Week 1

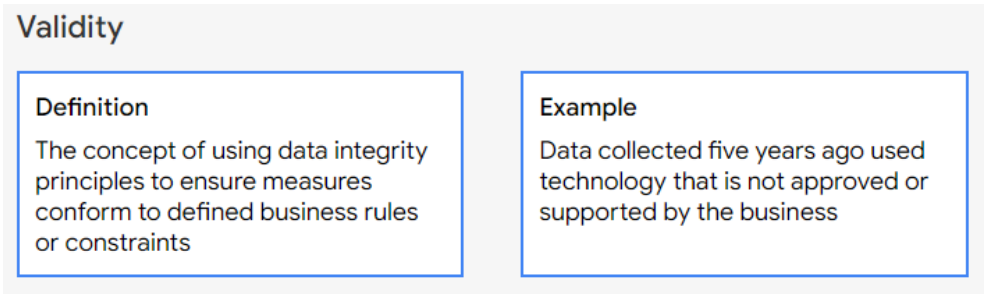
- Data integrity: the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.
- Data replication: the process of storing data in multiple locations
- Data transfer: the process of copying data from a storage device to memory, or from one computer to another.
- Data manipulation: the process of changing data to make it more organized and easier to read.
- ## Types of insufficient data

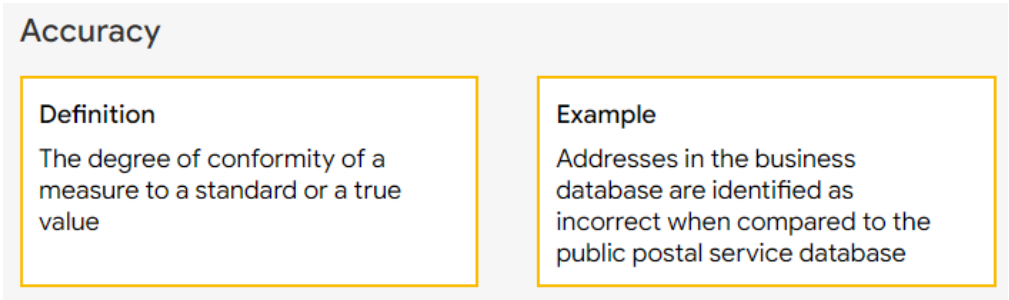
  - Data from only one source
  - Data that keeps updating
  - Outdated data
  - Geographically-limited data
- ## Ways to address insufficient data

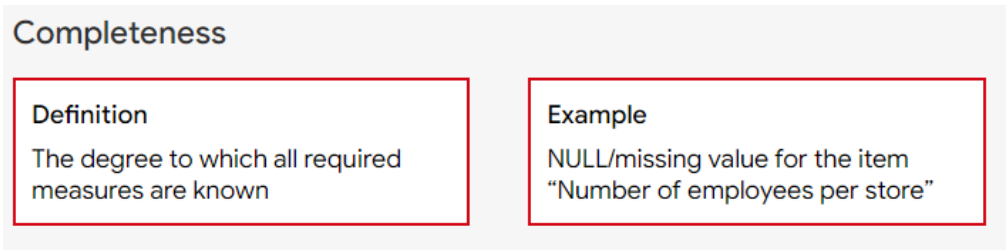
  - Identify trends with the available data
  - Wait for more data if time allows
  - Talk with stakeholders and adjust your objective
  - Look for a new dataset
Terminology	Definitions
Population	The entire group that you are interested in for your study. For example, if you are surveying people in your 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.
- Hypothesis testing: a way to see if a survey or experiment has meaningful results

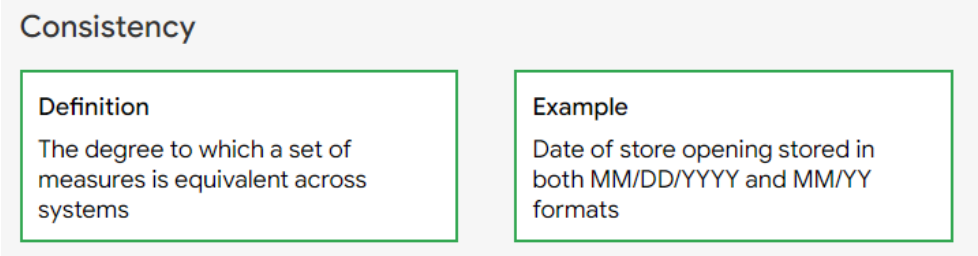
- A 0.8 or 80% statistical power is typically considered the minimum for statistical significance. But most industries hope for at least a 90% or 95% percent confidence level
- Confidence level: the probability that your sample size accurately reflects the greater population
- To calculate margin of error you need: population size, sample size, confidence level.

Week 2

- Dirty data: data that is incomplete, incorrect, or irrelevant to the problem you're trying to solve.
- Data warehousing specialists: develop processes and procedures to effectively store and organize data
- Data engineers transform data into a useful format for analysis; give it a reliable infrastructure; and develop, maintain, and test databases and related systems.
- Dirty data types: duplicate data, outdated data, incomplete data, incorrect data, inconsistent data.
- Data validation: a tool for checking the accuracy and quality of data before adding or importing it.
- 

- 

- 



- Merger: an agreement that unites two organizations into a single new one.
- Data merging: the process of combining two or more datasets into a single dataset.
  - Do I have all the data I need?
  - Does the data I need exist within these datasets?
  - Does the data need to be cleaned, or are they ready for me to use?
  - Are the datasets cleaned to the same standard?

Common mistakes to avoid

- **Not checking for spelling errors**: Misspellings can be as simple as typing or input errors. Most of the time the wrong spelling or common grammatical errors can be detected, but it gets harder with things like names or addresses. For example, if you are working with a spreadsheet table of customer data, you might come across a customer named “John” whose name has been input incorrectly as “Jon” in some places. The spreadsheet’s spellcheck probably won’t flag this, so if you don’t double-check for spelling errors and catch this, your analysis will have mistakes in it.
- **Forgetting to document errors**: Documenting your errors can be a big time saver, as it helps you avoid those errors in the future by showing you how you resolved them. For example, you might find an error in a formula in your spreadsheet. You discover that some of the dates in one of your columns haven’t been formatted correctly. If you make a note of this fix, you can reference it the next time your formula is broken, and get a head start on troubleshooting. Documenting your errors also helps you keep track of changes in your work, so that you can backtrack if a fix didn’t work.
- **Not checking for misfielded values**: A misfielded value happens when the values are entered into the wrong field. These values might still be formatted correctly, which makes them harder to catch if you aren’t careful. For example, you might have a dataset with columns for cities and countries. These are the same type of data, so they are easy to mix up. But if you were trying to find all of

the instances of Spain in the country column, and Spain had mistakenly been entered into the city column, you would miss key data points. Making sure your data has been entered correctly is key to accurate, complete analysis.

- **Overlooking missing values**: Missing values in your dataset can create errors and give you inaccurate conclusions. For example, if you were trying to get the total number of sales from the last three months, but a week of transactions were missing, your calculations would be inaccurate. As a best practice, try to keep your data as clean as possible by maintaining completeness and consistency.
- **Only looking at a subset of the data**: It is important to think about all of the relevant data when you are cleaning. This helps make sure you understand the whole story the data is telling, and that you are paying attention to all possible errors. For example, if you are working with data about bird migration patterns from different sources, but you only clean one source, you might not realize that some of the data is being repeated. This will cause problems in your analysis later on. If you want to avoid common errors like duplicates, each field of your data requires equal attention.
- **Losing track of business objectives**: When you are cleaning data, you might make new and interesting discoveries about your dataset-- but you don't want those discoveries to distract you from the task at hand. For example, if you were working with weather data to find the average number of rainy days in your city, you might notice some interesting patterns about snowfall, too. That is really interesting, but it isn't related to the question you are trying to answer right now. Being curious is great! But try not to let it distract you from the task at hand.
- **Not fixing the source of the error:** Fixing the error itself is important. But if that error is actually part of a bigger problem, you need to find the source of the issue. Otherwise, you will have to keep fixing that same error over and over again. For example, imagine you have a team spreadsheet that tracks everyone's progress. The table keeps breaking because different people are entering different values. You can keep fixing all of these problems one by one, or you can set up your table to streamline data entry so everyone is on the same page. Addressing the source of the errors in your data will save you a lot of time in the long run.
- **Not analyzing the system prior to data cleaning:** If we want to clean our data and avoid future errors, we need to understand the root cause of your dirty data. Imagine you are an auto mechanic. You would find the cause of the problem before you started fixing the car, right? The same goes for data. First, you figure out where the errors come from. Maybe it is from a data entry error, not setting up a spell check, lack of formats, or from duplicates. Then, once you understand where bad data comes from, you can control it and keep your data clean.
- **Not backing up your data prior to data cleaning**: It is always good to be proactive and create your data backup before you start your data clean-up. If

your program crashes, or if your changes cause a problem in your dataset, you can always go back to the saved version and restore it. The simple procedure of backing up your data can save you hours of work-- and most importantly, a headache.

- **Not accounting for data cleaning in your deadlines/process**: All good things take time, and that includes data cleaning. It is important to keep that in mind when going through your process and looking at your deadlines. When you set aside time for data cleaning, it helps you get a more accurate estimate for ETAs for stakeholders, and can help you know when to request an adjusted ETA.
- https://support.google.com/a/users/answer/9604139?hl=en#zippy=%2Clearn-how
- Conditional formatting: a spreadsheet tool that changes how cells appear when values meet specific conditions
- Data mapping: the process of matching fields from one data source to another

Week 4

- Verification: a process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable
- Changelog: a file containing a chronologically ordered list of modifications made to a project
- Correct the most common problems:
    - **Sources of errors**: Did you use the right tools and functions to find the source of the errors in your dataset?
    - **Null data**: Did you search for NULLs using conditional formatting and filters?
    - **Misspelled words**: Did you locate all misspellings?
    - **Mistyped numbers**: Did you double-check that your numeric data has been entered correctly?
    - **Extra spaces and characters**: Did you remove any extra spaces or characters using the **TRIM** function?
    - **Duplicates**: Did you remove duplicates in spreadsheets using the **Remove Duplicates** function or **DISTINCT** in SQL?
    - **Mismatched data types**: Did you check that numeric, date, and string data are typecast correctly?
    - **Messy (inconsistent) strings**: Did you make sure that all of your strings are consistent and meaningful?
    - **Messy (inconsistent) date formats**: Did you format the dates consistently throughout your dataset?
    - **Misleading variable labels (columns)**: Did you name your columns meaningfully?
    - **Truncated data**: Did you check for truncated or missing data that needs correction?
    - **Business Logic**: Did you check that the data makes sense given your knowledge of the business?
        - Documentation: the process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort

- Documentation can :
    1. Recover data-cleaning errors
    2. Inform other users of changes
    3. Determine quality of data
- Example of a changelog:

```
# Changelog
This file contains the notable changes to the project

Version 1.0.0 (02-23-2019)
## New
    - Added column classifiers (Date, Time, PerUnitCost, TotalCost, etc. )
    - Added Column “AveCost” to track average item cost

## Changes
    - Changed date format to MM-DD-YYYY
    - Removal of whitespace (cosmetic)

## Fixes
    - Fixed misalignment in Column "TotalCost" where some rows did not match with correct
    - Fixed SUM to run over entire column instead of partial

```

- Common data errors: human error in data entry, flawed processes, system issues.

Week 5

-