# The Data Analysis Process: Takeaways



## Concepts

### THE DATA ANALYSIS PROCESS

- The data analysis process is an established methodology for optimizing data for improved insights and effective decision-making.
- The stages of the data analysis process include the following:
    - **Ask a Question**
    - **Get the Data**
    - **Prepare the Data**
    - **Explore the Data**
    - **Analyze the Data**
    - **Communicate the Results**

### ASK A QUESTION

- The data analysis process starts with asking the right question(s).
- Data questions often come from reflecting on a problem or issue that a business would like to change or improve.
- These types of questions are commonly called business questions.
- A business question is a question about business performance that is key to the business's overall strategy.
- Good data or business questions shouldn't be ambiguous or open to interpretation.

### GET THE DATA

- After identifying one or more good data questions, it's time to get data that can address the question(s).
- It's crucial to understand the context of the data before taking any action.
- Here are some questions to ask the next time you encounter data or are asked to have confidence in data:
    - **Who**: who collected the data? What perspectives or agenda might that person or group bring to the data-collection process?
    - **What**: what data is included? What data, if any, is missing?
    - **When**: when was the data collected? Is it current? Is it still relevant?
    - **Where**: if the data is from an external source, where did it come from? Is the source reputable? Trustworthy?
    - **Why**: why was the data collected? Can data collected for one purpose be appropriately used for another purpose?

## PREPARE THE DATA

- Preparing the data for exploration and analysis generally includes the following:
    - **Extracting the data from its source**: this may involve writing SQL queries to extract and transform the data from a database, or downloading it from an external source.
    - **Transforming/organizing the data**: this involves summarizing the data, joining data from different sources, and putting it into a format that is accessible in the tool you'll use to explore it.
    - **Cleaning the data**: this involves things like removing duplicate data, correcting inaccurate or missing data, reformatting columns, and dealing with outliers.

## EXPLORE THE DATA

- Once relevant data is obtained to address a data question, it's time to explore the data.
- Data exploration uncovers initial patterns and interesting features in the dataset.
- This step of the process typically includes the following:
    - Looking for duplicate records, missing data, and/or very unusual cases (also called “outliers”).
    - Data wrangling (or data cleaning) to delete duplicate records, resolve missing data, and decide whether to keep or exclude the outliers.
    - Transforming existing variables or creating new variables based on the original data.

## ANALYZE THE DATA

- Analyzing the data follows data exploration.
- There are four levels of data analysis. Each level represents a different way to extract information from data, which combined with knowledge, can contribute to better decisions. In increasing order of complexity or sophistication, the four levels of data analysis are as follows:
    - **Descriptive analysis** uses past data to answer the question, “What happened?”
    - **Diagnostic analysis** seeks reasons or explanations for what has happened, asking and answering the question *Why did this occur*?
    - **Predictive analysis** uses past data to predict what will occur, asking and answering the question *What is likely to happen in the future*?
    - **Prescriptive analysis** is the most sophisticated of the four levels, using historical data to address the question *What should we do*?
- Descriptive and diagnostic analyses are often “one and done.” In other words, they are conducted once and never again in the same context.
- Other times, data analysis is conducted to “build a model” that is beyond the scope of normal data analysis in spreadsheets.
    - This is characteristic of predictive and prescriptive data analysis.

## COMMUNICATE THE RESULTS

- The final phase of the data analysis process is to communicate the results to stakeholders.
- These results can take many forms, such as reports, dashboards, presentations, visualizations, or even simple things like emails and conversations.

- One of the most powerful tools for communcating results is data visualization. Good data visualizations can help your audience quickly grasp the key takeaways from an analysis.
- We should remember the following principles when creating data visualizations and communicating results:
    - **Design for an audience**: results and visualizations should be tailored to the appropriate audience. This includes using visuals and metrics that the audience will easily understand.
    - **Less is more**: the takeaway from each visual should be clear. Adding too much data or unnecessary design features can distract our audience from the main point.
    - **Clear labels and annotations**: The chart titles, axis titles, legends, etc. should work together so a viewer can read the chart or report and know what the data represents without guessing.
    - **Tell a story with data**: *Data storytelling* involves communicating data insights via narratives and visualizations. Using elements from storytelling, such as a hook, a narrative, and a moral of the story, can make our analysis more memorable and engaging.

## Resources

- What Is the Data Analysis Process? 5 Key Steps to Follow
- A Step-by-Step Guide to the Data Analysis Process
- A Beginners Guide To The Data Analysis Process
- What is Data Analysis? Research | Types | Methods | Techniques
- Evaluating Data Sources