Machine Learning topic: data handling basics

By

Prof. Dr. Sourav Saha

DATA: EXPERT'S LOAN APPROVAL RECORDS

	A	B	C	D	E	F	G	H	I	J	K
1	Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	Credit History	Property_Area	Status
2	LP001002	Male	No	0	No	$5849.00		60	1	Urban	Y
3	LP001003	Male	Yes	1	No	$4583.00	120		1	Rural	N
4	LP001005	Male	Yes	0	Yes	$3000.00	$66.00	60	1	Urban	Y
5	LP001006	Male	Yes	2	No	$2583.00	$120.00	60	1	Urban	Y

Can we replace the expert by creating a model to determine whether a customer loan should be approved or not based on the expert's past approval records???

	A	B	C	D	E	F	G	H	I	J	K
1	Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	Credit History	Property_Area	Status
2	LP001002	Male	No	0	No	$5849.00		60	1	Urban	Y
3	LP001003	Male	Yes	1	No	$4583.00	120		1	Rural	N
4	LP001005	Male	Yes	0	Yes	$3000.00	$66.00	60	1	Urban	Y
5	LP001006	Male	Yes	2	No	$2583.00	$120.00	60	1	Urban	Y

Types of Variable

- **Predictor / Independent**
  - Gender
  - Married
  - Dependents
  - Self_Implyed
  - Income
  - LoanAmt
  - Term
  - Credit History
  - Property_Area
- **Target / Dependent**
  - Status

The value of status will be dependent on these variables.

This is what we are trying to determine.

	A	B	C	D	E	F	G	H	I	J	K
1	Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	Credit History	Property_Area	Status
2	LP001002	Male	No	0	No	$5849.00		60	1	Urban	Y
3	LP001003	Male	Yes	1	No	$4583.00	120		1	Rural	N
4	LP001005	Male	Yes	0	Yes	$3000.00	$66.00	60	1	Urban	Y
5	LP001006	Male	Yes	2	No	$2583.00	$120.00	60	1	Urban	Y

Data Type

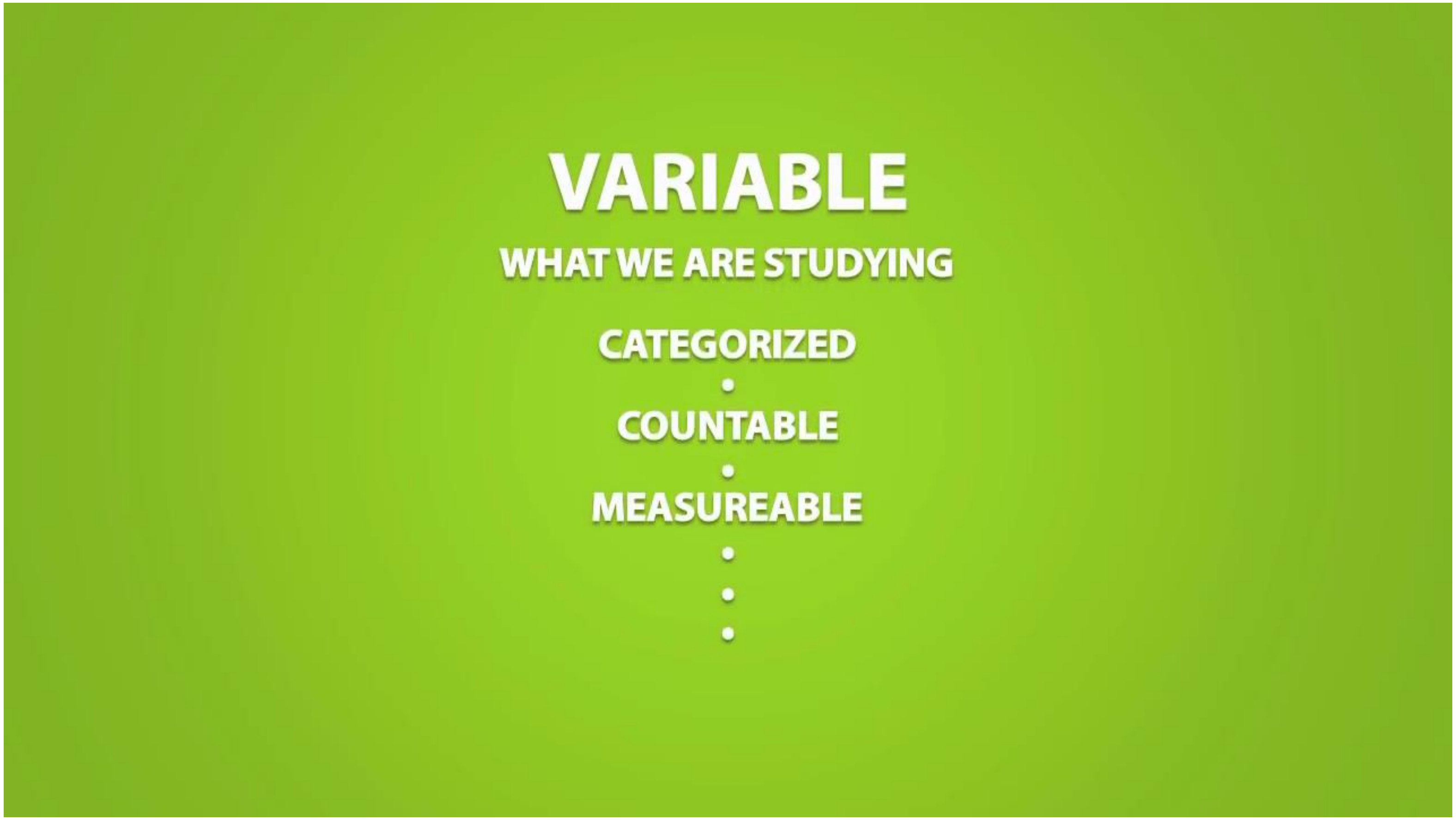
- **Character / String**
  - Gender
  - Married
  - Self_Implyed
  - Property_Area
  - Status
- **Numeric**
  - Dependents
  - Income
  - LoanAmt
  - Term
  - Credit History

	A	B	C	D	E	F	G	H	I	J	K
1	Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	Credit History	Property_Area	Status
2	LP001002	Male	No	0	No	$5849.00		60	1	Urban	Y
3	LP001003	Male	Yes	1	No	$4583.00	120		1	Rural	N
4	LP001005	Male	Yes	0	Yes	$3000.00	$66.00	60	1	Urban	Y
5	LP001006	Male	Yes	2	No	$2583.00	$120.00	60	1	Urban	Y

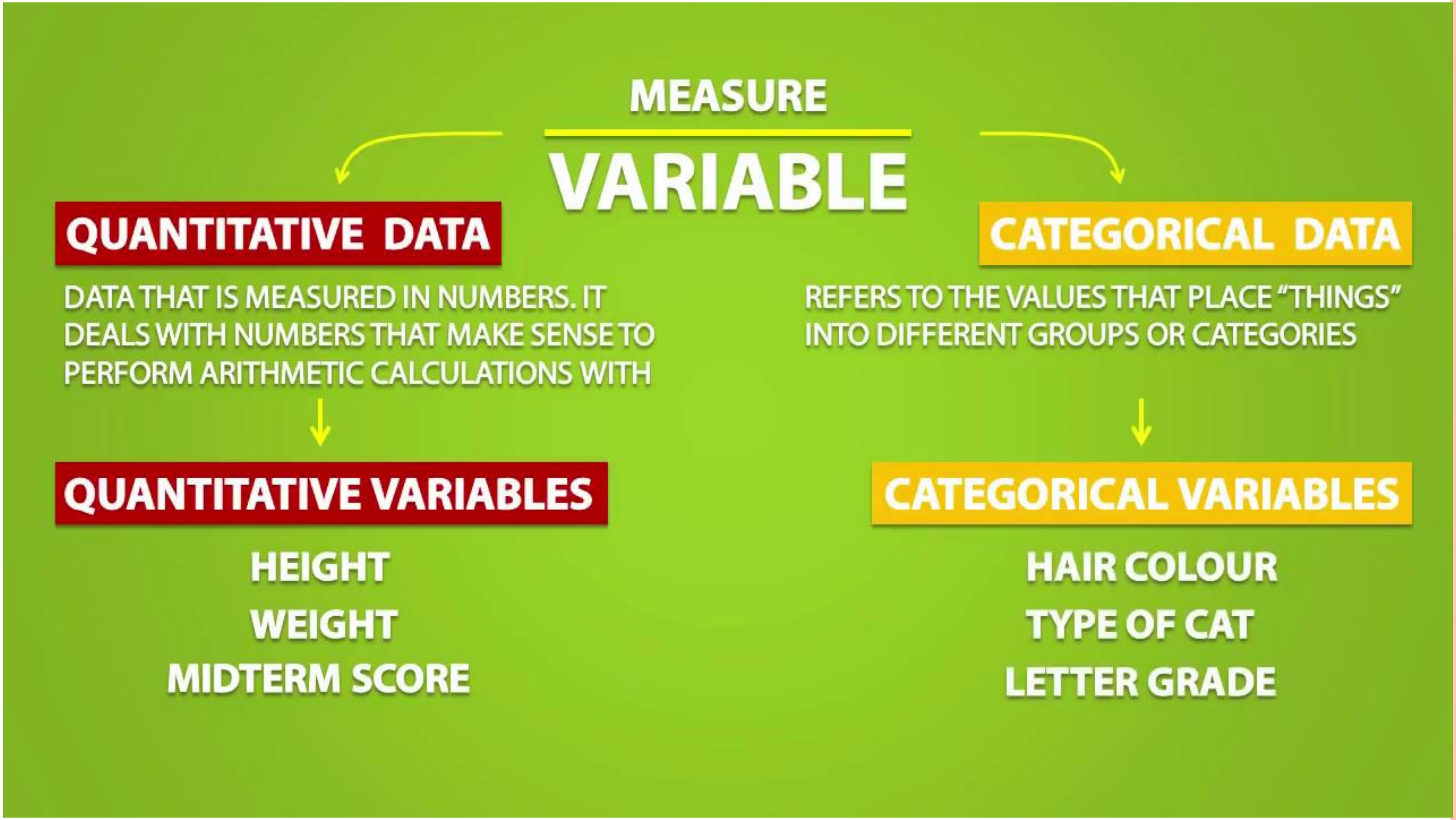
Category

- **Categorical**
  - Gender
  - Married
  - Self_Imployed
  - Credit History
  - Property_Area
  - Status
- **Continuous**
  - Dependents
  - Income
  - LoanAmt
  - Term

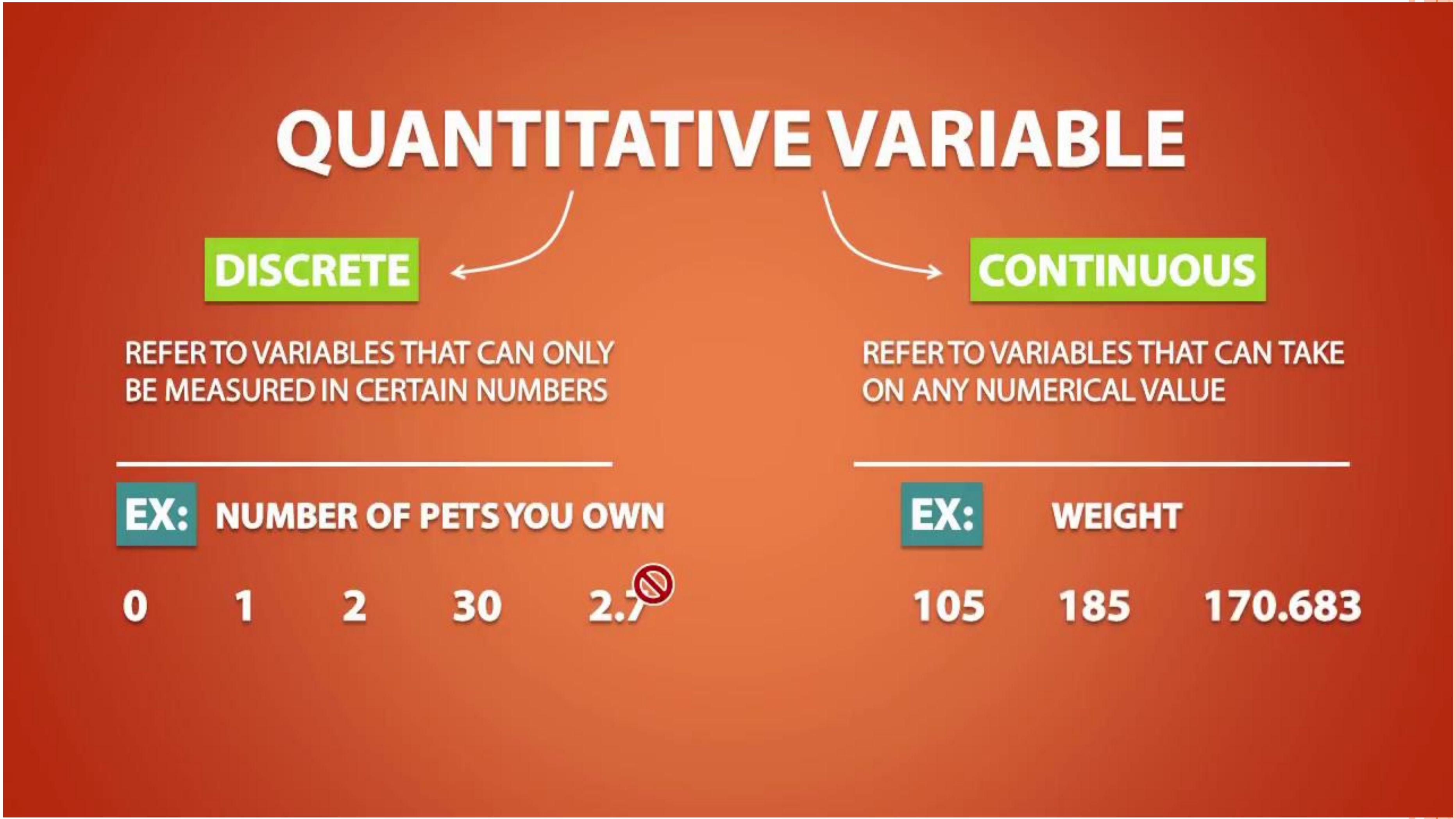
Variable data



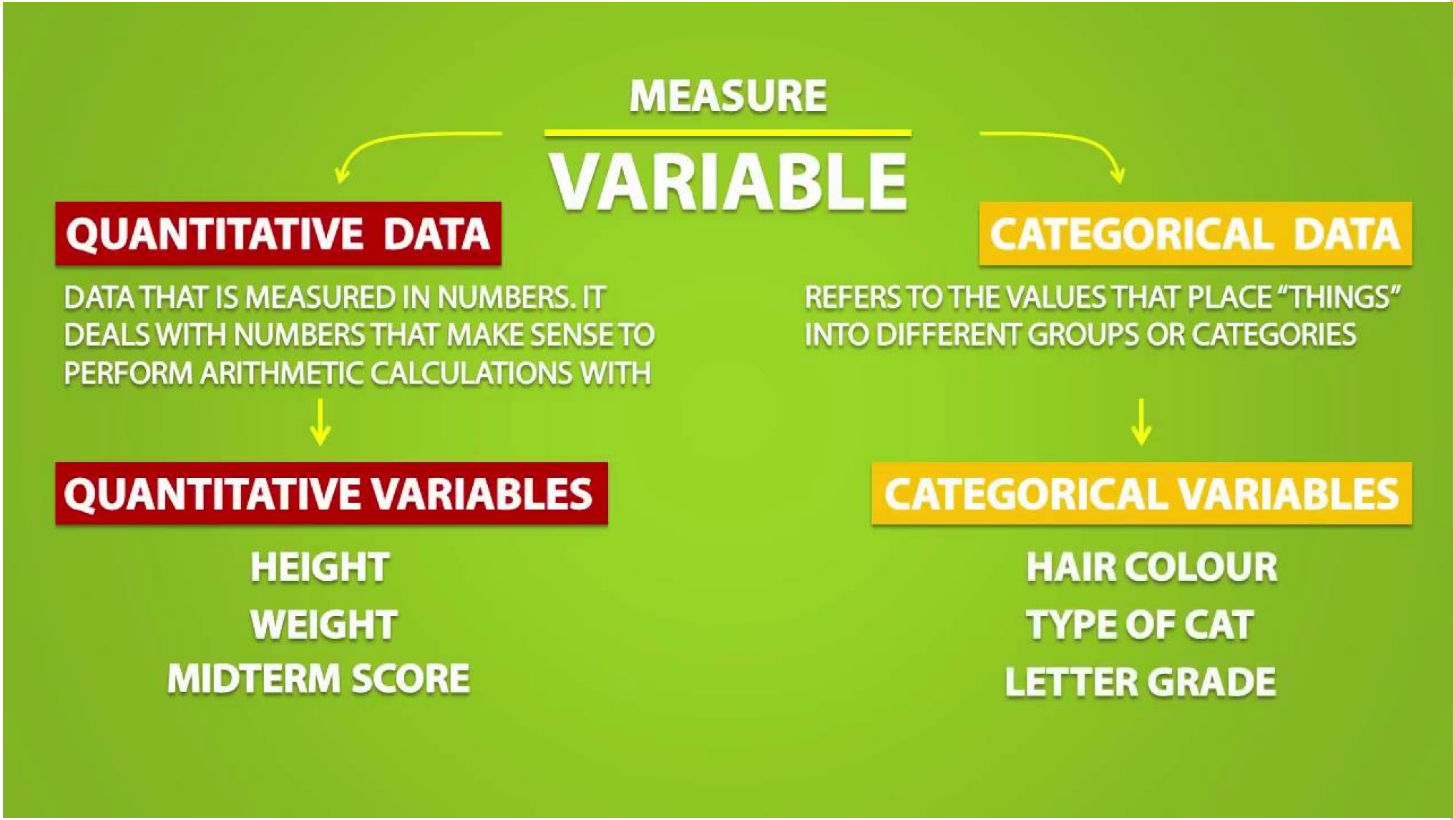
DATA TYPES


MEASURE
VARIABLE
QUANTITATIVE DATA
DATA THAT IS MEASURED IN NUMBERS. IT DEALS WITH NUMBERS THAT MAKE SENSE TO PERFORM ARITHMETIC CALCULATIONS WITH
QUANTITATIVE VARIABLES
HEIGHT
WEIGHT
MIDTERM SCORE
CATEGORICAL DATA
REFERS TO THE VALUES THAT PLACE "THINGS" INTO DIFFERENT GROUPS OR CATEGORIES
CATEGORICAL VARIABLES
HAIR COLOUR
TYPE OF CAT
LETTER GRADE

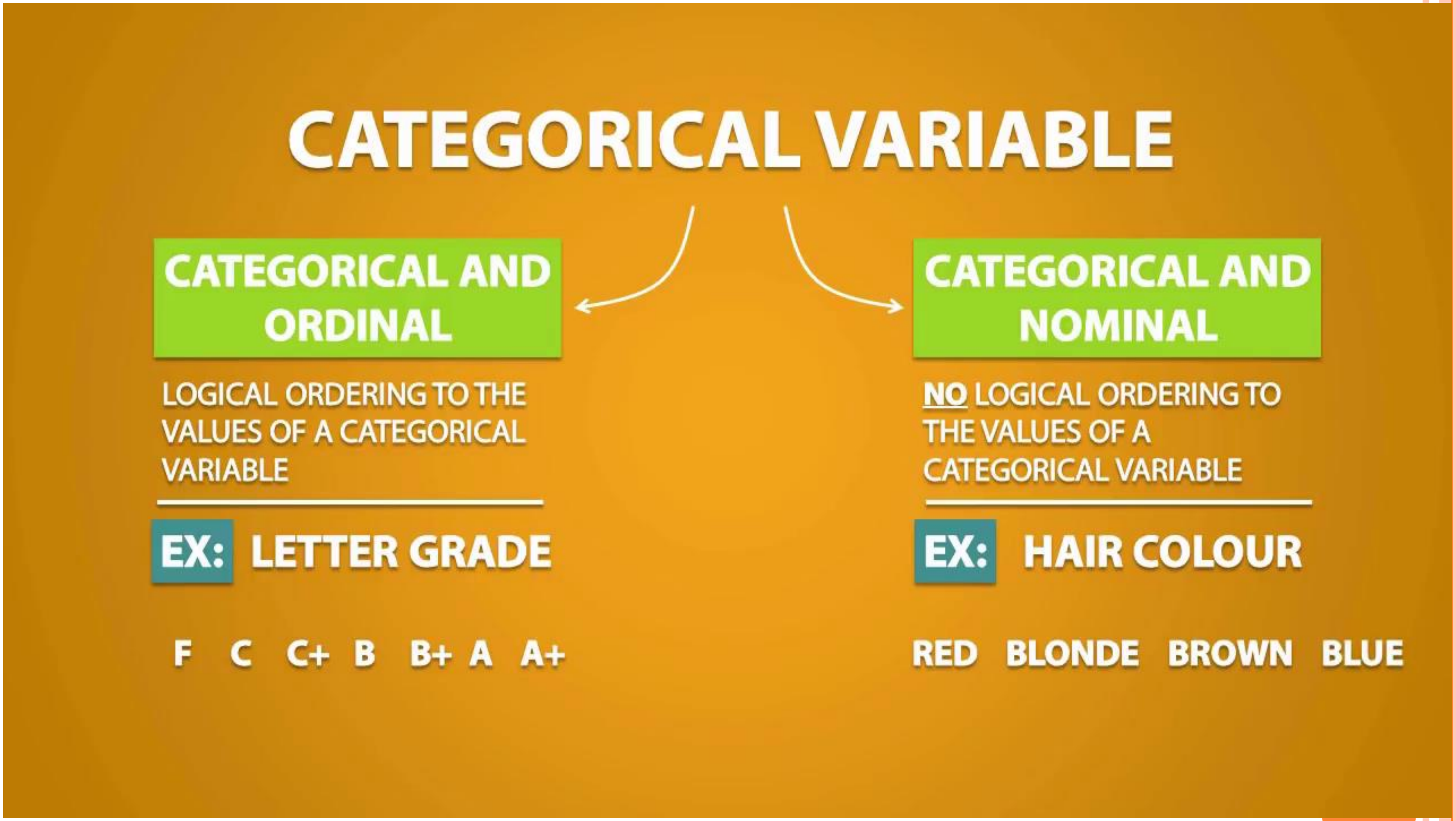
Data types


QUANTITATIVE VARIABLE
DISCRETE
REFER TO VARIABLES THAT CAN ONLY BE MEASURED IN CERTAIN NUMBERS
EX: NUMBER OF PETS YOU OWN
0 1 2 30 2.7
CONTINUOUS
REFER TO VARIABLES THAT CAN TAKE ON ANY NUMERICAL VALUE
EX: WEIGHT
105 185 170.683

DATA TYPES


MEASURE
VARIABLE
QUANTITATIVE DATA
DATA THAT IS MEASURED IN NUMBERS. IT DEALS WITH NUMBERS THAT MAKE SENSE TO PERFORM ARITHMETIC CALCULATIONS WITH
QUANTITATIVE VARIABLES
HEIGHT
WEIGHT
MIDTERM SCORE
CATEGORICAL DATA
REFERS TO THE VALUES THAT PLACE "THINGS" INTO DIFFERENT GROUPS OR CATEGORIES
CATEGORICAL VARIABLES
HAIR COLOUR
TYPE OF CAT
LETTER GRADE

DATA TYPES


CATEGORICAL VARIABLE
CATEGORICAL AND ORDINAL
LOGICAL ORDERING TO THE VALUES OF A CATEGORICAL VARIABLE
EX: LETTER GRADE
F C C+ B B+ A A+
CATEGORICAL AND NOMINAL
NO LOGICAL ORDERING TO THE VALUES OF A CATEGORICAL VARIABLE
EX: HAIR COLOUR
RED BLONDE BROWN BLUE

Qualitative vs Quantitative Data

Categorical Data

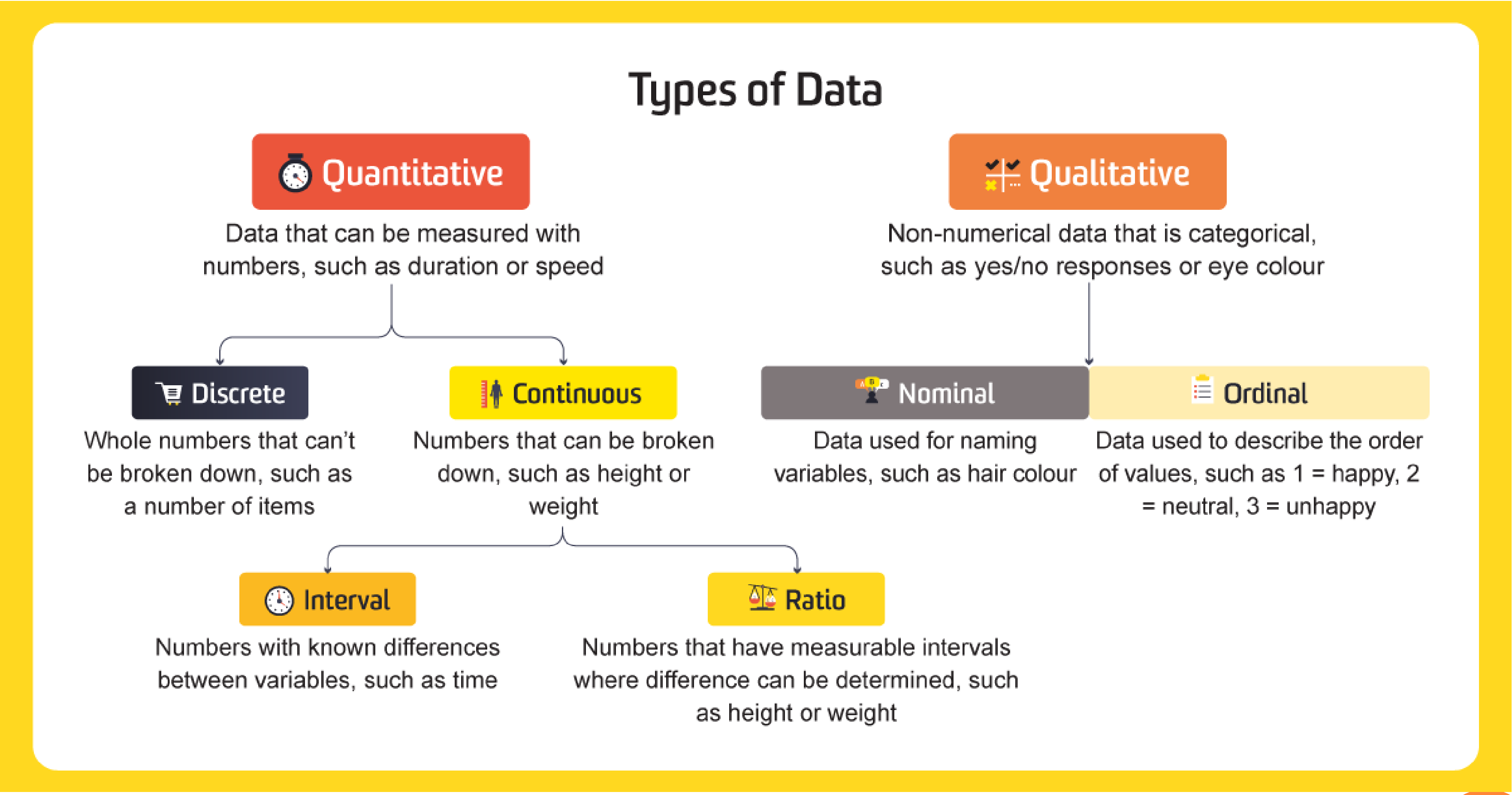
Overview:

- Deals with descriptions.
- Data can be observed but not measured.
- Colors, textures, smells, tastes, appearance, beauty, etc.
- Qualitative → Quality

Numerical Data

Overview:

- Deals with numbers.
- Data which can be measured.
- Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc.
- Quantitative → Quantity


Types of Data
Quantitative
Data that can be measured with numbers, such as duration or speed
Qualitative
Non-numerical data that is categorical, such as yes/no responses or eye colour
Discrete
Whole numbers that can't be broken down, such as a number of items
Continuous
Numbers that can be broken down, such as height or weight
Nominal
Data used for naming variables, such as hair colour
Ordinal
Data used to describe the order of values, such as 1 = happy, 2 = neutral, 3 = unhappy
Interval
Numbers with known differences between variables, such as time
Ratio
Numbers that have measurable intervals where difference can be determined, such as height or weight

Scales of Measurement

Data	Nominal	Ordinal	Interval	Ratio
Labeled	✓	✓	✓	✓
Meaningful Order	✗	✓	✓	✓
Measurable Difference	✗	✗	✓	✓
True Zero Starting Point	✗	✗	✗	✓

Classification dataset

case ID predictors target

CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

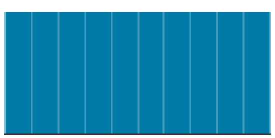
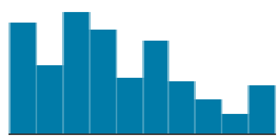
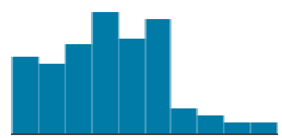
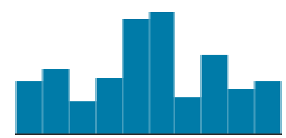
Clustering dataset

Mall_Customers.csv (3.89 KB)

Detail Compact Column 5 of 5 columns

About this file

This file contains the basic information (ID, age, gender, income, spending score) about the customers

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (...
Unique ID assigned to the customer	Gender of the customer	Age of the customer	Annual Income of the customee	Score assigned by the mall based on customer behavior and spending nature
 1 200	Female 56% Male 44%	 18 70	 15 137	 1 99
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77

Regression dataset

Multiple features (variables).

Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

Notation:

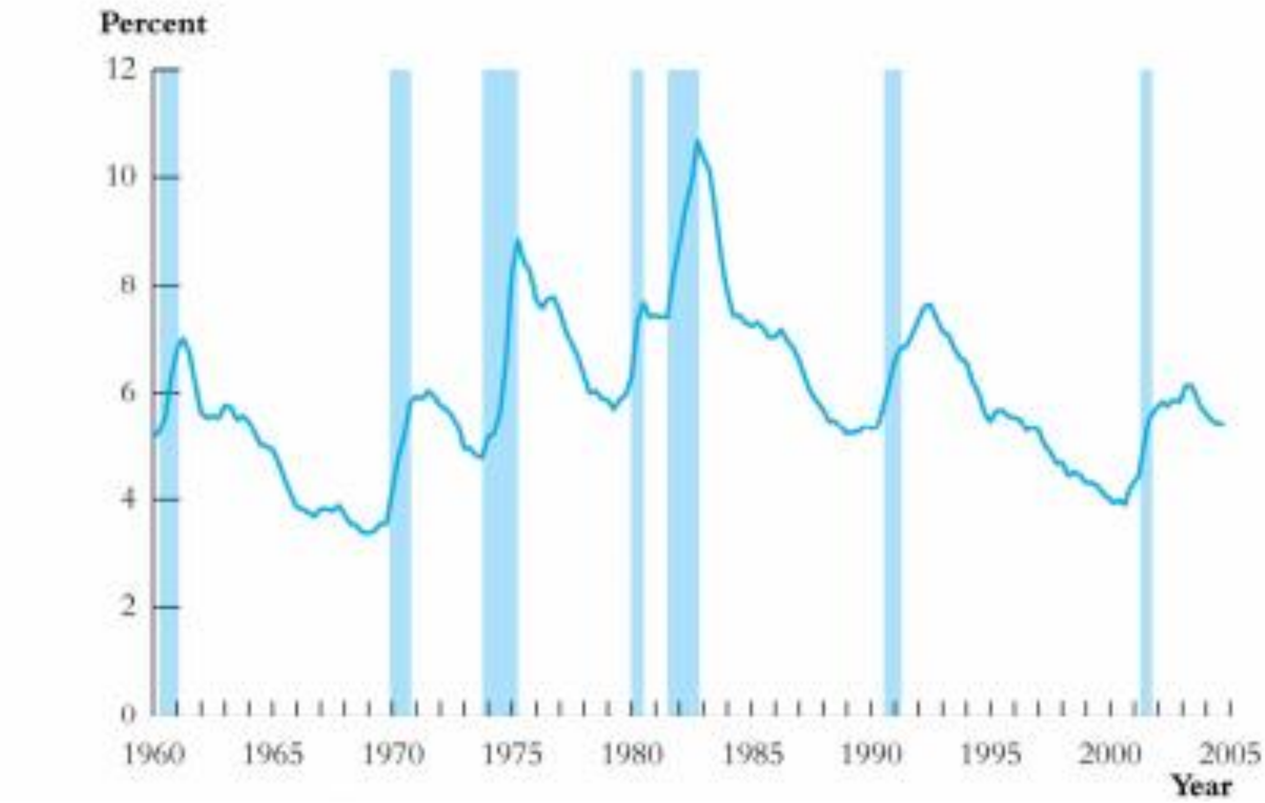
n = number of features $n = 4$

$x^{(i)}$ = input (features) of i^{th} training example.

$x_j^{(i)}$ = value of feature j in i^{th} training example.

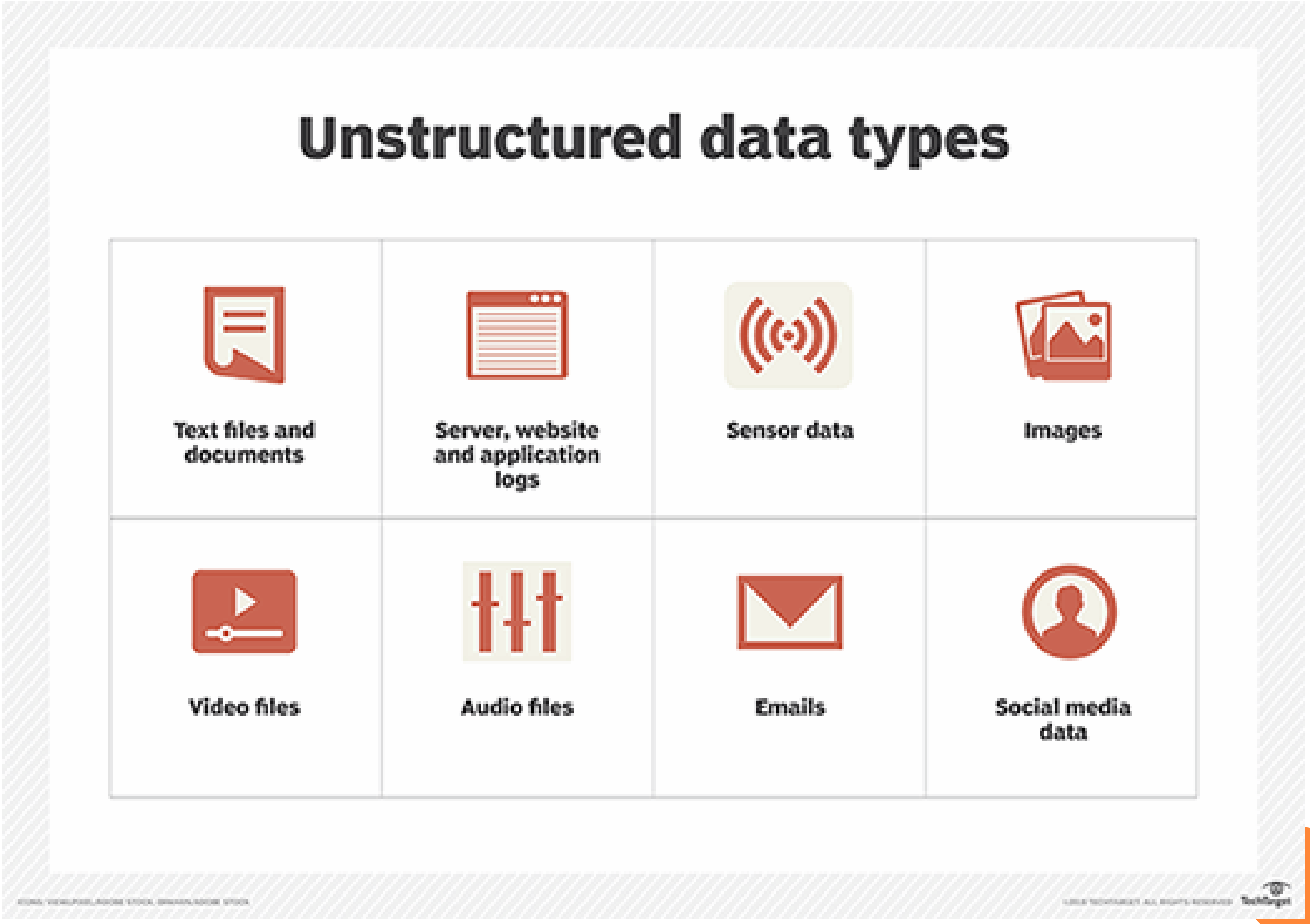
Dataset for time series analysis

Example #2: US rate of unemployment



Unstructured data

- Video
- Image
- Audio
- Text

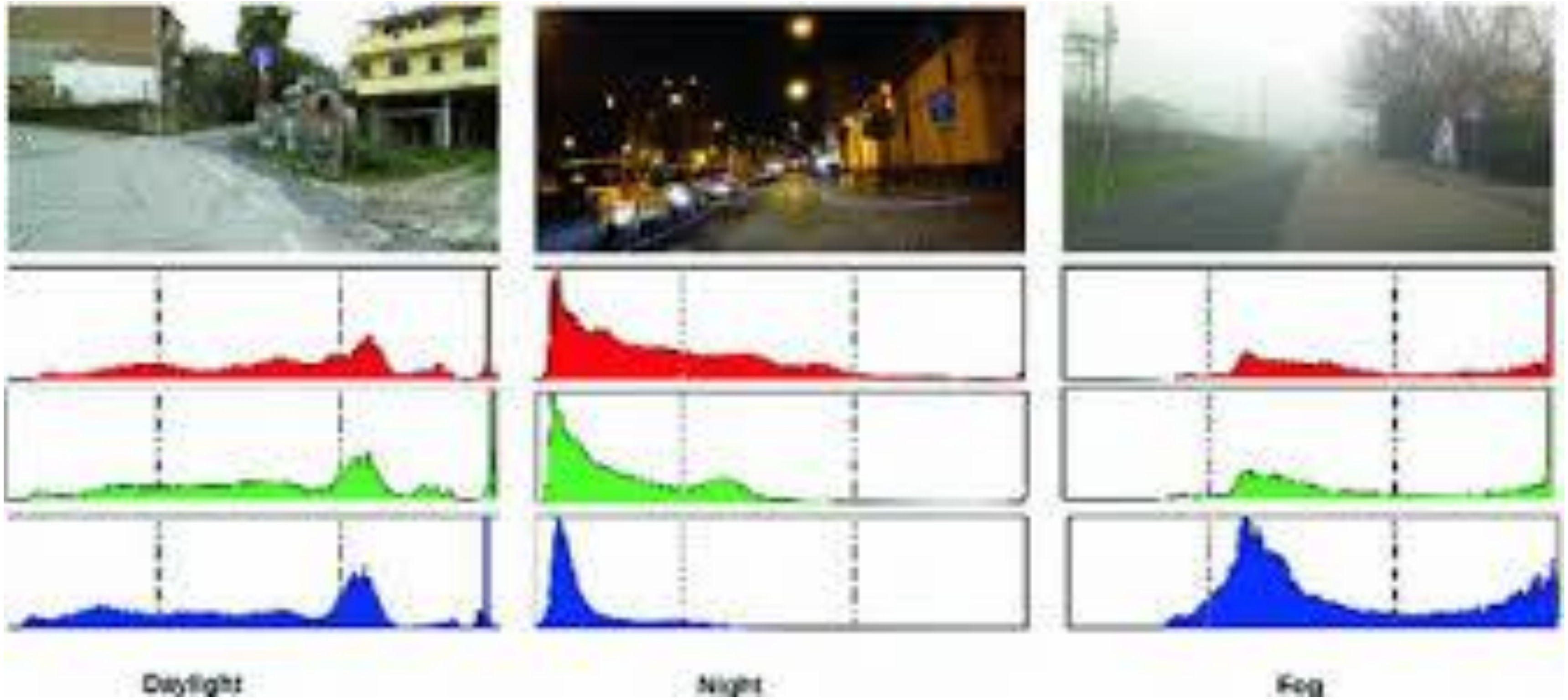


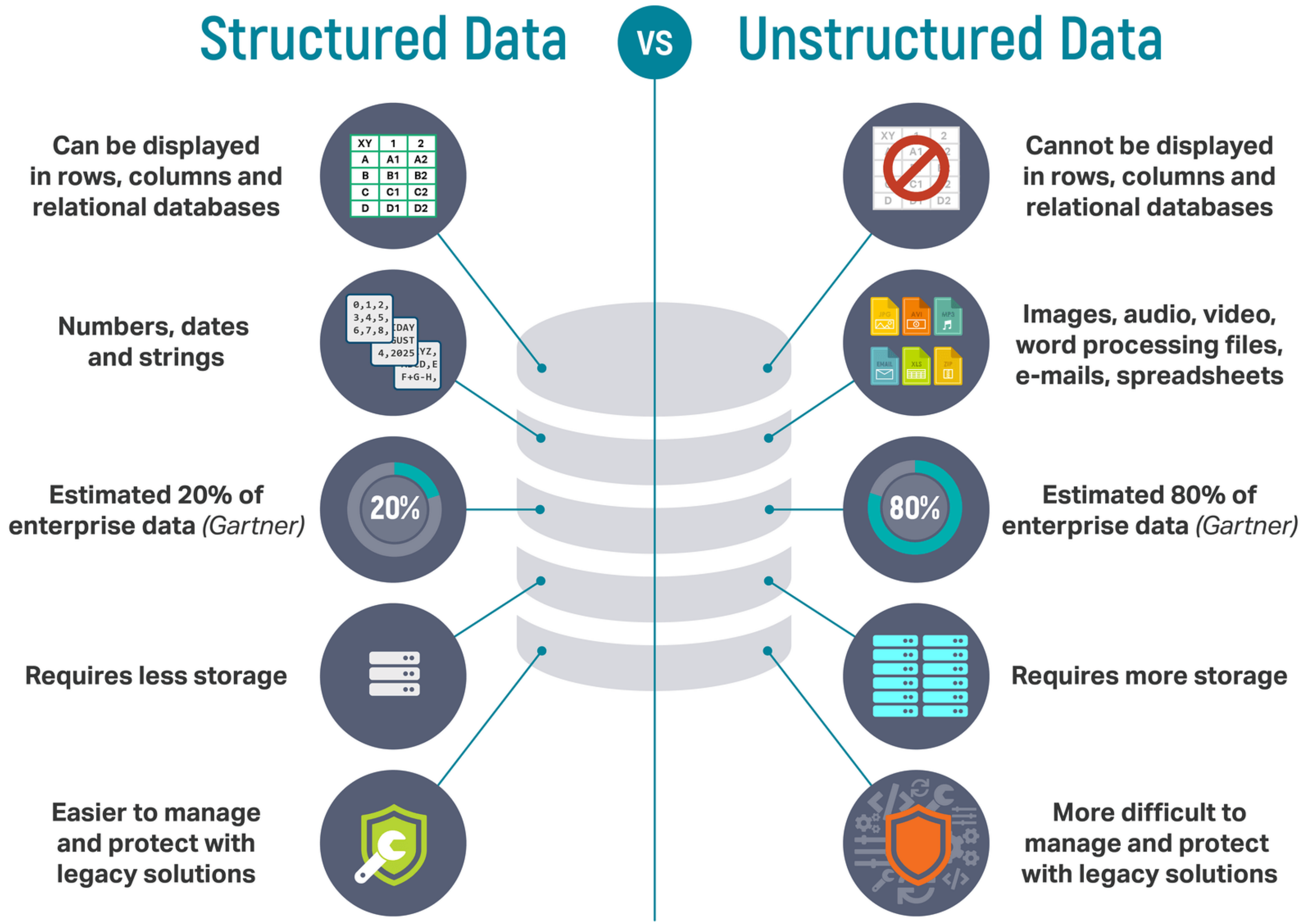
BAG-OF-WORDS FOR TEXT

Review1: This movie is very scary. Review2: This movie is scary and not slow. Review3: This movie is spooky and good.

Term	Review 1	Review 2	Review 3
This	1	1	1
movie	1	1	1
is	1	2	1
very	1	0	0
scary	1	1	0
and	1	1	1
long	1	0	0
not	0	1	0
slow	0	1	0
spooky	0	0	1
good	0	0	1

Histogram for image




Structured Data
VS
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Univariate data analysis – E.g. age, weight, salary

- Understanding central tendency: Mean, Median, Mode

Mean	Average value
Median	Middle value
Mode	Most frequent value

Mean

Mean

It is the sum of the value observation divide by the sample size.

The mean of the values 5,6,6,8,9,9,9,9,10,10 is

(5+6+6+8+9+9+9+9+10+10)/10 = 8.1

Limitation :

It is affected by enormous values. Very large or very small numbers can distort the answer

MEDIAN

It is the middle value of the list. It splits the data in half. Half of the data are above the median; half of the data are below the median.

Advantage :

*It is **NOT** affected by enormous values. large or small numbers doesn't make a impact.*

Median Odd
23
21
18
16
15
13
12
10
9
7
6
5
2

	Median Even
	40
	38
	35
	33
	32
	30
28	29
	27
	26
	24
	23
	22
	19
	17

MODE

It is the value that occurs most frequently in a data-set.

Advantage :

It can be used when the data is not numerical.

Disadvantage :

1. There may be no mode at all if none of the data is the same

2. There may be more than one mode

Mode
5
5
5
4
4
3
2
2
1

Univariate data analysis

- Understanding data Spread/Variability

Range	Difference between max and min in a distribution
Standard Deviation	Average distance of scores in a distribution from their mean
Variance	Square of the standard deviation
Skewness	Degree to which scores in a distribution are spread out.
Kurtosis	Flatness or peakness of the curve

Univariate Data Analysis: Range

Range: defined as a single number representing the spread of the data

***Range* = maximum score — minimum score**

In order to figure out the range, A) arrange your data set in order from lowest to highest and B) subtract the lowest number from the highest number.

A) When arranged in order, **4, 6, 3, 7, 9, 4, 2, 1, 4, 2** becomes: **1, 2, 2, 3, 4, 4, 4, 6, 7, 9**

B) The **lowest number** is **1** and the **highest number** is **9**.

Therefore, R = 9–1 = 8

Univariate data analysis: mean deviation

The mean deviation gives us a measure of the typical difference (or deviation) from the mean. If most data values are very similar to the mean, then the mean deviation score will be low, indicating high similarity within the data. If there is great variation among scores, then the mean deviation score will be high, indicating low similarity within the data.

$$MD = \frac{\sum |x_i - \bar{x}|}{N}$$

Mean Deviation: An Example

Data: X = {6, 10, 5, 4, 9, 8} $\bar{X} = 42 / 6 = 7$

$\bar{X} - X_i$	Abs. Dev.
7 – 6	1
7 – 10	3
7 – 5	2
7 – 4	3
7 – 9	2
7 – 8	1
Total:	12

1. Compute X (Average)
2. Compute X – X and take the Absolute Value to get Absolute Deviations
3. Sum the Absolute Deviations
4. Divide the sum of the absolute deviations by N

12 / 6 = 2

UNIVARIATE DATA ANALYSIS: VARIANCE & STANDARD DEVIATION

Variance is defined as a number indicating how spread out the data is. **Standard Deviation** is the square root of the variance.

B8 f_x =(E2+E3+E4+E5+E6)/COUNT(A2:A6)

	A	B	C	D	E
1	No.of Items (N)	Observations(x)	μ	x- μ	(x- μ)^2
2	1	50		0	0
3	2	55		5	25
4	3	45	50	-5	25
5	4	60		10	100
6	5	40		-10	100
7					
8	σ^2	50			
9					

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

Measure of consistency

A large variance indicates that numbers in the set are far from the mean and far from each other. A small variance, on the other hand, indicates the opposite i.e more consistency

BATSMAN A : 70, 80, 75, 90;

mean = 78; variance = 54; std-dev = 7

BATSMAN B : 25, 175, 85, 35;

mean = 80; variance = 3525; std-dev = 59

BATSMAN A is more consistent as the variance of his scores is less as compared to BATSMAN B

UNIVARIATE DATA ANALYSIS: COEFFICIENT OF VARIATION

The coefficient of variation (CV) is a relative measure of variability that indicates the size of a standard deviation in relation to its mean. It is a standardized, unitless measure that allows you to compare variability between disparate groups and characteristics. It is also known as the relative standard deviation (RSD).

$$CV = \frac{Standard\ deviation}{Mean}$$

A data set of [100, 100, 100] has constant values. Its standard deviation is 0 and average is 100, giving the coefficient of variation as
0 / 100 = 0
A data set of [90, 100, 110] has more variability. Its sample standard deviation is 10 and its average is 100, giving the coefficient of variation as
10 / 100 = 0.1
A data set of [1, 5, 6, 8, 10, 40, 65, 88] has still more variability. Its standard deviation is 30.78 and its average is 27.9, giving a coefficient of variation of
30.78 / 27.9 = 1.10

Example of Coefficient of Variation

A Restaurant owner wants to open a new outlet. There are two territories to choose from. The choice mainly depends on the rental value, and the best option would be to open the restaurant in the territory that has lesser variation in the rentals.

Territory A	Territory B
Average Rental is around Rs 120,000	Average rental is around Rs 200,000
Standard Deviation is 2,000	Standard deviation is 3,000
Co-efficient of Variation = 2000/120000 = 0.016 or 1.60%	**Coefficient of variation = 3000/200000 = 0.015 or 1.50%**

The data available to us is as follows.
(i) If you look at the rental values, Territory A seems to be a better bet as the average rental cost is considerably lower when compared to Territory B.
(ii) However, it is not the right option because the variation in the rental values is lower in Territory B as compared to Territory A
(iii) The CV of Territory B is 1.50% whereas the CV of Territory A is 1.60%
(iv) **Thus, the better option for the restaurant owner is to open the QSR in Territory B.**

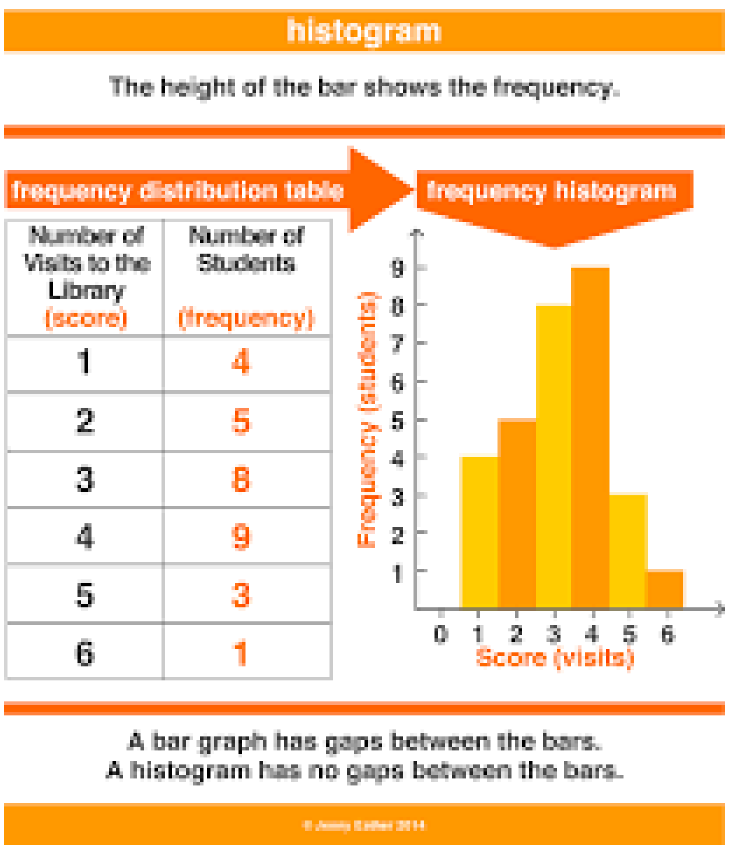
A lower value of Coefficient of Variation is preferable because of the lesser degree of volatility. Thus, the lower the CV, the better is the option.

Data visualization

- Histogram
- Box Plot
- Scatter Plot

Univariate data analysis

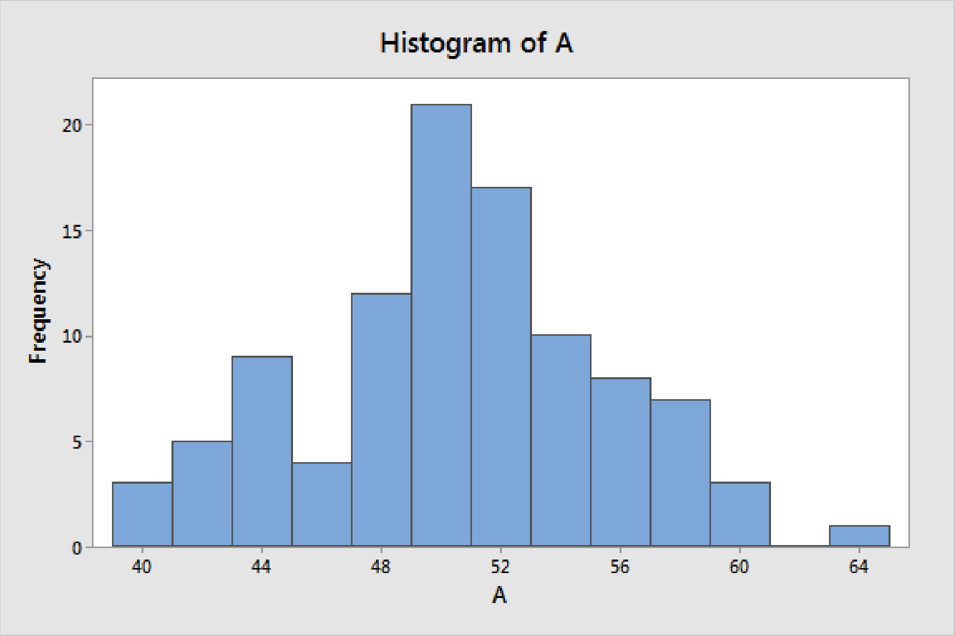
- Understanding data distribution: Histogram



A histogram is a graphical representation of the distribution of numerical data.

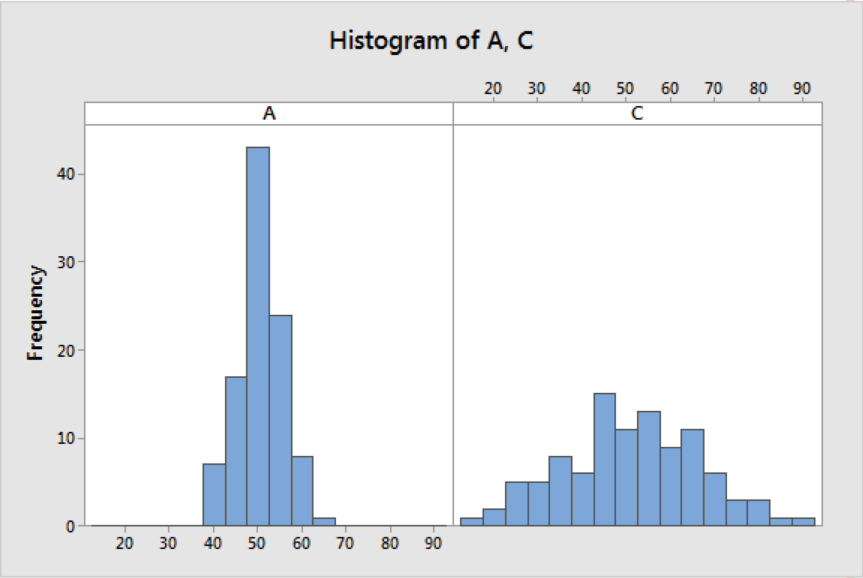
Histograms & central tendency

Use histograms to understand the center of the data. In the histogram below, you can see that the center is near 50. Most values in the dataset will be close to 50, and values further away are rarer. The distribution is roughly symmetric and the values fall between approximately 40 and 64.



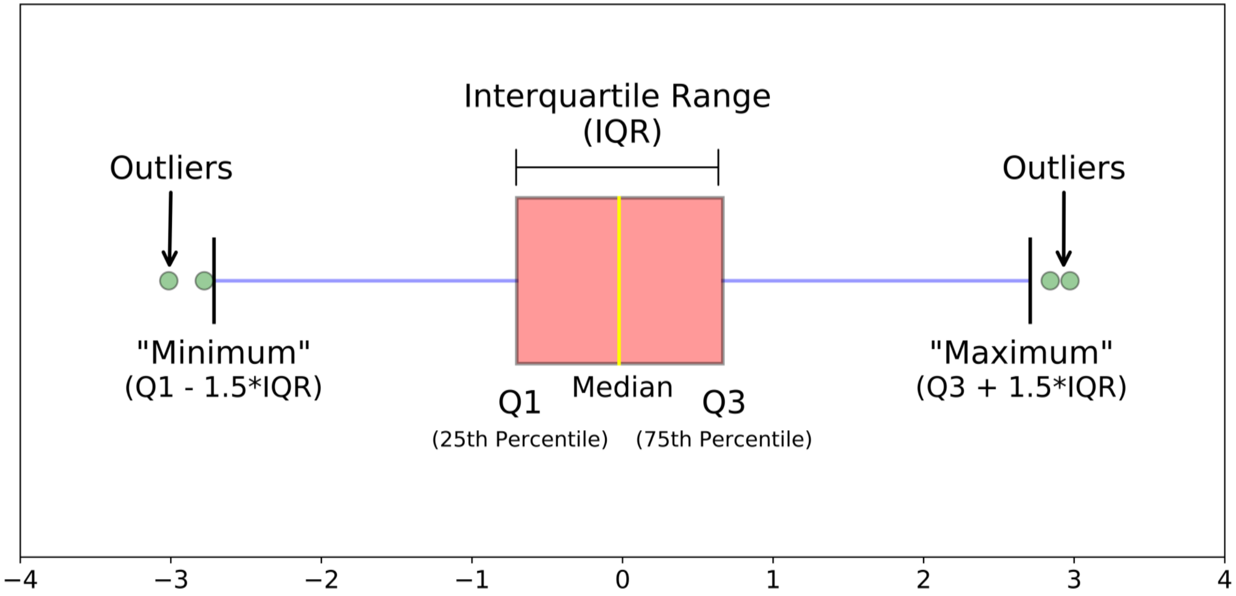
Histograms & variability

Suppose you hear that two groups have the same mean of 50. It sounds like they're practically equivalent. However, after you graph the data, the differences become apparent, as shown below. The histograms center on the same value of 50, but the spread of values is notably different. The values for group A mostly fall between 40 – 60 while for group B that range is 20 – 90. The mean does not tell the entire story! At a glance, the difference is evident in the histograms.



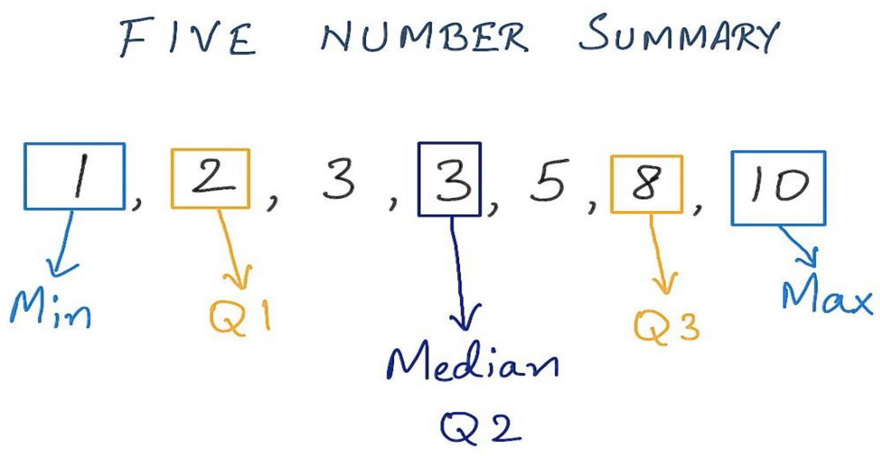
Box-plot

Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).



BOXPLOT: FIVE NUMBER SUMMARY

- Mean
- Median
- Quartile (Q)
- Minimum
- Maximum



Example: 5, 8, 3, 2, 1, 3, 10

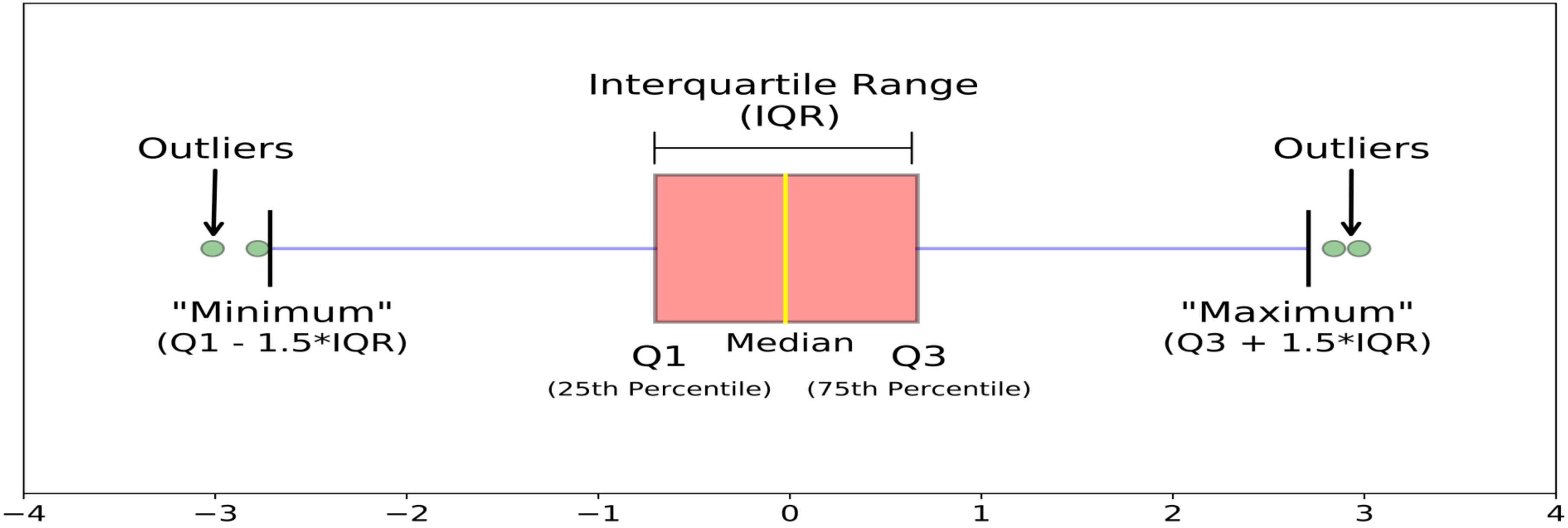
If the data-set has an even number of values, the value of Q2 (median), will be the mean of the middle 2 values. The value of Q1 will be the median of all values to the left of calculated Q2 and the value of Q3 will be the median of all values to the right of calculated Q2.

Once the Five Number Summary values have been computed, finding the Range and Interquartile Range is easy.

Range = Maximum — Minimum = 10–1 = **9**

Interquartile Range (IQR) = Q3 — Q1 = 8–2 = **6**

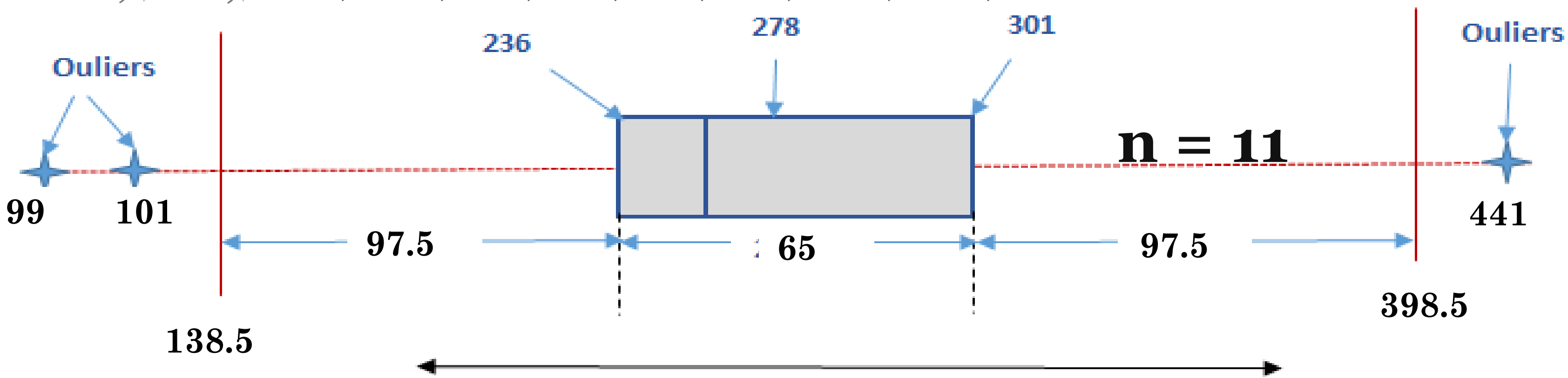
BOX-PLOT



median (Q2): the middle value of the dataset.
first quartile (Q1): the middle number between the smallest number (not the "minimum") and the median of the dataset.
third quartile (Q3): the middle value between the median and the highest value (not the "maximum") of the dataset.
interquartile range (IQR): 25th to the 75th percentile.
whiskers (shown in blue); outliers (shown as green circles)
"maximum": Q3 + 1.5*IQR
"minimum": Q1 -1.5*IQR

BOX-PLOT EXAMPLE:

99, 101, 236, 269,271,278,283,291, 301, 303, 441



Median (Q2) $= \frac{1}{2}\ (11 + 1)\ th\ term = 6th\ Term$

Q2 = 278

Lower Quartile (Q1) $= \frac{1}{4}\ (11 + 1)\ th\ term = 3rd\ Term$

Q1 = 236

Upper Quartile (Q3) $= \frac{3}{4}\ (11 + 1)\ th\ term = 9th\ Term$

Q3 = 301

Quartile Range (IQR) = Q3 – Q1 = 301 – 236

IQR = 65

Lower Limit = Q1 – 1.5 IQR = 236 – 1.5 * 65

Lower Limit = 138.5

Upper Limit = Q3 + 1.5 IQR = 301 + 1.5 * 65

Upper Limit = 398.5

Handing outliers

Height
5.5
5
5.8
5.11
8
5
4

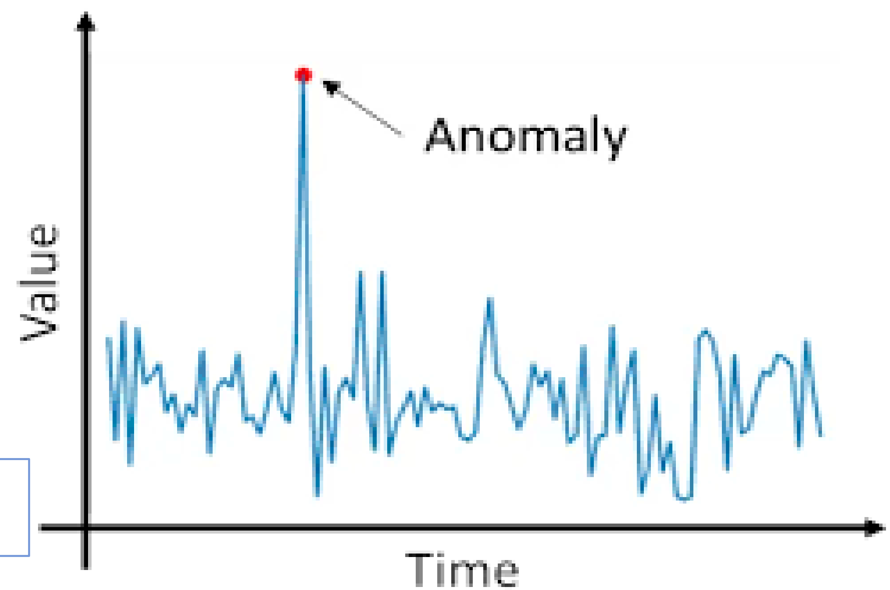
$\mu = 5.4$
$\sigma^2 = 1.3$

Height
5.5
5
5.8
5.11
~~8~~
5
~~4~~

Remove

Height
5.5
5
5.8
5.11
5.4
5
5.4

Replace



HANDING MISSING VALUES

Height
5.5
5
5.8
5.11
?
5
?

Height
5.5
5
5.8
5.11
5

Height
5.5
5
5.8
5.11
5.4
5
5.4

Remove | Replace

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

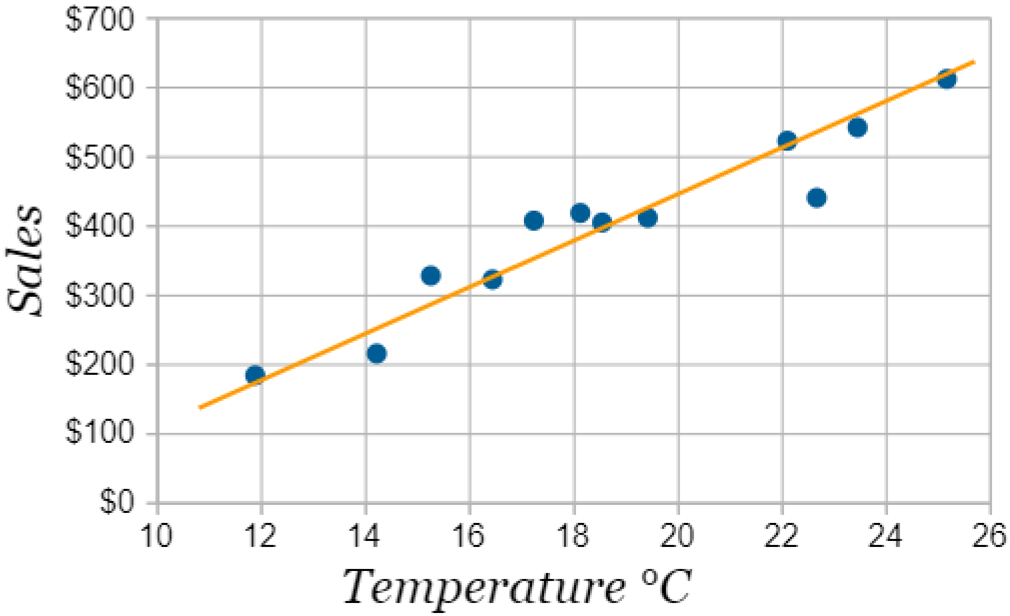
mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

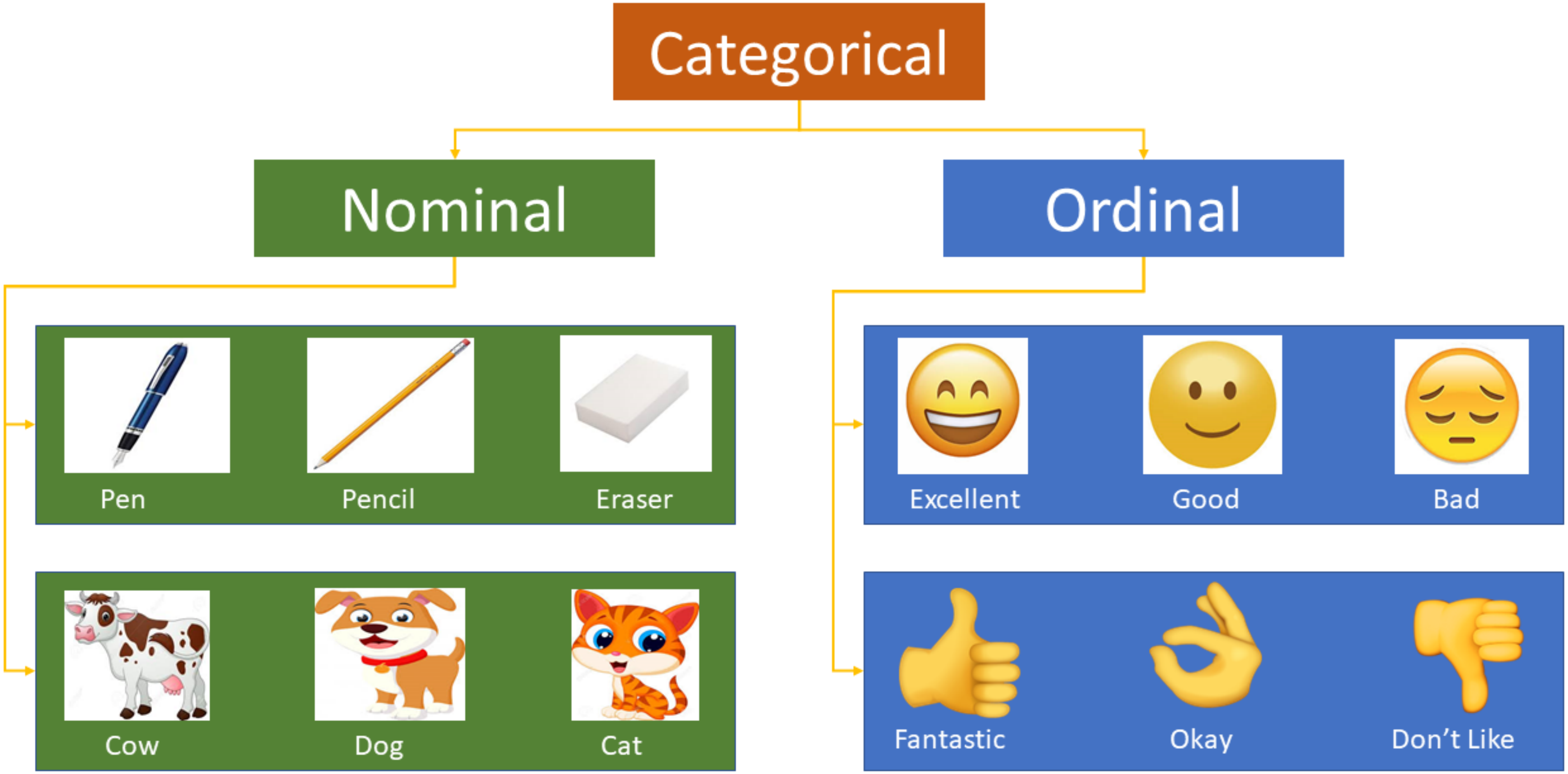
Gender	Manpower	Sales
M	25	343
~~F~~	~~.~~	~~280~~
M	33	332
~~M~~	~~.~~	~~272~~
~~F~~	~~25~~	~~.~~
M	29	326
	~~26~~	~~259~~
M	32	297

Scatter plot

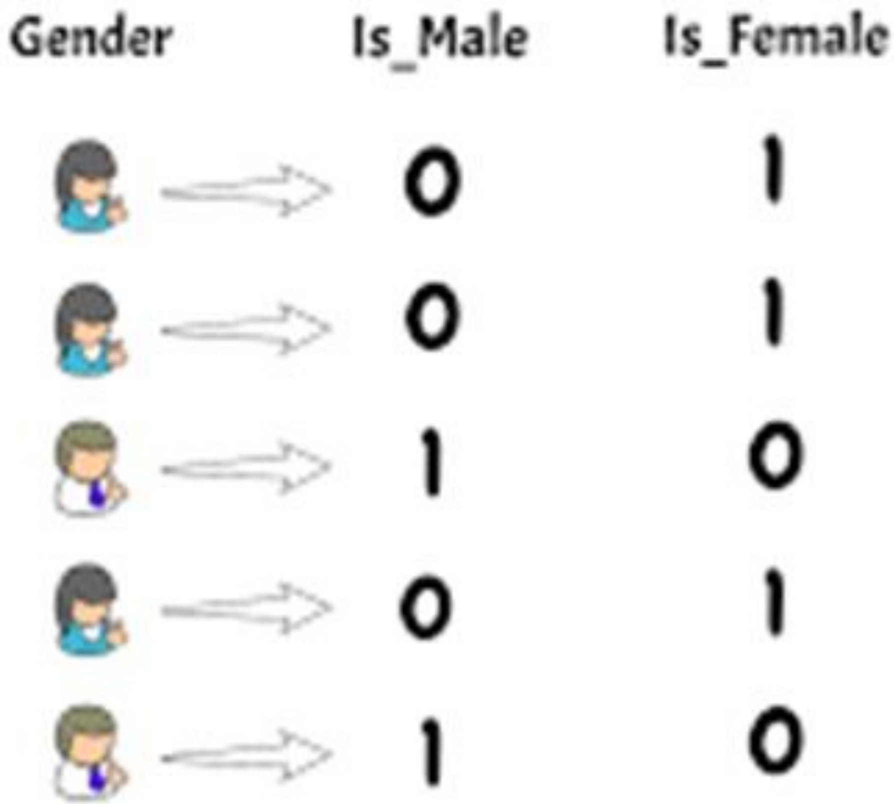
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Categorical Data encoding



Categorical Data encoding: Nominal



Categorical Data encoding: ordinal

	Temperature	Temp_Ordinal
0	Hot	3
1	Cold	1
2	Very Hot	4
3	Warm	2
4	Hot	3
5	Warm	2
6	Warm	2
7	Hot	3
8	Hot	3
9	Cold	1

If we consider in the temperature scale as the order, then the ordinal value should from cold to “Very Hot. “ Ordinal encoding will assign values as

Cold(1) <Warm(2)<Hot(3)<”Very Hot(4)

Usually, we Ordinal Encoding is done starting from 1.

Categorical Data encoding: one hot

	weather	..
0	sunny	..
1	rainy	..
2	rainy	..
3	sunny	..

$\xrightarrow{\hspace{2em}}$ *OH encoding*

	is_sunny	is_rainy	..
0	1	0	..
1	0	1	..
2	0	1	..
3	1	0	..

Categorical Data encoding: ordinal plus One hot

Temperature
Hot
Cold
Very Hot
Warm
Hot
Warm
Warm
Hot
Hot
Cold

Order
1
2
3
4
1
4
4
1
1
2

Binary
001
010
011
100
001
100
100
001
001
010

Temperature_0	Temperature_1	Temperature_2
0	0	1
0	1	0
0	1	1
1	0	0
0	0	1
1	0	0
1	0	0
0	0	1
0	0	1
0	1	0

Data Assessment

The first step in data understanding is a Data Assessment. This should be undertaken before the kick-off of a project as it is an important step to validate its feasibility. This task evaluates what data is available and how it aligns to the business problem. It should answer the following questions:

- What data is available?
- How much data is available?
- Do you have access to the ground truth, the values you're trying to predict?
- What format will the data be in?

Data Assessment

- Count the number of records — is this what you expected?
- What are the datatypes? Will you need to change these for a machine learning model?
- Look for missing values — how should you deal with these?
- Verify the distribution of each column — are they matching the distribution you expect (e.g. normally distributed)?

Data Assessment

- Search for outliers — are there anomalies in your data? Are all values valid (e.g. no ages less than 0)?
- Validated if your data is balanced — are different groups represented in your data? Are there enough examples of each class you wish to predict?
- Is there bias in your data — are subgroups in your data treated more favorable than others?

Thank You