Data Mart Analysis

INTRODUCTION:

Data Dart is latest venture and I want to analyze the sales and performance of 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.

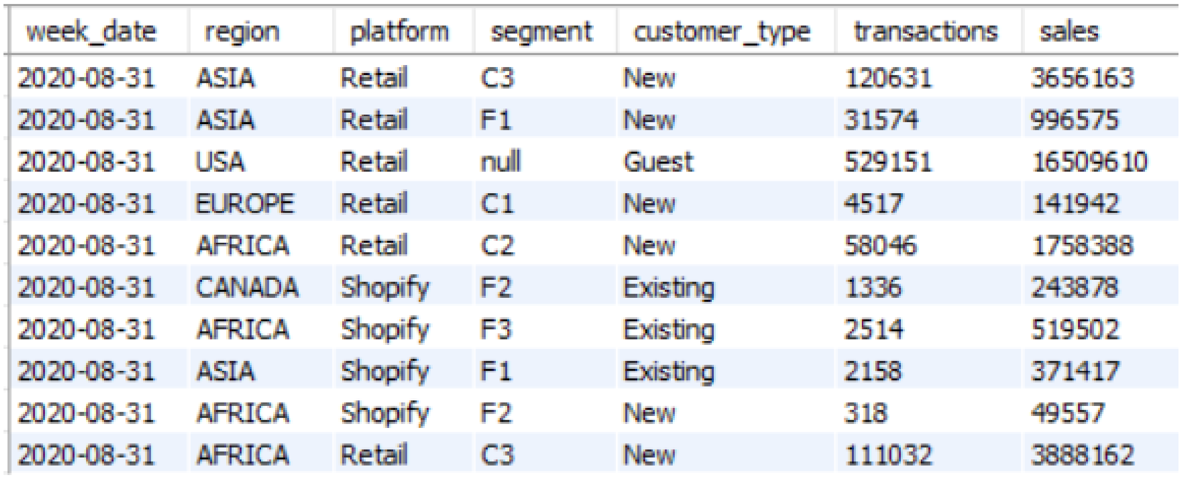
I need to quantify the impact of this change on the sales performance for Data Mart and its separate business areas.

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

Dataset:

select * from weekly_sales limit 10;



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

A. Data Cleaning

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

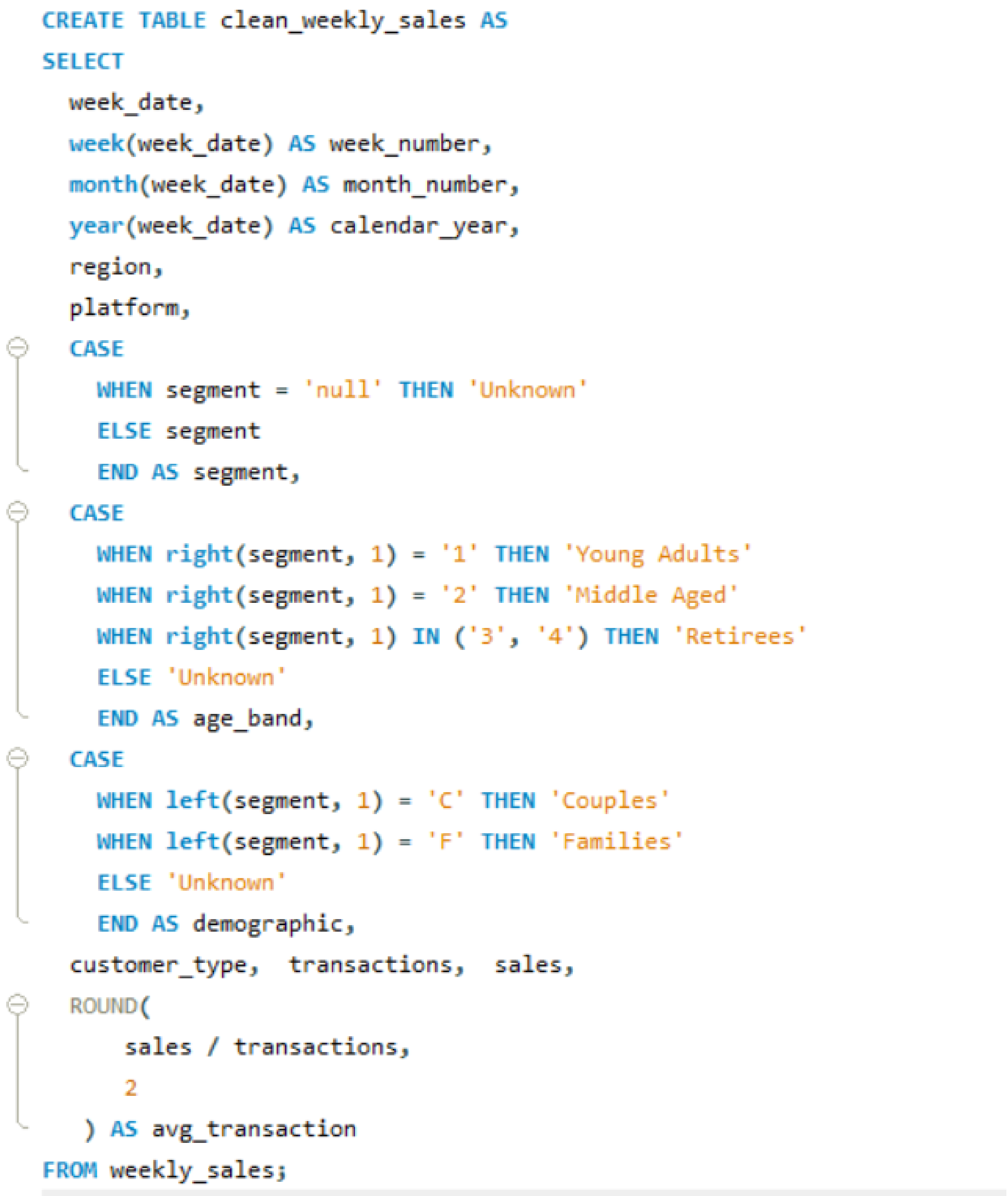
5. Add a new demographic column using the following mapping for the first letter in the segment values:

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

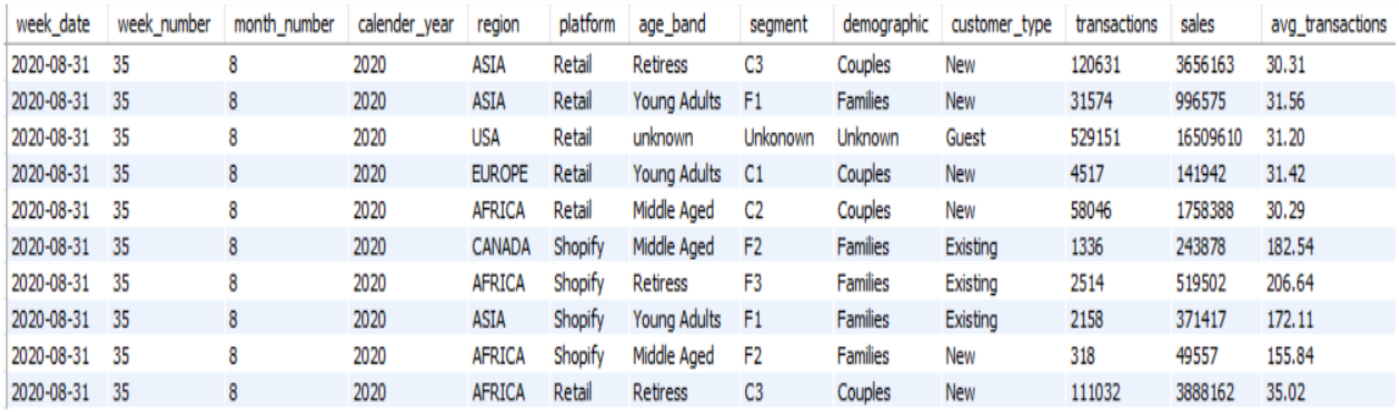
Solution:



```sql
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;
```

Cleaned Dataset:

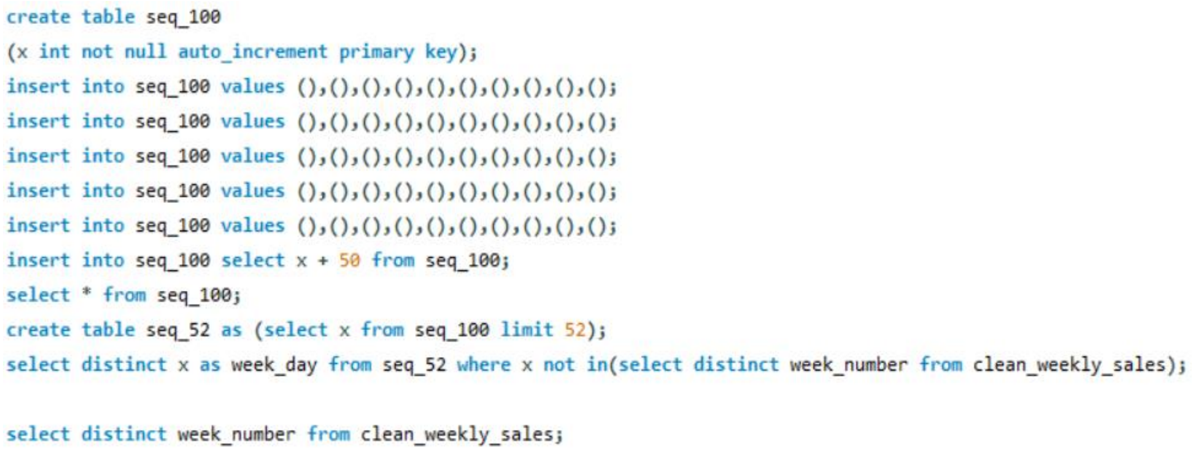
select * from clean_weekly_sales limit 10;



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

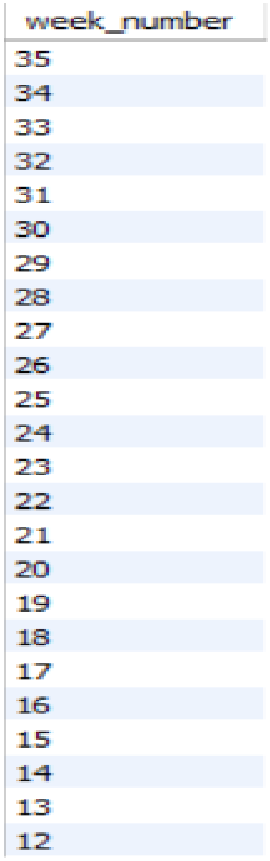
B. Data Exploration

1. Which week numbers are missing from the dataset?



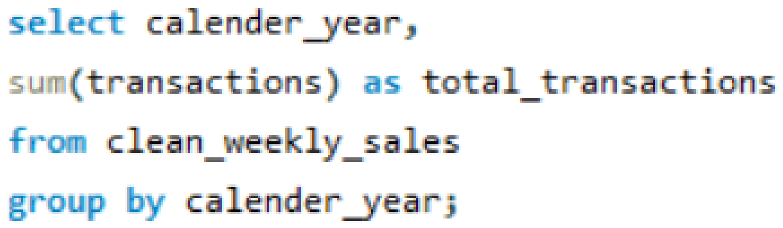
```
create table seq_100
(x int not null auto_increment primary key);
insert into seq_100 values (),(),(),(),(),(),(),(),(),();
insert into seq_100 values (),(),(),(),(),(),(),(),(),();
insert into seq_100 values (),(),(),(),(),(),(),(),(),();
insert into seq_100 values (),(),(),(),(),(),(),(),(),();
insert into seq_100 values (),(),(),(),(),(),(),(),(),();
insert into seq_100 select x + 50 from seq_100;
select * from seq_100;
create table seq_52 as (select x from seq_100 limit 52);
select distinct x as week_day from seq_52 where x not in(select distinct week_number from clean_weekly_sales);

select distinct week_number from clean_weekly_sales;
```

Output :



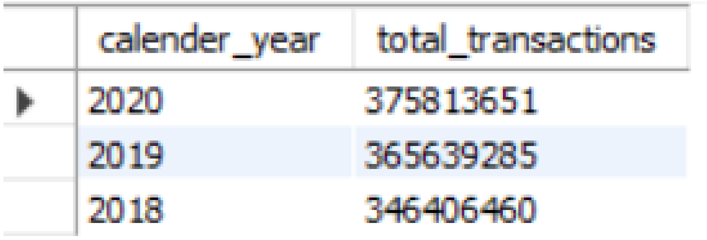
```
week_number
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
```

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



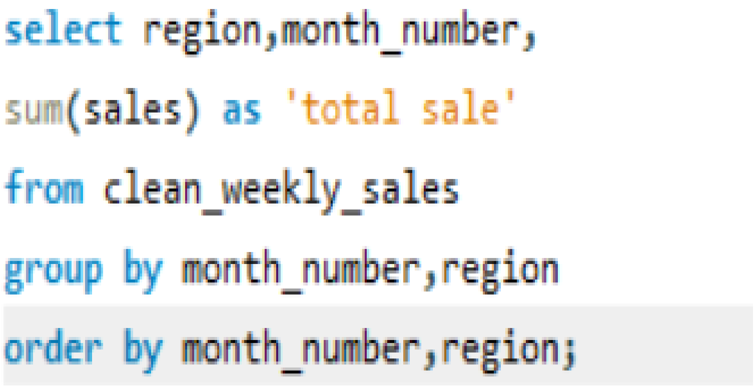
```
select calender_year,
sum(transactions) as total_transactions
from clean_weekly_sales
group by calender_year;
```

Output:



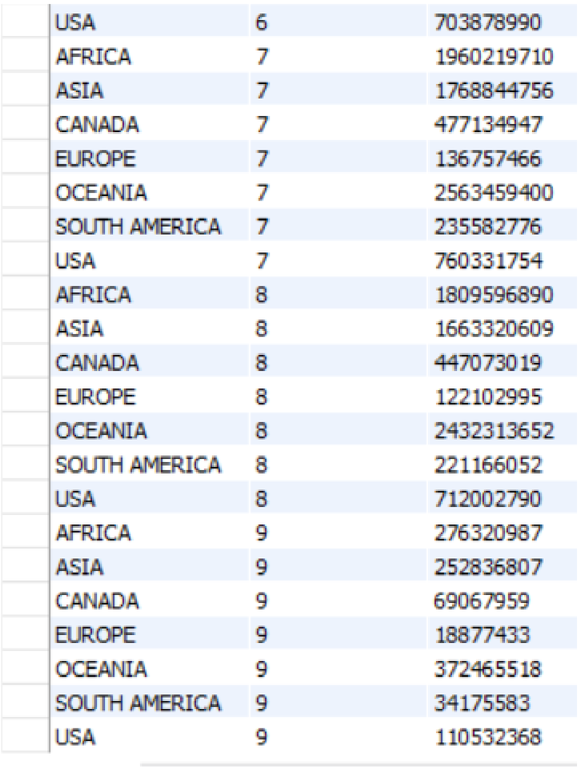
	calender_year	total_transactions
▶	2020	375813651
	2019	365639285
	2018	346406460

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



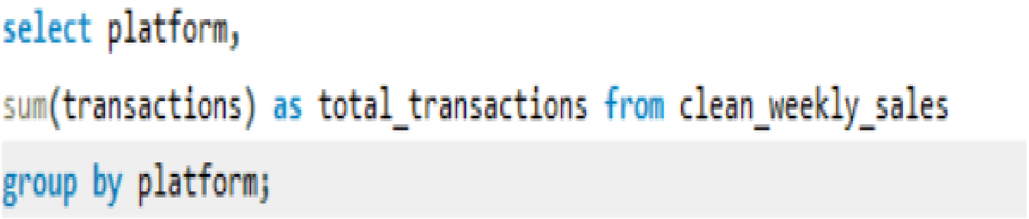
```
select region,month_number,
sum(sales) as 'total sale'
from clean_weekly_sales
group by month_number,region
order by month_number,region;
```

Output :



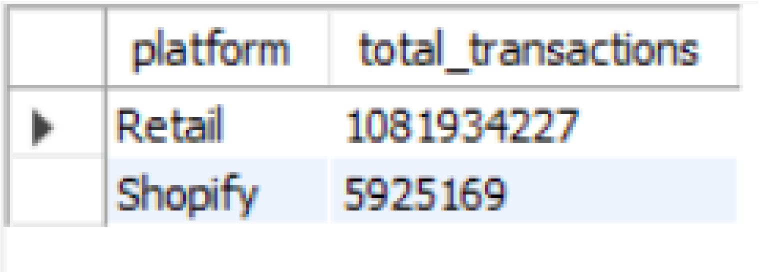
USA	6	703878990
AFRICA	7	1960219710
ASIA	7	1768844756
CANADA	7	477134947
EUROPE	7	136757466
OCEANIA	7	2563459400
SOUTH AMERICA	7	235582776
USA	7	760331754
AFRICA	8	1809596890
ASIA	8	1663320609
CANADA	8	447073019
EUROPE	8	122102995
OCEANIA	8	2432313652
SOUTH AMERICA	8	221166052
USA	8	712002790
AFRICA	9	276320987
ASIA	9	252836807
CANADA	9	69067959
EUROPE	9	18877433
OCEANIA	9	372465518
SOUTH AMERICA	9	34175583
USA	9	110532368

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



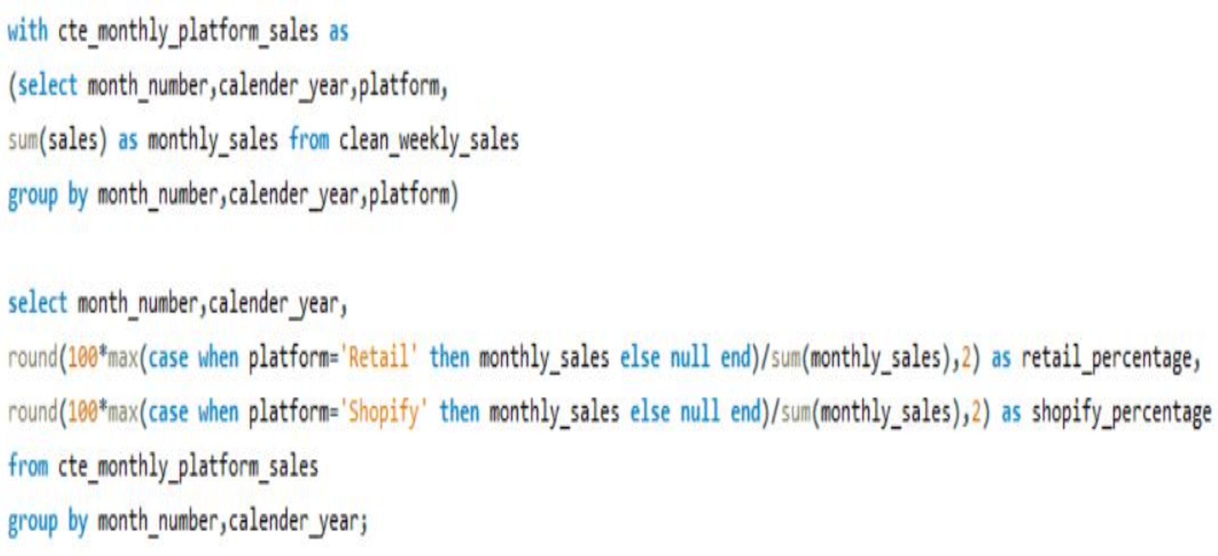
```
select platform,
sum(transactions) as total_transactions from clean_weekly_sales
group by platform;
```

Output :



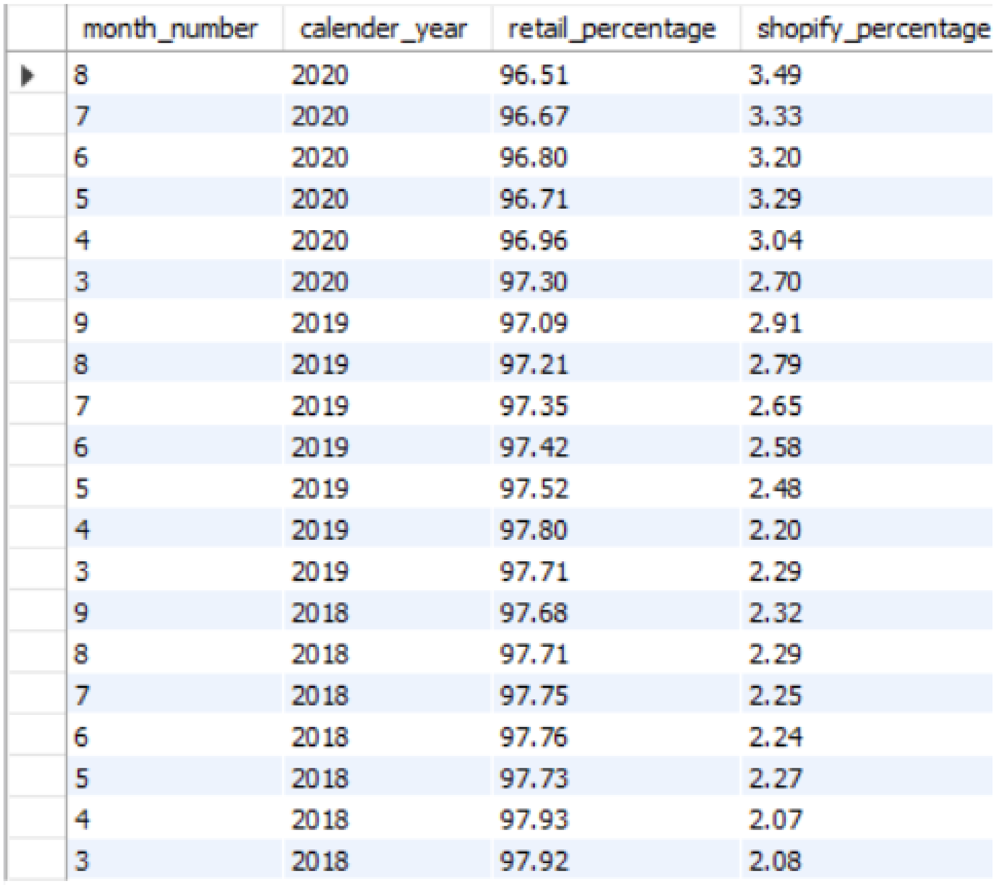
	platform	total_transactions
▶	Retail	1081934227
	Shopify	5925169

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



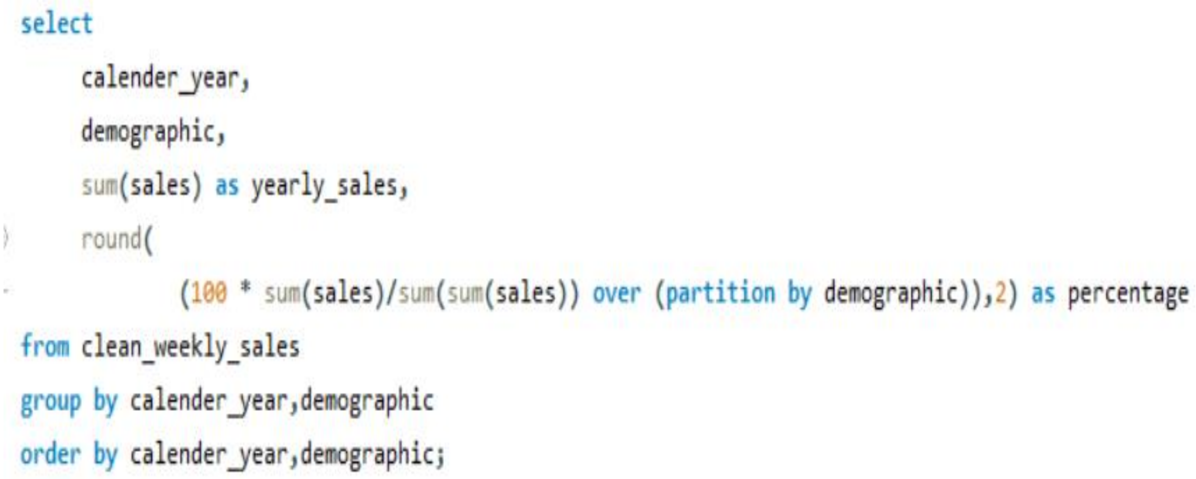
```
with cte_monthly_platform_sales as
(select month_number,calender_year,platform,
sum(sales) as monthly_sales from clean_weekly_sales
group by month_number,calender_year,platform)

select month_number,calender_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,calender_year;
```

Output :



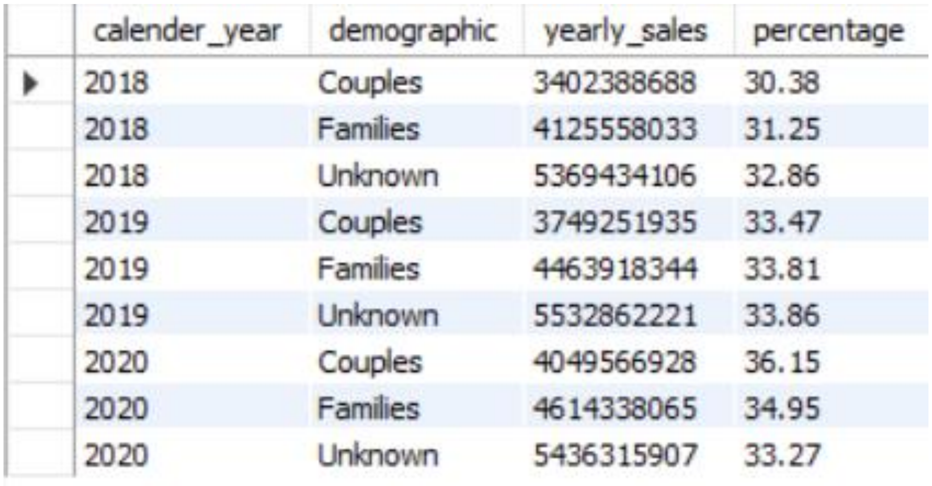
month_number	calender_year	retail_percentage	shopify_percentage
8	2020	96.51	3.49
7	2020	96.67	3.33
6	2020	96.80	3.20
5	2020	96.71	3.29
4	2020	96.96	3.04
3	2020	97.30	2.70
9	2019	97.09	2.91
8	2019	97.21	2.79
7	2019	97.35	2.65
6	2019	97.42	2.58
5	2019	97.52	2.48
4	2019	97.80	2.20
3	2019	97.71	2.29
9	2018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

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



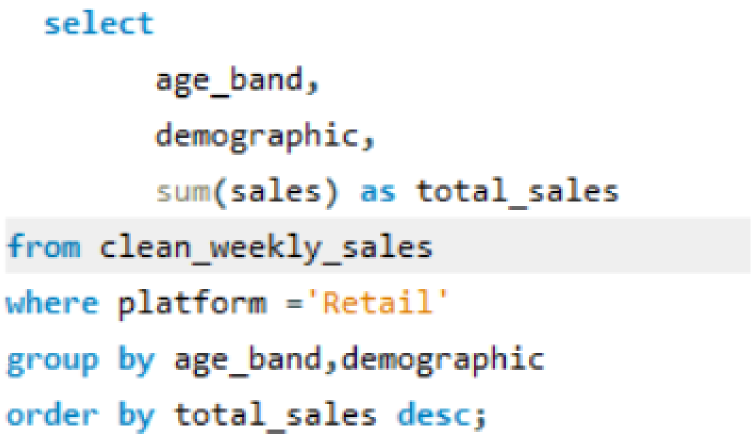
```
select
    calender_year,
    demographic,
    sum(sales) as yearly_sales,
    round(
            (100 * sum(sales)/sum(sum(sales)) over (partition by demographic)),2) as percentage
from clean_weekly_sales
group by calender_year,demographic
order by calender_year,demographic;
```

Output:



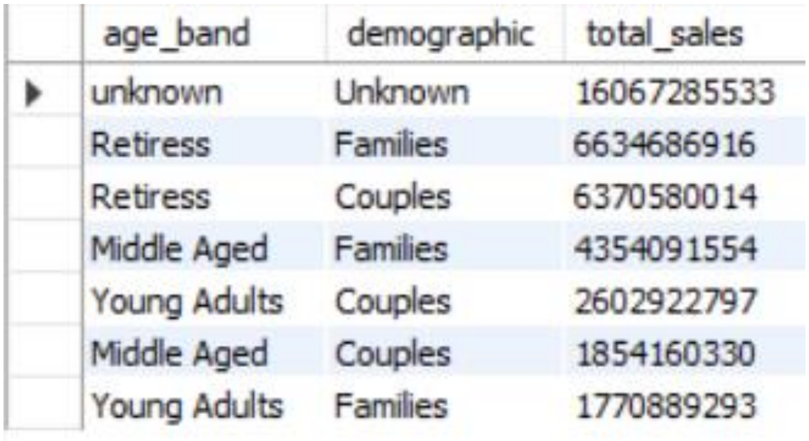
calender_year	demographic	yearly_sales	percentage
2018	Couples	3402388688	30.38
2018	Families	4125558033	31.25
2018	Unknown	5369434106	32.86
2019	Couples	3749251935	33.47
2019	Families	4463918344	33.81
2019	Unknown	5532862221	33.86
2020	Couples	4049566928	36.15
2020	Families	4614338065	34.95
2020	Unknown	5436315907	33.27

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



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

Output:



	age_band	demographic	total_sales
▶	unknown	Unknown	16067285533
	Retiress	Families	6634686916
	Retiress	Couples	6370580014
	Middle Aged	Families	4354091554
	Young Adults	Couples	2602922797
	Middle Aged	Couples	1854160330
	Young Adults	Families	1770889293

Conclusion :

1) Year 2020 occupied more transactions than previous 2 years.So sustainable packaging method made positive impact on the increasing sales performance of the Data Mart.
2) Customers is more favour to buy products from retail platform than shopify.
3) ‘Unknown’ age_band and ‘Unknown’ demographic segment contribute the most to Retail sales .Where ‘Unknown’ is a missing record ,so we take care for recording customers information.
4) Out of total sales for 3 years ,Retail platform made 97% of sales and Shopify platform made only 3% of sales.So we have to find reason for why customers are not buying products from online platform and take required actions to increasing sales from online platform.