dunnhumby USA
essential customer genius

dunnhumby source files

Introduction. Structure. Details.

Overview and Legal Notice

This document provides you details on how to access dunnhumby's Source Files. Source Files is intended for use for academic study and research.

Before accessing this information, you and your organization are required to agree to the data license and confidentiality agreement with dunnhumbyUSA, found at www.us.dunnhumby.com/sourcefiles.aspx. While we do not charge for accessing this data, this agreement ensures obligations regarding the use of the data are adhered to.

Contact Information

If you have any general questions about dunnhumbyUSA, please contact:

Michelle Easton
Direct: 513.632.1177
michelle.easton@us.dunnhumby.com

If you have any technical questions regarding the use of dunnhumby's Source Files, please contact:

Brian Sampsel
Direct: 770.373.8019
brian.sampsel@us.dunnhumby.com

Accessing Source Files

All Source Files are available as either SAS® data sets or as comma separated files.

Each Source File has a relational database structure. The key data set contains retail transaction records at the household level, while the other relational databases are lookup files from which other descriptive data can be joined.

All data sets can be downloaded from:
www.us.dunnhumby.com/sourcefiles.aspx.

To access Source Files, click on each of the files and save them to your computer.

The files will need to be un-zipped once saved to your computer. If using UNIX®/Linux®, then the gunzip command can be used. In a Windows®-based environment, the program WinZip® can be used. If you do not have WinZip, a trial version can be downloaded from: www.winzip.com.

If you have any difficulty downloading the data sets, please contact us.

Source Files Overview

There are two Source Files from the grocery class of trade for use today which are described within this section. Each Source File has its own section to describe in detail the contents. In addition, we suggest possible questions to use in an academic setting with each Source File.

Stay tuned for future additions to Source Files, as dunnhumbyUSA has plans to add Source Files from industries beyond grocery as well as beyond the USA.

Source File # 1: Carbo-Loading

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.

Carbo-Loading 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 the search for insights.

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

- 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?
- Are there customers who first purchased an item or a category using a coupon? If so, how many of these customers made additional purchases of the item or category?
- In two complementary categories, what are products commonly purchased together? For example, are there any pasta and pasta sauce products often purchased together? Are there cross-promotional opportunities?
- What percent of category customers are loyal to a product or a 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? Is it price, coupon usage, etc.?

Source File # 2 – The Complete Journey

This database is similar to the first as it contains household level transactions over two years; however, it has two major differences:

1. It contains transactions for a small group of households (2,500) who are frequent shoppers of the store
2. It contains all of these households' purchases within the store, not just those from a limited number of categories

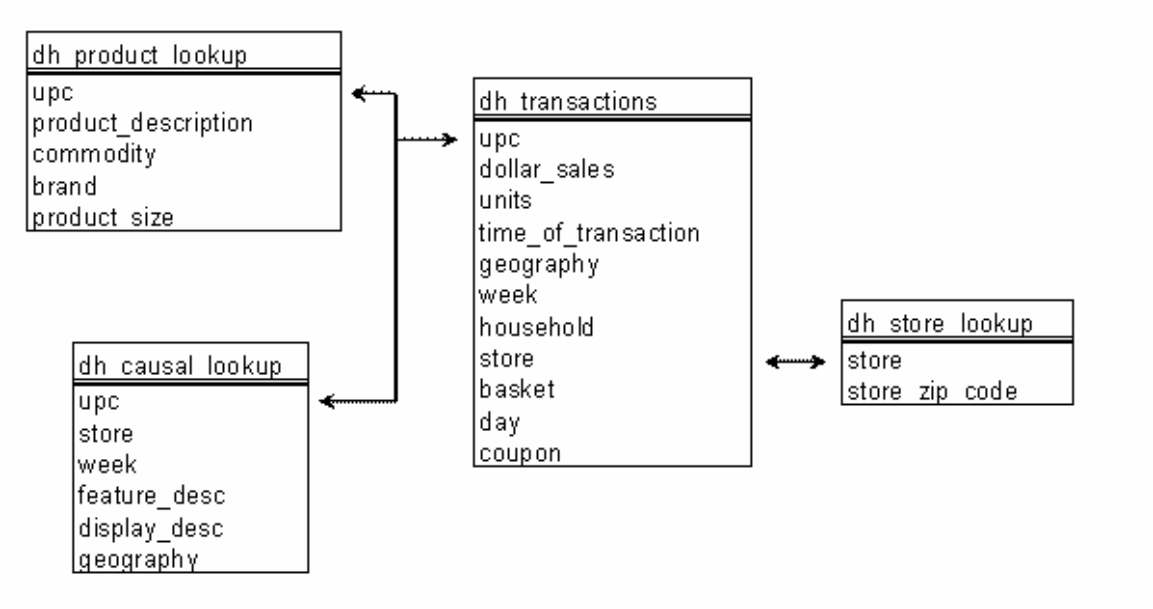
For certain households, demographic information as well as direct marketing contact history are included.

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 in an advanced classroom setting and/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, which categories are they becoming less engaged with?
- Which demographic factors (e.g. HH size, presence of children, income) appear to affect spend of the customer? Engagement with certain categories?
- Is there evidence to suggest that direct marketing improves overall customer engagement?

Carbo-Loading: Source File 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.

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.

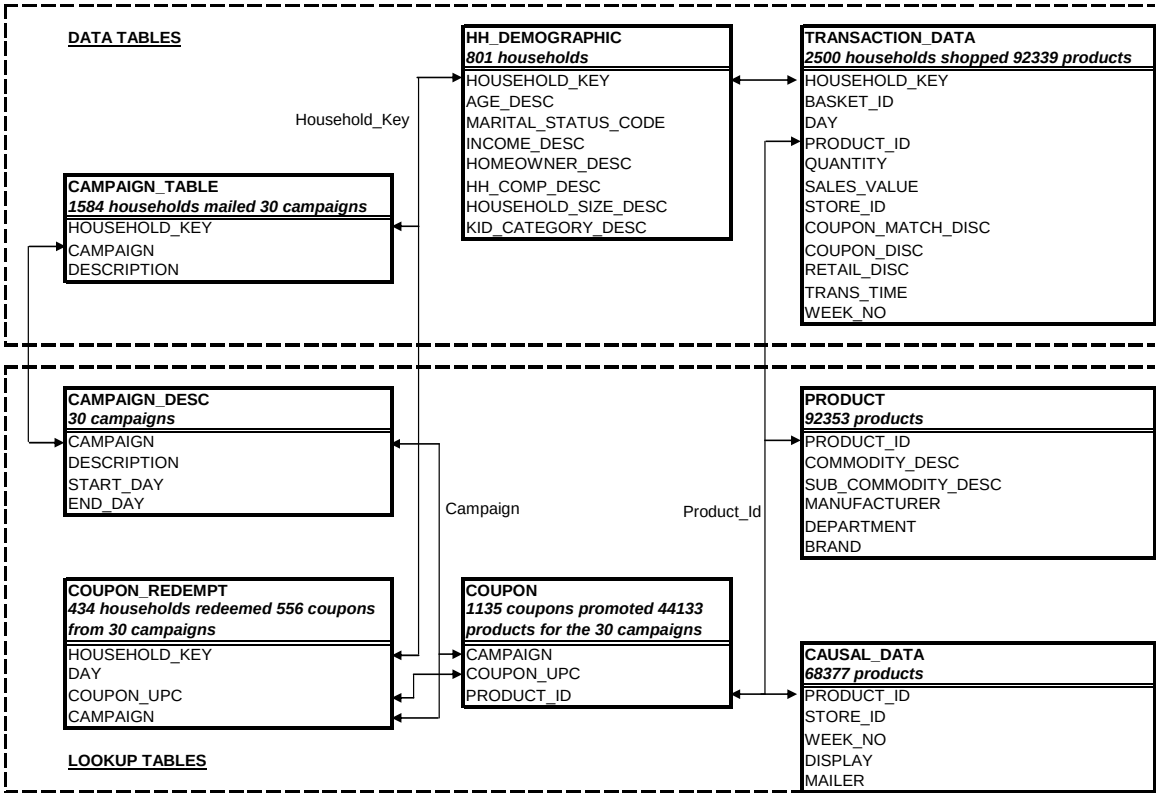
dh_product_lookup

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.

The Complete Journey – Source File Details



hh_demographic

This table contains demographic information for a portion of households. Due to nature of the data, the demographic information is not available for all households.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
AGE_DESC	Estimated age range
MARITAL_STATUS_CODE	Marital Status (A - Married, B- Single, U - Unknown)
INCOME_DESC	Household income
HOMEOWNER_DESC	Homeowner, renter, etc.
HH_COMP_DESC	Household composition
HOUSEHOLD_SIZE_DESC	Size of household up to 5+
KID_CATEGORY_DESC	Number of children present up to 3+

transaction_data

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

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 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 program
TRANS_TIME	Time of day when the transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 - 102

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

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

campaign_table

This table lists the campaigns received by each household in the study. Each household 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

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.

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

coupon

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 behavior. 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.

Variable	Description
CAMPAIGN	Uniquely identifies each campaign. Ranges 1 - 30
COUPON_UPC	Uniquely identifies each coupon (unique to household and campaign)
PRODUCT_ID	Uniquely identifies each product

coupon_redempt

This table identifies the coupons that each household redeemed.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
DAY	Day when 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

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.

To learn a little about John, we can obtain his demographic information by looking at the record in the *hh_demographic* table where *household_key* = 208. The table below shows the information we receive, and tells us that he is a homeowner, who makes between $50,000 and $74,000 a year and is between 45 and 54 years old.

Age Desc	Maritial Status Code	Income Desc	Homeowner Desc	HH Comp Desc	Household Size Desc	Kid Category Desc	Household Key
45-54	U	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	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 of the study. 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
72717	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
78466	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
98340	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
441607	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
502673	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
618203	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
822690	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
865156	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS
904813	236	GROCERY	National	REFRGRATD DOUGH PRODUCTS	REFRIGERATED SPECILATY ROLLS

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 behavior. 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 behavior provides clarity around The Complete Journey database, and inspires your own investigation into the purchasing behavior of these customers.