Data Quality Issues

Stacy, Liu

2022-06-10

```
library(R.utils)
library(jsonlite)
library(tidyverse)
library(DescTools)
```

Brand Data

Data Input

```
brands = stream_in(file("brands.json"))
```

```
##  Found 500 records... Found 1000 records... Found 1167 records... Imported 1167 records. Simplifying
```

```
summary(brands)
```

```
##       _id.$oid        barcode            category         categoryCode
##  Length:1167        Length:1167        Length:1167        Length:1167
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##  cpg.$id.$oid             cpg.$ref          name             topBrand
##  Length:1167         Length:1167        Length:1167        Mode :logical
##  Class :character    Class :character   Class :character   FALSE:524
##  Mode  :character    Mode  :character   Mode  :character   TRUE :31
##                                                            NA's :612
##   brandCode
##  Length:1167
##  Class :character
##  Mode  :character
##
```

```
# quick review data structure
brands %>%
  as_tibble() %>%
  unnest(cols = c(`_id`)) -> brands_unnest

str(brands_unnest)
```

```
## tibble [1,167 x 8] (S3: tbl_df/tbl/data.frame)
##  $ $oid        : chr [1:1167] "601ac115be37ce2ead437551" "601c5460be37ce2ead43755f" "601ac142be37ce2e
##  $ barcode     : chr [1:1167] "511111019862" "511111519928" "511111819905" "511111519874" ...
##  $ category    : chr [1:1167] "Baking" "Beverages" "Baking" "Baking" ...
##  $ categoryCode: chr [1:1167] "BAKING" "BEVERAGES" "BAKING" "BAKING" ...
##  $ cpg         :'data.frame':    1167 obs. of  2 variables:
##   ..$ $id :'data.frame': 1167 obs. of  1 variable:
##   .. ..$ $oid: chr [1:1167] "601ac114be37ce2ead437550" "5332f5fbe4b03c9a25efd0ba" "601ac142be37ce2ea
##   ..$ $ref: chr [1:1167] "Cogs" "Cogs" "Cogs" "Cogs" ...
##  $ name        : chr [1:1167] "test brand @1612366101024" "Starbucks" "test brand @1612366146176" "t
##  $ topBrand    : logi [1:1167] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ brandCode   : chr [1:1167] NA "STARBUCKS" "TEST BRANDCODE @1612366146176" "TEST BRANDCODE @161236
```

Data Quality Issues

- Duplicate Data

I expected brand_id and barcode should be a unique value. Therefore, I evaluated these 2 columns first. Duplicate records were found in the Barcode column.

```
# 6 barcode are duplicate
brands_unnest[duplicated(brands_unnest$barcode), ]
```

```
## # A tibble: 7 x 8
##   ‘$oid‘ barcode category categoryCode cpg$‘$id‘$‘$oid‘ name  topBrand brandCode
##   <chr>  <chr>   <chr>    <chr>        <chr>            <chr> <lgl>    <chr>
## 1 5a8c3~ 511111~ Grocery  <NA>         5a734034e4b0d58~ Pace  FALSE    PACE
## 2 5ccb2~ 511111~ Condime~ <NA>         559c2234e4b06ac~ The ~ NA       PIONEER ~
## 3 5d602~ 511111~ Snacks   <NA>         5332f5fbe4b03c9~ CHES~ NA       CHESTERS
## 4 5d642~ 511111~ Magazin~ <NA>         5d5d4fd16d5f3b2~ Rach~ NA       51111130~
## 5 5c463~ 511111~ Dairy    <NA>         5c45f8b087ff355~ Bran~ TRUE     09090909~
## 6 5a7e0~ 511111~ <NA>     <NA>         55b62995e4b0d8e~ Diet~ NA       DIETCHRI~
## 7 5cdac~ 511111~ Condime~ <NA>         559c2234e4b06ac~ Bitt~ NA       BITTEN
## # ... with 1 more variable: cpg$‘$ref‘ <chr>
```

```
# Duplicate brand_id were not found
brands_unnest[duplicated(brands_unnest$`$oid`),]
```

```
## # A tibble: 0 x 8
## # ... with 8 variables: $oid <chr>, barcode <chr>, category <chr>,
## #   categoryCode <chr>, cpg <df[,2]>, name <chr>, topBrand <lgl>,
## #   brandCode <chr>
```

- Missing data

CategoryCode and topBrand column have large percentages of missing values, which are 55.7% and 52.4% respectively.

```
# dataframe overview
Abstract(brands_unnest)
```

```
## ------------------------------------------------------------------------------
## brands_unnest
##
## data frame:  1167 obs. of  8 variables
##      NA complete cases (NA)
##
##   Nr  ColName       Class       NAs          Levels
##   1   $oid          character     .
##   2   barcode       character     .
##   3   category      character   155 (13.3%)
##   4   categoryCode  character   650 (55.7%)
##   5   cpg           data.frame    .
##   6   name          character     .
##   7   topBrand      logical     612 (52.4%)
##   8   brandCode     character   234 (20.1%)
```

After a quick data overview, I recommend re-designing the way brandCode encoding is since it currently contains a mess of information in there without any encoding rule. I also found some values in brandCode are the same as the barcode.

```
# found 54 records which brandCode are the same with barcode.
brands_unnest %>%
  filter(brandCode == barcode) %>%
  select(c(1:2), brandCode)
```

```
## # A tibble: 54 x 3
##    ‘$oid‘                   barcode      brandCode
##    <chr>                    <chr>        <chr>
##  1 5d6413156d5f3b23d1bc790a 511111205012 511111205012
##  2 5d66d71fa3a018093ab34728 511111105329 511111105329
##  3 5d66d94d6d5f3b6188d4f04b 511111505365 511111505365
##  4 5da609991dda2c3e1416ae90 511111805854 511111805854
##  5 5da60576a60b87376833e349 511111305569 511111305569
##  6 5da608131dda2c3e1416ae8a 511111505716 511111505716
##  7 5d658ff3a3a018514994f432 511111005216 511111005216
##  8 5d642dbfa3a018514994f42e 511111005148 511111005148
##  9 5da6094ca60b87376833e357 511111605829 511111605829
## 10 5da608dfa60b87376833e354 511111805786 511111805786
## # ... with 44 more rows
```

Users Data

Data Input

```
users = stream_in(file("users.json"))
```

```
##  Found 495 records... Imported 495 records. Simplifying...
```

```
summary(users)
```

```
##       _id.$oid        active         createdDate
##  Length:495         Mode :logical   Min.   :2014-12-19 09:21:22
##  Class :character   FALSE:1         1st Qu.:2021-01-04 14:30:17
##  Mode  :character   TRUE :494       Median :2021-01-13 15:19:38
##                                     Mean   :2020-08-05 21:34:47
##                                     3rd Qu.:2021-01-25 12:31:59
##                                     Max.   :2021-02-12 09:11:06
##
##    lastLogin                       role           signUpSource
##  Min.   :2018-05-07 13:23:40   Length:495         Length:495
##  1st Qu.:2021-01-08 13:14:53   Class :character   Class :character
##  Median :2021-01-21 08:57:48   Mode  :character   Mode  :character
##  Mean   :2021-01-23 02:48:00
##  3rd Qu.:2021-02-03 10:34:11
##  Max.   :2021-03-05 11:52:23
##  NA's   :62
##     state
##  Length:495
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# quick review data
users %>%
  as_tibble() %>%
  unnest(cols = c(`_id`)) -> users_unnest

str(users_unnest)
```

```
## tibble [495 x 7] (S3: tbl_df/tbl/data.frame)
##  $ $oid        : chr [1:495] "5ff1e194b6a9d73a3a9f1052" "5ff1e194b6a9d73a3a9f1052" "5ff1e194b6a9d73a
##  $ active      : logi [1:495] TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ createdDate : POSIXct[1:495], format: "2021-01-03 10:24:04" "2021-01-03 10:24:04" ...
##  $ lastLogin   : POSIXct[1:495], format: "2021-01-03 10:25:37" "2021-01-03 10:25:37" ...
##  $ role        : chr [1:495] "consumer" "consumer" "consumer" "consumer" ...
##  $ signUpSource: chr [1:495] "Email" "Email" "Email" "Email" ...
##  $ state       : chr [1:495] "WI" "WI" "WI" "WI" ...
```

Data Quality Issues

- Duplicate data

I expected user ID should be a unique value. Therefore, I evaluated the user ID first. Duplicate records were found in the user ID column.

```r
# 283 records are duplicate
dim(users_unnest[duplicated(users_unnest$`$oid`), ])
```

```
## [1] 283   7
```

```r
# quick review the duplicate data
head(users_unnest[duplicated(users_unnest$`$oid`), ])
```

```
## # A tibble: 6 x 7
##   ‘$oid‘ active createdDate         lastLogin           role  signUpSource state
##   <chr>  <lgl>  <dttm>              <dttm>              <chr> <chr>        <chr>
## 1 5ff1e~ TRUE   2021-01-03 10:24:04 2021-01-03 10:25:37 cons~ Email        WI
## 2 5ff1e~ TRUE   2021-01-03 10:24:04 2021-01-03 10:25:37 cons~ Email        WI
## 3 5ff1e~ TRUE   2021-01-03 10:24:04 2021-01-03 10:25:37 cons~ Email        WI
## 4 5ff1e~ TRUE   2021-01-03 10:24:04 2021-01-03 10:25:37 cons~ Email        WI
## 5 5ff1e~ TRUE   2021-01-03 10:24:04 2021-01-03 10:25:37 cons~ Email        WI
## 6 5ff1e~ TRUE   2021-01-03 10:24:04 2021-01-03 10:25:37 cons~ Email        WI
```

```r
# duplicate user ID
unique(users_unnest[duplicated(users_unnest$`$oid`), ][1])
```

```
## # A tibble: 70 x 1
##    ‘$oid‘
##    <chr>
##  1 5ff1e194b6a9d73a3a9f1052
##  2 5ff1e1eacfcf6c399c274ae6
##  3 5ff370c562fde912123a5e0e
##  4 5ff36d0362fde912123a5535
##  5 5ff36be7135e7011bcb856d3
##  6 5ff36a3862fde912123a4460
##  7 5ff47392c3d63511e2a47881
##  8 5ff4ce33c3d63511e2a484b6
##  9 5ff4ce3dc3d63511e2a484dc
## 10 5ff5d15aeb7c7d12096d91a2
## # ... with 60 more rows
```

- Missing data

```r
# dataframe overview
Abstract(users_unnest)
```

```
## ------------------------------------------------------------------------------
## users_unnest
##
## data frame:  495 obs. of  7 variables
##      364 complete cases (73.5%)
##
##   Nr  ColName       Class            NAs         Levels
##   1   $oid          character         .
##   2   active        logical           .
##   3   createdDate   POSIXct, POSIXt   .
##   4   lastLogin     POSIXct, POSIXt  62 (12.5%)
##   5   role          character         .
##   6   signUpSource  character        48 (9.7%)
##   7   state         character        56 (11.3%)
```

If the data we are supposed to collect is from 2014 to 2021, then 2016, 2018, and 2019 user records are missing in the user's data.

```
unique(format(as.Date(users_unnest$createdDate, format="%d-%m-%Y"),"%Y"))
```

```
## [1] "2021" "2020" "2015" "2017" "2014"
```

Receipts Data

Data Input

```
receipts = stream_in(file("receipts.json"))
```

```
##  Found 500 records... Found 1000 records... Found 1119 records... Imported 1119 records. Simplifying
```

```
receipts %>%
  as_tibble() %>%
  unnest(cols = c(`_id`)) -> receipts_unnest
# quick overview data
glimpse(receipts_unnest)
```

```
## Rows: 1,119
## Columns: 15
## $ ‘$oid‘                   <chr> "5ff1e1eb0a720f0523000575", "5ff1e1bb0a720f052~
## $ bonusPointsEarned        <int> 500, 150, 5, 5, 5, 750, 5, 500, 5, 250, 100, 7~
## $ bonusPointsEarnedReason <chr> "Receipt number 2 completed, bonus point sched~
## $ createDate               <dttm> 2021-01-03 10:25:31, 2021-01-03 10:24:43, 202~
## $ dateScanned              <dttm> 2021-01-03 10:25:31, 2021-01-03 10:24:43, 202~
## $ finishedDate             <dttm> 2021-01-03 10:25:31, 2021-01-03 10:24:43, NA,~
## $ modifyDate               <dttm> 2021-01-03 10:25:36, 2021-01-03 10:24:48, 202~
## $ pointsAwardedDate        <dttm> 2021-01-03 10:25:31, 2021-01-03 10:24:43, NA,~
## $ pointsEarned             <chr> "500.0", "150.0", "5", "5.0", "5.0", "750.0", ~
## $ purchaseDate             <dttm> 2021-01-02 19:00:00, 2021-01-02 10:24:43, 202~
## $ purchasedItemCount       <int> 5, 2, 1, 4, 2, 1, 1, 1, 5, 3, 1, 5, 10, 11, 1,~
## $ rewardsReceiptItemList   <list> [<data.frame[1 x 12]>], [<data.frame[2 x 18]>~
## $ rewardsReceiptStatus     <chr> "FINISHED", "FINISHED", "REJECTED", "FINISHED"~
## $ totalSpent               <chr> "26.00", "11.00", "10.00", "28.00", "1.00", "3~
## $ userId                   <chr> "5ff1e1eacfcf6c399c274ae6", "5ff1e194b6a9d73a3~
```

Data Quality Issues

- Duplicate Observations

I expected the receipts ID should be a unique value. Therefore, I evaluated the receipts ID first. Duplicate records were not found.

```
receipts_unnest[duplicated(receipts_unnest$`$oid`), ]
```

```
## # A tibble: 0 x 15
## # ... with 15 variables: $oid <chr>, bonusPointsEarned <int>,
## #   bonusPointsEarnedReason <chr>, createDate <dttm>, dateScanned <dttm>,
## #   finishedDate <dttm>, modifyDate <dttm>, pointsAwardedDate <dttm>,
## #   pointsEarned <chr>, purchaseDate <dttm>, purchasedItemCount <int>,
## #   rewardsReceiptItemList <list>, rewardsReceiptStatus <chr>,
## #   totalSpent <chr>, userId <chr>
```

- Missing data

Receipts data has large proportion missing value, especially in bonusPointsEarned(51.4%) column and pointsAwardedDate(52.0%) column. Also, I've noted that the pointsEarned should be an integer class rather than a character class

```
# dataframe overview
Abstract(receipts_unnest)
```

```
## ------------------------------------------------------------------------------
## receipts_unnest
##
## data frame:  1119 obs. of  15 variables
##      NA complete cases (NA)
##
##   Nr  ColName                  Class            NAs          Levels
##   1   $oid                     character          .
##   2   bonusPointsEarned        integer          575 (51.4%)
##   3   bonusPointsEarnedReason  character        575 (51.4%)
##   4   createDate               POSIXct, POSIXt    .
##   5   dateScanned              POSIXct, POSIXt    .
##   6   finishedDate             POSIXct, POSIXt  551 (49.2%)
##   7   modifyDate               POSIXct, POSIXt    .
##   8   pointsAwardedDate        POSIXct, POSIXt  582 (52.0%)
##   9   pointsEarned             character        510 (45.6%)
##   10  purchaseDate             POSIXct, POSIXt  448 (40.0%)
##   11  purchasedItemCount       integer          484 (43.3%)
##   12  rewardsReceiptItemList   list               .
##   13  rewardsReceiptStatus     character          .
##   14  totalSpent               character        435 (38.9%)
##   15  userId                   character          .
```

The range of bonusPointsEarned and pointsEarned value is unreasonable

```
summary(receipts_unnest$bonusPointsEarned)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     5.0     5.0    45.0   238.9   500.0   750.0     575
```

```
summary(as.numeric(receipts_unnest$pointsEarned))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0       5     150     586     750   10200     510
```