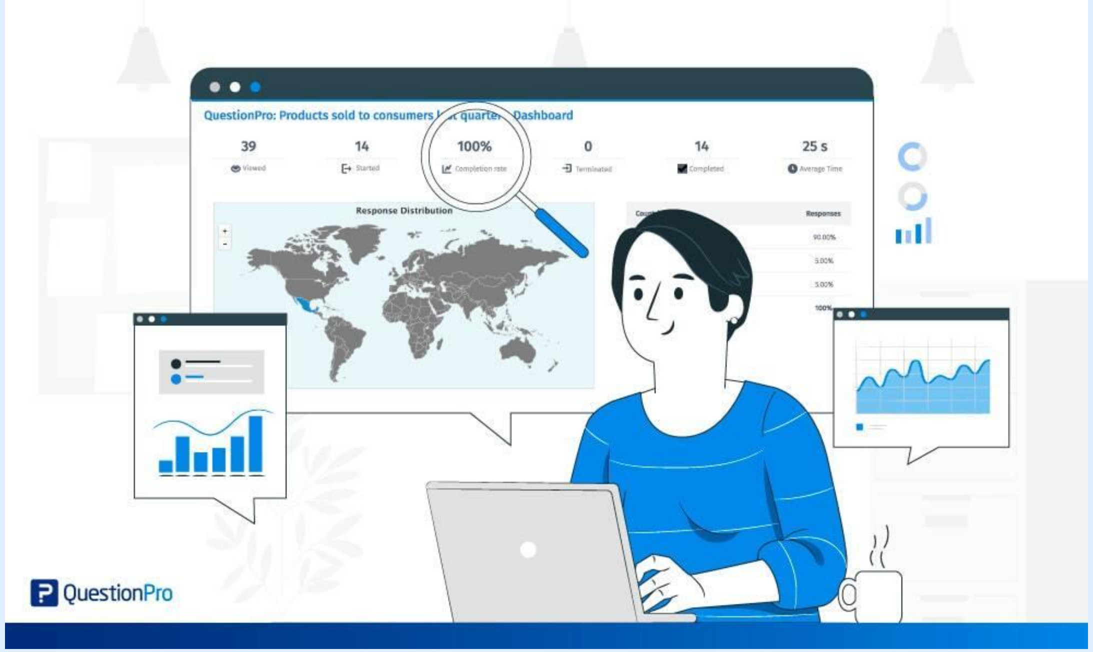
Data Analysis: The First Steps

Omar Mohamed Elgharib

What is data analysis

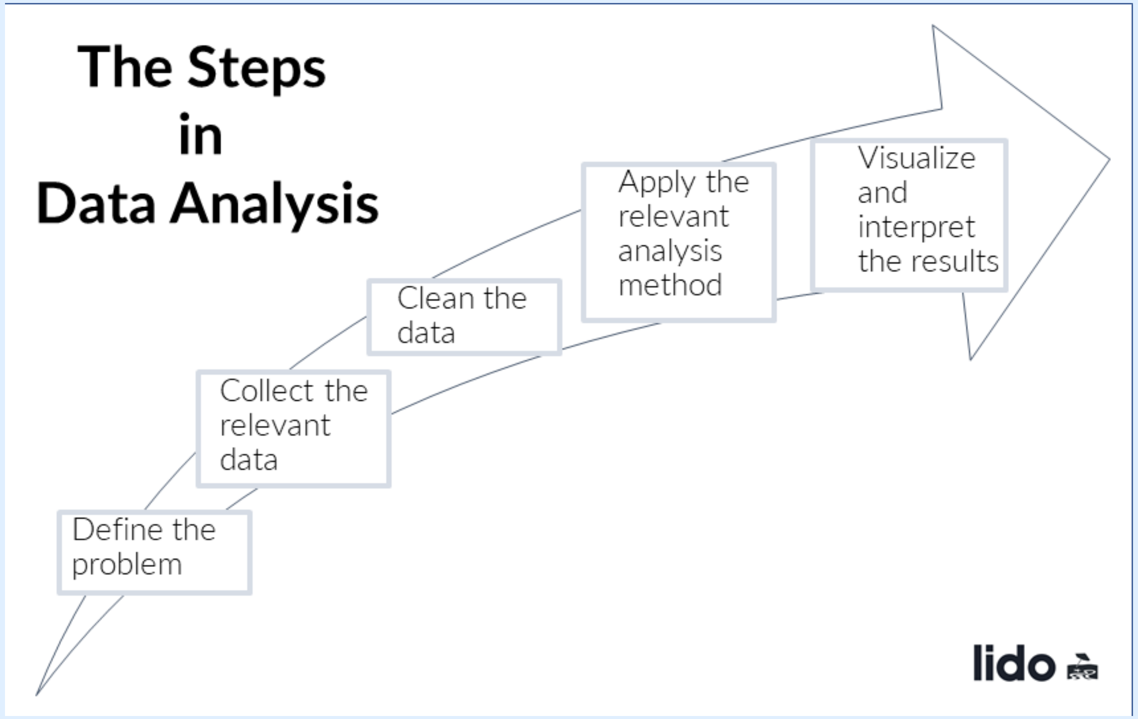
- Data analysis is the process of **inspecting**, **cleansing**, **transforming**, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.



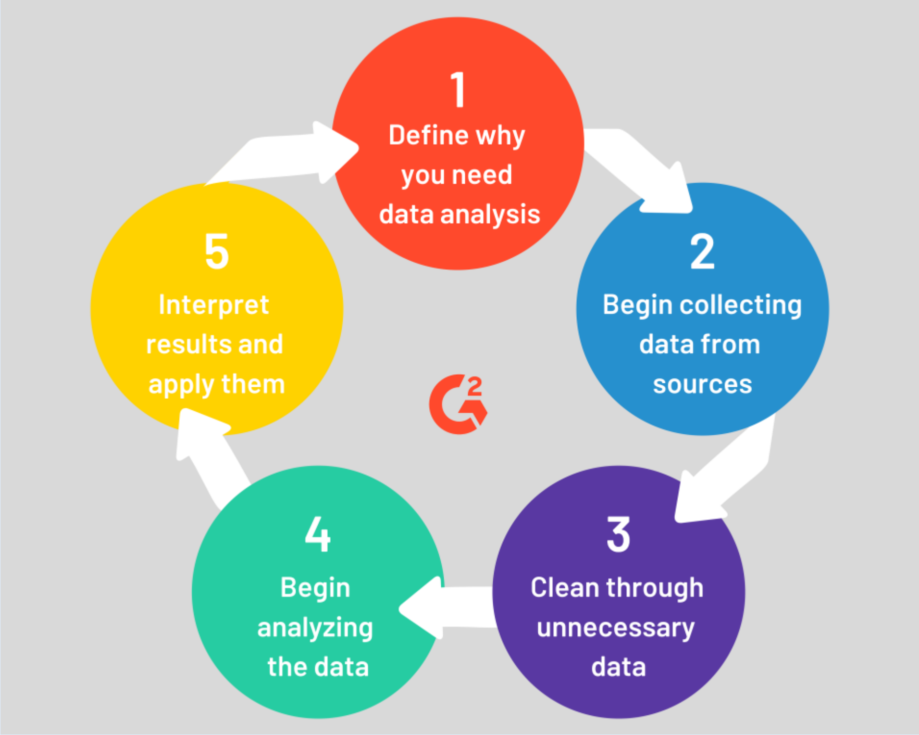
Why data analysis is important

- From small businesses to global enterprises, the amount of data businesses generate today is simply staggering, and this is why the term “big data” has become so buzzwordy.
- However, this mountain of data hardly does much other than clog up cloud storage and databases without proper data analysis.
- To uncover a variety of insights that sit within your systems, learn more about data analytics and implement it to extract valuable insights.

Steps in data analysis

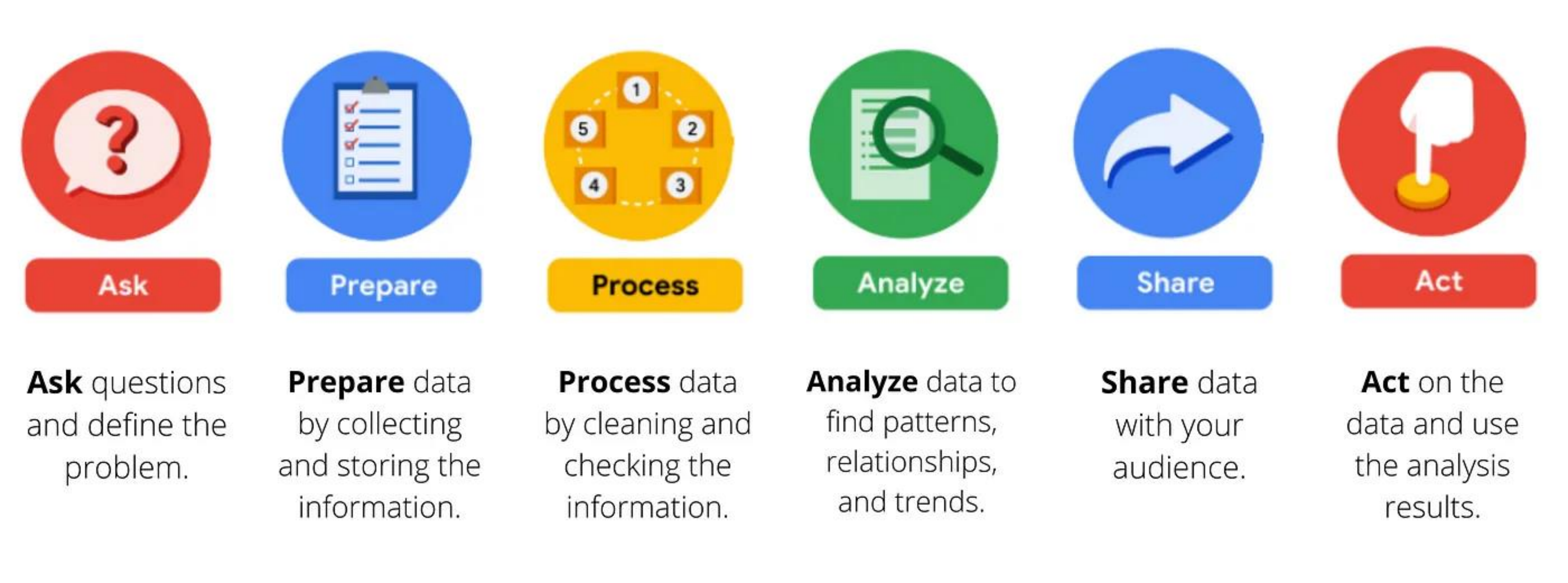

The Steps
in
Data Analysis
Define the problem
Collect the relevant data
Clean the data
Apply the relevant analysis method
Visualize and interpret the results
lido

Source: https://www.lido.app/post/steps-in-conducting-data-analysis


1
Define why you need data analysis
2
Begin collecting data from sources
3
Clean through unnecessary data
4
Begin analyzing the data
5
Interpret results and apply them

Source: https://www.g2.com/articles/data-analysis-process

Steps in data analysis



Ask questions and define the problem.

Prepare data by collecting and storing the information.

Process data by cleaning and checking the information.

Analyze data to find patterns, relationships, and trends.

Share data with your audience.

Act on the data and use the analysis results.

Steps in data analysis

1- Ask

First and foremost, the analysts needed to define what the project would entail and what would constitute a successful outcome. As a result, they asked effective questions and collaborated with leaders and managers who were interested in the outcome of their people analysis to determine these things.

Steps in data analysis

They asked the following types of questions:

- What do you think new employees need to learn to be successful in their first year on the job?
- Have you gathered data from new employees before? If so, may we have access to the historical data?
- Do you believe managers with higher retention rates offer new employees something extra or unique?
- What do you suspect is a leading cause of dissatisfaction among new employees?
- By what percentage would you like employee retention to increase in the next fiscal year?

Steps in data analysis

2. Prepare

It all started with careful planning. The group established a three-month timeline and decided how they would communicate their progress to interested parties. During this step, the analysts also determined what data they required to achieve the successful outcome identified in the previous step; in this case, the analysts chose to collect the data from an online survey of new employees. They prepared by doing the following:

Steps in data analysis

- They developed specific questions to ask about employee satisfaction with different business processes, such as hiring and onboarding, and their overall compensation.
- They established rules for who would have access to the data collected in this case, anyone outside the group wouldn't have access to the raw data, but could view summarized or aggregated data. For example, an individual's compensation wouldn't be available, but salary ranges for groups of individuals would be viewable.

Steps in data analysis

- They finalized what specific information would be gathered, and how best to present the data visually. The analysts brainstormed possible project- and data-related issues and how to avoid them.
- How quickly does a decision need to be made?

Steps in data analysis

3. Process

- The survey was distributed by the group. Great analysts understand how to respect their data as well as the people who provide it. Because employees provided the data, it was critical that all employees gave their consent to participate. The data analysts also ensured that employees were aware of how their data would be collected, stored, managed, and safeguarded. They took the following steps to maintain confidentiality and effectively protect and store the data:

Steps in data analysis

- They restricted access to the data to a limited number of analysts.
- They cleaned the data to make sure it was complete, correct, and relevant. Certain data was aggregated and summarized without revealing individual responses.
- They uploaded raw data to an internal data warehouse for an additional layer of security.

Steps in data analysis

4. Analyze

- The analysts then did what they do best: they analyzed! The data analysts discovered that an employee’s experience with certain processes was a key indicator of overall job satisfaction based on the completed surveys. These were their findings:
- Employees who experienced a long and complicated hiring process were most likely to leave the company.
- Employees who experienced an efficient and transparent evaluation and feedback process were most likely to remain with the company.

Steps in data analysis

5. Share

- The analysts were as cautious in sharing the report as they were in protecting the data. They presented their findings in the following manner:
- They shared the report with managers who met or exceeded the minimum number of direct reports with submitted responses to the survey.
- They presented the results to the managers to make sure they had the full picture.

Steps in data analysis

- They asked the managers to personally deliver the results to their teams.
- This process gave managers an opportunity to communicate the results with the right context. As a result, they could have productive team conversations about next steps to improve employee engagement.

Steps in data analysis

6. Act

- The team of analysts' final stage of the process was to collaborate with company leaders to determine how best to implement changes and take actions based on the findings. Their recommendations were as follows:
- Standardize the hiring and evaluation process for employees based on the most efficient and transparent practices.

Steps in data analysis

- Conduct the same survey annually and compare results with those from the previous year.
- The same survey was distributed to employees a year later. Analysts predicted that a comparison of the two sets of results would show that the action plan was effective. As it turns out, the changes improved new employee retention, and leaders’ actions were successful!

Steps in data analysis

Data + business knowledge = mystery solved

Blending data with business knowledge, plus maybe a touch of gut instinct, will be a common part of your process as a junior data analyst. The key is figuring out the exact mix for each particular project. A lot of times, it will depend on the goals of your analysis. That is why analysts often ask, “How do I define success for this project?”

Steps in data analysis

- In addition, try asking yourself these questions about a project to help find the perfect balance:
- What kind of results are needed?
- Who will be informed?
- Am I answering the question being asked?

Link to source: https://medium.com/codex/life-cycle-of-a-data-analytics-project-954d0e6926fe

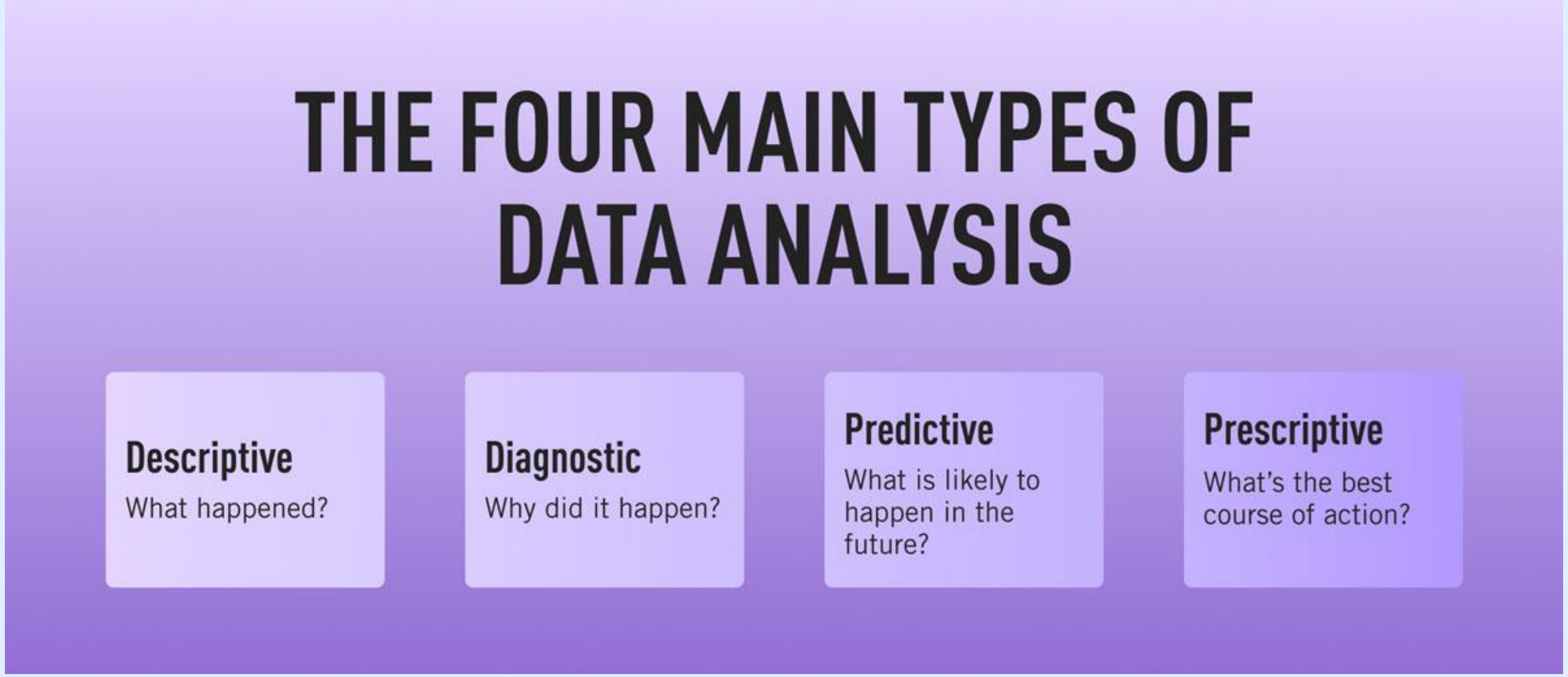
Steps in data analysis in short

Data analysis process

- **Identify** the business question you'd like to answer. What problem is the company trying to solve? What do you need to measure, and how will you measure it?
- **Collect** the raw data sets you'll need to help you answer the identified question. Data collection might come from internal sources, like a company's client relationship management (CRM) software, or from secondary sources, like government records or social media application programming interfaces (APIs).
- **Clean** the data to prepare it for analysis. This often involves purging duplicate and anomalous data, reconciling inconsistencies, standardizing data structure and format, and dealing with white spaces and other syntax errors.
- **Analyze** the data. By manipulating the data using various data analysis techniques and tools, you can begin to find trends, correlations, outliers, and variations that tell a story. During this stage, you might use data mining to discover patterns within databases or data visualization software to help transform data into an easy-to-understand graphical format.
- **Interpret** the results of your analysis to see how well the data answered your original question. What recommendations can you make based on the data? What are the limitations to your conclusions?

Link to source: https://www.coursera.org/articles/what-is-data-analysis-with-examples

Types of data analysis



THE FOUR MAIN TYPES OF DATA ANALYSIS

Descriptive
What happened?

Diagnostic
Why did it happen?

Predictive
What is likely to happen in the future?

Prescriptive
What's the best course of action?

Source: Link

Types of data analysis

- **Descriptive analytics** looks at what has happened in the past.

- **Diagnostic analytics** seeks to delve deeper in order to understand why something happened. The main purpose of diagnostic analytics is to identify and respond to anomalies within your data. For example: If your descriptive analysis shows that there was a 20% drop in sales for the month of March, you’ll want to find out why. The next logical step is to perform a diagnostic analysis.

Types of data analysis

- **Predictive analytics** seeks to predict what is likely to happen in the future. Based on past patterns and trends, data analysts can devise predictive models which estimate the likelihood of a future event or outcome. This is especially useful as it enables businesses to plan ahead.
- **Prescriptive analytics** looks at what has happened, why it happened, and what might happen in order to determine what should be done next.

Data quality

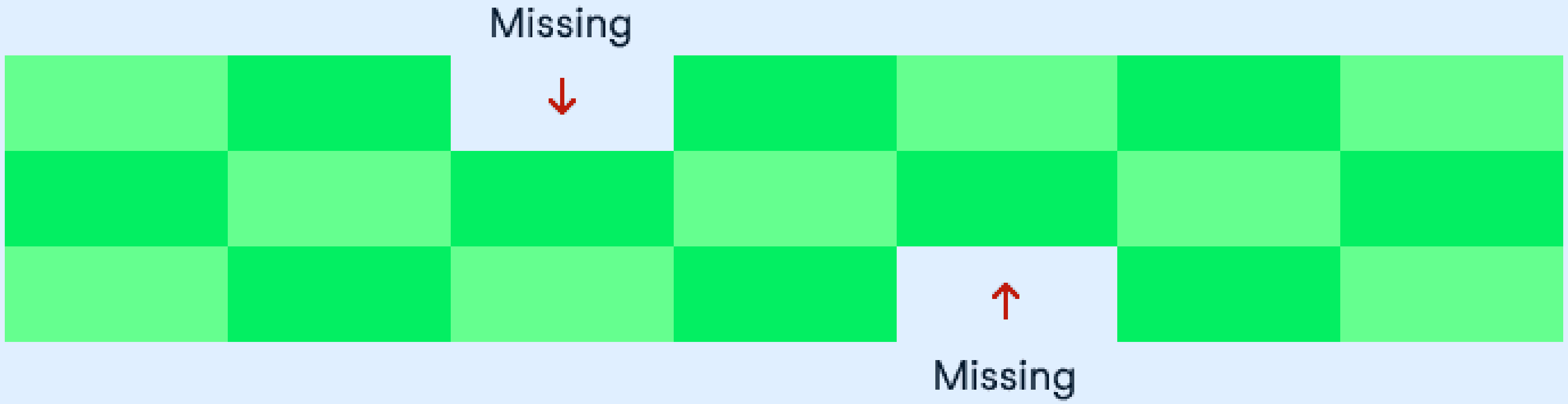
What are Data Quality Dimensions?

Data Quality is a measurement of the degree to which data is fit for purpose. Good data quality generates trust in data. Data Quality Dimensions are a measurement of a specific attribute of a data's quality.

Data quality

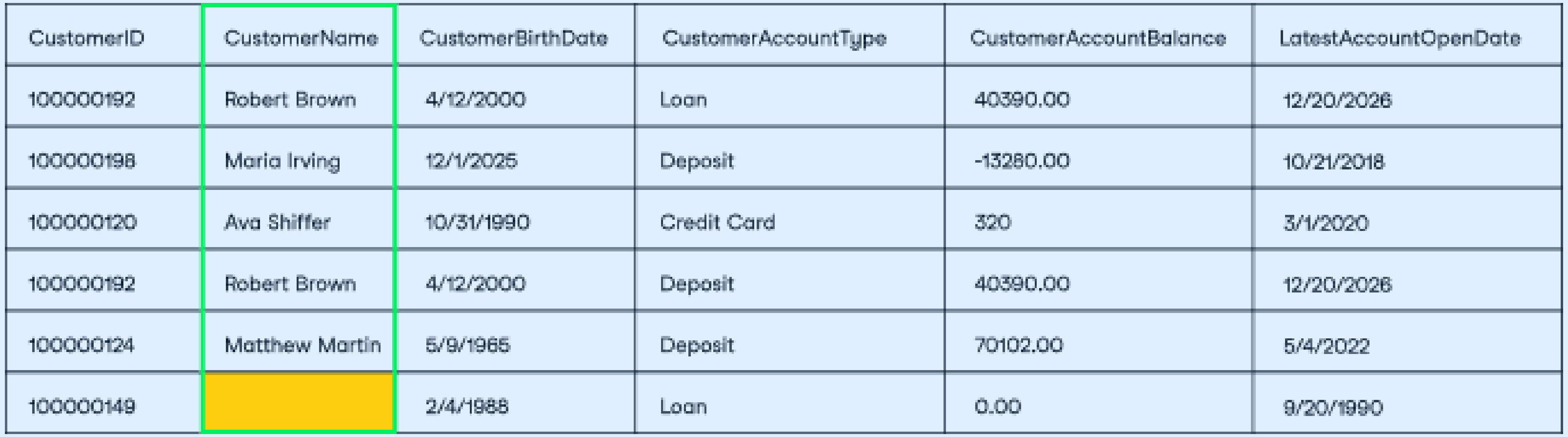
Completeness

Completeness measures the degree to which all expected records in a dataset are present. At a data element level, completeness is the degree to which all records have data populated when expected.



Data quality

- **Completeness Example**



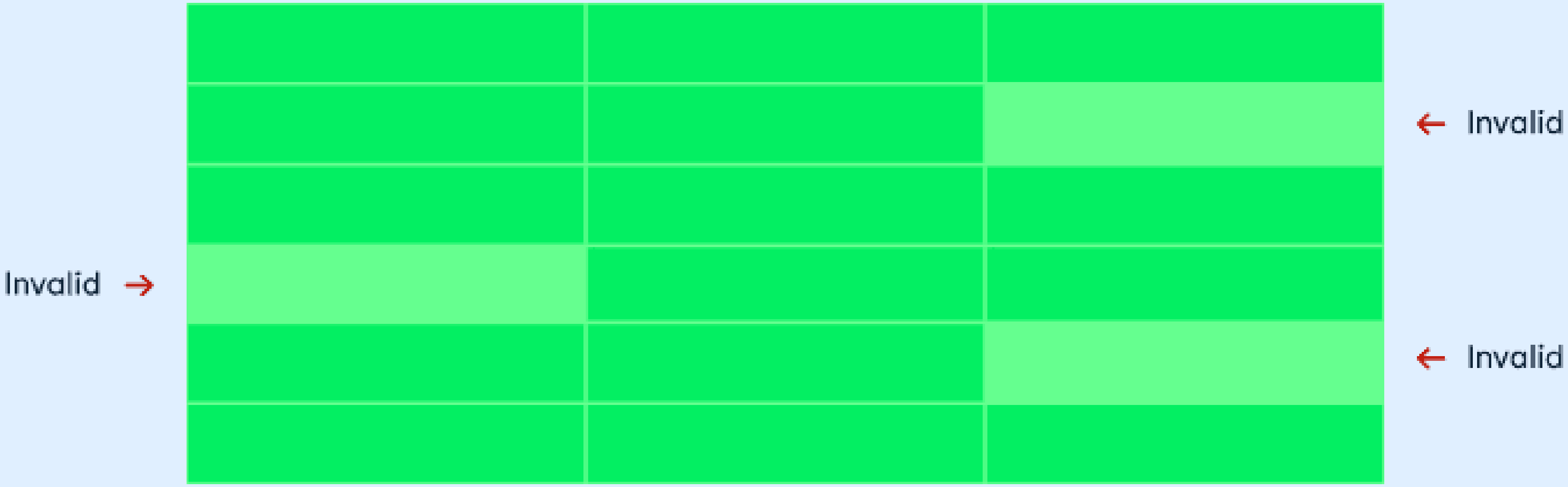
CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

All records must have a value populated in the CustomerName field.

Data quality

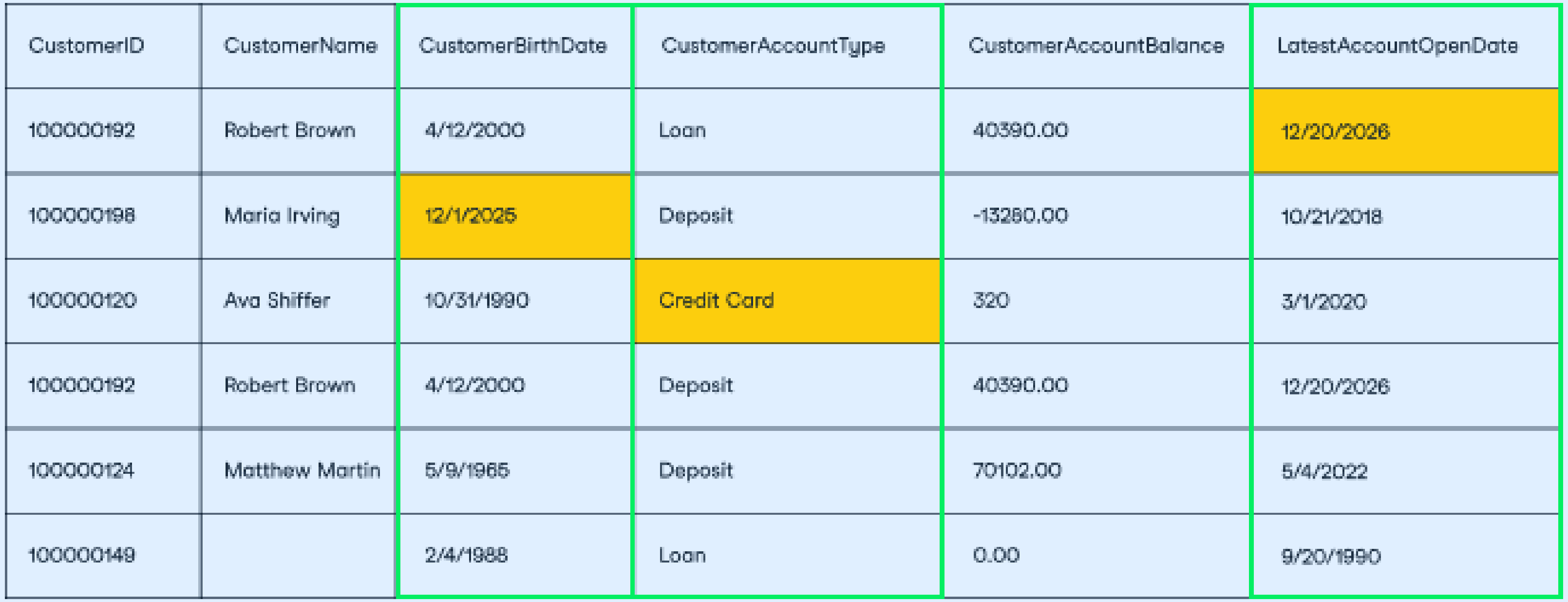
- **Validity**

Validity measures the degree to which the values in a data element are valid.



Data quality

- **Validity Example**
- CustomerBirthDate value must be a date in the past.
- CustomerAccountType value must be either Loan or Deposit.
- LatestAccountOpenDate value must be a date in the past.

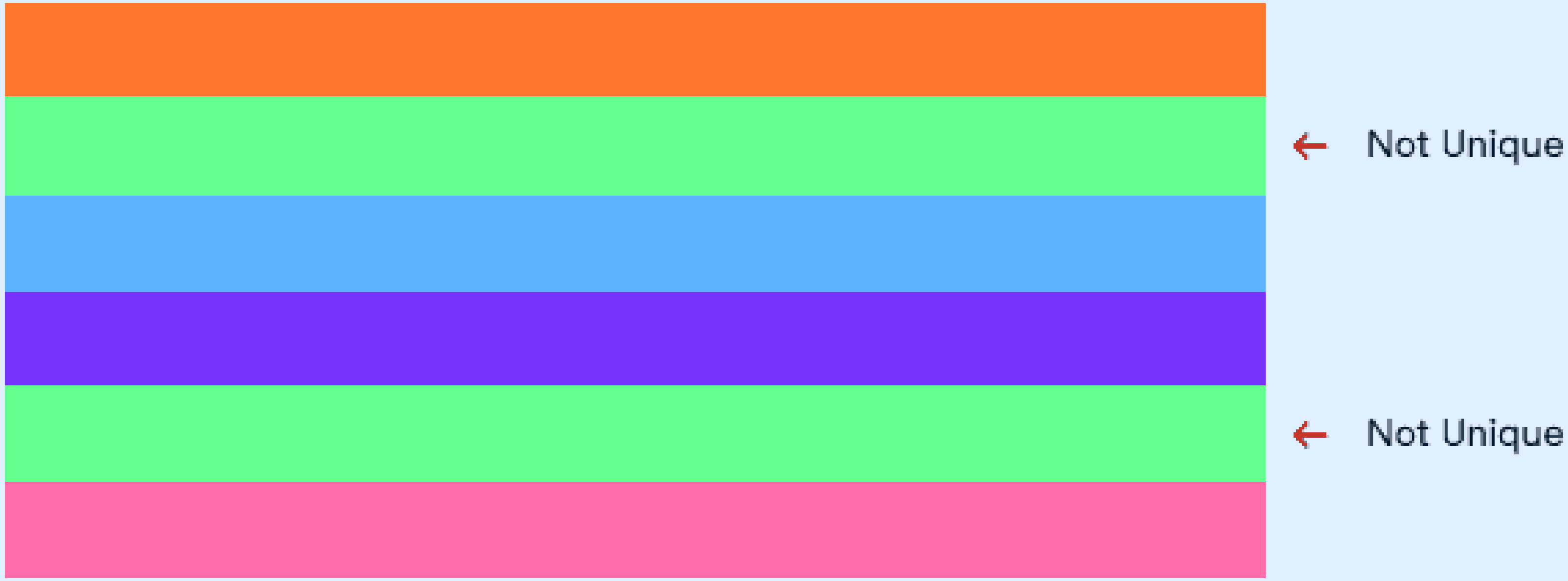


CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

Data quality

- **Uniqueness**

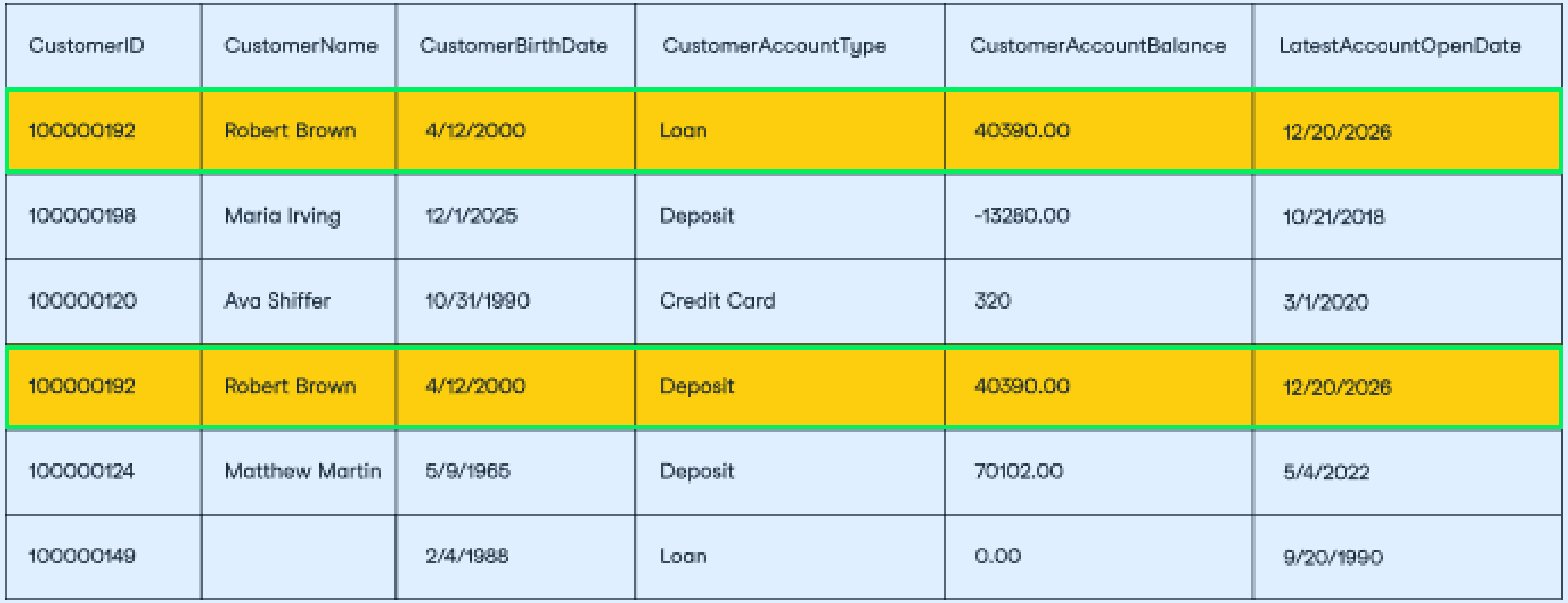
Uniqueness measures the degree to which the records in a dataset are not duplicated.



Data quality

- **Uniqueness Example**

All records must have a unique CustomerID and CustomerName.

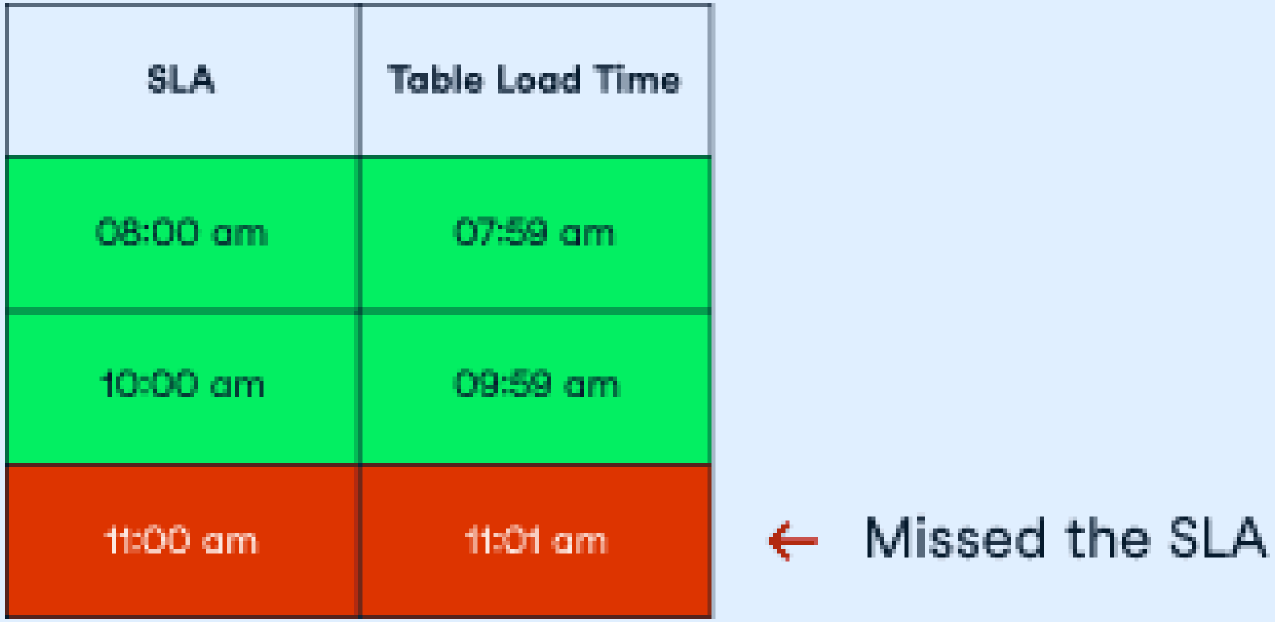


CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

Data quality

Timeliness

Timeliness is the degree to which a dataset is available when expected and depends on service level agreements being set up between technical and business resources.



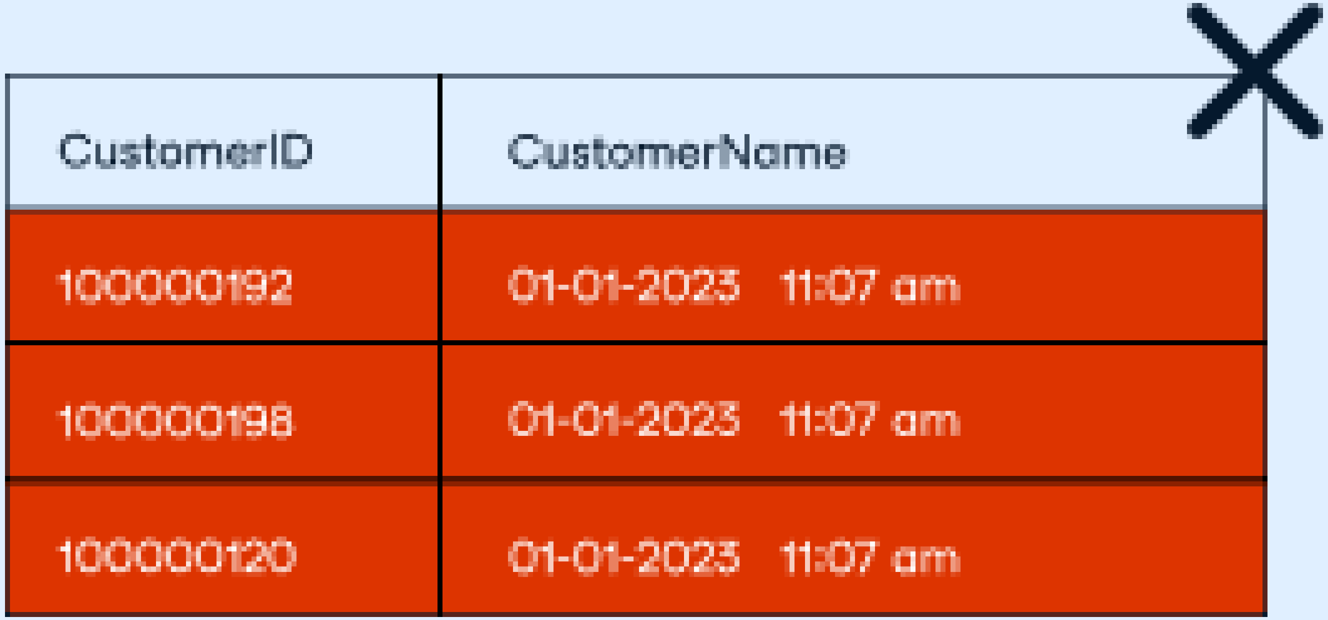
SLA	Table Load Time
08:00 am	07:59 am
10:00 am	09:59 am
11:00 am	11:01 am

← Missed the SLA

Data quality

- **Timeliness Example**

All records in the customer dataset must be loaded by the 9:00 am.

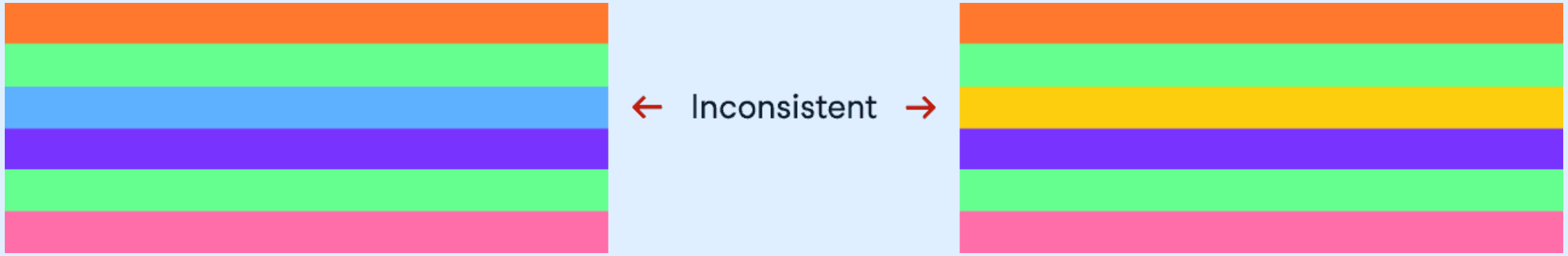


CustomerID	CustomerName
100000192	01-01-2023 11:07 am
100000198	01-01-2023 11:07 am
100000120	01-01-2023 11:07 am

Data quality

- **Consistency**

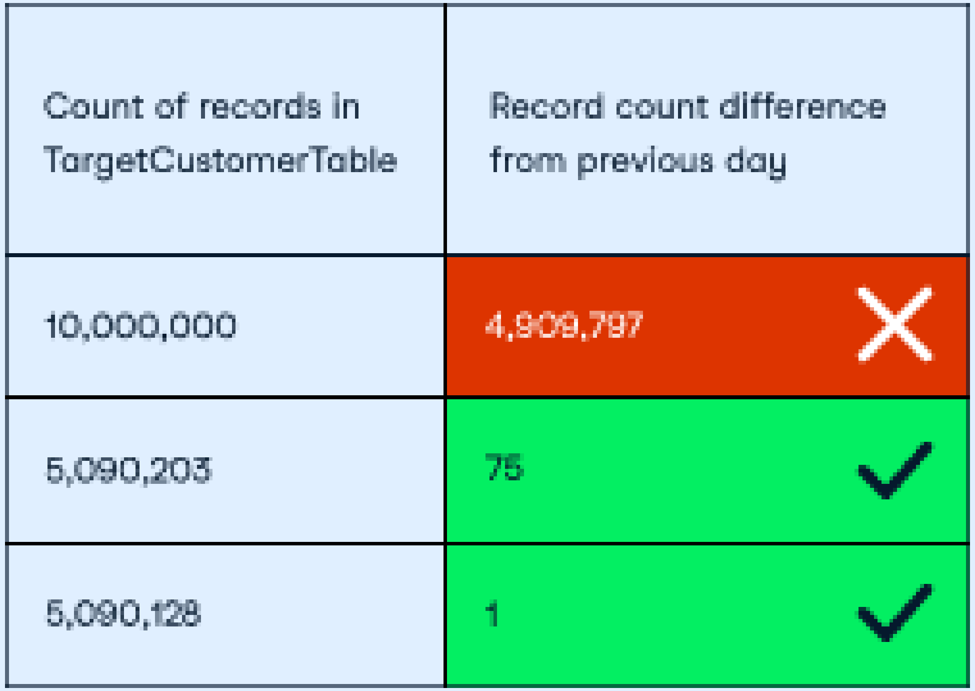
Consistency is a data quality dimension that measures the degree to which data is the same across all instances of the data. Consistency can be measured by setting a threshold for how much difference there can be between two datasets.



Data quality

- **Consistency Example**

The count of records loaded today must be within +/- 5% of the count of records loaded yesterday.

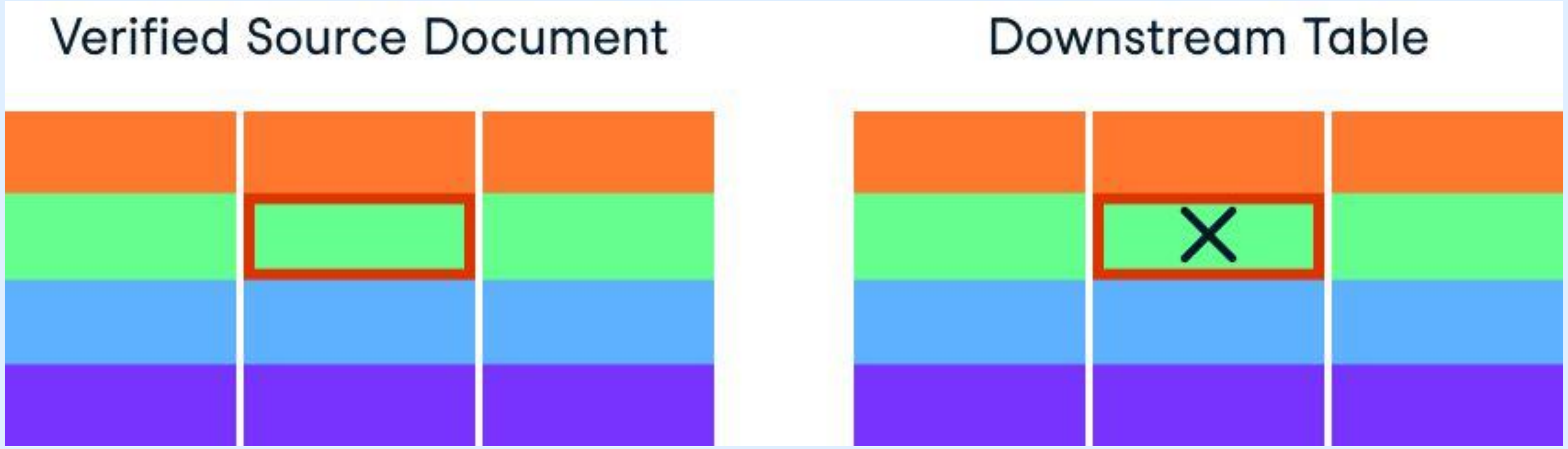


Count of records in TargetCustomerTable	Record count difference from previous day
10,000,000	4,909,797 ✗
5,090,203	75 ✓
5,090,128	1 ✓

Data quality

- **Accuracy**

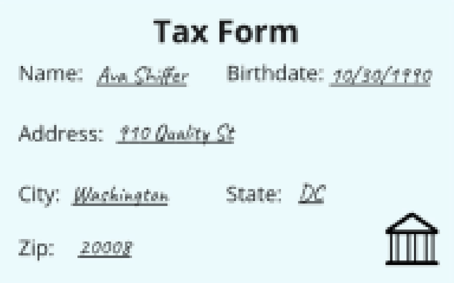
All records in the Customer Table must have accurate Customer Name, Customer Birthdate, and Customer Address fields when compared to the Tax Form.



Data quality

- **Accuracy Example**

All records in the Customer Table must have accurate Customer Name, Customer Birthdate, and Customer Address fields when compared to the Tax Form.



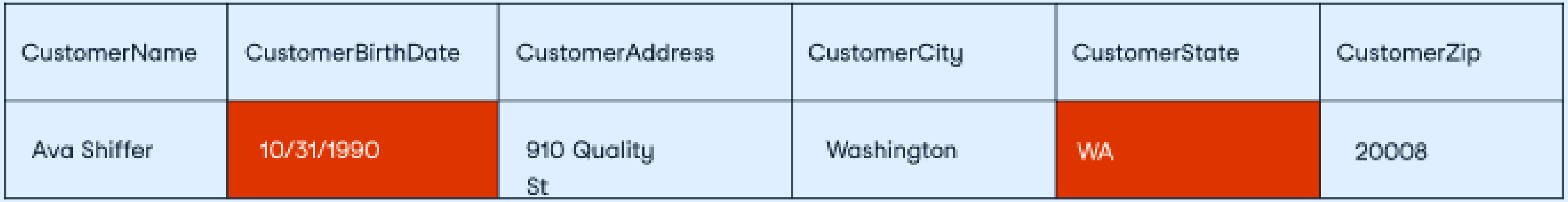
Tax Form

Name: Ava Shiffer Birthdate: 10/30/1990

Address: 910 Quality St

City: Washington State: DC

Zip: 20008



CustomerName	CustomerBirthDate	CustomerAddress	CustomerCity	CustomerState	CustomerZip
Ava Shiffer	10/31/1990	910 Quality St	Washington	WA	20008

Link to source: https://www.datacamp.com/cheat-sheet/data-quality-dimensions-cheat-sheet

Data cleaning

Data cleaning (sometimes also known as **data cleansing** or **data wrangling**) is an important early step in the data analytics steps. This crucial exercise, which involves preparing and validating data, usually takes place before your core analysis. Data cleaning is not just a case of removing erroneous data, although that’s often part of it. The majority of work goes into detecting rogue data and (wherever possible) correcting it.

Data cleaning

- The dirty data includes things like **incomplete**, **inaccurate**, **irrelevant**, **corrupt** or **incorrectly formatted data**. The process also involves **deduplicating**, or ‘**deduping**’. This effectively means merging or removing identical data points.

- Note: Data cleaning is time-consuming with great importance comes great time investment. Data analysts spend anywhere from 60-80% of their time cleaning data.

More videos and tutorials

- https://www.youtube.com/watch?v=kCP-H8VRDCw
- https://www.youtube.com/watch?v=_jmiEGZ6PIY
- https://www.youtube.com/watch?v=5HcDJ8e9NwY
- https://github.com/AlexTheAnalyst/Excel-Tutorial/blob/main/Data%20Cleaning%20Excel%20Tutorial.xlsx