Course: Artificial Intelligence and Machine Learning

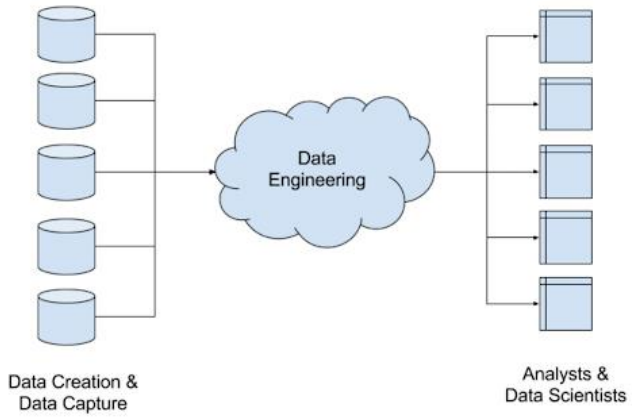
Code: 20CS51I WEEK- 4

SESSION – 1 Data engineering pipeline

- **Data engineering pipeline**
- **Data Collection**
  - **Different Methods of Data Collection**
  - **Population and sample**
- **Types of Data**

4.1 Data engineering pipeline

Data engineering is the process of designing and building systems that let people collect and analyze raw data from multiple sources and formats.



Data engineering pipeline is a transformation or processing of the data. This is because raw data loaded from a source might not be error-free or usable. And thus, it requires some alteration to be useful at its next node. The data engineering pipeline removes errors and resists bottlenecks or holdups, thus increasing end-to-end speed.

1. Extract — retrieving incoming data. At the start of the pipeline, we're dealing with raw data from numerous separate sources. Data engineers write pieces of code – jobs – that run on a schedule extracting all the data gathered during a certain period.

2. Transform — standardizing data. Data from disparate sources is often inconsistent. So, for efficient querying and analysis, it must be modified. Having data extracted, engineers execute another set of jobs that transforms it to meet the format requirements (e.g., units of measure, dates, attributes like color or size.) Data transformation is a critical function, as it significantly improves data discoverability and usability.

Activities

- A data engineering pipeline is the design and structure of algorithms and models that copy, cleanse, or modify data as needed.
- a data pipeline streamlines and automates the flow of data from one point to another, and automates all the data-related activities in the pipeline. These include data extraction, data ingestion, data transformation, and data loading.
- the primary purpose is to transfer raw data from database sources to data warehouses for use.

4.2 Data Collection

Data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

During data collection, the researchers must identify the data types, the sources of data, and what methods are being used.

Before an analyst begins collecting data, they must answer three questions first:

- What's the goal or purpose of this research?
- What kinds of data are they planning on gathering?
- What methods and procedures will be used to collect, store, and process the information?

Primary Data Sources

- The data sources which provide primary data are known as primary data sources, and information gathered directly from first-hand experience is referred to as preliminary data. This is the information you collect for the aim of a particular research endeavour.

- Primary data gathering is a straightforward method suited to a company's particular requirements

- E.g., in Census data collected by the government, Stock prices are taken from the stock market.

Secondary Data Sources

- These data sources provide secondary data. Secondary data has previously been gathered for another reason but is relevant to your investigation. Additionally, the data is collected by someone other than the team who needs the data.

- Second hand information is referred to as secondary data. It is not the first time it has been used, and that's why it's referred to as secondary.

- Secondary data sources contribute to the interpretation and analysis of main data. They may describe primary materials in-depth and frequently utilize them to promote a certain thesis or point of view.

Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity.

During data collection, the researchers must identify the data types, the sources of data, and what methods are being used.

Different Methods of Data Collection

Data collection breaks down into two methods.

- Primary.

As the name implies, this is original, first-hand data collected by the data researchers. This process is the initial information gathering step, performed before anyone carries out any further or related research. Primary data results are highly accurate provided the researcher collects the information. However, there's a downside, as first-hand research is potentially time-consuming and expensive.

Different methods to collect primary data,

- Interviews.

The researcher asks questions of a large sampling of people, either by direct interviews or means of mass communication such as by phone or mail.

- Projective Technique.

With projective data gathering, the interviewees get an incomplete question, and they must fill in the rest, using their opinions, feelings, and attitudes.

- Delphi Technique.

researchers use the Delphi technique by gathering information from a panel of experts. Each expert answers questions in their field of specialty, and the replies are consolidated into a single opinion.

- Focus Groups.

Focus groups, like interviews, are a commonly used technique. The group consists of anywhere from a half-dozen to a dozen people, led by a moderator, brought together to discuss the issue.

- Questionnaires.

Questionnaires are a simple, straightforward data collection method. Respondents get a series of questions, either open or close-ended, related to the matter at hand.

- Secondary.

Secondary data is second-hand data collected by other parties and already having undergone statistical analysis. This data is either information that the researcher has tasked other people to collect or information the researcher has looked up. Secondary information raises concerns regarding accuracy and authenticity. Quantitative data makes up a majority of secondary data.

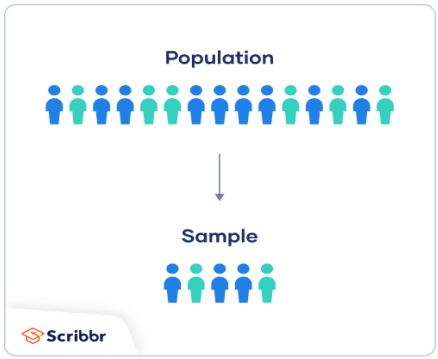
Unlike primary data collection, there are no specific collection methods. Instead, since the information has already been collected, the researcher consults various data sources, such as:

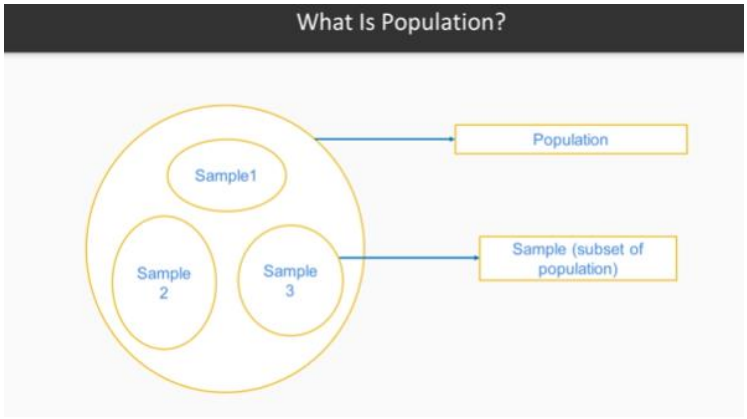
- Financial Statements
- Sales Reports
- Retailer/Distributor/Deal Feedback
- Customer Personal Information (e.g., name, address, age, contact info)
- Business Journals
- Government Records (e.g., census, tax records, Social Security info)
- Trade/Business Magazines
- The internet

4.2.1 Population and sample

A **population** is the entire group that you want to draw conclusions about.

A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.





Types of Data

Statistical Data types

1.2.1 Cross-sectional Data

Cross-section data is collected in a single time period and is characterized by individual units - people, companies, countries, etc. Some examples include:

- Student grades at the end of the current semester;
- Household data of the previous year - expenditure on food, unemployment, income, etc.
- Car data - average speed, horsepower, color, etc.

With cross-sectional data the ordering of the data does not matter. In other words, we can order the data by ascending, descending or even randomized order and this will not affect out modeling results.

1.2.2 Time Series Data

Data collected at a number of specific points in time is called time series data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc. Time series data can be observed at many different frequencies (hourly, daily, weekly, monthly, quarterly, anually, etc.).

Unlike cross-sectional data, the ordering of the data is important in time-series data. Each point represents the values at specific points in time. As such, time series data are typically presented in chronological order. Changing the order of the data ignores the time-dimensionality of the data.

Other Data types

- Image data

Various datasets based on image data (products and product placement, lighting conditions, color of advertisements, etc. An example of a practical use of image/video and other types of data can be seen in Amazon Go store chain (another link).

- Sentiment analysis text data

Also known as opinion mining. Used in extracting and analysing reviews, survey responses, online forum and social media posts. It is used to determine the attitude (positive, negative, emotional, etc.) of the subject (customer, reviewer, respondent, etc.) of the data.

Univariate, Bivariate and Multivariate data and its analysis

1. Univariate data –

This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

Suppose that the heights of seven students of a class is recorded, there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

2. Bivariate data –

This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.Example of bivariate data can be temperature and ice cream sales in summer season.

Suppose the temperature and ice cream sales are the two variables of a bivariate data. Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase. Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

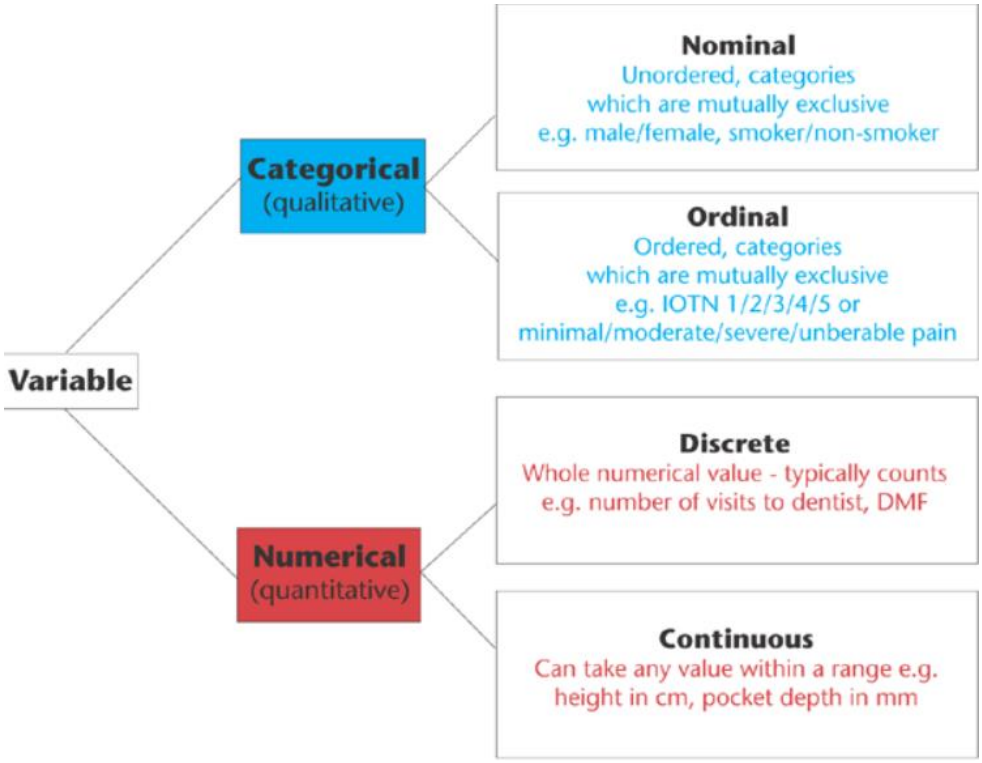
3. Multivariate data –

When the data involves three or more variables, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

Quantitative (Numerical) vs Qualitative (Categorical)

Qualitative variables are descriptive/categorical. Many statistics, such as mean and standard deviation, do not make sense to compute with qualitative variables. quantitative means you can count it and it's numerical (think quantity - something you can count).

Quantitative variables have numeric meaning, so statistics like means and standard deviations make sense. Qualitative means you can't, and it's not numerical (think quality - categorical data instead).



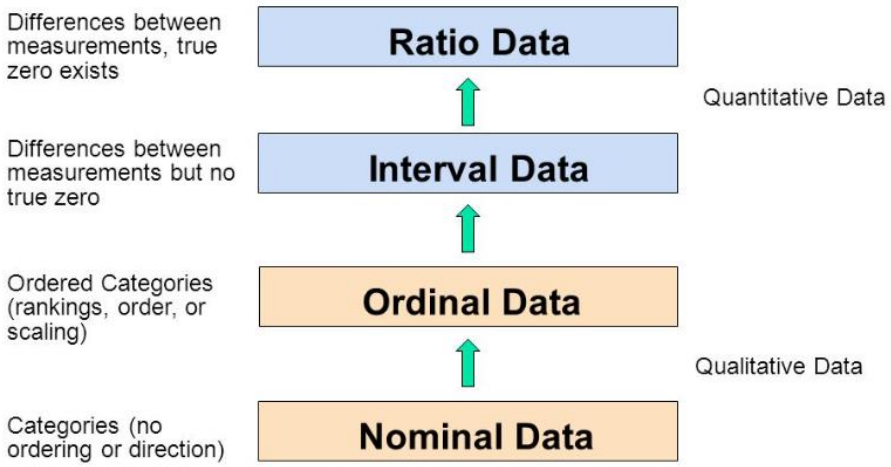
There's one more distinction we should get straight before moving on to the actual data types, and it has to do with quantitative (numbers) data: discrete vs. continuous data.

Discrete data involves whole numbers (integers - like 1, 356, or 9) that can't be divided based on the nature of what they are.

Like the number of people in a class, the number of fingers on your hands, or the number of children someone has. You can't have 1.9 children in a family .

Continuous data, on the other hand, is the opposite. It can be divided up as much as you want, and measured to many decimal places.

Like the weight of a car (can be calculated to many decimal places), temperature (32.543 degrees, and so on), or the speed of an airplane.



Nominal

A nominal scale describes a variable with categories that do not have a natural order or ranking. You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless.

Examples of nominal variables include:

- genotype, blood type, zip code, gender, race, eye color, political party

Ordinal

An ordinal scale is one where the order matters but not the difference between values.

Examples of ordinal variables include:

- socio economic status (“low income”,”middle income”,”high income”), education level (“high school”,”BS”,”MS”,”PhD”), income level (“less than 50K”, “50K-100K”, “over 100K”), satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).

Interval

An interval scale is one where there is order and the difference between two values is meaningful.

Examples of interval variables include:

- temperature (Farenheit), temperature (Celcius), pH, SAT score (200-800), credit score (300-850).

Ratio

A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.

Examples of ratio variables include:

- enzyme activity, dose amount, reaction rate, flow rate, concentration, pulse, weight, length, survival time.

Data Collection

data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Case Study

Federal government ensures the quality of education across the country every year through a surveying agency. The agency conducts surprise tests for primary school students to identify the quality of education.

The agency provides two options for conducting this exercise:

Option 1:

Conduct surprise test for every student throughout the country, analyse their performances and then conclude on the quality of primary school education.

Option 2:

Conduct surprise test for a selected group of students across various schools in the country. Analyse the marks obtained by the selected students and draw inferences on the quality of primary school education.

Option 1 collects all the required data and conducts statistical operations to arrive at conclusions. This type of statistical analysis falls under Descriptive Statistics.

Option 2 collects a subset of data, known as Sample in statistical terminology, from the entire data known as Population. The sample is analysed and conclusions are drawn about the population. This type of analysis falls under Statistical Inference (also known as Inferential Statistics).

To conduct descriptive statistics, one must analyse the entire population. However, it may not be feasible to collect all the data because of the complexity and cost involved in data collection. For example, it is not feasible to conduct surprise test for every student across all the schools in the country.

The surveying agency prefers Option 2.

The surveying agency uses various statistical inference techniques to draw conclusions about the population, by analysing the sample. You will learn such techniques throughout this course.