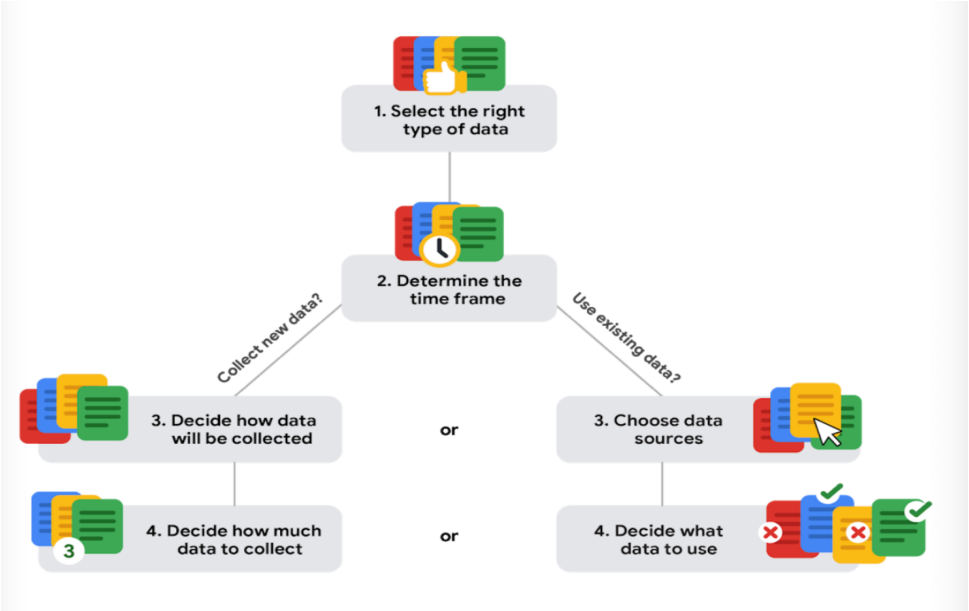
Course 3 - Prepare Data for Exploration

- How data is collected - interviews, observations, forms, questionnaires, surveys, cookies (how online ads make really accurate suggestions, how websites remember your preferences)
    - Cookies are small files stored on computers that contain info about users
    - Helps inform advertisers about your personal interests and habits based on surfing without personally identifying you
- Observation is the method of data-collection most often used by scientists
- **Factors to consider when collecting data**
    - **How data will be collected**
    - **Choose data sources**
        - First party data (preferred) - data collected by an individual or group using their own resources - know exactly where it came from
        - Second party data - data collected by a group directly from its audience and then sold
        - Third party data - data is sold by a provider that didn't collect the data themselves - may have come from a number of difference sources so it may not be reliable but can still be useful
        - Check for accuracy, bias, credibility
        - The data you choose should apply to your needs and must be approved for use
    - **Decide what data to use** - choose data that can help you find answers and solve problems and not get distracted by other data
    - **How much data to collect**
        - **Population** - all possible data values in a certain dataset
        - **Sample** - part of a population that is representative of the population
    - **Select the right data type** eg. time series data - data that includes dates
    - **Determine the time frame** - decide how long you will need to collect it, esp if tracking trends over a long period of time - have to use` historical if you need an answer immediately
    -



    - **Qualitative data** - usually listed as a name, category or description
        - **Nominal data -** qualitative data without a set order (yes/no answer, first time vs return vs regular customer)
        - **Ordinal data** - qualitative data with a set order or scale (rank movie from 1-5)
    - **Quantitative data** - can be measured, counted and then expressed as a number
        - **Discrete data** - is counted and has a limited number of values (dollar amounts, stars, points - when partial measurements aren't allowed - only full stars or points)
        - **Continuous data** - is measured and can have almost any numeric value - can have decimal with several places (run time for movie - 110.0356 minutes)
    - **Internal data -** lives within company's own systems
        - More reliable and easier to collect
    - **External data** - lives and is generated outside of an organisation
        - More valuable when your analysis depends on as many sources as possible
    - **Structured data** - organised in a certain format such as rows and columns (spreadsheets, relational databases) - makes it easier to search, analyse and store
    - **Unstructured data** - not organised in any easily identifiable manner, may have an internal structure (audio, video files, social media, emails - no clear way to identify or organise content - can't neatly fit in rows and columns)
- **Data modelling** - process of creating diagrams that visually represent how data is organised and structured
- **Data model -** a visual representation that is used for organising data elements and how they related to one another (Eg. blueprint of a house)
    - Diff users might have diff data needs but the data model gives them an understanding of the structure as a whole
- **Data elements** - pieces of info - people's name, account numbers, addresses
- **Three most common types of data modelling**
    - **Conceptual** - gives high level view of data structure - how data interacts across an organisation (define business requirements, doesn't contain technical details)
    - **Logical -** focuses on technical details - relationships, attributes, entities (defines how individual records are uniquely identified in a database, but does not' spell out actual names of database tables)
    - **Physical** - depicts how a database operates; defines all entities and attributes used (includes table names, column, data types)
- **Data-modelling techniques**
    - **Entity relationship diagram (ERD)** - visual way to understand the relationship between entities in the data model
    - **Unified modelling language (UML)** - detailed diagrams that describe the structure of a system by showing system's entities, attributes, operations and relationships

- **Data type** - a specific kind of data attribute that tells what kind of value the data is - tells you what kind of data you're working with
- **Data types in spreadsheets**
    - **number**
    - **Text or string** - a sequence of characters and punctuation that contains textual information (eg. people's names, phone numbers - treated like text)
    - **Boolean** - data type with only 2 possible values - True or False/ Yes or no
- **Rows = records**
- **Columns = fields ;** field can also refer to a single piece of data
    - Eg. music playlist - each song is a record, each song characteris - title, artist, is a field
- **Wide data** - data in which every data subject has a single row with split columns to hold the values of various attributes of the subject - easily compare different columns
- **Long data** - data in which each row is one time point per subject, so each subject will have data in multiple rows (each row contains a single data point)
    - great for storing and organising data when there's multiple variables for each subject at each time point
    - When there are multiple dates for one subject
    -

	Country Name	Country	Series Name	Year	Population
2	Antigua and Barb	ATG	Population, total	2010	88028
3	Antigua and Barb	ATG	Population, total	2011	89253
4	Antigua and Barb	ATG	Population, total	2012	90409
5	Antigua and Barb	ATG	Population, total	2013	91516
6	Antigua and Barb	ATG	Population, total	2014	92562
7	Antigua and Barb	ATG	Population, total	2015	93566
8	Antigua and Barb	ATG	Population, total	2016	94527
9	Antigua and Barb	ATG	Population, total	2017	95426
10	Antigua and Barb	ATG	Population, total	2018	96286
11	Antigua and Barb	ATG	Population, total	2019	97118
12	Argentina	ARG	Population, total	2010	40788453

-

Symbol	Date	Open
AAPL	2018-09-18	217.79
AAPL	2018-09-17	222.15
AAPL	2018-09-14	225.75
AAPL	2018-09-13	223.52
AMZN	2018-09-18	1918.65
AMZN	2018-09-17	1954.73
AMZN	2018-09-14	1992.93
AMZN	2018-09-13	2000
GOOGL	2018-09-18	1162.66
GOOGL	2018-09-17	1177.77

-

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** - process of changing the data's format, structure, or values
    - Involves adding, copying or replicating data, deleting fields or records, standardising the names of variables, renaming, moving or combining columns in a database, joining one set of data with another, saving a file in adiff format
    - Goals for data transformation - **organisation, compatibility** (with diff systems), **migration, merging, enhancement, comparison**
    - Eg. a company buying another company - need to make the data compatible - remove duplicate rows for customers in common, transform the format of one

- **Bias** - a preference in favor of or against a person, group of people or thing - can be conscious or subconscious
- **Data bias -** a type of error that systematically skews results in a certain direction
    - The way you collect data can also bias a data set - rushing people
- **Sampling bias** - when a sample isn't representative of the population as a whole
    - Avoid by choosing sample at random - so all parts of population of equal chance of being included
- **Unbiased sampling** - when a sample is representative of the population being measured
    - Can make sure by using visualisation to bring the results to life
- **Observer bias (experimenter/researcher)** - tendency for diff people to observe things differently
- **Interpretation bias** - the tendency to always interpret ambiguous situations in a positive or negative way
- **Confirmation bias -** tendency to search for or interpret info in a way that confirms pre-existing beliefs - wanting to confirm a gut feeling - socialise with ppl because they hold similar views
- **Identifying good data sources**
    - **Reliable** - trust that you're getting accurate, complete and unbiased info
    - **Original** - make sure to validate with original source esp if discovering data through 2nd/3rd sources
    - **Comprehensive** - contains all critical info needed to answer the questions
    - **Current** - and relevant
    - **Cited** - a source - more credible - think about who created the data set, is it part of a credible organisation, when was the data last refreshed - vetted public data sets, academic paper
- **Identifying bad data** - inaccurate, biased, incomplete, can't find original source, out of date, irrelevant, not cited
- **Ethics -** well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of right, obligations, benefits to society, fairness or specific virtues
- **GDPR (General Protection Regulation of the European Union)** - data privacy/protection legislation to help protect people and their data

- **Data ethics** - well-founded standards of right and wrong that dictate how data is collected,shared and used
    - Gets to the root of accountability companies have in protecting and responsibly using the data they collect
    - **Aspects of data ethics**
        - **Ownership** - individuals that own the raw data they provide and they have primary control over its usage, how it's processed, how it's shared (not the organisation that collected it)
        - **Transaction transparency** - all data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data
        - **Consent** - an individual's right to know explicit details about how and why their data will be used before agreeing to provide it ;how long stored
        - **Currency** - individuals should be aware of financial transaction resulting from the use of their personal data and the scale of these transactions
        - **Privacy -** data are people - preserving a data subject's info and activity any time a data transaction occurs
            - Person's legal right to their data - should have protection from unauthorised access to our private data
            - Freedom from inappropriate use of our data
            - The right to inspect, update, or correct our data
            - Ability to give consent to use our data
            - Legal right to access the data
        - **Openness -** free access, usage and sharing of data (gov't activity)
- **Data anonymization -** types of data that should be anonymized
    - Process of protecting people's private or sensitive data by eliminating that kind of info - blanking, masking personal info
    - Health care and financial data - most sensitive type
    - Usually go through **de-identification** - process to wipe data clean of all personally identify info
        - Eg. number, name, licence plate, SIN, address, medical records, account number, photographs
- **Personally identifiable information (PII)** - data that can be used by itself or with other data to track down a person's identity and make info known about them (address, credit card, etc.)

- **Open data** - must be available and accessible to the public as a complete dataset, must be provided under terms that allow reuse and redistribution, must have universal participation - everyone can use, reuse and redistribute ; no discrimination
    - Benefits of open data - good data is more widely available and combining data from diff fields of knowledge
- **Data interoperability** - ability of data systems and services to openly connect and share data

    - Important for health care systems - where multiple org - hospitals, clinics, pharmacies need to acces and share data - all have compatible databases

Sites, Resources for open data:

- **U.S. government data site**: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.
- **U.S. Census Bureau**: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
- **Open Data Network**: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
- **Google Cloud Public Datasets**: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
- **Dataset Search:** The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

- **Database -** a collection of data stored in a computer system; enables analysts to manipulate, store, and process data
- **Relational database** - a database that contains a series of related tables that can be connected via their relationship
    - For two tables to have a relationship, one or more of the same fields must exist inside both
    - **Normalisation** - process of organising data in a relational database (creating tables and establishing relationship btw those tables - eliminate data redundancy, increase data integrity and reduce complexity)
- **Primary key** - an identifier that reference a column in which each value is unique
    - Ensures data in a specific column is unique
    - Uniquely identifies a record in a relational database table
    - Only one primary key allowed in a table
    - Can't contain null or blank values
    - **Composite key** - a primary key constructed using multiple columns of a table
    - Some tables don't require a primary key - revenue
- **Foreign key -** a field within a table that is a primary key in another table
    - A column or group of columns ina relational database table that provides a link between the data in two tables
    - Refers to the field in a table that's the primary key of another table
    - A table can have multiple foreign keys
- **Structure Query Language -** a type of query language that lets data analyst communicate with a database
    - In a relational database, can write queries to get data from the related tables
- **Metadata -** data about data - reference guide about where the data comes from, when and how it was created, what it's all about

    - Describe the data that's contained in something, like a photo or email
    - Used in a database management to help data analysts interpret the contents of the data within the database (it is not the data itself)
    - Used to standardise data and evaluate 3rd party data's quality and credibility
- **3 common types of metadata**
    - **Descriptive** - metadata that describes a piece of data and can be used to identify it at a later point in time
        - Eg. descriptive metadata of a book in a library - the code on its spine (ISBN - International Standard Book Number, author, title)
    - **Structural** - indicates how a piece of data is organised and whether it is part of one, or more than one, data collection
        - Eg. how the pages of a book are put together to create diff chapters
        - Also keeps track of the relationship between two things - show the digital document of a book manuscript was actually the original version of a now printed book
    - **Administrative** - indicates the technical source of a digital asset
        - Eg. photo - shows you the type of file, data and time taken
- **Elements of metadata**:
    - **Title and description -** name of the file/website, what type of content
    - **Tags and categories** - general overview of the data, is it indexed or described
    - **Who created it and when -** where it came from, recent or existed a long time
    - **Who last modified and when** - were there any changes, were they recent
    - **Who can access or update it** - public or need specific permissions to customise
- **Examples of metadata** - photos (camera filename, date, time, geolocation), emails(subject line, sender, recipient), spreadsheets and documents (title, creation data, author, user comments, tabs), websites, digital files, books

- Metadata creates a single source of truth by keeping things consistent and uniform
    - Make it easier to discover relationships between the data inside it and elsewhere
    - Makes data more reliable - making sure it's accurate, precise, relevant, timely
- **Metadata repository** - a database specifically created to store metadata
    - Can be physical or virtual like the cloud
    - Describes the state and location,structure of the tables inside, and how data flows through
    - Kept it in an accessible form so it can be used quickly and easily
    - Keep track of who accessed it
    - If it's external data - metadata is useful to ensure that the data is clean, accurate, relevant and timely; make sure we can access or purchase it
- **Metadata management** - Stored in a single, central location and gives the company standardised info about all of its data vs. using second or third party data where each system has its own rules so data is organised in a different way
    - Includes info about where each system is located and where the data sets are locaetd within those systems

    - Describes how all the data is connected between the various systems
    - **Data governance -** a process to ensure the formal management of a company's data assets
        - Gives organisation better control of their data, helps manage issues related to security, privacy, integrity, usability, internal and external data flows
        - Metadata specialists organise and maintain company data, ensuring that it's of the highest possible quality ; they create basic metadata identification and discovery info, describe the way different data sets work together, and explain the many diff types of data resources

- Importing data from other spreadsheets - IMPORTRANGE function
    - Can specify a range
- Importing data from **CSV(Comma-separated values - saves data in a table format)** files - file, import, upload
    - CSV uses plain text and are delineated by characters, such as a comma (delineator indicates a boundary or separation between two things)
    - Makes it easier for data analysts to examine a small part of a large dataset, import data to a new spreadsheet, and distinguish values from one another
- Importing HTML tables from web page - IMPORTHTML function
- **Sorting** - arranging data into a meaningful order to easier understand, analyse and visualise
- Spreadsheet more suited for self-contained data, where data exists in one place such as internal revenue data
- Databases are more suitable for external tables, allowing you to change data in several places by editing in only one place, such as dynamic consumer info

- **Bigquery** - a data warehouse on google cloud that data analysts can use to query, filter large datasets, aggregate results, and perform complex operations
- **Best practices when organising data:**
    - **Name conventions** - consistent guidelines that describe the content, date, or version of a file in its name - use logical and descriptive names for your files, make them easier to find and use
        - Align with team's, agree on a file name, include date (YYYYMMDD) and version, keep it short and sweet
        - Lead revision numbers with 0
        - create a sample text file with content that breaks down naming convention
        - avoid spaces - use underscores or dashes (spaces and special characters causes errors in some applications) - better for SQL
    - **Foldering, subfolders** - create them in a logical hierarchy so related files are together; separate ongoing from completed work
    - **Archiving older files** - in a separate folder or external storage location, back up files

    - **Align your aming and storage practices with your team**
    - **Develop metadata practices**
        - Think about how often you're making copies - data can contradict itself and make mistakes later on
        - Relational databases help avoid duplication

- **SELECT** is the section of a query that indicates what data you want SQL to return to you
- **FROM** is the section of a query that indicates which table the desired data comes from.
    - Backticks are optional ``
- **WHERE** is the section of a query that indicates any **filters** you'd like to apply to your dataset
- The SELECT DISTINCT statement is **used to return only distinct (different) values**.
- SQL help data analysts interact with databases and view the data they need

SQL Notes

- Use all caps for **clause starters** (SELECT) and **functions** (SUM)
- **Column names** - all lower case (snake_case), use underscore between words
    - SQL defaults to f0,f1, etc if you don't name them
- **Table names** - CamelCase - capitalise start of each word
    - Some people don't capitalise first word camelCase
    - Key is to be consistent and make it easy and professional to read
- BigQuery is case sensitive - US vs us (some SQl dialects are not and will return all entries that have any variation of "us"
- Use **single quotes** for the most part ('US')
- Use **double quotes** when your string has quotes inside it ("Shepherd's pie")
- **Comments** as reminders, use **two dashes** – , the database will ignore everything in the same line after –
    - Can also use # but MYSQL doesn't recognize it
- **Indentation** - general rule, keep each line <= 100 characters

```
SELECT
      CASE
            WHEN genre = 'horror' THEN 'Will not watch'
            WHEN genre = 'documentary' THEN 'Will watch alone'
            ELSE 'Watch with others'
            END AS watch_category, COUNT(movie_title) AS number_of_movies
FROM
      MovieTheater
GROUP BY
      1
```

- **Multi-line comments -** use /* to start, */ to close the comment - cleaner looking

4. In the following FROM clause, what is the table name in the SQL query?

   FROM bigquery-public-data.sunroof_solar.solar_potential_by_postal_code

   - ○ solar_potential_by_postal_code
   - ◉ public-data.sunroof
   - ○ solar.solar
   - ○ sunroof_solar

   ⊗ **Incorrect**
   The table name in the SQL query is solar_potential_by_postal_code. This table is in the sunroof_solar dataset, a public dataset in BigQuery.

- **Data security** - protecting data from unauthorised access or corruption by adopting safety measures
  - Locking down formulas so they aren't accidentally broken
  - Password protection, user permissions
  - Google sheets - sharing menu
- Companies need to balance their data security measures with their data access needs
- **Encryption** - uses algorithm to alter data and make it unusable by users and applications that don't know the algorithm
  - Algorithm is saved as a key which can be used to reverse the encryption
- **Tokenization** - replaces the data elements you want to protect with randomly generated data referred to as a token
  - Original data is stored in a separate location and mapped to the tokens
  - To access, the user needs to have permission to use the tokenized data and the token mapping
  - If tokenized data is hacked, the original data is still and safe and secure in a separate location

Course 4 - Process Data from Dirty to Clean

- clean data is essential to data integrity and reliable solutions and decisions
- **Data integrity** - the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle
  - Can be compromised every time it's replicated, transferred or manipulated in any way
  - **Data replication** - the process of storing data in multiple locations - data can become out of sync

- **Data transfer** - process of copying data from a storage device to memory, or from one computer to another
- **Data manipulation** - process of changing data to make it more organised and easier to read
- Other threats - human error, viruses, malware, hacking, system failures

-

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.

- **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**
    - If data only partially aligns with objective, think about how you can modify constraints to make the data align better with the business objective

- **Types of insufficient data**
    - Data only from one source
    - Data that keeps updating (it's still incoming)
    - Outdated data
    - Geographically limited data
- **Ways to address insufficient data**
    - Identify trends with the available data
    - Wati for ore dat if time allows
    - Talk with stakeholders and adjust your objective
    - Look for a new dataset

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.

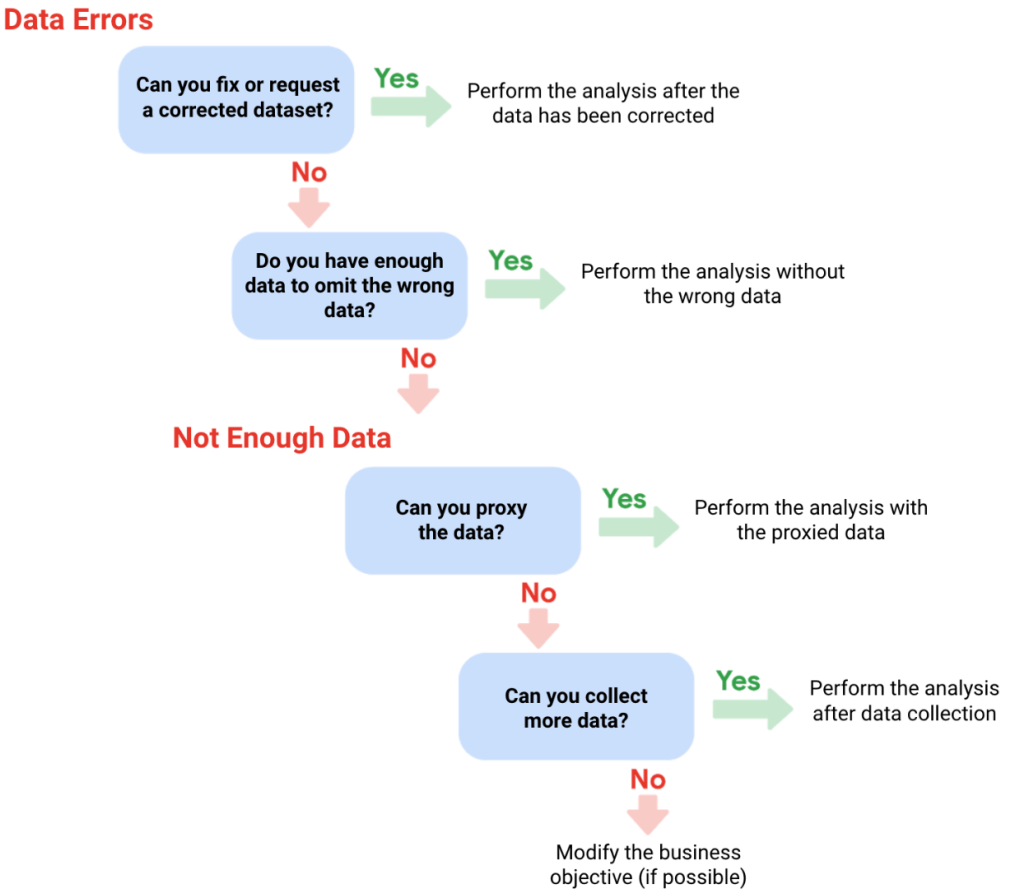
Data issue 2: too little data

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.*



-
-

- **Population** - all possible data values in a certain dataset
- **Sample size** - a part of a population that is representative of the population
- **Population** - the total number you hope to pull your sample from
- **Sampling bias** - sample isn't representative of the population as a whole
- **Random sampling** - way of selecting a sample from a populations o that every possible type of the sample has an equal chance of being chosen
- **Margin of error** - the max amt expected to differ between the sample results and the results from the actual population
    - Helps you understand how reliable the data from your hypothesis is
    - The closer to zero the marin of error, the closer your results is to the population
    - It is counted in both directions eg. 10% MOE - results is 60% - it means your result is 50%-70% - survey would be inconclusive (MOE defines a range of values below and above the avg result for the sample)
        - Avg result for the entire population is expected to be within that range
- **Confidence level -** how confident you are in the survey results - a 95% confidence level means if you were to run the same survey 100 times, you would get similar results 95/100
    - This 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
- **Estimated response rate** - percentage of people you expect will complete your survey out of those who received it
    - If you need a sample of 100 individuals and your estimated response rate is 10%, you will need to send your survey out to 1000 individuals
- **Statistical significance** - the determination of whether you result could be due to random chance or not
    - The greater the significance, the less due to chance
    - If a test is statistically significant it means that the results of the test are real and not an error caused by random chance
        - If it's 60% - it means there's a 60% chance that the results are reliable, 40% chance that the result of the test is wrong
        - Usually need 80% to consider your results statistically significant

- Confidence level and margin of error are independent of each other/ don't need to add up to 100
- The smaller the margin of error, the larger the sample size needs to be

- Don't use a sample size less than 30 - been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the avg result of a population
    - Based on **Central Limit Theorem** (CLT) - as sample size increases, the results more closely resemble the normal bell-shaped distribution from a large number of samples
- Confidence level most commonly used is 95%, 90% can work in some cases
- **A larger sample size results in a higher confidence level, decrease the margin of error, greater statistical significance**
- **Sample sizes vary by business problem**

    - Trying to find out how residents feel about a new library vs how residents would vote to fund the library
    - **Larger sample sizes will cost more** - benefits need to outweigh the cost - understanding consumer preferences vs the effects of a new drug; small size would be good enough for consumer preferences
    - https://docs.google.com/spreadsheets/d/1dPNj-gLL3tf5QPLePYOW6l0VCgHjJZC8d8dE4Rw1VW4/edit#gid=0

- **Open data** - info published on gov't sites - structured, open-liscnesed, maintained
- **Public data -** data that everywhere else - freely available but not accessible on the web - usually unconstructed, unruly
- Pre-cleaning increases the efficiency and success of your data analysis tasks
- Knowing your data is accurate, consistent and complete will give you confidence that your results are valid
- Stakeholders will be pleased if you connect the data to business objectives
- Knowing when to stop collecting data allows you to finish your tasks in a timely manner

- **Statistical power** - the probability of getting meaningful results from a test
- **Hypothesis testing** - a way to see if a survey or experiment has meaningful results

- Calculating margin of error - need population size, sample size and confidence level
    - https://docs.google.com/spreadsheets/d/19yFoyzOvQegK2jrIUTNJY8KQMO7WXEsv5PxflJgfnm4/edit?resourcekey=0-N9j7vCwAF8HuJNxuKdIxLw

- In marketing - a/b testing - split testing - tests two variations of the same email,web page, etc. to see which is more successful in attracting traffic
    - **Conversion rate** - user traffic that gets monetized
    - Eg. A gets 5%, B gets 3%, if MOE is 2%, the confidence interval is 3%-7% ; lowr end is overlapping with B's results - can't conclude there is a statistical significant difference btw A and B

- **Dirty data** - data that is incomplete, incorrect, or irrelevant to the problem you're trying to solve
    - Includes outdated, inconsistent data
    - Causes skewed metrics, inflated or inaccurate counts, inaccurate insights, decision making, decreased productivity, contradiction leading to confusion
    - Usually caused by inconsistent formatting, typos, null fields, duplication, differences in currency
    -

- **Data engineers -** transform data into useful format for analysis and give it a reliable infrastructure
    - Develop, maintain and test databases, data processors, etc.
- **Data warehousing specialists -** develop processes and procedures to effectively store and organise data
    - Make sure data is available, secure and backed up to prevent loss
- Internal data from these colleagues are much more reliable and likely to be clean data

- **Null** - an indication that a value does not exist (customer skipped question); doesn't mean it's zero
- **Field length -** a tool for determining how many characters can be keyed into a field
    - Assigning a certain length helps avoid errors
- **Field -** is a single piece of information from a row or column of a spreadsheet.
- **Data validation** - a tool for checking the accuracy and quality of data before adding or importing it/ form of data cleaning

- **Validity -** using data integrity principles to ensure measures conform to defined business rules or constraints
    - eg.Data collected five years ago used technology that is not approved or supported by the business
- **Accuracy** - The degree of conformity of a measure to a standard or a true value

- Merging data - questions to ask for **data compatibility** (how well two or more datasets are able to work together)
    - Do i have al the data I need
    - Does the data I need exist within these datasets
    - Does the data need to be cleaned or are they ready for me to be used
    - Are the datasets cleaned to the same standard
- Common mistakes to avoid:
    - **Not checking for spelling errors**
    - **Forgetting to document errors** - helps keep track of changes so you can backtrack if a fix didn't work
    - **Not Checking for misfielded values** - correct values but in the wrong field
    - **Overlooking missing values** - maintain completeness and consistency
    - **Only looking at a subset of the data** - each filed requires equal attentions
    - **Losing track of business objectives** - keep discoveries relevant to the question
    - **Not fixing the source of the error**
    - **Not analysing the system prior to data cleaning** - to find the root cause of dirty data
    - **Not aching up your data prior to data cleaning**
    - **Not accounting for data cleaning in your deadlines/process** - essential to get a more accurate timeline estimate

- *Deselecting columns - commandA x 2 - then deselect cells with command*
- **Conditional formatting** - spreadsheet tool that changes how cells appear when values meet specific conditions
- **Remove duplicates** - data->data clean up->remove duplicates - automatically searches for and eliminate duplicate entries
- **format->number->date** - sorts all data into consistent format
- **Text string** - a group of characters within a cell, most often composed of letters
    - **Sub string** - smaller subset of a text string
- **Split** - tool that divides text around a specified character and puts each fragment into a new, separate cell
    - Eg. If a cell has full name and want to split first and last name to two separate columns
    - **Highlight column->data->split text to columns**
        - Can also convert numbers stored as text into just numbers

    - **Specified text separator = delimiter** - eg. comma (indicates beg or end of data item)
- **Concatenate** - function that joins multiple text strings into a single string

- **Syntax -** predetermined structure that includes all required info and its proper placement
- **Function** - set of instructions that performs a specific calculation using the data in a spreadsheet
    - **COUNTIF -** returns the number of cells that match a specified value
        - Syntax - **=COUNTIF(range, "value")**
    - **LEN** - function that tells you the length of a text string by counting the number of characters it contains
        - Syntax **=LEN(RANGE)**
    - **LEFT** - A function that gives you a set number of characters from the left side of a text string
        - Syntax **=LEFT(range, number of characters)**
    - **RIGHT** - function that gives you a set number of characters from the right side of a text string
        - Syntax **=RIGHT(range, number of characters)**
    - **MID** - function that gives you a segment from the middle of a text string
        - Syntax **=MID(range, reference starting point, number of middle characters)**
    - **CONCATENATE** - A function that joins together two or more text strings
        - Syntax **= CONCATENATE (item 1, item 2)**
    - **TRIM** - function that removes leading, trailing, and repeated spaces in data
        - Syntax = **TRIM(range)**

- **Pivot table** - a data summarization tool that is used in data processing
    - Select data ->insert->pivot table->
- **Vertical lookup = VLOOKUP** - function that searches for a certain value in a column to return a corresponding piece of info
    - Syntax **=VLOOKUP**(**data to look up, 'where to look'!Range, column, false)**
        - ! - indicates we're referring to a diff sheet than the current one
        - VLOOKUP searches for the value in the first argument in the **leftmost** column of the specified location
        - The value of the third argument tells VLOOKUP to return the value in the same row from the specified column
        - False - asks for an exact match
- **Plotting** - putting data in a graph chart, table or other visual to help you quickly find what it looks like
    - Useful to identify skewed data or outliers

- **Data mapping** - process of matching fields from one data source to another
    - Evaluates how well two or more data sources work together
    - First step - identify what data needs to be moved
    - Identify desired format
    - Test - inspect a sample piece of data to confirm that it's clean properly formatted
        - Data validation, conditional formatting, COUNTIF, sorting, filtering
- **Schema -** way of describing how something is organised
- **Primary key -** references a column in which each value is unique
- **Foreign key** - a field within a table that is a primary key in another table

- Cleaning up data with SQL - A language used to interact with database programs
    - Can pull out info from diff sources in the database, good for larger datasets, working in a team, useful across multiple programs
    - Can also join data, use formulas and perform arithmetic like spreadsheets
- Vs spreadsheet - generated with a program, can only access data you input, stored locally, smaller dataset, good for working solo
- Standard SQL works with a majority of databases and requires a small number of syntax changes to adapt to other dialects.

A **byte** is a collection of 8 bits. Take a moment to examine the table below to get a feel for the difference between data measurements and their relative sizes to one another.

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)

-
- querying a dataset with billions of items isn't feasible without tools such as relational databases and SQL.
- Performing large queries by hand would take years and years of manual work. The ability to query large datasets is an extremely helpful tool for data analysts.

Widely used SQL Queries

- **INSERT INTO** customer_data.customer_address
  (customer_id, address)
    - add data
  VALUES
  (2645, '333 SQL Road')

- **UPDATE** customer_data.customer_address
  SET address = '123 address'
  (what you want to change it to)
  From Where

- **SELECT** LastName, Country
  FROM customers
  WHERE Country = 'Germany'
    - (where sort of acts as a filter like in a spreadsheet)

- **SELECT**
  DISTINCT customer_id
  FROM customer_data.customer_address
  - When you don't want duplicates to show up in your query

- **SELECT**
  LENGTH/ LEN (country) **as** letters_in_country
  FROM customer_data.customer_address
  - Quickly shows how many letters in a each country column to make sure they are all consistently a certain number

- **SELECT** country
  FROM customer_data.customer_address
  WHERE
  LENGTH (country)>2

- **SELECT**
  DISTINCT customer_id
  FROM customer_data.customer_address
  WHERE
  SUBSTR(country,1,2) = 'US'
  (column, which letter to start with, how many letters to pul)
  - Pull all the countries that start with US including USA

- **SELECT**
  DISTINCT customer_id
  FROM customer_data.customer_address
  WHERE
  TRIM (state) = 'OH'
  (column we want to remove spaces from)

- **SELECT**
  MIN(compression_ratio) AS min_compression_ratio,
  MAX(compression_ratio) AS max_compression_ratio
  FROM
  Cars.car_info;
  WHERE
  compression_ratio <> 70
  <> means not equal

- **DELETE** cars.car_info
  WHERE compression_ratio = 70;

- **SELECT**
  COUNT(*) AS num_of_rows_to_delete
  - returns the number of rows including null and duplicates
  FROM
  cars.car_info
  WHERE
  compression_ratio = 70

- **SELECT**
  DISTINCT drive_wheels,
  LENGTH(drive_wheels) AS string_length
  FROM

- **UPDATE**
  cars.car_info
  SET
  drive_wheels = **TRIM**(drive_wheels)
  WHERE TRUE;

- **CAST** - can be used to convert anything from on data type to another
- **ORDER BY DESC**
- **Float -** number that contains a decimal
- **Typecasting** - converting data from one type to another
    - **SELECT CAST**(purchase_price **AS FLOAT64**)
        - Float64 just means we're casting numbers in the 64 bit system as floats (may be different in other sql platforms)
        - Converting string to float
      FROM customer_data.customer_purchase
        - From the customer_purchase table in the customer_data dataset
      ORDER BY
      CAST(purchase_price **AS FLOAT64**) **DESC**

Advanced data-cleaning functions, part 2 video (week 3)

- **SELECT CAST(date as date) as date_only, purchase_price**
    - To return the date with just the date and no time
  FROM customer_data.customer_purchase
  WHERE
  date between '2020-12-01' and '2020-12-31'

- **CONCAT() -** adds strings together to create new text strings that can be used as **unique keys**
    - **SELECT CONCAT**(product_code, product_color)
      FROM `customer_data.customer_purchase`
      WHERE
      product = 'couch'
        - To see which color is most popular

- **COALESCE()** - used to return non-null values in a list
    - **SELECT COALESCE**(product, product_code)
      FROM `customer_data.customer_purchase`
        - Tells SQL which column to check first, then check the next column if the first one was null (product column was optional so some were null)

- SELECT, FROM, WHERE - pull data from a specific place in a table, typically a table column
- SELECT, FROM - pull data from any table in database

-

- **Verification** - process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable
    - 1st step - going back to original unclean data set and compare it to what you have now (ensure there are no nulls, misspelling if that's what you found originally - use conditional formatting, filters, FIND)
    - Big picture view - confirm you're focusing on the business problem you're trying to solve
        - 1 consider the business problem you're trying to solve
        - 2 consider the goal of the project - eg. trying to make improvements on the product - also need to know whether the data you've collected and cleaned will actually help your company achieve that goal
        - 3 consider where the data came from, testing your data collection to see whether your data is capable of solving the problem

- **COUNTA -** function that counts the total number of values within a specified range (eg. the number of times a supplier's name appears in Column C)
    - VS Count - only counts the numerical values within a specified range
- **CASE statement** - goes through one or more conditions and returns a value as soon as a condition is met
    - **SELECT** customer_id
        CASE
            WHEN first_name = 'Tnoy' **THEN '**Tony'
            WHEN first_name = 'Tmo' **THEN** 'Tom'
            WHEN first_name = 'Rachl' **THEN** 'Rachel'
            ELSE first_name
            END AS cleaned_name
        FROM customer_data.customer_name

- **Data-cleaning verification checklist:**
    - **Sources of errors**: Did you use the right tools and functions to find the source of the er 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 co
    - **Extra spaces and characters**: Did you remove any extra spaces or characters using the function?
    - **Duplicates**: Did you remove duplicates in spreadsheets using the **Remove Duplicates** **DISTINCT** in SQL?
    - **Mismatched data types**: Did you check that numeric, date, and string data are typecas correctly?
    - **Messy (inconsistent) strings**: Did you make sure that all of your strings are consistent meaningful?
    - **Messy (inconsistent) date formats**: Did you format the dates consistently throughout 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
- **Review the goal of your project**
    - Confirm the business problem
    - Confirm the goal of the project
    - Verify that data can solve the problem and is aligned to the goal

Capturing cleaning changes

- **Documentation -** process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort; benefits: (first two assume the errors aren't fixable)
    - Can recover data-cleaning errors
    - Inform other users of change
    - Determine quality of data
- Documenting makes it possible to be transparent about your process, keep members on the same page and demonstrate to stakeholders that you are accountable
- **Changelog -** a file containing a chronologically ordered list of modifications made to a project
    -
    - spreadsheet - file - version history ; or right click a cell and choose 'show edit history'

- SQL - add comments or query history - which track all - the queries you've run - project history in the bottom tab
- Can record:
    - 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
        - Let other users know your reason for change
        - Undo a formula
        - Can also jump to revert a change in any order - made 4 changes, can undo change #2 without clicking undo 3x

Google Sheets	1. Right-click the cell and select **Show edit history**. 2. Click the left-arrow < or right arrow > to move backward and forward in the history as needed.
Microsoft Excel	1. If Track Changes has been enabled for the spreadsheet: click **Review**. 2. Under **Track Changes**, click the **Accept/Reject Changes** option to accept or reject any change made.
BigQuery	Bring up a previous version (without reverting to it) and figure out what changed by comparing it to the current version.

- Engineers use **engineering change orders (ECOs) -** keeps track of new product design details and proposed changes
- Writers use **document revision histories** - keep track of changes to document flow and edit

- **Version control system** IRL- analyst makes changes to an existing SQL query shared across the company:

Best practices for changelogs

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

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

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

- Added: new features introduced
- Changed: changes in existing functionality
- Deprecated: features about to be removed
- Removed: features that have been removed
- Fixed: bug fixes
- Security: lowering vulnerabilities

1. A company has official versions of important queries in their **version control system**.
2. An analyst makes sure the most up-to-date version of the query is the one they will change. This is called **syncing**
3. The analyst makes a change to the query.
4. The analyst might ask someone to review this change. This is called a **code review** and can be informally or formally done. An informal review could be as simple as asking a senior analyst to take a look at the change.
5. After a reviewer approves the change, the analyst submits the updated version of the query to a repository in the company's version control system. This is called a **code commit**. A best practice is to document exactly what the change was and why it was made in a comments area. Going back to our example of a query that pulls daily revenue, a comment might be: *Updated revenue to include revenue coming from the new product, Calypso*.
6. After the change is **submitted**, everyone else in the company will be able to access and use this new query when they **sync** to the most up-to-date queries stored in the version control system.
7. If the query has a problem or business needs change, the analyst can ***undo*** the change to the query using the version control system. The analyst can look at a chronological list of all changes made to the query and who made each change. Then, after finding their own change, the analyst can **revert** to the previous version.
8. The query is back to what it was before the analyst made the change. And everyone at the company sees this reverted, original query, too.

Function	Syntax (Google Sheets)	Menu Options (Microsoft Excel)	Primary Use
IMPORTRANGE	=IMPORTRANGE(spreadsheet_url, range_string)	Paste Link (copy the data first)	Imports (pastes) data from one sheet to another and keeps it automatically updated.
QUERY	=QUERY(Sheet and Range, "Select *")	Data > From Other Sources > From Microsoft Query	Enables pseudo SQL (SQL-like) statements or a wizard to import the data.
FILTER	=FILTER(range, condition1, [condition2, ...])	Filter (conditions per column)	Displays only the data that meets the specified conditions.

Advanced functions for cleaning (google sheets):

- **= IMPORTRANGE** - allows you to insert data from one sheet to another
    - Can cherry pick the data you want to analyse, leave behind the others
- **= QUERY** - pulling data from another spreadsheet - like SQL, can extract specific data within a spreadsheet
    - eg.QUERY(A2:E6,"select avg(A) pivot B")
      QUERY(A2:E6,F2,FALSE)
- **=FILTER**