Data Cleansing & Processing Documentation

Ling Dai

dailing616@qq.com

Manual Preparation:

- **Download '2014-daylevel' and '2015-daylevel' folders.**
- Delete the 'MISC Items' tab and 'MISC Items Uniq and Sorted' tab of '1-13-15__Law_Jan13.xlsx'. (pivot tables)
- Delete the last two tabs of '2015-daylevel/2_Harris DONE/1-13-15__Harris_Jan13.xlsx' (pivot tables)
- Delete the first row of sheet 3 of '1-13-15__Harris_Jan13.xlsx'.
- Delete the first row of sheet 3 of '1-13-15__Law_Jan13.xlsx'.
- For '2015-daylevel/4_missing txndtls allcafes done/Law DONE/Mar 10th.xlsx', add Check Total (9.99)
- Add check total (5.42) for check 760 of '2015-daylevel/4_missing txndtls allcafes done/HarrisOld DONE/Jan 14th.xlsx'
- Readjust the space for Check 573 and Check 601 of '2015-daylevel/4_missing txndtls allcafes done/Gordon DONE/Jan 6th.xlsx'
- Manually correct check 888 and 987 of '2015-daylevel/4_missing txndtls allcafes done/Gordon DONE/Jan 7th.xlsx'
- Readjust space for check 982 of '2014-daylevel/4_missing txndtls allcafes/Gordon_Missing_2014/Dec18.xlsx'
- Add Check Total for Check 230, 702, 854, 990 of 'GCIS Data Jan07 - Jan28 2014 – MULTIPLEDAYS.xlsx'
- Add Subtotal, Tax, and Check Total for check 632 (Jan 23) of 'GCIS Data Jan07 - Jan28 2014 – MULTIPLEDAYS.xlsx'
- A Lot of Corrupted Data in 1/6/2015 and 1/7/2015 of GCIS:
    - Change the price of 'Breakfast Crossa' from 4.85 to 4.95.
    - For check 686, change the price of 'twix reg sz' from 1.19 to 1.09.
    - For check 743, change price of 'BOAR MTO deli' from 1.89 to 1.49.
    - For check 913 of 1/7/2015, change the price of 'fountain med' from 1.89 to 1.79.
    - For check 957, change the price of 'fresh fruit' from 1.99 to 0.99 and the price of 'gatr' from 2.09 to 2.29.
- **After manual preparation, there are still 11 checks of which the Subtotal value does not match the sum of item prices. The discrepancy in prices are mostly due to the incorrectly recorded prices of miscellaneous items. However, because the differences are not large, these checks are not likely to have a significant influence on the analysis results and are therefore included in the analysis for completeness.**

Text Parsing for All Files

- The function *parse_all_files()* can handle all three different formats.
- **Put all directories in the list *test_dirs* before running the code. It should take several minutes to parse all the files.**
- The resulting check-level dataframe should contain 138,829 observations.

Data Processing and Cleaning

- Clean date format (to 'yyyy-mm-dd') and create a variable 'DateClean'.
- **Create Unique ID (UID) using a concatenated string of 'DateClean' + 'cafe' + 'Check#' + 'Subtotal'. Subtotal is used because in some rare cases a file would contain who different checks with the same check number.**
- **Drop duplicates based on UID. There are 132,467 remaining observations.**
- Calculate number of items and create the variable *'#Items*'.
- Code week, week of quarter, and day of week.
- Code intervention based on week.
- Lookup total calories.
- Clean time and calculate hour.
- **Impute calorie content for miscellaneous items:**
    - Create an indicator variable 'Misc_Ind' to label all observations that include at least a miscellaneous item.
    - For all the observations, calculate both subtotal and total calorie content after excluding miscellaneous items.
    - Impute with linear regression model (without an intercept so that the predicted value is non-negative), with 'subtotal after excluding miscellaneous items' of all the observations as train_x and 'total calories after excluding miscellaneous items' of all the observations as train_y.
    - **The resulting slope should be approximately 69.14.**
- Label high/low-calorie bottled drinks:
    - Label all the bottled drinks in the item database.
    - Label bottled drinks with calorie content <= 20 kcal as low-calorie bottled drinks, and those with calorie content > 20 kcal as high-calorie bottled drinks.
    - Calculate the number of low-calorie bottled drinks and high-calorie bottled drinks for each observation.
- **Save the data**
    - Save the data to pkl format to keep the correct format of the vectors.
    - Save the data to csv format for subsequent analysis in R.

Aggregate the Data to Date-Cafe Level for Subsequent Analysis

- **Sampling Criteria:**
    - **Include observations with '#Items' <= 5.**
    - **Include observations with 'Subtotal' > 0.**
    - **Note: because the first sampling criterion is relatively arbitrary, a sensitivity analysis should be performed.**
- Aggregate the check-level dataframe to a date-café level dataframe, grouping by '*DateClean*' and '*cafe*'.
- Code date of week, week of quarter, and intervention.
- Calculate calorie per dollar for each observation.
- Save the data to csv format.