Fetch Rewards Coding Exercise - Data Analyst

Ashley Fairley
May 2021

Before reading the files into R for exploratory analysis, I used a couple lines of python to convert the json files:

```
import pandas as pd
import json
df = pd.read_json('PATH/receipts.json', lines = True)
df.to_excel('PATH/receipts.xlsx')
```

```
options(scipen = 999)
library(dlookr)
```

```
##
## Attaching package: 'dlookr'
```

```
## The following object is masked from 'package:base':
##
##     transform
```

```
receipts <- read.csv("receipts.csv")
users <- read.csv("users.csv")
brands <- read.csv("brands.csv")

str(receipts)
```

```
## 'data.frame':    1185 obs. of  15 variables:
##  $ receipt_id            : chr  "5ff1e1eb0a720f0523000575" "5ff1e1bb0a720f052300056b" "5ff1e1f10a720f05230005
7a" "5ff1eee0a7214ada100056f" ...
##  $ bonusPointsEarned     : chr  "500" "150" "5" "5" ...
##  $ bonusPointsEarnedReason: chr  "Receipt number 2 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50
e83a36)" "Receipt number 5 completed, bonus point schedule DEFAULT (5cefdcacf3693e0b50e83a36)" "All-receipts rece
ipt bonus" "All-receipts receipt bonus" ...
##  $ createDate            : chr  "1609690000000" "1609690000000" "1609690000000" "1609690000000" ...
##  $ dateScanned           : num  1609690000000 1609690000000 1609690000000 1609690000000 1609690000000 ...
##  $ finishedDate          : num  1609690000000 1609690000000 NA 1609690000000 1609690000000 ...
##  $ modifyDate            : num  1609690000000 1609690000000 1609690000000 1609690000000 1609690000000 ...
##  $ pointsAwardedDate     : num  1609690000000 1609690000000 NA 1609690000000 1609690000000 ...
##  $ pointsEarned          : num  500 150 5 5 5 750 5 500 5 250 ...
##  $ purchaseDate          : num  1609630000000 1609600000000 1609630000000 1609630000000 1609600000000 ...
##  $ purchasedItemCount    : int  5 2 1 4 2 1 1 1 5 3 ...
##  $ rewardsReceiptItemList : chr  "{'barcode': '4011', 'description': 'ITEM NOT FOUND', 'finalPrice': '26.00',
'itemPrice': '26.00', 'needsFetchRe"| __truncated__ "{'barcode': '4011', 'description': 'ITEM NOT FOUND', 'finalP
rice': '1', 'itemPrice': '1', 'partnerItemId': '1',"| __truncated__ "{'needsFetchReview': False, 'partnerItemId':
'1', 'preventTargetGapPoints': True, 'userFlaggedBarcode': '4011',"| __truncated__ "{'barcode': '4011', 'descript
ion': 'ITEM NOT FOUND', 'finalPrice': '28.00', 'itemPrice': '28.00', 'needsFetchRe"| __truncated__ ...
##  $ rewardsReceiptStatus   : chr  "FINISHED" "FINISHED" "REJECTED" "FINISHED" ...
##  $ totalSpent            : num  26 11 10 28 1 3.25 2.23 10 20 20 ...
##  $ userId                : chr  "5ff1e1eacfcf6c399c274ae6" "5ff1e194b6a9d73a3a9f1052" "5ff1e1f1cfcf6c399c274b
0b" "5ff1e1eacfcf6c399c274ae6" ...
```

```
sum(is.na(receipts))
```

```
## [1] 3560
```

```
receipts_na <- sapply(receipts, function(x) any(is.na(x)))
names(receipts)[receipts_na]
```

```
## [1] "dateScanned"        "finishedDate"       "modifyDate"
## [4] "pointsAwardedDate"  "pointsEarned"       "purchaseDate"
## [7] "purchasedItemCount" "totalSpent"
```

```
diagnose(receipts)
```

variables <chr>	types <chr>	missing_count <int>	missing_percent <dbl>	unique_count <int>	unique_rate <dbl>
receipt_id	character	0	0.00000	1120	0.945147679
bonusPointsEarned	character	0	0.00000	15	0.012658228
bonusPointsEarnedReason	character	0	0.00000	32	0.027004219
createDate	character	0	0.00000	243	0.205063291
dateScanned	numeric	66	5.56962	221	0.186497890
finishedDate	numeric	617	52.06751	84	0.070886076
modifyDate	numeric	66	5.56962	212	0.178902954
pointsAwardedDate	numeric	648	54.68354	83	0.070042194
pointsEarned	numeric	576	48.60759	120	0.101265823
purchaseDate	numeric	514	43.37553	154	0.129957806

1-10 of 15 rows Previous 1 2 Next

```
str(users)
```

```
## 'data.frame':    495 obs. of  7 variables:
##  $ user_id   : chr  "5ff1e194b6a9d73a3a9f1052" "5ff1e194b6a9d73a3a9f1052" "5ff1e194b6a9d73a3a9f1052" "5ff1e1
eacfcf6c399c274ae6" ...
##  $ active    : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ createdDate : num  1609687444800 1609687444800 1609687444800 1609687530554 1609687444800 ...
##  $ lastLogin   : num  1609687537858 1609687537858 1609687537858 1609687530597 1609687537858 ...
##  $ role      : chr  "consumer" "consumer" "consumer" "consumer" ...
##  $ signUpSource: chr  "Email" "Email" "Email" "Email" ...
##  $ state     : chr  "WI" "WI" "WI" "WI" ...
```

```
sum(is.na(users))
```

```
## [1] 62
```

```
users_na <- sapply(users, function(x) any(is.na(x)))
names(users)[users_na]
```

```
## [1] "lastLogin"
```

```
diagnose(users)
```

variables <chr>	types <chr>	missing_count <int>	missing_percent <dbl>	unique_count <int>	unique_rate <dbl>
user_id	character	0	0.00000	212	0.428282828
active	logical	0	0.00000	2	0.004040404
createdDate	numeric	0	0.00000	212	0.428282828
lastLogin	numeric	62	12.52525	173	0.349494949
role	character	0	0.00000	2	0.004040404
signUpSource	character	0	0.00000	3	0.006060606
state	character	0	0.00000	9	0.018181818

7 rows

```
str(brands)
```

```
## 'data.frame':    1167 obs. of  8 variables:
##  $ brand_id    : chr  "5332f5ebe4b03c9a25efd0a7" "5332f5f2e4b03c9a25efd0a9" "5332f5f2e4b03c9a25efd0ab" "5332f5
f3e4b03c9a25efd0ad" ...
##  $ barcode     : num  511111304050 511111804048 511111604037 511111104025 511111904014 ...
##  $ category    : chr  "" "" "" "" ...
##  $ categoryCode: chr  "5332f5ebe4b03c9a25efd0a8" "5332f5f2e4b03c9a25efd0aa" "5332f5f3e4b03c9a25efd0ac" "5332f5
f3e4b03c9a25efd0ae" ...
##  $ cpg         : chr  "Cpgs" "Cpgs" "Cpgs" "Cpgs" ...
##  $ name        : chr  "Monster" "Eggo" "Our Family" "Gree Giant" ...
##  $ topBrand    : logi  NA NA NA NA NA NA ...
##  $ brandCode   : chr  "" "" "" "" ...
```

```
sum(is.na(brands))
```

```
## [1] 612
```

```
brands_na <- sapply(brands, function(x) any(is.na(x)))
names(brands)[brands_na]
```

```
## [1] "topBrand"
```

```
diagnose(brands)
```

variables <chr>	types <chr>	missing_count <int>	missing_percent <dbl>	unique_count <int>	unique_rate <dbl>
brand_id	character	0	0.00000	1167	1.000000000
barcode	numeric	0	0.00000	1160	0.994001714
category	character	0	0.00000	24	0.020565553
categoryCode	character	0	0.00000	196	0.167952014
cpg	character	0	0.00000	2	0.001713796
name	character	0	0.00000	1156	0.990574122
topBrand	logical	612	52.44216	3	0.002570694
brandCode	character	0	0.00000	845	0.724078835

8 rows

All of the files contain missing values:

- receipts: 3560
- users: 62
- brands: 612

The Receipts file is missing values in multiple columns: “dateScanned”, “finishedDate”, “modifyDate”, “pointsAwardedDate”, “pointsEarned”, “purchaseDate”, “purchasedItemCount”, “totalSpent”

Users is missing values in one column: “lastLogin”

Brands is missing values in one column: “topBrand”

```
str(receipts$rewardsReceiptItemList)
```

```
##  chr [1:1185] "{'barcode': '4011', 'description': 'ITEM NOT FOUND', 'finalPrice': '26.00', 'itemPrice': '26.00
', 'needsFetchRe"| __truncated__ ...
```

One of the first issues I noticed was that the column, “rewardsReceiptItemList”, contains too much information in each entry to be useful and should be split up in the database and stored separately. I created a table for receipt (purchased) items in my database diagram. If possible, it may be useful to create an additional table that lists all possible items that could earn awards in addition to the receipt_items table.

```
str(users$createdDate)
```

```
##  num [1:495] 1609687444800 1609687444800 1609687444800 1609687530554 1609687444800 ...
```

Another issue would be to determine a preferred, unified way to store the dates for all tables in the database so that it is easy to read and use in queries. The json files have the dates formatted as a string of numbers. It might be best to correct how the dates are being retrieved.