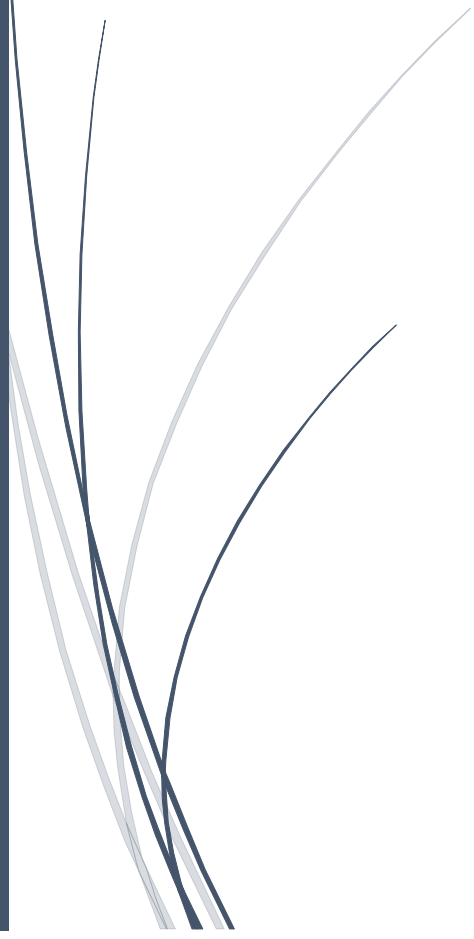
PROJECT REPORT

1st April 2023

CAPSTONE GROUP G6

The Complete Journey - Dunnhumby



PGP – DSE OCT`22

PROJECT SUMMARY

Batch Details	PGP-DSE OCT'22
Team Members	Samarth Kumar, Deepali Joshi, Himani Kandpal, Astha Aggarwal, Harsh Vardhan Chaudhary, Aman Saxena, Ishaan Nagwani
Domain of Project	Retail
Proposed Project Title	"Coupon Foresight" - Seeing Coupons in a New Light
Group Number	G6
Team Leader	Ishaan Nagwani
Mentor Name	Mr. Jatinder Bedi

Date: 1st April 2023

Signature of the Mentor

Signature of the Team Leader

Contents

EXECUTIVE SUMMARY

OVERVIEW

The data provided includes two years' worth of transactional information for 2500 households, as well as demographic information for some of those households.

Additionally, information on product categories and promotional marketing for 30 campaigns that took place within the same two-year period is also included.

APPROACH

To understand how customers, interact with a business, a comprehensive analysis of sales data is conducted. This analysis aims to identify trends that may negatively impact the business's growth, as well as trends that may contribute to its growth.

The insights gained from this analysis are data-driven and can assist management in making informed decisions about the future direction of the business.

An in-depth analysis of direct marketing data is carried out to study the effectiveness of promotional activities done by the retailer.

We've tried to mine all types of patterns from the engagement of customers in those marketing campaigns. We've also attempted to bring out loopholes and ineffective practices that lead to underutilization of resources.

OBJECTIVES

- ✓ To study consumer behavior
- ✓ To mine important business trends or patterns (negative or positive)
- ✓ To study effectiveness of promotional activities
- ✓ Gain deeper insights on campaign's and coupon strategy
- ✓ To identify the households with highest sensitivity to redemption of coupons through medium of a Machine Learning Model.
- ✓ To identify products and product categories to bring out insights on top performing and low performing products, with respect to promotional activities, through medium of a Machine Learning Model.
- ✓ Building models to predict a coupon redemption and to design better promotions

METHODOLOGY FOLLOWED

1. Business Understanding - Wherein we've delved into the various ways that this data can be leveraged to drive sales growth and improve overall business performance. Analyzing the engagement of customers with the business was our motive, build insights into the business audience and help the client to make more informed and data-driven decisions.

2. Data Understanding – The whole dataset is divided into 8 different tables. The schema below shows that it is a relational database and all the tables. The target variable was not given and hence we've combined data from multiple tables to obtain our target variable. We see that in any scenario we just want to predict whether the coupon is redeemed (1) or not (0). Hence, it is a classification problem.

3. Data Preparation – We've addressed the coupon redemption predictive modelling with respect to household demographics and product details.

   As we were dealing with a relational database with 8 tables, our first and most important task was to consider all the given data and carefully merge all the tables.

   After combining the data, we created the target variable. And after creation of target variable we started to mine features from other tables.

   We've used multiple tables to mine features for the final variable creation for the problem statements.

   Due to lack of unique identifiers in most of the tables, the data from such tables needs to be aggregated by a keyword such as the mean_purchase_value of each household using customer_id from customer transaction data and demographic data.

   We've used aggregation throughout the process of mining the features for predictive modelling data.

   The most important part of feature engineering process was to combine information from multiple tables. For example, if we were to get information from Customer Transaction Data about products, we need to aggregate the transaction table by product_id.

4. Modeling – While proceeding further with the modelling process we first figured out the answer to the following question:

   Q. What was more important to predict between two classes in the target variable?

   The answer is positive class. Hence, the scoring parameter we need to look into for correct judgement is recall score.

KEY FINDINGS

Coupons have been a popular marketing tool for decades, but the dataset tells us that they are not always effective. In fact, over 70% of coupons go unused or unredeemed or simply got wasted.

This is a significant waste of time and money for the retailer who invested capital in promotional campaigns and promotions.

However, by analyzing the transactions where coupons were redeemed, we can identify the characteristics of consumers and products that are highly sensitive to coupon redemption.

This information can inform future campaigns and help companies target the right audience with their promotions. The dataset also shows that the campaigns were not distributed effectively, which further highlights the need for a more targeted approach.

By optimizing promotional campaigns based on coupon redemption data, companies can improve their ROI and avoid underutilization of resources.

INTRODUCTION TO THE DATASET

The dataset contains household level transactional data of 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 some selected categories. The entire transactional activity of those households at the retailer’s outlet is included.

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

Besides actual transactional data of all these households, the dataset covers every possible related sales detail, such as product details & product store-display data.

It also covers data about the promotional marketing activities of the retailer, spanning over a period of 2 years.

These activities include running campaigns and giving out coupons. The coupons are given to households for redemption, and the final goal of the campaign is to promote certain products and boost revenue.

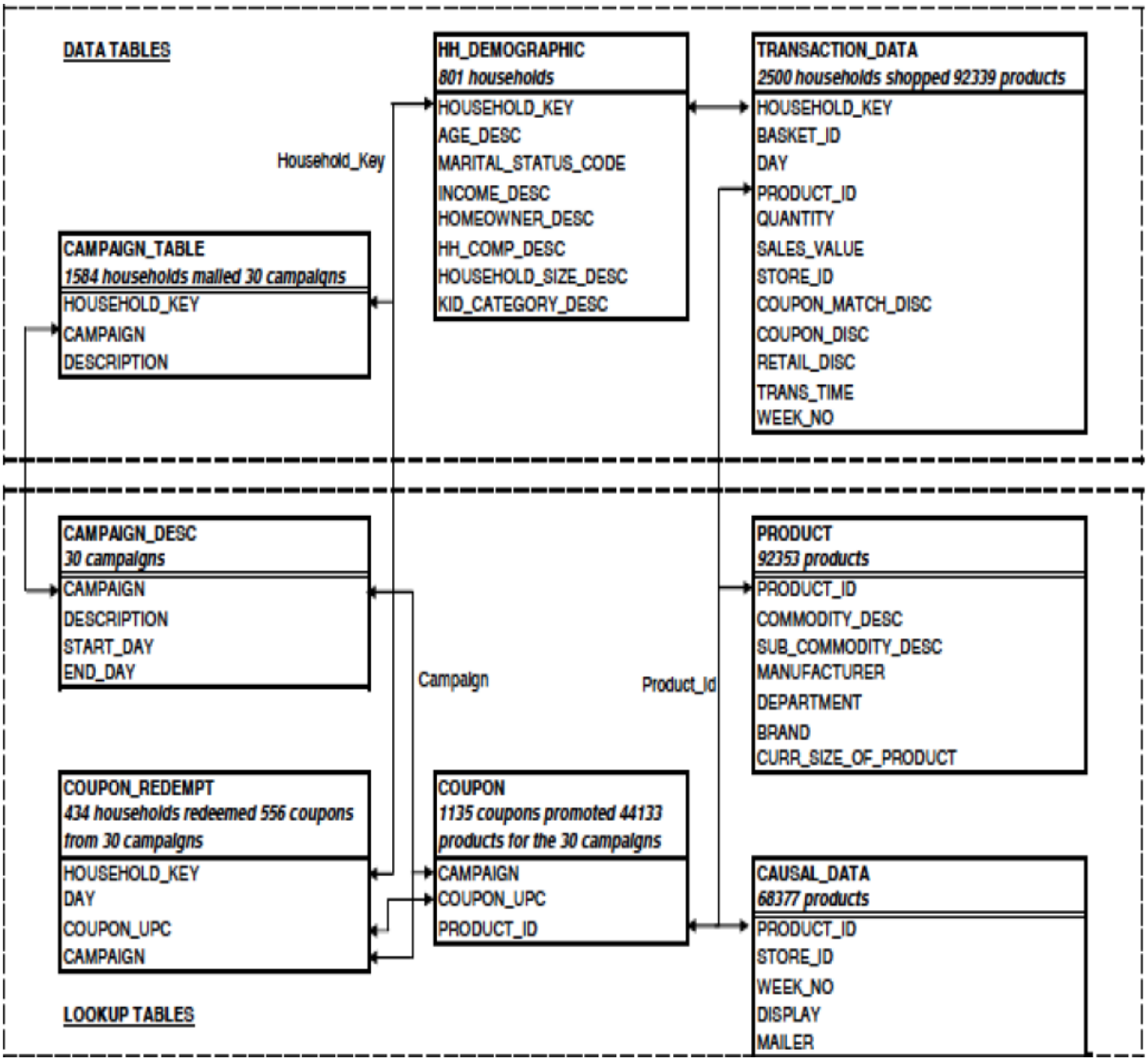
In today's digitally driven world, data is the cornerstone of successful businesses. While analyzing consumer behavior, having access to comprehensive data is crucial.

Without it, businesses can miss out on key insights that could help them improve their sales strategies and target their marketing efforts more effectively.

By looking into this kind of a data, companies can understand their target audience better and make more informed decisions.

This can eventually lead to growth and more revenue. So, let's take a closer look at what this dataset has to offer and how businesses can benefit from it.

DATA DIAGRAM



DIRECTION OF ANALYSIS & BUSINESS PROBLEM STATEMENT

Throughout the course of our analysis we've addressed the data in the following ways:

- General Data Analysis –

  We've delved into the various ways that this data can be leveraged to drive sales growth and improve overall business performance. Analyzing the engagement of customers with the business was our motive, build insights into the business audience and help the client to make more informed and data-driven decisions.

- Prediction of Coupon Redemption (with respect to households) –

  Here we are analyzing the retailer's direct marketing efforts through promotional campaigns that the business runs to boost sales and promote specific products.

  Every marketing and promotional activity requires infusion of funds because it involves additional costs and investment into the medium (here campaigns and coupons).

  In order to be successful with the promotional activities, any business needs to make sure that the audience they are trying to reach through their campaigns is the right audience. They need to make sure that the audience they are trying to reach is actually interested in the promotions.

  Here, if the retailer is able to identify the audience, that can actively engage in redemption of coupons, the marketing efforts can then bring out a high success rate and the retailer can achieve a better return on invested capital, that has been made on running promotional campaigns and distribution of discount coupons.

  The dataset shows that 70% of the customers never use the coupons they receive and this leads to a waste of money and time for the company. Hence, if there was a way to identify the ideal audience for promotions, it would help to use resources more effectively and would significantly help grow the business.

- Prediction of Coupon Redemption (with respect to products) –

  This side of the problem statement entirely focuses on the selection of products or product categories that ought to be highlighted within promotional campaigns.

  This problem statement highlights the importance of effectively selecting the right products for promotional campaigns in order to maximize their impact and drive sales.

  It is crucial to identify the products that are in high demand or have the potential to become popular. Proper selection will ensure effective marketing and allocation of resources, leading to successful campaigns.

  As such, careful analysis and research of market trends, consumer behavior, company goals, product margins and competitor strategies are essential in determining which products or product categories should be promoted for maximum revenue generation.

  By thoroughly analyzing these factors and researching the market, companies can develop effective promotional campaigns that target specific products or categories that are projected to yield the best results.

  The proper selection of products for promotional campaigns can lead to increased brand awareness, higher sales, and overall success of the campaign.

  Therefore, it is imperative for the retailer to carefully evaluate and consider which products or product categories to promote within their campaigns.

METHODOLOGY

BUSINESS UNDERSTANDING

In today's digital age, businesses are generating vast amounts of data, and it is crucial to analyze this data to gain insights into customer behavior and preferences. Analyzing customer engagement with a business can provide valuable insights that can be leveraged to drive sales growth and improve overall business performance. This data can help businesses understand what their customers are looking for, what they like and dislike, and how they interact with the business. By analyzing the engagement of customers with a business, businesses can build insights into their audience and make more informed and data-driven decisions.

There are various ways that this data can be leveraged to drive sales growth and improve overall business performance. For example, companies can use this data to find patterns and trends in customer behavior. This information can help them better target their audience. Additionally, businesses can use this data to personalize their customer experience by understanding their preferences and providing them with -offers. By leveraging this data, businesses can optimize their promotional activities, reduce customer churn, and ultimately drive sales growth.

In conclusion, analyzing the engagement of customers with a business is crucial in today's digital age. It provides valuable insights that businesses can leverage to drive sales growth and improve overall business performance. By understanding their audience's behavior and preferences, businesses can make more informed and data-driven decisions that can help them stay ahead of the competition.

PROJECT SUMMARY

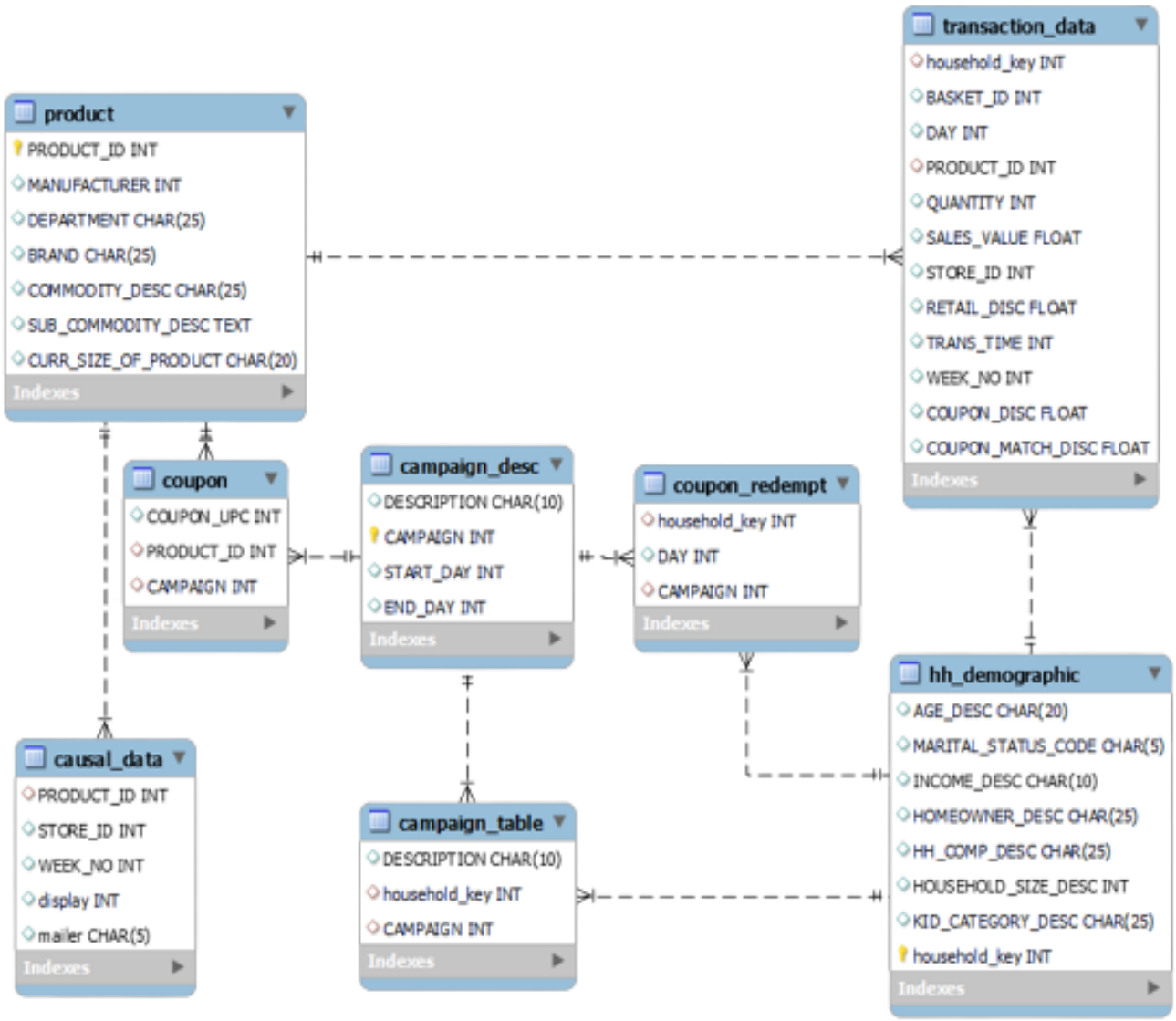
Running campaigns and giving out coupons are two common activities that businesses engage in to increase product promotion and revenue.

Coupons are often distributed to households for redemption, providing a tangible incentive for customers to purchase a product or service. Campaigns, on the other hand, may involve a variety of marketing tactics, such as social media advertising, email marketing, or influencer partnerships. Both strategies can be effective in capturing consumer attention and driving sales.

In conclusion, running campaigns and offering coupons are just two of the many tactics that businesses can use to promote their products and boost revenue. However, in order to succeed in today's competitive landscape, companies must also prioritize data analysis and use insights to guide their marketing efforts. By doing so, businesses can create more effective campaigns, improve customer engagement, and ultimately drive greater revenue.

DATA UNDERSTANDING

ER DIAGRAM



PROJECT SUMMARY

The dataset in question has been divided into 8 different tables, each containing unique and relevant information. As evident from the schema, it is a relational database where all the tables are interconnected.

However, the target variable was not provided, which led to the need for combining data from multiple tables to obtain the target variable.

After careful analysis, it was determined that the primary objective was to predict whether the coupon would be redeemed or not.

This essentially makes it a classification problem, where the output is binary, i.e., either 1 or 0.

The process of combining data from different tables to obtain the target variable can be complex, but with the right approach and techniques, it can be accomplished with high accuracy.

We carefully combined the data and figured out the right tables for target variable creation in both the datasets.

Feature creation also required merging and combining the data in a right manner so as to mine intelligence from related data for obtaining our final variable creation.

DATA PREPERATION

Throughout the course of our analysis we were required to prepare the data twice:

1. First while creation of data for Coupon Redemption Prediction (with respect to households)

2. Second while creation of data for Coupon Redemption Prediction (with respect to products)

Data Preparation for Coupon Redemption Prediction (with respect to households)

Data Considered:

1. campaign_table
2. coupon
3. coupon_redempt
4. hh_demographic
5. transaction_data

Based on the dataset, it is evident that 70% of customers fail to redeem the coupons they receive, leading to a significant waste of resources for the company. To address this issue, our primary goal is to develop a classification model that can accurately predict whether or not a customer will redeem their coupon. While identifying customers who are likely to redeem their coupons is important, it's equally crucial for companies to pinpoint those who won't redeem them. This information can help companies tailor their marketing and communication strategies accordingly or even decide to withhold coupons altogether and save money in the process.

```
Total No. Of Households : 2500
No. Of Households To Which Coupons Were Provided : 1584
No. Of Households Who Redmeed Coupons : 434
Percentage Of Households Where Coupons Were Unused : 72.6 %
```

Creation of Target Variable:

The coupon_redempt table consists of data of 434 unique households that redeemed a coupon when they received them in a particular campaign.

```
In [189]:
1 ##  coupon redempt by  households
2 coupon_redempt.household_key.nunique()

Out[189]: 434
```

```
In [23]:
1 coupon_redempt.head(7)
```

Out[23]:

	household_key	DAY	COUPON_UPC	CAMPAIGN
0	1	421	10000085364	8
1	1	421	51700010076	8
2	1	427	54200000033	8
3	1	597	10000085476	18
4	1	597	54200029176	18
5	8	422	53600000078	8
6	13	396	53700048182	5

```
In [19]:
1 hh_demographic.head(2)
```

Out[19]:

	AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC	household_k
0	65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	
1	45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	

We combined the coupon_redempt table and the hh_demographic table and hence found out that out of 434 households who there are 311 such households whose demographic details are included in the hh_demographic table. Hence all those households whose coupon redemption details were not included in coupon_redempt table we consider them as households who were not sensitive to coupons.

Hence, our target variable is formed.

1	Households that redeemed coupons
0	Households that didn't redeemed coupons

```
hh_demographic.iloc[:,[7,6,5,4,3,2,1,0]].head(7)
```

	household_key	KID_CATEGORY_DESC	HOUSEHOLD_SIZE_DESC	HH_COMP_DESC	HOMEOWNER_DESC	INCOME_DESC	MARITAL_STATUS_CODE	AGE_DES
0	1	None/Unknown	2	2 Adults No Kids	Homeowner	35-49K	A	6
1	7	None/Unknown	2	2 Adults No Kids	Homeowner	50-74K	A	45-
2	8	1	3	2 Adults Kids	Unknown	25-34K	U	25-
3	13	2	4	2 Adults Kids	Homeowner	75-99K	U	25-
4	16	None/Unknown	1	Single Female	Homeowner	50-74K	B	45-
5	17	None/Unknown	2	2 Adults No Kids	Homeowner	Under 15K	B	6
6	18	None/Unknown	2	2 Adults No Kids	Homeowner	100-124K	A	45-

```
# creating target variable.
hh_demographic['Target']=hh_demographic.household_key.apply(lambda x: 1 if x in coupon_redempt.household_key.unique() else 0)
hh_demographic.Target.value_counts()
```

```
0    490
1    311
Name: Target, dtype: int64
```

Creation of Features:

Next step after creation of target variable was to mine features from other tables.

In the process of mining features for predictive modelling data, we have used multiple tables. However, due to a lack of unique identifiers in most tables and the need to use data from such tables, we had to aggregate it using a keyword such as "cnt_camp_received_per_hsld". For instance, we aggregated the mean purchase value of each household by household_key from both hh_demographics and transaction_data to create the final variables. Aggregation has been a crucial part of our feature mining process.

1. Feature from campaign_table:

   Number of Campaigns received by unique households:

	household_key	cnt_camp_recieved_per_hsld
0	1	8
1	2	1
2	3	3
3	4	1
4	6	4

Statistical Test: Feature Significant

	F_onewayresult
Test_Statistics	2.112970e+03
pvalue	9.050725e-290

2. Feature from coupon_redempt

   household wise count of campaigns in which the coupon were redeemed

	household_key	distinct_camprdmptn_per_hsld
0	1	2
1	8	1
2	13	7
3	14	1
4	18	3

Statistical Test: Feature Significant

	F_onewayresult
Test_Statistics	7.039228e+01
pvalue	1.097288e-16

3. Coupon Redemption Rate : it is the ratio of no. of campaigns received and coupon redemption on those campaigns.

	household_key	cnt_camp_recieved_per_hsld	distinct_camprdmptn_per_hsld	camp_rdmptn_rate
2316	2317	17.0	3.0	0.176471
2488	2489	16.0	6.0	0.375000
717	718	15.0	5.0	0.333333

Statistical Test: Feature Significant

	F_onewayresult of camp_rdmptn_rate
Test_Statistics	1.750850e+02
pvalue	6.600895e-38

4. From transaction_data table:

   Total sales value per household: it is the sum of the purchased value made by each household.

	household_key	TOTAL_SALES_VALUE_hsld_wise
0	1	4330.16
1	2	1954.34
2	3	2653.21

Statistical Test: Feature Significant

	F_onewayresult of TOTAL_SALES_VALUE_hsld_wise
Test_Statistics	1.797826e+03
pvalue	7.944291e-260

5. Mean items purchase per transaction : average numbers of items purchased in a single transaction.

	household_key	mean_items_purch_per_trans
0	1	23.0
1	2	19.0
2	3	182.0

Statistical Test: Feature Significant

	F_onewayresult of mean_items_purch_per_trans
Test_Statistics	4.597601e+02
pvalue	2.609885e-89

6. Number of Total visits:

   Total number of visits made by unique households over the span of two years.

	household_key	No_of_total_visits
0	1	85
1	2	45
2	3	47

   Statistical Test: Feature Significant

	F_onewayresult of No_of_total_visits
Test_Statistics	1.403534e+03
pvalue	4.536211e-218

7. MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS:
   Average purchase made by unique households in a single transaction.

	household_key	MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
0	1	50.94
1	2	43.43
2	3	56.45

   Statistical Test: Feature Significant

	F_onewayresult
Test_Statistics	1.598726e+03
pvalue	2.108284e-239

8. MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS:

   Median purchase made by unique households in a single transaction.

	household_key	MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
0	1	49.33
1	2	26.94
2	3	36.38

   Statistical Test: Feature Significant

	F_onewayresult
Test_Statistics	1.598726e+03
pvalue	2.108284e-239

9. Price per product purchased:
   It is the average price of product generally purchased by unique households.

Out[225]:

	household_key	Price_per_product_purchased
0	1	2.31
1	2	2.53
2	3	1.98
3	4	3.40
4	5	3.34

   Statistical Test: Feature Significant

	F_onewayresult of Price_per_product_purchased
Test_Statistics	6913.179884
pvalue	0.000000

Summary of Final Variable Creation:

Out[227]:

	Tables Extracted	Information Derived
0	cnt_camp_recieved_per_hsld	count of campaign recieved per household
1	distinct_camprdmptn_per_hsld	coupons of distinct campaign redempt by specif...
2	cnt_of_camp_recieved_by_num_of_hsld	number of campaign recieved by number of house...
3	camp_rdmptn_rate	rate of redemption of each household
4	cpnrdmpt_campwise	Coupon redemption campaign wise
5	hsld_wise_ttlsales	total purchase done by each household during t...
6	No_of_items_in_single_trans	total number of transactions and number of pro...
7	mean_items_purch_per_trans	average number of products purchased per trans...
8	No_of_total_visits	total number of visits of each households
9	amt_spent_single_trans	amount spent in a single transaction
10	mean_spent_single_trans_hsldwise	mean spent of each household in single transac...
11	median_spent_single_trans_hsldwise	median spent of each household in single trans...
12	avg_productcst_purchased_hsldwise	average cost of products purchased by specific...

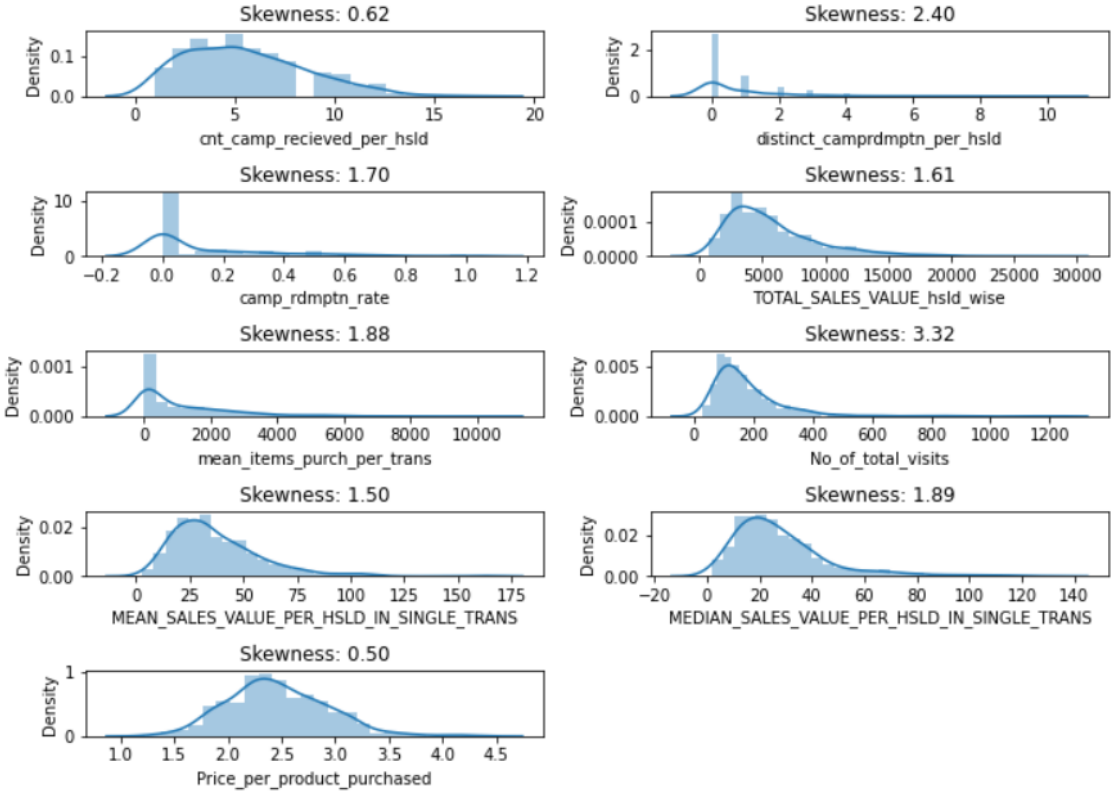
In [228]:
```
1 # Merging of all the features obtained during feature engineering to the obtain the main file for modeling.
```

Modeling:

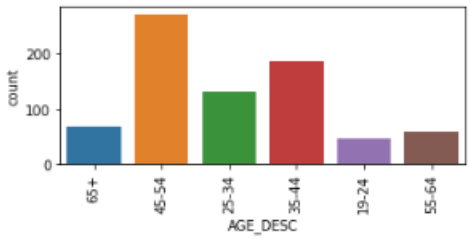
Analysis of the final dataset through which we are going to build the model.

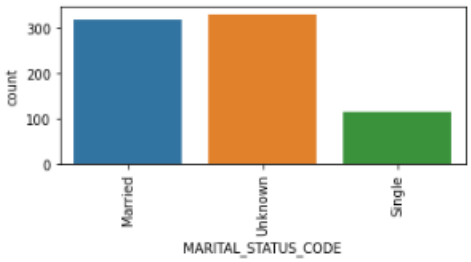
Univariate Analysis

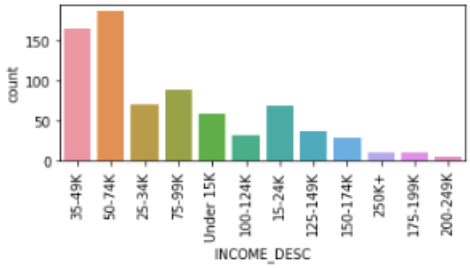
NUM COLS

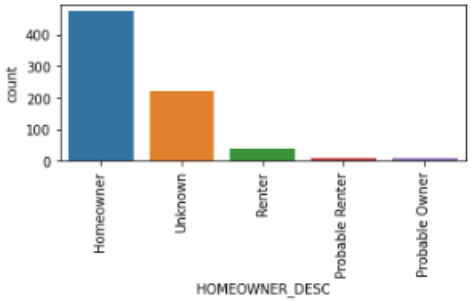


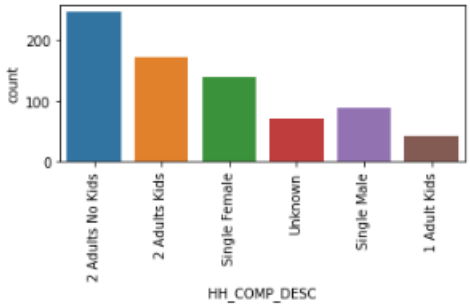
CAT COLS

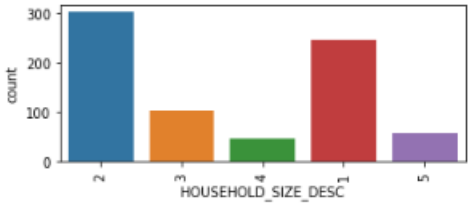

count
200
100
0
65+
45-54
25-34
35-44
19-24
55-64
AGE_DESC

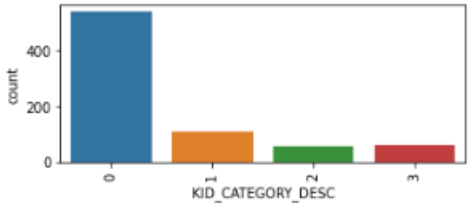

count
300
200
100
0
Married
Unknown
Single
MARITAL_STATUS_CODE


count
150
100
50
0
35-49K
50-74K
25-34K
75-99K
Under 15K
100-124K
15-24K
125-149K
150-174K
250K+
175-199K
200-249K
INCOME_DESC

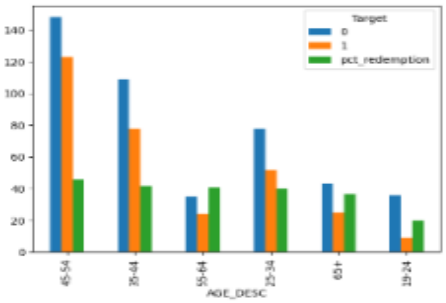

count
400
300
200
100
0
Homeowner
Unknown
Renter
Probable Renter
Probable Owner
HOMEOWNER_DESC

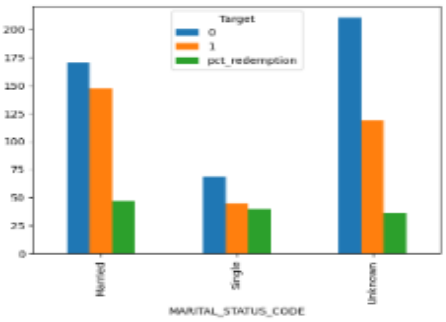

count
200
100
0
2 Adults No Kids
2 Adults Kids
Single Female
Unknown
Single Male
1 Adult Kids
HH_COMP_DESC

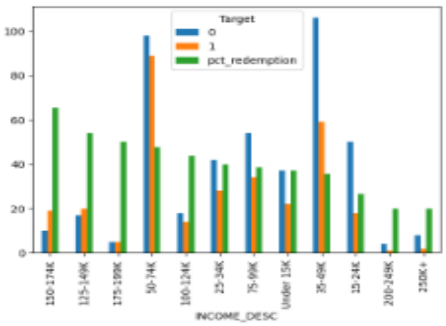

count
300
200
100
0
2
3
4
1
5
HOUSEHOLD_SIZE_DESC

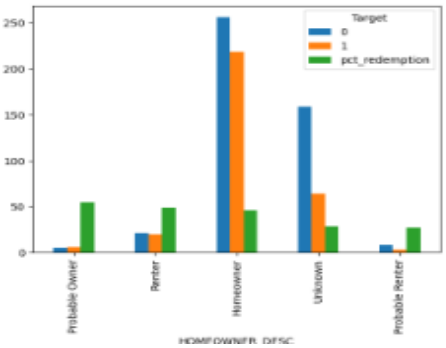

count
400
200
0
0
1
2
3
KID_CATEGORY_DESC

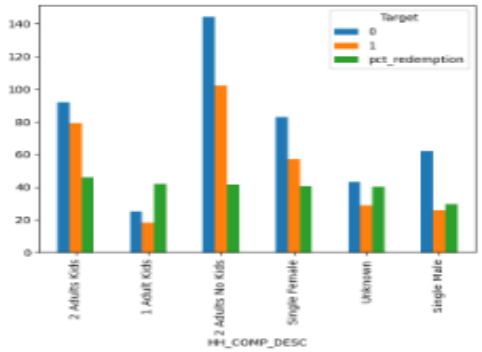
Bivariate Analysis

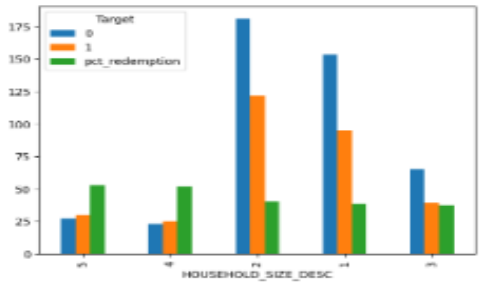


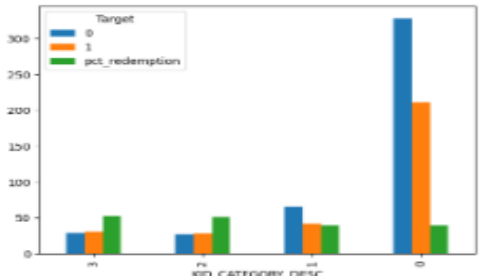


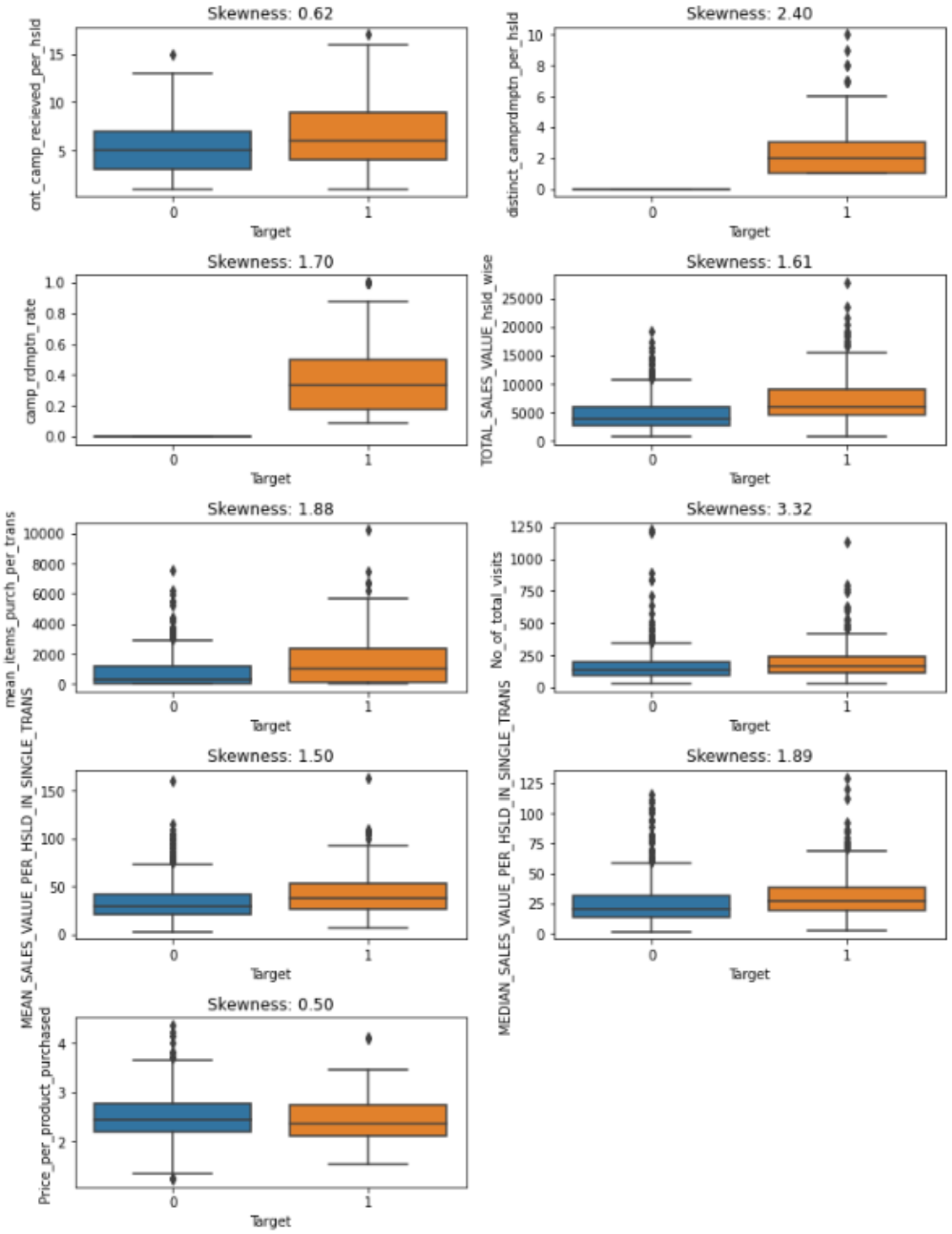




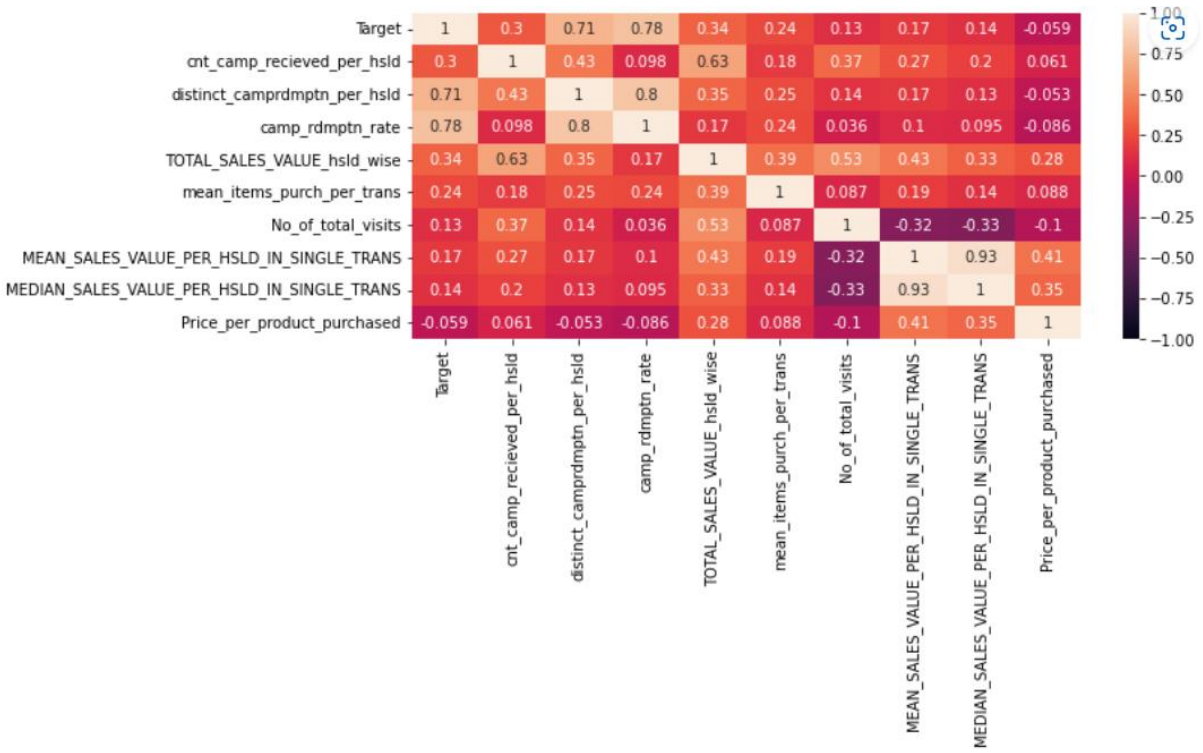







Skewness: 0.62
cnt_camp_recieved_per_hsld
Target
Skewness: 2.40
distinct_camprdmptn_per_hsld
Target
Skewness: 1.70
camp_rdmptn_rate
Target
Skewness: 1.61
TOTAL_SALES_VALUE_hsld_wise
Target
Skewness: 1.88
mean_items_purch_per_trans
Target
Skewness: 3.32
No_of_total_visits
Target
Skewness: 1.50
MEAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
Target
Skewness: 1.89
MEDIAN_SALES_VALUE_PER_HSLD_IN_SINGLE_TRANS
Target
Skewness: 0.50
Price_per_product_purchased
Target

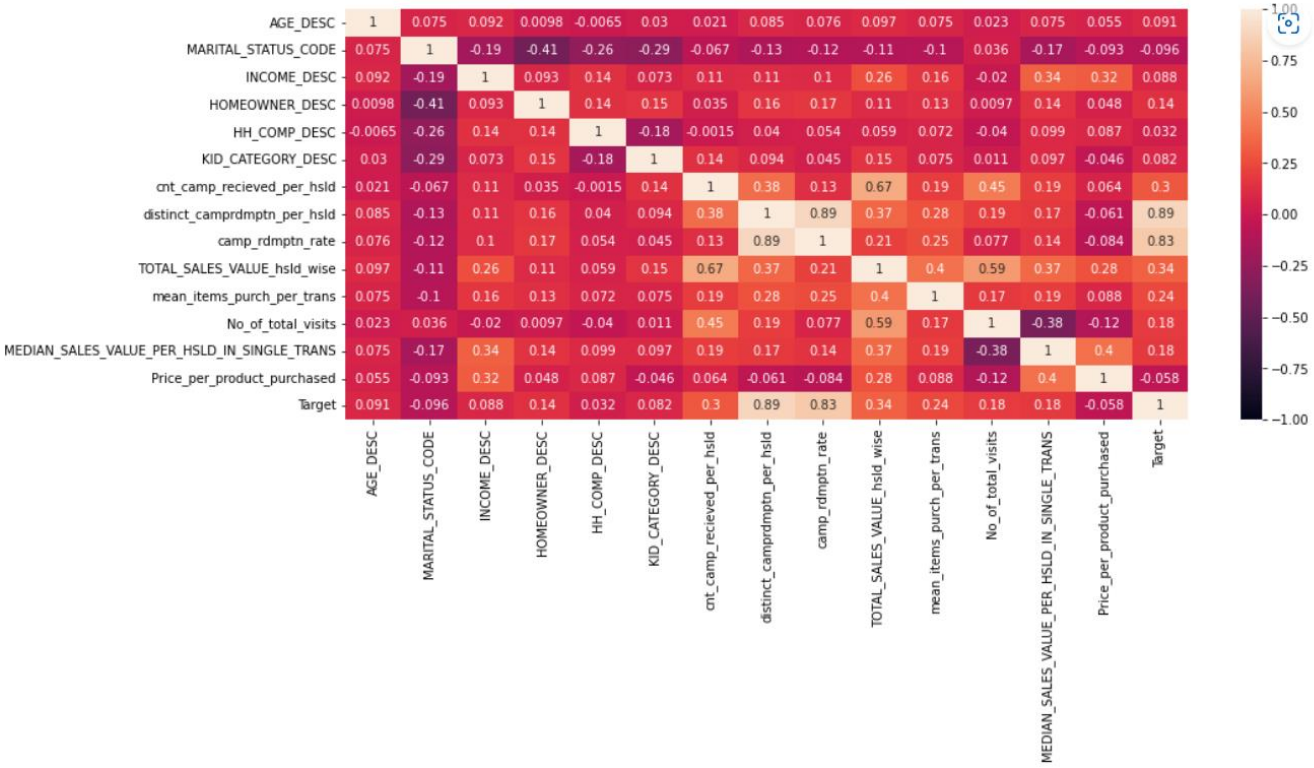
Correlation Plot Before Encoding and Scaling:



Data Preprocessing:

- ✓ Outliers Treatment
- ✓ Encoding of Categorical Variables
- ✓ Scaling of Numerical Variables
- ✓ Multicollinearity Treatment
- ✓ Train Test Split

Correlation Plot After Data Preprocessing:



Modeling:

Due to creation of many strong features the model tends to overfit.

```
In [309]: 1 dic1.sort_values(['accuracy_score','f1_score'],ascending=False)
```

Out[309]:

	Model	accuracy_score	f1_score	recall_score	precision_score
0	dt	1.000000	1.000000	1.000000	1.000000
1	nb	1.000000	1.000000	1.000000	1.000000
3	rfc	1.000000	1.000000	1.000000	1.000000
4	ada	1.000000	1.000000	1.000000	1.000000
5	gbm	1.000000	1.000000	1.000000	1.000000
6	xgb	1.000000	1.000000	1.000000	1.000000
2	knn	0.973684	0.967213	0.951613	0.983333

Data Preparation for Coupon Redemption Prediction (with respect to products)

Data Considered:

1. product
2. coupon
3. hh_demographic
4. transaction_data

This problem statement emphasizes the significance of strategically choosing products or product categories for promotional campaigns to optimize their impact and boost sales, particularly in terms of coupon redemption.

It is imperative to identify the products that are in high demand or have the potential to become popular and can be effectively promoted through coupons. Proper selection will ensure efficient marketing and allocation of resources, leading to successful campaigns with higher coupon redemption rates.

Therefore, thorough analysis and research of market trends, consumer behavior, and competitor strategies are essential for selecting the right products for promotional campaigns that maximize coupon redemption.

Target Variable Creation:

- The `products` table contains data of around 92 thousand products.
- For `around 44` thousand products, coupons have been provided for.
- Around `48 thousand` were such products for which the coupons were not provided for.
- We are `uncertain` how the retailer determined the products for which coupons were available.
- To develop a `challenge scenario`, we recognize that the retailer has selected the products that have coupons, randomly.
- Researching the `likely influence of coupons` on the product and its sales data is our goal.
- Now, all those products that were randomly selected for the promotions and campaigns will be our `training set`.
- Those products that we did not provide with coupons beforehand => will turn out into `real test set`.

```
In [12]:
product.Target.fillna(-1).value_counts()
```

```
Out[12]: -1.0    48220
          0.0    37995
          1.0     6138
         Name: Target, dtype: int64
```

```
In [13]:
# checking for null values.
product.isnull().sum()
# The null values contained by Target are the actual test set on which we will do the final testing of our mode.
```

```
Out[13]: PRODUCT_ID                0
         MANUFACTURER              0
         DEPARTMENT                0
         BRAND                     0
         COMMODITY_DESC            0
         SUB_COMMODITY_DESC        0
         CURR_SIZE_OF_PRODUCT      0
         Target                48220
         dtype: int64
```

Feature Engeneering:

```
:
# Feature Engineering..
# unique count of households that has purchased that particular product over the span of 2 years.
d['num_hsld_prch_prd']=transaction_data.groupby('PRODUCT_ID')['household_key'].nunique()

# unique count of stores that sells that particlar product.
d['num_stores_has_prd']=transaction_data.groupby('PRODUCT_ID')['STORE_ID'].nunique()

# count of products sold during the span of 2 years.
d['quantity_sold_total']=transaction_data.groupby('PRODUCT_ID')['QUANTITY'].sum()
```

num_hsld_prch_prd	num_stores_has_prd	quantity_sold_total
3.0	3.0	6.0
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	1.0
1.0	1.0	2.0

Statistical Tests of Predictors

Statistics

n [22]:

```
# statistical approval for categorical columns.
for i in d.select_dtypes('object').columns:
    tbl=pd.crosstab(d[i],d.Target)
    # Ho: has no effect, Target variable is independent of the respected column.
    # Ha: has an effect, Target variable is dependant on the respected column.
    ts,p,_,_=stats.chi2_contingency(tbl)
    print(i,':',p)
```

```
MANUFACTURER : 0.0
DEPARTMENT : 0.0
BRAND : 4.249617475100515e-168
COMMODITY_DESC : 0.0
SUB_COMMODITY_DESC : 0.0
CURR_SIZE_OF_PRODUCT : 5.7140933727928e-281
```

n [23]:

```
# statistical approval for numerical columns.
for i in d.select_dtypes(np.number).columns[1:]:
    zero=d.loc[d.Target==0,i]
    ones=d.loc[d.Target==1,i]
    # Ho: Data is normal
    # Ha: Data not normal
    p1=stats.shapiro(zero)[1]
    p2=stats.shapiro(ones)[1]
    if (p1<0.05)&(p2<0.05):
        # Ho: (has no effect) => Target variable is independent of the respected column.
        # Ha: (has an effect) => Target variable is dependant on the respected column.
        print(i,':',stats.mannwhitneyu(zero,ones)[1])
    else:
        print(i,':',stats.ttest_ind(zero,ones)[1])
```

```
num_hsld_prch_prd : 0.0
num_stores_has_prd : 0.0
quantity_sold_total : 0.0
```

.

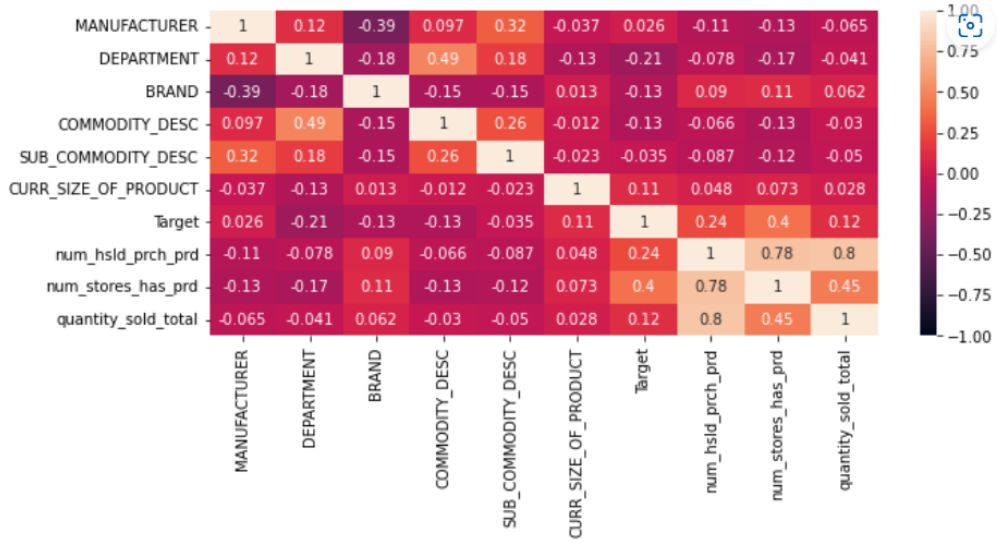
Data Pre-processing:

- ✓ Encoding of Categorical Variables
- ✓ Multicollinearity Treatment
- ✓ Train Test Split

Correlation Plot After Data Pre-processing:

[36]:
```
train_df2=df2[df2.Target.notna()]
test_df2=df2[df2.Target.isnull()]
```

[37]:
```
plt.figure(figsize=(10,4))
sns.heatmap(train_df2.corr(),annot=True,vmin=-1,vmax=+1)
plt.show()
```



Modelling:

Scores after testing on train data:

In [51]:
```
dic_train.sort_values(['accuracy_score','f1_score'],ascending=False)
# here the best performance is given by => RandomForestClassifier(), DecisionTreeClassifier().
```

Out[51]:

	Model	accuracy_score	f1_score	recall_score	precision_score
2	dt	0.992716	0.973320	0.950961	0.996755
3	rfc	0.992671	0.973242	0.954057	0.993216
6	xgbm	0.942185	0.774623	0.711144	0.850546
7	lgbm	0.927139	0.711336	0.642555	0.796607
0	knn	0.921153	0.687647	0.621212	0.769992
5	gbm	0.903171	0.608720	0.539101	0.698986
4	ada	0.893497	0.571795	0.508961	0.652328
1	nb	0.876243	0.420795	0.321766	0.607879

Scores after testing on test data:

In [53]:
```
dic_test.sort_values(['accuracy_score','f1_score'],ascending=False)
# LGBMClassifier(), XGBClassifier() => performs best.
# => tuning of hyperparameter might improve performance.
```

Out[53]:

	Model	accuracy_score	f1_score	recall_score	precision_score
6	xgbm	0.921703	0.700609	0.655537	0.752336
7	lgbm	0.917606	0.683566	0.636808	0.737736
3	rfc	0.906908	0.638373	0.587948	0.698259
5	gbm	0.900193	0.610395	0.559446	0.671554
4	ada	0.890634	0.558161	0.494300	0.640971
0	knn	0.889951	0.561849	0.504886	0.633299
2	dt	0.881074	0.578119	0.583062	0.573259
1	nb	0.877660	0.440396	0.344463	0.610390

EVALUATION

Thoroughly evaluating the model and review the steps executed to construct the model to be certain it properly achieves our business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, final model will be created.

DEPLOYMENT

Project currently doesn’t involve deployment.

REFRENCES

https://towardsdatascience.com/are-you-using-enough-coupons-d18c2d18dd5f

Dunnhumby - The Complete Journey | Kaggle

https://www.researchgate.net/publication/262688139_Factors_Affecting_Coupon_Redemption_Rates

https://itstherealdyl.com/2020/02/09/coupon-redemption-competition/

Predictive Analytics Case Study on Propensity to Redeem a Coupon | by Satyam Kumar | Geek Culture | Medium