Springboard - Data Science Track
Capstone Project 2:
Customer Segmentation & Churn Analysis

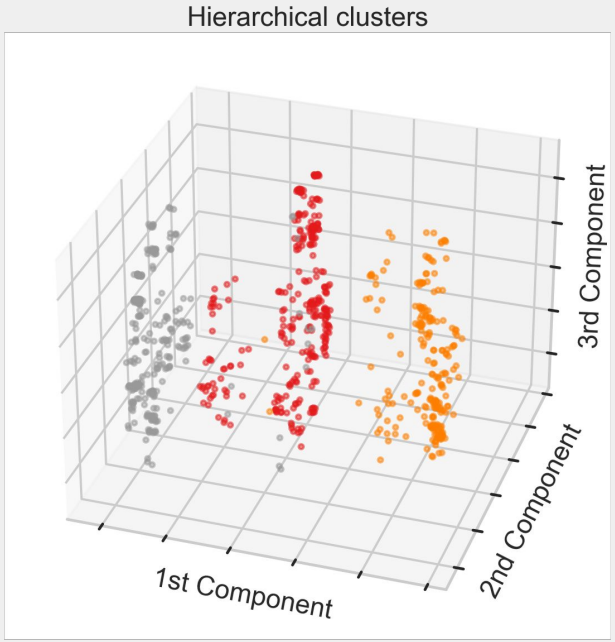
By: Nicholas Dean

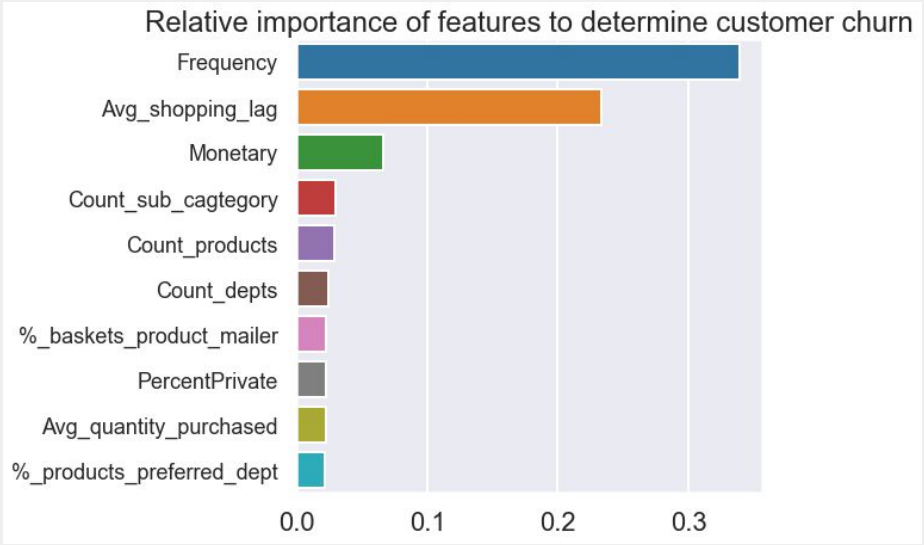
September 2021

Defining the problem & stakeholders

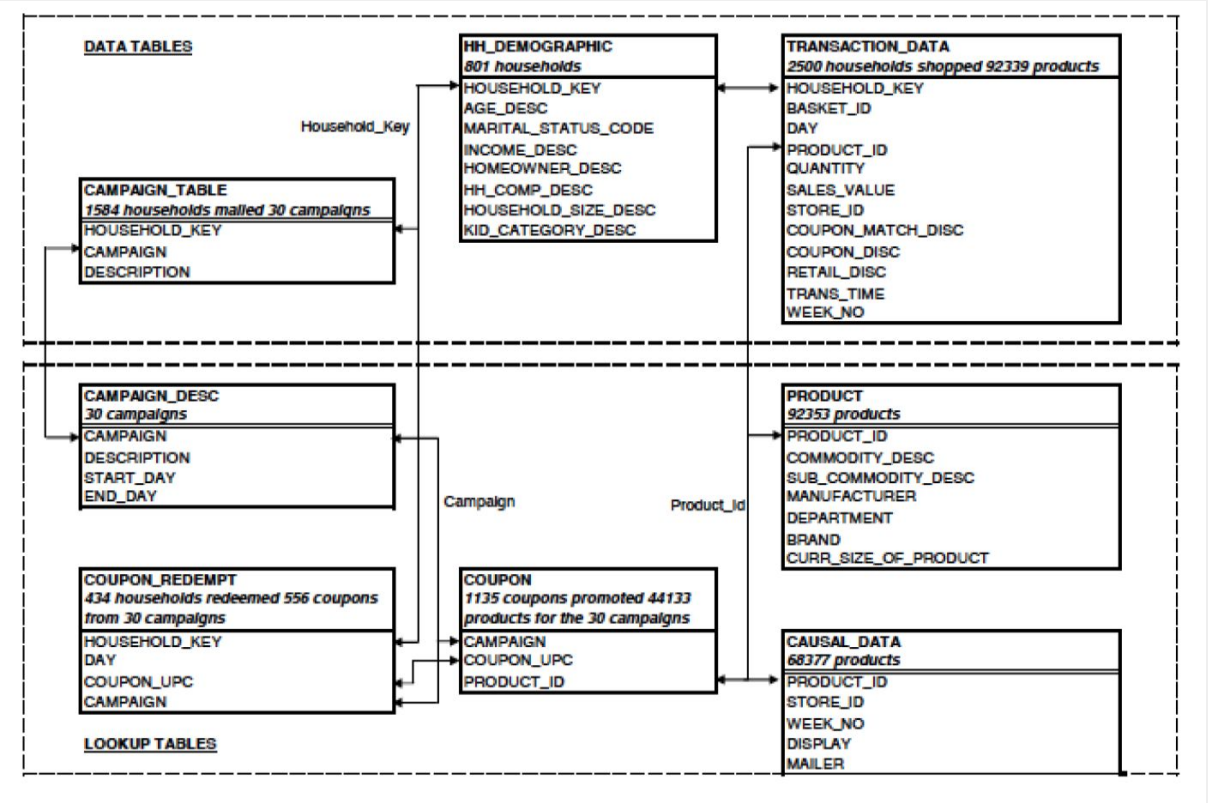
- According to HubSpot: “Customer churn is the percentage of customers that stopped using your company's product or service during a certain time frame”
- Additionally, Client Success estimates that about 65% of existing customers can be upsold to, vs a 13% conversion rate for new customers (a conversion 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
3rd Component
2nd Component
1st Component


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

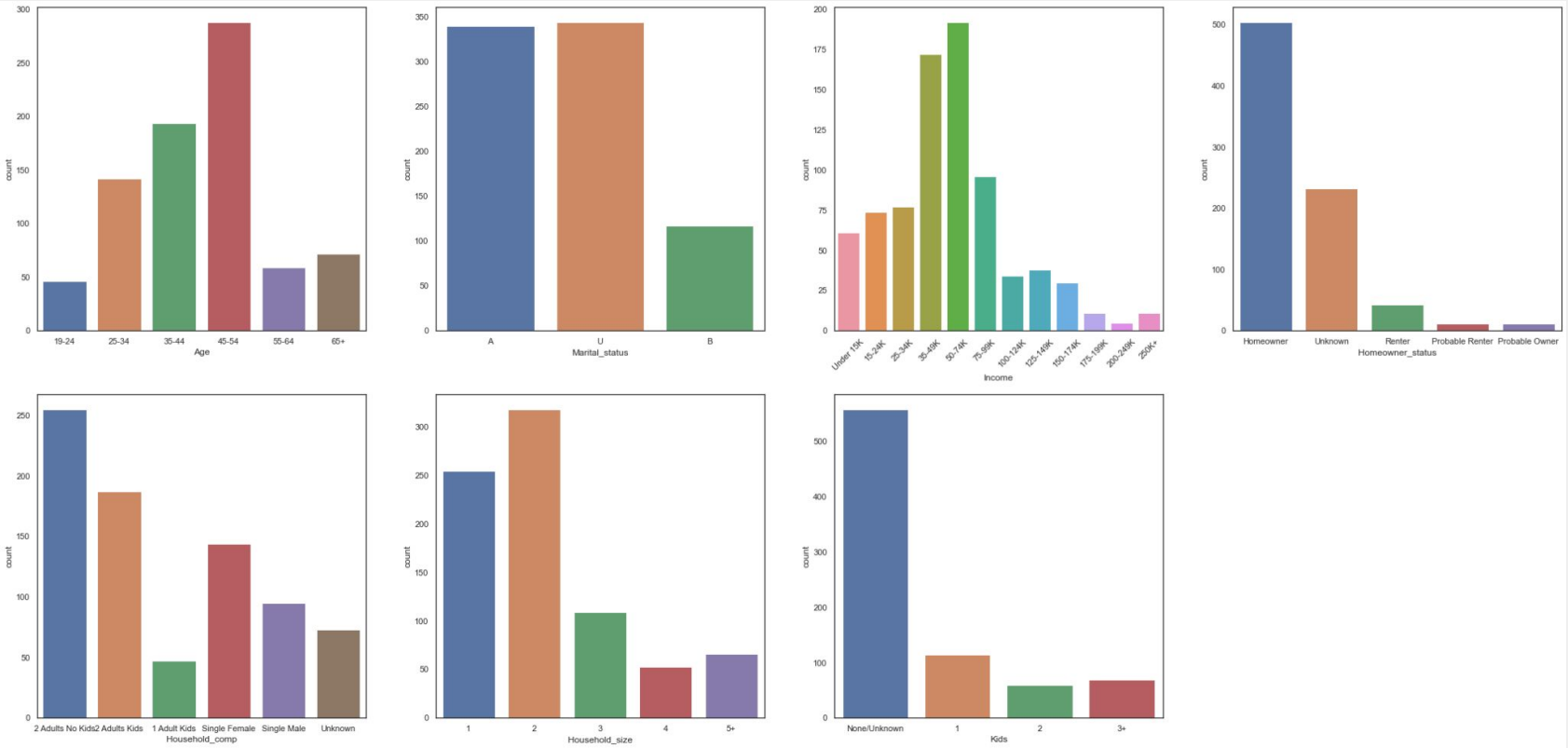
Data Acquisition



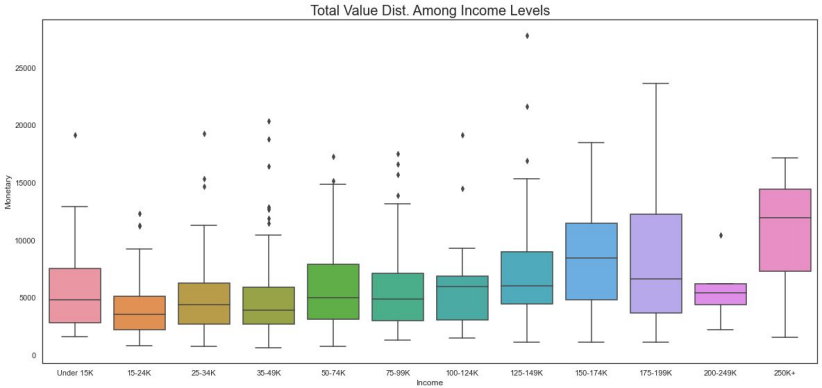
Outliers and Missing Data

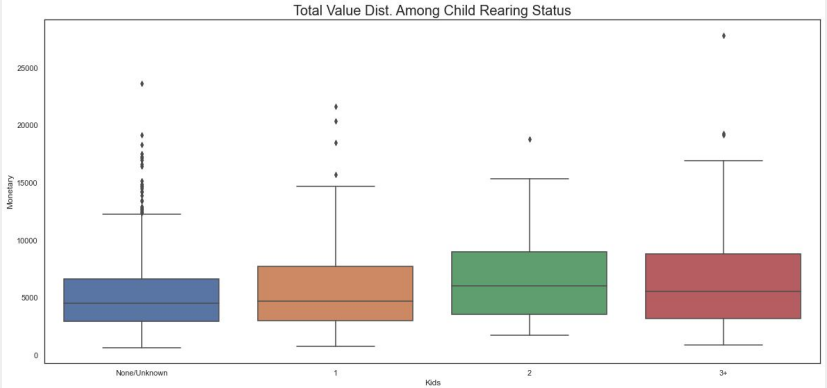
- Significant number of outliers with outrageous purchasing numbers
- It appeared that while 85 households were outliers in all noted qualities, over 700 were an outlier in one of those respects
- Potential for ‘corporate’ customers driving up the numbers, but need actual stakeholders to consult with 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)

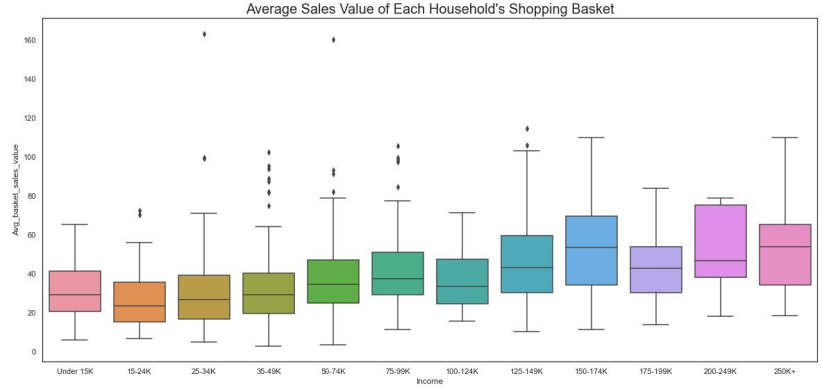
Demographic Data

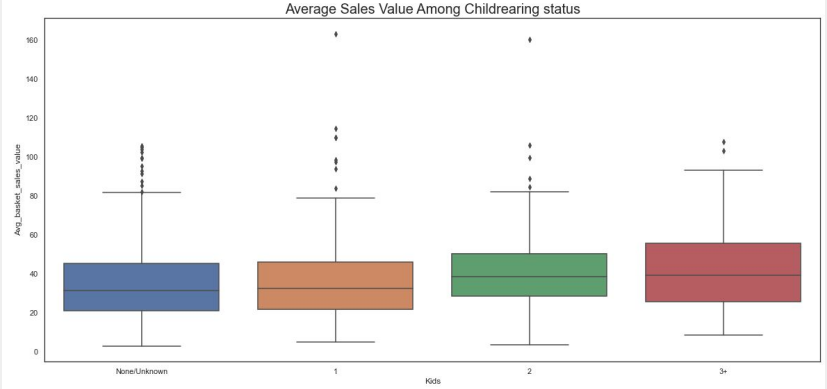


Household Income & Child-rearing Status

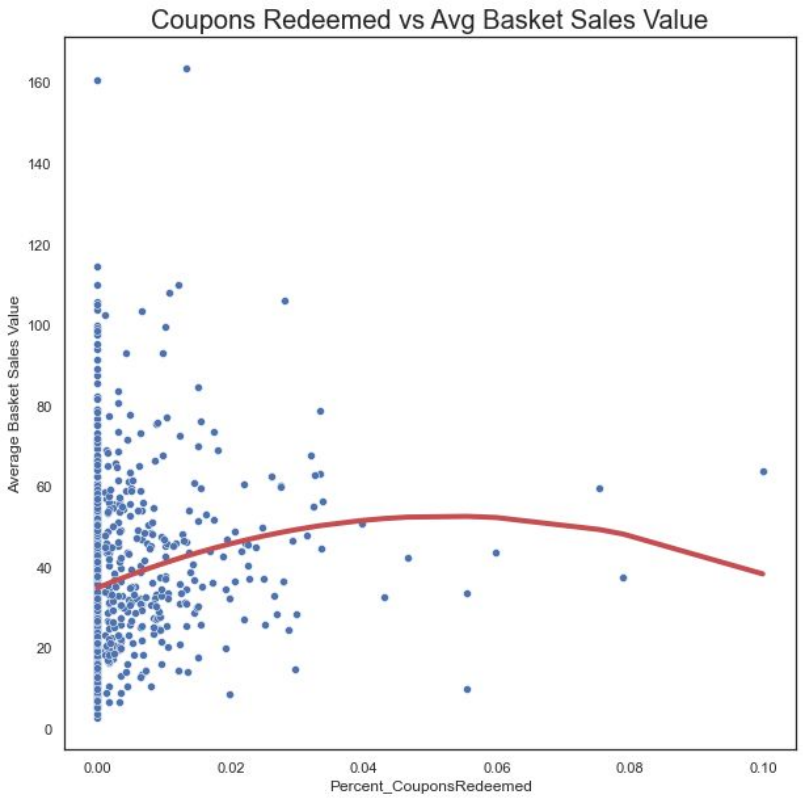


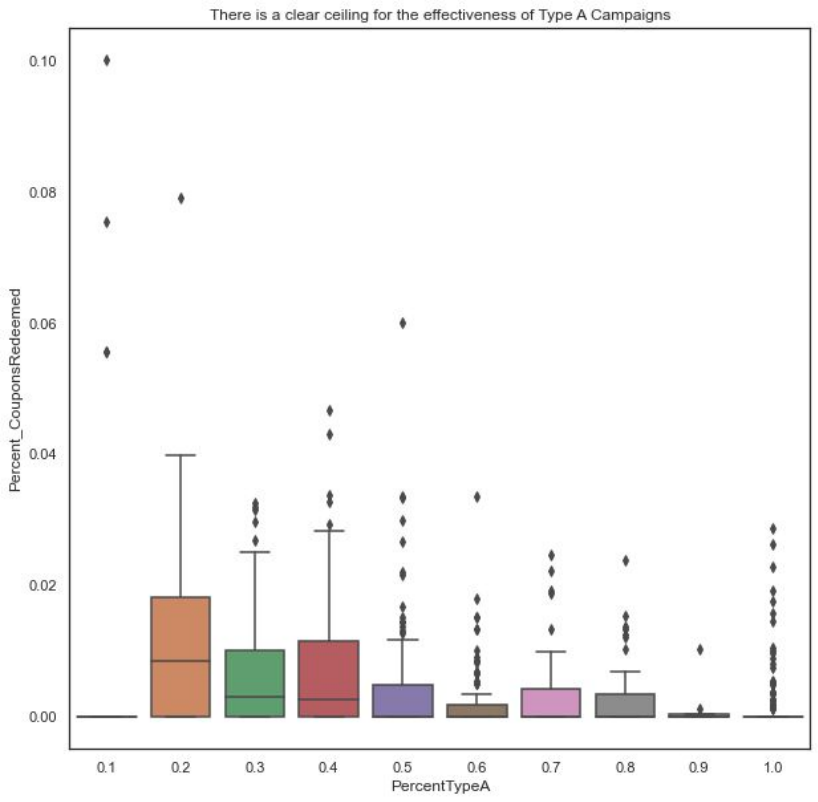




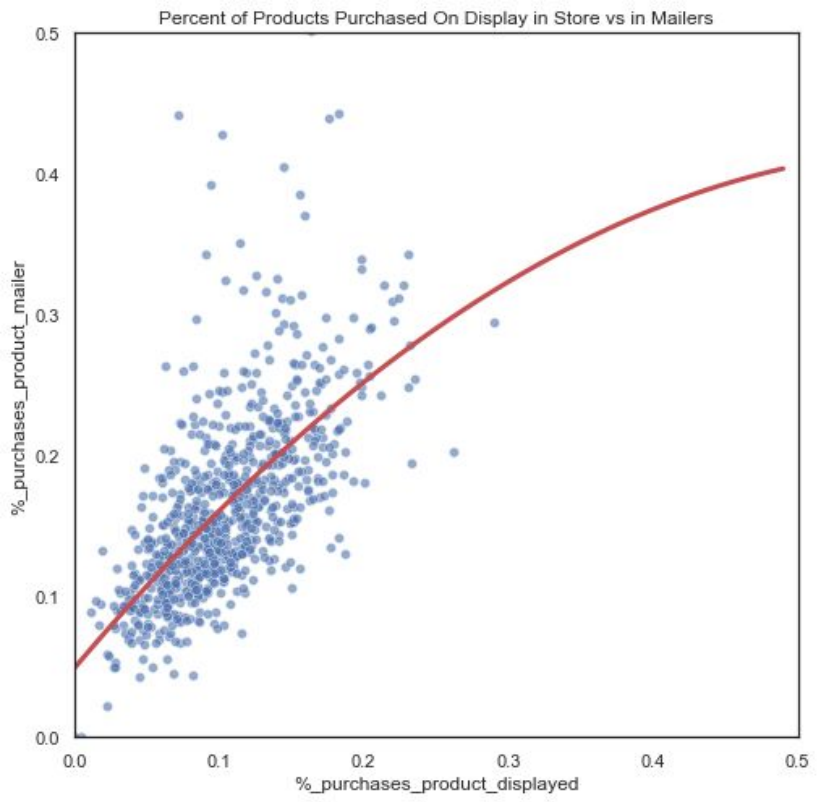


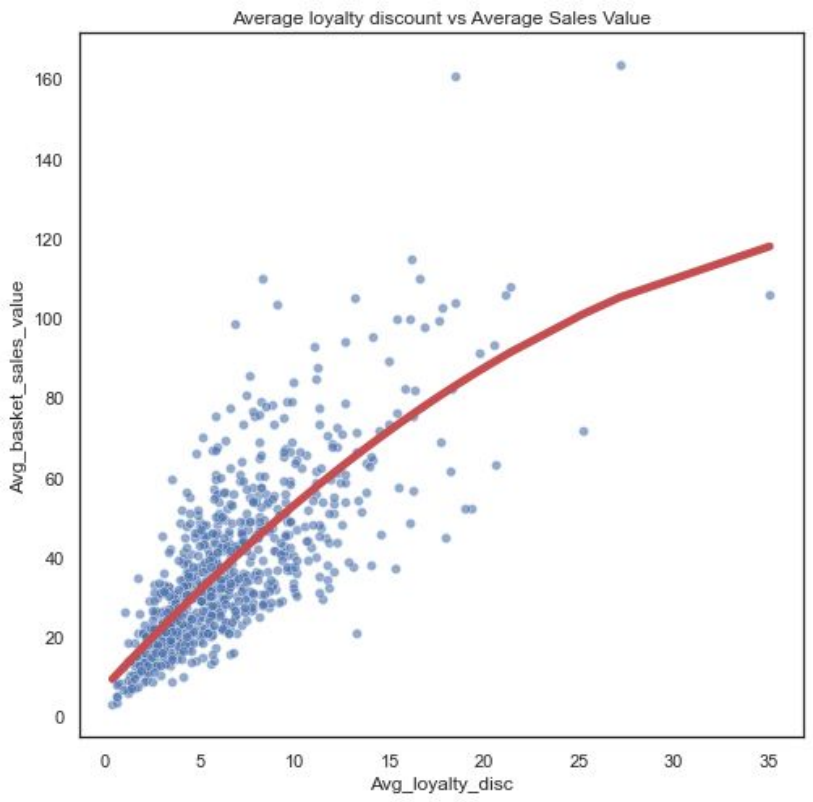
Coupon Usage & Campaign Response



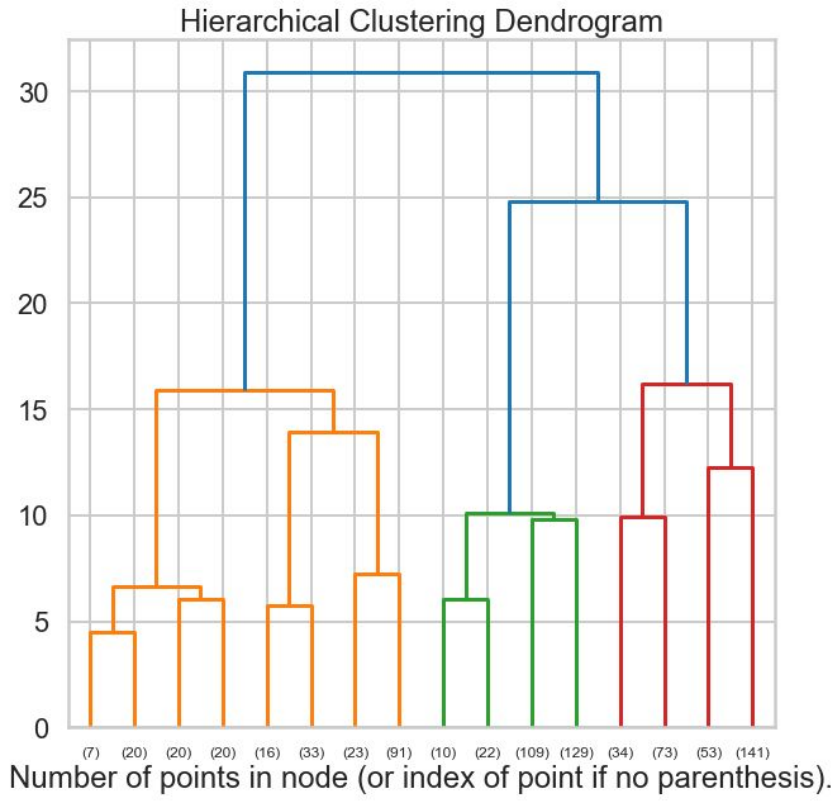


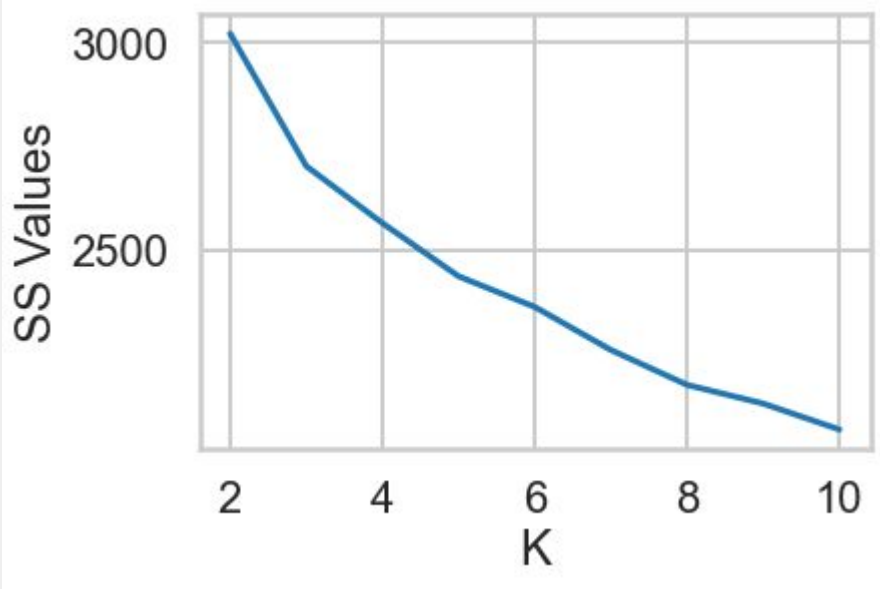
Efficacy of Marketing & Loyalty Discounts



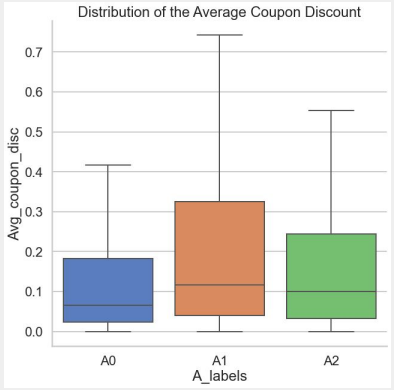


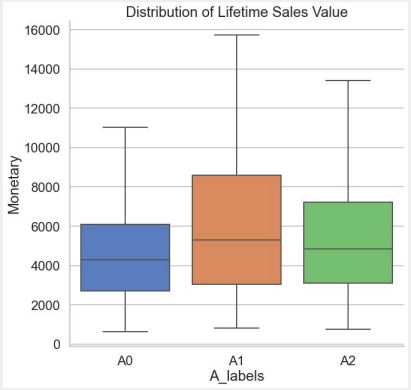
Baseline Clustering

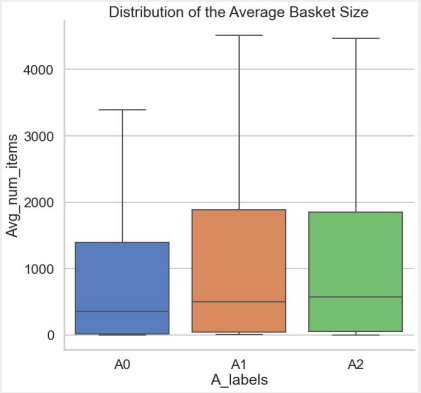


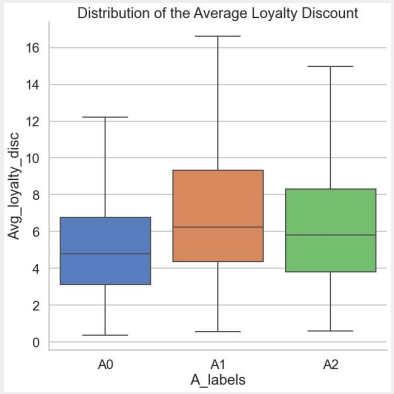


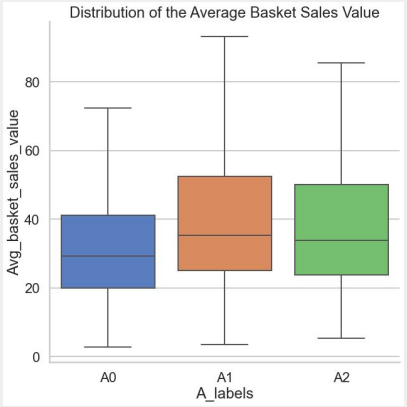
Key Differences With Segments

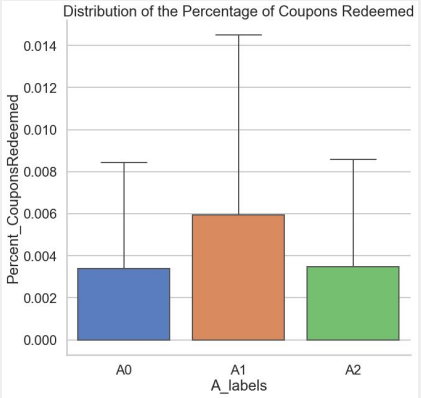










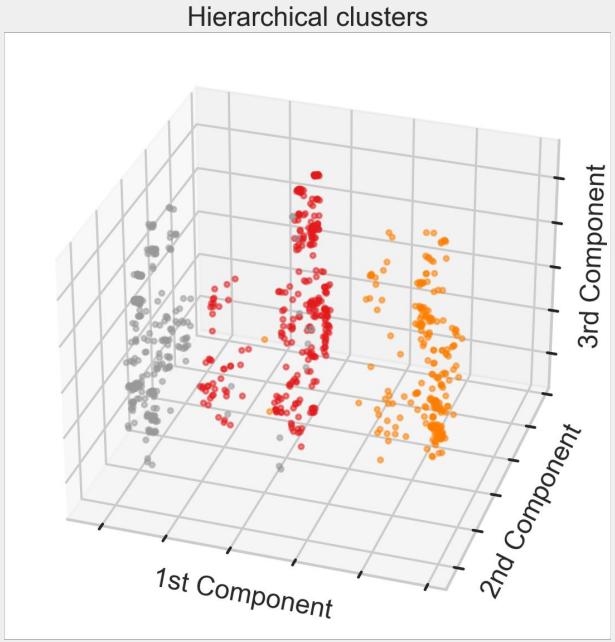


Customer Segment Summaries

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.



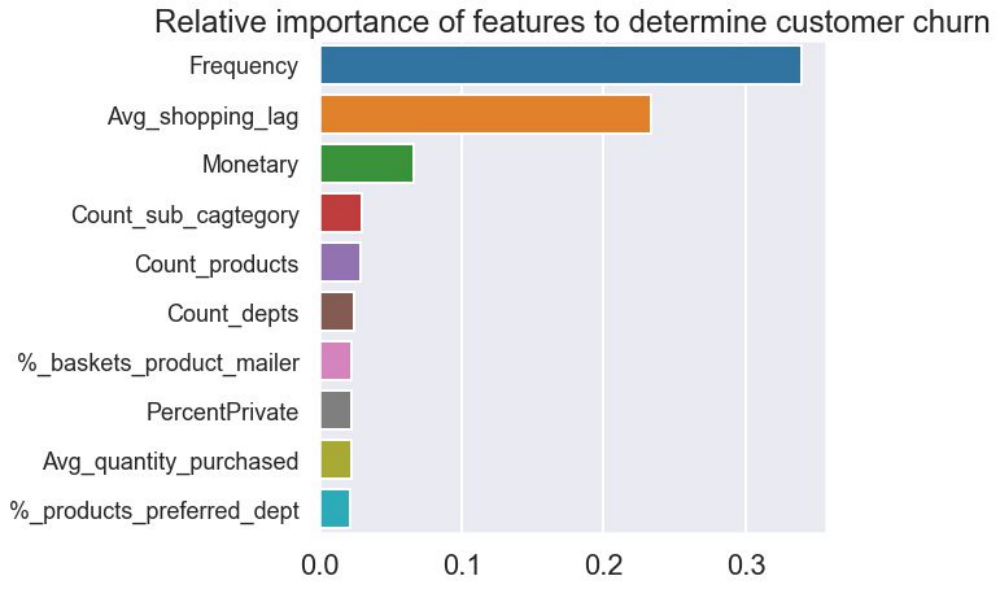
Defining Churn

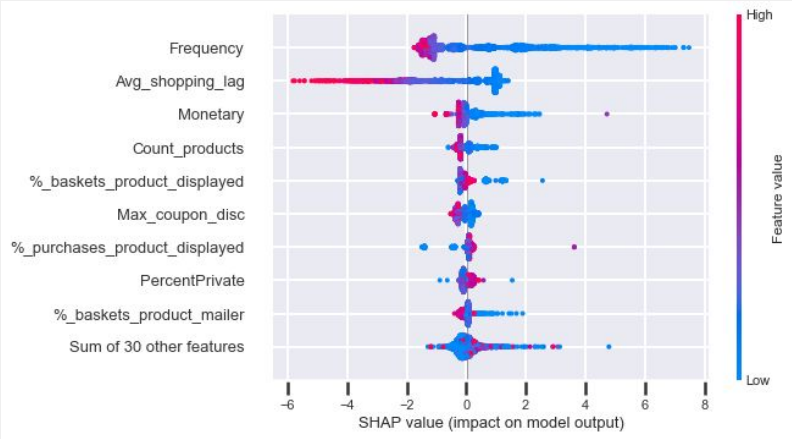
Baseline Modeling

- 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





Questions?