some items blanked out to protect identity

Background:
The Data Services team at _____ collects sales data from a wide variety of_____ companies. We also collect a broad range of descriptive and contextual data (master data, geographical data, population, etc.) that we integrate with sales and transactional data to create a robust reporting warehouse and suite of reporting solutions.

We leverage a T-SQL client to query our main data warehouse (currently Teradata), as well as a variety of reporting tools (Alteryx, Power Bi, etc.) to transform, analyze, support, and report on this data.

As a member of the _____ team, you will interact with the sales and master data stored in Teradata on a regular basis. As with any company, _____'s data is unique and nuanced, however, our approach to data architecture and management is consistent with other SQL-based environments.

Data Overview:
- Fact Data (invoice/sales data)
- Customer Master
- Product Master
- Zip-to-population data

Additional documents:
- _____ – Customer report 1 + SQL code
- _____ – Customer report 2 + SQL code

Questions

1. How many unique customers made purchases within the dataset? Which customer has the highest sales volume?

    a. There are **3** unique customers. The first query tells you how many and the second query gives you the name of the customers and how many times each customer appears.

```
select count(distinct(customer_no))
from fact_data;

Select customer_name, count(f.customer_no)
From fact_data f
join customer_master c on c.customer_no=f.customer_no
group by 1;
```

    b. _____ has the highest sales value. This query selects the name of the customer, and the total sum of the sales volume grouped by customer. Once grouped, it is sorted by descending order to get the highest total sales volume and then limited by 1 to have only the value wanted showing.

```
Select c.customer_name, sum(volume_cases) as total_volume_cases
From customer_master c
join fact_data f on f.customer_no=c.customer_no
Group by customer_name
Order by total_volume_cases desc
limit 1;
```

2. Each customer has several stores reporting within this dataset. For the customer_____, how would you identify which stores purchased_____but *did not* purchase_____?

    I selected the customer_name, store_number, brand_name and count of brand name just to see how many times that store number received that brand from the fact_data table. I then joined the customer_master table using customer_no just to get _____' to be in the output as well as joined the product_master table using product_no.
    From those joined tables, I was able to filter for only the customer_____ (customer code 853) and only the brands of interest.
    Using group by allowed me to see each distinct store that purchased either _____or_____and it was only 13 different stores. Since the amount was so small, I could visually see which stores purchased_____but not_____.

```
select c.customer_name,f.store_number, brand_name, count(brand_name)
from fact_data f
join customer_master c on c. customer_no=f.customer_no
join product_master p on p. product_no= f.product_no
where f.customer_no=853
and brand_name like '%[redacted]%'
or brand_name like '[redacted]%'
group by 2,3,1
order by 2;
```

From those results I created a CTE to eliminate all values containing _____ to come up with the final answer.

```
with [redacted] as (select c.customer_name,f.store_number, brand_name, count(brand_name)
from fact_data f
join customer_master c on c. customer_no=f.customer_no
join product_master p on p. product_no= f.product_no
where f.customer_no=853
and brand_name like '[redacted]%'
or brand_name like '[redacted]%'
group by 2,3,1
order by 2)
select *
from dasanibody
where brand_name not like '%Body%';
```

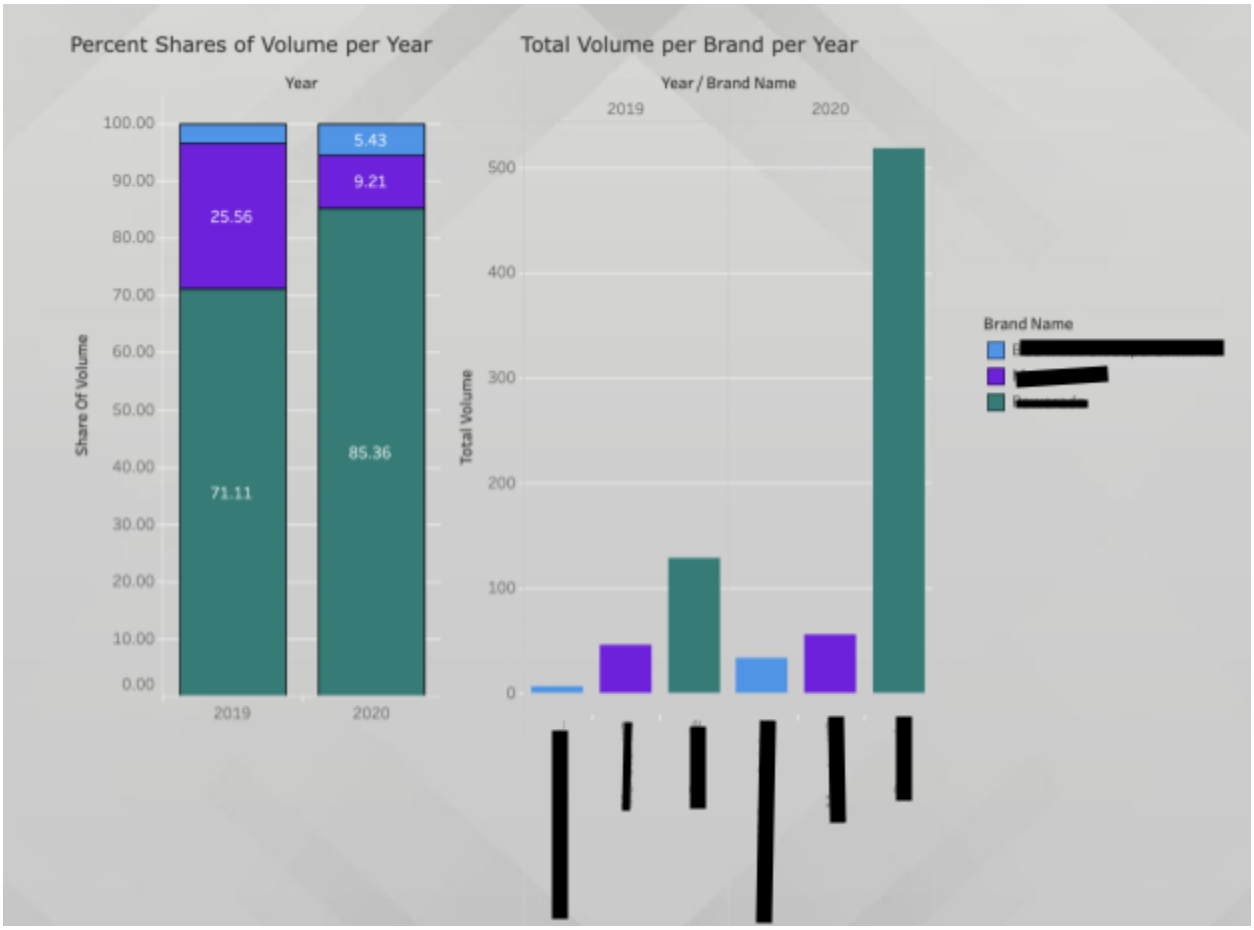
3. The CFO wants to understand which_____ brands are growing vs declining over time, and how these changes impact overall performance for this customer group. Create a data visualization showing share of volume by brand, and year over year change for total volume.

I created a new column in the fact_data table which contained the 2 digit year that was parsed out from invoice_date, I then created a temporary table to select all the data from the fact_data table and included another column of the 4 digit year which I obtained by using a case when statement. In order to find the share of volume per year, I needed to find the total volume_cases per year. We only want to focus on _____brands, so I added a where clause to have the energy_brand_flag =1. The total ended up being 180 for 2019 and 608 for 2020. After that I selected the year, brand_name, and formula for share volume which was just the total volume per brand divided by the total amount per year for each year. Grouping it by the year and then the name outputs the share of volume, then ordered by year to clearly see the percentages for each year.

```sql
alter table fact_data
add column 2_digit_year int;

update fact_data
set year= right(invoice_date,2);

create temporary table fact_data2
select *,
case when 2_digit_year='19' then 2019
when 2_digit_year='20' then '2020'
end as year
from fact_data;

select sum(volume_cases)
from fact_data2 f
join product_master p on p.product_no=f.product_no
where energy_brand_flag=1
and year = 2019;

select sum(volume_cases)
from fact_data2 f
join product_master p on p.product_no=f.product_no
where energy_brand_flag=1
and year = 2020;
```

```sql
Select year, p.brand_name, sum(volume_cases) as total_volume_per_brand_per_year,
case when year = 2019 then sum(volume_cases)/180*100
when year = 2020 then sum(volume_cases)/608*100
end as share_of_volume
from product_master p
join fact_data2 f on f.product_no=p.product_no
where energy_brand_flag=1
group by 1,2
order by year;
```

I was then able to insert this table into tableau and create a visualization of shares of volume by brand per year, from the visualization we notice that the_____ Brand decreases but still has the second highest percentage of shares per volume between the brands, and we can visually see the total volume increasing for each brand by year with_____ showing only a slight increase.



4. A sales associates asks for help reconciling two reports for the same customer. One report shows the customer's total volume is 108K cases, but the second report shows total volume is 318K cases. They sent the SQL code along with both reports. Refer to Customer Reports 1 and 2.

a. How would you reconcile the differences between the two reports?

I would first look at the different reports to determine if there's any trends in the difference of the numbers and not just random differences. That could give me an idea of where in the sql code there could have been something missing or added. I then would look at the sql code itself and look at the similarities and differences between them to reconcile them.

b. How would you explain your findings to the North America sales associate?

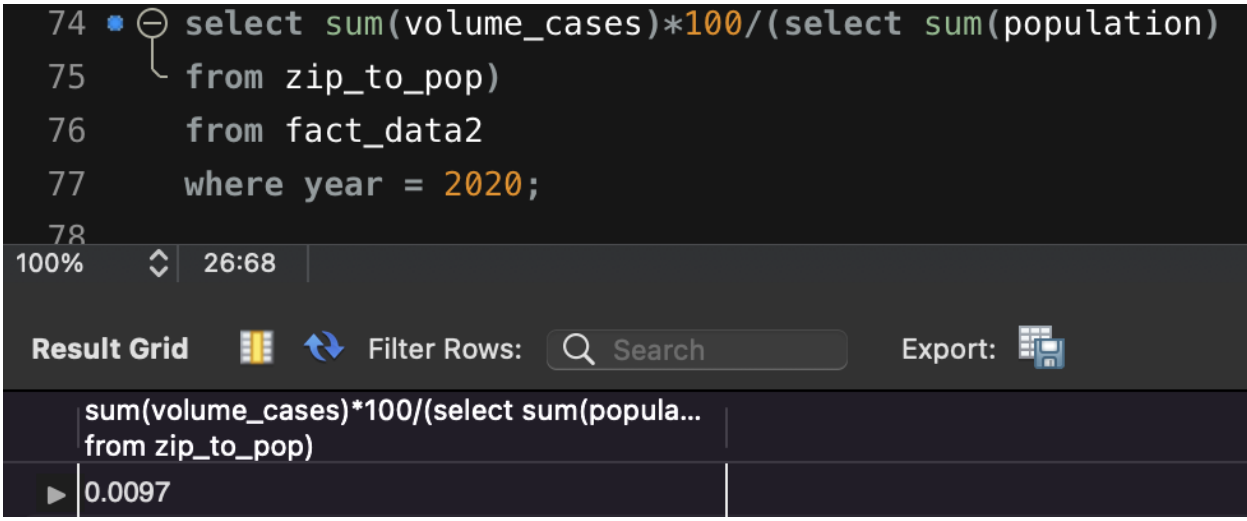
Starting with just the reports itself, customer report 2 has ~3 times as much volume than report 1. When looking at the sql code, report 1 filters on product type being in 2. Without looking at the actual dataset I can assume that filtering the product type would have an overall decline in total volume thus noting the difference between the 2 reports.

Also worth noting that customer report 1 filters on ____ account while report 2 filters on ____. This could possibly be the same thing, but again without looking at the dataset nor having additional information, I would not be able to tell.
The last thing to note is the items in the group by clause. The order in which you group by matters so it is possible that that could have had an affect on the total volume as well.

5. The brand team wants to understand per capita consumption (aka the volume sold per person) during 2020.

a. Using the datasets provided, describe how you would calculate this.

I believe the results I get would be inaccurate. The zip_to_pop table does not specify which year it was for the given population. Was this population taken in 2019 or 2020? or is it a depiction of both years together. Also, some of the populations seem very large for a zip code. For this problem, I'm just going to assume that the population for 2019 and 2020 were the same and the populations for zip code is accurate. If I'm understanding the question right, it wants the total volume per the total population. Filtered on total volume_cases from only the year 2020.The query below gives that. And has a per capita of **0.0097%**



b. Which zip has the highest per capita consumption?

Selected the population, store_zip and per_capita from the year 2020. Grouped by store_zip first, then population to get the per_capita per zip. Ordered by per capita in a descending order and limited by 3 to see the difference between the top 3 zip codes but the zip code with the highest consumption was the first result which is **40507**

```
79  select population, store_zip,sum(volume_cases)/population as per_capita
80  from fact_data2 f
81  join customer_master c on c.customer_no=f.customer_no
82  join zip_to_pop z on z.zip=c.store_zip
83  where year = 2020
84  group by 2,1
85  order by per_capita desc
86  limit 3;
```

100% 1:73

Result Grid Filter Rows: Search Export:

population	store_zip	per_capita
15440	40507	1.5762
40468	40202	1.1152
2317	41559	0.6888