CPG Data Description

in progress

Yifan Xu, Owen Eagen

2019/1/1

Contents

This file documents observations made during the data cleaning process for the CPG project.

1 Replicating Previous Data Cleaning Process

This section deals with the process and results of replicating the cleaning code provided by Professor Joo in the RA2/data_cleaning directory.

2 Exploring Files in the "externals" Directory

This section stores our findings when exploring the files in the "externals" directory. It mainly includes our interpretation of the variable names and the comparison we make between the private and public data.

We converted the data from sas format to csv first. The files described below can be found in the directory externals/POG/data_csv.

2.1 PID_csv.zip

2.1.1 Files in PID_csv.zip

There are in total 48 files in the folder.

There are two types of file names:

- `namexxxi` where `xxx` is the abbreviation for the product and `i` an index from `a` to `e` (a map of the identifiers to the product names is stored in `externals/POG/doc/category_desc.txt`)
- `pidxxxi` where `xxx` and `i` same as above

Only 7 products are covered in this folder, whose abbreviations are `cer, cra, frj, ptw, rfj, sdr, tti`.

In the files named with `namexxxi`, there are two columns, `NAME` and `PID`; in those named with `pidxxxi`, there are also two columns, `UPC` and `PID`.

2.2 src__csv.zip

The original src.zip is not organized very well, therefore when we converted the files in the zip file to csv, we organized the files according to their directory and subdirectory. Now the csv formatted files are stored under `externals/POG/data_csv/src_csv`. The names of the new zip files represent their original location in src.zip. For example, `src_DICT_csv.zip` contains all the non-folder files under the original `src/DICT` directory, while `src_AGG_PID_csv.zip` contains all non-folder files under the original `src/AGG/PID` directory.

2.2.1 src__DICT__csv.zip

There are in total 6 files in the zip file, namely `dict9xxx.csv`, with the `xxx` being `tna, coo, did, ana, oat, cso` respectively.

There are 115 variables in each file as follows:

```
  [1] "UPC"      "WKSTART"  "WKSTOP"   "WKSALES"  "MKTSHARE" "IRCAT"    "VOLFAC"   "DESC1"
  [9] "DESC2"    "DESC3"    "DESC4"    "DESC5"    "DESC6"    "DESC7"    "DESC8"    "DESC9"
 [17] "EUPC"     "VOLUME"   "UPCSYS"   "UPCGEN"   "PRTFLAG"  "BPRTFLAG" "SVOLUME"  "PKGBONUS"
 [25] "STBRECIP" "CHVOL"    "STLETTER" "IND2"     "STBUNIT"  "NUMBVOL"  "SUBSTVOL" "VOLSLASH"
 [33] "SIZECHEC" "VOLPLUS"  "STBDASH"  "SCALE"    "VOLSIZE"  "ATTRIBC"  "GEN"      "RECIPE"
 [41] "BRAND"    "PARENT"   "VENDOR"   "KEYCAT"   "CATDES"   "BMB"      "BONUS"    "BMB2"
 [49] "BONUS2"   "SIZE1U"   "SIZE2U"   "SIZE3U"   "SIZE4U"   "SIZE5U"   "SIZE6U"   "ATTRIB1"
 [57] "ATTRIB2"  "ATTRIB3"  "ATTRIB4"  "ATTRIB5"  "ATTRIB6"  "SIZE1"    "SIZE2"    "SIZE3"
 [65] "SIZE4"    "SIZE5"    "SIZE6"    "COM_CODE" "DESCRIP"  "SIZE"     "CASE"     "NITEM"
 [73] "ATTRIB7"  "ATTRIB8"  "RETNUM"   "RETDES"   "PLAN"     "WSTART"   "BONUSMB"  "BONUSMB2"
 [81] "SIZE7"    "SIZE7U"   "DCAT"     "WKMOVE"   "UPCSIZE"  "UPCDESC"  "CHUPC"    "UPCUNIT"
 [89] "UPCSIZE2" "NEWSIZE"  "DUPLICAT" "SUBCAT"   "SUBTYPE"  "STDSIZE"  "PROBLEM"  "PROBLEM1"
 [97] "SIZEPROB" "MISMATCH" "SIZE7YES" "SIZFOUND" "CHSIZE"   "LETTSIZE" "IND"      "SIZEUNIT"
[105] "NUMBSIZE" "SUBSIZE1" "SIZSLASH" "SIZEPLUS" "SIZEDASH" "SIZSCALE" "SIZESIZE" "SIZPROB2"
[113] "MSMATCH2" "PROBLEM2" "PROBLEM3"
```

2.2.2 src__AGG__PID__csv.zip

There are 6 files in the zip file, namely `pidxxxa.csv`, with the `xxx` being `ana, did, cso, tna, oat, coo` respectively, which is the same as those in `src_DICT_csv.zip`.

In `pidanaa.csv, pidcsoa.csv, pidtnaa.csv, pidoata.csv`, there are 3 variables, namely `UPC, PID, MKTSHARE`.

In `piddida.csv`, there is an extra variable, `RECIPE`; in `pidcooa.csv`, there are 4 more variables, `RECIPE, DESCRIP, TRY, DUPLIC`. `RECIPE` contains product descriptions in abbreviations as well as unit size, while `DESCRIP` contains more readable descriptions. `TRY` and `DUPLIC` only appear in `pidcooa.csv`, with `TRY` taking either `0` or `1` and `DUPLIC` taking `0` only.

2.2.3 src__AGG__old__csv.zip

There are in total 48 files in the zip file, centering 6 products in total. Product abbreviations that appear in the zip file are `cso, did, coo, oat, tna, ana`. For each, there are 8 files, namely `afxxxa, axxxa, avxxxa, afvxxxa, afcxxxa, acxxxa, afcvxxxa, acvxxxa`.

Variables in the files are as follows:

afxxxa, axxxa

```
[1] "NITEM"  "STORE"  "WEEK"   "SALES"  "MOVE"   "NSALE"  "LPRICE" "PROFIT"
```

From some samples, it seems although afxxxa and acsoa have the same variable names, the data are mildly different for the same STORE-WEEK-SALES-MOVE combination.

avxxxa

```
 [1] "STORE"    "WEEK"     "SALES1"   "SALES2"   "SALES3"   "SALES4"   "SALES5"   "SALES6"
 [9] "SALES7"   "SALES8"   "SALES9"   "SALES10"  "SALES11"  "SALES12"  "MOVE1"    "MOVE2"
[17] "MOVE3"    "MOVE4"    "MOVE5"    "MOVE6"    "MOVE7"    "MOVE8"    "MOVE9"    "MOVE10"
[25] "MOVE11"   "MOVE12"   "NSALE1"   "NSALE2"   "NSALE3"   "NSALE4"   "NSALE5"   "NSALE6"
[33] "NSALE7"   "NSALE8"   "NSALE9"   "NSALE10"  "NSALE11"  "NSALE12"  "LPRICE1"  "LPRICE2"
[41] "LPRICE3"  "LPRICE4"  "LPRICE5"  "LPRICE6"  "LPRICE7"  "LPRICE8"  "LPRICE9"  "LPRICE10"
[49] "LPRICE11" "LPRICE12" "PROFIT1"  "PROFIT2"  "PROFIT3"  "PROFIT4"  "PROFIT5"  "PROFIT6"
[57] "PROFIT7"  "PROFIT8"  "PROFIT9"  "PROFIT10" "PROFIT11" "PROFIT12"
```

SALES1, MOVE1, NSALE1, LPRICE1, PROFIT1 are identical to SALES, MOVE, NSALE, LPRICE, PROFIT in axxxa.

afvxxxa

Variable nmaes same as in avxxxa, but SALES1, MOVE1, NSALE1, LPRICE1, PROFIT1 now correspond to SALES, MOVE, NSALE, LPRICE, PROFIT in afxxxa.

afcxxxa

```
[1] "NITEM"  "WEEK"   "SALES"  "MOVE"   "NSALE"  "LPRICE" "PROFIT"
```

Basically everything in afxxxa except "STORE".

acxxxa

Variable names same as in afcxxxa.

afcvxxxa

Variable names same as in avxxxa excpet "STORE".

acvxxxa

Variable names same as in afcvxxxa. Data differ from afcvxxxa on WEEK-MOVE-SALE level.

2.2.4 src_AGG_csv.zip

The file names are exactly the same as in src_AGG_old_csv.zip. I suppose they are mostly the same, but have not checked yet.

2.3 dff_csv.zip

DFF, Dominick's Finer Foods, was a leading supermarket chain in Chicago.

2.3.1 Files in dff_csv.zip

There are in total 72 files in the folder. There are

- upcxxx (29) w/ xxx being fsf, coo, tbr, bjc, tti, sha, frd, fre, sdr, tna, lnd, rfj, tpa, sna, bat, soa, cra, cer, cig, gro, cso, che, oat, fec, frj, ber, ana, ptw, did

- `wxxxsh` (12) w/ `xxx` being `cso, cer, ptw, did, tpa, rfj, bjc, tna, ana, tti, ora, tbr,`
- `wxxx` (29) w/ `xxx` being `cer, cig, cso, che, gro, oat, fec, frj, ber, ana, ptw, did, fsf, bjc, tbr, coo, tti, frd, fre, sdr, tna, lnd, tpa, sna, rfj, bat, soa, cra, sha`
- `wanavsh` (1)
- `wanabsh` (1)

The `upcxxx` files contain UPC/product information, the `wxxxsh` files contain shelf/planogram information, and the `wxxx` files contain movement/sales information.

The following product categories have corresponding planogram files:

- `cso`: Canned Soups
- `cer`: RTE Cereal
- `ptw`: Paper Towels
- `did`: Dish Detergent
- `tpa`: Toothpaste
- `rfj`: Refrigerated Juice
- `bjc`: Bottled Juice
- `tna`: Canned Tuna
- `ana`: Analgesics
- `tti`: Toilet Paper
- `ora`: ???
- `tbr`: Tooth Brushes

2.3.2 Variable names

A detailed description of the variable names can be found in `externals/POG/doc/struct_dff.txt`.

2.4 IRI_csv.zip

2.4.1 A bit about IRI

IRI, Information Resources Inc, is a market research company founded in Chicago in 1979, and was acquired by Symphony Technology Group in 2003.[1] IRI developed Apollo system, which provides desktop-based solutions that cover category management process, including assortment management and on-shelf planogram (IRI is mentioned on page 304 in Dreze's paper).

2.4.2 Files in IRI_csv.zip

There are in total 28 files in the folder. Names of all files have 6 characters, starting with "acv." The last 3 letters are the abbreviation for the product documented.

2.4.3 Variable Names

```
[1] "UPC"      "WEEK"     "DACVFD"   "DACVF"    "DACVD"    "DACVP"    "DINCREM"  "DACVFA"
[9] "DACVFB"   "DKEYCAT"  "MINCREM"  "DCINCREM" "MCINCREM" "DBVOL"    "DBP"      "DBF"
[17] "DBFD"     "DVNP"     "DBVNP"    "DBD"      "BD"       "DAVGFD"   "DAVGF"    "DAVGD"
[25] "DAVGP"    "DWTAVG"
```

We have not yet figured out the exact meanings of the variables yet. There are some descriptions in `externals/POG/doc/struct_iri.txt` but they are not clear enough. We tried grouping them according to observed patterns and made some guesses based on the patterns.

[1] Source: Wikipedia

2.4.3.1 Grouping of variable names

- DACVF, DACVP, DACVD
  - ACVFB, DACVFD, DACVFA
- DINCREM, MINCREAM, DCINCREM, MCINCREM
- DBVOL, DBP, DBF, DBD, DBFD
- DVNP, DBVNP
- BD
- DAVGFD, DAVGF, DAVGD, DAVGP, DWTAVG

2.4.3.2 Guesses of meanings of patterns

- D-: Daily? Dominick's?
- -AC-: Accumulated? Average cost?
- -ACV-:
- V: Volume
- -AVG-: Average
- WTAVG: Weighted average
- -P: Price
- -BP: Bundle price?
- -D: Deal
- -F: Feature
- -INCREM: Increment
- -C-: Cumulative? Cost?
- M-: Monthly

2.5 OMNI_csv.zip

Omni Resources is a technology consulting firm found in 1984.

2.5.1 Files in OMNI_csv.zip

There are in total 54 files in the folder, the names of half of which start with ow and the other half oupc. The last 3 letters of the file names, as in IRI.zip, are still abbreviations of the product documented. Files starting with oupc contain information of products while those starting with ow contain movement records.

2.5.2 Variable Names

Variable names for oupc-

```
[1] "COM_CODE"  "NITEM"  "UPC"  "DESCRIP"  "SIZE"  "CASE"
```

These variables are consistent with those in the previously cleaned data and can therefore be cleaned/handcoded in a similar way. They can be matched to variables in the product_char files as follows:

- COM_CODE: com_code
- NITEM: dominick_id
- UPC: UPC
- DESCRIP: description
- SIZE: package_size
- CASE: boxsize_seller

Variable names for ow-

```
[1] "STORE"  "WEEK"   "UPC"   "MOVE"   "QTY"  "PRICE"  "SALE"
[9] "PROFIT"  "OK"
```

These variables are also consistent with those in the previously cleaned data and can therefore be cleaned in a similar way.

2.6 OMNIIRI_csv.zip

There are in total 23 files in the folder, whose names follow the pattern of `oacvxxx`, with abbreviations of product names as the last three letters. The variable names are the same with those in `IRI.zip`.

3 Replicating results from Dreze et al. paper

3.1 Category-level results

3.1.1 Methodology

As noted in the paper, "For each product category, we compared average weekly sales during the test period to sales in a pre-experimental baseline period. All sales, regular and promotional, were included in our performance measures. Baselines were computed over historical periods spanning 86 to 99 weeks depending on the category. Each experimental period lasted 16 weeks."

3.1.2 Space-to-movement

According to the paper, there should be 30 stores receiving space-to-movement planograms and 30 received a control one.

3.1.2.1 Example: dish detergents (did)

We used dish detergents as an example. After unzipping `wdid.csv` and `wdidsh.csv`, it is worth noticing that the former is 172.4 MB, whereas the latter is only 7.6 MB.

1. Our first guess would be `wdidsh.csv` contains experimental data, whereas `wdid.csv` is the source of control. In order to do some sanity checks, we looked at the weeks and stores covered in each dataset. Our findings are as follows:
    - `wdid.csv` covers 393 unique weeks, which is much larger than the number of weeks used as baseline in the paper, which is 86-99. `wdidsh.csv` covers 33 unique weeks, which also seems larger than the 16-week experimental periods. Every week number that appears in `wdidsh.csv` is covered in `wdid.csv`.
    - Number of unique stores covered, on the other hand, appears much more consistent with the paper. In `wdid.csv`, there are 93 unique stores, and in `wdidsh.csv` there are 58, which is close to the 60 claimed in the paper (30 experiment + 30 control). Every store number that appears in `wdidsh.csv` is covered in `wdid.csv`.
2. With the two observations, our first assumption has been proved wrong. However, given the consistency between the number of stores in the `wdidsh.csv` and that is claimed in paper, we updated our assumption. It is possible that the actual experiment and control periods for each store are different, with the length of the periods being the same. So we need to check the weeks covered on a store level in both datasets. Our findings are as follows:

- In `wdidsh`, after filtering by the condition `EXPER == 1`, there are 12-15 unique weeks of data per store, which is close to the 16 weeks of experiment as claimed.
- We also observed that after the filtering above, week numbers in `wdidsh` range from 92 to 106 (w/o filtering, the range is 74 to 106).
- We tried filtering `wdid` as well using different conditions such as `MOVE != 0` and `OK = 1`, but it seems like they have little effect on the number of weeks covered.
- But with the second point, we are able to make a new guess, that is, the historical reference weeks are simply the ones with number smaller than 92 (this number might change on a store level, but should be roughly the same). Since `wdid` contains data for each store in almost every single week from week 1 to week 399, there are data for the selected stores in weeks prior to week 92(ish), which is consistent with 86-99 weeks of historical data claimed.