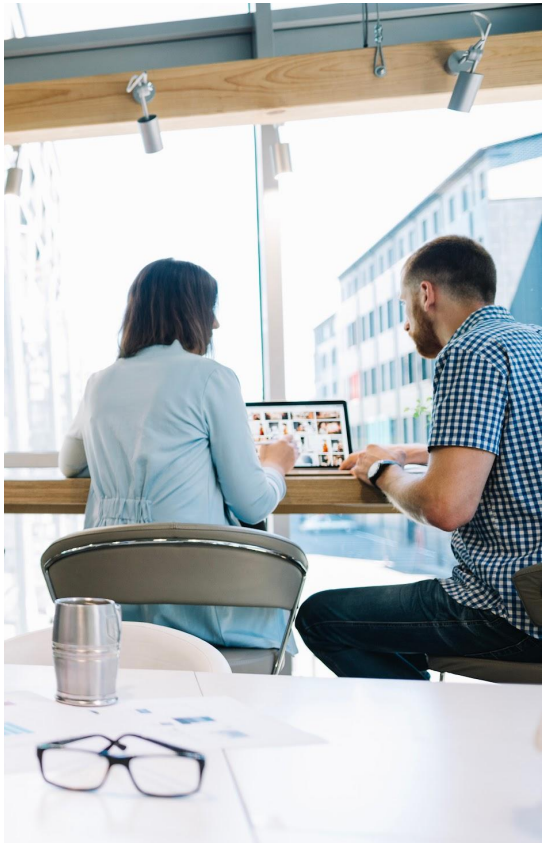


Market Campaign Analysis

Jiazhou (Jay) Pang
April 2024

01

Problem Identification

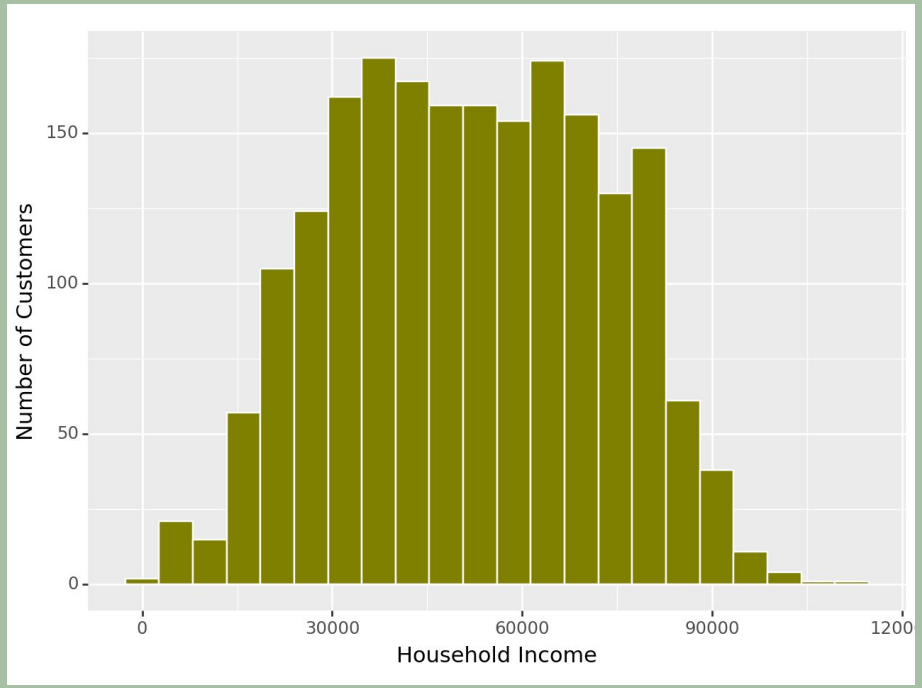
Identify the problems that can be addressed using the dataset

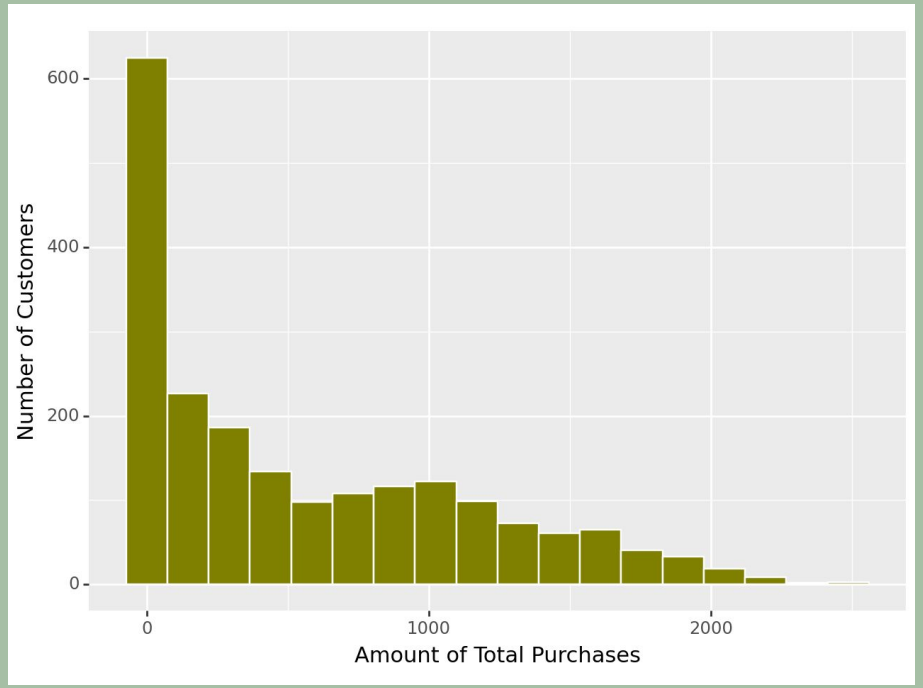
Understanding the Dataset

After removing null values and duplicate entries, the dataset :

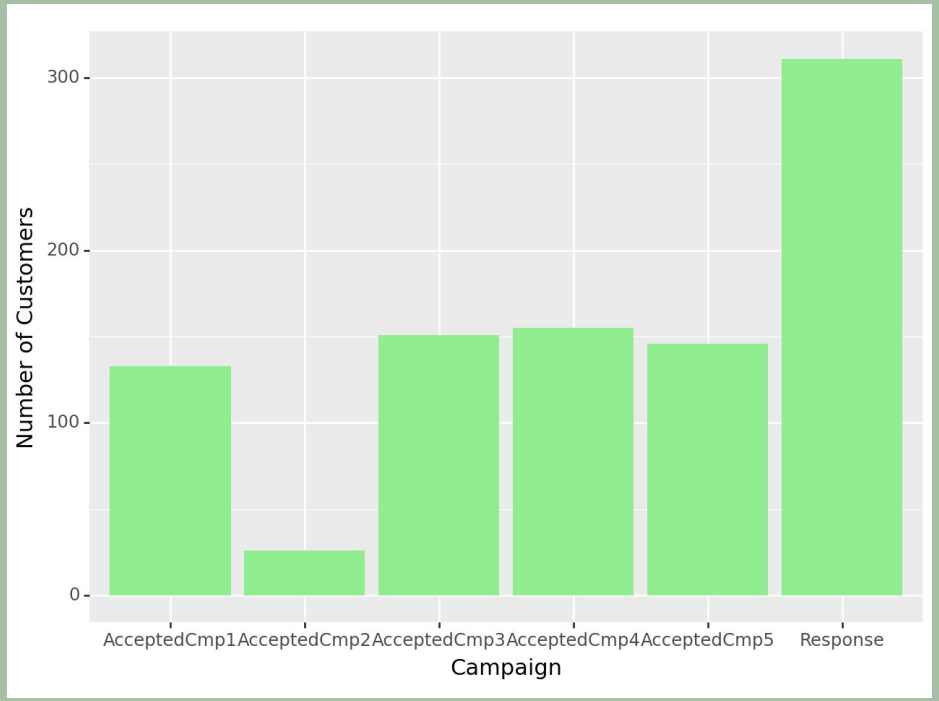
- Contains **2021** entries
- **12** categorical variables, **24** numerical variables
- Doe **not** contain timestamps

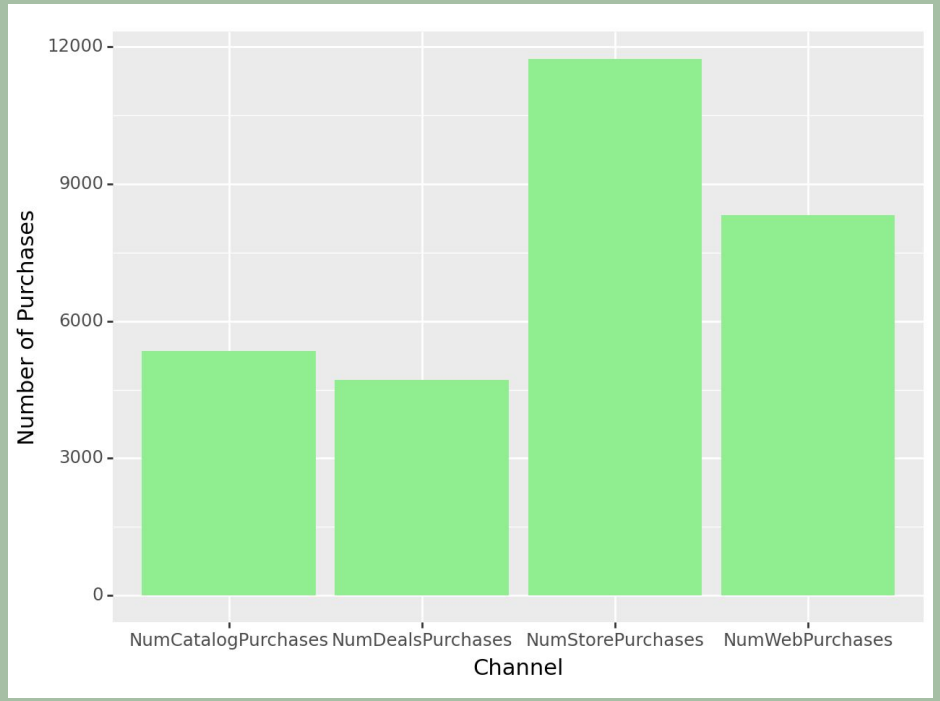
Customer Distribution in terms of income and purchases





Customer breakdown by campaigns and channels





Defining a “successful” campaign

There are several metrics to evaluate a campaign.
Here are some possible options:

- **Retention**— how well a campaign retains customer
- **Revenue** — how much customer spending increased.
- **Satisfaction** — Amount of customer complaints

Potential Challenges

- Relatively small dataset. Might be difficult to detect any statistical relationships
- Mixed features. Clustering models should be chosen with care
- Accessing multiple campaigns based on multiple metrics at the same time
- Primacy bias: the more early campaigns might be favored due to the fact that more time has past. The bias cannot be eliminated from the dataset due to the lack of timestamps

With these challenges in mind. We now can propose what can be achieved from analyzing this dataset.

The Goal of the project

Identify the most successful marketing campaign in terms of promoting customer spending (MntTotal). Estimate the treatment effect of said campaign

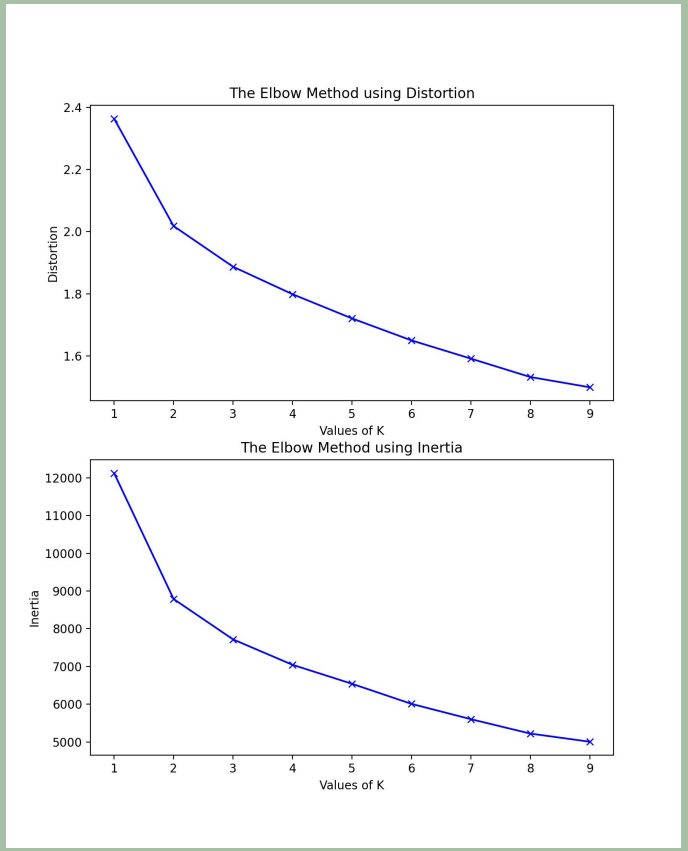
02

Data Analysis

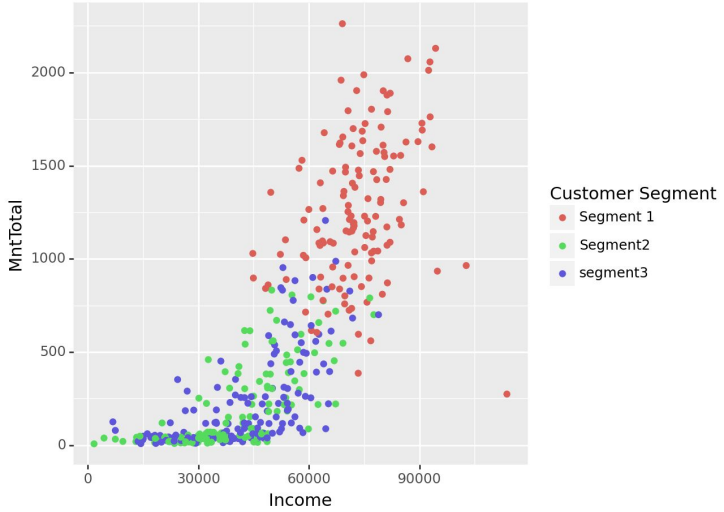
Glean information through statistical techniques

Creating Customer Segments

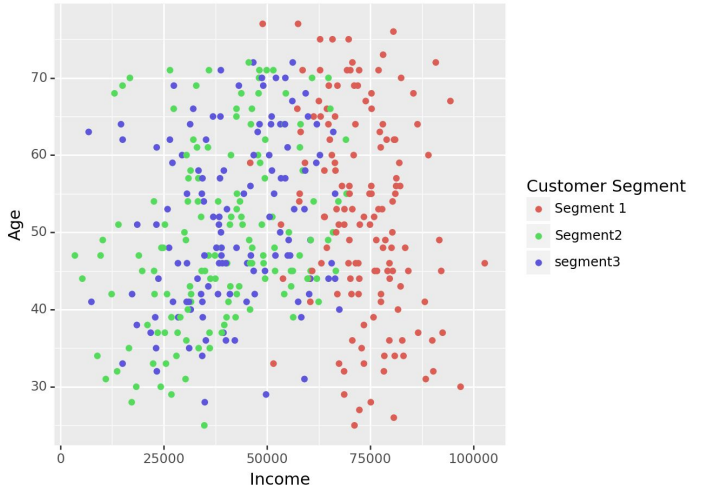
- Customers were clustered using observable variables (Income, MntTotal, Age, Dependants, Recency, and Cusstomer_Days)
- The elbow method were used to select a K Means algorithm of K=3 as both inertia and distortion become linear after 3.
- The three segments created have 687, 675, and 659 customers Respectively
- Silhouette score of 0.183 was achieved



Understand the customer segments



- Segment 1 has both the highest average MntTotal and Income
- All three segments have similar average age (53, 48, 51)
- Inverse relationship between dependants and MntTotal



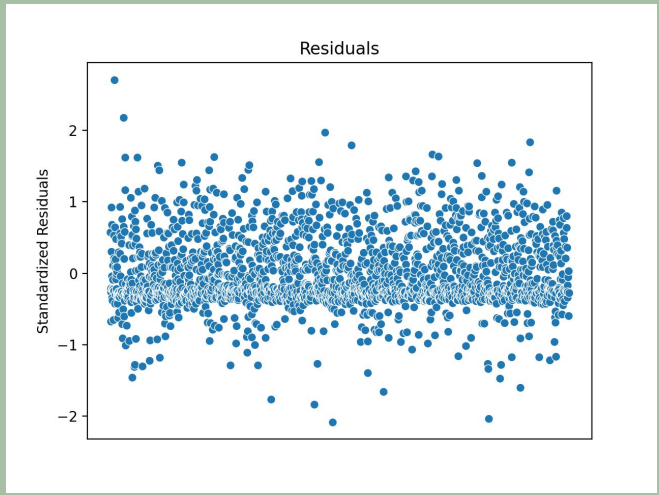
Regression Analysis

Using linear regression, we can determine the extent of association between variables. Assuming there are no missing variables or confounding factors, the association can be interpreted causally.

Variables Used:
Target: MntTotal
Independent: AcceptedCmp variables, Response, Customer Segment

I tested the multicollinearity of the data using both Variance Inflation Factor (VIF) and condition index. Both tests indicated low multicollinearity.

This implies that the coefficients from the Regression analysis are robust. The treatment effect of each campaign can be approximated despite using observational data.



Results from Regression Analysis

- R-squared value of 0.749 was achieved
- Campaign 5 is associated with the greatest increase in MntTotal, followed by campaign 1 and the most recent campaign.
- Weak statistical evidence that campaign 2 and 3 improved MntTotal
- Complain rate is not associated with any of the explanatory variables
- Campaign 5 is the most effective on all customer segments except segment 3, whose most significant campaign is 1.

OLS Regression Results

Dep. Variable:	MntTotal	R-squared:	0.749
Model:	OLS	Adj. R-squared:	0.748
Method:	Least Squares	F-statistic:	749.4
Date:	Tue, 23 Jul 2024	Prob (F-statistic):	0.00
Time:	17:55:36	Log-Likelihood:	-1472.0
No. Observations:	2021	AIC:	2962.
Df Residuals:	2012	BIC:	3013.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9713	0.023	42.083	0.000	0.926	1.017
C(AcceptedCmp1)[T.1]	0.3061	0.052	5.925	0.000	0.205	0.407
C(AcceptedCmp2)[T.1]	-0.1385	0.105	-1.314	0.189	-0.345	0.068
C(AcceptedCmp3)[T.1]	0.0371	0.044	0.836	0.404	-0.050	0.124
C(AcceptedCmp4)[T.1]	0.1305	0.046	2.825	0.005	0.040	0.221
C(AcceptedCmp5)[T.1]	0.4701	0.052	9.077	0.000	0.369	0.572
C(Response)[T.1]	0.1315	0.035	3.788	0.000	0.063	0.200
C(Pred)[T.1]	-1.6562	0.030	-56.078	0.000	-1.714	-1.598
C(Pred)[T.2]	-1.5439	0.030	-51.812	0.000	-1.602	-1.486

Omnibus:	201.856	Durbin-Watson:	1.884
Prob(Omnibus):	0.000	Jarque-Bera (JB):	358.666
Skew:	0.678	Prob(JB):	1.31e-78
Kurtosis:	4.556	Cond. No.	10.8

03

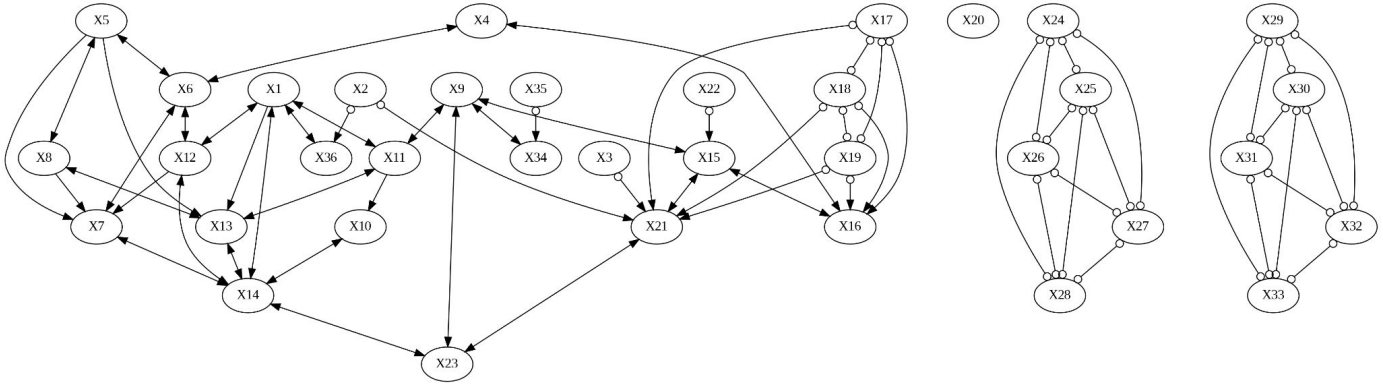
Causal Inference

Identify the effectiveness of the most recent campaign

Causal relationships discovered using FCI

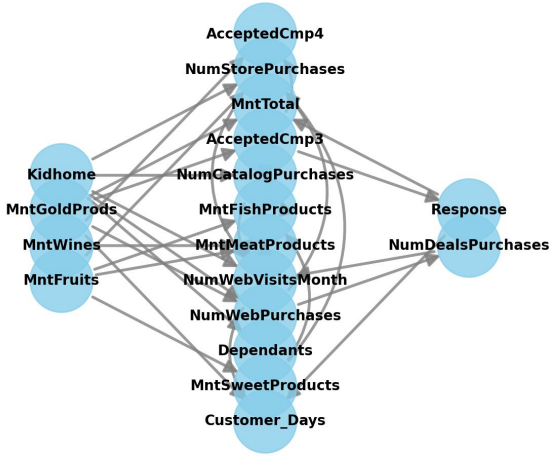
From results from the regression analysis, we know that campaigns do affect MntTotal collectively. Now, we can identifying whether the recent campaign is needed a cause for increased customer spending by itself.

To achieve that, a causal discovery algorithm was implemented to identify potential causal relationships. The output causal map is as follows:



Causal Inference

- Previous causal map was loaded into DoWhy to identify whether response (treatment) causes MntTotal (target) to increase.
- Average treatment effect is 0.540, p-value close to zero
- Refuted result using a placebo treatment, low p-value obtained, indicates causal effect likely exists



04

Evaluation

Summarize findings and food for thought

Summary

- The recent campaign and campaign 5 are effective in generating revenue
- Campaign 1 is relatively more effectively on customer with lower income and higher dependants
- More data is needed to compare the extent of effectiveness between campaign 1, 5, and recent.
- Catalog channel most associated with campaigns

Areas of Improvement

- Get timestamps for acceptance of campaigns to better estimate the individual effectiveness of campaigns.
- More information on the types of campaigns conducted
- Gather information from a greater population of customers
- All participants provided income information, this may induce a self-selection bias
- More Data