Table of Contents

[1.0 Introduction 2](#_Toc18407795)

[2.0 Data Overview 3](#_Toc18407796)

[2.1 Glossary of each column 3](#_Toc18407797)

[3.0 Data Pre-Processing 5](#_Toc18407798)

[3.1 Headers 5](#_Toc18407799)

[3.2 Defining column types 5](#_Toc18407800)

[3.3 Null Values 6](#_Toc18407801)

[4.0 Data Analysis 7](#_Toc18407802)

[4.1 Product Popularity 7](#_Toc18407803)

[4.2 Store popularity 8](#_Toc18407804)

[4.3 Income relationship with age 8](#_Toc18407805)

[4.2 Distribution of data set 9](#_Toc18407806)

[5.0 Correlations of Dataset 10](#_Toc18407807)

[6.0 Modelling 10](#_Toc18407808)

2

# 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]](#footnote-1)

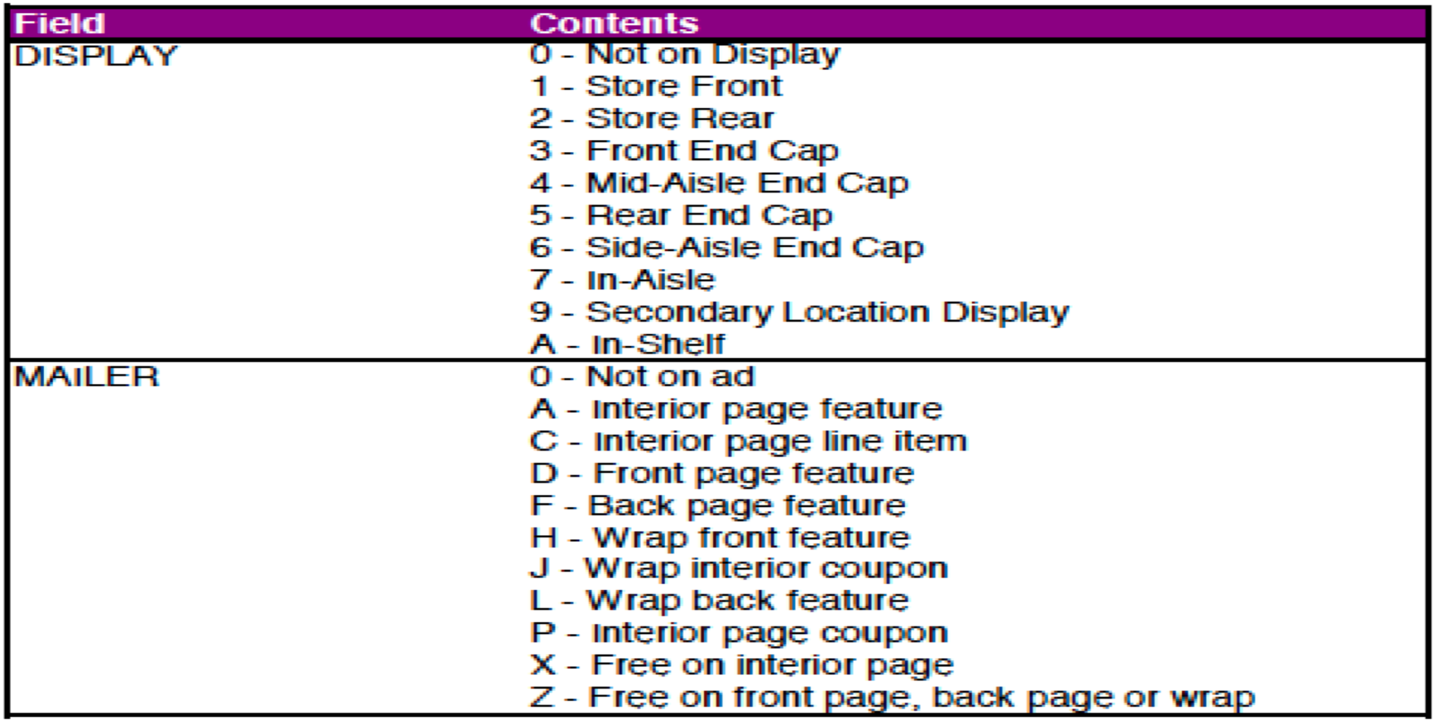
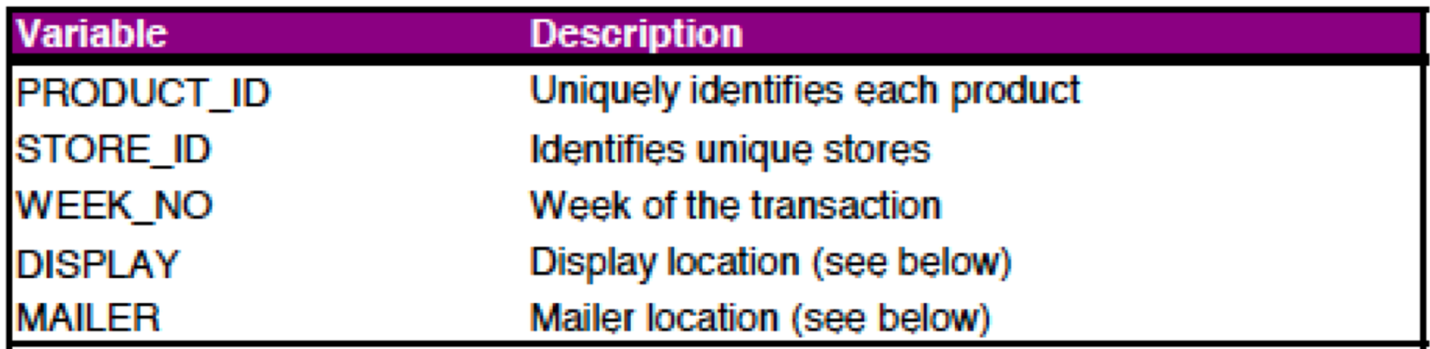
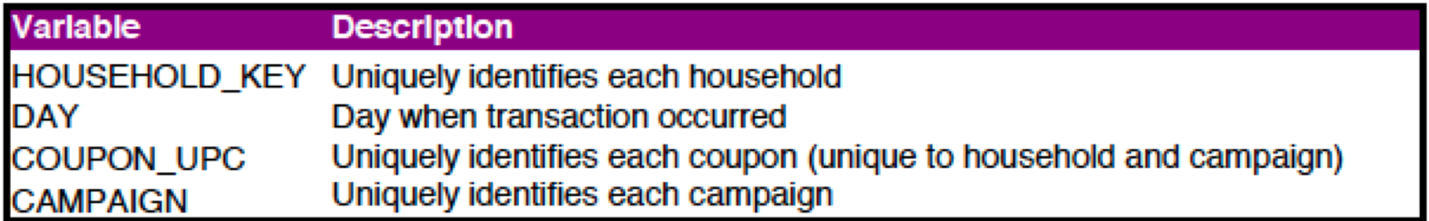
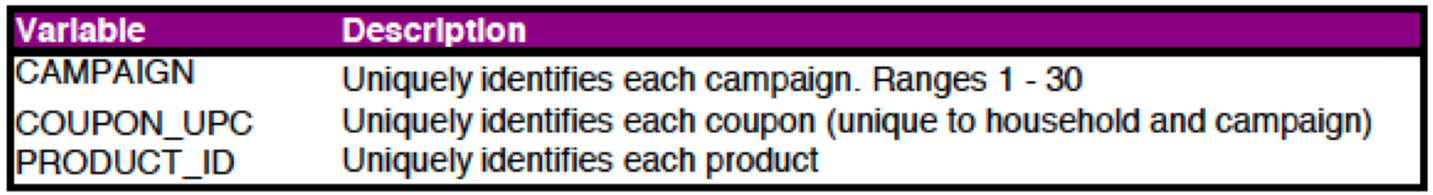
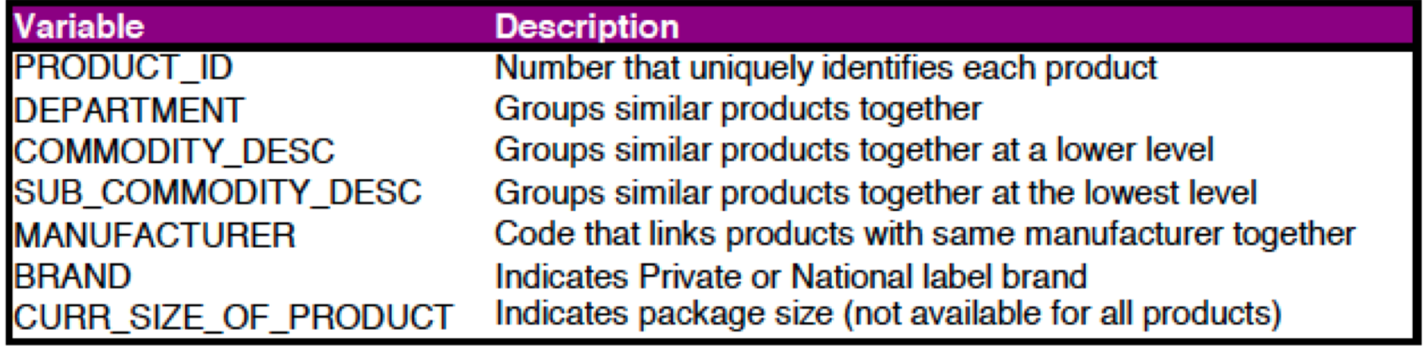
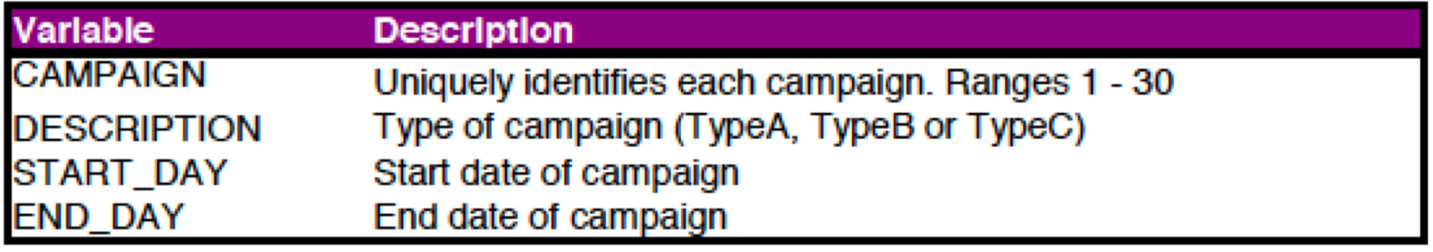
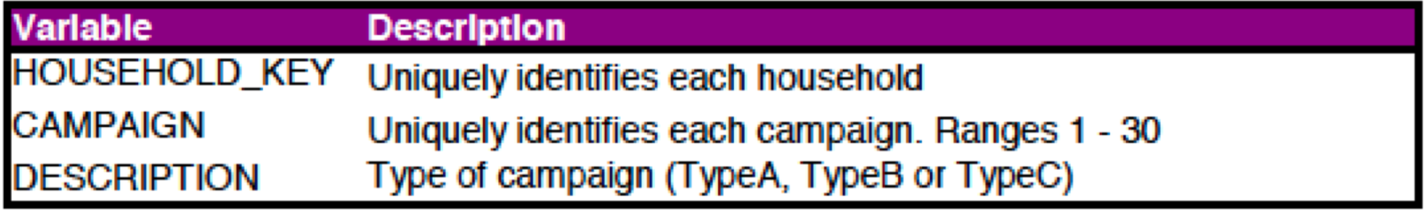
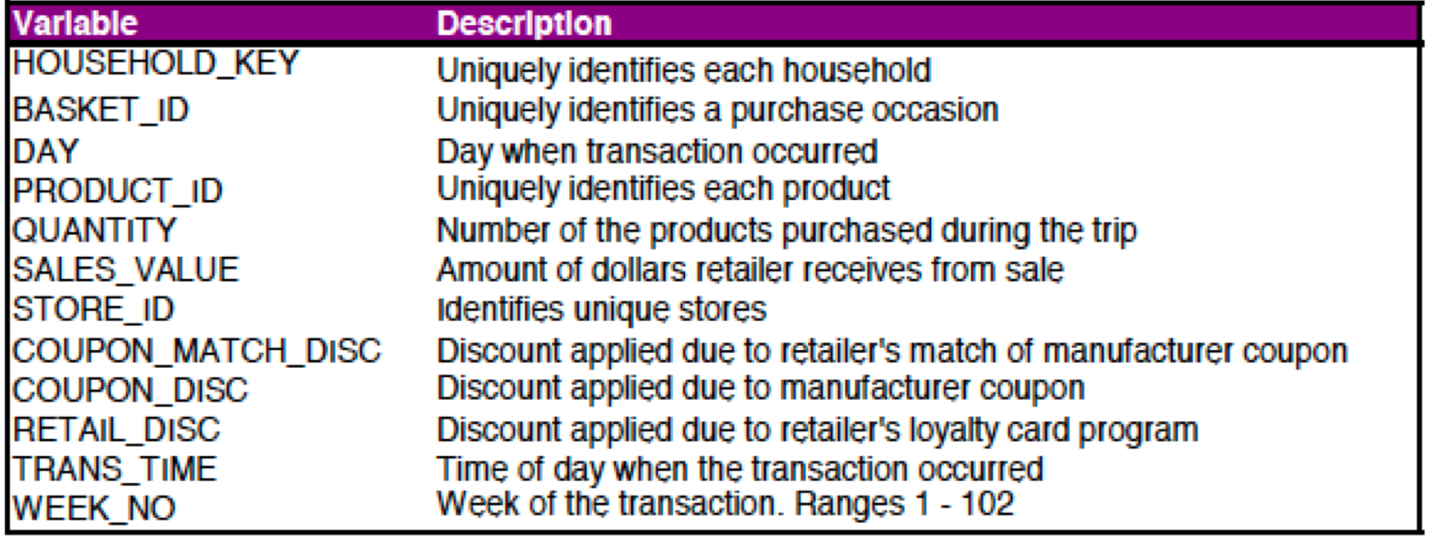
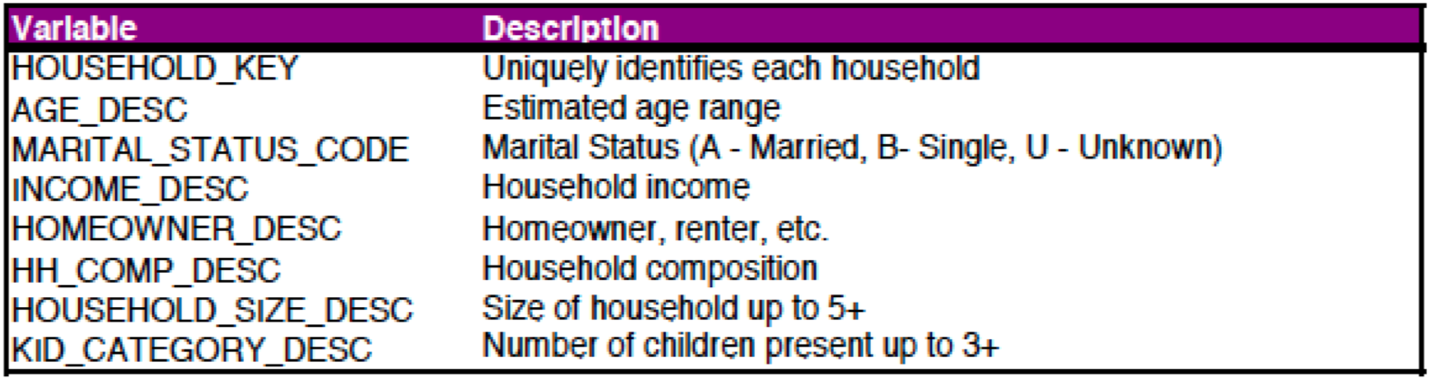
This analysis tries to understand human behavior and their purchasing options when they go to a new store. We will call them cold starter shoppers. Shoppers we do not have any information except after they make their first purchase.

# 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:

[[2]](#footnote-2)

# 3.0 Data Pre-Processing

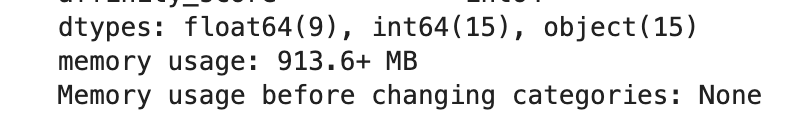
## 3.1 Headers

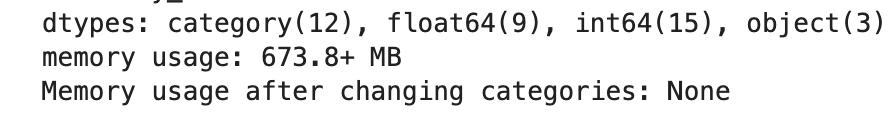
Changing all headers to lower case for a uniform and easy usage for future reference.



## 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.

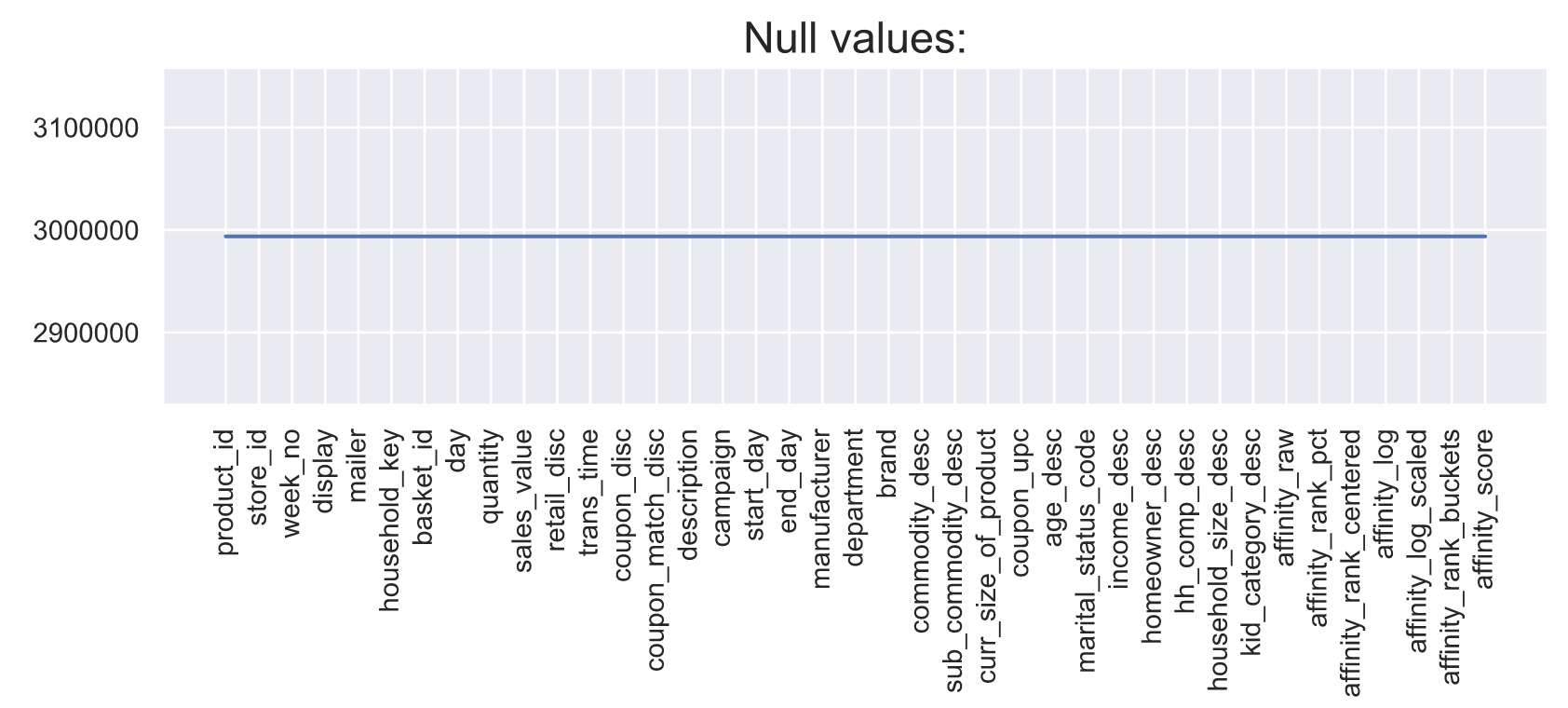




By changing some object columns to categories, I was able to save 250GBs that allows for a faster render of the dataset.

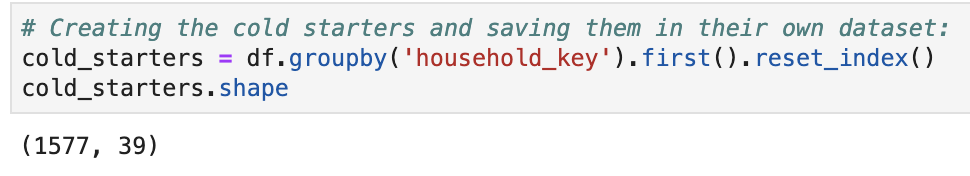
## 3.3 Null Values

The dataset does not have any null values present



## 3.4 Creating Cold Starters Dataset

Focusing on the first purchase of a customer, by extracting the 1st purchases recorded in the dataset. As a result, we gather 1577 unique customer that can be compared to the different relationships that will be extracted next.

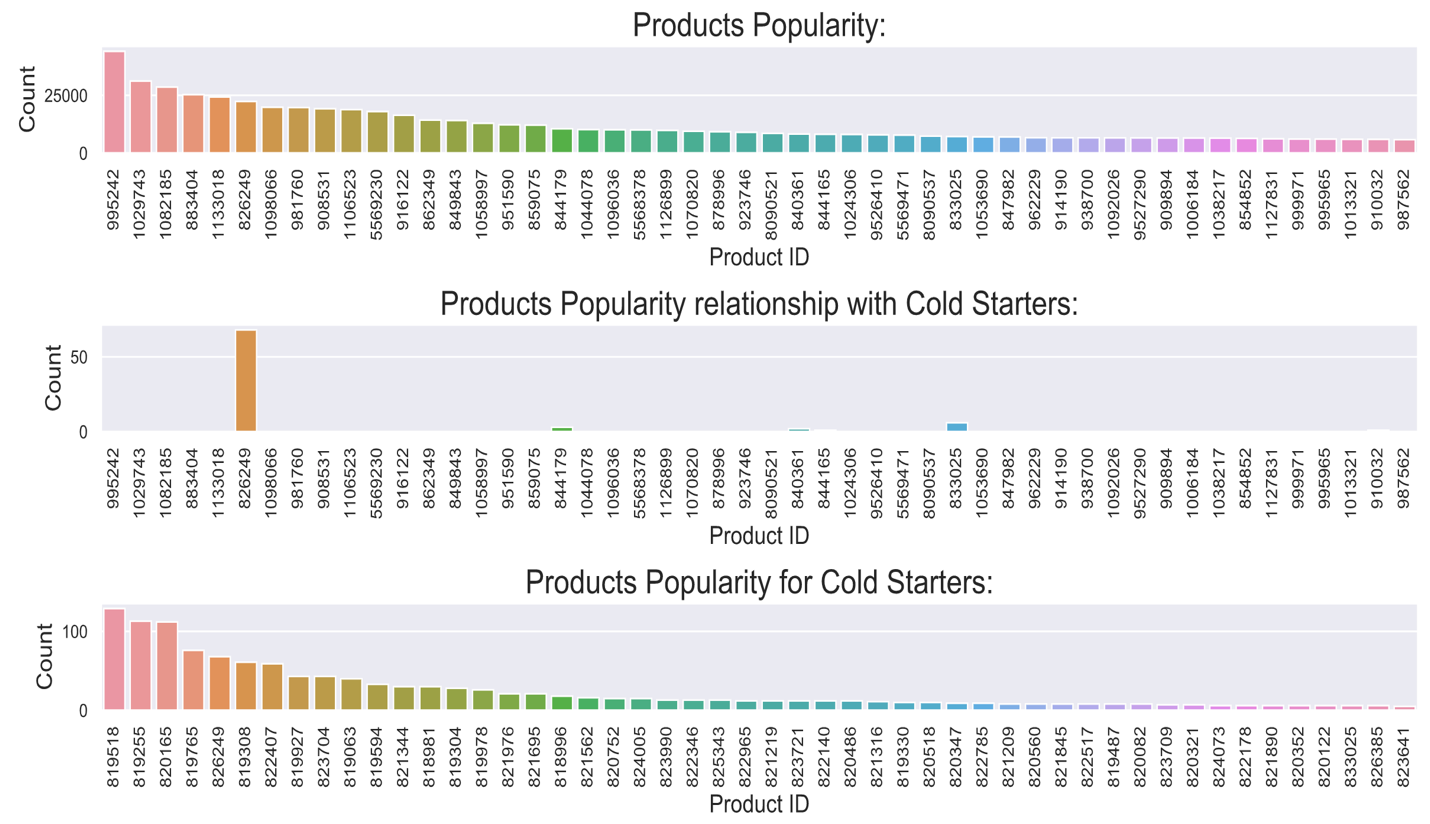


# 4.0 Data Analysis

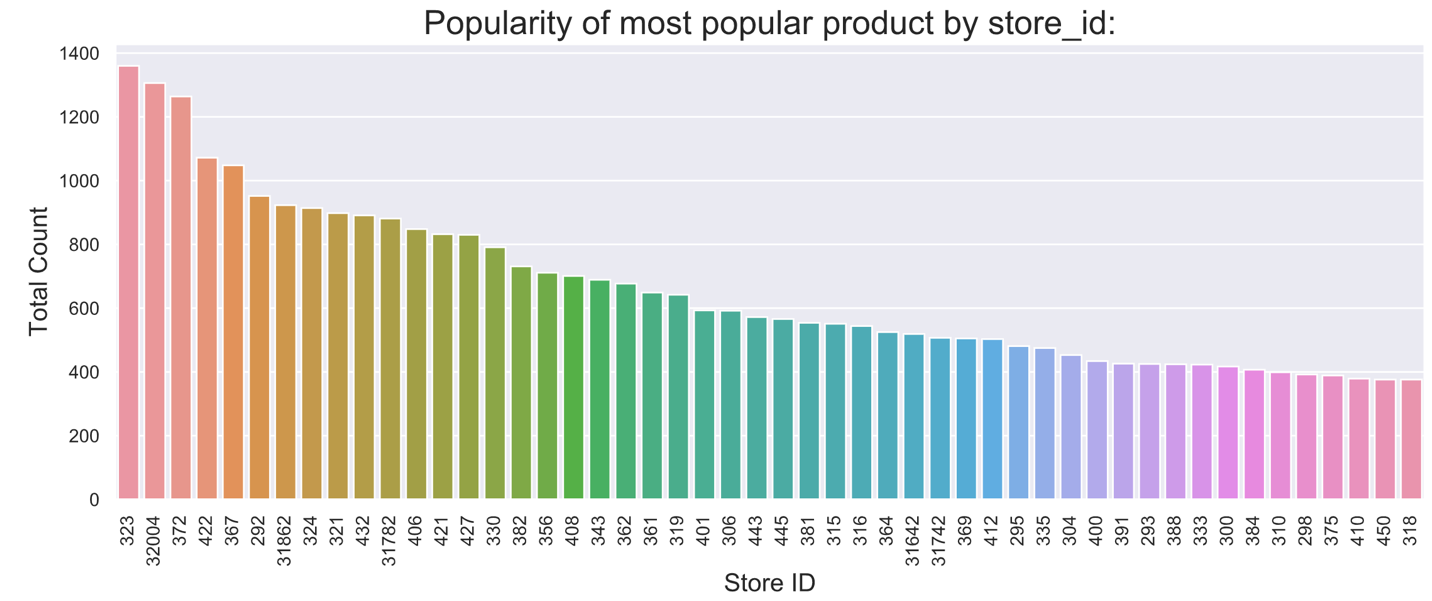
In order to better understand our dataset, we want to understand the data and what relations some columns have with each other’s. Especially if cold starter users tend to shop in specific shops, for specific products, and if some decisions are swayed by the income, age and location of the product within the store.

## 4.1 Product Popularity

Product 995242 tends to sell at least 25% overall better than the 2nd most popular product. However, this product is usually not purchased by cold starters. The most popular is product 819518 with over 130+ counts. Focusing on this product will help us better understand if there is a correlation with necessity, locations of products and whether or not coupons were the cause to help sway the decision to buy this product in the first place.



Even though the most popular product outsells the 2nd product by over 15K; it is not enough to boost the store that makes the most sells to be the one that makes the most transactions in the dataset.



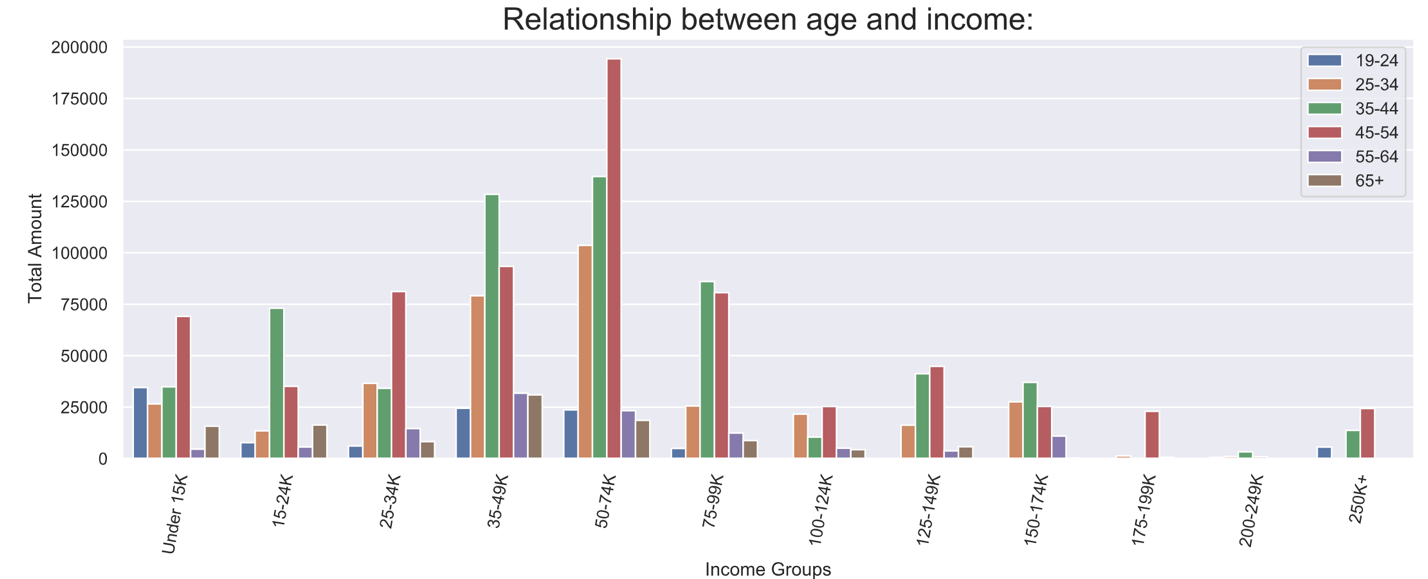
## 4.2 Store popularity

Store 292 made close to 78K transactions in the dataset, followed closely by store 406. The rest of the stores fall behind by 12K and slowly descend in a linear decline. Later one we’ll check whether or not there’s a relation between the stores that make the mos



## 4.3 Income relationship with age

One of the big factors that we want to test is whether or not 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.

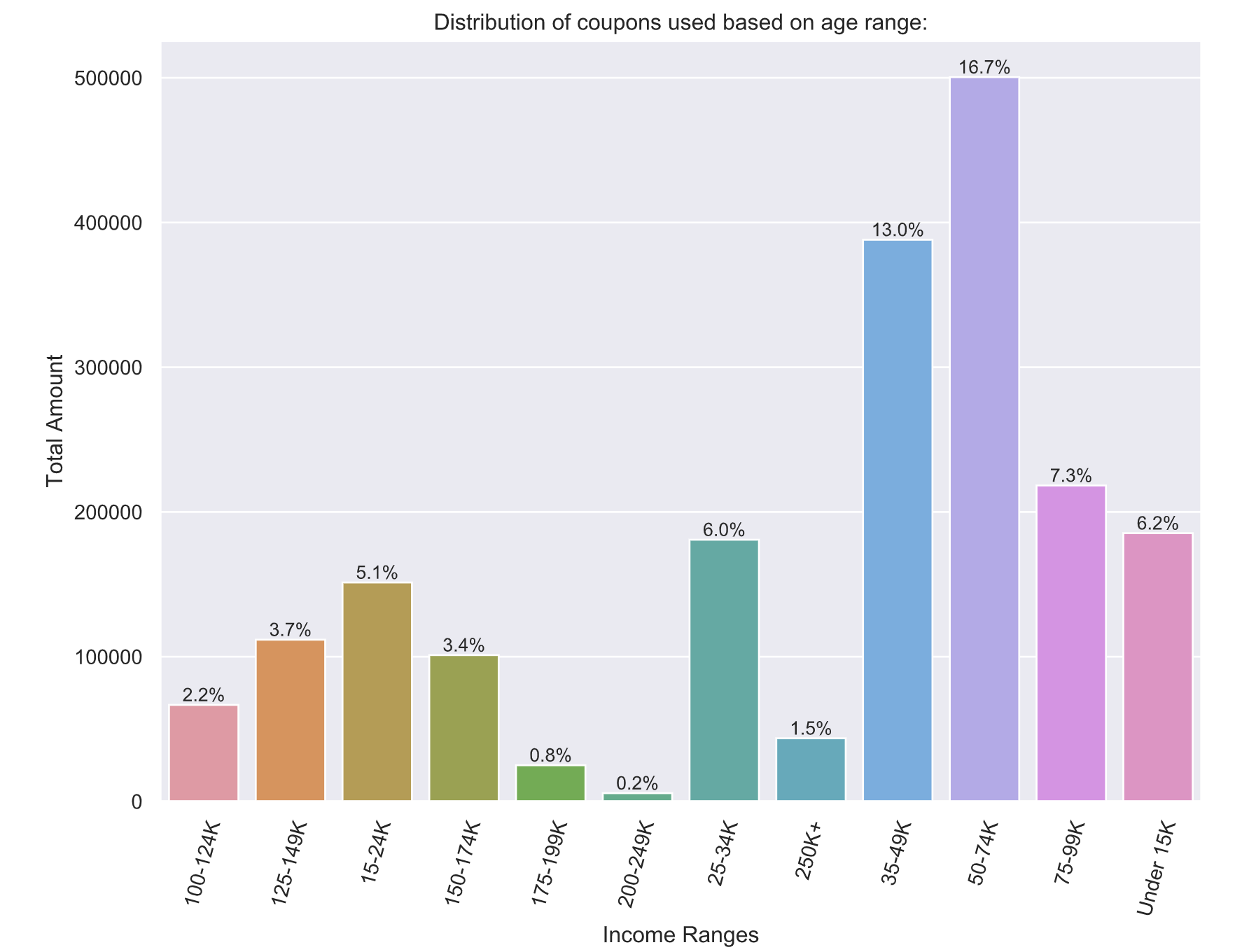


As displayed, lower income customers tend to use coupons more often than customer with much higher wages (all incomes over 100k are displayed in the later half part of the graph). Also, customers between the age group of 35-54 tend to be the ones that use the most coupons. Some irregularities are:

* 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?

## 4.4 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

1. <https://www.ers.usda.gov/webdocs/publications/42215/5838_aib750i_1_.pdf?v=41055> [↑](#footnote-ref-1)
2. dunnhumby - The Complete Journey User Guide.pdf [↑](#footnote-ref-2)