SQL Case Study 1: 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.

Done By : L. ANUSH BHARATHWAJ

LINK

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.**

```
create table clean_weekly_sales as
select
week_date,
week(week_date) as week_number ,
month(week_date) as month_number,
monthname(week_date) as month_name,
year(week_date) as calendar_year,
        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',
platform,
region,
Round(sales/transactions,2) as 'avg_transaction',
transactions,
sales
from weekly_sales;

select * from clean_weekly_sales limit 10;
```

week_date	week_number	month_number	month_name	calendar_year	segment	age_band	demographic	platform	region	avg_transaction	transactions	sales
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Retail	ASIA	30.31	120631	3656163
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Retail	ASIA	31.56	31574	996575
2020-08-31	35	8	August	2020	Unknown	Unknown	Unknown	Retail	USA	31.20	529151	16509610
2020-08-31	35	8	August	2020	C1	Young Adults	Couples	Retail	EUROPE	31.42	4517	141942
2020-08-31	35	8	August	2020	C2	Middle Aged	Couples	Retail	AFRICA	30.29	58046	1758388
2020-08-31	35	8	August	2020	F2	Middle Aged	Families	Shopify	CANADA	182.54	1336	243678
2020-08-31	35	8	August	2020	F3	Retirees	Families	Shopify	AFRICA	206.64	2514	519502
2020-08-31	35	8	August	2020	F1	Young Adults	Families	Shopify	ASIA	172.11	2158	371417
2020-08-31	35	8	August	2020	F2	Middle Aged	Families	Shopify	AFRICA	155.84	318	49557
2020-08-31	35	8	August	2020	C3	Retirees	Couples	Retail	AFRICA	35.02	111032	3888162

B. Data Exploration

1. Which week numbers are missing from the dataset?

```
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;

select * from seq100;

create table seq52 as select x from seq100 limit 52;

select * from seq52;

select x as 'Miss_week_numbers' from seq52 where x not in (select
distinct week_number from clean_weekly_sales);
```

Miss_week_numbers
1
2
3
4
5
6
7
8
9
10
11
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

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

```
select calendar_year as 'Year' ,sum(avg_transaction) as 'Total
Transactions' from clean_weekly_sales group by Year;
```

Year	Total Transactions
2020	615332.54
2019	626260.68
2018	657630.28

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

```
select month_name , month_number ,region ,sum(sales) from
clean_weekly_sales group by month_name,
month_number,region;
```

month_name	month_number	region	sum(sales)
August	8	ASIA	1663320609
August	8	USA	712002790
August	8	EUROPE	122102995
August	8	AFRICA	1809596890
August	8	CANADA	447073019
August	8	OCEANIA	2432313652
August	8	SOUTH AMERI…	221166052
July	7	AFRICA	1960219710
July	7	CANADA	477134947
July	7	USA	760331754
July	7	EUROPE	136757466
July	7	OCEANIA	2563459400
July	7	SOUTH AMERI…	235582776
July	7	ASIA	1768844756
June	6	OCEANIA	2371884744
June	6	USA	703878990
June	6	SOUTH AMERI…	218247455

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

```
select platform , count(transactions) from clean_weekly_sales group
by platform;
```

platform	count(transactio...
Retail	8568
Shopify	8549

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

platform	sum(transactions)
Retail	1081934227
Shopify	5925169

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

```
with cte_montly_sales as
( select month_number , calendar_year , platform , SUM(sales)
as saless from clean_weekly_sales group by month_number ,
calendar_year , platform )
select month_number , calendar_year ,
        ROUND( 100 * MAX(case when platform = 'Retail' then
saless else NULL end) / SUM(saless),2) as 'retail_perc',
        ROUND( 100 * MAX(case when platform = 'Shopify' then
saless else NULL end) / SUM(saless),2) as 'shopify_perc'
from cte_montly_sales
group by month_number , calendar_year
order by month_number , calendar_year ;
```

month_number	calendar_year	retail_perc	shopify_perc
3	2018	97.92	2.08
3	2019	97.71	2.29
3	2020	97.30	2.70
4	2018	97.93	2.07
4	2019	97.80	2.20
4	2020	96.96	3.04
5	2018	97.73	2.27
5	2019	97.52	2.48
5	2020	96.71	3.29
6	2018	97.76	2.24
6	2019	97.42	2.58
6	2020	96.80	3.20
7	2018	97.75	2.25
7	2019	97.35	2.65
7	2020	96.67	3.33
8	2018	97.71	2.29
8	2019	97.21	2.79
8	2020	96.51	3.49
9	2018	97.68	2.32
9	2019	97.09	2.91

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

```
select calendar_year, demographic , sum(sales) as year_sales,
round( 100 * sum(sales)/sum(sum(sales)) over(partition by
demographic),2)as perc from clean_weekly_sales
group by calendar_year, demographic
order by calendar_year, demographic;
```

calendar_year	demographic	year_sales	perc
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
2019	Unknown	5532862221	33.86
2020	Couples	4049566928	36.15
2020	Families	4614338065	34.95
2020	Unknown	5436315907	33.27

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

```
select age_band , demographic , sum(sales) as Sales from clean_weekly_sales
where platform = 'Retail' group by age_band , demographic order by  Sales
desc;
```

age_band	demographic	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