SQL CAPSTONE PROJECT

**create database amazon;**

create table amazondata(

invoice\_id VARCHAR(30),

branch VARCHAR(5),

city VARCHAR(30),

customer\_type VARCHAR(30),

gender VARCHAR(10),

product\_line VARCHAR(100),

unit\_price DECIMAL(10,2),

quantity INT,

VAT FLOAT(6,4),

total DECIMAL(10,2),

date DATE,

time TIME,

payment\_method VARCHAR(200),

cogs DECIMAL(10,2),

gross\_margin\_percentage FLOAT(11,9),

gross\_income DECIMAL(10,2),

rating FLOAT(2,1)

);

SELECT \* FROM amazondata;

**# Add a new column named timeofday to give insight of sales in the Morning,**

**#Afternoon and Evening. This will help answer the question on which part of the day most sales are made.**

ALTER TABLE amazondata

ADD COLUMN timeofday VARCHAR(10);

UPDATE amazondata

SET timeofday =

IF(HOUR(time) >= 0 AND HOUR(time) < 12, 'Morning',

IF(HOUR(time) >= 12 AND HOUR(time) < 18, 'Afternoon', 'Evening'));

**#Add a new column named dayname that contains the extracted days of the week on which the given transaction took place (Mon, Tue, Wed, Thur, Fri).**

**#This will help answer the question on which week of the day each branch is busiest.**

ALTER TABLE amazondata

ADD COLUMN dayname VARCHAR(10);

UPDATE amazondata

SET dayname = DAYNAME(date);

**#Add a new column named monthname that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar).**

**#Help determine which month of the year has the most sales and profit.**

ALTER TABLE amazondata

ADD COLUMN monthname VARCHAR(10);

UPDATE amazondata

SET monthname = MONTHNAME(date);

**Business Questions To Answer:**

**#1-What is the count of distinct cities in the dataset? ANSWER-3**

SELECT

COUNT(DISTINCT city)

FROM

amazondata;

**#2-For each branch, what is the corresponding city?**

SELECT DISTINCT

branch,city

FROM

amazondata

ORDER BY branch;

**#3-What is the count of distinct product lines in the dataset?ANSWER-6**

SELECT

COUNT(DISTINCT product\_line)

FROM

amazondata;

**#4-Which payment method occurs most frequently?ANSWER=Cash**

SELECT

payment\_method, COUNT(\*) AS count\_pm

FROM

amazondata

GROUP BY payment\_method

ORDER BY count\_pm DESC

LIMIT 1;

**#5-Which product line has the highest sales?ANSWER=Fashion accessories**

SELECT

product\_line, COUNT(invoice\_id) AS sales\_count

FROM

amazondata

GROUP BY product\_line

ORDER BY sales\_count desc

limit 1;

**#6=How much revenue is generated each month?**

SELECT

monthname, SUM(total) as revenue

FROM

amazondata

GROUP BY monthname

ORDER BY SUM(total) DESC;

**#7-In which month did the cost of goods sold reach its peak?**

SELECT

monthname, SUM(cogs)

FROM

amazondata

GROUP BY monthname

ORDER BY SUM(cogs) DESC

LIMIT 1;

**#8-Which product line generated the highest revenue?**

SELECT

product\_line, SUM(total) AS revenue

FROM

amazondata

GROUP BY product\_line

ORDER BY SUM(total) DESC

LIMIT 1;

**#9-In which city was the highest revenue recorded?**

SELECT

city, SUM(total) as revenue

FROM

amazondata

GROUP BY city

ORDER BY revenue DESC

LIMIT 1;

**#10-Which product line incurred the highest Value Added Tax?**

SELECT

product\_line, SUM(vat) as VAT

FROM

amazondata

GROUP BY product\_line

ORDER BY VAT DESC

LIMIT 1;

**#11-For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."**

SELECT

product\_line,

SUM(total) AS revenue,

CASE

WHEN

SUM(total) < (SELECT AVG(total) FROM amazondata) THEN 'bad'

ELSE 'good'

END AS category\_of\_sales

FROM

amazondata

GROUP BY product\_line

ORDER BY revenue DESC;

**#12-Identify the branch that exceeded the average number of products sold**

SELECT

branch, SUM(quantity) AS total\_quantity

FROM

amazondata

GROUP BY branch

HAVING SUM(quantity) > (SELECT AVG(quantity) FROM amazondata);

**#13-Which product line is most frequently associated with each gender?**

WITH freq AS (

SELECT

product\_line,

gender,

ROW\_NUMBER() OVER (PARTITION BY gender ORDER BY COUNT(\*) DESC) AS rank1

FROM

amazondata

GROUP BY

product\_line, gender

)

SELECT

product\_line,

gender

FROM

freq

WHERE

rank1 = 1;

**#14-Calculate the average rating for each product line.**

SELECT

product\_line, AVG(rating) as avg\_rating

FROM

amazondata

GROUP BY product\_line

ORDER BY avg\_rating DESC;

**#15-Count the sales occurrences for each time of day on every weekday.**

SELECT

dayname,timeofday,COUNT(invoice\_id) AS sales\_occurence

FROM

amazondata

WHERE dayname NOT IN ("saturday","sunday")

GROUP BY dayname,timeofday

ORDER BY dayname DESC,COUNT(invoice\_id) DESC;

**#16-Identify the customer type contributing the highest revenue.**

SELECT

customer\_type, SUM(total) as revenue

FROM

amazondata

GROUP BY customer\_type

ORDER BY revenue DESC

LIMIT 1;

**#17-Determine the city with the highest VAT percentage.**

SELECT

city, MAX(vat)

FROM

amazondata

GROUP BY city

ORDER BY MAX(vat) DESC

LIMIT 1;

**#18-Identify the customer type with the highest VAT payments.**

SELECT

customer\_type, MAX(vat)

FROM

amazondata

GROUP BY customer\_type

ORDER BY MAX(vat) DESC

LIMIT 1;

**#19-What is the count of distinct customer types in the dataset?**

SELECT

COUNT(DISTINCT customer\_type) AS unique\_customer\_type

FROM

amazondata;

**#20-What is the count of distinct payment methods in the dataset?**

SELECT

COUNT(DISTINCT payment\_method) AS unique\_payment\_method

FROM

amazondata;

**#21-Which customer type occurs most frequently?**

SELECT

COUNT(\*) AS occurence, customer\_type

FROM

amazondata

GROUP BY customer\_type

ORDER BY COUNT(\*) DESC

LIMIT 1;

**#22-Identify the customer type with the highest purchase frequency.**

**SELECT**

customer\_type, COUNT(invoice\_id) as purchase\_freq

FROM

amazondata

GROUP BY customer\_type

ORDER BY COUNT(invoice\_id) DESC

LIMIT 1;

**#24-Examine the distribution of genders within each branch.**

SELECT

branch, gender, COUNT(\*) AS gender\_count

FROM

amazondata

GROUP BY branch , gender

ORDER BY branch , gender\_count DESC;

**#25-Identify the time of day when customers provide the most ratings.**

SELECT

timeofday,

COUNT(\*) AS rating\_count

FROM

amazondata

GROUP BY

timeofday

ORDER BY

rating\_count DESC

LIMIT 1;

**#26-Determine the time of day with the highest customer ratings for each branch**.

SELECT

branch, timeofday, COUNT(\*) AS rating\_count

FROM

amazondata

GROUP BY branch , timeofday

ORDER BY branch , rating\_count DESC;

**#27-Identify the day of the week with the highest average ratings.**

SELECT

dayname, AVG(rating) AS avg\_rating

FROM

amazondata

GROUP BY dayname

ORDER BY avg\_rating DESC

LIMIT 1;

**#28-Determine the day of the week with the highest average ratings for each branch.**

**WITH high\_avg\_r AS (**

SELECT

branch,timeofday,AVG(rating) AS avg\_rating,

row\_number() OVER (PARTITION BY branch ORDER BY AVG(rating) DESC) AS rownum

FROM

amazondata

GROUP BY

branch,timeofday

)

SELECT

branch,timeofday,avg\_rating FROM high\_avg\_r

WHERE

rownum=1

ORDER BY

branch;