Case Study Questions

The following case study questions require some data cleaning steps before we start to unpack Danny's key business questions in more depth.

1. 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:

- 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

segment	age_band
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:

segment	demographic
C	Couples
F	Families

- 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
- Generate a new avg_transaction column as the sales value divided by transactions rounded to 2 decimal places for each record

```sql
Select
str_to_date(week_date, '%d/%m/%Y') AS week_date,
week(str_to_date(week_date, '%d/%m/%Y')) as week_number,
month(str_to_date(week_date, '%d/%m/%Y')) as month_number,
year(str_to_date(week_date, '%d/%m/%Y')) as calendar_year,
region,
platform,
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,
round(sales/transactions,2) as avg_transaction,
sales
From weekly_sales;
```

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

2. Data Exploration

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

```sql
Select dayname(week_date) as day_of_the_week
From clean_weekly_sales
Group by day_of_the_week;
```

day_of_the_week
Monday

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

```
CREATE TABLE numbers (n INT);
INSERT INTO numbers VALUES (1),(2),(3),(4),(5), (6), (7), (8),(9),
(10), (11),(12),(13),(14),(15), (16), (17), (18),(19),
(20), (21),(22),(23),(24),(25), (26), (27), (28),(29),
(30), (31),(32),(33),(34),(35), (36), (37), (38),(39),
(40), (41),(42),(43),(44),(45), (46), (47), (48),(49),
(50), (51), (52);

Select n.n as missing_week_number
From numbers as n
LEFT OUTER JOIN clean_weekly_sales as s
ON n.n = s.week_number
WHERE s.week_number IS NULL;
```

missing_week_number
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

28 row(s) returned

3. *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
Order by calendar_year;
```

calendar_year	total_transactions
2018	346406460
2019	365639285
2020	375813651

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

```
Select region, month_number, sum(sales) as total_sales
From clean_weekly_sales
Group by region, month_number
Order by region, month_number;
```

region	month_number	total_sales
AFRICA	3	567767480
AFRICA	4	1911783504
AFRICA	5	1647244738
AFRICA	6	1767559760
AFRICA	7	1960219710
AFRICA	8	1809596890
AFRICA	9	276320987
ASIA	3	529770793

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

```
Select platform, count(transactions) as total_count_of_transactions
From clean_weekly_sales
Group by platform
Order by total_count_of_transactions;
```

platform	total_count_of_transactions
Shopify	8549
Retail	8568

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

```
Create View Sales_platform As
Select calendar_year, month_number, platform, sum(sales) as total_sales
From clean_weekly_sales
Group by calendar_year, month_number, platform;
```

calendar_year	month_number	platform	total_sales
2020	8	Retail	2810210216
2020	8	Shopify	101583216
2020	7	Shopify	77642565
2020	7	Retail	2255852981
2020	6	Shopify	92714414
2020	6	Retail	2807693824
2020	5	Retail	2284387029

```
Select calendar_year, month_number,
Round(100*Max(Case When platform = 'Retail' Then total_sales Else Null End)/sum(total_sales),2) as retail_percentage,
Round(100*Max(Case When platform = 'Shopify' Then total_sales Else Null End)/sum(total_sales),2) as shopify_percentage
From Sales_platform
Group by calendar_year, month_number
Order by calendar_year, month_number;
```

calendar_year	month_number	retail_percentage	shopify_percentage
2018	3	97.92	2.08
2018	4	97.93	2.07
2018	5	97.73	2.27
2018	6	97.76	2.24
2018	7	97.75	2.25
2018	8	97.71	2.29
2018	9	97.68	2.32

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

```
Create View Sales_demographic As
Select calendar_year, demographic, sum(sales) as total_sales
From clean_weekly_sales
Group by calendar_year, demographic
Order by calendar_year;
```

calendar_year	demographic	total_sales
2018	Couples	3402388688
2018	Families	4125558033
2018	unknown	5369434106
2019	Couples	3749251935
2019	Families	4463918344
2019	unknown	5532862221
2020	Couples	4049566928

```
Select calendar_year,
Round(100*Max(Case When demographic = 'Couples' Then total_sales Else Null End)/sum(total_sales),2) as couples_percentage,
Round(100*Max(Case When demographic = 'Families' Then total_sales Else Null End)/sum(total_sales),2) as families_percentage,
Round(100*Max(Case When demographic = 'unknown' Then total_sales Else Null End)/sum(total_sales),2) as unknow_percentage
From Sales_demographic
Group by calendar_year
Order by calendar_year;
```

calendar_year	couples_percentage	families_percentage	unknow_percentage
2018	26.38	31.99	41.63
2019	27.28	32.47	40.25
2020	28.72	32.73	38.55

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

```
Select age_band, demographic, platform, sum(sales) as retail_sales
From clean_weekly_sales
Where platform = 'Retail'
Group by age_band, demographic
Order by retail_sales DESC;
```

age_band	demographic	platform	retail_sales
unknown	unknown	Retail	16067285533
Retirees	Families	Retail	6634686916
Retirees	Couples	Retail	6370580014
Middle Aged	Families	Retail	4354091554
Young Adults	Couples	Retail	2602922797
Middle Aged	Couples	Retail	1854160330
Young Adults	Families	Retail	1770889293

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?

```
Select calendar_year, platform, round(avg(avg_transaction),2) as average_transaction_by_row,
round(sum(sales)/sum(transactions),2) as average_transaction_by_group
From clean_weekly_sales
Group by calendar_year, platform
Order by calendar_year, platform;
```

calendar_year	platform	average_transaction_by_row	average_transaction_by_group
2018	Retail	42.91	36.56
2018	Shopify	188.28	192.48
2019	Retail	41.97	36.83
2019	Shopify	177.56	183.36
2020	Retail	40.64	36.56
2020	Shopify	174.87	179.03

3. Before & 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?*

Before we start, we find out the **week_number** of '2020-06-15':

week_number
24

Then, we need a table with data 4 weeks before and after '2020-06-15':

```
Create View week_number_20_27 As
Select week_date, week_number, sum(sales) as total_sales
From clean_weekly_sales
Where (week_number between 20 and 27) and calendar_year = 2020
Group by week_date, week_number;
```

week_date	week_number	total_sales
2020-07-06	27	590335394
2020-06-29	26	575390599
2020-06-22	25	583242828
2020-06-15	24	570025348
2020-06-08	23	586283390
2020-06-01	22	585466073
2020-05-25	21	589120804

And changes during these periods

```
Create View changes As
Select sum(case when week_number between 20 and 23 Then total_sales End) as before_change,
sum(case when week_number between 24 and 27 Then total_sales End) as after_change
From week_number_20_27;
```

before_change	after_change
2345878357	2318994169

Solution:

```
Select before_change, after_change, (after_change - before_change) as growth_or_reduction,
round(100*(after_change-before_change)/before_change,2) as percentage
From changes;
```

before_change	after_change	growth_or_reduction	percentage
2345878357	2318994169	-26884188	-1.15

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

Similar code as above.

Solution:

before_change	after_change	growth_or_reduction	percentage
7126273147	6973947753	-152325394	-2.14

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

```sql
Create View week_number_12_35_years As
Select calendar_year, week_date, week_number, sum(sales) as total_sales
From clean_weekly_sales
Where (week_number between 12 and 35)
Group by calendar_year, week_date, week_number;

Create View changes_3 As
Select sum(case when week_number between 12 and 23 Then total_sales End) as before_change,
sum(case when week_number between 24 and 35 Then total_sales End) as after_change, calendar_year
From week_number_12_35_years
Group by calendar_year;


Select calendar_year, before_change, after_change, (after_change - before_change) as growth_or_reduction,
round(100*(after_change-before_change)/before_change,2) as percentage
From changes_3
order by calendar_year;
```

calendar_year	before_change	after_change	growth_or_reduction	percentage
2018	6396562317	6500818510	104256193	1.63
2019	6883386397	6862646103	-20740294	-0.30
2020	7126273147	6973947753	-152325394	-2.14