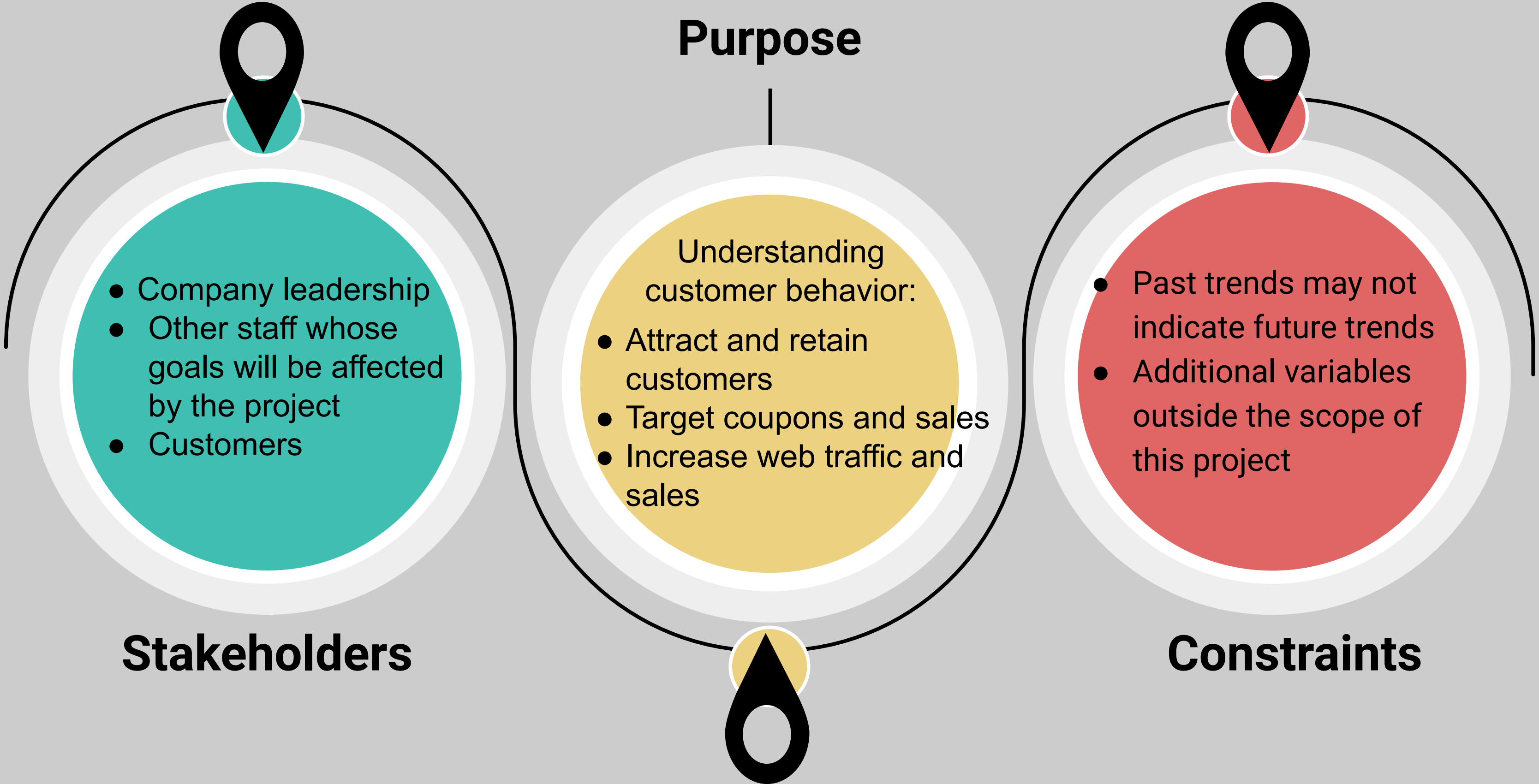
Customer

Segmentation

Unsupervised machine learning in Python

Context



Stakeholders

- Company leadership
- Other staff whose goals will be affected by the project
- Customers

Purpose

Understanding customer behavior:

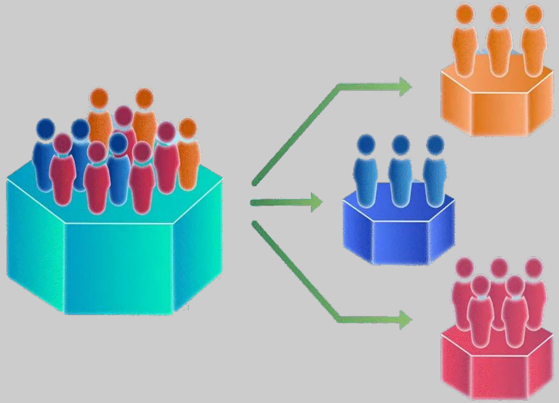
- Attract and retain customers
- Target coupons and sales
- Increase web traffic and sales

Constraints

- Past trends may not indicate future trends
- Additional variables outside the scope of this project

Problem Statement:

Grouping customers based on personal attributes and purchasing history can help companies understand their customers and aid in targeted marketing and other key business decisions.



About the data:

"Customer personality analysis" by Dr. Omar Romero-Hernandez, uploaded by Akash Patel

Csv file containing 2240 observations with 29 variables:

Demographics

ID

Year_Birth

Education

Marital_Status

Income

Kidhome

Teenhome

Dt_customer: date enrolled

Recency: last purchase

Complain: 1=yes/2=no

Purchases

Spent in the past 2 years:

MntWines: on wine

MntFruits: on fruit

MntMeatProducts: meat

MntFishProducts: fish

MntSweetProducts: sweets

MntGoldProds: gold

Discounts

Customer participated in:

NumDealsPurchases: total discount purchases

AcceptedCmp1: 1st campaign

AcceptedCmp2: 2nd campaign

AcceptedCmp3: 3rd

AcceptedCmp4: 4th

AcceptedCmp5: 5th

Response: last campaign

Shopping habits

Location of purchases:

NumWebPurchases

NumCatalogPurchases

NumStorePurchases

NumWebVisitsMonth

Data Wrangling

NaN

Problem: missing data

Imputed missing values with mean value for the column

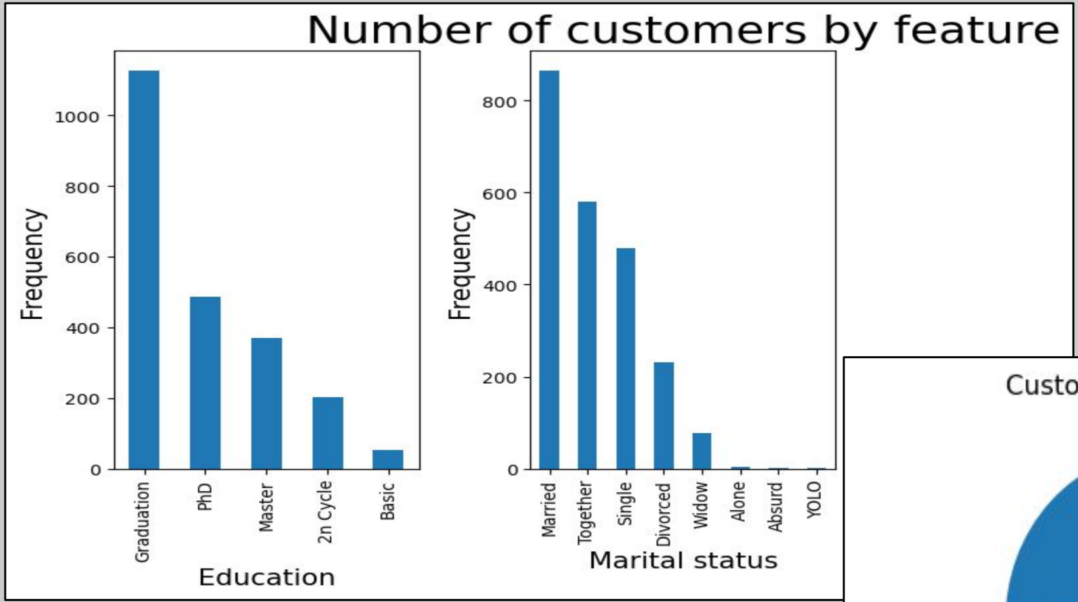
Problem: different formats

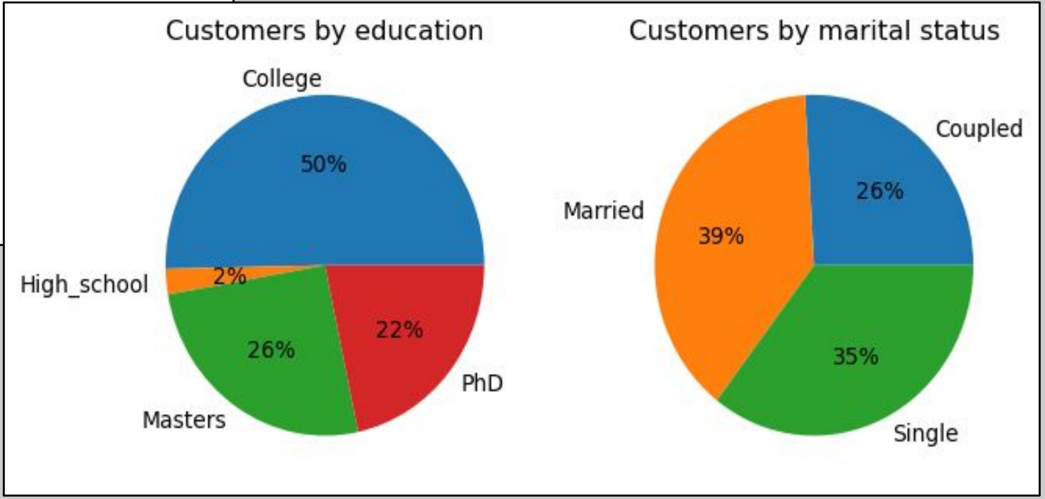
Dropped unusable categorical responses and combined extraneous categories into fewer, logical groups

Problem: extra variables

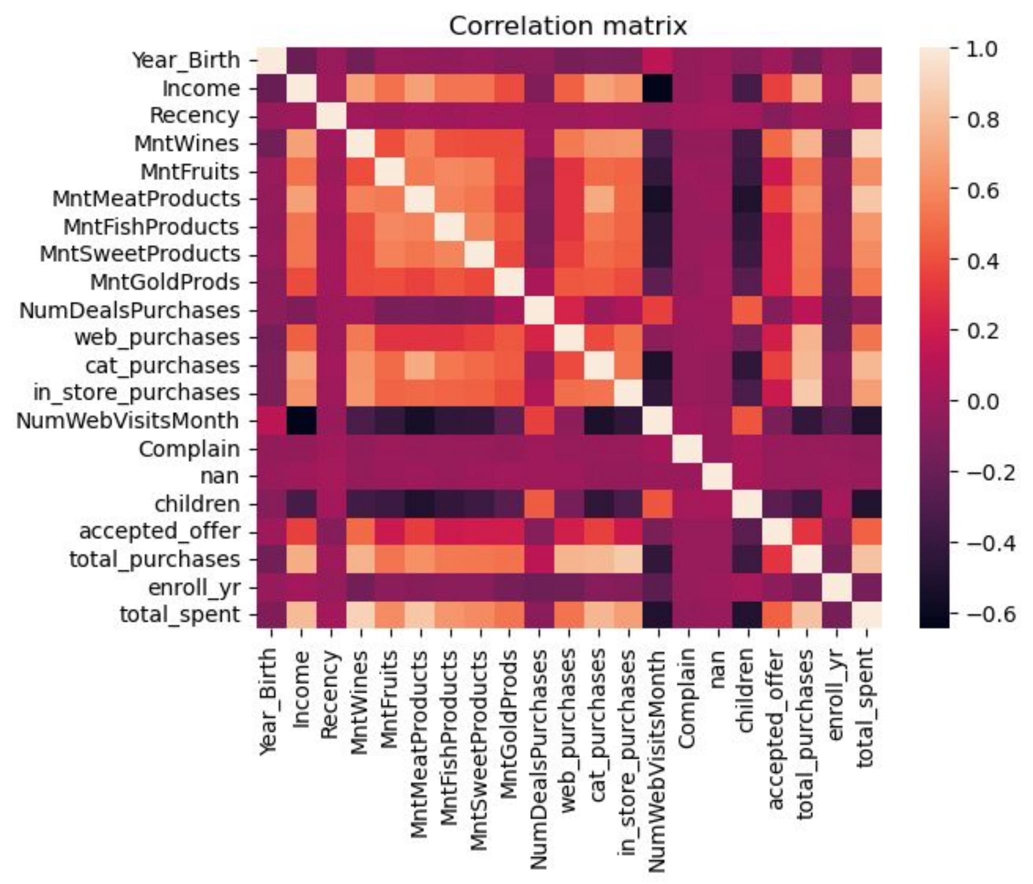
Combined logical features like kidhome and teenhome to children

Exploratory Data Analysis





Exploratory Data Analysis



Exploratory Data Analysis

Key findings

Univariate analysis

Examining rows with missing values show no significant different compared with other rows.

Univariate distribution

The distribution of each variable is fairly uniform with many features heavily right skewed.

Distribution

Right skewed features suggest many customers purchase or participate minimally or not at all and less customers purchase or participate at higher levels.

Outliers

There are 3 customers whose birth years before 1940, all of which are around 1900.

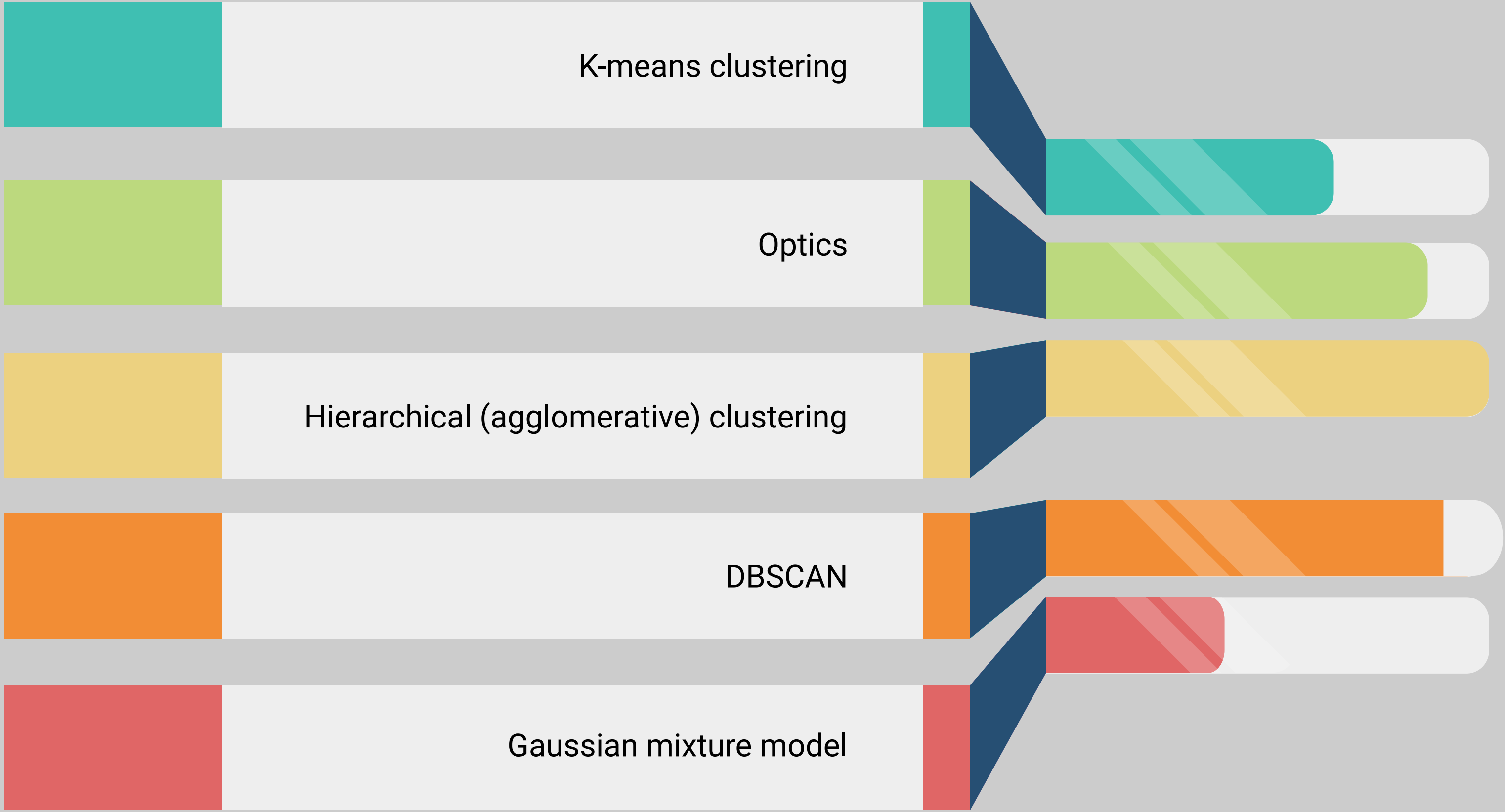
Bivariate analysis

No two variables appear to be so highly correlated as to either create cause for excitement or concern in modeling.

Correlation matrix

High correlations seem to be between total purchases and in store purchases, total spent and wine purchases, and number of purchases and total spent.

Machine Learning Algorithms



K-means clustering

Optics

Hierarchical (agglomerative) clustering

DBSCAN

Gaussian mixture model

Refining Models

8	K-Means
38	Optics
2	Agglomerative
1	DBSCAN
2	GMM

Overall best number of clusters **2**

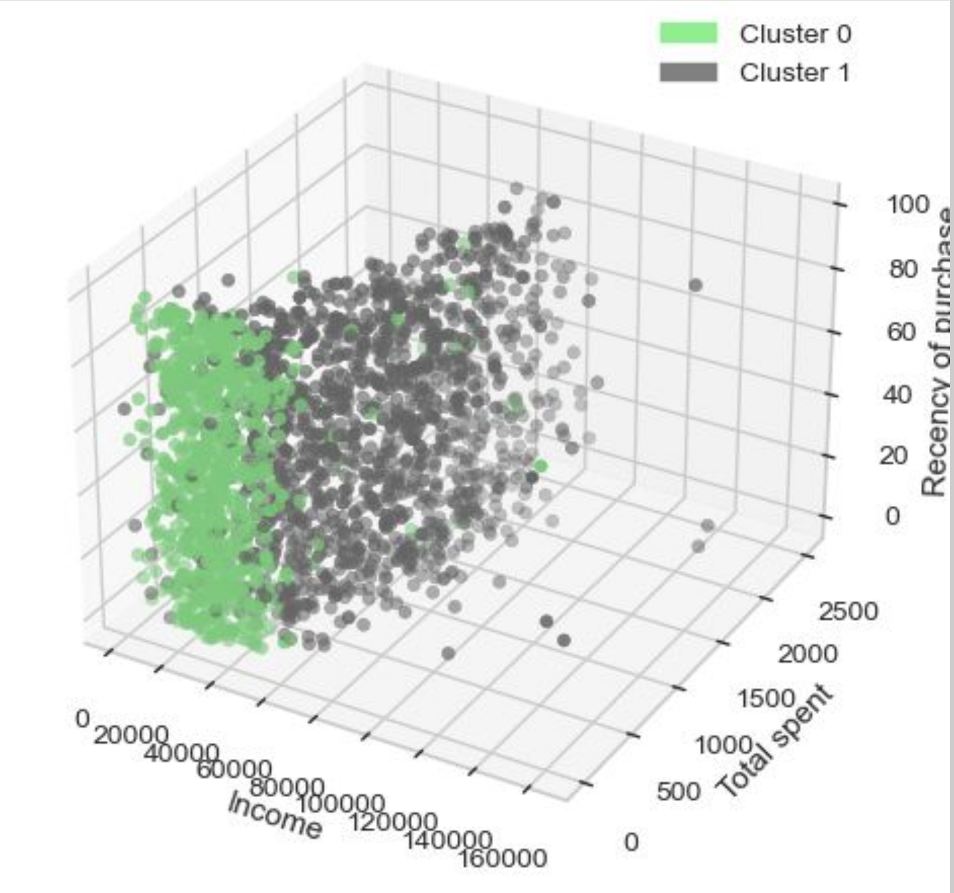
Model Metrics

Model name	Cluster #	Other hyperparameters	Silhouette score
Agglomerative	2	Complete linkage, euclidean metric	0.522831
	2	Average linkage, euclidean metric	0.522831
	2	Average linkage, manhattan metric	0.522831
	2	Single linkage, manhattan metric	0.522831
	3	Single linkage, manhattan metric	0.476398
	2	Single linkage, cosine metric	0.446398
K-Means	2	Lloyd algorithm	0.474022
	2	Elkan algorithm	0.474022
	3	Lloyd algorithm	0.282738
	3	Elkan algorithm	0.282738
	5	Lloyd algorithm	0.062095

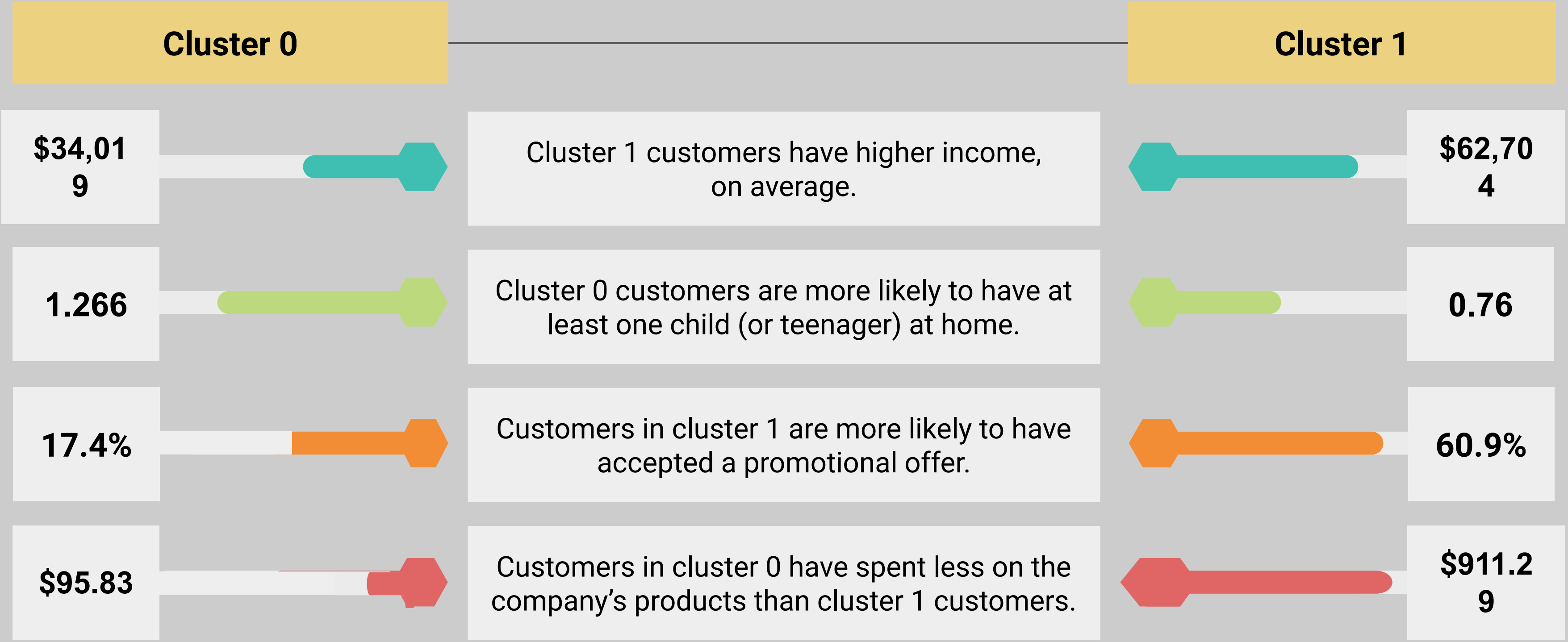
Model Metrics

Model name	Cluster #	Other hyperparameters	Silhouette score
GMM	2	Tied covariance	0.456527
	2	Full covariance	0.401546
	3	Tied covariance	0.287145
	3	Full covariance	0.1668
DBSCAN	—------	epsilon=10, p=2	0.446398
	—------	epsilon=10, p=1	0.446398
	—------	epsilon=8, p=2	0.443647
	—------	epsilon=9, p=2	0.438544
Optics	3	p=2	-0.317792
	2	p=2	-0.320906
	3	p=1	-0.345548

Visualizing Clusters



Conclusions from Clusters



Cluster 0		Cluster 1
$34,019	Cluster 1 customers have higher income, on average.	$62,704
1.266	Cluster 0 customers are more likely to have at least one child (or teenager) at home.	0.76
17.4%	Customers in cluster 1 are more likely to have accepted a promotional offer.	60.9%
$95.83	Customers in cluster 0 have spent less on the company's products than cluster 1 customers.	$911.29

Recommendations

For cluster 0:

the group that has thus-far been purchasing less items:

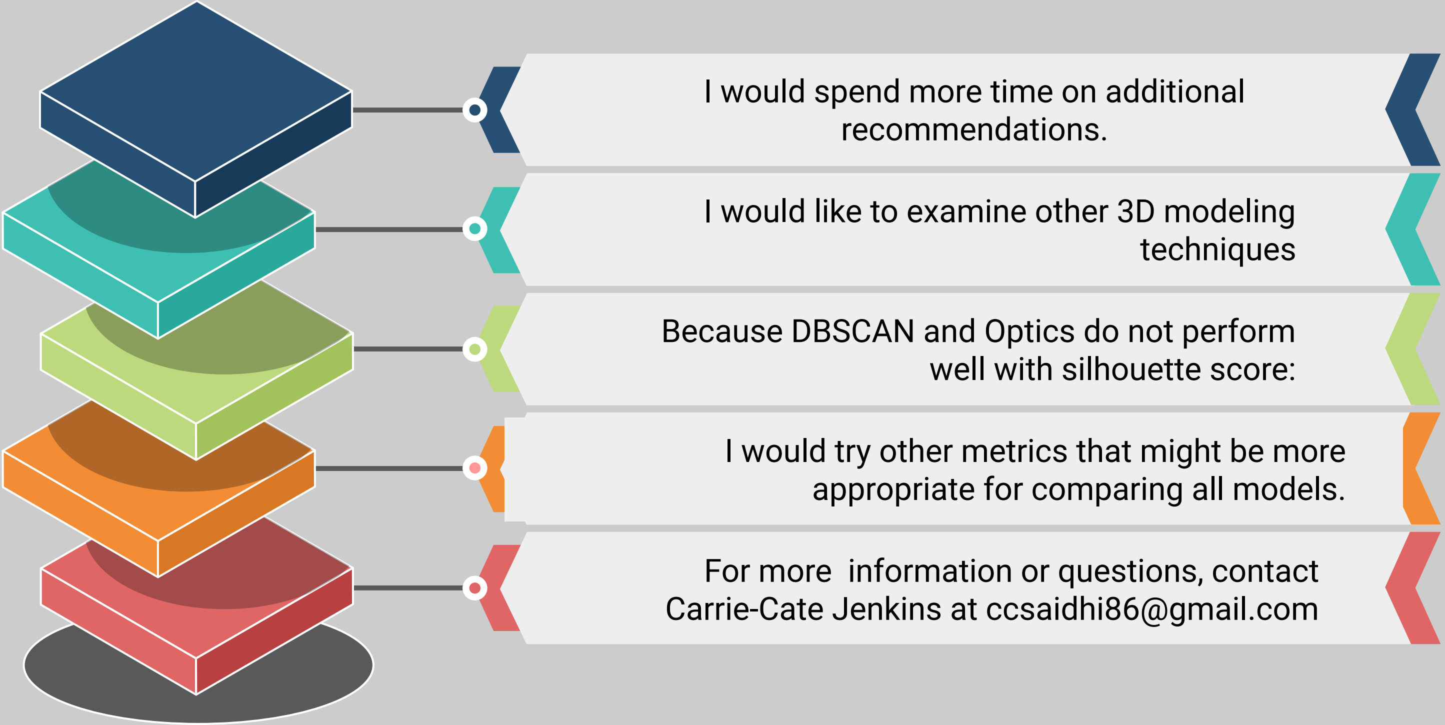
The company may want to examine this group to see whether they can increase sales within this set of their customers.

For cluster 1:

the group that tends to purchase more items:

The company may want to target this group when they are having sales or are pushing any initiatives toward customers who are more reliable buyers.

Additional Thoughts and Future Research



I would spend more time on additional recommendations.

I would like to examine other 3D modeling techniques

Because DBSCAN and Optics do not perform well with silhouette score:

I would try other metrics that might be more appropriate for comparing all models.

For more information or questions, contact Carrie-Cate Jenkins at ccsaidhi86@gmail.com

Thanks and Credits

- Special thanks to Silvia Seceleanu for her guidance and support.
- Credit to Springboard for curriculum and project ideas.
- Slide template designed by Slidesgo school.