**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**

## 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 |

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

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

1. 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
2. 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

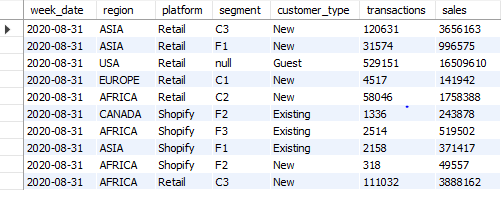
Then we use this Databaes:

Syntax: Use Database Name

* create database case1;
* use case1;

-----------------------------------------------------------------------------

Data:



**Data Cleansing**

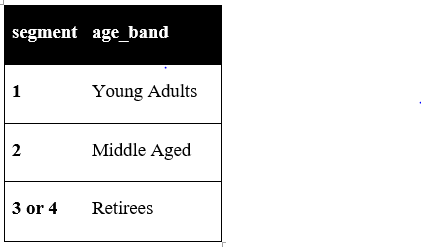
In a single query, perform the following operations and generate a new table in the data\_mart schema named clean\_weekly\_sales:

**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



**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,

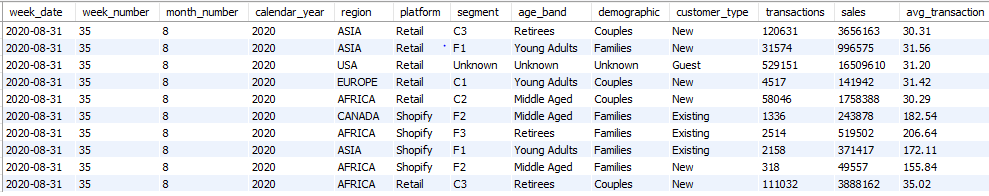
customer\_type,transactions,sales,

ROUND(sales / transactions,2) AS avg\_transaction

FROM weekly\_sales;

select \* from clean\_weekly\_sales limit 10;

Output:-



**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

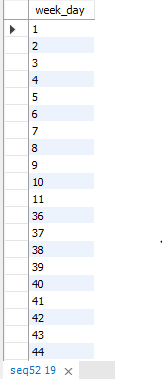
where x not in (select distinct week\_number from clean\_weekly\_sales);

select distinct week\_number

from clean\_weekly\_sales

order by 1 asc;

Output:



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

Solution:-

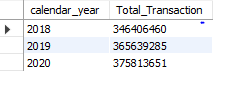
select calendar\_year,sum(transactions) as Total\_Transaction

from clean\_weekly\_sales

group by 1

order by 2 asc;

Output:



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

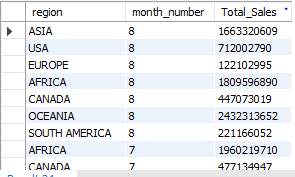
Solution:-

select region,month\_number, sum(sales) as Total\_Sales

from clean\_weekly\_sales

group by 1,2;

Output:



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

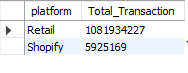
Solution:-

select platform,sum(transactions) as Total\_Transaction

from clean\_weekly\_sales

group by 1;

Output:



**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

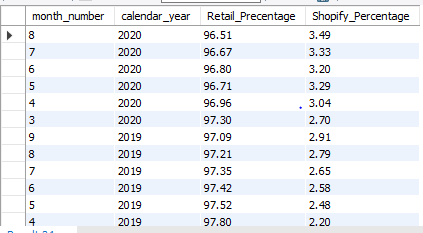
group by 1,2;

**Hint:-**

**Monthly\_Sales :- Retail -🡪 Retail % = R** (Sales Made by Retail) **/ Monthly Sales \* 100**

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

Output:



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

Solution:-

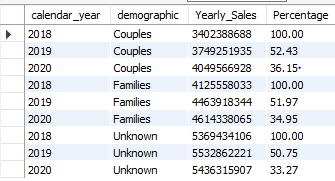
select calendar\_year,demographic,sum(sales) as Yearly\_Sales,round(100\*sum(sales)/sum(sum(sales))

over (partition by demographic order by calendar\_year),2) as Percentage

from clean\_weekly\_sales

group by 1,2;

Output:-



**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:-

