Hello Team,  
Hope you are doing well!  
  
I am working on three datasets ‘rececipts’, ‘brands’, and ‘users’ to obtain answers for given analytical questions. Here is the approach I went with:

The data is in semi structured JSON format. Firstly, I flattened the data to convert it into a structured and simple way in Python. Then I investigated data issues in these datasets. We can solve these issues after a quick discussion on the same. Then I went on to write SQL queries to answer 5 questions.

**Data Issues:**

1. A lot of important columns such as id, brandCode have missing data.
2. Brandcode is missing in receipts data but it is present in brands data. Which means the analysis outputs might be a little off.
3. Duplicate rows in Receipts and Users data due to un-nesting. This causes difficulty to find the primary key for the tables.
4. Invalid and testing data is present in production dataset. (Assuming ‘Test brand’ is a valid brand for this analysis)

**Answers:**

**Q1** What are the top 5 brands by receipts scanned for most recent month?

**Q2** How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?

A screenshot of a computer

Description automatically generated

* In the above table we can see the three brands with their latest month(February 2021) receipt count and their previous month(January 2021) receipt count.
* There are only 3 when 5 were asked, as the latest month has limited data.

**Q3** When considering average spend from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

A close up of a number

Description automatically generated

* Assuming the question ment 'Finished' and 'Rejected', above are the average amounts for both of them.
* Here Finished amount is far greater than Rejected amount based on total spent.

**Q4** When considering total number of items purchased from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

A close up of a number

Description automatically generated

* Again, assuming the question ment 'Finished' and 'Rejected', above are the item counts for both of them.
* Here Finished item count is far greater than Rejected count based on purchased Item Count.

**Q5** Which brand has the most spend among users who were created within the past 6 months?[¶](http://localhost:8889/notebooks/Downloads/Fetch%20Analytics%20Engineer%20Data%20Challenge/Fetch%20Data%20Challenge.ipynb#Q5-Which-brand-has-the-most-spend-among-users-who-were-created-within-the-past-6-months?)

A black text on a white background

Description automatically generated

* 'BEN AND JERRYS' has the most spend among users who were created within the last 6 months from the max date

Regards,  
Ajinkya