Assignment: MegaMart Customer Segmentation Analysis (Full Business Case)

1. Project Objective

You are a data analytics consultant for **ShopSmart Analytics**, hired by **MegaMart Retail**, a national retail chain. MegaMart has collected extensive customer behavioral data but lacks a structured understanding of their customer base.

Your objective is to analyze the provided retail_customer_data.csv to **discover natural customer segments** using unsupervised clustering techniques. Unlike classification problems where groups are predefined, you must identify meaningful patterns in customer behavior without any labeled categories.

You will compare **Hierarchical Clustering** (with multiple linkage methods) and **K-Means Clustering** to determine the optimal number and composition of customer segments.

Your final submission will be a complete consulting engagement, consisting of your technical analysis, a non-technical business report, and a stakeholder presentation.

2. The Business Case

MegaMart faces several strategic challenges:

- **Generic Marketing**: Current campaigns treat all customers the same, resulting in low engagement rates
- **Resource Waste**: Marketing budget is spread evenly without targeting high-value segments
- **Missed Opportunities**: Lack of understanding about customer needs prevents personalized experiences
- **Churn Risk**: Inability to identify at-risk segments before they leave

Your clustering analysis will provide the foundation for:

1. **Targeted Marketing Campaigns**: Segment-specific messaging and offers
2. **Resource Prioritization**: Focus retention efforts on high-value customers

3. **Personalized Experiences**: Tailored product recommendations and customer journeys
4. **Strategic Planning**: Data-driven decisions about product mix and pricing

Critical Note: This is **unsupervised learning**. There are no "correct" cluster labels in the dataset. Your job is to discover meaningful patterns and justify your segment definitions.

3. Deliverables

You will submit three distinct deliverables:

IMPORTANT: All documents must include in the header or first page:

- Team number (as assigned by the instructor)
- Names of all team members

1. **Deliverable 1: The Code (MegaMart_Segmentation.ipynb)**
    - A single Jupyter Notebook containing your complete technical analysis, all code, visualizations, and technical interpretations. This is your "analytical appendix."
    - Include team number and member names in the first markdown cell.
2. **Deliverable 2: The Business Report (Executive_Summary.pdf)**
    - A 2-3 page, code-free document written for a non-technical executive. It must summarize the business problem, your discovered segments, and marketing recommendations.
    - Include team number and member names in the document header.
    - **Must include the link to your video presentation** in the document (e.g., in the footer or references section).
3. **Deliverable 3: The Pitch (Video & Slides)**
    - **Slides (Presentation.pdf or .pptx):** A 6-8 slide deck that visually supports your presentation. Include team number and member names on the title slide.
    - **Video (Presentation.mp4 or link):** A 7-10 minute recorded video presentation "pitching" your segmentation findings to the MegaMart executive team.

4. Step-by-Step Instructions

Part A: The Jupyter Notebook (Your Analytical Work)

Your notebook should be well-organized and tell a clear story, similar to the customer_clustering_analysis.ipynb example. Use Markdown cells extensively to explain your steps and findings.

Section 1: Project Setup & Data Loading

- **First Markdown Cell**: Create a title cell with the project title, team number (as assigned by instructor), and names of all team members.
- Import all necessary libraries (pandas, numpy, sklearn, scipy, matplotlib, seaborn).
- Load the retail_customer_data.csv file.
- Perform an initial inspection: .head(), .info(), .describe().
- Write a brief summary of your initial findings (e.g., "3,000 customers," "9 behavioral variables," "No missing values").

Section 2: Exploratory Data Analysis (EDA)

- **Variable Distributions**: Create histograms or KDE plots for each of the 9 behavioral variables.
- **Correlation Analysis**: Generate a correlation matrix heatmap. Note any highly correlated variables.
- **Outlier Detection**: Use box plots to identify potential outliers in key variables like total_spend or monthly_transactions.
- **Scatter Plots**: Create 2-3 scatter plots showing relationships between key variables (e.g., total_spend vs monthly_transactions, colored by other variables if helpful).

Section 3: Data Preprocessing

- **Check for Missing Data**: Verify there are no missing values (there should be none).
- **Standardization**: Initialize a StandardScaler and fit_transform the entire dataset. This is **critical** for clustering because distance-based algorithms are sensitive to variable scales.
- **Explain your choice**: In a Markdown cell, explain why standardization is necessary for clustering (e.g., "Without standardization, variables with larger ranges like total_spend would dominate the distance calculations").

Section 4: Hierarchical Clustering Analysis

- **Compute Linkage Matrices**: Calculate hierarchical clustering using all four linkage methods: 'single', 'complete', 'average', and 'ward'.
- **Create Dendrograms**: Plot dendrograms for each linkage method (2x2 subplot grid).
- **Interpret Dendrograms**: In a Markdown cell, discuss:
    - Which linkage method appears most suitable? (Hint: Ward's typically performs best for customer segmentation)
    - Where would you cut the dendrogram? What does the height of merges tell you?
    - What is the "chaining effect" and which linkage method is most susceptible to it?

Section 5: Determining Optimal Number of Clusters (Hierarchical)

- **Focused Dendrogram**: Create a detailed dendrogram using Ward's linkage with a horizontal line showing potential cuts.
- **Extract Clusters**: Use scipy.cluster.hierarchy.fcluster to extract clusters for k = 3, 4, 5, and 6.
- **Calculate Silhouette Scores**: For each k, calculate the silhouette_score on the standardized data.
- **Create a Table**: Display a summary table showing k (number of clusters) vs silhouette score for hierarchical clustering.

Section 6: K-Means Clustering - Elbow Method

- **Inertia Calculation**: Run K-Means for k = 2 through 10, storing the inertia (within-cluster sum of squares) for each k.
- **Elbow Plot**: Create a line plot of k vs inertia.
- **Silhouette Scores**: Also calculate silhouette scores for each k value.
- **Dual Plot**: Create a 1x2 subplot showing both the elbow plot and silhouette scores vs k.
- **Interpretation**: In a Markdown cell, identify where the "elbow" occurs. Does the silhouette score agree?

Section 7: Final Cluster Selection

- **Choose Optimal k**: Based on both hierarchical and k-means analyses, select your final number of clusters (e.g., k = 4 or k = 5).
- **Justify Your Choice**: Write a clear explanation citing:
    - Dendrogram structure (large vertical gaps)
    - Elbow plot inflection point

    - Silhouette scores
    - Business considerations (too few clusters = overgeneralization, too many = not actionable)

Section 8: Apply Final K-Means Model

- **Fit K-Means**: Fit a KMeans model with your chosen k on the standardized data.
- **Extract Cluster Labels**: Get the cluster assignments for all 3,000 customers.
- **Cluster Sizes**: Print the size (count and percentage) of each cluster.

Section 9: Cluster Profiling and Interpretation

- **Add Clusters to Original Data**: Merge cluster labels with the original (unstandardized) data.
- **Calculate Cluster Means**: Group by cluster and calculate mean values for all 9 behavioral variables.
- **Heatmap**: Create a heatmap showing cluster profiles (clusters as columns, variables as rows).
- **Characterization**: For each cluster, write a 2-3 sentence profile identifying its distinctive characteristics:
    - Example: "Cluster 0 (15%): High-Value Loyalists - Very high total spend, frequent monthly transactions, low return rate, high email engagement. These are our best customers."

Section 10: Cluster Validation - Silhouette Analysis

- **Silhouette Plot**: Create a silhouette plot showing the silhouette coefficient for each customer, grouped by cluster.
- **Interpretation**: Discuss:
    - Which clusters are well-defined (high silhouette values)?
    - Are there any customers poorly matched to their cluster (negative silhouette)?
    - Does this validate your choice of k?

Section 11: Cluster Visualization (PCA Projection)

- **Apply PCA**: Use PCA to reduce the 9 dimensions to 2 principal components.
- **Scatter Plot**: Create a scatter plot of PC1 vs PC2, colored by cluster assignment.
- **Add Centroids**: Plot the cluster centroids as larger markers.

- **Variance Explained**: Report how much variance is captured by the first 2 PCs.
- **Note Limitation**: Explain that this 2D view is a projection and actual clusters exist in 9-dimensional space.

Section 12: Technical Conclusion

- **Summary**: Summarize your technical findings:
    - Optimal number of clusters and justification
    - Hierarchical vs K-means comparison
    - Quality metrics (silhouette scores)
    - Key characteristics of each discovered segment

Part B: The Executive Summary (Your Business Report)

Write a 2-3 page professional report. **This document must not contain any Python code or technical jargon.** It should be structured as follows:

Document Header (First Page):

- Team number (as assigned by instructor)
- Names of all team members
- Date
- **Video presentation link** (include the URL to your recorded presentation)

1. **Business Problem:**
    - State MegaMart's challenge (e.g., "MegaMart lacks customer segmentation, leading to generic marketing and resource waste").
    - State your project objective (e.g., "Our team analyzed 3,000 customers to discover natural behavioral segments").
2. **Discovered Customer Segments:**
    - Describe each segment in plain English with a business-friendly name:
        - Bad Example: "Cluster 2 has mean total_spend = 5847.32 and monthly_transactions = 11.7"
        - Good Example: "**High-Value Loyalists (15% of customers):** These customers shop frequently (nearly weekly), spend significantly above average, and have the highest engagement with email campaigns. They represent our most valuable segment."
    - Create a customer persona for each segment (name, brief description, key behaviors).

3. **Marketing Strategy Recommendations:**
    - For each segment, provide 2-3 specific, actionable marketing strategies:
        - Example for High-Value Loyalists: "Launch a VIP rewards program with exclusive early access to sales and personalized concierge service."
        - Example for Bargain Hunters: "Send weekly deal emails highlighting clearance items and bundle offers with free shipping thresholds."
    - Prioritize which segments should receive the most marketing resources and why.
4. **Expected Business Impact:**
    - Estimate the potential benefits of segment-based marketing:
        - "By targeting High-Value Loyalists with retention campaigns, we project a 25% reduction in churn among this critical segment."
        - "Personalized messaging to each segment is expected to increase email open rates by 40% and conversion rates by 20%."
    - Discuss trade-offs and implementation considerations.
5. **Next Steps & Recommendations:**
    - Recommend clear next steps (e.g., "Deploy segmentation model to production CRM system," "A/B test segment-specific campaigns").
    - Suggest future enhancements (e.g., "Incorporate product category preferences," "Track segment migration over time").

Part C: The Video Presentation (Your Stakeholder Pitch)

Record a **7-10 minute video** presenting your findings to the MegaMart executive team. Your accompanying slide deck should be 6-8 slides, clean, professional, and visual.

- **Slide 1: Title Page** (Project Title, Team Number, All Team Member Names, Date)
- **Slide 2: The Business Problem** (The challenge: No customer segmentation. The goal: Discover actionable segments. Use simple visuals and icons.)
- **Slide 3: Our Analytical Approach** (Brief overview: "We analyzed 3,000 customers across 9 behavioral metrics using advanced clustering techniques." Show a simple flowchart: Data -> Clustering -> Segments -> Strategies)

- **Slide 4-5: Meet Your Customer Segments** (For each segment, show:
    - Segment name and icon/image
    - % of customer base
    - 3-4 key characteristics in bullet points
    - Consider splitting across 2 slides if you have 4+ segments)
- **Slide 6: Marketing Strategy Overview** (High-level summary of your segment-specific recommendations. Use a table or matrix showing segment names vs key strategy themes)
- **Slide 7: Expected Business Impact** (Show projected benefits: "25% churn reduction among VIP segment," "40% increase in email engagement," etc. Use simple metrics and charts)
- **Slide 8: Next Steps** (Your recommendation: Deploy these segments, test campaigns, monitor results. Clear, actionable next steps.)

Presentation Tips:

- Speak clearly and confidently, as if presenting to actual executives
- Avoid technical jargon (no mention of "silhouette scores," "Ward's linkage," etc.)
- Use storytelling: "We discovered MegaMart has five distinct customer types..."
- Show enthusiasm for your findings
- Stay within the 7-10 minute time limit

5. Grading Rubric (100 Points Total)

Deliverable	Section	Criteria	Points
Deliverable 1: Jupyter Notebook (40 total)	EDA & Setup	Data loaded correctly. Comprehensive EDA with distributions, correlations, outlier analysis.	8
Deliverable 1: Jupyter Notebook	Preprocessing	**Standardization applied correctly** (StandardScaler). Clear justification for why standardization is necessary.	6
Deliverable 1: Jupyter Notebook	Hierarchical Clustering	Multiple linkage methods applied. Dendrograms plotted. Clear discussion of linkage method differences and chaining effect.	8

Deliverable	Section	Criteria	Points
Deliverable 1: Jupyter Notebook	K-Means & Optimal k	Elbow method and silhouette analysis applied. **Well-justified selection of optimal k** based on multiple criteria (not just one metric).	8
Deliverable 1: Jupyter Notebook	Cluster Profiling & Validation	Cluster means calculated. Heatmap created. Silhouette plot generated. PCA visualization included. **Clear cluster characterization** in plain language.	10
Deliverable 2: Executive Summary (30 total)	Clarity & Professionalism	2-3 pages, **no code**, well-formatted, free of technical jargon and typos.	5
Deliverable 2: Executive Summary	Problem & Objective	Business problem and project goal clearly stated in non-technical terms.	5
Deliverable 2: Executive Summary	Segment Descriptions	**Each discovered segment has a business-friendly name and clear, plain-English description** (not just cluster statistics). Customer personas are compelling.	10
Deliverable 2: Executive Summary	Marketing Strategies	**Specific, actionable, creative marketing recommendations for each segment**. Strategies are realistic and address segment needs.	10
Deliverable 3: Video & Slides (30 total)	Slide Deck	Slides are professional, **visual (not text-heavy), and support the narrative** effectively.	10
Deliverable 3: Video & Slides	Presentation Quality	Clear, confident, and persuasive delivery. Stays within the 7-10 minute time limit. Good pacing and energy.	10

Deliverable	Section	Criteria	Points
Deliverable 3: Video & Slides	Content & Storytelling	Presentation tells a compelling story (Problem -> Segments -> Strategies -> Impact). **Successfully translates technical findings** for a non-technical audience. Segments are brought to life.	10
Total			**100**

6. Key Success Factors

Do:

- Standardize your data before clustering
- Try multiple values of k and justify your final choice
- Give each segment a memorable, business-friendly name
- Make your marketing strategies specific and actionable
- Focus on storytelling in your presentation

Don't:

- Skip standardization (major error)
- Choose k arbitrarily without justification
- Use technical jargon in the business report or presentation
- Create too many clusters that aren't actionable
- Ignore silhouette scores if they indicate poor clustering

7. Deliverable Checklist

Before submission, verify:

Identification Requirements:

◯ Team number (assigned by instructor) is included in all three deliverables
◯ All team member names are listed in all three deliverables
◯ Video presentation link is included in the Executive Summary document

Technical Requirements:

◯ Jupyter Notebook runs from top to bottom without errors

- ◯ All 9 sections in the notebook are complete with code AND markdown explanations
- ◯ Data is standardized before clustering
- ◯ Multiple linkage methods are compared for hierarchical clustering
- ◯ Elbow method AND silhouette analysis are used to select optimal k
- ◯ Each cluster has a clear profile and business-friendly name

Business Report Requirements:

- ◯ Executive Summary is 2-3 pages, no code, professional formatting
- ◯ Marketing strategies are specific and actionable for each segment

Presentation Requirements:

- ◯ Presentation slides are visual and support the narrative
- ◯ Video presentation is 7-10 minutes, clear, and persuasive
- ◯ All file names follow the convention: [TeamName]_[DeliverableName].[ext]

8. Recommended Timeline

Days 1-3: Data exploration and hierarchical clustering

- Load data, perform EDA
- Standardize data
- Run hierarchical clustering with multiple linkage methods

Days 4-6: K-means and optimal cluster selection

- Implement elbow method and silhouette analysis
- Compare hierarchical vs k-means results
- Select final number of clusters

Days 7-9: Cluster profiling and business interpretation

- Calculate cluster characteristics
- Develop customer personas
- Create cluster visualizations

Days 10-12: Marketing strategy development

- Research retail marketing best practices
- Develop segment-specific strategies

- Draft business impact projections

Days 13-14: Deliverable finalization

- Complete Jupyter Notebook with final polish
- Write Executive Summary
- Create presentation slides and record video
- Peer review and final edits

Good luck! Remember: There is no single "correct" answer in clustering. You will be evaluated on the rigor of your analysis, the quality of your justifications, and the actionability of your business recommendations.