DATAMART ANALYSIS

By Akash Dutta

Objective

Data Dart is the latest venture and assistance is needed to analyze its sales and performance. In June 2020, large-scale supply changes were implemented at Data Mart, ensuring all products now use sustainable packaging methods at every step, from the farm to the customer. Help is required to quantify the impact of this change on the overall sales performance for Data Mart and its separate business areas.

Data Cleaning

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

Data Cleaning Codes

```
## 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,
```

```sql
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 'Retires'
    else 'UNKNOWN'
    End as age_band,
```

```sql
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_transactions'
from weekly_sales;

select * from clean_weekly_sales limit 10;
```

week_date	week_number	month_number	calendar_year	region	platform	segment	age_band	Demographic	customer_type	transactions	sales
2020-08-31	35	8	2020	ASIA	Retail	C3	Retires	Couples	New	120631	365616
2020-08-31	35	8	2020	ASIA	Retail	F1	Young Adults	Families	New	31574	996575
2020-08-31	35	8	2020	USA	Retail	null	UNKNOWN	UNKNOWN	Guest	529151	165096
2020-08-31	35	8	2020	EUROPE	Retail	C1	Young Adults	Couples	New	4517	141942

Data Exploration

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

```
SELECT
  calendar_year,
  SUM(transactions) AS total_transactions
FROM clean_weekly_sales group by calendar_year;
```

calendar_year	total_transactions
2020	375813651
2019	365639285
2018	346406460

2.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 month_number,region;
```

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

3.What is the total count of transactions for each platform

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

platform	total_trasactions
Retail	1081934227
Shopify	5925169

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

First, we create a CTE..

```
## first, we create a CTE

WITH cte_monthly_platform_sales AS (
  SELECT
    month_number,calendar_year,
    platform,
    SUM(sales) AS monthly_sales
  FROM clean_weekly_sales
  GROUP BY month_number,calendar_year, platform
)
```

```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 cte_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	97.30	2.70
4	2018	97.93	2.07
4	2019	97.80	2.20

5.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

6. 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
Retires	Families	6634686916
Retires	Couples	6370580014
Middle Aged	Families	4354091554
Young Adults	Couples	2602922797

Thank you..