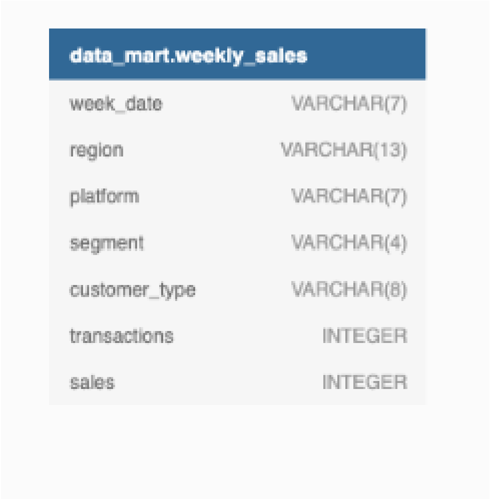
Case Study #1 - Data Mart

This case study actually is based on a real life change in Australian retailers where plastic bags were no longer provided for free - as you can expect, some customers would have changed their shopping behavior because of this change ! Shareholders need help to quantify the impact of this change on the sales performance for Data Mart and its separate business areas.

The key business question he wants you to know are the following:

- What was the quantifiable impact of the changes introduced in June 2020?
- Which platform, region, segment and customer types were the most impacted by this change?
- What can we do about future introduction of similar sustainability updates to the business to minimize impact on sales?

For this case study there is only a single table: data_mart.weekly_sales



data_mart.weekly_sales	
week_date	VARCHAR(7)
region	VARCHAR(13)
platform	VARCHAR(7)
segment	VARCHAR(4)
customer_type	VARCHAR(8)
transactions	INTEGER
sales	INTEGER

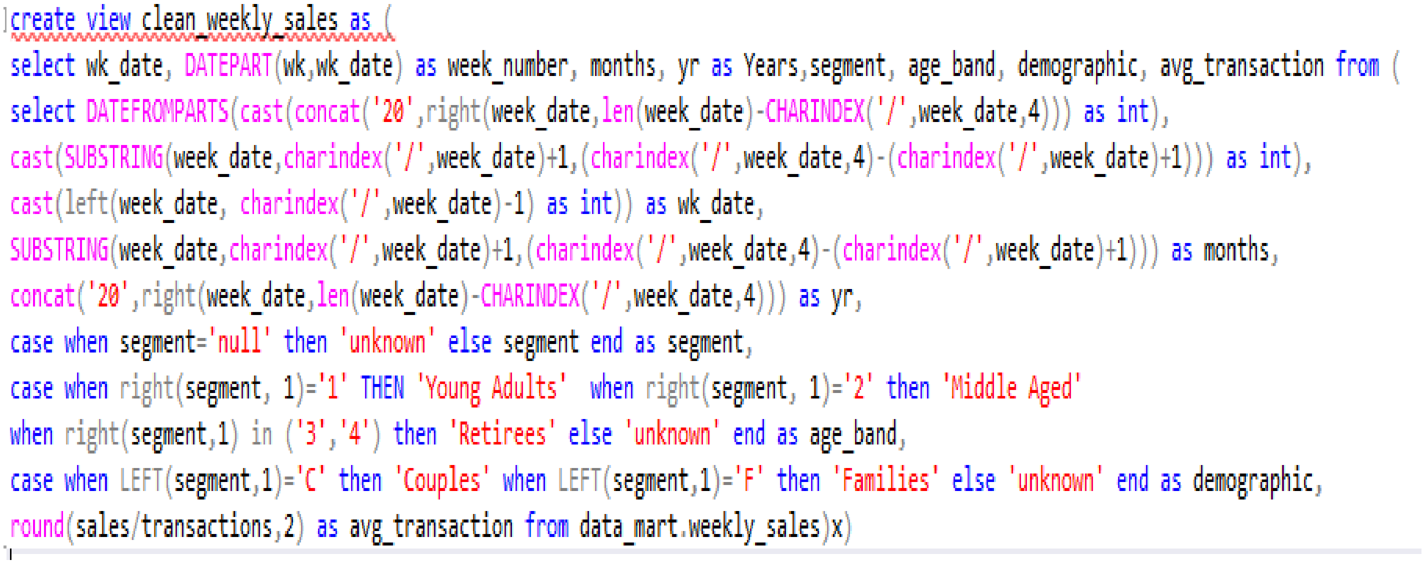
Case Study Questions

A. Data CLEANING Steps

In a single query, perform the following operations and generate a new table in the data_mart schema named clean_weekly_sales:

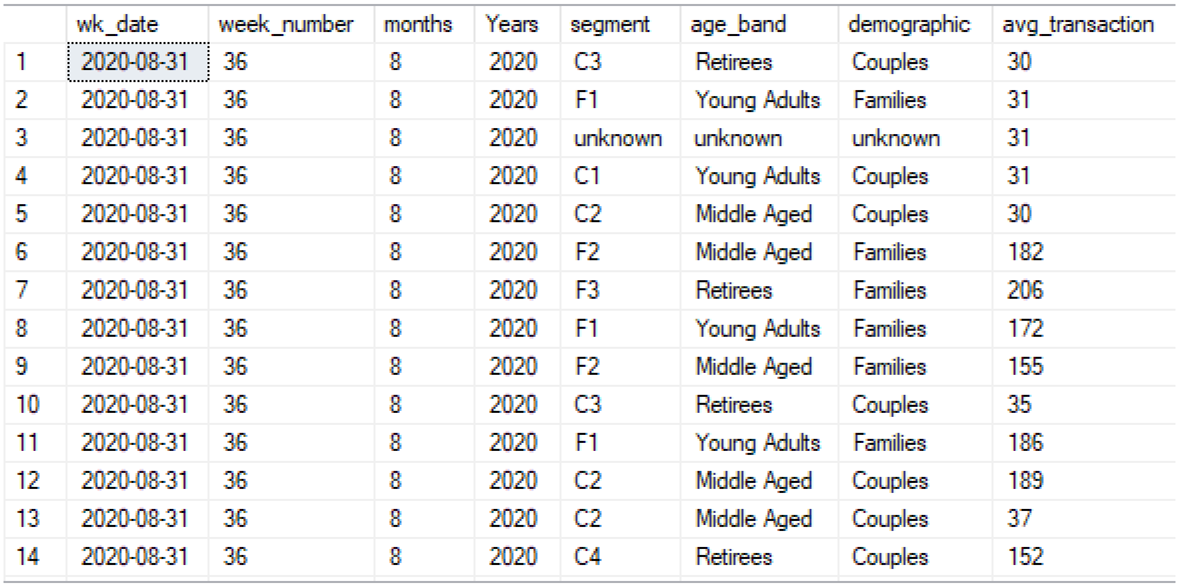
- Convert the week_date to a DATE format
- Add a week_number as the second column for each week_date value, for example any value from the 1st of January to 7th of January will be 1, 8th to 14th will be 2 etc
- Add a month_number with the calendar month for each week_date value as the 3rd column
- Add a calendar_year column as the 4th column containing either 2018, 2019 or 2020 values
- Add a new column called age_band after the original segment column using the following mapping on the number inside the segment value: 1 - Young Adults; 2 - Middle Aged; 3 or 4 - Retirees
- Add a new demographic column using the following mapping for the first letter in the segment values: C - Couples ; F- Families

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



```
create view clean_weekly_sales as (
select wk_date, DATEPART(wk,wk_date) as week_number, months, yr as Years,segment, age_band, demographic, avg_transaction from (
select DATEFROMPARTS(cast(concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4))) as int),
cast(SUBSTRING(week_date,charindex('/',week_date)+1,(charindex('/',week_date,4)-(charindex('/',week_date)+1))) as int),
cast(left(week_date, charindex('/',week_date)-1) as int)) as wk_date,
SUBSTRING(week_date,charindex('/',week_date)+1,(charindex('/',week_date,4)-(charindex('/',week_date)+1))) as months,
concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4))) as yr,
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,
round(sales/transactions,2) as avg_transaction from data_mart.weekly_sales)x
```

select * from clean_weekly_sales;

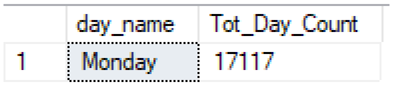


	wk_date	week_number	months	Years	segment	age_band	demographic	avg_transaction
1	2020-08-31	36	8	2020	C3	Retirees	Couples	30
2	2020-08-31	36	8	2020	F1	Young Adults	Families	31
3	2020-08-31	36	8	2020	unknown	unknown	unknown	31
4	2020-08-31	36	8	2020	C1	Young Adults	Couples	31
5	2020-08-31	36	8	2020	C2	Middle Aged	Couples	30
6	2020-08-31	36	8	2020	F2	Middle Aged	Families	182
7	2020-08-31	36	8	2020	F3	Retirees	Families	206
8	2020-08-31	36	8	2020	F1	Young Adults	Families	172
9	2020-08-31	36	8	2020	F2	Middle Aged	Families	155
10	2020-08-31	36	8	2020	C3	Retirees	Couples	35
11	2020-08-31	36	8	2020	F1	Young Adults	Families	186
12	2020-08-31	36	8	2020	C2	Middle Aged	Couples	189
13	2020-08-31	36	8	2020	C2	Middle Aged	Couples	37
14	2020-08-31	36	8	2020	C4	Retirees	Couples	152

B. Data Exploration

1. What day of the week is used for each week_date value?

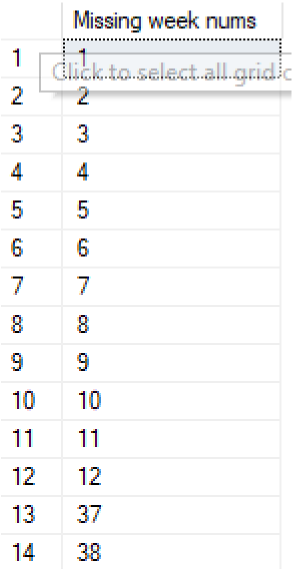
select day_name, count(*) Tot_Day_Count from (
select wk_date, DATENAME(DW,wk_date) as day_name from clean_weekly_sales)p
group by day_name



	day_name	Tot_Day_Count
1	Monday	17117

2. What range of week numbers are missing from the dataset?

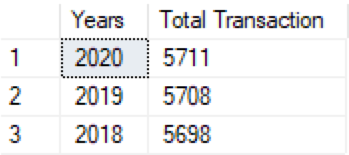
```
with wk_num as (
select 1 as n
union all
select n+1 from wk_num where n<53)

select n as [Missing week nums] from wk_num where n not in (select week_number from
clean_weekly_sales)
```



	Missing week nums
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	37
14	38

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

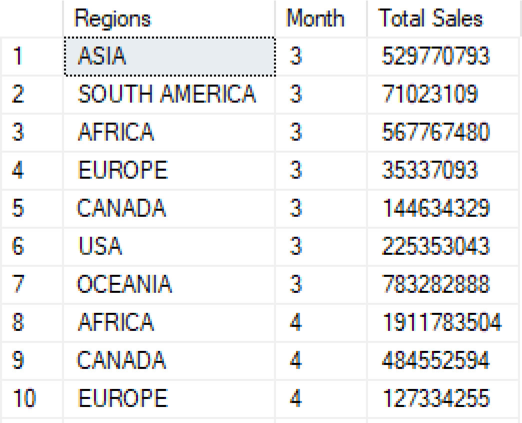
```
select Years, count(avg_transaction) as [Total Transaction] from clean_weekly_sales group by Years
```



	Years	Total Transaction
1	2020	5711
2	2019	5708
3	2018	5698

4. What is the total sales for each region for each month?

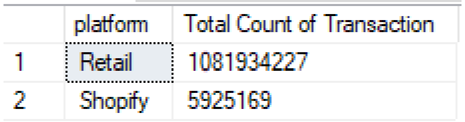
```
select region [Regions], mnth [Month], sum(cast(sales as bigint)) [Total Sales] from (
select
cast(SUBSTRING(week_date,charindex('/',week_date)+1,(charindex('/',week_date,4)-(charindex('/',week_date)+1))) as int) as Mnth,region,sales
from data_mart.weekly_sales)x group by region, Mnth
order by Mnth
```



	Regions	Month	Total Sales
1	ASIA	3	529770793
2	SOUTH AMERICA	3	71023109
3	AFRICA	3	567767480
4	EUROPE	3	35337093
5	CANADA	3	144634329
6	USA	3	225353043
7	OCEANIA	3	783282888
8	AFRICA	4	1911783504
9	CANADA	4	484552594
10	EUROPE	4	127334255

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

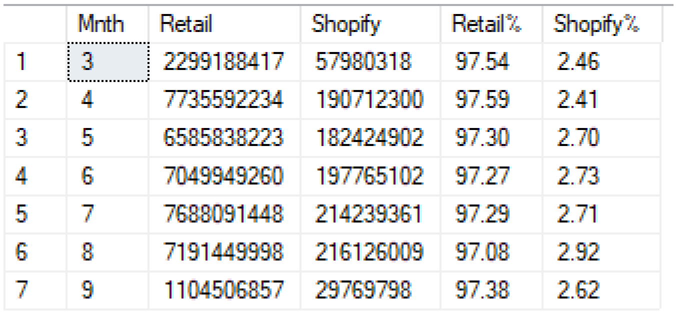
select platform, sum(transactions) as [Total Count of Transaction] from data_mart.weekly_sales group by platform



	platform	Total Count of Transaction
1	Retail	1081934227
2	Shopify	5925169

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

select *, cast(Retail*100.0/(Retail+Shopify) as dec(5,2)) as 'Retail%',
cast(Shopify*100.0/(Retail+Shopify) as dec(5,2)) as 'Shopify%' from (
select * from (
select
cast(SUBSTRING(week_date,charindex('/',week_date)+1,(charindex('/',week_date,4)-(charindex('/',week_date)+1))) as int) as Mnth, platform, cast(sales as bigint) as sales
from data_mart.weekly_sales)x
pivot
(sum(x.sales) for platform in ([Retail],[Shopify])) pivot_data)p order by Mnth

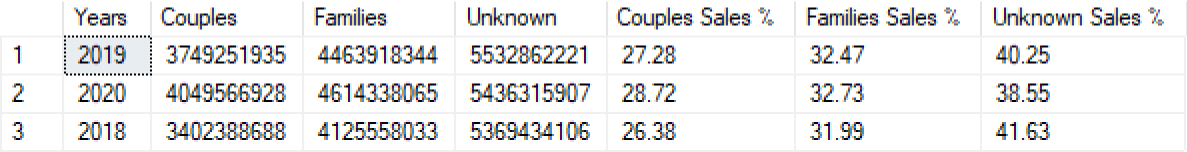


	Mnth	Retail	Shopify	Retail%	Shopify%
1	3	2299188417	57980318	97.54	2.46
2	4	7735592234	190712300	97.59	2.41
3	5	6585838223	182424902	97.30	2.70
4	6	7049949260	197765102	97.27	2.73
5	7	7688091448	214239361	97.29	2.71
6	8	7191449998	216126009	97.08	2.92
7	9	1104506857	29769798	97.38	2.62

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

select *, cast(couples*100.0/(couples+families+unknown) as dec(5,2)) as [Couples Sales %],
cast(families*100.0/(couples+families+unknown) as dec(5,2)) as [Families Sales %],
cast(unknown*100.0/(couples+families+unknown) as dec(5,2)) as [Unknown Sales %] from (

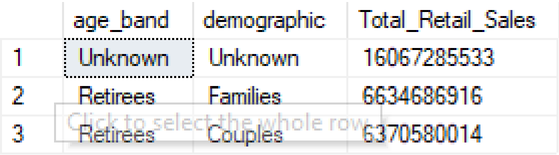
```
select * from (
select convert(int,concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4)))) as
[Years],
case when left(segment, 1)='C' then 'Couples' when left(segment,1)='F' then 'Families' else 'Unknown'
end as demographic, convert(bigint,sales) sales
from data_mart.weekly_sales)a
pivot
(sum(sales) for demographic in ([Couples],[Families],[Unknown]))pivot_data)p
```



	Years	Couples	Families	Unknown	Couples Sales %	Families Sales %	Unknown Sales %
1	2019	3749251935	4463918344	5532862221	27.28	32.47	40.25
2	2020	4049566928	4614338065	5436315907	28.72	32.73	38.55
3	2018	3402388688	4125558033	5369434106	26.38	31.99	41.63

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

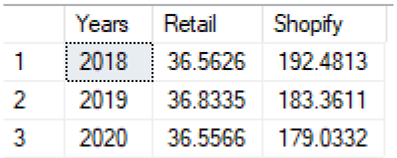
```
select top 3 age_band, demographic, sum(sales) as Total_Retail_Sales from (
select case when right(segment,1)='1' then 'Young Adults' when right(segment,1)='2' then 'Middle
Aged'
when right(segment,1) in ('3','4') then 'Retirees' else 'Unknown' end as age_band,
case when left(segment,1)='C' then 'Couples' when left(segment,1)='F' then 'Families' else 'Unknown'
end as demographic, platform ,
convert(bigint,sales) as sales from data_mart.weekly_sales where platform='Retail')x group by
age_band, demographic
order by Total_Retail_Sales desc
```


Click to select the whole row

	age_band	demographic	Total_Retail_Sales
1	Unknown	Unknown	16067285533
2	Retirees	Families	6634686916
3	Retirees	Couples	6370580014

9. Can we use the avg_transaction column to find the average transaction size for each year for Retail vs Shopify? If not - how would you calculate it instead?
-- Average Transaction Size = Total Sales by Period/ Total No. of Transactions by same period

```
select Years, max(case when platform='Retail' then avg_trans_size end) as [Retail],
max(case when platform='Shopify' then avg_trans_size end)as [Shopify] from
(select Years, platform, convert(dec(14,4),tot_sales*1.0/tot_trans) as avg_trans_size from (
select Years, platform, sum(transactions) as tot_trans, sum(cast(sales as bigint)) as tot_sales from (
select convert(int,concat('20',right(week_date,len(week_date)-charindex('/',week_date,4)))) as
[Years],platform, transactions, sales
from data_mart.weekly_sales)x group by Years, platform)y)z group by Years
```



	Years	Retail	Shopify
1	2018	36.5626	192.4813
2	2019	36.8335	183.3611
3	2020	36.5566	179.0332

C. Before and After Analysis

This technique is usually used when we inspect an important event and want to inspect the impact before and after a certain point in time.

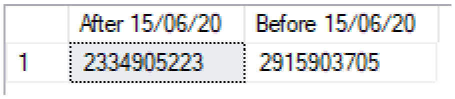
Taking the week_date value of 2020-06-15 as the baseline week where the Data Mart sustainable packaging changes came into effect.

We would include all week_date values for 2020-06-15 as the start of the period **after** the change and the previous week_date values would be **before**

Using this analysis approach - answer the following questions:

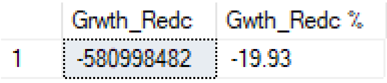
1. What is the total sales for the 4 weeks before and after 2020-06-15? What is the growth or reduction rate in actual values and percentage of sales?

```
with aft_bef as (
select * from (
select case when wk_date between DATEADD(week,-4,'2020-06-15') and '2020-06-15' then 'Before
15/06/20' else 'After 15/06/20' end as Changes, Sales from (
select * from (
select
DATEFROMPARTS(cast(concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4)))
as int),
cast(SUBSTRING(week_date,charindex('/',week_date)+1,(charindex('/',week_date,4)-(charindex('/',we
ek_date)+1))) as int),
cast(left(week_date, charindex('/',week_date)-1) as int)) as wk_date, cast(sales as bigint) as Sales
from data_mart.weekly_sales)a where wk_date between DATEADD(week,-4,'2020-06-15') and
DATEADD(week,4,'2020-06-15'))b)c
pivot
(sum(sales) for changes in ([After 15/06/20],[Before 15/06/20]))pivot_data)

select * from aft_bef;
```



	After 15/06/20	Before 15/06/20
1	2334905223	2915903705

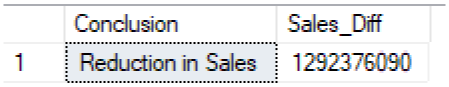
```
select [after 15/06/20]-[before 15/06/20] as Grwth_Redc,
cast(([after 15/06/20]-[before 15/06/20])*100.0/[before 15/06/20] as dec(5,2)) as [Gwth_Redc %] from
aft_bef;
```



	Grwth_Redc	Gwth_Redc %
1	-580998482	-19.93

2. What about the entire 12 weeks before and after?

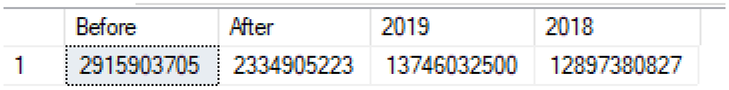
```
select case when result>0 then 'Growth in Sales' else 'Reduction in Sales' end as Conclusion,
abs(result) as Sales_Diff from (
select [after 12 wk]-[before 12 wk] as result from (
select * from (
select case when wk_date between dateadd(week, -12,'2020-06-15') and '2020-06-15' then 'Before 12
Wk' else 'After 12 Wk' end as After_Before, Sales from(
select * from(
select
DATEFROMPARTS(cast(concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4)))
as int),
cast(SUBSTRING(week_date,charindex('/',week_date)+1,(charindex('/',week_date,4)-(charindex('/',we
ek_date)+1))) as int),
cast(left(week_date, charindex('/',week_date)-1) as int)) as wk_date, cast(sales as bigint) as Sales
from data_mart.weekly_sales)a where wk_date between DATEADD(week,-12,'2020-06-15') and
DATEADD(week,12,'2020-06-15'))b)c
pivot
(sum(sales) for After_Before in ([After 12 Wk],[Before 12 Wk]))pivot_data)p)q
```



	Conclusion	Sales_Diff
1	Reduction in Sales	1292376090

3. How do the sales metrics for these 2 periods before and after compare with the previous years in 2018 and 2019?

```
select * from (
select case when wk_date between DATEADD(week,-4,'2020-06-15') and '2020-06-15'  then 'Before'
when  wk_date between '2020-06-15' and DATEADD(week,4,'2020-06-15') then 'After'
when year(wk_date)=2019 then '2019' else '2018' end as Comparison, sales from (
select wk_date, sales from (
select
DATEFROMPARTS(cast(concat('20',RIGHT(week_date,len(week_date)-CHARINDEX('/',week_date,4)
)) as int),
SUBSTRING(week_date,CHARINDEX('/',week_date)+1,charindex('/',week_date,4)-(charindex('/',week
_date)+1)),
cast(LEFT(week_date,CHARINDEX('/',week_date)-1) as int)) as
wk_date,(cast(concat('20',RIGHT(week_date,len(week_date)-CHARINDEX('/',week_date,4))) as int))
as [Years],
cast(sales as bigint) as Sales from data_mart.weekly_sales)p
where years in (2019,2018) or wk_date between DATEADD(week,-4,'2020-06-15') and
DATEADD(week,4,'2020-06-15'))q)r
pivot
(sum(Sales) for Comparison in ([Before],[After],[2019],[2018]))pivot_data
```



	Before	After	2019	2018
1	2915903705	2334905223	13746032500	12897380827

D. BONUS Question

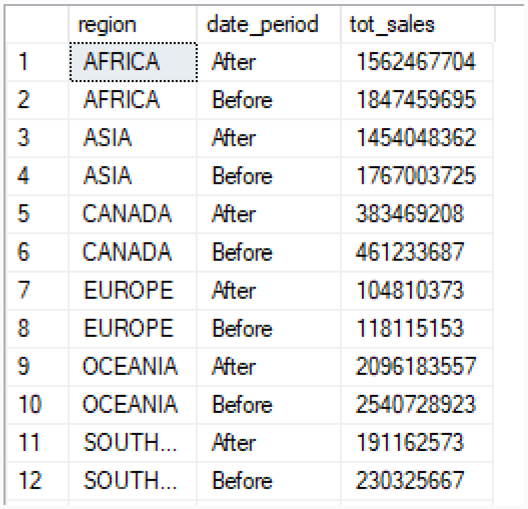
Which areas of the business have the highest negative impact in sales metrics performance in 2020 for the 12 weeks before and after period?

- region
- platform
- age_band
- demographic
- customer_type

```
with filt_tble as (
select wk_date,region, platform,
case when right(segment,1)='1' then 'Young Adult' when right(segment,1)='2' then 'Mid 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 demographics,
customer_type,sales, case when wk_date between DATEADD(week,-12,wk_date) and '2020-06-15'
then 'Before' else 'After' end as date_period
from (
select
DATEFROMPARTS(convert(int,concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4)))),
convert(int,SUBSTRING(week_date, CHARINDEX('/',week_date)+1,
CHARINDEX('/',week_date,4)-(charindex('/',week_date)+1))),
convert(int,left(week_date,charindex('/',week_date)-1))) as wk_date,
convert(int,concat('20',right(week_date,len(week_date)-CHARINDEX('/',week_date,4)))) years,
region, platform, customer_type,segment, convert(bigint,sales) sales
from data_mart.weekly_sales)p where years=2020 and wk_date between
DATEADD(week,-12,wk_date) and DATEADD(week,12,wk_date))
```

REGION:

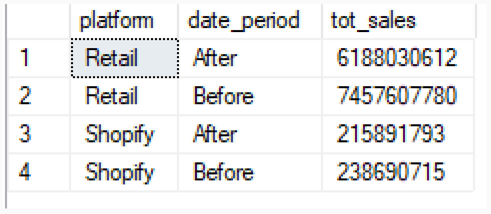
```
select region, date_period, sum(sales) tot_sales from filt_tble group by region, date_period
order by region;
```



	region	date_period	tot_sales
1	AFRICA	After	1562467704
2	AFRICA	Before	1847459695
3	ASIA	After	1454048362
4	ASIA	Before	1767003725
5	CANADA	After	383469208
6	CANADA	Before	461233687
7	EUROPE	After	104810373
8	EUROPE	Before	118115153
9	OCEANIA	After	2096183557
10	OCEANIA	Before	2540728923
11	SOUTH...	After	191162573
12	SOUTH...	Before	230325667

PLATFORM

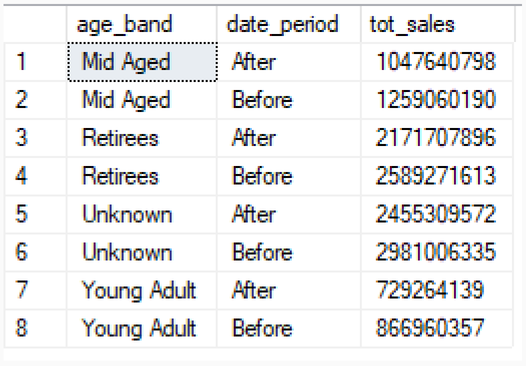
```
select platform, date_period, sum(sales) tot_sales from filt_tble group by platform,
date_period order by platform;
```



	platform	date_period	tot_sales
1	Retail	After	6188030612
2	Retail	Before	7457607780
3	Shopify	After	215891793
4	Shopify	Before	238690715

AGE BAND

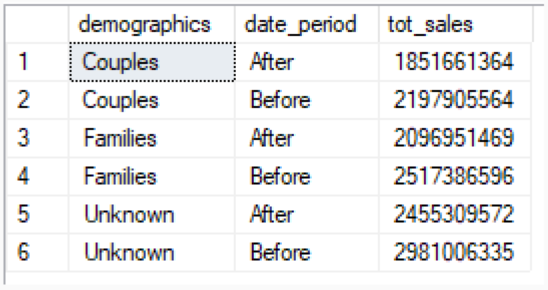
```
select age_band, date_period, sum(sales) as tot_sales from filt_tble group by age_band,
date_period order by age_band;
```



	age_band	date_period	tot_sales
1	Mid Aged	After	1047640798
2	Mid Aged	Before	1259060190
3	Retirees	After	2171707896
4	Retirees	Before	2589271613
5	Unknown	After	2455309572
6	Unknown	Before	2981006335
7	Young Adult	After	729264139
8	Young Adult	Before	866960357

DEMOGRAPHICS

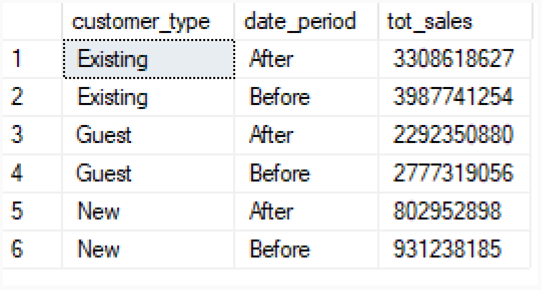
```
select demographics, date_period, sum(sales) tot_sales from filt_tble group by demographics,
date_period order by demographics;
```



	demographics	date_period	tot_sales
1	Couples	After	1851661364
2	Couples	Before	2197905564
3	Families	After	2096951469
4	Families	Before	2517386596
5	Unknown	After	2455309572
6	Unknown	Before	2981006335

CUSTOMER TYPE

```
select customer_type, date_period, sum(sales) tot_sales from filt_tble group by
customer_type, date_period order by customer_type;
```



	customer_type	date_period	tot_sales
1	Existing	After	3308618627
2	Existing	Before	3987741254
3	Guest	After	2292350880
4	Guest	Before	2777319056
5	New	After	802952898
6	New	Before	931238185