

# Customer Analytics: Segmentation and Lifetime Value Modelling

**Author:** [Denis Kapesa]

## 1. Introduction

This project applies advanced customer analytics to identify high-value segments and estimate Customer Lifetime Value (CLV) using five years of transactional retail data. Through a structured process combining RFM segmentation, K-Means clustering, and BG/NBD probabilistic modelling, the analysis quantifies customer behaviour, highlights profit concentration among a small cohort of loyal buyers, and provides clear recommendations for retention and growth. The findings reveal that most customers are single-purchase buyers, while Loyal Customers and Champions drive the majority of predicted value. Actionable strategies are proposed to prioritise marketing investment, automate reactivation for lower-value segments, and build scalable loyalty initiatives grounded in data-driven evidence. The analysis is implemented in Python (Google Colab) and focuses on interpretable, business-facing outputs.

## 2. Data preparation and feature engineering

### 2.1 Dataset summary

The dataset contains 1,194 transaction records covering the period March 2020 to March 2025. After aggregation to the customer level there are 802 unique customers. The dataset includes order amount, profit, quantity, product category, payment mode, location and timestamps for each order.

#### Key aggregated metrics

- Total customers: 802
- Total revenue: \$6,182,639.00
- Average revenue per customer: \$7,709.03

### 2.2 Cleaning and feature engineering

All numeric fields were validated and cast to appropriate types. Order dates were converted to datetime and used to derive temporal fields. Transactions were aggregated to customer level to produce the following engineered features:

## Retail Sales Analytics

Feature	Description
RecencyDays	Days since last purchase
Frequency	Number of unique orders per customer
Monetary	Total revenue generated by the customer
TotalProfit	Total profit earned from the customer
AOV	Average Order Value = Monetary / Frequency
ProfitMargin	Profit-to-Revenue ratio = TotalProfit / Monetary

### Descriptive statistics (customer level)

Metric	Mean	Std Dev	Min	25%	Median	75%	Max
Recency (days)	896.99	511.14	0	463.25	905.00	1306.00	1819.00
Frequency	1.006	0.079	1	1.00	1.00	1.00	2.00
Monetary (\$)	7,709.03	5,229.21	523	3,882.75	6,937.50	9,908.00	28,557.00
TotalProfit (\$)	2,008.35	1,693.77	52	727.00	1,575.00	2,826.00	8,840.00
AOV (\$)	7,655.74	5,162.23	523	3,877.00	6,933.00	9,902.50	28,557.00
Profit Margin	0.26	0.13	0.01	0.16	0.25	0.37	0.50

### Key observations from feature engineering

- Repeat purchases are rare: almost all customers made only one purchase (Frequency  $\approx 1.0$ ).
- Monetary values are right skewed; a small group of high-value customers drives a large portion of revenue.
- Profit margins vary considerably across customers, indicating variation in product mix or discounting.

These engineered features form the basis for segmentation and lifetime value modelling.

### 3. Exploratory data analysis (selected visuals and observations)

Only the most business-relevant visuals are included in the final report. Insert the charts from the notebook at the locations shown.

#### Figure 1. Total revenue and total customers

Total customers: 802

Total revenue: \$6,182,639.00

Average revenue per customer: \$7,709.03

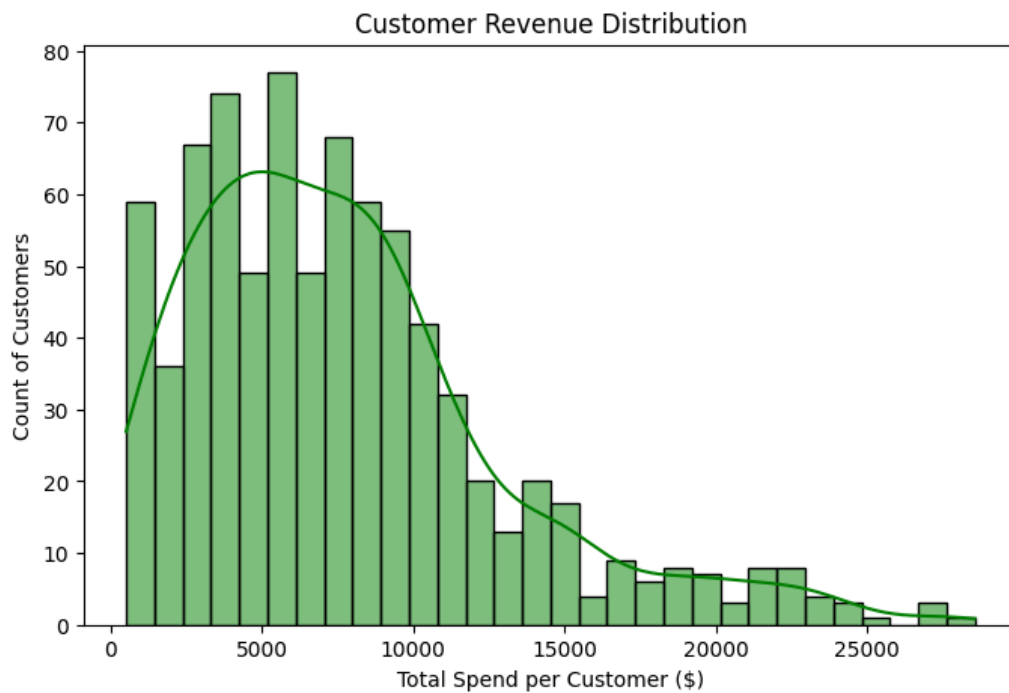
#### Figure 2. Top 10 customers by total spend

Customer Name	Monetary (\$)	Frequency	Recency (Days)
Cory Evans	28,557	1	272
Emily Ellison	27,352	1	941
George Foster	27,352	1	1009
Nicholas Anderson	27,352	1	1215
Katherine Williams	25,121	1	1020
Jacqueline Harris	24,433	1	771
Randy Johnson	24,295	1	881
Tammy Bell	23,895	1	725
Brian Green	23,737	1	1189
Zachary Perez	23,737	1	1136

Table showing the top 10 customers by monetary value. The top customer values range up to \$28,557.

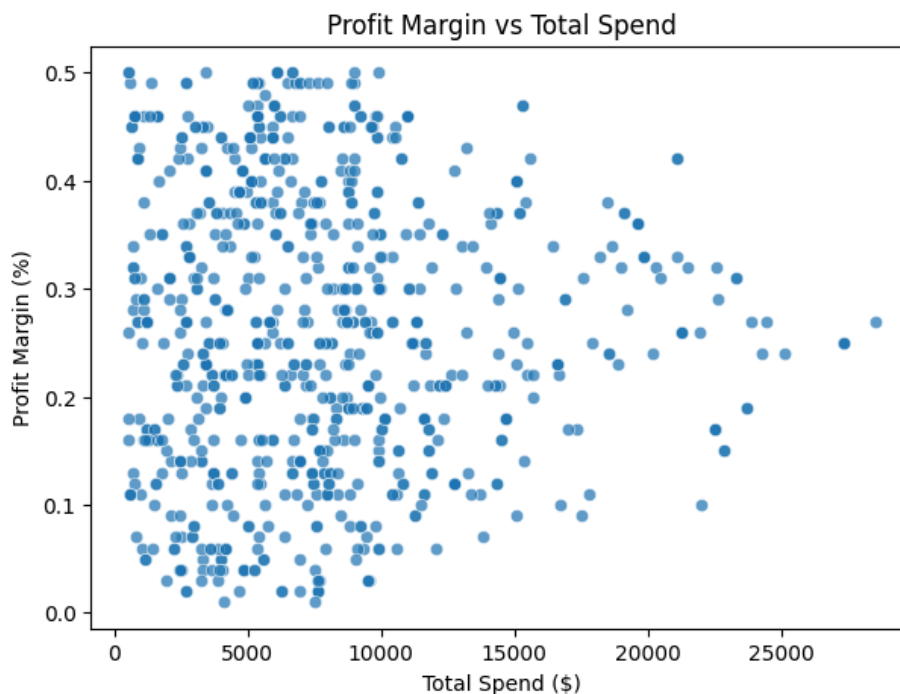
## Retail Sales Analytics

**Figure 3. Distribution of customer revenue**



Histogram with KDE overlay showing a peak around \$4,000–\$5,000 and a long right tail to ~\$28,0

**Figure 4. Profit margin versus total spend**



Scatterplot of Profit Margin against Monetary. This highlights the lack of strong positive correlation between spend and margin; high spenders are not necessarily high margin.

## **Retail Sales Analytics**

### **Narrative summary**

- Revenue is concentrated: a small cohort of customers contributes disproportionately to overall revenue.
- The customer base is largely transactional, with minimal repeat behaviour.
- Profit margin heterogeneity suggests that revenue alone is not a sufficient indicator of long-term profitability.

These EDA observations motivate a segmentation and CLV approach that balances frequency, recency and monetary measures, and that explicitly considers profitability when prioritising interventions.

## 4. Customer segmentation

Segmentation was performed with two complementary approaches: quantile RFM scoring for interpretability and K-Means clustering for latent behavioural patterns.

### 4.1 RFM quantile scoring

Each RFM variable was divided into quartiles and assigned a 1–4 score. The three scores were summed to produce an RFM score (range 3–12). Based on the RFM score, customers were assigned to five descriptive segments: Champions, Loyal Customers, Potential Loyalists, At Risk and Lost.

#### Segment counts (approximate)

- Champions: 130
- Loyal Customers: 266
- Potential Loyalists: 288
- At Risk: 104
- Lost: 14

**Figure 5. Customer distribution by RFM segment**



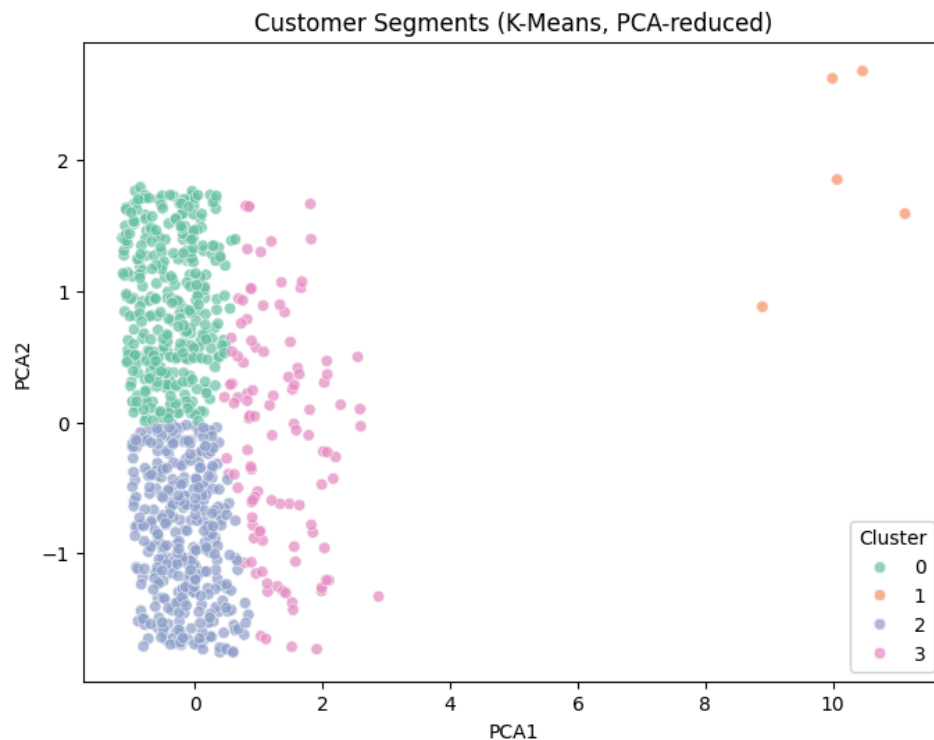
Bar chart showing counts per segment.

### 4.2 K-Means clustering

K-Means was applied to standardised RFM features. The Elbow method and silhouette analysis supported a four-cluster solution. Cluster profiling revealed the following high-level groups:

Cluster	Profile
Cluster 1	Small elite group: frequent and high spenders
Cluster 3	High spenders, moderate recency
Cluster 2	Recently active, moderate spend
Cluster 0	Dormant or low spend customers

**Figure 6. PCA scatterplot of clusters**



2-D PCA scatter with clusters indicated.

### 4.3 Segmentation insight

RFM quantiles provide interpretability for business stakeholders, while K-Means highlights latent structures not captured by simple quantiles. Both methods align: Champions and high-value clusters are small but strategically significant. Potential Loyalists are the largest cohort and represent the most scalable growth opportunity.

## 5. CLV modelling and retention forecasting

### 5.1 Approach and justification

For probabilistic lifetime estimation we used the BG/NBD model to forecast purchase frequency and the Gamma-Gamma framework to estimate monetary value per transaction where sufficient repeat data exist. BG/NBD is widely used in non-contractual settings to estimate the probability a customer will make future purchases from recency and frequency history; it is a practical and well validated choice for this task. See Fader, Hardie and Lee (2005) for a formal description of the BG/NBD model and the behavioural assumptions underpinning it. [brucehardie.com+1](https://brucehardie.com/notes/010/)

In practice this dataset has very low repeat purchase rates (most customers have frequency = 1). The Gamma-Gamma monetary model requires sufficient repeat purchase observations to produce stable estimates. For customers without repeat purchases the model cannot reliably estimate conditional average order value. The Python lifetimes package documents this requirement and recommends fitting the monetary model on customers with frequency > 0 only; for customers without repeat transactions it is appropriate to use observed order values or global averages as fallbacks. For transparency, the implementation in Colab followed this guidance and used observed average order values where Gamma-Gamma predictions were not feasible. [lifetimes.readthedocs.io+1](https://lifetimes.readthedocs.io/)

### 5.2 Implementation notes

- Observation period ends: 15 March 2025.
- Unit of time: days.
- Horizon for CLV: 180 days (six months). The horizon was chosen to reflect a near-term marketing planning window and to limit extrapolation given the sparse repeat behaviour.
- BG/NBD was fitted with a small penaliser to stabilise estimation for this modest dataset; customers with frequency > 1 were used to support fit diagnostics when required.

### 5.3 Customer-level CLV results (top examples)

Figure 7. Top predicted CLV customers

Customer Name	Frequency	Avg order value (\$)	Expected purchases (180d)	Predicted CLV (\$)
---------------	-----------	----------------------	---------------------------	--------------------

## Retail Sales Analytics

Dr. Terry Alvarado	0.0	5,178.09	0.0029	15.13
Elizabeth Hernandez	0.0	5,178.09	0.0029	15.05
Vanessa Bauer	0.0	5,178.09	0.0028	14.73
William Martin	0.0	5,178.09	0.0028	14.57

Interpretation: low expected repeat purchases drive small 180-day CLV values for single-purchase customers, even when their AOV is high. The result is a realistic, conservative short-term forecast that reflects the dataset's behaviour.

### 5.4 Segment-level CLV aggregation

CLV estimates were merged with the segmentation to produce segment summaries.

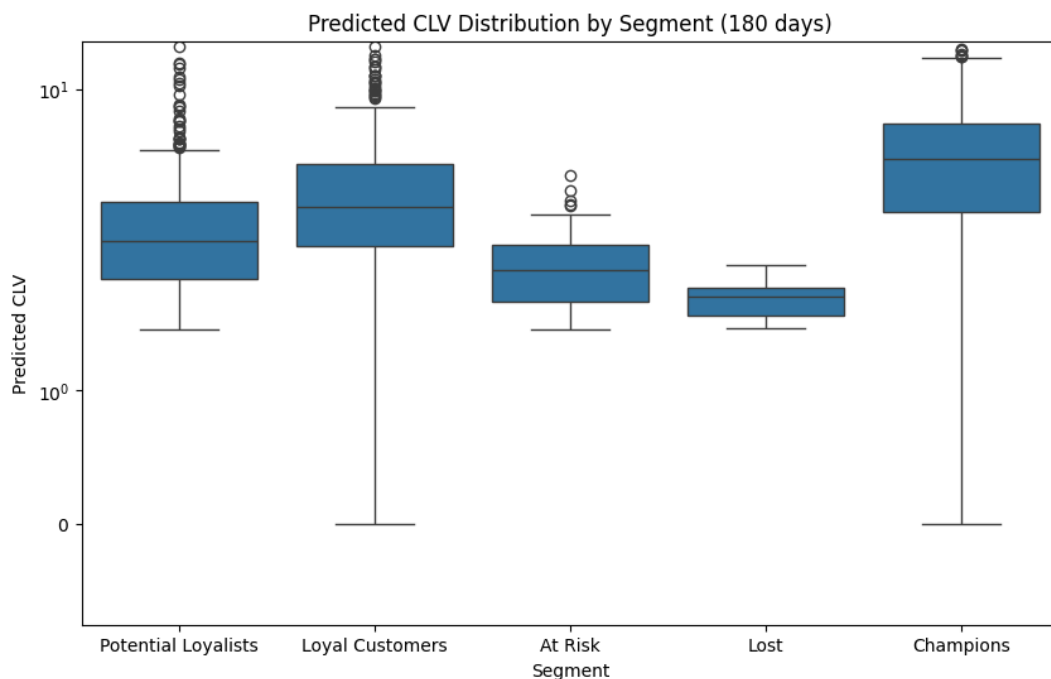
**Table: CLV by segment (180 day horizon)**

Segment	Customer s	Avg CLV (\$)	Total CLV (\$)	Avg expected purchases
Loyal Customers	266	4.13	1,098.78	0.000798
Potential Loyalists	288	3.11	896.99	0.000601
Champions	130	5.80	753.59	0.001119
At Risk	104	2.05	213.17	0.000396
Lost	14	1.69	23.70	0.000327

## Retail Sales Analytics

- Loyal Customers: deliver the highest total CLV, driven by their large population and consistent engagement. Their moderate average CLV and above-average expected purchases make them ideal for scalable retention programs.
- Potential Loyalists are the largest segment and contribute significantly to total CLV. Their average CLV is lower than Champions and Loyal Customers, but their volume makes them a strategic growth opportunity. Targeted nurturing could elevate their value.
- Champions lead in average CLV and expected purchases, despite being a smaller group. Their high transaction value and engagement make them ideal for premium loyalty programs and personalised upsell campaigns.
- At Risk customers show declining engagement and low CLV. While they still represent a sizable group, reactivation efforts should be selective and cost-efficient.
- Lost customers have minimal future value. Their low CLV and purchase probability suggest limited returns from re-engagement, unless tied to broader win-back campaigns or product-specific triggers.

**Figure 8. Boxplot of predicted CLV by segment (180 days)**



Interpretation: Loyal Customers generate the highest aggregate CLV due to their larger population, while Champions have the highest average CLV. Potential Loyalists present scale opportunities; At Risk and Lost segments offer limited short-term returns

## 6. Insights and recommendations

The analysis translates into clear prioritisation and tactical recommendations.

### 6.1 Strategic priorities

1. **Prioritise retention for Loyal Customers and Champions.**  
Loyal Customers generate the largest share of predicted CLV in aggregate, and Champions have high average CLV. A targeted retention programme for these groups will likely produce the best short-term ROI.
2. **Nurture Potential Loyalists at scale.**  
Potential Loyalists are the largest segment. Low-cost marketing automation—such as personalised follow-up emails, product recommendations and time-limited offers—could move a portion of this group into higher CLV segments.
3. **Use automated, cost-efficient reactivation for At Risk customers.**  
Given their lower CLV, manual high-touch interventions are unlikely to be cost effective. Automated campaigns with modest incentives or cross-sell messaging are recommended.
4. **Be selective with win-back for Lost customers.**  
Lost customers show minimal predicted value. Consider product-triggered reactivation or low cost offers under a broader win-back programme.

### 6.2 Tactical recommendations

- Implement a tiered loyalty programme: premium benefits for Champions, points or discounts for Loyal Customers and tailored onboarding for Potential Loyalists.
- Use CLV ranking to allocate marketing budget: run A/B tests to measure uplift before scaling.
- Monitor CLV trends monthly to detect changes in repetition and to reallocate spend dynamically.

### 6.3 Example ROI simulation

A simple simulation shows incremental CLV uplift if expected purchases increase by 20 per cent for Potential Loyalists. This calculation can be run in Colab and compared to campaign cost to estimate ROI. (Include simulation outputs if you run the test.)

# 7. Limitations and further work

## 7.1 Limitations

- Low repeat-purchase density. Most customers are single-purchase, which reduces the signal available to estimate future purchases and monetary distributions over short horizons.
- Model assumptions. BG/NBD assumes stationarity in purchasing behaviour and that dropout is unobserved; major changes in business or market conditions can violate these assumptions. [brucehardie.com+1](https://brucehardie.com)
- No marketing or channel covariates. Adding campaign, acquisition channel or product affinity data would improve targeting and causal understanding.

## 7.2 Next steps

- Backtest the BG/NBD predictions by training on an earlier cutoff and comparing predicted purchases to realized purchases.
- Integrate basic marketing data (email opens, campaign exposure) to estimate uplift and attribution.
- If the business acquires more repeat behaviour over time, fit the Gamma-Gamma model formally for improved monetary predictions. The Gamma-Gamma framework is standard for spend modelling where repeat transactions are sufficient. [brucehardie.com+1](https://brucehardie.com)

# 8. Conclusion

This project produced a robust, business-facing CLV framework built on RFM segmentation and probabilistic lifetime modelling. The results highlight the strategic importance of a relatively small group of high-value customers and a large cohort of Potential Loyalists who represent scalable growth opportunities. The modelling approach is conservative given sparse repeat behaviour and provides a solid, actionable basis for prioritising retention and marketing investments.

### References

1. Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). Counting your customers the easy way: An alternative to the Pareto/NBD model. *Marketing Science*. [brucehardie.com+1](https://brucehardie.com/1)
2. Fader, P. S. & Hardie, B. G. S. (2013). The Gamma-Gamma model of monetary value. Technical note. [brucehardie.com](https://brucehardie.com)
3. Lifetimes documentation. Quickstart and model usage notes for BetaGeoFitter and GammaGammaFitter. [lifetimes.readthedocs.io+1](https://lifetimes.readthedocs.io+1)