**Data Cleaning Using SQL**

* Creating a Database to import a data table in this database

CREATE DATABASE data\_cleaning\_eda;

* importing loptop dataset manually by clicking table import wizard

SELECT \* FROM data\_cleaning\_eda.laptop;

* STEP-1 : create a backup or copy of the data

USE data\_cleaning\_eda;

CREATE TABLE laptop\_backup LIKE laptop;

INSERT INTO data\_cleaning\_eda.laptop\_backup

SELECT \* FROM data\_cleaning\_eda.laptop;

SELECT \* FROM data\_cleaning\_eda.laptop\_backup;

* STEP-2 : Count number of rows in the dataset

SELECT COUNT(\*) FROM data\_cleaning\_eda.laptop;

* STEP-3 : Check memory consumtion for reference

SELECT \* FROM information\_schema.TABLES

WHERE TABLE\_SCHEMA = 'data\_cleaning\_eda'

AND TABLE\_NAME = 'laptop';

SELECT DATA\_LENGTH AS 'bytes\_length', DATA\_LENGTH/1024 AS 'kb'

FROM information\_schema.TABLES

WHERE TABLE\_SCHEMA = 'data\_cleaning\_eda'

AND TABLE\_NAME = 'laptop';

* How to rename a Column Name

ALTER TABLE `data\_cleaning\_eda`.`laptop`

CHANGE COLUMN `Unnamed: 0` `index` INT NULL DEFAULT NULL ;

SELECT \* FROM data\_cleaning\_eda.laptop;

* Drop rows where all column values are null

DELETE FROM data\_cleaning\_eda.laptop

WHERE `index` IN(SELECT \* FROM data\_cleaning\_eda.laptop WHERE

`index` IS NULL AND 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);

* # Drop Duplicates in SQL : You can do that by groupby, windows function and all
* # Applying DISTINCT() function to see categorical columns all item

SELECT DISTINCT(Company) FROM data\_cleaning\_eda.laptop;

SELECT DISTINCT(TypeName) FROM data\_cleaning\_eda.laptop;

* # How to modify DataTypes of a particular column

-- modifying `Inches` column

ALTER TABLE data\_cleaning\_eda.laptop

MODIFY COLUMN Inches DECIMAL(10, 1);

-- modify Price Column

ALTER TABLE data\_cleaning\_eda.laptop

MODIFY COLUMN Price DECIMAL(10, 1);

SELECT \* FROM data\_cleaning\_eda.laptop;

* # Modify values from Ram column. Every Ram are like 4GB, 8GB, 16GB convert them into 4, 8, 16

UPDATE data\_cleaning\_eda.laptop

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

SELECT \* FROM data\_cleaning\_eda.laptop;

* now modify ram DataTypes

ALTER TABLE data\_cleaning\_eda.laptop

MODIFY COLUMN Ram INTEGER;

* # Modify values from Weight column. Every Weight are like 2kg, 1.37 kg. convert them into 2, 1.37

SELECT Weight, REPLACE(Weight, 'kg', '') FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

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

* # now modify weight DataTYpes

ALTER TABLE data\_cleaning\_eda.laptop

MODIFY COLUMN Weight DECIMAL(10, 2);

-- this code is not running because weight have some non numerical value lets handle it

UPDATE data\_cleaning\_eda.laptop

SET Weight = NULL

WHERE Weight NOT REGEXP '^[0-9]+(\.[0-9]+)?$';

-- now run the previous code

ALTER TABLE data\_cleaning\_eda.laptop

MODIFY COLUMN Weight DECIMAL(10, 2);

* # Modify Operating System (OpSys)

-- view query

SELECT OpSys,

CASE

WHEN OpSys LIKE '%mac%' THEN 'macos'

WHEN OpSys LIKE '%Windows%' THEN 'windows'

WHEN OpSys LIKE '%linux%' THEN 'linux'

WHEN OpSys LIKE 'No OS' THEN 'N/A'

ELSE 'other'

END AS 'Operating\_system'

FROM data\_cleaning\_eda.laptop;

-- lets update operating\_system

UPDATE data\_cleaning\_eda.laptop

SET OpSys = CASE

WHEN OpSys LIKE '%mac%' THEN 'macos'

WHEN OpSys LIKE '%Windows%' THEN 'windows'

WHEN OpSys LIKE '%linux%' THEN 'linux'

WHEN OpSys LIKE 'No OS' THEN 'N/A'

ELSE 'other'

END;

SELECT \* FROM data\_cleaning\_eda.laptop;

* # Create 2 new column(gpu\_brand, gpu\_name) from GPU column

-- lets at first create 2 new column

ALTER TABLE data\_cleaning\_eda.laptop

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

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

-- Updating gpu\_brand column

SELECT Gpu, SUBSTRING\_INDEX(Gpu, ' ', 1) FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

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

-- Updating gpu\_name column

SELECT Gpu, gpu\_brand, REPLACE(Gpu, gpu\_brand, ' ') FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

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

-- Now delete / drop Gpu column

ALTER TABLE data\_cleaning\_eda.laptop DROP COLUMN Gpu;

SELECT \* FROM data\_cleaning\_eda.laptop;

* # Create 3 new column(cpu\_brand, cpu\_name, cpu\_speed) from Cpu column

-- creating 3 new column first

ALTER TABLE data\_cleaning\_eda.laptop

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;

-- extract cpu\_brand info from Cpu column and insert into cpu\_brand column

SELECT Cpu, SUBSTRING\_INDEX(Cpu, ' ', 1) FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

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

-- extract cpu\_speed info from Cpu column and insert into cpu\_speed column

SELECT Cpu, REPLACE(SUBSTRING\_INDEX(Cpu, ' ', -1),'GHz',' ') FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

SET cpu\_speed = REPLACE(SUBSTRING\_INDEX(Cpu, ' ', -1),'GHz',' ');

-- extract cpu\_name info from Cpu column and insert into cpu\_name column

SELECT Cpu, REPLACE(REPLACE(Cpu,cpu\_brand, ''), cpu\_speed, ''),

REPLACE(REPLACE(Cpu,cpu\_brand,''),SUBSTRING\_INDEX(REPLACE(Cpu,cpu\_brand,''),' ',-1),'')

FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

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

-- Now delete / drop Cpu column

ALTER TABLE data\_cleaning\_eda.laptop DROP COLUMN Cpu;

SELECT \* FROM data\_cleaning\_eda.laptop;

* # Screenresolution Column have multiple information, Extract all into diff column .alter

-- we will breakdown our screenresolution column into three main column

-- lets create 2 column first resolution\_width, resolution\_height

ALTER TABLE data\_cleaning\_eda.laptop

ADD COLUMN resolution\_width INTEGER AFTER ScreenResolution,

ADD COLUMN resolution\_height INTEGER AFTER resolution\_width;

-- extracting both of information and updating

SELECT ScreenResolution,

SUBSTRING\_INDEX(ScreenResolution, ' ', -1),

SUBSTRING\_INDEX(SUBSTRING\_INDEX(ScreenResolution, ' ', -1),'x',1) AS 'resolution\_width',

SUBSTRING\_INDEX(SUBSTRING\_INDEX(ScreenResolution, ' ', -1),'x',-1) AS 'resolution\_height'

FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

SET resolution\_width = SUBSTRING\_INDEX(SUBSTRING\_INDEX(ScreenResolution, ' ', -1),'x',1),

resolution\_height = SUBSTRING\_INDEX(SUBSTRING\_INDEX(ScreenResolution, ' ', -1),'x',-1);

-- Create one more column from ScreenResolution that is tourch\_scree or not

ALTER TABLE data\_cleaning\_eda.laptop

ADD COLUMN touch\_screen INTEGER AFTER resolution\_height;

SELECT ScreenResolution, ScreenResolution LIKE '%Touch%' FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

SET touch\_screen = ScreenResolution LIKE '%Touch%';

SELECT \* FROM data\_cleaning\_eda.laptop

* # Extract information from Memory column. I will breakdown my Memory column into 3 new column

-- Create 3 new column(memory\_type, primary\_storage, secondary\_storage) from GPU column

SELECT \* FROM data\_cleaning\_eda.laptop;

ALTER TABLE data\_cleaning\_eda.laptop

ADD COLUMN memory\_type VARCHAR(255) AFTER Memory,

ADD COLUMN primary\_storage INTEGER AFTER memory\_type,

ADD COLUMN secondary\_storage INTEGER AFTER primary\_storage;

-- extracting and updating memory\_type column from Memory

UPDATE data\_cleaning\_eda.laptop

SET memory\_type = CASE

WHEN Memory LIKE '%SSD%' AND Memory LIKE '%HDD%' THEN 'Hybrid'

WHEN Memory LIKE '%SSD%' THEN 'SSD'

WHEN Memory LIKE '%HDD%' THEN 'HDD'

WHEN Memory LIKE '%Flash Storage%' THEN 'Flash Storage'

WHEN Memory LIKE '%Hybrid%' THEN 'Hybrid'

WHEN Memory LIKE '%Flash Storage%' AND Memory LIKE '%HDD%' THEN 'Hybrid'

ELSE NULL

END ;

-- extracting and updating primary storage and secondary storage info from memroy

SELECT Memory,

REGEXP\_SUBSTR(SUBSTRING\_INDEX(Memory,' ', 1),'[0-9]+') AS 'primary\_storage',

CASE WHEN Memory LIKE '%+%' THEN REGEXP\_SUBSTR(SUBSTRING\_INDEX(Memory, '+', -1),'[0-9]+') ELSE 0 END AS 'secondary\_storage'

FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

SET primary\_storage = REGEXP\_SUBSTR(SUBSTRING\_INDEX(Memory,' ', 1),'[0-9]+'),

secondary\_storage = CASE WHEN Memory LIKE '%+%' THEN REGEXP\_SUBSTR(SUBSTRING\_INDEX(Memory, '+', -1),'[0-9]+') ELSE 0 END;

-- Primary storage and secondary storage have tb like 1 2 lets convert it into gb

SELECT primary\_storage, secondary\_storage,

CASE WHEN primary\_storage <= 2 THEN primary\_storage\*1024 ELSE primary\_storage END,

CASE WHEN secondary\_storage <= 2 THEN secondary\_storage\*1024 ELSE secondary\_storage END

FROM data\_cleaning\_eda.laptop;

UPDATE data\_cleaning\_eda.laptop

SET primary\_storage = CASE WHEN primary\_storage <= 2 THEN primary\_storage\*1024 ELSE primary\_storage END,

secondary\_storage = CASE WHEN secondary\_storage <= 2 THEN secondary\_storage\*1024 ELSE secondary\_storage END;

SELECT \* FROM data\_cleaning\_eda.laptop

* Saving The cleaned laptop Data

USE data\_cleaning\_eda;

CREATE TABLE cleaned\_laptop LIKE data\_cleaning\_eda.laptop;

-- inserting clean data

INSERT INTO data\_cleaning\_eda.cleaned\_laptop

SELECT \* FROM data\_cleaning\_eda.laptop;