# **DATA CLEANING**

1. **Create Backup**

CREATE TABLE laptop\_backup LIKE laptops

Insert into laptops\_backup

select \* from laptops

1. **Check number of rows**

select count(\*) from laptops

1. **Check memory consumption for reference**

*[ Information\_schema.TABLES  contains information about all tables in the database.*

*Estimates data length by multiplying TABLE\_ROWS with AVG\_ROW\_LENGTH (average row length) and then dividing by 1024 for kilobytes.]*

SELECT SUM (TABLE\_ROWS \* AVG\_ROW\_LENGTH) / 1024 AS total\_data\_length\_kb

FROM information\_schema.TABLES

WHERE TABLE\_SCHEMA = 'campusx' AND TABLE\_NAME = 'laptops';

1. **Drop non important cols**

Removing ”Unnammed Column”

ALTER TABLE laptops

DROP COLUMN “Unnamed: 0

1. **Drop null values**

DELETE FROM laptops

WHERE Company IS NULL

AND TypeName IS NULL

AND Inches IS NULL

AND ScreenResolution IS NULL

AND Cpu IS NULL

AND Ram IS NULL

AND Memory IS NULL

AND Gpu IS NULL

AND OpSys IS NULL

AND Weight IS NULL

AND Price IS NULL;

1. **Drop duplicates**

Not Applied in this Case as there are no duplicates in data

1. **Change data type of Inch column to Decimal**

UPDATE laptops

MODIFY COLUMN Inches DECIMAL(10,1)

1. **Clean the Ram column**

UPDATE laptops

SET Ram = REPLACE(Ram, 'GB', '')

ALTER TABLE laptops

MODIFY COLUMN Ram INTEGER

1. **Clean Weight Column & Change Column Type**

UPDATE laptops

SET Weight = REPLACE(Weight,'kg','')

delete from laptops

where Weight = '?'

ALTER TABLE laptops

MODIFY COLUMN Weight decimal(10,2) DEFAULT 0;

1. **ROUND Price Column and change to integer**

Update laptops

SET Price = ROUND(Price)

Alter TABLE laptops

MODIFY COLUMN Price INTEGER;

1. Change the OpSys col

[mac, windows, linux, no os, (android, chrome) -> others]

Update laptops

SET OpSys = CASE

when OpSys Like '%mac%' then 'macos'

when OpSys Like 'windows%' then 'windows'

when OpSys Like '%linux%' then 'linux'

when OpSys = 'No OS' then 'N/A'

ELSE "other"

END

1. **Change the Gpu Column**

Alter TABLE laptops

ADD COLUMN gpu\_brand VARCHAR(255) AFTER Gpu,

ADD COLUMN gpu\_name VARCHAR(255) AFTER gpu\_brand;

UPDATE laptops

SET gpu\_brand = SUBSTRING\_INDEX(Gpu, " ", 1);

UPDATE laptops

SET gpu\_name = REPLACE(Gpu,gpu\_brand,'')

ALTER table laptops

DROP COLUMN Gpu

1. **Change the Cpu Column**

ALTER TABLE laptops

ADD COLUMN cpu\_brand VARCHAR(255) AFTER Cpu,

ADD COLUMN cpu\_name VARCHAR(255) After cpu\_brand,

ADD COLUMN cpu\_speed DECIMAL(10,1) After cpu\_name;

UPDATE laptops

SET cpu\_brand = SUBSTRING\_INDEX(Cpu, ' ', 1);

UPDATE laptops

SET cpu\_speed = cast(REPLACE(SUBSTRING\_INDEX(Cpu, ' ', -1), 'GHz','') AS decimal(10,2));

UPDATE laptops

SET Cpu = REPLACE(Cpu, cpu\_brand,'');

UPDATE laptops

SET cpu\_name = REPLACE(Cpu, Substring\_index(Cpu,' ', -1),'');

ALTER TABLE laptops

DROP COLUMN Cpu;