TYPES OF DATA/ DATA CHARACTERISTIC

STRUCTURED VS. SEMI-STRUCTURED VS. UNSTRUCTURED DATA

- **Structured Data**
- It comes with a predefined format and structure. Structured Data is usually stored in Relational Databases. It is easy to deal with in the Data Science domain.

Sepal_length	Sepal_width	Petal_length	Petal_width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	versicolor
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	virginica

SEMI-STRUCTURED DATA

- It comes with a predefined format and structure but is not stored in the Relational Database.

- JSON (Javascript Object Notation)

```json
{
    name: "Linear Algebra for Machine Learning",
    author: "Json Brownlee",
    pages: 211,
    parts: 5,
    format: "PDF",
    total_codes: 92
}
```

- XML (Extensible Markup Language)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<book>
   <author>Json Brownlee</author>
   <format>PDF</format>
   <name>Linear Algebra for Machine Learning</name>
   <pages>211</pages>
   <parts>5</parts>
   <total_codes>92</total_codes>
</book>
```

UNSTRUCTURED DATA

- It does not have a specific format and lacks structure. It is the type of data that presents many challenges to handle in the Data Science domain
- **Examples**:
  - Images
  - Videos
  - Speech
  - web logs

DATA

- Collection of objects and their attributes

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

Tuple

Variables/features/attribues

Observations /records

price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
$42,000.00	5850	3	1	2	yes	no	yes
$38,500.00	4000	2	1	1	yes	no	no
$49,500.00	3060	3	1	1	yes	no	no
$60,500.00	6650	3	1	2	yes	yes	no
$61,000.00	6360	2	1	1	yes	no	no
$66,000.00	4160	3	1	1	yes	yes	yes
$66,000.00	3880	3	2	2	yes	no	yes
$69,000.00	4160	3	1	3	yes	no	no
$83,800.00	4800	3	1	1	yes	yes	yes
$88,500.00	5500	3	2	4	yes	yes	no
$90,000.00	7200	3	2	1	yes	no	yes
$30,500.00	3000	2	1	1	no	no	no
$27,000.00	1700	3	1	2	yes	no	no

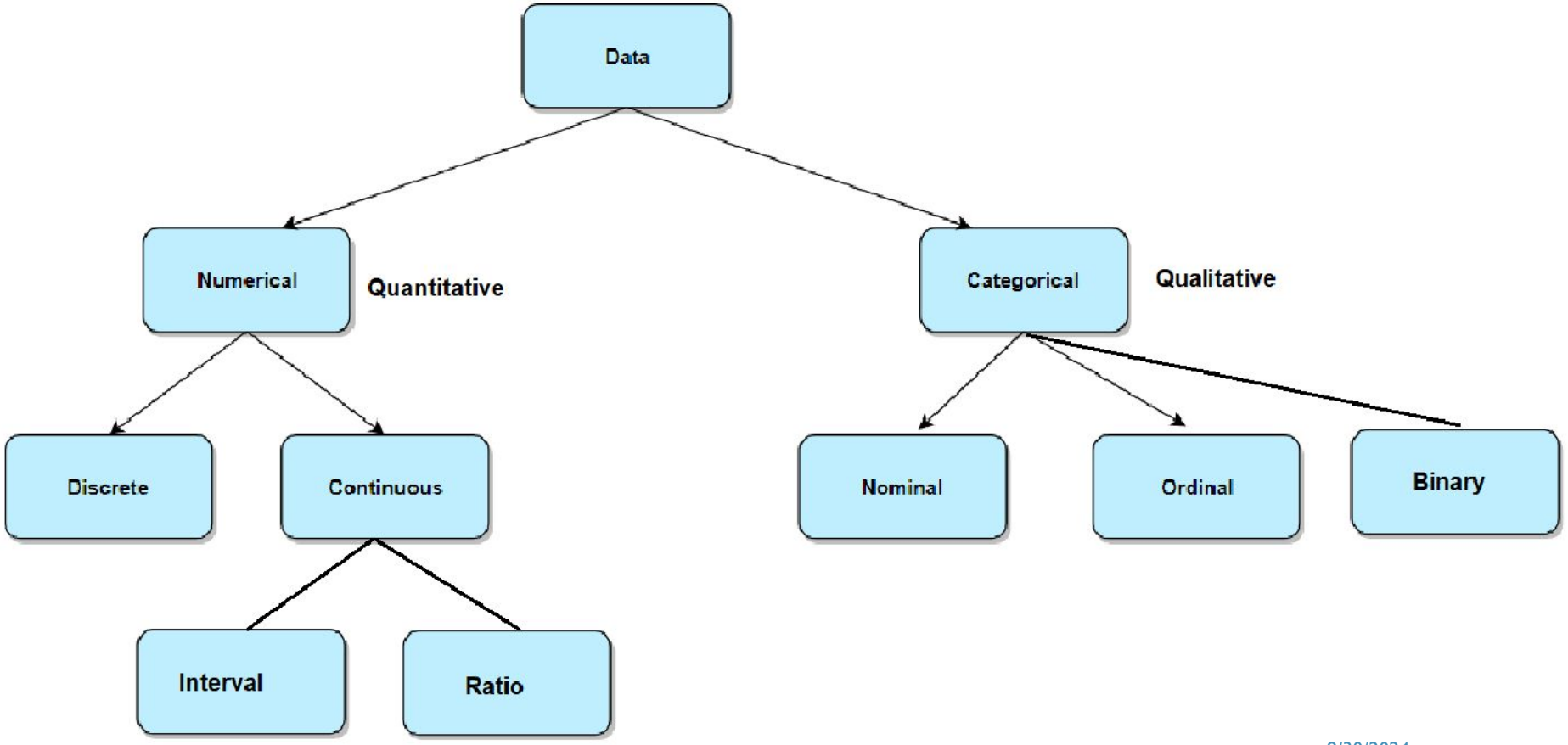
Quantitative Variables
- Numeric

Qualitative Variables
- Categorical

DESCRIBE THE DATASET

- Need to know in what form the data is present to analyze it properly and apply different statistical methods on it
- What do your records represent?
- What does each attribute mean?
- What type of attributes?
  - Categorical
  - Numerical

DATA TYPES


Data
Numerical
Quantitative
Categorical
Qualitative
Discrete
Continuous
Nominal
Ordinal
Binary
Interval
Ratio

CATEGORICAL DATA

- Categorical data as the name suggests, represent categories or characteristics such as gender, language, level of education, marital status, the genre of a movie, etc
- It is also known as *Qualitative Data*.
- We can associate numerical values with categorical data, but they would not have any mathematical meaning, e.g., 0/1 for male/female.

CATEGORICAL DATA
1. NOMINAL DATA

- Nominal data is categorical data that has no order
- It can be thought of as *labels,* have no quantitative value
  - *G*ender of a person as male or female
  - Language a person speaks
  - Eye color

- Nominal Data can be dealt with using frequencies, proportions, pie charts, bar plots, etc.

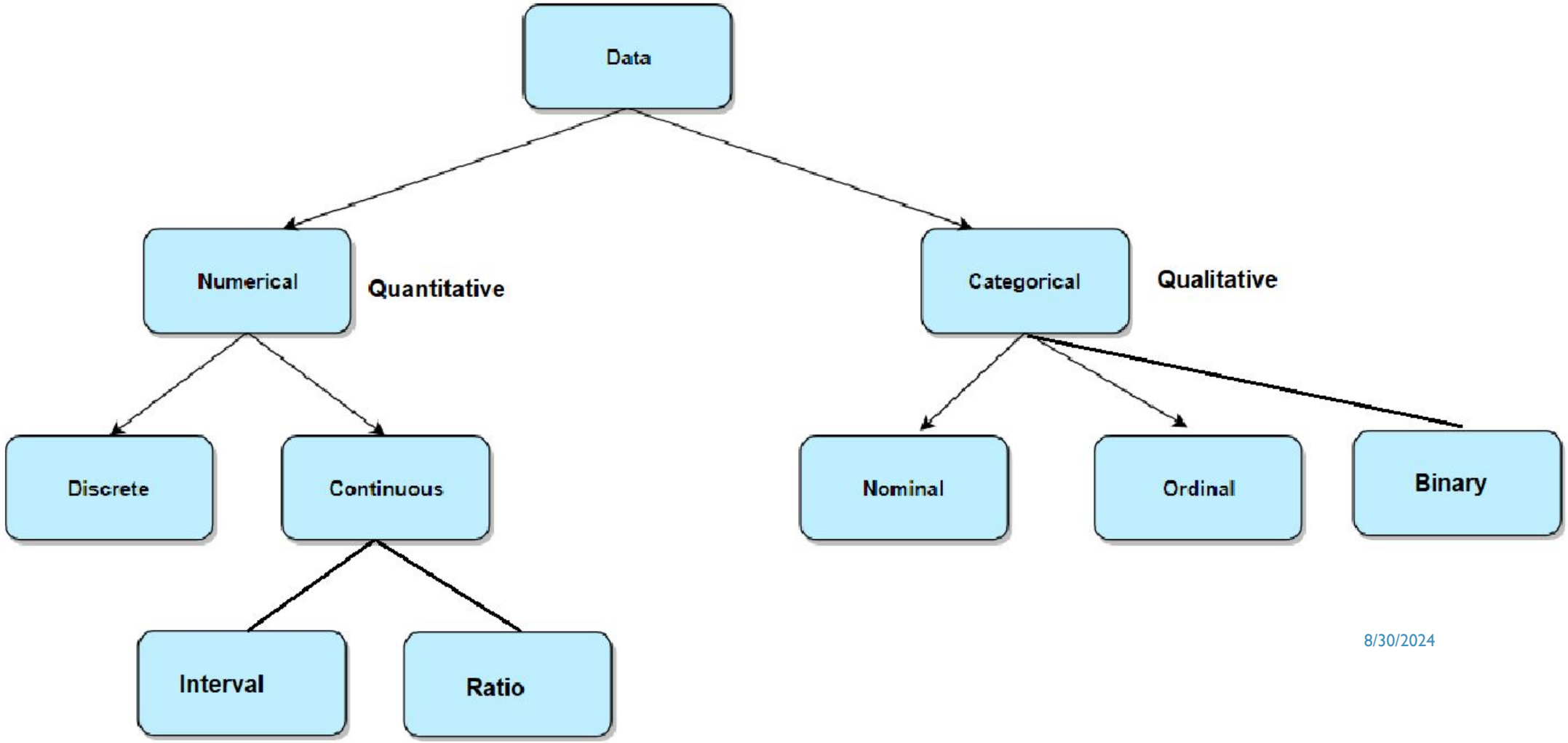
CATEGORICAL DATA
2. ORDINAL DATA

- Ordinal data is categorical data that has a sense of order to it
- Ordinal values represent discrete and ordered units
- It is therefore nearly the same as nominal data, except that it's ordering matters
  - Happiness level of a customer
  - Level of education (Higher, Secondary, Primary)
  - Rating of a movie on a scale of 0−5
  - Height in {tall, medium,short}
  - rRnkings (e.g. Taste of potato chips on scale from 1-10)
  - Letter grades in exam (A,B,C,D etc.)
- We can summarize ordinal data with percentiles, frequencies, median, mean, etc. For visualization, we can use pie charts and bar charts.

CATEGORICAL DATA
3. BINARY

- Special type of categorical data called binary
- Binary data types only have two values – yes or no.
- This can be represented in different ways such as “True” and “False” or 1 and 0
- Binary data is used heavily for classification machine learning models.

DATA TYPES


Data
Numerical
Quantitative
Categorical
Qualitative
Discrete
Continuous
Nominal
Ordinal
Binary
Interval
Ratio

NUMERICAL /QUANTITATIVE DATA

- Expressed as a number, so it can be quantified
- Represents the numerical value integer or real values
  - Height of a person
  - Price of a product
  - IQ of a person
  - Number of lessons in this course
- **Discrete data**
- **Continuous data**

NUMERICAL DATA
1. DISCRETE DATA

- Has only a finite or countably infinite set of values
- Data is discrete if the values of data are distinct and separate
- Data can only take on certain values
- This type of data can't be measured but it can be counted
  - Zip codes
  - Set of words in a collection of documents
  - Number of heads in 100 tosses of a coin flip
  - Number of students in a classroom
  - Number of cars in a showroom

  Often represented as integer variables

  We can use statistical methods such as mean, median, quartiles, Box plots, and Histograms to describe numerical data.

NUMERICAL /QUANTITATIVE DATA (CONT.)
2. CONTINUOUS DATA

- Continuous data cannot be counted, but it can be measured, It represents measurements
  - Market share Price (Money)
  - Height/weight of a person
  - Amount of rainfall
  - Speed of a car
  - Wi-Fi Frequency

It can be divided into further meaningful parts

Has real numbers as attribute values

CONTINUOUS DATA
1. INTERVAL DATA

- The data can be categorized and ranked and evenly spaced
- Interval-scaled attributes are measured on a scale of equal-size units
- The values of interval-scaled attributes have order and can be positive, 0, or negative.
- Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values
  - temperatures in celcius or Fahrenheit
  - calendar dates, the years 2002 and 2010 are eight years apart

CONTINUOUS DATA
2. RATIO DATA

- A ratio-scaled attribute is a numeric attribute with an inherent zero-point.
- That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value
- In addition, the values are ordered, and we can also compute the difference between values
  - Unlike temperatures in Celsius and Fahrenheit, the Kelvin (K) temperature scale has what is considered a true zero-point ($0^\circ K = -273.15^\circ C$): It is the point at which the particles that comprise matter have zero kinetic energy.
  - Other examples of ratio-scaled attributes include count attributes such as years of experience (e.g., the objects are employees) and number of words (e.g., the objects are documents).

PROPERTIES OF ATTRIBUTE VALUES:

- The type of an attribute depends on which of the following properties it possesses:
- Distinctness: = □
- Order: < >
- Addition: + -
- Multiplication: * /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties

PROPERTIES OF ATTRIBUTE VALUES:

- **Nominal Attribute:**
- **Distinctness**: Nominal data are categories that are distinct from each other but do not have any inherent order or numerical meaning.
- **Addition & Multiplication**: **Not Applicable**. You cannot perform addition or multiplication on nominal data because they are purely categorical with no numerical value or order. For example, you can't add "Red" + "Blue" or multiply "Apple" × "Banana."

PROPERTIES OF ATTRIBUTE VALUES:

- **Ordinal Attribute:**
- **Distinctness & Order**: Ordinal data have a clear order, but the intervals between the categories are not consistent or meaningful.
- **Addition & Multiplication**: **Not Applicable**. While ordinal data have order, you cannot meaningfully add or multiply these values because the differences between categories are not uniform. For example, the difference between "Neutral" and "Satisfied" is not necessarily the same as between "Satisfied" and "Very Satisfied," so operations like addition or multiplication are not appropriate.

PROPERTIES OF ATTRIBUTE VALUES:

- **Interval Attribute:**
- **Distinctness, Order & Addition**: Interval data have meaningful order, and the intervals between values are consistent.
- **Addition**: **Applicable**. You can say that the difference between 2:00 PM and 4:00 PM is 2 hours, just as the difference between 4:00 PM and 6:00 PM is also 2 hours.
- **Multiplication**: **Not Applicable**. However, saying that 4:00 PM is "twice as late" as 2:00 PM doesn't make sense because time on a clock does not start from a true zero point.

PROPERTIES OF ATTRIBUTE VALUES:

- **Ratio Attribute:**
- **Distinctness, Order, Addition & Multiplication**: Ratio data have all the properties of interval data but also include a true zero point, which signifies the complete absence of the quantity being measured.
- **Addition**: **Applicable**. Like interval data, you can add and subtract ratio data meaningfully. For example, 50 kg + 20 kg = 70 kg.
- **Multiplication**: **Applicable**. Ratio data have a true zero, so multiplication is meaningful. For example, if you weigh 60 kg, and someone weighs 30 kg, you can say that you are "twice as heavy" as that person because 0 kg represents no weight at all.

ACTIVITY

Attribute	Example Values
Blood Type	A, B, AB, O
Movie Rating	1 star, 2 stars, 3 stars, 4 stars, 5 stars
Temperature (°C)	10°C, 20°C, 30°C
Height (cm)	150 cm, 160 cm, 170 cm
ZIP Code	10001, 90210, 30301
Education Level	High School, Bachelor's, Master's, Ph.D.
Year of Birth	1990, 2000, 2010
Weight (kg)	50 kg, 75 kg, 100 kg
Phone Number	555-1234, 555-5678, 555-9876
Military Rank	Colonel, Major, Captain
Time of Day (24hr)	12:00, 14:30, 18:00
Distance (km)	5 km, 10 km, 20 km

ACTIVITY

Attribute	Example Values
Eye Color	Blue, Green, Brown, Hazel
Pain Scale (0-10)	0, 3, 7, 10
IQ Score	90, 100, 110, 120
Income ($)	$30,000, $50,000, $70,000
Social Security Number	123-45-6789, 987-65-4321
Satisfaction Rating	Very Dissatisfied, Neutral, Very Satisfied
Calendar Year	2015, 2020, 2025
Age (years)	25, 40, 60
Vehicle License Plate	ABC-1234, XYZ-5678
Customer Loyalty Level	Bronze, Silver, Gold
Date (MM/DD/YYYY)	01/15/2022, 12/31/2023
Weight of a Backpack (kg)	2 kg, 4 kg, 6 kg

SOLUTION

Attribute	Type	Properties	Example Values
Blood Type	Nominal	Distinctness	A, B, AB, O
Movie Rating	Ordinal	Distinctness, Order	1 star, 2 stars, 3 stars, 4 stars, 5 stars
Temperature (°C)	Interval	Distinctness, Order, Addition	10°C, 20°C, 30°C
Height (cm)	Ratio	Distinctness, Order, Addition, Multiplication	150 cm, 160 cm, 170 cm
ZIP Code	Nominal	Distinctness	10001, 90210, 30301
Education Level	Ordinal	Distinctness, Order	High School, Bachelor's, Master's, Ph.D.
Year of Birth	Interval	Distinctness, Order, Addition	1990, 2000, 2010
Weight (kg)	Ratio	Distinctness, Order, Addition, Multiplication	50 kg, 75 kg, 100 kg
Phone Number	Nominal	Distinctness	555-1234, 555-5678, 555-9876
Military Rank	Ordinal	Distinctness, Order	Private, Sergeant, Captain
Time of Day (24hr)	Interval	Distinctness, Order, Addition	12:00, 14:30, 18:00
Distance (km)	Ratio	Distinctness, Order, Addition, Multiplication	5 km, 10 km, 20 km

Attribute	Type	Properties	Example Values
Eye Color	Nominal	Distinctness	Blue, Green, Brown, Hazel
Pain Scale (0-10)	Ordinal	Distinctness, Order	0, 3, 7, 10
IQ Score	Interval	Distinctness, Order, Addition	90, 100, 110, 120
Income ($)	Ratio	Distinctness, Order, Addition, Multiplication	$30,000, $50,000, $70,000
Social Security Number	Nominal	Distinctness	123-45-6789, 987-65-4321
Satisfaction Rating	Ordinal	Distinctness, Order	Very Dissatisfied, Neutral, Very Satisfied
Calendar Year	Interval	Distinctness, Order, Addition	2015, 2020, 2025
Age (years)	Ratio	Distinctness, Order, Addition, Multiplication	25, 40, 60
Vehicle License Plate	Nominal	Distinctness	ABC-1234, XYZ-5678
Customer Loyalty Level	Ordinal	Distinctness, Order	Bronze, Silver, Gold
Date (MM/DD/YYYY)	Interval	Distinctness, Order, Addition	01/15/2022, 12/31/2023
Weight of a Backpack (kg)	Ratio	Distinctness, Order, Addition, Multiplication	2 kg, 4 kg, 6 kg