**MODULE 2 - DATA RESPONSIBILITY**

**UNBIASED AND OBJECTIVE DATA**

**[INTRODUCTION TO BIAS, CREDIBILITY, PRIVACY, AND ETHICS](https://www.coursera.org/learn/data-preparation/lecture/qALoI/introduction-to-bias-credibility-privacy-and-ethics)**

[**BIAS: FROM QUESTIONS TO CONCLUSIONS**](https://www.coursera.org/learn/data-preparation/lecture/D3P1G/bias-from-questions-to-conclusions)

Our brains are biologically designed to streamline thinking and make quick judgments. Bias has evolved to become a preference in favor of or against a person, group of people, or thing. It can be conscious or subconscious. The good news is 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. It's important to know that bias can also find its way into the world of data. Data bias is a type of error that systematically skews results in a certain direction.

As a data analyst, you have to think about bias and fairness from the moment you start collecting data to the time you present your conclusions. After all, those conclusions can have serious implications.

[**BIASED AND UNBIASED DATA**](https://www.coursera.org/learn/data-preparation/lecture/eqg98/biased-and-unbiased-data)

When data is biased, it can systematically skew results in a certain direction, making them unreliable.

If you'd use a more randomized sample of the population that included all genders, you'd have an unbiased sample. 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.

**[ACCOUNT FOR BIAS](https://www.coursera.org/learn/data-preparation/discussionPrompt/g4GCA/account-for-bias)**

[**UNDERSTAND BIAS IN DATA**](https://www.coursera.org/learn/data-preparation/lecture/NGtoT/understand-bias-in-data)

**Sampling bias is when a sample isn't representative of the population as a whole.** For example, if you're doing research on commuters, and only survey people walking by on the sidewalk, you'll miss out on input from people who ride bicycles, drive, or take the subway. You need all sides of the story to avoid sampling bias.

We'll explore three more types of data bias, **observer bias, interpretation bias, and confirmation bias**, and we'll learn how to avoid them.

**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. You might remember earlier, we learned that scientists use observations a lot in their work, like when they're looking at bacteria under a microscope to gather data. While two scientists looking into the same microscope might see different things, that's observer bias. Another time observer bias might happen is during manual blood pressure readings. Because the pressure meter is so sensitive, health care workers often get pretty different results. Usually, they'll just round up to the nearest whole number to compensate for the margin of error. But if doctors consistently round up, or down the blood pressure readings on their patients, health conditions may be missed, and any studies involving their patients wouldn't have precise, and accurate data.

Another common type of data bias is **interpretation bias**. **The tendency to always interpret ambiguous situations in a positive, or negative way.** Here's an example. Let's say you're having lunch with a colleague, when you get a voicemail from your boss, asking you to call her back. You put the phone down in a huff, certain that she's angry, and you're on the hot seat for something. But when you play the message for your friend, he doesn't hear anger at all, he actually thinks she sounds calm and straightforward. **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**. Your history with your boss made you interpret the call one way, while your friend interpreted it in another way, because they're strangers. Add these interpretations to a data analysis, and you can get bias results.

The last type of bias we'll cover, reminds me of the saying, **people see what they want to see**. That pretty much sums up confirmation bias in a nutshell. **Confirmation bias, is the tendency to search for, or interpret information in a way that confirms preexisting beliefs**. Someone might be so eager to confirm a gut feeling, that they only notice things that support it, ignoring all other signals. This happens all the time in everyday life. We might get our news from a certain website because the writers share our beliefs, or we socialize with people because we know that they hold similar views. After all, conflicting viewpoints might cause us to question our worldview, which can lead us to changing our whole belief system, and let's face it, change is tough. **But you know what's even tougher?** Doing good work when you have bad data, so **it's important to keep bias out of it**.

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. Unfortunately, they're also just a small sample, pun intended, of the types of bias you may encounter in your career as a data analyst. But the good news is, once you know a few, you'll find yourself constantly on guard for bias in any form. It's also important to remember, that no matter what kind of data you use, all of it needs to be inspected for accuracy, and trustworthiness.

**[TEST YOUR KNOWLEDGE ON UNBIASED AND OBJECTIVE DATA](https://www.coursera.org/learn/data-preparation/quiz/eQwG4/test-your-knowledge-on-unbiased-and-objective-data)**

**ACHIEVE DATA CREDIBILITY**

[**IDENTIFY GOOD DATA SOURCES**](https://www.coursera.org/learn/data-preparation/lecture/vEobk/identify-good-data-sources)

Hey, what's good!? No, really, I want to know: What is good? Let me put it this way. If I asked you to name a good song, I might not like it. That's because good is subjective. What I think is good and what you think is good might be different. So **what about good data sources?** Are those subjective, too? In some ways they are, but luckily, there's some **best practices to follow that'll help you measure the reliability of data sets before you use them**.

I think we can all agree that we all want good data. The more high quality data we have, the more confidence we can have in our decisions. Let's learn how we can go about finding and identifying good data sources. First things first, we need to learn how to identify them. A **process** I like to **call ROCCC, R-O-C-C-C**. Okay. I just made that up, but I think acronyms are a really great way to help new information to stick in the brain. Kicking things off is **R for reliable**. Like a good friend, **good data sources are reliable**. With this data you can trust that you're getting accurate, complete and unbiased information that's been vetted and proven fit for use.

**O is for 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.**

Time for the **first C**. **C is for 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.

The next **C is for 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 lis**t. The same goes for data. The best data sources are current and relevant to the task at hand.

The **last C is for cited.** If you've ever told a friend where you heard that a new movie sequel was in the works, you've cited a source. **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?** If you have original data from a reliable organization and it's comprehensive, current, and cited, it ROCCCs!

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**. Now that you know how to spot the good data, which ROCCCs, you're ready to learn about the mountain of bad data and how to avoid it. Let's get moving.

[**WHAT IS "BAD" DATA?**](https://www.coursera.org/learn/data-preparation/lecture/lHirM/what-is-bad-data)

We found that if the data set is **reliable, original, comprehensive, current and cited, it ROCCCs** (or more seriously: **it's** **good**).

Hopefully this is refreshing your memory. Now it's time to pull from what we learned about good data and apply it to today's lesson: **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.

We'll start again with R. **R is for 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. Check out these 2 bar graphs, for example. The one on the left uses a y-axis starting point of 3.14%. And the one on the right uses 0. This makes it seem like interest rates have skyrocketed over a four year period when they've actually remained pretty flat.

Okay, onto O. **O is for 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.

Now, **C is for 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.

The **next C is for 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. For example, you can always trust Data.gov, which is home to the U.S. government's open data.

The **last C is for not cited.** If your source hasn't been cited or vetted, it's a no-go. So to sum up, good data should be original data from a reliable organization, comprehensive, current, and cited. It should ROCCC! Otherwise, it's bad data.

**If you need a great reliable data source**, check out the U.S. Census Bureau, which regularly updates their information. It's important for data analysts to understand and keep an eye out for bad data because it can have serious and lasting impacts. Whether it's an incorrect conclusion leading to one bad business decision, or inaccurate information causing processes to fail and putting populations at risk, 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**. And with that, we've come to the end of our adventure with bias and credibility. After a few more exercises, you'll be ready for what lies ahead. I look forward to your progress.

[**TEST YOUR KNOWLEDGE ON DATA CREDIBILITY**](https://www.coursera.org/learn/data-preparation/quiz/5ZRVN/test-your-knowledge-on-data-credibility)

**DATA ETHICS AND PRIVACY**

[**ESSENTIAL DATA ETHICS**](https://www.coursera.org/learn/data-preparation/lecture/AWqwt/essential-data-ethics)

What comes to your mind when you think of the word, ethics? For me, it's a set of principles to live by. Most people have a personal code of ethics that helps them navigate the world. When we're young, it could be as simple as never lie, cheat or steal, but as we get older, it's a much broader list of do's and don'ts. Our personal ethics evolve and become more rational, giving us a moral compass to use as we face life's questions, challenges, and opportunities. **When we analyze data, we're also faced with questions, challenges, and opportunities, but we have to rely on more than just our personal code of ethics to address them.** As we learned earlier, we all have our own personal biases, not to mention subconscious biases that make ethics even more difficult to navigate. That's why we have data ethics, an important aspect of analytics that we'll explore right here. But first, let's go back to the **general idea of ethics**. While an exact definition is still under discussion in philosophy, one practical view is that 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. **Just like humans, data has standards to live up to as well**.

**Data ethics** **refers to well- founded standards of right and wrong that dictate how data is collected, shared, and used**. Since the ability to collect, share and use data in such large quantities is relatively new, the rules that regulate and govern the process are still evolving. The importance of data privacy has been recognized by governments worldwide and they started creating data protection legislation to help protect people and their data. The GDPR of the European Union was created to do just this. While policy makers continue their work, companies like Google have a responsibility to lead the effort and will do so in the same spirit we always have by offering products that make privacy a reality for everyone. **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. There are lots of different aspects of data ethics but we'll cover six: ownership, transaction transparency, consent, currency, privacy, and openness. We'll explore data privacy and openness a bit later.

First up is **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**.

Next, we have **transaction transparency**, which is 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.

Now let's talk about **another aspect of data ethics**, **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.

Next, there's **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. The **last two aspects of data ethics**, **privacy and openness**, deserve their own spotlight on this data stage. Coming up, you'll see why.

[**OPTIONAL REFRESHER: ALEX AND THE IMPORTANCE OF DATA ETHICS**](https://www.coursera.org/learn/data-preparation/lecture/BVQjE/optional-refresher-alex-and-the-importance-of-data-ethics)

Hi, I'm Alex. I'm a research scientist at Google. My team is called the Ethical AI Team. We're a group of folks that really are concerned not only about how AI, the technology operates, but how it interacts with society and how it might help or harm marginalized communities. So when we talk about **data ethics,** we think about, **What is the good and right way of using data? What are going to be ways that are going to be uses of data that are going to be beneficial to people?** When it comes to data ethics it's not just about minimizing harm, but it's actually this concept of beneficence. **How do we actually improve the lives of people by using data?** **When we think about data ethics we're thinking about who's collecting the data? Why are they collecting it? How are they collecting it? And for what purpose?** Because of the way that organizations have imperatives to make money, or to report to somebody, or provide some analysis, we also have to keep strongly in mind how this is actually going to benefit people at the end of the day. **Are the people represented in this data going to be benefited by this?** I think that's the thing you never want to lose sight of as a data scientist or a data analyst. I think aspiring data analysts need to keep in mind that a lot of the data that you're going to encounter is data that comes from people. **So at the end of the day, data are people**. And you want to have a responsibility to those people that are represented in those data.

Second, is **thinking about how to keep aspects of their data protected and private**. We don't want to go through our practice thinking about data instances as something we can just throw on the web. No, there needs to be considerations about how to keep that information and likenesses, like their images, or their voices or their text. How do we keep that private? We also need to think about how we can have mechanisms of giving users and giving consumers more control over their data. It's not going to be sufficient just to say, we collected all this data and trust us with all this data, but **we need to ensure that there's actionable ways in which people can consent to giving those data and ways that they can ask for it to be revoked or removed**.

Data's growing, and at the same time, **we need to empower people to have control over their own data**. The future is that data is always growing. We haven't seen any kind of evidence that data is actually shrinking. And with the knowledge that data's growing, these issues become more and more piqued and more and more important to think about.

[**PRIORITIZE DATA PRIVACY**](https://www.coursera.org/learn/data-preparation/lecture/OPSzY/prioritize-data-privacy)

We've been exploring some **important aspects of data ethics**, and **one of the most personal areas** **involves privacy. Privacy is personal.** We may all define privacy in our own way, and we're all entitled to it. Whether it's family members wanting privacy when using a shared computer, a teenager wanting to share a selfie with only specific people, or a company wanting to keep their customers' credit card info secure, we're all concerned how our data is used and shared. Data privacy is big in today's culture, so let's explore it fully.

When talking about data, **privacy means preserving a data subject's information and activity any time a data transaction occurs**. This is **sometimes called 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. Data privacy is important, even if you're not someone who thinks about it on a day-to-day basis. The importance of data privacy has been recognized by governments worldwide, and they've started creating data protection legislation to help protect people and their data. **Being able to trust companies with your data is important.** **It's what makes people want to use a company's product, share their information, and more.** Trust is a really big responsibility that can't be taken lightly. The final aspect involving data ethics is one that's constantly being discussed. The idea of openness, free access, usage, and sharing of data.

[**DATA ANONYMIZATION**](https://www.coursera.org/learn/data-preparation/supplement/rtTel/data-anonymization)

## **WHAT IS DATA ANONYMIZATION?**

You have been learning about the importance of privacy in data analytics. Now, it is time to talk about data anonymization and **what types of data should be anonymized**. **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.**

### 

### Your role in data anonymization. Organizations have a responsibility to protect their data and the personal information that data might contain. **As a data analyst,** you might be expected to **understand** what data **needs to be anonymized**, but **you generally wouldn't be responsible for the data anonymization itself**. A **rare exception** might be if you **work with a copy of the data for testing or development purposes**. In this case, you could be required to anonymize the data before you work with it.

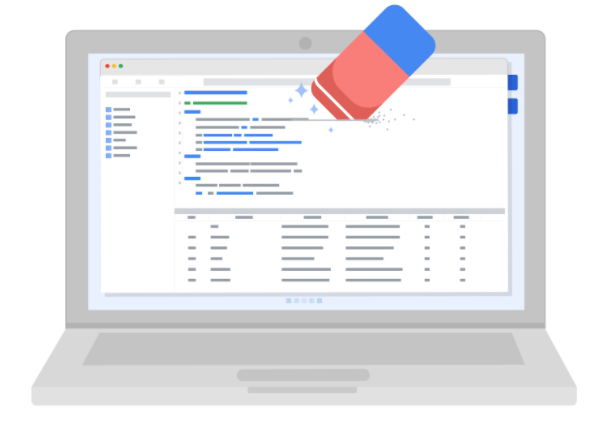
## 

## 

## 

## **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, which is 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

For some people, it just makes sense that this type of data should be anonymized. For others, we have to be very specific about what needs to be anonymized. Imagine a world where we all had access to each other’s addresses, account numbers, and other identifiable information. That would invade a lot of people’s privacy and make the world less safe. **Data anonymization is one of the ways we can keep data private and secure!**

[**ANDREW: THE ETHICAL USE OF DATA**](https://www.coursera.org/learn/data-preparation/lecture/a38HL/andrew-the-ethical-use-of-data)

My name is Andrew. I am a senior developer advocate on the ethical AI research group at Google. As a senior developer advocate, I try and help the larger community build socially responsible AI systems. **One consequence of not using this technology responsibly is the possibility of amplifying or reinforcing unfair biases**. Now these algorithms, these data sets, are often being used in settings where they are deciding the outcome. Whether it's curating content for an individual or determining whether or not they're eligible for a loan, all these different decision making processes depend on the algorithms and the data sets that are being used in that context. And so if this were to be handled irresponsibly, then the very outcomes of these systems could potentially harm underrepresented communities, minority groups. There's a lot that the field, the industry, the community, is learning about the responsible use of data and AI. So what I try to do is I try to correlate all those different elements, whether it's working with various research groups within Google, working with various product teams at Google, engaging with the larger community. We have to go above and beyond and actually educate those that are striving to build this technology for good but may not necessarily have the resources or the institutional community wisdom to actually carry out their good intentions. So the truth of the matter is that AI, data, and any technology that's built around that, there's a lot of great benefits to that. It's improving the lives of many people out there. It's enabling us to do things we couldn't ordinarily do. It's providing us with affordances to think about other things in life. This is all the more reason why it's important that we together, collectively, not just one organization, but the entire community, and even non-technologists, too, everyone needs to be involved. That's the role that I play here, is that I try to help AI evolve ethically together, and to do that is contingent on the democratization of the responsible use of AI.

[**TEST YOUR KNOWLEDGE ON DATA ETHICS AND PRIVACY**](https://www.coursera.org/learn/data-preparation/quiz/UCgeu/test-your-knowledge-on-data-ethics-and-privacy)

**UNDERSTAND OPEN DATA**

[**FEATURES OF OPEN DATA**](https://www.coursera.org/learn/data-preparation/lecture/4TTkQ/features-of-open-data)

There's just something so liberating about being able to find information on any subject at all on the Internet. Can't remember the 3rd line of your favorite childhood song, curious who had the most home runs in 1986, want to teach yourself sign language? Just pop open your laptop, type up some text and poof, you have what you need. Many groups think we should also have this level of access to data. There's even a global movement that believes the openness of data can transform society and how decisions are made. So far, we've talked a lot about the **power of data** and the importance of **data ethics concerns** including **ownership, transaction transparency, consent, currency, and privacy**.

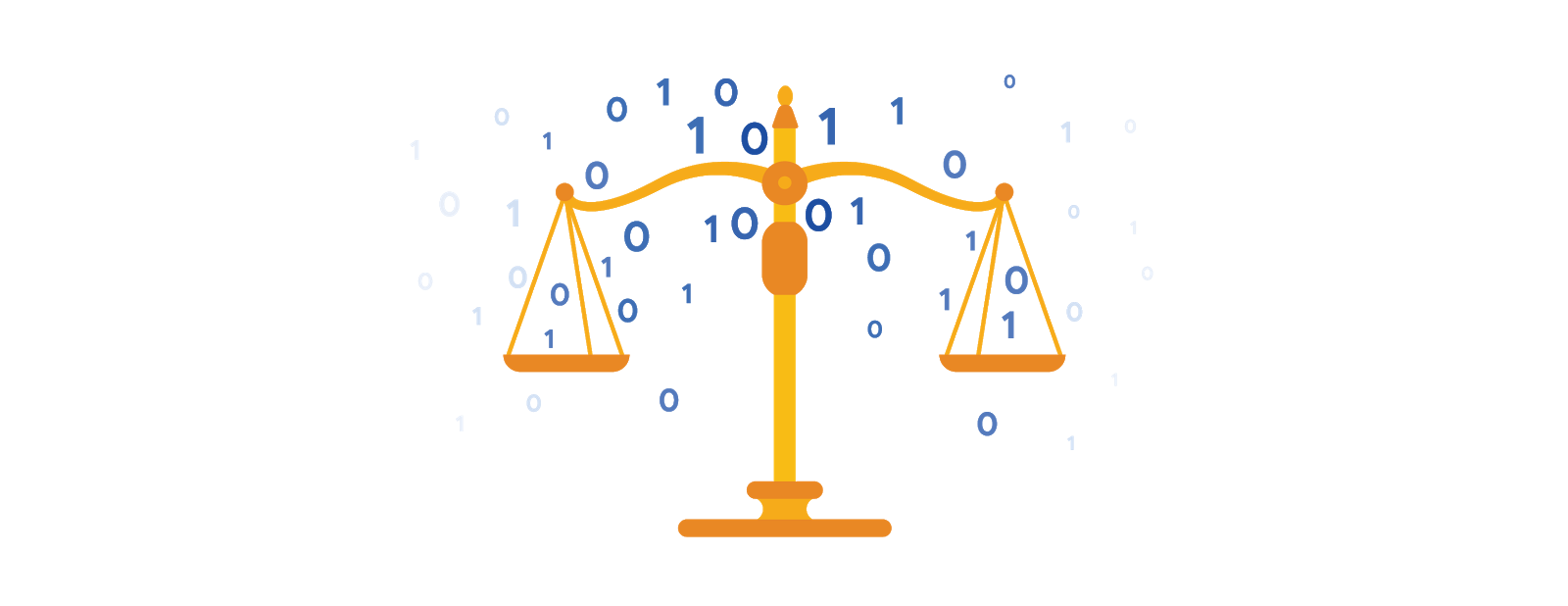
Now, let's talk about **openness**. When referring to data, **openness refers to free access, usage and sharing of data**. Sometimes we refer to this as **open data**, but it doesn't mean we ignore the other aspects of data ethics we covered. We should still be transparent, respect privacy, and make 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. For example, 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 form. The website data.gov is a great example. You can download science and research data for a wide range of industries in simple file formats like a spreadsheet. Another standard surrounds **reuse and redistribution**. **Open data must be provided under terms that allow reuse and redistribution including the ability to use it with other datasets**. And the **last area is 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.

Now let's talk a little more about **why open data is such a great thing and how it can help you as a data analyst.** One of the **biggest benefits of open data is that credible databases can be used more widely.** More importantly, all of that good data can be leveraged, shared, and combined with other data. Just imagine the impact that would have on scientific collaboration, research advances, analytical capacity, and decision-making. For example, 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. The possibilities and benefits are almost endless. But of course, every big idea has its challenges. A whole lot of resources are needed to make the technological shift to open data. **Interoperability is key to open data's success**. **Interoperability is the ability of data systems and services to openly connect and share data.** For example, data interoperability is important for health care information systems where multiple organizations such as hospitals, clinics, pharmacies, and laboratories need to access and share data to ensure patients get the care that they need. This is why your doctor is able to send your prescription directly to your pharmacy to fill. They have compatible databases that allow them to share information. But this kind of interoperability requires a lot of cooperation. While there is serious potential in the open, timely, fair, and simple sharing of data, its future will depend on how effectively larger challenges are addressed. As a data analyst, I say the sooner the better.

Now that you've learned all about data ethics, you have some important principles to guide you on your data journey. Anytime you're not sure of your data, remember what you've learned here.

[**THE OPEN DATA DEBATE**](https://www.coursera.org/learn/data-preparation/supplement/dj6K0/the-open-data-debate)

Just like data privacy, open data is a widely debated topic in today’s world. Data analysts think a lot about open data, and as a future data analyst, you need to understand the basics to be successful in your new role.



## **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.

[**ANDREW: STEPS FOR ETHICAL DATA USE**](https://www.coursera.org/learn/data-preparation/lecture/Bcpoq/andrew-steps-for-ethical-data-use)

My name is Andrew. I am a Senior Developer Advocate on the ethical AI research group at Google. As an analyst, there's quite a few 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.

One of it is 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.

And then also, 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?

And going beyond that, also 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.

But as a data analyst, you stand in the intersection between the very people that will stand to benefit from the technology that's being developed and those in your organization that are trying to make a more informed decision as to whether or not to move forward with the productionization of the technology. It may feel like there's a lot of weight there, and there is, but it's also very pivotal, and it speaks to the volume of the impact of your work.

[**RESOURCES FOR OPEN DATA**](https://www.coursera.org/learn/data-preparation/supplement/3hAmz/resources-for-open-data)

Luckily for data analysts, there are lots of trustworthy 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**](https://www.data.gov/): 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**](https://www.census.gov/data.html): 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**](https://www.opendatanetwork.com/): 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**](https://cloud.google.com/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**](https://datasetsearch.research.google.com/)**:** The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

[**HANDS-ON ACTIVITY: KAGGLE DATASETS**](https://www.coursera.org/learn/data-preparation/quiz/HXUbK/hands-on-activity-kaggle-datasets)

**Activity overview**

**Empty alt text.**

In the last activity, you got set up on Kaggle and explored the Notebooks feature. In this activity, we will work with a different feature of the Kaggle platform: datasets.

Kaggle has tens of thousands of datasets that are available for public use. Anyone can upload a dataset to Kaggle. If they choose to make it public, other Kagglers can use that dataset to create their own projects.

First, you’ll take a tour of a specific dataset. Then, you’ll have a chance to choose your own datasets to work with. Finally, you’ll use what you’ve learned in this module to determine the kind of data in your datasets, and whether the data is biased or unbiased.

By the time you complete this activity, you will be able to use many of the helpful features Kaggle has to offer. This will enable you to find data for projects and engage with the data community, which is important for developing skills and networking in your career as a data analyst.

## 

## 

## **Explore Kaggle datasets**

**Empty alt text.**

Let’s explore the datasets feature!

### 

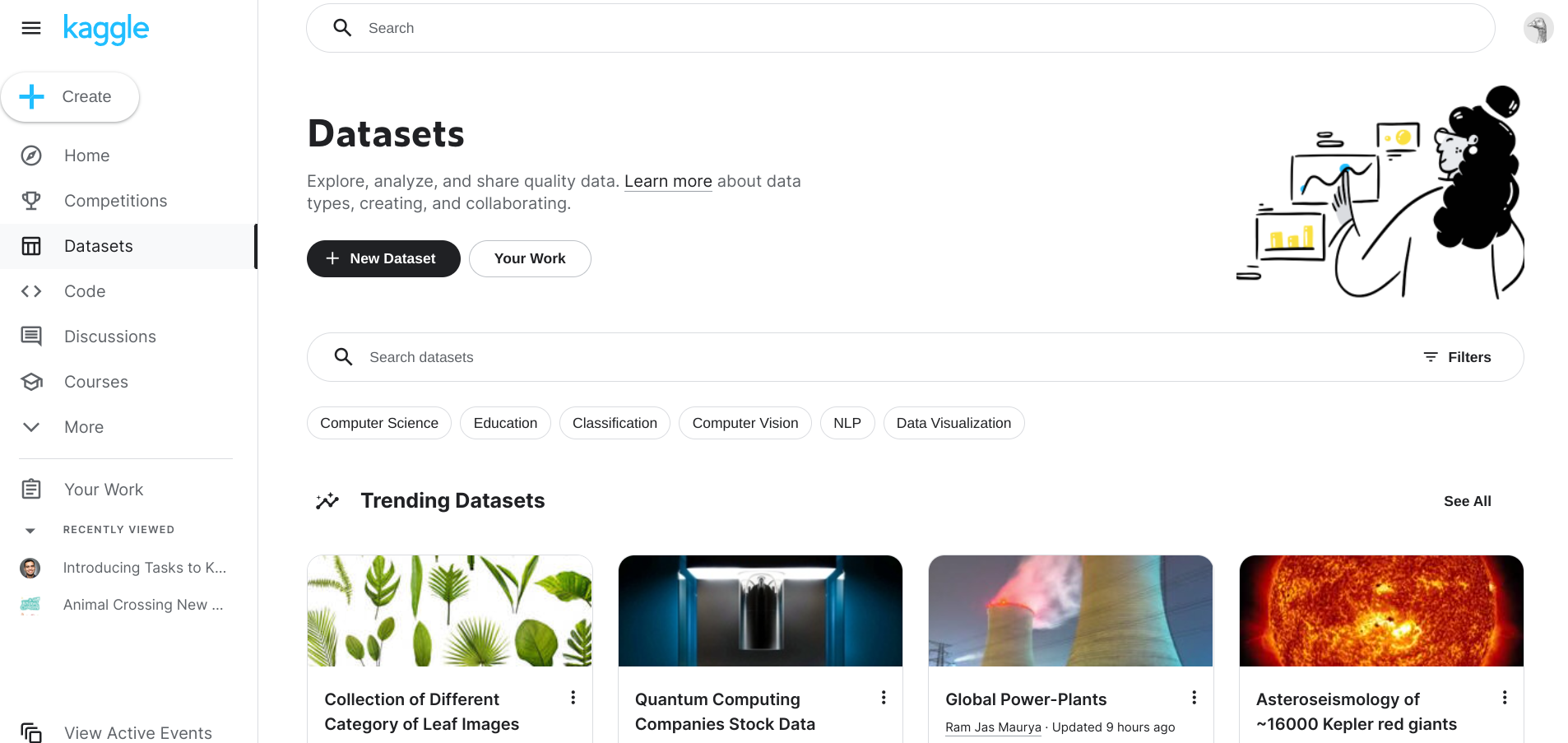
### **Find a dataset**

****

1. To start, log in to your Kaggle account.

* Note: Kaggle frequently updates its user interface. The latest changes may not be reflected in the screenshots, but the principles in this activity remain the same. Adapting to changes in software updates is an essential skill for data analysts, and we encourage you to practice troubleshooting. You can also reach out to your community of learners on the discussion forum for help.

2. Then, click on the Datasets icon in the vertical Navigation bar on the left. This takes you to the Datasets home page. From here, you can create a new dataset or search for datasets created by other Kagglers.



3. Now, check out a specific dataset. Enter Animal Crossing in the search bar to find datasets related to the Nintendo video game *Animal Crossing*.

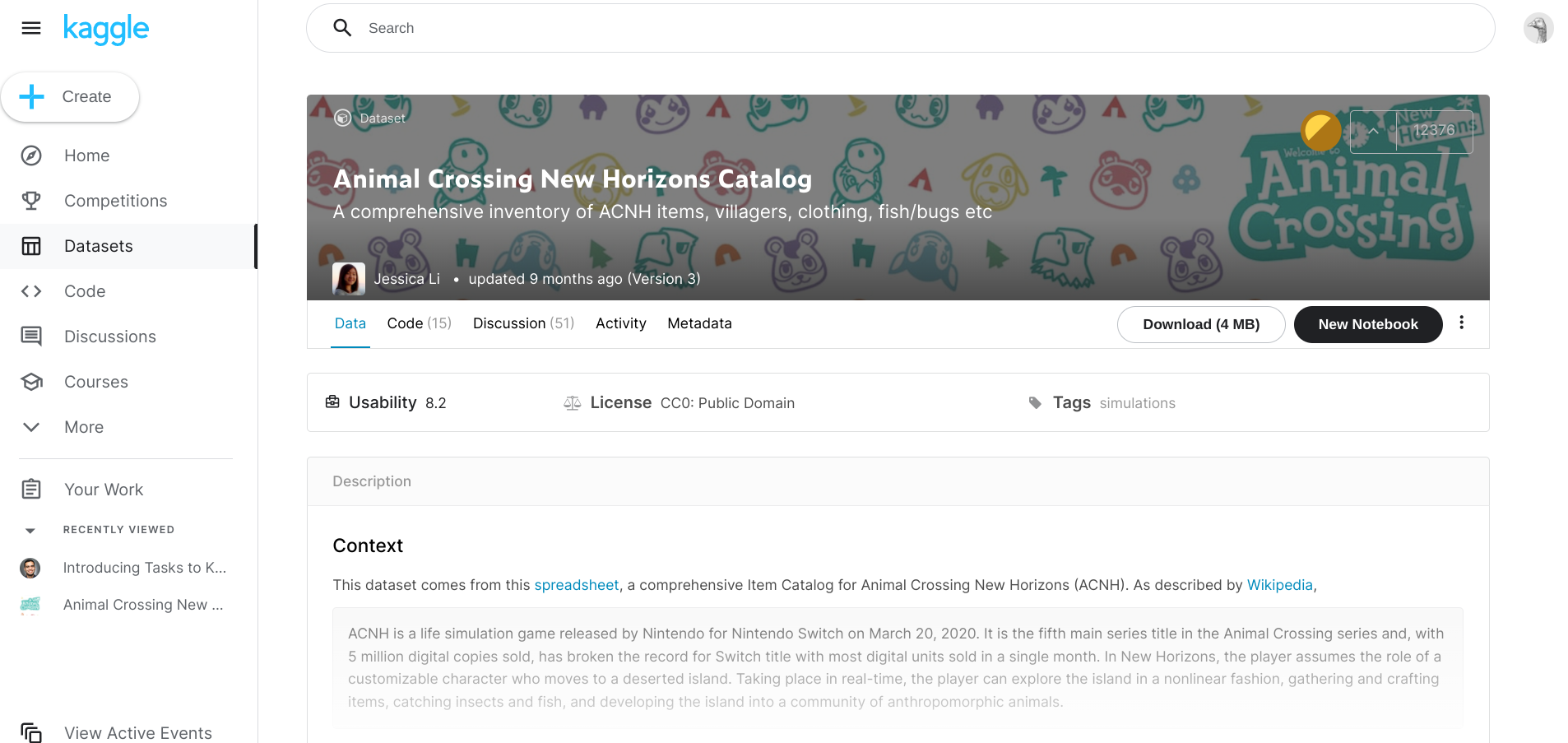
4. There’s more than one option, so click on the Animal Crossing New Horizons Catalog. This takes you to the landing page for this dataset.

### 

### 

### **Tour a dataset landing page**

****

****

Header: The header at the top of the page contains the following information about the dataset:

* Its title
* A brief description of its contents
* The name of its creator
* When it was last updated
* Its current version

Badge: In the top-right corner of the header, you’ll find three more items:

* A badge in the shape of a circle
* An icon in the shape of a caret symbol ( ^ )
* A number

The badge is related to Kaggle’s progression system. If you want, you can read more about it [here](https://www.kaggle.com/progression/).

Upvotes: Clicking on the caret lets you “upvote” the dataset. The number shows the number of times this dataset has been upvoted by the Kaggle community.

Tabs: Beneath the header is a bar with three tabs: Data Card, Code, and Discussion. Take a moment to click on each of these tabs and explore their contents. Afterwards, navigate back to the Datasets tab!

Now, you can move down the page. You’ll find a box that contains three terms: Usability, License, and Tags.

Usability shows how complete the dataset *webpage* is (and not the dataset itself). Kaggle encourages the community to add information to the dataset webpage to make the dataset itself easier to understand. For example, a brief description or a column header. Hover your cursor over the Usability score to discover what the dataset page contains.

Licenses govern how a dataset can be used. Click on the license name to learn more about that specific license.

Expected update frequency will report when the data set was updated with new attributes, amendments, or expansions of the data. You may also see when any expected future dates are scheduled for a data set update. In this case, there will never be an update scheduled.

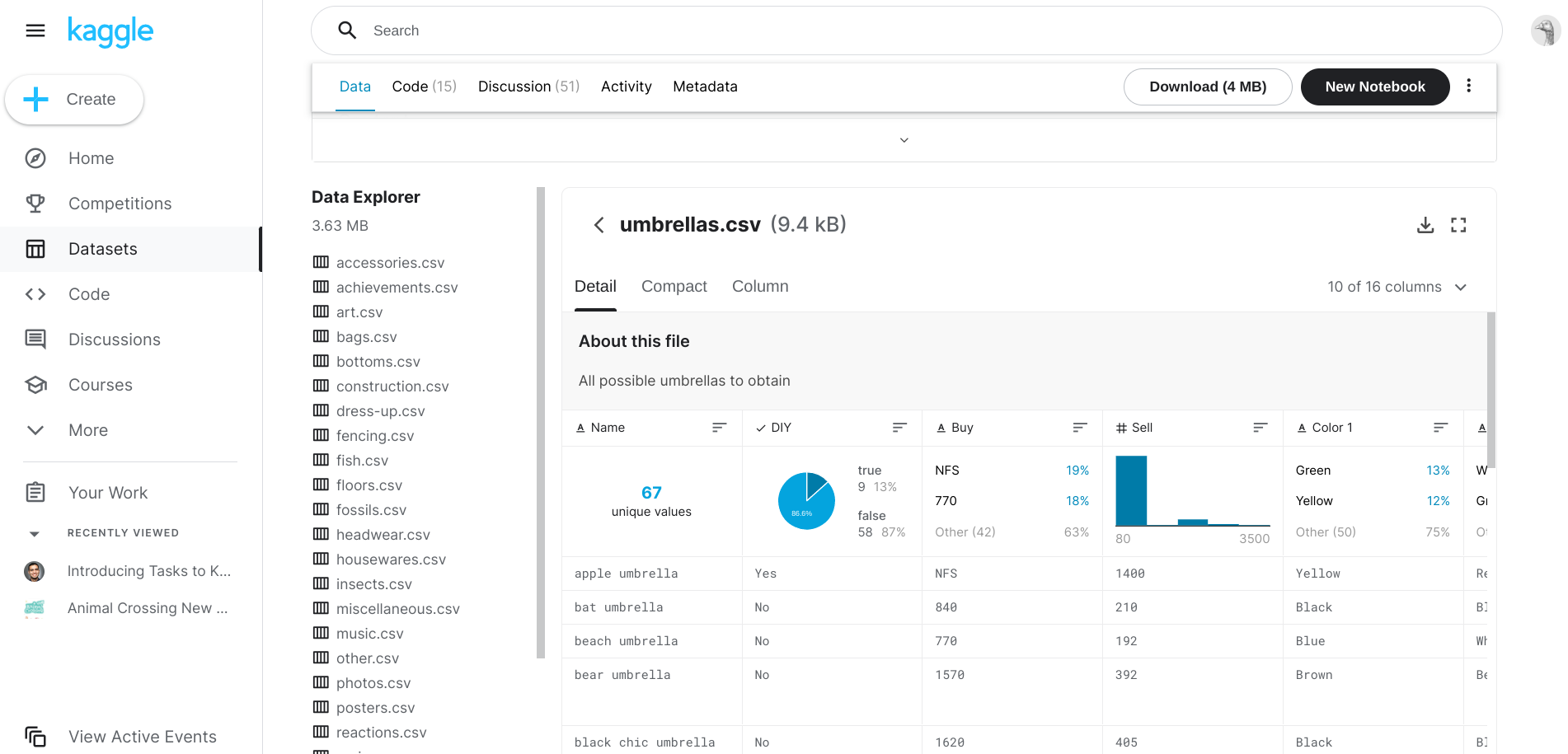
The next box down contains a detailed description of the dataset. Kagglers often include information on where the dataset came from and how the dataset was prepared.

And last—but certainly not least—is the Data Explorer!

## **Use the data explorer**

**Empty alt text.**

The Data Explorer menu shows that the *Animal Crossing* dataset contains 30 .csv files. If you click on a file name, the window to the right will display information from that specific file. Try clicking on umbrellas.csv to check it out!



### 

### 

### 

### 

### **Tour the dataset explorer**

****

Notice that the Data Explorer has three viewing options: Detail, Compact, and Column. For now, we’ll focus on the Detail tab.

The description at the top of the Detail tab shows that the umbrellas.csv file contains data on all the umbrellas in the video game. Let’s check out the columns. Each column header has three items:

* A small icon on the left that shows the data type
* The name of the column
* An icon with three bars that lets you sort the data if you click on it

Below each column header is a box that contains a summary of the data. This lets you quickly get an idea of what’s in the dataset. For example, the summary for the Name column shows there are 67 unique values for the umbrella names. The summary for the DIY column shows that 9 of the umbrella recipes are DIY, or “do it yourself.” Take a moment to explore the summaries for the other columns.

And that completes our tour! That’s a lot of information. Feel free to go back and review.

## **Access a dataset**

**Empty alt text.**

After you’ve explored a dataset, you can link it to a Kaggle notebook or download it to access it for your own use. Linking a dataset to a Kaggle notebook means you create a new notebook from the existing dataset so that it is available for you to use.

### **Find your own datasets**

****

Now, you’ll get a chance to choose your own datasets to work with! Use the following steps to find datasets that interest you:

1. When you’re ready, click on the Data icon on the left to return to the Datasets landing page.

2. Note that datasets can exist in a variety of formats. If you want to make sure your dataset is in a .csv format, click on the Filter button on the right side of the Datasets search bar. Then, choose CSV from the menu.

3. Find 2-3 datasets that you’re interested in exploring further.

4. Create notebooks from them, download them, or check them out in the Data Explorer. Keep these datasets in mind for your upcoming reflection.

### 

### **Link or download a dataset**

****

Here are the options to create a notebook or download the dataset:

* Create a Kaggle notebook: To link a dataset to a Kaggle notebook, you click on the New Notebook button in the dataset header. This will create a notebook in your Kaggle account that links to the dataset.
* Download the dataset: To download a copy of the dataset to your computer, click on the Download button in the dataset header at the top of the page.
* Open the file in Google Sheets: To open a Google Sheets view of the file, click on the download icon at the top-right of the Data Explorer. You can then download the file.

[**TEST YOUR KNOWLEDGE ON OPEN DATA**](https://www.coursera.org/learn/data-preparation/quiz/VbUWF/test-your-knowledge-on-open-data)

**MODULE 2 CHALLENGE**

[**GLOSSARY TERMS FROM MODULE 2**](https://www.coursera.org/learn/data-preparation/supplement/HHhBb/glossary-terms-from-module-2)

**Terms and definitions for Course 3, Module 2**

**Bad data source:** A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Open data:** Data that is available to the public

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

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

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

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

**[MODULE 2 CHALLENGE](https://www.coursera.org/learn/data-preparation/exam/jZOOG/module-2-challenge)**