Course 3:

Prepare Data for Exploration

Week 1: Data Types and Structures

23-01-2022

Introduction to data exploration

- Understanding the different types of data and data structures
- What type of data is right for the question you're answering
- Practical skills about how to extract, use, organize and protect your data

Learning objectives

- How data is generated
- Different formats, types, and structures of data
- Analyze data for bias and credibility
- What "clean data" means
- Databases
- Extract your own data using spreadsheets and SQL
- The basics of data organization
- The process of protecting your data

Collecting Data

Data Collection in our world

Every piece of information is data. All that data is usually generated as a result of our activity in the world.

How data is collected:

- interviews
- observations
- forms
- questionnaires
- surveys
- cookies

Knowing how the data is generated can help add context to the data, and knowing how to collect it can make the data analysis process more efficient.

Determining What Data to Collect

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

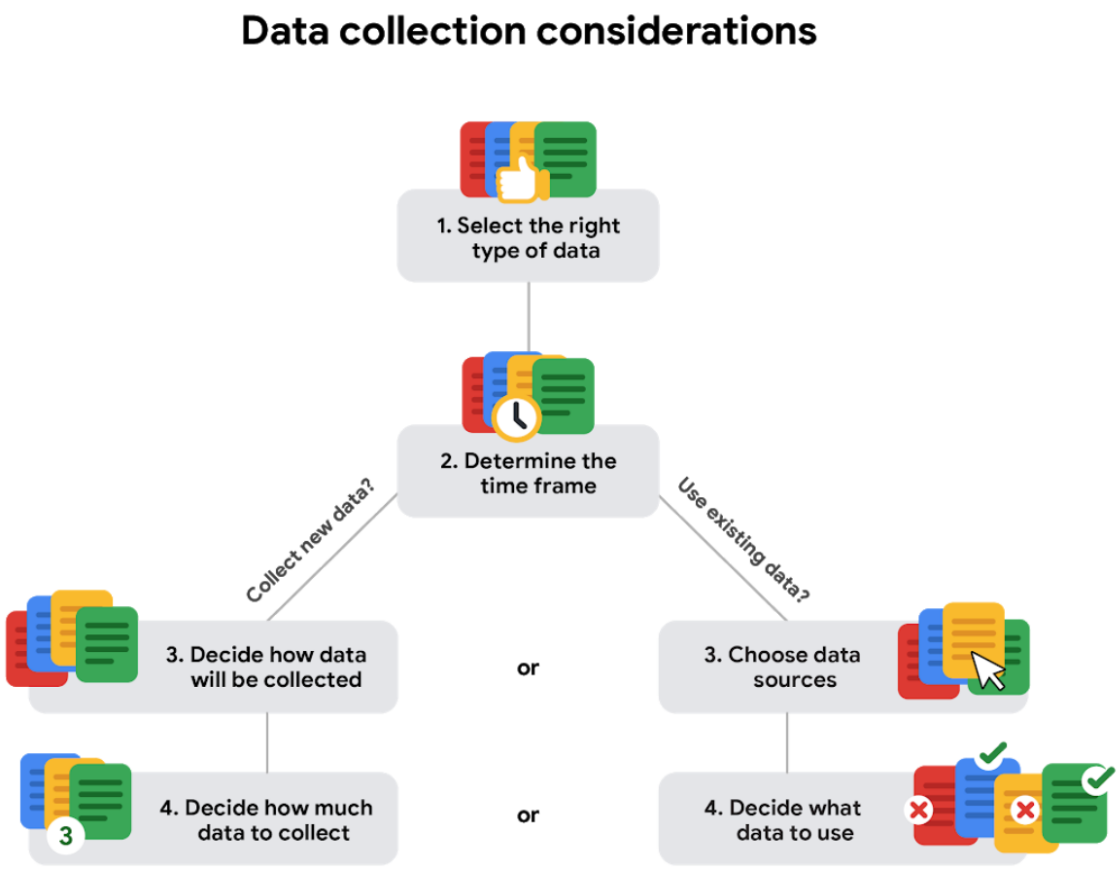
Third-party data: Data collected by outside sources who did not collect it directly

Population: All possible data values in a certain dataset

Sample: A part of a population that is representative of the population

Data collection considerations:

- how the data will be collected
- choose data sources
- decide what data to use
- how much data to collect
- select the right data type
- determine the time frame



Differentiate between data formats and structures

Discover data formats

Qualitative and Quantitative data

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

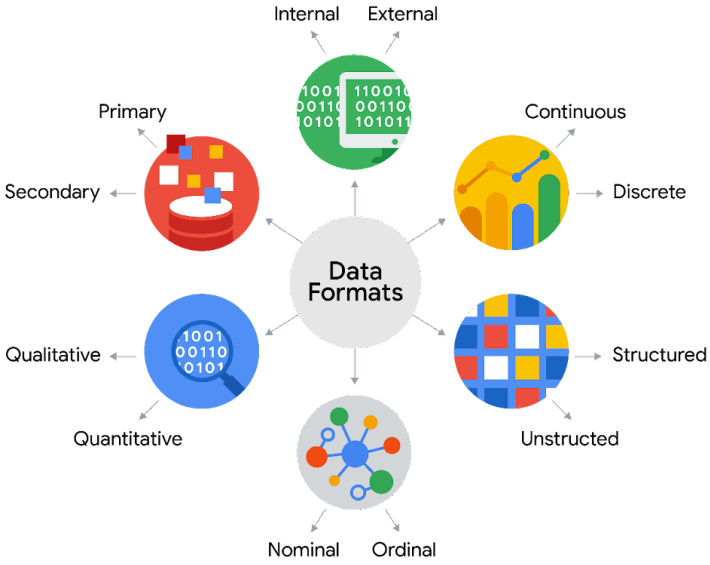
Ordinal data: A type of qualitative data with a set order or scale

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

Unstructured data: Data that is organized in any easily identifiable manner



Understanding structured data

Examples:

- audio files
- video files
- emails
- photos
- social media

Data model: A model that is used for organizing data elements and how they relate to one another.

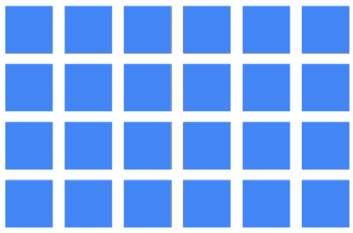
Structured data works within a data model.

Data elements: Pieces of information, such as people's names, account numbers, and addresses

Sources of structured data:

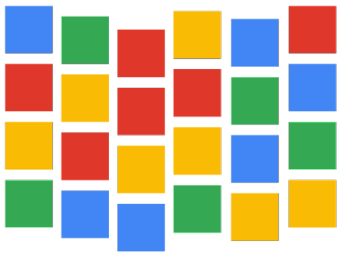
- spreadsheets
- databases that store data

Structured data



- Defined data types
- Most often quantitative data
- Easy to organize
- Easy to search
- Easy to analyze
- Stored in relational databases & data warehouses
- Contained in rows and columns
- Examples: Excel, Google Sheets, SQL, customer data, phone records, transaction history

Unstructured data



- Varied data types
- Most often qualitative data
- Difficult to search
- Provides more freedom for analysis
- Stored in data lakes, data warehouses, and NoSQL databases
- Can't be put in rows and columns
- Examples: Text messages, social media comments, phone call transcriptions, various log files, images, audio, video

31-01-2022

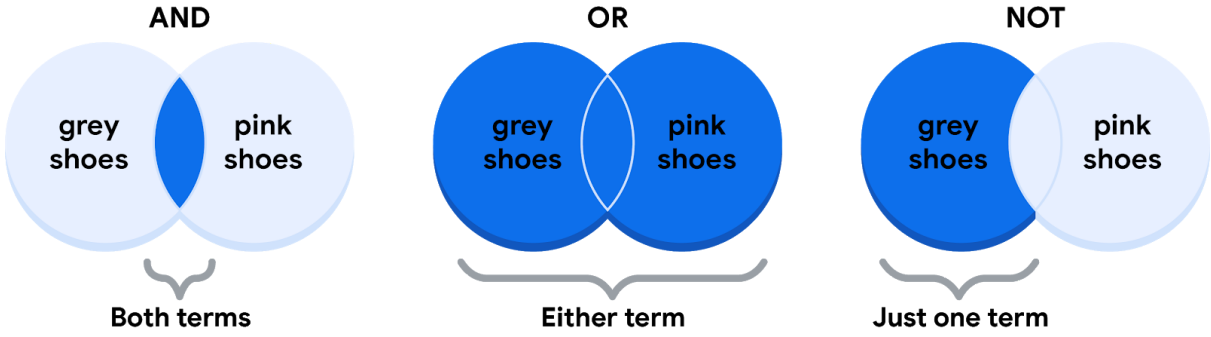
Explore data types, fields, and values

Data Type: A specific kind of data attribute that tells what kind of value the data is

Data Types in spreadsheets

- number
- text or string: a sequence of characters and punctuation that contains textual information
- boolean: a data type with only two possible values such as TRUE or FALSE

Boolean Logic



AND, OR, and NOT operators are helpful in forming conditions. Boolean Logic allows stacking multiple conditions together to filter results.

Data Table Components

Rows → Records

Columns → Fields

Meet wide and long data

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: It is the process of changing the data's format, structure, or values. It 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

Goals for data transformation might be:

- data organization
- data compatibility
- data migration
- data merging
- data enhancement

- data comparison

Wide data is preferred when

- creating tables and charts with a few variables about each subject
- comparing straightforward line graphs

Long data is preferred when

- storing a lot of variables about each subject
- performing advanced statistical analysis or graphing

Week 2: Prepare data for Exploration

31-01-2022

Ensuring data integrity

analyze data for bias and credibility

good data vs. bad data

data ethics, privacy, and access

Bias: From questions to conclusions

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

Biased and unbiased data

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

Understanding bias in data

- **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

Identifying good data sources

Reliable

Original

Comprehensive

Current

Cited

What is "bad" data?

Do not follow ROCCC.

Data ethics and privacy

Introduction to data ethics

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.

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

GDPR: General Data Protection Regulation of the European Union

Aspects of data ethics:

- Ownership: Individuals own the raw data they provide and they have primary control over its usag, 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.

    - 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: It is the process of protecting people's private or sensitive data by eliminating that kind of information. Typically, it involves blanking, hashing, or masking personal information.

Personally identifiable information: PII is 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. Healthcare and financial data are two of the most sensitive types of data.

Features of Open Data

Data interoperability: The ability of data systems and services to openly connect and share data

For data to be considered open it has to be:

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

Sites for trustworthy open data:

- https://www.data.gov/
- https://www.census.gov/data.html
- https://www.opendatanetwork.com/
- https://cloud.google.com/public-datasets
- https://datasetsearch.research.google.com/

05-02-2022

Week 3: Prepare Data For Exploration

Database: Collection of data stored in a computer system.

Metadata: Data about data

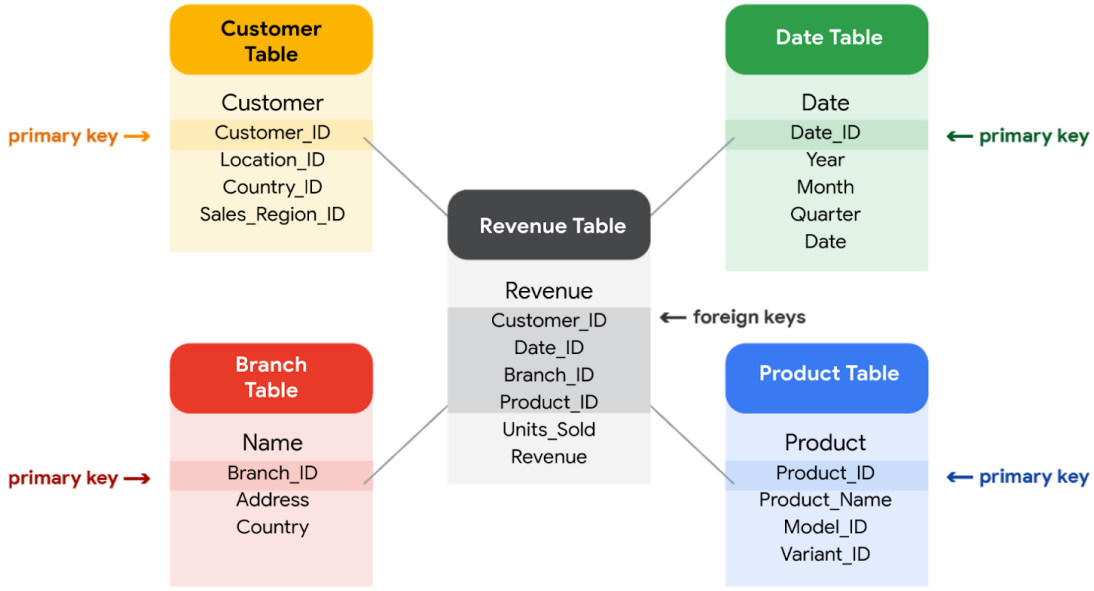
Relational Database: A database that contains a series of related tables that can be connected via their relationships

Primary key: An identifier that references a column in which each value is unique.

- 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 of blank values

Foreign key: A field within a table is a primary key in another table.

- 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



Databases use a special language to communicate called a query language. **Structured Query Language (SQL)** is a type of query language that lets data analysts communicate with a database.

Exploring MetaData

Metadata:

- Data about Data
- Metadata is used in databases management to help data analysts interpret the contents of the data within the database
- Metadata creates a single source of truth by keeping things consistent and uniform
- Metadata also makes data more reliable by making sure it's accurate, precise, relevant, and timely

Common types of metadata:

- descriptive
- structural
- administrative

Descriptive metadata:

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

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.

Administrative metadata:

Metadata that 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 repositories make it easier and faster to bring together multiple sources for data analysis.

- describe the state and location of the metadata
- describe the structures of the tables inside
- describe how the data flows through the repository
- keep track of who accesses the metadata and when

Metadata management

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.

09-02-2022

Accessing different data sources

Internal data: data that lives within a company's own systems.

External data: data that lives and is generated outside an organization

Importing data from spreadsheets and databases

CSV = Comma - separated values

A CSV file saves data in a table format.

Sorting and Filtering

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

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

Working with large datasets in SQL

BigQuery: Sandbox

09-02-2022

Week 4: Organizing and protecting your data

Benefits of organizing data

- makes it easier to find and use
- helps you avoid making mistakes during your analysis
- helps to protect your data

Let's get organized

Best practices when organizing data:

- 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
- Foldering
  - organizing your files into folders
  - breakdown folders into subfolders
- Archiving old files
  - move old projects to a separate location to create an archive and cut down on clutter
- Align your naming and storage practices with your team
- Develop metadata practices
- Think about how often you're making copies of data and storing it in different places

Security features in spreadsheets

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

10-02-2022

Week 5: Engaging in the data community

A professional online presence can

- help potential employers find you
- make connections with other analysts
- learn and share data findings
- participate in community events

LinkedIn

Github

Networking: Professional relationship building

Podcasts: Partially Derivative, O’Reilly Data Show

Blogs: O’Reilly, Kaggle, KDnuggets, Github, Medium