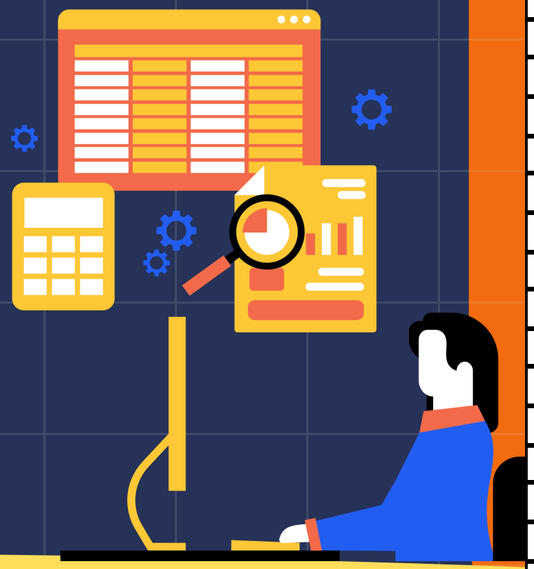
GROUP 4

DIRTY DATASET CHALLENGE

Employee Salary Dataset

SYARAH AQILAH BINTI YA'ACOB A24CS0195
NIVEETHITA A/P PANDIA RAJAN A24CS0148
KAVIVARTHAN A/L MANNIVANAN A24CS0093



EmpID	Name	Salary	Dept
0	Amin	-2000	Finance
5	Sara	-2000	IT
2	Sara	-2000	HR
3	Sara	4000	HR
5	Lee	-2000	hr
5	Lee	5000	IT
6	Sara	3000	hr
5	Sara	5000	IT
8	Lee	3000	hr
5	Lee	4000	hr
10		3000	Finance
5	Amin	3000	Finance
12	Amin	-2000	IT
5		4000	hr
14	Lee	3000	Finance
5	Sara	5000	Finance
16		4000	hr
17		4000	hr
18	Sara	-2000	HR
5	Lee	-2000	HR
20	Lee	3000	IT
21	Amin	3000	Finance
22	Lee	4000	HR
23	Sara	3000	hr
5	Lee	-2000	IT
25	Lee	3000	IT
5	Sara	4000	IT
5	Lee	-2000	hr
28	Lee	3000	HR
5	Sara	3000	Finance

DATASET OVERVIEW

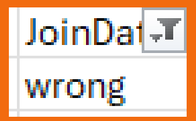
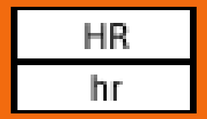
JoinDate	Type	Age	Gender	Level	Status
1/1/2020	Part	30	M	1	Active
1/1/2020	Full	30	M	1	Inactive
wrong	Full	30	Male	1	Inactive
1/1/2020	Full	200	Male	3	Active
wrong	Full	30	Male	2	Active
1/2/2020	Full	200	Male	3	Active
1/1/2020	Full	30	M	3	Inactive
1/2/2020	Full	30	M	2	Inactive
wrong	Full	200	F	3	Active
1/1/2020	Full	200	M	3	Active
1/1/2020	Full	25	M	2	Inactive
1/1/2020	Part	200	M	3	Inactive
wrong	Full	200	F	3	Inactive
wrong	Full	25	F	3	Inactive
1/2/2020	Part	25	Male	3	Active
wrong	Part	25	F	2	Inactive
1/1/2020	Part	200	Male	3	Inactive
wrong	Full	30	Male	1	Inactive
1/2/2020	Full	30	Male	3	Inactive
1/1/2020	Part	25	Male	3	Active
1/2/2020	Full	200	F	2	Inactive
wrong	Full	30	Male	1	Inactive
1/1/2020	Part	25	Male	2	Active
1/2/2020	Part	25	M	2	Active
1/1/2020	Part	30	M	3	Active
1/2/2020	Part	25	Male	3	Inactive
1/2/2020	Part	200	M	1	Inactive
wrong	Part	200	Male	2	Inactive
1/2/2020	Full	200	M	2	Active
wrong	Part	25	F	1	Active

Item	Description
Dataset Name	Employee Salary Dataset
Number of Rows	30
Number of Columns	10
Data Type	Structured Data

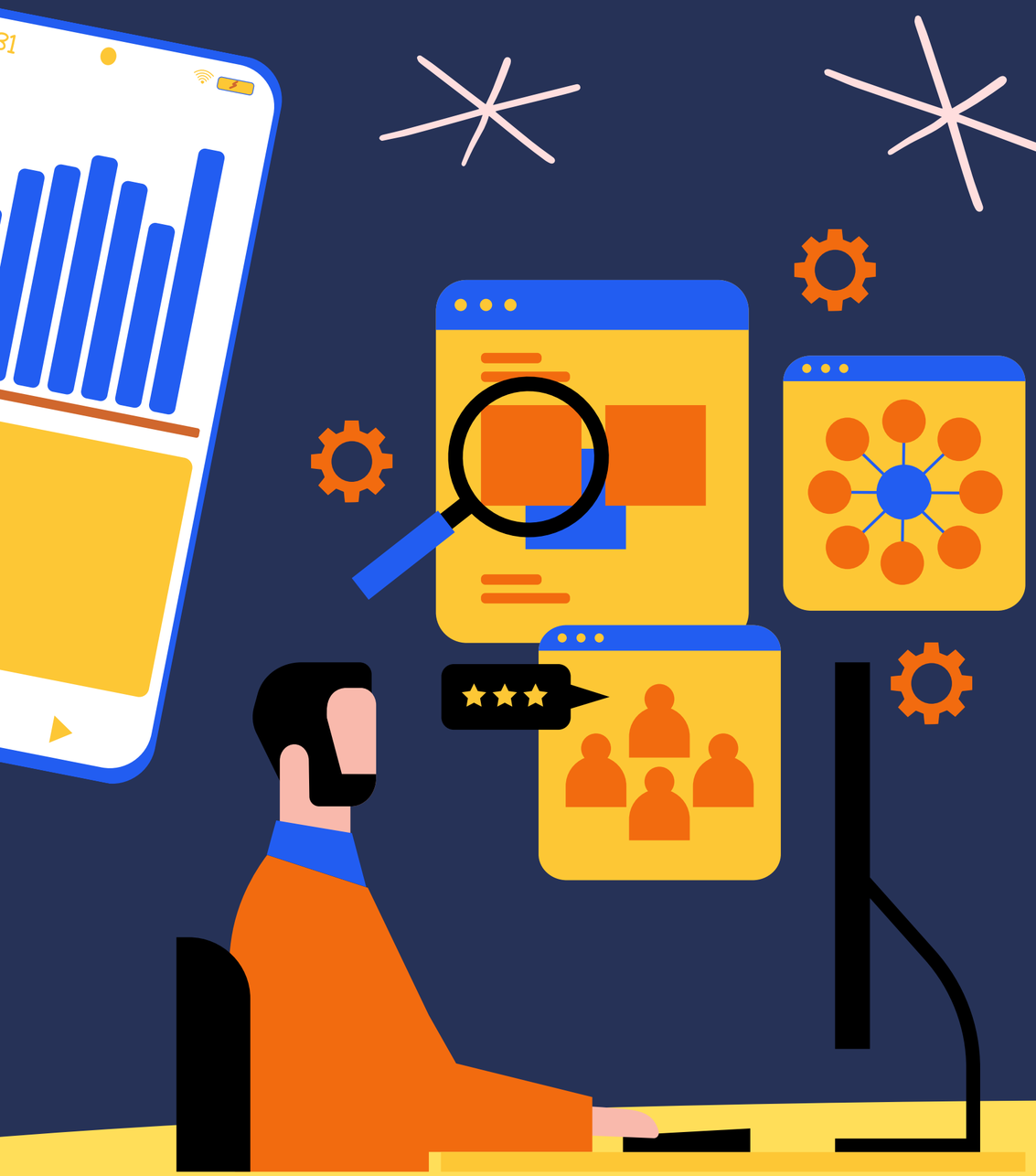
ATTRIBUTES DECRRIPTIONS

Attribute	Data Type	Meaning
EmpID	Int	Unique ID as identity
Name	String	Employee name
Salary	Int	Monthly Salary
Dept	String	Department name
JoinDate	Date	Date employeejoined the company
Type	String	Employee type
Age	Int	Employee age
Gender	String	Employee gender
Level	Int	Job level
Status	String	Employee status

ISSUES IDENTIFIED

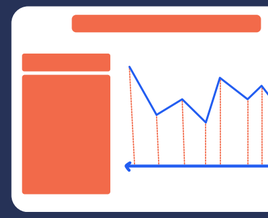
Issue Type	Example		Problem
Missing Value		Missing employee name	Reduces data quality
Invalid Value		Salary cannot be negative	Invalid data
		Unrealistic human age	
		Wrong data	
Inconsistent Data		The system treat the data as a different categories even it is actually the same data with different format.	Incorrect grouping/statistics
			
Duplicate ID		Same employee ID for 2 different employee	Primary Key should be unique

TOOLS & TECHNIQUES USED



- Microsoft Excel
- Data filtering and sorting
- Duplicate detection
- Formatting and standardization techniques

METHODS/TECHNIQUES APPLIED



Handle Invalid/Noisy Data →
1- Salary :-2000
Technique : replace with median. (Median = 3000) Use formula median in excel
Justification : median not affected by extreme values.

2- Age = 200
Outlier and doesn't make sense
Technique : same as salary

3- "wrong"
Replace with mode

Handle Missing Values→
Use filter in excel then click blank, then type unknown

Handle inconsistent data (Dept Hr, Gender M).→
standardization

Handle duplicate data →
Assign new number for each data

JUSTIFICATION OF METHODS USED

Handle Invalid/Noisy Data

improves the accuracy and reliability of analysis

Handle Missing Values

maintain data completeness and avoid deleting record.

Handle inconsistent data

ensure consistency and improve data accuracy

Handle duplicate data

prevents biased or inflated results

COMPARISON

BEFORE

EmpID	Name	Salary	Dept	JoinDate	Type	Age	Gender	Level	Status
0	Amin	-2000	Finance	1/1/2020	Part	30	M	1	Active
5	Sara	-2000	IT	1/1/2020	Full	30	M	1	Inactive
2	Sara	-2000	HR	wrong	Full	30	Male	1	Inactive
3	Sara	4000	HR	1/1/2020	Full	200	Male	3	Active
5	Lee	-2000	hr	wrong	Full	30	Male	2	Active
5	Lee	5000	IT	1/2/2020	Full	200	Male	3	Active
6	Sara	3000	hr	1/1/2020	Full	30	M	3	Inactive
5	Sara	5000	IT	1/2/2020	Full	30	M	2	Inactive
8	Lee	3000	hr	wrong	Full	200	F	3	Active
5	Lee	4000	hr	1/1/2020	Full	200	M	3	Active
10		3000	Finance	1/1/2020	Full	25	M	2	Inactive
5	Amin	3000	Finance	1/1/2020	Part	200	M	3	Inactive
12	Amin	-2000	IT	wrong	Full	200	F	3	Inactive
5		4000	hr	wrong	Full	25	F	3	Inactive
14	Lee	3000	Finance	1/2/2020	Part	25	Male	3	Active
5	Sara	5000	Finance	wrong	Part	25	F	2	Inactive
16		4000	hr	1/1/2020	Part	200	Male	3	Inactive
17		4000	hr	wrong	Full	30	Male	1	Inactive
18	Sara	-2000	HR	1/2/2020	Full	30	Male	3	Inactive
5	Lee	-2000	HR	1/1/2020	Part	25	Male	3	Active
20	Lee	3000	IT	1/2/2020	Full	200	F	2	Inactive
21	Amin	3000	Finance	wrong	Full	30	Male	1	Inactive
22	Lee	4000	HR	1/1/2020	Part	25	Male	2	Active
23	Sara	3000	hr	1/2/2020	Part	25	M	2	Active
5	Lee	-2000	IT	1/1/2020	Part	30	M	3	Active
25	Lee	3000	IT	1/2/2020	Part	25	Male	3	Inactive
5	Sara	4000	IT	1/2/2020	Part	200	M	1	Inactive
5	Lee	-2000	hr	wrong	Part	200	Male	2	Inactive
28	Lee	3000	HR	1/2/2020	Full	200	M	2	Active
5	Sara	3000	Finance	wrong	Part	25	F	1	Active

- Missing employee names
- Duplicate Employee IDs
- Negative or unrealistic salary values
- Inconsistent department naming (e.g., “HR”, “hr”, “Human Resource”)
- Invalid age values (e.g., too high or too low)
- Unstructured and unreliable dataset

Data was inaccurate, inconsistent, and not suitable for analysis.

AFTER DATA CLEANING

EmpID	Name	Salary	Dept	JoinDate	Type	Age	Gender	Level	Status
0	Amin	3000	Finance	1/1/2020	Part	30	M	1	Active
1	Sara	3000	IT	1/1/2020	Full	30	M	1	Inactive
2	Sara	3000	HR	1/1/2020	Full	30	M	1	Inactive
3	Sara	4000	HR	1/1/2020	Full	30	M	3	Active
4	Lee	3000	HR	1/1/2020	Full	30	M	2	Active
5	Lee	5000	IT	1/2/2020	Full	30	M	3	Active
6	Sara	3000	HR	1/1/2020	Full	30	M	3	Inactive
7	Sara	5000	IT	1/2/2020	Full	30	M	2	Inactive
8	Lee	3000	HR	1/1/2020	Full	30	F	3	Active
9	Lee	4000	HR	1/1/2020	Full	30	M	3	Active
10	Unknown	3000	Finance	1/1/2020	Full	25	M	2	Inactive
11	Amin	3000	Finance	1/1/2020	Part	30	M	3	Inactive
12	Amin	3000	IT	1/1/2020	Full	30	F	3	Inactive
13	Unknown	4000	HR	1/1/2020	Full	25	F	3	Inactive
14	Lee	3000	Finance	1/2/2020	Part	25	M	3	Active
15	Sara	5000	Finance	1/1/2020	Part	25	F	2	Inactive
16	Unknown	4000	HR	1/1/2020	Part	30	M	3	Inactive
17	Unknown	4000	HR	1/1/2020	Full	30	M	1	Inactive
18	Sara	3000	HR	1/2/2020	Full	30	M	3	Inactive
19	Lee	3000	HR	1/1/2020	Part	25	M	3	Active
20	Lee	3000	IT	1/2/2020	Full	30	F	2	Inactive
21	Amin	3000	Finance	1/1/2020	Full	30	M	1	Inactive
22	Lee	4000	HR	1/1/2020	Part	25	M	2	Active
23	Sara	3000	HR	1/2/2020	Part	25	M	2	Active
24	Lee	3000	IT	1/1/2020	Part	30	M	3	Active
25	Lee	3000	IT	1/2/2020	Part	25	M	3	Inactive
26	Sara	4000	IT	1/2/2020	Part	30	M	1	Inactive
27	Lee	3000	HR	1/1/2020	Part	30	M	2	Inactive
28	Lee	3000	HR	1/2/2020	Full	30	M	2	Active
29	Sara	3000	Finance	1/1/2020	Part	25	F	1	Active

- All missing values handled or replaced
- Duplicate records removed
- Salary values corrected and validated
- Standardized department names
- Valid and realistic age values
- Clean, structured, and reliable dataset

Dataset is now ready for analysis and decision-making.

THANK YOU !