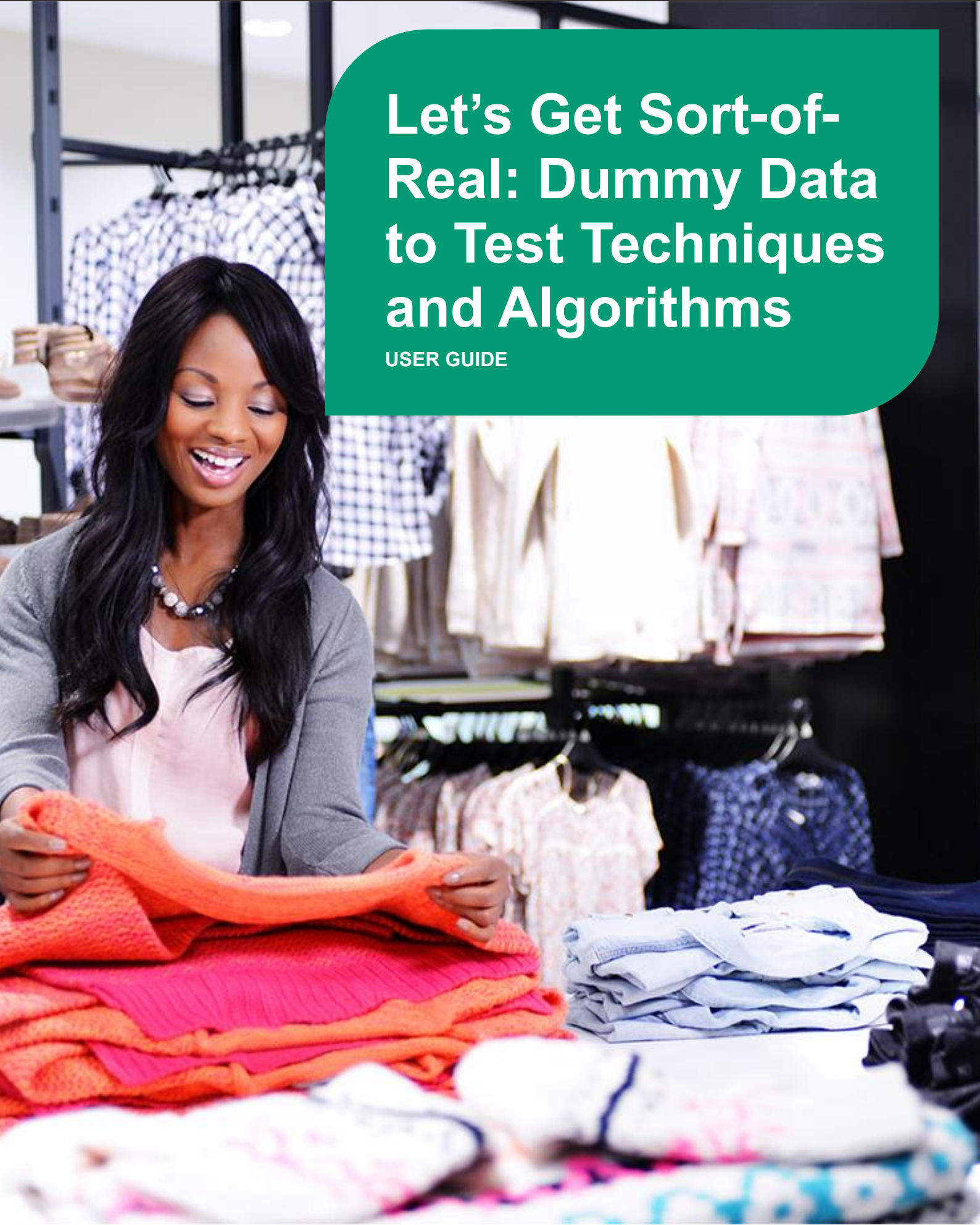


Let's Get Sort-of-Real: Dummy Data to Test Techniques and Algorithms

USER GUIDE



LET'S GET SORT-OF-REAL: DUMMY DATA TO TEST TECHNIQUES AND ALGORITHMS

This dataset includes dummy data(*) of transactions at till, spanning over a period of 117 weeks (two years and a quarter). The dataset includes both transactions with a loyalty card associated and transactions made without a loyalty card.

This dataset is made available to enable data scientists to experiment algorithms and techniques on top of sort-of-real data with a considerable size, which albeit not being real still contain real patterns and correlations. We've made a significant effort to replicate the typical patterns found in real in-store sales data to enable curious data scientists to test their techniques and algorithms using considerably expansive, sort-of-real data.

Customers' and baskets' information is also included (such as basket size, basket price sensitivity, basket dominant mission, customer price sensitivity, customer lifestage) which makes the data interesting for analysis to be run on.

BY THE NUMBERS

- 4.12 GB : total size of compressed data (split in 9 files, each between 450 and 500 MB)
 - 40.7 GB : total size of data, when uncompressed
 - 117: Weeks of transactions at till dummy data
 - ~300M: total number of transactions
 - ~47M: total number of baskets
- 400,000: Average number of baskets per week
 - 2.6M: Average number of transactions per week
 - ~500,000: Distinct number of customers
 - ~5,000: Distinct number of products
 - ~760: Distinct number of stores

SAMPLES FROM THE FULL DATASET

In order to give the option to work with smaller sets of data, some samples have been created and made available:

- A sample of 2,000 baskets randomly selected over a period of two weeks
- A sample of all transactions (for the whole 117 weeks period) for a randomly selected sample of 5,000 customers. A randomly selected sample, with a consistent size, of baskets without loyalty card has been added.
- A sample of all transactions (for the whole 117 weeks period) for a randomly selected sample of 50,000 customers. A randomly selected sample, with a consistent size, of baskets without loyalty card has been added.
- Also each of the nine zip files the full dataset is split into can be used as an independent dataset which includes all baskets in a period of 13 weeks (a quarter)

LET'S GET SORT-OF-REAL: DATASET DETAILS

All datasets (the full one and also the samples created from it) have been split in weekly files to be more manageable and to give more flexibility when loading the data.

Each row in the files corresponds to one unique product in a basket (e.g. if there are three occurrences of the same product in a basket, the file has one row for the product in that basket, with quantity equal to three). Each file has the following structure:

| Column name | Description | Type | Sample values |
|------------------------|---|------|---|
| shop_week | Identifies the week of the basket | Char | Format is YYYYWW where the first 4 characters identify the fiscal year and the other two characters identify the specific week within the year (e.g. 200735). Being the fiscal year, the first week doesn't start in January. (See time.csv file for start/end dates of each week) |
| shop_date | Date when shopping has been made. Date is specified in the yyymmdd format | Char | 20060413, 20060412 |
| shop_weekday | Identifies the day of the week | Num | 1=Sunday, 2=Monday, ..., 7=Saturday |
| shop_hour | Hour slot of the shopping | Num | 0=00:00-00:59, 1=01:00-01:59, ...23=23:00-23:59 |
| Quantity | Number of items of the same product bought in this basket | Num | Integer number |
| spend | Spend associated to the items bought | Num | Number with two decimal digits |
| prod_code | Product Code | Char | PRD0900001, PRD0900003 |
| prod_code_10 | Product Hierarchy Level 10 Code | Char | CL00072, CL00144 |
| prod_code_20 | Product Hierarchy Level 20 Code | Char | DEP00021, DEP00051 |
| prod_code_30 | Product Hierarchy Level 30 Code | Char | G00007, G00015 |
| prod_code_40 | Product Hierarchy Level 40 Code | Char | D00002, D00003 |
| cust_code | Customer Code | Char | CUST0000001624, CUST0000001912 |
| cust_price_sensitivity | Customer's Price Sensitivity | Char | LA=Less Affluent, MM=Mid Market, UM=Up Market, XX=unclassified |

| Column name | Description | Type | Sample values |
|--------------------------|--|------|--|
| cust_lifestage | Customer's Lifestage | Char | YA=Young Adults, OA=Older Adults, YF=Young Families, OF=Older Families, PE=Pensioners, OT=Other, XX=unclassified |
| basket_id | Basket ID. All items in a basket share the same basket_id value. | Num | 994100100000020, 994100100000344 |
| basket_size | Basket size | Char | L=Large, M=Medium, S=Small |
| basket_price_sensitivity | Basket price sensitivity | Char | LA=Less Affluent, MM=Mid Market, UM=Up Market, XX=unclassified |
| basket_type | Basket type | Char | Small Shop, Top Up, Full Shop, XX |
| basket_dominant_mission | Shopping dominant mission | Char | Fresh, Grocery, Mixed, Non Food, XX |
| store_code | Store Code | Char | STORE00001, STORE00002 |
| store_format | Format of the Store | Char | LS, MS, SS, XLS |
| store_region | Region the store belongs to | Char | E02, W01, E01, N03 |

The TIME table


This table contains information regarding the time periods (weeks). Each row in the file corresponds to one week. Please note that the periods are referring to fiscal years. This means, for instance, that the first week of the year doesn't fall in January.

The file has the following structure:

| Column name | Description | Type | Sample values |
|-------------|---|------|--|
| shop_week | Week code | Char | Format is YYYYWW where the first 4 characters identify the year and the other two characters identify the specific week within the fiscal year (e.g. 200735) |
| date_from | Start date for the week. Dates are specified in yyyyymmdd format | Char | 20060413, 20060412 |
| date_to | End date for the week. Dates are specified in yyyyymmdd format | Char | 20060413, 20060412 |



CONTACT INFORMATION



For general questions about dunnhumby or the Source Files programme, or for technical questions regarding the use of this dataset, please contact:

DUNNHUMBY SOURCE FILES SUPPORT

E: sourcefiles@dunnhumby.com