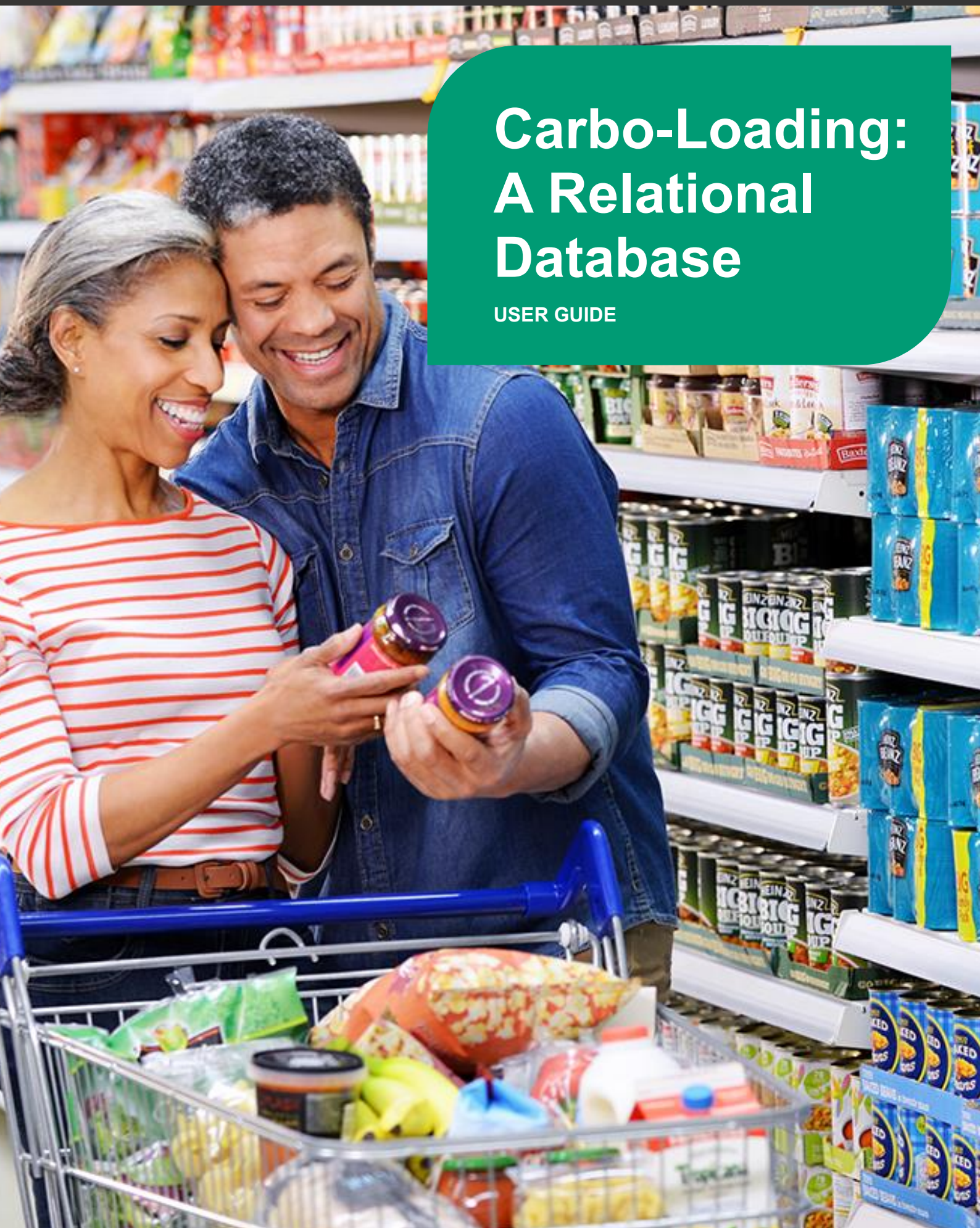
dunnhumby | SOURCE FILES



Carbo-Loading: A Relational Database

USER GUIDE

CARBO-LOADING: A RELATIONAL DATABASE

Carbo-Loading contains household level transactions over a period of two years from four categories: Pasta, Pasta Sauce, Syrup, and Pancake Mix. These categories were chosen so that interactions between the categories can be detected and studied.

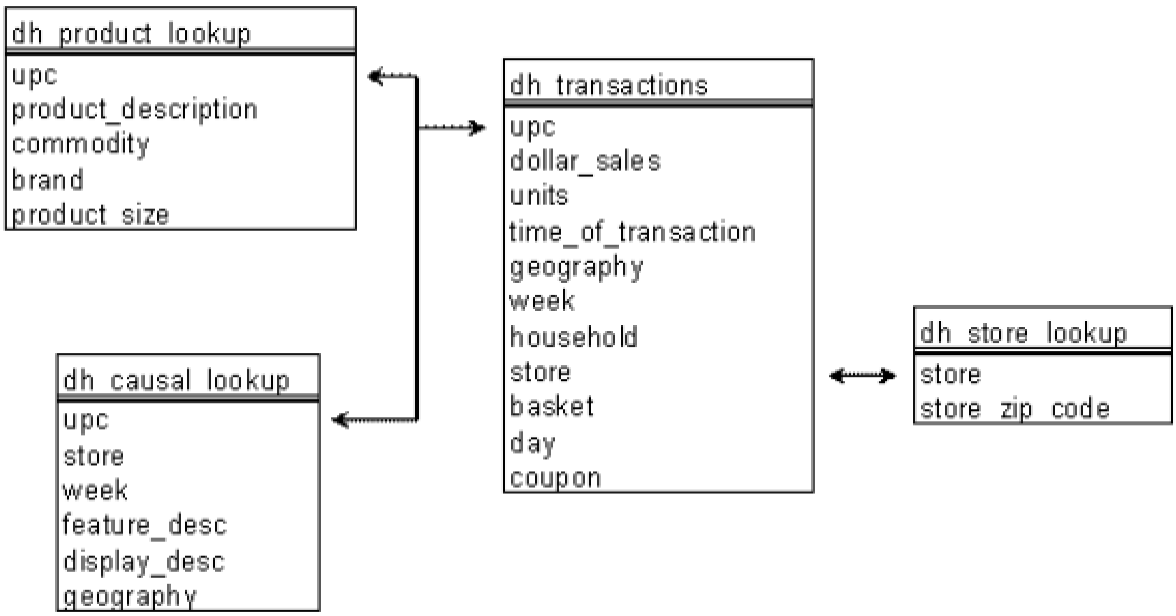
The dataset has successfully been used in classroom projects and case studies, and is ideally suited for this use. It allows students to interact with real-world data and search for their own insights. The richness of this data and the potential analyses it enables makes it a valuable teaching tool.

It also provides students with the opportunity to understand the process required to mine data. Since this is a relational database, students will need to merge multiple data tables together and aggregate data in search for insights.

The following are examples of questions that could be submitted to students from the Carbo-Loading data set:

- What is the household penetration of Product X? That is, out of all customers purchasing Pasta Sauce, what percent purchase Product X or Brand Z?
- Did any customers first purchase an item or category using a coupon? If so, how many of these customers made additional purchases of the item or category?
- In two complementary categories (e.g. Pasta and Pasta Sauce), what products, if any, are commonly purchased together? Are there cross-promotional opportunities?
- What percent of category customers are loyal to a product or brand? Is there a large percent of customers who purchase similar products within the same category but do not appear to be brand loyal? If so, can the reason for switching be detected?

CARBO-LOADING: DATASET DETAILS



dh_transactions

Description: This table contains a sample of 2 years of Pasta, Pasta Sauce, Syrup and Pancake Mix transactions, at the household level, obtained through the loyalty card program of a leading US grocer.	
# of Records: 5,197,681	
Variable	**Description**
upc	Standard 10 digit UPC.
dollar_sales	Amount of dollars spent by the consumer.
units	Number of products purchased by the consumer.
time_of_transaction	The time of transaction expressed in military time.
geography	Distinguishes between two large geographical regions. Each region typically contains portions of several states. Possible values are 1 or 2.
week	Expresses week of the transaction. Possible values are 1 through 104. Values are assigned in a chronological order.
household	Identifies unique households.
store	Identifies unique stores.
basket	Identifies unique baskets/trips to store.
day	Expresses day of the transaction. Possible values are 1 to 728. When 'day' has values 1 through 7, then 'week' will be 1. When 'day' has values 8 through 14, then 'week' will be 2, etc.
coupon	Indicates coupon usage. 1 if used, 0 for no coupon.

CARBO-LOADING: DATASET DETAILS

dh_store_lookup

Description: Provides each store's zip code.	
# of Records: 387	
Variable	**Description**
store	Identifies unique stores.
store_zip_code	5 digit zip code.

dh_product_lookup

Description: Provides detailed product information for each upc in 'dh_transactions'.	
# of Records: 927	
Variable	**Description**
upc	Standard 10 digit UPC.
product_description	Description of product.
commodity	Specifies 1 of 4 categories: Pasta, Pasta Sauce, Pancake Mix or Syrup.
brand	Specifies brand of item.
product_size	Specifies package size of product.

dh_causal_lookup

Description: Provides trade activity for each UPC/week. If a UPC is missing a record for a week then no trade activity occurred for that item. Note that weeks 1 - 42 do not have any causal data.	
# of Records: 351,372	
Variable	**Description**
upc	Standard 10 digit UPC.
store	Identifies unique stores.
week	Expresses the week of the transaction. Possible values are 1 through 104. The values are assigned in a chronological order.
feature_desc	Describes location of product on weekly mailer.
display_desc	Describes location of temporary in-store display containing the product.
geography	Distinguishes between two large geographical regions. Each region typically contains portions of several states. Possible values are 1 or 2.

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