The complete journey

- Dataset includes household level transactions over two years.
- Data from 2,500 households who are frequent shoppers at a retailer.
- Contains all purchases made by each household, not limited to specific categories.
- Includes demographic information and direct marketing contact history for certain households.
- Suggested for advanced classroom settings due to the complexity and number of tables.
- Ideal for academic research, particularly for studying the effects of direct marketing to customers.

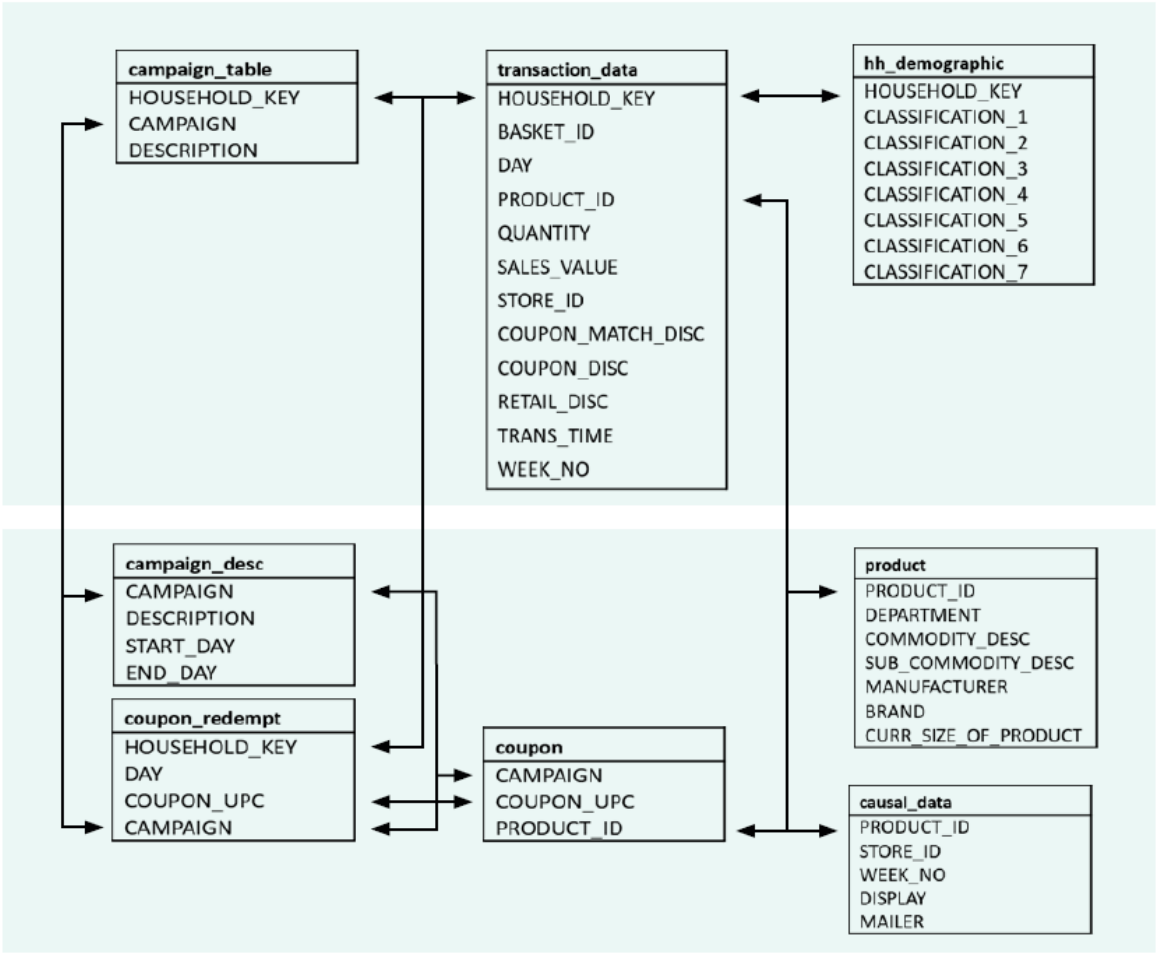
The following are examples of questions that could be submitted to students or considered for academic research:

- 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 spending? Engagement with certain categories?
- Is there evidence to suggest that direct marketing improves overall engagement?

The complete journey: dataset details

Data tables

Lookup tables



 https://www.dunnhumby.com/source-files/

transaction_data

- Table contains all products purchased by households in the dataset.

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

- Table contains all products purchased by households in the dataset.
- Each line is similar to a store receipt line.
- The `sales_value` represents the dollars received by the retailer for the product, including coupon match and loyalty card discount.
- The `sales_value` is not the actual price paid by the customer.
- If a coupon is used, the actual price paid by the customer is less than the `sales_value`.
- Manufacturer reimburses the retailer for the amount of the coupon.
- Formulas for calculating actual product prices:

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

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

Example calculations:

- **Line 1**: No discounts; Qty = 1, SALES_VALUES = 1.67,
  RETAIL_DISC = COUPON_DISC = COUPON_MATCH_DISC = 0
  - Total customer payment = `sales_value` = $1.67.
- **Line 2**: Two items with retail discount; Qty = 2, SALES_VALUES = 2, RETAIL_DISC =-1.34,
  COUPON_DISC = COUPON_MATCH_DISC = 0
  - regular shelf price = (SALES_VALUE - RETAIL_DISC)/QUANTITY= (2+1.34)/2=1.67,
  - discounted price = SALES_VALUE/QTY = $2/2 = $1,
  - total customer payment = (discounted price)(QUANTITY) = $1*2 = $2.
- **Line 3**: Qty = 2, SALES_VALUES = 2.89, RETAIL_DISC = 0, COUPON_DISC = -0.55, COUPON_MATCH_DISC = -0.45
  - Actual shelf price = (SALES_VALUE - COUPON_MATCH_DISC)/QUANTITY
    = (2.89-(-0.45))/2=1.67,
  - customer payment after discount = SALES_VALUE + COUPON_DISC = $2.89 + (−$0.55) = $2.34,
  - retailer receives = SALES_VALUE = $2.89
    = $1.67*2 - $0.45 = Actual shelf price * QUANTITY + COUPON_MATCH_DISC

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

Conclusions about **SALES_VALUE** from calculation samples:

1. **SALES_VALUE** reflects the **total amount** the retailer receives from the sale of the product. This means:
    - **It does not account for customer-facing discounts** directly (e.g., coupon discounts or retail loyalty discounts) as the amount the retailer receives includes reimbursements from manufacturers or promotions.
    - **SALES_VALUE** is the amount after the discounts have been applied on the customer's side but before any reimbursement adjustments.
2. **RETAIL_DISC** and **COUPON_MATCH_DISC** are negative because they represent **discounts** given to the customer, reducing the amount the customer pays, but they do **not reduce the retailer's earnings** (because the retailer is often reimbursed for these discounts).
3. In your examples:
    - **Line 1**: No discount applied, so the customer pays exactly the **SALES_VALUE**.
    - **Line 2**: There's a retail discount, so the **regular shelf price** (what the customer would have paid without a discount) is calculated by adding the **RETAIL_DISC** back to the **SALES_VALUE** (since **RETAIL_DISC** is negative). The **discounted price** is simply the **SALES_VALUE** divided by the quantity. The total customer payment reflects the reduced price, but the retailer still receives the full **SALES_VALUE**.
    - **Line 3**: Both a coupon and a coupon match discount apply, but **SALES_VALUE** reflects the total amount the retailer receives before coupon reimbursements are taken into account. After discounts, the customer pays less than the **SALES_VALUE**, but the retailer is reimbursed for the coupon value, so the retailer effectively receives the full **SALES_VALUE**.

Key Reconclusions:

- **SALES_VALUE includes the total value the retailer receives** for the sale, including the impact of discounts (since the retailer is often reimbursed for them), but **excludes the actual price the customer pays**.
- When **discounts** are applied, the **customer payment is less than SALES_VALUE**, but the retailer still receives the total **SALES_VALUE**.
- The negative values in **RETAIL_DISC** and **COUPON_MATCH_DISC** reflect discounts but do not subtract from **SALES_VALUE**; they only reduce the amount the customer pays.

Therefore, **SALES_VALUE** represents the **retailer's gross income** from the sale, factoring in any store discounts but excluding manufacturer-paid coupon reimbursements.

- **SALES_VALUE** represents the **real revenue** the retailer receives from the sale of the product.
- It **includes the cost** from **RETAIL_DISC** (loyalty or store discount) because the retailer absorbs that discount.
- It **excludes the manufacturer coupon** (**COUPON_DISC** and **COUPON_MATCH_DISC**) because the retailer will be **reimbursed** by the manufacturer for that amount. Hence, the final revenue from the sale includes what the customer pays plus the reimbursement from the manufacturer for any coupon discounts.

hh_demographic

- Table represents demographic information for a portion of households.
- Fields have generic names (e.g., `classification_1`, `classification_2`).
- Values are labeled generically (e.g., `classification_1` values: Group1 to Group6).
- Values are chosen to provide meaningful information.
- Ordinality is important, meaning values are ordered logically.
- Trends can be investigated based on the ordered values.

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

- 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

- Table provides the duration of each campaign.
- Coupons received as part of a campaign are valid within the dates specified in this table.

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

product

- 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

- Table lists all coupons sent to customers as part of a campaign.
- Specifies the products each coupon is redeemable for.
- Some coupons are redeemable for multiple products (e.g., any private label frozen vegetable).
- **Campaign TypeA**:
    - Table provides a pool of possible coupons.
    - Each customer received 16 coupons selected based on prior purchase behavior.
    - Specific 16 coupons received by each customer are not identified in the database.
- **Campaign TypeB and TypeC**:
    - All participating customers receive all coupons related to the campaign.

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

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

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

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

- John Smith is a valued customer at a national grocery retailer.
- Identified with a `household_key` of 208 in the database.
- John's records from the `campaign_table` show he received 8 different campaigns.
  - 5 campaigns were TypeA.
  - 3 campaigns were TypeB.

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

- The campaigns received by John Smith were spread out over the 2-year period represented by the data.
- To understand the time periods of these campaigns, refer to the records in the `campaign_desc` table for John's campaigns.

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

- Campaign 22 had 21 distinct coupons.
- To find these, look at the distinct `coupon_upc`s in the `coupon` table where `campaign = 22`.

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

- Deep dive into a specific coupon under campaign 22.
- Focus on coupon with coupon_upc = 51800000050.
- Review all records from the coupon table.
- Note potential redemption for multiple products.

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

- Coupon valid for 38 distinct products.
- Products include refrigerated specialty rolls.
- National brand associated with the products.
- Products identified by specific product_id's (e.g., 72717, 78466).

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 received multiple campaigns with numerous coupons over two years.
- He likely did not redeem every received coupon.
- Examining coupon_redemption records for household_key 208.
- John redeemed a total of 7 coupons from 3 campaigns.

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 participated in campaign 22, active from day 624 to 656.
- He redeemed coupon 51800000050 on day 654.
- Coupon 51800000050 valid for various products, including product 1017772.
- Transaction data indicates John purchased product 1017772 and received a discount through coupon redemption.

Understanding when John redeems coupons can provide insights into his behavior:

- **Impact of Campaign Receipt**: Analyzing if receiving campaigns prompts John to purchase more items than usual.
- **Coupon Redemption Patterns**: Examining whether John tends to redeem coupons for products he already purchases or if he tries new products.

These insights can help in understanding how promotional campaigns influence John's shopping behaviors and preferences over time.

- Check causal_data table for product_id 72717.
- Identify weeks and stores where the product was featured:
    - Store 421, week 12: Featured on a display in the rear of the store and on an interior page of the weekly mailer.

This information helps assess if John's purchase of product 72717 might be influenced by its featured status in the store, either through in-store displays or promotional mailers.

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