Business Research and Data Analytics

Lecture 2: Data analysis workflow, data type, and data collection

Iegor Vyshnevskyi, Ph.D.

Woosong University

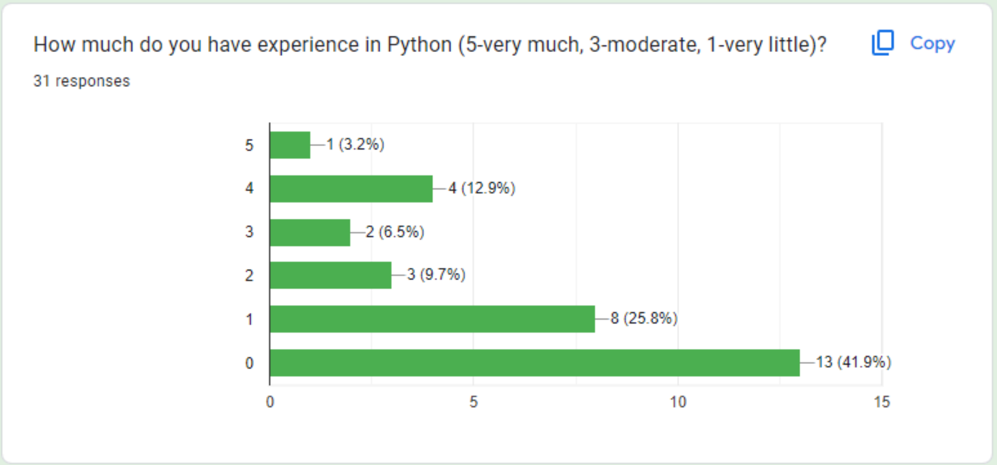
March 13, 2024

Agenda

1. Business application of Data Science
2. Data workflow
3. Data types
4. Types of data collection
5. Steps of secondary data collection
6. Defining the research questions
7. Types and sources of secondary data
8. Challenges in secondary data collection
9. Methods of accessing secondary data
10. Real-life examples of data collection
11. In-class assignment

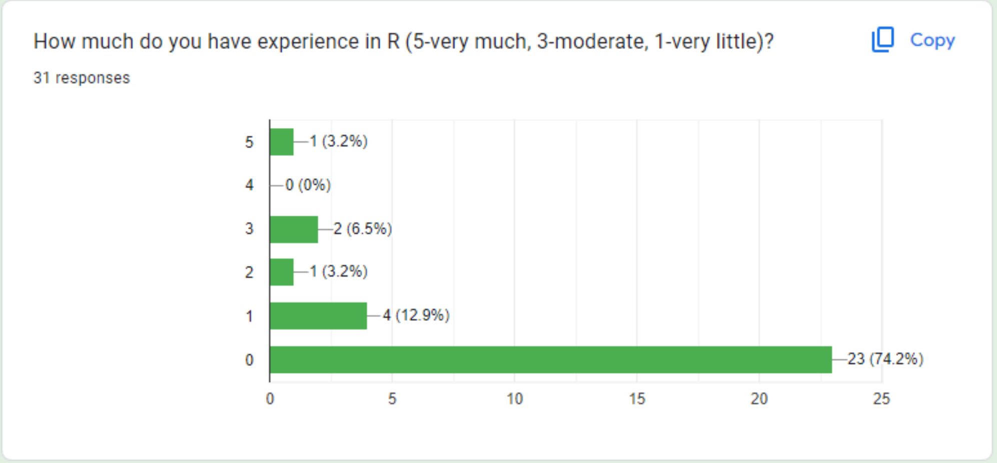
The surveys outcome

Background survey



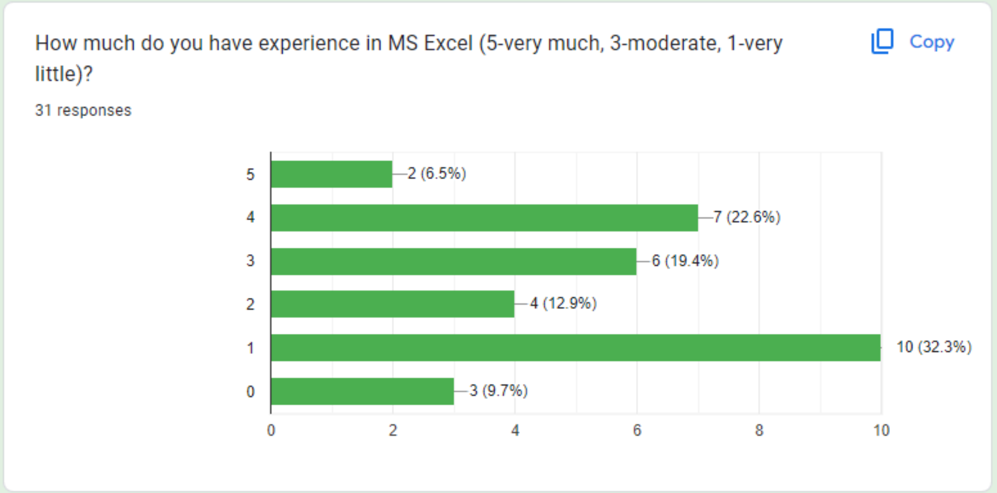
How much do you have experience in Python (5-very much, 3-moderate, 1-very little)?

31 responses



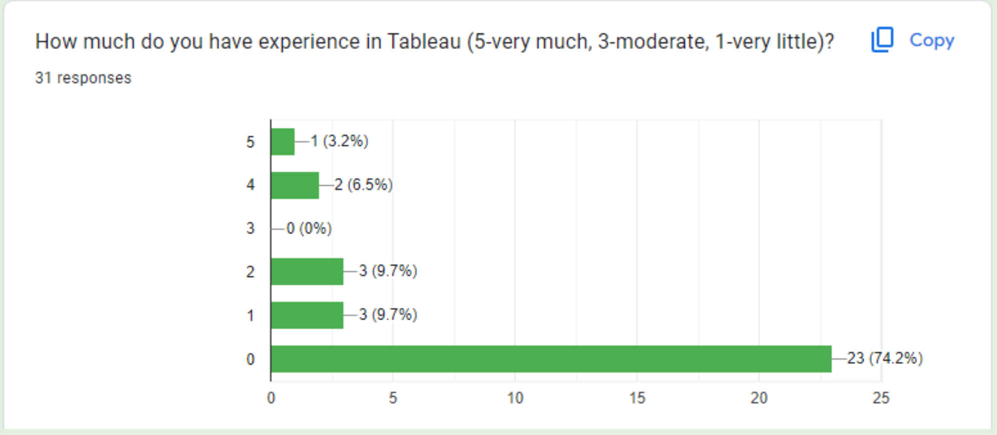
How much do you have experience in R (5-very much, 3-moderate, 1-very little)?

31 responses



How much do you have experience in MS Excel (5-very much, 3-moderate, 1-very little)?

31 responses



How much do you have experience in Tableau (5-very much, 3-moderate, 1-very little)?

31 responses

1. Business application of Data Science

Netflix: Netflix heavily relies on data science to personalize its content recommendations for users. By analyzing viewing habits, preferences, and engagement metrics, Netflix can suggest relevant content to each individual user, leading to higher user satisfaction and retention. This data-driven approach has contributed to Netflix's significant growth and increased valuation over the years.

Amazon: Amazon utilizes data science across various aspects of its business, from product recommendations and pricing optimization to inventory management and supply chain optimization. By analyzing vast amounts of data, Amazon can anticipate customer demand, optimize its logistics operations, and offer competitive pricing, resulting in improved profitability and increased market share.

Facebook: Facebook leverages data science to enhance user engagement and monetization on its platform. Through sophisticated algorithms, Facebook can deliver personalized content, advertisements, and recommendations to its users, maximizing user interaction and advertising revenue. This data-driven approach has played a significant role in Facebook's continued success and high valuation.

Tesla: Tesla utilizes data science and machine learning in its autonomous driving technology. By collecting and analyzing data from sensors and cameras installed in its vehicles, Tesla can improve the performance and safety of its self-driving features. This data-driven approach has not only enhanced Tesla's reputation as a leader in autonomous driving but also contributed to its increased valuation as investors recognize the potential of its technology.

Uber: Uber relies on data science to optimize its ride-hailing operations and improve the overall user experience. By analyzing data on rider demand, traffic patterns, and driver availability, Uber can efficiently match drivers with passengers, reduce wait times, and optimize routes, leading to increased customer satisfaction and loyalty. This data-driven approach has helped Uber maintain its competitive edge in the ride-hailing industry and attract investors, contributing to its high valuation.

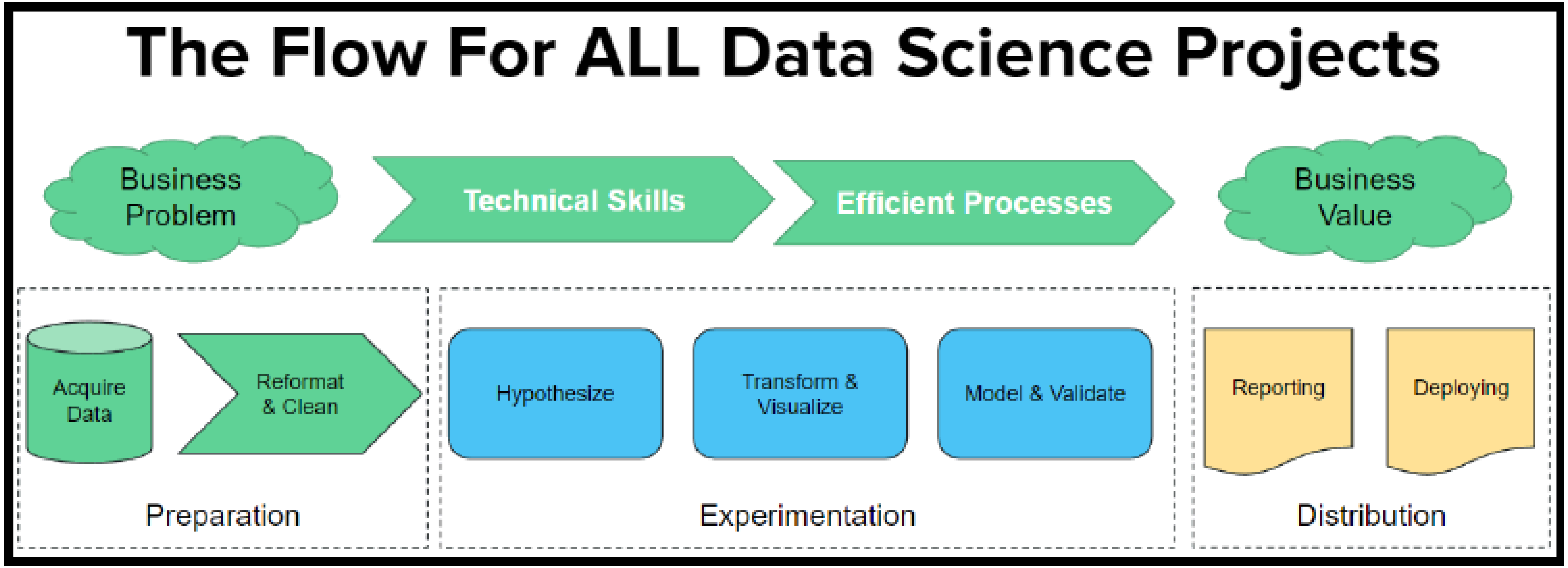
What conclusion can be made from these examples:

Data is a valuable asset that can enable companies to build **sustainable competitive advantages** and **justify high market valuations**.

By leveraging data to understand customers' behavior and preferences, companies can make better decisions about what products and services to offer, which can lead **to increased growth and profitability**.

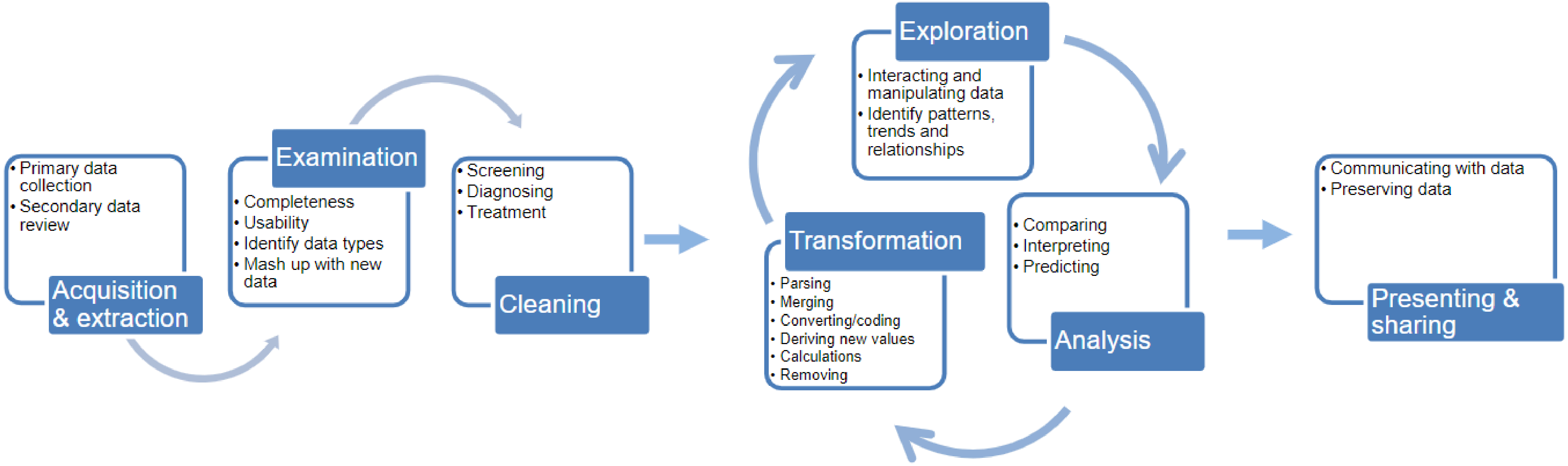
2. Data Science Workflow

Data Science workflow



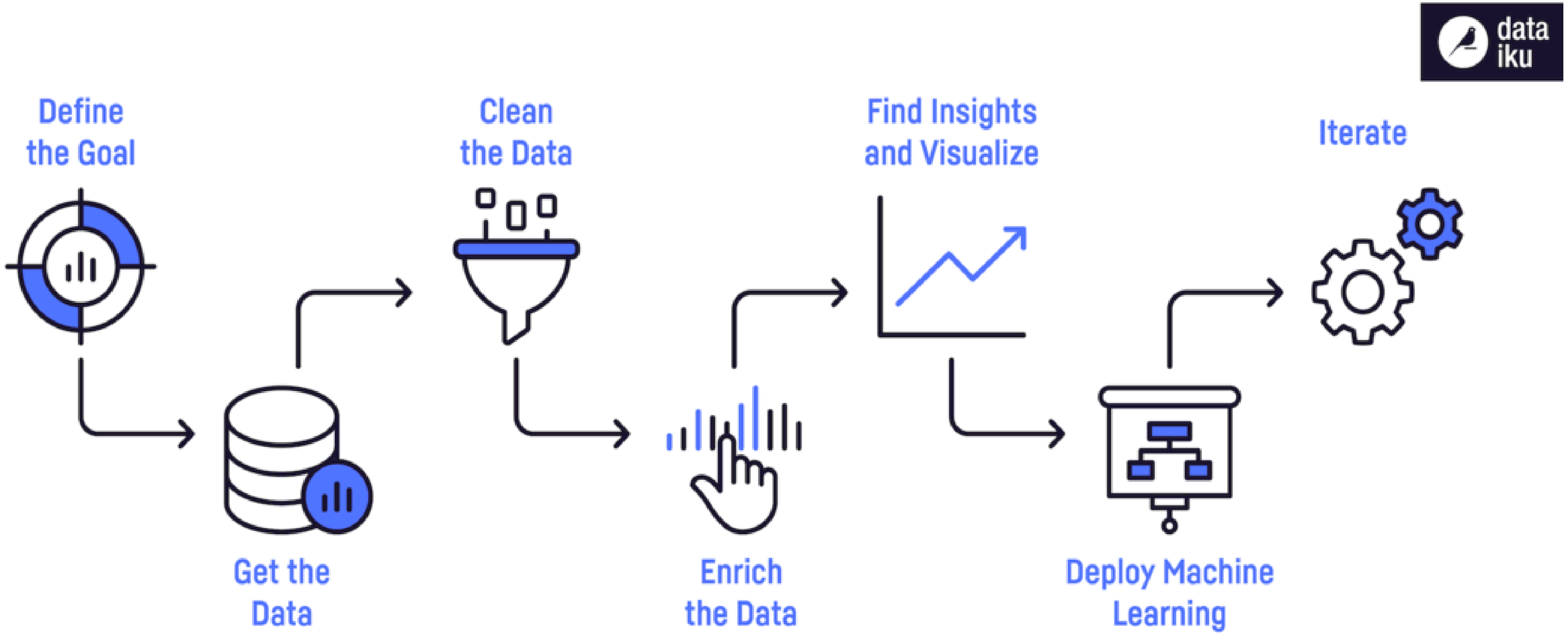
Source: Matt Dancho. Everything You Should Already Know About Data Science]

Data Analytics workflow (1)



Source: Kirk 2012, ACAPS 2013

Data Analytics workflow (2)



Source: A Comprehensive Guide on Microsoft Excel for Data Analysis]

Analytics Workflow Stages

1. *Generate*: All the ways data is generated and the systems of record where it is stored or originates from, also referred to as data ingress
2. *Collect*: All the ways data is collected or ingested
3. *Prepare*: All the ways data is transformed, including ETL (extract, transform, load), ELT (extract, load, transform), reverse ETL (from a warehouse into business applications), and ML (machine learning)
4. *Store*: All the ways data is stored, organized, and secured for analytics purposes
5. *Analyze*: All the ways data is analyzed
6. *Deliver*: All the ways data is delivered and how it is consumed, also referred to as data egress or data products

Source: Gary Stafford blog. Capturing Data Analytics Workflows and System Requirements]

Rules for Data Science

From the board room to the shipping dock, decisions are made every moment of the day using quantifiable, fact-based, trustworthy data (Heine Krog Iversen, Forbes)

1. Usefulness > Complexity
2. Data Quality > Hyperparameter Tuning (i.e., set of optimal values)
3. Simplicity > Novelty
4. Communication > Everything

Source: 4 Powerful Rules for Better Data Science

Rules for Data Analysis

Three rules for data analysis:

PLOT THE DATA, PLOT THE DATA, PLOT THE DATA

Source: Uwe H. Kaufmann

3. Data Types

Types

- Although different programming languages have different data types, there are some general types.

- Numeric Types:

Integer (int): Represents whole numbers. Example: 42, -10, 0

Floating-Point (float): Represents decimal numbers. Example: 3.14, -0.5, 123.456

Complex (complex): Represents numbers with real and imaginary parts. Example: 2 + 3j, -1.5 + 0.5j

- Textual Types:

String (str): Represents sequences of characters. Example: "Hello, World!", 'Python/R', "42"

- Boolean Type:

Boolean (bool): Represents true or false values. Example: True, False

- Collection Types:

List: Represents ordered sequences of elements. Example: [1, 2, 3], ['apple', 'banana', 'cherry']

Tuple: Represents ordered, immutable sequences of elements. Example: (1, 2, 3), ('red', 'green', 'blue')

Set: Represents unordered collections of unique elements. Example: {1, 2, 3}, {'apple', 'banana', 'cherry'}

Dictionary: Represents key-value pairs for efficient lookup. Example: {'name': 'Alice', 'age': 30}

- Special Types:

NoneType (None): Represents the absence of a value. Example: None

- Function: Represents executable code.
- Class and Object: Represents custom data types and instances.
- Other: images, video, geolocation, etc.

4. Types of Data Collection

Two Types of Data Collection

Primary Data Collection

the process of collecting original data, directly from the source.

- Data is designed to meet the specific needs of the research project.

Examples: surveys, interviews, offline quizzes, delphi technique, focus groups and observations.

Secondary Data Collection

the process of gathering information that has already been collected and analyzed by others for purposes other than the research at hand.

- Data has already been gathered for a different purpose other than the present research project.

Examples: sales records, industry reports, interview transcripts, APIs, web scraping.

**The Delphi technique%% is a structured process for collecting and summarizing opinions and feedback from a group of experts or stakeholders on a particular topic or question, with the goal of arriving at a reliable and accurate answer or consensus. It typically involves multiple rounds of feedback and anonymous responses to ensure unbiased and reliable results.

Advantage and Disadvantages of Both Types(1)

Primary Data Collection

Advantages

- Data is specific and relevant to the research question.
- Researchers can control the data collection process.
- Researchers can choose the method of data collection.
- Researchers can establish a relationship with respondents.

Limitations

- Can be time-consuming and costly to collect.
- May suffer from respondent bias.
- Sample size may be too small to make generalizations.
- Researcher must ensure the accuracy of the data collected.

Advantage and Disadvantages of Both Types(2)

Secondary Data Collection

Advantages

- Saves time and money.
- Large data sets available for analysis.
- Data has already been collected, so there is no need to go through the entire research process again.
- Can be used to compare data from different sources.

Limitations

- Data may not be specific to the research question.
- May not be up-to-date or relevant.
- Researchers do not have control over the original data collection process for secondary data collection since the data has already been collected by someone else.
- May suffer from biases introduced by the original data collectors.

In our lecture, we focus mainly on secondary data collection.

5. Steps of Secondary Data Collection

Steps of Secondary Data Collection

1. Define the research question. Determine the specific information needed and the research objectives. This will help to identify the types of data required and the sources that should be explored.

2. Identify potential sources. Search for sources of secondary data that can provide the information needed.

3. Evaluate the sources. Determine if the data is relevant and suitable for the research objectives.

4. Select the sources. Choose the most appropriate sources for the research objectives. Consider the cost, time, and effort required to access and analyze the data.

5. Obtain the data. Collect the data from the selected sources. This may involve downloading data from online sources, purchasing data from a vendor, or obtaining permission to access proprietary data.

6. Organize and store the data. Organize the data into a format that is easy to use and analyze. This may involve cleaning and transforming the data, removing duplicates and errors, and formatting the data for analysis.

7. Document the data. Keep track of the sources of the data, including the dates, locations, and authors. This will help to ensure the accuracy and reliability of the data and will also allow for easy citation of the sources in reports or other publications.

6. Defining the Research Questions

Defining the research questions

to make data-driven decisions

Why is it important to answer right questions?

People say: “To ask the right question, you need to know 80% of the answer”

- Well-defined questions assure the access to obtain relevant and actionable insights from data, and lead to informed and effective decision-making.

Steps to Define the Research Question

1. Identify what problem you want to solve

2. Formulate the effective research question

Identify what problem you want to solve

1. Making predictions

2. Categorizing things

3. Spotting something unusual

4. Identifying themes

5. Discovering connections

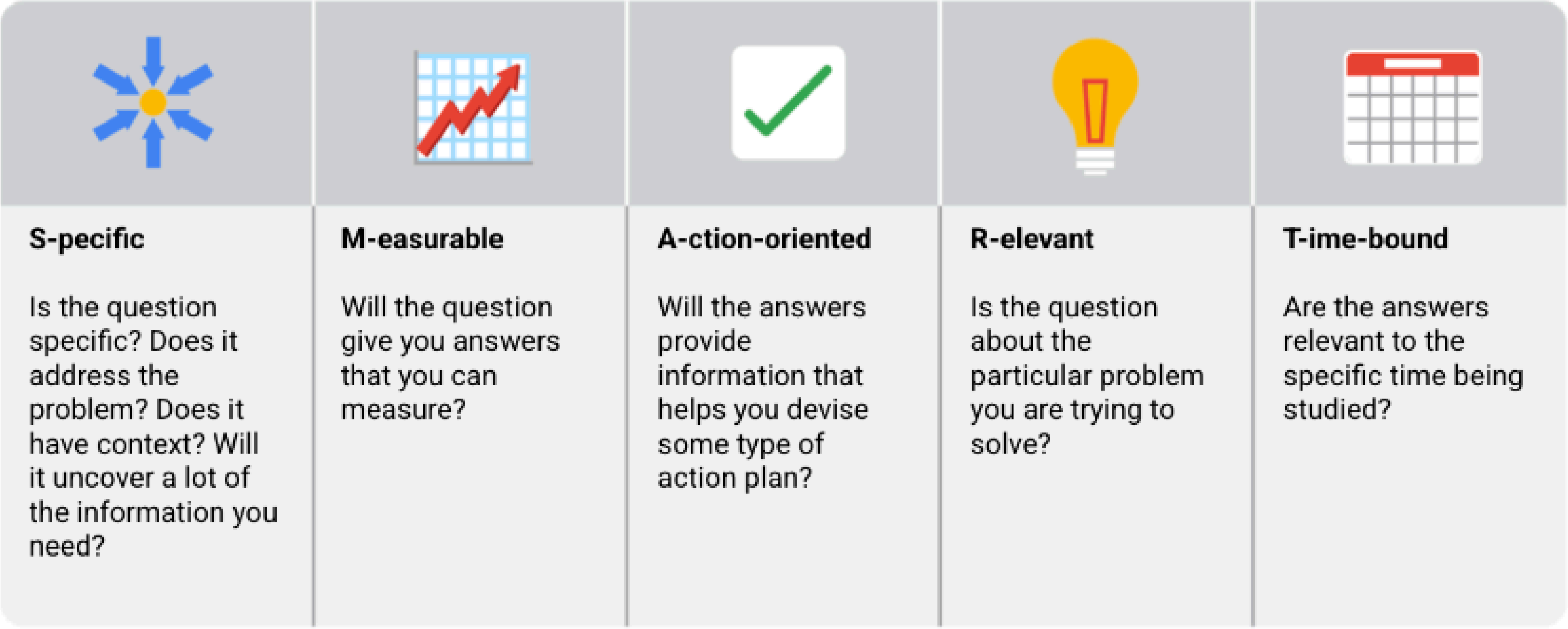
6. Finding patterns

Source: Coursera course “Ask Questions to Make Data-Driven Decisions”

Formulate the effective research question

SMART questions-approach

turns a problem question into one or more SMART questions



S-pecific	M-easurable	A-ction-oriented	R-elevant	T-ime-bound
Is the question specific? Does it address the problem? Does it have context? Will it uncover a lot of the information you need?	Will the question give you answers that you can measure?	Will the answers provide information that helps you devise some type of action plan?	Is the question about the particular problem you are trying to solve?	Are the answers relevant to the specific time being studied?

Source: Coursera course “Ask Questions to Make Data-Driven Decisions”

Example: *How does customer satisfaction affect repeat business in the restaurant industry?*

Specific: Does the question focus on a particular aspect of customer satisfaction, such as food quality or customer service?

Measurable: Does the question include a metric for measuring customer satisfaction, such as customer feedback or ratings?

Action-oriented: Does the question provide insights into how restaurants can improve customer satisfaction and increase repeat business, such as by improving their menu or training their staff?

Relevant: Does the question identify which factors of customer satisfaction are most important in driving repeat business, such as overall experience or value for money?

Time-bound: Does the question consider how customer satisfaction has changed over time and how it has affected repeat business in the restaurant industry in recent years?

7. Types and Sources of Secondary Data

Types of Secondary Data

Internal data: data that is generated and collected within an organization, such as sales records, financial data, and employee records.

External data: data that is collected from sources outside of an organization, such as government reports, industry publications, and market research reports.

Published data: data that is made available to the public through books, journals, and other publications.

Unpublished data: data that is not made publicly available, such as proprietary data, internal company reports, and personal communications.

Sources of Secondary Data

Government sources: census data, surveys, statistical databases, and reports published by government agencies.

Commercial sources: market research reports, industry reports, financial reports, and data from information services.

Non-profit sources: research reports, academic papers, and data from non-profit organizations.

Online sources: websites, social media, databases, and online publications.

8. Challenges in Secondary Data Collection

Major Challenges in Secondary Data Collection

- Data Quality
- Data Relevance
- Data Availability
- Data Limitations

Data Quality

What it refers to: The accuracy, completeness, and reliability of the data.

Why it is important: Poor data quality can lead to incorrect insights and decisions, which can be costly and damaging for organizations.

Examples of pure data quality: Inaccurate data, missing values, inconsistent data, duplicated data, etc.

Methods for evaluating: Checking the source of the data, assessing the methodology used to collect the data, considering potential biases in the data, and conducting statistical tests to identify errors and inconsistencies.

Possible ways to improve the situation: Using data cleaning and standardization techniques, implementing data quality checks during the data collection process, and choosing reputable data sources.

Data Relevance

What it refers to: The degree to which the data is useful for the intended research or analysis.

Why it is important: Collecting and analyzing irrelevant data can waste time, money, and resources, and can lead to incorrect conclusions.

Examples of pure data relevance: Data that is not directly related to the research question, outdated data, or data that is too broad or too narrow for the research needs.

Methods for evaluating: Conducting a thorough review of the research question and objectives to determine what data is needed, and carefully selecting and evaluating the data sources to ensure they meet those needs.

Possible ways to improve the situation: Clearly defining the research question and objectives, conducting a thorough review of available data sources before starting the data collection process, and using targeted and specific data collection techniques.

Data Availability

What it refers to: The ability to access and obtain the desired data.

Why it is important: Limited data availability can hinder research and analysis, and may lead to incomplete or inaccurate conclusions.

Examples of pure data availability: Data that is not publicly available, or data that requires permission or payment to access.

Methods for evaluating: Researching the availability of the desired data before beginning the data collection process, and assessing the cost and effort required to obtain the data.

Possible ways to improve the situation: Using publicly available data sources, negotiating access to restricted data, and considering alternative data sources or methods of data collection.

Data Limitations

What it refers to: The inherent limitations and biases of the data.

Why it is important: Understanding the limitations and biases of the data is important for interpreting and analyzing the data accurately.

Examples of pure data limitation: Biases in the data due to the sampling method used, limitations in the scope of the data, or limitations in the time period covered by the data.

Methods for evaluating: Understanding the limitations of the data source, and assessing the potential biases of the data through statistical analysis or other methods.

Possible ways to improve the situation: Using multiple data sources to complement each other and fill gaps in the data, and carefully documenting any limitations or biases in the data.

9. Methods of Accessing Secondary Data

Online databases: accessing data through online resources, such as databases and data repositories. Examples: FRED (Federal Reserve Economic Data),Bloomberg Terminal, PitchBook, S&P Global Market Intelligence, Statista, IBISWorld

Web scraping: using software to extract data from websites. Examples: BeautifulSoup (Python library), Scrapy (Python framework), RSelenium (R interface to Selenium), rvest (R package), Octoparse (web-based tool), WebHarvy (Windows desktop app), ParseHub (web-based tool)

API: accessing data through Application Programming Interfaces provided by data sources. Examples: Twitter API, Facebook API, and Google Maps API.

Web search: searching for data on the internet using search engines. Examples: Google, Bing.

Social media monitoring: collecting data from social media platforms. Examples: Twitter, Facebook, and Instagram.

Electronic surveys: conducting surveys online through email, websites, or social media. Examples: SurveyMonkey, Google Forms, and Qualtrics.

Electronic health records: accessing and analyzing medical data electronically, such as patient health information and medical histories.

10. Real-Life Examples of Data Collection

Electronic databases

- World Bank Databank link
- IMF data link

11. In-class assignment

Instructions

It's an individual assignment.

1. Define the research question.
2. Identify potential sources.
3. Evaluate the sources.
4. Select the sources.
5. Obtain the data.
6. Organize and store the data.
7. Document the data.

Write down the report on your activity and upload it together with your data.

Max score: 10 points

For Excel beginners: I highly recommend to complete the Introduction to Excel course with DataCamp (optional)

Any QUESTIONS?

Thank you!