**## Exploratory Data Analysis Using SQL ##**

SELECT \* FROM data\_cleaning\_eda.cleaned\_laptop;

* How to see data : Head, Tail, Sample

-- Head

SELECT \* FROM data\_cleaning\_eda.cleaned\_laptop

ORDER BY `index` ASC LIMIT 10;

-- Tail

SELECT \* FROM data\_cleaning\_eda.cleaned\_laptop

ORDER BY `index` DESC LIMIT 10;

-- Sample

SELECT \* FROM data\_cleaning\_eda.cleaned\_laptop

ORDER BY RAND() LIMIT 5;

* # Univariate Analysis Numeric features #

-- 8 number summary : lets do in one numeric column Price

SELECT

COUNT(\*) OVER() ,

MIN(Price) OVER(),

MAX(Price) OVER(),

AVG(Price) OVER(),

STDDEV(Price) OVER(),

PERCENTILE\_CONT(0.25) WITHIN GROUP(ORDER BY Price) AS 'Q1',

PERCENTILE\_CONT(0.5) WITHIN GROUP(ORDER BY Price) AS 'Median',

PERCENTILE\_CONT(0.75) WITHIN GROUP(ORDER BY Price) AS 'Q3'

FROM data\_cleaning\_eda.cleaned\_laptop;

* # Is there any missing value in a particular column (Price)

SELECT COUNT(Price) FROM data\_cleaning\_eda.cleaned\_laptop

WHERE Price IS NULL;

* # Is there any outliers in a particular column

-- Extracting Outliers : IQR method

SELECT \* FROM (SELECT \*,

PERCENTILE\_CONT(0.25) WITHIN GROUP(ORDER BY Price) OVER() AS 'Q1',

PERCENTILE\_CONT(0.75) WITHIN GROUP(ORDER BY Price) OVER() AS 'Q3'

FROM datacleaning.laptopdata) t

WHERE t.price < t.Q1 - (1.5\*(t.Q3 - t.Q1)) OR

t.Price > t.Q3 + (1.5\*(t.Q3 - 5.Q2));

* # Create a histogram of a numeric Column

SELECT Price\_range, REPEAT("\*",COUNT(Price)/10) FROM

(SELECT Price,

CASE

WHEN Price BETWEEN 0 AND 25000 THEN '0-25k'

WHEN Price BETWEEN 25001 AND 50000 THEN '25-50k'

WHEN Price BETWEEN 50001 AND 75000 THEN '50-75k'

WHEN Price BETWEEN 75001 AND 100000 THEN '75-100k'

ELSE '>100k'

END AS 'Price\_range'

FROM data\_cleaning\_eda.cleaned\_laptop) t

GROUP BY t.Price\_range;

-- Take a challange Create a vertical Histogram

* # Univariate Analysis Categorical features #

-- How to deal with Categorical Column in SQL

SELECT Company, COUNT(Company) FROM data\_cleaning\_eda.cleaned\_laptop

GROUP BY Company;

-- copy the output and paste in online google sheet. Select the output on sheet and insert chart (pie)

* -- # Bivariate Analysis #

-- side by side 8 number analysis can be done

-- scatterplot

-- correlation

* # Scatte plot creating Data

SELECT Price, cpu\_speed FROM data\_cleaning\_eda.cleaned\_laptop;

-- select the output paste it into google sheet then apply scatter chart

* # Correlation

SELECT CORR(Inches,Price) FROM data\_cleaning\_eda.cleaned\_laptop;

* # Bivariate analysis on cateogical-categorical column

-- Contingency table

SELECT Company,

SUM(CASE WHEN touch\_screen = 1 THEN 1 ELSE 0 END) AS 'touch\_screen\_yes',

SUM(CASE WHEN touch\_screen = 0 THEN 1 ELSE 0 END) AS 'touch\_screen\_no'

FROM data\_cleaning\_eda.cleaned\_laptop

GROUP BY Company;

* # Bivariate analysis on numerical\_categorical column

SELECT Company, MIN(Price), MAX(Price), AVG(Price), STD(Price)

FROM data\_cleaning\_eda.cleaned\_laptop

GROUP BY Company;

-- How to treat missing value : replacing missing price with avg(price)

UPDATE data\_cleaning\_eda.cleaned\_laptop

SET Price = AVG(Price)

WHERE Price IS NULL;

-- Properly treat missing value like corresponding

UPDATE data\_cleaning\_eda.cleaned\_laptop l1

SET Price = (SELECT AVG(Price) FROM data\_cleaning\_eda.cleaned\_laptop l2

WHERE l2.Company = l1.Company AND l2.cpu\_name = l1.cpu\_name)

WHERE Price IS NULL;

* # Feature Engineering : Create New Feature

-- at first create the column with assigning data types

ALTER TABLE data\_cleaning\_eda.cleaned\_laptop

ADD COLUMN ppi INTEGER AFTER resolution\_height;

-- inserting info into the column,

UPDATE data\_cleaning\_eda.cleaned\_laptop

SET ppi = ROUND(SQRT((resolution\_width\*resolution\_width) + (resolution\_height\*resolution\_height)));

* # -- Create a feature from screensize(Inches) column like by some size range assign small, mid, large

-- Creating column at first

ALTER TABLE data\_cleaning\_eda.cleaned\_laptop

ADD COLUMN screen\_type VARCHAR(255) AFTER Inches;

-- now updating

UPDATE data\_cleaning\_eda.cleaned\_laptop

SET screen\_type = CASE

WHEN Inches < 14.0 THEN 'small'

WHEN Inches >= 14.0 AND Inches <= 17.0 THEN 'medium'

ELSE 'large'

END;

SELECT \* FROM data\_cleaning\_eda.cleaned\_laptop;

* # One Hot Encoding on gpu\_brand Column

SELECT gpu\_brand,

CASE WHEN gpu\_brand = 'Intel' THEN 1 ELSE 0 END AS 'intel',

CASE WHEN gpu\_brand = 'AMD' THEN 1 ELSE 0 END AS 'amd',

CASE WHEN gpu\_brand = 'nvidia' THEN 1 ELSE 0 END AS 'nvidia',

CASE WHEN gpu\_brand = 'arm' THEN 1 ELSE 0 END AS 'arm'

FROM data\_cleaning\_eda.cleaned\_laptop;