Course 3-Prepare Data for Exploration

Week – 1 Data types and structures

How data is collected

- Interviews, Observations(most often used by scientists), Forms, Questionnaires, Surveys, Cookies.

Data collection considerations

- How the data will be collected
    - Decide if you will collect the data using your own resources or receive (and possibly purchase it) from another party.
- Choose the data sources
    - **First-party data**: Data collected by an individual or group using their own resources. Collecting first-party data is typically the preferred method 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. This data might have come from a number of different sources before you investigated it. It might not be as reliable, but that doesn't mean it can't be useful.

  No matter what kind of data you use, it needs to be inspected for accuracy, bias, and credibility.
- Decide what data to use
    - Choosing the data that can help you find answers and solve problems and not getting distracted by other data.
- How much data to collect
    - A **population** refers to all possible data values in a certain data set.
    - In instances when collecting data from an entire population is challenging, data analysts may choose to use a sample. A **sample** is a part of a population that is representative of that population.

- Select the right data type
- Determine the time frame
    - If you are collecting your own data, decide how long you will need to collect it, especially if you are tracking trends over a long period of time. If you need an immediate answer, you might not have time to collect new data. In this case, you would need to use historical data that already exists.

▼ Data formats

Primary	Secondary
Collected by a researcher from first-hand sources	Gathered by other people or from other research
Examples: • Data from an interview you conducted • Data from a survey returned from 20 participants • Data from questionnaires you got back from a group of workers	Examples: • Data you bought from a local data analytics firm's customer profiles • Demographic data collected by a university • Census data gathered by the federal government

Internal	External
Data that lives inside a company's own systems. Internal data is usually more reliable and easier to collect	Data that lives outside of a company or organization. It is structured.
Examples: • Wages of employees across different business units tracked by HR • Sales data by store location • Product inventory levels across distribution centers	Examples: • National average wages for the various positions throughout your organization • Credit reports for customers of an auto dealership

Continuous	Discrete
Data that is measured and can have almost any numeric value	Data that is counted and has a limited number of values
Examples: • Height of kids in third grade classes (52.5 inches, 65.7 inches) • Runtime markers in a video • Temperature	Examples: • 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

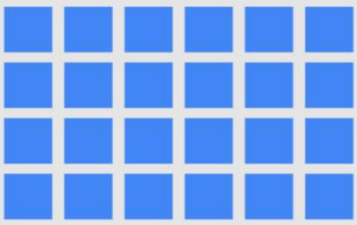
GOOGLE DATA ANALYTICS – COURSE 3

Qualitative	Quantitative
Subjective and explanatory measures of qualities and characteristics	Specific and objective measures of numerical facts
Examples: • Exercise activity most enjoyed • Favorite brands of most loyal customers • Fashion preferences of young adults	Examples: • Percentage of board certified doctors who are women • Population of elephants in Africa • Distance from Earth to Mars

Nominal	Ordinal
A type of qualitative data that isn't categorized with a set order	A type of qualitative data with a set order or scale
Examples: • First time customer, returning customer, regular customer • New job applicant, existing applicant, internal applicant • New listing, reduced price listing, foreclosure	Examples: • 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)

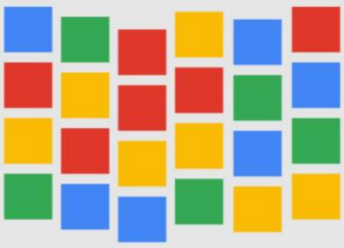
Structured	Unstructured
Data organized in a certain format, like rows and columns	Data that isn't organized in any easily identifiable manner
Examples: • Expense reports • Tax returns • Store inventory	Examples: • Social media posts • Emails • Videos

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

Data modeling

Data modeling is the process of creating diagrams that visually represent how data is organized and structured. These visual representations are called data models. You can think of data modeling as a blueprint of a house. At any point, there might be electricians, carpenters, and plumbers using that blueprint. Each one of these builders has a different relationship to the blueprint, but they all need it to understand the overall structure of the house. Data models are similar; different users might have different data needs, but the data model gives them an understanding of the structure as a whole.

Data model is a model that is used for organizing data elements and how they relate to one another. **Data elements** are pieces of information, such as people's names, account numbers, and addresses. Data models help to keep data consistent and provide a map of how data is organized. This makes it easier for analysts and other stakeholders to make sense of their data and use it for business purposes.

Each level of data modeling has a different level of detail.

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.

Data-modeling techniquesThere 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. As a junior data analyst, you will need to understand that there are different data modeling techniques, but in practice, you will probably be using your organization’s existing technique.

Data modeling can help you explore the high-level details of your data and how it is related across the organization’s information systems. Data modeling sometimes requires data analysis to understand how the data is put together; that way, you know how to map the data. And finally, data models make it easier for everyone in your organization to understand and collaborate with you on your data.

Data modeling keeps data consistent, provides a map of how data is organized, and makes data easier to understand. Data modeling is the process of creating a model that is used for organizing data elements and how they relate to one another.

Data type

A specific kind of data attribute that tells what kind of value the data is.
Data types can be different depending on the query language you're using. For example, SQL allows for different data types depending on which database you're using.

Data types in spreadsheets: Number, Text or String, and Boolean.

A data table, or tabular data, has a very simple structure. It's arranged in rows and columns. You can call the rows **records** and the columns **fields**. They basically mean the same thing, but records and fields can be used for any kind of data table, while rows and columns are usually reserved for spreadsheets. Sometimes a **field** can also refer to a single piece of data, like the value in a cell.

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 lets you easily identify and quickly compare different columns. Wide data is preferred when:

- Creating tables and charts with a few variables about each subject.
- Comparing straightforward line graphs.

Long data

Data in which each row is one time point per subject, so each subject will have data in multiple rows. Long data is a great format for storing and organizing data when there's multiple variables for each subject at each time point that we want to observe. Long data is preferred when:

- Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank.
- Performing advanced statistical analysis or graphing.

Data transformation

Data transformation is the process of changing the data's format, structure, or values.

Data transformation 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

Week2-Bias, credibility, privacy, ethics, and access

Our brains are biologically designed to streamline thinking and make quick judgments.

Bias

A preference in favor of or against a person, group of people, or thing. It can be conscious or subconscious. Once we know and accept that we have bias, we can start to recognize our own patterns of thinking and learn how to manage it.

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. You can avoid this by making sure the sample is chosen at random, so that all parts of the population have an equal chance of being included. **Unbiased sampling** results in a sample that's representative of the population being measured. Another great way to discover if you're working with unbiased data is to bring the results to life with visualizations. This will help you easily identify any misalignment with your sample.
- **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 preexisting beliefs.

How we can go about finding and identifying ***good data*** sources: **Reliable**(Good data sources are reliable), **Original**(To make sure you're dealing with good data, be sure to validate it with the original source), **Comprehensive**(The best data sources contain all critical information needed to answer the question or find the solution), **Current**(The best data sources are current and relevant to the task at hand), **Cited**(Citing makes the information you're providing more credible). Every good solution is found by avoiding bad data. For good data, stick with vetted public data sets, academic papers, financial data and governmental agency data.

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.

Some aspects of data ethics

- **Ownership**: Individuals who own the raw data they provide, and they have primary control over its usage, how it's processed and how it's shared.
- **Transaction transaparency**: 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. This is sometimes called information privacy or 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 our data
- **Openness**: Free access, usage, and sharing of data.

Data anonymization

Personally identifiable information, or 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 that kind of information. Typically, data anonymization involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values. Data anonymization applies to all personally identifiable information, including text and images.

De-identification is a process used to wipe data clean of all personally identifying information.

For data to be considered open, it has to meet all three of these standards:

- Availability and access: Open data must be available as a whole, preferably by downloading over the Internet in a convenient and modifiable form.
- Reuse and redistribution: Open data must be provided under terms that allow reuse and redistribution including the ability to use it with other datasets.
- 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

Interoperability is key to open data's success. It is the ability of data systems and services to openly connect and share data. Different databases using common formats and terminology is an example of interoperability.

One of the biggest benefits of open data is that credible databases can be used more widely. Basically, this means that all of that good data can be leveraged, shared,

and combined with other data. But it is important to think about the individuals being represented by the public, open data, too.

Sites and resources for open data

1. U.S. government data site: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.
2. U.S. Census Bureau: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
3. Open Data Network: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
4. Google Cloud Public Datasets: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
5. Dataset Search: The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

Kaggle's datasets and Data Explorer allow you to search for, access, and upload your own datasets. You can use Kaggle to conduct research, complete data projects, and share your accomplishments with other members of the data science community. Online platforms like Kaggle allow you to search for, view, explore, upload, and work with datasets from a variety of sources and perspectives.

Week3-Databases: Where data lives

A **database** is a collection of data stored in a computer system. **Metadata** is data about data. Metadata tells you where the data comes from, when and how it was created, and what it's all about.

Relational database

A relational database is 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. They also present the same information to each collaborator by keeping data consistent regardless of where it's accessed.

In a non-relational table, you will find all of the possible variables you might be interested in analyzing all grouped together. This can make it really hard to sort through. This is one reason why relational databases are so common in data analysis:

they simplify a lot of analysis processes and make data easier to find and use across an entire database.

There are two types of keys that connect tables in relational databases.:

- A **primary key** is 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 or blank values
  - A primary key may also be constructed using multiple columns of a table. This type of primary key is called a **composite key**.
- A **foreign key** is a field within a table that's a primary key in another table.
  - A column or group of columns in a relational database table that provides a link between the data and 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

Metadata

Metadata is used in database management to help data analysts interpret the contents of the data within the database. Regardless of whether you are working with a large or small quantity of data, metadata is the mark of a knowledgeable analytics team, helping to communicate about data across the business and making it easier to reuse data. In essence, metadata tells the who, what, when, where, which, how, and why of data. Metadata ensures that you are able to find, use, preserve, and reuse data in the future. Data analysts use metadata to combine data, evaluate data, and interpret a database.

3 common types of metadata:

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

Putting data into context is probably the most valuable thing that metadata does, but there are still many more benefits of using metadata.

- 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.

Metadata repository

A database specifically created to store metadata. These repositories describe where metadata came from, keep it in an accessible form so it can be used quickly and easily, and keep it in a common structure for everyone who may need to use it. Using a metadata repository, a data analyst can find it easier to bring together multiple sources of data, confirm how or when data was collected, and verify that data from an outside source is being used appropriately.

Metadata repositories make it easier and faster to bring together multiple sources for data analysis. They do this by describing the state and location of the metadata, the structure of the tables inside, and how data flows through the repository. They even keep track of who accesses the metadata and when.

Metadata is stored in a single, central location and it gives the company standardized information about all of its data. This is done in two ways. First, metadata includes information about where each system is located and where the data sets are located within those systems. Second, the metadata 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. This gives an organization better control of their data and helps a company manage issues related to data security and privacy, integrity, usability, and internal and external data flows.

Metadata specialists organize and maintain company data, ensuring that it's of the highest possible quality. These people create basic metadata identification and discovery information, describe the way different data sets work together, and explain the many different types of data resources. Metadata specialists also create very important standards that everyone follows and the models used to organize the data.

CSV = Comma-separated values. A CSV file saves data in a table format. CSV files use plain text and are delineated by characters, such as a comma. A delineator indicates a boundary or separation between two things. A CSV file makes it easier for data analysts to examine a small part of a large dataset, import data to a new spreadsheet, and distinguish values from one another.

When you work with spreadsheets, there are a few different ways to import data: Other spreadsheets [In Google Sheets, you can use the IMPORTRANGE function], CSV files [In Google Sheets, you can use the IMPORTDATA function in a spreadsheet cell to import data using the URL to a CSV file], HTML tables (in web pages) [In Google Sheets, you can use the IMPORTHTML function].

Sorting data

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. A filter simplifies a spreadsheet by only showing us the information we need.

BigQuery is a data warehouse on Google Cloud that data analysts can use to query, filter large datasets, aggregate results, and perform complex operations.

In-depth guide: SQL best practices

Week4-Organizing and protecting your data

Best practices when organizing data:

- Naming conventions: These are consistent guidelines that describe the content, date, or version of a file in its name. Basically, this means you want to use logical and descriptive names for your files to make them easier to find and use. Naming conventions help us organize, access, process, and analyze our data.
- Foldering: Organizing your files into folders helps keep project-related files together in one place
- Archiving older 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 to avoid any confusion.
- Develop metadata practices: Your team might also develop metadata practices like creating a file that outlines project naming conventions for easy reference.

A data analytics team uses metadata to indicate consistent naming conventions for a project.

File naming recommendations:

- Work out and agree on file naming conventions early on in a project to avoid renaming files again and again.
- Align your file naming with your team's or company's existing file-naming conventions.
- Ensure that your file names are meaningful; consider including information like project name and anything else that will help you quickly identify (and use) the file for the right purpose.
- Include the date and version number in file names; common formats are YYYYMMDD for dates and v## for versions (or revisions).
- Create a text file as a sample file with content that describes (breaks down) the file naming convention and a file name that applies it.
- Avoid spaces and special characters in file names. Instead, use dashes, underscores, or capital letters. Spaces and special characters can cause errors in some applications.

Good file organization includes making it easy to find current, related files that are backed up regularly.

Data security

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

Google sheets and excel have features to:

- Protect spreadsheets from being edited
- Control access like password protection and user permissions
- Tabs can also be hidden and unhidden in Sheets and Excel, allowing you to change what data is being viewed. But even hidden tabs can be unhidden by someone else, so be sure you're okay with those tabs still being accessible.

When using data security measures, analysts can choose between protecting an entire spreadsheet or protecting certain cells within the spreadsheet. Data security can be used to protect an entire spreadsheet, specific parts of a spreadsheet, or even just a single cell. Data analysts use encryption and sharing permissions to control who can access or edit a spreadsheet.

Some data security options:

- **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; so if you have the key, you can still use the data in its original form.
- **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 separate 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. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

Week5-Optional: Engaging in the data community

A professional online presence can:

- Help potential employers find you
- Make connections with other data analysts in your field
- Learn and share data findings
- Participate in community events

LinkedIn is specifically designed to help people make connections with other people in their field. It's a great way to follow trends in your industry, learn from industry leaders, and stay engaged with the wider professional community.

GitHub is part code-sharing site, part social media. It has an active community collaborating and sharing insights to build resources. You can talk with other GitHub users on the forum, use the community-driven wikis, or even use it to manage team projects. GitHub also hosts community events where you can meet other people in the field and learn some new things.

Networking is the most effective way to connect with fellow data analysts. Networking can be called professional relationship building. When you're networking, you can meet other professionals and participate in industry-related groups. Your connections will help you increase your knowledge and skills.

A mentor is a professional who shares their knowledge, skills, and experience to help you develop and grow. A mentor helps you skill up. A sponsor helps you move up. A sponsor is a professional advocate who's committed to moving a sponsee's career forward within an organization.