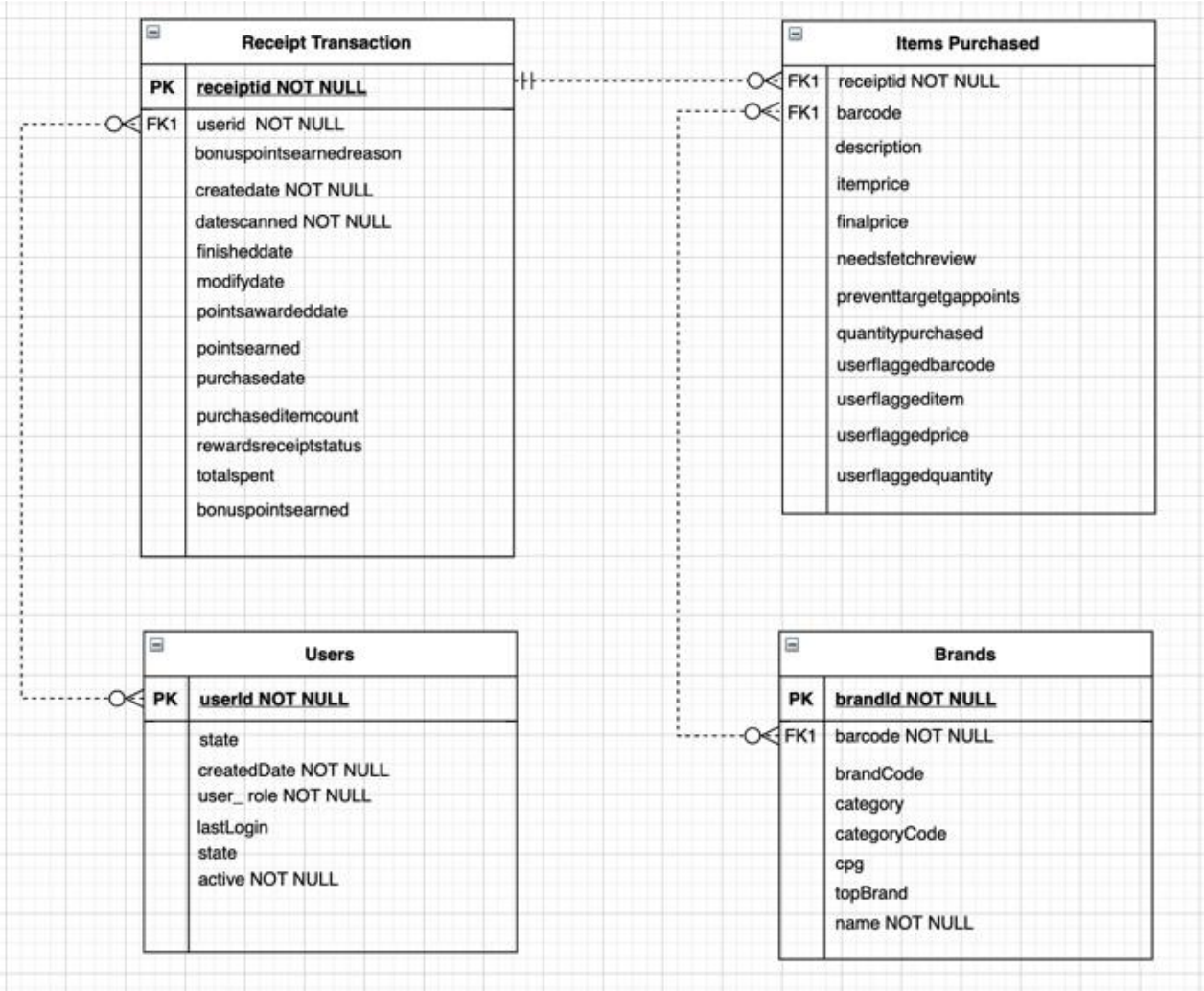
Fetch Rewards Coding Exercise - Data Analyst

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

The ERD Diagram is as shown below:



**Please refer to the drawio file in the git repo for the original ERD file*

Second: Write a query that directly answers a predetermined question from a business stakeholder

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

```
select count(*),b.brandname from
items_purchased i
join brands b on (i.barcode = b.barcode)
join receipts_transactions r on (i.receiptid = r.receiptid)
where r.createdate>='2021-01-01 00:00:00'
group by b.brandname
order by count(*) desc limit 5;
```

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

```
SELECT
   AVG (CAST (totalspent AS FLOAT)),rewardsreceiptstatus
FROM receipts_transactions
where LOWER(rewardsreceiptstatus) in ('accepted','finished','rejected')
group by rewardsreceiptstatus;
```

****Please refer to the .sql files in the git repo for all the queries***

Third: Evaluate Data Quality Issues in the Data Provide

The overall data quality for all datasets was evaluated based on missingness, date format, duplicate values and Outliers.
All the dates were converted to **date_time** format from **Javascript** date format.

The Receipts Data is divided into two separate tables, namely **Items purchased** and **Receipt transactions** to simplify data understanding and analyze the data quality issues.

1. Receipt transactions

- **Overview**

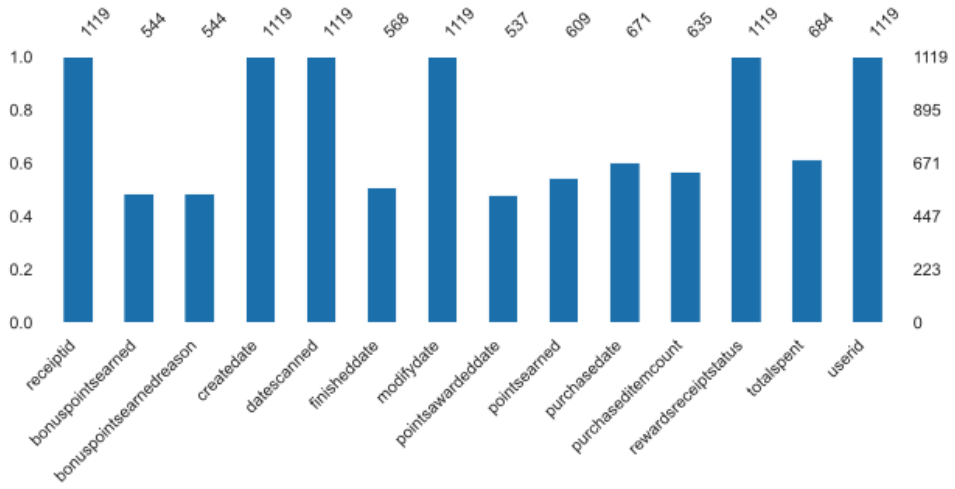
Dataset statistics

Number of variables	14
Number of observations	1119
Missing cells	4160
Missing cells (%)	26.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	122.5 KiB
Average record size in memory	112.1 B

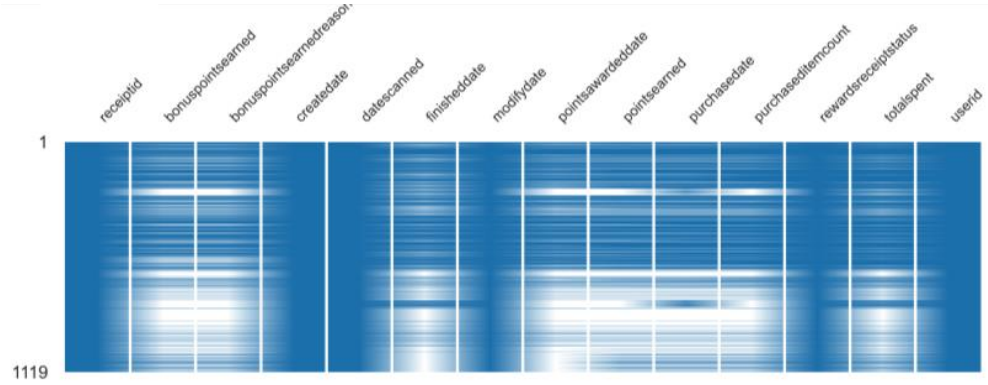
There are **1119 unique receipts** in the Receipt Data.

- **Missingness**

Bar plot below shows the count of values in each column



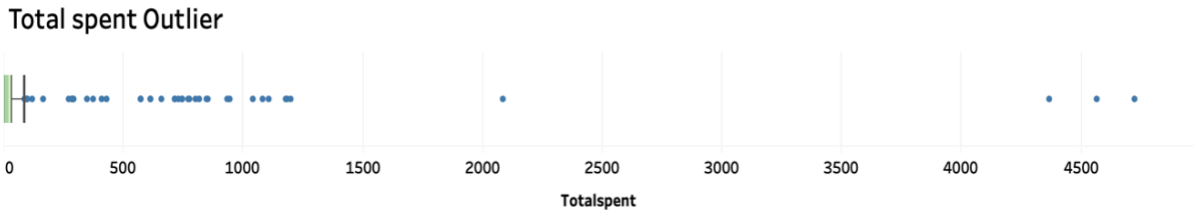
Heatmap showing the location of the missing values in the data

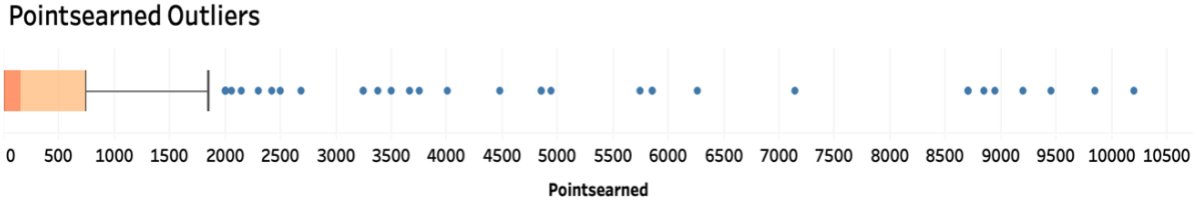


Sr. no.	Column	Count of Missing values	Percent Missing
1.	bonuspointsearned	575	51.4%
2.	bonuspointsearnedreason	575	51.4%
3.	finisheddate	551	49.2%
4.	pointsawardeddate	582	52%
5.	pointsearned	510	45.6%
6.	purchasedate	448	40%
7.	purchaseditemcount	484	43.3%
8.	totalspent	435	38.9%

- **<u>Outliers</u>**

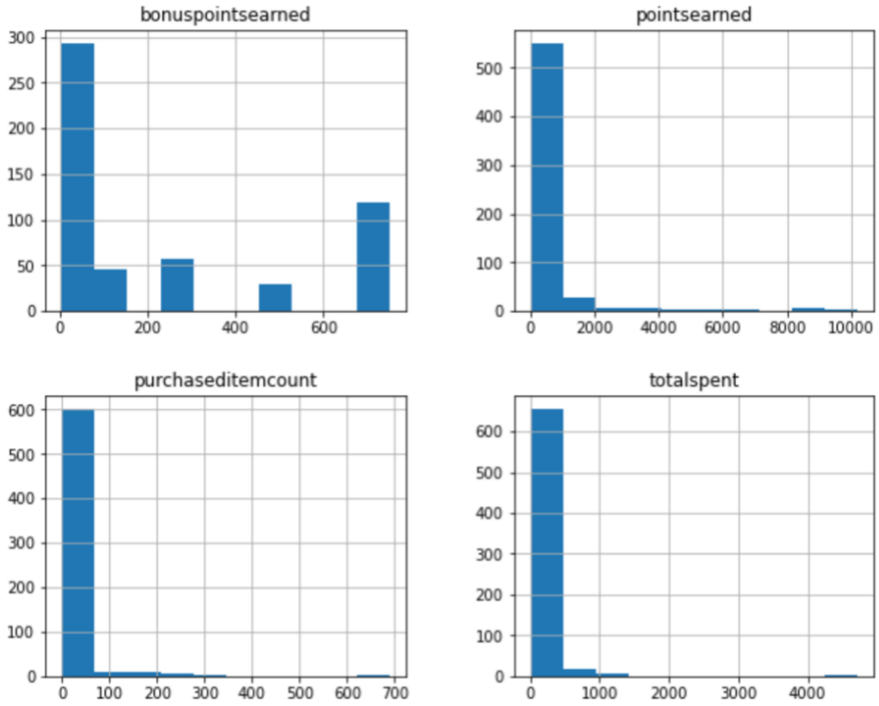
  The plot below shows the outliers in the **Total spent** and **Pointsearned** columns.





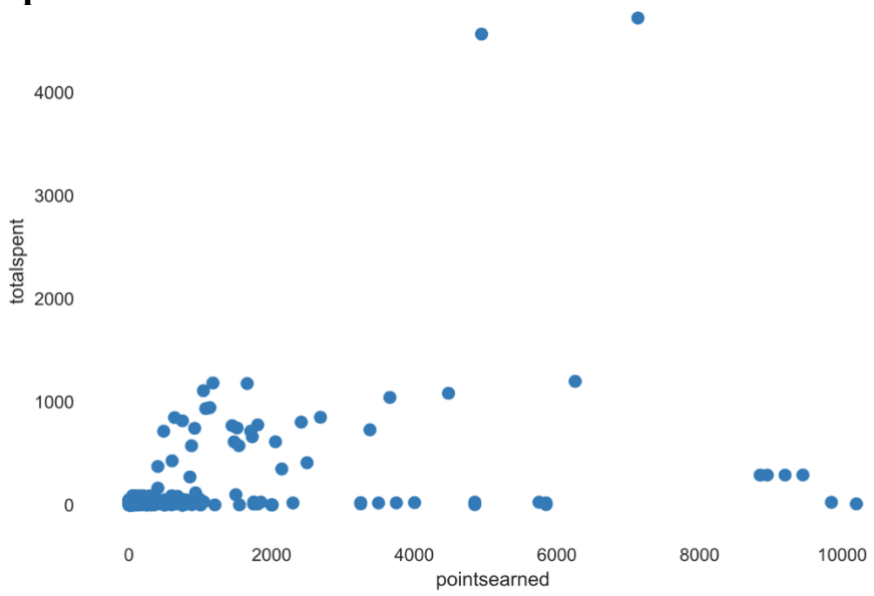
- **Distributions**

  The distribution of all the numeric variables is as shown below:



- **Insights**

  The plot below shows the interaction between **pointsearned** and **totalspent**.



2. Items purchased

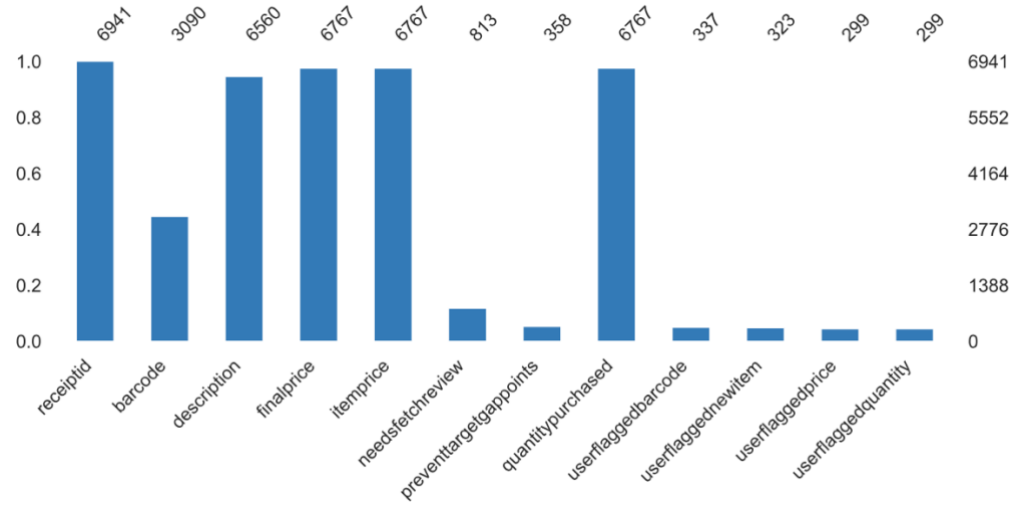
- **Overview**

Dataset statistics

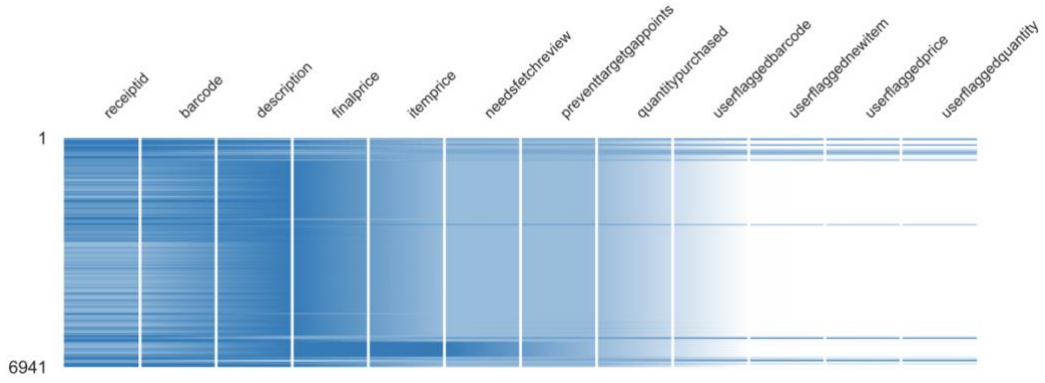
Number of variables	12
Number of observations	6941
Missing cells	43971
Missing cells (%)	52.8%
Duplicate rows	1462
Duplicate rows (%)	21.1%
Total size in memory	650.8 KiB
Average record size in memory	96.0 B

- **Missingness**

Bar plot below shows the count of values in each column



Heatmap showing the location of the missing values in the data



Sr. no.	Column	Count of Missing values	Percent Missing
1.	barcode	3851	55.5%
2.	description	381	5.5%
3.	finalprice	174	2.5%
4.	itemprice	174	2.5%
5.	needsfetchreview	6128	88.3%
6.	preventtargetgappoints	6583	94.8%
7.	quantitypurchased	174	2.5%
8.	userflaggedbarcode	6604	95.1%
9.	userflaggednewitem	6618	95.3%
10.	useflaggedprice	6642	95.7%
11.	userflaggedquantity	6642	95.7%

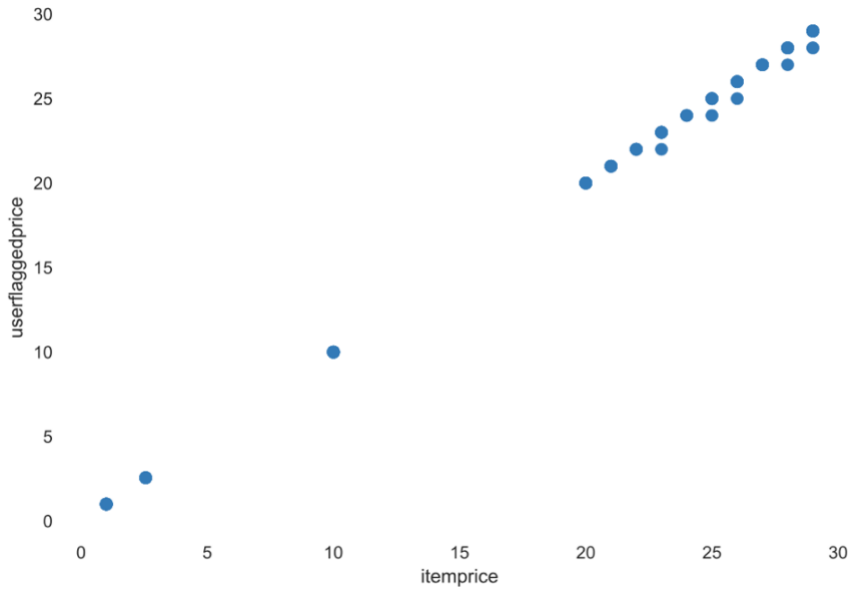
- **Duplicates**

There are **1462 duplicate rows**, thus **52.8%** of the Items purchased table is duplicate. With one receipt_id repeated as much as 10 times, which happens to be the purchase of **Miller Lite 24 Pack.**

There are **679 unique receipts** in this table.

- **Insights**

  The plot below shows the interaction between the **itemprice** and **userflaggedprice.** Users tend to flag items priced over $15. However, this cannot be generalized because over **95%** of **userflagged** price is missing.



3. Users Data

- **Overview**

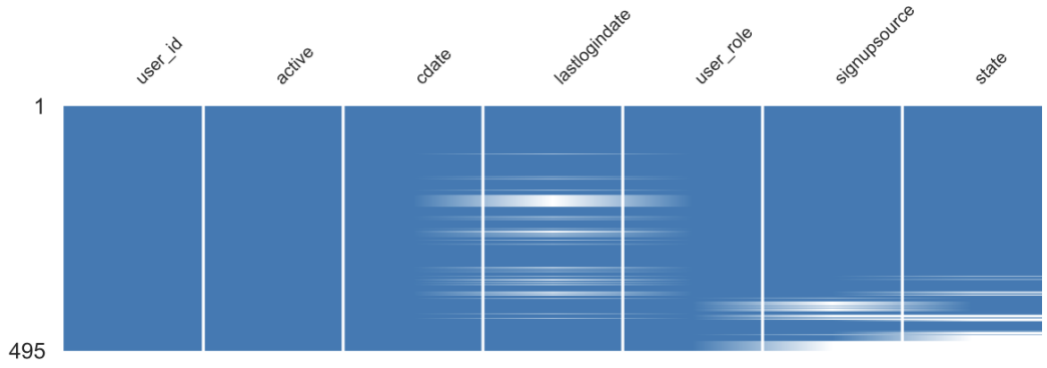
  There are **212 unique users** in this dataset.

Dataset statistics

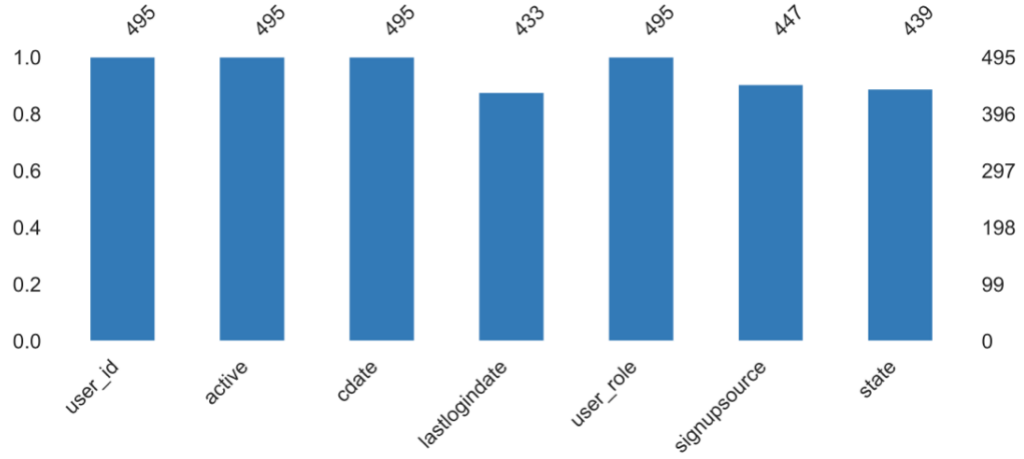
Number of variables	7
Number of observations	495
Missing cells	166
Missing cells (%)	4.8%
Duplicate rows	283
Duplicate rows (%)	57.2%
Total size in memory	23.8 KiB
Average record size in memory	49.3 B

- **Missingness**

Heatmap showing the location of the missing values in the data



Bar plot below shows the count of values in each column



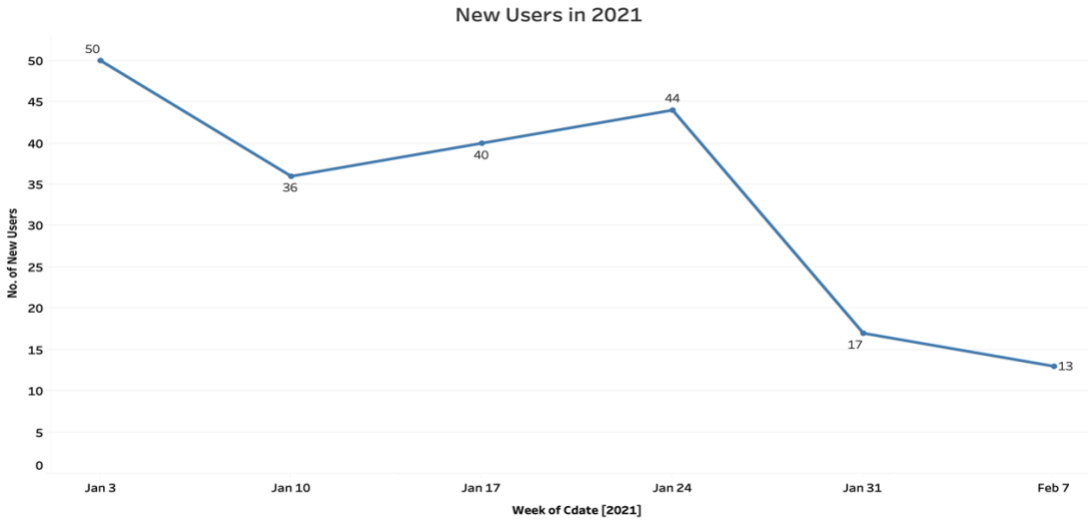
Sr. no.	Column	Count of Missing values	Percent Missing
1.	lastlogindate	62	12.5%
2.	signupsource	48	9.7%
3.	State	56	11.3%

- **Duplicates**

There are **283 duplicate rows**, thus **57.2%** of the Users data is duplicate. With one User_id repeated as much as 20 times.

- **Insights**

The Line graph below depicts no. of users who joined Fetch Rewards in 2021



4. Brands Data

- **Overview**

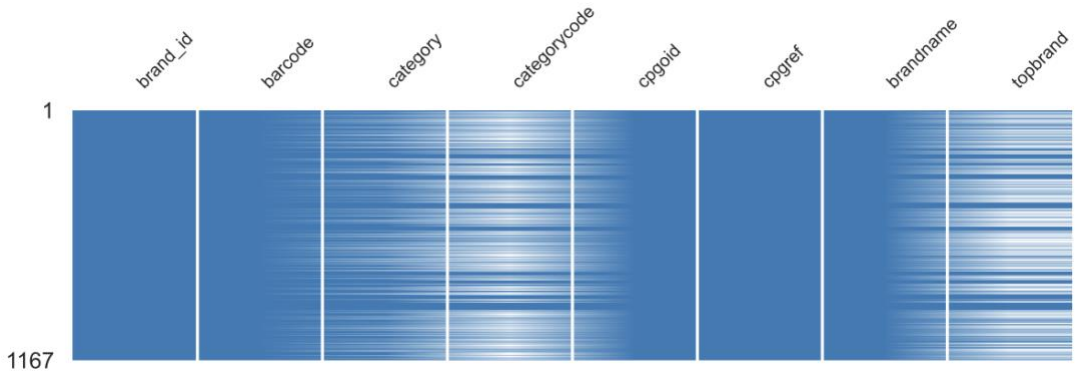
There are **1167 unique brands** in the dataset.

Dataset statistics

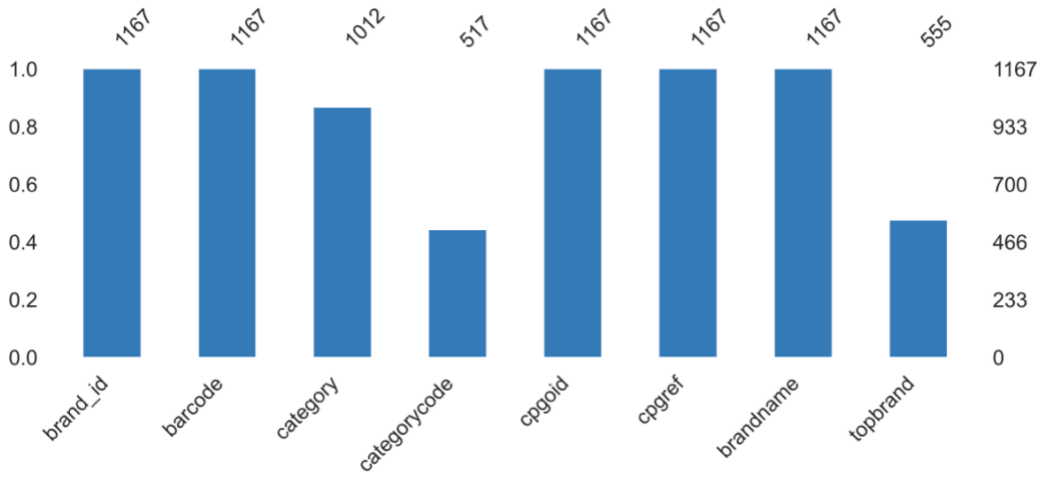
Number of variables	8
Number of observations	1167
Missing cells	1417
Missing cells (%)	15.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	73.1 KiB
Average record size in memory	64.1 B

- **Missingness**

Heatmap showing the location of the missing values in the data



Bar plot below shows the count of values in each column



Sr. no.	Column	Count of Missing values	Percent Missing
1.	category	155	13.3%
2.	category code	650	55.7%
3.	topbrand	612	52.4%

- **Duplicates**

  The duplicate brand names are shown below:

Value	Count
ONE A DAY® WOMENS	2
Huggies	2
V8 Hydrate	2
Health Magazine	2
I CAN'T BELIEVE IT'S NOT BUTTER!	2
Baken-Ets	2
Sierra Mist	2
Pull-Ups	2
Caleb's Kola	2
Dippin Dots® Cereal	2

  There are **23 distinct categories** and **14 distinct category codes**.

  Please refer to the Tableau Dashboards for more insights

Fourth: Communicate with Stakeholders

Hello,

I hope you are doing well. After taking a look at the data I have a few questions/thoughts that I would like to share with you. I consolidated the datasets into CSV files and used a combination of Python and Tableau to generate a data quality report. I primarily looked at the missingness, duplicate values, outliers and consistency of date formats. I have the following questions about the data:

- What decision are you planning to make using this data?
- The user data indicates that most of the users are concentrated in the state of Wisconsin, is it due to any specific marketing campaigns targeted for Wisconsin?
- The metadata does not define user_id to be a unique identifier. Is there any unique identifier for the user, perhaps a combination of multiple fields?
- The metadata does not mention anything about the 'fetch staff' user role. Does it represent the fetch employees who use fetch?
- What is the meaning of userflagged features in the dataset? Are they important?

In order to solve some of the data quality issues it would be important to know the following things:

- It is always good to know the source through which the data was collected.
- If the data was collected through multiple sources, there needs to be an ETL process in place to perform some preliminary processing on data before bringing it over to the data warehouse. This will include elimination of duplicates and imputation of missing variables
- Which features are irrelevant? Irrelevant features with missing values can be dropped from the analysis completely.

Furthermore, I need to know the following things in order to start building data assets/applications based on this dataset.

- What are the most important metrics in the entire dataset?
- What decisions are you looking to drive from the analysis of this dataset?

- I believe Fetch rewards operates on a business model that makes money via affiliate commissions paid by the brands they partner with. Is the 'top brands' feature an indication of its partner brands? If not, we need a feature that indicates our brand partners.
- What are the KPIs that you want to track on a day-to-day basis? It would be best to have an automated dashboard which updates everyday with these KPIs.
- If the users were to be segmented into different groups, what would be the ideal number of groups you would like the users to be divided in?

I would also like to mention some of the performance/scaling concerns I anticipate in production and the steps that need to be taken in order to avoid these issues. They are as follows:

- **User tracking issues due to duplicate user ids and rows in the user data.**
  It is important to have a unique identifier for users because duplicate user ids will cause problems if we decide to perform customer segmentation. The duplicate rows in the user data will drop the accuracy of the analysis.

- **Missing values in the important features**
  Missing values in the important features will cause problems while training machine learning/forecasting models on the dataset. The important features should be set to NOT NULL while collecting the data.

- **Inconsistent barcode data**
  The barcodes consist of alphanumeric characters that can get messy when dealing with large datasets. Thus, there should either be a new unique identifier for brands that can join the brands data to the other tables or, the existing barcodes should be mapped to numeric values.

I apologize for the long email. Please feel free to reach out if you have any other questions or need any additional information.

Thank you,

Akash Bhoite