Data Analysis Course 3: Prepare Data for Exploration

Module 1

Data collection
Every piece of info is data.

How data is collected

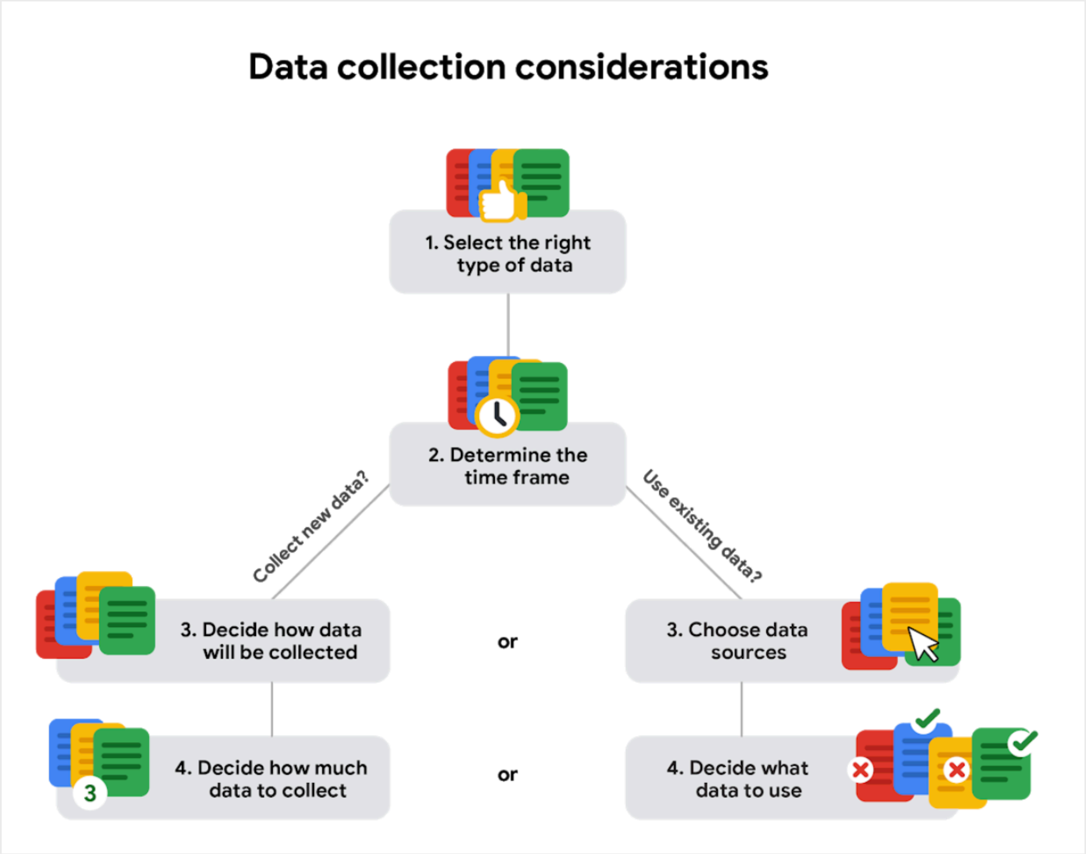
- Interviews
- Observations
- Forms
- Questionnaires
- Surveys
- Cookies

Data collection considerations

- Know 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 timeframe

First-party data: data collected by an individual group using their own resources. Preferred because you 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 collected from outside sources who did not collect it directly.

Population: all possible values in a certain data set.
Sample: a part of a population that is representative of the population.


Data collection considerations

1. Select the right type of data
2. Determine the time frame
Collect new data?
Use existing data?
3. Decide how data will be collected
or
3. Choose data sources
3
4. Decide how much data to collect
or
4. Decide what data to use

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 the 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

- Spreadsheets and relational databases store data in a structured way.
- Organized in a certain formate, such as rows and columns.
- Most often quantitative

Unstructured data: data that is not organized in any easily identifiable manner.

- Audio or video files, photos, emails, social media posts.
- Not organized in any easy-to-identify way.
- Most often qualitative
- Advancements in AI and machine learning algorithms are making this easier to search, manage, and analyze, but this is typically harder to do those things with.

Data model: a model that is used for organizing data elements and how they relate to one another.
Data elements: pieces of info, such as people’s names, account numbers, and addresses.

Data models help keep data consistent and enable people to map out how data is organized.

- 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.
- 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.
- 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.

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.

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
- Boolean

Text or string data type: a sequence of characters and punctuation that contains textual information.

Boolean data type: a data type with only two possible values: true or false.

- Boolean statements include AND, OR, NOT.
    - Work similar to mathematical operators and are used to create logical

statements that filter results.

Tabular data
Rows are records
Columns are fields
Values

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.

Long data: data in which each row is one time point per subject, so each subject will have data in multiple rows.

Data transformation: the process of changing the data’s format, structure, or values.

- 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.
- 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

Terms and definitions for Course 3, Module 1
Agenda: A list of scheduled appointments

Audio file: Digitized audio storage usually in an MP3, AAC, or other compressed format

Boolean data: A data type with only two possible values, usually true or false

Continuous data: Data that is measured and can have almost any numeric value

Cookie: A small file stored on a computer that contains information about its users

Data element: A piece of information in a dataset

Data model: A tool for organizing data elements and how they relate to one another

Digital photo: An electronic or computer-based image usually in BMP or JPG format

Discrete data: Data that is counted and has a limited number of values

External data: Data that lives, and is generated, outside of an organization

Field: A single piece of information from a row or column of a spreadsheet; in a data table, typically a column in the table

First-party data: Data collected by an individual or group using their own resources

Long data: A dataset in which each row is one time point per subject, so each subject has data in multiple rows

Nominal data: A type of qualitative data that is categorized without a set order

Ordinal data: Qualitative data with a set order or scale

Ownership: The aspect of data ethics that presumes individuals own the raw data they provide and have primary control over its usage, processing, and sharing

Pixel: In digital imaging, a small area of illumination on a display screen that, when combined with other adjacent areas, forms a digital image

Population: In data analytics, all possible data values in a dataset

Record: A collection of related data in a data table, usually synonymous with row

Sample: In data analytics, a segment of a population that is representative of the entire population

Second-party data: Data collected by a group directly from its audience and then sold

Social media: Websites and applications through which users create and share content or participate in social networking

String data type: A sequence of characters and punctuation that contains textual information (Refer to Text data type)

Structured data: Data organized in a certain format such as rows and columns

Text data type: A sequence of characters and punctuation that contains textual information (also called string data type)

United States Census Bureau: An agency in the U.S. Department of Commerce that serves as the nation's leading provider of quality data about its people and economy

Unstructured data: Data that is not organized in any easily identifiable manner

Video file: A collection of images, audio files, and other data usually encoded in a compressed format such as MP4, MV4, MOV, AVI, or FLV

Wide data: A dataset in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject

Module 2

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

unbiased sampling: when a sample is representative of the population being measured

observer bias (aka experimenter bias/research bias): the tendency for different people to observe things differently.

interpretation bias: tendency to always interpret ambiguous situations in a positive or negative way.

confirmation bias: tendency to search for or interpret information in a way that confirms pre-existing beliefs.

ROCCC (Good data sources)

- Reliable
- Original
- Comprehensive
- Current
- Cited

Bad data is the opposite of ROCCC
- Unreliable
- Unoriginal
- Missing important information
- Out of date, irrelevant
- No cited

Every good solution if found by avoiding bad data.

Ethics: well-founded standards of right and wrong the prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues.

Data ethics: well-founded standards of right and wrong that dictate how data is collected, shared, and used.

GDPR

Aspects of data ethics
- **ownership**: who owns data? individuals own the raw data they provide and they have primary control over its usage, how it's processed, and how it's shared.
- **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.
- **currency**: individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions.
- **privacy**: preserving a data subject's information and activity any time a data transaction occurs. (information privacy/data protection)
  - 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

- **openness**: free access, usage, and sharing of data

data anonymization: the process of protecting people's private or sensitive d at a by eliminating personally identifiable information.

- typically involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values.
- health care and financial data are two of the most sensitive types.

personally identifiable information (PII): information that can be used by itself or with other data to track down a person's identity.
de-identification: a process used to wipe data clean of all personally identifying information.

Understanding Open Data
In order for data to be considered open, it has to:

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

Open data allows credible data to be available more widely.

data interoperability: the ability of data systems and services to openly connect and share data.

Terms and definitions for Course 3, Module 2
Bad data source: A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

Bias: A conscious or subconscious preference in favor of or against a person, group of people, or thing

Confirmation bias: The tendency to search for or interpret information in a way that confirms pre-existing beliefs

Consent: The aspect of data ethics that presumes an individual's right to know how and why their personal data will be used before agreeing to provide it

Cookie: A small file stored on a computer that contains information about its users

Currency: The aspect of data ethics that presumes individuals should be aware of financial transactions resulting from the use of their personal data and the scale of those transactions

Data anonymization: The process of protecting people's private or sensitive data by eliminating identifying information

Data bias: When a preference in favor of or against a person, group of people, or thing systematically skews data analysis results in a certain direction

Data ethics: Well-founded standards of right and wrong that dictate how data is collected, shared, and used

Data interoperability: A key factor leading to the successful use of open data among companies and governments

Data privacy: Preserving a data subject’s information any time a data transaction occurs

Ethics: Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues

Experimenter bias: The tendency for different people to observe things differently (also called observer bias)

Fairness: A quality of data analysis that does not create or reinforce bias

First-party data: Data collected by an individual or group using their own resources

General Data Protection Regulation of the European Union (GDPR): Policy-making body in the European Union created to help protect people and their data

Good data source: A data source that is reliable, original, comprehensive, current, and cited (ROCCC)

Interpretation bias: The tendency to interpret ambiguous situations in a positive or negative way

Observer bias: The tendency for different people to observe things differently (also called experimenter bias)

Open data: Data that is available to the public

Openness: The aspect of data ethics that promotes the free access, usage, and sharing of data

Sampling bias: Overrepresenting or underrepresenting certain members of a population as a result of working with a sample that is not representative of the population as a whole

Transaction transparency: The aspect of data ethics that presumes all data-processing activities and algorithms should be explainable and understood by the individual who provides the data

Unbiased sampling: When the sample of the population being measured is representative of the population as a whole

Module 3

database: a collection of data stored in a computer system.

metadata: data about data

database features and components

relational database: a database that contains a series of related tables that can be connected to form relationships.

- When using a relational database, data analysts write queries to request data from the related tables.

primary key: an identifier that references a column in which each value is unique

- should be unique, no two rows should have the same 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

foreign key: a field within a table that is a primary key in another table

- how one table can be connected to another
- a column or group of columns in a 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
- more than one foreign key is allowed to exist in a table

normalization: process of organizing data in a relational database.

- creating tables and establishing relationships between those tables.

composite key: a primary key constructed using multiple columns of a table.

metadata is used in database management to help data analysts interpret the contents of the data within the database. Puts data into context and makes the data more understandable. metadata is stored in a single, central location and it gives the company standardized information about all of its data.
three types:

- descriptive metadata: describes a pieces of data and can be used to identify it at a later point in time. (ISBN #, author, title of book)
- structural metadata: indicates how a piece of data is organized and whether it is part of one, or more than one, data collection. (how pages make chapters)
- administrative metadata: indicates the technical source of a digital asset. (type of file, date and time of photo)

metadata helps analysts confirm reliability by making sure it is:

- accurate
- precise
- relevant
- timely

when data is uniform it is:

- organized: Data analysts can easily find tables and files, monitor the creation and alteration of assets, and store metadata.
- classified: Data analysts can categorize data when it follows a consistent format, which is beneficial in cleaning and processing data.
- stored: Consistent and uniform data can be efficiently stored in various data repositories. This streamlines storage management tasks such as managing a database.
- accessed: Users, applications, and systems can efficiently locate and use data.

Metadata repositories are specialized databases specifically created to store and manage metadata. They can be kept in a physical location or a virtual environment—like data that exists in the cloud.

- help analysts ensure their data is reliable and consistent.

data governance: a process to ensure the formal management of a company's data assets.

In Google Sheets, the **IMPORTRANGE** function can import all or part of a dataset from another Google Sheet.

In Google Sheets, you can use the **IMPORTHTML** function to import the data from an HTML table (or list) on a web page

You can use the **IMPORTDATA** function in a Google Sheet to import data into a Google Sheet.

Sorting data: arranging data into a meaningful order to make it easier to understand, analyze, and visualize.

Freeze header row: select top row, go to View, Freeze, 1 Row

Select whole table, Data, Advanced, Sort range

- Can add more than one sorting

Filtering: showing only the data that meets a specific criteria while hiding the rest.

BigQuery

- Sandbox
    - 12 projects at a time
    - Cannot add new records, update records, has limits on amount of data you can process
- Free trial
    - More access, 90 day trial

Data manipulation language (DML)

When you select a dataset in BigQuery you will have a few tabs that describe it:

- Schema, which displays the column names in the dataset
- Details, which contains additional metadata, such as the creation date of the dataset
- Preview, which shows the first rows from the dataset

Terms and definitions for Course 3, Module 3

Administrative metadata: Metadata that indicates the technical source of a digital asset

CSV (comma-separated values) file: A delimited text file that uses a comma to separate values

Data governance: A process for ensuring the formal management of a company’s data assets

Descriptive metadata: Metadata that describes a piece of data and can be used to identify it at a later point in time

Foreign key: A field within a database table that is a primary key in another table (Refer to primary key)

FROM: The section of a query that indicates where the selected data comes from

Geolocation: The geographical location of a person or device by means of digital information

Metadata: Data about data

Metadata repository: A database created to store metadata

Naming conventions: Consistent guidelines that describe the content, creation date, and version of a file in its name

Normalized database: A database in which only related data is stored in each table

Notebook: An interactive, editable programming environment for creating data reports and showcasing data skills

Primary key: An identifier in a database that references a column in which each value is unique (Refer to foreign key)

Redundancy: When the same piece of data is stored in two or more places

Schema: A way of describing how something, such as data, is organized

SELECT: The section of a query that indicates the subset of a dataset

Structural metadata: Metadata that indicates how a piece of data is organized and whether it is part of one or more than one data collection

WHERE: The section of a query that specifies criteria that the requested data must meet

World Health Organization: An organization whose primary role is to direct and coordinate international health within the United Nations system

Module 4

Best practices when organizing data

- Naming conventions
- Foldering
- Archiving older files
- Align your naming and storage practices with your team
- Develop metadata practices

Naming conventions: consistent guidelinesthat describe the content, date, or version of a file in its name

File-naming conventions help you organize, access, process, and analyze data because they act as quick reference points to identify what's in a file.

data security: protecting data from unauthorized access of corruption by adopting safety measures.

Encryption uses a unique algorithm to alter data and make it unusable by users and applications that don't know the algorithm.

Tokenization replaces the data elements you want to protect with randomly generated data referred to as a "token."

Version control enables all collaborators within a file to track changes over time. You can understand who made what changes to a file, when they were made, and why.

Terms and definitions for Course 3, Module 4

Access control: Features such as password protection, user permissions, and encryption that are used to protect a spreadsheet
Data security: Protecting data from unauthorized access or corruption by adopting safety measures
Inbox: Electronic storage where emails received by an individual are held