



# Recommendation Engine Proposal for eCommerce Bookstore



*Technical Presentation for Chief Technology Officer*



# Our Team



**Anusha Bhat**  
*Data Engineer*



**Sara Chaker**  
*Team Lead*



**Louis Gao**  
*Data Analyst*



**Hritik Jhaveri**  
*Business Analyst*



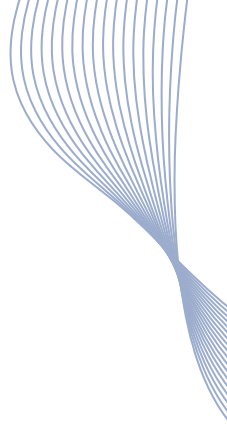
**Q Li**  
*Data Scientist*



**John Melel**  
*Associate Partner, SME*



**Olivia Rumere**  
*Data Scientist*



# Agenda

**01**

## Business Problem & EDA

*Summary of the eCommerce  
Bookstore's primary challenge  
and brief review of dataset*

**02**

## Model Choice & Performance

*Overview of the selected model  
and model performance*

**03**

## Segment Pairing

*Demonstration of Customer  
Segmentation from Model*

**04**

## Business Impact

*Credible estimate of the model's  
business value and financial impact*

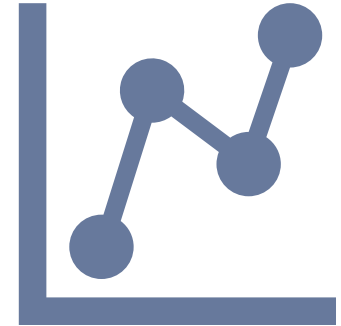
# Personalizing Book Recommendations for Increased Sales

- Your Ecommerce Bookstore promotes the same books to all visitors, **limiting potential sales** by not accounting for individual customer preferences.
- The challenge is to develop more **advanced recommendation system** that **segments users** by behaviors and purchase patterns to allow for **personalized product suggestions**.



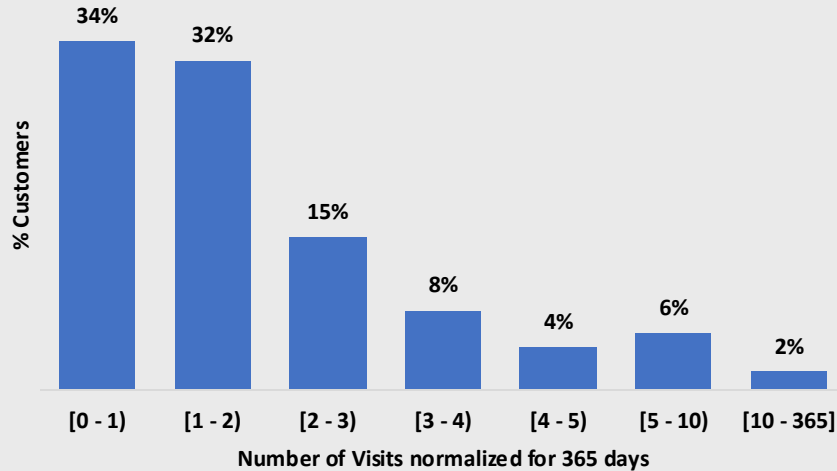
# Leveraging Analytics to Drive Personalized Sales

- We will use **unsupervised machine learning techniques** to create this recommendation engine.
- This will drive higher engagement, increase the average order value, and enhance customer satisfaction through more personalized suggestions, improving overall **sales performance by 15%** in the next 6 months.

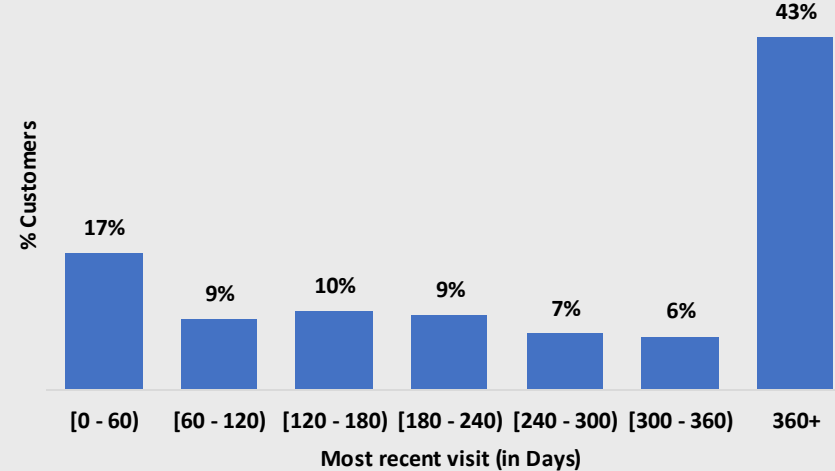


# A majority of customers visit the site infrequently and have less recent purchases.

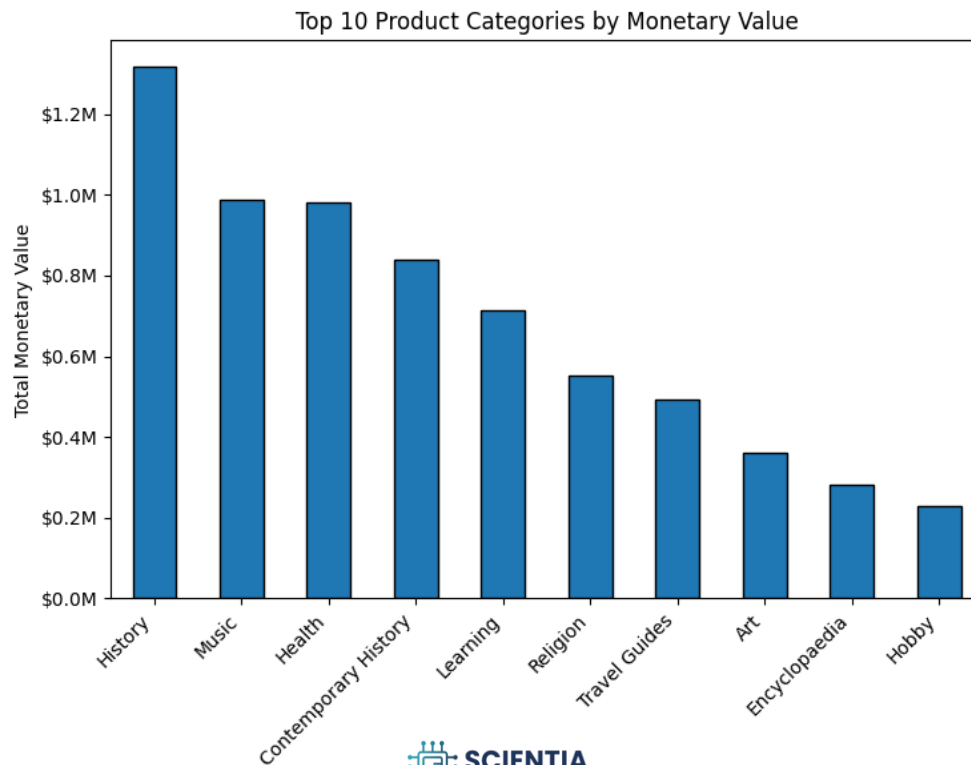
## Visits per year



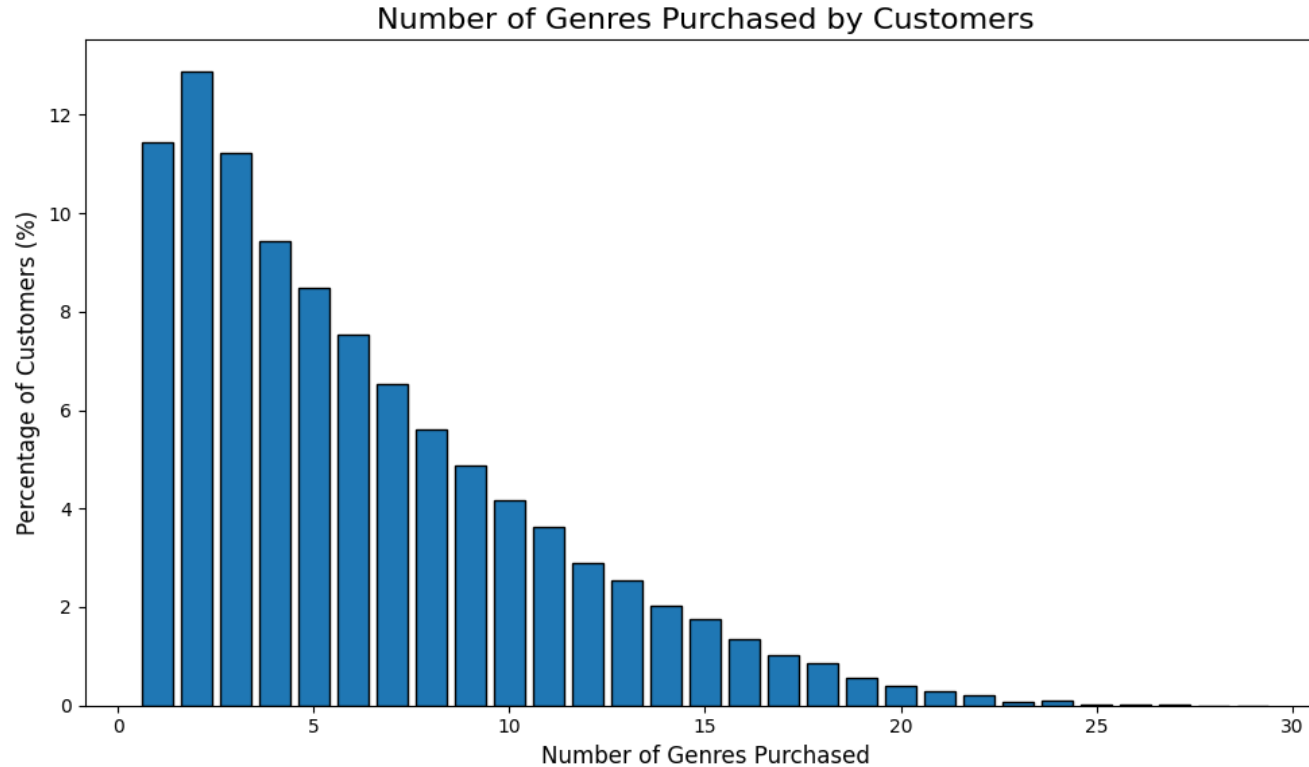
## Recency



# History, Music, and Health books contribute the most to sales.



# Most customers purchase fewer than 5 genres.





# K-Means Offers the Most Comprehensive Model Features

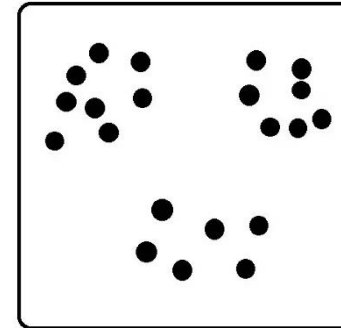
	DBSCAN	GMM	K-Means
<b>Suitable for Large Datasets</b>	✗	✗	✓
<b>High Interpretability</b>	✗	✗	✓
<b>Easy Parameter Tuning</b>	✗	✗	✓
<b>Fast and Efficient</b>	✗	✗	✓
<b>Handles Arbitrary Shapes</b>	✓	✗	✗
<b>Handles Overlapping Clusters</b>	✗	✓	✗
<b>Handles Noise/Outliers</b>	✓	✗	✗

DBSCAN = Density-Based Spatial Clustering of Applications with Noise  
GMM = Gaussian Mixture Models

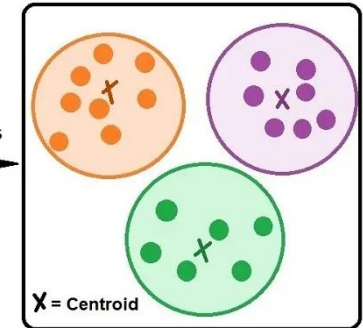
# Utilizing K-Means to Segment Customers for Personalized Recommendations

K-Means Clustering is an Unsupervised Machine Learning algorithm, which groups an unlabeled dataset into different clusters.

Unlabeled Data



Labeled Clusters



K-Means

By grouping customers into clusters, we can identify patterns and similarities that allow us to make personalized and targeted recommendations.

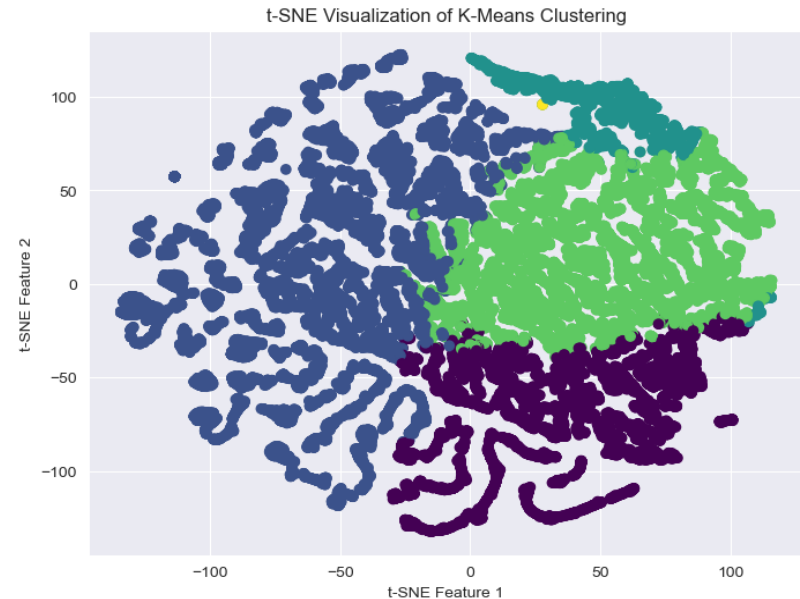
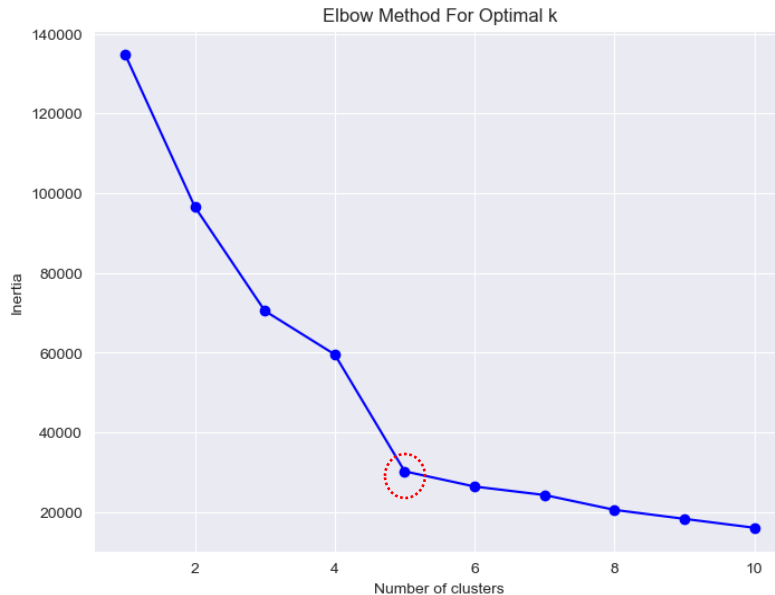
# Benefits of K-Means as a Practical Solution for Effective Segmentation

## Key Advantages of K-Means

- ***Simplicity and Efficiency:*** K-Means is computationally efficient for large datasets, making it ideal for creating a recommendation engine that needs to handle a significant volume of data in a timely manner.
- ***Interpretability:*** The clusters formed by K-Means are easy to interpret and visualize, which helps in clearly understanding the distinct segments of book categories.
- ***Scalability:*** As our dataset grows, K-Means remains a scalable option, adapting well to higher dimensional data.
- ***Objective Alignment:*** The algorithm's goal of minimizing intra-cluster variance aligns perfectly with the need to group similar customers together to make accurate recommendations.

# Customers Can Be Segmented Into 5 Clusters

- We selected **k=5** based on within-cluster sum-of-squares elbow plot for our K-means model.
- The clusters are fairly differentiated as seen in the t-SNE plot.



# K-Means Model Clustering Provides Advertisement Strategies

- We observe 5 clusters which are segmented on key metrics, each falling into different advertisement strategies.

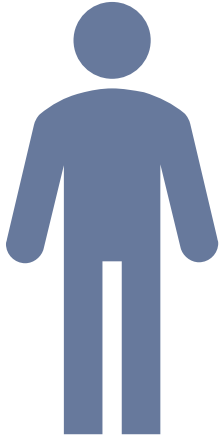
	Clusters				
	[1] Long Lost	[2] New Cheap	[3] Loyal	[4] Potential Loyal	[5] VIP
#customers	6,980	14,676	1,919	10,135	3
%customers	21%	44%	6%	30%	0%
avg <b>Recency</b>	1422	255	154	313	117
avg <b>Frequency</b>	3.0	2.5	27.1	8.5	10.7
avg <b>Amount</b>	\$ 119	\$ 84	\$ 1,104	\$ 310	\$ 389,662
avg <b>Time on File</b>	1846	476	2165	1940	1636
avg <b>Visits per Year</b>	0.6	3.2	4.8	1.7	2.9

# The "Long Lost Customer"

## Segment Pairing - Cluster 1

*High recency (haven't interacted recently), low frequency, and low monetary value.*

Customer ID 8907



*Personalized Incentive*



**Special discount or coupon** to encourage re-engagement by sending email/message.

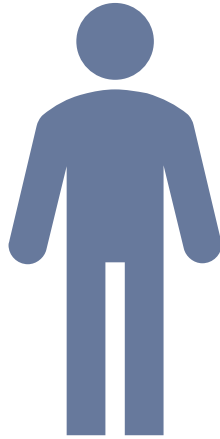
**"We miss you! Enjoy 20% off your next purchase using coupon GETBACK20 "**

# The "New and Frugal Customer"

## Segment Pairing - Cluster 2

*Low recency values but low frequency and monetary value, indicating recent but minimal engagement.*

Customer ID 2234785



*Personalized Incentive*



Send a personalized **welcome email** thanking them for their first purchase and introducing your brand's value propositions.

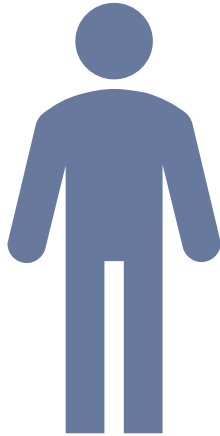
"Thank you for making your first purchase with us! To show our appreciation, we'd love to offer you **10% off your next purchase!** Just use the coupon code **WELCOME10** at checkout."

# The "Loyal Customer"

## Segment Pairing - Cluster 3

*Very loyal to the brand, making frequent and recent purchases.*

Customer ID 4565



*Personalized Incentive*



Offer **exclusive rewards** such as **early access** to products, **loyalty perks** like reward points programs, and **personalized incentives**.

"Celebrate your birthday with an exclusive **15% discount** – because you deserve it!"

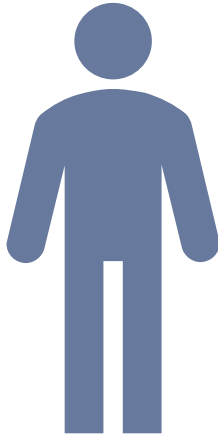


# The "Potential Loyal Customer"

## Segment Pairing - Cluster 4

*Moderate recency, frequency, and monetary value.*

Customer ID 68283



*Personalized Incentive*



Provide them with **bundled offers** on complementary items to increase average order value. Based on their purchase history, also **recommend similar items**.

"Pair your [recent purchase] with [complementary product] for a perfect match – **save 15%** when you buy them together!

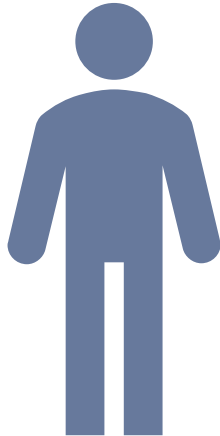
Based on your purchase history, we think you might love:  
**[Recommended Product #1]."**

# The "VIP Customer"

## Segment Pairing - Cluster 5

*Extreme spenders, very recent*

Customer ID 543



*Personalized Incentive*



**Early access to limited edition products** or collections before they are available to the public. Provide also a **personalized shopping concierge service**.

"Be the first to explore our **limited-edition collections** and **new arrivals** before they go live to the public!"

"Need help finding the perfect product? Our dedicated concierge team is here for you!"

# The Personalized Recommendation Model Can Lead to a \$1,601,342 Increase in Revenue

## Projected Growth with Personalized Recommendation

Customer Segment	Growth Rate	Estimated Revenue	Revenue Increase	Profit Increase
Long Lost	10%	\$913,682	\$83,062	\$24,919
New and Frugal	15%	\$1,417,702	\$8,184,918	\$55,475
Loyal	25%	\$2,648,220	\$529,644	\$158,893
Potential Loyal	20%	\$3,770,220	\$628,370	\$188,511
VIP	15%	\$1,344,334	\$175,348	\$52,604
<b>Total</b>	<b>18.8%</b>	<b>\$10,094,158</b>	<b>\$1,601,342</b>	<b>\$480,402</b>

**Current Revenue: \$8,492,816**

Growth Assumptions:

- Purchase Frequency Increase: 10%
- Spending Per Purchase Increase: 20%
- Profit Margin: 30%

Higher rates for loyal customers (25%), medium rates for new frugal, potential loyal and extreme spenders (15-20%), lower rates for long lost customers (10%).

Estimated Revenue = Current Revenue + Revenue Increase

Profit Increase = Current Revenue x Growth rate x Profit Margin

# Key Takeaways



## Intelligent Segmentation

5 distinct clusters enabling personalized recommendations by K-Means clustering.

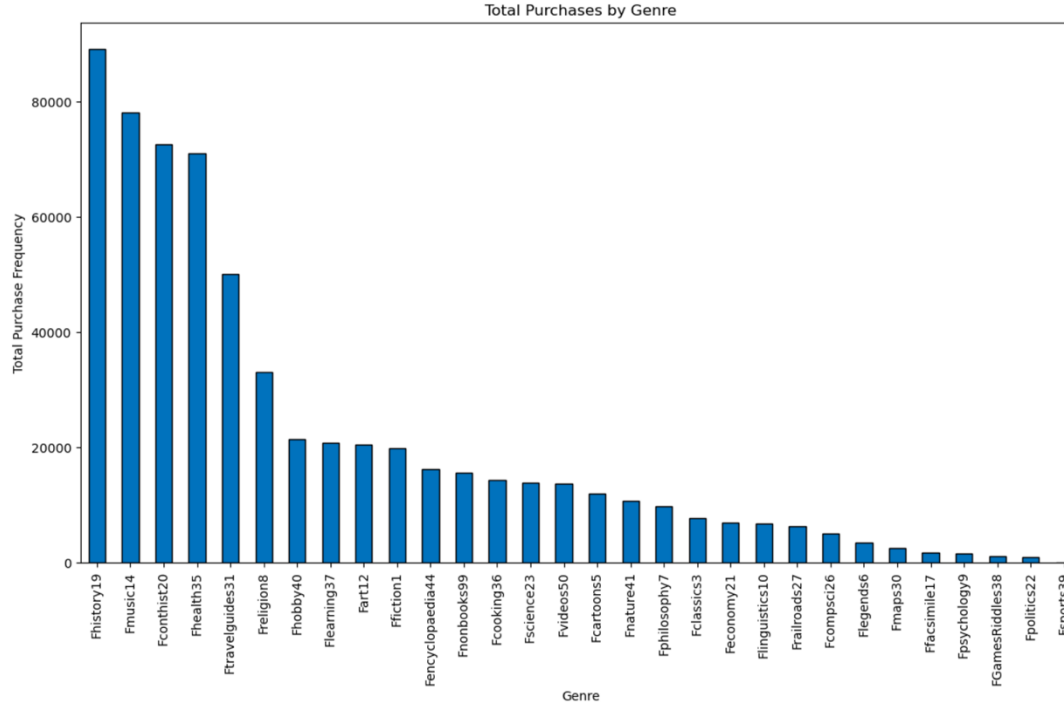


## Revenue Impact

Increase overall revenue by 19% and an additional profit of half a million dollars.

# Appendix

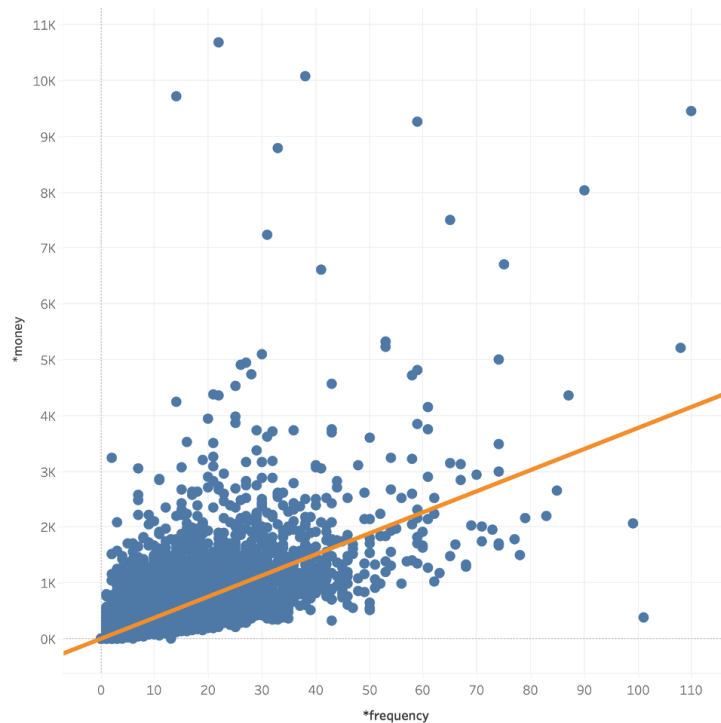
History is the most frequently bought genre  
meanwhile sports books are the least frequently  
bought.





There is an  
increase in money  
spent as frequency  
increases.

Scatter Plot of f and m (after removing outliers)

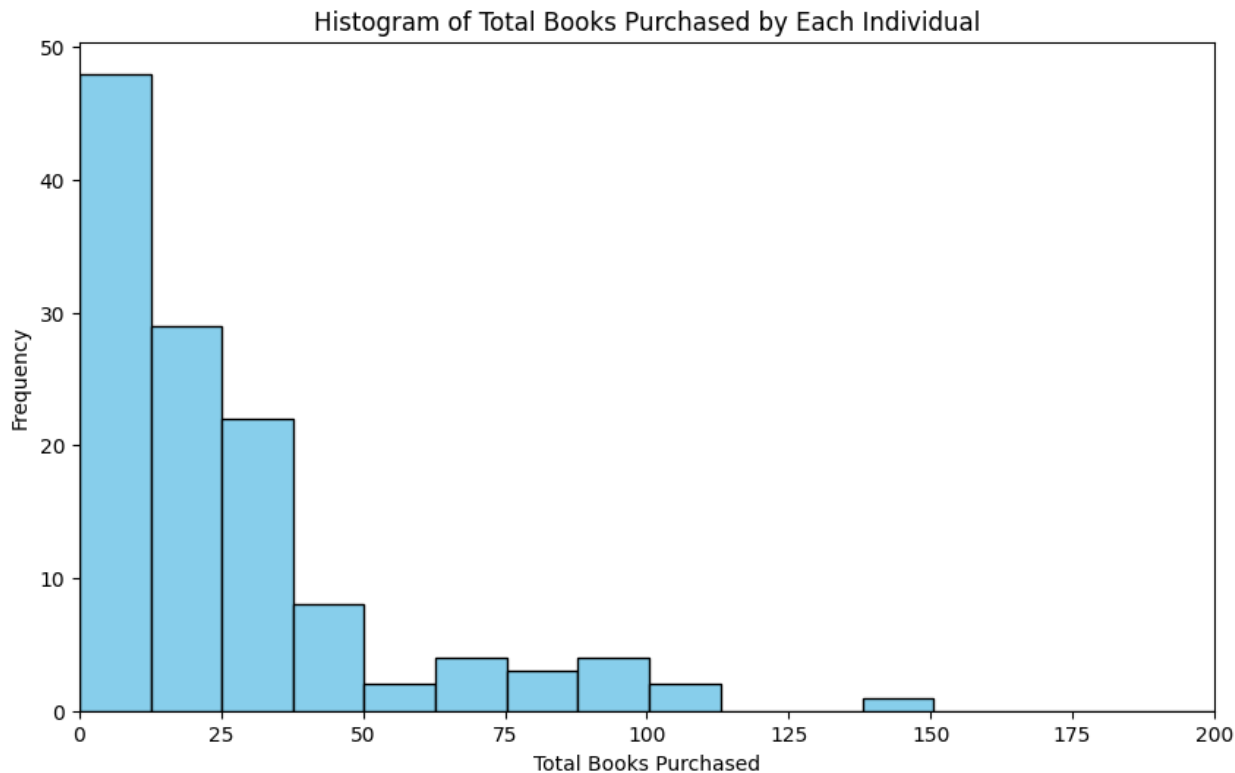


Sum of \*frequency vs. sum of \*money. Details are shown for Id. The view is filtered on Id, which excludes 5405602, 5900190, 8192553, 13729756 and 14158205.





**Most  
customers buy  
less than 25  
books.**







# Advertisement Strategies From Clustering

Cluster ID	Segment	Total Customer	Recommendation
1	Long lost customers	6,980 (21%)	Offer reactivation incentives such as "Welcome Back" discounts or special campaigns to entice them to re-engage.
2	New and Frugal Customers	14,676 (44%)	Offer upselling opportunities, such as discounts on higher-value items or loyalty program enrollment.
3	Loyal customers	1,919 (6%)	Provide exclusive rewards, loyalty perks, or early access to products to strengthen their loyalty.
4	Potential loyal customers	10,135 (30%)	Encourage higher spending through bundle offers.
5	VIP Customers	3 (<1%)	Early access to new product before launching or sell on public for same product or product in same genre





# Business Impact Estimate

## Current Baseline Revenue

Segment	Customers	Avg. Spending	Current Revenue	% of Revenue
Long Lost	6,980	\$119	\$830,620	9.8%
New and Frugal	14,676	\$84	\$1,232,784	14.5%
Loyal	1,919	\$1,104	\$2,118,576	24.9%
Potential Loyal	10,135	\$310	\$3,141,850	37.0%
VIP	3	\$389,662	\$,1168,986	13.8%
<b>Total</b>	<b>33,713</b>	<b>\$251.92</b>	<b>\$8,492,816</b>	<b>100%</b>

Current Revenue = Number of Customers x Average Spending

