SQL CASE STUDY : DATA MART ANALYSIS

BY- NITESH KUMAR

INTRODUCTION:

Data Dart is a venture and I will be analyzing the sales and performance of the venture. In June 2020 - large scale supply changes were made at Data Mart. All Data Mart products now use sustainable packaging methods in every single step from the farm all the way to the customer.

I will quantify the impact of this change on the sales performance for Data Mart and its separate business areas.

SCHEMA USED: WEEKLY_SALES TABLE

Column name	Data type
week_date	date
region	varchar(20)
platform	varchar(20)
segment	varchar(10)
customer	varchar(20)
transactions	int
sales	int

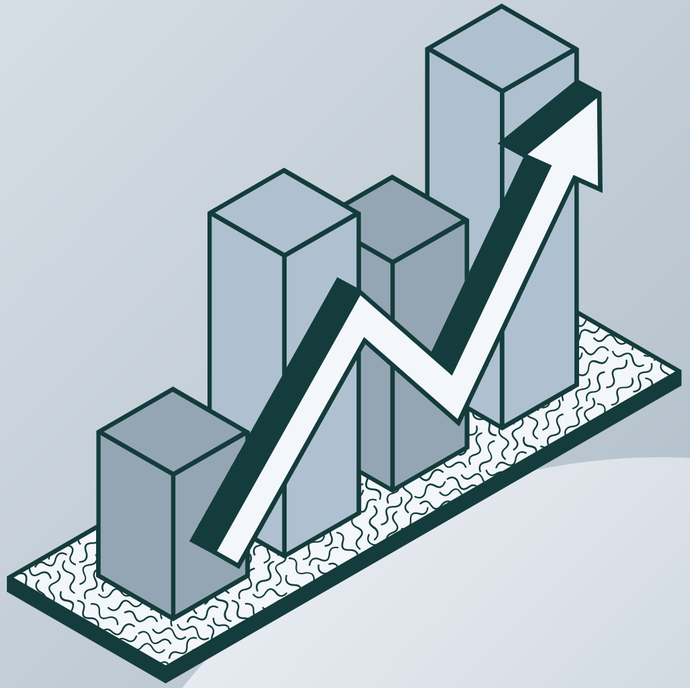
CASE STUDY QUESTIONS

A. Data Cleansing Steps

In a single query, perform the following operations and generate a new table in the data_mart schema named clean_weekly_sales:

1. Add a week_number as the second column for each week_date value, for example any value from the 1st of January to 7th of January will be 1, 8th to 14th will be 2, etc.
2. Add a month_number with the calendar month for each week_date value as the 3rd column
3. Add a calendar_year column as the 4th column containing either 2018, 2019 or 2020 values
4. Add a new column called age_band after the original segment column using the following mapping on the number inside the segment value

segment	age_band
1	Young Adults



DATA EXPLORATION

B. Data Exploration

1. Which week numbers are missing from the dataset?
2. How many total transactions were there for each year in the dataset?
3. What are the total sales for each region for each month?
4. What is the total count of transactions for each platform
5. What is the percentage of sales for Retail vs Shopify for each month?
6. What is the percentage of sales by demographic for each year in the dataset?
7. Which age_band and demographic values contribute the most to Retail sales?

1. WHICH WEEK NUMBERS ARE MISSING FROM THE DATASET?

```sql
create table seq100 (x int auto_increment primary key)

insert into seq100 values (),(),(),(),(),(),(),(),(),();
insert into seq100 values (),(),(),(),(),(),(),(),(),();
insert into seq100 values (),(),(),(),(),(),(),(),(),();
insert into seq100 values (),(),(),(),(),(),(),(),(),();
insert into seq100 values (),(),(),(),(),(),(),(),(),();

insert into seq100 select x+50 from seq100

create table seq52 as (select x from seq100 limit 52)

select distinct x as week_day from seq52
where x not in (select distinct week_number from clean_weekly_sales)
```

week_day
1
2
3
4
5

2. HOW MANY TOTAL TRANSACTIONS WERE THERE FOR EACH YEAR IN THE DATASET?

```sql
SELECT
    calendar_year, sum(transactions) AS total_transactions
FROM
    clean_weekly_sales
GROUP BY 1
```

Result Grid | Filter Rows:

calendar_year	total_transactions
2020	375813651
2019	365639285
2018	346406460

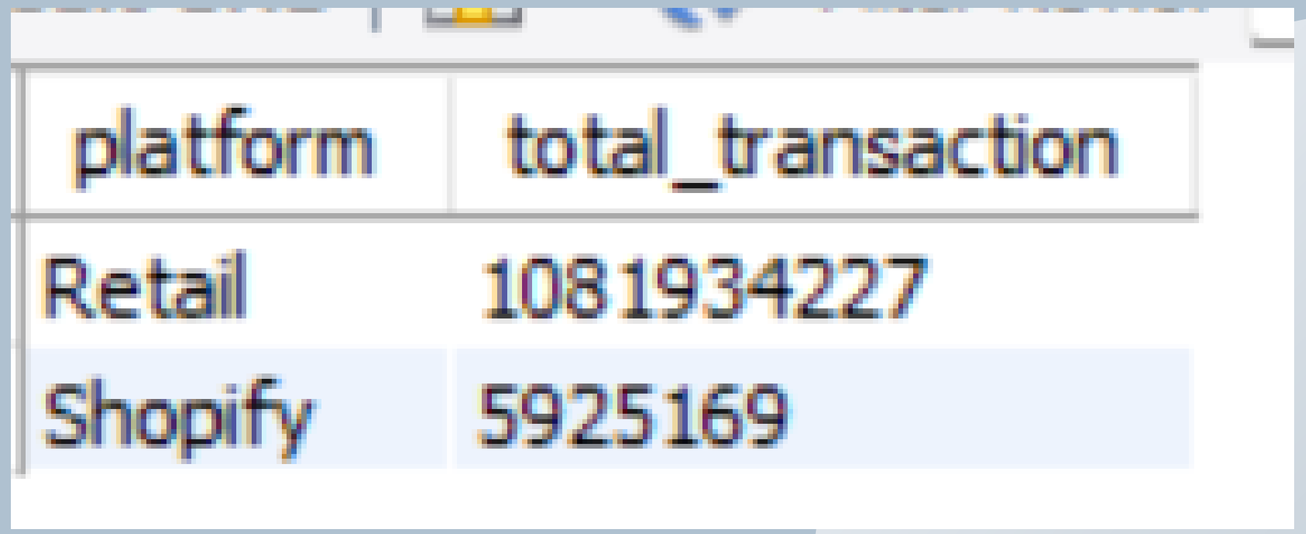
3. WHAT ARE THE TOTAL SALES FOR EACH REGION FOR EACH MONTH?

```sql
SELECT
   region, month_number, SUM(sales) AS total_sales
FROM
    clean_weekly_sales
GROUP BY 1,2
```

region	month_number	total_sales
ASIA	8	1663320609
USA	8	712002790
EUROPE	8	122102995
AFRICA	8	1809596890
CANADA	8	447073019

4. WHAT IS THE TOTAL COUNT OF TRANSACTIONS FOR EACH PLATFORM?

```sql
SELECT
    platform, SUM(transactions) AS total_transaction
FROM
    clean_weekly_sales
GROUP BY platform
```



platform	total_transaction
Retail	1081934227
Shopify	5925169

5. WHAT IS THE PERCENTAGE OF SALES FOR RETAIL VS SHOPIFY FOR EACH MONTH?

```sql
SELECT
  month_number,calendar_year,
  ROUND(
    100 * MAX(CASE WHEN platform = 'Retail' THEN monthly_sales ELSE NULL END) /
      SUM(monthly_sales),
    2
  ) AS retail_percentage,
  ROUND(
    100 * MAX(CASE WHEN platform = 'Shopify' THEN monthly_sales ELSE NULL END) /
      SUM(monthly_sales),
    2
  ) AS shopify_percentage
FROM monthly_platform_sales
GROUP BY month_number,calendar_year
ORDER BY month_number,calendar_year;
```

month_number	calendar_year	retail_percentage	shopify_percentage
3	2018	97.92	2.08
3	2019	97.71	2.29
3	2020	97.3	2.7
4	2018	97.93	2.07
4	2019	97.8	2.2

6. WHAT IS THE PERCENTAGE OF SALES BY DEMOGRAPHIC FOR EACH YEAR IN THE DATASET?

```sql
SELECT
  calendar_year,
  demographic,
  SUM(SALES) AS yearly_sales,
  ROUND(
    (
      100 * SUM(sales)/
        SUM(SUM(SALES)) OVER (PARTITION BY demographic)
    ),
    2
  ) AS percentage
FROM clean_weekly_sales
GROUP BY
  calendar_year,
  demographic
ORDER BY
  calendar_year,
  demographic;
```

calendar_year	demographic	yearly_sales	percentage
2018	Couples	3402388688	30.38
2018	Families	4125558033	31.25
2018	Unknown	5369434106	32.86
2019	Couples	3749251935	33.47
2019	Families	4463918344	33.81

7. WHICH AGE_BAND AND DEMOGRAPHIC VALUES CONTRIBUTE THE MOST TO RETAIL SALES?

```sql
SELECT
    age_band, demographic, SUM(sales) AS total_sales
FROM
    clean_weekly_sales
WHERE
    platform = 'Retail'
GROUP BY age_band , demographic
ORDER BY total_sales DESC
```

age_band	demographic	total_sales
Unknown	Unknown	16067285533
Retirees	Families	6634686916
Retirees	Couples	6370580014
Middle Aged	Families	4354091554
Young Adults	Couples	2602922797

IMPORTANT FINDINGS

- Maximum customer prefer Retail than Shopify.
- Sales is more in Asia than other Continents.
- Sales are increasing year by year.
- People prefer to shop more in the beginning and the end of the year.