The Complete Journey

User guide

dunnhumby

The complete journey

This dataset contains a representation of household level transactions over two years from a group of 2,500 households who are frequent shoppers at a retailer. It contains all of each household's purchases, not just those from a limited number of categories. For certain households, demographic information as well as direct marketing contact history are included.

Due to the number of tables and the overall complexity of The Complete Journey, it is suggested that this database be used in more advanced classroom settings. Further, The Complete Journey would be ideal for academic research as it should enable one to study the effects of direct marketing to customers.

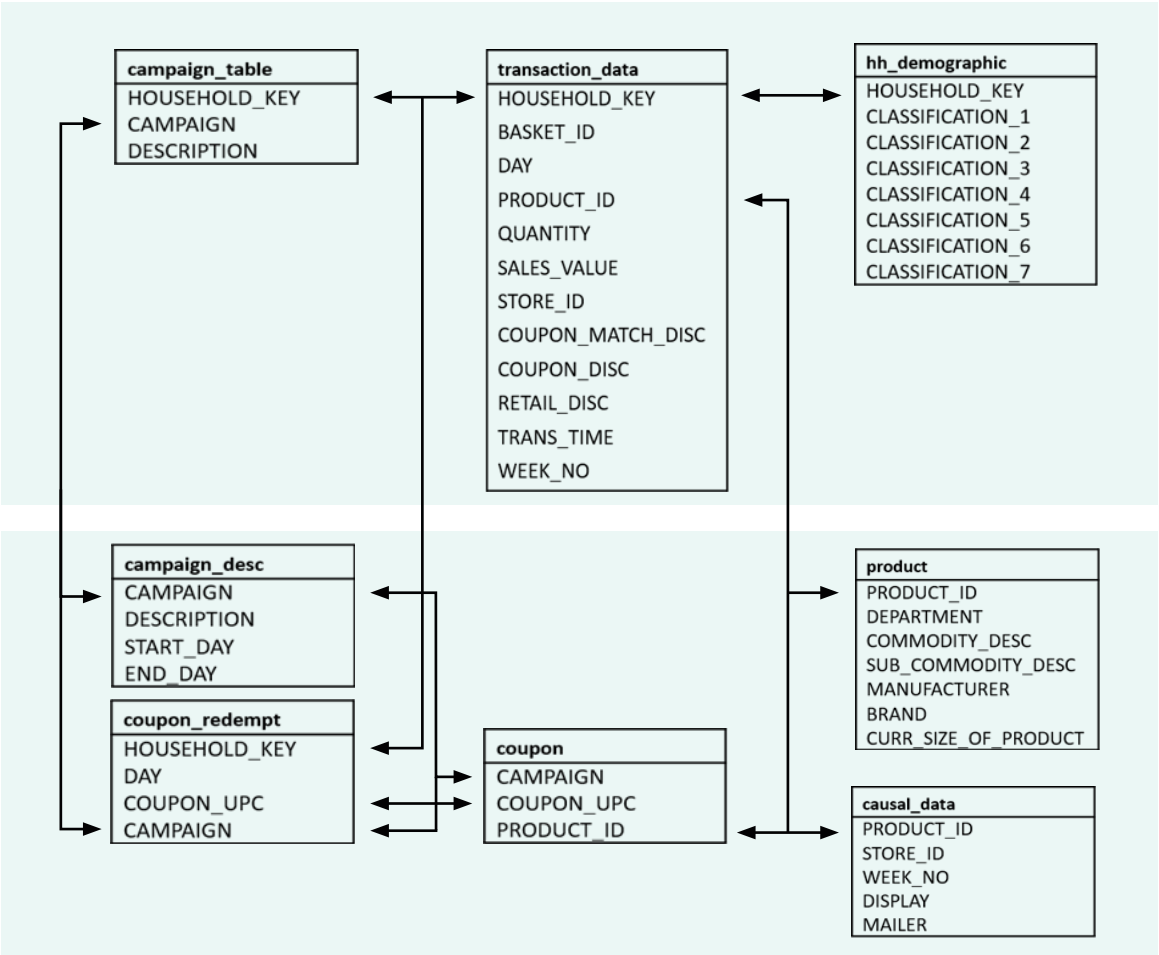
The following are examples of questions that could be submitted to students or considered for academic research:

- **How many customers are spending more over time? Less over time? Describe these customers.**
- **Of those customers who are spending more over time, which categories are growing at a faster rate?**
- **Of those customers who are spending less over time, with which categories are they becoming less engaged?**
- **Which demographic factors appear to affect customer spend? Engagement with certain categories?**
- **Is there evidence to suggest that direct marketing improves overall engagement?**

The complete journey: dataset details

Data tables

Lookup tables



transaction_data

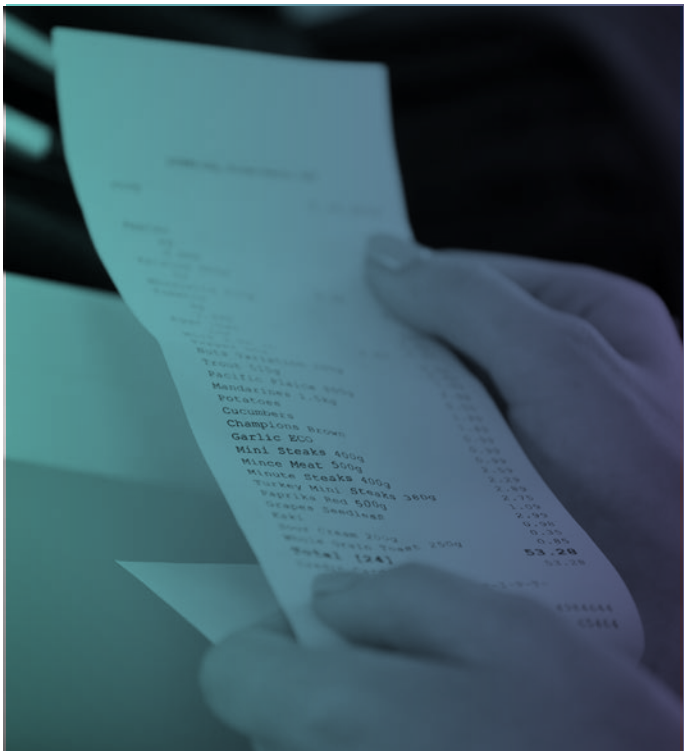
Description: This table contains all products purchased by households within the dataset. Each line found in this table is essentially the same line that would be found on a store receipt.

The variable sales_value in this table is the amount of dollars received by the retailer on the sale of the specific product, taking the coupon match and loyalty card discount into account. It is not the actual price paid by the customer. If a customer uses a coupon, the actual price paid will be less than the sales_value because the manufacturer issuing the coupon will reimburse the retailer for the amount of the coupon.

To calculate the actual product prices, use the formulas below:

Loyalty card price = (sales_value – (retail_disc + coupon_match_disc))/quantity

Non-loyalty card price = (sales_value – (coupon_match_disc))/quantity



Variable	Description
household_key	Uniquely identifies each household
basket_id	Uniquely identifies a purchase occasion
day	Day when transaction occurred
product_id	Uniquely identifies each product
quantity	Number of the products purchased during the trip
sales_value	Amount of dollars retailer receives from the sale
store_id	Identifies unique stores
coupon_match_disc	Discount applied due to retailer's match of manufacturer coupon
coupon_disc	Discount applied due to manufacturer coupon
retail_disc	Discount applied due to retailer's loyalty card programme
trans_time	Time of day when transaction occurred
week_no	Week of the transaction. Ranges 1 – 102

The example below demonstrates how to calculate the actual shelf price of the product:

Line 1 – When this product was purchased the retail_disc and coupon_disc were both zero, meaning the price of the product is the same as the amount received by the retailer.

Line 2 – Two items of this product were purchased, and there was a retail discount applied due to a loyalty card. To determine the regular shelf price of the product (exclusive of loyalty card discount) we take the sum of the amount paid and the discount, then divide it by the quantity. ($2 + $1.34)/2 = $1.67. The shelf price of the product including loyalty card discount is $2 / 2 = $1. Also, the customer paid $2 for both of these products which is the same amount the retailer received.

Line 3 – The actual shelf price of each product here is ($2.89 + $0.45)/2 = $1.67. Also, the customer paid $2.34 ($2.89 - $0.55) for these products, but the retailer will receive $2.89 due to the manufacturer discount.

Household Key	Basket ID	Day	Product ID	Quantity	Sales Value	Store ID	Retail Disc	Trans Time	Week No	Coupon Disc	Coupon Match Disc
2381	35730137393	534	819063	1	1.67	32004	0	2025	77	0	0
1431	41756231898	671	819063	2	2	446	-1.34	1740	97	0	0
888	36027750817	540	819063	2	2.89	401	0	1254	78	-0.55	-0.45

hh_demographic

Description: This table provides a representation of demographic information for a portion of households. The fields have been given generic names (classification_1, classification_2, etc.) and values (eg. classification_1 has values Group1, Group2, though to Group6). The values, however, have been chosen such that they provide meaningful information: ordinality is important. In other words, values are ordered in a logical fashion such that trends can be investigated.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
BASKET_ID	Household level demographic segmentation. Values have meaningful order. Possible values: Group1 through to Group6.
DAY	Household level demographic segmentation. Possible values: X, Y and Z.
PRODUCT_ID	Household level demographic segmentation. Values have meaningful order. Possible values: Level1 through to Level12.
QUANTITY	Household level demographic segmentation. Values have meaningful order. Possible values: 1 through to 5+.
SALES_VALUE	Household level demographic segmentation. Values have meaningful order. Possible values: Group1 through to Group6.
STORE_ID	Household level demographic segmentation. Values have meaningful order. Possible values: Group1 through to Group5.
COUPON_MATCH_DISC	Household level demographic segmentation. Values have meaningful order. Possible values: 1, 2, 3, None/Unknown.
COUPON_DISC	Discount applied due to manufacturer coupon
RETAIL_DISC	Discount applied due to retailer's loyalty card programme
TRANS_TIME	Time of day when transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 – 102

campaign_table

Description: This table lists the campaigns received by each household in the dataset. Each household may have received a different set of campaigns.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
CAMPAIGN	Uniquely identifies each campaign. Ranges 1-30
DESCRIPTION	Type of campaign (TypeA, TypeB or TypeC)

campaign_desc

Description: This table gives the length of time for which a campaign runs. So, any coupons received as part of a campaign are valid within the dates contained in this table.de

Variable	Description
CAMPAIGN	Uniquely identifies each campaign. Ranges 1-30
DESCRIPTION	Type of campaign (TypeA, TypeB or TypeC)
START_DAY	Start date of campaign
END_DAY	End date of campaign

假设每次 campaign 都是在一开始发券

product

Description: This table contains information on each product sold such as type of product, national or private label and a brand identifier.

Variable	Description
PRODUCT_ID	Number that uniquely identifies each product
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
MANUFACTURER	Code that links products with same manufacturer together
BRAND	Indicates Private or National label brand
CURR_SIZE_OF_PRODUCT	Indicates package size (not available for all products)

coupon

Description: This table lists all the coupons sent to customers as part of a campaign, as well as the products for which each coupon is redeemable. Some coupons are redeemable for multiple products. One example is a coupon for any private label frozen vegetable. There are a large number of products where this coupon could be redeemed.

For campaign TypeA, this table provides the pool of possible coupons. Each customer participating in a TypeA campaign received 16 coupons out of the pool. The 16 coupons were selected based on the customer's prior purchase behaviour. Identifying the specific 16 coupons that each customer received is outside the scope of this database.

For campaign TypeB and TypeC, all customers participating in a campaign receives all coupons pertaining to that campaign. 做不了 Type A 的

Variable	Description
CAMPAIGN	Uniquely identifiew each campaign. Ranges 1-30
COUPON_UPC	Uniquely identifies each coupon (unique to household and campaign)
PRODUCT_ID	Uniquely identifies the product for which the coupon is redeemable.

coupon_redempt

Description: This table identifies the coupons that each household redeemed.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
DAY	Day when the transaction occurred
COUPON_UPC	Uniquely identifies each coupon (unique to household and campaign).
CAMPAIGN	Uniquely identifies each campaign.

causal_data

Description: This table signifies whether a given product was featured in the weekly mailer or was part of an in-store display (other than regular product placement).

Variable	Description
product_id	Uniquely identifies each product
store_id	Identifies unique stores
week_no	Week of the transaction
display	Display location(see below)
mailer	Mailer location (see below)

Field	Contents
display	0 – Not on Display
	1 – Store Front
	2 – Store Rear
	3 – Front End Cap
	4 – Mid-Aisle End Cap
	5 – Rear End Cap
	6 – Side-Aisle End Cap
	7 – In-Aisle
	9 – Secondary Location Display
	A – In-Shelf
mailer	0 – Not on ad
	A – Interior page feature
	C – Interior page line item
	D – Front page feature
	F – Back page feature
	H – Wrap front feature
	J – Wrap interior coupon
	L – Wrap back feature
	P – Interior page coupon
	X – Free on interior page
	Z – Free on front page, back page or wrap

The complete journey: case study

John Smith is a valued customer at a national grocery retailer for which we have detailed transaction data. Throughout all the tables in the database, he is identified with a household_key of 208.

If we look at John's records from campaign_table, we can see that he received 8 different campaigns. Five of the campaigns were TypeA, and three were TypeB.

Description	Household Key	Campaign
TypeA	208	8
TypeA	208	13
TypeB	208	17
TypeA	208	18
TypeB	208	22
TypeA	208	26
TypeB	208	29
TypeA	208	30

These campaigns were spread out over the 2 year period represented by the data. To understand the time periods of these campaigns, look at the records in the campaign_desc table for the campaigns listed above for John.

Description	Campaign	Start Day	End Day
TypeA	8	412	460
TypeA	13	504	551
TypeB	17	575	607
TypeA	18	587	642
TypeB	22	624	656
TypeA	26	224	264
TypeB	29	281	334
TypeA	30	323	369

Let us take a closer look at campaign 22. When we look at all the distinct coupon_upc's from the coupon table where campaign = 22, we see that there were 21 distinct coupons sent out as part of that campaign.

Coupon UPC
10000085486
10000085487
10000089316
51312010033
51450050050
51800000050
52100000031
52113100077
52732670076
52800031032
54132220050
54400021032
54450000076
54850010033
55100090033
55150081028
55150081060
56233833793
57045970076
57100771033
57797520075

Let us take an even deeper look at one of the specific coupons offered as part of the campaign. If we print out all records from the coupon table where campaign = 22 and coupon_upc = 51800000050, we see that this coupon could actually be redeemed for a number of products.

Coupon UPC	Product ID	Campaign
51800000050	72717	22
51800000050	78466	22
51800000050	98340	22
51800000050	441607	22
51800000050	502673	22
51800000050	618203	22
51800000050	822690	22
51800000050	865156	22
51800000050	904813	22

Although all the products are not displayed above, we find that this coupon is actually valid on 38 distinct products.

If we go to the product table and print out all records for the product_id's above (72717, 78466, etc.), we see that this coupon is valid for refrigerated specialty rolls from a national brand.

Product ID	Manufacturer	Department	Brand	Commodity Desc	Sub Commodity Desc	Product Size
72717	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	8 OZ
78466	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	8 OZ
98340	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	12.4 OZ
441607	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	11 OZ
502673	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	8 CT
618203	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	8 OZ
822690	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	13.9 OZ
865156	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	8 OZ
904813	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS	4 OZ

As we've seen, John received a number of campaigns over the past two years that contained many coupons. Chances are, he did not redeem every coupon he received. So, let us take a look to see what coupons he did redeem. To do this, we need to view all records from the coupon_redempt table where household_key is 208. This shows us that he redeemed 7 coupons from 3 of the campaigns.

Household Key	Day	Coupon UPC	Campaign
208	606	10000085475	18
208	606	10000085475	18
208	654	51800000050	22
208	597	51800015050	18
208	597	51920021576	18
208	427	55100090033	8
208	601	55410000076	18

John's coupon redemptions are only part of the overall picture of his purchasing behavior. If we look at the records from the transaction_data table where household_key equals 208, we can view everything that John purchased.

Household Key	Basket ID	Day	Product ID	Quantity	Sales Value	Store ID	Retail Disc	Trans Time	Week No	Coupon Disc	Coupon Match Disc
208	31097480262	276	919534	1	1	327	-0.89	0923	40	-0.5	-0.5
208	31424115725	300	919534	1	1.99	327	-0.1	1248	44	-0.3	-0.3
208	31424115725	300	1017772	1	1.99	327	-0.1	1248	44	-0.3	-0.3
208	34749055907	503	1017772	1	1.69	327	0	1325	73	-0.3	-0.3
208	34749055907	503	1085095	1	1.69	327	0	1325	73	-0.3	-0.3
208	40666652921	597	1017772	1	0.88	327	-0.61	2039	86	-0.5	-0.5
208	40765530992	606	1017772	2	2.36	327	-1.22	1206	87	-0.4	-0.4
208	41008341062	622	919534	1	0.98	327	-0.45	1345	90	-0.4	-0.4
208	41008341062	622	1017772	1	1.38	327	-0.45	1345	90	0	0
208	41531980403	654	1017772	1	1.33	327	0	1043	94	-0.5	-0.5
208	41665840886	664	919534	1	1.83	327	0	1257	96	0	0
208	41665840886	664	1017772	1	1.83	327	0	1257	96	0	0

This gets a bit complicated, but we can combine the transaction data with the other tables to understand John's behavior when he was redeeming a coupon (and when he wasn't redeeming a coupon).

- **John received offers as part of campaign 22, which occurred between days 624 and 656**
- **We know he redeemed coupon 51800000050 on day 654**
- **Through the coupon table, we know that the coupon is actually valid for a number of products, including product 1017772**
- **From the table above, we can see (3rd line from the bottom) where John purchased this item and received a discount from using a coupon**

Knowing when John redeemed a coupon can help us learn a lot about him, and how the receipt of certain campaigns affected his behaviour. Does the receipt of campaigns cause him to purchase more items than he did previously? Is John more likely to redeem coupons for products he already purchases, or does it entice him to try products he has never purchased before?

There is one bit of information we have not talked about yet – what is happening in the rest of the store? Is it possible that John purchased the item above because of other events occurring in the store in addition to his coupon?

We obviously do not know a customer's reason for purchasing an item, but we do know whether an item was featured during the time of the purchase. To do this, let us look at product 72717. If we view all records from the causal_data table where product_id equals 72717, we see the weeks and stores where this product was featured in the weekly mailer and where it was featured as part of an in-store display. If we look at the first line, we can tell that in store 421 and week 12, the product was featured on a display in the rear of the store and was featured on an interior page of the mailer.

Product ID	Store ID	Week No	Display	Mailer
72717	421	12	2	A
72717	424	12	2	A
72717	299	12	7	A
72717	359	12	7	A
72717	400	12	7	A
72717	375	17	7	0
72717	424	24	2	A
72717	306	29	7	A
72717	333	29	7	A

We hope that this quick look at John Smith's behaviour provides clarity around The Complete Journey database, and inspires your own investigation into the purchasing behaviour of these customers.

CONTACT INFORMATION

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

sourcefiles@dunnhumby.com

dunnhumby THE WORLD'S FIRST CUSTOMER DATA SCIENCE PLATFORM

dunnhumby is the global leader in Customer Data Science, empowering businesses everywhere to compete and thrive in the modern data-driven economy. We always put the Customer First. Our mission: to enable businesses to grow and reimagine themselves by becoming advocates and champions for their Customers.

With deep heritage and expertise in retail — one of the world's most competitive markets, with a deluge of multi-dimensional data — dunnhumby today enables businesses all over the world, across industries, to be Customer First.

The dunnhumby Customer Science Platform is our unique mix of technology, software and consulting enabling businesses to increase revenue and profits by delivering exceptional experiences for their Customers – in-store, offline and online. dunnhumby employs over 2,000 experts in offices throughout Europe, Asia, Africa, and the Americas working for transformative, iconic brands such as Tesco, Coca-Cola, Meijer, Procter & Gamble, Raley's and L'Oreal.

Connect with us to start the conversation

dunnhumby.com