Prepare Data for Exploration

Saturday, January 28, 2023 22:21

Data Exploration

Wednesday, September 21, 2022 18:50

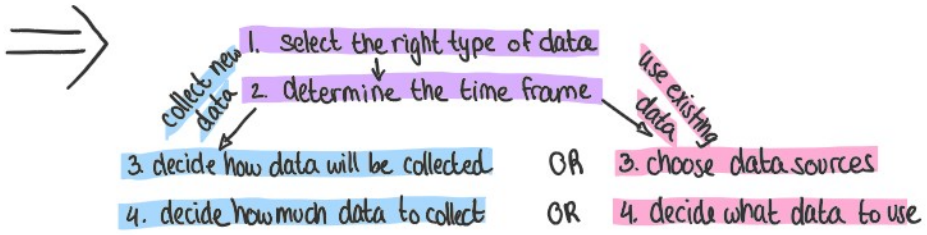
A lot of data is generated every moment.

How data is collected:

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

Determining what data to collect

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



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

Population: All possible data values in a certain dataset

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

Data formats and structures

Friday, September 23, 2022 22:57

qualitative data: can't be counted, measured, or easily expressed using numbers, usually listed as a name, category, or description

↳ **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

quantitative data: can be measured or counted and then expressed as a number, data with a certain quantity, amount, or range

↳ **discrete data**: data that is counted and has a limited number of values

↳ **continuous data**: that is measured and can have almost any numeric value

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 not organized in any easily identifiable manner → audio & video files, emails, photos, social media

data model: a model that is used for organizing data elements and how they relate to one another, diagrams that visually represent how data is

elements and how they relate to one another, diagrams that visually represent how data is organized and structured (blueprint of a house)

data elements: pieces of information, such as people's names, account numbers, and addresses

Levels of data modeling

1. **Conceptual data modeling**: high-level view of the data structure, eg may be used to define the business requirements for a new database, doesn't contain technical details
2. **logical data modeling**: focuses on the technical details of a database such as relationships, attributes and entities, eg. defines how individual records are uniquely identified in a database, doesn't spell out actual names of dataset tables
3. **Physical data modeling**: depicts how a database operates, defines all entities and attributes used, includes table names, column names, and data types for the data base

Data modeling techniques

- Entity Relationship Diagram (ERD)
- Unified Modeling Language (UML)
- Data Dictionary

Data types, fields, and values

Saturday, September 24, 2022 15:29

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
- Boolean

text or string: a sequence of characters and punctuation that contains textual information

boolean data type: a data type with only two possible values, such as TRUE or FALSE

data table components: Rows → records
Columns → fields

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 (can make data more compact)

Transforming data

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

why transform data?

- 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 comparison can then be made

Data Integrity

Sunday, September 25, 2022 14:35

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 -> use random sampling to avoid

↳ **observer bias** (experimenter/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 - provides accurate, unbiased, and complete information that's been vetted and proven fit for use

Original - if data comes from second or third party, validate it with the original source

Comprehensive - contains all critical information needed to answer the question or find the solution

Current - the usefulness of data decreases as time passes

Cited - citing makes the information more credible > who created the dataset? > is it part of a credible organization? > when was the data last refreshed?

What is "bad" data?

not reliable - inaccurate, incomplete or biased

not original - can't locate first-party information

not original - can't locate first-party information

not comprehensive - missing important information, may contain human error

not current - out of date and irrelevant

not cited - not cited or vetted

Data Ethics

Sunday, September 25, 2022 15:29

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
- Transaction transparency
- Consent
- Currency
- Privacy
- Openness

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

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

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

Personally Identifiable Information (PII) is information that can be used by itself or with other data to track down a person's identity

Data anonymization is the process of protecting people's private or sensitive data by eliminating PII. usually involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values.

Healthcare and financial data usually goes through de-identification, a process to wipe data clean of all PII. Every industry uses data anonymization, data that is often anonymized:

- phone numbers
- names
- email addresses
- license plates
- Social security numbers
- Photographs
- IP addresses
- medical records
- account numbers

Open Data

Sunday, September 25, 2022 19:12

Free access, usage, and sharing of data.

There are standards around **availability and access**

- open data must be available as a whole, preferably by downloading over the internet in a convenient and modifiable format

around **reuse and distribution**

- open data must be provided under terms that allow redistribution and reuse including the ability to use it with other data sets

around **universal participation**

- everyone must be able to use, reuse, and redistribute the data
- there shouldn't be any discrimination against fields, persons, or groups

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

=> open data is difficult to implement

Databases

Tuesday, September 27, 2022 17:37

Database: a collection of data stored in a computersystem.

Metadata: data about data

Database features

Relational database: A database that contains a series of related tables that can be connected via their relationships. For two tables to have a relationship, one or more of the same fields must exist inside both tables.

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
- 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.

- provides a link between the data in two tables
- more than one foreign key is allowed to exist in a table

Database Normalization: the process of organizing data in a relational database, e.g. creating tables and establishing relationships between those tables, applied to eliminate redundancy, increase data integrity, and reduce complexity in a database

Composite key: a primary key constructed using multiple columns of a table

Metadata

Tuesday, September 27, 2022 18:29

Metadata is used in database management to help data analysts interpret the contents of the data within the database. 3 common types of metadata:

- **Descriptive**: describes a piece of data and can be used to identify it at a later point in time
- **Structural**: indicates how a piece of data is organized and whether it is part of one, or more than one, data collection
- **Administrative**: indicates the technical source of a digital asset

Elements of metadata:

- Title and description
- who created the data and when
- who last modified the data and when
- Tags and categories
- Who can access or updat it

Metadata creates a single source of truth by keeping things consistent and uniform. It also makes data more reliable by making sure it's accurate, precise, relevant, and timely.

Metadata repository: a database specifically created to store metadata. They make it easier and faster to bring together multiple sources for data analysis.

- describe state and location of 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. Metadata includes information about where each system is located and where the datasets are located within those systems. The Metadata also describes how all of the data is connected between the various systems.

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

Sorting and Filtering

Wednesday, September 28, 2022 14:48

Sorting data: arranging data into a meaningful order to make it easier to understand, analyze, and visualize. Sort alphabetically, numerically, in ascending or descending order.

Freeze header row before sorting: highlight row → view → freeze → 1 row

To sort: select column by which to sort → drop down arrow → sort sheet A→Z

multiple criteria sorting: select entire dataset → Data → Sort range → Advanced range sorting options → check "data has header row" → select sort by → add another sort column

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

Data → create a Filter → choose the column with the data we need → click filter button in the header and make selection

Working with large datasets in SQL

Wednesday, September 28, 2022 15:41

Setting up BigQuery

- Sandbox
  - no charge, available to anyone with a Google account
  - max 12 projects at a time
  - cannot insert new records to a database
  - cannot update field values of existing records
- Free trial
  - $300 in credit during the first 90 days
  - fewer limitations

Organizing Data

Thursday, September 29, 2022 17:59

Best practices

Naming conventions: consistent guidelines that describe the content, date, or version of a file in its name, create meaningful names, avoid spaces and special characters in filenames

Foldering: Organizing files into folders. > Break folders down into subfolders

> move old projects to a separate location to create an archive and cut down on clutter

Align naming and storage practices with your team

Develop metadata practices

Avoid data duplication (storing the same data in several locations)

Organization can be:
- categorical
- hierarchical
- chronological
- by location

File naming

- work out conventions early
- align file naming with team
- make sure filenames are meaningful
  > reference project name, creation date, revision version, etc
- keep file names short and sweet
- format dates yyyymmdd (international standard)
- lead revision numbers with 0, so that double digit numbers are already built-in
- use hyphens, underscores, or capitalized letters instead of spaces
- create a text file that lays out all naming conventions on a project

Securing Data

Thursday, September 29, 2022 18:51

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

Excel & Google sheets

- let you protect your spreadsheets from being edited
- access control features (password protection, user permissions)

Encryption

- uses a unique algorithm to alter data and make it UNUSABLE by users and applications that DON'T know the algorithm
- this 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"
- the original data is stored in a seperate location and mapped to the tokens
- to access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping
- even if the tokenized data is hacked, the original data is still safe and secure in a separate location