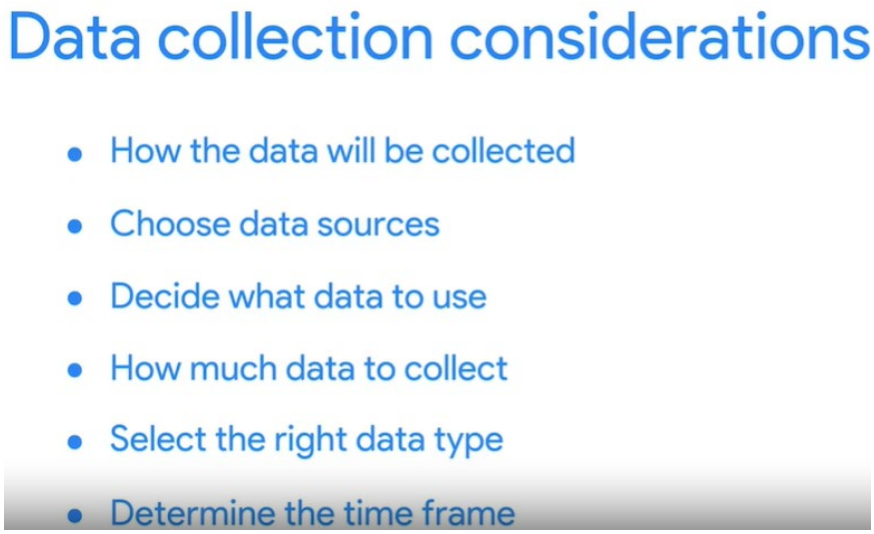
Week 1

- First-party data: data collected by an individual or group using their own resources
- Second-party data: data collected by a group directly from its audience and then sold.
- Third-party data: data collected from outside sources who did not collect it directly



- In data analytics, a population refers to all possible data values in a certain dataset.
- Discrete data: data that is counted and has a limited number of values.
- Continuous data: data that is measured and can have almost any numeric value.

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

- Nominal data: a type of qualitative data that is categorized without a set order
- Ordinal data: a type of qualitative data with a set order or scale

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)

-
- Internal data: data that lives within a company's own systems.
- External data: data that lives and is generated outside of an organization.

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

-
- Structured data: data organized in a certain format such as rows and columns. Structured data that is grouped together to form relations enables analysts to more easily store, search, and analyze the data.
- Unstructured data: data that is not organized in any easily identifiable manner (such as video, audio)

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

-

- Data model: a model that is used for organizing data elements and how they relate to another
- Data elements: pieces of information, such as people's names, account numbers, and addresses.
- Three most common types of data modeling:

    1. **Conceptual data modeling** gives a high-level view of the data structure, such as how data interacts across an organization. For example, a conceptual data model may be used to define the business requirements for a new database. A conceptual data model doesn't contain technical details.
    2. **Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
    3. **Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

- Long data: data in which each row is one time point per subject, so each subject will have data in multiple rows.
- Wide data: data in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject.

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

-
- Why transform data?

    ### Why transform data?

    Goals for data transformation might be:

    - 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

- Bias: a preference in favor of or against a person, group of people, or thing.
- Data bias: a type of error that systematically skews results in a certain direction
- Sampling bias: when a sample isn't representative of the population as a whole.

- Observer bias (experimenter bias/research bias): the tendency for different people to observe things differently
- Interpretation bias: the tendency to always interpret ambiguous situations in a positive or negative way
- Confirmation bias: the tendency to search for or interpret information in a way that confirms pre-existing beliefs
- Good data: reliable, original, comprehensive, current, cited
- Vetted public datasets, academic papers, and governmental agency data are usually good data sources.
- Data ethics: well-founded standards of right and wrong that dictate how data is collected, shared, and used.

-
    1. Ownership: individuals own the raw data they provide and they have primary control over its usage, how it's processed, and how it's shared.
    2. Consent: an individual's right to know explicit details about how and why their data will be used before agreeing to provide it.
    3. Currency: individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions.
    4. Privacy: preserving a data subject's information and activity any time a data transaction occurs
        - Protection from unauthorized 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
    5. Openness: free access, usage, and sharing of data.
    6. Transaction transparency: Transaction transparency states that all data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data.

- Data anonymization is the process of protecting people's private or sensitive data by eliminating that kind of 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
- Data interoperability: the ability of data systems and services to openly connect and share data

What is 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

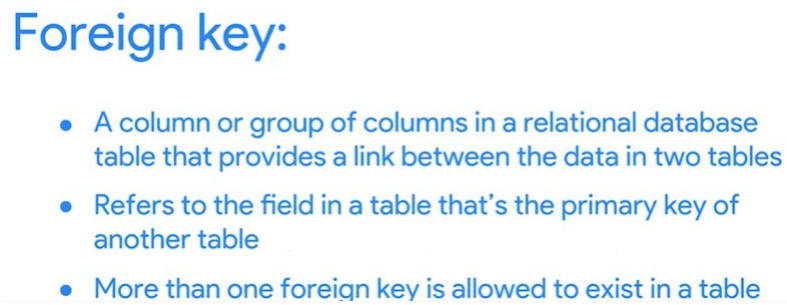
- Data can only be considered open when it meets all three of these standards.
- Reliable data sources:
    1. U.S. government data site: https://data.gov/
    2. U.S. Census Bureau: https://www.census.gov/data.html
    3. Open data network: https://www.opendatanetwork.com/
    4. Google cloud public datasets: https://cloud.google.com/datasets
    5. Dataset search: https://datasetsearch.research.google.com/

Week 3

- Relational database: a database that contains a series of related tables that can be connected via their relationships

Primary key:

- Used to ensure data in a specific column is unique
- Uniquely identifies a record in a relational database table
- Only one primary key is allowed in a table
- Cannot contain null or blank values



- Metadata:
  Three common types:
  1. Descriptive: describes a piece of data and can be used to identify it at a later point in time,
  2. Structural: indicates how a piece of data is organized and whether it is part of one, or more than one, data collection,
  3. Administrative: indicates the technical source of a digital asset
- Elements of metadata: title and description, tags and categories, who created it and when, who last modified it and when, who can access or update it
- Metadata repository: a database specifically created to store metadata.

- Metadata is stored in a single, central location, and gives the company standardized information about all of its data.
- Data governance: a process to ensure the formal management of a company's data assets.

Week 4

- Best practices when organizing data
  1. Naming conventions: consistent guidelines that describe the content, date, or version of a file in its name, use logical and descriptive names for your files to make them easier to find and use.
  2. Foldering
  3. Achieving older files
  4. Align your naming and storage practices with your team

5. Develop metadata practices

- Data security: protecting data from unauthorized access or corruption by adopting safety measures.

Week 5