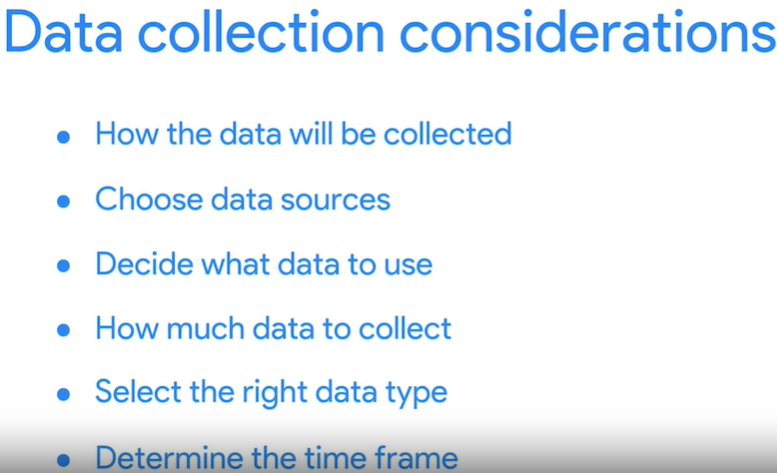
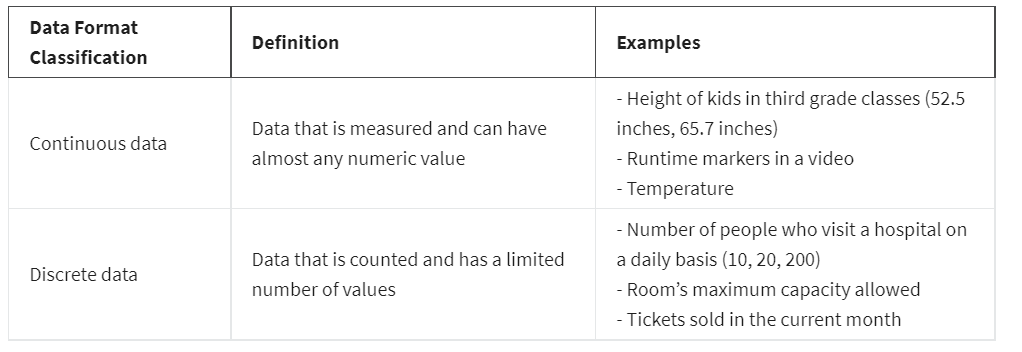
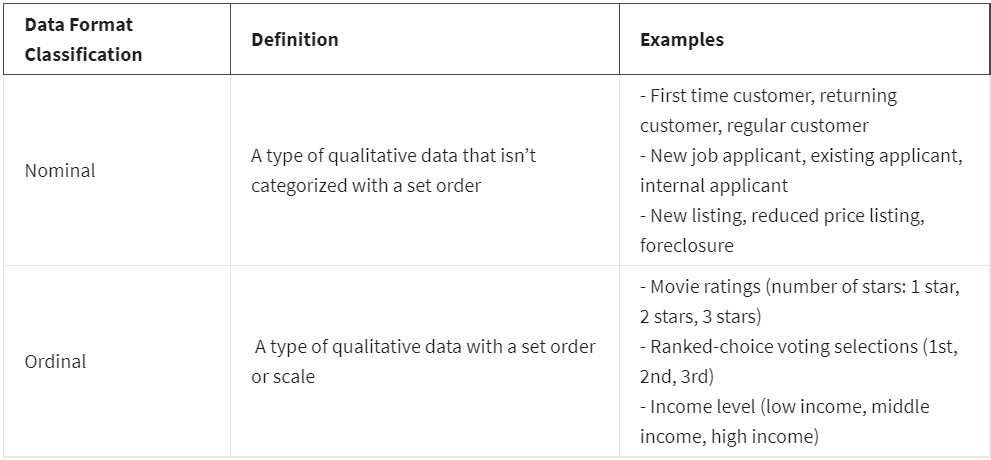
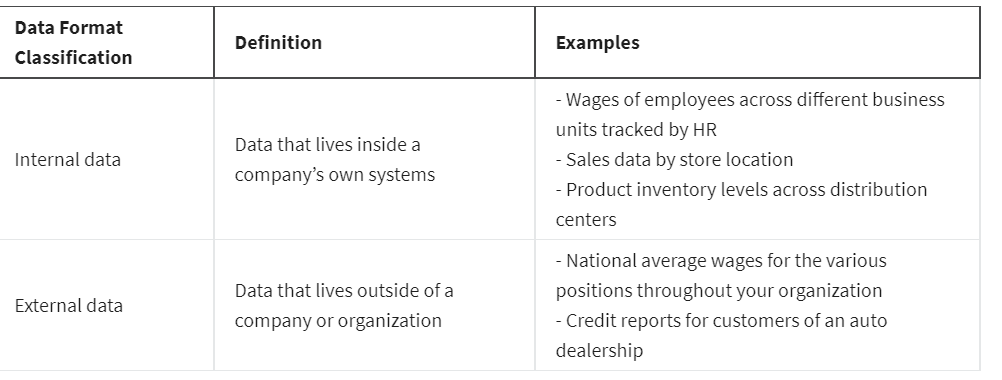
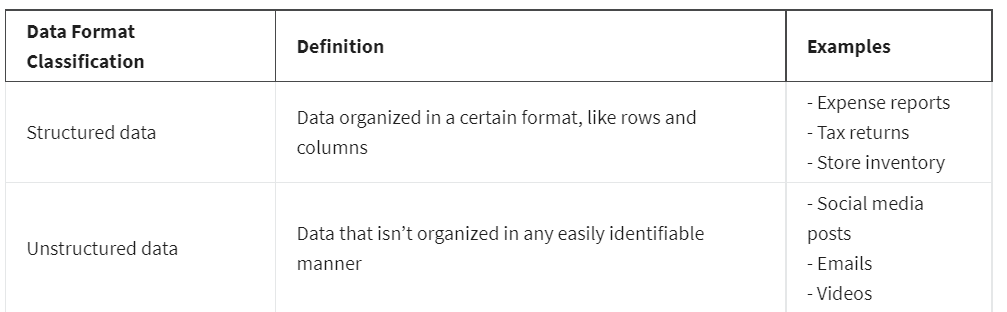
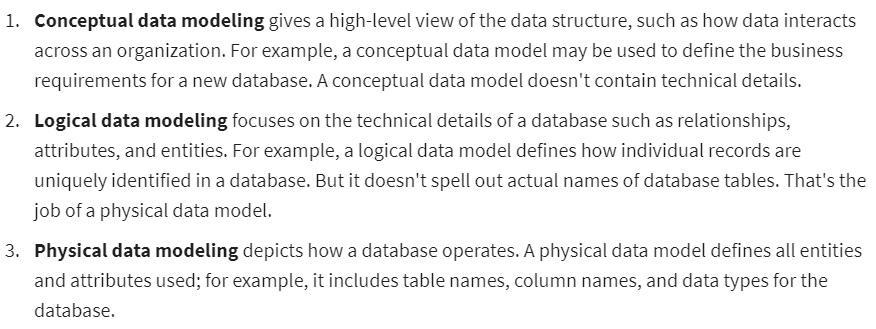
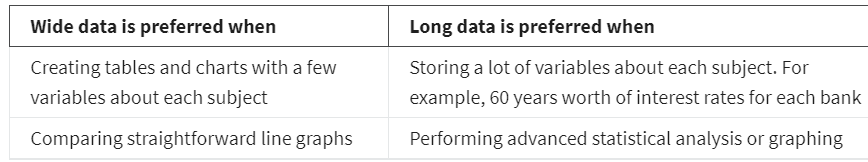
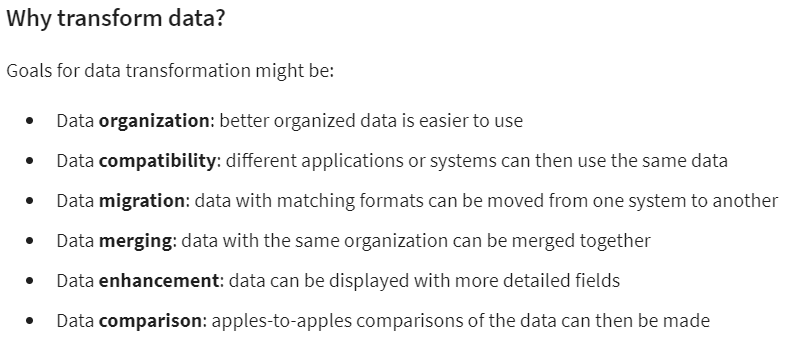
# Week 1

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

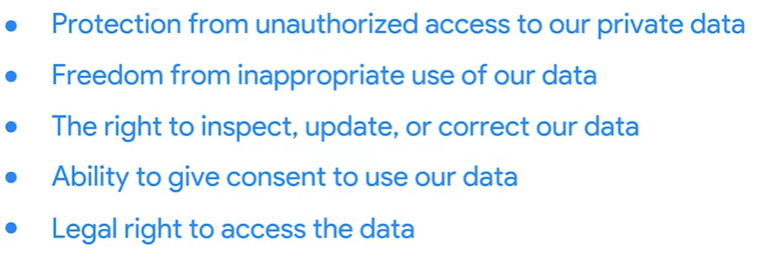


* Long data: data in which each row is one time point per subject, so each subject will have data in multiple rows.
* Wide data: data in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject.
* 
* Why transform data?

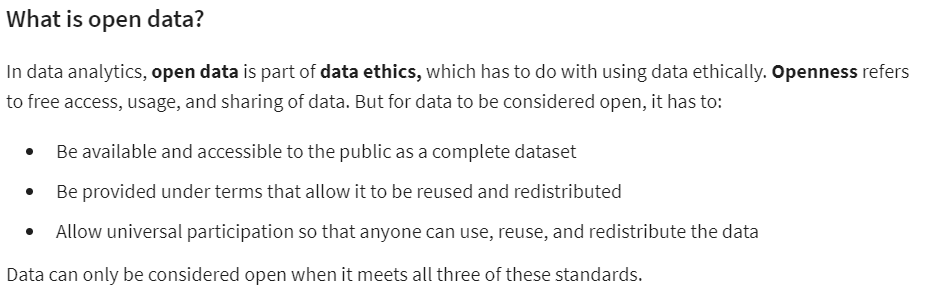


* Bias: a preference in favor of or against a person, group of people, or thing.
* Data bias: a type of error that systematically skews results in a certain direction
* Sampling bias: when a sample isn’t representative of the population as a whole.
* Observer bias (experimenter bias/research bias): the 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

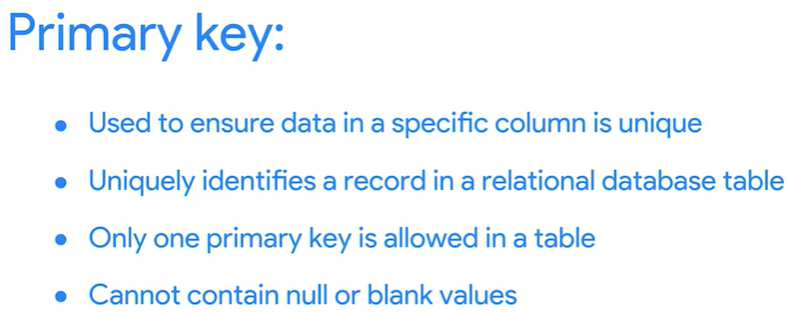
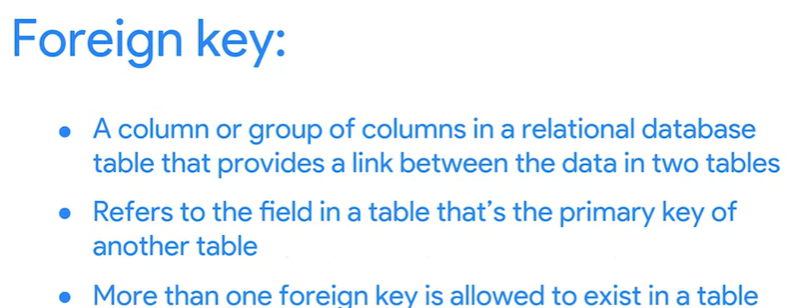


1. Openness: free access, usage, and sharing of data.
2. Transaction transparency: Transaction transparency states that all data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data.

* 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
* 
* 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
* 
* 
* 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

* 