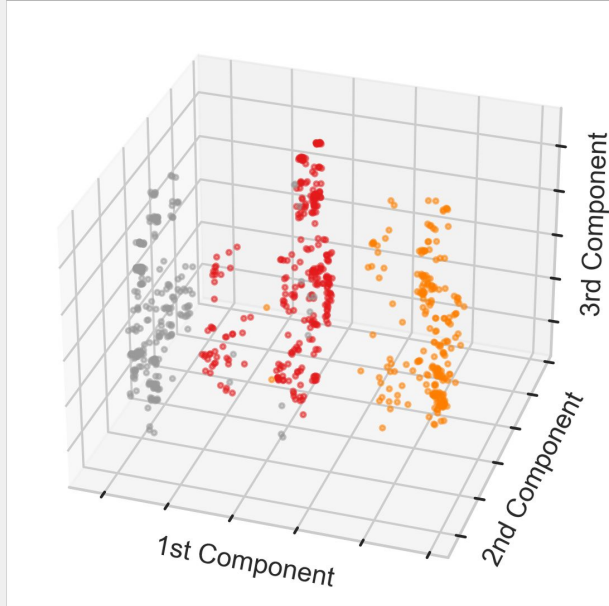Springboard - Data Science Track
Capstone Project 2:
Customer Segmentation & Churn Analysis
By: Nicholas Dean
September 2021

Defining the problem & stakeholders
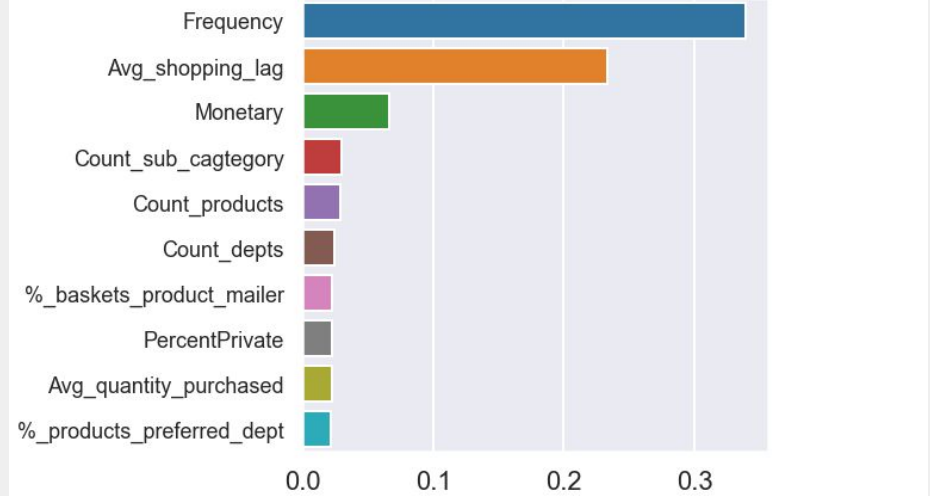
- According to HubSpot: "Customer churn is the percentage of customers that stopped using your company's product or service during a certain time frame"

- Additionally, Client Success estimates that about 65% of existing customers can be upsold to, vs a 13% conversion rate for new customers (a conversion rate that is already extremely generous).

- The intended stakeholders for a project like this are decision makers on product, customer, and marketing strategy teams for the grocery store chain being examined
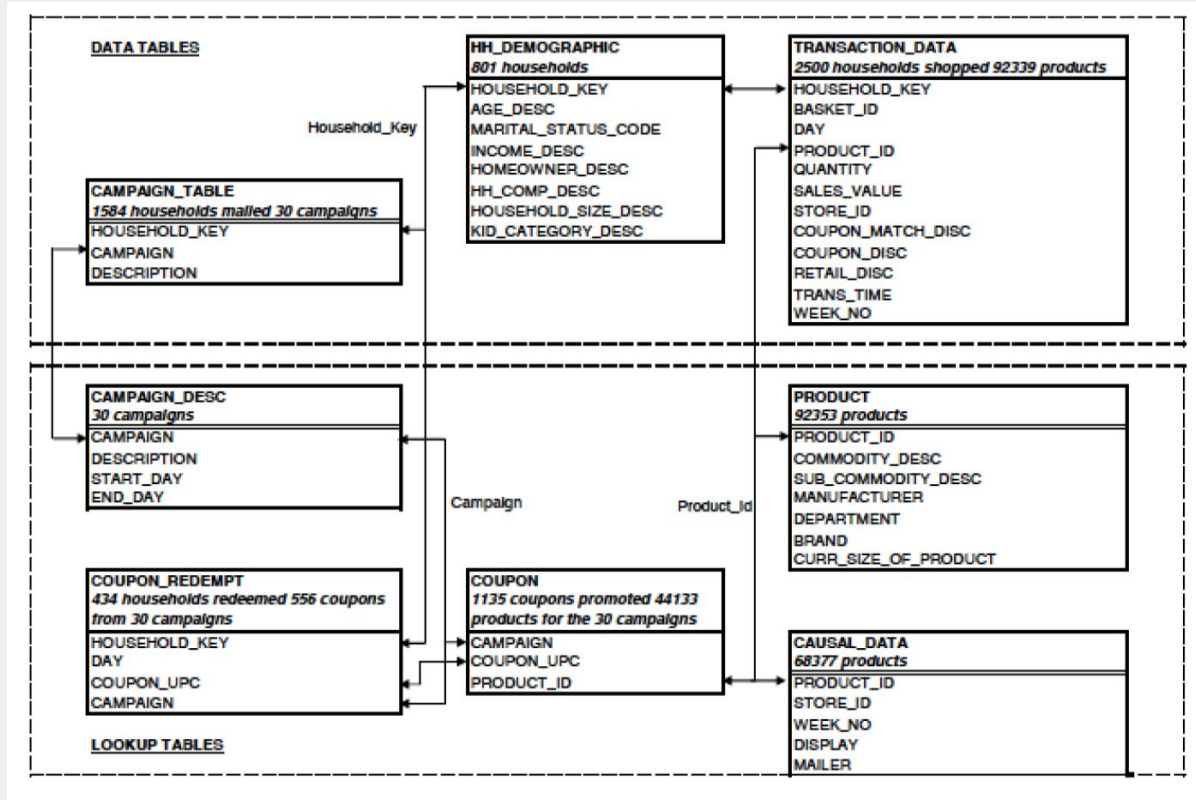
Bottom Line

Hierarchical clusters

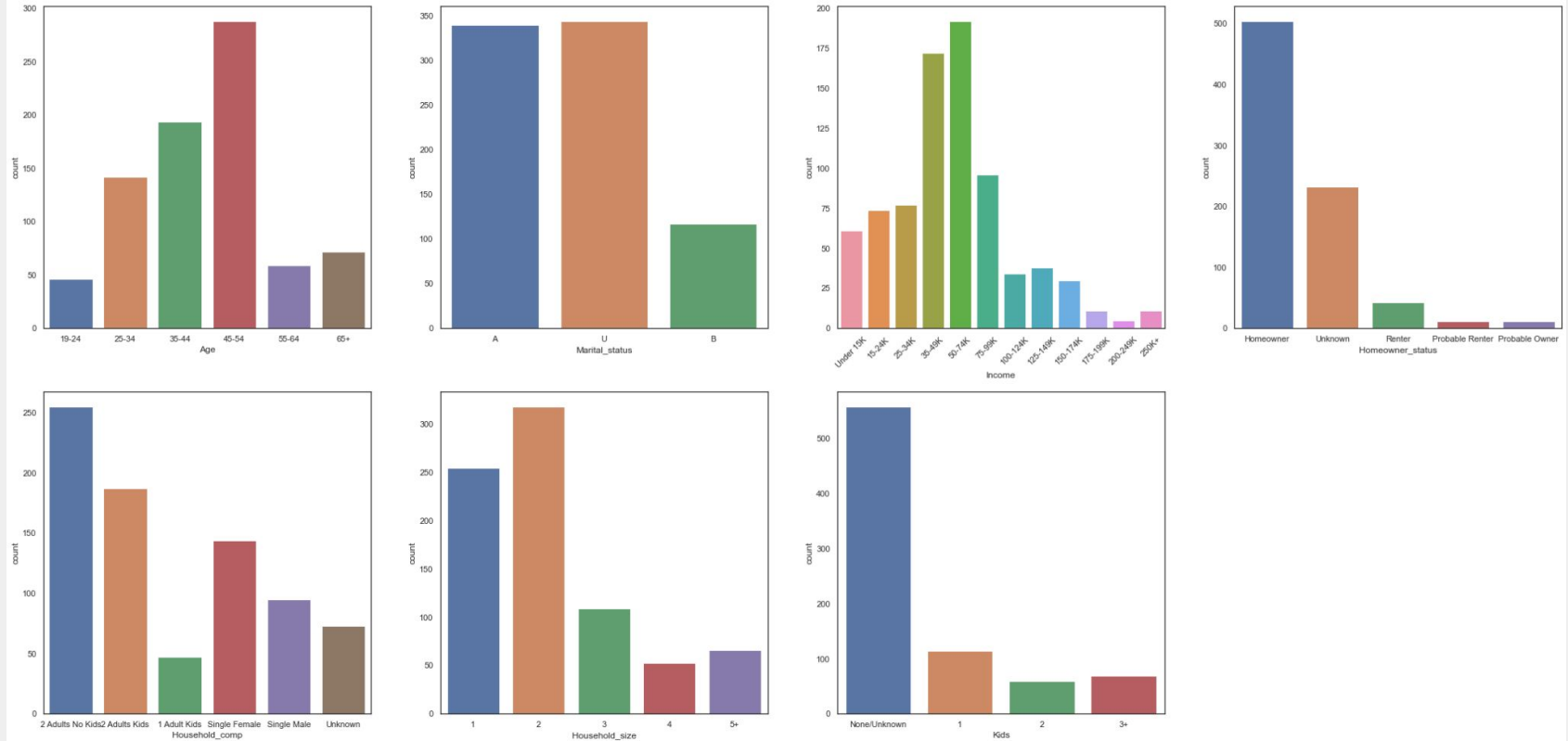Relative importance of features to determine customer churn

Frequency
Avg_shopping_lag
Monetary
Count_sub_cagtegory
Count_products
Count_depts
%_baskets_product_mailer
PercentPrivate
Avg_quantity_purchased
%_products_preferred_dept

0.0   0.1   0.2   0.3

Data Acquisition

DATA TABLES

HH_DEMOGRAPHIC
801 households
HOUSEHOLD_KEY
AGE_DESC
MARITAL_STATUS_CODE
INCOME_DESC
HOMEOWNER_DESC
HH_COMP_DESC
HOUSEHOLD_SIZE_DESC
KID_CATEGORY_DESC

TRANSACTION_DATA
2500 households shopped 92339 products
HOUSEHOLD_KEY
BASKET_ID
DAY
PRODUCT_ID
QUANTITY
SALES_VALUE
STORE_ID
COUPON_MATCH_DISC
COUPON_DISC
RETAIL_DISC
TRANS_TIME
WEEK_NO

Household_Key

CAMPAIGN_TABLE
1584 households mailed 30 campaigns
HOUSEHOLD_KEY
CAMPAIGN
DESCRIPTION

CAMPAIGN_DESC
30 campaigns
CAMPAIGN
DESCRIPTION
START_DAY
END_DAY

PRODUCT
92353 products
PRODUCT_ID
COMMODITY_DESC
SUB_COMMODITY_DESC
MANUFACTURER
DEPARTMENT
BRAND
CURR_SIZE_OF_PRODUCT

Campaign

Product_Id

COUPON_REDEMPT
434 households redeemed 556 coupons from 30 campaigns
HOUSEHOLD_KEY
DAY
COUPON_UPC
CAMPAIGN

COUPON
1135 coupons promoted 44133 products for the 30 campaigns
CAMPAIGN
COUPON_UPC
PRODUCT_ID

CAUSAL_DATA
68377 products
PRODUCT_ID
STORE_ID
WEEK_NO
DISPLAY
MAILER
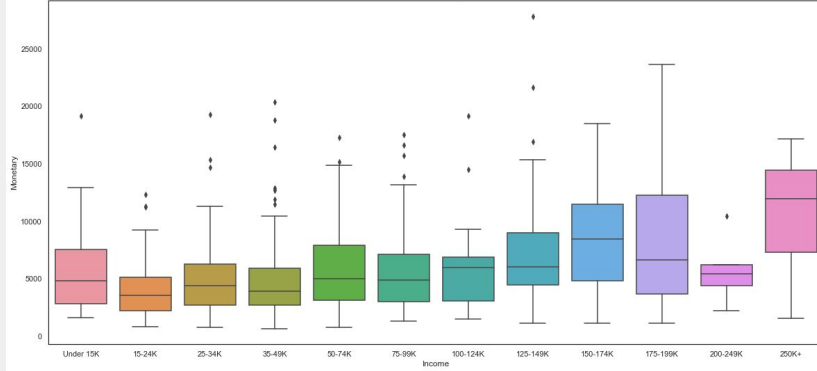
LOOKUP TABLES

Outliers and Missing Data

- Significant number of outliers with outrageous purchasing numbers

- It appeared that while 85 households were outliers in all noted qualities, over 700 were an outlier in one of those respects

- Potential for 'corporate' customers driving up the numbers, but need actual stakeholders to consult with on this kind of problem

- 1699 households were missing any demographic data

- Ultimately I decided to proceed with only households for which I had demographic information (801 total) for customer segmentation, while I'd work with all of the household data and exclude demographic data for churn analysis (2500)

Demographic Data

Household Income & Child-rearing Status

Coupon Usage & Campaign Response

Coupons Redeemed vs Avg Basket Sales Value
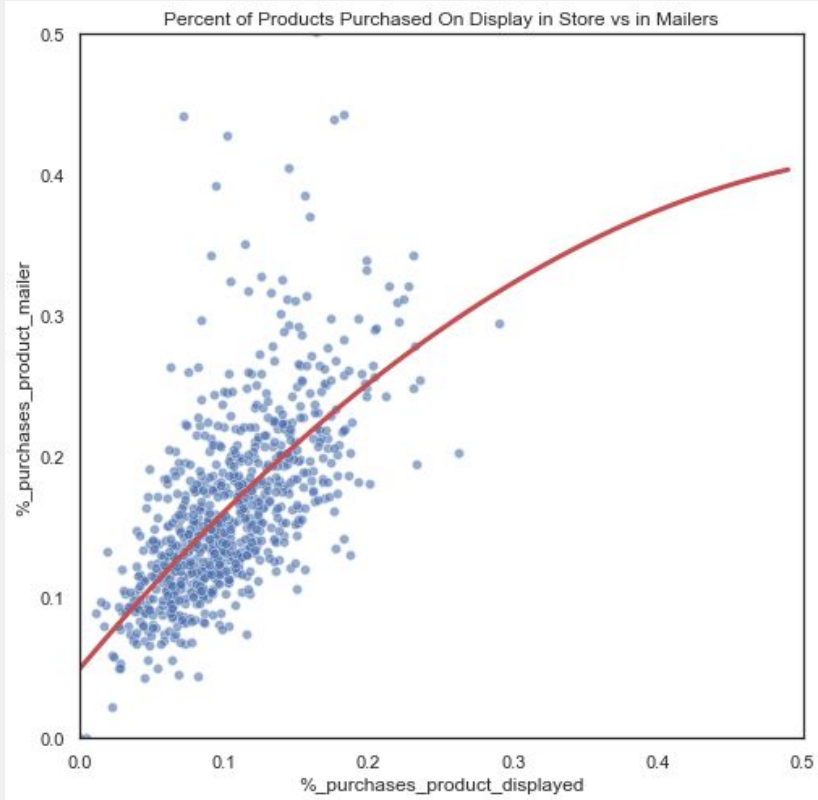

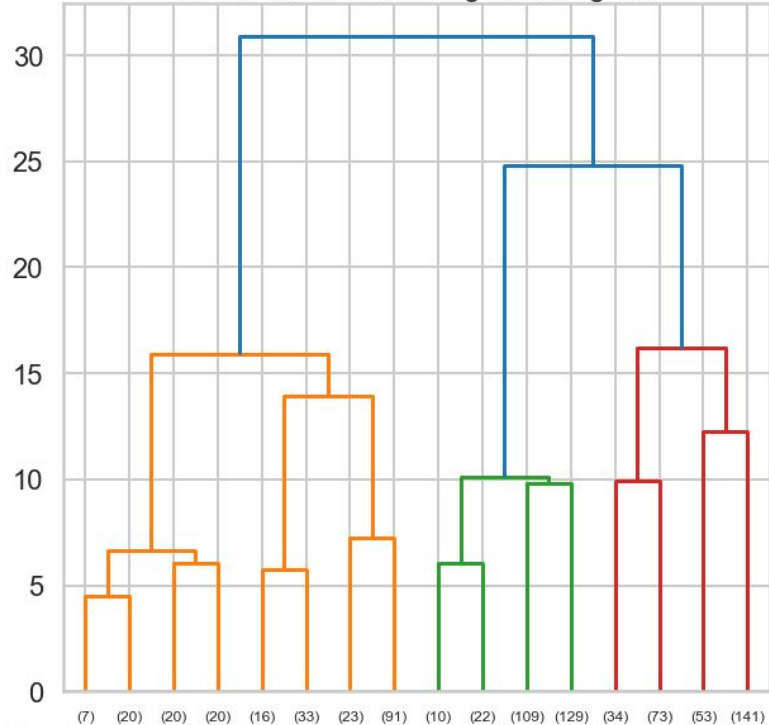There is a clear ceiling for the effectiveness of Type A Campaigns
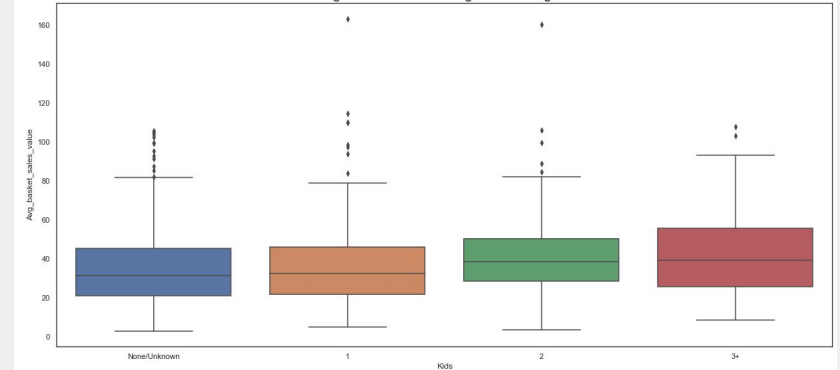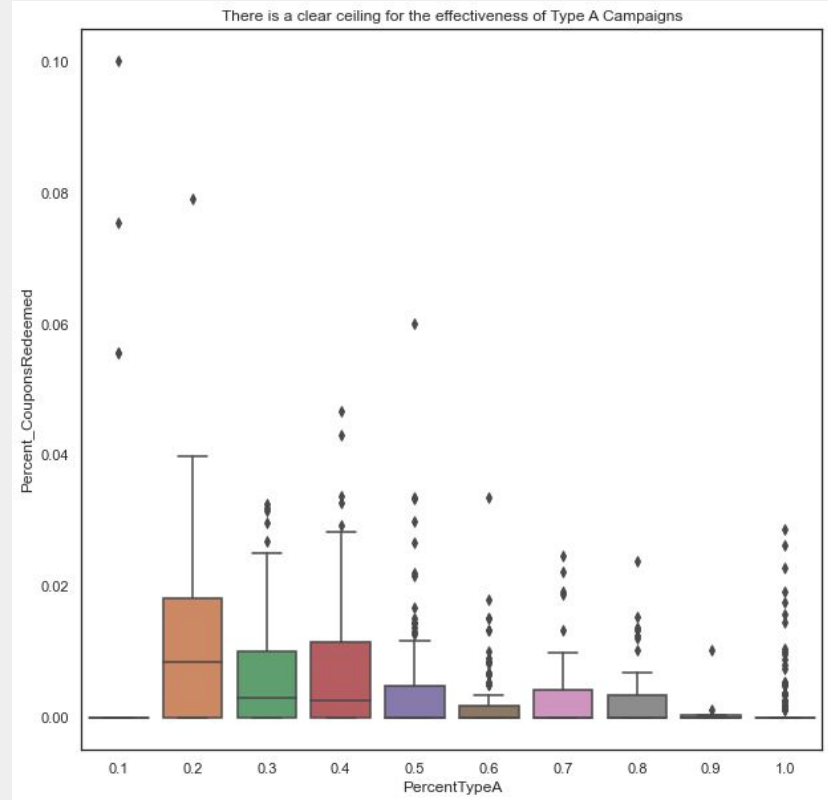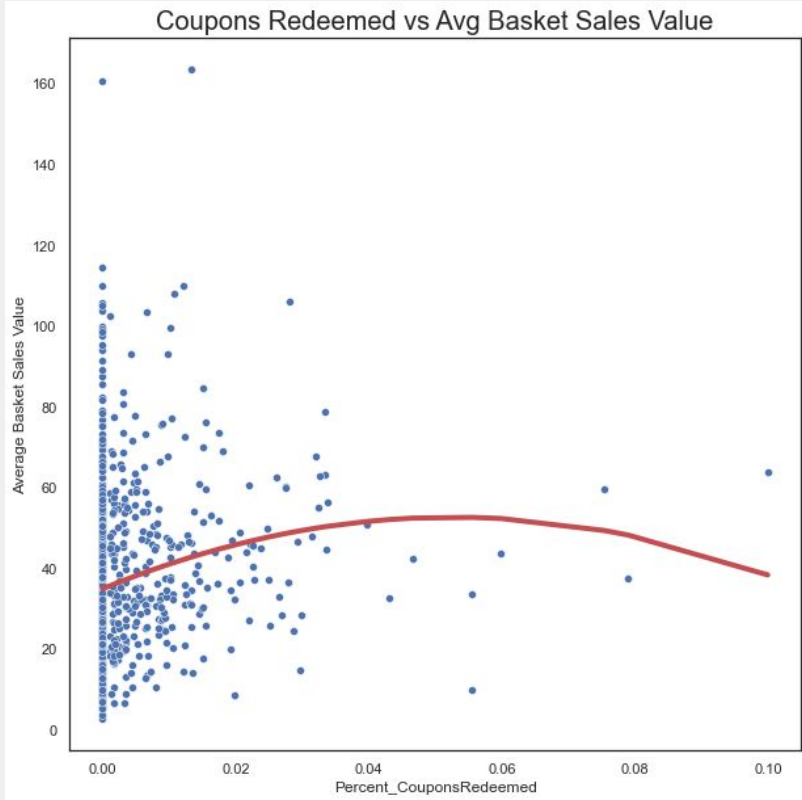
Efficacy of Marketing & Loyalty Discounts
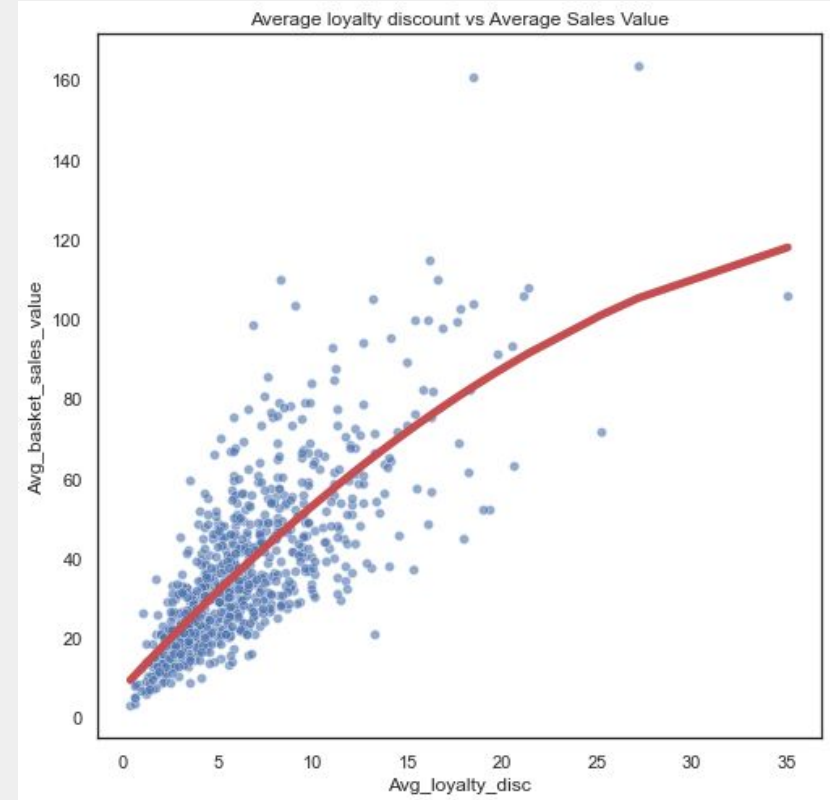
Baseline Clustering
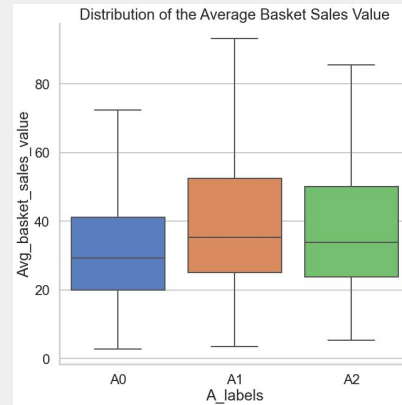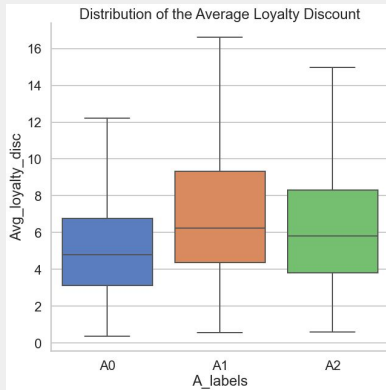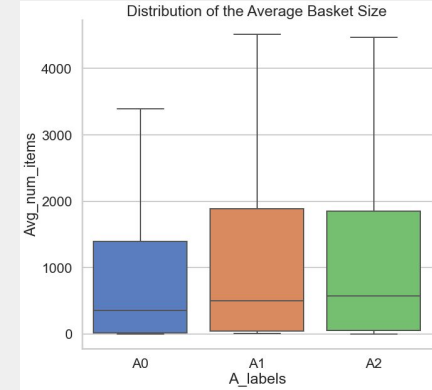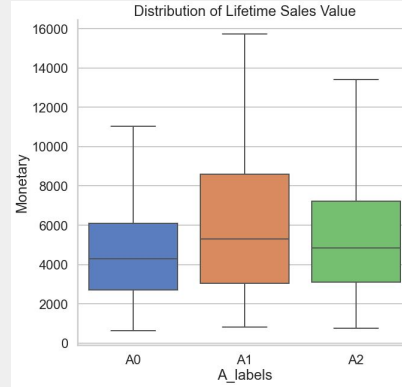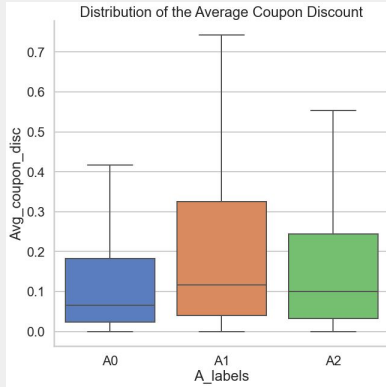

Hierarchical Clustering Dendrogram

Number of points in node (or index of point if no parenthesis).
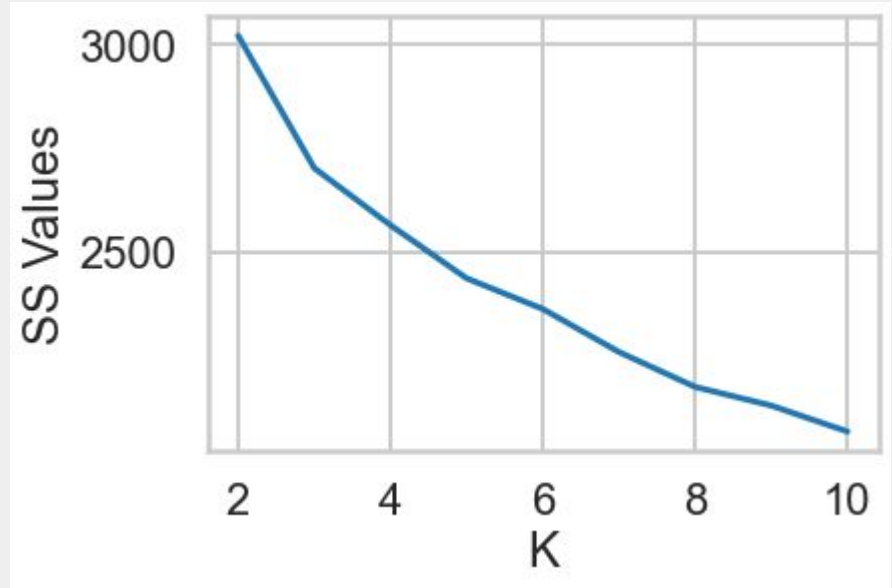
Key Differences With Segments
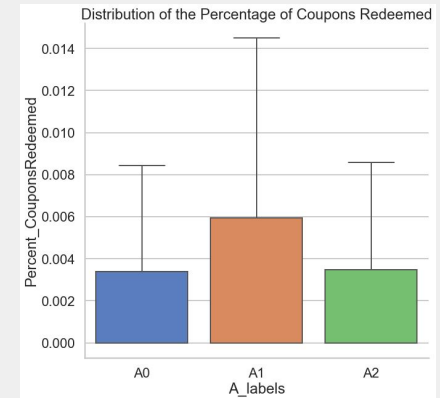
Customer Segment Summaries

1. **A0**: It's overwhelmingly made up single person households which are poorer on average than those belonging to another cluster.

2. **A1**: Is primarily made up of people between the ages of 25 and 44, and is also the only cluster to be made up of households with children, and not a single member has fewer than 3 members.

3. **A2**: is exclusively made up of 2-person households with no children. Generally speaking this cluster is older on average than the other two, with the largest number of households over the age of 65.

Hierarchical clusters

Defining Churn

Plot showing the number of days since a customer shopped relative to the day they first shopped

The red line represents churn definition #1, churn definition #2 is the shaded gree area, and the purple line is churn definition #3
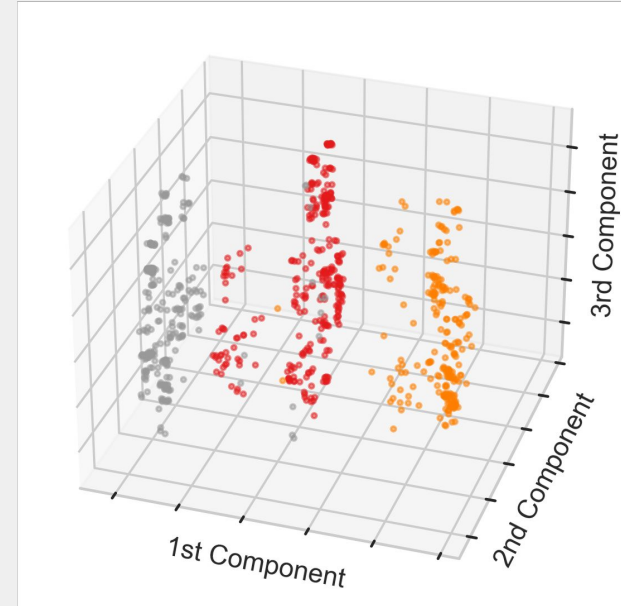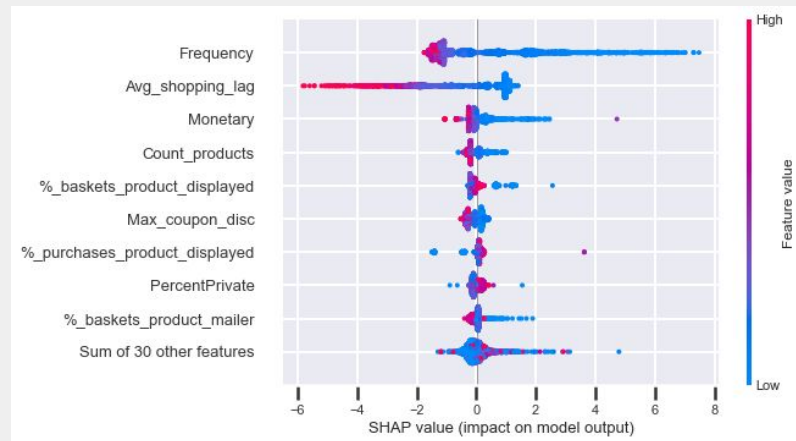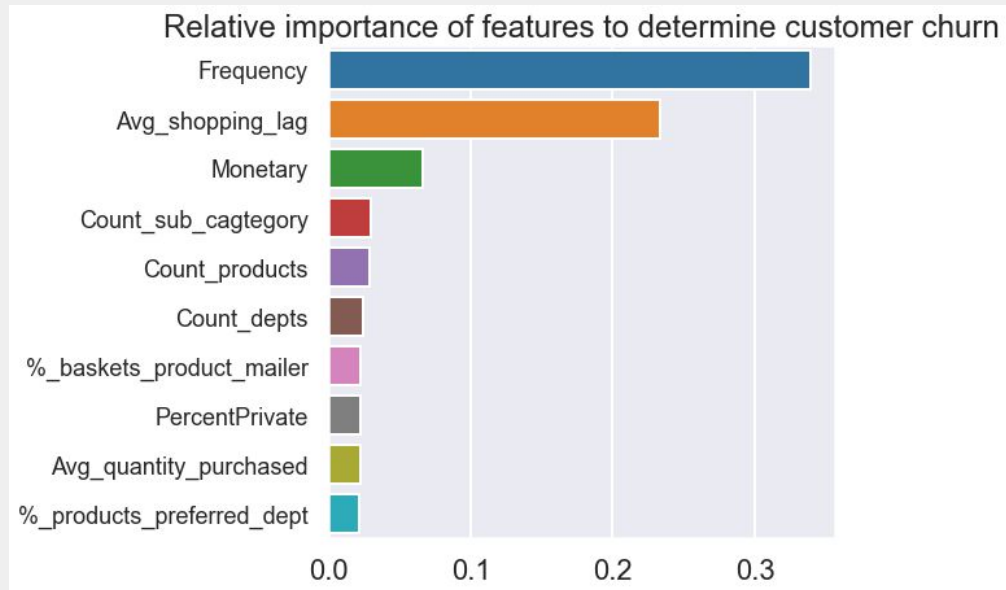
Baseline Modeling

- Primarily focused on the recall metric for the churn class

- Churned households make up 6.84% of the dataset.

- Focus of this project is to understand what features impact churn determinations the most

- Precision is a good secondary metric to use

	precision	recall	f1-score	support
0	0.93	1.00	0.96	466
1	0.00	0.00	0.00	34
accuracy			0.93	500
macro avg	0.47	0.50	0.48	500
weighted avg	0.87	0.93	0.90	500

Gradient Boosting Model

Relative importance of features to determine customer churn

Questions?

Coupon Usage & Campaign Response

Coupons Redeemed vs Avg Basket Sales Value

There is a clear ceiling for the effectiveness of Type A Campaigns