Assignment By Romy Wadhwa

First Part

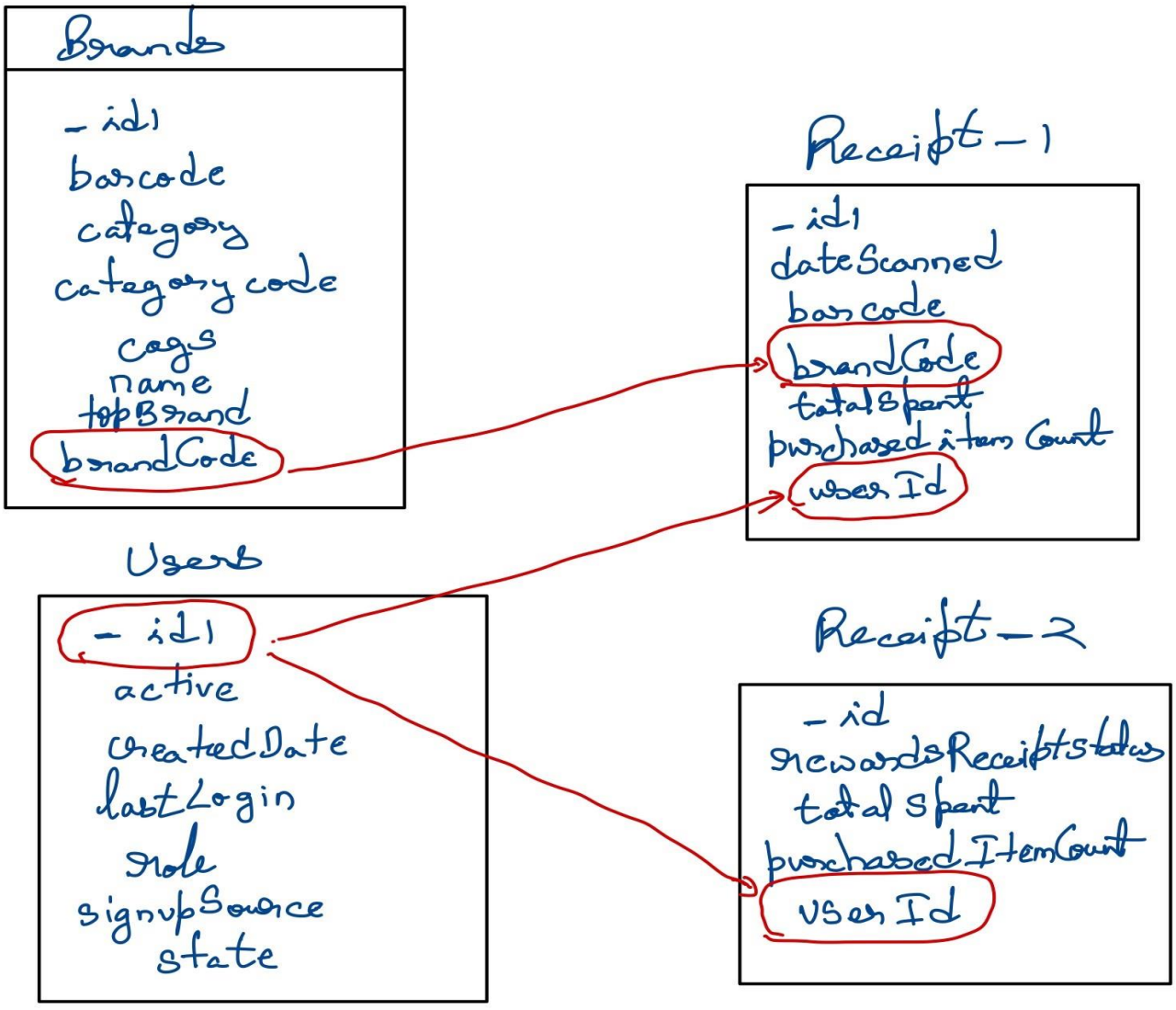
First: Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

After Loading JSON data in Python many operations have been performed for cleaning the data after that dataframes were converted to csv and were loaded to POSTGRE SQL. All SQL queries have been written in the POSTGRE SQL. Please refer Jupyter notebook for the data cleaning code in Python. Screenshots of SQL queries used for solving the questions and their results have been presented in this report.

CSV files named are

Users.csv , brands.csv , receipt_1.csv, receipt_2.csv

Depending upon the data requirement I have used receipt_1.csv or receipt_2.csv



Second Part

Question 1) What are the top 5 brands by receipts scanned for most recent month?

Question 2) How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?

Question 1 and question 2 have been done in combined way

Brand code with BRAND and MISSION are two brands in 2021-02 BRAND ranked number 1 in 2021-01 MISSION ranked 4 in 2021-01

Query | Query History

```
with cte1 as(
select to_date(to_char(datescanned, 'mm/yyyy'), 'mm/yyyy') as y,brandCode, count(brandCode) as cnt
from receipts_1
where brandCode is not null
group by 1,2
),
cte2 as(
select y,brandcode,cnt,
dense_rank() over(partition by y order by y desc,cnt desc) as r1
from cte1
)
select * from cte2
where r1 <= 5
```

	y date	brandcode character varying (100)	cnt bigint	r1 bigint
1	2021-02-01	BRAND	3	1
2	2021-02-01	MISSION	2	2
3	2021-01-01	BRAND	19	1
4	2021-01-01	BEN AND JERRYS	8	2
5	2021-01-01	HY-VEE	8	2
6	2021-01-01	WINGSTOP	7	3
7	2021-01-01	MISSION	5	4
8	2021-01-01	BORDEN	5	4
9	2021-01-01	BETTY CROCKER	4	5
10	2021-01-01	PEPSI	4	5

Question 3) When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

```
with cte as (
select rewardsreceiptstatus,coalesce(sum(totalspent),0) as s from receipts_2
group by rewardsreceiptstatus
  )
select * ,
dense_rank() over(order by s desc) as r1
from cte
order by r1
```

	rewardsreceiptstatus character varying (25)	s double precision	r1 bigint
1	FINISHED	41882.53	1
2	FLAGGED	8300.78	2
3	REJECTED	1656.15000000000	3
4	PENDING	1373.59000000000	4
5	SUBMITTED	0	5

Question 4) When considering *total number of items purchased* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

```
with cte as (
select rewardsreceiptstatus,coalesce(sum(purchasedItemCount),0) as item_cnt from receipts_2
group by rewardsreceiptstatus
  )
select * ,
dense_rank() over(order by item_cnt desc) as r1
from cte
order by r1
```

	rewardsreceiptstatus character varying (25)	item_cnt double precision	r1 bigint
1	FINISHED	8184	1
2	FLAGGED	1014	2
3	REJECTED	173	3
4	SUBMITTED	0	4
5	PENDING	0	4

Question 5) Which brand has the most *spend* among users who were created within the past 6 months?

```
with cte1 as (
select *,to_date(to_char(createddate, 'mm/yyyy'), 'mm/yyyy') as y
from users
),
cte2 as(
select *,
dense_rank() over(order by y desc) as r1
from cte1
order by r1
  ),
cte3 as (
select brandcode, sum(totalSpent) as s ,
dense_rank() over(order by sum(totalSpent) desc) as r2 from cte2
inner join receipts_1 on cte2._id1=receipts_1.userId
where r1<=6 and brandcode is not null
group by brandcode
order by sum(totalSpent) desc
)
select * from cte3
where r2 <=5
order by r2
```

	brandcode character varying (100)	s double precision	r2 bigint
1	HY-VEE	2586	1
2	BETTY CROCKER	1875	2
3	PEPSI	1028	3
4	DORITOS	309	4
5	BEN AND JERRYS	254	5

Question 6) Which brand has the most *transactions* among users who were created within the past 6 months?

```
select *,to_date(to_char(createddate, 'mm/yyyy'), 'mm/yyyy') as y
from users
),
cte2 as(
select *,
dense_rank() over(order by y desc) as r1
from cte1
order by r1
  ),
cte3 as (
select brandcode, count(*) as cnt ,
dense_rank() over(order by count(*) desc) as r2 from cte2
inner join receipts_1 on cte2._id1=receipts_1.userId
where r1<=6 and brandcode is not null
group by brandcode
order by count(*) desc
)
select * from cte3
where r2 <=5
order by r2
```

	brandcode character varying (100)	cnt bigint	r2 bigint
1	BRAND	61	1
2	MISSION	31	2
3	HY-VEE	23	3
4	ORAL-B GLIDE	6	4
5	BEN AND JERRYS	5	5

Third: Evaluate Data Quality Issues in the Data Provided

1) There are numerous data quality concerns, primarily related to missing data. The extent of missing data has been quantified in Jupyter Notebook.
2) The data format itself poses issues. For instance, the TotalSpent column is currently in string format, but it should be converted to an integer format.
3) Within the receipts JSON, the "rewardsReceiptItemList" column contains data in the form of a list, where each element is a single dictionary. However, this crucial column suffers from a significant amount of missing data. Ideally, each element in this column should have a dictionary data type.
4) Inconsistencies in data types are prevalent throughout the dataset, further complicating the data analysis process.

Part 4) Email to Stakeholders:

Subject: Data Quality Concerns and Collaboration Request

Dear [Stakeholder's Name],

I hope this email finds you well. As we continue to explore the data, I have come across a few questions and concerns that I believe require the attention of the data engineering and analytics teams. I would greatly appreciate your assistance in addressing these issues.

Firstly, I have noticed that several columns in the data are missing valuable information. This has had a significant impact on the accuracy of our calculations regarding brand sales. To ensure the reliability of our analyses, it is crucial that we address these missing data points promptly.

Additionally, while working with the data in Python, I have identified various inconsistencies in the format of certain columns. Notably, the prices column is currently in string format, which may pose challenges for accurate calculations. Furthermore, we have encountered numerous instances of missing values in the columns pertaining to item count and scanned items.

In light of these data quality concerns, I would like to propose a collaborative effort between the data engineering and analytics teams. Firstly, I would reach out to the data engineering team to confirm the missing data and work towards finding a solution. Secondly, I believe it would be beneficial to have a discussion with the data engineering team regarding the format inconsistencies. By jointly reviewing the data originating systems, we can identify and rectify the underlying causes of these quality issues.

In order to enhance the overall data quality and reduce the occurrence of missing data, it is essential that the data engineering and analytics teams collaborate closely. By doing so, we can ensure that the data we work with is accurate, complete, and reliable.

I kindly request your support and involvement in facilitating this collaboration between the teams. Together, we can address these data quality concerns and optimize the effectiveness of our analyses.

Thank you for your attention to this matter. I am available to discuss further details or provide any additional information that may be required. Looking forward to your response.

Best regards,

Romy Wadhwa

Data Analyst

513-879-8749