Customer Segmentation Analysis

Turning customer profiles into testable marketing segments

K-Means analysis of a 200-customer mall dataset using age, gender, annual income, and Spending Score.

REPORT STRUCTURE

Contents

Executive findings first. Methodology and interpretation limits follow in the appendices.

01 Executive Summary 03

02 Business Problem & Analytical Approach 04

03 Segmentation Methodology 05

04 Segment Portfolio 06

05 Key Customer Findings 07

06 Strategic Recommendations & Measurement Plan 08

A Model Methodology & Definitions 09

B Supporting Analysis & Limitations 10

REPORT SCOPE

This report describes five groups in a 200-customer sample. It does not establish permanent customer types or commercial outcomes.

01 / EXECUTIVE SUMMARY

Executive Summary

The analysis used K-Means clustering to group 200 customers using four characteristics: Age, Gender, Annual Income, Spending Score.

The final model produced five clusters with different average customer profiles. Three findings are most useful for deciding what to test next.

HIGH-INCOME CONTRAST

$82.1K / 54.4

CLUSTER 3 · INCOME / SCORE

$85.2K / 14.1

CLUSTER 1 · INCOME / SCORE

The relatively small income difference is accompanied by a much larger difference in the mall-assigned Spending Score.

HIGHEST SPENDING SCORE

70.2

Cluster 2 · 42 customers

Average age: 28.7. Average income: $60.9K. Average Spending Score: 70.2. This makes Cluster 2 a useful starting population for testing whether differentiated messaging or offers produce a measurable response.

NEXT DECISION

Test whether the clearest segment contrasts produce meaningfully different customer responses.

SEGMENT → TEST → MEASURE → VALIDATE

MANAGEMENT IMPLICATION

Keep the segmentation only if measured customer behavior shows that the groups respond differently in a useful way.

WHAT THE DATA DOES NOT ESTABLISH

The dataset has no revenue, profit, customer lifetime value, campaign response, purchase frequency, or product preference. The five clusters are testable customer groups, not finished marketing personas or proven strategies.

02 / BUSINESS PROBLEM

Business Problem & Analytical Approach

Customers with similar incomes can show very different spending profiles.

BUSINESS QUESTION

Can multiple customer characteristics identify groups worth treating differently in a marketing experiment?

The dataset contains 200 customer records with four attributes available for segmentation: Age, Annual Income, Spending Score, and Gender.

Spending Score is a 1-100 index assigned by the mall, rather than a monetary spending value.

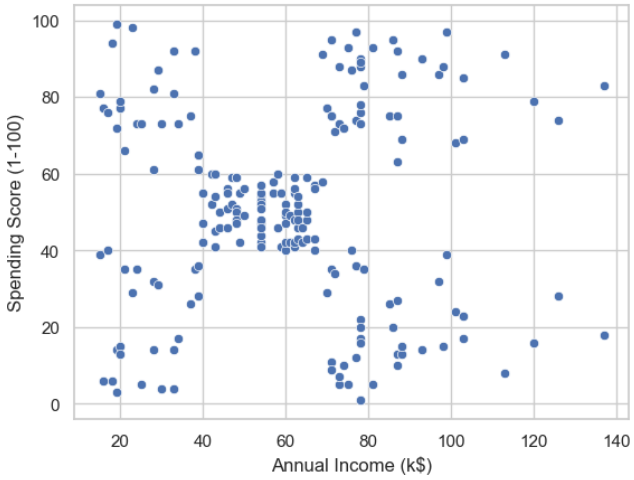


Exhibit 1 / Similar income levels contain very different Spending Scores.

ANALYTICAL APPROACH

1 Simple clustering

Income used by itself to understand the basic K-Means workflow.

2 Two-variable clustering

Annual Income + Spending Score used to examine visible customer differences.

3 Final multivariate model

Age + encoded Gender + Annual Income + Spending Score.

READING THE EXHIBIT

Customers at similar income levels appear across very different Spending Scores. Income alone provides an incomplete picture.

03 / MODEL CONSTRUCTION

Segmentation Methodology

The final model uses four standardized customer characteristics.

FINAL MODEL INPUTS

The final segmentation uses: Age, Annual Income, Spending Score, Encoded Gender. Gender was converted into a numeric dummy variable. The four inputs were then standardized using StandardScaler.

Standardization gives the inputs comparable scale before K-Means calculates distances between observations. Without that step, a feature with larger raw numeric values, such as annual income, could have disproportionate influence on the clustering.

SELECTING K

K-Means was tested across multiple values of K. The multivariate elbow curve was used to select K = 5 for the final model.

Five clusters should be treated as a practical model choice for this 200-customer sample. The elbow method does not establish that five permanent customer types exist in a broader population.

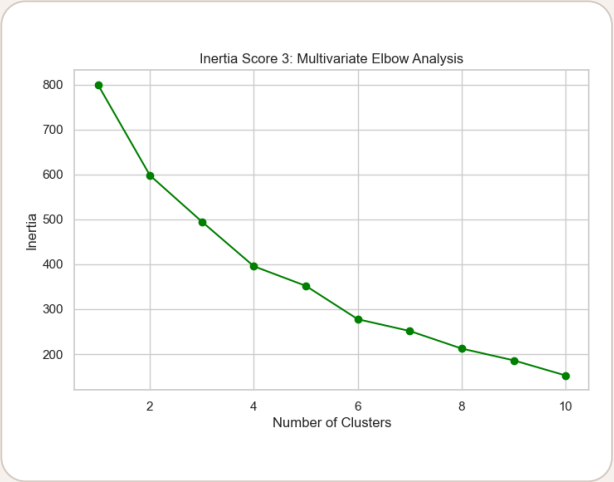
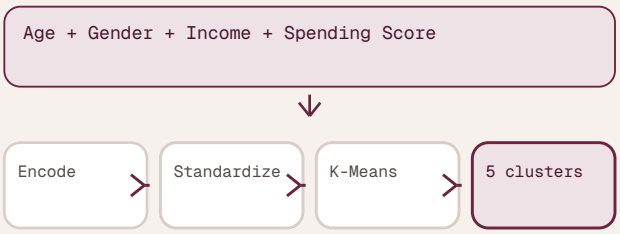


Exhibit 2 / Multivariate elbow curve used to select K = 5.

MODEL FLOW



SELECTION BOUNDARY

No silhouette-score or cluster-stability analysis was used in the final model selection documented here.

04 / PROFILE MATRIX

Segment Portfolio

The five clusters have distinct average profiles.

The rows below describe the final K = 5 model in the 200-customer sample. Cluster numbers are model identifiers, not customer persona names.

CLUSTER	DESCRIPTIVE SUMMARY	CUSTOMERS	AGE	INCOME	SCORE
2	Younger · mid-income · highest Spending Score	42	28.7	$60.9K	70.2
4	Younger · lower-income · mid-high Spending Score	38	27.3	$38.8K	56.2
3	Higher-income · mid-range Spending Score	49	37.9	$82.1K	54.4
0	Older · moderate-income · lower Spending Score	51	56.5	$46.1K	39.3
1	Higher-income · lowest Spending Score	20	39.5	$85.2K	14.1

WORKING DESCRIPTIONS

These labels stay descriptive. They do not infer motivations, commercial value, or permanent customer types.

05 / KEY CUSTOMER FINDINGS

Key Customer Findings

The strongest contrast appears between two higher-income groups.

Cluster 3 averages $82.1K income and 54.4 Spending Score. Cluster 1 averages $85.2K income and 14.1 Spending Score.

CLUSTER 3

$82.1K

AVERAGE ANNUAL INCOME

54.4 AVERAGE SPENDING SCORE

Higher-income · mid-range Spending Score

CLUSTER 1

$85.2K

AVERAGE ANNUAL INCOME

14.1 AVERAGE SPENDING SCORE

Higher-income · lowest Spending Score

ABOUT $3.1K INCOME GAP

> 40 SCORE POINTS

Their average incomes differ by only about $3.1K, while their average Spending Scores differ by more than 40 points.

BUSINESS IMPLICATION

Income alone would place these customers relatively close together. The multivariate segmentation separates groups whose observed profiles differ substantially on Spending Score. The dataset cannot explain the cause of that difference. Additional behavioral data would be required to determine whether product preference, visit frequency, purchase history, customer tenure, or another factor explains the contrast.

CLUSTER 2 / INITIAL TEST POPULATION

Cluster 2 contains 42 customers with the highest average Spending Score, 70.2, and an average income of $60.9K. This makes the group useful for testing whether a differentiated treatment produces a measurable change in behavior. It does not establish that Cluster 2 generates the most revenue or profit.

06 / NEXT TESTS

Strategic Recommendations

The five clusters are useful only if management tests whether they predict different responses.

01

Start with the clearest segment contrasts

Use Cluster 2 and Cluster 1 as initial testing populations. Cluster 2 provides the highest average Spending Score. Cluster 1 combines relatively high income with the lowest average Spending Score. These differences provide a reasonable basis for testing different treatments.

02

Test treatments before assigning permanent personas

Use defined messages, offers, or experiences and measure whether customer response differs across segments. Possible outcomes include response rate, conversion, visit frequency, basket value. These metrics are not present in the current dataset and would need to be collected.

03

Add behavioral variables to future segmentation

A stronger customer model could incorporate transaction history, recency, purchase frequency, product/category behavior, campaign response. These variables would connect customer profiles to actual commercial behavior.

04

Validate the cluster structure before operational use

Re-run the segmentation on a larger customer population. Use additional validation techniques beyond the elbow method and check whether the cluster profiles remain stable.

MEASUREMENT FRAMEWORK

MANAGEMENT CONCLUSION

The five clusters provide hypotheses for customer strategy. Measured customer behavior determines whether they become useful business segments.

APPENDIX A / MODEL METHODOLOGY

Appendix A - Model Methodology & Definitions

ANALYSIS POPULATION

200 customer records. Each record contains: CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100). The final labeled dataset also preserves intermediate cluster assignments and the final Multivariate_Cluster value.

SPENDING SCORE

Spending Score is a 1-100 index contained in the source dataset. It is not a dollar amount. A difference of ten Spending Score points therefore represents a difference in the source index, not a known dollar difference in customer spending.

FINAL MODEL INPUTS

The final K-Means model uses: Age, encoded Gender, Annual Income, Spending Score.

GENDER ENCODING

The categorical Gender variable was converted into a numeric dummy variable using pd.get_dummies before inclusion in the final model.

FEATURE STANDARDIZATION

The four final inputs were transformed using scikit-learn's StandardScaler. For each feature, standardization centers the values around the feature mean and scales them relative to its standard deviation. The purpose is to prevent the raw numerical scale of one feature from dominating the Euclidean-distance calculation used by K-Means.

K-MEANS

K-Means partitions observations into a selected number of clusters by iteratively assigning observations to cluster centers and updating those centers. The final analysis uses K = 5.

MODEL SELECTION

The project used the elbow method to inspect within-cluster variation across candidate values of K. The multivariate elbow curve supported five clusters as the working model. The elbow method was the primary cluster-count diagnostic in this project. No silhouette-score or cluster-stability analysis was used as part of the final selection documented here.

> **APPENDIX NOTE**
>
> Definitions in this appendix keep the model's inputs, selection rule, and interpretation boundary visible alongside the executive findings.

APPENDIX B / INTERPRETATION LIMITS

Appendix B - Supporting Analysis & Interpretation Limits

EARLIER CLUSTERING STAGES

The analysis progressed through: income-only clustering, followed by Annual Income + Spending Score, followed by the final Age + Gender + Income + Spending Score model. The earlier models were exploratory steps and should not be confused with the final five-cluster multivariate segmentation.

SCORE-TO-INCOME INDEX

The notebook calculates Spending Score ÷ Annual Income. The final CSV preserves this value as Spend_Income_Ratio. Cluster 4 had the highest average value reported during the project: 1.85. This value is best described as an exploratory score-to-income index. Spending Score is an index rather than dollars spent. The value therefore does not represent a percentage of income spent, a financial result, return on investment, profitability, or customer value.

WHAT THE SOURCE DATA DOES NOT CONTAIN

The dataset does not provide: revenue, profit, customer lifetime value, campaign response, purchase frequency, basket value, product preference, reason for Spending Score. The clustering cannot directly support claims about those outcomes.

SEGMENT LABELS

Cluster numbers are model identifiers. The report uses descriptive summaries based on the observed means rather than behavioral persona names. For example: Cluster 1: Higher-income · lowest Spending Score rather than an inferred motivation. This avoids assigning motivations or commercial value that the dataset does not measure.

VALIDATION BOUNDARY

The segmentation describes this 200-customer sample. Before operational use, a stronger validation process would include: a larger customer population, additional behavioral variables, more than one cluster-quality measure, cluster-stability testing, measured customer-response outcomes.

> INTERPRETATION NOTE
>
> 1.85 is a secondary exploratory score-to-income index. It is not a spending amount, percentage of income, financial result, or customer value measure.