

Preparing & Analyzing Data

1.1: A Brief History of Data Analytics

Learning Goals

- **Goal:** Understand the importance of data analysis in today's world.
- **Key Insight:** Data analysis impacts every industry, from business to healthcare and beyond.

2. Introduction

- **What you've done:** You've already learned to analyze and visualize data in Excel.
- **What's next:** In this new course, you'll dive deeper by using advanced tools like **SQL**, **Python**, and **Tableau**.
- **Project Alert:** You'll work on real-world projects, like helping a medical agency prepare for an upcoming season by figuring out what resources they'll need.

3. A Chronology of Data Analysis

- **Pre-Computer Era:**
 - **20,000 B.C.:** Simple tally sticks to record trade.
 - **100 A.D.:** **Antikythera mechanism** (analog computer to predict eclipses).
 - **1636:** John Gaunt used mortality records for early demography.
- **Key Invention:** **Herman Hollerith** created a **tabulating machine** in 1890 to process census data faster. Data = useful!
- **Computer Era:**
 - Computers allowed for faster, more complex analysis.
 - **Relational databases (SQL)** in the 1970s revolutionized data storage and access.

4. Data Analysis Applications

- **Business & Corporate Use:**
 - Analysts improve profits, develop products, understand customers, etc.
 - **Cool Fact:** The movie "Moneyball" is based on the Oakland A's use of data (sabermetrics) to build a competitive baseball team.
- **Government/Non-Profits:**

- Analysts help plan infrastructure and improve education, healthcare, etc.
- **Obama's 2012 Campaign** used data to tailor messaging to voters and increase donations.

5. Data-Driven Decision Making

- **What it is:** Making decisions based on actual data instead of gut feelings.
- **Example:** A city government prioritizes road repairs based on traffic data and road conditions, not just geography.
- **Data-Centric Organizations:**
 - These organizations treat data as an asset, shared across departments for smarter decisions.
 - **Intel's CEO** said data is “the new oil,” showing its value in today's world.

6. Technical Toolkit of a Data Analyst

- **Tools you'll use:**
 1. **Excel:** You already know this one! You'll continue using it for your first project.
 2. **Tableau:** A data visualization tool that makes your insights pop with dynamic dashboards.
 3. **SQL:** A language for managing large datasets in relational databases.
 4. **Python:** A flexible, powerful programming language for cleaning and analyzing data.

7. Spotlight on AI

- **AI & Data Analysis:**
 - Tools like Microsoft's **AI-aided Formula Editor** help streamline your data analysis by suggesting formulas, finding errors, and even recommending visualizations.
 - **How it helps:** AI reduces errors and saves time, allowing you to focus on insights rather than manual tasks.

8. Summary

- **You covered:**
 - How data analysis evolved from tally sticks to AI-powered tools.
 - Applications of data analysis across different fields, like healthcare, sports, and business.

- An intro to tools like **SQL**, **Tableau**, and **Python**.
- **Next Step:** Start your first project and dive into **Achievement 1**, focusing on forming research hypotheses and integrating datasets.

Quiz: 1.1: A Brief History of Data Analytics

1. How does the goal of data analysis differ between the corporate world and government organizations?
 - a. Increasing profits is typically not a goal in governmental data analysis.
 2. Data has existed for centuries, dating back to ancient civilizations. Which of the following data collection methods came first?
 - a. The tally stick
 3. Why was data analysis, such as with the census in the United States, significantly more labor-intensive prior to the 20th century?
 - a. A lack of machines to tabulate data
 4. In a data-driven approach, how are business decisions made?
 - a. Decisions are based on the analysis of evidence.
 5. Why was the advent of the internet so important for data analysis?
 - a. It increased the range of data that could be analyzed.
-

1.2 Starting with Requirements

Learning Goals

- **Goal:** Learn how to turn business requirements into data questions that guide your analysis.
- **Key Insight:** Business requirements are the foundation for what questions you'll ask as a data analyst.

2. Introduction

- **What's this about?:** You'll explore the types of business requirements that inform data projects and how to break them down into questions.
- **Why it matters:** Analysts often work with unclear or incomplete requirements, so knowing how to clarify and translate them is crucial to making impactful analyses.

3. Understanding Business Requirements

- **What are they?:** Business requirements are the starting point of any project. They outline the **goals, objectives**, and **outcomes** of a project.
- **Real-world challenge:** You won't always get clear instructions. Sometimes, you have to figure out what's needed by asking the right questions.
- **Best-case scenario:** You receive a **Business Requirements Document (BRD)**, which acts as a blueprint for the project. If not, you might need to create your own or clarify vague directions.

4. Business Requirements Breakdown

- **Main Sections of a BRD:**
 1. **Project Overview:** The snapshot of the project that explains its purpose (why), objectives (what), and scope (how).
 - **Example:** A school district wants to reduce dropout rates by 5%.
 2. **Stakeholders:** People involved in or impacted by the project.
 - **Example:** Superintendents, principals, teachers, parents.
 3. **Success Factors:** What defines success for the project.
 - **Example:** Dropout rates decrease by 5%, project finishes within a year.
 4. **Assumptions & Constraints:** Assumptions are things believed to be true but not confirmed. Constraints are the limitations, like budget or time.
 - **Example:** Assuming the school has some control over dropout rates is an assumption. The constraint is limited staff time.
 5. **Requirements:** Steps or deliverables needed to achieve the project goals.
 - **Example:** Analyze dropout data and suggest an actionable solution.
 6. **Glossary of Terms:** Defines important terms used in the BRD to avoid confusion.
 - **Example:** "Dropout" means leaving school without earning credits.

5. Translating Business Requirements Into Questions

- **Why ask questions?:** Questions help you focus your analysis on the right goals and dig deeper into the data.
- **Curiosity is key:** Good analysts are always asking questions—it's how you clarify and explore business needs.

6. Types of Questions

1. **Clarifying Questions:** Clear up any confusion about the project.
 - **Example:** “Why are students dropping out? Which students drop out?”
2. **Adjoining Questions:** Look at related areas that might affect the project.
 - **Example:** “How do dropout rates compare to other schools?”
3. **Funneling Questions:** Start broad, then narrow down into specifics.
 - **Example:** “Do boys and girls drop out for the same reasons?”
4. **Elevating Questions:** Take a high-level, big-picture view.
 - **Example:** “Why is reducing dropout rates important now?”

7. Privacy & Ethics

- **Privacy Considerations:** Ensure that any personal data you use is compliant with laws, such as education or medical privacy regulations.
 - **Example:** Do we need consent to use student data?
- **Ethical Considerations:** Be aware of how your data analysis might affect people. Don’t make unfair assumptions or target groups based on predictions.
 - **Example:** What ethical concerns are there in identifying at-risk students?

8. Writing Good Questions

- **Be specific:** Make sure your questions are clear and precise to avoid confusion.
 - **Example:** “When do students drop out?” → “In which months do most students drop out?”
- **Domain knowledge matters:** Understanding the subject you’re analyzing helps you ask better questions. If you don’t know, ask someone who does.
- **Flip the question:** Reframe your questions to open new insights.
 - **Example:** “Why do students drop out?” → “Why are so many students graduating?”

9. Summary

- **What you learned:** Business requirements guide your analysis, but you have to translate them into clear data questions to stay on track.
- **Next step:** Start working with a business requirements document and create questions that guide your project analysis.

Quiz: 1.2: Starting with Requirements

1. Kelly, a data analyst, recently transferred from her company's sales department to marketing. She's been tasked with measuring the performance of digital ads, an area with which she's unfamiliar. How can she effectively address her lack of domain knowledge in this new business area?
 - a. Shadow coworkers in the marketing department to gain domain knowledge
 2. What question can an analyst ask to make sure data privacy is considered during an analysis?
 - a. Does the analysis adhere to applicable data laws?
 3. Which action can an analyst take to make sure their questions aren't biased or reliant on unproven assumptions?
 - a. Challenge the premise of questions.
 4. What's the official term for a document that outlines crucial project information and lists stakeholders for a project?
 - a. Business requirements document
 5. Which of the following questions can an analyst ask to identify a complete list of stakeholders in a project?
 - a. Who would be interested in, involved in, or impacted by the project?
-

1.3 Designing a Data Research Project

Learning Goals

- **Goal:** Design a data research project based on business requirements and create a research hypothesis.
- **Key Insight:** You'll build a project plan and form a hypothesis to guide your analysis.

2. Introduction

- **What's this about?:** You've written questions to guide your analysis based on business requirements. Now, you'll learn how to plan a project and create a hypothesis.
- **Why it matters:** Project management and hypotheses are critical for keeping your analysis on track and ensuring it addresses the business requirements.

3. Planning Your Project

- **Why plan?:** A solid project plan saves time, helps manage expectations, and ensures success.
- **Project management basics:**
 1. **Stakeholder Communication:** Keep everyone updated through regular meetings, calls, or emails.
 2. **Schedule & Milestones:** Break the project into clear tasks with deadlines.
 3. **Project Deliverables:** Define what you'll deliver (reports, presentations, dashboards).
 4. **Audience Definition:** Tailor your presentation to the audience's understanding of data.

4. Stakeholder Communication

- **Keep stakeholders informed:** Update key players at milestones (25%, 50%, and 75% completion).
 - Use **meetings** for lots of feedback, **calls** for quick updates, and **emails** for info-only updates.
- **Emergency plans:** If delays happen, send an email followed by a meeting or call.

5. Schedule and Milestones

- **Why milestones matter:** They help track your progress and ensure you stay on schedule.
 - **Example:** Complete dropout rate analysis by week 2, compile research by week 4, and hold a meeting to choose an intervention by week 8.

6. Project Deliverables

- **Types of deliverables:**
 1. **Verbal presentation:** For small projects, just explain your analysis verbally.
 2. **Meeting presentation:** Use slides (PowerPoint) to visually show your analysis.
 3. **Written report:** Detailed report for research-heavy projects.
 4. **Dashboard:** Visual tool (e.g., Tableau) that stakeholders can use to track data.
- **Example:** For the school district dropout project, you'll present results with slides to stakeholders.

7. Audience Definition

- **Know your audience:** Tailor your presentation based on their data knowledge.
 - **Example:** For parents, teachers, and school officials, explain graduation rates and keep things simple.

8. Forming a Hypothesis

- **What is a hypothesis?:** It's a **testable prediction** about the relationship between two variables (e.g., drug x lowers blood pressure).
 - **Structure:** "If [independent variable], then [dependent variable]."
- **Example:** For the dropout project:
 - **Motivation:** Students drop out due to family financial issues.
 - **Hypothesis:** "If students don't have to work to help their families, they won't drop out."

9. Creating a Data Wishlist

- **Why it's important:** Your wishlist outlines the data needed to test your hypothesis.
 - **Example:**
 - For the dropout hypothesis: You'll need data on students' work status, dropout rates, and family finances.

10. Summary

- **What you learned:** You've designed a project plan and formed a hypothesis based on business requirements. Now, you're ready to start sourcing data in the next step!

Quiz: 1.3: Designing a Data Research Project

1. Rick is tasked with developing a project that can track and monitor sales performance as it happens. The goal is to create one deliverable that will update dynamically as new data becomes available. What type of deliverable should he use?
 - a. Tableau dashboard
2. Why is it important to define a project's audience when planning a data research project?
 - a. To determine an appropriate output style for the final deliverable
3. At a minimum, what should a schedule for an analysis project include?
 - a. A final completion date
4. Which of the following is an example of a hypothesis?
 - a. If customers are happier, then they'll purchase more products

5. Which one of the following steps is typically included in a project management plan?
 - a. Crafting a schedule and milestones
-

1.4 Sourcing the Right Data

Learning Goals:

- The main goal here is to **identify relevant data** for your project while being able to distinguish **irrelevant data** that won't help your analysis.

Introduction:

- After you've interpreted the **requirements of your project** and developed a **hypothesis** (which is an educated guess you want to test), it's time to **find the right data** to back it up.
- In the previous exercise, you created a **wishlist** of the data you *think* you need based on your hypothesis.
- Now, you need to **source** (find) the data. But **where do you find it**? That's what this exercise focuses on.

Why This is Important:

- Finding the right data is crucial because if you **don't have enough of the correct data**, you can't properly support your hypothesis, and your project might fall apart.
- Many times, you may need to **combine data from different places** to get everything you need.

Data Sourcing:

- **Data sourcing** is just a fancy way of saying, "How do I find the data I need?"
- Most of the time, you're not going to be collecting new data. Instead, you'll need to find **existing data**.
 - **Pro Tip:** Talk to different departments in an organization or **other analysts** to see what data they use. You can also do an online search to see what data is commonly used for similar projects.

Two Types of Data Sources:

1. **Internal Data:** Data that's created and collected **inside the organization**. This can include things like sales data, customer behavior, or marketing data.
 - **Example:** Netflix tracks everything its customers do—what they watch, for how long, and when. This data is **internal** because Netflix owns and collects it.
2. **External Data:** Data that comes from **outside** the organization. Often, this type of data is used to **compare** performance across industries, regions, or groups.
 - **Example:** A hospital might use **benchmarking data** to compare how much they charge for services compared to other hospitals.
 - **APIs** are a common way to get external data. **Google Maps** or **Twitter data** is accessed using APIs. But **you won't usually be the one fetching the data—developers handle that part.**

Why This is Important:

- You'll likely use both **internal and external data** in your projects, depending on the information you need. Internal data gives insight into how an organization runs, while external data allows for **comparisons**.

Internal Data Sources:

- **Internal data** is often specific to an organization and is **not publicly available**. This includes information about a company's **customers, sales, and operations**.
 - **Example:** Netflix tracks its customer behavior internally—everything from what they watch to when and how long they watch. This is internal data because it's private to Netflix and helps them improve their service.
- **Survey Data:** Surveys are another example of internal data. Even though surveys gather information from external people (like customers), the **survey results** are stored internally by the organization.

Why This is Important:

- Internal data helps organizations understand their **day-to-day operations** and improve decision-making, such as deciding what **new features** to add or which marketing efforts are working.

External Data Sources:

- **External data** comes from outside the organization and is used for **benchmarking** or comparing against industry standards.

- **Benchmarking:** You can compare data across multiple companies or regions. For example, an HR department might compare salaries across different companies to make sure their employees are **fairly compensated**.
- Governments provide **census data** that organizations use to make decisions.
 - **Example:** A company can use population data to figure out where their next store should be located.
- **APIs** are often used to access external data. For example, Google provides location data through its **Google Maps API**, and Twitter provides user activity data through its API.

Why This is Important:

- **External data** allows organizations to see how they compare to others and to use large, reliable data sets (like census data) to guide decisions.

Data Collection Methods:

1. Administrative Data:

- **Administrative data** refers to the data an organization needs to run its daily operations.
 - **Example:** An insurance company tracks **enrollments and claims** to keep their business running smoothly.
- Governments also collect administrative data, such as **birth records, death certificates, and census data**.

2. Usage Data:

- **Usage data** shows how people interact with a product or service.
 - **Example:** Netflix tracks **what you click on**, what you watch, and how long you watch it.
- Businesses use **usage data** to improve their products. For example, if Netflix sees that customers rarely click on a specific feature, they may decide to remove or change it.

3. Interviews and Surveys:

- **Interviews and surveys** are used to collect data about customer **opinions** or **experiences** with a product.
 - **Example:** A company might send out surveys asking customers what they think about their products and how they could improve.

Why This is Important:

- Different types of data collection help you understand various aspects of a business, such as **customer behavior** or **operational efficiency**. It's important to know **how the data was collected** to assess its accuracy and **reliability**.

Data Privacy and Ethics:

- **Data privacy** refers to protecting people's personal information. Sometimes, data is **suppressed** to protect individuals' identities.
 - **Example:** If only one person in a census category fits certain criteria, their data might be left out to **protect their privacy**.

Describing a Data Set:

When working with data, it's important to be able to describe it in terms of:

1. **Source:** Where does the data come from? Is it **internal** or **external**?
2. **Collection Method:** How was the data collected? Was it **automated**, **manual**, or through **surveys**?
3. **Contents:** What's in the data? Instead of listing hundreds of variables, **summarize** the most important ones.
 - **Example:** In student data, instead of listing everything (e.g., name, height, weight, grades), just say the dataset contains **student demographics** and **performance data**.

Why This is Important:

- Summarizing data helps you (and others) **understand it quickly**. Keeping a **cheat sheet** of data contents can save you time when you're working on a project with **multiple data sets**.

Determining Data Relevance:

- Not all data is relevant to your project. You need to focus on data that **directly supports your hypothesis**.
 - **Example:** In a project about **student dropout rates**, you'd want **graduation data** and **rental assistance data** (because it impacts housing stability).
 - You **wouldn't** need data about **student performance** or **satisfaction** because they aren't directly tied to your hypothesis.

Why This is Important:

- Using **irrelevant data** can distract from your analysis and **waste time**. Focus on the data that will **help you answer your hypothesis** and meet your **project objectives**.

Student Project: Influenza Deaths Data Set:

For your **student project**, you'll use data from the **Centers for Disease Control (CDC)**. Here's how to evaluate its relevance:

1. Data Source:

- This data comes from the **CDC**, so it's an **external** source, and you can trust its reliability because it's **government-collected**.

2. Data Collection:

- It's **administrative data**, collected from **death certificates** where doctors record the **cause of death** (e.g., influenza or pneumonia).
- One thing to watch out for is that **only one cause of death** is listed on the death certificate, so there could be some **discrepancies**, especially in vulnerable populations like people with AIDS.

3. Data Contents:

- This data includes **monthly death counts** for influenza in the U.S. from **2009 to 2017**, broken down by **state** and **age**.

4. Data Relevance:

- This data is relevant because it helps track **when and where influenza deaths** occur, which can be used to **plan resources**.
- You can also **prioritize vulnerable populations** by looking at the **age breakdown** to see who's most affected.

Why This is Important:

- The **CDC influenza data** is a key part of your project because it helps you understand **where and when to allocate resources** based on past influenza seasons.

Summary:

- In this exercise, you learned how to **assess data** in terms of its **source, collection method, and contents** to determine its **relevance** to your project.
- Only use data that's **relevant** to your hypothesis. **Ignore** data that **doesn't help** answer your project questions.

- Apply these principles to the rest of the data sets in your project to confirm which ones you.

Quiz: 1.4: Sourcing the Right Data

1. Which of the following is an example of internal data?
 - a. Organization-specific data about a product, service, or operations
 2. Kate is embarking on a data research project for which she already has a project objective and hypothesis. Next, she needs to find some relevant data. What can she examine to determine whether or not a data set is relevant to her project?
 - a. The data's source, content, and collection method
 3. Which of the following is an example of usage data?
 - a. Data gathered by a car on a customer's driving habits
 4. Data that benchmarks statistics about a specific population or industry across multiple organizations is which kind of data?
 - a. External data
 5. What role does a data analyst play in the collection, management, and usage of data within an organization?
 - a. An analyst uses data collected by others in the organization.
-

1.5: Data Profiling & Integrity

Learning Goals:

1. Profile and clean a data set to prepare it for analysis.
2. Improve the integrity (reliability, consistency, and accuracy) of data.

Introduction:


In the previous exercise, you focused on sourcing relevant data. Now it's time to **clean** and **profile** your data.

Goal: Ensure your data is high-quality (accurate, consistent, and reliable) before analysis. Think of it as cleaning your room—organized data is easier to work with!

- A good **data profile** provides a summary of:






Variables (data columns) and their types

 Integrity issues (errors or inconsistencies)

 Statistical summaries (e.g., averages, counts)

What's Data Profiling?

It's like a **“check-up” for your data** to make sure everything looks right. You'll:

-  **Identify what kind of data** you have (e.g., numbers, dates, or text).
-  **Spot errors or inconsistencies** (like typos or weird values).
-  **Summarize the content** to know exactly what's in your dataset.

Getting to Know Your Data Types, Data Terminology and Structure:

✨ Understanding Data Types

1. Structured Data

- Neatly organized into **rows and columns** (like an Excel table or database).
- Easy to analyze using tools like **Excel or SQL**.

Examples of Structured Data:

- Employee Records: Name, Salary, Date of Birth.
- Survey Results: Age group (18-24, 25-34), Satisfaction rating (1-5)

2. Unstructured Data

- No fixed format - more like **random notes** on paper.
- Harder to analyze - requires special tools like **Natural Language Processing (NLP)** for text).

Examples of Unstructured Data:

- Customer Reviews: “The product was amazing, but delivery was slow.”
- Emails or Social Media Posts: Random text without a pattern.


Semi-Structured Data:

A mix of both **structured and unstructured** data.

- Example:** JSON or XML files (organized but still messy).

Qualitative Data vs. Quantitative Data:

1. Qualitative Data (Descriptive)

- Describes **characteristics or categories** (non-numerical).
- Types:**
 -  **Binary:** Two values (e.g., Yes/No, True/False).

- **Nominal:** No order to categories (e.g., colors like Red and blue).
- ▼ **Ordinal:** Categories with an order (e.g., Gold, Silver, Bronze medals).

2. Quantitative Data (Numerical)

- Can be **measured or counted** (e.g., income, height).
- **Types:**
 - **Discrete:** Whole numbers (e.g., number of siblings).
 - **Continuous:** Can be divided further (e.g., time or weight).

Time-Variant vs. Time-Invariant Variables

- **Time-Variant:** Changes over time (e.g., weight, income).
- **Time-Invariant:** Stays the same (e.g., birthdate, eye color).






Example:

Your driver's license shows time-invariant variables (like height) and time-variant ones (like hair color).



How are Variables Recorded?

- **Variables = Columns** in a dataset (e.g., Purchase Date, Platform, or Item).
- **Records (or Observations) = Rows** representing a single data entry (e.g., one customer purchase).

Excel Functions to Identify Data Issues



-  **MIN():** Find the **smallest value** in a range (e.g., to detect outliers).
Example: =MIN(A1:A50)
-  **MAX():** Find the **largest value** (e.g., to spot unusual maximums).
Example: =MAX(A1:A50)
-  **AVERAGE():** Calculate the **mean** (useful for spotting unreasonable averages).
Example: =AVERAGE(A1:A50)
-  **COUNT():** Ensure there are no **missing values** in a dataset.
Example: =COUNT(A1:A50)
-  **SUM():** Verify if **totals make sense** for key metrics.
Example: =SUM(A1:A50)

Nominal vs. Ordinal Data:

- **Nominal Data** 
 - No inherent order (e.g., Colors: Red, Green, Blue).
 - **Example from 1.5:** Platform (Phone vs. Website).
- **Ordinal Data** 
 - Has an order, but differences between values are not meaningful (e.g., Star Ratings: 1 to 5 stars).

- **Example:** Olympic medals: Bronze < Silver < Gold.

Discrete vs. Continuous Data in Quantitative Analysis:

-  **Discrete Data:** Whole numbers only (e.g., Number of items sold).
 - **Example:** Customer purchases (3 cameras, 2 lenses).
-  **Continuous Data:** Can have decimals (e.g., weight, time).
 - **Example:** Product weight (1.5 kg), Time taken (2.75 hours).

Answer to Question:

- **Discrete and Continuous Data** describe **Quantitative Data** (not qualitative).

Data Integrity:

What is Data Integrity?

- Ensures that data is **accurate** (correct) and **consistent** (same format).
 - **Accurate data:** Correct and free of errors.
 - **Consistent data:** Uniform format (e.g., all dates follow YYYY-MM-DD).

Why Data Integrity Matters:

- **Example:** If student grades are listed wrong, your analysis will **give incorrect results**.

What Causes Poor Data Integrity?

1. Human Errors:

- **Manual data entry** can lead to typos, missing data, or inconsistent formatting.
- **Example:** Entering #203 instead of just 203 in an address field may confuse the system.

2. Computer Issues:

- Systems may store **different currencies** under the same variable (e.g., euros and pounds).
- Changing standards (like new **industry codes** (NAICS)) may change over time, leading to inconsistencies.

3. Corrupt Data:

- Errors in **storing or transmitting data** can make it unreadable or incomplete.

How to Check Your Data's Integrity:

- **Accuracy:** Is the data correct? (e.g., no one should have a score of 970 out of 100!)
- **Consistency:** Is the format the same everywhere? (e.g., all dates as YYYY-MM-DD).

Quick Tips for Spotting Issues:

- Use Excel's **=MIN()** and **=MAX()** to find the smallest and largest values.
- Use **=AVERAGE()** to calculate the mean and see if anything looks out of place.
- Check for outliers (values that seem **too high or too low**).

Resolving Data Integrity Issues:

Accuracy Issues:

- **Example:** If a student has a grade of 970%, it's clearly wrong.

How to Fix:

- Use **min/max values** to find errors.
- **Cross-reference** with other data sources.
 - **Excel Tips:**
 - Use **=MIN(A1:A50)** or **=MAX(A1:A50)** to find the smallest and largest values. (aka to quickly spot outliers)
 - Use **=AVERAGE(A1:A50)** to calculate the mean (average).

Consistency Issues:

- **Example:** “Mobile” and “Phone” are listed separately when they mean the same thing.
- **Use:** Pivot Tables to find inconsistencies.
- In Excel, a **frequency table** is essentially a **pivot table** where you use the **count option** to summarize how often each value or category appears. It's one of the quickest ways to **identify patterns, detect errors, and summarize data!**

Excel Pivot Table Steps:

- Insert → Pivot Table → Drag the column to **“Rows”** and **“Values”** to review **duplicates** or inconsistencies.

Creating a Data Profile:

Data profiling helps **summarize and identify problems** in your dataset before analysis.

- **Steps:**
 1. **Identify variables** and data types (e.g., qualitative or quantitative).
 2. **Check for errors** or inconsistencies.

3. **Summarize the data** using statistics like count, mean, and mode.

Statistical Measures to Use:

- **Median:** The middle value in a sorted dataset.
 - Excel formula: `=MEDIAN(A1:A50)`
- **Mode:** The most frequent value in a dataset.
 - Excel formula: `=MODE(A1:A50)`
- **Record count:** Total number of rows or entries.
- **Value count:** Total number of unique values in a column.

How to Clean Data:

1. **Fix Errors:**
 - Example: If you see a date like “208-01-01,” change it to “2018-01-01.”
 - Use tools like **Pivot Tables** to spot inconsistencies (e.g., multiple names for the same thing).
2. **Remove or Handle Missing Data:**
 - If you can’t correct a wrong value, **delete it or mark it as missing**.
3. **Create a Data Profile:**

A **data profile** is a summary of your dataset. It includes things like:

 - Variables and their types (e.g., “Platform” is a categorical variable).
 - Data counts (how many entries or unique values you have).
 - Any issues you fixed (e.g., changing “Mobile” to “Phone”).

Example Data Profile: Photography Sales Data

Consider this sample data set about **photography equipment sales**:

Customer	Purchase Date	Platform	Item	Purchase Price	Currency
123	2019-09-09	Website	Tripod	\$108	Euro
839	2019-03-04	Phone	Lens	\$534	Euro
478	2018-01-01	Phone	Flash	\$225	Euro
478	208-01-01	Mobile	Lens	\$888	Euro

Fixes Made:

- Changed “Mobile” to “Phone.”
- Fixed incorrect date: “208-01-01” → “2018-01-01.”

Summary:

- **6 Variables, 4 Records**
- **Qualitative Variables:** Platform, Item, Currency
- **Quantitative Variable:** Price (min = \$108, max = \$888, mean = \$439)

Profiling Notes:

- **Errors Found:**
 - Date Error: B5 "208" should be "2018."
 - Platform Issue: C5 "Mobile" should be changed to "Phone."
- **Summary of Variables:**
 - **Customer:** Nominal, time-invariant (ID numbers).
 - **Purchase Date:** Continuous, time-invariant.
 - **Platform:** Nominal, time-invariant (Phone, Website).
 - **Item:** Nominal (Tripod, Lens, Flash).
 - **Purchase Price:** Discrete, time-variant.
 - **Currency:** Nominal (Euro).
- **Summary Statistics:**
 - **Purchase Price:**
 - Min: \$108
 - Max: \$888
 - Mean: \$439
 - Mode: N/A (all values are unique)

Why Data Integrity Matters:

Good data = good analysis. If your data isn't **accurate and consistent**, you'll get **bad results** that could lead to poor decisions (like missing students who should graduate).

Why Data Profiling is Important:

- **Helps you understand** the structure and content of your data
- identifying errors or inconsistencies **before analysis**.
- **Finds errors** before they affect your results.
- **Document your changes** to ensure transparency and repeatability in your analysis.
- **Document everything** to make your work traceable.

Summary:

- **Data profiling** involves:
 1. **Classifying variables** (e.g., qualitative or quantitative).
 2. **Checking for data integrity issues** (accuracy and consistency).
 3. **Using statistics** (mean, mode, etc.) to summarize the data.
- High-quality data ensures your **analysis is accurate and reliable**.
- In your project, **create a data profile** to track variables, errors, and improvements.

Final Takeaways:

🔧 **Profile your data:** Know what's in your dataset and if anything looks odd.

🔧 **Clean your data:** Fix errors, fill in missing values, or remove bad data.

Classify Each Variable:

Below are some hints and steps to help you fill in the **"Data Profile of Raw Data"** tab in your template:

1. **Structured or Unstructured?**

- Since this data is organized into rows and columns with clear labels, it's likely **structured**.

2. **Qualitative or Quantitative?**

- **Qualitative:** Categories or names (e.g., County, State).
- **Quantitative:** Numbers representing counts or measurements (e.g., Total population, Vulnerable Population Rank).

3. **Nominal or Ordinal (for Qualitative data)?**

- **Nominal:** No order or ranking (e.g., County names).
- **Ordinal:** Ordered categories (e.g., Vulnerable Populations Rank).

4. **Discrete or Continuous (for Quantitative data)?**

- **Discrete:** Whole numbers (e.g., population counts).
- **Continuous:** Measurable data that can take finer increments (e.g., percentage, though not present here).

5. **Time-Variant ⌚ or Time-Invariant 🕒?**

- **Time-Variant:** Data that changes over time (e.g., Year, Total population).
- **Time-Invariant:** Data that stays constant (e.g., County, State).

Quiz: 1.5: Data Profiling & Integrity

1. Angie asks visitors to her recipe blog a one-question survey when they leave her site. The question asks visitors to write in an open text field how appetizing they feel her recipes are. What sort of data is Angie collecting?
 - a. Unstructured data

Because Angie is asking an open-ended question to which a website visitor could enter any answer, this is unstructured data. It is also qualitative data, as respondents are writing out their answers in words. Should Angie have asked visitors to rate her recipes using a star scale, with only 1-5 options, this would be structured data. Quantitative data refers to a quantity or statistic, such as a 1-5

scale, not an opinion. This data is also not ordinal as there is no order to opinions, such as largest to smallest.

2. Which Excel function might an analyst use to help identify accuracy issues within a data set?
 - a. MIN

An analyst might use the MIN function to identify the lowest value within a data set to identify if this value is lower than expected, or an outlier. The COUNT function will only calculate the number of records, while the SUM value shows only the total, neither of which will reveal if record is inaccurate. MEAN can be used to help identify inaccurate data; however, MEAN is not a valid Excel function, and the analyst should use AVERAGE instead.

3. Dante has examined his data set and has identified inconsistent data, specifically customer phone numbers, which lack an area code. How should Dante proceed?
 - a. Reach out to the person who collected the data to see if consistent data exists elsewhere.

Whenever an analyst identifies data that is inconsistent, inaccurate, or otherwise doesn't make sense, the first step is to see if this data can be procured from the original source. Dante should reach out to whomever collected the data to see if that person can resolve the issue.

4. Qualitative data can be narrowed into categories of nominal and ordinal data. What is the key difference between these categories of qualitative data?
 - a. Ordinal data has a set sequence by which it can be ranked.

The key difference between nominal and ordinal data is if the data can be ranked or ordered using a set sequence. Ordinal data can be ranked in size, quality, or other measurements, such as smallest to largest or most to least favorable. Nominal data has no order or set sequence. Pizza topping data is an example of nominal data: despite personal taste preferences, one topping is not assumed to be "better than" or "before" another. Whether data is structured or unstructured is irrelevant to determining nominal versus ordinal data.

5. A data profile may contain variables and data types, data integrity issues, changed records, and summary statistics. Why do analysts create data profiles?
 - a. To document any data cleaning efforts performed

A data profile is a tool an analyst can create for themselves to provide an overview of the data and what was done to it, including any data cleaning efforts. This record is important to make the process trackable and repeatable.

A data profile is typically not shared with clients and does not detail any connection to a hypothesis or data sources.

1.6: Data Quality Measures

Learning Goals 🎯:

- **Determine and improve data quality** from a given source.
- Ensure your data is **ready for analysis** by addressing quality issues.

Introduction 🙌:

In the **previous exercise (1.5: Data Profiling & Integrity)**, you worked on **accuracy and consistency** to improve the integrity of your data. Now, you'll explore **other data quality measures** such as completeness, uniqueness, and timeliness. These steps will ensure **your data sets are fully cleaned** before moving on to data integration.

What is Data Quality? 🏆

Data quality measures whether your **data is reliable and fit for use**. It ensures you can **trust** your analysis. Think of it like a health checkup! Instead of measuring temperature or pulse, you're checking:

- **Accuracy** ✓: Is the data correct?
- **Consistency** 🔄: Is the format uniform across all entries?
- **Completeness** ✖: Are any values missing?
- **Uniqueness** 🔑: Are there duplicate records?
- **Timeliness** ⌚: Is the data available fast enough for your needs?

💡 Data Quality Formula:

Data Quality = Data Integrity (Accuracy + Consistency) + Completeness + Uniqueness + Timeliness

Data Quality Measures 🔍:

1. Completeness ✖:

- **Do any variables have missing values?**
- Example: In a library's check-out records, are **all transactions included?**
- **How to find it:** Use **frequency tables** to spot missing data patterns (e.g., 0 values or placeholders like "Unknown").

2. Uniqueness 🔑:

- **Is the data distinct?**
- Example: In a library, **each item should only be checked out once** per transaction.
- **Duplicates skew counts**, making your analysis inaccurate. Use **pivot tables** to find and remove duplicates.

3. Timeliness 🕒:

- **How quickly is data available?**
- Example: If sales managers use Monday reports to assign work, but data is entered on Tuesday, it's **not timely enough**.
- **Check timestamps** to ensure your data is updated in time for analysis.

Why Data Quality Matters 🚨:

Poor data quality can lead to **inaccurate analysis** and **wrong business decisions**. It also wastes **time and effort** and erodes stakeholder trust.

● **Case Study: Apple Maps Fail** 🚗

In 2012, Apple Maps launched with errors: mislabeled towns, missing landmarks, and incorrect business listings. These **data quality issues** resulted in customer dissatisfaction and damaged Apple's reputation.

How to Address Missing Data 🧹:

1. **Do nothing** 🛑:

- If the data is **not critical**, leave it as-is but mark it properly (e.g., with "NA" or "Unknown").

2. **Remove records or variables** 🗑️:

- If **less than 5% of data is missing**, you can **delete rows or columns** without significantly impacting your analysis.

3. **Impute values** 🧠:

- **Educated guesses**: Use related data to fill gaps (e.g., infer gender from a name).
- **Average or Median**: For numerical data, use **=AVERAGE()** or **=MEDIAN()** to fill missing values.
- **Interpolation**: For time series data, use **linear trends** or **last observation carried forward (LOCF)**.

Educated Guess: Reasoning

- Since Wes has access to ZIP code data, an educated guess implies using external knowledge, such as a ZIP-to-county reference table, to accurately infer the missing counties.
- While it's not the perfect term for a systematic approach, "educated guess" is the closest match to what a data analyst would do: leveraging relevant data sources to make an informed decision.

Using Random Values 🎲:

! [Lottery balls in a tumbler]

A **less common method of imputation** is using **random values**. This method works only for **numeric data**.

- **How it works:**

- Generate a random number within a specific **range** based on your dataset.
- For example, when imputing age, use the **minimum** and **maximum** ages from the data to create a **random value** within that range.

Excel Formula to Generate Random Values:

Scss Copy code:

```
=RANDBETWEEN(low_number, high_number)
```

Example: To generate a random age between 0 and 100:

Scss Copy code:

```
=RANDBETWEEN(0, 100)
```

- **Guidelines for Using Random Values:**

- Only use **random values or central tendency measures** when the **missing data is $\leq 5\%$** of the dataset.
- If **more than 5%** of the data is missing, filling in values with random numbers may **alter the dataset's characteristics** (e.g., skew, outliers).

Handling Missing Value Example Methods 🛠️

Methods	Pros ✓	Cons ✗
<ul style="list-style-type: none">• Method 1: Remove missing data	<ul style="list-style-type: none">• Easiest method to apply	<ul style="list-style-type: none">• Doesn't work well when data size is small since you'll lose data
<ul style="list-style-type: none">• Method 2: Replace missing data with average value	<ul style="list-style-type: none">• Easy to implement & understand	<ul style="list-style-type: none">• Requires checking that data is normally distributed. If not, use median instead of average
<ul style="list-style-type: none">• Method 3: Replace missing data with random value	<ul style="list-style-type: none">• Simple and easy to apply	<ul style="list-style-type: none">• Requires justification for using random values to stakeholders
<ul style="list-style-type: none">• Method 4: Replace missing data with interpolated value	<ul style="list-style-type: none">• Advanced method that retains trends in the data	<ul style="list-style-type: none">• Can be difficult to implement and explain to stakeholders

When dealing with missing data, the method you choose will depend on the size of the data set, the nature of the data, and the goals of your analysis. Each method comes with its own benefits and challenges, and it's essential to document and justify your choice, especially for stakeholders.

Measuring Central Tendencies

Central tendency measures help summarize data and fill in missing values effectively. Here's a breakdown of the key measures:

1. Mean (Average)

- **Formula:** (Sum of values) / (Count of values)

Excel Formula:

scss

Copy code

```
=AVERAGE(A1:A50)
```

-
- **Use Case:** If data is **symmetrical**, you can use the **mean** to fill in missing values.
- **Example:** In a dataset of ages, replace missing values with the average age.

2. Median (Middle Value)

- **Definition:** The **middle value** of a sorted dataset.

Excel Formula:

scss

Copy code

```
=MEDIAN(A1:A50)
```

-
- **Use Case:** If the data is **skewed** or has outliers, use the **median**.
- **Example:** For income data with extreme values, the **median** is more representative than the mean.

3. Mode (Most Frequent Value)

- **Definition:** The **most common value** in the dataset.

Excel Formula:

scss



Copy code

```
=MODE(B1:B50)
```



-

- **Use Case:** Use the **mode** for categorical data or when identifying the most frequent value.
- **Example:** Fill missing survey responses with the most common answer.

How to Handle Duplicates

1. **Find duplicates using pivot tables** :
 - Use **Transaction ID - Item** as your data grain to find unique records.
 - Identify duplicate rows by counting occurrences.
2. **Remove duplicates** :
 - Use **Excel's Remove Duplicates** function under the **Data tab**.
 - If needed, use **VLOOKUP()** to recode similar entries (e.g., “Jaxson Bailey” vs. “Sir Jaxson Bailey”).

How to Evaluate Timeliness

1. **Look for timestamps** :
 - Check the **most recent date** in the dataset to ensure the data is up-to-date.
2. **Ask the data owner** :
 - If you don't see a timestamp, **consult with the data provider** to understand when the data was last updated.

In Practice: Photography Sales Data Example

Here's an example of how to apply these data quality measures:

Customer	Purchase Date	Platform	Item	Price	Currency
123	2019-09-09	Website	Tripod	\$108	Euro
478	2018-01-01	Mobile	Lens	\$888	Euro
555	2019-10-10	Website		\$1050	Euro

Profiling Notes

- **Completeness:**
 - The last row has a missing item. You checked the catalog and filled it in with **"Camera."**
- **Uniqueness:**
 - Customer **478's flash purchase** had a duplicate entry. You removed the duplicate and standardized “Mobile” to “Phone.”
- **Timeliness:**

- No indication of timeliness provided. Ask the **data owner** for more information.

Changed Records

1. **Removed duplicate flash purchase** for customer 478.
2. **Recoded "Mobile" to "Phone"** for consistency.
3. **Filled in the missing item** with "Camera" from product catalog research.

Final Data Table

Customer	Purchase Date	Platform	Item	Price	Currency
123	2019-09-09	Website	Tripod	\$108	Euro
478	2018-01-01	Phone	Lens	\$888	Euro
555	2019-10-10	Website	Camera	\$1050	Euro

Summary

In this chapter, you:

1. **Identified missing values, duplicates, and timeliness issues** to improve data quality.
2. **Learned methods to handle missing data**, through deletion, imputation, and interpolation.
3. **Used pivot tables and VLOOKUP to find and remove duplicates.**
4. **Assessed timeliness** to ensure data aligns with project needs.
5. Applied **measures of central tendency** to fill in missing values appropriately.

These steps help **improve your data's fitness for use**—ensuring you produce **accurate, reliable results** for your analysis.

Consistency Check Pivot Table vs. Frequency Table

What is a Frequency Table?

A **frequency table** provides a **summary of how often unique values occur** in a dataset. It's used to:

- Identify **patterns** or **outliers**.
- Spot **missing values** or **unexpected duplicates**.

- Understand the **distribution of data** (e.g., most and least frequent values).

Example of a Frequency Table:

Item	Count of Transactions
The Da Vinci Code	15
Harry Potter and the Deathly Hallows	25
The Hobbit	10

- **What it tells you:**
 - "Harry Potter" was checked out **25 times**, making it the most popular item.
 - It can also highlight **anomalies** like missing titles or unusually high transaction counts.

How to Create a Frequency Table in Excel 🧑:

1. Select your data range.
2. Go to **Insert** → **Pivot Table**.
3. In the Pivot Table Field List:
 - Drag the column you want to count (e.g., **Item**) to the **Rows area**.
 - Drag the same column (**Item**) to the **Values area** and change the **Value Field Setting** to **Count**.

What is a Consistency Check Pivot Table? 🤖

A **consistency check pivot table** ensures that **related data across multiple columns** matches logically. It's used to:



- **Verify relationships** between fields (e.g., Item Title vs. Item Type).
- Detect **mismatched or inconsistent values** across related fields.
- Ensure data quality by checking for **logical alignment**.

Example Use Case:

In a library system:

- Check if **every DVD** (item type) is correctly marked with a **matching title** (e.g., "Harry Potter" → DVD).
- Confirm that no book titles are mistakenly labeled as DVDs.

Comparison: Frequency Table vs. Consistency Check Pivot Table

Aspect	Frequency Table 	Consistency Check Pivot Table 
Purpose	Counts how often each unique value occurs.	Verifies data consistency across columns.
Primary Use	Identifies distributions, patterns, and outliers.	Detects inconsistencies or mismatches between fields.
Example Use Case	Count transactions per book title.	Check if item types align with item titles.
Data Structure	Uses a single column (e.g., transactions).	Compares multiple columns (e.g., Title vs. Type).
When to Use	When you need a quick summary of occurrences.	When validating consistency across related fields.

When to Use Each Table

- **Frequency Table:**
 - Use when you need to know **how often specific values occur**.
 - **Example:** Find out how many times each library item was checked out.
- **Consistency Check Pivot Table:**
 - Use when you need to **validate relationships across multiple columns**.
 - **Example:** Ensure every **DVD title** is correctly labeled as a **DVD type**.

How They Work Together

- **Frequency tables** can help **spot inconsistencies** by highlighting unusual counts or missing values.
- **Consistency check pivot tables** confirm those inconsistencies by validating relationships across different fields.
 - **Example:** A frequency table shows a high count of "Harry Potter." The consistency check pivot table ensures that "Harry Potter" is consistently marked as a DVD.

Key Takeaway

- **Frequency Tables:** Focus on **counts and distributions**.

- **Consistency Check Pivot Tables:** Focus on **validating relationships** between fields.
- Both tools are **essential for data cleaning**—you’ll often use them together to identify and resolve data quality issues.
-

What is Interpolation?

Interpolation is a **data imputation technique** used to estimate unknown values between two known values in **time series data**. It assumes that data changes **gradually** over time and tries to fill in missing values by predicting what they would have been.




- **Best used for:** Time-based or sequential data (e.g., stock prices, temperatures, sales trends).
- **Goal:** Ensure the trend or pattern in the data remains intact.

What are Measures of Central Tendency?

Measures of central tendency provide a **summary statistic** that represents the center point or typical value of a dataset. These measures are often used to:

- **Summarize data** quickly.
- **Identify trends or patterns.**
- **Impute missing values** with a typical data point when needed.

The three primary measures are:

1. **Mean (Average)** 
2. **Median (Middle Value)** 
3. **Mode (Most Frequent Value)** 

Data Grain Explained

Imagine you have a **Lego set**. Each **Lego piece** is the smallest part of your set. If you want to **describe the set in detail**, you can’t just say, “It’s made of blocks.” You need to say, “**Each piece has a specific size, color, and shape.**”

In the same way, **data grain** is about **how specific** you need to be when describing **one row of data** in your dataset. It’s asking:

- **What does each row actually represent?**
- **How much detail** do you need to describe what each row means?

Quiz: 1.6: Data Quality Measures

1. In the context of data quality, what does completeness measure?

- a. How much data is unavailable
 2. Veronica is analyzing product sales data which includes variables for the date each item was purchased, a unique identification number for each product sold, the customers' names, and the products purchased. Which of these variables is an example of a data grain?
 - a. The unique identification number
 3. Data quality is a measurement of a data set's fitness for use. Data integrity, completeness, uniqueness, and which other factor are assessed when measuring data quality?
 - a. Timeliness
 4. Wes is analyzing data on volunteers of a statewide grassroots political organization. He finds the "County of Residence" variable cells mostly blank due to the question only recently being added to the volunteer signup form. The data, however, includes each volunteer's ZIP code. Which of the following methods could Wes use to impute the missing county values?
 - a. Educated guess
 5. In the context of data quality, what does uniqueness measure?
 - a. If the data contains duplicate records
-

1.7: Data Transformation & Integration

Learning Goals

- Integrate data from different sources.
- Prepare data for final analysis through transformation and integration.

Introduction

- So far, you've been working with separate data sets. Now, you'll combine them through **data integration** to create a **single unified dataset**.
- Integration helps you answer questions that need data from multiple sources.

What is Data Integration?

- **Data integration:** Combining data from different sources into one view.
- Often necessary because data lives in **silos** (separate, unconnected sources).

Why is integration necessary?

- Single data sets can give an incomplete picture.
- Example: Using both population and health data gives a full picture of influenza's impact in each state.

Key Steps for Data Integration

1. **Identify common variables** (called key variables) between the datasets.
2. **Map data sets** to find relationships and discrepancies between variables.
3. **Transform data** to ensure common variables match in format and structure.
4. **Use VLOOKUP()** in Excel to integrate the data sets into one.

Relational Data

- **Relational data:** Data sets with **common variables** (e.g., “Year” and “Country”).
- These variables serve as **keys** to combine datasets correctly.

Data Mapping

- **Data mapping:** Manually **matching variables** between different datasets.
- **Goal:** Identify **overlapping variables** and **resolve discrepancies** (e.g., mismatched formats).
- **Example:** “USA” in uppercase vs. “usa” in lowercase must be standardized.

Deciding on a Primary and Secondary Dataset

- **Primary dataset:** The most **relevant, complete, or accurate** dataset.
- **Secondary dataset:** Adds additional variables to the primary dataset.

Data Grain (Level of Detail)

- **Data grain:** The **lowest level of detail** in a dataset.
- Both datasets must be at the **same grain** for successful integration.
- Example: One dataset at a movie-title level, and the other at a country-year level.

Solution: Aggregate data to the same level (e.g., **Country-Year**).

Data Transformations

1. Format Transformations

- Ensure consistent formats (e.g., dates and text) between datasets.
- **Use Excel functions:**
 - **TRIM()**: Remove whitespace.
 - **PROPER()**: Capitalize text properly (e.g., "usa" → "USA").
 - **RIGHT()**, **FIND()**, and **LEN()**: Extract date segments.

2. Aggregation Transformation

- Aggregate data using **pivot tables** to match grain (e.g., **Country-Year**).

Integrating Data Using VLOOKUP()

- **VLOOKUP() formula:** Copies variables from a secondary dataset to a primary dataset.

Formula Breakdown:

=VLOOKUP(C2, Ratings!\$C\$2:\$D\$15, 2, FALSE)

- **C2:** Key value to look up.
- **Ratings!\$C\$2:\$D\$15:** Data range to search.
- **2:** Column number to return.
- **FALSE:** Exact match required.

Post-Integration Transformations

1. **Remove formulas:** Convert VLOOKUP() columns to values.
2. **Normalize data:** Use **ratios** to compare values across different scales (e.g., population percentages).
3. **Calculations:** Create **time-based windows** (e.g., sales in the last 30 days) or currency conversions.

Data Aggregation

What is Data Aggregation?

Data aggregation involves **summarizing or grouping data** to a higher level of detail. It helps reduce complex data into **meaningful summaries**, like totals, averages, or counts, often grouped by one or more key variables.

When to Use Data Aggregation?

- **Simplify large datasets** by summarizing key metrics.
- **Prepare data for analysis** by aligning different datasets (ensuring matching data grain).
- **Answer business questions** that require grouped insights (e.g., total revenue by year or region).
- **Data integration:** Aggregating data help

Why Aggregation is Important for Data Integration

- **Aligning the grain:** When working with different datasets (e.g., sales and social media data), you often need to **aggregate both to the same level** (like **Country-Year**).
- **Combining datasets:** Aggregated data helps reduce mismatches and ensures **clean, consistent integration**.
- **Answering business questions:** Aggregation lets you **focus on the variables** that matter most to your analysis.

Aggregation Example for Integration:

In your scenario of **movie revenues and Facebook likes**, aggregation helps:

1. **Movies data:** Aggregate gross revenue by **Country-Year**.
2. **Ratings data:** Aggregate Facebook likes by **Country-Year**.
3. **Integrate both datasets** using **Country-Year** as the common key.

Key Takeaways

- **Data aggregation** simplifies complex datasets by summarizing key metrics.
- Use **pivot tables** in Excel to aggregate data effectively.
- Aggregation helps **align the data grain** across different datasets, ensuring smooth **data integration**.
- It's an essential step to **prepare data for meaningful analysis**.

Summary ✨

In this chapter, you learned:

- **How to map, transform, and integrate data** from different sources.
- **How to align datasets** by matching their **grain and format**.
- **How to use VLOOKUP()** to combine datasets into a **unified view**.
- **How post-integration transformations** (e.g., normalization) improve data usability.

Quiz: 1.5: Data Profiling & Integrity

1. What is the first step in the data integration process?
 - a. Identify relational data
2. What is the purpose of data integration?
 - a. To combine multiple data sets into one
3. “VLOOKUP” is a powerful Excel function that allows for quick integration of two data sets when a relational variable exists. What is the correct syntax for the “VLOOKUP” function?
 - a. `=VLOOKUP(cell to lookup, where to look, what to return, if a fuzzy match is okay)`
4. Which of the following is an example of a data aggregation?

- a. Numeric calculations such as counts and summations
 - 5. Normalization is one type of calculation used to transform data. What is the purpose of normalization?
 - a. To make units uniform, such as converting the weight of a product from pounds to kilograms
-

1.8: Conducting Statistical Analyses

Learning Goals 🎯

- Utilize statistical methods such as standard deviation, variance, and correlation to analyze data effectively.

Introduction 🙌

- Statistical analysis helps in summarizing, interpreting, and finding patterns in data.
- In this section, we'll dive into measures of variability (variance, standard deviation), data distribution (normal distribution), and identifying relationships (correlation).

Descriptive Statistics 📊

- **Purpose:** Summarize data to detect trends, identify anomalies, and verify data quality.
- **Examples in Use:**
 - **Retail:** Stores predict product demand based on historical data (like stocking Pop-Tarts before hurricanes).
 - **Weather:** Temperature highs and lows help forecast trends.
- **In Analysis:** Descriptive statistics offer insights into patterns, confirming data quality and spotting outliers that might skew results.

Probability Distributions 🎲

- **Definition:** Displays the likelihood of various outcomes in a dataset.
- **Common Shapes:**
 - **Right Skewed:** Tail on the right (e.g., income distribution with high earners).
 - **Left Skewed:** Tail on the left.

- **Symmetrical/Normal:** Bell-shaped, balanced around a central mean.
- **Normal Distribution (Bell Curve):**
 - **Characteristics:** Mean, median, and mode align at the center.
 - **Applications:** Often found in natural data like heights or test scores.

Central Limit Theorem 🌐

The Central Limit Theorem is essential in statistics because it lets us treat data with more flexibility:

1. **What It Says:** As the sample size grows (usually above 30), the distribution of sample means becomes approximately **normal (bell-shaped)**, regardless of the population's shape.
2. **Why It Matters:** This lets analysts use **parametric statistics**, which are statistical methods that assume normal distribution. So, even if your data isn't perfectly symmetrical, you can apply these methods if:
 - Your data is **normally distributed**.
 - Or you have **over 30 samples**, allowing you to assume approximate normality.

The theorem helps in applying standard statistical methods confidently, giving accurate results when conditions are met.

Variance and Standard Deviation 📐

- **Variance and Standard Deviation** measure data spread, indicating how similar or different data points are within a dataset.

Variance

- **Purpose:** Quantifies how much each data point deviates from the mean.
- **Population Formula:** $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$ $\sigma^2 = N \sum (x - \mu)^2$
- **Sample Formula:** $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$ $s^2 = \frac{1}{n - 1} \sum (x - \bar{x})^2$

Standard Deviation

- **Purpose:** The average distance of data points from the mean, easier to interpret than variance.
- **Formulas:**
 - Population: $\sigma = \sqrt{\sigma^2}$ $\sigma = \sqrt{\sigma^2}$
 - Sample: $s = \sqrt{s^2}$ $s = \sqrt{s^2}$

Excel Calculations

- **Population Variance:** =VAR.P(range)
- **Sample Variance:** =VAR.S(range)
- **Population Standard Deviation:** =STDEV.P(range)
- **Sample Standard Deviation:** =STDEV.S(range)

Interpreting Results

- **Low Standard Deviation:** Data points are close to the mean (low variability).
- **High Standard Deviation:** Data points are widely spread (high variability).

The Empirical Rule

- **Purpose:** Guides interpretation of standard deviation in normal distributions.
 - **1 Standard Deviation ($\pm 1\sigma$):** 68% of data points.
 - **2 Standard Deviations ($\pm 2\sigma$):** 95% of data points.
 - **3 Standard Deviations ($\pm 3\sigma$):** 99.7% of data points.
- **Outliers:** Points more than two standard deviations from the mean are often considered outliers, helping identify potential anomalies.

Z-Scores

- **Definition:** Measures the number of standard deviations a value is from the mean. Useful for standardizing data across different scales.
- **Formula:** $Z = \frac{X - \mu}{\sigma}$ (population) or $Z = \frac{X - \bar{X}}{s}$ (sample).
- **Application:** Enables comparison across different units, and identifies extreme values (outliers).

Example

- If the mean drug dosage is 80 mg with a 6 mg standard deviation, a dosage of 86 mg has a z-score of:
 - $Z = \frac{86 - 80}{6} = 1$
- **Interpretation:** A z-score of 1 means this dosage is one standard deviation above the mean.

Correlations

- **Purpose:** Measures the strength and direction of a relationship between two variables.
- **Pearson's Correlation Coefficient (r):**
 - **Range:** -1 to +1.
 - **+1:** Perfect positive relationship.
 - **-1:** Perfect negative relationship.
 - **0:** No relationship.
- **Excel Formula:** =CORREL(range1, range2)

Interpreting Correlation Coefficients

- **0.0 - 0.3:** Weak relationship.
- **0.3 - 0.5:** Moderate relationship.
- **0.5 - 1.0:** Strong relationship.

Example in Analysis

- If analyzing drug dosage and weight, a positive correlation would suggest that higher weights correlate with higher dosages.

Summary ✨

- **Key Concepts Covered:**
 - **Variance & Standard Deviation:** Show data spread and variability.
 - **Empirical Rule & Outliers:** Identify outliers based on normal distribution.
 - **Z-Scores:** Standardize data for cross-variable comparison.
 - **Correlation:** Quantifies relationships between variables.
- **Importance:** These statistical tools help to interpret data meaningfully, identify patterns, detect anomalies, and compare data points, providing clarity in analysis.

Formula

Variance (Sample)

- Formula: =VAR.S(data range)
- Meaning: Measures how spread out the values in a data set are. For samples, it calculates the "average of squared differences" from the mean to see if values vary widely or stay close together.

Variance (Population)

- Formula: =VAR.P(data range)

- Meaning: Similar to sample variance, but used when you have data for an entire group (population). This also checks how spread out the values are from the average.

Standard Deviation (Sample)

- Formula: `=STDEV.S(data range)`
- Meaning: Shows the "average distance" of each data point from the mean (sample). It's more interpretable than variance since it's in the same units as the data.

Standard Deviation (Population)

- Formula: `=STDEV.P(data range)`
- Meaning: Same as sample standard deviation but used when analyzing the entire population. It shows data spread for the whole group.

Mean Absolute Error (MAE)

- Formula: `=AVERAGE(ABS(data - actual value))`
- Meaning: Calculates how far predictions (or guesses) are from actual values on average. Good for evaluating overall accuracy in predictions.

Empirical Rule (Not a Formula)

- Explanation: A guideline stating that in a normal distribution, about 68% of data falls within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3. Helps spot if data points (outliers) deviate too far from the mean.

Z-Score

- Formula (Sample): $(\text{data point} - \text{mean}) / \text{standard deviation}$
- Meaning: Tells you how far (in standard deviations) a data point is from the mean. Positive or negative values show if it's above or below the average, useful for comparing different data points.

Correlation Coefficient (Pearson's r)

- Formula: `=CORREL(data range 1, data range 2)`
- Meaning: Measures how closely two variables move together, ranging from -1 (opposite) to +1 (same direction). Helps to see if one variable's changes are related to another's.

Quiz: 1.8: Conducting Statistical Analyses

1. Variance and standard deviation measure what about a data set?
 - a. Spread
2. How does the Empirical Rule classify outliers?
 - a. Values more than two standard deviations away from the mean

3. What's the purpose of a Z-score?
 - a. To normalize data based on the number of standard deviations away from the mean
 4. Normal distribution is one way in which data can be distributed. What's unique about the normal distribution?
 - a. Normal distribution has a single symmetrical curve.
 5. Which of the following is true about variance?
 - a. Variance uses squared units
-

1.9: Statistical Hypothesis Testing

Learning Goals

- Conduct an inferential analysis using hypothesis testing
- Interpret results from a hypothesis test




1 Introduction to Inferential Statistics


- **Descriptive vs. Inferential:** While **descriptive statistics** describe data (like spread and relationships), **inferential statistics** allow generalizations from a sample to a population.
- **Key Uses:**
 - **Surveys:** Using a subset (sample) to represent a whole group (population).
 - **Quality Testing:** Testing some items to infer quality for all products.
 - **Medical Studies:** Comparing sample groups to see treatment effects.

2 Hypothesis Testing Process

Goal: Test if your hypothesis about a sample holds true for the population.

Steps:

1. **Formulate Hypotheses** 
 - **Null Hypothesis (H_0):** The statement you want to **disprove**.
 - **Alternative Hypothesis (H_1):** The statement you want to **prove**.
2. **Identify the Test Statistic** 
 - Measures how much your sample aligns with the population (e.g., **z-score** or **t-score**).
3. **Compute the p-value** 

- Indicates the likelihood that your results are due to **chance**.
4. **Compare p-value to Significance Level (α)** 
- Determines if your results are **statistically significant**.

Hypothesis Testing Details

Step 1: Formulating Hypotheses

- **Null Hypothesis (H_0):** The default assumption for the population.
- **Alternative Hypothesis (H_1):** Opposes H_0 ; what you think is true.
- **One-Tailed vs. Two-Tailed Tests:**
 - **One-tailed:** Tests for an effect in a specific direction.
 - **Two-tailed:** Tests for any difference, regardless of direction.

Example:

- Testing a new drug to reduce blood pressure:
 - H_0 : Drug has no effect on blood pressure.
 - H_1 : Drug lowers blood pressure.

Step 2: Identifying the Test Statistic

- **z-test:** Compares sample mean to a population mean if **population standard deviation is known**.
- **t-test:** Similar to z-test but used when **population standard deviation is unknown**.
 - **Purpose:** Accounts for variability and makes conclusions beyond just the mean.

Step 3: Computing the p-value

- **p-value:** Shows the likelihood your sample results happened by chance.
 - **Small p-value (low p-value):** Less chance results are random (supports H_1).
 - **High p-value:** Results may be random (supports H_0).

Step 4: Comparing the p-value to the Significance Level (α)

- **Significance Level (α):** Sets the cutoff for rejecting H_0 .
 - **Common α values:** 0.05 (95% confidence), 0.10 (90% confidence).
- **Decision Rule:**
 - If $p \leq \alpha$: Reject H_0 (significant result).
 - If $p > \alpha$: Fail to reject H_0 (not significant).

Types of T-tests

1. **Two-Sample t-test:** Compares means between two independent groups.
2. **One-Sample t-test:** Compares a sample mean to a known population mean.
3. **Paired t-test:** Compares means of the same group at different times.
4. **Two-Sample Assuming Unequal Variances:** A more conservative two-sample t-test (reduces the chance of falsely finding significance).

Running a t-test in Excel

1. **Enable Analysis ToolPak.**
2. **Set up null and alternative hypotheses.**
3. **Choose your t-test type** (one-sample, two-sample, or paired).
4. **Interpret Results:**
 - **T-score:** Shows the difference magnitude between groups.
 - **p-value:** Compared to α to determine if results are significant.

Summary

1. Identify hypotheses (**null** and **alternative**).
2. Select the appropriate **test statistic** (z or t).
3. Calculate **p-value** and compare to α .
4. **Reject or fail to reject H_0** based on results.

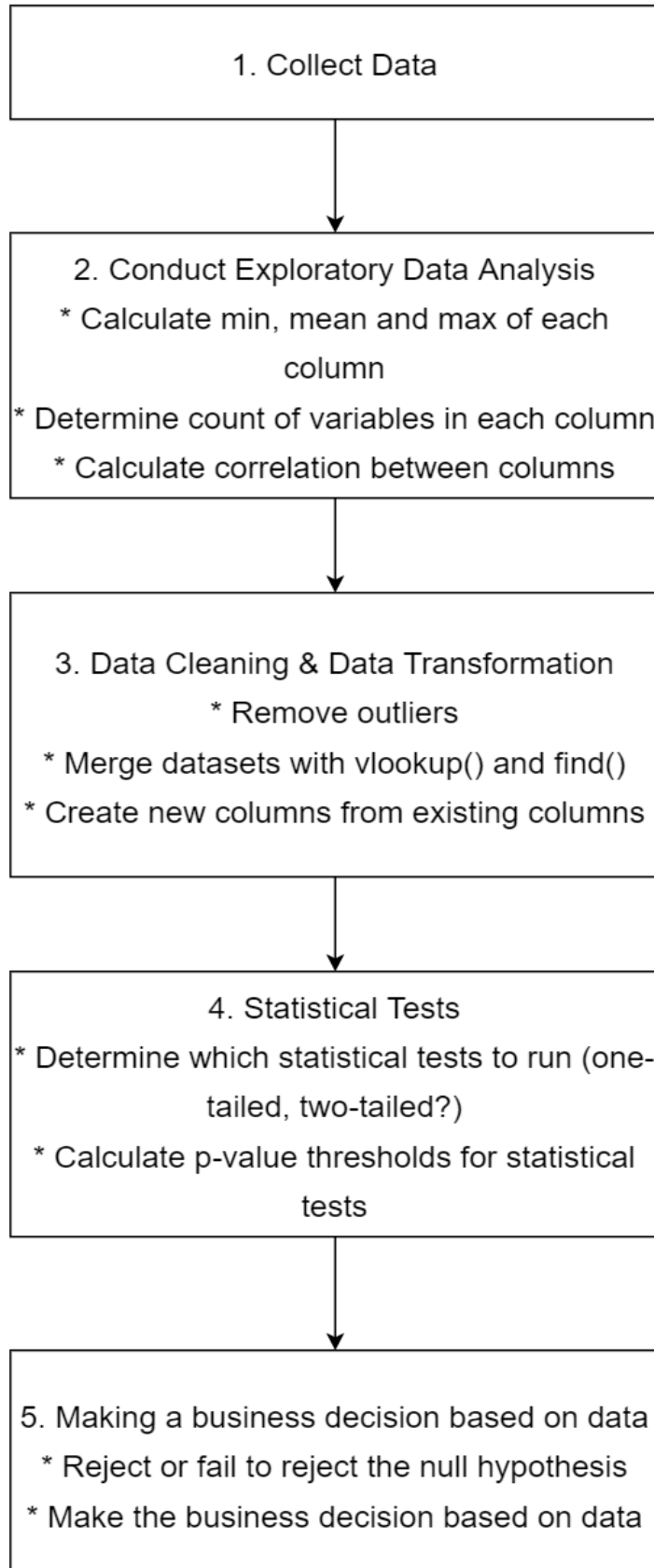
Outcome Statement Example: "At an α of 0.05 (95% confidence), there was no significant difference in rental assistance between students who graduated and those who dropped out."

Formula Chart

Formula Name	Formula	Meaning
Z-score	$=(X - \mu) / \sigma$	Measures how far a value is from the mean in terms of standard deviations.
T-score	$=(X - \mu) / (s/\sqrt{n})$	Similar to Z-score but uses sample standard deviation; ideal for smaller samples.
P-value	Excel output from test	Indicates the probability of observing the result if the null hypothesis is true.
One-sample t-test	T.TEST (range1, μ , tails, 1)	Compares means of two independent groups. Use "2" for two-sample test in Excel.

Paired t-test	T.TEST (range1, range2, tails, 1)	Compares means of the same group at two different times.
---------------	--	--

Flow Chart for Statistical Analysis



Quiz: 1.9: Statistical Hypothesis Testing

1. The test statistic is the random variable that helps assess the similarity between the sample data and what?
 - a. **Statistical hypothesis**
 - i. The test statistic is the random variable that helps assess the similarity between the sample data and the statistical hypothesis. The test statistic is used to test the null hypothesis and can be calculated in one of several different ways, for instance, by way of the z-score.
2. Which of the following statements about statistical hypotheses is true?
 - a. The alternative hypothesis represents what an analyst believes to be true
 - i. When forming statistical hypotheses, the null hypothesis represents what an analyst believes to be false, and the alternative hypothesis represents what an analyst believes to be true. Both hypotheses are mutually exclusive and tested by data. A null hypothesis can only be proven false—not true.
3. The p-value is a measure of the strength of the evidence against the null hypothesis. What does a low p-value indicate?
 - a. More-reliable or significant results
 - i. The p-value measures the probability of something happening by chance. Therefore, a lower p-value indicates more-reliable, or more-significant, results (i.e., results that are less likely to happen by chance). While this alone doesn't prove a null hypothesis false (and a null hypothesis can't be proven true), what it can do is increase the confidence level that a null hypothesis COULD be proven false.
4. Which type of t-test compares the means from the same group at two different times?
 - a. Paired t-test
 - i. The paired t-test compares the means from the same group at two different times. Conversely, a one-sample t-test compares the mean from a single group against a known mean, and a two-sample t-test compares the means for two independent groups. There's no such thing as a Student's t-test (only the Student's t-distribution).
5. What's the process of examining two opposing characteristics about a sample to determine which one is true within the entire population?
 - a. Hypothesis testing

- i. Hypothesis testing is the term for the process of examining two characteristics about a sample to determine which one is true about a population. First, null and alternative hypotheses are created. Then, a test is conducted to determine which of the two hypotheses is supported by the sample data. Whichever is deemed to be true for the sample can then be inferred to be true for the population.
-

1.10: Consolidation Analytical Insights

Learning Goals

- **Summarize research findings** for stakeholders
- **Communicate insights** effectively

1 Introduction to Data Analysis Reporting

- **Purpose:** Share an update with stakeholders in a concise, accessible way.
- **Format:** 1–2 page **executive summary** with visuals if needed.
- **Audience:** Present a high-level summary for stakeholders; avoid unnecessary details.

2 Components of a Report

1. Project Overview

- **Motivation:** Why the project is needed (e.g., rising dropout rates).
- **Objective:** Clear project goal (e.g., reduce dropout rate by 5%).
- **Scope:** Who/what the project impacts (e.g., the current graduating class).

2. *Tip:* Include a “tl;dr” summary at the top to make key points clear in a sentence or two.





3. Hypothesis

- Restate your **research hypothesis**.
- *Example:* “If families have rental assistance, students won’t drop out.”

4. Data Overview

- Summarize **data sources** and **basic info**.
- Avoid deep technical details; focus on what the data represents.

5. Data Limitations

- Briefly mention **any data issues** that could impact results (e.g., manual entry errors).
 - *Example:* Rental assistance data may have inaccuracies due to manual entry.
6. **Descriptive Analysis** 
- Include **means, standard deviations**, and any important correlations.
 - *Example:* "The average rental assistance was \$100, and family size correlated with assistance received ($r = 0.7$)."
7. **Results & Insights** 
- **Highlight key findings** from your hypothesis testing.
 - Be clear about any **significant or null results**.
 - *Tip:* If results were negative or null, present them objectively.
8. *Example:* At a 95% confidence level, rental assistance did not impact dropout rates.
9. **Remaining Analysis and Next Steps**  17
- Outline any **remaining analysis** and **next steps**.
 - *Example:* "Hold stakeholder meetings to explore potential reasons for null results."
10. **Appendix** 
- Include **additional documentation** for interested stakeholders.
 - Possible items: project brief, data profiles, hypothesis development notes.

Key Tips for Presenting Results

- **Be Objective:** Especially for negative or null results.
- **Use Simple Language:** Avoid too much jargon; make insights clear for non-technical readers.
- **Connect to Project Goals:** Tie findings back to the original objectives to help stakeholders understand the relevance.

Formula Chart

Formula Name	Formula	Meaning
Mean (Average)	=AVERAGE(data range)	Finds the average value of a data set.
Variance (Sample)	=VAR.S(data range)	Measures data spread for a sample.
Variance (Population)	=VAR.P(data range)	Measures data spread for an entire population.
Standard Deviation (Sample)	=STDEV.S(data range)	Shows the average distance of each point from the mean (for a sample).

Standard Deviation (Population)	=STDEV.P(data range)	Shows the average distance of each point from the mean (for a population).
Correlation	=CORREL(array1, array2)	Measures the strength and direction of a relationship between two variables.
P-value	Excel test output	Shows the probability of getting results due to chance; a lower p-value indicates significance.

Quiz: 1.10: Consolidation Analytical Insights

1. Dan is writing a report for an analysis project that measures how parents reading to children at home affects student literacy. What should he include in the data overview portion of his report?
 - a. **Basic information about the relevant data sources used in his analysis**
 - i. The data overview portion of a written report should include a quick overview of data sources. If Dan's student literacy data came from test results, for instance, he'd include a brief description of the information (e.g., reading, writing, and speaking test results for students of a certain age over a certain number of years). Data limitations are discussed in their own section, while data cleaning and quality information are better suited for the appendix. The project's motive, objective, and scope should be included in the project overview report section.
2. What's the term for the attachment to a written report that includes further details, explanations, and resources related to an analysis project?
 - a. Appendix
 - i. The additional section attached to a written report that includes more in-depth relevant explanations is called the appendix. This section can include the business requirements document, a hypothesis development section, profiles on data sets, and additional results and insights.
3. Data analysis project deliverables can include written reports, meeting presentations, and dynamic dashboards, among others. Why would an analyst create a written report instead of the other options?
 - a. **Written reports are concise, easily shareable project status updates**
 - i. Written reports can update stakeholders without requiring a meeting, which can be helpful when gathering stakeholders is

not easy. Written reports aren't necessarily easier or harder than other deliverables, nor are they required in a project. Additionally, they don't necessarily summarize findings better than other deliverables.

4. What's the purpose of the "hypothesis" section of a written report?
 - a. **To remind readers what prediction was tested during the analysis**
 - i. The hypothesis section reminds readers of the prediction that was made and tested during an analysis project. For example, an analyst may hypothesize that increasing the staffing of emergency room nurses would lead to fewer patient fatalities. The analyst would then collect and analyze data to either prove or disprove this hypothesis during the project. This process—the stated hypothesis, the data used to test it, and the results—would then be included in a report for stakeholders.
5. What's the best practice when presenting null or negative results for a hypothesis test?
 - a. **Be as objective as possible**
 - i. When presenting null or negative results, it's best practice to be as objective as possible. Focus on the facts and what the data suggests—not on anyone's individual contributions to a project or whether their actions were ineffective or negative. Encouraging discussion on why these results were unexpected can also ensure that your stakeholders are engaged, which can help decide the direction of the project.