Dunnhumby - The Complete Journey

Summary	Marketing Analytics Assignment on Dunnhumby - The Complete Journey dataset
URL	https://www.dunnhumby.com/careers/engineering/sourcefiles
Category	Web
Author	Gayatri and Rohan

Dunnhumby - The Complete Journey

The Complete Journey

Contributors :

- Gayatri Reddiar
- Rohan Kapadnis

About the dataset

Dunnhumby - The Complete Journey

- This dataset contains household level transactions over two years from a group of 2,500 households who are frequent shoppers at a retailer.
- It also includes various marketing contact history for select households.
- The dataset is broken down into eight data frames, which can be categorized into two groups: data tables and look-up tables.
- Within the data tables, there are three data frames: hh_demographic contains demographic information on 801 households; transaction_data contains all products purchased by households participating in the study; and campaign_table lists the marketing campaigns received by each household.
- Within the look-up tables there are five data frames: campaign_desc is a table which gives the length of time for which the campaign runs; product contains all the product information for each product ID
- Coupon lists all the products each coupon is good for (note that most coupons are good for more than one product)
- Coupon_redempt identifies which households redeemed a coupon and when they did; and causal_data signifies whether a given product was featured in the weekly mailer or part of an in-store display.

Data Wrangling

The process of cleaning, structuring and enriching raw data into a desired format has become increasingly significant today for better decision making in less time.

We have used the following tools for cleaning, joining, filtering, aggregating the dataset and handling missing values:

1. XSV Tool
2. Trifacta Wrangler
3. Pandas Library

XSV Tool

XSV is a command line program for indexing, slicing, analyzing, splitting and joining CSV files

Tasks performed using XSV

- Stats : Shows basic types and statistics of each column in the CSV file. (i.e., mean, standard deviation, median, range, etc.)

```
(base) Rohans-MacBook-Pro:DH rohankapadnis$ xsv stats transaction_data.csv
field,type,sum,min,max,min_length,max_length,mean,stddev
household_key,Integer,3301647851,1,2500,1,4,1271.952517055037,726.0658250296798
BASKET_ID,Integer,88322893943181368,26984851472,42305362535,11,11,34026199138.886086,4711648130.286996
DAY,Integer,1009106965,1,711,1,3,388.7562217517156,189.72095876712996
PRODUCT_ID,Integer,7505390769685,25671,18316298,5,8,2891435.1595946625,3837402.949779842
QUANTITY,Integer,260685622,0,89638,1,5,100.42855811000898,1153.435988792271
SALES_VALUE,Float,8057463.08008801,0,840,1,12,3.104119793568993,4.182273313779992
STORE_ID,Integer,8157537419,1,34280,1,5,3142.673210870742,8937.11157415614
RETAIL_DISC,Float,-1398334.8399981533,-180,3.99,1,12,-0.5387053979378078,1.24919120216716
TRANS_TIME,Integer,4053459112,0,2359,4,4,1561.586139092985,399.8377428678983
WEEK_NO,Integer,145935944,1,102,1,3,56.221498983716984,27.10222267907501
COUPON_DISC,Float,-42611.54000000039,-55.93,0,1,6,-0.01641600134374444,0.2168409481344479
COUPON_MATCH_DISC,Float,-7575.809999999634,-7.7,0,1,5,-0.0029185640120011455,0.039690035512505634
```

```
(base) Rohans-MacBook-Pro:DH rohankapadnis$ xsv stats product.csv
field,type,sum,min,max,min_length,max_length,mean,stddev
PRODUCT_ID,Integer,492089369545,25671,18316298,5,8,5328352.8368867505,5359908.059201377
MANUFACTURER,Integer,160622954,1,6477,1,4,1739.228330427837,1818.259725375221
DEPARTMENT,Unicode,, ,VIDEO RENTAL,1,15,,
BRAND,Unicode,,National,Private,7,8,,
COMMODITY_DESC,Unicode,, ,YOGURT,1,30,,
SUB_COMMODITY_DESC,Unicode,, ,YOGURT NOT MULTI-PACKS,1,30,,
CURR_SIZE_OF_PRODUCT,Unicode,, ,XLRG,1,10,,
(base) Rohans-MacBook-Pro:DH rohankapadnis$
```

Pros :

- Really fast and powerful.
- Performs the task of joining large datasets in under 10 seconds
- Ability to generate stats and frequency to get overall profile of the data

Cons :

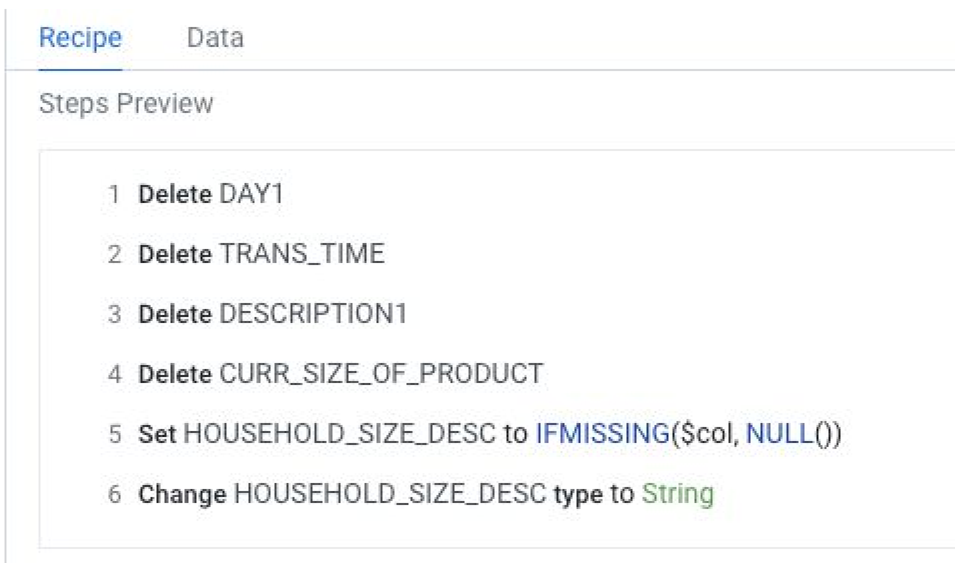
- Lack of UI
- Low community support
- Limited number of functionalities

Trifacta Wrangler

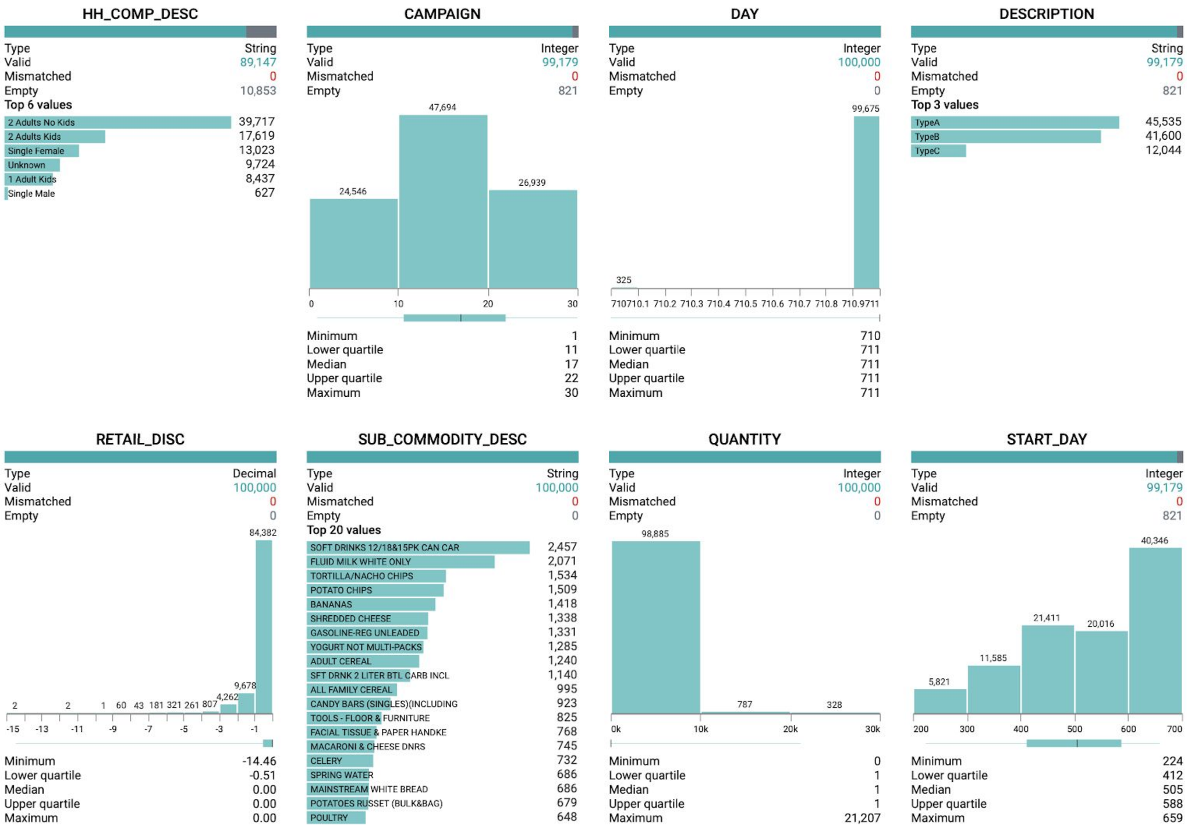
Tool used to visually explore, transform, clean and join together

Tasks performed using Trifacta Wrangler :

- Created recipes to remove unnecessary columns
- Changed data types
- Set value to NULL for missing values



- Generate Reports



Pros :

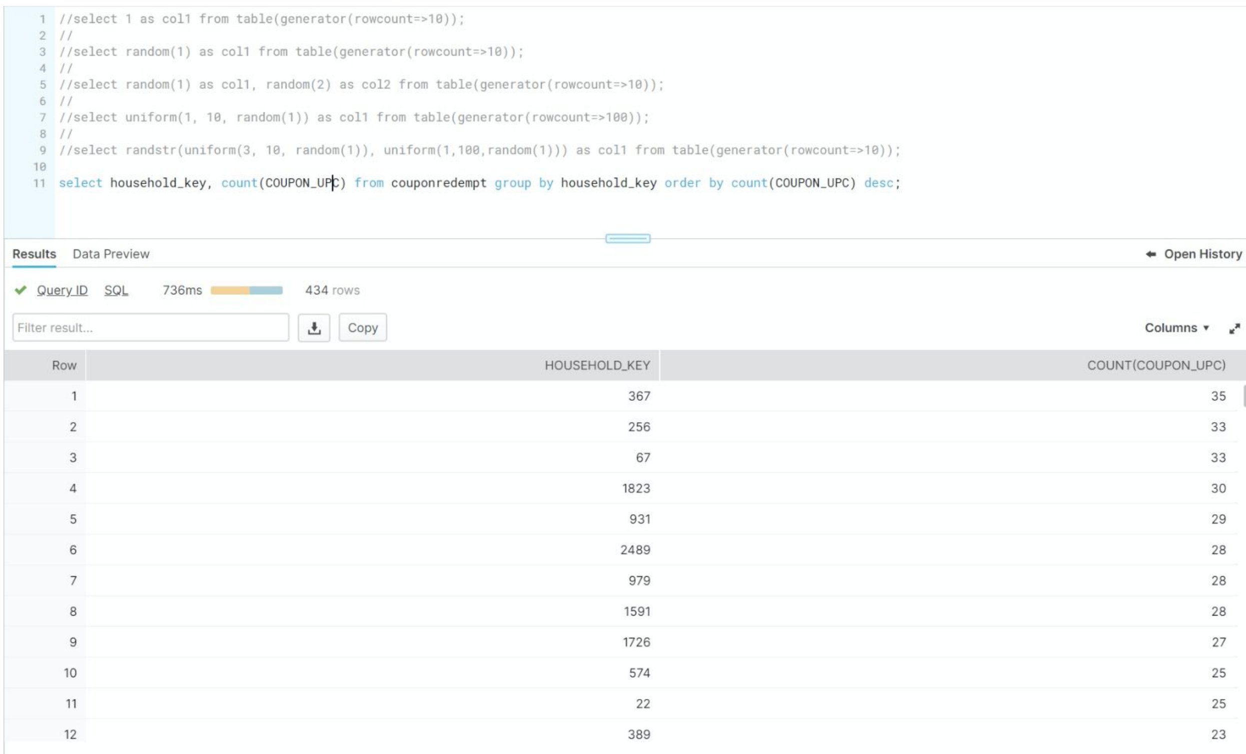
- Provides an easy UI to generate recipes
- Recipe Suggestions for quick wrangling
- Ability to download data report to get the overview

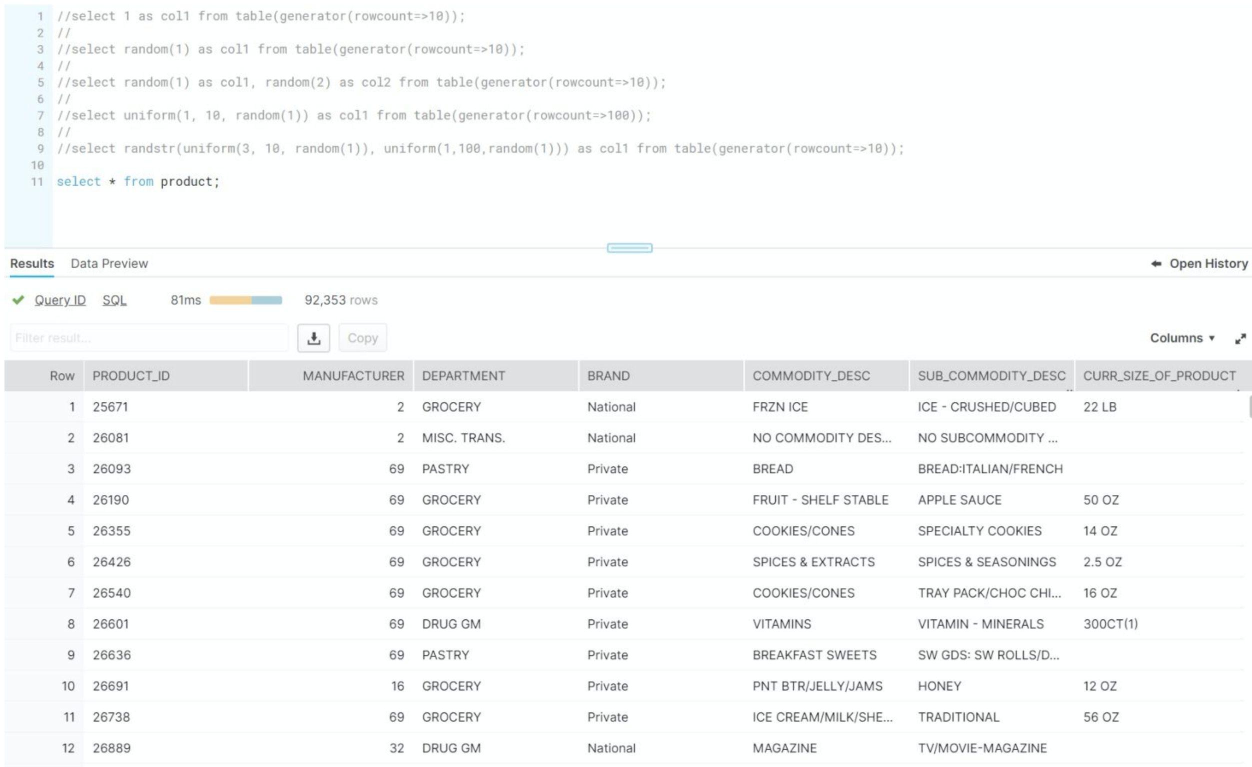
Cons :

- Free version allows only 100 MB file limit
- Pro version is expensive

Snowflake

Snowflake is an **analytic cloud based data warehouse** provided as Software-as-a-Service (SaaS).





```
//select 1 as col1 from table(generator(rowcount=>10));
//
//select random(1) as col1 from table(generator(rowcount=>10));
//
//select random(1) as col1, random(2) as col2 from table(generator(rowcount=>10));
//
//select uniform(1, 10, random(1)) as col1 from table(generator(rowcount=>100));
//
//select randstr(uniform(3, 10, random(1)), uniform(1,100,random(1))) as col1 from table(generator(rowcount=>10));

select * from product;
```


Row	PRODUCT_ID	MANUFACTURER	DEPARTMENT	BRAND	COMMODITY_DESC	SUB_COMMODITY_DESC	CURR_SIZE_OF_PRODUCT
1	25671	2	GROCERY	National	FRZN ICE	ICE - CRUSHED/CUBED	22 LB
2	26081	2	MISC. TRANS.	National	NO COMMODITY DES...	NO SUBCOMMODITY ...	
3	26093	69	PASTRY	Private	BREAD	BREAD:ITALIAN/FRENCH	
4	26190	69	GROCERY	Private	FRUIT - SHELF STABLE	APPLE SAUCE	50 OZ
5	26355	69	GROCERY	Private	COOKIES/CONES	SPECIALTY COOKIES	14 OZ
6	26426	69	GROCERY	Private	SPICES & EXTRACTS	SPICES & SEASONINGS	2.5 OZ
7	26540	69	GROCERY	Private	COOKIES/CONES	TRAY PACK/CHOC CHI...	16 OZ
8	26601	69	DRUG GM	Private	VITAMINS	VITAMIN - MINERALS	300CT(1)
9	26636	69	PASTRY	Private	BREAKFAST SWEETS	SW GDS: SW ROLLS/D...	
10	26691	16	GROCERY	Private	PNT BTR/JELLY/JAMS	HONEY	12 OZ
11	26738	69	GROCERY	Private	ICE CREAM/MILK/SHE...	TRADITIONAL	56 OZ
12	26889	32	DRUG GM	National	MAGAZINE	TV/MOVIE-MAGAZINE	

Pros :

- Simple UI with great functionalities
- Data Loading is easy
- Integration with other BI Tools such as Einstein Analytics

Cons :

- File loading limit is restricted to 50MB for free version

Pandas Library

Pandas is a Python Programming Language based library for Data Analysis and Manipulation. The Pandas, NumPy, Matplotlib and Mlxtend libraries of Python have been predominantly used for analysing the given dataset. Below analysis was possible using the given libraries-

1. Exploratory Data Analysis

1. EXPLORATORY DATA ANALYSIS

TRANSACTION SUMMARY

In [10]:
```
buyers = transaction_data.groupby('household_key').agg({'SALES_VALUE': 'sum','PRODUCT_ID': 'count','BASKET_ID': 'nunique
buyers.columns = ['tot_spend','num_prods','trips']
buyers.describe().transpose()
```

Out[10]:

	count	mean	std	min	25%	50%	75%	max
tot_spend	2500.0	3222.985232	3349.026076	8.17	970.74	2157.75	4413.32	38319.79
num_prods	2500.0	1038.292800	999.097354	4.00	325.00	734.00	1454.50	6851.00
trips	2500.0	110.593600	115.669368	1.00	39.00	79.00	142.25	1300.00

From the above table we see that over the two-year period, the average consumer:

spent a total $3,223 purchased 1,038 products visited the supermarket 111 times

In [30]:
```
product.PRODUCT_ID.nunique()
```

Out[30]: 92353

There are 92353 products offered by the retailer across all its stores

In [31]:
```
transaction_data.PRODUCT_ID.nunique()
```

Out[31]: 92339

Out of the 92353 products, 92339 have been bought by customers over a 2 year period

In [32]:
```
causal_data.PRODUCT_ID.nunique()
```

Out[32]: 68377

The various stores have promoted 68377 products via mailers or different display options in the store

2. Marketing Effectiveness (Campaigns, Mailers, Displays)

In [26]:

```
mt = mailer_transaction.groupby(['STORE_ID','mailer']).agg({'SALES_VALUE': 'sum'})
mt.to_csv('C:/NEU/Courses/Summer/Assignment/Assignment1/dunnhumby_The-Complete-Journey/MailerSales.csv')
mt
```

Out[26]:

STORE_ID	mailer	SALES_VALUE
286	0	22434.94
	A	65827.60
	C	1715.45
	D	45046.55
	F	8917.47
...	...	...
34280	H	2948.06
	J	955.36
	L	490.81
	X	1118.12
	Z	54.87

1155 rows × 1 columns

3. Purchase Behaviour as a factor of Income

In [87]:

```
custinc = income.merge(customer, on=['household_key'], how='inner')
custinc.sort_values('CUSTOMER_PAID', ascending=False).head(10)
```

Out[87]:

	household_key	INCOME_DESC	CUSTOMER_PAID
517	1609	125-149K	27781.20
745	2322	175-199K	23593.02
470	1453	125-149K	21565.43
463	1430	35-49K	20298.49
235	707	100-124K	19179.82
526	1653	Under 15K	19115.97
241	718	25-34K	18958.96
331	982	35-49K	18583.98
139	400	150-174K	18382.42
405	1229	150-174K	18272.01

4. Recency, Frequency, Monetary Analysis to gauge customer loyalty

To see the top 10 customers, we'll filter the data based on RFM score of 111

In [80]:

```
segmented_rfm[segmented_rfm['RFMClass']=='111'].sort_values('monetary_value', ascending=False).head(10)
```

Out[80]:

household_key	recency	frequency	monetary_value	r_segment	f_segment	m_segment	RFMClass
1023	2	4403	38240.65	1	1	1	111
1609	1	6625	27781.20	1	1	1	111
2322	1	5692	23593.02	1	1	1	111
1453	2	6561	21565.43	1	1	1	111
1430	1	5372	20298.49	1	1	1	111
707	1	4310	19179.82	1	1	1	111
1653	2	5347	19115.97	1	1	1	111
982	2	5806	18583.98	1	1	1	111
400	1	4678	18382.42	1	1	1	111
1229	1	4455	18272.01	1	1	1	111

5. Cohort Analysis to understand Customer Retention

```
cohort_pivot = df_cohort.pivot_table(index = 'cohort',
                                     columns = 'period_number',
                                     values = 'n_customers')
cohort_pivot
```

period_number / cohort	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	2
2018-01	540.0	435.0	430.0	425.0	429.0	435.0	438.0	435.0	436.0	433.0	...	426.0	445.0	439.0	442.0	440.0	461.0	444.0	452.0	442.0	351
2018-02	508.0	422.0	418.0	407.0	404.0	397.0	416.0	401.0	410.0	404.0	...	409.0	430.0	409.0	411.0	418.0	405.0	415.0	399.0	310.0	Na
2018-03	714.0	573.0	571.0	566.0	571.0	565.0	575.0	564.0	561.0	569.0	...	598.0	575.0	580.0	594.0	575.0	584.0	597.0	451.0	NaN	Na
2018-04	729.0	600.0	597.0	586.0	589.0	597.0	602.0	581.0	578.0	584.0	...	590.0	595.0	613.0	594.0	610.0	600.0	482.0	NaN	NaN	Na
2018-05	2.0	2.0	NaN	2.0	2.0	2.0	2.0	2.0	2.0	1.0	...	1.0	1.0	1.0	1.0	1.0	1.0	NaN	NaN	NaN	Na
2018-06	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	1.0	1.0	1.0	1.0	1.0	NaN	NaN	NaN	NaN	Na
2018-07	1.0	1.0	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
2018-08	1.0	1.0	1.0	1.0	1.0	1.0	NaN	NaN	NaN	NaN	...	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
2018-10	1.0	NaN	NaN	NaN	NaN	1.0	1.0	NaN	1.0	1.0	...	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
2019-03	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
2019-09	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
2019-11	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

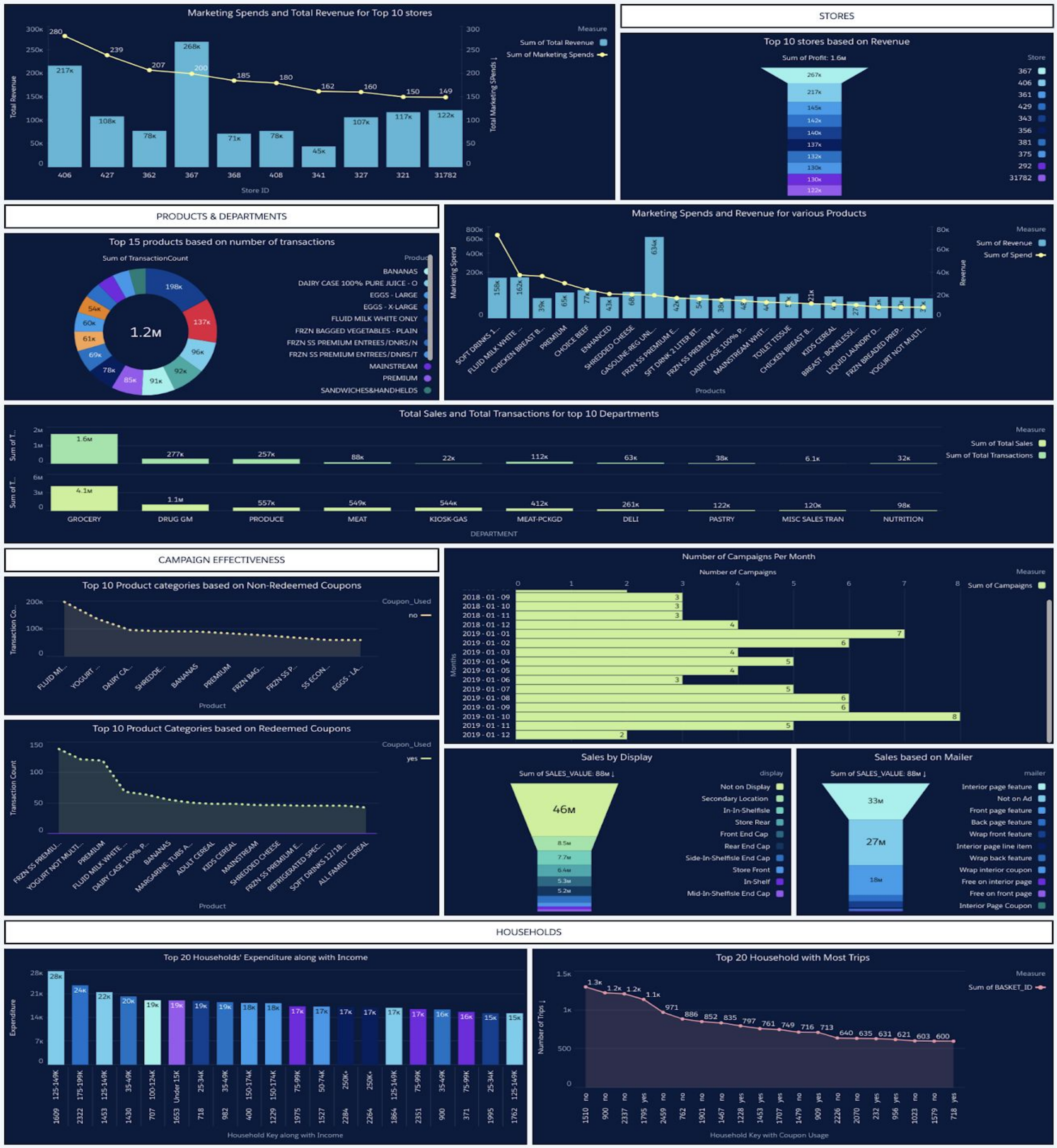
12 rows × 24 columns

6. Market Basket Analysis to implement targeted marketing

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
rules.head()
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(FLUID MILK WHITE ONLY)	(ALL FAMILY CEREAL)	0.221749	0.037098	0.024081	0.108596	2.927266	0.015855	1.080208
1	(ALL FAMILY CEREAL)	(FLUID MILK WHITE ONLY)	0.037098	0.221749	0.024081	0.649118	2.927266	0.015855	2.217983
2	(FLUID MILK WHITE ONLY)	(BANANAS)	0.221749	0.109601	0.056571	0.255113	2.327649	0.032267	1.195348
3	(BANANAS)	(FLUID MILK WHITE ONLY)	0.109601	0.221749	0.056571	0.516154	2.327649	0.032267	1.608467
4	(MAINSTREAM WHITE BREAD)	(BANANAS)	0.096989	0.109601	0.022905	0.236165	2.154765	0.012275	1.165695

Dashboard


Marketing Spends and Total Revenue for Top 10 stores
STORES
Top 10 stores based on Revenue
Sum of Profit: 1.6M
PRODUCTS & DEPARTMENTS
Top 15 products based on number of transactions
Sum of TransactionCount
1.2M
Marketing Spends and Revenue for various Products
Total Sales and Total Transactions for top 10 Departments
CAMPAIGN EFFECTIVENESS
Top 10 Product categories based on Non-Redeemed Coupons
Number of Campaigns Per Month
Top 10 Product Categories based on Redeemed Coupons
Sales by Display
Sum of SALES_VALUE: 88M
Sales based on Mailer
Sum of SALES_VALUE: 88M
HOUSEHOLDS
Top 20 Households' Expenditure along with Income
Top 20 Household with Most Trips

Insights

1. **Purchase Summary:** Based on the transactions over a two-year period, below is the purchase summary of all the 2500 households

	count	mean	std	min	25%	50%	75%	max
tot_spend	2500.0	3222.985232	3349.026076	8.17	970.74	2157.75	4413.32	38319.79
num_prods	2500.0	1038.292800	999.097354	4.00	325.00	734.00	1454.50	6851.00
trips	2500.0	110.593600	115.669368	1.00	39.00	79.00	142.25	1300.00

From the above table we see that over the two-year period, the average consumer spent a total $3,223, purchased 1,038 products and visited the supermarket 111 times

2. **Product Analysis:** In terms of transaction size, fluid milk white sells the most with around 198K transactions whereas GASOLINE-REG UNLEADED tops the list in terms of revenue share amounting to 634K. On the other hand, WHOLE - TOM (16 LBS & OVER FRZ, WHOLE TOMS (OVER 15LBS), INFANTS DRESSES RTW are few of the products with more spends than revenue resulting in losses for the retailer.
3. **Departmental Analysis:** Grocery as a department has the highest share in terms of both transaction size and revenue with a whopping 4.1M in sales and 1.6 M transactions. On the other hand, ELECT & PLUMBING, GRO BAKERY, HOUSEWARES are few of the departments with little to no sales.
4. **Store Analysis:** In terms of profit share, stores 367, 406, 361 contribute the most to the retailer's earnings whereas 1235, 551, 765, 2760 have negligible contribution in the retailer's earnings. The retailer should analyse performance of each store and take measures to address the issue in such cases.
5. **Campaign Analysis:** Out of the total 1135 coupons promoted as part of the 30 different campaigns launched, 556 coupons were used by a total of 434 households out of the 2500 transacting with the retailer stores. However about 99% of the transactions did not involve use of coupons. On a bright note, about 38.5% of coupons were used to buy an unfamiliar product.

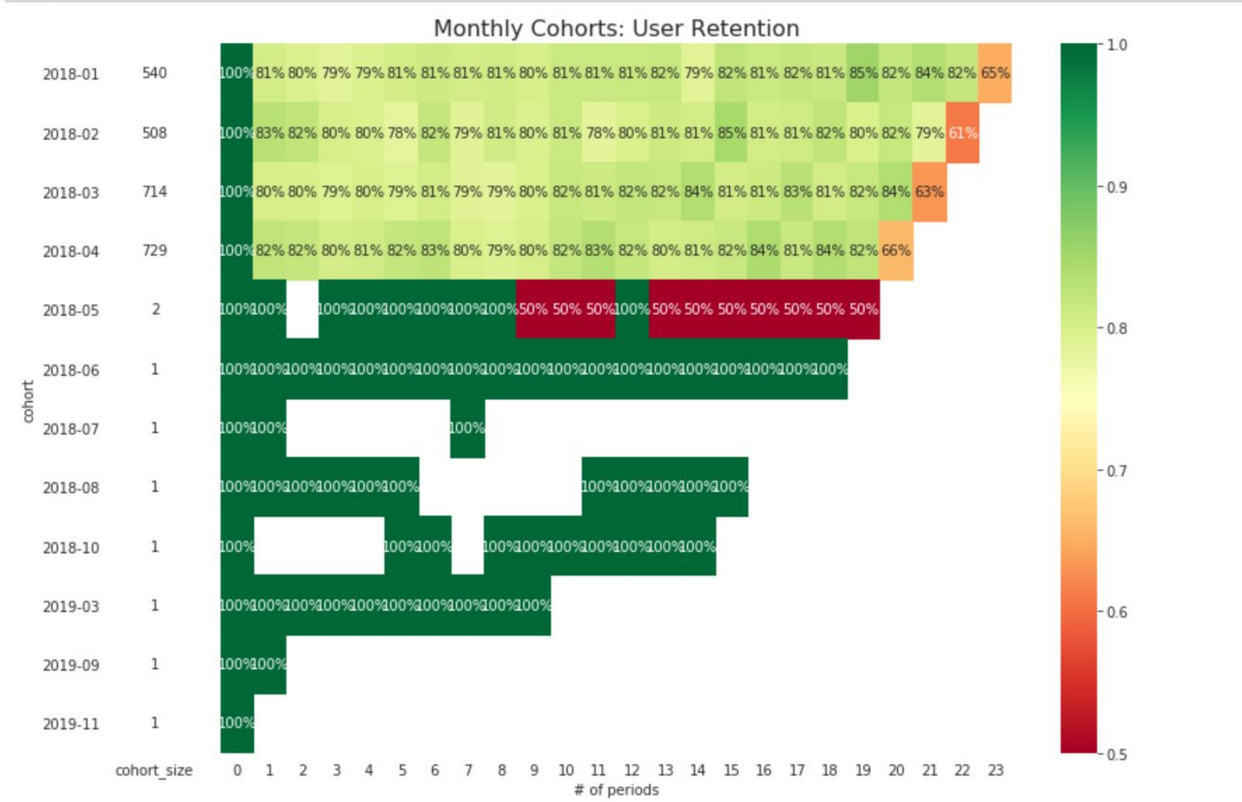
6. **Mailer & Store Display Analysis:** About 68377 products were promoted either in mailers or through different display options in the store. About 70% of the revenue share amounting to about 61M was contributed by products featured through ads while about 50% of the revenue share totaling 42M was contributed by products on different displays in the store.
7. **Household Analysis:** A total of 2500 households transacted with the retailer over a 2-year period. The household's income is directly proportional to their spending at the stores. About 80% of the 2500 households shopped with the retailer without redeeming coupons and made the most trips to the stores.
8. **RFM Analysis:** RFM analysis is a customer segmentation technique that uses past **purchase behavior** to segment customers into groups according to the distribution of values for **recency**, **frequency**, and **monetary** value.
    a. **RECENCY (R):** Time since last purchase
    b. **FREQUENCY (F):** Total number of purchases
    c. **MONETARY VALUE (M):** Total monetary value

Segment	RFM	Description	Marketing
Best Customers	111	Bought most recently and most often, and spend the most	No price incentives, new products, and loyalty programs
Loyal Customers	X1X	Buy most frequently	Use R and M to further segment
Big Spenders	XX1	Spend the most	Market your most expensive products
Almost Lost	311	Haven't purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Customers	411	Haven't purchased for some time, but purchased frequently and spend the most	Aggressive price incentives
Lost Cheap Customers	444	Last purchased long ago, purchased few, and spent little	Don't spend too much trying to re-acquire

household_key	recency	frequency	monetary_value	r_segment	f_segment	m_segment	RFMClass
1023	2	4403	38240.65	1	1	1	111
1609	1	6625	27781.20	1	1	1	111
2322	1	5692	23593.02	1	1	1	111
1453	2	6561	21565.43	1	1	1	111
1430	1	5372	20298.49	1	1	1	111
707	1	4310	19179.82	1	1	1	111
1653	2	5347	19115.97	1	1	1	111
982	2	5806	18583.98	1	1	1	111
400	1	4678	18382.42	1	1	1	111
1229	1	4455	18272.01	1	1	1	111

The lowest recency, highest frequency and monetary value are our best customers. The above table highlights these top 10 customers based on their RFM class. These customers are loyal to the store and shall continue with the retailer irrespective of any promotions. Based on the segment each customer falls in, the retailer can pitch targeted campaigns to retain existing customers or reel in new customers.

9. **Cohort Analysis:** Cohort analysis helps to track groups of people over a period of time, to identify common patterns. One such application of cohort analysis is customer retention, or the stickiness of a customer, which would help evaluate health of a business.



In the above image, we can see that for the first few months there's a good influx of customers who regularly purchase from the retailer. A year after the first purchase, there is a 60% retention. This might be a cohort of dedicated customers, who first joined the platform based on some already-existing connections with the retailer. After the fourth month, however, we do not see any new customers purchasing from the retailer. The retailer may need to focus on effective marketing campaigns to gain new customers in future.

10. **Market Basket Analysis:** Market basket analysis is one of the applications of Association Analysis which looks at retail sales data and determines what products are purchased together. It is a key technique to analyze customer buying habits by finding associations between the different items that customers place in their "shopping baskets".The discovery of these associations can help retailers develop targeted marketing strategies by gaining insight into which items are frequently purchased

together by customers. In order to find out interesting rules out of multiple possible rules, we will be using the following matrices:

1. **Support:** It's the default popularity of an item. In mathematical terms, the support of item A is the ratio of transactions involving A to the total number of transactions.
2. **Confidence:** Likelihood of the customer buying both A and B. It divides the number of transactions involving both A and B by the number of transactions involving B.
3. **Lift :** Increase in the sale of A on selling B.So, likelihood of a customer buying both A and B together is 'lift-value' times more than the chance of purchasing alone.

Out[49]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1	(ALL FAMILY CEREAL)	(FLUID MILK WHITE ONLY)	0.037098	0.221749	0.024081	0.649118	2.927266	0.015855	2.217983
3	(BANANAS)	(FLUID MILK WHITE ONLY)	0.109601	0.221749	0.056571	0.516154	2.327649	0.032267	1.608467
7	(CONDENSED SOUP)	(FLUID MILK WHITE ONLY)	0.041608	0.221749	0.021318	0.512344	2.310468	0.012091	1.595900
9	(DAIRY CASE 100% PURE JUICE - O)	(FLUID MILK WHITE ONLY)	0.058665	0.221749	0.034505	0.588163	2.652383	0.021496	1.889706
11	(EGGS - LARGE)	(FLUID MILK WHITE ONLY)	0.047822	0.221749	0.027333	0.571547	2.577454	0.016728	1.816423
13	(EGGS - X-LARGE)	(FLUID MILK WHITE ONLY)	0.034360	0.221749	0.020041	0.583263	2.630288	0.012422	1.867488
14	(IWS SINGLE CHEESE)	(FLUID MILK WHITE ONLY)	0.041221	0.221749	0.022504	0.545933	2.461944	0.013363	1.713957
17	(KIDS CEREAL)	(FLUID MILK WHITE ONLY)	0.046466	0.221749	0.030403	0.654316	2.950709	0.020100	2.251338
18	(MAINSTREAM WHEAT/MULTIGRAIN BR)	(FLUID MILK WHITE ONLY)	0.050748	0.221749	0.026533	0.522842	2.357813	0.015280	1.631015
21	(MAINSTREAM WHITE BREAD)	(FLUID MILK WHITE ONLY)	0.096989	0.221749	0.050813	0.523904	2.362599	0.029306	1.634651
33	(SNACK CAKE - MULTI PACK)	(FLUID MILK WHITE ONLY)	0.039377	0.221749	0.020388	0.517773	2.334955	0.011656	1.613871
41	(YOGURT NOT MULTI-PACKS)	(FLUID MILK WHITE ONLY)	0.046661	0.221749	0.024396	0.522828	2.357747	0.014049	1.630965

As can be seen above, market basket analysis can help the retailer find interesting product relationships. A product association with a high value of confidence and lift means that the combo occurs more frequently than would be expected given the number of transaction and product combinations. In such instances, retailers can make use of the above analysis to target promising product combinations to increase sales. This can be done by changing the store layout based on trends, cross marketing, promotion of trending itemsets or customized emailing.