

Week1 – Introduction to Data Science

Data Sources

Evolution of
Technology

IOT

Social Media

Other factors



1,736,111 pictures



347,222 tweets



204,000,000
emails



4,166,667 likes &
200,000 photos



300 hours of video
uploaded

Data Sources



Walmart Use Case

Data Analysis At Walmart

Halloween and cookie sales



Data scientist at Walmart found a connection between Halloween and the sales of cookies.

Data Analysis At Walmart

Hurricane and strawberry pop tarts



Data scientist at Walmart found that sales of Strawberry pop-tarts increased by 7 times before a Hurricane.

Data Analysis At Walmart

Social media and cake pops



Walmart is leveraging social media data to find about the trending products so that they can be introduced to the Walmart stores across the world

What Is Data Science?

What is Data Science?



“Torture the data, and it will confess to anything.”
~ Ronald Coase, Economics, Nobel Prize

Data Science is the process of extracting knowledge and insights from data by using *scientific methods*.



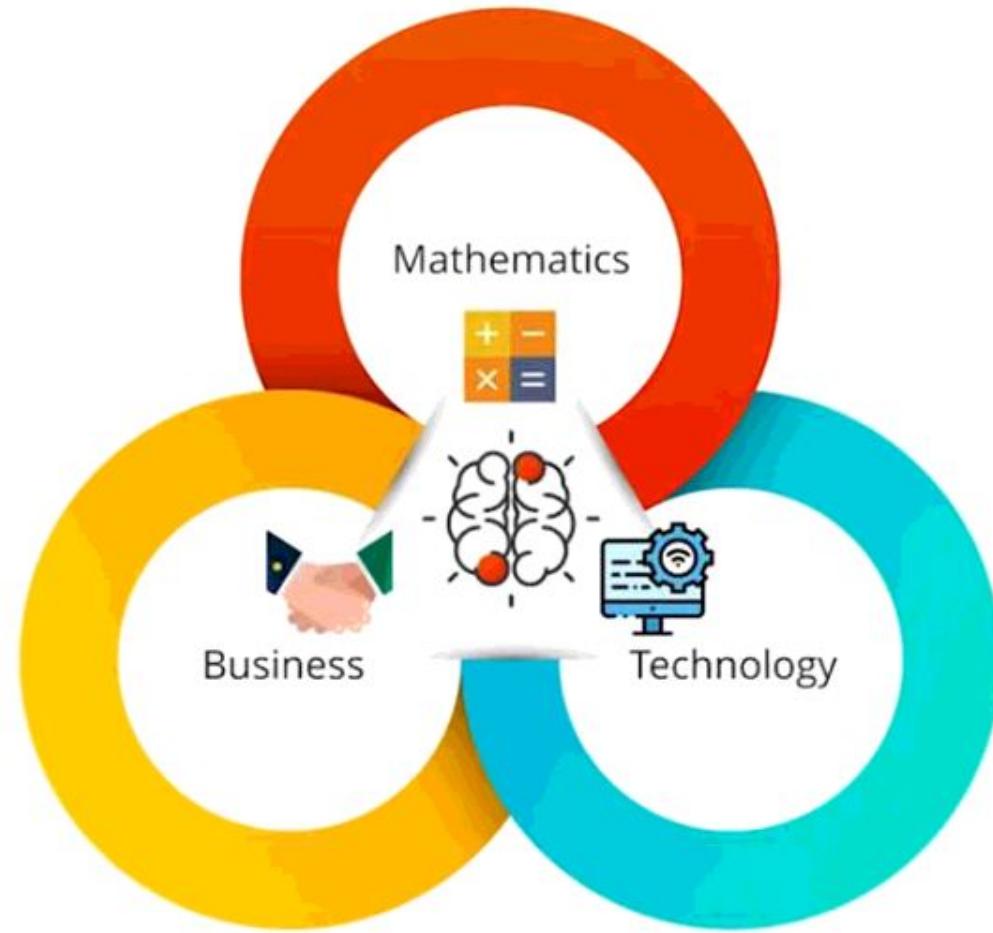
Scientific methods:

Programming + Statistics + Business



Who Is A Data Scientist?

Who Is A Data Scientist?



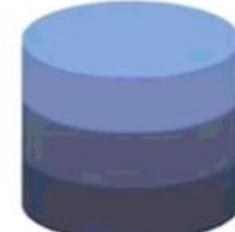
Data Science – Skill Set



Statistics



Programming languages



Data extraction & processing



Data wrangling & exploration



Machine Learning



Big Data processing frameworks



Data visualisation

Data Science Job Roles

Data Science Job Roles



Data Scientist



Data Analyst



Data Architect



Data Engineer



Statistician



Database
Administrator



Business Analyst



Data & Analytics
Manager

Data Life Cycle

25 ▲



Data Life Cycle



- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

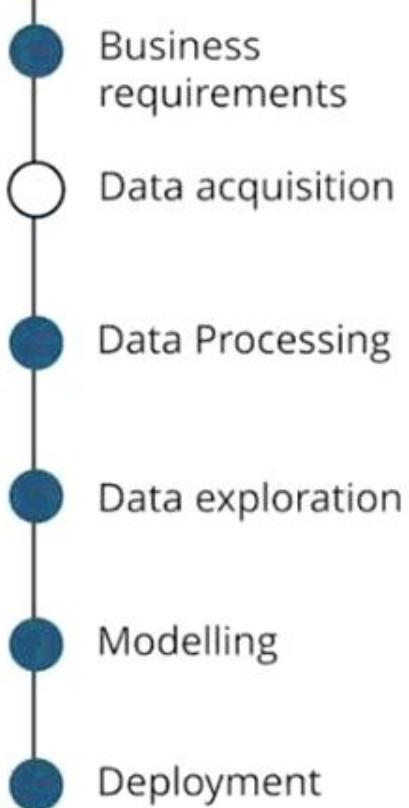
Understand the problem

Identify central objectives

Identify variables that need to be predicted



Data Life Cycle



What data do I need for my project?
What are the data sources?
How can I obtain the data?
What is the most efficient way to store and access all of it?



Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Transform data into desired format
Data cleaning

- Missing values
- Corrupted data
- Remove unnecessary data



Data Life Cycle

25 ▲

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

understand the patterns in the data
Retrieve useful insight
form hypotheses



Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Determine optimal data features
for the machine-learning model

Create a model that predicts the
target most accurately

Evaluate & test the efficiency of
the model



Data Life Cycle

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Check the deployment environment for dependency issues
Deploy the model in a pre-production/ test environment

Monitor the performance



Week2 – Probability & Statistics

WHAT IS DATA?

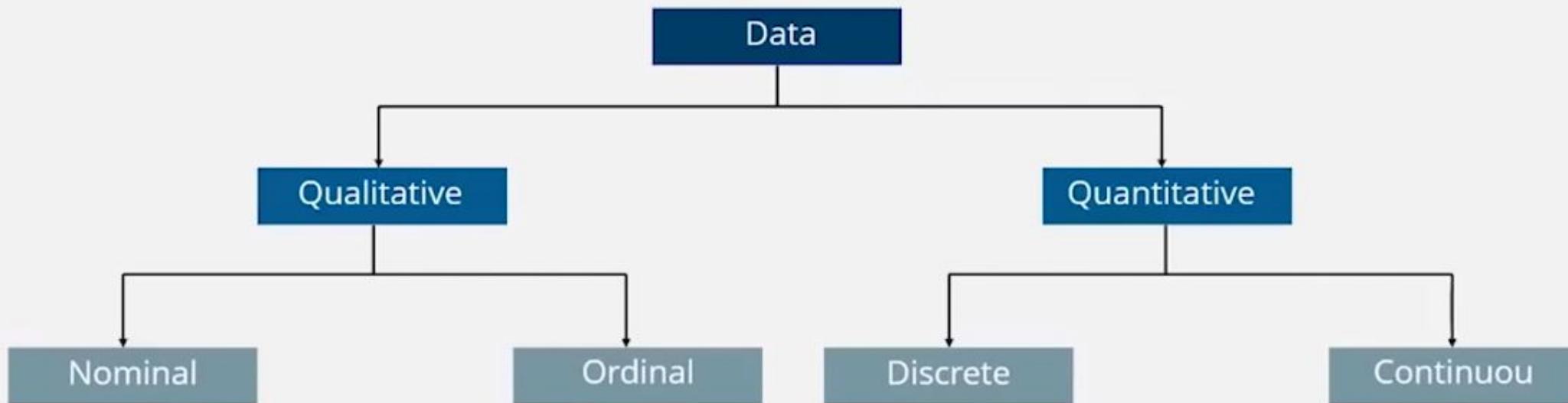
Data refers to facts and statistics collected together for reference or analysis.



02

CATEGORIES OF DATA

Types of Data



QUALITATIVE DATA

Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively.

Nominal Data

Data with no inherent order or ranking such as gender or race, such kind of data is called Nominal data



Gender
Male
Female
Male
Male

Ordinal Data

Data with an ordered series, such as shown in the table, such kind of data is called Ordinal data

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

QUANTITATIVE DATA

Quantitative data deals with numbers and things you can measure objectively.

Discrete Data

Also known as categorical data, it can hold finite number of possible values.

Example: Number of students in a class



Continuous Data

Data that can hold infinite number of possible values.

Example: Weight of a person



03

WHAT IS STATISTICS?

WHAT IS STATISTICS?

Statistics is an area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.



Your company has created a new drug that may cure cancer. How would you conduct a test to confirm the drug's effectiveness?



WHAT IS STATISTICS?

Statistics is an area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.

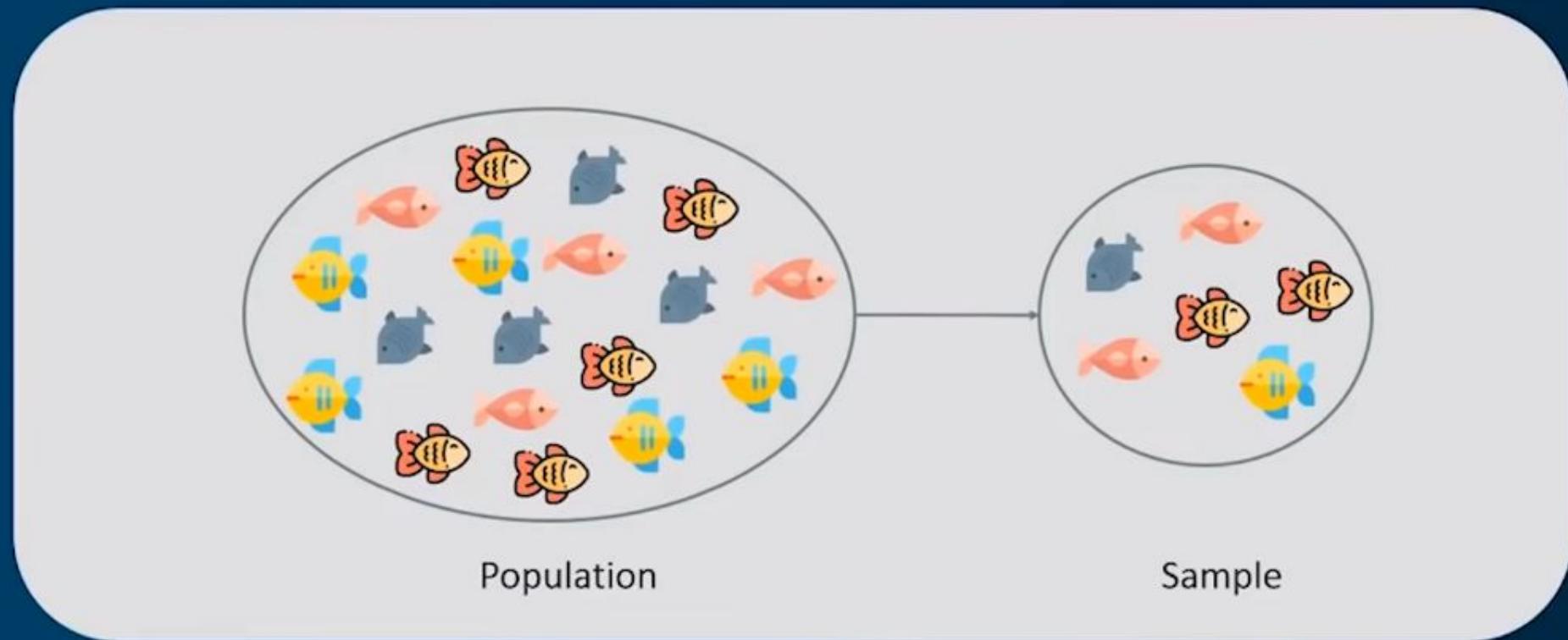


The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?



04

BASIC TERMINOLOGIES IN STATISTICS



Statistics Terminologies

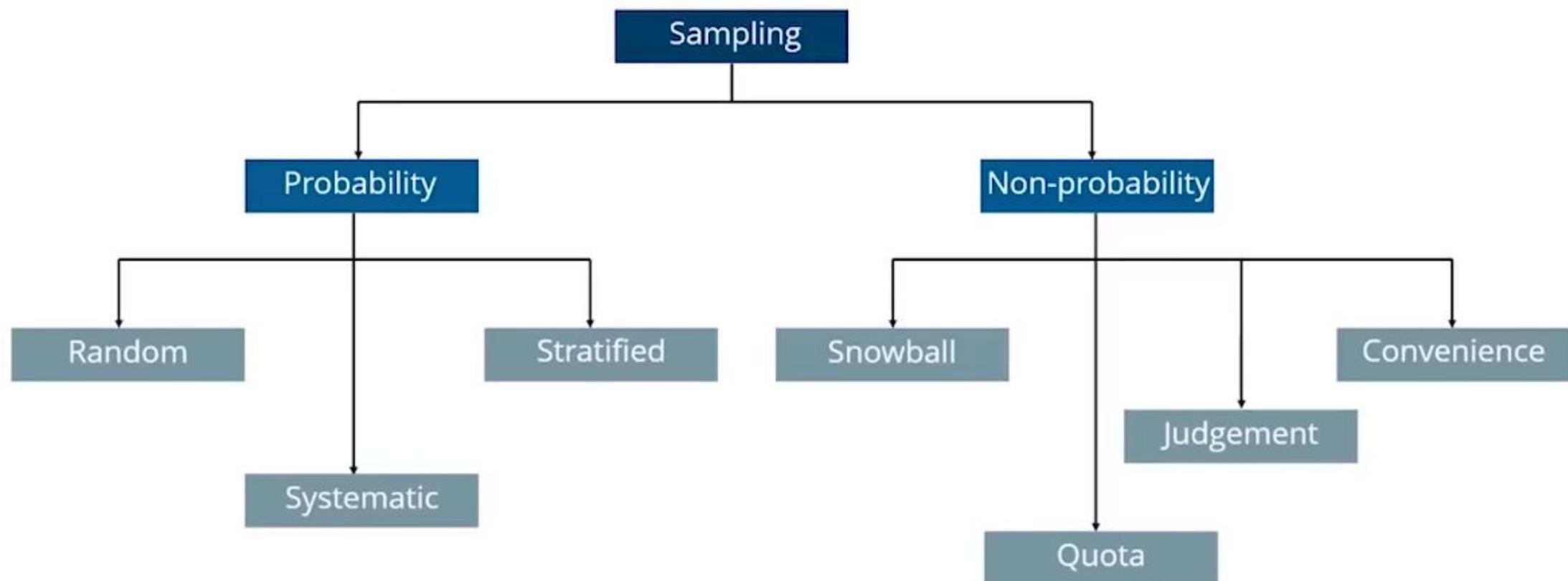
Population: A collection or set of individuals or objects or events whose properties are to be analyzed.

Sample: A subset of population is called 'Sample'. A well chosen sample will contain most of the information about a particular population parameter

05

SAMPLING TECHNIQUES

SAMPLING TECHNIQUES



RANDOM SAMPLING

Random Sampling

Systematic Sampling

Stratified Sampling



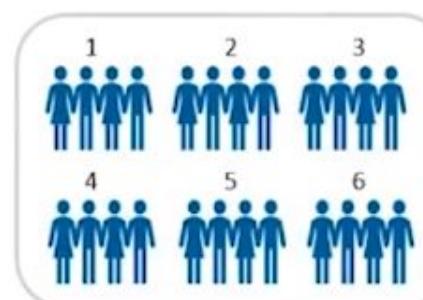
Each member of the population has equal chance of being selected in the sample.

SYSTEMATIC SAMPLING

Random Sampling

Systematic Sampling

Stratified Sampling



Every nth record is chosen



Every 2nd record is chosen

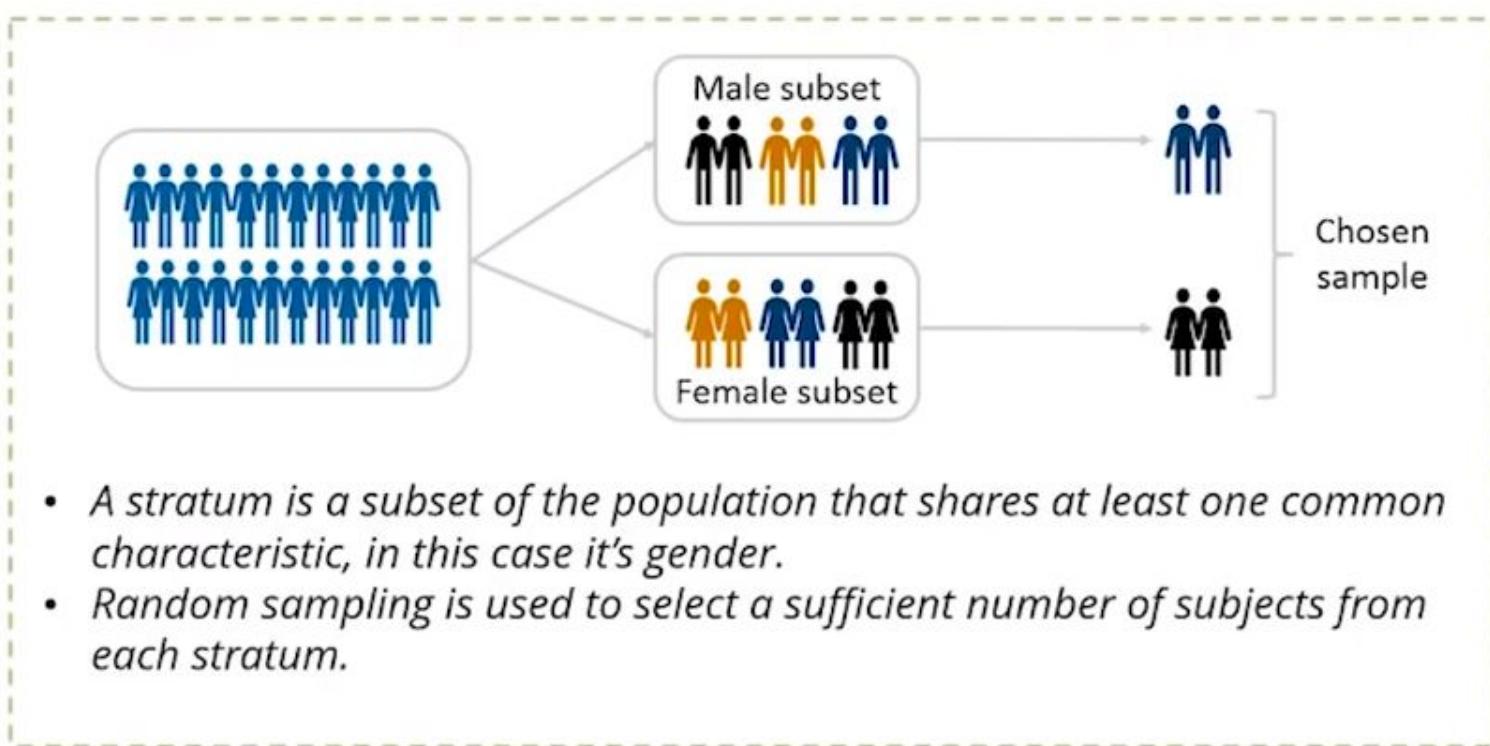
In Systematic sampling every nth record is chosen from the population to be a part of the sample.

STRATIFIED SAMPLING

Random Sampling

Systematic Sampling

Stratified Sampling



06

TYPES OF STATISTICS

DESCRIPTIVE STATISTICS

Descriptive statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.



Maximum

Average

Minimum



Descriptive Statistics is mainly focused upon the main characteristics of data. It provides graphical summary of the data.

INFERRENTIAL STATISTICS

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.



Large

Medium

Small



Inferential statistics, generalizes a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using a sample data.

07

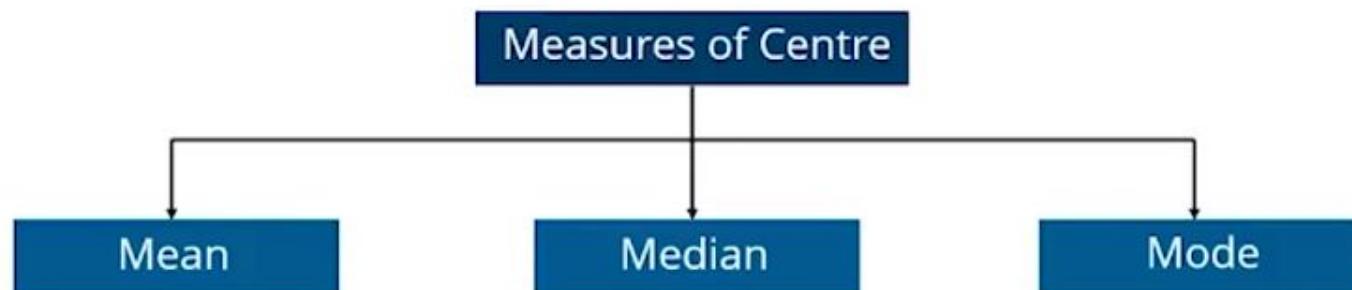
DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- **Measures of Central tendency**
- Measures of Variability (spread)

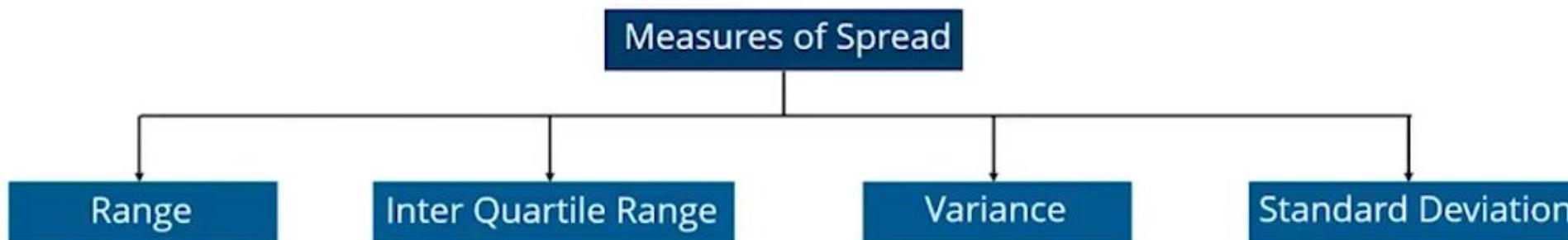


DESCRIPTIVE STATISTICS

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- Measures of Central tendency
- **Measures of Variability (spread)**



08

MEASURES OF CENTRE

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mean

Measure of average of all the values in a sample is called Mean.

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mean

To find out the average horsepower of the cars among the population of cars, we will check and calculate the average of all values:

$$\frac{110 + 110 + 93 + 96 + 90 + 110 + 110 + 110}{8} = 103.625$$

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Median

Measure of the central value of the sample set is called **Median**.

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Median

To find out the center value of mpg among the population of cars, arrange records in *Ascending order*, i.e., **21, 21, 21.3, 22.8, 23, 23, 23, 23**

In case of even entries, take average of the two middle values, i.e. $(22.8+23)/2 = 22.9$

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mode

The value most recurrent in the sample set is known as Mode.

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mode

To find the most common type of cylinder among the population of cars, check the value which is repeated most number of times, i.e., cylinder type 6

09

MEASURES OF SPREAD

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Range is the given measure of how spread apart the values in a dataset are.

$$\text{Range} = \text{Max}(x_i) - \text{Min}(x_i)$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

Q1	Q2	Q3
1 2	3 4	5 6

MEASURES OF SPREAD

Consider the marks of the 100 students below, ordered from the lowest to the highest scores

The first quartile (Q1) lies between the 25th and 26th.
 $Q1 = (45 + 45) \div 2 = 45$

Order	Score								
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

The second quartile (Q2) between the 50th and 51st.
 $Q2 = (58 + 59) \div 2 = 58.5$

The third quartile (Q3) between the 75th and 76th.
 $Q3 = (71 + 71) \div 2 = 71$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Inter Quartile Range(IQR) is the measure of variability, based on dividing a dataset into quartiles.

- Quartiles divide a rank-ordered data set into four equal parts, denoted by Q_1 , Q_2 , and Q_3 , respectively
- The interquartile range is equal to Q_3 minus Q_1 , i.e.. $IQR = Q_3 - Q_1$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation



MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Variance describes how much a random variable differs from its expected value.
It entails computing squares of deviations.

$$s^2 = \frac{\sum_{i=1}^{n=1} (x_i - \bar{x})^2}{n}$$

x : Individual data points

n : Total number of data points

\bar{x} : Mean of data points

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Deviation is the difference between each element from the mean.

$$\text{Deviation} = (x_i - \mu)$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Population Variance is the average of squared deviations.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Sample Variance is the average of squared differences from the mean.

$$s^2 = \frac{1}{(n - 1)} \sum_{i=1}^N (x_i - \bar{x})^2$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

Standard Deviation is the measure of the dispersion of a set of data from its mean.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 1

Find out the mean for your sample set.

The Mean is:

$$\frac{9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4}{20}$$

$$\mu=7$$

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 2

Then for each number, subtract the Mean and square the result.

$$(x_i - \mu)^2$$

$$(9-7)^2 = 2^2 = 4$$

$$(2-7)^2 = (-5)^2 = 25$$

$$(5-7)^2 = (-2)^2 = 4$$

And so on...

□ We get the following results:

4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 3

Then work out the mean of those squared differences.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

$$\frac{4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9}{20}$$

□ $\sigma^2 = 8.9$

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 4

Take square root of σ^2 .

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

□ $\sigma = 2.983$

10

INFORMATION GAIN & ENTROPY

INFORMATION GAIN & ENTROPY

Entropy

Entropy measures the impurity or uncertainty present in the data.

$$H(S) = - \sum_{i=1}^N p_i \log_2 p_i$$

where:

- S – set of all instances in the dataset
- N – number of distinct class values
- p_i – event probability

Information Gain (IG)

IG indicates how much “information” a particular feature/variable gives us about the final outcome.

$$\text{Gain}(A, S) = H(S) - \sum_{j=1}^v \frac{|S_j|}{|S|} \cdot H(S_j) = H(S) - H(A, S)$$

where:

- $H(S)$ – entropy of the whole dataset S
- $|S_j|$ – number of instance with j value of an attribute A
- $|S|$ – total number of instances in dataset S
- v – set of distinct values of an attribute A
- $H(S_j)$ – entropy of subset of instances for attribute A
- $H(A, S)$ – entropy of an attribute A



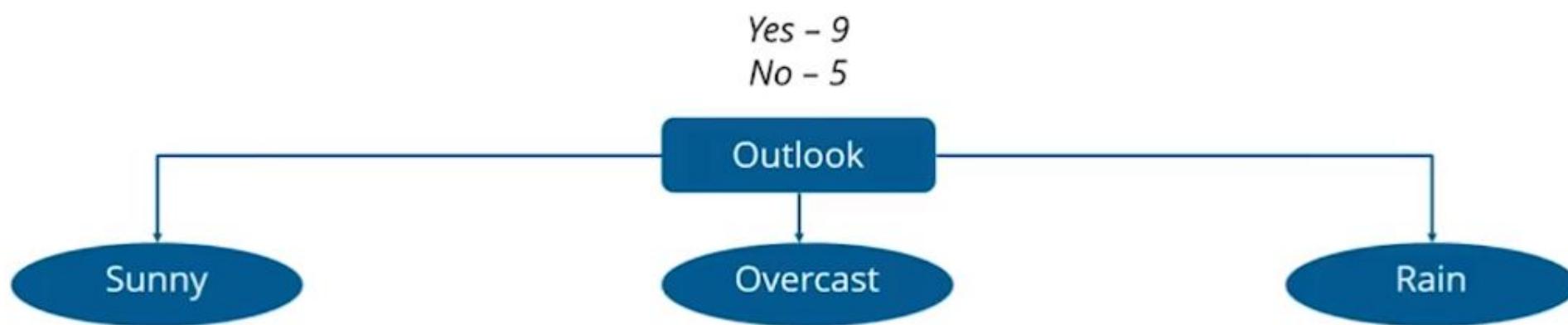
USE CASE

USE CASE



Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

USE CASE



Day	Outlook	Humidity	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

Yes - 2
No - 3

Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Yes - 4
No - 0

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

Yes - 3
No - 2

USE CASE

From the total of 14 instances we have:

- 9 instances "yes"
- 5 instances "no"

The Entropy is:

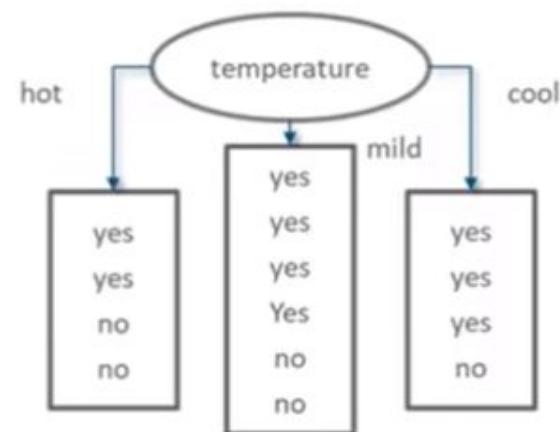
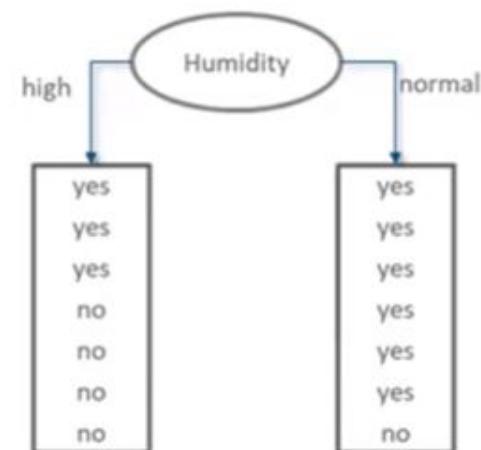
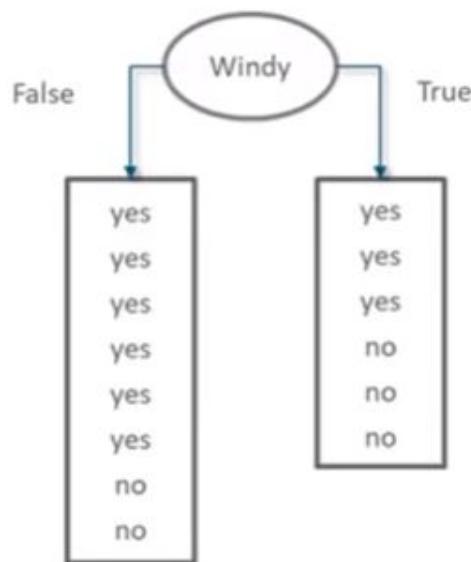
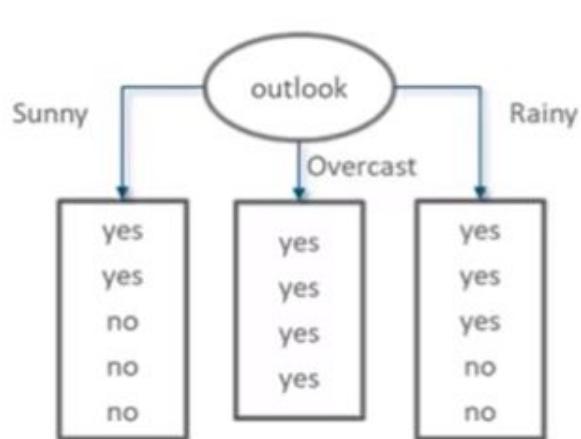
$$H(S) = - \sum_{i=1}^N p_i \log_2 p_i$$

$$H(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940$$



USE CASE

Selecting the root variable



USE CASE

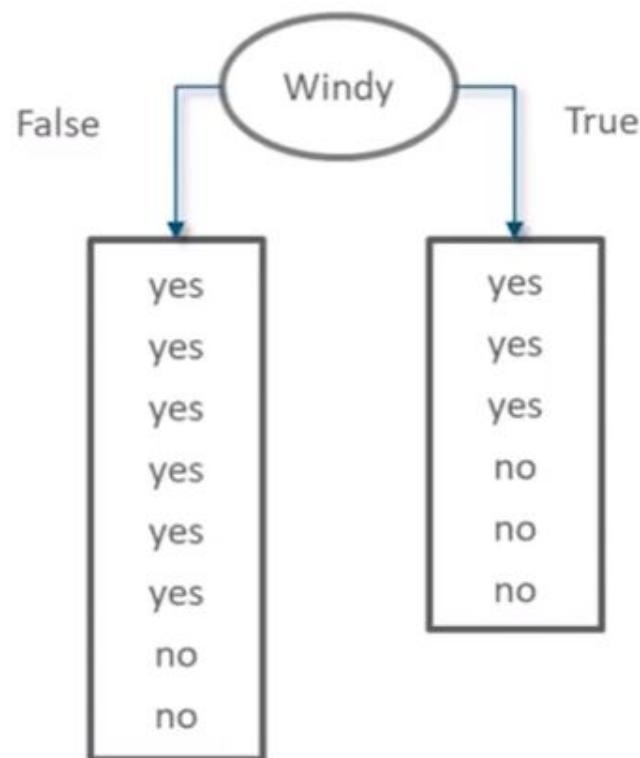
Information Gain of attribute "windy"

From the total of 14 instances we have:

- 6 instances "true"
- 8 instances "false"

$$Gain(A, S) = H(S) - \sum_{j=1}^v \frac{|S_j|}{|S|} \cdot H(S_j)$$

$$\begin{aligned}Gain(A_{Windy}, S) &= 0.940 - \\&\frac{8}{14} \cdot \left(-\left(\frac{6}{8} \cdot \log_2 \frac{6}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} \right) \right) + \\&\frac{6}{14} \cdot \left(-\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6} \right) \right) = 0.048\end{aligned}$$



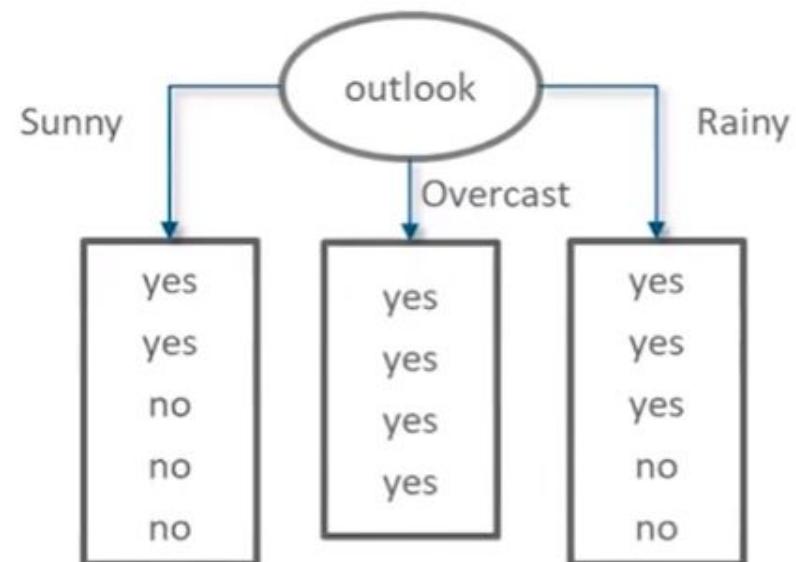
USE CASE

Information Gain of attribute "outlook"

From the total of 14 instances we have:

- 5 instances "sunny"
- 4 instances "overcast"
- 5 instances "rainy"

$$\begin{aligned} \text{Gain}(A_{\text{outlook}}, S) = & 0.940 - \\ & \frac{5}{14} \cdot \left(-\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \right) + \\ & \frac{4}{14} \cdot \left(-\left(\frac{4}{4} \log_2 \frac{4}{4} \right) \right) + \\ & \frac{5}{14} \cdot \left(-\left(\frac{3}{5} \cdot \log_2 \frac{3}{5} + \frac{2}{5} \cdot \log_2 \frac{2}{5} \right) \right) = 0.247 \end{aligned}$$



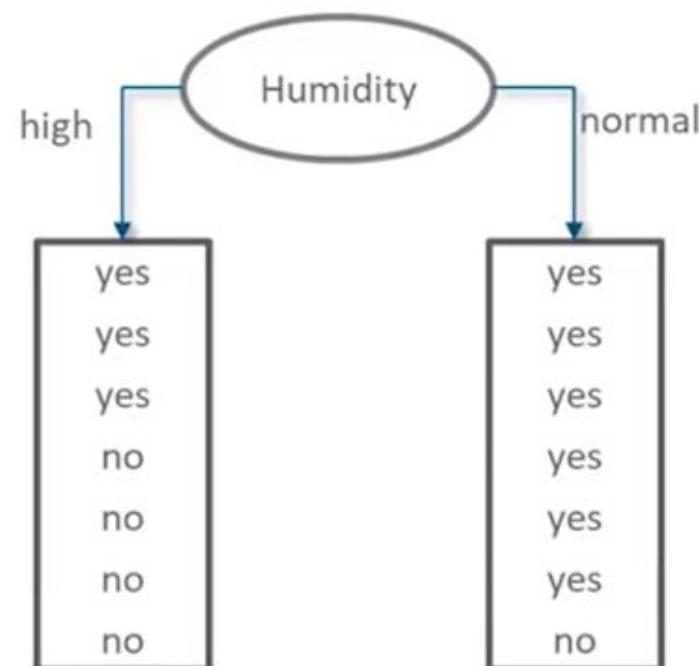
USE CASE

Information Gain of attribute "humidity"

From the total of 14 instances we have:

- 7 instances "high"
- 7 instances "normal"

$$\begin{aligned} \text{Gain}(A_{\text{Humidity}}, S) &= 0.940 - \\ &\frac{7}{14} \cdot \left(-\left(\frac{3}{7} \cdot \log_2 \frac{3}{7} + \frac{4}{7} \cdot \log_2 \frac{4}{7} \right) \right) + \\ &\frac{7}{14} \cdot \left(-\left(\frac{6}{7} \cdot \log_2 \frac{6}{7} + \frac{1}{7} \cdot \log_2 \frac{1}{7} \right) \right) = 0.151 \end{aligned}$$



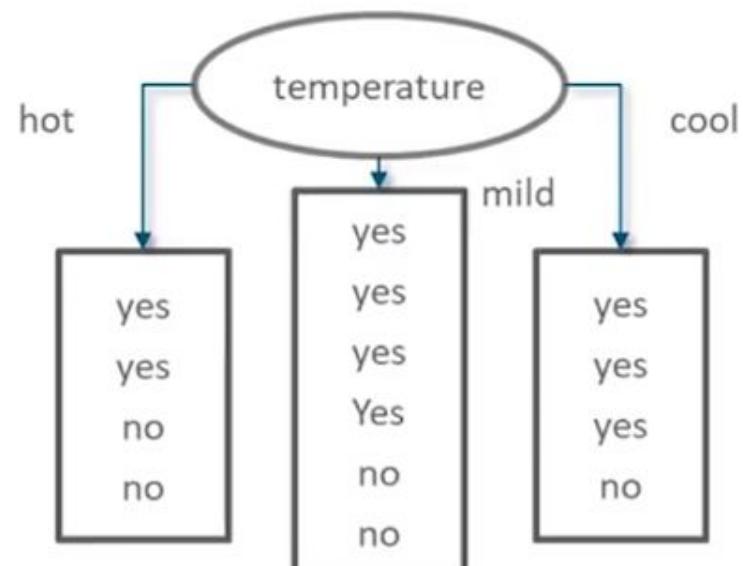
USE CASE

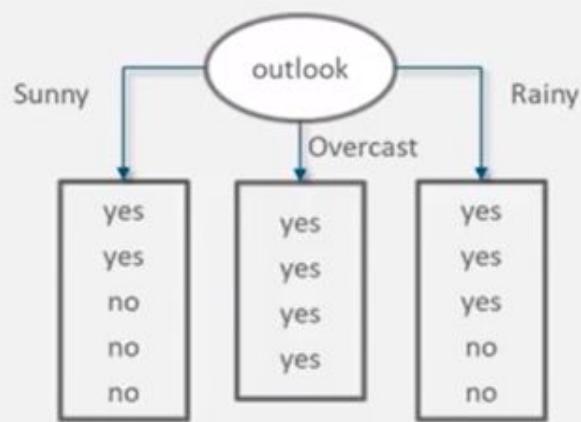
Information Gain of attribute "temperature"

From the total of 14 instances we have:

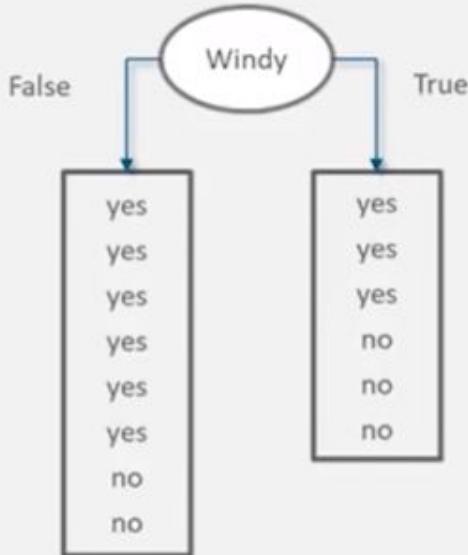
- 4 instances "hot"
- 6 instances "mild"
- 4 instances "cool"

$$\begin{aligned} \text{Gain}(A_{\text{Temperature}}, S) = & 0.940 - \\ & \frac{4}{14} \cdot \left(- \left(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4} \right) \right) + \\ & \frac{6}{14} \cdot \left(- \left(\frac{4}{6} \cdot \log_2 \frac{4}{6} + \frac{2}{6} \cdot \log_2 \frac{2}{6} \right) \right) + \\ & \frac{4}{14} \cdot \left(- \left(\frac{3}{4} \cdot \log_2 \frac{3}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4} \right) \right) = 0.029 \end{aligned}$$

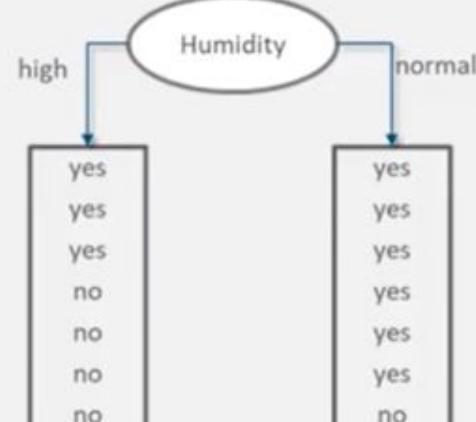




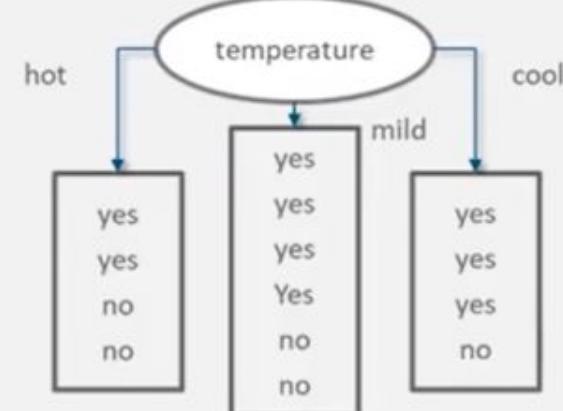
Gain = 0.247



Gain = 0.048

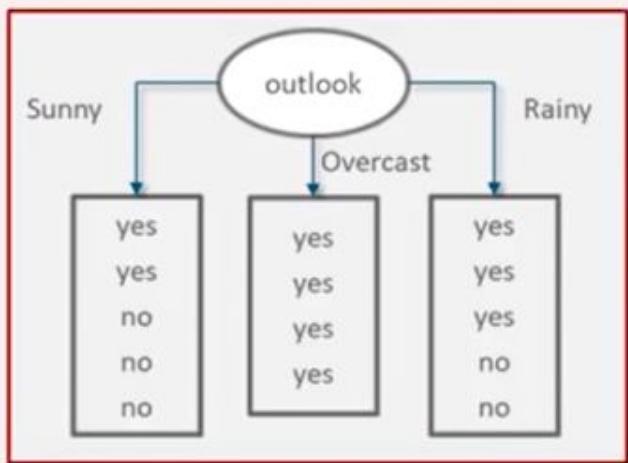


Gain = 0.151

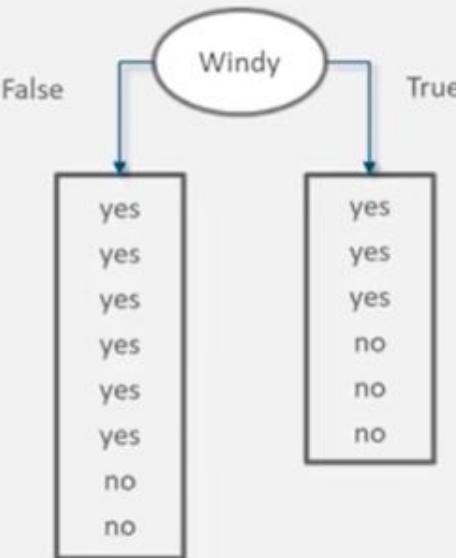


Gain = 0.029

The variable with the highest IG is used to split the data at the root node.



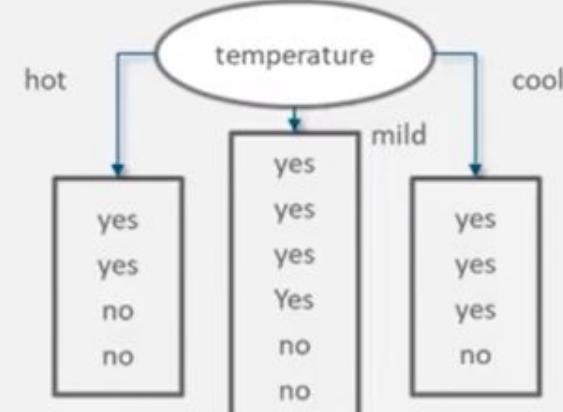
Gain = 0.247



Gain = 0.048



Gain = 0.151



Gain = 0.029

The variable with the highest IG is used to split the data at the root node. The 'Outlook' variable has the highest IG, therefore it can be assigned to the root node.

12

CONFUSION MATRIX

CONFUSION MATRIX

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

*Confusion Matrix represents a tabular representation of Actual vs Predicted values
You can calculate the accuracy of your model with:*

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

CONFUSION MATRIX

- There are two possible predicted classes: "yes" and "no"
- The classifier made a total of 165 predictions
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times
- In reality, 105 patients in the sample have the disease, and 60 patients do not



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

CONFUSION MATRIX

Predicted Values

		Actual Values	
		Positive (1)	Negative (0)
Positive (1)	Positive (1)	TP	FP
	Negative (0)	FN	TN



true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease

true negatives (TN): We predicted no, and they don't have the disease

false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

Week2 (B)

Class Activity01

Problem Identification

- Find out problem and Understand it. Identify its requirements and class/labels to be predicted.

Data Collection

- Acquire data, there are many data sources, studied in class, collect data from any of those.

Data Preprocessing

- Make your data into desired format, find and fill missing values, remove unnecessary data.

Class Activity01 – Cont.

Data Exploration

- Do some analysis and visualization for finding out some meaningful patterns from data.

ML Model Selection

- Split the data into train and test (80%, 20% ration respectively)
- Determine model to be applied on data (whether classification/regression).
- Apply some models and find out which model is the best amongst others.

Model Evaluation

- Train and test the efficiency of the model.

Welcome to Data Science

- Your work MUST be your own
- **Zero tolerance for cheating**
 - You get an F for the course if you cheat in anything
 - however small
- **NO DISCUSSION !!!**
 - Try to enjoy studies ☺
 - Don't overwhelm yourself!

Welcome to Data Science

- Never miss classes
- Take notes
- Be (inter-) active
- Start working on assignments as soon as you get.
- Submit assignment through classroom.google.com
- **There is NO late submission policy !!!**
- Always be prepared for a quiz
- Try making a habit of reading