**CUSTOMER SEGMENTATION USING DATA SCIENCE**

**Introduction:**

Customer segmentation is a fundamental practice in marketing and business strategy that

involves categorizing a company's diverse customer base into distinct groups or segments

based on shared characteristics, behaviors, and preferences. The primary goal of customer

segmentation is to better understand customers, target them more effectively, and tailor

marketing efforts, products, and services to meet their specific needs.In an era of increased

market competition and the availability of vast amounts of customer data, segmentation has

become a crucial tool for businesses seeking to gain a competitive edge. By dividing the

customer base into identifiable segments, organizations can create more personalized and

relevant marketing campaigns, leading to higher customer satisfaction and increased

profitability.

**Problem Definition:**

Defining the problem for a customer segmentation project is a crucial first step in ensuring its success. Here's how you can formulate a problem statement:

\*\*Problem Statement:\*\*

"Develop a customer segmentation strategy to improve marketing and service personalization, thereby enhancing customer satisfaction and business performance."

\*\*Key Elements in the Problem Statement:\*\*

1. \*\*Objective:\*\* Start by clearly stating the primary objective of the project, which, in this case, is to create a customer segmentation strategy.

2. \*\*Benefit:\*\* Explain why this project is essential. In this case, it's to improve marketing and service personalization.

3. \*\*Outcomes:\*\* Describe what you aim to achieve with this project - in this case, enhancing customer satisfaction and business performance.

\*\*Additional Considerations:\*\*

1. \*\*Data Availability:\*\* Mention whether the required data is available and in what format (structured, unstructured, etc.).

2. \*\*Scope:\*\* Define the boundaries of the project. For example, specify the target customer group, geographic location, or specific business units involved.

3. \*\*Metrics:\*\* State the key performance indicators (KPIs) that will be used to measure success. In the context of customer segmentation, this could include metrics like increased conversion rates, reduced customer churn, or improved customer feedback scores.

4. \*\*Constraints:\*\* Identify any limitations or constraints that may affect the project, such as budget, timeline, or regulatory compliance.

5. \*\*Ethical Considerations:\*\* Highlight the importance of handling customer data responsibly and ethically, ensuring compliance with data protection laws and privacy regulations.

6. \*\*Stakeholders:\*\* List the key stakeholders involved in the project, including the marketing team, data scientists, IT department, and management.

By defining the problem clearly, you provide a foundation for the project's direction, which makes it easier to plan and execute the customer segmentation initiative effectively.

**Phases of project development:**

Designing a machine learning-based customer segmentation strategy involves several key steps:

1. \*\*Data Collection\*\*: Gather relevant data from various sources, such as customer demographics, purchase history, online behavior, and more.

2. \*\*Data Preprocessing\*\*: Clean, preprocess, and normalize the data to ensure it's ready for machine learning algorithms.

3. \*\*Feature Engineering\*\*: Create meaningful features that can help differentiate customers, such as RFM (Recency, Frequency, Monetary) scores, customer lifetime value, and customer behavior attributes.

4. \*\*Algorithm Selection\*\*: Choose appropriate machine learning algorithms for customer segmentation. Common choices include k-means clustering, hierarchical clustering, DBSCAN, or more advanced techniques like Gaussian Mixture Models or t-SNE.

5. \*\*Training the Model\*\*: Use historical data to train the selected machine learning model to segment customers.

6. \*\*Validation and Testing\*\*: Validate the model's performance using metrics like silhouette score, Davies-Bouldin index, or visual inspections of clustering results.

7. \*\*Segmentation\*\*: Apply the trained model to segment customers into distinct groups based on their attributes and behavior.

8. \*\*Interpretation\*\*: Understand what each segment represents in terms of customer characteristics and behaviors. This step can involve feature importance analysis and visualization.

9. \*\*Innovation and Personalization\*\*: Develop innovative strategies for each customer segment. Tailor marketing, product offerings, and customer experiences to meet the unique needs and preferences of each segment.

10. \*\*Feedback Loop\*\*: Continuously monitor and refine the segmentation and personalization strategies based on customer feedback and changing market conditions.

11. \*\*Compliance and Privacy\*\*: Ensure that your segmentation process complies with data privacy regulations and ethical considerations.

To enhance innovation in customer segmentation:

- Leverage deep learning and neural networks for more sophisticated patterns.

- Incorporate natural language processing for sentiment analysis of customer feedback.

- Implement reinforcement learning to optimize marketing strategies dynamically.

- Explore novel data sources, such as social media or IoT data, to gain deeper insights into customer behavior.

- Consider using AI-driven chatbots or recommendation engines to enhance customer interactions.

Remember that customer segmentation should be an ongoing process, and innovation should be at its core to adapt to changing customer preferences and market dynamics.

**(1)Data Preprocessing**

**I) Data Selection and loading:**

Data selection and loading are vital steps in the data preparation process for machine learning. Data selection involves choosing the right dataset that aligns with your objectives, ensuring it contains relevant information. Data loading refers to the process of importing the chosen dataset into your machine learning environment, such as a Python notebook or database. This step includes reading data from files (e.g., CSV, Excel) or databases, making it ready for further analysis. Accurate data selection and loading set the foundation for successful machine learning tasks by providing the right information for model training and evaluation.

**Code:**

#Importing the necessary libraries

importnumpyasnp

import pandas as pd

importmatplotlib.pyplotasplt

importseabornassns

from mpl\_toolkits.mplot3d import Axes3D

%matplotlibinline

#Reading the excel file

data=pd.read\_excel("Mall\_Customers.xlsx")

data.head()

data.describe()

Output:

|index|CustomerID|Gender|Age|Annual Income \(k$\)|Spending Score \(1-100\)|

|---|---|---|---|---|---|

|0|1|Male|19|15|39|

|1|2|Male|21|15|81|

|2|3|Female|20|16|6|

|3|4|Female|23|16|77|

|4|5|Female|31|17|40|

<class'pandas.core.frame.DataFrame'>

RangeIndex: 200 entries, 0 to 199

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 CustomerID 200 non-null int64

1 Gender 200 non-null object

2 Age 200 non-null int64

3 Annual Income (k$) 200 non-null int64

4 Spending Score (1-100) 200 non-null int64

dtypes: int64(4), object(1)

memory usage: 7.9+ KB

**II) Data Cleaning:**

- Data often contains missing values, outliers, and inconsistencies that need to be addressed. Missing values can be filled in through imputation or removed if necessary. Outliers can be adjusted or flagged.

- Importance: Cleaning data ensures that the model is not influenced by errors or outliers, leading to more robust and accurate predictions.

**Code:**

# Data cleaning and handling missing values

print(df.drop\_duplicates())

print(df.dropna())

print(df.isna().any())

print(df.isna().sum())

cf=df.copy()

Output:

CustomerID Gender Age Annual Income (k$) Spending Score (1-100)

0 1 Male 19 15 39

1 2 Male 21 15 81

2 3 Female 20 16 6

3 4 Female 23 16 77

4 5 Female 31 17 40

.. ... ... ... ... ...

195 196 Female 35 120 79

196 197 Female 45 126 28

197 198 Male 32 126 74

198 199 Male 32 137 18

199 200 Male 30 137 83

[200 rows x 5 columns]

CustomerID Gender Age Annual Income (k$) Spending Score (1-100)

0 1 Male 19 15 39

1 2 Male 21 15 81

2 3 Female 20 16 6

3 4 Female 23 16 77

4 5 Female 31 17 40

.. ... ... ... ... ...

195 196 Female 35 120 79

196 197 Female 45 126 28

197 198 Male 32 126 74

198 199 Male 32 137 18

199 200 Male 30 137 83

[200 rows x 5 columns]

CustomerID False

Gender False

Age False

Annual Income (k$) False

Spending Score (1-100) False

dtype: bool

CustomerID 0

Gender 0

Age 0

Annual Income (k$) 0

Spending Score (1-100) 0

dtype: int64

**III) Feature Selection:**

In some cases, not all variables or features in the data may be equally relevant for segmentation. Feature selection identifies and retains the most valuable attributes for the segmentation process while discarding less informative ones.

**Code:**

#We take just the Annual Income and Spending score

df1=data[["CustomerID","Gender","Age","Annual Income (k$)","Spending Score (1-100)"]]

X=df1[["Annual Income (k$)","Spending Score (1-100)"]]

#The input data

X.head()

**Output:**

|index|Annual Income \(k$\)|Spending Score \(1-100\)|

|---|---|---|

|0|15|39|

|1|15|81|

|2|16|6|

|3|16|77|

|4|17|40|

**IV) Data Exploration:**

Data exploration is a crucial initial step in the process of customer segmentation. It involves gaining a deeper understanding of your customer data, identifying patterns, trends, and characteristics that can inform your segmentation strategy.

**Data correlation:**

Code:

data.corr()

dataplot = sns.heatmap(data.corr(),cmap="YlGnBu",annot=True)

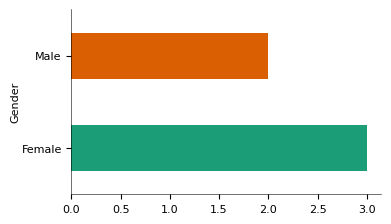
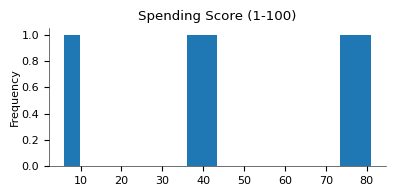
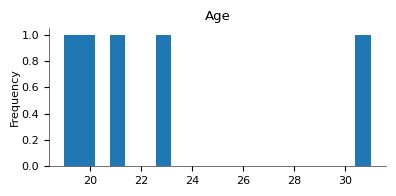
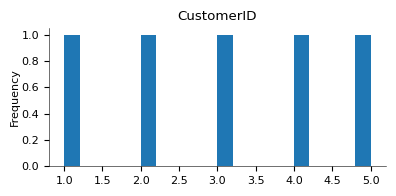
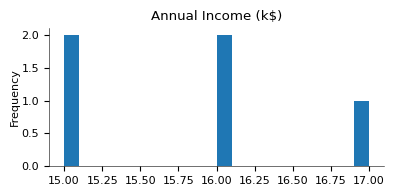
# displaying heatmap

plt.show()

Output:



**Feature distributions:**

** **

**V)Data visualization:**

Visualize your data to reveal patterns and relationships. Create histograms, bar charts, scatter plots, and other relevant charts to examine the distribution of variables and correlations between them. Tools like Matplotlib, Seaborn, or ggplot2 are useful for data visualization.

**Code:**

#Scatterplot of the input data

plt.figure(figsize=(10,6))

sns.scatterplot(x = 'Annual Income (k$)',y = 'Spending Score (1-100)',  data = X  ,s = 60)

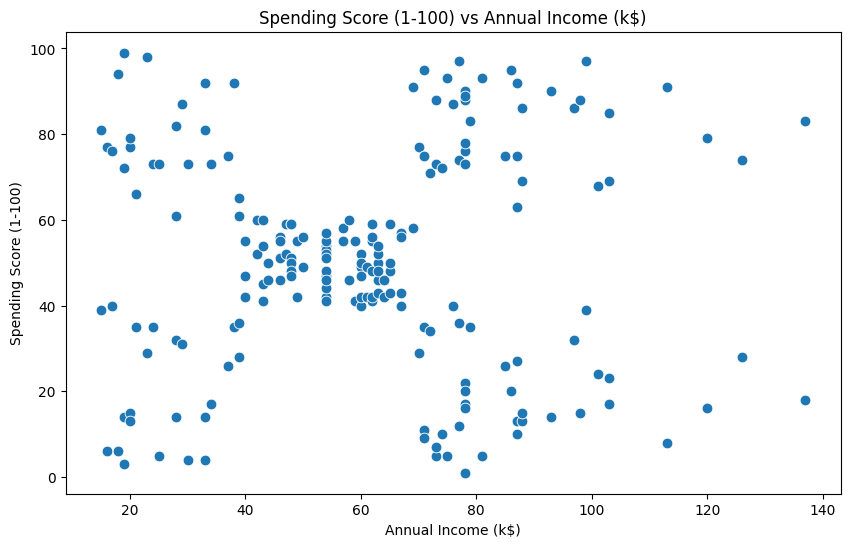
plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.title('Spending Score (1-100) vs Annual Income (k$)')

plt.show()

**Output:**

****

**2)Building a Unsupervised Machine learning model**

**Feature Selection:**

Identify the relevant features (variables) that are likely to influence your segmentation goals. This might include demographic information, purchase history, website interactions, location data, etc. Start with a comprehensive set and then narrow it down as needed.

**Code:**

#We take just the Annual Income and Spending score

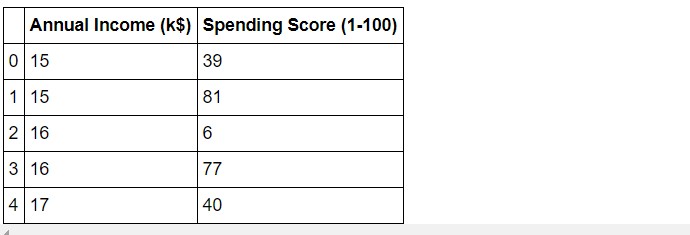
df1=data[["CustomerID","Gender","Age","Annual Income (k$)","Spending Score (1-100)"]]

X=df1[["Annual Income (k$)","Spending Score (1-100)"]]

#The input data

X.head()

**Output:**

****

**Applying K-Means Algorithm:**

Model Building and Evaluation:

Use a suitable segmentation algorithm (e.g., k-means, hierarchical clustering, or machine learning models) to segment your customers based on the engineered features. Evaluate the model's performance using appropriate metrics such as Silhouette Score, Davies-Bouldin Index, or business-specific KPIs.

#Taking 5 clusters

km1=KMeans(n\_clusters=5)

#Fitting the input data

km1.fit(X)

#predicting the labels of the input data

y=km1.predict(X)

#adding the labels to a column named label

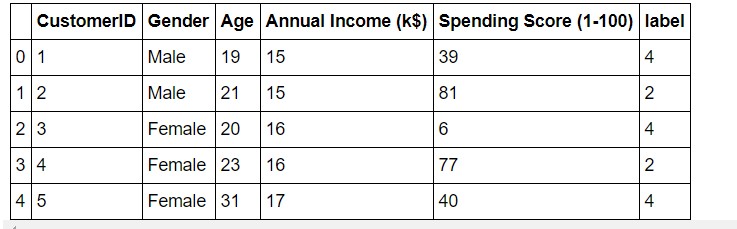
df1["label"] = y

#The new dataframe with the clustering done

df1.head()

#The labels added to the data.

**Output:**



**Visualization:**

Visualizing customer segmentation is essential for understanding the distinct groups and their characteristics. Here are several common visualization techniques you can use to represent customer segmentation results in Python:

**Code:**

#Scatterplot of the input data

plt.figure(figsize=(10,6))

sns.scatterplot(x = 'Annual Income (k$)',y = 'Spending Score (1-100)', data = X ,s = 60 )

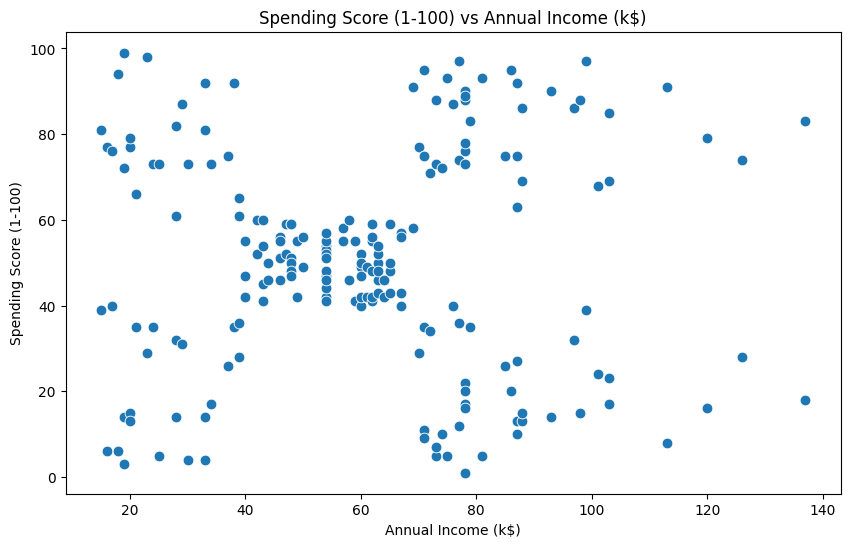
plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.title('Spending Score (1-100) vs Annual Income (k$)')

plt.show()

**Output:**

****

**Code:**

#Scatterplot of the clusters

plt.figure(figsize=(10,6))

sns.scatterplot(x = 'Annual Income (k$)',y = 'Spending Score (1-100)',hue="label",

palette=['green','orange','brown','dodgerblue','red'], legend='full',data = df1 ,s = 60 )

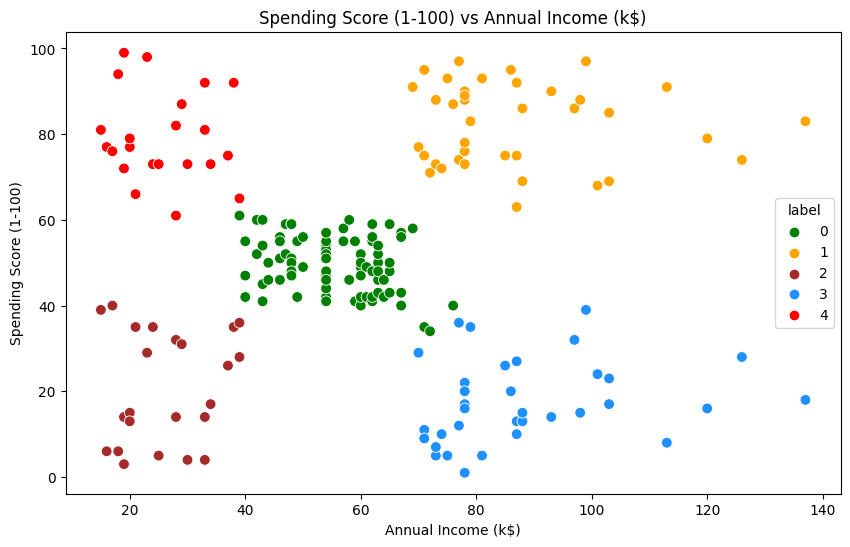
plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.title('Spending Score (1-100) vs Annual Income (k$)')

plt.show()

**Output:**

****

**3D Visualization:**

**Code:**

#3D Plot as we did the clustering on the basis of 3 input features

fig = plt.figure(figsize=(20,10))

ax = fig.add\_subplot(111, projection='3d')

ax.scatter(df2.Age[df2.label == 0], df2["Annual Income (k$)"][df2.label == 0], df2["Spending Score (1-100)"][df2.label == 0], c='purple', s=60)

ax.scatter(df2.Age[df2.label == 1], df2["Annual Income (k$)"][df2.label == 1], df2["Spending Score (1-100)"][df2.label == 1], c='red', s=60)

ax.scatter(df2.Age[df2.label == 2], df2["Annual Income (k$)"][df2.label == 2], df2["Spending Score (1-100)"][df2.label == 2], c='blue', s=60)

ax.scatter(df2.Age[df2.label == 3], df2["Annual Income (k$)"][df2.label == 3], df2["Spending Score (1-100)"][df2.label == 3], c='green', s=60)

ax.scatter(df2.Age[df2.label == 4], df2["Annual Income (k$)"][df2.label == 4], df2["Spending Score (1-100)"][df2.label == 4], c='yellow', s=60)

ax.view\_init(35, 185)

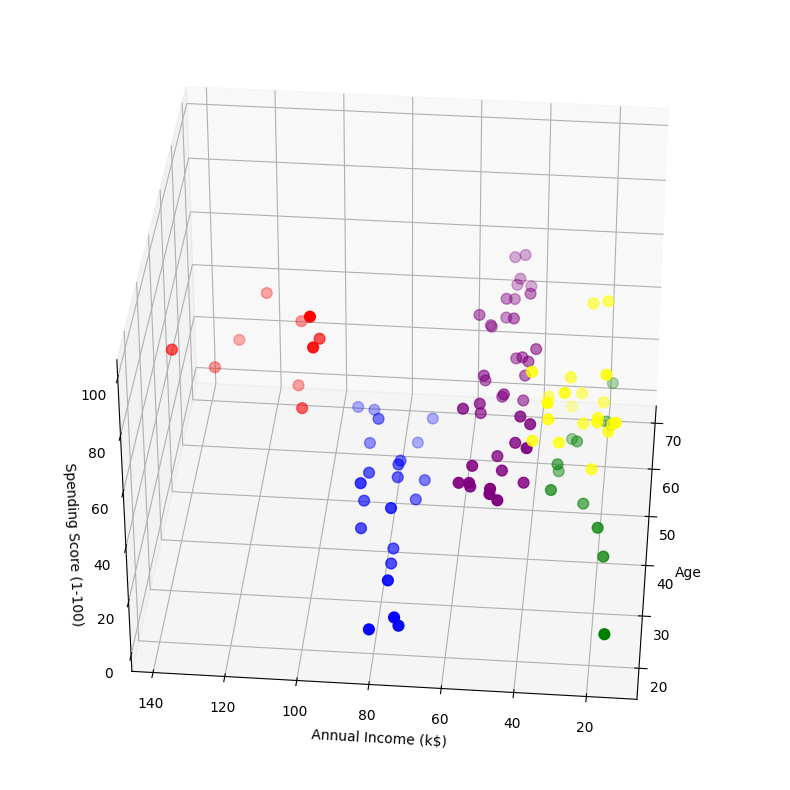
plt.xlabel("Age")

plt.ylabel("Annual Income (k$)")

ax.set\_zlabel('Spending Score (1-100)')

plt.show()

**Output:**

****

**Interpretation:**

Finally, you interpret the clusters by analyzing the statistics and the visualizations, give them meaningful names, and use the insights to drive business actions specific to each segment. Adjust the code to match your specific dataset and clustering results, and be sure to perform a more in-depth analysis based on your domain knowledge and segmentation goals.

**Code:**

cust1=df2[df2["label"]==1]

print('Number of customer in 1st group=', len(cust1))

print('They are -', cust1["CustomerID"].values)

print("--------------------------------------------")

cust2=df2[df2["label"]==2]

print('Number of customer in 2nd group=', len(cust2))

print('They are -', cust2["CustomerID"].values)

print("--------------------------------------------")

cust3=df2[df2["label"]==0]

print('Number of customer in 3rd group=', len(cust3))

print('They are -', cust3["CustomerID"].values)

print("--------------------------------------------")

cust4=df2[df2["label"]==3]

print('Number of customer in 4th group=', len(cust4))

print('They are -', cust4["CustomerID"].values)

print("--------------------------------------------")

cust5=df2[df2["label"]==4]

print('Number of customer in 5th group=', len(cust5))

print('They are -', cust5["CustomerID"].values)

print("--------------------------------------------")

**Output:**

Number of customer in 1st group= 10

They are - [182 184 186 188 190 192 194 196 198 200]

--------------------------------------------

Number of customer in 2nd group= 22

They are - [129 131 135 137 139 141 145 149 151 153 155 157 159 163 165 167 169 171

173 175 177 179]

--------------------------------------------

Number of customer in 3rd group= 40

They are - [44 46 47 48 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70

71 72 73 74 75 76 77 78 79 81 82 85 86 87 88 89]

--------------------------------------------

Number of customer in 4th group= 12

They are - [ 3 7 9 11 13 15 23 25 31 33 35 37]

--------------------------------------------

Number of customer in 5th group= 21

They are - [ 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42] --------------------------------------------

**(3)Key Findings & Insights**

To present key findings, insights, and recommendations based on customer segments, you should follow a structured approach. The process typically involves data analysis and understanding customer behavior. Below, I'll outline a step-by-step guide to help you present your findings effectively:

**Data Collection and Segmentation:**

Collect relevant customer data, including demographics, behavior, purchase history, and any other pertinent information.

Segment the customer data into distinct groups based on common characteristics. Common segmentation variables include age, gender, location, purchase frequency, or product preferences.

**Key Findings:**

Identify trends and patterns within each customer segment. For example:

Segment A (e.g., young adults) may show a higher preference for online shopping.

Segment B (e.g., parents) may have a tendency to purchase in-store.

Highlight any significant differences or similarities between the segments.

**Insights:**

Provide insights into the behaviors, preferences, and needs of each customer segment. For example:

Segment A values convenience and quick delivery.

Segment B prioritizes product quality and is willing to pay more for it.

Explain why these insights are important for your business.

**Recommendations:**

Based on the insights, offer specific recommendations for marketing, product development, or customer service tailored to each segment. For example:

For Segment A, invest in a user-friendly online shopping experience and offer expedited shipping options.

For Segment B, focus on enhancing product quality and providing in-store experiences that cater to family needs.

**Personalization:**

Suggest strategies for personalizing interactions with customers within each segment. This may involve personalized email campaigns, product recommendations, or special offers.

**Measurement and Feedback:**

Develop key performance indicators (KPIs) to measure the success of your recommendations. This could include customer satisfaction scores, conversion rates, or revenue growth.

Continuously gather feedback from customers within each segment to refine your strategies.

**Implementation Plan:**

Create a clear and actionable plan for implementing the recommendations. Assign responsibilities and set timelines for execution.

**Monitoring and Iteration:**

Continuously monitor the performance of your strategies and be prepared to adjust them as needed based on evolving customer behaviors and market dynamics.

**Reporting:**

Provide a comprehensive report that summarizes the key findings, insights, and recommendations for each customer segment. Use visuals like charts and graphs to make the information more digestible.

**Stakeholder Communication:**

Share your findings, insights, and recommendations with relevant stakeholders within your organization, such as marketing teams, product development teams, and customer service teams.

By following this structured approach, you can effectively present key findings, insights, and recommendations based on customer segments, ultimately leading to more targeted and successful business strategies.

**Conclusion:**

In conclusion, a well-executed customer segmentation project is a strategic asset for businesses. It enables targeted marketing, personalized services, and improved customer satisfaction. By continually refining and adapting the segmentation strategy, businesses can stay competitive and better serve their diverse customer base. The ability to understand and cater to the unique needs of different customer segments is a key driver of success in today's market.