Nielsen Data Build

1. RMS

Objective

Here is how we would like to rebuild the Nielsen Retail Scanner Data. In the original Nielsen extracts, the movement files are in the hierarchy of **Year - Product Group - Product Module.tsv**. We would like the structure to be **Product Module - UPC / All Others.RData** instead. Note that some UPCs have switched modules over the years. Here is how we have dealt with them: For example suppose UPC 104829 appeared in module 1042 from 2006 to 2010 and in module 1043 from 2011 to 2014. Then there will be separate records of the UPC in two modules; the movement data from 2006 to 2010 will be stored in module 1042, and the rest will be stored in module 1043.

In the process of rebuilding the movement data, we would also like to produce a meta data table which should include some summary statistics for each UPC such as the revenue, the date it is first/last observed, and the number of stores/retail chains it has ever appeared in each year. The original movement files also do not have upc versions column, so we should also add that information as well. Finally, the original movement files have date in "YYYYMMDD" which are usually read in as plain number format, and it would be more desirable to convert them into the Date format in R for future use.

Process

1) Building Meta Data

The process of building the meta data is pretty simple. We obtain a list of product modules (and groups to locate where the module files are) by searching through all the files in the `nielsen_extracts` folder. Then for each module, we read in the movement files for each year and calculate the summary statistics for each upc. We produce the summary file for each year using parallelization, and after we have obtained the separate summary files for all years, we stack them and append upc version and product module code.

2) Building Movement Data

Using the meta data produced in step 1), we can calculate the rank in terms of the revenue a upc made by year. Then, we calculate the highest rank for each upc and choose upcs that have ever made top 75,000. Next, we loop over the modules and years, reading in and stacking the movement data. Then for upc's that are in the list of top upc's, we separate out the rows and store them as a separate file with name "UPC.RData". The rest goes into the "All-Other.RData" file. This part can be memory-consuming as we are stacking all the years.

3) UPC Version Correction & Base Price Imputation

In Nielsen data, a UPC can be reassigned to an entirely different product, or to a similar product with different size or packaging. However, there are cases where different versions of UPC represent exactly the same product. Hence, we correct those cases in our RMS data build. Details for UPC version correction can be found in "UPC-Version-Correction" folder and the related documentation there.

Moreover, RMS data are recorded when the product is actually sold. This can induce repeatedly missing price records for products with low sales volume. To deal with the problem, we follow base price imputation algorithm in Hitsch et al. 2019. Scripts and documentation regarding the base price imputation can be found in "Base-Price-Imputation" folder.

2. Homescan

Objective

Unlike the RMS data, the hierarchy of this dataset is relatively simple; there are three main files annually: `panelists`, `trips`, and `purchases`. So the main purpose of this build is to stack up the data for all the years starting from 2004 (Note that the RMS data starts from 2006). We also relabel some columns in the `panelists` file and append some additional information to the `purchases` file.

Some Notes on the Extracts File

There are a few things to note:

- Many demographic variables in the `panelists` file use numerical codes, the codebook of which is in the data manual. You can easily adapt the script written by James Sams to turn them into factors.
- Some columns such as `fips_state_descr` and `fips_county_descr` in the `panelists` file are read in as plain characters and was forced into factors.
- The household income in a year refers to the income made 2 years prior to it is listed. For example, the household income in panel year 2014 refers to the income made by the end of 2012.
- The column `method_of_payment_cd` column in the `trips` file appears only after 2013.
- The column `deal_type` was dropped from `purchases` file for all the years in 2013.
- There is a special category of items called the *Magnet* products, which don't have standard UPCs. These appear in the product module 0750 starting 2007. From 2011, product modules 0445-0468 also has Magnet products in them.
- `Products`, `Products_Extra`, and `Brand_Variations` files can be shared with the RMS data.

Process

1) Panelists File

This file is produced by stacking up the panelists file for each year and keying with `household_code` and `panel_year`. Note that, now variable labels are capitalized in the HMS extract file, for example, `household_size` in the raw data is now `Household_Size`. We should change the label accordingly.

2) Trips File

This file is produced in the same way as the panelists file with keys set to `household_code`, `retailer_code`, and `purchase_date`. Note that the `deal_type` variable was dropped from the Nielsen extracts file in 2013 and no longer exists for years prior to 2013 also.

3) Purchases File

This file is produced by stacking the purchases file for each year and merging with the `trips` file to add some relevant columns (like `household_code` and `retailer_code`, etc.). It also contain purchase data of the Magnet UPCs unlike in the original build.

References

Hitsch, Günter J., Ali Hortaçsu, and Xiliang Lin. 2019. “Prices and Promotions in U.S. Retail Markets: Evidence from Big Data.” *Manuscript*.