Receipts_Quality

May 29, 2024

1 Import Libraries

[1]:
```
import pandas as pd
import numpy as np
import json
```

2 Convert JSON into DataFrame Object

[2]:
```
receipts = pd.read_json('receipts.json', lines=True)
```

3 Get overview of data

[3]:
```
receipts.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1119 entries, 0 to 1118
Data columns (total 15 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   _id                      1119 non-null   object
 1   bonusPointsEarned        544 non-null    float64
 2   bonusPointsEarnedReason  544 non-null    object
 3   createDate               1119 non-null   object
 4   dateScanned              1119 non-null   object
 5   finishedDate             568 non-null    object
 6   modifyDate               1119 non-null   object
 7   pointsAwardedDate        537 non-null    object
 8   pointsEarned             609 non-null    float64
 9   purchaseDate             671 non-null    object
 10  purchasedItemCount       635 non-null    float64
 11  rewardsReceiptItemList   679 non-null    object
 12  rewardsReceiptStatus     1119 non-null   object
 13  totalSpent               684 non-null    float64
 14  userId                   1119 non-null   object
dtypes: float64(4), object(11)
memory usage: 131.3+ KB
```

[4]:
```
receipts.head()
```

[4]:

```
                                      _id  bonusPointsEarned  \
0  {'$oid': '5ff1e1eb0a720f0523000575'}              500.0
1  {'$oid': '5ff1e1bb0a720f052300056b'}              150.0
2  {'$oid': '5ff1e1f10a720f052300057a'}                5.0
3  {'$oid': '5ff1e1ee0a7214ada100056f'}                5.0
4  {'$oid': '5ff1e1d20a7214ada1000561'}                5.0

                            bonusPointsEarnedReason  \
0  Receipt number 2 completed, bonus point schedu…
1  Receipt number 5 completed, bonus point schedu…
2                        All-receipts receipt bonus
3                        All-receipts receipt bonus
4                        All-receipts receipt bonus

                 createDate               dateScanned  \
0  {'$date': 1609687531000}  {'$date': 1609687531000}
1  {'$date': 1609687483000}  {'$date': 1609687483000}
2  {'$date': 1609687537000}  {'$date': 1609687537000}
3  {'$date': 1609687534000}  {'$date': 1609687534000}
4  {'$date': 1609687506000}  {'$date': 1609687506000}

                finishedDate                modifyDate  \
0  {'$date': 1609687531000}  {'$date': 1609687536000}
1  {'$date': 1609687483000}  {'$date': 1609687488000}
2                       NaN  {'$date': 1609687542000}
3  {'$date': 1609687534000}  {'$date': 1609687539000}
4  {'$date': 1609687511000}  {'$date': 1609687511000}

          pointsAwardedDate  pointsEarned              purchaseDate  \
0  {'$date': 1609687531000}         500.0  {'$date': 1609632000000}
1  {'$date': 1609687483000}         150.0  {'$date': 1609601083000}
2                       NaN           5.0  {'$date': 1609632000000}
3  {'$date': 1609687534000}           5.0  {'$date': 1609632000000}
4  {'$date': 1609687506000}           5.0  {'$date': 1609601106000}

   purchasedItemCount                                rewardsReceiptItemList  \
0                 5.0  [{'barcode': '4011', 'description': 'ITEM NOT …
1                 2.0  [{'barcode': '4011', 'description': 'ITEM NOT …
2                 1.0  [{'needsFetchReview': False, 'partnerItemId': …
3                 4.0  [{'barcode': '4011', 'description': 'ITEM NOT …
4                 2.0  [{'barcode': '4011', 'description': 'ITEM NOT …

  rewardsReceiptStatus  totalSpent                    userId
0             FINISHED        26.0  5ff1e1eacfcf6c399c274ae6
1             FINISHED        11.0  5ff1e194b6a9d73a3a9f1052
```

```
2             REJECTED        10.0  5ff1e1f1cfcf6c399c274b0b
3             FINISHED        28.0  5ff1e1eacfcf6c399c274ae6
4             FINISHED         1.0  5ff1e194b6a9d73a3a9f1052
```

4 Evaluation

4.1 Mean/Median/Quartiles

[5]:
```
receipts.describe()
```

[5]:

	bonusPointsEarned	pointsEarned	purchasedItemCount	totalSpent
count	544.000000	609.000000	635.00000	684.000000
mean	238.893382	585.962890	14.75748	77.796857
std	299.091731	1357.166947	61.13424	347.110349
min	5.000000	0.000000	0.00000	0.000000
25%	5.000000	5.000000	1.00000	1.000000
50%	45.000000	150.000000	2.00000	18.200000
75%	500.000000	750.000000	5.00000	34.960000
max	750.000000	10199.800000	689.00000	4721.950000

Standard deviation is fairly large, but the max values aren't large enough for me to think the numbers are corrupted.

4.2 Unique Values

[6]:
```
# Column rewardsReceiptItemList contains lists of dictionaries, we can convert␣
 ↪them to check for full uniqueness, but we'll drop that column
# with a copy of the dataframe for time
receipts_no_items = receipts.copy()
receipts_no_items = receipts_no_items.drop(columns=['rewardsReceiptItemList'])
```

[7]:
```
# Converting all dictionary columns to JSON so we can use nunique() to check␣
 ↪for uniqueness
receipts_no_items._id = receipts_no_items._id.apply(json.dumps)
receipts_no_items.createDate = receipts_no_items.createDate.apply(json.dumps)
receipts_no_items.dateScanned = receipts_no_items.dateScanned.apply(json.dumps)
receipts_no_items.finishedDate = receipts_no_items.finishedDate.apply(json.
 ↪dumps)
receipts_no_items.modifyDate = receipts_no_items.modifyDate.apply(json.dumps)
receipts_no_items.pointsAwardedDate = receipts_no_items.pointsAwardedDate.
 ↪apply(json.dumps)
receipts_no_items.purchaseDate = receipts_no_items.purchaseDate.apply(json.
 ↪dumps)
```

[8]:
```
receipts_no_items.nunique()
```

[8]:
```
_id                         1119
bonusPointsEarned             12
bonusPointsEarnedReason        9
createDate                  1107
dateScanned                 1107
finishedDate                 554
modifyDate                  1104
pointsAwardedDate            524
pointsEarned                 119
purchaseDate                 359
purchasedItemCount            50
rewardsReceiptStatus           5
totalSpent                    94
userId                       258
dtype: int64
```

The primary key '_id' contains as many unique values as there are rows, meaning every row is unique

4.3 Null/NaN/Missing Values

[9]:
```
# Get the total number of records with Null values
receipts.isnull().sum()
```

[9]:
```
_id                         0
bonusPointsEarned         575
bonusPointsEarnedReason   575
createDate                  0
dateScanned                 0
finishedDate              551
modifyDate                  0
pointsAwardedDate         582
pointsEarned              510
purchaseDate              448
purchasedItemCount        484
rewardsReceiptItemList    440
rewardsReceiptStatus        0
totalSpent                435
userId                      0
dtype: int64
```

[10]:
```
# Rewards Receipt Item list has 440 missing values? Let's look at few rows with␣
 ↪missing items and see why
receipts[receipts['rewardsReceiptItemList'].isnull()]
```

[10]:
```
                                    _id  bonusPointsEarned  \
71    {'$oid': '5ff475820a7214ada10005cf'}                NaN
93    {'$oid': '5ff5ecb90a7214ada10005f9'}                NaN
```

```
149   {'$oid': '5ff726860a720f05230005ec'}                NaN
175   {'$oid': '5ff8da570a720f05c5000015'}                NaN
212   {'$oid': '5ffce8570a7214ad4e003e6f'}                NaN
…                                      …                  …
1110  {'$oid': '603c6adf0a720fde1000039a'}                NaN
1111  {'$oid': '603c9e6e0a720fde100003c7'}                NaN
1115  {'$oid': '603d0b710a720fde1000042a'}                NaN
1116  {'$oid': '603cf5290a720fde10000413'}                NaN
1118  {'$oid': '603c4fea0a7217c72c000389'}                NaN

     bonusPointsEarnedReason                createDate  \
71                       NaN  {'$date': 1609856386000}
93                       NaN  {'$date': 1609952440000}
149                      NaN  {'$date': 1610032774000}
175                      NaN  {'$date': 1610144343000}
212                      NaN  {'$date': 1610410071000}
…                        …                          …
1110                     NaN  {'$date': 1614572255736}
1111                     NaN  {'$date': 1614585454307}
1115                     NaN  {'$date': 1614613361873}
1116                     NaN  {'$date': 1614607657664}
1118                     NaN  {'$date': 1614565354962}

                   dateScanned finishedDate                 modifyDate  \
71    {'$date': 1609856386000}          NaN  {'$date': 1609856386000}
93    {'$date': 1609952440000}          NaN  {'$date': 1609952440000}
149   {'$date': 1610032774000}          NaN  {'$date': 1610032774000}
175   {'$date': 1610144343000}          NaN  {'$date': 1610144344000}
212   {'$date': 1610410071000}          NaN  {'$date': 1610410071000}
…                          …            …                        …
1110  {'$date': 1614572255736}          NaN  {'$date': 1614572255736}
1111  {'$date': 1614585454307}          NaN  {'$date': 1614585454307}
1115  {'$date': 1614613361873}          NaN  {'$date': 1614613361873}
1116  {'$date': 1614607657664}          NaN  {'$date': 1614607657664}
1118  {'$date': 1614565354962}          NaN  {'$date': 1614565354962}

     pointsAwardedDate  pointsEarned purchaseDate  purchasedItemCount  \
71                 NaN           NaN          NaN                 NaN
93                 NaN           NaN          NaN                 NaN
149                NaN           NaN          NaN                 NaN
175                NaN           NaN          NaN                 0.0
212                NaN           NaN          NaN                 NaN
…                  …             …            …                   …
1110               NaN           NaN          NaN                 NaN
1111               NaN           NaN          NaN                 NaN
1115               NaN           NaN          NaN                 NaN
1116               NaN           NaN          NaN                 NaN
```

```
1118                NaN           NaN          NaN                   NaN

     rewardsReceiptItemList rewardsReceiptStatus  totalSpent  \
71                      NaN            SUBMITTED         NaN
93                      NaN            SUBMITTED         NaN
149                     NaN            SUBMITTED         NaN
175                     NaN             REJECTED         0.0
212                     NaN            SUBMITTED         NaN
…                       …                    …          …
1110                    NaN            SUBMITTED         NaN
1111                    NaN            SUBMITTED         NaN
1115                    NaN            SUBMITTED         NaN
1116                    NaN            SUBMITTED         NaN
1118                    NaN            SUBMITTED         NaN

                        userId
71    5a43c08fe4b014fd6b6a0612
93    5a43c08fe4b014fd6b6a0612
149   5ff7264e8f142f11dd189504
175   5ff8da28b3348b11c9337ac6
212   59c124bae4b0299e55b0f330
…                          …
1110  5fc961c3b8cfca11a077dd33
1111  5fc961c3b8cfca11a077dd33
1115  5fc961c3b8cfca11a077dd33
1116  5fc961c3b8cfca11a077dd33
1118  5fc961c3b8cfca11a077dd33

[440 rows x 15 columns]
```

Rows without items look to be receipts without any items on them. That makes some sense because it's the column with the 2nd-least amount of missing values. Let's check to see if a row exists with both bonusPointsEarned & rewardsReceiptItemList

[11]:
```
receipts_2 = receipts.copy()
```

[12]:
```
receipts_2[receipts_2['rewardsReceiptItemList'].isnull()]
```

[12]:
```
                                        _id  bonusPointsEarned  \
71    {'$oid': '5ff475820a7214ada10005cf'}                NaN
93    {'$oid': '5ff5ecb90a7214ada10005f9'}                NaN
149   {'$oid': '5ff726860a720f05230005ec'}                NaN
175   {'$oid': '5ff8da570a720f05c5000015'}                NaN
212   {'$oid': '5ffce8570a7214ad4e003e6f'}                NaN
…                                      …                  …
1110  {'$oid': '603c6adf0a720fde1000039a'}                NaN
1111  {'$oid': '603c9e6e0a720fde100003c7'}                NaN
1115  {'$oid': '603d0b710a720fde1000042a'}                NaN
```

```
1116  {'$oid': '603cf5290a720fde10000413'}                NaN
1118  {'$oid': '603c4fea0a7217c72c000389'}                NaN

     bonusPointsEarnedReason                createDate  \
71                       NaN  {'$date': 1609856386000}
93                       NaN  {'$date': 1609952440000}
149                      NaN  {'$date': 1610032774000}
175                      NaN  {'$date': 1610144343000}
212                      NaN  {'$date': 1610410071000}
…                        …                          …
1110                     NaN  {'$date': 1614572255736}
1111                     NaN  {'$date': 1614585454307}
1115                     NaN  {'$date': 1614613361873}
1116                     NaN  {'$date': 1614607657664}
1118                     NaN  {'$date': 1614565354962}

                   dateScanned finishedDate                modifyDate  \
71    {'$date': 1609856386000}          NaN  {'$date': 1609856386000}
93    {'$date': 1609952440000}          NaN  {'$date': 1609952440000}
149   {'$date': 1610032774000}          NaN  {'$date': 1610032774000}
175   {'$date': 1610144343000}          NaN  {'$date': 1610144344000}
212   {'$date': 1610410071000}          NaN  {'$date': 1610410071000}
…                          …          …                          …
1110  {'$date': 1614572255736}          NaN  {'$date': 1614572255736}
1111  {'$date': 1614585454307}          NaN  {'$date': 1614585454307}
1115  {'$date': 1614613361873}          NaN  {'$date': 1614613361873}
1116  {'$date': 1614607657664}          NaN  {'$date': 1614607657664}
1118  {'$date': 1614565354962}          NaN  {'$date': 1614565354962}

     pointsAwardedDate  pointsEarned purchaseDate  purchasedItemCount  \
71                 NaN           NaN          NaN                 NaN
93                 NaN           NaN          NaN                 NaN
149                NaN           NaN          NaN                 NaN
175                NaN           NaN          NaN                 0.0
212                NaN           NaN          NaN                 NaN
…                  …             …            …                   …
1110               NaN           NaN          NaN                 NaN
1111               NaN           NaN          NaN                 NaN
1115               NaN           NaN          NaN                 NaN
1116               NaN           NaN          NaN                 NaN
1118               NaN           NaN          NaN                 NaN

     rewardsReceiptItemList rewardsReceiptStatus  totalSpent  \
71                      NaN            SUBMITTED         NaN
93                      NaN            SUBMITTED         NaN
149                     NaN            SUBMITTED         NaN
175                     NaN             REJECTED         0.0
```

```
212                     NaN            SUBMITTED         NaN
…                       …                …          …
1110                    NaN            SUBMITTED         NaN
1111                    NaN            SUBMITTED         NaN
1115                    NaN            SUBMITTED         NaN
1116                    NaN            SUBMITTED         NaN
1118                    NaN            SUBMITTED         NaN

                        userId
71    5a43c08fe4b014fd6b6a0612
93    5a43c08fe4b014fd6b6a0612
149   5ff7264e8f142f11dd189504
175   5ff8da28b3348b11c9337ac6
212   59c124bae4b0299e55b0f330
…                          …
1110  5fc961c3b8cfca11a077dd33
1111  5fc961c3b8cfca11a077dd33
1115  5fc961c3b8cfca11a077dd33
1116  5fc961c3b8cfca11a077dd33
1118  5fc961c3b8cfca11a077dd33

[440 rows x 15 columns]
```

[13]:
```
receipts_2 = receipts_2.dropna(axis=0, subset=['bonusPointsEarned'])
```

[14]:
```
receipts_2[receipts_2['rewardsReceiptItemList'].isnull()]
```

[14]:
```
                                      _id  bonusPointsEarned  \
396  {'$oid': '6009eb000a7214ada2000003'}              250.0
424  {'$oid': '600aff160a720f053500000c'}              500.0

                            bonusPointsEarnedReason  \
396  Receipt number 3 completed, bonus point schedu…
424  Receipt number 2 completed, bonus point schedu…

                  createDate               dateScanned  \
396  {'$date': 1611262720000}  {'$date': 1611262720000}
424  {'$date': 1611333398000}  {'$date': 1611333398000}

                finishedDate                modifyDate  \
396  {'$date': 1611262746000}  {'$date': 1611262755000}
424  {'$date': 1611333421000}  {'$date': 1611333433000}

           pointsAwardedDate  pointsEarned              purchaseDate  \
396  {'$date': 1611262746000}         250.0  {'$date': 1611187200000}
424  {'$date': 1611333421000}         500.0  {'$date': 1611273600000}
```

```
     purchasedItemCount rewardsReceiptItemList rewardsReceiptStatus  \
396                 0.0                    NaN             FINISHED
424                 0.0                    NaN             FINISHED

     totalSpent                     userId
396         0.0  6009e60450b3311194385009
424         0.0  600afb2a7d983a124e9aded0
```

[15]:
```
# 2 receipts with a Items list when all rows with missing 'bonusPointsEarned'␣
↪are dropped, let's check the reason:
print(receipts.iloc[396]['bonusPointsEarnedReason'])
print(receipts.iloc[424]['bonusPointsEarnedReason'])
```

```
Receipt number 3 completed, bonus point schedule DEFAULT
(5cefdcacf3693e0b50e83a36)
Receipt number 2 completed, bonus point schedule DEFAULT
(5cefdcacf3693e0b50e83a36)
```

My hypothesis is either: 1. Once a given number of receipts are scanned a set number of points will be rewarded. 2. These are errored receipts.

The 1st hypothesis is supported searching the receipts.json file. Every row where bonusPointsEarnedReason contains 'Receipt number 3 completed' has bonusPointsEarned == 250.0.

[16]:
```
# For good measure, let's see every receipt with 'Receipt number 3 completed...
↪':
receipts_3 = receipts.copy()
receipts_3 = receipts_3.groupby('bonusPointsEarnedReason', as_index=False)
```

[17]:
```
receipts_3.get_group('Receipt number 3 completed, bonus point schedule DEFAULT␣
↪(5cefdcacf3693e0b50e83a36)')
```

[17]:
```
                                        _id  bonusPointsEarned  \
9    {'$oid': '5ff1e1eb0a7214ada100056b'}              250.0
51   {'$oid': '5ff36c570a720f0523000593'}              250.0
121  {'$oid': '5ff7946f0a720f052300063c'}              250.0
147  {'$oid': '5ff726640a720f05230005e6'}              250.0
162  {'$oid': '5ff8cea10a7214adca00000b'}              250.0
166  {'$oid': '5ff8da390a7214adca000013'}              250.0
168  {'$oid': '5ff8da7f0a7214adca000022'}              250.0
173  {'$oid': '5ff873f90a720f0523000651'}              250.0
185  {'$oid': '5ffc9daa0a7214adca000054'}              250.0
190  {'$oid': '5ffcb4900a720f0515000002'}              250.0
204  {'$oid': '5ffc8fa10a7214adca00002e'}              250.0
252  {'$oid': '5fff4c810a720f05f3000023'}              250.0
293  {'$oid': '6000d4900a7214ad4c000064'}              250.0
296  {'$oid': '6000d4ac0a7214ad4c000069'}              250.0
317  {'$oid': '60020af60a720f05f3000089'}              250.0
340  {'$oid': '600742490a720f05fa000004'}              250.0
```

```
362  {'$oid': '600887560a720f05fa000098'}              250.0
396  {'$oid': '6009eb000a7214ada2000003'}              250.0
399  {'$oid': '600988820a7214ad8900012b'}              250.0
422  {'$oid': '600affe60a7214ada2000009'}              250.0
462  {'$oid': '600edb570a720f053500001d'}              250.0
487  {'$oid': '600fb1fd0a720f053500004a'}              250.0
490  {'$oid': '6010bdfa0a7214ada200005a'}              250.0
510  {'$oid': '60118be80a7214ada2000075'}              250.0
526  {'$oid': '60132ae40a7214ad50000007'}              250.0
550  {'$oid': '601448f30a720f05f80000d9'}              250.0
561  {'$oid': '601442dd0a7214ad50000054'}              250.0
574  {'$oid': '60144adf0a720f05f80000ef'}              250.0
649  {'$oid': '60183c850a7214ad50000308'}              250.0
654  {'$oid': '60182d8c0a720f05f8000318'}              250.0
912  {'$oid': '6024024b0a7214d8e900018d'}              250.0

                              bonusPointsEarnedReason  \
9    Receipt number 3 completed, bonus point schedu…
51   Receipt number 3 completed, bonus point schedu…
121  Receipt number 3 completed, bonus point schedu…
147  Receipt number 3 completed, bonus point schedu…
162  Receipt number 3 completed, bonus point schedu…
166  Receipt number 3 completed, bonus point schedu…
168  Receipt number 3 completed, bonus point schedu…
173  Receipt number 3 completed, bonus point schedu…
185  Receipt number 3 completed, bonus point schedu…
190  Receipt number 3 completed, bonus point schedu…
204  Receipt number 3 completed, bonus point schedu…
252  Receipt number 3 completed, bonus point schedu…
293  Receipt number 3 completed, bonus point schedu…
296  Receipt number 3 completed, bonus point schedu…
317  Receipt number 3 completed, bonus point schedu…
340  Receipt number 3 completed, bonus point schedu…
362  Receipt number 3 completed, bonus point schedu…
396  Receipt number 3 completed, bonus point schedu…
399  Receipt number 3 completed, bonus point schedu…
422  Receipt number 3 completed, bonus point schedu…
462  Receipt number 3 completed, bonus point schedu…
487  Receipt number 3 completed, bonus point schedu…
490  Receipt number 3 completed, bonus point schedu…
510  Receipt number 3 completed, bonus point schedu…
526  Receipt number 3 completed, bonus point schedu…
550  Receipt number 3 completed, bonus point schedu…
561  Receipt number 3 completed, bonus point schedu…
574  Receipt number 3 completed, bonus point schedu…
649  Receipt number 3 completed, bonus point schedu…
654  Receipt number 3 completed, bonus point schedu…
```

```
912  Receipt number 3 completed, bonus point schedu…

                    createDate                dateScanned  \
9    {'$date': 1609687531000}  {'$date': 1609687531000}
51   {'$date': 1609788503000}  {'$date': 1609788503000}
121  {'$date': 1609866511000}  {'$date': 1609866511000}
147  {'$date': 1610032740000}  {'$date': 1610032740000}
162  {'$date': 1609968545000}  {'$date': 1609968545000}
166  {'$date': 1610144313000}  {'$date': 1610144313000}
168  {'$date': 1610144383000}  {'$date': 1610144383000}
173  {'$date': 1610118137000}  {'$date': 1610118137000}
185  {'$date': 1610390954000}  {'$date': 1610390954000}
190  {'$date': 1610396816000}  {'$date': 1610396816000}
204  {'$date': 1610387361000}  {'$date': 1610387361000}
252  {'$date': 1610566785000}  {'$date': 1610566785000}
293  {'$date': 1610667152000}  {'$date': 1610667152000}
296  {'$date': 1610667180000}  {'$date': 1610667180000}
317  {'$date': 1610746614000}  {'$date': 1610746614000}
340  {'$date': 1611088457000}  {'$date': 1611088457000}
362  {'$date': 1611171670000}  {'$date': 1611171670000}
396  {'$date': 1611262720000}  {'$date': 1611262720000}
399  {'$date': 1611237506000}  {'$date': 1611237506000}
422  {'$date': 1611333606000}  {'$date': 1611333606000}
462  {'$date': 1611586391000}  {'$date': 1611586391000}
487  {'$date': 1611641341000}  {'$date': 1611641341000}
490  {'$date': 1611709946000}  {'$date': 1611709946000}
510  {'$date': 1611762664000}  {'$date': 1611762664000}
526  {'$date': 1611868900000}  {'$date': 1611868900000}
550  {'$date': 1611942131000}  {'$date': 1611942131000}
561  {'$date': 1611940573000}  {'$date': 1611940573000}
574  {'$date': 1611942623000}  {'$date': 1611942623000}
649  {'$date': 1612201093000}  {'$date': 1612201093000}
654  {'$date': 1612197260000}  {'$date': 1612197260000}
912  {'$date': 1612972619000}  {'$date': 1612972619000}

                   finishedDate                modifyDate  \
9    {'$date': 1609687531000}  {'$date': 1609687536000}
51   {'$date': 1609788503000}  {'$date': 1609788503000}
121  {'$date': 1610039315000}  {'$date': 1610060917000}
147  {'$date': 1610033642000}  {'$date': 1610033642000}
162  {'$date': 1610141355000}  {'$date': 1610141355000}
166  {'$date': 1610144314000}  {'$date': 1610144314000}
168  {'$date': 1610144383000}  {'$date': 1610144388000}
173  {'$date': 1610118137000}  {'$date': 1610118137000}
185  {'$date': 1610390955000}  {'$date': 1610390955000}
190                      NaN  {'$date': 1610396817000}
204  {'$date': 1610387368000}  {'$date': 1610387368000}
```

```
252  {'$date': 1610566786000}  {'$date': 1610566786000}
293  {'$date': 1610667153000}  {'$date': 1610667153000}
296  {'$date': 1610667181000}  {'$date': 1610667185000}
317  {'$date': 1610746985000}  {'$date': 1610759601000}
340  {'$date': 1611088458000}  {'$date': 1611088458000}
362  {'$date': 1611171671000}  {'$date': 1611171671000}
396  {'$date': 1611262746000}  {'$date': 1611262755000}
399  {'$date': 1611237507000}  {'$date': 1611237507000}
422  {'$date': 1611333622000}  {'$date': 1611333632000}
462  {'$date': 1611586821000}  {'$date': 1611586909000}
487  {'$date': 1611641342000}  {'$date': 1611641342000}
490  {'$date': 1611710849000}  {'$date': 1611710849000}
510                       NaN  {'$date': 1611762665000}
526  {'$date': 1611868901000}  {'$date': 1611868901000}
550  {'$date': 1611942131000}  {'$date': 1611942131000}
561  {'$date': 1611940573000}  {'$date': 1611940574000}
574  {'$date': 1611942624000}  {'$date': 1611942624000}
649  {'$date': 1612201094000}  {'$date': 1612201098000}
654                       NaN  {'$date': 1612197260000}
912  {'$date': 1612972620000}  {'$date': 1612972620000}

           pointsAwardedDate  pointsEarned              purchaseDate  \
9    {'$date': 1609687531000}         250.0  {'$date': 1609632000000}
51   {'$date': 1609788503000}         250.0  {'$date': 1609702103000}
121  {'$date': 1610039315000}        3250.0  {'$date': 1609848000000}
147  {'$date': 1610032741000}         250.0  {'$date': 1609891200000}
162  {'$date': 1610141355000}        1999.6  {'$date': 1609882145000}
166  {'$date': 1610144314000}         350.0  {'$date': 1609459200000}
168  {'$date': 1610144383000}         250.0  {'$date': 1610064000000}
173  {'$date': 1610118137000}         250.0  {'$date': 1610031737000}
185  {'$date': 1610390955000}         350.0  {'$date': 1609718400000}
190                       NaN        8950.0  {'$date': 1613075216000}
204  {'$date': 1610387361000}         250.0  {'$date': 1610300961000}
252  {'$date': 1610566786000}         250.0  {'$date': 1610480385000}
293  {'$date': 1610667153000}         350.0  {'$date': 1609977600000}
296  {'$date': 1610667181000}         250.0  {'$date': 1610582400000}
317  {'$date': 1610746985000}         250.0  {'$date': 1610668800000}
340  {'$date': 1611088458000}         250.0  {'$date': 1610776800000}
362  {'$date': 1611171671000}         250.0  {'$date': 1613850070000}
396  {'$date': 1611262746000}         250.0  {'$date': 1611187200000}
399  {'$date': 1611237507000}         300.0  {'$date': 1609891200000}
422  {'$date': 1611333622000}         250.0  {'$date': 1611273600000}
462  {'$date': 1611586821000}         487.7  {'$date': 1611532800000}
487  {'$date': 1611641342000}         250.0  {'$date': 1610323200000}
490  {'$date': 1611709948000}         250.0  {'$date': 1611619200000}
510                       NaN         250.0  {'$date': 1611676264000}
526  {'$date': 1611868900000}         250.0  {'$date': 1611187200000}
```

```
550  {'$date': 1611942131000}         375.0  {'$date': 1611468000000}
561  {'$date': 1611940573000}         250.0  {'$date': 1611854173000}
574  {'$date': 1611942624000}         250.0  {'$date': 1611856223000}
649  {'$date': 1612201094000}         250.0  {'$date': 1612137600000}
654                       NaN         250.0  {'$date': 1612110860000}
912  {'$date': 1612972620000}         250.0  {'$date': 1612504800000}

     purchasedItemCount                                rewardsReceiptItemList  \
9                   3.0  [{'barcode': '4011', 'description': 'ITEM NOT …
51                  1.0  [{'barcode': '021000002917', 'competitiveProdu…
121                 3.0  [{'barcode': '4011', 'description': 'ITEM NOT …
147                 5.0  [{'barcode': '013562300631', 'description': 'A…
162                 2.0  [{'barcode': '4011', 'description': 'ITEM NOT …
166                 1.0  [{'barcode': '305210154278', 'brandCode': 'BRA…
168                 4.0  [{'barcode': '4011', 'description': 'ITEM NOT …
173                 1.0  [{'barcode': '4011', 'description': 'ITEM NOT …
185                 1.0  [{'barcode': '001111147332', 'brandCode': 'BRA…
190                10.0  [{'barcode': '034100573065', 'description': 'M…
204                 2.0  [{'barcode': '4011', 'description': 'ITEM NOT …
252                 1.0  [{'barcode': '025800026302', 'description': 'S…
293                 1.0  [{'barcode': '822142991523', 'brandCode': 'BRA…
296                 3.0  [{'barcode': '4011', 'description': 'ITEM NOT …
317                 1.0  [{'barcode': '037000979715', 'brandCode': 'ORA…
340                 5.0  [{'barcode': '686924113172', 'competitiveProdu…
362                 1.0  [{'barcode': '4011', 'description': 'ITEM NOT …
396                 0.0                                              NaN
399                10.0  [{'barcode': '013000798952', 'description': 'H…
422                21.0  [{'brandCode': 'BLUE DIAMOND', 'description': …
462               161.0  [{'description': 'GHRRDL V IDNIGHT REV #', 'di…
487                10.0  [{'barcode': '025400076363', 'competitiveProdu…
490                 5.0  [{'barcode': '013562300631', 'description': 'A…
510                 1.0  [{'barcode': '4011', 'description': 'ITEM NOT …
526                 1.0  [{'barcode': '070470400235', 'brandCode': 'BRA…
550                 5.0  [{'barcode': '043000006283', 'description': 'C…
561                 1.0  [{'barcode': '4011', 'description': 'ITEM NOT …
574                 1.0  [{'barcode': '021000024582', 'competitiveProdu…
649                 3.0  [{'barcode': '4011', 'description': 'ITEM NOT …
654                 1.0  [{'barcode': '4011', 'description': 'ITEM NOT …
912                 5.0  [{'barcode': '013000006187', 'description': 'H…

    rewardsReceiptStatus  totalSpent                    userId
9               FINISHED       20.00  5ff1e1eacfcf6c399c274ae6
51              FINISHED       10.00  5ff36be7135e7011bcb856d3
121             FINISHED       23.00  5ff79464b3348b11c933738b
147             FINISHED       50.00  5ff7264e8f142f11dd189504
162             FINISHED        1.00  5ff8ce8504929111f6e913cb
166             FINISHED       10.00  5ff8da28b3348b11c9337ac6
```

168	FINISHED	26.00	5ff8da7eb3348b11c9337b72
173	FINISHED	1.00	5ff873d1b3348b11c9337716
185	FINISHED	10.00	5ffc9d87b3348b11c9338920
190	FLAGGED	290.00	5ffcb47d04929111f6e9256c
204	FINISHED	1.00	5ffc8f9704929111f6e922bf
252	FINISHED	10.00	5fff4beedf9ace121f0c17ea
293	FINISHED	10.00	6000d46cfb296c121a81b20c
296	FINISHED	23.00	6000d4abe2571211db395b5c
317	FINISHED	3.99	6000b75bbe5fc96dfee1d4d3
340	FINISHED	25.00	600741d06e6469120a787853
362	FINISHED	1.00	6008873eb6310511daa4e8eb
396	FINISHED	0.00	6009e60450b3311194385009
399	FINISHED	10.00	600987d77d983a11f63cfa92
422	FINISHED	83.85	600afb2a7d983a124e9aded0
462	FINISHED	715.23	600ed95043298911ce45e82c
487	FINISHED	10.00	600fb1ac73c60b12049027bb
490	FINISHED	50.00	6010bddaa4b74c120bd19dfb
510	FLAGGED	1.00	60118bcfa4b74c18d3a8c0d7
526	FINISHED	10.00	60132acb73c60b3ca7f3ba32
550	FINISHED	25.00	6014485b84231211ce793d79
561	FINISHED	1.00	601442ce67804a1228b1dc41
574	FINISHED	10.00	6014499c67804a1228b1f5b0
649	FINISHED	27.00	60183c839a1b091205b61aca
654	FLAGGED	1.00	60182d6dc8b50e11d84547b2
912	FINISHED	25.00	602401f1c081001222409ad6

The 1st hypothesis holds. These receipts are markers that a person has scanned a given number of receipts.

4.4 Deeper Dive into rewardsReceiptItemList Array

[18]:
```
# From our earlier uniqueness checks, we know that column␣
 ↪rewardsReceiptItemList contains lists of dictionaries.
# Let's take a look into it.
rewards = receipts['rewardsReceiptItemList']
```

[19]:
```
print(rewards.iloc[0])
```

```
[{'barcode': '4011', 'description': 'ITEM NOT FOUND', 'finalPrice': '26.00',
'itemPrice': '26.00', 'needsFetchReview': False, 'partnerItemId': '1',
'preventTargetGapPoints': True, 'quantityPurchased': 5, 'userFlaggedBarcode':
'4011', 'userFlaggedNewItem': True, 'userFlaggedPrice': '26.00',
'userFlaggedQuantity': 5}]
```

[20]:
```
rewards.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 1119 entries, 0 to 1118
Series name: rewardsReceiptItemList
```

```
Non-Null Count  Dtype
--------------  -----
679 non-null    object
dtypes: object(1)
memory usage: 8.9+ KB
```

[21]:
```
rewards.head()
```

[21]:
```
0    [{'barcode': '4011', 'description': 'ITEM NOT …
1    [{'barcode': '4011', 'description': 'ITEM NOT …
2    [{'needsFetchReview': False, 'partnerItemId': …
3    [{'barcode': '4011', 'description': 'ITEM NOT …
4    [{'barcode': '4011', 'description': 'ITEM NOT …
Name: rewardsReceiptItemList, dtype: object
```

[22]:
```
rewards_no_nan = rewards.copy()
```

[23]:
```
rewards_no_nan = rewards_no_nan.dropna()
```

4.4.1 Get the embedded element names:

[24]:
```
# Quick and dirty method to get all the rewardsReceiptItemList elements:
# Not performant, but needed a way to get all the element names for the␣
 ↪relational diagram.
rewards_set = set()
for i in rewards_no_nan:
    for j in i:
        for key in j:
            rewards_set.add(key)
```

[25]:
```
# Note: This set will be used for the relational diagram.
print(rewards_set)
```

```
{'itemPrice', 'originalReceiptItemText', 'targetPrice', 'userFlaggedQuantity',
'partnerItemId', 'brandCode', 'needsFetchReview', 'rewardsProductPartnerId',
'itemNumber', 'originalFinalPrice', 'originalMetaBriteBarcode',
'userFlaggedPrice', 'originalMetaBriteItemPrice', 'description', 'pointsEarned',
'userFlaggedNewItem', 'pointsPayerId', 'preventTargetGapPoints',
'pointsNotAwardedReason', 'competitiveProduct',
'originalMetaBriteQuantityPurchased', 'userFlaggedDescription', 'finalPrice',
'originalMetaBriteDescription', 'discountedItemPrice', 'competitorRewardsGroup',
'metabriteCampaignId', 'needsFetchReviewReason', 'quantityPurchased',
'priceAfterCoupon', 'deleted', 'barcode', 'userFlaggedBarcode', 'rewardsGroup'}
```

4.5 What's With The Points?

Why are the categories bonusPointsEarned and pointsEarned both floats instead of integers?

[26]:
```
receipts['bonusPointsEarned'].sum()
```

[26]: 129958.0

[27]:
```
receipts['pointsEarned'].sum()
```

[27]: 356851.39999999997

Looks like fractions of points can be earned based on the items themselves, but not for bonus awards.

5 Conclusions:

More unstructured data coming through == more problems.

In terms of quality issues, there's some small nitpicks here and there, but the biggest problem is embedding a receipt's list of items purchased as its own dataset within a column. I can understand the reason why it comes in like this, but was there no other way to send rewardsReceiptItemList as a standalone dataset? A fair amount of quality issues that can, or are already, arising can be much more easily rectified if that data was sent separately.