MALL CUSTOMER SEGMENTATION

Project Summary Report

PAGE 1: EXECUTIVE SUMMARY

Project Overview

Objective: Segment mall customers into distinct groups to enable targeted marketing strategies

Dataset: Mall Customer Segmentation Data (Kaggle)

- 200 customers
- 5 features: CustomerID, Gender, Age, Annual Income, Spending Score

Methodology: K-Means Clustering with optimal K determination

Key Result: Successfully identified **5 distinct customer segments** with actionable business insights

---

Key Findings

Model Performance

- **Optimal K:** 5 clusters
- **Silhouette Score:** 0.555 (indicates good cluster separation)
- **Davies-Bouldin Index:** 0.572 (low score = better clustering)

Customer Distribution

Cluster	Size	Percentage	Profile
0	39	19.5%	Low Income, Low Spenders
1	35	17.5%	High Income, Low Spenders
2	46	23.0%	Low Income, High Spenders
3	40	20.0%	High Income, High Spenders
4	40	20.0%	Moderate Spenders

Business Impact

- **20% of customers** (Cluster 3) are high-value VIP segment
- **23% of customers** (Cluster 2) are impulse buyers needing payment plans
- **Clear marketing strategies** identified for each segment
- **Income and spending are independent** (correlation: ~0.01), validating segmentation approach

---

Deliverables

1. **Data Exploration Notebook** - Comprehensive EDA with visualizations
2. **Model Training Notebook** - K-Means clustering (K=5) + DBSCAN
3. **Streamlit Dashboard** - 5-page interactive application
4. **Clustered Dataset** - Customer data with segment labels

Task Completion

Core Requirements

- Cluster customers by income and spending score
- Perform feature scaling
- Visual exploration of data
- K-Means clustering with optimal K determination
- 2D cluster visualizations

Bonus Tasks

- DBSCAN algorithm implementation
- Average spending analysis per cluster

PAGE 2: METHODOLOGY & DATA ANALYSIS

Data Overview

Dataset Characteristics

- **Total Customers:** 200
- **Features:** 5 (CustomerID, Gender, Age, Annual Income, Spending Score)
- **Missing Values:** 0 (clean dataset)
- **Duplicates:** 0

Feature Statistics

Feature	Mean	Min	Max	Std Dev
Age	38.9	18	70	14.0
Annual Income (k$)	60.6	15	137	26.3
Spending Score	50.2	1	99	25.8

Gender Distribution

- **Female:** 112 customers (56%)
- **Male:** 88 customers (44%)

Exploratory Data Analysis

Key Insights

1. **Age Distribution**
   - Normally distributed around mean of 39 years
   - Wide range (18-70) indicates diverse customer base
2. **Income Distribution**
   - Mean: $60.6k, Median: $61.5k
   - Range: $15k-$137k
   - Relatively uniform distribution
3. **Spending Score Distribution**
   - Mean: 50.2, Median: 50.0
   - Uniform distribution across range 1-99
   - No clear peaks (good for clustering)
4. **Correlation Analysis**
   - Age vs Income: 0.002 (no correlation)
   - Age vs Spending: -0.327 (weak negative)
   - **Income vs Spending: 0.009 (virtually independent)**
   - This independence validates using both features for clustering

---

Clustering Methodology

Feature Selection

Selected Features: Annual Income (k$) and Spending Score (1-100)

Rationale:

- Both are continuous variables
- Virtually independent (good for creating distinct segments)
- Business-relevant for marketing strategies

Preprocessing

1. **Feature Scaling:** StandardScaler applied
   - Ensures equal weight to both features
   - Required for distance-based algorithms

Algorithm Selection

Primary: K-Means Clustering

- Partition-based algorithm
- Works well with continuous data
- Scalable and efficient

Bonus: DBSCAN (Density-Based Spatial Clustering)

- Alternative approach for comparison

- Identifies outliers/noise points

Optimal K Determination

Evaluation Metrics Used

1. **Elbow Method** - Inertia (within-cluster sum of squares)
2. **Silhouette Score** - Measures cluster separation (-1 to 1, higher is better)
3. **Davies-Bouldin Index** - Measures cluster similarity (lower is better)

Results (K=2 to K=10)

K	Inertia	Silhouette	Davies-Bouldin	Decision
2	269.69	0.321	1.267	Too simple
3	157.70	0.467	0.716	Suboptimal
4	108.92	0.494	0.710	Good
5	**65.57**	**0.555**	**0.572**	**OPTIMAL**
6	55.06	0.540	0.655	Declining
7	44.86	0.528	0.715	Overfitting
8-10	<38	<0.46	>0.76	Poor

Selection Rationale

K=5 selected because:

- Highest Silhouette Score (0.555)
- Lowest Davies-Bouldin Index (0.572)
- Clear elbow point in inertia plot
- Balanced cluster sizes
- Business interpretability

PAGE 3: CLUSTERING RESULTS & SEGMENTS

Cluster Profiles

Cluster 0: Low Income, Low Spenders

Size: 39 customers (19.5%)

Metric	Value
Avg Age	42.7 years
Avg Income	$26.3k
Avg Spending Score	20.9
Income Range	$15k - $40k
Spending Range	1 - 40

Characteristics:

- Budget-conscious customers
- Price-sensitive segment
- Focus on necessities over luxuries
- Low purchasing power

Cluster 1: High Income, Low Spenders

Size: 35 customers (17.5%)

Metric	Value
Avg Age	41.1 years
Avg Income	$88.2k
Avg Spending Score	17.1
Income Range	$60k - $137k
Spending Range	1 - 40

Characteristics:

- Wealthy but selective buyers
- High savings tendency
- Quality over quantity mindset
- Research before purchasing

Cluster 2: Low Income, High Spenders

Size: 46 customers (23.0%)

Metric	Value
Avg Age	25.3 years
Avg Income	$25.7k
Avg Spending Score	79.3
Income Range	$15k - $40k
Spending Range	60 - 99

Characteristics:

- Enthusiastic shoppers
- Impulse buying tendencies
- Trend-focused consumers
- Despite limited income, high spending frequency

Cluster 3: High Income, High Spenders

Size: 40 customers (20.0%)

Metric	Value
Avg Age	32.7 years
Avg Income	$87.8k
Avg Spending Score	82.1
Income Range	$60k - $137k
Spending Range	60 - 99

Characteristics:

- Premium VIP customers (MOST VALUABLE)
- High disposable income
- Frequent high-value purchases
- Brand loyal

Cluster 4: Moderate Income, Moderate Spenders

Size: 40 customers (20.0%)

Metric	Value
Avg Age	45.2 years
Avg Income	$55.3k
Avg Spending Score	49.5
Income Range	$40k - $70k
Spending Range	40 - 60

Characteristics:

- Balanced middle-market segment
- Pragmatic decision makers
- Value quality-price balance
- Moderate purchasing frequency

Visual Analysis

Cluster Separation

- **2D Scatter Plot:** Clear separation between clusters in Income-Spending space
- **No Overlap:** Minimal overlap between adjacent clusters
- **Centroid Distance:** Centroids are well-separated

Cluster Characteristics

- **Age Distribution:** Similar across most clusters (except Cluster 2 is younger)
- **Income Split:** Clear high/low income divide
- **Spending Split:** Clear high/moderate/low spending divide
- **Gender:** Relatively balanced across all clusters

Model Validation

Silhouette Analysis

- **Overall Score:** 0.555 (good separation)
- **Per-Cluster Scores:** All clusters above 0.4 (acceptable threshold)
- **No negative scores:** No misclassified points

Davies-Bouldin Index

- **Score:** 0.572 (low is good)
- Indicates clusters are compact and well-separated

Business Validation

- All clusters have clear business meaning
- Segments are actionable for marketing
- Size distribution is balanced (17-23%)
- Each segment has distinct characteristics

PAGE 4: BUSINESS RECOMMENDATIONS

Marketing Strategies by Segment

Cluster 0: Low Income, Low Spenders

Target: Budget-conscious customers (19.5%)

Marketing Strategy:

1. **Pricing:**
    - Discount programs (10-30% off)
    - Loyalty rewards and points
    - Bundle deals and bulk discounts
2. **Products:**
    - Value-based offerings
    - Generic/store brands
    - Essential items focus

3. **Communication:**
    - "Save Money" messaging
    - Price comparison campaigns
    - Budget-friendly tips
4. **Channels:**
    - Email newsletters with deals
    - SMS alerts for flash sales
    - In-store signage

Expected ROI: Low per-customer, high volume

Cluster 1: High Income, Low Spenders

Target: Selective wealthy customers (17.5%)

Marketing Strategy:

1. **Positioning:**
    - Premium quality emphasis
    - Durability and longevity
    - Investment value messaging
2. **Products:**
    - High-end product lines
    - Extended warranties
    - Professional/business items
3. **Communication:**
    - "Quality Investment" messaging
    - Expert reviews and testimonials
    - Educational content
4. **Engagement:**
    - Exclusive membership tiers
    - Private shopping events
    - Concierge services

Expected ROI: High per-transaction, low frequency

Cluster 2: Low Income, High Spenders

Target: Enthusiastic shoppers (23.0%)

Marketing Strategy:

1. **Payment Solutions:**
    - Buy Now Pay Later (BNPL)
    - Installment plans (3-12 months)
    - Low-interest financing
2. **Products:**
    - Trendy, fashion-forward items
    - Affordable luxury alternatives
    - Limited edition releases
3. **Communication:**

    - "Treat Yourself" messaging
    - FOMO-driven campaigns
    - Social media influencer partnerships
4. **Tactics:**
    - Flash sales and limited offers
    - Early access to new arrivals
    - Gamified shopping experiences

Expected ROI: Medium per-customer, high frequency

Cluster 3: High Income, High Spenders

Target: VIP premium customers (20.0%) - HIGHEST VALUE

Marketing Strategy:

1. **VIP Treatment:**
    - Dedicated account managers
    - Priority customer service
    - Exclusive shopping appointments
2. **Products:**
    - Luxury product lines
    - Designer collaborations
    - Limited edition exclusives
3. **Perks:**
    - Early/exclusive access
    - Complimentary services
    - Personal styling/consultation
4. **Retention:**
    - Ultra-premium loyalty program
    - Invitation-only events
    - Personalized recommendations

Expected ROI: HIGHEST per-customer and lifetime value

Cluster 4: Moderate Spenders

Target: Middle-market balance (20.0%)

Marketing Strategy:

1. **Balanced Approach:**
    - Seasonal promotions
    - Mid-tier product focus
    - Quality-value balance
2. **Products:**
    - Mid-range brands
    - Good-better-best options
    - Reliable everyday items
3. **Communication:**
    - "Smart Choice" messaging

    - Value + quality emphasis
    - Family-oriented campaigns
4. **Loyalty:**
    - Standard loyalty program
    - Birthday/anniversary rewards
    - Referral bonuses

Expected ROI: Steady medium returns, reliable segment

Resource Allocation

Priority Ranking

1. **Cluster 3** (VIP) - 35% of marketing budget
2. **Cluster 2** (High Spenders) - 25% of marketing budget
3. **Cluster 4** (Moderate) - 20% of marketing budget
4. **Cluster 0** (Budget) - 15% of marketing budget
5. **Cluster 1** (elective) - 5% of marketing budget (low maintenance)

Cross-Selling Opportunities

- **Cluster 2 → Cluster 3:** Upsell with credit/payment plans
- **Cluster 0 → Cluster 4:** Graduate customers with income growth
- **Cluster 4 → Cluster 3:** Premium product introductions

Key Performance Indicators

By Cluster

- **Cluster 0:** Basket size increase, repeat visit frequency
- **Cluster 1:** Average transaction value, product quality ratings
- **Cluster 2:** Conversion rate, BNPL adoption rate
- **Cluster 3:** Lifetime value, retention rate, NPS score
- **Cluster 4:** Overall satisfaction, referral rate

Overall Metrics

- Customer retention rate by segment
- Revenue per segment
- Marketing ROI by segment
- Cross-segment migration rates

PAGE 5: DASHBOARD & CONCLUSIONS

Streamlit Dashboard

Application Features

5 Interactive Pages:

1. **Overview**
    - Real-time metrics dashboard
    - Dataset preview and statistics
    - Project methodology
    - K=5 selection rationale
2. **Data Exploration**
    - Distribution plots (Age, Income, Spending, Gender)
    - Interactive scatter plots
    - Correlation heatmap
    - Statistical insights
3. **Clustering Results**
    - 5-cluster visualization
    - Cluster size distribution
    - Characteristics table
    - Box plots by cluster
    - Model performance metrics
4. **Customer Insights**
    - Average spending by cluster
    - Detailed cluster selector
    - Marketing strategy recommendations
    - Customer demographics
    - Business interpretations
5. **Predict Cluster**
    - Interactive customer classification
    - Real-time prediction tool
    - Visual position in cluster space
    - Similar customer finder
    - Marketing recommendations

Technical Implementation

- Built with Streamlit framework
- Interactive Plotly visualizations
- Real-time K-Means predictions
- Responsive design
- Easy-to-use interface

Project Conclusions

Technical Success

Data Quality: Clean dataset with no missing values or duplicates
Model Performance: Silhouette Score of 0.555 indicates good clustering
Optimal K: K=5 outperformed all other values (K=2 to K=10)
Validation: Multiple metrics confirm cluster quality
Visualization: Clear separation visible in 2D plots

Business Value

Actionable Segments: 5 distinct customer profiles identified
Marketing Strategies: Specific tactics for each segment
Resource Allocation: Clear priority ranking established
High-Value Identification: 20% VIP customers identified (Cluster 3)
Growth Opportunities: Payment plan potential for 23% (Cluster 2)

Key Insights

1. **Income ≠ Spending:** Weak correlation validates two-feature approach
2. **Balanced Distribution:** No dominant cluster (17-23% each)
3. **Age Independence:** Spending behavior not age-dependent
4. **Clear Segments:** No ambiguous or overlapping clusters

Bonus Achievements

DBSCAN Analysis

- Alternative clustering approach implemented
- Identified outliers and noise points
- Validated K-Means results
- Provided additional perspective on data structure

Average Spending Analysis

- Detailed per-cluster spending metrics
- Statistical breakdown (mean, median, std)
- Visualization of spending patterns
- Business implications documented

Limitations & Future Work

Current Limitations

- Limited to 2 features (Income, Spending)
- No temporal data (shopping frequency, seasonality)
- No product category information
- Small dataset (200 customers)

Future Enhancements

1. **Additional Features:**
   - Shopping frequency
   - Product categories purchased
   - Time of year preferences
   - Online vs in-store behavior
2. **Advanced Modeling:**
   - Hierarchical clustering
   - Customer lifetime value prediction
   - Churn prediction
   - Recommendation systems
3. **Business Integration:**
   - CRM system integration
   - Real-time segmentation API
   - A/B testing framework
   - Automated campaign triggers
4. **Deployment:**
   - Cloud deployment (AWS, Azure, GCP)
   - Mobile app integration
   - Email automation connection
   - Sales team dashboard

Final Recommendations

Immediate Actions

1. **Deploy dashboard** for marketing team use
2. **Implement Cluster 3** VIP program immediately
3. **Launch BNPL** payment options for Cluster 2
4. **Create segment-specific** email campaigns
5. **Train sales team** on cluster characteristics

6-Month Goals

- Increase Cluster 3 retention by 15%
- Grow Cluster 2 conversion rate by 20%
- Migrate 10% of Cluster 0 to Cluster 4
- Achieve 25% revenue increase from targeted campaigns

Success Metrics

- Monitor KPIs per cluster monthly
- Track cross-segment migration
- Measure marketing ROI by segment
- Customer satisfaction scores by cluster

Project Summary

Mission Accomplished: Successfully segmented 200 mall customers into 5 actionable groups using K-Means clustering with optimal performance metrics.

Deliverables Complete:

- 2 Jupyter Notebooks (EDA + Modeling)
- Interactive Streamlit Dashboard
- Comprehensive Documentation
- Bonus Tasks Completed

Business Impact: Clear, data-driven customer segmentation strategy ready for implementation with specific marketing tactics for each of the 5 segments.
