**C4 - PROCESS DATA FROM DIRTY TO CLEAN**

**DATA INTEGRITY AND ANALYTICS OBJECTIVES**

**Data integrity is critical to successful analysis. In this part of the course, you’ll explore methods and steps that analysts take to check their data for integrity. This includes knowing what to do when you don’t have enough data. You’ll also learn about random samples and understand how to avoid sampling bias. All of these methods will also help you ensure your analysis is successful.**

### 

### Learning Objectives

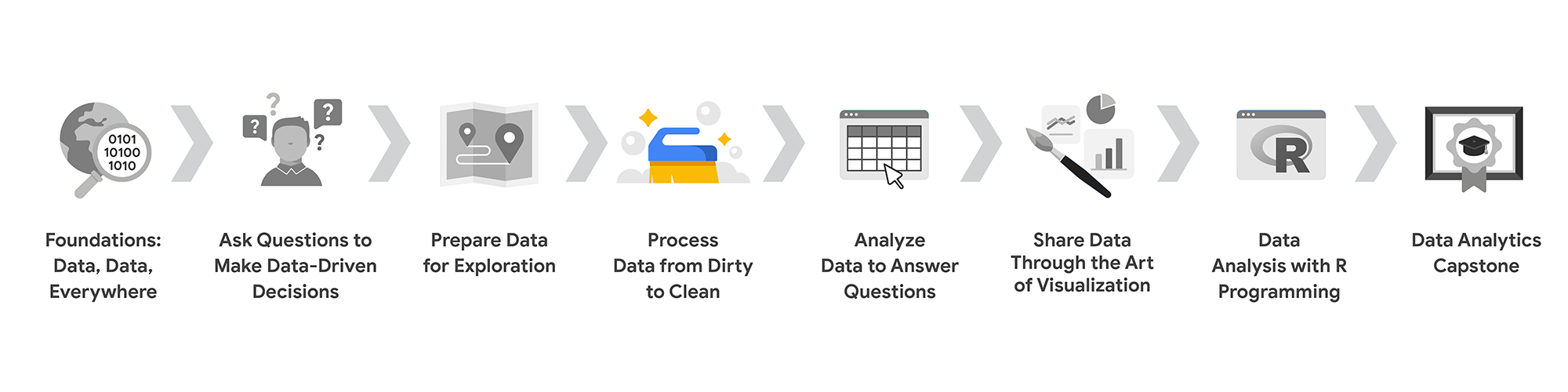
* **Describe statistical measures associated with data integrity including statistical power, hypothesis testing, and margin of error**
* **Describe strategies that can be used to address insufficient data**
* **Discuss the importance of sample size with reference to sample bias and random samples**
* **Describe the relationship between data and related business objectives**
* **Define data integrity with reference to types and risks**
* **Discuss the importance of pre-cleaning activities**

# 

# Course 4 overview: Set your expectations

**Welcome to the fourth course in the program! As you make your way through the certificate, this course and others that follow will begin to focus more on practical, skills-based assignments and projects.**

**In this course, you’ll learn to clean data by checking it for completeness and correctness. You’ll review a variety of approaches to clean data in spreadsheets and databases. Then, you’ll gain essential troubleshooting skills that will enable you to fix any errors. An important step in cleaning data is creating reports to communicate the changes you’ve made to others. You’ll understand how to do that in order to ensure the accuracy and reliability of data. Together, these skills will help ensure your data analysis is successful.**

****

## 

## 

## **Course Overview**

1. [**Foundations: Data, Data, Everywhere**](https://www.coursera.org/learn/foundations-data/home/welcome)
2. [**Ask Questions to Make Data-Driven Decisions**](https://www.coursera.org/learn/ask-questions-make-decisions/home/welcome)
3. [**Prepare Data for Exploration**](https://www.coursera.org/learn/data-preparation/home/welcome)
4. **Process Data from Dirty to Clean (this course)**
5. [**Analyze Data to Answer Questions**](https://www.coursera.org/learn/analyze-data/home/welcome)
6. [**Share Data Through the Art of Visualization**](https://www.coursera.org/learn/visualize-data/home/welcome)
7. [**Data Analysis with R Programming**](https://coursera.org/learn/data-analysis-r/home/welcome)
8. [**Google Data Analytics Capstone: Complete a Case Study**](https://coursera.org/learn/google-data-analytics-capstone/home/welcome)

## **Course 4 content**

**Each course is broken into modules. Here’s a quick overview of the skills you’ll gain in each of the five Course 4 modules.**

### **Module 1: The importance of integrity**

**Data integrity is critical to successful analysis. In this part of the course, you’ll explore methods and steps that analysts take to check their data for integrity. This includes knowing what to do when you don’t have enough data. You’ll also learn about random samples and understand how to avoid sampling bias. All of these methods will also help you ensure your analysis is successful.**

### **Module 2: Clean data for more accurate insights**

**Every data analyst wants to analyze clean data. In this part of the course, you’ll learn the difference between clean and dirty data. Then, you’ll practice cleaning data in spreadsheets and other tools.**

### **Module 3: Data cleaning with SQL**

**Knowing a variety of ways to clean data can make a data analyst’s job much easier. In this part of the course, you’ll use SQL to clean data from databases. In particular, you’ll explore how SQL queries and functions can be used to clean and transform your data before an analysis.**

### **Module 4: Verify and report cleaning results**

**When you clean data, you make changes to the original dataset. It’s important to verify the changes you make are accurate and to let your teammates know about the changes. In this part of the course, you’ll learn to verify that data is clean and report your data cleaning results. With verified clean data, you’re ready to begin analyzing!**

### **Module 5: Optional: Add data to your resume**

**Creating an effective resume will help you in your data analytics career. In this part of the course, you’ll learn all about the job application process. Your focus will be on building a resume that highlights your strengths and relevant experience.**

### **Module 6: Course wrap-up**

**Review the course glossary and prepare for the next course in the Google Data Analytics Certificate program.**

## **What to expect**

**Each module includes a series of lessons with many types of learning opportunities. These include:**

* **Videos for instructors to teach new concepts and demonstrate the use of tools**
* **In-video questions that pop up from time to time to help you to check your understanding of key concepts and skills**
* **Step-by-step guides you can use to follow along with instructors as they demonstrate tools**
* **Readings to explore topics more in-depth and build on the concepts from the videos**
* **Discussion forums to share, explore, and reinforce lesson topics**
* **Discussion prompts to promote thinking and engagement in the discussion forums**
* **Practice quizzes to prepare you for graded quizzes**
* **Graded quizzes to measure your progress and give you valuable feedback**

**This program was designed to let you work at your own pace—your personalized deadlines are just a guide. There is no penalty for late assignments. To earn your certificate, you simply need to complete all of the work.**

**If you miss two assessment deadlines in a row, or if you miss an assessment deadline by two weeks, you'll see a Reset deadlines option on the Grades page. Click it to switch to a new course schedule with updated deadlines. You can use this option as many times as you need—it won’t remove any progress you’ve already made in the course, but you may find new course content if the instructor updated the course after you started. If you cancel a subscription and then reactivate it, your deadlines will automatically reset.**

**In this course, you'll be assessed with graded quizzes and activities. Both are based on the wide variety of learning materials and activities that reinforce the important skills you’ll develop. And both can be taken more than once.**

## **Tips for success**

* **It is strongly recommended that you go through the items in each lesson in the order they appear because new information and concepts build on previous knowledge.**
* **Participate in all learning opportunities to gain as much knowledge and experience as possible.**
* **If something is confusing, don’t hesitate to replay a video, review a reading, or repeat a self-review activity.**
* **Use the additional resources that are referenced in this course. They are designed to support your learning. You can find all of these resources in the** [**Resources**](https://www.coursera.org/learn/ask-questions-make-decisions/resources/xW7lI) **tab.**
* **When you encounter useful links in this course, bookmark them so you can refer to the information later for study or review.**
* **Understand and follow the** [**Coursera Code of Conduct**](https://www.coursera.support/s/article/208280036-Coursera-Code-of-Conduct?) **to ensure that the learning community remains a welcoming, friendly, and supportive place for all members.**

**Updates to the course**

**As you complete this course, you may notice updates to the content, like new practice materials and additional examples. These updates ensure the program provides up-to-date skills and guidance that will help you in your data analytics career. If you previously completed a graded activity, you *may* need to repeat the assessment in order to complete this course. For more information, check out** [**the course discussion forum.**](https://www.coursera.org/learn/process-data/discussions)

[**WHY DATA INTEGRITY IS IMPORTANT**](https://www.coursera.org/learn/process-data/lecture/rvfKH/why-data-integrity-is-important)

We're going to discuss data integrity and some risks you might run into as a data analyst.

A strong analysis depends on the integrity of the data. If the data you're using is compromised in any way, your analysis won't be as strong as it should be.

**Data integrity is the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle**.

That might sound like a lot of qualities for the data to live up to. But trust me, it's worth it to check for them all before proceeding with your analysis. Otherwise, your analysis could be wrong. Not because you did something wrong, but because the data you were working with was wrong to begin with. When data integrity is low, it can cause anything from the loss of a single pixel in an image to an incorrect medical decision. In some cases, one missing piece can make all of your data useless.

Data integrity can be compromised in lots of different ways.

There's a chance data can be compromised every time it's replicated, transferred, or manipulated in any way.

**Data replication** is the process of storing data in multiple locations. If you're replicating data at different times in different places, there's a chance your data will be out of sync. This data lacks integrity because different people might not be using the same data for their findings, which can cause inconsistencies.

There's also the issue of **data transfer**, which is the process of copying data from a storage device to memory, or from one computer to another. If your data transfer is interrupted, you might end up with an incomplete data set, which might not be useful for your needs.

The **data manipulation process** involves changing the data to make it more organized and easier to read. Data manipulation is meant to make the data analysis process more efficient, but an error during the process can compromise the efficiency.

Finally, **data can also be compromised through** human error, viruses, malware, hacking, and system failures, which can all lead to even more headaches.

I'll stop there. That's enough potentially bad news to digest. Let's move on to some potentially good news. In a lot of companies, the data warehouse or data engineering team takes care of ensuring data integrity. Coming up, we'll learn about checking data integrity as a data analyst. But rest assured, someone else will usually have your back too. After you've found out what data you're working with, it's important to double-check that your data is complete and valid before analysis. This will help ensure that your analysis and eventual conclusions are accurate.

**Checking data integrity is a vital step in processing your data to get it ready for analysis, whether you or someone else at your company is doing it.**

[**MORE ABOUT DATA INTEGRITY AND COMPLIANCE**](https://www.coursera.org/learn/process-data/supplement/uWR9E/more-about-data-integrity-and-compliance)

This reading illustrates the importance of data integrity using an example of a global company’s data. Definitions of terms that are relevant to data integrity will be provided at the end.

## 

## **Scenario: calendar dates for a global company**

Calendar dates are represented in a lot of different short forms. Depending on where you live, a different format might be used.

* In some countries, **12/10/20** (DD/MM/YY) stands for October 12, 2020.
* In other countries, the national standard is YYYY-MM-DD so October 12, 2020 becomes **2020-10-12**.
* In the United States, (MM/DD/YY) is the accepted format so October 12, 2020 is going to be **10/12/20**.

Now, think about what would happen if you were working as a data analyst for a global company and didn’t check date formats. Well, your data integrity would probably be questionable. Any analysis of the data would be inaccurate. Imagine ordering extra inventory for December when it was actually needed in October!

A good analysis depends on the integrity of the data, and data integrity usually depends on using a common format. So it is important to double-check how dates are formatted to make sure what you think is December 10, 2020 isn’t really October 12, 2020, and vice versa.

Here are some other things to watch out for:

* **Data replication compromising data integrity:** Continuing with the example, imagine you ask your international counterparts to verify dates and stick to one format. One analyst copies a large dataset to check the dates. But because of memory issues, only part of the dataset is actually copied. The analyst would be verifying and standardizing incomplete data. That partial dataset would be certified as compliant but the full dataset would still contain dates that weren't verified. Two versions of a dataset can introduce inconsistent results. A final audit of results would be essential to reveal what happened and correct all dates.
* **Data transfer compromising data integrity:** Another analyst checks the dates in a spreadsheet and chooses to import the validated and standardized data back to the database. But suppose the date field from the spreadsheet was incorrectly classified as a text field during the data import (transfer) process. Now some of the dates in the database are stored as text strings. At this point, the data needs to be cleaned to restore its integrity.
* **Data manipulation compromising data integrity:** When checking dates, another analyst notices what appears to be a duplicate record in the database and removes it. But it turns out that the analyst removed a unique record for a company’s subsidiary and not a duplicate record for the company. Your dataset is now missing data and the data must be restored for completeness.

## 

## 

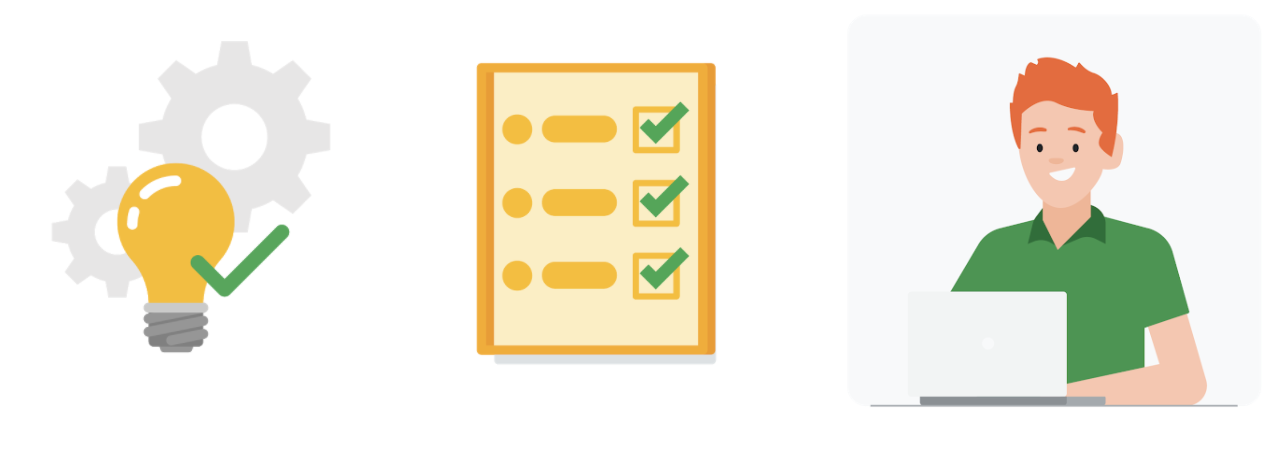
## 

## 

## 

## **Conclusion**

Fortunately, with a standard date format and compliance by all people and systems that work with the data, data integrity can be maintained. But no matter where your data comes from, always be sure to check that it is valid, complete, and clean before you begin any analysis.



## **Reference: Data constraints and examples**

As you progress in your data journey, you'll come across many types of data constraints (or criteria that determine validity). The table below offers definitions and examples of data constraint terms you might come across.

| **Data constraint** | **Definition** | **Examples** |
| --- | --- | --- |
| **Data type** | Values must be of a certain type: date, number, percentage, Boolean, etc. | If the data type is a date, a single number like 30 would fail the constraint and be invalid |
| **Data range** | Values must fall between predefined maximum and minimum values | If the data range is 10-20, a value of 30 would fail the constraint and be invalid |
| **Mandatory** | Values can’t be left blank or empty | If age is mandatory, that value must be filled in |
| **Unique** | Values can’t have a duplicate | Two people can’t have the same mobile phone number within the same service area |
| **Regular expression (regex) patterns** | Values must match a prescribed pattern | A phone number must match ###-###-#### (no other characters allowed) |
| **Cross-field validation** | Certain conditions for multiple fields must be satisfied | Values are percentages and values from multiple fields must add up to 100% |
| **Primary-key** | (Databases only) value must be unique per column | A database table can’t have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program. |
| **Set-membership** | (Databases only) values for a column must come from a set of discrete values | Value for a column must be set to Yes, No, or Not Applicable |
| **Foreign-key** | (Databases only) values for a column must be unique values coming from a column in another table | In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table |
| **Accuracy** | The degree to which the data conforms to the actual entity being measured or described | If values for zip codes are validated by street location, the accuracy of the data goes up. |
| **Completeness** | The degree to which the data contains all desired components or measures | If data for personal profiles required hair and eye color, and both are collected, the data is complete. |
| **Consistency** | The degree to which the data is repeatable from different points of entry or collection | If a customer has the same address in the sales and repair databases, the data is consistent. |

[**BALANCE OBJECTIVES WITH DATA INTEGRITY**](https://www.coursera.org/learn/process-data/lecture/hVFpS/balance-objectives-with-data-integrity)

t's also important to check that the data you use aligns with the business objective. This adds

another layer to the maintenance of data integrity because the data you're using might have limitations that you'll need to deal with. The process of matching data to business objectives can actually be pretty straightforward. Here's a quick example. Let's say you're an analyst for a business that produces and sells auto parts.

If you need to address a question about the revenue generated by the sale of a certain part, then you'd pull up the revenue table from the data set.

If the question is about customer reviews, then you'd pull up the reviews table to analyze the average ratings. But before digging into any analysis, you need to consider a few limitations that might affect it. If the data hasn't been cleaned properly, then you won't be able to use it yet. You would need to wait until a thorough cleaning has been done. Now, let's say you're trying to find how much an average customer spends. You notice the same customer's data showing up in more than one row. This is called duplicate data. To fix this, you might need to change the format of the data, or you might need to change the way you calculate the average. Otherwise, it will seem like the data is for two different people, and you'll be stuck with misleading calculations. You might also realize there's not enough data to complete an accurate analysis. Maybe you only have a couple of months' worth of sales data. There's a slim chance you could wait for more data, but it's more likely that you'll have to change your process or find alternate sources of data while still meeting your objective. I like to think of a data set like a picture. Take this picture. What are we looking at?



Unless you're an expert traveler or know the area, it may be hard to pick out from just these two images.

Visually, it's very clear when we aren't seeing the whole picture. When you get the complete picture, you realize... you're in London!



With incomplete data, it's hard to see the whole picture to get a real sense of what is going on. We sometimes trust data because if it comes to us in rows and columns, it seems like everything we need is there if we just query it. But that's just not true. I remember a time when I found out I didn't have enough data and had to find a solution.

I was working for an online retail company and was asked to figure out how to shorten customer purchase to delivery time. Faster delivery times usually lead to happier customers. When I checked the data set, I found very limited tracking information. We were missing some pretty key details. So the data engineers and I created new processes to track additional information, like the number of stops in a journey. Using this data, we reduced the time it took from purchase to delivery and saw an improvement in customer satisfaction. That felt pretty great! Learning how to deal with data issues while staying focused on your objective will help set you up for success in your career as a data analyst. And your path to success continues. Next step, you'll learn more about aligning data to objectives. Keep it up!

[**WELL-ALIGNED OBJECTIVES AND DATA**](https://www.coursera.org/learn/process-data/supplement/Rhj9e/well-aligned-objectives-and-data)

You can gain powerful insights and make accurate conclusions when data is well-aligned to business objectives. As a data analyst, alignment is something you will need to judge. Good alignment means that the data is relevant and can help you solve a business problem or determine a course of action to achieve a given business objective.

In this reading, you will review the business objectives associated with three scenarios. You will explore how clean data and well-aligned business objectives can help you come up with accurate conclusions. On top of that, you will learn how new variables discovered during data analysis can cause you to set up data constraints so you can keep the data aligned to a business objective.

## 

## **Clean data + alignment to business objective = accurate conclusions**

### 

### **Business objective**

Account managers at Impress Me, an online content subscription service, want to know how soon users view content after their subscriptions are activated.



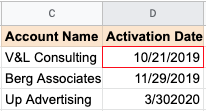
To start off, the data analyst verifies that the data exported to spreadsheets is clean and confirms that the data needed (when users access content) is available. Knowing this, the analyst decides there is good alignment of the data to the business objective. All that is missing is figuring out exactly how long it takes each user to view content after their subscription has been activated.

Here are the data processing steps the analyst takes for a user from an account called V&L Consulting. (These steps would be repeated for each subscribing account, and for each user associated with that account.)

### **Step 1**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Look up the activation date for V&L Consulting | Account spreadsheet |

**Relevant data in spreadsheet:**

****

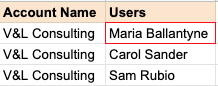
**Result**: October 21, 2019

### 

### **Step 2**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Look up the name of a user belonging to the V&L Consulting account | Account spreadsheet (users tab) |

**Relevant data in spreadsheet**:

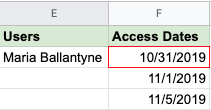


**Result**: Maria Ballantyne

### **Step 3**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Find the first content access date for Maria B. | Content usage spreadsheet |

**Relevant data in spreadsheet:**

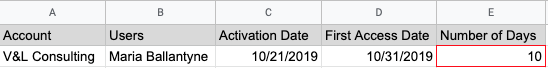
****

**Result**: October 31, 2019

### **Step 4**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Calculate the time between activation and first content usage for Maria B. | New spreadsheet calculation |

**Relevant data in spreadsheet**:



**Result**: 10 days

### 

### 

### 

### **Pro tip 1**

In the above process, the analyst could use **VLOOKUP** to look up the data in Steps 1, 2, and 3 to populate the values in the spreadsheet in Step 4. [VLOOKUP](https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1) is a spreadsheet function that searches for a certain value in a column to return a related piece of information. Using **VLOOKUP** can save a lot of time; without it, you have to look up dates and names manually.

Refer to the [VLOOKUP](https://support.google.com/docs/answer/3093318?hl=en) page in the Google Help Center for how to use the function in Google Sheets.

### 

### **Pro tip 2**

In Step 4 of the above process, the analyst could usethe **DATEDIF** function to automatically calculate the difference between the dates in column C and column D. The function can calculate the number of days between two dates.

Refer to the Microsoft Support [DATEDIF](https://support.microsoft.com/en-us/office/datedif-function-25dba1a4-2812-480b-84dd-8b32a451b35c) page for how to use the function in Excel. The [DAYS360](https://support.microsoft.com/en-us/office/days360-function-b9a509fd-49ef-407e-94df-0cbda5718c2a) function does the same thing in accounting spreadsheets that use a 360-day year (twelve 30-day months).

Refer to the [DATEDIF](https://support.google.com/docs/answer/6055612?hl=en) page in the Google Help Center for how to use the function in Google Sheets.

## **Alignment to business objective + additional data cleaning = accurate conclusions**

### **Business objective**

Cloud Gate, a software company, recently hosted a series of public webinars as free product introductions. The data analyst and webinar program manager want to identify companies that had five or more people attend these sessions. They want to give this list of companies to sales managers who can follow up for potential sales.



The webinar attendance data includes the fields and data shown below.

| **Name** | **<*First name*> <*Last name*>** | **This was required information attendees had to submit** |
| --- | --- | --- |
| **Email Address** | xxxxx@*company*.com | This was required information attendees had to submit |
| **Company** | <*Company name*> | This was optional information attendees could provide |

### 

### 

### 

### **Data cleaning**

The webinar attendance data seems to align with the business objective. But the data analyst and program manager decide that some data cleaning is needed before the analysis. They think data cleaning is required because:

* The company name wasn’t a mandatory field. If the company name is blank, it might be found from the email address. For example, if the email address is username@google.com, the company field could be filled in with Google for the data analysis. This data cleaning step assumes that people with company-assigned email addresses attended a webinar for business purposes.
* Attendees could enter any name. Since attendance across a series of webinars is being looked at, they need to validate names against unique email addresses. For example, if Joe Cox attended two webinars but signed in as Joe Cox for one and Joseph Cox for the other, he would be counted as two different people. To prevent this, they need to check his unique email address to determine that he was the same person. After the validation, Joseph Cox could be changed to Joe Cox to match the other instance.

## 

## **Alignment to business objective + newly discovered variables + constraints = accurate conclusions**

### **Business objective**

An after-school tutoring company, A+ Education, wants to know if there is a minimum number of tutoring hours needed before students have at least a 10% improvement in their assessment scores.



The data analyst thinks there is good alignment between the data available and the business objective because:

* Students log in and out of a system for each tutoring session, and the number of hours is tracked
* Assessment scores are regularly recorded

### **Data constraints for new variables**

After looking at the data, the data analyst discovers that there are other variables to consider. Some students had consistent weekly sessions while other students had scheduled sessions more randomly even though their total number of tutoring hours was the same. The data doesn’t align as well with the original business objective as first thought, so the analyst adds a data constraint to focus only on the students with consistent weekly sessions. This modification helps to get a more accurate picture about the enrollment time needed to achieve a 10% improvement in assessment scores.

## **Key takeaways**

Hopefully these examples give you a sense of what to look for to know if your data aligns with your business objective.

* When there is clean data and good alignment, you can get accurate insights and make conclusions the data supports.
* If there is good alignment but the data needs to be cleaned, clean the data before you perform your analysis.
* If the data only partially aligns with an objective, think about how you could modify the objective, or use data constraints to make sure that the subset of data better aligns with the business objective.

**OVERCOME THE CHALLENGES OF INSUFFICIENT DATA**

[**DEAL WITH INSUFFICIENT DATA**](https://www.coursera.org/learn/process-data/lecture/i8V8h/deal-with-insufficient-data)

Every analyst has been in a situation where there is insufficient data to help with their business objective. Considering how much data is generated every day, it may be hard to believe, but it's true. So let's discuss what you can do when you have insufficient data. We'll cover how to set limits for the scope of your analysis and what data you should include.

At one point, I was a data analyst at a support center. Every day, we received customer questions, which were logged in as support tickets.

I was asked to forecast the number of support tickets coming in per month to figure out how many additional people we needed to hire. It was very important that we had sufficient data spanning back at least a couple of years because I had to account for year-to-year and seasonal changes. If I just had the current year's data available, I wouldn't have known that a spike in January is common and has to do with people asking for refunds after the holidays. Because I had sufficient data, I was able to suggest we hire more people in January to prepare. Challenges are bound to come up, but the good news is that once you know your business objective, you'll be able to recognize whether you have enough data. And if you don't, you'll be able to deal with it before you start your analysis. Now, let's check out some of those limitations you might come across and how you can handle different types of insufficient data.

Say you're working in the tourism industry, and you need to find out which travel plans are searched most often. If you only use data from one booking site, you're limiting yourself to data from just one source. Other booking sites might show different trends that you would want to consider for your analysis. If a limitation like this impacts your analysis, you can stop and go back to your stakeholders to figure out a plan. If your data set keeps updating, that means the data is still incoming and might not be complete. So if there's a brand new tourist attraction that you're analyzing interest and attendance for, there's probably not enough data for you to determine trends. For example, you might want to wait a month to gather data. Or you can check in with the stakeholders and ask about adjusting the objective. For example, you might analyze trends from week to week instead of month to month. You could also base your analysis on trends over the past three months and say, "Here's what attendance at the attraction for month four could look like."

You might not have enough data to know if this number is too low or too high. But you would tell stakeholders that it's your best estimate based on the data that you currently have. On the other hand, your data could be older and no longer be relevant. Outdated data about customer satisfaction won't include the most recent responses. So you'll be relying on the ratings for hotels or vacation rentals that might no longer be accurate. In this case, your best bet might be to find a new data set to work with. Data that's geographically-limited could also be unreliable. If your company is global, you wouldn't want to use data limited to travel in just one country. You would want a data set that includes all countries. So that's just a few of the most common limitations you'll come across and some ways you can address them. You can identify trends with the available data or wait for more data if time allows; you can talk with stakeholders and adjust your objective; or you can look for a new data set.

The need to take these steps will depend on your role in your company and possibly the needs of the wider industry. But learning how to deal with insufficient data is always a great way to set yourself up for success. Your data analyst powers are growing stronger. And just in time. After you learn more about limitations and solutions, you'll learn about statistical power, another fantastic tool for you to use. See you soon!

[**WHEN YOU FIND AN ISSUE WITH YOUR DATA**](https://www.coursera.org/learn/process-data/supplement/NQPE4/when-you-find-an-issue-with-your-data)

When you are getting ready for data analysis, you might realize you don’t have the data you need or you don’t have enough of it. In some cases, you can use what is known as proxy data in place of the real data. Think of it like substituting oil for butter in a recipe when you don’t have butter. In other cases, there is no reasonable substitute and your only option is to collect more data.

Consider the following data issues and suggestions on how to work around them.

## **Data issue 1: no data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data. | If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees. |
| If there isn’t time to collect data, perform the analysis using proxy data from other datasets.  *This is the most common workaround.* | If you are analyzing peak travel times for commuters but don’t have the data for a particular city, use the data from another city with a similar size and demographic. |

## **Data issue 2: too little data**

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Do the analysis using proxy data along with actual data. | If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors. |
| Adjust your analysis to align with the data you already have. | If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: *this conclusion applies to adults 25 years and older* *only*. |

## 

## 

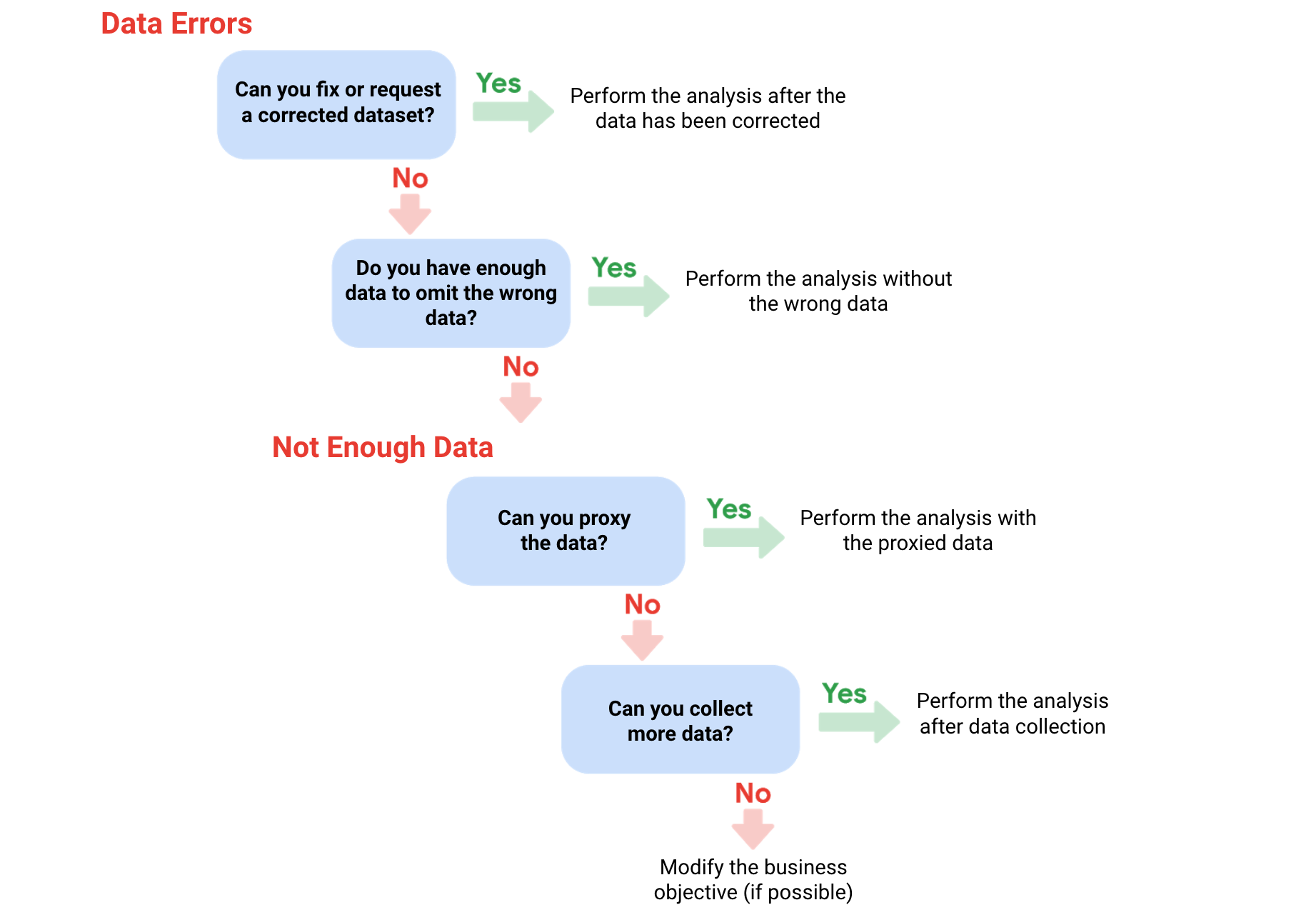
## 

## **Data issue 3: wrong data, including data with errors\***

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| If you have the wrong data because requirements were misunderstood, communicate the requirements again. | If you need the data for female voters and received the data for male voters, restate your needs. |
| Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors. | If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values. |
| If you can’t correct data errors yourself, you can ignore the  wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won’t cause systematic bias. | If your dataset was translated from a different language and some of the translations don’t make sense, ignore the data with bad translation and go ahead with the analysis of the other data. |

***\* Important note:*** *Sometimes data with errors can be a warning sign that the data isn’t reliable. Use your best judgment.*

### **Use the following decision tree as a reminder of how to deal with data errors or not enough data:**

****

[**THE IMPORTANCE OF SAMPLE SIZE**](https://www.coursera.org/learn/process-data/lecture/48VAS/the-importance-of-sample-size)

Okay, earlier we talked about having the right kind of data to meet your business objective and the importance of having the right amount of data to make sure your analysis is as accurate as possible. You might remember that for data analysts, a population is all possible data values in a certain dataset. If you're able to use 100 percent of a population in your analysis, that's great. But sometimes collecting information about an entire population just isn't possible. It's too time-consuming or expensive.

For example, let's say a global organization wants to know more about pet owners who have cats. You're tasked with finding out which kinds of toys cat owners in Canada prefer. But there's millions of cat owners in Canada, so getting data from all of them would be a huge challenge. Fear not! Allow me to introduce you to... sample size!

When you use sample size or a sample, you use a part of a population that's representative of the population. The goal is to get enough information from a small group within a population to make predictions or conclusions about the whole population. The sample size helps ensure the degree to which you can be confident that your conclusions accurately represent the population. For the data on cat owners, a sample size might contain data about hundreds or thousands of people rather than millions.

Using a sample for analysis is more cost-effective and takes less time. If done carefully and thoughtfully, you can get the same results using a sample size instead of trying to hunt down every single cat owner to find out their favorite cat toys. There is a potential downside, though. When you only use a small sample of a population, it can lead to uncertainty.

You can't really be 100 percent sure that your statistics are a complete and accurate representation of the population. This leads to sampling bias, which we covered earlier in the program. Sampling bias is when a sample isn't representative of the population as a whole. This means some members of the population are being overrepresented or underrepresented.

For example, if the survey used to collect data from cat owners only included people with smartphones, then cat owners who don't have a smartphone wouldn't be represented in the data. Using random sampling can help address some of those issues with sampling bias.

Random sampling is a way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen. Going back to our cat owners again, using a random sample of cat owners means cat owners of every type have an equal chance of being chosen. Cat owners who live in apartments in Ontario would have the same chance of being represented as those who live in houses in Alberta.

As a data analyst, you'll find that creating sample sizes usually takes place before you even get to the data. But it's still good for you to know that the data you are going to analyze is representative of the population and works with your objective. It's also good to know what's coming up in your data journey. In the next video, you'll have an option to become even more comfortable with sample sizes. See you there.

[**CALCULATE SAMPLE SIZE**](https://www.coursera.org/learn/process-data/supplement/blyd3/calculate-sample-size)

Before you dig deeper into sample size, familiarize yourself with these terms and definitions:

| **Terminology** | **Definitions** |
| --- | --- |
| **Population** | The entire group that you are interested in for your study. For example, if you are surveying people in your company, the population would be all the employees in your company. |
| **Sample** | A subset of your population. Just like a food sample, it is called a sample because it is only a taste. So if your company is too large to survey every individual, you can survey a representative sample of your population. |
| **Margin of error** | Since a sample is used to represent a population, the sample’s results are expected to differ from what the result would have been if you had surveyed the entire population. This difference is called the margin of error. The smaller the margin of error, the closer the results of the sample are to what the result would have been if you had surveyed the entire population. |
| **Confidence level** | How confident you are in the survey results. For example, a 95% confidence level means that if you were to run the same survey 100 times, you would get similar results 95 of those 100 times. Confidence level is targeted before you start your study because it will affect how big your margin of error is at the end of your study. |
| **Confidence interval** | The range of possible values that the population’s result would be at the confidence level of the study. This range is the sample result +/- the margin of error. |
| **Statistical significance** | The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance. |

## 

## 

## 

## 

## 

## 

## 

## **Things to remember when determining the size of your sample**

When figuring out a sample size, here are things to keep in mind:

* Don’t use a sample size less than 30. It has been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the average result of a population.
* The confidence level most commonly used is 95%, but 90% can work in some cases.

Increase the sample size to meet specific needs of your project:

* For a **higher** confidence level, use a larger sample size
* To **decrease** the margin of error, use a larger sample size
* For **greater** statistical significance, use a larger sample size

**Note:** Sample size calculators use statistical formulas to determine a sample size.

### **Why a minimum sample of 30?**

This recommendation is based on the **Central Limit Theorem (CLT)** in the field of probability and statistics. As sample size increases, the results more closely resemble the normal (bell-shaped) distribution from a large number of samples. A sample of 30 is the smallest sample size for which the CLT is still valid. Researchers who rely on **regression analysis** –statistical methods to determine the relationships between controlled and dependent variables –also prefer a minimum sample of 30.

Still curious? Without getting too much into the math, check out these articles:

* [Central Limit Theorem (CLT)](https://www.investopedia.com/terms/c/central_limit_theorem.asp): This article by Investopedia explains the Central Limit Theorem and briefly describes how it can apply to an analysis of a stock index.
* [Sample Size Formula](https://www.statisticssolutions.com/dissertation-resources/sample-size-calculation-and-sample-size-justification/sample-size-formula/): This article by Statistics Solutions provides a little more detail about why some researchers use 30 as a minimum sample size.

## **Sample sizes vary by business problem**

Sample size will vary based on the type of business problem you are trying to solve.

For example, if you live in a city with a population of 200,000 and get 180,000 people to respond to a survey, that is a large sample size. But without actually doing that, what would an acceptable, smaller sample size look like?

Would 200 be alright if the people surveyed represented every district in the city?

**Answer**: It depends on the stakes.

* A sample size of 200 might be large enough if your business problem is to find out how residents felt about the new library
* A sample size of 200 might not be large enough if your business problem is to determine how residents would vote to fund the library

You could probably accept a larger margin of error surveying how residents feel about the new library versus surveying residents about how they would vote to fund it. For that reason, you would most likely use a larger sample size for the voter survey.



## **Larger sample sizes have a higher cost**

You also have to weigh the cost against the benefits of more accurate results with a larger sample size. Someone who is trying to understand consumer preferences for a new line of products wouldn’t need as large a sample size as someone who is trying to understand the effects of a new drug. For drug safety, the benefits outweigh the cost of using a larger sample size. But for consumer preferences, a smaller sample size at a lower cost could provide good enough results.



## **Knowing the basics is helpful**

Knowing the basics will help you make the right choices when it comes to sample size. You can always raise concerns if you come across a sample size that is too small. A sample size calculator is also a great tool for this. Sample size calculators let you enter a desired confidence level and margin of error for a given population size. They then calculate the sample size needed to statistically achieve those results.

Refer to the [Determine the Best Sample Size](https://www.coursera.org/learn/process-data/lecture/mSj5A/determine-the-best-sample-size) video for a demonstration of a sample size calculator, or refer to the [Sample Size Calculator](https://www.coursera.org/learn/process-data/supplement/ZqcDw/sample-size-calculator) reading for additional information.



## **Key takeaways**

As you continue on your data analytics journey, be sure to familiarize yourself with key terms including population, sample, margin of error, confidence level, and confidence interval before calculating sample size.

**Remember that a minimum sample size of 30 is recommended and that sample size varies depending on the specific business problem**. Also consider the trade-off between accuracy and cost when determining sample size, as larger sample sizes provide more accurate results but at a higher cost. Finally, use sample size calculators to determine the appropriate sample size for your study.

[**SELF-REFLECTION: PRE-CLEANING ACTIVITIES**](https://www.coursera.org/learn/process-data/quiz/dB0CC/self-reflection-pre-cleaning-activities)

[**TEST YOUR KNOWLEDGE ON INSUFFICIENT DATA**](https://www.coursera.org/learn/process-data/quiz/rZuAA/test-your-knowledge-on-insufficient-data)

**TEST YOUR DATA**

[**USING STATISTICAL POWER**](https://www.coursera.org/learn/process-data/lecture/LpFCZ/using-statistical-power)

[A Gentle Introduction to Statistical Power and Power Analysis in Python](https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/) sums it up nicely:

"**Statistical power** can be calculated and reported for a completed experiment to comment on the confidence one might have in the conclusions drawn from the results of the study. It can also be used as a tool to estimate the number of observations or sample size required in order to detect an effect in an experiment."

**Optional:** If you want a more detailed explanation of statistical power and power analysis, the above link is a tutorial that also lists additional references.

[**WHEN DATA ISN'T READILY AVAILABLE**](https://www.coursera.org/learn/process-data/supplement/dOyv7/when-data-isn-t-readily-available)

Earlier, you learned how you can still do an analysis using proxy data if you have no data. You might have some questions about proxy data, so this reading will give you a few more examples of the types of datasets that can serve as alternate data sources.

## **Proxy data examples**

Sometimes the data to support a business objective isn’t readily available. This is when proxy data is useful. Take a look at the following scenarios and where proxy data comes in for each example:

| **Business scenario** | **How proxy data can be used** |
| --- | --- |
| A new car model was just launched a few days ago and the auto dealership can’t wait until the end of the month for sales data to come in. They want sales projections now. | The analyst proxies the number of clicks to the car specifications on the dealership’s website as an estimate of potential sales at the dealership. |
| A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years. | The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years. |
| The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren’t publicly available yet. | The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier. |

## **Open (public) datasets**

If you are part of a large organization, you might have access to lots of sources of data. But if you are looking for something specific or a little outside your line of business, you can also make use of open or public datasets. (You can refer to this [Medium article](https://medium.com/thinkdata/is-there-a-difference-between-open-data-and-public-data-2b8d5608b2f1) for a brief explanation of the difference between open and public data.)

Here's an example. A nasal version of a vaccine was recently made available. A clinic wants to know what to expect for contraindications, but just started collecting first-party data from its patients. A **contraindication** is a condition that may cause a patient not to take a vaccine due to the harm it would cause them if taken. To estimate the number of possible contraindications, a data analyst proxies an open dataset from a trial of the injection version of the vaccine. The analyst selects a subset of the data with patient profiles most closely matching the makeup of the patients at the clinic.

There are plenty of ways to share and collaborate on data within a community. Kaggle ([kaggle.com](https://www.kaggle.com/)) which we previously introduced, has datasets in a variety of formats including the most basic type, Comma Separated Values (CSV) files.



### **CSV, JSON, SQLite, and BigQuery datasets**

* CSV: Check out this [Credit card customers](https://www.kaggle.com/sakshigoyal7/credit-card-customers) dataset, which has information from 10,000 customers including age, salary, marital status, credit card limit, credit card category, etc. (CC0: Public Domain, Sakshi Goyal).
* JSON: Check out this JSON dataset for [trending YouTube videos](https://www.kaggle.com/datasnaek/youtube-new) (CC0: Public Domain, Mitchell J).
* SQLite: Check out this SQLite dataset for 24 years worth of [U.S. wildfire data](https://www.kaggle.com/rtatman/188-million-us-wildfires) (CC0: Public Domain, Rachael Tatman).
* BigQuery: Check out this [Google Analytics 360](https://www.kaggle.com/bigquery/google-analytics-sample) sample dataset from the Google Merchandise Store (CC0 Public Domain, Google BigQuery).

Refer to the Kaggle [documentation for datasets](https://www.kaggle.com/docs/datasets) for more information and search for and explore datasets on your own at [kaggle.com/datasets](https://www.kaggle.com/datasets).

As with all other kinds of datasets, be on the lookout for duplicate data and ‘Null’ in open datasets. Null most often means that a data field was unassigned (left empty), but sometimes Null can be interpreted as the value, 0. It is important to understand how Null was used before you start analyzing a dataset with Null data.

## **Key takeaways**

As you work on data analysis projects, proxy data can often be used to estimate or predict outcomes when actual data is not available. Open or public datasets can be used as proxy data sources, and there are many available online repositories for finding relevant datasets. But be cautious when using proxy data and ensure that it is well-suited for the intended purpose. Finally, check for duplicate data and null values in open datasets before using them for analysis.

[**DETERMINE THE BEST SAMPLE SIZE**](https://www.coursera.org/learn/process-data/lecture/mSj5A/determine-the-best-sample-size)

Great to see you again. In this video, we'll go into more detail about sample sizes and data integrity. If you've ever been to a store that hands out samples, you know it's one of life's little pleasures. For me, anyway! Those small samples are also a very smart way for businesses to learn more about their products from customers without having to give everyone a free sample. A lot of organizations use sample size in a similar way. They take one part of something larger. In this case, a sample of a population. Sometimes they'll perform complex tests on their data to see if it meets their business objectives.

We'll focus on a "big picture" look at the process and what it involves.

As a quick reminder, sample size is a part of a population that is representative of the population.

For businesses, it's a very important tool. It can be both expensive and time-consuming to analyze an entire population of data.

Using sample size usually makes the most sense and can still lead to valid and useful findings.

There are handy calculators online that can help you find sample size. You need to input the confidence level, population size, and margin of error. We've talked about population size before.

To build on that, we'll learn about confidence level and margin of error. Knowing about these concepts will help you understand why you need them to calculate sample size. The confidence level is the probability that your sample accurately reflects the greater population. You can think of it the same way as confidence in anything else. It's how strongly you feel that you can rely on something or someone. Having a 99 percent confidence level is ideal. But **most industries hope** for at least a **90 or 95 percent confidence level**. Industries like pharmaceuticals usually want a confidence level that's as high as possible when they are using a sample size. This makes sense because they're testing medicines and need to be sure they work and are safe for everyone to use.

For other studies, organizations might just need to know that the test or survey results have them heading in the right direction. For example, if a paint company is testing out new colors, a lower confidence level is okay.

You also want to consider the **margin of error** for your study.

You'll learn more about this soon, but it **basically tells you how close your sample size results are to what your results would be if you use the entire population that your sample size represents**. Think of it like this. Let's say that the principal of a middle school approaches you with a study about students' candy preferences. They need to know an appropriate sample size, and they need it now. The school has a student population of 500, and they're asking for a confidence level of 95 percent and a margin of error of 5 percent. We've set up a calculator in a spreadsheet, but you can also easily find this type of calculator by searching "sample size calculator" on the internet. Just like those calculators, our spreadsheet calculator doesn't show any of the more complex calculations for figuring out sample size.

All we need to do is input the numbers for our population, confidence level, and margin of error. And when we type 500 for our population size, 95 for our confidence level percentage, 5 for our margin of error percentage, the result is about 218. That means for this study, an appropriate sample size would be 218. If we surveyed 218 students and found that 55 percent of them preferred chocolate, then we could be pretty confident that would be true of all 500 students. 218 is the minimum number of people we need to survey based on our criteria of a 95 percent confidence level and a 5 percent margin of error.

In case you're wondering, the **confidence level and margin of error** don't have to add up to 100 percent. They'**re independent of each other**.

So let's say we change our margin of error from 5 percent to 3 percent. Then we find that our sample size would need to be larger, about 341 instead of 218, to make the results of the study more representative of the population.

[**SAMPLE SIZE CALCULATOR**](https://www.coursera.org/learn/process-data/supplement/ZqcDw/sample-size-calculator)

In this reading, you will learn the basics of sample size calculators, how to use them, and how to understand the results. A **sample size calculator** tells you how many people you need to interview (or things you need to test) to get results that represent the target population. Let’s review some terms you will come across when using a sample size calculator:

* **Confidence level**: The probability that your sample size accurately reflects the greater population.
* **Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population.
* **Population**: This is the total number you hope to pull your sample from.
* **Sample**: A part of a population that is representative of the population.
* **Estimated response rate**: If you are running a survey of individuals, this is the percentage of people you expect will complete your survey out of those who received the survey.

## **How to use a sample size calculator**

In order to use a sample size calculator, you need to have the population size, confidence level, and the acceptable margin of error already decided so you can input them into the tool. If this information is ready to go, check out these sample size calculators below:

* [Sample size calculator by surveymonkey.com](https://www.surveymonkey.com/mp/sample-size-calculator/)
* [Sample size calculator by raosoft.com](http://www.raosoft.com/samplesize.html)

## **What to do with the results**

After you have plugged your information into one of these calculators, it will give you a recommended sample size. Keep in mind, the calculated sample size is the **minimum** number to achieve what you input for confidence level and margin of error. If you are working with a survey, you will also need to think about the estimated response rate to figure out how many surveys you will need to send out. For example, if you need a sample size of 100 individuals and your estimated response rate is 10%, you will need to send your survey to 1,000 individuals to get the 100 responses you need for your analysis.

Now that you have the basics, try some calculations using the sample size calculators and refer back to this reading if you need a refresher on the definitions.

**CONSIDER THE MARGIN OF ERROR**

[**EVALUATE DATA RELIABILITY**](https://www.coursera.org/learn/process-data/lecture/MrYOj/evaluate-data-reliability)

**As a data analyst, it's important for you to figure out sample size and variables like confidence level and margin of error before running any kind of test or survey**.

It's the best way to make sure your results are objective, and it gives you a better chance of getting statistically significant results. But if you already know the sample size, like when you're given survey results to analyze, you can calculate the margin of error yourself. Then you'll have a better idea of how much of a difference there is between your sample and your population.

**Margin of error is the maximum that the sample results are expected to differ from those of the actual population.**

It would be great to survey or test an entire population, but it's usually impossible or impractical to do this. So instead, we take a sample of the larger population.

Based on the sample size, the resulting margin of error will tell us how different the results might be compared to the results if we had surveyed the entire population.

**Margin of error helps you understand how reliable the data from your hypothesis testing is.**

The closer to zero the margin of error, the closer your results from your sample would match results from the overall population.

For example, let's say you completed a nationwide survey using a sample of the population. You asked people who work five-day workweeks whether they like the idea of a four-day workweek. So your survey tells you that 60% prefer a four-day workweek. The margin of error was 10%, which tells us that between 50 and 70% like the idea. So if we were to survey all five-day workers nationwide, between 50 and 70% would agree with our results.

Keep in mind that our range is between 50 and 70%. That's because the margin of error is counted in both directions from the survey results of 60%. If you set up a 95% confidence level for your survey, there'll be a 95% chance that the entire population's responses will fall between 50 and 70% saying, yes, they want a four-day workweek.

Since your margin of error overlaps with that 50% mark, you can't say for sure that the public likes the idea of a four-day workweek. In that case, you'd have to say your survey was inconclusive.

Now, if you wanted a lower margin of error, say 5%, with a range between 55 and 65%, you could increase the sample size. But if you've already been given the sample size, you can calculate the margin of error yourself.

Then you can decide yourself how much of a chance your results have of being statistically significant based on your margin of error. In general, the more people you include in your survey, the more likely your sample is representative of the entire population.

Decreasing the confidence level would also have the same effect, but that would also make it less likely that your survey is accurate.

**So to calculate margin of error, you need three things: population size, sample size, and confidence level.**

And just like with sample size, you can find lots of calculators online by searching "margin of error calculator."

But we'll show you in a spreadsheet, just like we did when we calculated sample size.

Let's say you're running a study on the effectiveness of a new drug. You have a sample size of 500 participants whose condition affects 1% of the world's population. That's about 80 million people, which is the population for your study.

Since it's a drug study, you need to have a confidence level of 99%. You also need a low margin of error. Let's calculate it. We'll put the numbers for population, confidence level, and sample size, in the appropriate spreadsheet cells. And our result is a margin of error of close to 6%, plus or minus. When the drug study is complete, you'd apply the margin of error to your results to determine how reliable your results might be.

Calculators like this one in the spreadsheet are just one of the many tools you can use to ensure data integrity.

And it's also good to remember that checking for data integrity and aligning the data with your objectives will put you in good shape to complete your analysis.

Knowing about sample size, statistical power, margin of error, and other topics we've covered will help your analysis run smoothly. That's a lot of new concepts to take in. If you'd like to review them at any time, you can find them all in the glossary, or feel free to rewatch the video! Soon you'll explore the ins and outs of clean data. The data adventure keeps moving! I'm so glad you're moving along with it. You got this!

[**ALL ABOUT MARGIN OF ERROR**](https://www.coursera.org/learn/process-data/supplement/INdVZ/all-about-margin-of-error)

**Margin of error** is the maximum amount that the sample results are expected to differ from those of the actual population. More technically, the margin of error defines a range of values below and above the average result for the sample. The average result for the entire population is expected to be within that range. We can better understand the margin of error by using some examples below.

## 

## **Margin of error in baseball**

****

Imagine you are playing baseball and that you are up at bat. The crowd is roaring, and you are getting ready to try to hit the ball. The pitcher delivers a fastball traveling about 90-95 mph, which takes about 400 milliseconds (ms) to reach the catcher’s glove. You swing and miss the first pitch because your timing was a little off. You wonder if you should have swung slightly earlier or slightly later to hit a home run. That time difference can be considered the margin of error, and it tells us how close or far your timing was from the average home run swing.

## **Margin of error in marketing**

The margin of error is also important in marketing. Let’s use A/B testing as an example. **A/B testing** (or split testing) tests two variations of the same web page to determine which page is more successful in attracting user traffic and generating revenue. User traffic that gets monetized is known as the **conversion rate**. A/B testing allows marketers to test emails, ads, and landing pages to find the data behind what is working and what isn’t working. Marketers use the **confidence interval** (determined by the conversion rate and the margin of error) to understand the results.

For example, suppose you are conducting an A/B test to compare the effectiveness of two different email subject lines to entice people to open the email. You find that subject line A: “Special offer just for you” resulted in a 5% open rate compared to subject line B: “Don’t miss this opportunity” at 3%.

Does that mean subject line A is better than subject line B? It depends on your margin of error. If the margin of error was 2%, then subject line A’s actual open rate or confidence interval is somewhere between 3% and 7%. Since the lower end of the interval overlaps with subject line B’s results at 3%, you can’t conclude that there is a statistically significant difference between subject line A and B.

Examining the margin of error is important when making conclusions based on your test results.

## **Want to calculate your margin of error?**

All you need is population size, confidence level, and sample size. In order to better understand this calculator, review these terms:

* **Confidence level**: A percentage indicating how likely your sample accurately reflects the greater population
* **Population**: The total number you pull your sample from
* **Sample**: A part of a population that is representative of the population
* **Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population

In most cases, a 90% or 95% confidence level is used. But, depending on your industry, you might want to set a stricter confidence level. A 99% confidence level is reasonable in some industries, such as the pharmaceutical industry.

After you have settled on your population size, sample size, and confidence level, plug the information into a margin of error calculator like the ones below:

* [Margin of error calculator by Good Calculators (free online calculators)](https://goodcalculators.com/margin-of-error-calculator/)
* [Margin of error calculator by CheckMarket](https://www.checkmarket.com/sample-size-calculator/#sample-size-margin-of-error-calculator)

## **Key takeaways**

Margin of error is used to determine how close your sample’s result is to what the result would likely have been if you could have surveyed or tested the entire population. **Margin of error helps you understand and interpret survey or test results in real-life.**

Calculating the margin of error is particularly helpful when you are given the data to analyze. After using a calculator to calculate the margin of error, you will know how much the sample results might differ from the results of the entire population.

[**GLOSSARY TERMS FROM MODULE 1**](https://www.coursera.org/learn/process-data/supplement/CF2ej/glossary-terms-from-module-1)

**Terms and definitions for Course 4, Module 1**

**Accuracy:** The degree to which the data conforms to the actual entity being measured or described

**Completeness:** The degree to which the data contains all desired components or measures

**Confidence interval:** A range of values that conveys how likely a statistical estimate reflects the population

**Confidence level:** The probability that a sample size accurately reflects the greater population

**Consistency:** The degree to which data is repeatable from different points of entry or collection

**Cross-field validation:** A process that ensures certain conditions for multiple data fields are satisfied

**Data constraints:** The criteria that determine whether a piece of a data is clean and valid

**Data integrity:** The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

**Data manipulation:** The process of changing data to make it more organized and easier to read

**Data range:** Numerical values that fall between predefined maximum and minimum values

**Data replication:** The process of storing data in multiple locations

**DATEDIF:** A spreadsheet function that calculates the number of days, months, or years between two dates

**Estimated response rate**: The average number of people who typically complete a survey

**Hypothesis testing:** A process to determine if a survey or experiment has meaningful results

**Mandatory:** A data value that cannot be left blank or empty

**Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population

**Random sampling:** A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

**Regular expression (RegEx):** A rule that says the values in a table must match a prescribed pattern

**SID**

[**DATA**](https://www.coursera.org/learn/data-preparation/lecture/BJn55/feel-confident-in-your-data)

h