Week I

Introduction to data exploration

Preparing data correctly ✧.*

- *This is where understanding the different types of data and data structures comes in.*
- *Knowing this lets you figure out what type of data is right for the question you're answering.*
- *Plus, you'll gain practical skills about how to extract, use, organize, and protect your data.*

Data collection in our world

How data is collected ✧.*

- Interview
- Observations - studying behavior
- Forms
- Questionnaires
- Surveys
- Cookies - small files stored on computers that contain information about users

Determining what data to collect

Data collection consideration ✧.*

- How the data will be collected
- Choose data sources
    - **First-party 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** - which is 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.

- Decide what data to use
- How much data to collect
    - **Population** - all possible data values in a certain dataset
    - **Sample** - part of a population that is representative of the population
- Select the right data type
- Determine the time frame
    - **Historical data** - data that already exists.

Selecting the right data

Following are some data-collection considerations to keep in mind for your analysis:

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. Data that you collect yourself is called first-party data.

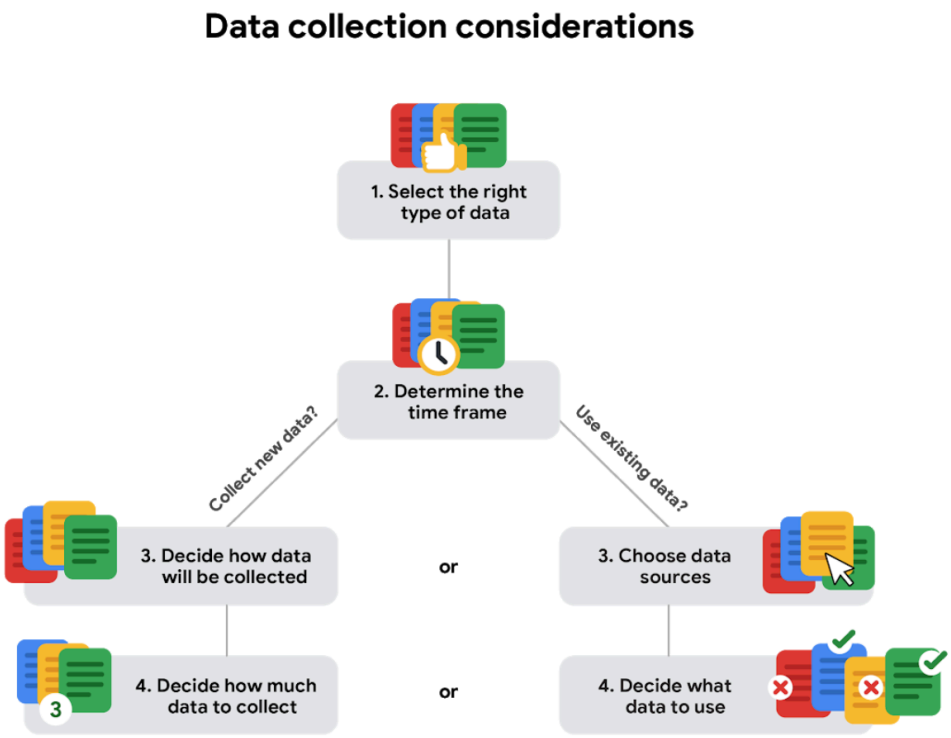
Data sources ✧.*
If you don’t collect the data using your own resources, you might get data from second-party or third-party data providers. Second-party data is collected directly by another group and then sold. Third-party data is sold by a provider that didn’t collect the data themselves. Third-party data might come from a number of different sources.

Solving your business problem ✧.*
Datasets can show a lot of interesting information. But be sure to choose data that can actually help solve your problem question. For example, if you are analyzing trends over time, make sure you use time series data — in other words, data that includes dates.

How much data to collect ✧.*
If you are collecting your own data, make reasonable decisions about sample size. A random sample from existing data might be fine for some projects. Other projects might need more strategic data collection to focus on certain criteria. Each project has its own needs.

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.

Use the flowchart below if data collection relies heavily on how much time you have:



Discover data formats

Qualitative data ✧.*

it can't be counted, measured, or easily expressed using numbers. Qualitative data is usually listed as a name, category, or description.

Quantitative data ✧.*

which can be measured or counted and then expressed as a number. This is data with a certain quantity, amount, or range.

Discrete data ✧.*

data that's counted and has a limited number of values. Discrete data isn't limited to dollar amounts. Examples of other discrete data are stars and points. When partial measurements (half-stars or quarter-points) aren't allowed, the data is discrete. If you don't accept anything other than full stars or points, the data is considered discrete. (Whole value)

Continuous data ✧.*

can be measured using a timer, and its value can be shown as a decimal with several places.

Nominal data ✧.*

a type of qualitative data that's categorized without a set order. In other words, this data doesn't have a sequence.

Ordinal data ✧.*

a type of qualitative data with a set order or scale. If you asked a group of people to rank a movie from 1 to 5, some might rank it as a 2, others a 4, and so on.

Internal data ✧.*

which is data that lives within a company's own systems. The great thing about internal data is that it's usually more reliable and easier to collect

External data ✧.*

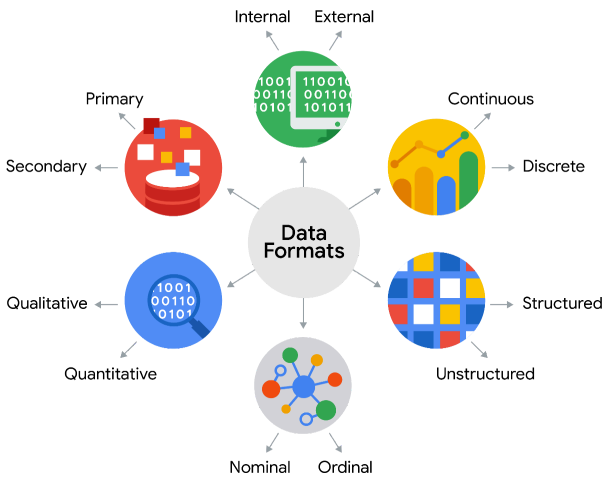
data that lives and is generated outside of an organization. External data becomes particularly valuable when your analysis depends on as many sources as possible. A great thing about this data is that it's structured.

Structured data✧.*

data that's organized in a certain format, such as rows and columns.

Unstructured data ✧.*

data that is not organized in any easily identifiable manner. Audio and video files are examples of unstructured data because there's no clear way to identify or organize their content. Unstructured data might have internal structure, but the data doesn't fit neatly in rows and columns like structured data.



Data formats in practice

Data Format Classification	Definition	Examples
Primary data	Collected by a researcher from first-hand sources	- 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
Secondary data	Gathered by other people or from other research	- 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 data	Data that lives inside a company's own systems	- Wages of employees across different business units tracked by HR - Sales data by store location - Product inventory levels across distribution centers
External data	Data that lives outside of a company or organization	- National average wages for the various positions throughout your organization - Credit reports for customers of an auto dealership
Continuous data	Data that is measured and can have almost any numeric value	- Height of kids in third grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature
Discrete data	Data that is counted and has a limited number of values	- 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
Qualitative	Subjective and explanatory measures of qualities and characteristics	- Exercise activity most enjoyed - Favorite brands of most loyal customers - Fashion preferences of young adults

Quantitative	Specific and objective measures of numerical facts	- Percentage of board certified doctors who are women - Population of elephants in Africa - Distance from Earth to Mars
Nominal	A type of qualitative data that isn't categorized with a set order	- First time customer, returning customer, regular customer - New job applicant, existing applicant, internal applicant - New listing, reduced price listing, foreclosure
Ordinal	A type of qualitative data with a set order or scale	- 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 data	Data organized in a certain format, like rows and columns	- Expense reports - Tax returns - Store inventory
Unstructured data	Data that isn't organized in any easily identifiable manner	- Social media posts - Emails - Videos

Understanding structured data

Unstructured data examples ✧.*

- *Audio files*
- *Video files*
- *Emails*
- *Photos*
- *Social media*

Structured data ✧.*

works nicely within a data model, which is a model that is used for organizing data elements and how they relate to one another.

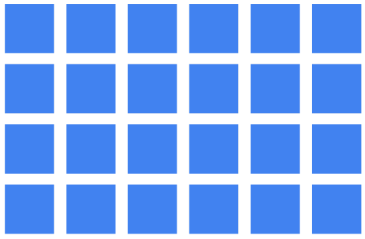
Data elements ✧.*

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. In

addition to working well within data models, structured data is also useful for databases. This makes it easy for analysts to enter, query, and analyze the data whenever they need to.

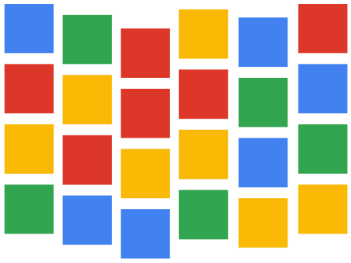
The structured of data

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

The fairness issue ✧.*

The lack of structure makes unstructured data difficult to search, manage, and analyze. But recent advancements in artificial intelligence and machine learning algorithms are beginning to change that. Now, the new challenge facing data scientists is making sure these tools are inclusive and unbiased. Otherwise, certain elements of a dataset will be more heavily weighted and/or represented than others. And as you're learning, an unfair dataset does not accurately represent the population, causing skewed outcomes, low accuracy levels, and unreliable analysis.

Data modeling levels and techniques

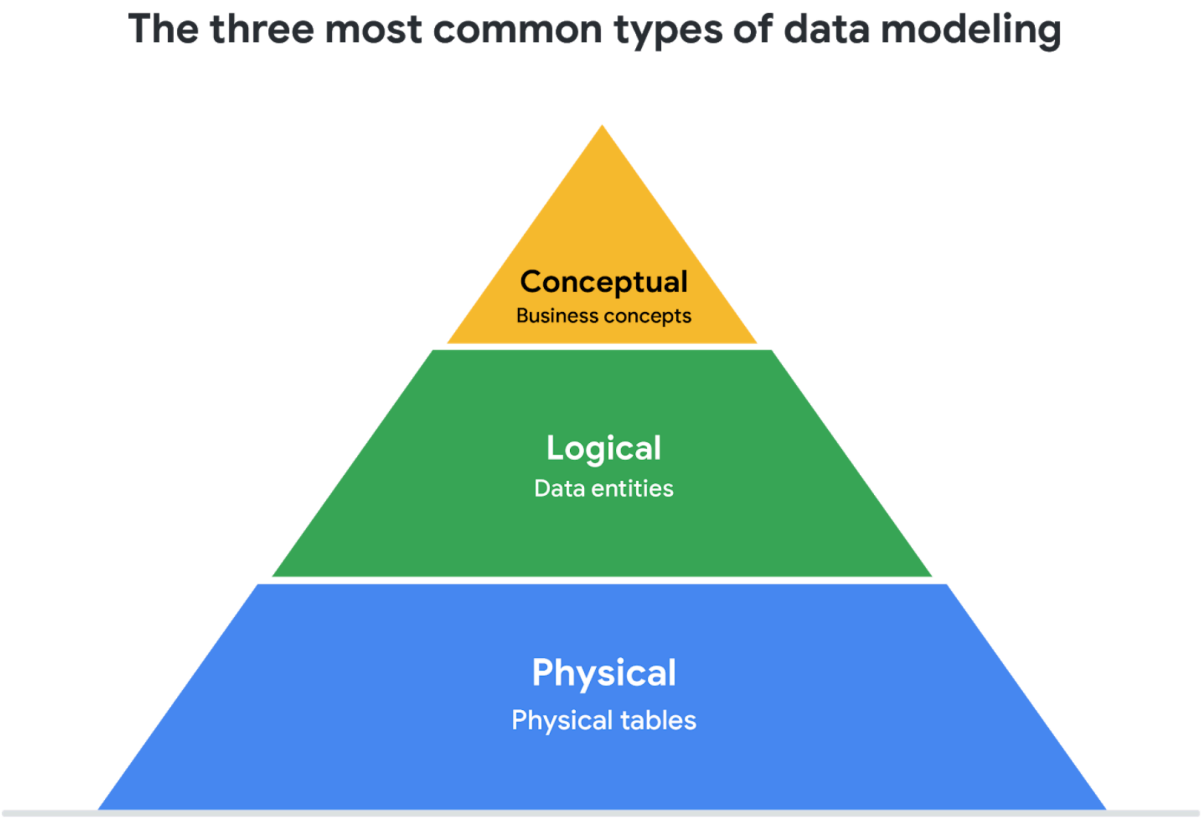
Data models help keep data consistent and enable people to map out how data is organized. A basic understanding makes it easier for analysts and other stakeholders to make sense of their data and use it in the right ways.

Data modeling ✧.*

*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.*

Levels of Data Modelling ✧.*

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 techniques ✧.*

*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. 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 analysis and data modeling ✧.*

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.

Know the type of data you're working with

Data type ✧.*

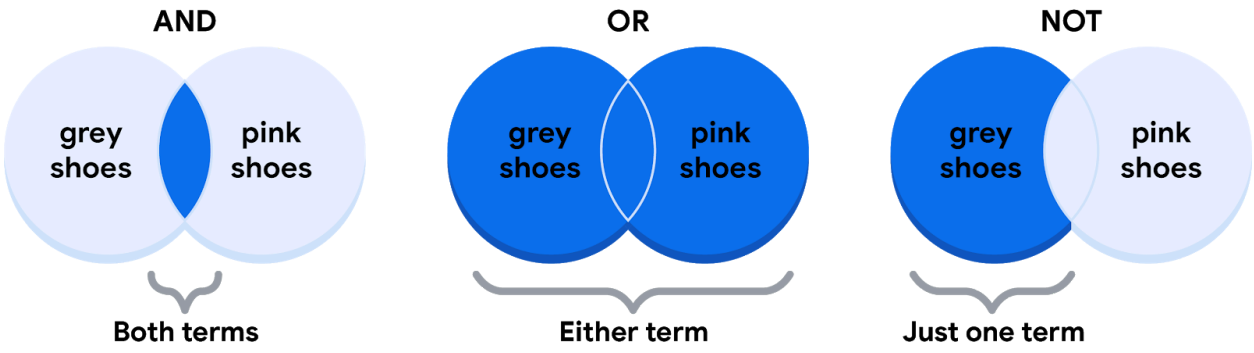
a specific kind of data attribute that tells what kind of value the data is. In other words, a data type tells you what kind of data you're working with.

In spreadsheets, it could be:

- Number
- Text or string - a sequence of characters and punctuation that contains textual information. In this example, that information would be the treats and people's names. These can also include numbers, like phone numbers or numbers in street addresses. But these numbers wouldn't be used for calculations. In this case they're treated like text, not numbers.

- Boolean - a data type with only two possible values: true or false. Boolean logic lets you create multiple conditions to filter your results.

Understanding Boolean Logic



The AND Operator - the AND operator lets you stack multiple conditions.

The OR Operator - the OR operator lets you move forward if either one of your two conditions is met.

The NOT Operator - the NOT operator lets you filter by subtracting specific conditions from the results.

A	B	A and B	A or B	Not A
False	False	False	False	True
False	True	False	True	True
True	False	False	True	False
True	True	True	True	False

Data table components

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. When talking about structured databases, people in data analytics usually go with "records" and "fields."

Rows - records
Columns - fields. Sometimes a field can also refer to a single piece of data, like the value in a cell.

Meet wide and long data

Wide data ✧.*
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.

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. With this long data format, we can store and analyze all of this data using fewer columns. The long data format keeps everything nice and compact.

Transforming data

Data transformation ✧.*
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

***Long data** is data where **each row contains a single data point** for a particular item. In the long data example below, individual stock prices (data points) have been collected for Apple (AAPL), Amazon (AMZN), and Google (GOOGL) (particular items) on the given dates.*

Symbol	Date	Open
AAPL	2018-09-18	217.79
AAPL	2018-09-17	222.15
AAPL	2018-09-14	225.75
AAPL	2018-09-13	223.52
AMZN	2018-09-18	1918.65
AMZN	2018-09-17	1954.73
AMZN	2018-09-14	1992.93
AMZN	2018-09-13	2000
GOOGL	2018-09-18	1162.66
GOOGL	2018-09-17	1177.77
GOOGL	2018-09-14	1188
GOOGL	2018-09-13	1179.7

Wide data *is data where* ***each row contains multiple data points*** *for the particular items identified in the columns.*

Symbol	AAPL	AMZN	GOOGL
Date			
2018-09-13	223.52	2000	1179.7
2018-09-14	225.75	1992.93	1188
2018-09-17	222.15	1954.73	1177.77
2018-09-18	217.79	1918.65	1162.66

Wide data is preferred when	Long data is preferred when
Creating tables and charts with a few variables about each subject	Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank
Comparing straightforward line graphs	Performing advanced statistical analysis or graphing

Week II

Ensuring data integrity

Bias ✧.*
a preference in favor of or against a person, group of people, or thing. Bias can also happen if a sample group lacks inclusivity.

Data bias ✧.*
a type of error that systematically skews results in a certain direction.

Biased and unbiased data

Bias ✧.*
it can systematically skew results in a certain direction, making them unreliable.

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.

Understanding bias in data

You need all sides of the story to avoid sampling bias. The four types of data bias we covered, sampling bias, observer bias, interpretation bias, and confirmation bias, are all unique, but they do have one thing in common. They each affect the way we collect, and make sense of the data.

Observer bias ✧.*
which is sometimes referred to as experimenter bias or research bias. Basically, it's the tendency for different people to observe things differently.

Interpretation bias ✧.*
The tendency to always interpret ambiguous situations in a positive, or negative way. Interpretation bias can lead to two people seeing or hearing the exact same thing, and

interpreting it in a variety of different ways, because they have different backgrounds, and experiences.

Confirmation bias ✧.*
the tendency to search for, or interpret information in a way that confirms preexisting beliefs.

Identifying good data sources

How to identify good data ✧.*

- **Reliable -** trust that you're getting accurate, complete and unbiased information that's been vetted and proven fit for use.
- **Original** - There's a good chance you'll discover data through a second or third party source. 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. Think about it like this. You wouldn't want to work for a company just because you found one great online review about it. You'd research every aspect of the organization to make sure it was the right fit. It's important to do the same for your data analysis.
- **Current** - The usefulness of data decreases as time passes. If you wanted to invite all current clients to a business event, you wouldn't use a 10-year-old client list. The same goes for data. The best data sources are current and relevant to the task at hand.
- **Cited** - Citing makes the information you're providing more credible. When you're choosing a data source, think about three things. Who created the data set? Is it part of a credible organization? When was the data last refreshed?

There's lots of places that are known for having good data. Your best bet is to go with the vetted public data sets, academic papers, financial data, and governmental agency data.

What is “bad” data?

Bad data ✧.*
bad data sources that don't ROCCC. They're not reliable, original, comprehensive, current or cited. Even worse, they could be flat-out wrong or filled with human error.

- **Not reliable -** Bad data can't be trusted because it's inaccurate, incomplete, or biased. This could be data that has sample selection bias because it doesn't reflect the overall population. Or it could be data visualizations and graphs that are just misleading.
- **Not original** - If you can't locate the original data source and you're just relying on second or third party information, that can signal you may need to be extra careful in understanding your data.
- **Not comprehensive** - Bad data sources are missing important information needed to answer the question or find the solution. What's worse, they may contain human error, too.
- **Not current** - Bad data sources are out of date and irrelevant. Many respected sources refresh their data regularly, giving you confidence that it's the most current info available.
- **Not cited** - If your source hasn't been cited or vetted, it's a no-go.

Introduction to data ethics

Ethics ✧.*
refers to 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 ✧.*
refers to well- founded standards of right and wrong that dictate how data is collected, shared, and used. The concept of data ethics and issues related to transparency and privacy are part of the process. Data ethics tries to get to the root of the accountability companies have in protecting and responsibly using the data they collect.

GDPR ✧.*
Stands for General Data Protection Regulation Of The European Union. They create data protection legislation to help protect people and their data.

Aspects of Data Ethics ✧.*
- **Ownership -** This answers the question who owns data? It isn't the organization that invested time and money collecting, storing, processing, and analyzing it. It's 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 transparency -** the idea that all data processing activities and algorithms should be completely explainable and understood by the individual who provides their data. This is in response to concerns over data bias, which we discussed earlier, is a type of error that systematically skews results in a certain direction. Biased outcomes can lead to negative consequences. To avoid them, it's helpful to provide transparent analysis

especially to the people who share their data. This lets people judge whether the outcome is fair and unbiased and allows them to raise potential concerns.

- **Consent -** This is an individual's right to know explicit details about how and why their data will be used before agreeing to provide it. They should know answers to questions like why is the data being collected? How will it be used? How long will it be stored? The best way to give consent is probably a conversation between the person providing the data and the person requesting it. But with so much activity happening online these days, consent usually just looks like a terms and conditions checkbox with links to more details. Let's face it, not everyone clicks through to read those details. Consent is important because it prevents all populations from being unfairly targeted which is a very big deal for marginalized groups who are often disproportionately misrepresented by biased data.
- **Currency -** Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions. If your data is helping to fund a company's efforts, you should know what those efforts are all about and be given the opportunity to opt out.
- **Privacy**
- **Openness**

Introduction to data privacy

Privacy ✧.*
means preserving a data subject's information and activity any time a data transaction occurs.

Information privacy or Data protection ✧.*
It's all about access, use, and collection of data. It also covers a person's legal right to their data. This means someone like you or me:

- should have 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, and
- legal right to access our data.

For companies, it means putting privacy measures in place to protect the individuals' data.

Openness ✧.*
free access, usage, and sharing of data.

Data anonymization

What is data anonymization?

Data anonymization ✧.*
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.

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

What types of data should be anonymized?

Healthcare and financial data are two of the most sensitive types of data. These industries rely a lot on data anonymization techniques. After all, the stakes are very high. That's why data in these two industries usually goes through de-identification.

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

Data anonymization is used in just about every industry. That is why it is so important for data analysts to understand the basics. 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

Features of open data

Openness ✧.*
refers to free access, usage and sharing of data. Sometimes we refer to this as open data, but it doesn't mean we ignore the other aspects of data ethics we covered. We should still be transparent, respect privacy, and make sure we have consent for data that's owned by others. This just means we can access, use, and share that data if it meets these high 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. No one can place restrictions on the data like making it only available for use in a specific industry.

Advantages of open data

- Credible databases can be used more widely.
- Good data can be leveraged, shared, and combined with other data.

- In human health, openness allows us to access and combine diverse data to detect diseases earlier and earlier.
In government, you can help hold leaders accountable and provide better access to community services.

Data Interoperability ✧.*
the ability of data systems and services to openly connect and share data.

The open-data debate

What is open data?
In data analytics, **open data** is part of **data ethics**, which has to do with using data ethically. **Openness** refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

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

Data can only be considered open when it meets all three of these standards.

The open data debate: What data should be publicly available?
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. This could have a huge impact on scientific collaboration, research advances, analytical capacity, and decision-making. But it is important to think about the individuals being represented by the public, open data, too.

Third-party data is collected by an entity that doesn't have a direct relationship with the data. You might remember learning about this type of data earlier. For example, third parties might collect information about visitors to a certain website. Doing this lets these third parties create

audience profiles, which helps them better understand user behavior and target them with more effective advertising.

Personal identifiable information (PII) is data that is reasonably likely to identify a person and make information known about them. It is important to keep this data safe. PII can include a person's address, credit card information, social security number, medical records, and more.

Everyone wants to keep personal information about themselves private. Because third-party data is readily available, it is important to balance the openness of data with the privacy of individuals.

Steps for ethical data use

Things you can do as you're evaluating your dataset in order to ensure that you're looking at it through the various ethical lenses.

1. **To self-reflect and understand what it is that you're doing and the impact that it has.** The best way to challenge that is to question who we are. We being, like, okay, we in this team are trying to build this because we think that that's going to help improve this product or that's going to help inform decisions about what we want to do next. Think about not just those that sit laterally next to you, but also think about those that are represented in this dataset and those that aren't represented in this dataset, and then use that intuition to then continue to question the integrity, the quality, the representation that is present in that dataset.
2. **Think about the various harms and risks associated with the work that you're doing** For example, if you think that you'll benefit from keeping the dataset longer, you may want to also understand what's the risk of holding onto this dataset? What's the potential harm that could arise if you continue to look at the dataset and continue to store it and continue to retrieve this data?
3. **Understanding what the consent process is like.** Are you informing those that you're collecting data from how it's going to be used? What's the communication channel like? Putting on the various ethical lenses, taking a more nuanced approach to your analysis, being cognizant of all the possible risks and harms that can arise when not just analyzing your dataset, but also presenting your dataset. How you portray the results, how they're being used in the decision-making process, whether you are presenting this to management, or presenting this to executives, or presenting this to a larger audience. All of that matters in the responsible use of the dataset.

Sites and resources for open data

Luckily for data analysts, there are lots of trustworthy sites and resources available for open data. It is important to remember that even reputable data needs to be constantly evaluated, but these websites are a useful starting point:

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.

Week III

All about databases

Databases ✧.*
a collection of data stored in a computer system, but storage is just the beginning.

Metadata ✧.*
Data about data. Metadata is extremely important when working with databases. Think of it like a reference guide. Without the guide all you have is a bunch of data with no context explaining what it means. Metadata tells you where the data comes from, when and how it was created, and what it's all about.

Databases 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. You can think of it as a unique identifier for each row in a table. If you do decide to include a primary key, it should be unique, meaning no two rows can have the same primary key. Also, it cannot be null or blank.

- 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
- they cannot contain null or blank values.

Foreign key ✧.*

a field within a table that's a primary key in another table. In other words, a foreign key is how one table can be connected to another.

- 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.

Databases in data analytics

Databases enable analysts to manipulate, store, and process data. This helps them search through data a lot more efficiently to get the best insights.

Relational databases

A **relational database** is a database that contains a series of tables that can be connected to show relationships. Basically, they allow data analysts to organize and link data based on what the data has in common.

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.
Database Normalization

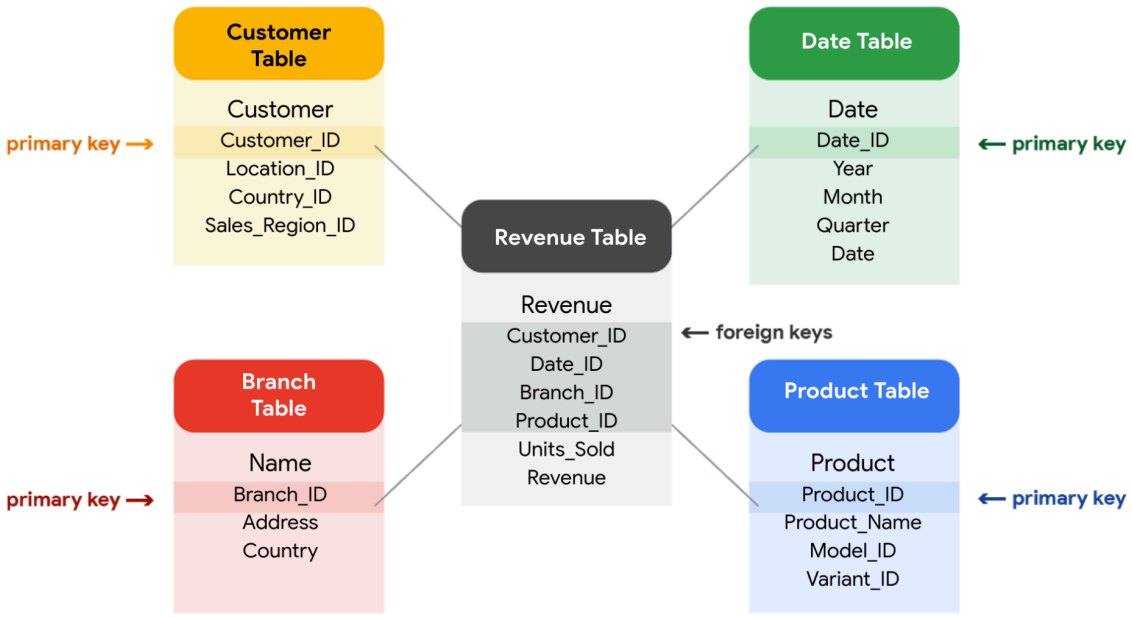
Normalization is a process of organizing data in a relational database. For example, creating tables and establishing relationships between those tables. It is applied to eliminate data redundancy, increase data integrity, and reduce complexity in a database.

The key to relational databases

Tables in a relational database are connected by the fields they have in common. You might remember learning about primary and foreign keys before. As a quick refresher, a **primary key** is an identifier that references a column in which each value is unique. In other words, it's a column of a table that is used to uniquely identify each record within that table. The value assigned to the primary key in a particular row must be unique within the entire table. For example, if customer_id is the primary key for the customer table, no two customers will ever have the same customer_id.

By contrast, a **foreign key** is a field within a table that is a primary key in another table. A table can have only one primary key, but it can have multiple foreign keys. These keys are what create the relationships between tables in a relational database, which helps organize and connect data across multiple tables in the database.

Some tables don't require a primary key. For example, a revenue table can have multiple foreign keys and not have a primary key. A primary key may also be constructed using multiple columns of a table. This type of primary key is called a **composite key**. For example, if customer_id and location_id are two columns of a composite key for a customer table, the values assigned to those fields in any given row must be unique within the entire table.



SQL? You're speaking my language

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. So, a data analyst will use SQL to create a query to view the specific data that they want from

within the larger set. In a relational database, data analysts can write queries to get data from the related tables. SQL is a powerful tool for working with databases

Exploring metadata

Metadata ✧.*
information that's used to describe the data that's contained in something, like a photo or an email. Keep in mind that metadata is not the data itself. Instead, it's data about the data. In data analytics, metadata helps data analysts interpret the contents of the data within a database.

Descriptive ✧.*
metadata that describes a piece of data and can be used to identify it at a later point in time. For instance, the descriptive metadata of a book in a library would include the code you see on its spine, known as a unique International Standard Book Number, also called the ISBN.

Structural ✧.*
metadata that indicates how a piece of data is organized and whether it's part of one or more than one data collection. An example of structural data would be how the pages of a book are put together to create different chapters.

Administrative ✧.*
metadata that indicates the technical source of a digital asset. When we looked at the metadata inside the photo, that was administrative metadata. It shows you the type of file it was, the date and time it was taken, and much more.

Metadata is as important as the data itself

Elements of metadata

Before looking at metadata examples, it is important to understand what type of information metadata typically provides.

Title and description ✧.*
What is the name of the file or website you are examining? What type of content does it contain?

Tags and categories ✧.*
What is the general overview of the data that you have? Is the data indexed or described in a specific way?

Who created it and when ✧.*
Where did the data come from, and when was it created? Is it recent, or has it existed for a long time?

Who last modified it and when ✧.*
Were any changes made to the data? If yes, were the modifications recent?

Who can access or update it ✧.*
Is this dataset public? Are special permissions needed to customize or modify the dataset?

Examples of metadata

In today's digital world, metadata is everywhere, and it is becoming a more common practice to provide metadata on a lot of media and information you interact with. Here are some real-world examples of where to find metadata:

Photos ✧.*
Whenever a photo is captured with a camera, metadata such as camera filename, date, time, and geolocation are gathered and saved with it.

Emails ✧.*
When an email is sent or received, there is lots of visible metadata such as subject line, the sender, the recipient and date and time sent. There is also hidden metadata that includes server names, IP addresses, HTML format, and software details.

Spreadsheets and documents ✧.*
Spreadsheets and documents are already filled with a considerable amount of data so it is no surprise that metadata would also accompany them. Titles, author, creation date, number of pages, user comments as well as names of tabs, tables, and columns are all metadata that one can find in spreadsheets and documents.

Websites ✧.*
Every web page has a number of standard metadata fields, such as tags and categories, site creator's name, web page title and description, time of creation and any iconography.

Digital files ✧.*
Usually, if you right click on any computer file, you will see its metadata. This could consist of file name, file size, date of creation and modification, and type of file.

Books ✧.*

Metadata is not only digital. Every book has a number of standard metadata on the covers and inside that will inform you of its title, author's name, a table of contents, publisher information, copyright description, index, and a brief description of the book's contents.

Using metadata as an analyst

Metadata creates a single source of truth by keeping things consistent and uniform ✧.*
After all, data that's uniform can be organized, classified, stored, accessed, and used effectively. Plus, when a database is consistent, it's so much easier to discover relationships between the data inside it and the data elsewhere.

Metadata also makes data more reliable by making sure it's accurate, precise, relevant, and timely ✧.*
This also makes it easier for data analysts to identify the root causes of any problems that might pop up. The bottom line is, when the data we work with is high quality, it makes things easier and improves our results.

Metadata repository ✧.*
a database specifically created to store metadata. Metadata repositories can be stored in a physical location, or they can be virtual, like data that exists in the cloud. 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. 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 management

Metadata ✧.*
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 ✧.*
the roles and responsibilities of the people who work with the metadata every day. and they 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.

Importing data from spreadsheets and databases

CSV ✧.*
CSV stands for comma-separated values. A CSV file saves data in a table format. CSV files use plain text and they're delineated by characters. So each column or field is clearly distinct from another when importing.

Exploring Public Datasets

Open data helps create a lot of **public datasets** that you can access to make data-driven decisions. Here are some resources you can use to start searching for public datasets on your own:

- The Google Cloud Public Datasets allow data analysts access to high-demand public datasets, and make it easy to uncover insights in the cloud.
- The Dataset Search can help you find available datasets online with keyword searches.
- Kaggle has an Open Data search function that can help you find datasets to practice with.
- Finally, BigQuery hosts 150+ public datasets you can access and use.

Public health datasets

1. Global Health Observatory data: You can search for datasets from this page or explore featured data collections from the World Health Organization.
2. The Cancer Imaging Archive (TCIA) dataset: Just like the earlier dataset, this data is hosted by the Google Cloud Public Datasets and can be uploaded to BigQuery.
3. 1000 Genomes: This is another dataset from the Google Cloud Public resources that can be uploaded to BigQuery.

Public climate datasets

1. National Climatic Data Center: The NCDC Quick Links page has a selection of datasets you can explore.
2. NOAA Public Dataset Gallery: The NOAA Public Dataset Gallery contains a searchable collection of public datasets.

Public social-political datasets

1. UNICEF State of the World's Children: This dataset from UNICEF includes a collection of tables that can be downloaded.
2. CPS Labor Force Statistics: This page contains links to several available datasets that you can explore.
3. The Stanford Open Policing Project: This dataset can be downloaded as a .CSV file for your own use.

Sorting and filtering

Sorting ✧.*
involves arranging data into a meaningful order to make it easier to understand, analyze, and visualize. Data can be sorted in ascending or descending order, and alphabetically or numerically. Sorting can be done across all of a spreadsheet or just in a single column or table. You can also sort by multiple variables.

Freeze ✧.*
This locks the row in place.

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.

Setting up BigQuery, including sandbox and billing options

BigQuerry Account Types:

- Sandbox - account is available at no charge and anyone with a Google account can log in and use it. There are a couple of limitations to this account type. For example, you get a maximum of 12 projects at a time. This means that if you want to make a 13th project, you'll have to delete one of your original 12. It also doesn't allow you to insert new records to a database or update the field values of existing records.
- Free trial - gives you access to more of what BigQuery has to offer with fewer overall limitations.

BigQuery in action

Sometimes a data set is too large to download, or it won't fit in a spreadsheet. So a data analyst will use SQL to create a query to view the specific data that they want from within the larger set.

Clicking can help you preview a data set.

Subset ✧.*
Where we can find details regarding the data.

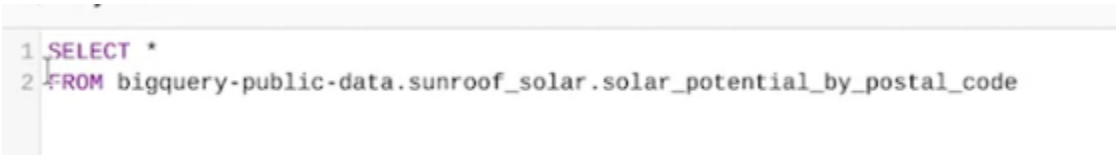
The query will still run without the backticks, (‘ ‘).

The words you see before the dot represent the database name. And the words after the dot represent the table name.

Most queries begin with the word **SELECT.**

Because we want to see the entire data set, we'll put an **asterisk (*)** next. The asterisk says we want to include all columns. This is a great shortcut because without it, we'd have to type in every single field name.

Next we'll press return and type **FROM**. FROM does just what it sounds like. It indicates where the data is coming from. After that, we'll add another space. Now, we paste in the name of the data set that we copied earlier.



WHERE also does exactly what it sounds like. It tells the database where to look for information. In this case, the state name column. So add a space and state underscore name, the name of the column.

Unsaved query Edited

```
SELECT *
FROM bigquery-public-data.sunroof_solar.solar_potential_by_postal_code
WHERE state_name = 'Pennsylvania'
```

Question

In an existing company database, the ***customers*** table contains the following columns: *CustomerId, FirstName, LastName, Company, Address, City, State, Country, PostalCode, Phone, Fax, Email,* and *SupportRepId.*

Create a query to return all the columns in the customer table for only customers in Germany.

```
SELECT *
FROM customers
WHERE Country = 'Germany'
```

Run

Reset

Week IV

Feel confident in your data

Benefits of organizing data ✧.*

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

Let's get organized

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.
- **Foldering**
    - Helps keep project-related files together in one place.
    - Break folder down to subfolders
- **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**
- **Develop metadata practices**
    - Like creating a file that outlines project naming conventions for easy reference.
    - Making copies of data and storing it in different places.

All about file naming

File naming conventions ✧.*
Consistent guidelines that describe the content, date, or version of a file in its name

File naming Do's ✧.*

- Work out your conventions early
- Align file naming with your team
- Make sure file names are meaningful
- Keep file names short and sweet
- Format dates yyyymmdd: SalesReport20201125
- Lead revision numbers with O: SalesReport20201125v02
- Use hyphens, underscores, or capitalized letters: SalesReport_2020_11_25_v02
- Create a test file that lays out all your naming conventions on a project.

Security features in spreadsheet

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

What spreadsheets and excel have in common ✧.*

- Protect spreadsheets from being edited

- Access control features

Balancing security and analytics

Data security ✧.*
means protecting data from unauthorized access or corruption by putting safety measures in place. Usually the purpose of data security is to keep unauthorized users from accessing or viewing sensitive data.

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.

Week V

Why an online presence is important

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 ✧.*
specifically designed to help people make connections with other people in their field.
- Make connections
- Follow industry trends
- Find job opportunities

GitHub ✧.*
part code-sharing site, part social media.

- Share insigh and resource
- Read forums and wikis
- Manage tear projects
- Hosts community events

Tips for enhancing your online presence

- Check the privacy settings on your accounts
- Posts should be family friendly

Networking know-how

Networking ✧.*
Professional relationship building. All about meeting people and making relationships with them.

- Search for public meetups in your area.
- Follow interesting companies or thought leaders on LinkedIn, Twitter, Facebook, and Instagram, interact with them, and share their content.
- Read blogs like O'Reilly, Kaggle, KDnuggets, GitHub and Medium, that can help you connect with peers and experts.

Developing a network

Online connections ✧.*

- **Subscriptions** to newsletters like Data Elixir. Not only will this give you a treasure trove of useful information on a regular basis, but you will also learn the names of data science experts who you can follow, or possibly even connect with if you have good reason to.
- **Hackathons** (competitions) like those sponsored by Kaggle, one of the largest data science and machine learning communities in the world. Participating in a hackathon might not be for everyone. But after joining a community, you typically have access to forums where you can chat and connect with other data analysts.
- **Meetups**, or online meetings that are usually local to your geography. Enter a search for 'data science meetups near me' to see what results you get. There is usually a posted

schedule for upcoming meetings so you can attend virtually to meet other data analysts. Find out more information about meetups happening around the world.

- **Platforms** like LinkedIn and Twitter. Use a search on either platform to find data science or data analysis hashtags to follow. You can also post your own questions or articles to generate responses and build connections that way. At the time of this writing, the LinkedIn #dataanalyst hashtag had 11,842 followers, the #dataanalytics hashtag had 98,412 followers, and the #datascience hashtag had 746,945 followers. Many of the same hashtags work on Twitter and even on Instagram.
- **Webinars** may showcase a panel of speakers and are usually recorded for convenient access and playback. You can see who is on a webinar panel and follow them too. Plus, a lot of webinars are free. One interesting pick is the Tableau on Tableau webinar series. Find out how Tableau has used Tableau in its internal departments.

Offline connections ✧.*

- **Conferences** usually present innovative ideas and topics. The cost of conferences vary, and some are pricey. But lots of conferences offer discounts to students and some conferences like Women in Analytics aim to increase the number of under-represented groups in the field. Leading research and advisory companies such as Gartner also sponsor conferences for data and analytics. The KDNuggets list of meetings and online events for AI, analytics, big data, data science, and machine learning is useful.
- **Associations** or societies gather members to promote a field like data science. The Digital Analytics Association. The KDNuggets list of societies and groups for analytics, data mining, data science, and knowledge discovery is useful.
- **User communities** and **summits** offer events for users of data analysis tools; this is a chance to learn from the best. Have you seen the Tableau community?
- **Non-profit organizations** that promote the ethical use of data science and might offer events for the professional advancement of their members. The Data Science Association is one example.

Developing a network

Mentorship ✧.*

A professional who shares their knowledge, skills, and experience to help you develop and grow. They can be can be trusted

- Advisors
- Sounding boards
- Critics
- Resources

For instance, websites like Score.org and MicroMentor.org and an app called Mentorship allow you to look for specific credentials that match your needs.

Sponsor ✧.*
A professional advocate who's committed to moving a sponsee's career forward within an organization

A mentor helps you skill up
A sponsor helps you move up

Unlike mentors, you don't get to choose the sponsor. The sponsor almost always chooses you. The best course of action is to commit yourself to doing your best work at all times.

Week VI

Why an online presence is important

A professional online presence can ✧.*