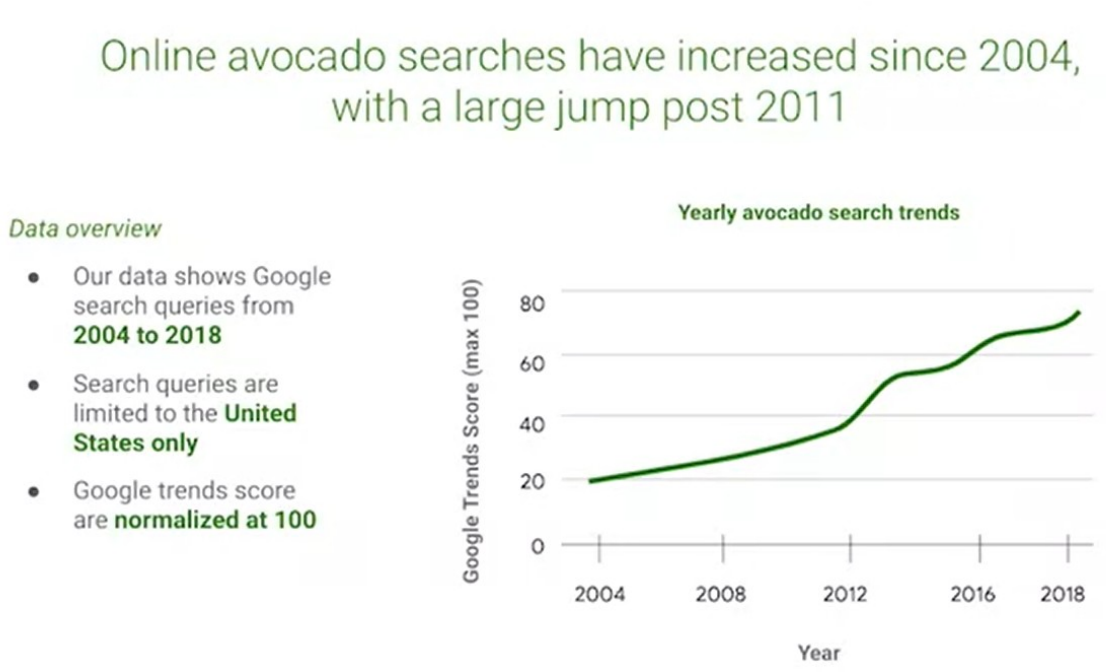
A general rule of thumb is to keep texts to **less than 5 lines** and **25 words per slide.**
Choose words carefully and keep it professional.
When presenting visuals, make sure only the data points relevant to the visual are present.

Strategic Framework for a successful presentation

The framework of your presentation starts with your understanding of the business task.

1. Understand the business task first before starting the presentation.
2. First slide should be the framing of the business task. An example would be:
   - Share an overview of the data.
   - Examine the trends using historical data.
   - Discuss any potential area for further exploration.

3. Outline the findings which would be required to solve the business problem so the viewers are not caught unexpected.
4. Showcase what business metrics we used in the data to make the audience understand the data.
5. Show data overview along with the visualization. An example:



6. Establish the initial hypothesis (the theory you are trying to prove or disapprove)
7. Explain your solution to the business problem using the examples and visualizations.
8. We can follow any method, generally McCandless method is the best:
    - Always make sure you have a title for the visualization.
    - Answer the obvious questions before the audience asks it. Work your way from high level information to minute details that is needed.
    - State the insight the visualization provides.
    - Call out data to support your insights.
    - Tell your audience why it matters.
        i. Present the possible business impact of the solution.
        ii. State the clear actions stakeholders can take.

Presentation tips

- Keep it straightforward and simple.
- Make your presentation fun.
- Make it as a story, you do not present it, you tell a story.
- Make sure you run through the presentation with someone who is going to be in the meet. Make sure to incorporate their feedback.
- Keep slides such as:
    - Cover page
    - Table of contents
    - Purpose statement

    - Tell your story
    - Conclusion
    - Appendix
- Channel your excitement
- Use the 5 second rule
- Start with broader ideas
- Preparation is key

Preparing for the Q&A

Before the presentation

1. Assemble and prepare your questions.
2. Discuss your presentation with your manager, other analysts, or other friendly contacts in your organization.
3. Ask a manager or other analysts what sort of questions were normally asked by your specific audience in the past.
4. Seek comments, feedback, and questions on the deck or the document of your analysis.
5. At least 24 hours ahead of the presentation, try and brainstorm tricky questions or unclear parts you may come across- this helps avoid surprises.
6. It never hurts to practice what you will be presenting, to account for any missing information or simply to calm your nerves.

During the presentation

1. Be prepared to respond to the things that you find and effectively and accurately explain your findings.
2. Address potential questions that may come up.
3. Avoid having a single question derail a presentation and propose following-up offline.
4. Put supplementary visualizations and content in the appendix to help answer questions.

- Listen to the whole question.
- Repeat the question if necessary
- Understand the context
- Involve the whole audience
- Keep your responses short and to the point

Python vs R language

Languages	R	Python
Common features	- Open-source - Data stored in data frames - Formulas and functions readily available - Community for code development and support	- Open-source - Data stored in data frames - Formulas and functions readily available - Community for code development and support
Unique advantages	- Data manipulation, data visualization, and statistics packages - "Scalpel" approach to data: *find packages to do what you want with the data*	- Easy syntax for machine learning needs - Integrates with cloud platforms like Google Cloud, Amazon Web Services, and Azure
Unique challenges	- Inconsistent naming conventions make it harder for beginners to select the right functions - Methods for handling variables may be a little complex for beginners to understand	- Many more decisions for beginners to make about data input/output, structure, variables, packages, and objects - "Swiss army knife" approach to data: *figure out a way to do what you want with the data*

Spreadsheets to SQL to R language

Key question	Spreadsheets	SQL	R
What is it?	A program that uses rows and columns to organize data and allows for analysis and manipulation through formulas, functions, and built-in features	A database programming language used to communicate with databases to conduct an analysis of data	A general purpose programming language used for statistical analysis, visualization, and other data analysis
What is a primary advantage?	Includes a variety of visualization tools and features	Allows users to manipulate and reorganize data as needed to aid analysis	Provides an accessible language to organize, modify, and clean data frames, and create insightful data visualizations
Which datasets does it work best with?	Smaller datasets	Larger datasets	Larger datasets
What is the source of the data?	Entered manually or imported from an external source	Accessed from an external database	Loaded with R when installed, imported from your computer, or loaded from external sources
Where is the data from my analysis usually stored?	In a spreadsheet file on your computer	Inside tables in the accessed database	In an R file on your computer
Do I use formulas and functions?	Yes	Yes	Yes
Can I create visualizations?	Yes	Yes, by using an additional tool like a database management system (DBMS) or a business intelligence (BI) tool	Yes

When to use R Studio

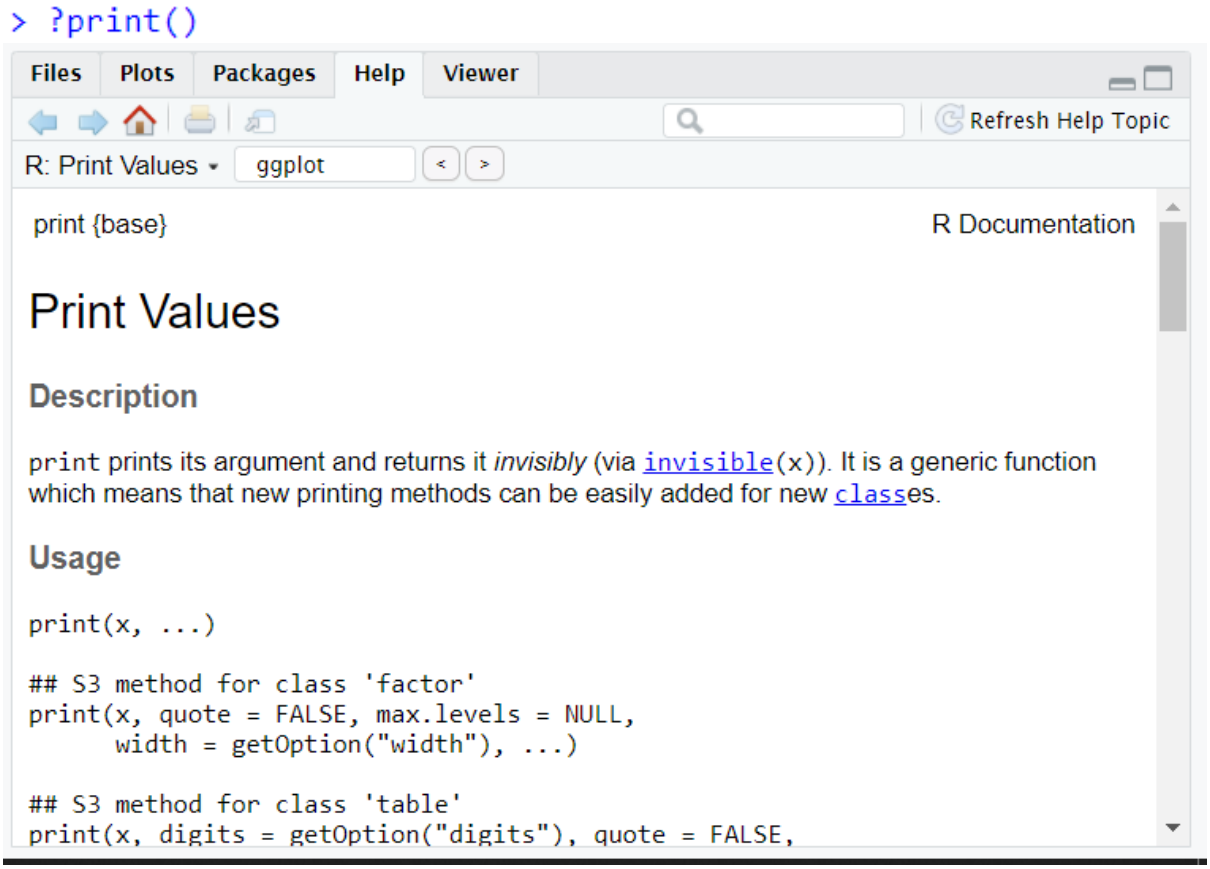
One of your core tasks as an analyst will be converting raw data into insights that are accurate, useful, and interesting. That can be tricky to do when the raw data is complex. R and RStudio are designed to handle large data sets, which spreadsheets might not be able to handle as well. RStudio also makes it easy to reproduce your work on different datasets. When you input your code, it's simple to just load a new dataset and run your scripts again.

You can also create more detailed visualizations using RStudio. When the data is spread across multiple categories or groups, it can be challenging to manage your analysis, visualize trends, and build graphics. And the more groups of data that you need to work with, the harder those tasks become. That's where RStudio comes in.

Basics of R Programming Language

```
> print("coding in R")
[1] "coding in R"
```

To print a string.



To get to know what a function does.

Fundamental elements of R Language

- Functions
  - A body of reusable code used to perform specific tasks in R. It begins with a function name and followed by arguments in parenthesis.
- Comments
  - A comment is a statement used to note something which will not affect the code in any way.
- Variables

- A representation of a value in R that can be stored for later use during programming. Variables can also be called objects.

- Data types
  - These are the categories of different data. It can be of different types such as integer, logical, date, date time, etc.
- Vectors
  - A vector is a group of data elements of the same type stored in a sequence in R. It can be made using the function “c”.
- Pipes
  - A pipe is a tool in R for expressing a sequence of multiple operations, represented by % > %. It is used to apply the output of one function to another function.

Lists and Atomic Vectors

Creating lists

Lists are different from atomic vectors because their elements can be of any type—like dates, data frames, vectors, matrices, and more. Lists can even contain other lists.

You can create a list with the **list()** function. Similar to the c() function, the list() function is just `list` followed by the values you want in your list inside parentheses: **list(x, y, z, ...)**. In this example, we create a list that contains four different kinds of elements: character (`"a"`), integer (`1L`), double (`1.5`), and logical (`TRUE`).

```
list("a", 1L, 1.5, TRUE)
```

Like we already mentioned, lists can contain other lists. If you want, you can even store a list inside a list inside a list—and so on.

```
list(list(list(1 , 3, 5)))
```

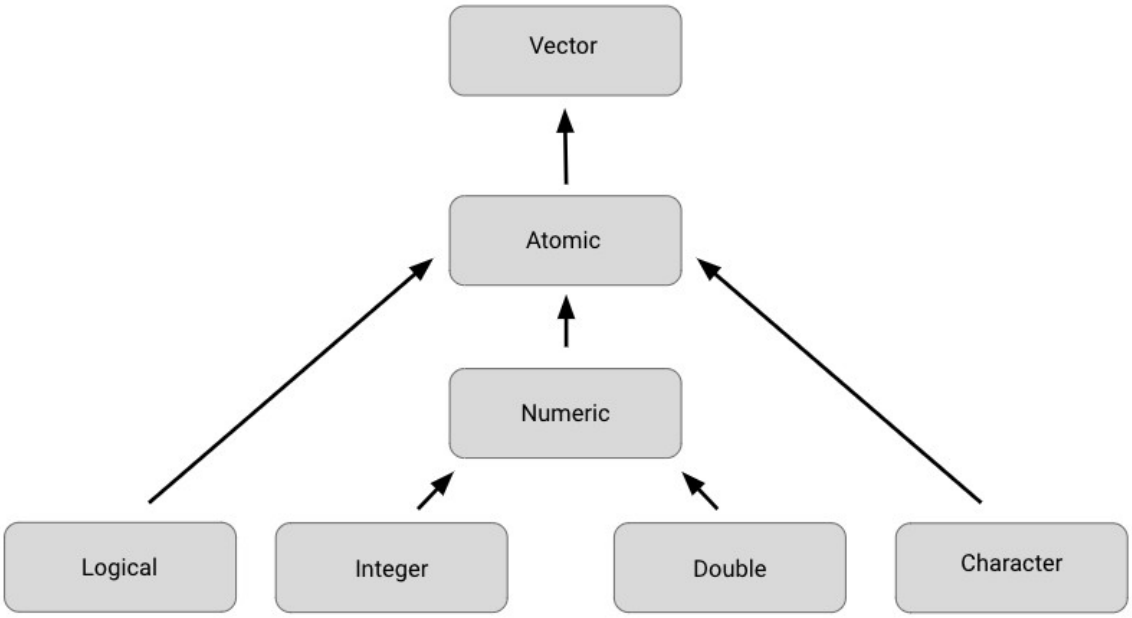
Atomic vectors

First, we will go through the different types of atomic vectors. Then, you will learn how to use R code to create, identify, and name the vectors.

Earlier, you learned that a **vector** is a group of data elements of the *same* type, stored in a sequence in R. You cannot have a vector that contains both logicals and numerics.

There are six primary types of atomic vectors: logical, integer, double, character (which contains strings), complex, and raw. The last two–complex and raw–aren’t as common in data analysis, so we will focus on the first four. Together, integer and double vectors are known as numeric vectors because they both contain numbers. This table summarizes the four primary types:

Type	Description	Example
Logical	True/False	`TRUE`
Integer	Positive and negative whole values	`3`
Double	Decimal values	`101.175`
Character	String/character values	`“Coding”`



Hierarchy of the relationships

Date and Time data types in R

Types

In R, there are three types of data that refer to an instant in time:

- A date `("2016-08-16")`
- A time within a day `("20:11:59 UTC")`
- And a date-time. This is a date plus a time `("2018-03-31 18:15:48 UTC")`

The time is given in UTC, which stands for Universal Time Coordinated, more commonly called Universal Coordinated Time. This is the primary standard by which the world regulates clocks and time.

Other common Data Structures in R

- Data Frames
  - Data frames are the most common way of storing and analyzing data in R, so it's important to understand what they are and how to create them. A data frame is a collection of columns–similar to a spreadsheet or SQL table. Each column has a name at the top that represents a variable, and includes one observation per row. Data frames help summarize data and organize it into a format that is easy to read and use.
- Files
  - R documentation is a tool that helps you easily find and browse the documentation of almost all R packages on CRAN. It's a useful reference guide for functions in R code.
- Matrices
  - A matrix is a two-dimensional collection of data elements. This means it has both rows and columns. By contrast, a vector is a one-dimensional sequence of data elements. But like vectors, matrices can only contain a single data type.

For example, you can't have both logical and numeric in a matrix. To create a matrix in R, you can use the matrix() function. One can create a matrix by using the following syntax:

```
matrix(c(3:8), nrow = 2)
```

Operators

A symbol that names the types of operation or calculation to be performed in a formula. A few basic types of operators are:

- Assignment operators
  - Used to assign value to variables and vectors.

Operator	Description	Example Code (after the sample code below, typing x will generate the output in the next column)	Result/ Output
<-	Leftwards assignment	x <- 2	[1] 2
<<-	Leftwards assignment	x <<- 7	[1] 7
=	Leftwards assignment	x = 9	[1] 9
->	Rightwards assignment	11 -> x	[1] 11
->>	Rightwards assignment	21 ->> x	[1] 21

- Arithmetic operators
  - Operators such as addition (+), subtraction (-), multiplication (*), etc. which are used for mathematical expressions/calculations are called arithmetic operators.
  -

Operator	Description	Example Code	Result/ Output
+	Addition	x + y	[1] 7
-	Subtraction	x - y	[1] -3
*	Multiplication	x * y	[1] 10
/	Division	x / y	[1] 0.4
%%	Modulus (returns the remainder after division)	y %% x	[1] 1
%/%	Integer division (returns an integer value after division)	y%/% x	[1] 2
^	Exponent	y ^ x	[1]25

- Relational Operators
  - Relational operators, also known as comparators, allow you to compare values. Relational operators identify how one R object relates to another—like whether an object is less than, equal to, or greater than another object. The output for relational operators is either TRUE or FALSE (which is a logical data type, or boolean).

Operator	Description	Example Code	Result/Output
<	Less than	x < y	[1] TRUE
>	Greater than	x > y	[1] FALSE
<=	Less than or equal to	x < = 2	[1] TRUE
>=	Greater than or equal to	y >= 10	[1] FALSE
==	Equal to	y == 5	[1] TRUE
!=	Not equal to	x != 2	[1] FALSE

- Logical Operators
    - Logical operators allow you to combine logical values. Logical operators return a logical data type or boolean (TRUE or FALSE).

Operator	Description
&	Element-wise logical AND
&&	Logical AND
|	Element-wise logical OR
||	Logical OR
!	Logical NOT

Data-frames in R

A data frame is a collection of columns. It's a lot like a spreadsheet or a SQL table. There's column names and rows and cells with data. The columns contain one variable, and the rows have a set of values that match each column. Somethings to know about the data-frames:

- Columns should be named
- Data stored can be many different types like numeric, factor or character.
- Each column should contain the same number of data items.

In the tidyverse, Tibbles are like streamlined data-frames. They are a bit different from the standard data-frames. Such as:

- Tibbles never change data types of the inputs.
- Tibbles never change the name of your variables.
- Tibbles never create row names.
- Tibbles makes printing in R easier.

Certain functions in R

Function	Description
head(*dataset_name*)	Gives a quick preview of the complete dataset.
str()	Gives the structure of the dataframe. Gives high level information such as the column name and the type of data it has.
colnames()	Lists out all the column names of the dataframe.

glimpse()	Gives a description of the dataset.
mutate()	Makes changes to the dataframe.
skim_without_charts()	Gives description of the dataset
select()	Selects a subset (column/s) of the data given in the parenthesis.
rename()	Renames the given column name.
rename_with(*dataset, toupper/tolower*)	Renames the column names to be consistent.
clean_names()	Cleans the column names and keeps it consistent
arrange()	Arranges the dataset in ascending order by default, on addition of "-" in the parenthesis, it will give descending order.
group_by()	Groups the data according to the given condition
separate()	Separate a column into two
unite()	Merge two columns together
pivot_wider()	Make a wide pivot table
pivot_longer()	Make a longer pivot table
bias()	Find out the bias between two datasets

File naming conventions

Do

- Keep your filenames to a reasonable length
- Use underscores and hyphens for readability
- Start or end your filename with a letter or number
- Use a standard date format when applicable; example: YYYY-MM-DD
- Use filenames for related files that work well with default ordering; example: in chronological order, or logical order using numbers first

Don't

- Use unnecessary additional characters in filenames
- Use spaces or "illegal" characters; examples: &, %, #, <, or >
- Start or end your filename with a symbol
- Use incomplete or inconsistent date formats; example: M-D-YY
- Use filenames for related files that do not work well with default ordering; examples: a random system of numbers or date formats, or using letters first

Common Problems when visualizing in R

- Case sensitivity of code
- Balancing parenthesis and quotation marks
- Using the "+" sign to add layers

Aesthetic attributes of R Visualizations

- **Color**: this allows you to change the color of all of the points on your plot, or the color of each data group
- **Size**: this allows you to change the size of the points on your plot by data group
- **Shape**: this allows you to change the shape of the points on your plot by data group

```
ggplot(data, aes(x=distance, y= dep_delay, color=carrier,
     size=air_time, shape = carrier)) + geom_point()
```

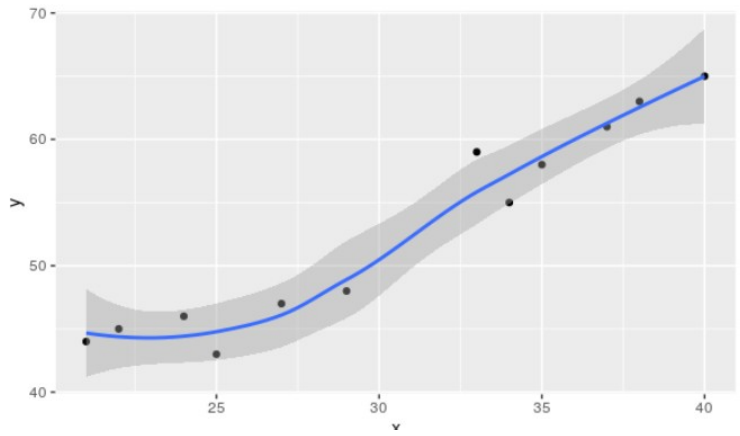
Smoothing in R Visualizations

Sometimes it can be hard to understand trends in your data from scatter plots alone. Smoothing enables the detection of a data trend even when you can't easily notice a trend from the plotted data points. Ggplot2's smoothing functionality is helpful because it adds a smoothing line as another layer to a plot; the smoothing line helps the data to make sense to a casual observer.

Example code

```
ggplot(data, aes(x=distance,
y= dep_delay)) +
    geom_point() +
    geom_smooth()
```

The example code creates a plot with a trend line similar to the blue line below.



Type of smoothing	Description	Example code
Loess smoothing	The loess smoothing process is best for smoothing plots with less than 1000 points.	`ggplot(data, aes(x=, y=))+ geom_point() + geom_smooth(method="loess")`
Gam smoothing	Gam smoothing, or generalized additive model smoothing, is useful for smoothing plots with a large number of points.	`ggplot(data, aes(x=, y=)) + geom_point() + geom_smooth(method="gam", formula = y ~s(x))`

R Markdown structure

1. Metadata
2. Text of the code
3. Visuals
4. Notes
5. Inline codes
6. Bullet points (add “*” before the statement)
7. Links (Add [link name] and the link url in parenthesis
8. Code chunks (Add ```{r}``` to get the code chunk in the R markdown)

The best Portfolios

- Are personal, unique and simple.
- Should reflect what you are interested in and what’s important to you.
- Keep a table of contents which leads to different pages keeping the landing page simple and easy to navigate.
- Keep the website fun and visually dynamic, but not so much as to distract viewers.
- Make the portfolio relevant and presentable.
- Keep the portfolio geared towards your desired field such as finance, technology, healthcare, etc.

Different jobs in the field of Data

	Data Analysts	Data Scientists	Data Specialists
Problem solving	Use existing tools and methods to solve problems with existing types of data	Invent new tools and models, ask open-ended questions, and collect new types of data	Use in-depth knowledge of databases as a tool to solve problems and manage data
Analysis	Analyze collected data to help stakeholders make better decisions	Analyze and interpret complex data to make business predictions	Organize large volumes of data for use in data analytics or business operations
Other relevant skills	• Database queries • Data visualization • Dashboards • Reports • Spreadsheets	• Advanced statistics • Machine learning • Deep learning • Data optimization • Programming	• Data manipulation • Information security • Data models • Scalability of data • Disaster recovery

Evaluate your portfolio

Now it's time to evaluate your portfolio. Select a portfolio piece to review and open it. Next, use the questions below as suggestions to help you review your work. As you answer each question, you will identify areas for improvement. When you're done, you can make these changes to improve your portfolio.

Is there anything missing? Are you missing steps in your projects, or details in your descriptions?

- If you have a website, are all the pages you need accounted for?
- If you are hosting your portfolio on an existing platform, are all your projects uploaded properly?

Is there too much info?

- Could any descriptions be revised for brevity?
- Are there places where you include more data than you need? Could something be cut without losing the meaning or context of your project?

Is there anything you think you shouldn't include?

- Have you included references to others' work that helped you without citing them? Can you remove them and instead include links to external work?
- Are there any other components that might seem extraneous or unprofessional?

Is your portfolio hosted on the most appropriate platform?

- There are many options for a data analytics platform, such as GitHub, Kaggle, and more. Is the one you're using (or plan on using) the most appropriate for your needs?