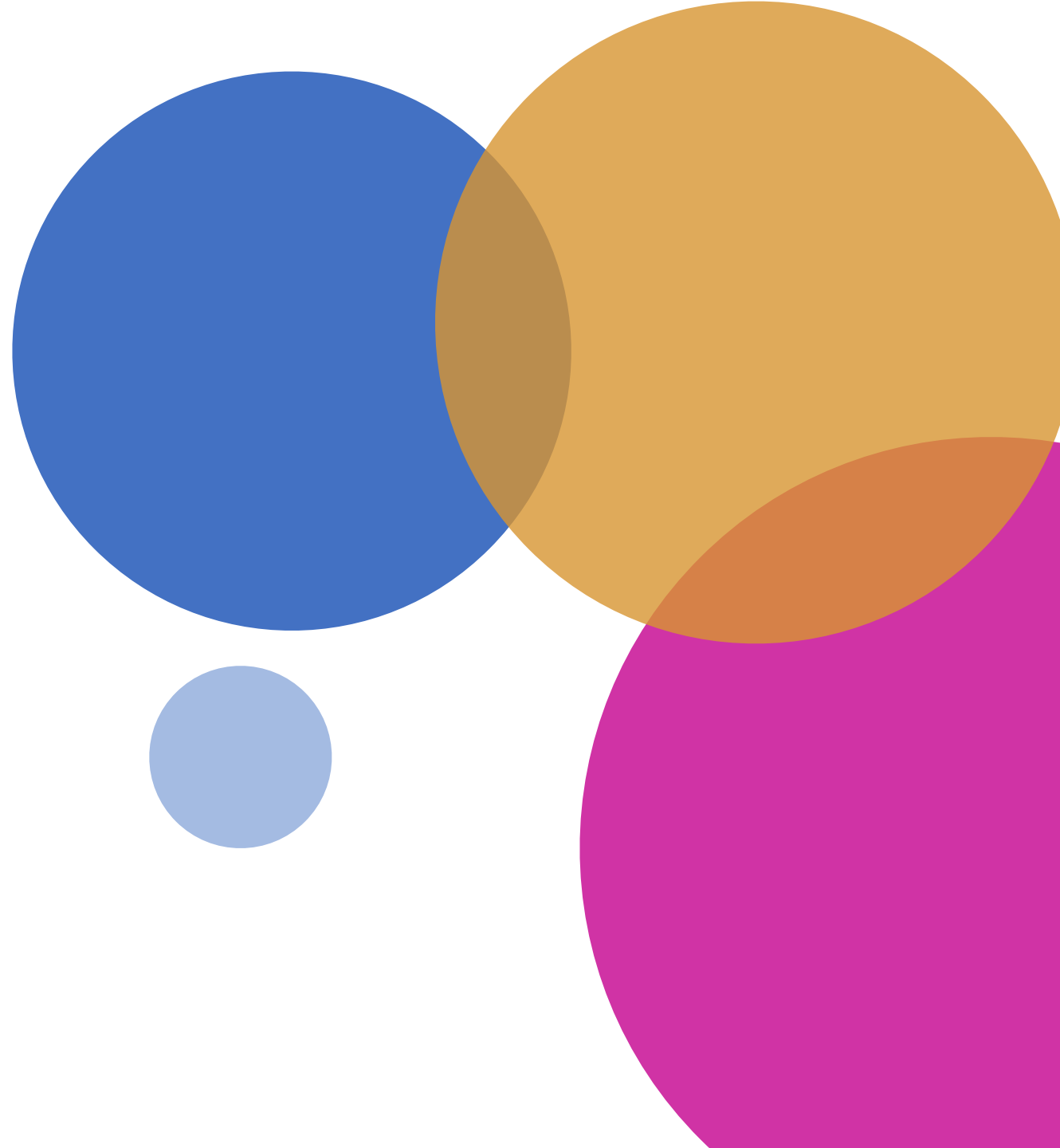


Customer Segmentation Using Data Science

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 “1” (the mean was 1.48, the mode was 1 and the median was 1)
- ***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

54.64% of Subscriber IDs have not made a purchase

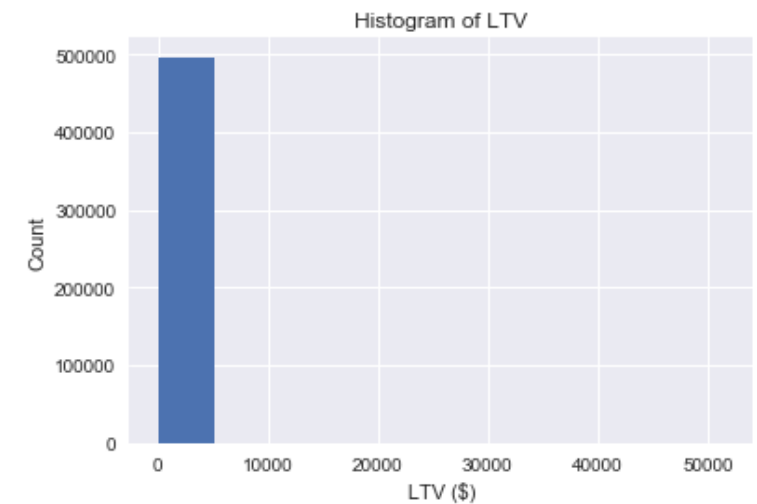
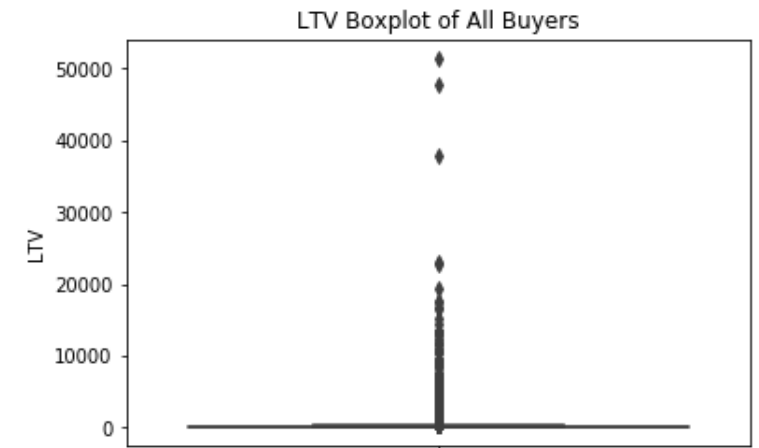
45.36% of Subscriber IDs have made at least one purchase

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



EDA: Lifetime Value (LTV)

Buyer Data Analysis (LTV) All Data	
Count	496,535
Mean	\$127.99
Standard Deviation	\$230.49
Min	\$0.97
25%	\$44.97
50%	\$78.98
75%	\$143.84
Max	\$51,496.07



EDA: Overview of Clustering Dataset

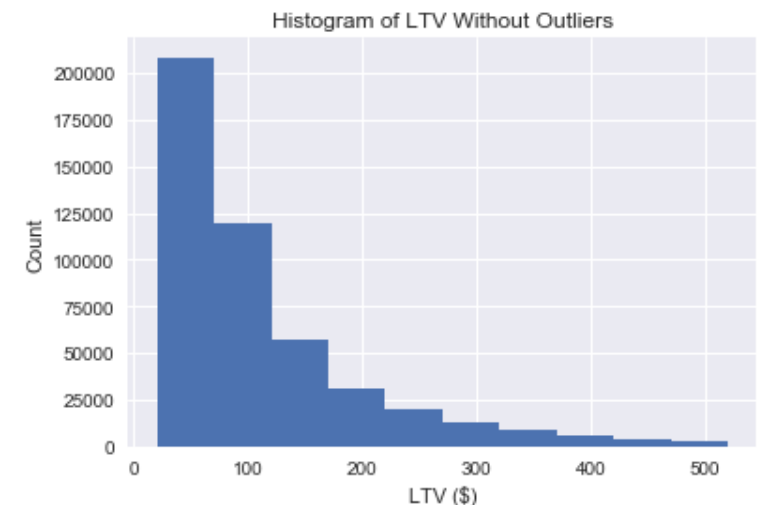
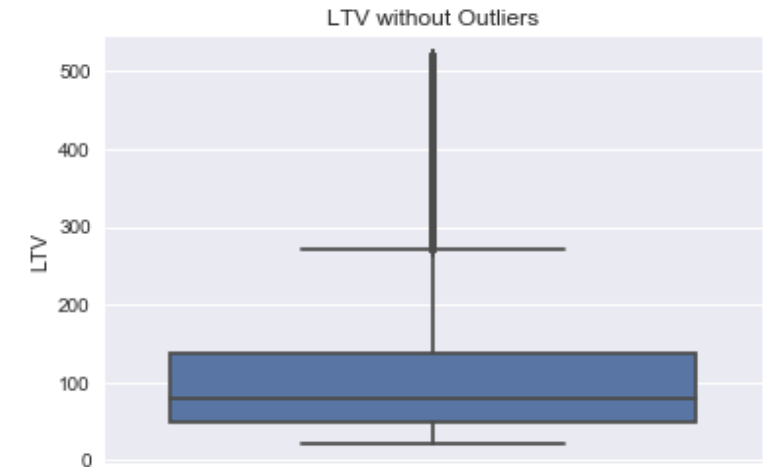
- ***Removal of Non purchasers***
- ***Removal of Outliers:*** Rejected Subscriber IDs with LTV outside of two standard deviations from the mean

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

Metric	Totals/Averages
Total Buyers (Subscriber IDs)	471,377
Total Orders	772,062
Total Revenue	\$52,242,809.05
Avg. Recency (days)	853
Avg. Frequency	1.64
Avg. LTV	\$110.83
AOV	\$67.67
Total Products	1,771,703
Avg. UPT	2.29
Avg. % of Orders Promo Used	30.26%

EDA: Revised Lifetime Value (LTV)

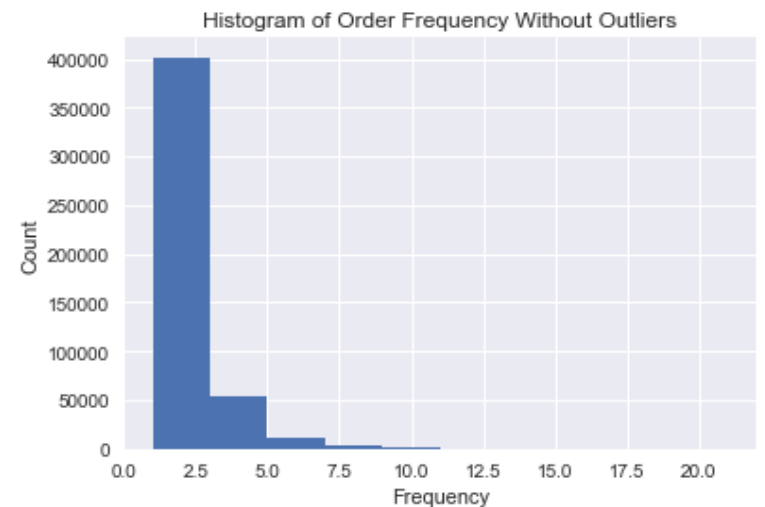
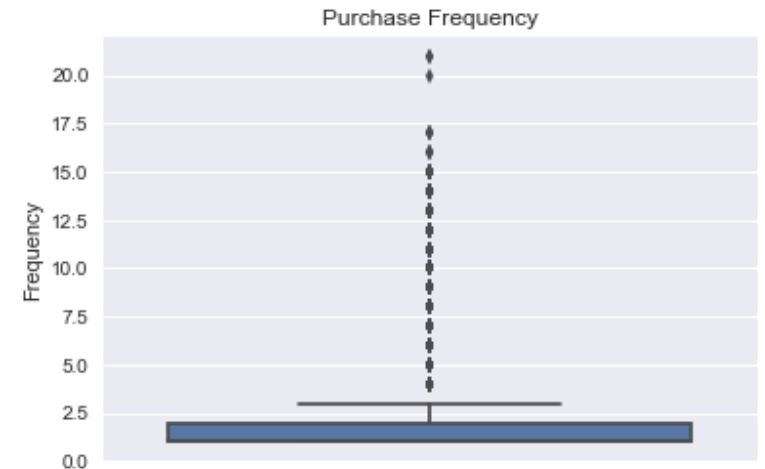
Buyer Data Analysis (LTV) Outliers Removed at 95%	
Count	471,377
Mean	\$110.83
Standard Deviation	\$91.73
Min	\$19.98
25%	\$47.94
50%	\$79.00
75%	\$136.96
Max	\$519.64



EDA: Frequency

Buyer Data Analysis (Frequency) Outliers Removed at 95%

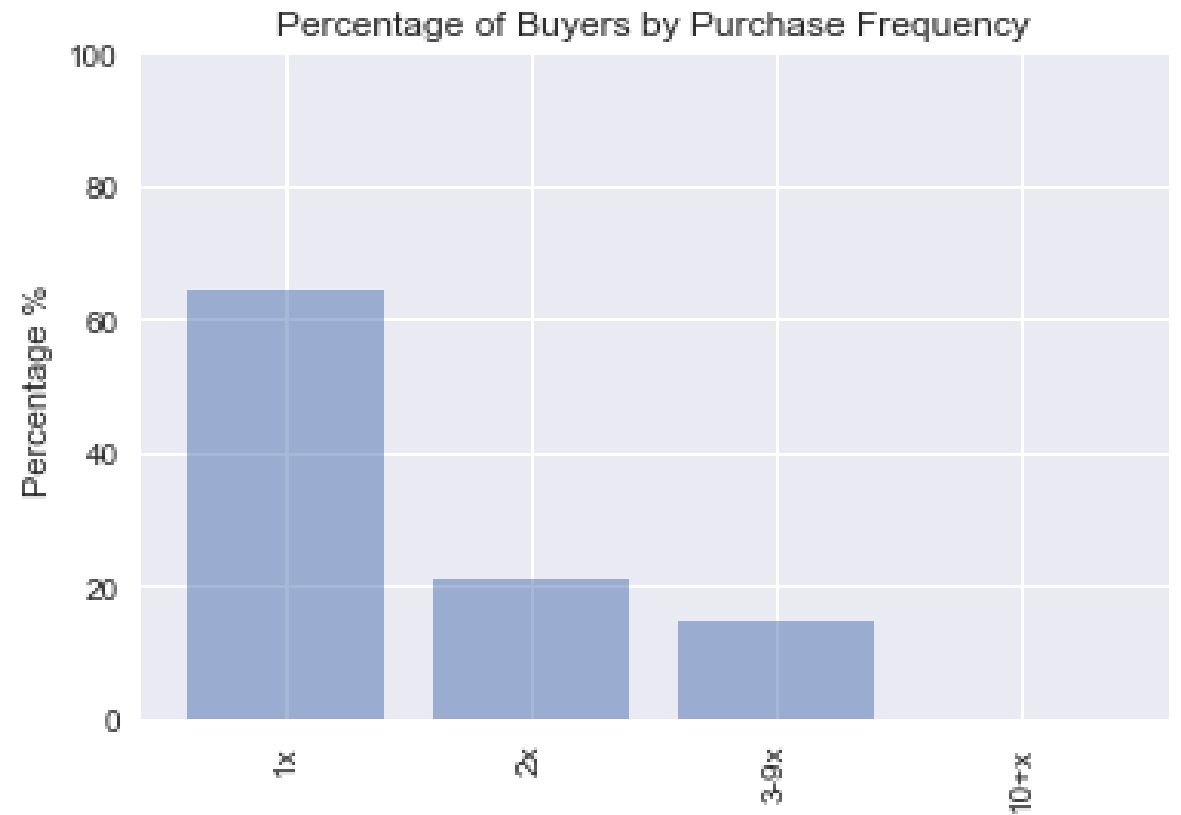
Count	471,377
Mean	1.63
Standard Deviation	1.14
Min	1.00
25%	1.00
50%	1.00
75%	2.00
Max	21.00



EDA: Frequency

Breakdown by Frequency:

- 61.16%: 1 x Buyer
- 21.03%: 2 x Buyer
- 14.72%: 3 – 9 x Buyer
- 0.09%: 10+ x Buyer



EDA: Recency

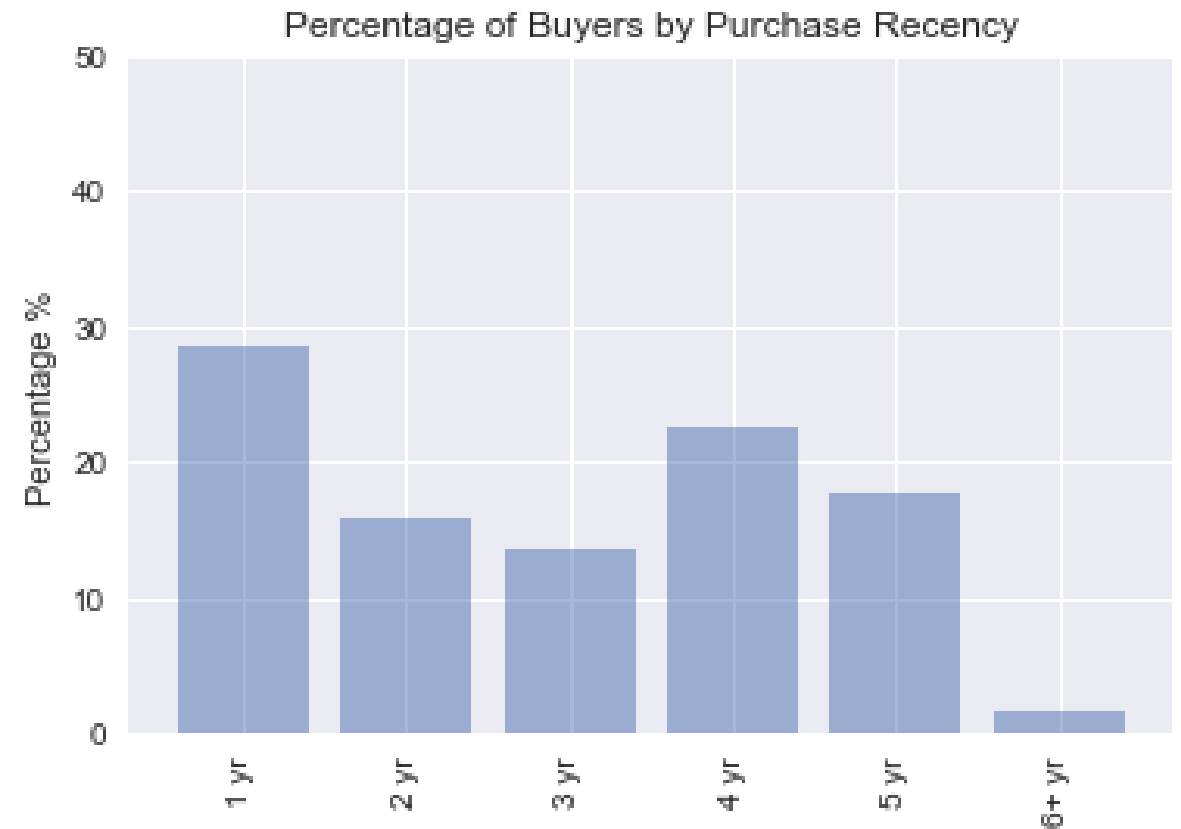
Buyer Data Analysis (Recency in Days) Outliers Removed at 95%	
Count	471,377
Mean	853
Standard Deviation	562
Min	0
25%	317
50%	851
75%	1,355
Max	2,227



EDA: Recency

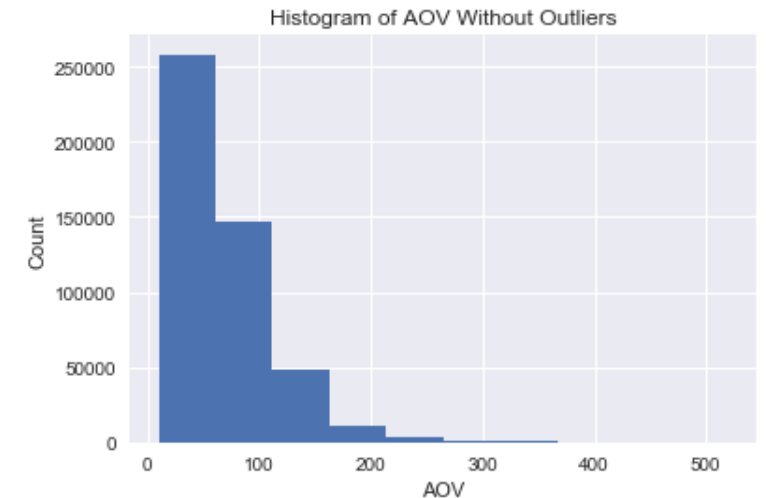
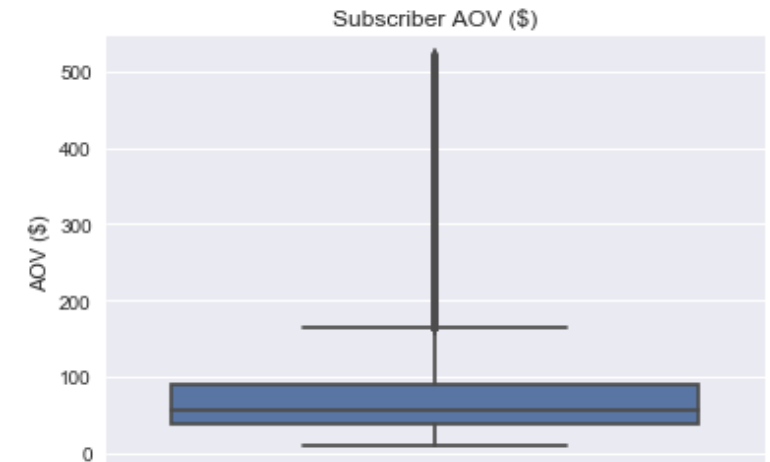
Breakdown by Recency:

- 28.47%: 1 year
- 15.95%: 2 year
- 13.71%: 3 year
- 22.62%: 4 year
- 17.70%: 5 year
- 1.56%: 6+ year



EDA: AOV

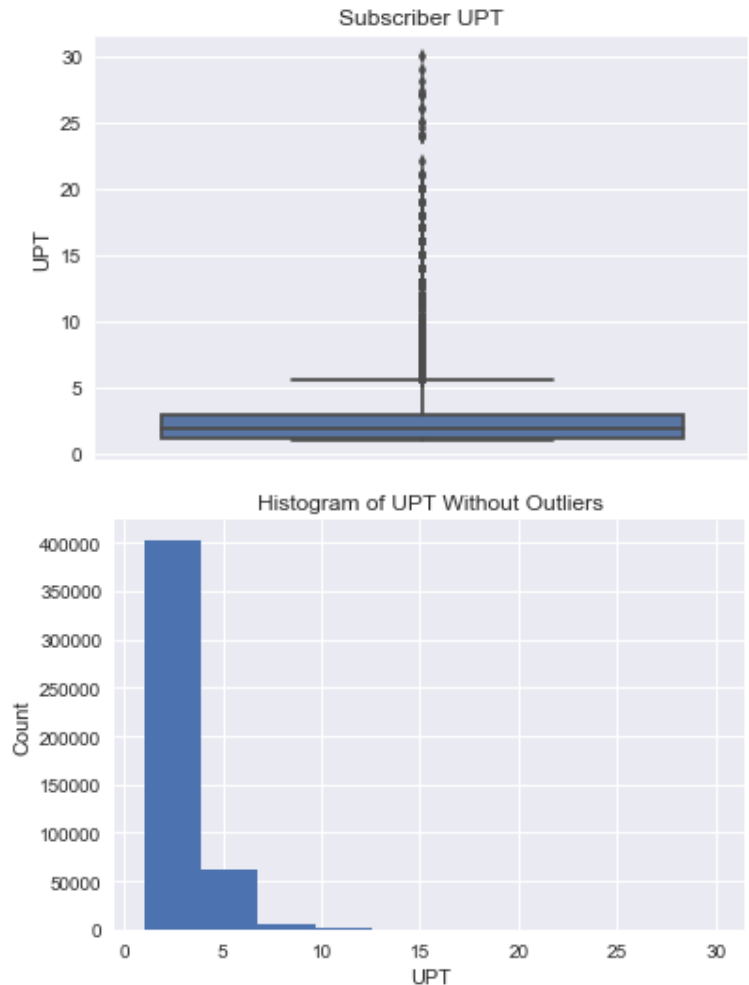
Buyer Data Analysis (Recency in Days) Outliers Removed at 95%	
Count	471,377
Mean	\$69.96
Standard Deviation	\$44.29
Min	\$9.99
25%	\$39.16
50%	\$56.91
75%	\$89.00
Max	\$519.54



EDA: UPT

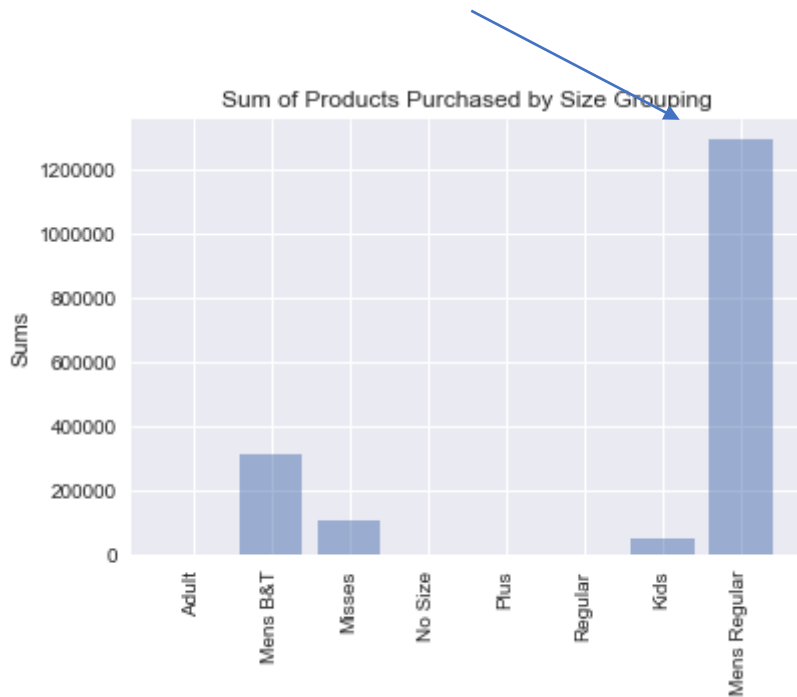
Buyer Data Analysis (Recency in Days) Outliers Removed at 95%

Count	471,377
Mean	2.34
Standard Deviation	1.39
Min	1.00
25%	1.25
50%	2.00
75%	3.00
Max	30.00

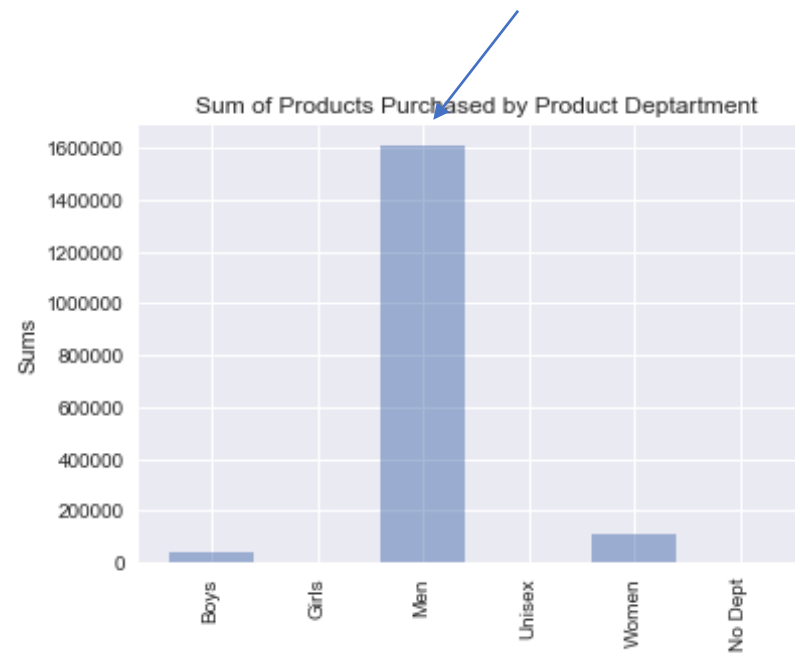


EDA: Product Data

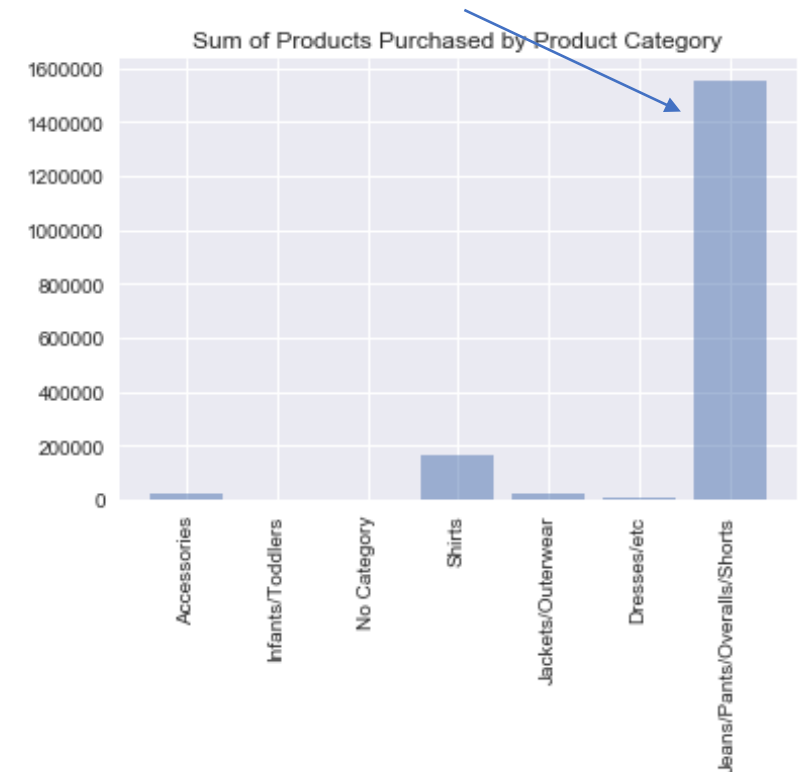
72.82% of products purchased were from the **Men's Regular** Size Group



90.80% of products purchased were from the **Men's** Product Department



87.76% of products purchased were from the **Jeans/Pants/Overalls/Shorts** Category

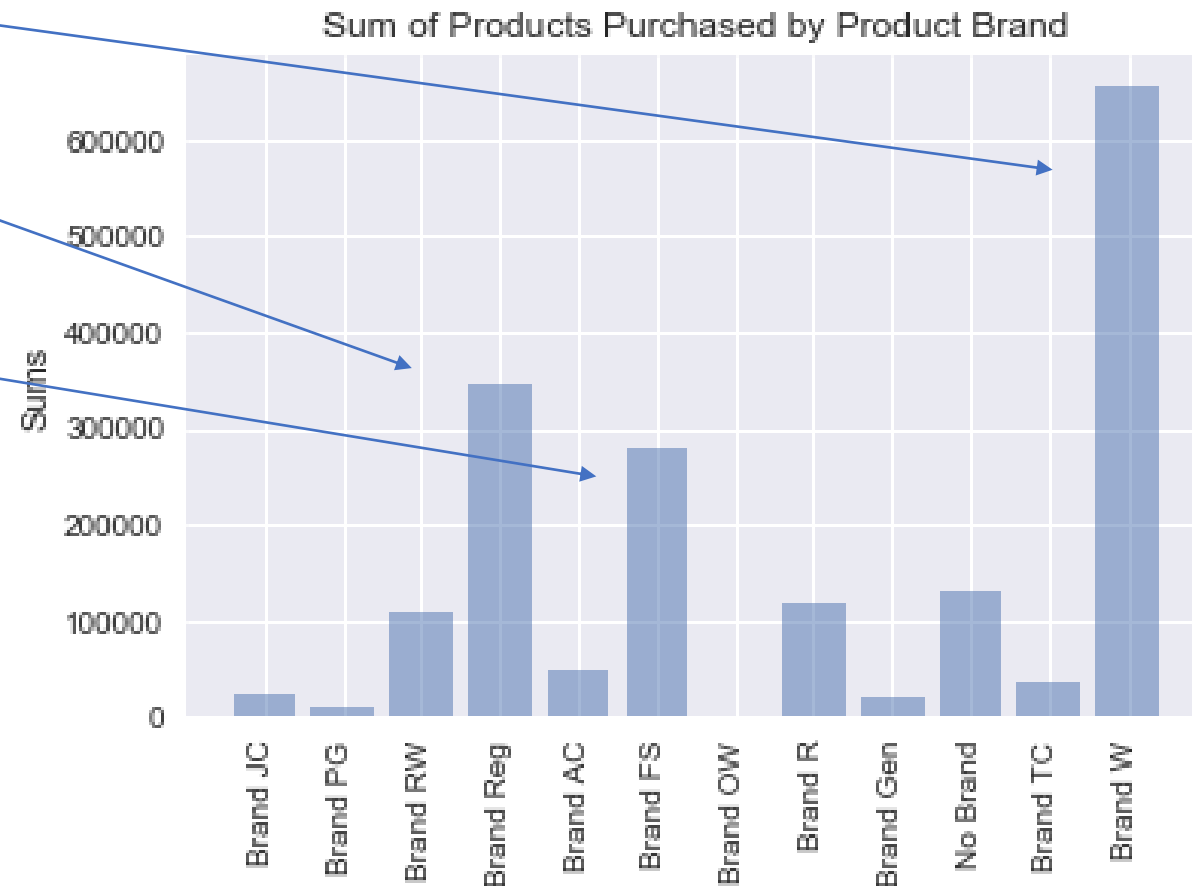


EDA: Product Data

37.03% of products purchased were from the **W** Brand group.

19.52% of products purchased were from the **Reg** Brand group.

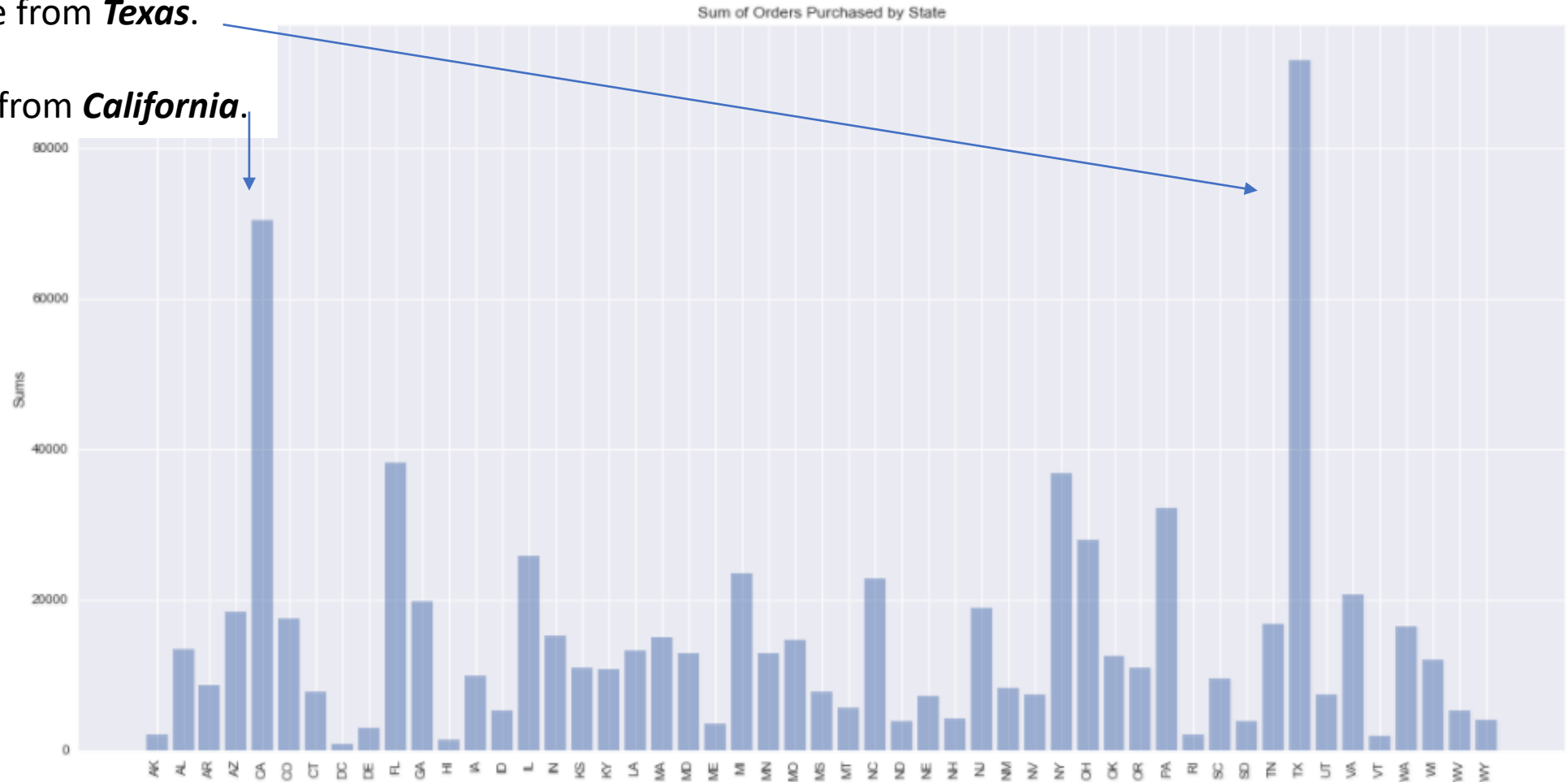
15.72% of products purchased were from the **FS** Brand group.



EDA: Order Location Data

11.87% of orders came from **Texas**.

9.12% of orders came from **California**.



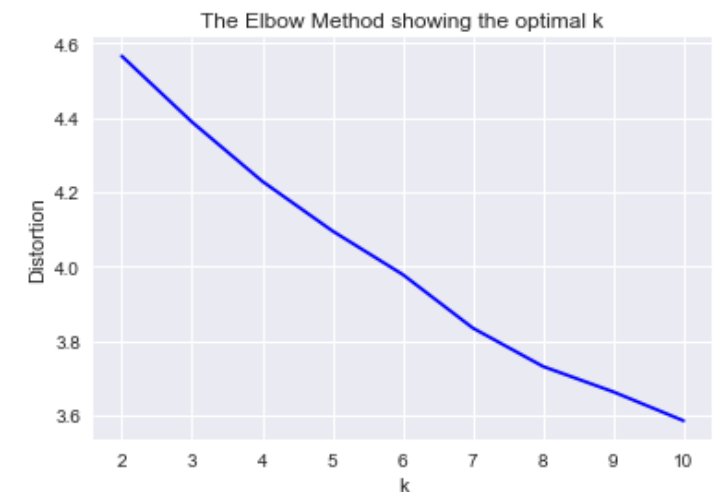
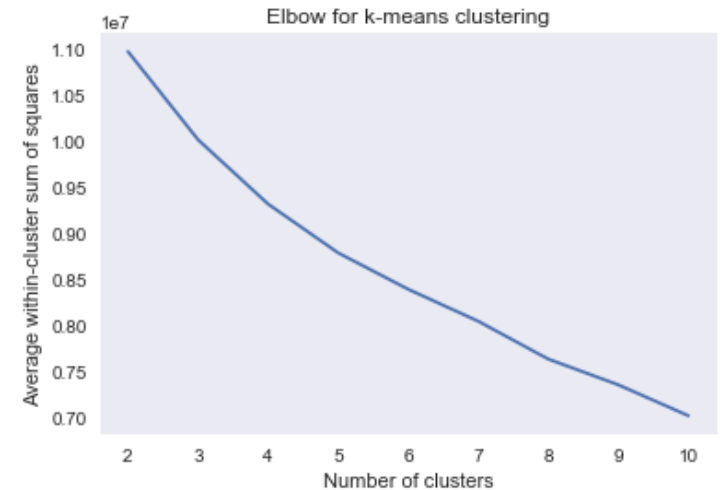
Clustering

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: 42,185 (8.95% of total buyers)

- ***Most recent purchasers*** of any cluster; on average their most recent purchase date was 369 days ago
- ***Most promotionally influenced*** of any cluster; average promotional offer use by this cluster was **49.84%** of orders utilized a promo
- ***Most engaged with email***, nearly **100%** have opened and clicked at least one email in the last 6 months
- **80%** of products purchased were **Men's Regular Size**
- **99%** of products purchased were from the **Men's Product Department**
- **95%** of products purchased were **Jeans/Pants/Overalls /Shorts**

Recency

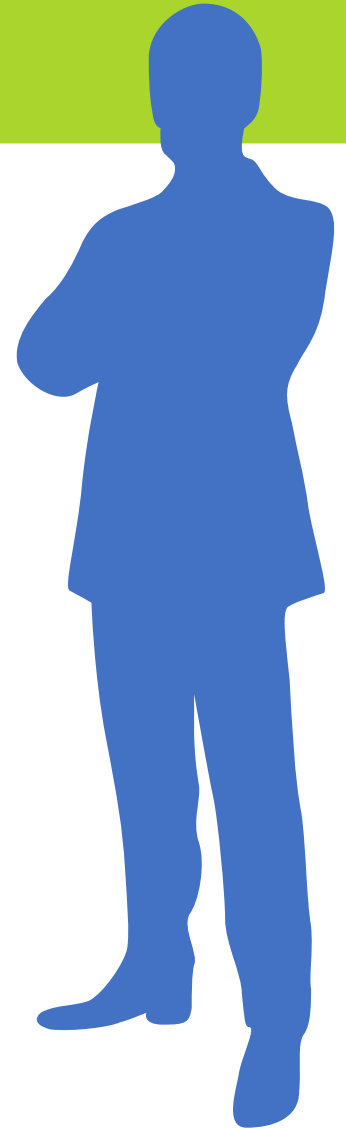
Average: 369 Days Ago

Frequency

Average: 1.69 Purchases

LTV

Average: \$97.95



Persona: (Cluster 1)

Number of Subscribers: 61,480 (13.04% of total buyers)

- ***Second longest recency*** of any cluster, on average they made their most recent purchase 974 days ago (between 2 and 3 years ago)
- ***93% of products purchased were from “Other Brands”***; this Brand category was the summation of products purchased in either Brand JC, Brand PG, Brand RW, Brand AC, Brand OW, Brand R, Brand Gen or Brand TC
- **70%** of products purchased were **Men’s Regular Size**
- **100%** of products purchased were from the **Men’s Product Department**
- **98%** of products purchased were **Jeans/Pants/Overalls /Shorts**

Recency

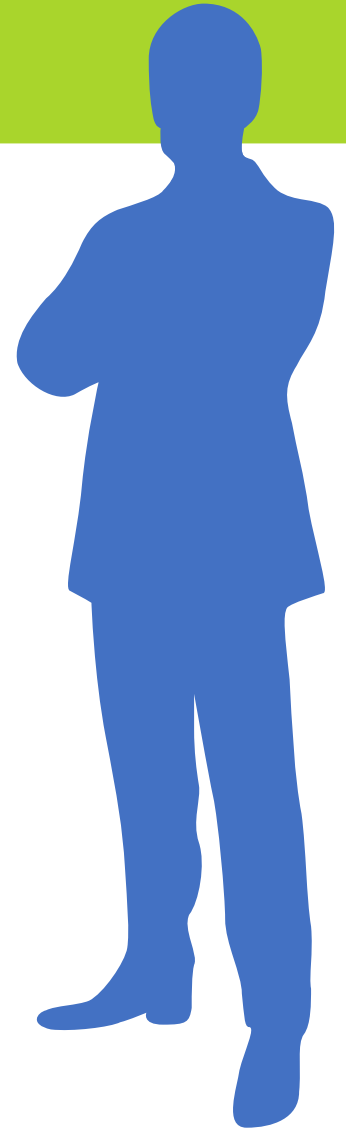
Average: 974 Days Ago

Frequency

Average: 1.35 Purchases

LTV

Average: \$86.55



Persona: Unengaged & Lapsed

(Cluster 2)

Number of Subscribers: 101,548 (21.54% of total buyers)

- ***Least recent purchasers*** of any cluster, on average their most recent purchase date was 1,048 days ago
- ***The largest cluster*** with 101,548 subscriber IDs
- ***Least engaged with email***, only **10%** have opened an email in the last 6 months and less than 1% have clicked
- **89%** of products purchased were **Men's Regular Size**
- **99%** of products purchased were from the **Men's Product Department**
- **98%** of products purchased were **Jeans/Pants/Overalls /Shorts**

Recency

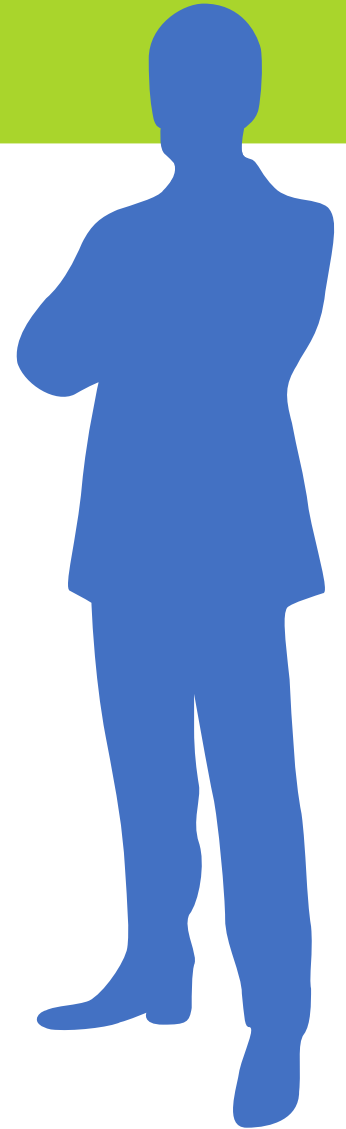
Average: 1,048 Days Ago

Frequency

Average: 1.31 Purchases

LTV

Average: \$100.94



Persona: Frequent Buyer & Big Spender

(Cluster 3)

Number of Subscribers: 47,331 (10.04% of total buyers)

- ***The highest frequency*** of any cluster, on average they've made 3.83 purchases in their lifetime
- ***The highest AOV*** of any cluster, on average a subscriber in this cluster spends \$102.51 per order
- ***The highest UPT*** of any cluster, on average they purchase 3.43 products/order
- ***The least promotionally influenced***; average promotional offer use by this cluster of **24.51%** was the lowest of any cluster
- Purchasing Men's products, but not necessarily the Regular Size
- **75%** of products purchased were **Men's Regular Size**
- **96%** of products purchased were from the **Men's Product Department**
- **89%** of products purchased were **Jeans/Pants/Overalls /Shorts**

Recency

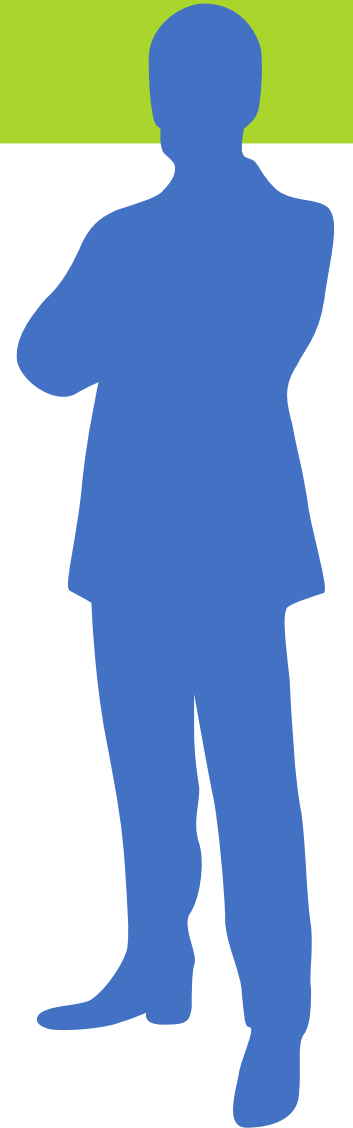
Average: 524 Days Ago

Frequency

Average: 3.83 Purchases

LTV

Average: \$313.69



Persona: Low Spenders

(Cluster 4)

Number of Subscribers: 78,145 (16.58% of total buyers)

- **The lowest LTV** of any cluster, on average they've spent \$69.95 in their lifetime
- **76% of products purchased were from Brand Reg**
- **18% of products purchased were unbranded products** (those that did not contain brand details in the product file); this was the *highest of any cluster*
- **84% of products purchased were Men's Regular Size**
- **100% of products purchased were from the Men's Product Department**
- **98% of products purchased were Jeans/Pants/Overalls /Shorts**

Recency

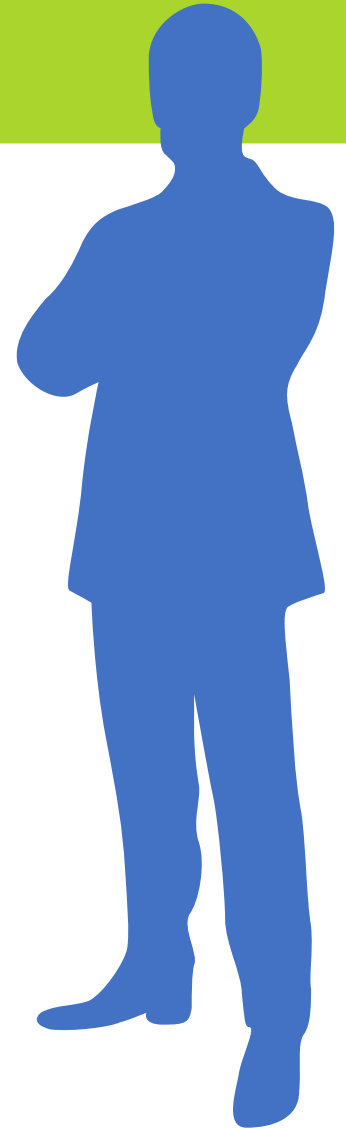
Average: 880 Days Ago

Frequency

Average: 1.38 Purchases

LTV

Average: \$69.95



Persona: Not Buying Men's

(Cluster 5)

Number of Subscribers: 48,201 (10.23% of total buyers)

- ***These shoppers ARE NOT purchasing Men's Products***
- **82% of products purchased were from Brand W**
- **1.50% of products purchased were Men's Regular Size**
- **2.09% of products purchased were from the Men's Product Department**
- **82% of products purchased were Jeans/Pants/Overalls /Shorts**

Recency

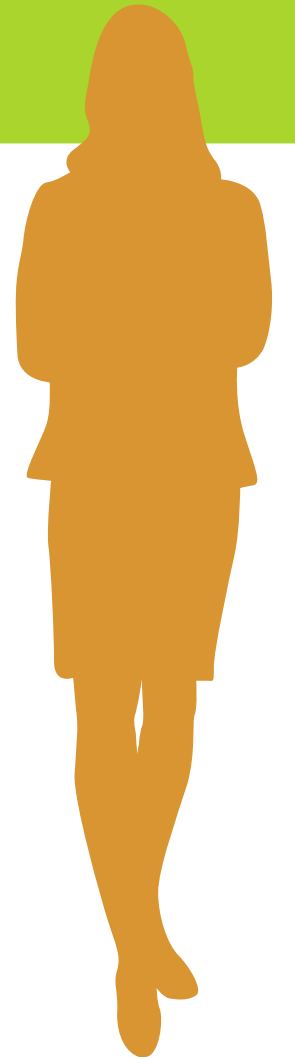
Average: 885 Days Ago

Frequency

Average: 1.40 Purchases

LTV

Average: \$104.40



Persona: Brand FS Buyers

(Cluster 6)

Number of Subscribers: 50,506 (10.71% of total buyers)

- ***The lowest LTV*** of any cluster, on average they've spent \$69.95 in their lifetime
- ***The lowest AOV*** of any cluster, on average a subscriber in this cluster spends \$51.01 per order, but
- ***The second highest UPT*** of any cluster, on average they purchase 2.56 products/order
- **89%** of products purchased by this group are from the **Brand FS**
- **80%** of products purchased were **Men's Regular Size**
- **99%** of products purchased were from the **Men's Product Department**
- **98%** of products purchased were **Jeans/Pants/Overalls /Shorts**

Recency

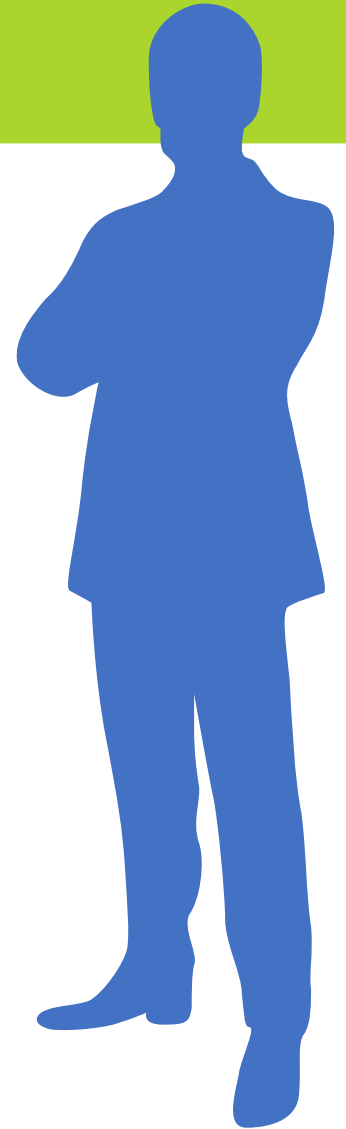
Average: 880 Days Ago

Frequency

Average: 1.48 Purchases

LTV

Average: \$71.61



Persona: Not Jean Shoppers

(Cluster 7)

Number of Subscribers: 41,981 (8.91% of total buyers)

- The **smallest** of all of the clusters with 41,981 subscribers (although close in size to cluster 0)
- Purchasing Men's products, but **NOT** from the Jeans/Pants/Overalls/Shorts categories.
- They are the only segment ***not likely to purchase jeans/pants/overalls/shorts***; 94% of the products purchased from this segment are from categories other than jeans etc
- ***The lowest frequency*** of any cluster, on average a subscriber in this cluster has purchased 1.28 times in their lifetime with the brand
- **68%** of products purchased were **Men's Regular Size**
- **98%** of products purchased were from the **Men's Product Department**
- **6%** of products purchased were **Jeans/Pants/Overalls /Shorts**

Recency

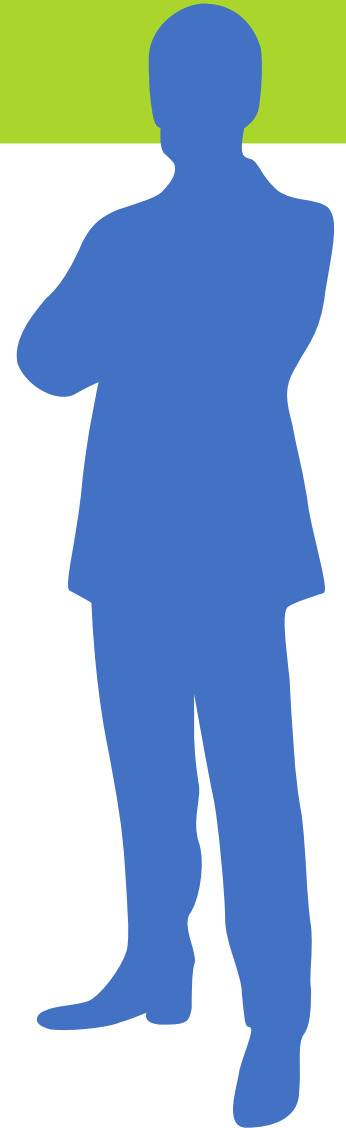
Average: 944 Days Ago

Frequency

Average: 1.28 Purchases

LTV

Average: \$85.22



Next Steps

- *What are the next steps?*

Client Recommendations

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	42,185	369	1.65	\$97.95	3.6	\$62.71	2.29	49.84%
1	61,480	974	1.35	\$86.55	2.8	\$67.17	2.15	28.62%
2	101,548	1,048	1.31	\$100.94	2.6	\$80.22	2.06	25.57%
3	47,331	524	3.83	\$313.69	10.8	\$102.51	3.43	24.51%
4	78,145	880	1.38	\$69.95	3.1	\$52.51	2.35	27.59%
5	48,201	885	1.40	\$104.40	2.8	\$75.44	2.01	30.53%
6	50,506	880	1.48	\$71.61	3.5	\$51.01	2.56	33.96%
7	41,981	944	1.28	\$85.22	2.7	\$68.85	2.17	31.07%

Average Product Behavior by Cluster

(Average Percentages of Products Purchased by Size Category (Men's Reg vs. Not Men's Reg), Product Department (PD Men vs. PD Not Men) & Category (Jeans/Pants/Overalls/Shorts vs. Other))

Cluster #	# of Subscribers	Men's Reg	Not Men's Reg	PD Men	PD Not Men	Category Jeans etc.	Category Other
0	42,185	79.93%	20.07%	98.77%	1.23%	95.01%	4.99%
1	61,480	69.57%	30.43%	99.60%	0.40%	98.21%	1.79%
2	101,548	88.96%	11.04%	99.26%	0.74%	97.82%	2.18%
3	47,331	74.82%	25.18%	95.88%	4.12%	89.01%	10.99%
4	78,145	83.97%	16.03%	99.70%	0.30%	98.49%	1.51%
5	48,201	1.50%	98.50%	2.09%	97.91%	82.47%	17.53%
6	50,506	79.67%	20.33%	99.45%	0.55%	98.07%	1.93%
7	41,981	68.08%	31.92%	98.44%	1.56%	6.07%	93.93%

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	42,185	28.90%	25.52%	18.07%	7.16%	20.35%
1	61,480	1.81%	1.81%	1.62%	1.45%	93.24%
2	101,548	97.09%	0.47%	0.54%	0.60%	1.30%
3	47,331	40.74%	18.24%	12.58%	6.23%	2.22%
4	78,145	1.51%	75.65%	1.47%	18.10%	3.27%
5	48,201	82.46%	1.54%	1.64%	11.86%	2.50%
6	50,506	3.24%	2.86%	88.73%	1.21%	3.96%
7	41,981	49.47%	5.19%	0.72%	13.25%	31.37%

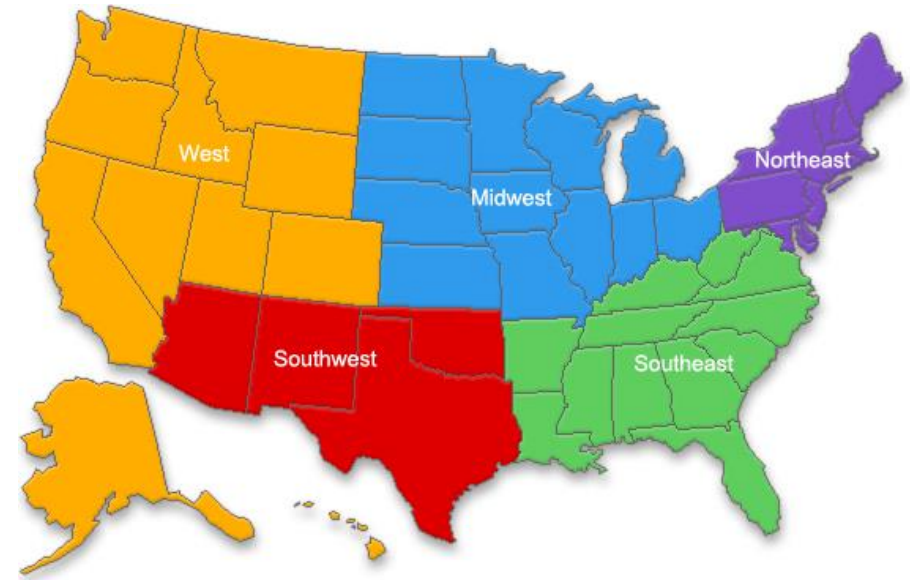
Average Email Behavior by Cluster

(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	42,185	41,986	98.95%	100.00%
1	61,480	59,089	12.44%	0.27%
2	101,548	97,671	9.70%	0.08%
3	47,331	46,020	25.17%	11.59%
4	78,145	74,726	13.46%	0.20%
5	48,201	46,480	24.44%	13.15%
6	50,506	48,476	13.94%	0.39%
7	41,981	40,444	20.13%	9.80%

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%



THANK YOU



Julianna Renaud

Web Analytics Strategist



Charlotte Werger, PhD

Springboard Mentor