SQL Case Study : Data Mart Analysis

INTRODUCTION:

Data Dart is my latest venture and I want your help to analyze the sales and performance of my 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 need your help to 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
2	Middle Aged
3 or 4	Retirees

5. Add a new demographic column using the following mapping for the first letter in the segment values:

segment | demographic |
C | Couples |
F | Families |

6. Ensure all null string values with an "unknown" string value in the original segment column as well as the
   new age_band and demographic columns
7. Generate a new avg_transaction column as the sales value divided by transactions rounded to 2 decimal places for each record

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?

First we have to create a Database :
Syntax: Create Database Database Name

Then we use this Databaes:
Syntax: Use Database Name

- create database case1;
- use case1;

Data:

week_date	region	platform	segment	customer_type	transactions	sales
2020-08-31	ASIA	Retail	C3	New	120631	3656163
2020-08-31	ASIA	Retail	F1	New	31574	996575
2020-08-31	USA	Retail	null	Guest	529151	16509610
2020-08-31	EUROPE	Retail	C1	New	4517	141942
2020-08-31	AFRICA	Retail	C2	New	58046	1758388
2020-08-31	CANADA	Shopify	F2	Existing	1336	243878
2020-08-31	AFRICA	Shopify	F3	Existing	2514	519502
2020-08-31	ASIA	Shopify	F1	Existing	2158	371417
2020-08-31	AFRICA	Shopify	F2	New	318	49557
2020-08-31	AFRICA	Retail	C3	New	111032	3888162

Data Cleansing

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
2	Middle Aged
3 or 4	Retirees

5. Add a new demographic column using the following mapping for the first letter in the segment values:

segment | demographic |

C | Couples |

F | Families |

6. Ensure all null string values with an "unknown" string value in the original segment column as well as the new age_band and demographic columns

7. Generate a new avg_transaction column as the sales value divided by transactions rounded to 2 decimal places for each record.

Solution of Data Cleansing:-

```
CREATE TABLE clean_weekly_sales AS
SELECT week_date,
            week(week_date) AS week_number,
            month(week_date) AS month_number,
            year(week_date) AS calendar_year,

           region,platform,

  CASE WHEN segment = 'null' THEN 'Unknown'
      ELSE segment
      END AS segment,

  CASE
    WHEN right(segment, 1) = '1' THEN 'Young Adults'
    WHEN right(segment, 1) = '2' THEN 'Middle Aged'
    WHEN right(segment, 1) IN ('3', '4') THEN 'Retirees'
    ELSE 'Unknown'
    END AS age_band,

  CASE WHEN left(segment, 1) = 'C' THEN 'Couples'
      WHEN left(segment, 1) = 'F' THEN 'Families'
      ELSE 'Unknown'
      END AS demographic,

  customer_type,transactions,sales,
  ROUND(sales / transactions,2) AS avg_transaction

FROM weekly_sales;

select * from clean_weekly_sales limit 10;
```

Output:-

week_date	week_number	month_number	calendar_year	region	platform	segment	age_band	demographic	customer_type	transactions	sales	avg_transaction
2020-08-31	35	8	2020	ASIA	Retail	C3	Retirees	Couples	New	120631	3656163	30.31
2020-08-31	35	8	2020	ASIA	Retail	F1	Young Adults	Families	New	31574	996575	31.56
2020-08-31	35	8	2020	USA	Retail	Unknown	Unknown	Unknown	Guest	529151	16509610	31.20
2020-08-31	35	8	2020	EUROPE	Retail	C1	Young Adults	Couples	New	4517	141942	31.42
2020-08-31	35	8	2020	AFRICA	Retail	C2	Middle Aged	Couples	New	58046	1758388	30.29
2020-08-31	35	8	2020	CANADA	Shopify	F2	Middle Aged	Families	Existing	1336	243878	182.54
2020-08-31	35	8	2020	AFRICA	Shopify	F3	Retirees	Families	Existing	2514	519502	206.64
2020-08-31	35	8	2020	ASIA	Shopify	F1	Young Adults	Families	Existing	2158	371417	172.11
2020-08-31	35	8	2020	AFRICA	Shopify	F2	Middle Aged	Families	New	318	49557	155.84
2020-08-31	35	8	2020	AFRICA	Retail	C3	Retirees	Couples	New	111032	3888162	35.02

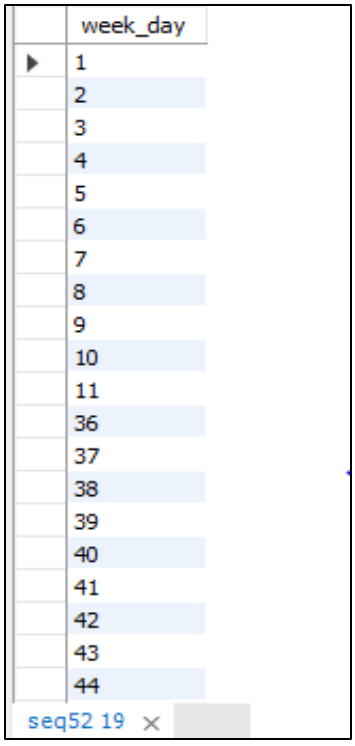
Data Exploration

1. Which week numbers are missing from the dataset?

Solution:-

```
-- .. (Create table and inserting values 1 to 100)

create table seq100(
x int not null 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;

select * from seq100;

-- .. In One Year we have total 52 Weeks. so creating 52 weeks table.
create table seq52 as (select x from seq100 limit 52);

-- .. The Output we are getting after running this query Conclusion:- These are week days which are not present in the Dataset.
select distinct x as week_day
from seq52
where x not in (select distinct week_number from clean_weekly_sales);

select distinct week_number
from clean_weekly_sales
order by 1 asc;
```

Output:



week_day
1
2
3
4
5
6
7
8
9
10
11
36
37
38
39
40
41
42
43
44

seq52 19

2. How many total transactions were there for each year in the dataset?

Solution:-

```
select calendar_year,sum(transactions) as Total_Transaction
from clean_weekly_sales
group by 1
order by 2 asc;
```

Output:

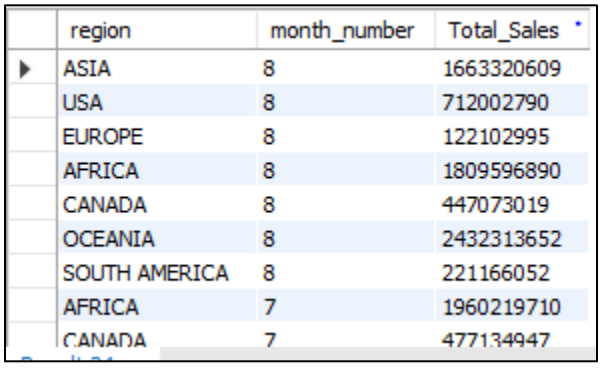
calendar_year	Total_Transaction
2018	346406460
2019	365639285
2020	375813651

3. What are the total sales for each region for each month?

Solution:-

```
select region,month_number, sum(sales) as Total_Sales
from clean_weekly_sales
group by 1,2;
```

Output:



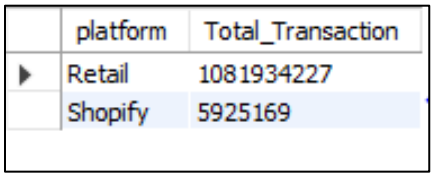
region	month_number	Total_Sales
ASIA	8	1663320609
USA	8	712002790
EUROPE	8	122102995
AFRICA	8	1809596890
CANADA	8	447073019
OCEANIA	8	2432313652
SOUTH AMERICA	8	221166052
AFRICA	7	1960219710
CANADA	7	477134947

4. What is the total count of transactions for each platform?

Solution:-

```
select platform,sum(transactions) as Total_Transaction
from clean_weekly_sales
group by 1;
```

Output:



platform	Total_Transaction
Retail	1081934227
Shopify	5925169

5. What is the percentage of sales for Retail vs Shopify for each month?

Solution:-

```
-- .. CTE (Create Temp Table) --> Commom table Expression
with cte_monthly_platform_sales as(
select month_number,calendar_year,platform, sum(sales) as monthly_slaes
from clean_weekly_sales
group by 1,2,3)

select month_number,calendar_year,
round(100*max(case when platform = "Retail"
                   then monthly_slaes
                  Else null End)/sum(monthly_slaes),2) as Retail_Precentage,
round(100*max(case when platform = "Shopify"
                   then monthly_slaes
                   Else null End)/sum(monthly_slaes),2) as Shopify_Percentage
```

```
from cte_monthly_platform_sales
group by month_number,calendar_year;

-- ..... Without CTE.....
select month_number,calendar_year,
round(100*max(case when platform = "Retail"
                   then monthly_slaes
                   Else null End)/sum(monthly_slaes),2) as Retail_Precentage,
round(100*max(case when platform = "Shopify"
                    then monthly_slaes
                    Else null End)/sum(monthly_slaes),2) as Shopify_Percentage
from
(select month_number,calendar_year,platform, sum(sales) as monthly_slaes
from clean_weekly_sales
group by 1,2,3) a
group by 1,2;
```

Hint:-

Monthly_Sales :- Retail -→ Retail % = R (Sales Made by Retail) **/ Monthly Sales * 100**

Shopify -→ Shopify% = S (Sales Made by Shopify) **/ Monthly Sales * 100**

Output:

month_number	calendar_year	Retail_Precentage	Shopify_Percentage
8	2020	96.51	3.49
7	2020	96.67	3.33
6	2020	96.80	3.20
5	2020	96.71	3.29
4	2020	96.96	3.04
3	2020	97.30	2.70
9	2019	97.09	2.91
8	2019	97.21	2.79
7	2019	97.35	2.65
6	2019	97.42	2.58
5	2019	97.52	2.48
4	2019	97.80	2.20

6. What is the percentage of sales by demographic for each year in the dataset?

Solution:-

```
select calendar_year,demographic,sum(sales) as
Yearly_Sales,round(100*sum(sales)/sum(sum(sales))
over (partition by demographic order by calendar_year),2) as Percentage
```

from clean_weekly_sales
group by 1,2;

Output:-

calendar_year	demographic	Yearly_Sales	Percentage
2018	Couples	3402388688	100.00
2019	Couples	3749251935	52.43
2020	Couples	4049566928	36.15
2018	Families	4125558033	100.00
2019	Families	4463918344	51.97
2020	Families	4614338065	34.95
2018	Unknown	5369434106	100.00
2019	Unknown	5532862221	50.75
2020	Unknown	5436315907	33.27

7. Which age_band and demographic values contribute the most to Retail sales?

Solution:-
select age_band,demographic,sum(sales) as Total_Sales
from clean_weekly_sales
where platform = "Retail"
group by 1,2
order by 3 desc;

Output:-

age_band	demographic	Total_Sales
Unknown	Unknown	16067285533
Retirees	Families	6634686916
Retirees	Couples	6370580014
Middle Aged	Families	4354091554
Young Adults	Couples	2602922797
Middle Aged	Couples	1854160330
Young Adults	Families	1770889293