data_quality_exploration

April 2, 2024

[106]:
```
import pandas

# read in the json files
udf = pd.read_json('users.json', lines=True)
bdf = pd.read_json('brands.json', lines=True)
rdf = pd.read_json('receipts.json', lines=True)

udf.head()
```

[106]:

	_id	active	createdDate	lastLogin	role	signUpSource	state
0	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Email	WI
1	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Email	WI
2	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Email	WI
3	{'$oid': '5ff1e1eacfcf6c399c274ae6'}	True	{'$date': 1609687530554}	{'$date': 1609687530597}	consumer	Email	WI
4	{'$oid': '5ff1e194b6a9d73a3a9f1052'}	True	{'$date': 1609687444800}	{'$date': 1609687537858}	consumer	Email	WI

[ ]:
```
# dates are in epoch form and need to be converted to timestamp
```

[28]:
```
# lastLogin, signUpSource, and state are missing for a small fraction of users
# why are only these columns missing for some users?

for c in udf.columns:
    print(f'{c} null count: {round(100*udf[udf[c].isnull()].shape[0] / udf.
 ↪shape[0],2)}%')
```

```
_id null count: 0.0%
active null count: 0.0%
createdDate null count: 0.0%
lastLogin null count: 12.53%
role null count: 0.0%
```

```
signUpSource null count: 9.7%
state null count: 11.31%
```

[31]:
```
# only one user has active=False

udf.groupby('active').agg({'_id': 'count'})
```

[31]:
```
        _id
active
False     1
True    494
```

[32]:
```
# nothing looks weird about this record though
udf[udf.active==False]
```

[32]:
```
                                      _id  active               createdDate  \
240  {'$oid': '6008622ebe5fc9247bab4eb9'}   False  {'$date': 1611162158662}

                     lastLogin      role signUpSource state
240  {'$date': 1611162158931}  consumer        Email    WI
```

[34]:
```
bdf.head()
```

[34]:
```
                                      _id       barcode        category  \
0  {'$oid': '601ac115be37ce2ead437551'}  511111019862          Baking
1  {'$oid': '601c5460be37ce2ead43755f'}  511111519928       Beverages
2  {'$oid': '601ac142be37ce2ead43755d'}  511111819905          Baking
3  {'$oid': '601ac142be37ce2ead43755a'}  511111519874          Baking
4  {'$oid': '601ac142be37ce2ead43755e'}  511111319917  Candy & Sweets

       categoryCode                                                cpg  \
0            BAKING  {'$id': {'$oid': '601ac114be37ce2ead437550'}, …
1         BEVERAGES  {'$id': {'$oid': '5332f5fbe4b03c9a25efd0ba'}, …
2            BAKING  {'$id': {'$oid': '601ac142be37ce2ead437559'}, …
3            BAKING  {'$id': {'$oid': '601ac142be37ce2ead437559'}, …
4  CANDY_AND_SWEETS  {'$id': {'$oid': '5332fa12e4b03c9a25efd1e7'}, …

                        name  topBrand                      brandCode
0  test brand @1612366101024       0.0                            NaN
1                  Starbucks       0.0                       STARBUCKS
2  test brand @1612366146176       0.0  TEST BRANDCODE @1612366146176
3  test brand @1612366146051       0.0  TEST BRANDCODE @1612366146051
4  test brand @1612366146827       0.0  TEST BRANDCODE @1612366146827
```

[145]:
```
# It looks like some records erroneously have barcode data in the brandCode␣
↪column
# these should be fixed
```

```
bdf[bdf.brandCode.str.isnumeric()==True]
```

[145]:

	_id	barcode	category	\
13	{'$oid': '5d6413156d5f3b23d1bc790a'}	511111205012	Magazines	
27	{'$oid': '5d66d71fa3a018093ab34728'}	511111105329	Magazines	
44	{'$oid': '5d66d94d6d5f3b6188d4f04b'}	511111505365	Magazines	
64	{'$oid': '5da609991dda2c3e1416ae90'}	511111805854	Health & Wellness	
134	{'$oid': '5da60576a60b87376833e349'}	511111305569	Health & Wellness	
137	{'$oid': '5da608131dda2c3e1416ae8a'}	511111505716	Health & Wellness	
143	{'$oid': '5d658ff3a3a018514994f432'}	511111005216	Magazines	
149	{'$oid': '5d642dbfa3a018514994f42e'}	511111005148	Magazines	
152	{'$oid': '5c45f91b87ff3552f950f027'}	511111204923	Grocery	
164	{'$oid': '5da6094ca60b87376833e357'}	511111605829	Health & Wellness	
177	{'$oid': '5da608dfa60b87376833e354'}	511111805786	Health & Wellness	
194	{'$oid': '5d6415d5a3a018514994f429'}	511111605058	Magazines	
257	{'$oid': '5d642de76d5f3b23d1bc7911'}	511111705161	Magazines	
263	{'$oid': '5da60932a60b87376833e356'}	511111405818	Health & Wellness	
268	{'$oid': '5da608a8a60b87376833e353'}	511111105763	Health & Wellness	
270	{'$oid': '5d66dda06d5f3b6188d4f050'}	511111005421	Magazines	
281	{'$oid': '5d6412f86d5f3b23d1bc7909'}	511111804994	Magazines	
326	{'$oid': '5d641306a3a018514994f427'}	511111705000	Magazines	
395	{'$oid': '5da608291dda2c3e1416ae8b'}	511111705727	Health & Wellness	
408	{'$oid': '5d642d946d5f3b23d1bc7910'}	511111805137	Magazines	
414	{'$oid': '5d6419746d5f3b23d1bc790f'}	511111605102	Magazines	
418	{'$oid': '5d6423ffa3a018514994f42c'}	511111105114	Magazines	
425	{'$oid': '5d6415c66d5f3b23d1bc790c'}	511111105046	Magazines	
492	{'$oid': '5d642dd1a3a018514994f42f'}	511111505150	Magazines	
506	{'$oid': '5d66d516a3a018093ab34725'}	511111305286	Magazines	
511	{'$oid': '5d66d8c86d5f3b6188d4f049'}	511111805342	Magazines	
531	{'$oid': '5da605bea60b87376833e34a'}	511111805571	Health & Wellness	
540	{'$oid': '5d66dad8a3a018093ab34729'}	511111205388	Magazines	
602	{'$oid': '5d66def56d5f3b6188d4f051'}	511111705444	Magazines	
604	{'$oid': '5da608c91dda2c3e1416ae8d'}	511111605775	Health & Wellness	
627	{'$oid': '5d6412d36d5f3b23d1bc7908'}	511111104971	Magazines	
651	{'$oid': '5d642d65a3a018514994f42d'}	511111305125	Magazines	
658	{'$oid': '5d6594a2a3a018514994f434'}	511111905240	Magazines	
676	{'$oid': '5da607daa60b87376833e350'}	511111405696	Health & Wellness	
682	{'$oid': '5d6417f56d5f3b23d1bc790d'}	511111305071	Magazines	
809	{'$oid': '5d659490a3a018514994f433'}	511111705239	Magazines	
820	{'$oid': '5da6097ea60b87376833e358'}	511111305842	Health & Wellness	
837	{'$oid': '5d6415b3a3a018514994f428'}	511111905035	Magazines	
936	{'$oid': '5d66d6a2a3a018093ab34727'}	511111905318	Magazines	
945	{'$oid': '5da607ef1dda2c3e1416ae89'}	511111005704	Health & Wellness	
947	{'$oid': '5d6417dda3a018514994f42a'}	511111805069	Magazines	
951	{'$oid': '5d66def56d5f3b6188d4f052'}	511111205456	Magazines	
956	{'$oid': '5d6415a26d5f3b23d1bc790b'}	511111405023	Magazines	

```
960   {'$oid': '5d66e03f6d5f3b6188d4f054'}  511111605492          Magazines
1006  {'$oid': '5d66d597a3a018093ab34726'}  511111805298          Magazines
1011  {'$oid': '5da60915a60b87376833e355'}  511111905806  Health & Wellness
1012  {'$oid': '5c4637ba87ff35681e840d57'}  511111605058              Dairy
1014  {'$oid': '5da609621dda2c3e1416ae8f'}  511111105831  Health & Wellness
1032  {'$oid': '5d66d4666d5f3b6188d4f046'}  511111105275          Magazines
1048  {'$oid': '5d66d38ca3a018093ab34724'}  511111605263          Magazines
1069  {'$oid': '5d66d81a6d5f3b6188d4f048'}  511111605331          Magazines
1130  {'$oid': '5d64191ea3a018514994f42b'}  511111505082          Magazines
1133  {'$oid': '5da608f61dda2c3e1416ae8e'}  511111305798  Health & Wellness
1141  {'$oid': '5d66d6016d5f3b6188d4f047'}  511111405306          Magazines
1150  {'$oid': '5d66da306d5f3b6188d4f04c'}  511111005377          Magazines
1151  {'$oid': '5d66dfe6a3a018093ab3472c'}  511111905479          Magazines

     categoryCode                                                    cpg  \
13            NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
27            NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
44            NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
64            NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
134           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
137           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
143           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
149           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
152           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5c45f8b087ff…
164           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
177           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
194           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
257           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
263           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
268           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
270           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
281           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
326           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
395           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
408           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
414           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
418           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
425           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
492           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
506           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
511           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
531           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
540           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
602           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
604           NaN  {'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
627           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
651           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
```

658	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
676	NaN	{'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
682	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
809	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
820	NaN	{'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
837	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
936	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
945	NaN	{'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
947	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
951	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
956	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
960	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1006	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1011	NaN	{'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
1012	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5c45f8b087ff…
1014	NaN	{'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
1032	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1048	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1069	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1130	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1133	NaN	{'$ref': 'Cogs', '$id': {'$oid': '53e10d6368ab…
1141	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1150	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…
1151	NaN	{'$ref': 'Cogs', '$id': {'$oid': '5d5d4fd16d5f…

	name	topBrand	\
13	Entertainment Weekly	NaN	
27	Elegant Homes Magazine	NaN	
44	Magnolia Journal Magazine	NaN	
64	ONE A DAY® WOMENS	0.0	
134	AFRIN® NO DRIP PUMP MISTS	NaN	
137	Claritin® ADULTS	NaN	
143	Cooking Light Magazine	NaN	
149	Shape	NaN	
152	Brand1	1.0	
164	ONE A DAY® MENS	NaN	
177	ONE A DAY® ENERGY	NaN	
194	Health Magazine	NaN	
257	Traditional Home	NaN	
263	ONE A DAY® KIDS AND TEENS	NaN	
268	MiraLAX® Laxatives	NaN	
270	Pillsbury Magazine	NaN	
281	Eating Well Magazine	NaN	
326	Family Circle	NaN	
395	Claritin® KIDS	NaN	
408	Real Simple	NaN	
414	People en Espanol	NaN	

```
418                                    People Magazine       NaN
425                                 Food & Wine Magazine     NaN
492                                      Southern Living     NaN
506                                Country Home Magazine     NaN
511                       LIFE Special Edition Magazine      NaN
531                          AFRIN® ORIGINAL NASAL SPRAY     NaN
540       National Geographic Special Edition Magazine      NaN
602   Rolling Stone Special Collectors Edition Magazine     NaN
604                                       ONE A DAY® 50+    NaN
627                           Travel and Leisure Magazine   NaN
651                                  Rachael Ray Everyday   NaN
658                                     Kitchen and Baths   NaN
676                         Bayer® ASPRIN EXTRA STRENGTH     NaN
682                         MARTHA STEWART LIVING MAGAZINE   0.0
809                                               Happy Paws NaN
820           ONE A DAY® VITACRAVES ADULT MULTIVITAMINS     NaN
837                       People Special Edition Magazine   NaN
936                           Eat This Not That Magazine    NaN
945                         Bayer® ASPRIN TAB/ CAPS/ CHEWS  NaN
947                                               InStyle   NaN
951                          TIME Special Edition Magazine  NaN
956                                Coastal Living Magazine  NaN
960                                  Yoga Journal Magazine  NaN
1006                               Diabetic Living Magazine NaN
1011                                ONE A DAY® HEART HEALTH NaN
1012                                              Brand2    1.0
1014                     ONE A DAY® PRENATAL AND PREGNANCY  NaN
1032                               Country Gardens Magazine NaN
1048                                  Clean Eating Magazine NaN
1069           History Channel Special Edition Magazine     NaN
1130                                       Midwest Living   NaN
1133                                    ONE A DAY® ESSENTIAL NaN
1141             Do It Yourself Special Edition Magazine    NaN
1150                           Modern Farmhouse Magazine    NaN
1151           Weight Watchers Special Edition Magazine     NaN

        brandCode
13    511111205012
27    511111105329
44    511111505365
64    511111805854
134   511111305569
137   511111505716
143   511111005216
149   511111005148
152     0987654321
164   511111605829
```

177	511111805786
194	511111605058
257	511111705161
263	511111405818
268	511111105763
270	511111005421
281	511111804994
326	511111705000
395	511111705727
408	511111805137
414	511111605102
418	511111105114
425	511111105046
492	511111505150
506	511111305286
511	511111805342
531	511111805571
540	511111205388
602	511111705444
604	511111605775
627	511111104971
651	511111305125
658	511111905240
676	511111405696
682	511111305071
809	511111705239
820	511111305842
837	511111905035
936	511111905318
945	511111005704
947	511111805069
951	511111205456
956	511111405023
960	511111605492
1006	511111805298
1011	511111905806
1012	09090909090
1014	511111105831
1032	511111105275
1048	511111605263
1069	511111605331
1130	511111505082
1133	511111305798
1141	511111405306
1150	511111005377
1151	511111905479

[146]:
```
# This affects 4.8% of brands
bdf[bdf.brandCode.str.isnumeric()==True].shape[0] / bdf.shape[0]
```

[146]:
```
0.04798628963153385
```

[140]:
```
bdf.groupby('topBrand').agg({'_id': 'count'})
```

[140]:

topBrand	_id
0.0	524
1.0	31

[139]:
```
# Categories and categoryCode's match appropriately

bdf.groupby(['category', 'categoryCode']).agg({'_id': 'count'})
```

[139]:

category	categoryCode	_id
Baby	BABY	7
Baking	BAKING	359
Beer Wine Spirits	BEER_WINE_SPIRITS	31
Beverages	BEVERAGES	1
Bread & Bakery	BREAD_AND_BAKERY	5
Candy & Sweets	CANDY_AND_SWEETS	71
Cleaning & Home Improvement	CLEANING_AND_HOME_IMPROVEMENT	6
Dairy & Refrigerated	DAIRY_AND_REFRIGERATED	5
Frozen	FROZEN	1
Grocery	GROCERY	11
Health & Wellness	HEALTHY_AND_WELLNESS	14
Magazines	MAGAZINES	1
Outdoor	OUTDOOR	1
Personal Care	PERSONAL_CARE	4

[36]:
```
# category, categoryCode, topBrand, and brandCode are missing for a substantial␣
 ↪amount of records
# brandCode missing is particularly bad because that column is needed to join␣
 ↪brands to receipts
# is there another way to join brands to receipts that doesn't rely on this␣
 ↪column?

for c in bdf.columns:
    print(f'{c} null count: {round(100*bdf[bdf[c].isnull()].shape[0] / bdf.
 ↪shape[0],2)}%')
```

```
_id null count: 0.0%
barcode null count: 0.0%
category null count: 13.28%
```

```
categoryCode null count: 55.7%
cpg null count: 0.0%
name null count: 0.0%
topBrand null count: 52.44%
brandCode null count: 20.05%
```

[37]:
```
rdf.head()
```

[37]:

	_id	bonusPointsEarned	bonusPointsEarnedReason	createDate	dateScanned	finishedDate	modifyDate	pointsAwardedDate	pointsEarned	purchaseDate	purchasedItemCount	rewardsReceiptItemList
0	{'$oid': '5ff1e1eb0a720f0523000575'}	500.0	Receipt number 2 completed, bonus point schedu…	{'$date': 1609687531000}	{'$date': 1609687531000}	{'$date': 1609687531000}	{'$date': 1609687536000}	{'$date': 1609687531000}	500.0	{'$date': 1609632000000}	5.0	[{'barcode': '4011', 'description': 'ITEM NOT …
1	{'$oid': '5ff1e1bb0a720f052300056b'}	150.0	Receipt number 5 completed, bonus point schedu…	{'$date': 1609687483000}	{'$date': 1609687483000}	{'$date': 1609687483000}	{'$date': 1609687488000}	{'$date': 1609687483000}	150.0	{'$date': 1609601083000}	2.0	[{'barcode': '4011', 'description': 'ITEM NOT …
2	{'$oid': '5ff1e1f10a720f052300057a'}	5.0	All-receipts receipt bonus	{'$date': 1609687537000}	{'$date': 1609687537000}	NaN	{'$date': 1609687542000}	NaN	5.0	{'$date': 1609632000000}	1.0	[{'needsFetchReview': False, 'partnerItemId': …
3	{'$oid': '5ff1e1ee0a7214ada100056f'}	5.0	All-receipts receipt bonus	{'$date': 1609687534000}	{'$date': 1609687534000}	{'$date': 1609687534000}	{'$date': 1609687539000}	{'$date': 1609687534000}	5.0	{'$date': 1609632000000}		
4	{'$oid': '5ff1e1d20a7214ada1000561'}	5.0	All-receipts receipt bonus	{'$date': 1609687506000}	{'$date': 1609687506000}	{'$date': 1609687511000}	{'$date': 1609687511000}	{'$date': 1609687506000}	5.0	{'$date': 1609601106000}		

```
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
```

[47]:
```
# Create receipt_items dataframe (from schema diagram)

t = pd.json_normalize(rdf.rewardsReceiptItemList)
t
```

[47]:
```
                                                      0    \
0     {'barcode': '4011', 'description': 'ITEM NOT F…
1     {'barcode': '4011', 'description': 'ITEM NOT F…
2     {'needsFetchReview': False, 'partnerItemId': '…
3     {'barcode': '4011', 'description': 'ITEM NOT F…
4     {'barcode': '4011', 'description': 'ITEM NOT F…
…                                                   …
1114  {'barcode': 'B076FJ92M4', 'description': 'muel…
1115                                               None
1116                                               None
1117  {'barcode': 'B076FJ92M4', 'description': 'muel…
1118                                               None

                                                      1     2     3     4    \
0                                                  None  None  None  None
1     {'barcode': '028400642255', 'description': 'DO…  None  None  None
2                                                  None  None  None  None
3                                                  None  None  None  None
4     {'barcode': '1234', 'finalPrice': '2.56', 'ite…  None  None  None
…                                                   …    …    …    …
1114  {'barcode': 'B07BRRLSVC', 'description': 'thin…  None  None  None
1115                                               None  None  None  None
1116                                               None  None  None  None
1117  {'barcode': 'B07BRRLSVC', 'description': 'thin…  None  None  None
1118                                               None  None  None  None

       5     6     7     8     9    …   449   450   451   452   453   454  \
0     None  None  None  None  None  …  None  None  None  None  None  None
1     None  None  None  None  None  …  None  None  None  None  None  None
2     None  None  None  None  None  …  None  None  None  None  None  None
3     None  None  None  None  None  …  None  None  None  None  None  None
4     None  None  None  None  None  …  None  None  None  None  None  None
```

```
…     …    …    …    …    …    …    …    …    …    …    …    …
1114  None  None  None  None  None  …  None  None  None  None  None  None
1115  None  None  None  None  None  …  None  None  None  None  None  None
1116  None  None  None  None  None  …  None  None  None  None  None  None
1117  None  None  None  None  None  …  None  None  None  None  None  None
1118  None  None  None  None  None  …  None  None  None  None  None  None

       455   456   457   458
0     None  None  None  None
1     None  None  None  None
2     None  None  None  None
3     None  None  None  None
4     None  None  None  None
…     …    …    …    …
1114  None  None  None  None
1115  None  None  None  None
1116  None  None  None  None
1117  None  None  None  None
1118  None  None  None  None

[1119 rows x 459 columns]
```

[ ]:
```
rrdf = pd.DataFrame()

for i in range(0,t.shape[0]):
    for j in range(0,t.shape[1]):
        try:
            t1 = pd.json_normalize(t.loc[i][j])
            rrdf = pd.concat([rrdf, t1])
        except:
            pass
```

[97]:
```
rrdf.head()
```

[97]:
```
        barcode                                        description finalPrice  \
0          4011                                      ITEM NOT FOUND      26.00
0          4011                                      ITEM NOT FOUND          1
0  028400642255  DORITOS TORTILLA CHIP SPICY SWEET CHILI REDUCE…      10.00
0           NaN                                                NaN        NaN
0          4011                                      ITEM NOT FOUND      28.00

  itemPrice needsFetchReview partnerItemId preventTargetGapPoints  \
0     26.00            False             1                    True
0         1              NaN             1                     NaN
0     10.00             True             2                    True
0       NaN            False             1                    True
0     28.00            False             1                    True
```

```
   quantityPurchased userFlaggedBarcode userFlaggedNewItem  … itemNumber  \
0                5.0               4011               True  …        NaN
0                1.0                NaN                NaN  …        NaN
0                1.0       028400642255               True  …        NaN
0                NaN               4011               True  …        NaN
0                4.0               4011               True  …        NaN

   originalMetaBriteQuantityPurchased pointsEarned targetPrice  \
0                                 NaN          NaN         NaN
0                                 NaN          NaN         NaN
0                                 NaN          NaN         NaN
0                                 NaN          NaN         NaN
0                                 NaN          NaN         NaN

  competitiveProduct originalFinalPrice originalMetaBriteItemPrice deleted  \
0                NaN                NaN                        NaN     NaN
0                NaN                NaN                        NaN     NaN
0                NaN                NaN                        NaN     NaN
0                NaN                NaN                        NaN     NaN
0                NaN                NaN                        NaN     NaN

  priceAfterCoupon metabriteCampaignId
0              NaN                NaN
0              NaN                NaN
0              NaN                NaN
0              NaN                NaN
0              NaN                NaN

[5 rows x 34 columns]
```

[98]:
```
# LOTS of missings from the rewardsReceiptItemList

for c in rrdf.columns:
    print(f'{c} null count: {round(100*rrdf[rrdf[c].isnull()].shape[0] / rrdf.
 ↪shape[0],2)}%')
```

```
barcode null count: 55.48%
description null count: 5.49%
finalPrice null count: 2.51%
itemPrice null count: 2.51%
needsFetchReview null count: 88.29%
partnerItemId null count: 0.0%
preventTargetGapPoints null count: 94.84%
quantityPurchased null count: 2.51%
userFlaggedBarcode null count: 95.14%
userFlaggedNewItem null count: 95.35%
```

```
userFlaggedPrice null count: 95.69%
userFlaggedQuantity null count: 95.69%
needsFetchReviewReason null count: 96.84%
pointsNotAwardedReason null count: 95.1%
pointsPayerId null count: 81.75%
rewardsGroup null count: 75.06%
rewardsProductPartnerId null count: 67.31%
userFlaggedDescription null count: 97.05%
originalMetaBriteBarcode null count: 98.98%
originalMetaBriteDescription null count: 99.86%
brandCode null count: 62.54%
competitorRewardsGroup null count: 96.04%
discountedItemPrice null count: 16.89%
originalReceiptItemText null count: 17.01%
itemNumber null count: 97.8%
originalMetaBriteQuantityPurchased null count: 99.78%
pointsEarned null count: 86.64%
targetPrice null count: 94.55%
competitiveProduct null count: 90.71%
originalFinalPrice null count: 99.87%
originalMetaBriteItemPrice null count: 99.87%
deleted null count: 99.87%
priceAfterCoupon null count: 86.23%
metabriteCampaignId null count: 87.57%
```

[ ]:
```
# many columns are only available for a few items.
# Makes sense for user flags, but not as much for other columns
# Also- what information is contained in this long list of columns?
# Some can be inferred, but none are documented
```

[107]:
```
bdf.barcode = bdf.barcode.astype('str')
```

[133]:
```
# Can join item data to brands on both brandCode and barcode...

bdf[bdf.barcode.notnull()].merge(rrdf[rrdf.barcode.notnull()][['barcode',␣
 ↪'brandCode']], how='inner', on='barcode')
```

[133]:

	_id	barcode	category	\
0	{'$oid': '5a8c36dbe4b0ccf165fac9e9'}	511111204206	Canned Goods & Soups	
1	{'$oid': '5a8c36dbe4b0ccf165fac9e9'}	511111204206	Canned Goods & Soups	
2	{'$oid': '5a8c36dbe4b0ccf165fac9e9'}	511111204206	Canned Goods & Soups	
3	{'$oid': '5a8c36dbe4b0ccf165fac9e9'}	511111204206	Canned Goods & Soups	
4	{'$oid': '5a8c36dbe4b0ccf165fac9e9'}	511111204206	Canned Goods & Soups	
..	…	…	…	
84	{'$oid': '5a7e0665e4b0aedb3b84afd4'}	511111704140	NaN	
85	{'$oid': '5a7e0665e4b0aedb3b84afd4'}	511111704140	NaN	
86	{'$oid': '5a7e0665e4b0aedb3b84afd4'}	511111704140	NaN	

```
87  {'$oid': '5a7e0665e4b0aedb3b84afd4'}  511111704140                  NaN
88  {'$oid': '5a7e0665e4b0aedb3b84afd4'}  511111704140                  NaN

   categoryCode                                                  cpg  \
0           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
1           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
2           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
3           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
4           NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
..          …                                                   …
84          NaN  {'$ref': 'Cogs', '$id': {'$oid': '55b62995e4b0…
85          NaN  {'$ref': 'Cogs', '$id': {'$oid': '55b62995e4b0…
86          NaN  {'$ref': 'Cogs', '$id': {'$oid': '55b62995e4b0…
87          NaN  {'$ref': 'Cogs', '$id': {'$oid': '55b62995e4b0…
88          NaN  {'$ref': 'Cogs', '$id': {'$oid': '55b62995e4b0…

              name  topBrand brandCode_x brandCode_y
0          Swanson       0.0     SWANSON     SWANSON
1          Swanson       0.0     SWANSON     SWANSON
2          Swanson       0.0     SWANSON     SWANSON
3          Swanson       0.0     SWANSON     SWANSON
4          Swanson       0.0     SWANSON     SWANSON
..             …        …          …          …
84  Diet Chris Cola       NaN  DIETCHRIS2       PREGO
85  Diet Chris Cola       NaN  DIETCHRIS2       PREGO
86  Diet Chris Cola       NaN  DIETCHRIS2       PREGO
87  Diet Chris Cola       NaN  DIETCHRIS2       PREGO
88  Diet Chris Cola       NaN  DIETCHRIS2       PREGO

[89 rows x 9 columns]
```

[131]:
```
#... but brandCode results in far more matches
# regardless, there is no way to match every item to a brand it seems with the␣
 ↪current data
# it seems we need more data to match brands to items

bdf[bdf.brandCode.notnull()].merge(rrdf[rrdf.brandCode.notnull()][['barcode',␣
 ↪'brandCode']], how='inner', on='brandCode')
```

[131]:
```
                                        _id     barcode_x              category  \
0    {'$oid': '57ebc2e7e4b0ac389136a34b'}  511111201915               Grocery
1    {'$oid': '5bd200a6965c7d66d92731ea'}  511111504627             Household
2    {'$oid': '5a8c36dbe4b0ccf165fac9e9'}  511111204206  Canned Goods & Soups
3    {'$oid': '5a8c36dbe4b0ccf165fac9e9'}  511111204206  Canned Goods & Soups
4    {'$oid': '5a8c36dbe4b0ccf165fac9e9'}  511111204206  Canned Goods & Soups
..                                    …           …                     …
630  {'$oid': '5bd2013f965c7d66d92731ec'}  511111904663             Household
```

```
631  {'$oid': '5bd2013f965c7d66d92731ec'}  511111904663            Household
632  {'$oid': '5bd201f090fa074576779a1a'}  511111204718            Household
633  {'$oid': '57d96112e4b0ac389136a2b8'}  511111102335               Frozen
634  {'$oid': '5887a290e4b02187f85cdad7'}  511111701132                  NaN

    categoryCode                                                    cpg  \
0            NaN  {'$ref': 'Cogs', '$id': {'$oid': '559c2234e4b0…
1            NaN  {'$ref': 'Cogs', '$id': {'$oid': '550b2565e4b0…
2            NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
3            NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
4            NaN  {'$ref': 'Cogs', '$id': {'$oid': '5a734034e4b0…
..           …                                                    …
630          NaN  {'$ref': 'Cogs', '$id': {'$oid': '550b2565e4b0…
631          NaN  {'$ref': 'Cogs', '$id': {'$oid': '550b2565e4b0…
632          NaN  {'$ref': 'Cogs', '$id': {'$oid': '550b2565e4b0…
633          NaN  {'$ref': 'Cogs', '$id': {'$oid': '559c2234e4b0…
634          NaN  {'$ref': 'Cogs', '$id': {'$oid': '559c2234e4b0…

           name  topBrand   brandCode     barcode_y
0     Taco Bell       0.0   TACO BELL  021000039340
1    Cottonelle       0.0  COTTONELLE  036000478044
2       Swanson       0.0     SWANSON  511111204206
3       Swanson       0.0     SWANSON  511111204206
4       Swanson       0.0     SWANSON  511111204206
..          …       …          …           …
630     Kleenex       0.0     KLEENEX  036000391718
631     Kleenex       0.0     KLEENEX  036000391718
632        Viva       0.0        VIVA  036000494129
633     Ore-Ida       0.0     ORE-IDA  013120002588
634   Stove Top       0.0   STOVE TOP  043000285213

[635 rows x 9 columns]
```

[ ]:
```
# It seems that there might be a missing schema here that would provide more␣
 ↪clarity
# There are references to a "partner product file", and corresponding columns␣
 ↪like:
# partnerItemId
# rewardsProductPartnerId
# perhaps this missing schema contains the "translation" needed to match␣
 ↪receipts to brands
# Some brands are grouped into CPGs as well, but it's not clear how
```