HUNT N' GATHER

an eclectic online store

CUSTOMER SEGMENTATION
K-MEANS ALGORITHM

Created by: Jenni Hawk, February 2023

OBJECTIVES

- Identify customer groups that haven't been uncovered by looking for signals in the data from a completely numeric perspective

- Utilize k-Means algorithm
  - Explore other algorithms as a next phase of work

WHY USE A CLUSTERING ALGORITHM?

Leverage advanced math and computational power

Identify new customer clusters and insights by looking for signals in the data from a completely numeric perspective

Ability to process more purchasing behaviors

Includes all customers, not just a small sample
Strategies are implemented on the exact customer

Utilize k-Means algorithm as starting point
Explore other algorithms as a next phase of work

PROCESS

01

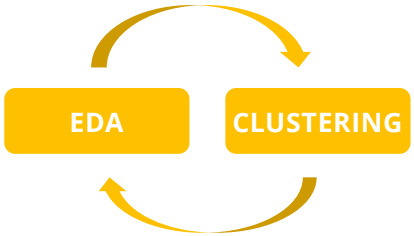
Exploratory Data Analysis

02

Define Clusters

03

Next Step Strategies



AVAILABLE DATA

Rows
541,909 Invoice Line Items

8 Columns
- Invoice No
- Stock Code
- Product Description
- Customer ID
- Country
- Quantity Ordered
- Invoice Date / Time
- Unit Price

Unique Values
- 25,900
- 4,070
- 4,223
- 4,372
- 38 Countries / 91% transactions come from UK
- Range: 1 – 80,995
- 2010 & 2011, Months 1-12
- NA

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

FIRST LOOK EDA OBSERVATIONS

Finding	Approach
2010: Only December Data Exists	Don't use 2010 Data
2011: Data for all 12 months, Incomplete Data for December (Missing day 10-31)	Use all 2011 data
Country: 91% of transactions from UK	Focus on UK
Appears there's a business customer in addition to retail customer because there are customers with extraordinary large spending and qty purchase	Don't remove these as outliers

STRATEGIC CONSIDERATIONS FOR K-MEANS FEATURE ENGINEERING

- Don’t create categorical variables
  - Not applicable to k-means - not relevant to Euclidean distance

- Be mindful to not creating features that would somehow replicate data that was already in another feature to prevent false strong signal

- Stay away from binning the math as to not create bias

ENGINEER FEATURES BASED ON BUYING BEHAVIORS

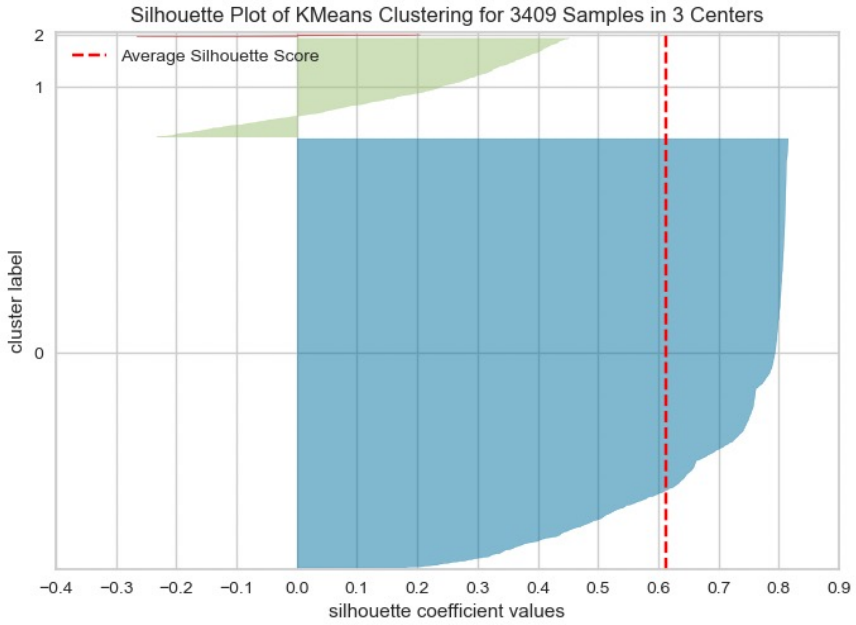
	BUYING BEHAVIORS	Feature(s) to Engineer
WHO Business and Retail Customer	**What role does quantity of purchased items play in clustering?** **What role does number of orders play in clustering?**	**Sum_Qty =** sum of total qty of items purchased per customer **Num_Orders** = Count number of invoices per customer
HOW MUCH MONEY Are they spending	**What role does annual spend play in the creation of clusters?** • Which segments, and corresponding behaviors, could be shifted based on spend level? • What characteristics do segments have based on annual spend level?	**Annual_Spend =** Sales Sum by Customer
WHEN Are they spending	**Does number of months they purchase provide us with insight?** • For example, does a customer who purchases in 3 unique months cluster differently from a customer that purchases in 6 unique months? • If so, what other characteristics define these people?	**Unique_ Month =** Count of number of unique months a customer purchased
WHAT ITEMS / HOW OFTEN	**Does the number of unique stock codes purchased play a role in clustering?**	**Num_Unique_Stock =** Number of unique stock codes that a customer purchased

01

OPTIMAL NUMBER OF CLUSTERS

SILHOUETTE SCORING: TWO VS THREE CLUSTERS

Three Clusters don’t work



Cluster isn’t Dense

- Cluster 1 has tail going left which means it's closer to other cluster than its own cluster

Clusters are not as clearly distinguished

- Silhouette coefficient for 3 clusters: 0.6
- Compare to coefficient for 2 clusters at 0.9

Scoring Key:

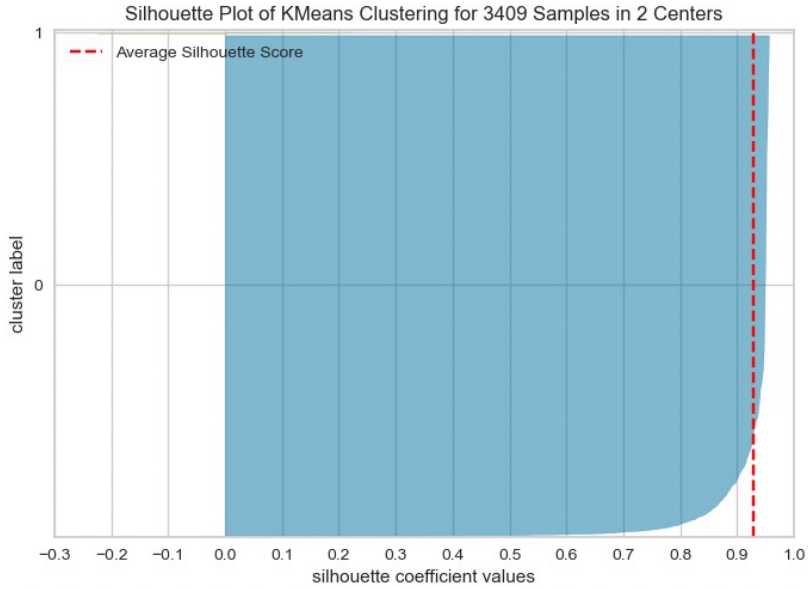
Score = 1: Clusters well apart and clearly distinguished

Score = 0: Distance between clusters not significant

Score = -1 Clusters are assigned in the wrong way

SILHOUETTE SCORING: TWO CLUSTERS WIN

Given the data / features it was provided



Clusters are Dense

- Don't have tails pointing left

Clusters are clearly distinguished

- Silhouette coefficient for 2 clusters: 0.9
- Note: Due to low volume of observations in cluster 1 the plot isn't colorizing

Scoring Key:

Score = 1: Clusters well apart and clearly distinguished

Score = 0: Distance between clusters not significant

Score = -1 Clusters are assigned in the wrong way

DEFINE CLUSTERS

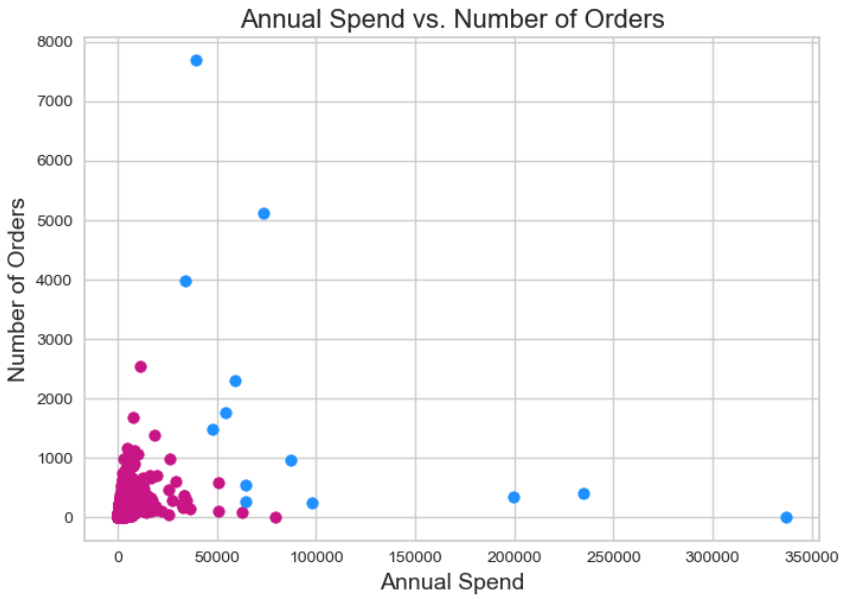
02

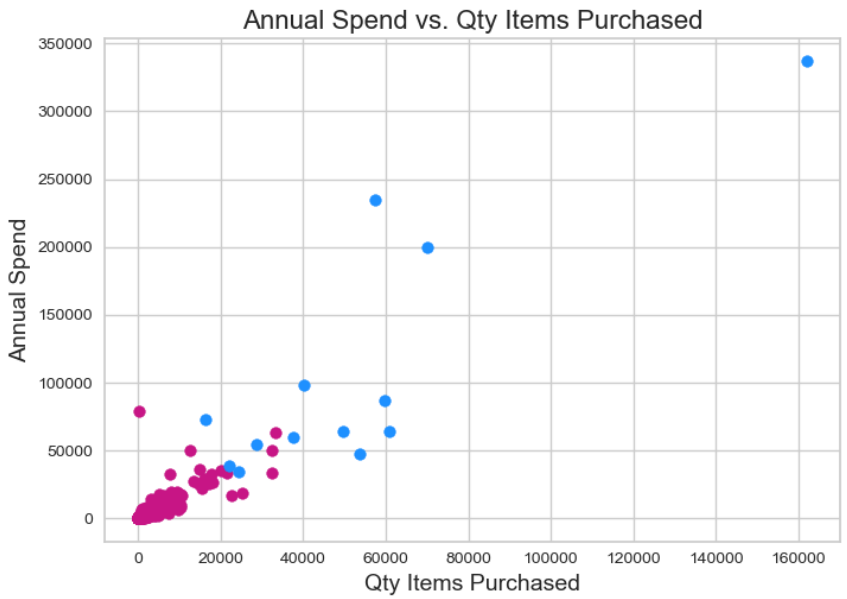
K-MEANS CLUSTER ASSIGNMENTS

Number of Customers	Cluster Assignment
3396	Cluster 0
13	Cluster 1

CLUSTERS VISUALIZED

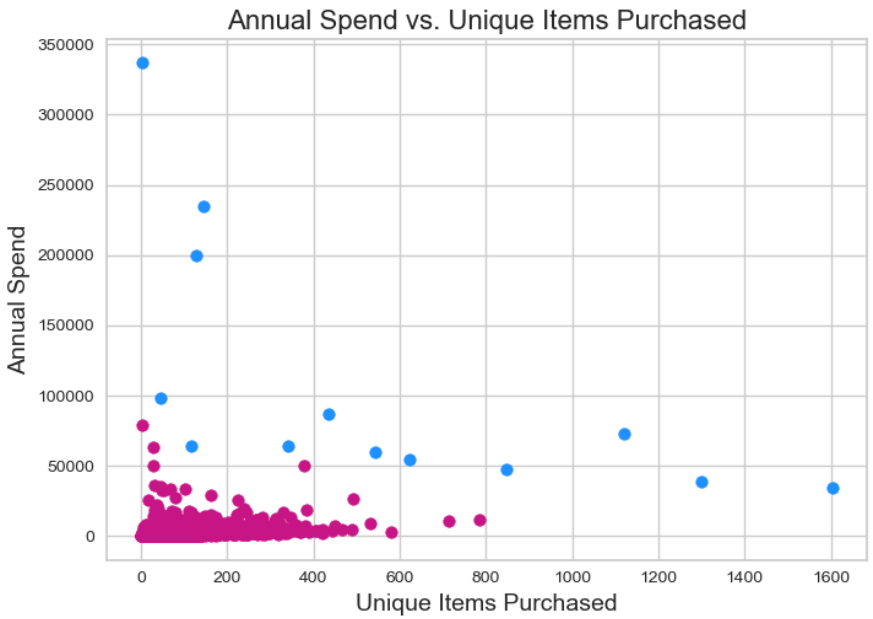
Cluster 1 is only 1% of total customers, but it greatly outperforms Cluster 0 in annual spend, ordering frequency, volume of items purchased

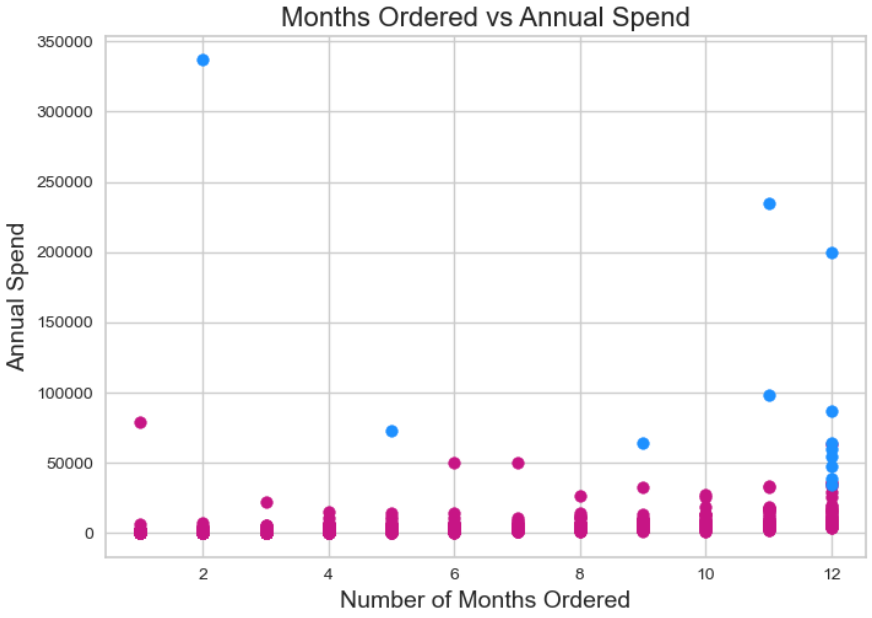




CLUSTERS VISUALIZED

Cluster 1 orders a lot more unique items and they order more frequently





NAME CLUSTERS BASED ON ANALYSIS

Name	Number of Customers	Cluster Assignment
Core Customer	3396	Cluster 0
Business Customer	13	Cluster 1

WHAT CHARACTERIZES THE CLUSTERS?

Core Customer Characteristics **Backbone of Business** **99% of Customers**	**Business Customer Characteristics** **Tiny but Mighty** **1% of Customers**
Spend • Min Annual Spend: $43 • Max Annual Spend: $7,918 Orders • 75% of customers bought a quantity of items between 13 – 996 • 50% of customers bought between 1 – 39 unique stock codes • 75% of customers had between 2 – 105 orders	Spend • Min Annual Spend: $3,408 • Max Annual Spend: $336,942 Orders • 75% of customers bought a quantity of items between 28,750 – 162,000 • 50% of customers bought between 434 – 1603 unique stock codes • 75% of customers had between 348 – 7692 orders

STRATEGIES

03

STRATEGY NEXT STEPS

Core Customer

Backbone of Business

- Understand how this group can be furthered clustered when the business customer is removed from the dataset

Business Customer

Tiny but Mighty

- Mark these people in the database as business customers and develop specific marketing communications based on their very specific buying behaviors
- To attract others, in order to expand this business segment, conduct interviews with this audience to understand why they purchase at this scale and to identify how the company can further address their needs

DISCUSSION