DS8003 - Final Project

The Complete Journey

Submitted By:-

Akshdeep Kaler

Li Gong

Problem

The retail stores sell products to customers, and they would like to retain their customers and sell more items. For that purpose, various campaign strategies were launched. To know if the employed strategy is working, it is important to know that what elements of the strategy steer the business in right direction and which of them are not producing the desired output.

Proposed Solution

By answering following questions:

1. What are the characteristics of customers whose spending at the store is increasing?
2. What are the categories of the products that are seeing increased/decreased sales?
3. What are the most profitable categories of the products over time?
4. Which day has the highest sales?
5. Are the marketing campaigns effective?
6. Which of the marketing campaigns was the most successful one?
7. What are the characteristics of customers who were attracted by each marketing campaign?

Work Distribution:
(1-4) Akshdeep Kaler
(5-7) Li Gong

Dataset

Title: **The Complete Journey**, made available by **Dunhumby**
The dataset contains household level, anonymized, transaction data (2500 households) including the demographics and marketing campaigns (30 campaigns). The transactional data include over 90 thousand products categorized in 44 departments. The dataset is available at https://www.kaggle.com/frtgnn/dunnhumby-the-complete-journey.

Datasets Descriptions:

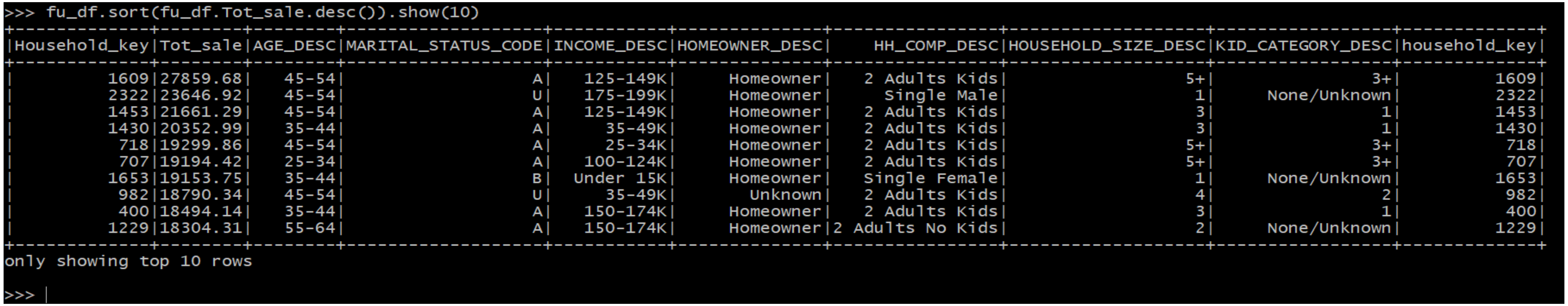
1. hh_demographic: The table contains demographic information for a portion of households.
2. transaction_data: This contains all products purchased by households.
3. campaign_table: This table lists the campaigns received by each household in the study.
4. campaign_desc: This table gives the length of time for which a campaign runs. Any coupons received as part of a campaign are valid within dates contained in this table.
5. product: This table contains information on each product sold such as type of product, national or private label and a brand identifier.
6. coupon: This table list all the coupons sent to customers as part of a campaign, as well as the products for which each coupon is redeemable.
7. coupon_redempt: This table identifies the coupons that each household redeemed.
8. casual_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).

What are the characteristics of customers whose spending at the store is increasing?

Datasets: **transaction_data.csv** , **hh_demographic.csv**

Tools:

- Hadoop Distributed File System (HDFS) (Storage of data file)
- Pyspark (file transformation)
- Pyspark SQL (creating table, Querying)



```
>>> fu_df.sort(fu_df.Tot_sale.desc()).show(10)
+-------------+--------+--------+-------------------+-----------+--------------+----------------+-------------------+-----------------+-------------+
|Household_key|Tot_sale|AGE_DESC|MARITAL_STATUS_CODE|INCOME_DESC|HOMEOWNER_DESC|    HH_COMP_DESC|HOUSEHOLD_SIZE_DESC|KID_CATEGORY_DESC|household_key|
+-------------+--------+--------+-------------------+-----------+--------------+----------------+-------------------+-----------------+-------------+
|         1609|27859.68|   45-54|                  A|   125-149K|     Homeowner|   2 Adults Kids|                 5+|               3+|         1609|
|         2322|23646.92|   45-54|                  U|   175-199K|     Homeowner|     Single Male|                  1|     None/Unknown|         2322|
|         1453|21661.29|   45-54|                  A|   125-149K|     Homeowner|   2 Adults Kids|                  3|                1|         1453|
|         1430|20352.99|   35-44|                  A|     35-49K|     Homeowner|   2 Adults Kids|                  3|                1|         1430|
|          718|19299.86|   45-54|                  A|     25-34K|     Homeowner|   2 Adults Kids|                 5+|               3+|          718|
|          707|19194.42|   25-34|                  A|   100-124K|     Homeowner|   2 Adults Kids|                 5+|               3+|          707|
|         1653|19153.75|   35-44|                  B|  Under 15K|     Homeowner|   Single Female|                  1|     None/Unknown|         1653|
|          982|18790.34|   45-54|                  U|     35-49K|       Unknown|   2 Adults Kids|                  4|                2|          982|
|          400|18494.14|   35-44|                  A|   150-174K|     Homeowner|   2 Adults Kids|                  3|                1|          400|
|         1229|18304.31|   55-64|                  A|   150-174K|     Homeowner|2 Adults No Kids|                  2|     None/Unknown|         1229|
+-------------+--------+--------+-------------------+-----------+--------------+----------------+-------------------+-----------------+-------------+
only showing top 10 rows

>>> |
```

What are the categories of the products that are seeing increased/decreased sales?

Datasets: **transaction_data.csv** and **product.csv**

Tools:

- Hadoop Distributed File System (HDFS) (Storage)
- Hive (Creating Table, Querying)

Output:

The categories which had the lowest sales are **Elect&Plumbing**, Gro Bakery, Housewares, Meat-WHSE, Prod-WHS Sales, HBC, Toys and Pork.

The categories which has the highest sales are **Grocery**, Drug GM, Produce, Meat, Kiosk-Gas, Meat-Pckgd, Deli, Pastry, Misc Sales Tran and Nutrition.

What are the most profitable categories of the products over time?

Grocery are most profitable category in the stores with sale of around 4,093,814 dollars.

Which day in two years period has the highest sales?

Datasets: **transaction_data.csv**

Tools:

- Hadoop Distributed File System (HDFS) (Storage)
- Hadoop map-reduce (processing)

Output:

The output shows that 641th day of 2-year period had the highest sales of 24,740.1 dollars. The 641th day falls in the month of October.

Insights Description:

- Mostly the customers who spend more on the store are of **45-54 year** age group are married and have kids. It could be the reason behind the higher sales of the **grocery** items in the stores.
- As the grocery covers around 50% of total sales and store has $24,740.1 dollar of highest sale in a day, the store need to originate up with the more ideas to increase the sales of other product.

Are the marketing campaigns effective?

Datasets: **transaction_data.csv**, **coupon_redempt.csv**

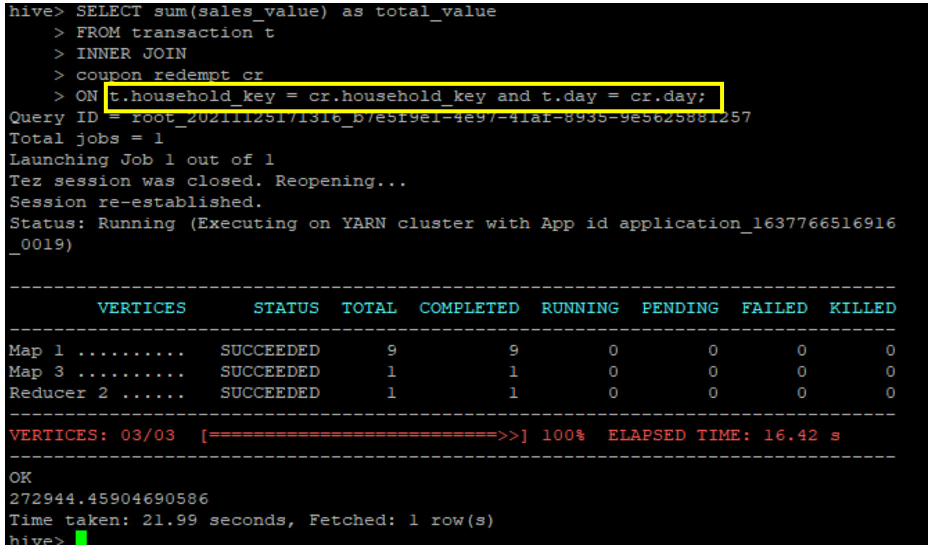
Tools:

- Hadoop Distributed File System (HDFS) (Storage)
- Hive (Querying)

Compare the purchases with redeemed coupons and without coupons.

- Find **total sales/quantities** of households (transaction_data.csv) who redeemed coupons (coupon_redempt.csv) -- A
- Find total sales/quantities of all households (transaction_data.csv) -- B
- Calculate the promotion rate by sales and quantities -- A/B%

Promotion Rate	
Sales	3.3%
Quantities	1.2%



```
hive> SELECT sum(sales_value) as total_value
    > FROM transaction t
    > INNER JOIN
    > coupon_redempt cr
    > ON t.household_key = cr.household_key and t.day = cr.day;
Query ID = root_20211125171316_b7e5f9e1-4e97-41af-8935-9e5625881257
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1637766516916
_0019)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      9          9        0        0       0       0
Map 3 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.42 s
----------------------------------------------------------------------------------------------
OK
272944.45904690586
Time taken: 21.99 seconds, Fetched: 1 row(s)
hive>
```

Insights:

- Marketing campaigns are more effective on higher price products.

Which of the marketing campaigns was the most successful one?

Datasets: **coupon_redempt.csv**, **campaign_desc.csv**, **transaction_data.csv**

Tools:

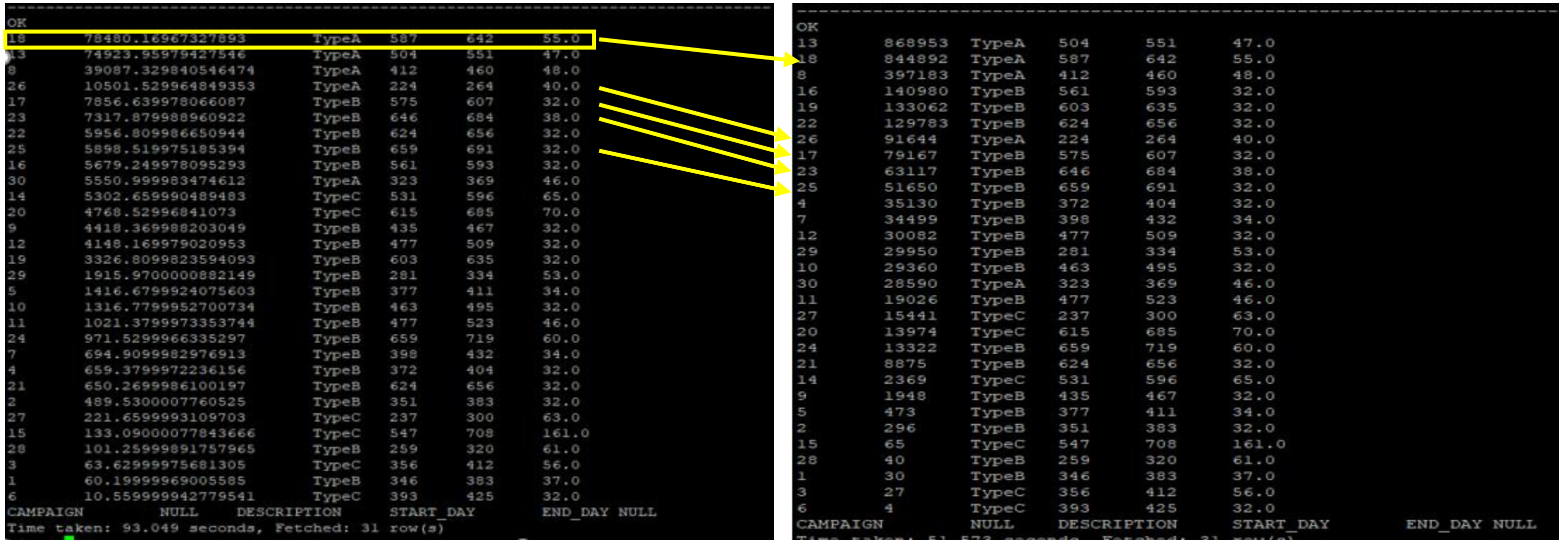
- Hadoop Distributed File System (HDFS) (Storage)
- Hive (Querying)

Compare the campaigns with their redeemed coupons

- Aggregate total sales/quantities (transaction_data.csv) of each campaign with redeemed coupons (coupon_redempt.csv)
- Link to campaign information (campaign_desc.csv)

Insights:

- No.18 campaign is the best. Top 4 campaigns were mostly type A. Type B campaigns performed in the medium level. Type C came last.
- Products with higher price are sensitive to marketing campaigns.



```
--------------------------------------------------------------------------------
OK
18      78480.16967327893       TypeA   587     642     55.0
13      74923.95979427546       TypeA   504     551     47.0
8       39087.329840546474      TypeA   412     460     48.0
26      10501.529964849353      TypeA   224     264     40.0
17      7856.639978066087       TypeB   575     607     32.0
23      7317.879988960922       TypeB   646     684     38.0
22      5956.809986650944       TypeB   624     656     32.0
25      5898.519975185394       TypeB   659     691     32.0
16      5679.249978095293       TypeB   561     593     32.0
30      5550.999983474612       TypeA   323     369     46.0
14      5302.659990489483       TypeC   531     596     65.0
20      4768.52996841073        TypeC   615     685     70.0
9       4418.369988203049       TypeB   435     467     32.0
12      4148.169979020953       TypeB   477     509     32.0
19      3326.8099823594093      TypeB   603     635     32.0
29      1915.9700000882149      TypeB   281     334     53.0
5       1416.6799924075603      TypeB   377     411     34.0
10      1316.7799952700734      TypeB   463     495     32.0
11      1021.3799973353744      TypeB   477     523     46.0
24      971.5299966335297       TypeB   659     719     60.0
7       694.9099982976913       TypeB   398     432     34.0
4       659.3799972236156       TypeB   372     404     32.0
21      650.2699986100197       TypeB   624     656     32.0
2       489.5300007760525       TypeB   351     383     32.0
27      221.6599993109703       TypeC   237     300     63.0
15      133.09000077843666      TypeC   547     708     161.0
28      101.25999891757965      TypeB   259     320     61.0
3       63.62999975681305       TypeC   356     412     56.0
1       60.19999969005585       TypeB   346     383     37.0
6       10.559999942779541      TypeC   393     425     32.0
CAMPAIGN        NULL    DESCRIPTION     START_DAY       END_DAY NULL
Time taken: 93.049 seconds, Fetched: 31 row(s)
```

By Sales

```
--------------------------------------------------------------------------------
OK
13      868953  TypeA   504     551     47.0
18      844892  TypeA   587     642     55.0
8       397183  TypeA   412     460     48.0
16      140980  TypeB   561     593     32.0
19      133062  TypeB   603     635     32.0
22      129783  TypeB   624     656     32.0
26      91644   TypeA   224     264     40.0
17      79167   TypeB   575     607     32.0
23      63117   TypeB   646     684     38.0
25      51650   TypeB   659     691     32.0
4       35130   TypeB   372     404     32.0
7       34499   TypeB   398     432     34.0
12      30082   TypeB   477     509     32.0
29      29950   TypeB   281     334     53.0
10      29360   TypeB   463     495     32.0
30      28590   TypeA   323     369     46.0
11      19026   TypeB   477     523     46.0
27      15441   TypeC   237     300     63.0
20      13974   TypeC   615     685     70.0
24      13322   TypeB   659     719     60.0
21      8875    TypeB   624     656     32.0
14      2369    TypeC   531     596     65.0
9       1948    TypeB   435     467     32.0
5       473     TypeB   377     411     34.0
2       296     TypeB   351     383     32.0
15      65      TypeC   547     708     161.0
28      40      TypeB   259     320     61.0
1       30      TypeB   346     383     37.0
3       27      TypeC   356     412     56.0
6       4       TypeC   393     425     32.0
CAMPAIGN        NULL    DESCRIPTION     START_DAY       END_DAY NULL
Time taken: 51.573 seconds, Fetched: 31 row(s)
```

By Quantities

What are the characteristics of customers who were attracted by each marketing campaign?

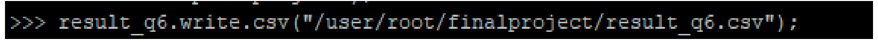
Datasets : **hh_demographic.csv, coupon_redempt.csv**

Tools:

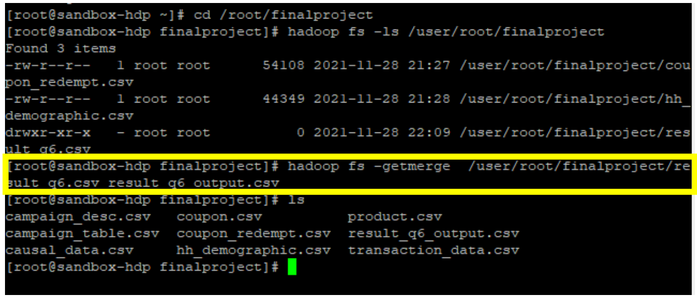
- Pyspark : create tables, query tables, output csv file
- Hadoop Distributed File System (HDFS) : store distributed data, output query result
- Kibana: visualize query result

Find campaigns with customer demographic

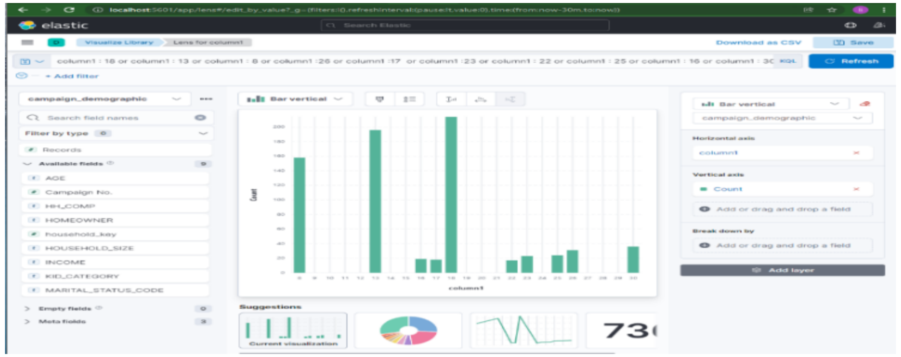
- Pyspark: Group data by campaign number and household ID (coupon_redempt.csv), Link to custom demographic (hh_demographic.csv)
- HDFS: Write query result into csv file. Merge and output result files from HDFS to local directory.
- Kibana: Import query result into Kibana. Set up Elasticsearch index pattern. Filter data by searching top 10 campaigns as we found in previous question, and apply the condition on dashboard.



```
>>> result_q6.write.csv("/user/root/finalproject/result_q6.csv");
```



```
[root@sandbox-hdp ~]# cd /root/finalproject
[root@sandbox-hdp finalproject]# hadoop fs -ls /user/root/finalproject
Found 3 items
-rw-r--r--   1 root root      54108 2021-11-28 21:27 /user/root/finalproject/cou
pon_redempt.csv
-rw-r--r--   1 root root      44349 2021-11-28 21:28 /user/root/finalproject/hh_
demographic.csv
drwxr-xr-x   - root root          0 2021-11-28 22:09 /user/root/finalproject/res
ult_q6.csv
[root@sandbox-hdp finalproject]# hadoop fs -getmerge  /user/root/finalproject/re
sult_q6.csv result_q6_output.csv
[root@sandbox-hdp finalproject]# ls
campaign_desc.csv   coupon.csv          product.csv
campaign_table.csv  coupon_redempt.csv  result_q6_output.csv
causal_data.csv     hh_demographic.csv  transaction_data.csv
[root@sandbox-hdp finalproject]# 
```



localhost:5601/app/lens#/edit_by_value?_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:now-30m,to:now))

column1 : 18 or column1 : 13 or column1 : 8 or column1 :26 or column1 :17 or column1 :23 or column1 : 22 or column1 : 25 or column1 : 16 or column1 : 3

Index pattern: campaign_demographic

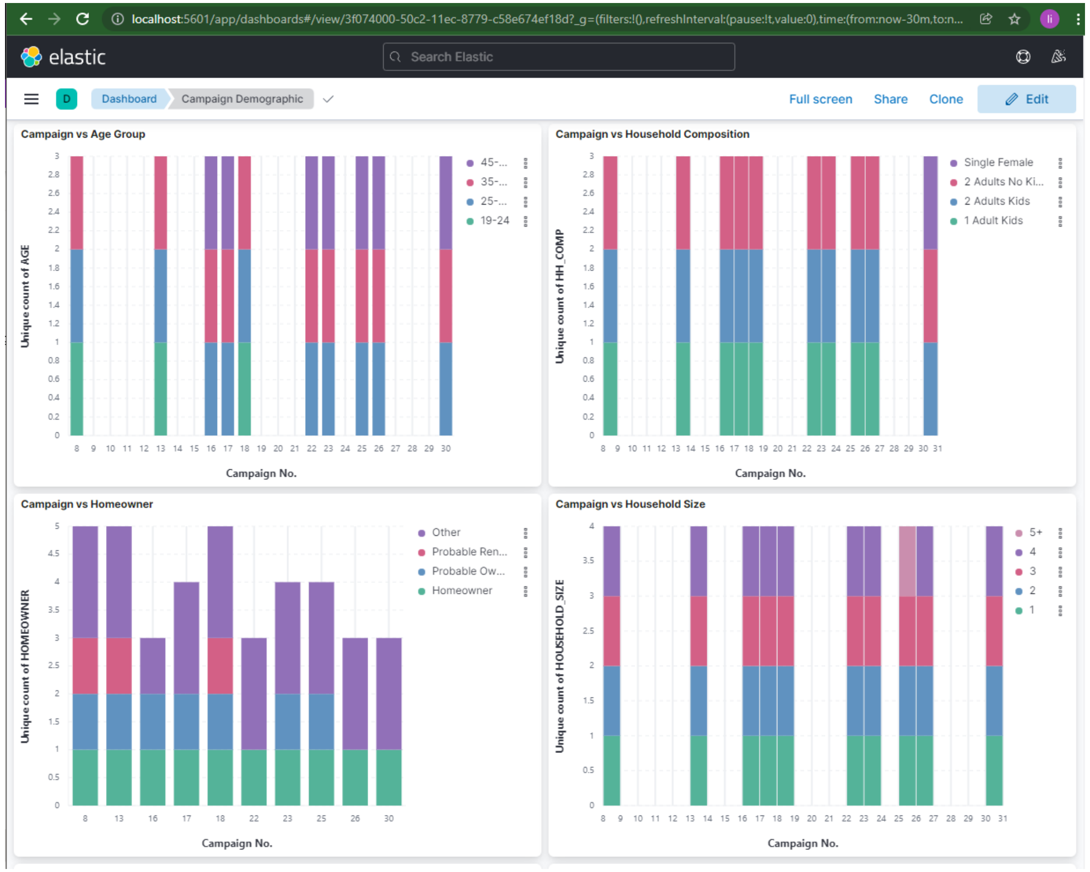
Available fields: AGE, Campaign No., HH_COMP, HOMEOWNER, household_key, HOUSEHOLD_SIZE, INCOME, KID_CATEGORY, MARITAL_STATUS_CODE

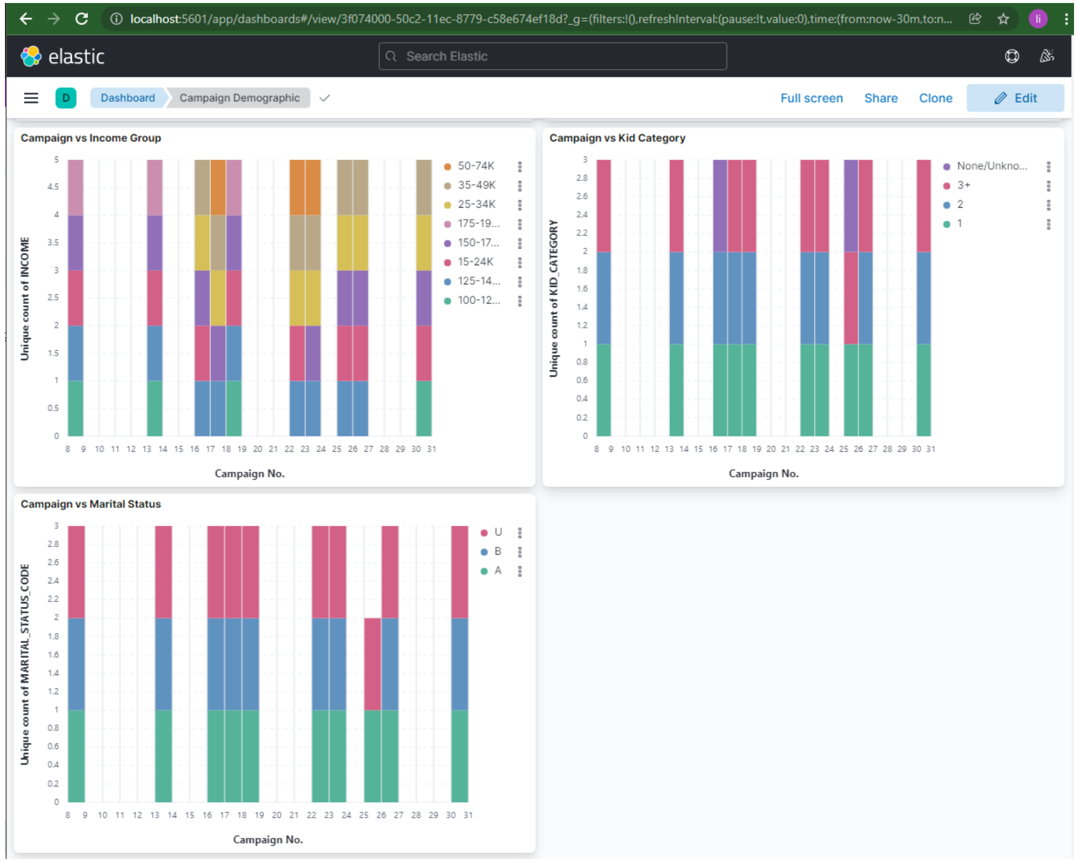
Bar vertical — Horizontal axis: column1 — Vertical axis: Count

What are the characteristics of customers who were attracted by each marketing campaign?

Insight:

Most of the campaign fans are between the ages of 45-54, have families consisting of two adults with more than three children, rent rather than own, and have incomes between $35k and $49k.





Insights

- Mostly the customers who spend more on the store are of 45-54 year age group have kids and are married. It could be the reason behind the higher sales of the grocery items in the stores.
- As the grocery covers around 50% of total sales and store has $24,740.1 dollar of highest sale in a day, the store need to originate up with the more ideas to increase the sales of other product.
- Marketing campaigns are more effective on higher price products due to the larger increase of sales than quantities.
- The campaigns with optimal performance mostly last about 45-50 days. We don’t recommend more than 55 days.
- Customers between age above 45, with three more children, and income between $35k and $49k are more likely to enrolled in the marketing campaigns.

Lessons we learned

- When joining tables with many_to_many keys, care should be taken with adding join keys to ensure that no more records are created to avoid expanding the total values.
- More query problems are best served by Hive, as it is an SQL interface operating on Hadoop. Obviously, it has its own database to store structured tables.
- Spark supports programming languages like python, making it easier and faster to do data analysis. For example, writing output to a csv file is much easier than Hive. However, it works with RDDs instead of tables, and we should convert them to views before querying.
- Kibana has limit on importing local files (not larger than 100MB), so it is better to import query results for visualization.

Future work

- We will consider causal dataset to exclude the effects of other occurring events, such as in-store display, weekly mailer, for each campaign.

Thank you !