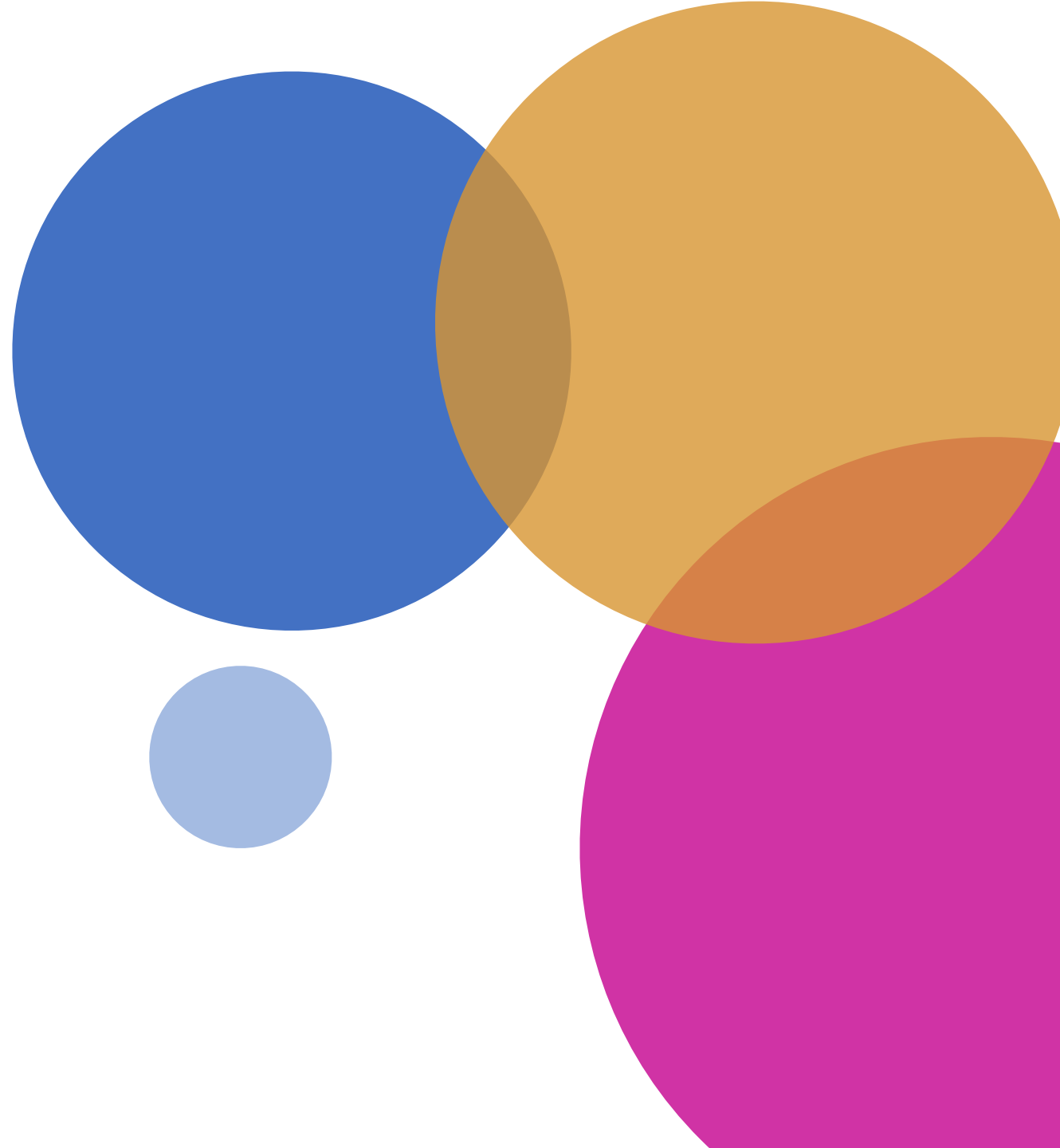


Customer Segmentation Using Data Science

by: Julianna Renaud




Project Overview

Utilize past purchase data and historical email engagement data to segment customers into distinct groups.



The Client



A ***fashion retailer*** selling a variety of men's, women's and children's clothing and accessories. This client operates both eCommerce and retail business, while also having a large percentage of product sales coming from third party retailers. This project focuses on the client's ***eCommerce business***.

**For confidentiality purposes, the name of the client will remain anonymous.*

The Data

SQL queries were utilized to pull data from the client's ESP*:

- **Account Master:** Total list of Subscriber IDs
- **Product File:** Product details such as category, sub-category, etc.
- **Purchase File:** Purchase details such as Order ID, Product ID (SKU), Subscriber ID (purchaser)
- **Subscriber Status:** Subscriber's email program status (Active, Held, Bounced)
- **Sent Count:** The total number of emails sent to the subscriber in the last six months
- **Open Count:** The total number of email opens in the last six months
- **Click Count:** The total number of email clicks in the last six months

* No Personally Identifiable Information (PII) was utilized. Subscriber ID was utilized in lieu of email address. City and state data were pulled for location of purchase, but not billing or shipping addresses. No payment data was utilized.

The Methodology

The project was treated as an *unsupervised learning classification* problem. Sklearn's *Mini Batch KMeans* and *KMeans* algorithms were utilized to build customer clusters. The *Elbow Method* was utilized to determine the appropriate number of clusters between two and ten.

Data Wrangling

- ***Date Fields:*** Required proper formatting
- ***Missing Data:*** The handling of missing data was dependent upon the column
 - Some columns were dropped
 - Some columns missing data was replaced with "0"
 - The Product Quantity column missing data was replaced with "x" (the mean was x.xx, the mode was x and the median was x (the mode and median were the same and rounding the mean down to the nearest whole number equaled the mode and median)
- ***Creation of Categorical Columns:*** Columns such as Product Category that contained text were pivoted into a column for each category. The columns were then populated with the sum of products purchased in that category for each Subscriber ID.
- ***Pivoting the Data:*** The data was pivoted to create a single row of data for each SubscriberID
- ***Formulation of Binary Email Columns:*** Email Status, Email Open Rate, and Email Click Rate were turned into binary columns

* More detailed explanations of each data wrangling step can be found in the project Code and Report document.

Data Wrangling

- ***Creation of Calculated Fields:*** Calculated fields were created to compare behavior from subscriber to subscriber:
 - Subscriber AOV
 - Subscriber UPT
 - % of items purchased by Product Category, Size Category, Product Department and Product Brand
 - % of orders purchased using a Promo Code (either shipping or product)
- ***Removed Non Purchasers:*** Since the project is focused on customer segments, removed non customers
- ***Removed Outliers:*** Based on LTV, removed outliers that were outside of the 95% (two standard deviations from the mean)
- ***Creation of Regional Columns:*** State data was consolidated into regional columns

* *More detailed explanations of each data wrangling step can be found in the project Code and Report document.*

EDA Overview

- ***Classified purchasers/non purchasers:*** Subsequently removed non purchasers from dataset
- ***Identified Outliers:*** Utilized lifetime value (LTV) to determine if any outliers existed in the dataset
- ***Removed Outliers:*** Removed subscribers with LTV greater than two standard deviations from the mean
- ***Analyzed KPIs:***
 - LTV
 - Recency
 - Frequency
 - AOV
 - UPT

EDA Overview

- ***Email Status:*** 1/0 binary column for active status (1) or not (0) in ESP
- ***Email Engagement:*** 1/0 binary columns indicating if they'd opened or clicked at least one email in the last six months
- ***Analyzed Specific Product Data:***
 - Product Size Category Purchased From
 - Product Department
 - Product Category
 - Product Brand
- ***Analyzed Order Location Data (by State of Purchase)***

EDA: Non Purchasers/Purchasers

xx.xx% of Subscriber IDs have not made a purchase

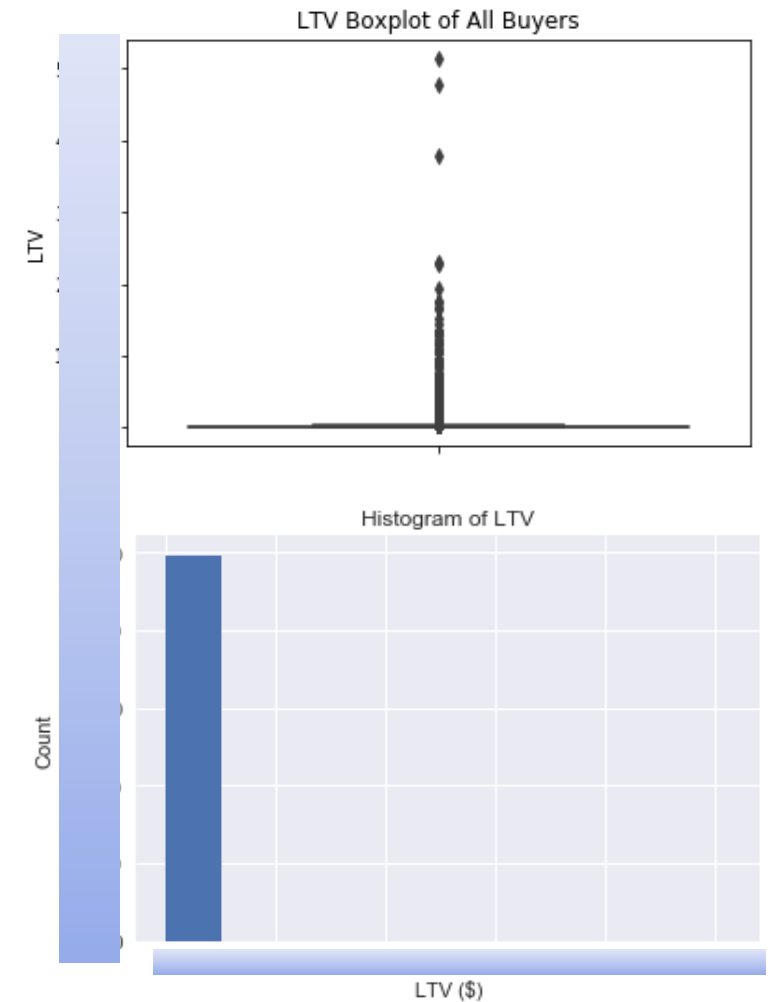
xx.xx% of Subscriber IDs have made at least one purchase

Non purchasers were removed for the final dataset utilized for clustering.



EDA: Lifetime Value (LTV)

Buyer Data Analysis (LTV) All Data	
Count	xxx,xxx
Mean	\$xxx.xx
Standard Deviation	\$xxx.xx
Min	\$x.xx
25%	\$xx.xx
50%	\$xx.xx
75%	\$xxx.xx
Max	\$xx,xxx.xx



EDA: Overview of Clustering Dataset

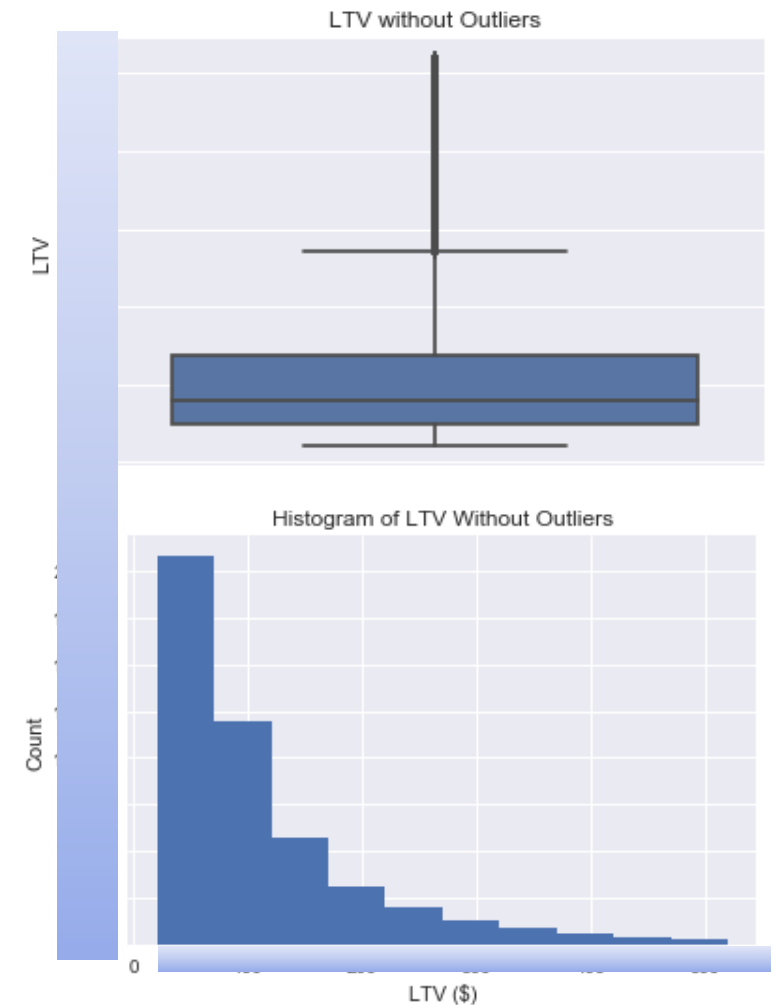
- ***Removal of Non purchasers***
- ***Removal of Outliers:*** Rejected Subscriber IDs with LTV outside of two standard deviations from the mean

* *All subsequent slides and data are representative of the new dataset created after removal of the outliers*

Metric	Totals/Averages
Total Buyers (Subscriber IDs)	xxx,xxx
Total Orders	xxx,xxx
Total Revenue	\$xx,xxx,xxx.xx
Avg. Recency (days)	xxx
Avg. Frequency	x.xx
Avg. LTV	\$xxx.xx
AOV	\$xx.xx
Total Products	x,xxx,xxx
Avg. UPT	x.xx
Avg. % of Orders Promo Used	xx.xx%

EDA: Revised Lifetime Value (LTV)

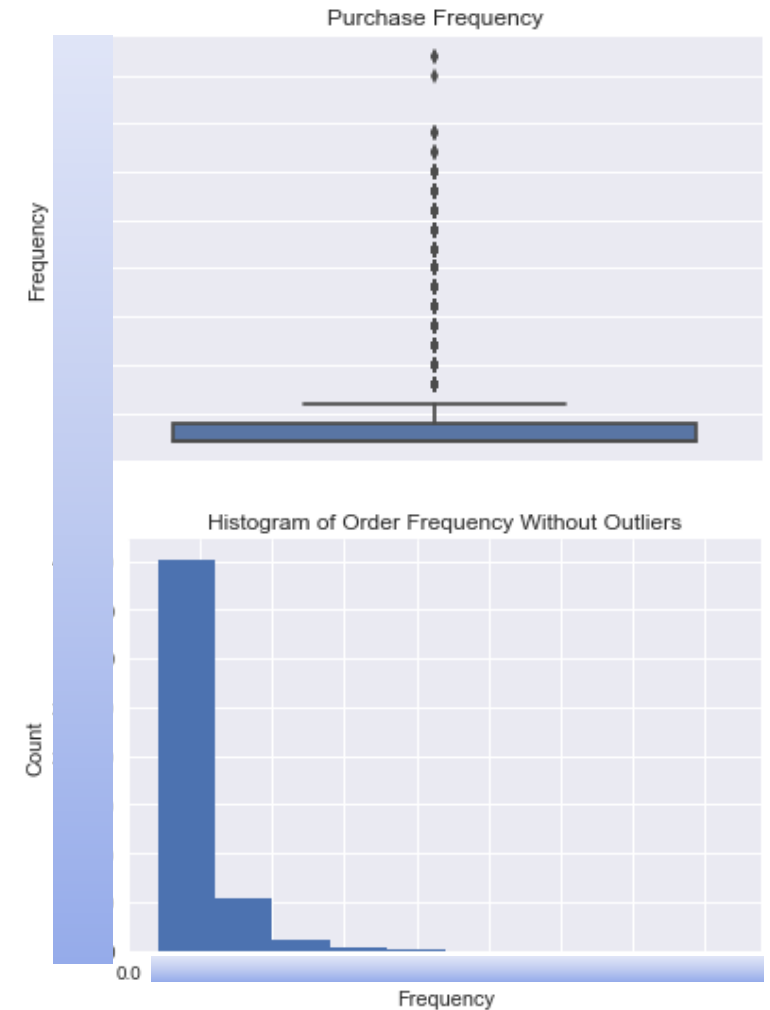
Buyer Data Analysis (LTV) Outliers Removed at 95%	
Count	xxx,xxx
Mean	\$xxx.xx
Standard Deviation	\$xx.xx
Min	\$xx.xx
25%	\$xx.xx
50%	\$xx.xx
75%	\$xxx.xx
Max	\$xxx.xx



EDA: Frequency

Buyer Data Analysis (Frequency) Outliers Removed at 95%

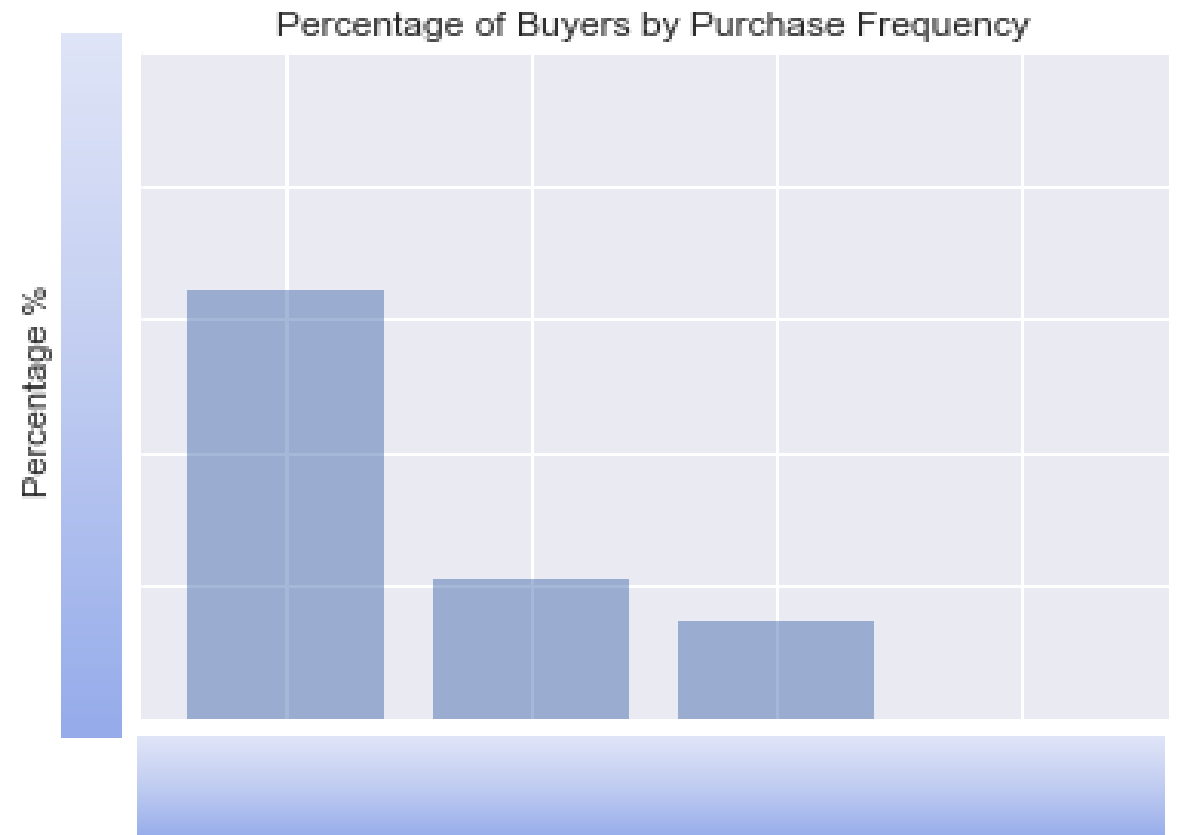
Count	xxx,xxx
Mean	x.xx
Standard Deviation	x.xx
Min	x.xx
25%	x.xx
50%	x.xx
75%	x.xx
Max	xx.xx



EDA: Frequency

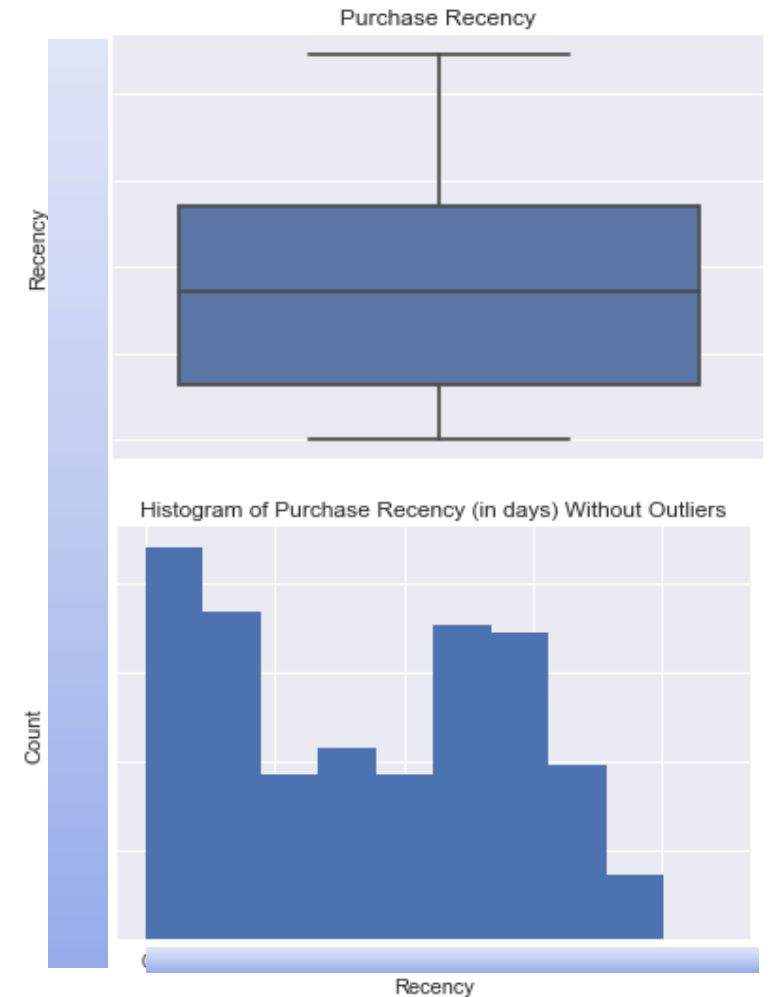
Breakdown by Frequency:

- xx.xx%: 1 x Buyer
- xx.xx%: 2 x Buyer
- xx.xx%: 3 – 9 x Buyer
- xx.xx%: 10+ x Buyer



EDA: Recency

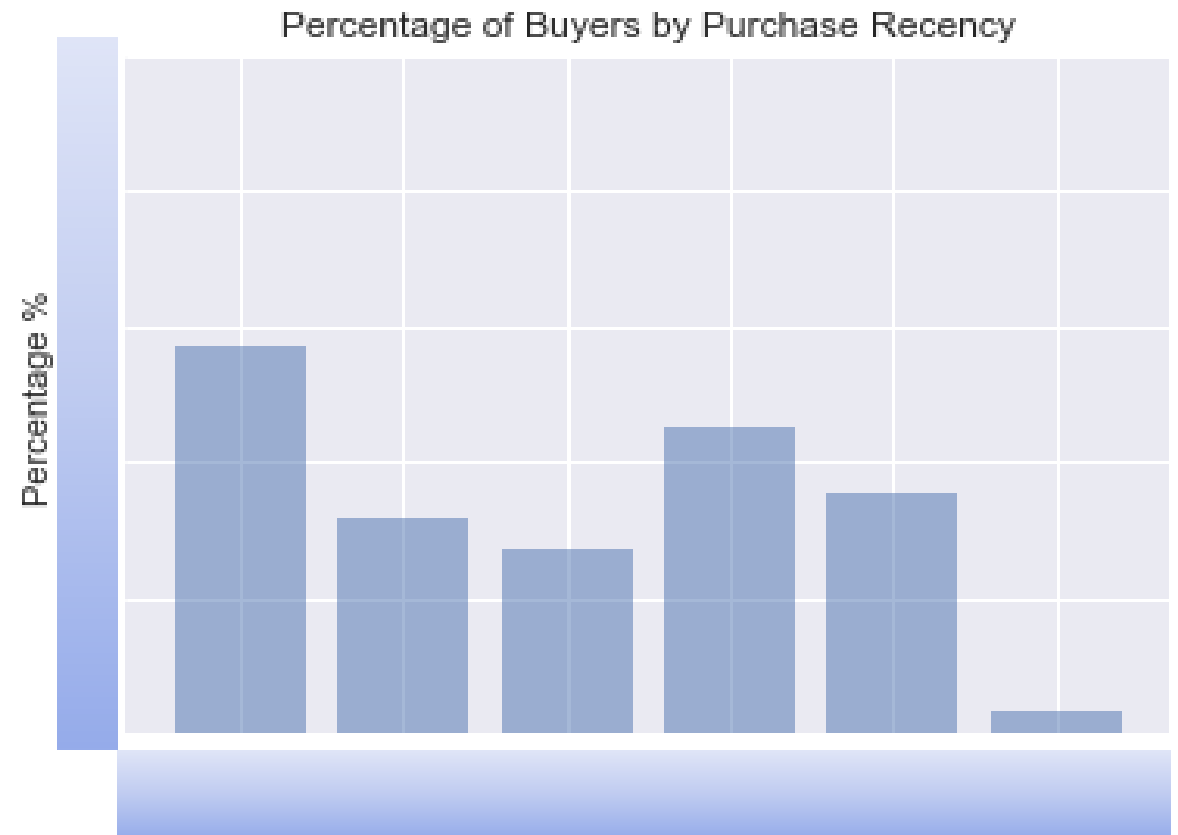
Buyer Data Analysis (Recency in Days) Outliers Removed at 95%	
Count	xxx,xxx
Mean	xxx
Standard Deviation	xxx
Min	x
25%	xxx
50%	xxx
75%	x,xxx
Max	x,xxx



EDA: Recency

Breakdown by Recency:

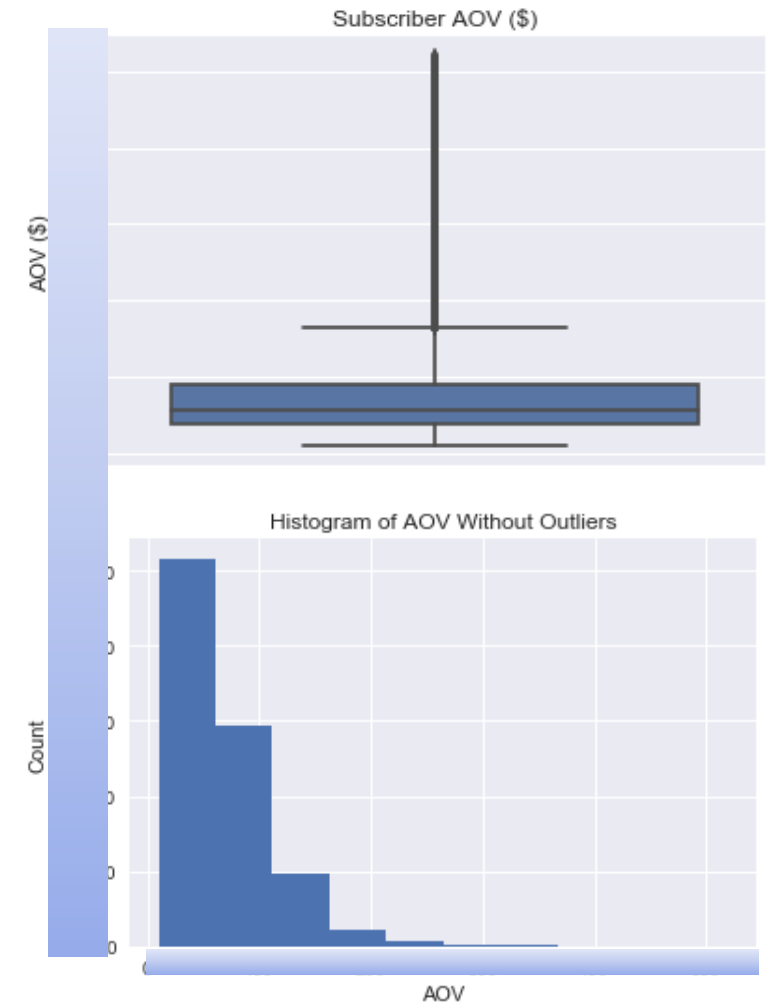
- xx.xx%: 1 year
- xx.xx%: 2 year
- xx.xx%: 3 year
- xx.xx%: 4 year
- xx.xx%: 5 year
- xx.xx%: 6+ year



EDA: AOV

Buyer Data Analysis (Recency in Days) Outliers Removed at 95%

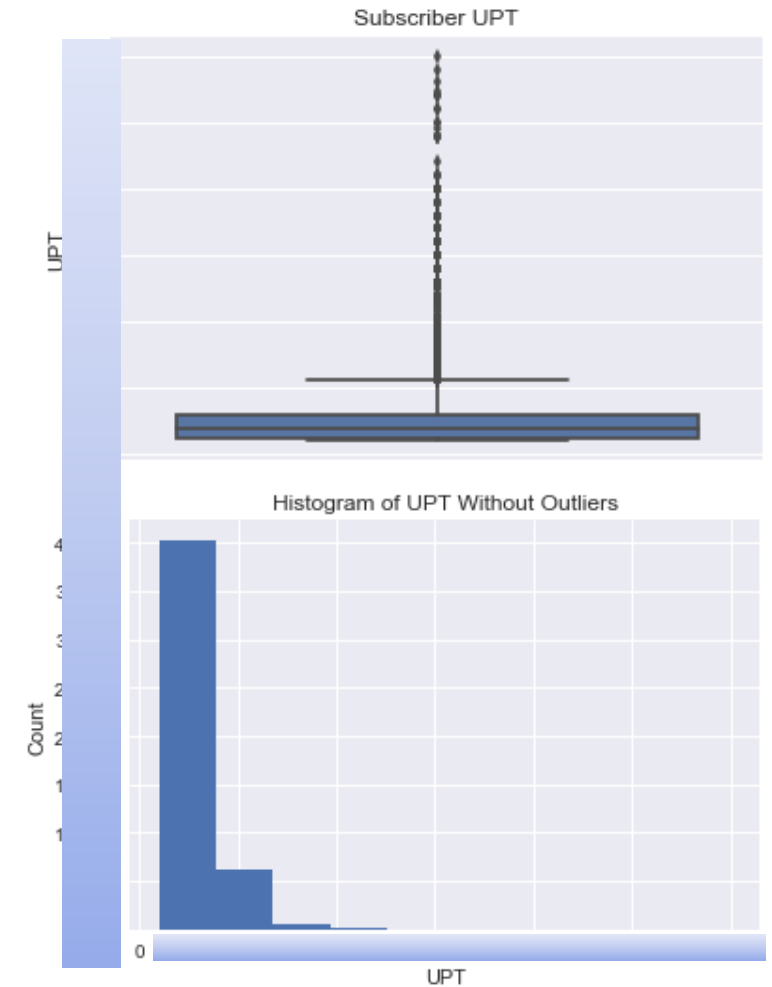
Count	xxx,xxx
Mean	\$xx.xx
Standard Deviation	\$xx.xx
Min	\$xx.xx
25%	\$xx.xx
50%	\$xx.xx
75%	\$xx.xx
Max	\$xxx.xx



EDA: UPT

Buyer Data Analysis (Recency in Days) Outliers Removed at 95%

Count	XXX,XXX
Mean	X.XX
Standard Deviation	X.XX
Min	X.XX
25%	X.XX
50%	X.XX
75%	X.XX
Max	XX.XX

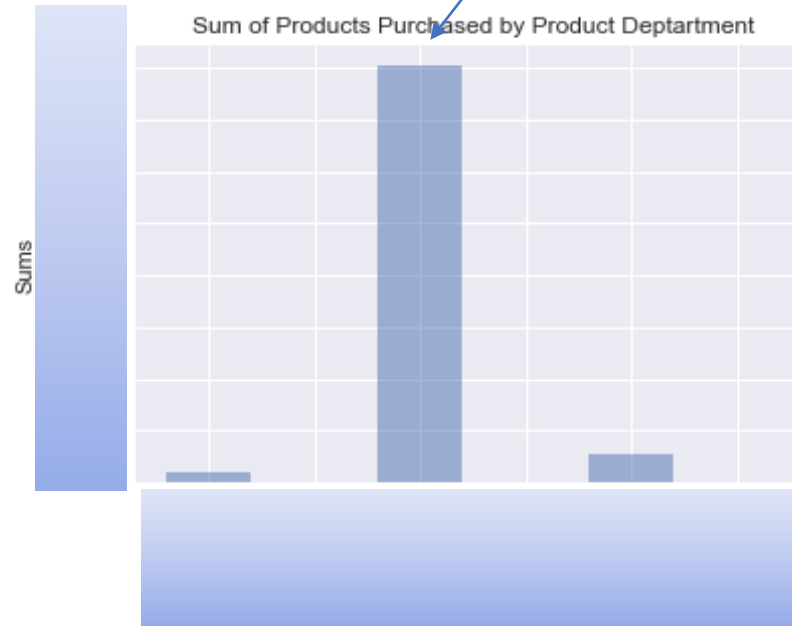


EDA: Product Data

xx.xx% of products purchased were from the **X** Size Group



xx.xx% of products purchased were from the **X** Product Department



xx.xx% of products purchased were from the **X** Category

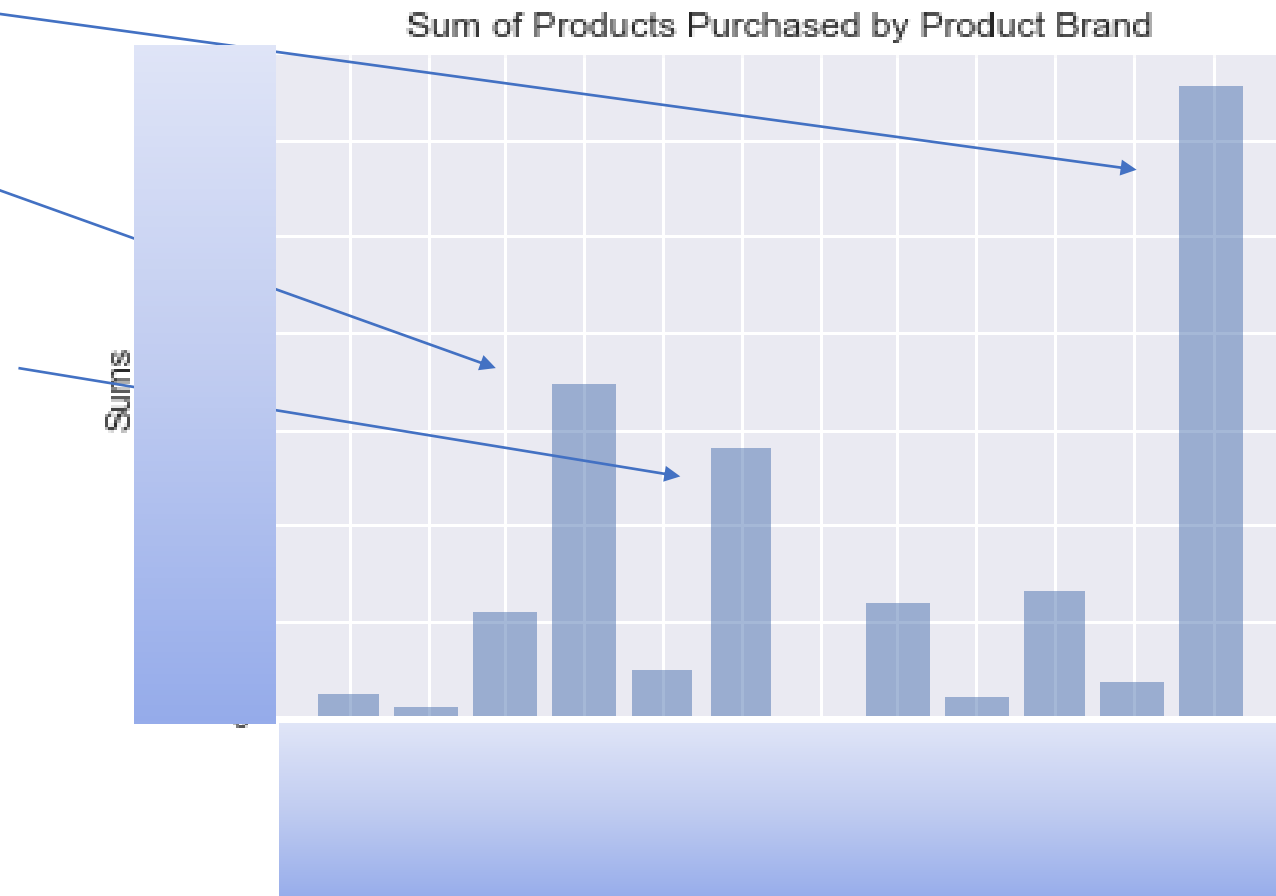


EDA: Product Data

xx.xx% of products purchased were from the **X** Brand group.

xx.xx% of products purchased were from the **Y** Brand group.

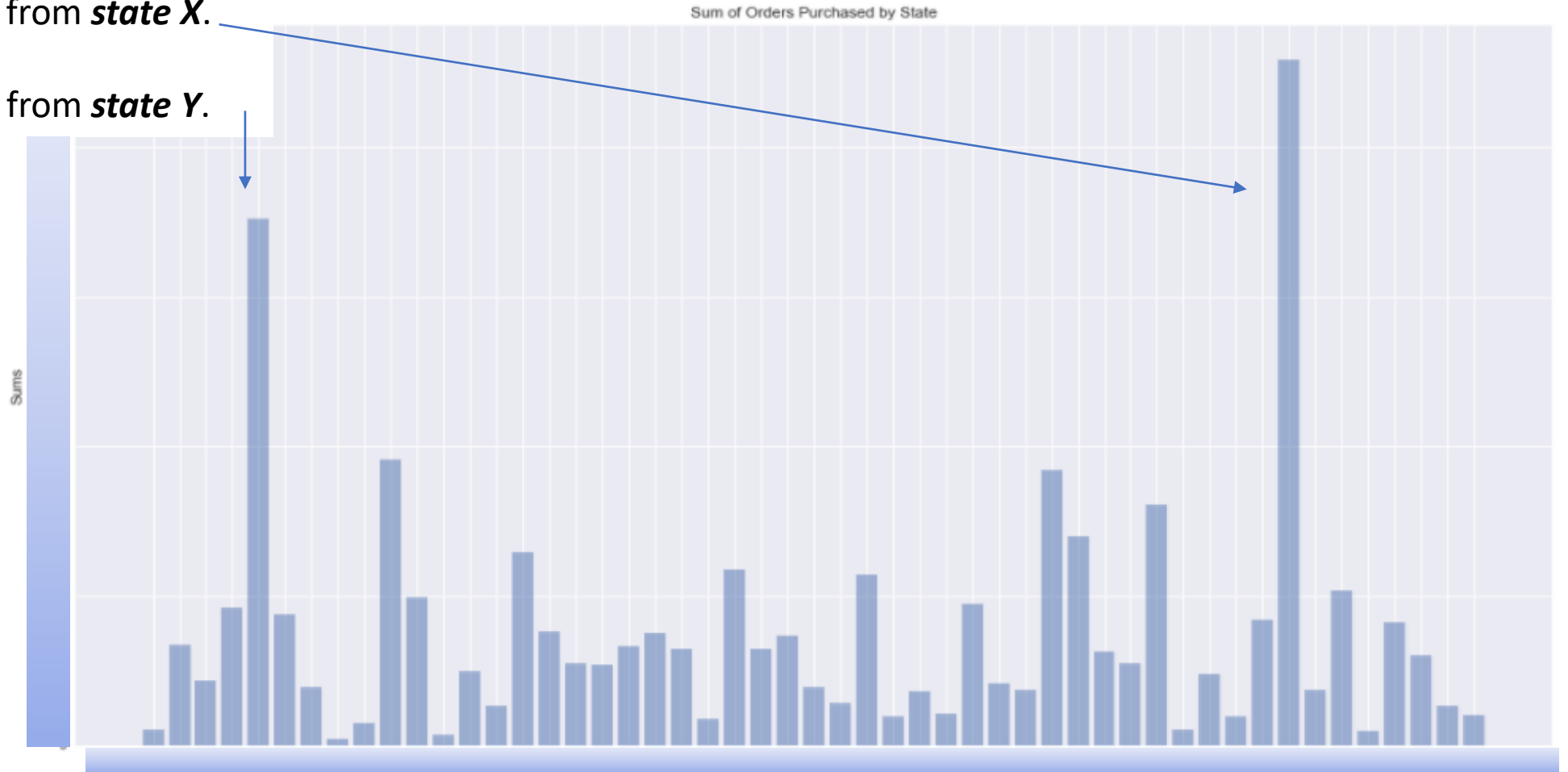
xx.xx% of products purchased were from the **Z** Brand group.



EDA: Order Location Data

xx.xx% of orders came from **state X**.

xx.xx% of orders came from **state Y**.



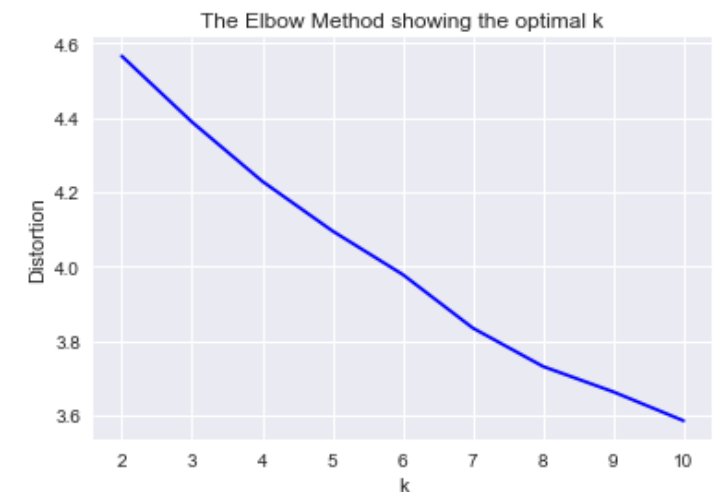
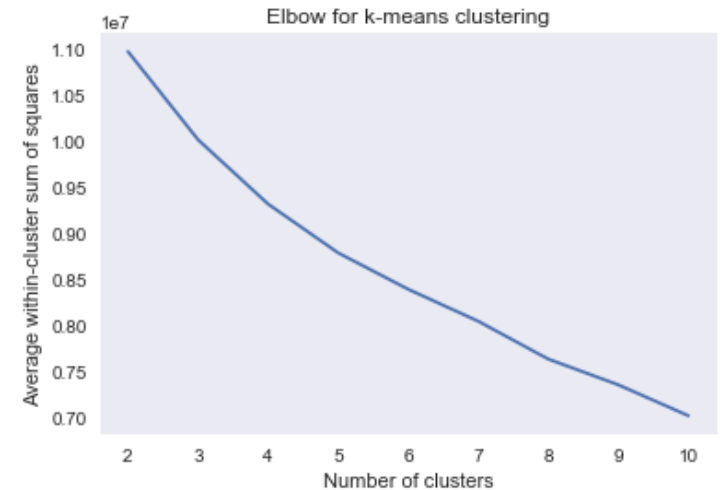
Clustering

The Elbow Method for Determining the Appropriate number of clusters.

As K increases, the centroids become closer to the cluster centroids and the improvements will begin to decline, in theory creating an elbow shape.

- Top Graph: Avg. Within Cluster Sum of Squares
- Bottom Graph: Distortion

Number of Clusters Selected = 8



Persona: Engaged Email & Promotional

(Cluster 0)

Number of Subscribers: xx,xxx (8.95% of total buyers)

- ***Most recent purchasers*** of any cluster; on average their most recent purchase date was xxx days ago
- ***Most promotionally influenced*** of any cluster; average promotional offer use by this cluster was **xx.xx%** of orders utilized a promo
- ***Most engaged with email***, nearly **100%** have opened and clicked at least one email in the last 6 months
- **xx%** of products purchased were **X Size**
- **xx%** of products purchased were from the **Department X**
- **xx%** of products purchased were **Product Category X**

Recency

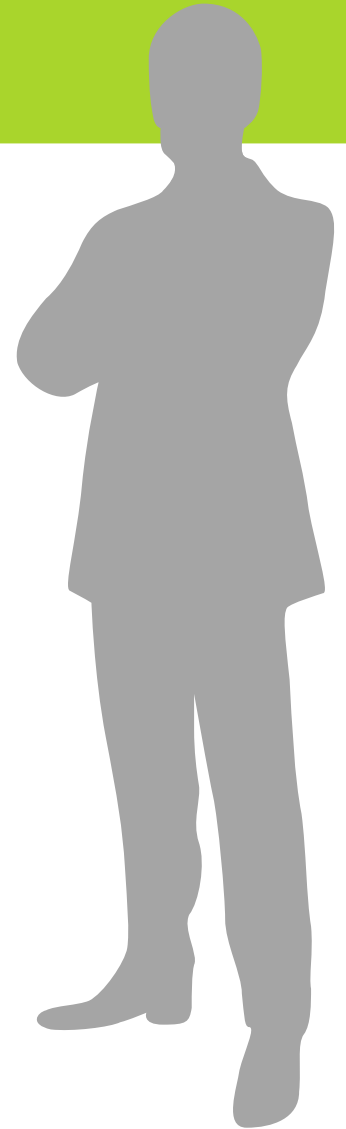
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xx.xx



Persona: “Other” Brands Buyers

(Cluster 1)

Number of Subscribers: xx,xxx (13.04% of total buyers)

- ***Second longest recency*** of any cluster, on average they made their most recent purchase xxx days ago (between 2 and 3 years ago)
- ***xx% of products purchased were from “Other Brands”***; this Brand category was the summation of products purchased in either Brand A, Brand B, Brand C, Brand D, Brand E, Brand F, Brand G or Brand H
- ***xx%*** of products purchased were **X Size**
- ***xx%*** of products purchased were from the **Department X**
- ***xx%*** of products purchased were **Product Category X**

Recency

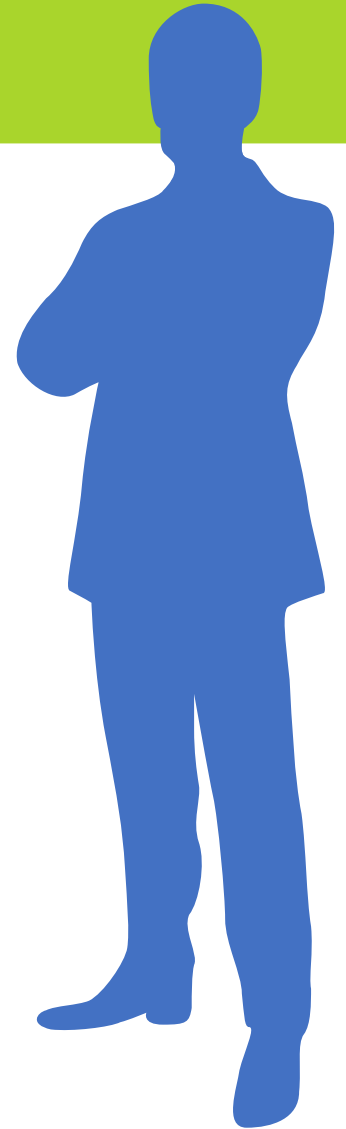
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xx.xx



Persona: Unengaged & Lapsed

(Cluster 2)

Number of Subscribers: xxx,xxx (21.54% of total buyers)

- ***Least recent purchasers*** of any cluster, on average their most recent purchase date was x,xxx days ago
- ***The largest cluster*** with xxx,xxx subscriber IDs
- ***Least engaged with email***, only xx% have opened an email in the last 6 months and less than xx% have clicked
- xx% of products purchased were **X Size**
- xx% of products purchased were from the **Department X**
- xx% of products purchased were **Product Category X**

Recency

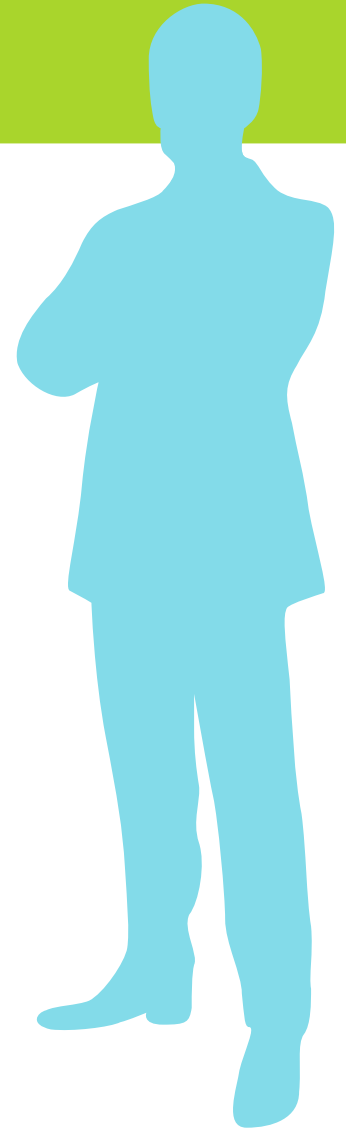
Average: x,xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xxx.xx



Persona: Frequent Buyer & Big Spender

(Cluster 3)

Number of Subscribers: xx,xxx (10.04% of total buyers)

- **The highest frequency** of any cluster, on average they've made x.xx purchases in their lifetime
- **The highest AOV** of any cluster, on average a subscriber in this cluster spends \$xxx.xx per order
- **The highest UPT** of any cluster, on average they purchase x.xx products/order
- **The least promotionally influenced**; average promotional offer use by this cluster of xx.xx% was the lowest of any cluster
- Purchasing products from Department X, but not necessarily Size X (meaning they likely shoppers of Department X, Size category Y)
- xx% of products purchased were **Size X**
- xx% of products purchased were from the **Department X**
- xx% of products purchased were **Category X**

Recency

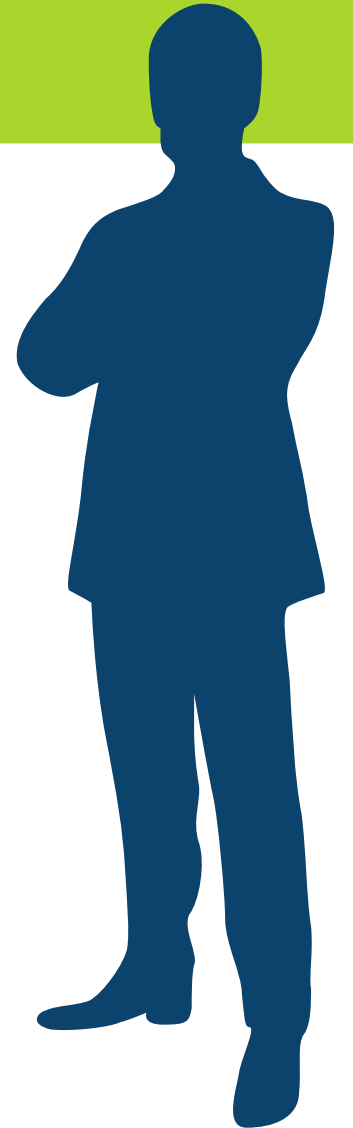
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xxx.xx



Persona: Low Spenders & Brand Reg

(Cluster 4)

Number of Subscribers: xx,xxx (16.58% of total buyers)

- ***The lowest LTV*** of any cluster, on average they've spent \$xx.xx in their lifetime
- **xx% of products purchased were from Brand X**
- **xx% of products purchased were unbranded products** (those that did not contain brand details in the product file); this was the *highest of any cluster*
- **xx% of products purchased were Size X**
- **xx% of products purchased were from the Department X**
- **xx% of products purchased were Category X**

Recency

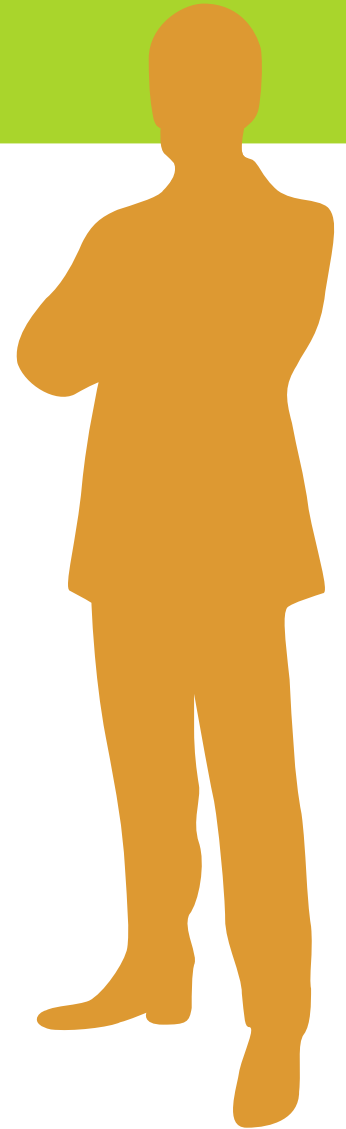
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xx.xx



Persona: Not Buying Men's

(Cluster 5)

Number of Subscribers: xx,xxx (10.23% of total buyers)

- ***These shoppers ARE NOT purchasing Products from Department X***
- **xx% of products purchased were from Brand W**
- **x.xx% of products purchased were Size X**
- **x.xx% of products purchased were from the Department X**
- **xx% of products purchased were Category X**

Recency

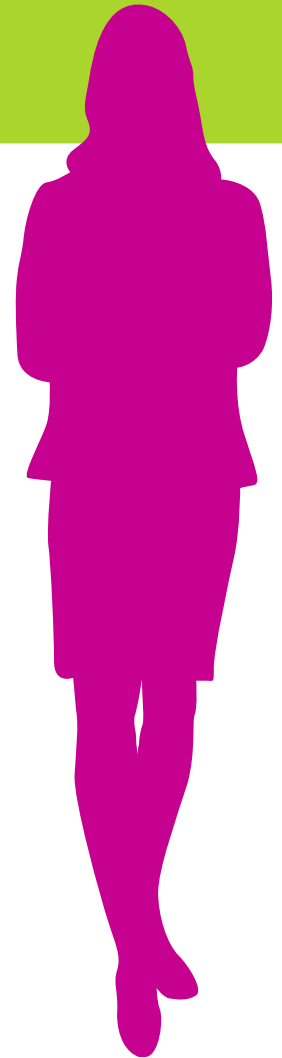
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xxx.xx



Persona: Brand FS Buyers

(Cluster 6)

Number of Subscribers: xx,xxx (10.71% of total buyers)

- ***The lowest LTV*** of any cluster, on average they've spent \$xx.xx in their lifetime
- ***The lowest AOV*** of any cluster, on average a subscriber in this cluster spends \$xx.xx per order, but have the second highest UPT, indicating that they are either purchasing sale items or low price point items
- ***The second highest UPT*** of any cluster, on average they purchase x.xx products/order
- **xx%** of products purchased by this group are from the **Brand FS**
- **xx%** of products purchased were **Size X**
- **xx%** of products purchased were from **Department X**
- **xx%** of products purchased were **Category X**

Recency

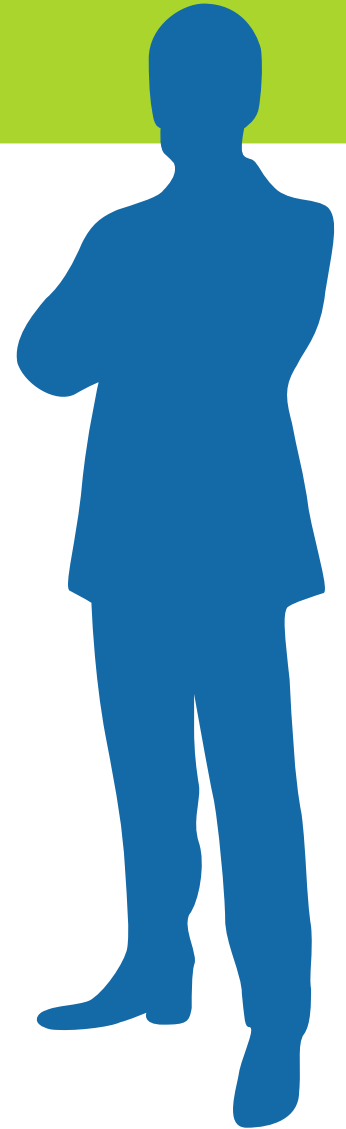
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xx.xx



Persona: Not Category X Shoppers

(Cluster 7)

Number of Subscribers: xx,xxx (8.91% of total buyers)

- The **smallest** of all of the clusters with xx,xxx subscribers (although close in size to cluster 0)
- Purchasing products from Department X, but **NOT** from Category X.
- They are the only segment ***not likely to purchase from Category X***; xx% of the products purchased from this segment are from categories other than Category X etc.
- ***The lowest frequency*** of any cluster, on average a subscriber in this cluster has purchased x.xx times in their lifetime with the brand
- **xx%** of products purchased were **Size X**
- **xx%** of products purchased were from the **Department X**
- **x%** of products purchased were from **Category X**

Recency

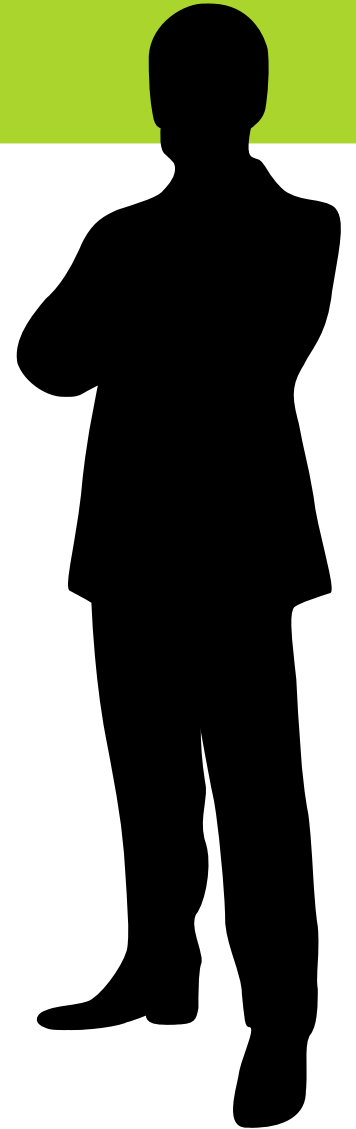
Average: xxx Days Ago

Frequency

Average: x.xx Purchases

LTV

Average: \$xx.xx



Next Steps/Client Recommendations

Leverage customer segments to encourage desired behavior from current customers as well as to reach new, similar customers:

- Look-Alike Campaigns
- Align Creative with Customer Interests
- Test Promotional Offers
- Content Placement
- Target Lapsed Customers
- Retention Strategies
- Survey Customers
- Test Marketing Channels
- Introduction of New Products
- Promote Products Based on Price Point



THANK YOU



Julianna Renaud

Web Analytics Strategist



Charlotte Werger, PhD

Springboard Mentor

Appendix



Average Purchase Behavior by Cluster

Cluster #	# of Subscribers	Recency (days)	Frequency	Monetary Value (LTV)	Products Purchased Lifetime	Subscriber AOV	Subscriber UPT	% of Orders Promo Utilized
0								
1								
2								
3								
4								
5								
6								
7								

Average Product Behavior by Cluster

(Average Percentages of Products Purchased by Size Category (X vs. Y), Product Department (X vs. Y) & Category (X vs.Y))

Cluster #	# of Subscribers	Size Cat. X	Size Cat. Y	PD X	PD Y	Category X	Category Y
0							
1							
2							
3							
4							
5							
6							
7							

Average Product Behavior by Cluster

(Average Percentages of Products Purchased by Brand)

Cluster #	# of Subscribers	Brand W	Brand Reg	Brand FS	No Brand	Other Brands
0						
1						
2						
3						
4						
5						
6						
7						

Average Email Behavior by Cluster

(Email Open and Email Click Percentages are based on the Number of Subscribers with an Active Status)

Cluster #	# of Subscribers	# of Subscribers with Active Status	% Opened Email in Last 6 Months	% Clicked an Email in Last 6 Months
0				
1				
2				
3				
4				
5				
6				
7				

Percent of Orders by Region

Cluster #	Northeast	Southeast	Midwest	Southwest	West
0	20.30%	24.99%	22.59%	14.88%	17.24%
1	21.03%	25.35%	25.20%	12.65%	15.77%
2	10.90%	25.23%	18.59%	24.61%	20.68%
3	17.88%	23.05%	21.36%	18.78%	18.94%
4	21.13%	25.79%	24.23%	11.69%	17.15%
5	13.02%	21.64%	21.09%	17.71%	26.54%
6	26.54%	22.97%	21.84%	11.13%	17.51%
7	15.79%	24.39%	19.95%	18.74%	21.13%
Unclustered	17.96%	24.19%	21.73%	16.94%	19.17%

