MODULE 01

DATA ANALYSIS: The collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision making.

DATA ANALYTICS: is the science of data.

SIX STEPS OF DATA ANALYSIS:

The data analysis process helps analysts break down business problems into a series of manageable tasks:

Ask phase: you'll work to understand the challenge to be solved or the question to be answered. It will likely be assigned to you by stakeholders. As this is the ask phase, you'll ask many questions to help you along the way.

Prepare phase: you'll find and collect the data you'll need to answer your questions. You'll identify data sources, gather data, and verify that it is accurate and useful for answering your questions.

Process phase: is when you will clean and organize your data. Tasks you perform here include removing any inconsistencies; filling in missing values; and, in many cases, changing the data to a format that's easier to work with. Essentially, you're ensuring the data is ready before you begin analysis.

Analyze phase: is when you do the necessary data analysis to uncover answers and solutions. Depending on the situation and the data, this could involve tasks such as calculating averages or counting items in categories so you can examine trends and patterns.

Share phase: when you present your findings to decision-makers through a report, presentation, or data visualizations. As part of the share phase, you decide which medium you want to use to share your findings and select the data to include. Tools for presenting data visually include charts made in Google Sheets, Tableau, and R.

Act phase, in which you and others in the company put the data insights into action. This could mean implementing a new business strategy, making changes to a website, or any other action that solves the initial problem.

PROCESS IN BREIF:

1. **Ask**: business challenge, objective, or question
2. **Prepare**: data generation, collection, storage, and data management
3. **Process**: data cleaning and data integrity
4. **Analyze**: data exploration, visualization, and analysis
5. **Share**: communicating and interpreting results
6. **Act**: putting insights to work to solve the problem

Putting the process into practice

Now, think about how the phases in this process can be applied to a business situation.

The retirement contribution dilemma

The management team at a fictional midsized tech company, Geo-Flow, Inc., noticed that employee participation in the company's retirement contribution program was lower than expected. The company had invested a lot of resources in establishing its world-class benefit program, with the goal of reducing employee turnover. Because so few employees were using the program, leaders wondered if they should develop educational training to explain the benefits to employees. They wanted to make a well-informed decision before committing to the investment, so they asked their data analytics department to make a recommendation.

The analysts used the six phases framework and began by defining the problem. They **asked**, "Are employees investing in the company's retirement contribution program?"And, if not, "Should we create an educational program to encourage participation?" Satisfied with their research questions, they **prepared** their analysis project by gathering data from HR, such as employee demographics, salary levels, and current retirement contributions.

Next, they **processed** the data by cleaning and organizing it. They removed duplicates and data from individuals who had retired or left the company, then sorted the data by the employees' ages, departments, and length of employment. Their **analysis** showed that some employee groups were less likely to contribute to the plan or to be aware that the company offers a matching contribution. They interpreted these results to mean that these employee groups were not receiving enough education on the company's retirement contribution matching

program. They also studied the data to find trends and insights and used data visualization to review their analysis by exploring it in different contexts.

The analysts **shared** their findings with the management team using visualizations including bar and pie charts that illustrated the facts clearly so decision-makers could easily interpret the data. The report showed that, while overall participation was decent, some employee groups were not taking full advantage of the retirement program—but they might, if they knew more about the program and the matching contribution the company offers.

Based on these findings, the company took **action**, creating a targeted educational program focusing on the benefits of retirement contributions, specifically aimed at the employee groups identified as low contributors. Results showed that a few months after implementing this training, there was a significant increase in retirement contributions among the targeted groups.

Iteration during the data analysis process

The data analysis process is designed to build on itself, so the results from each step are the inputs for the next step. Keep in mind, however, that you might not always move through the steps linearly. For example, you might be in the analyze phase and find out your data was pulled from the wrong database. Or, you could learn while cleaning the data that your original question didn't adequately define the problem.

In cases such as these, you may have to go back to an earlier stage and work through the process with new, better information. The important thing is not to skip steps and miss something that's important. In fact, the biggest mistake analysts make when using this framework is looking for quick and easy answers.

Finally, make sure to review your work in each phase of the analysis. This helps you learn more about the situation and your own skill set, which will lead to the kind of continuous growth that helps data professionals succeed.

DATA ECOSYSTEM: The various elements that interact with one another in order to produce, manage, store, organize, and share data.

DATA SCIENCE: Creating new ways of modeling and uderstanding the unknown by using raw data.

DATA-DRIVEN DECISION MAKING: Using facts to guide business strategy.

GUT INSTINCT: is an intuitive understanding of something with little or no explanation. This isn't always something conscious; we often pick up on signals without even realizing. You just have a "feeling" it's right.

Why gut instinct can be a problem

At the heart of data-driven decision making is data. Therefore, it's essential that data analysts focus on the data to ensure they make informed decisions. If you ignore data by preferring to make decisions based on your own experience, your decisions may be biased. But even worse, decisions based on gut instinct without any data to back them up can cause mistakes.

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?"

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?
- How quickly does a decision need to be made?

Data analysts and detectives share a similar approach to problem-solving, both relying on evidence and facts to make decisions. Data-driven decision-making is essential for analysts, but gut instinct can also play a role in identifying patterns and connections. Balancing data and gut instinct is crucial for making informed decisions, and the right mix depends on the project's goals and time constraints.

ANALYTICAL SKILLS: Qualities and characteristics associated with solving problems using facts.

- Curiosity
- Understanding of context
- Technical mindset
- Data design
- Data strategy

CURIOSITY: Curiosity is a skill that drives analysts to discover just how much information they can coax out of the data in expected or unexpected ways. Keep in mind that curiosity isn't the only skill that compels analysts to ask probing questions about their data.

UNDERSTANDING CONTEXT: Understanding context helps you solve problems by narrowing down variables that are most likely to influence the outcome, which in turn enables you to come up with more meaningful insights.

TECHNICAL MINDSET: means approaching problems (and datasets) in a systematic and logical manner. This starts with the way you clean, organize, and prepare your data. It can also guide the tools or software you use to break down data and help you identify and fix incorrect data that can skew your analysis.

DATA DESIGN: is an extension of your technical mindset. It deals with how information is organized. Suppose the dataset here is presented in a spreadsheet. You would be able to shift the cells to organize the data to find different patterns.

DATA STRATEGY: is the management of the people, processes, and tools used in data analysis. In this scenario, think of it as the approach you use to analyze your dataset. One element might be the tools you use.

ANALYTICAL THINKING: Identifying and defining a problem and then solving it by using data in an organised, step-by-step manner.

- Visualization
- Strategy
- Proble-orientation
- Correlation
- Big-picture and detail-oriented thinking

Use the five whys for root cause analysis

A **root cause** is the reason why a problem occurs. So, by identifying and eliminating the root cause, data professionals can help stop that problem from occurring again.

The **five whys** is a simple but effective technique for identifying a root cause. It involves asking "Why?" repeatedly until the answer reveals itself. This often happens at the fifth “why,” but sometimes you’ll need to continue asking more times, sometimes fewer.

Boost customer service

An online grocery store was receiving numerous customer service complaints about poor deliveries. To address this problem, a data analyst at the company asked their first “why?”

Why #1. “Customers are complaining about poor grocery deliveries. Why?”

The data analyst began by reviewing the customer feedback more closely. They noted the vast majority of complaints dealt with products arriving damaged. So, they asked “why?” again.

Why #2. “Products are arriving damaged. Why?”

To answer this question, the data analyst continued exploring the customer feedback. It turned out that many customers said products were not packaged properly.

Why #3. “Products are not packaged properly. Why?”

After asking their third “why,” the data analyst did some further detective work. They ultimately learned that their company’s grocery packers were not adequately trained on packing procedures.

Why #4. “Grocery packers are not adequately trained. Why?”

This “why” enabled the data analyst to uncover that nearly 35% of all packers were new to the company. They had not yet had the chance to complete all required training, yet they were already being asked to pack groceries for customer orders.

Why #5. “Packers have not completed required training. Why?”

This final “why?” led the data analyst to find out that the human resources department had not provided necessary training to any newly hired packers. This was because HR was in the middle of reworking the training program. Rather than training new hires using the old system, they had provided them with a quick one-page guide, which was insufficient.

So, in this example, the root cause of the problem was that HR had not completed the training program updates and was using a less-thorough guide to train new packers. Fortunately, this was a problem that the grocer could control. And thanks to the data analyst's work, they provided more support to the HR department to complete the training and retrain all newly hired grocery packers!

GAP ANALYSIS: a method for examining and evaluating how a process works currently in order to get where you want to be in the future.