1.3.1| Crafting Data Mining Problem Statement Part 1

You've learned in the previous session some practical applications of data mining. **How does data mining relevant to the work that you do or a problem that you want to solve? What is one question or problem that you would like to use data mining to solve?**

Please write it down and use it as a guide in meeting this session's learning outcome.

Whatever professional background you belong to, you most likely have access to data. Even in your everyday life, you unknowingly contribute to the generation of data. Wouldn't it be great if you can use these data to solve real-life problems?

In this session, you will learn how to identify and craft data-oriented problem statements. Crafting a problem statement is the initial step towards identifying the right data mining technique to use for model generation.

At the end of the session, you should be able:

1. Craft a data mining problem statement based on a pipeline that you have created in the previous session;
2. Make distinctions of descriptive and inferential analysis; and
3. Synthesize the steps in solving a data mining problem.

Steps in Solving a Data Mining Problem

Data mining begins with the question: **"How can we solve a problem using data?"** Go back to the previous session and refresh yourself with your data mining pipeline. Relate each stage of the pipeline to the following steps in solving a problem using data.

1. Practical motivation and sample collection
2. Problem formulation and data preparation

3. Statistical description and exploratory analysis
4. Pattern recognition and analytical visualization
5. Machine learning and algorithmic optimization
6. Statistical inference an information presentation.
7. Intelligence decision and ethical consideration

Step 1.1 Practical motivation and sample collection

Go back you to your answer to the question, What is one question or problem that you would like to use data mining to solve?

To check whether it is a good problem to solve using data mining, answer the following questions again.

1. Is the problem related to data?
2. If yes, can you solve this problem using data in practice?

Please write it down and use it as a guide in answering the next step of solving a problem using data.

Step 1.2 Data Collection

After crafting a problem that can be solved using data, you need to think about how to collect the relevant data. It's always good to review a lesson in statistics about getting a sample of a population.

In collecting data, you must answer the following questions.

1. Does the data you are trying to collect match the problem?
2. Does the data represent reality?

For example, if a poll is used to survey people's opinion about COVID 19 massive swab testing, how should this be conducted such that it can reflect reality?

Please write down your answer and use it as a guide in answering the next step of solving a problem using data.

Hover on the image for a hint.

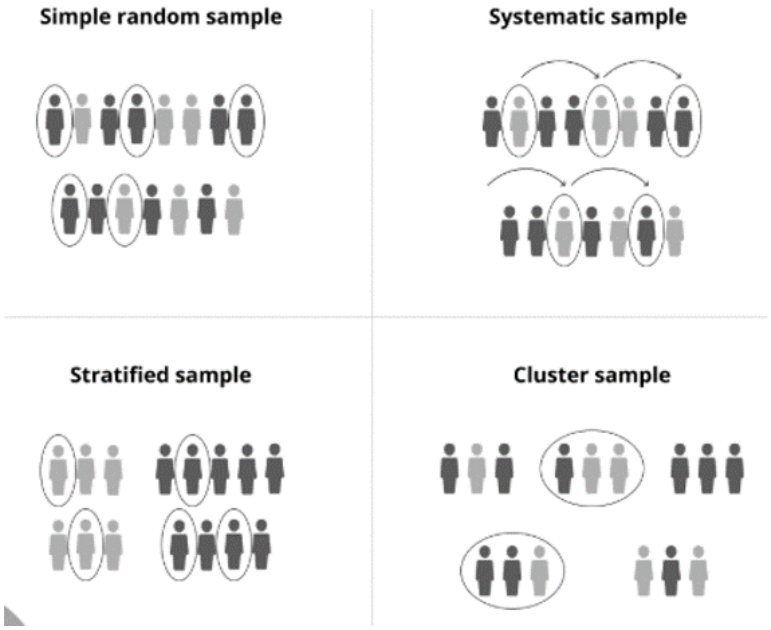

Figure 1.3.1 Sampling Methods

Click to review the details (https://www.scribbr.com/methodology/sampling-methods/)

Step 2.1 Problem Formulation

Which of the following questions describes a data mining problem?

1. Can age, blood type, location, and travel history predict if someone would most likely be infected with COVID 19?
2. How do I know whether someone has infected with COVID 19?

Hover on the image for a hint.

In case you didn't write it correctly, you can always rewrite your data mining problem. ✍☺

Step 2.2 Data Preparation

With the data mining problem that you have formulated in the previous step, you can now prepare your data for analysis.

Which of the following is the best kind of data?

1. Data from different sources
2. Data in a grid format

3. Noisy data

Hover on the image for a hint.

Step 3. Exploring the Data

How can we clearly express the data with some representations? What statistical insights are relevant? What kind of algorithm do you need to apply to find connections between data points? How to effectively aggregate the data?

In this stage, exploratory analysis is essential to have initial insights that we can get from the data. During exploratory analysis, some statistical properties of the dataset(mean, median, variance, distribution) are studied. What initial insights can you extract from the distribution of COVID cases in Figure 1.3.2?

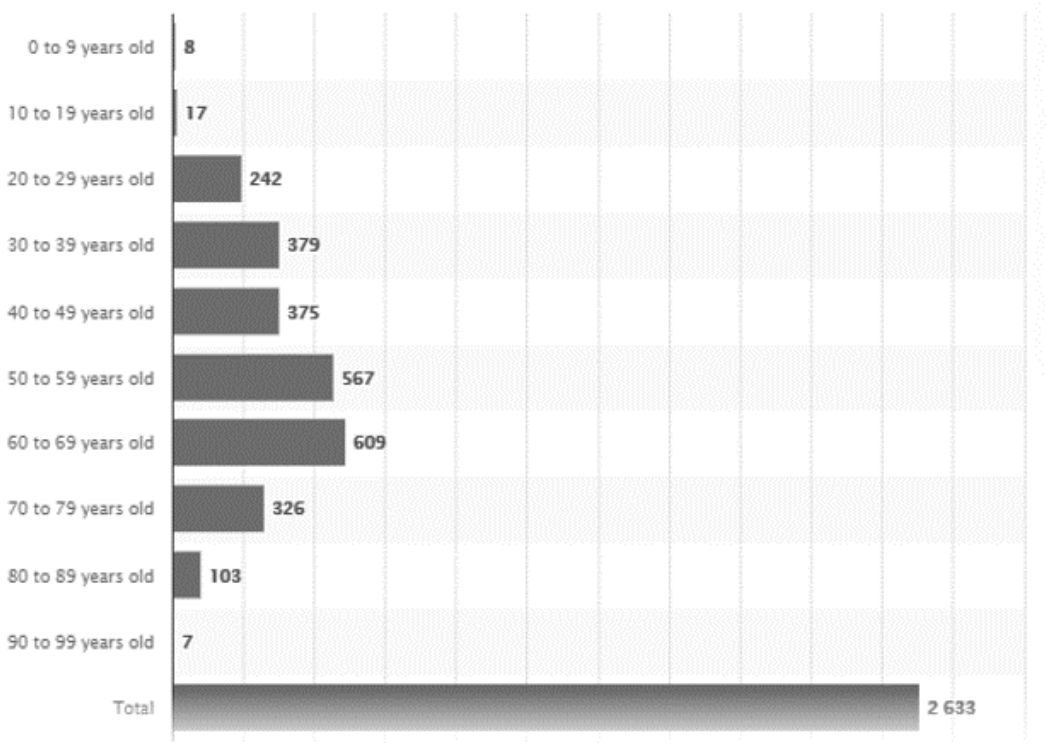

Figure 1.3.2 Distribution of COVID cases in the Philippines by Age Group

---

Step 4. Pattern Recognition and Analytical Visualization

Exploratory analysis is commonly referred to as descriptive analytics. In this type of analytics, statistical characteristics of data are presented in the form of visualization. This is the presentation of information using spatial or graphical representations for:

a. comparison facilitation;
b. pattern recognition; and
c. initial decision making

Types of Visualization

1. **Explore or Calculate**. This type of visualization requires a further analysis which requires a reason about a conveyed information.
2. **Communicate**. This type explains the information and suggests a hint for decision making.

Visualization can also be classified according to purposes, such as comparison, composition, distribution, and relationship. Figure 1.3 3 presents the different graphical presentations according to these classifications. The following questions can serve as a guide in determining which among the types is best suited for specific data.

- How many variables do you want to show in a single chart?
- How many data points will you display for each variable?
- Will you display values over a while, or among items or groups?

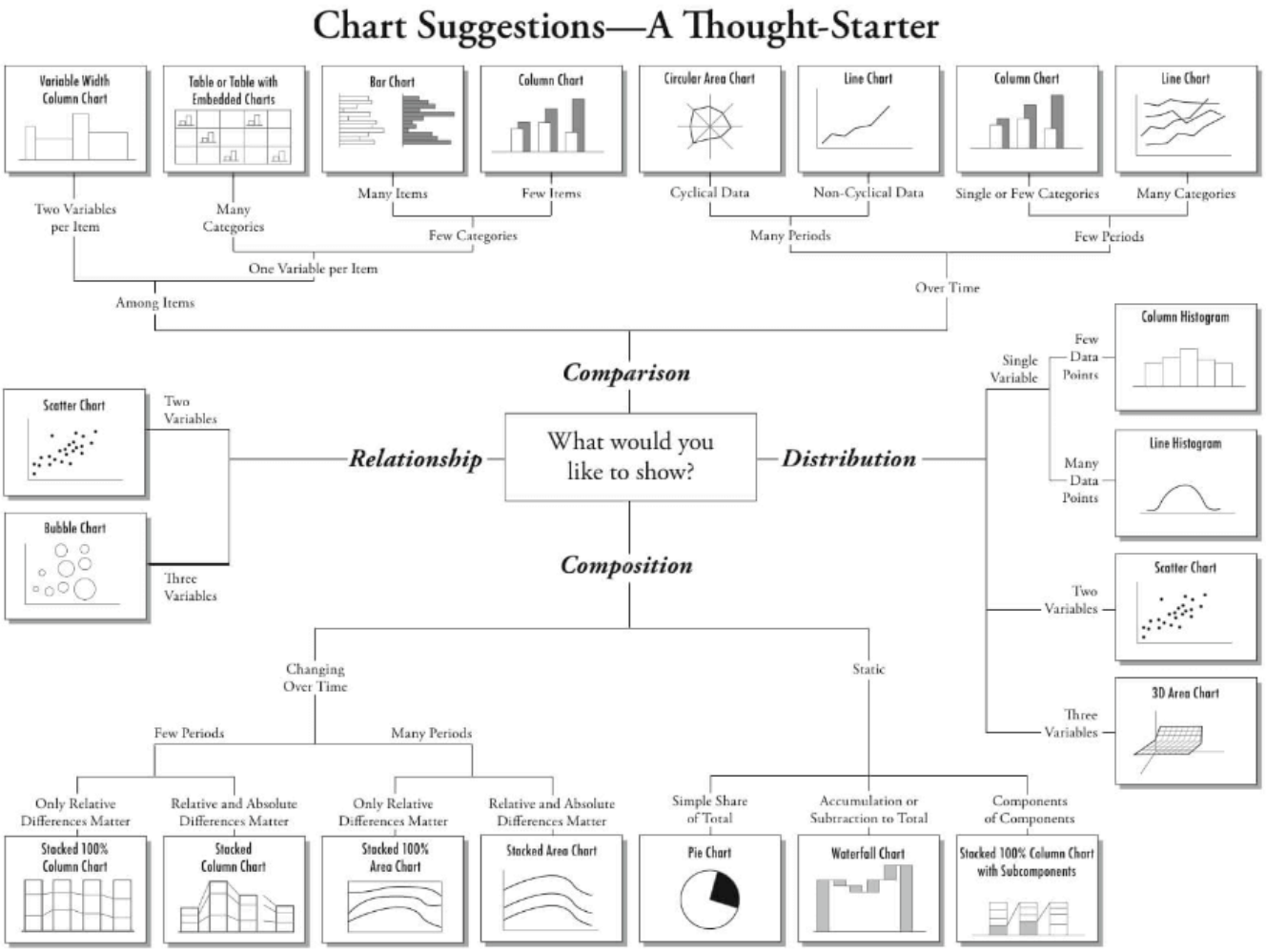

Figure 1.3 3 Visualization Types

Click here for more details **(https://www.tatvic.com/blog/7-visualizations-learn-r/)**

Upload examples of COVID 19 visualization. Describe the samples according to the type of visualisation.

Tap the button to share your answer to the other learners.

(https://drive.google.com/drive/folders/1dHJBVRkcfpaFC5cU2JVu4PesHX76dFgM?usp=sharing)

Step 5. Unleash the Power Data

In the previous steps, data was used for statistics and data visualization. This step emphasizes machine learning and algorithmic optimization.

The tasks in steps 1 to 5 can be broadly categorized as descriptive and inferential analytics.

The goal of **descriptive analytics** is to describe the population or dataset under study. The description can be used to generalized to any other group or population.

The **inferential analytics** is produced through complex mathematical calculations. The findings can be taken from the sample group and be generalized to a larger population.

Tap the image for more details of the difference between descriptive and inferential statistics

 (https://sciencestruck.com/descriptive-vs-inferential-statistics)

Let's assess your understanding of descriptive and inferential analytics. Which of the following examples are inferential analytics?

1. Use the model to a new set of data to generalize the observations.
2. Extract initial information from data using visualization.
3. Aggregate and summarize data for presentation
4. Interpret the general patterns that are extracted from data.

Hover on the image to check your answer

Step 6. Statistical Inference

It is also important to understand that learning algorithms are not perfect. So, how to confidently infer from statistical results?

Statistical inference can become certain in two ways:

1. Generalize the learning procedure or model
2. Estimate the confidence of prediction by the genera model

For example, a claim that COVID 19 cases will go down by **exactly** 10% in the coming month is certainly not dependable. However, if the claim that COVID 19 cases will decrease in a range of 3%-5% in the following month with the confidence of 95%. In statistics, this is referred to as confidence interval.

Share an example of a confidence interval to other learners.

Tap the button to upload your answer to the course bank.

(https://drive.google.com/drive/folders/1CfmKuOh1GkNHT5BUw-TtfdxA5J3TKvSD?usp=sharing)

Step 7. Actionable Intelligence and Ethical Considerations.

Data science is a team sport. Most of the time, it is essential to make decisions on real-life problems with the assistance of a domain expert.

It is also essential to keep in mind to safeguard data that needs utmost confidentiality. Hence, in analyzing data, it is necessary to consider the legality and privacy of individuals or organizations.

Share legal and ethical standards that you need to be aware of in solving a data mining problem.

Tap the button to upload your answer to the course bank.

(https://drive.google.com/drive/folders/1X2S8x8zLeMPGIVAJQLz1MMWXDTOj7J-Z?usp=sharing)

Tap the button to assess your understanding of crafting data mining problem statement.

(https://tip.instructure.com/courses/9953/assignments/156974)

References:

1. **Tatvic Visualization** **(https://www.tatvic.com/blog/7-visualizations-learn-r/)**
2. University of the Philippines Center of Analytics