Table of Contents

1.0 Introduction ***2***

2.0 Data Overview ***3***

2.1 Glossary of each column **3**

3.0 Data Pre-Processing ***5***

3.1 Headers **5**

3.2 Defining column types **6**

3.3 Null Values **6**

4.0 Data Analysis ***7***

4.1 Income with Age **8**

4.2 Distribution of data set **9**

5.0 Correlations of Dataset ***9***

6.0 Modelling ***9***

1.0 Introduction

Advertising and promotion are pivotal to the marketing of the American food supply. The U.S. food marketing system is the second largest advertiser in the American economy, and a leading supporter of network, spot, and cable television, newspapers, magazines, billboards, and commercial radio. Groceries account for about 70 percent of all manufacturers' coupons. Food manufacturers also spend massive amounts promoting the product to the retailer through discounts and allowances, incentives, and actual slotting allowances in order to secure scarce space on the Nation's grocery shelves.

Why so much advertising? There are several reasons for it. First, the food market is huge, capturing about 12.5 percent of consumer income, and there is vigorous competition among food firms to compete for this market. Second, food is a repeat-purchase item, lending itself to swift changes in consumer opinions. Third, food is one of the most highly branded items in the American economy, thus lending itself to major advertising. [1]

This analysis tries to understand human behavior and their purchasing options when they go to a new store. We will call them cold starter shoppers. Shoppers we do not have any information except after they make their first purchase.

[1] https://www.ers.usda.gov/webdocs/publications/42215/5838_aib750i_1_.pdf?v=41055

2.0 Data Overview

2.1 Glossary of each column

The dataset contains 39 columns with 2993708 rows. Our target columns will the client information and splitting the data with clients that only have one purchase and the first purchase made.

Overview of the different columns and their descriptions:

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+

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

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
CAMPAIGN	Uniquely identifies each campaign. Ranges 1 - 30
DESCRIPTION	Type of campaign (TypeA, TypeB or TypeC)

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

Variable	Description
PRODUCT_ID	Number that uniquely identifies each product
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
MANUFACTURER	Code that links products with same manufacturer together
BRAND	Indicates Private or National label brand
CURR_SIZE_OF_PRODUCT	Indicates package size (not available for all products)

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

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

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

[2]

3.0 Data Pre-Processing

3.1 Headers

Changing all headers to lower case for a uniform and easy usage for future reference.

```
Dataset headers: Index(['PRODUCT_ID', 'STORE_ID', 'WEEK_NO', 'display', 'mailer',
       'household_key', 'BASKET_ID', 'DAY', 'QUANTITY', 'SALES_VALUE',
       'RETAIL_DISC', 'TRANS_TIME', 'COUPON_DISC', 'COUPON_MATCH_DISC',
       'DESCRIPTION', 'CAMPAIGN', 'START_DAY', 'END_DAY', 'MANUFACTURER',
       'DEPARTMENT', 'BRAND', 'COMMODITY_DESC', 'SUB_COMMODITY_DESC',
       'CURR_SIZE_OF_PRODUCT', 'COUPON_UPC', 'AGE_DESC', 'MARITAL_STATUS_CODE',
       'INCOME_DESC', 'HOMEOWNER_DESC', 'HH_COMP_DESC', 'HOUSEHOLD_SIZE_DESC',
       'KID_CATEGORY_DESC', 'affinity_raw', 'affinity_rank_pct',
       'affinity_rank_centered', 'affinity_log', 'affinity_log_scaled',
       'affinity_rank_buckets', 'affinity_score'],
      dtype='object')
Lowercase headers: Index(['product_id', 'store_id', 'week_no', 'display', 'mailer',
       'household_key', 'basket_id', 'day', 'quantity', 'sales_value',
       'retail_disc', 'trans_time', 'coupon_disc', 'coupon_match_disc',
       'description', 'campaign', 'start_day', 'end_day', 'manufacturer',
       'department', 'brand', 'commodity_desc', 'sub_commodity_desc',
       'curr_size_of_product', 'coupon_upc', 'age_desc', 'marital_status_code',
       'income_desc', 'homeowner_desc', 'hh_comp_desc', 'household_size_desc',
       'kid_category_desc', 'affinity_raw', 'affinity_rank_pct',
       'affinity_rank_centered', 'affinity_log', 'affinity_log_scaled',
       'affinity_rank_buckets', 'affinity_score'],
      dtype='object')
```

[2] dunnhumby - The Complete Journey User Guide.pdf

3.2 Defining column types

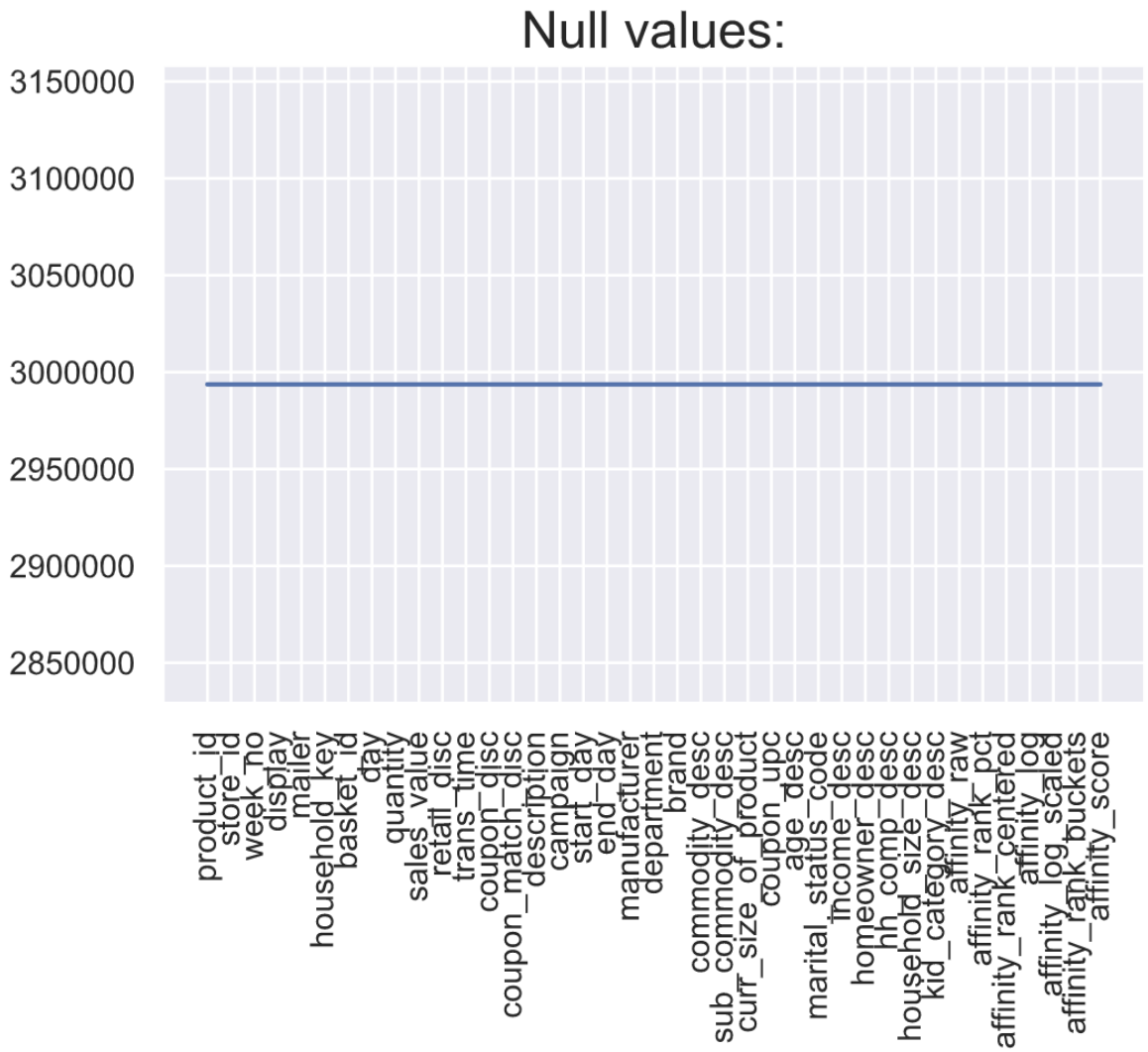
The dataset has almost 3 million rows with 39 rows. The original dataset almost uses 1GB of memory to process the entire dataset that delays simple process like plotting scatter plots of the entire dataset to find correlations between the different entries.

```
dtypes: float64(9), int64(15), object(15)
memory usage: 913.6+ MB
Memory usage before changing categories: None
dtypes: category(12), float64(9), int64(15), object(3)
memory usage: 673.8+ MB
Memory usage after changing categories: None
```

By changing some object columns to categories, I was able to save 250GBs that allows for a faster render of the dataset.

3.3 Null Values

The dataset does not have any null values present

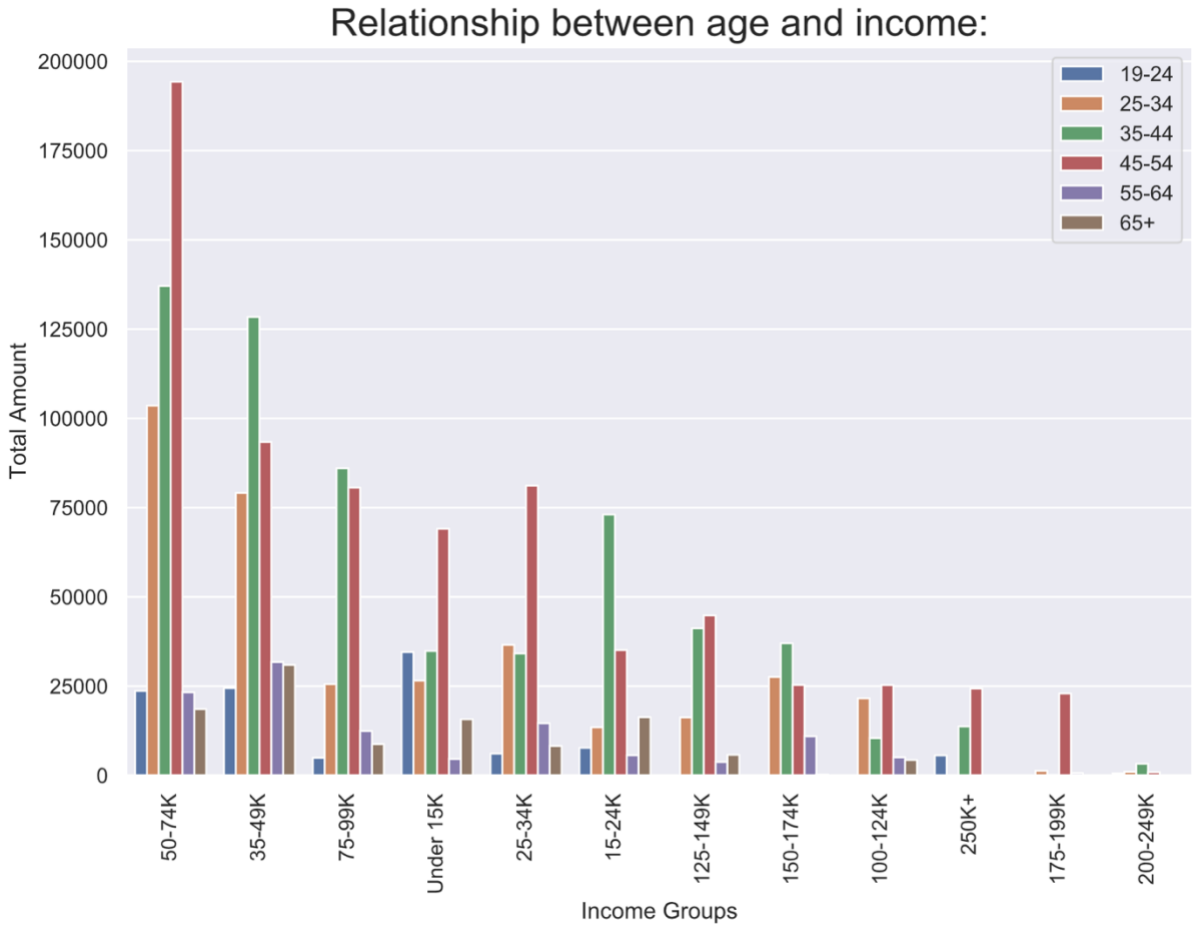


4.0 Data Analysis

In order to better understand our dataset, we want to understand the data and what relations some columns have with others.

4.1 Income with Age

One of the big factors that we want to test is that cold starter users will use coupons based on their income and age. The relation is as the income is higher, the customer will spend less time shopping for coupons, since that customer has more disposable income than the lower income customer. Also, we want to test as customers get older, they are more susceptible to use coupons. Age relationship is based that we are better aware of our finances as we get older and know what can be afforded.



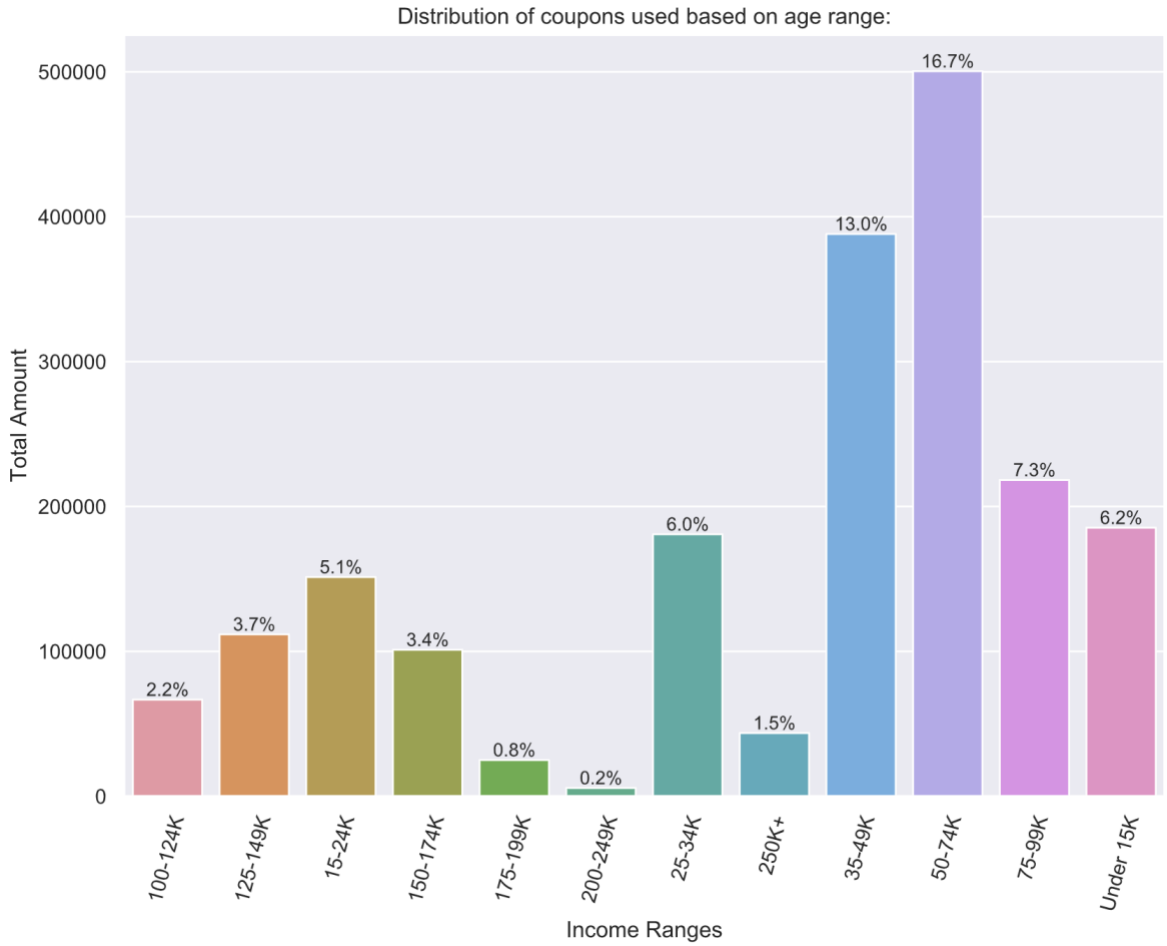
As displayed, lower income customers tend to use coupons more often than customer with much higher wages (all incomes over 100k are displayed in the later half part of the graph). Also, customers between the age group of 35-54 tend to be the ones that use the most coupons. Some irregularities are:

- Why aren't the lowest income customer using the most? Why customer that make 75-99K use more coupons that customers that make less than 15K?
- Overall, people that are between the ages 35-44 tend to use coupons more often than any other age group. Why is this?

- Why is there a big spike of usage for customers between the age of 45-54 when making 50-74K?

4.2 Distribution of data set

Most of the dataset is distributed between the lower income ranges; especially between 35-49K and 50-74K, compromising 29.7% of the total dataset. We will focus more on these group since we have more data to test on.



5.0 Correlations of Dataset

6.0 Modelling