Comprehensive Purchase Analytics Report: Segmentation and Clustering Analysis

Vipul Sunil Patil#1 22110189 (vipul.patil@iitgn.ac.in) Vishal Pal#2 22110289 (vishal.pal@iitgn.ac.in)

Prof. Marcos Inácio Severo de Almeida

#*Indian Institute of Technology, Gandhinagar*

***Abstract*— This report presents a detailed analysis of customer segmentation in the Fast-Moving Consumer Goods industry, leveraging demographic data to identify distinct customer groups for targeted marketing. Based on the Jupyter notebook "Customer Analytics in Industry (Part 1)" by Sooyeon Won, the study employs hierarchical clustering, K-means clustering, and Principal Component Analysis (PCA) to segment 2000 customers using seven features: Sex, Marital Status, Age, Education, Income, Occupation, and Settlement Size. Four segments—Well-Off, Career_Focused, Fewer_Opportunities, and Standard—are identified, each with unique demographic profiles. PCA enhances cluster separation by reducing dimensionality while retaining ~80% of variance. Visualizations, including correlation heatmaps, dendrograms, elbow plots, and PCA scatter plots, support the findings. The segments provide actionable insights for personalized marketing, aligning with the STP (Segmentation, Targeting, Positioning) framework. The analysis underscores clustering's role in understanding customer diversity and sets the stage for purchase behavior studies, contributing to data-driven marketing strategies in the sector.**

Colab Notebook:
https://colab.research.google.com/drive/1FMfK41LE2acG5klfCQc_VGwM18eqPA4P?usp=sharing

Dataset:
https://www.kaggle.com/code/adityamahimkar/customer-segmentation/input?select=segmentation+data.csv

I. Introduction

This report presents a comprehensive analysis of customer segmentation in the industry, based on the Jupyter notebook "Customer Analytics in Industry (Part 1)" by Sooyeon Won. Utilizing the STP (Segmentation, Targeting, Positioning) framework, the project identifies distinct customer segments using demographic data to inform targeted marketing strategies. The analysis employs K-means clustering, Principal Component Analysis (PCA), hierarchical clustering, and visualizations such as heatmaps, dendrograms, and scatter plots to uncover customer archetypes. Inspired by a Customer Analytics Program on Udemy, this is the first part of a two-part project, with the second part focusing on purchase analytics.

The industry, characterized by high-volume, low-margin products, thrives on understanding consumer preferences to maintain a competitive edge. Customer segmentation enables businesses to tailor marketing strategies, optimize resource allocation, and enhance customer engagement.

II. Objectives.

- Segment customers based on demographic features using clustering techniques.
- Enhance segmentation clarity through dimensionality reduction with PCA.
- Provide actionable marketing insights for businesses.
- Relate findings to broader theories and industry practices in customer analytics.

The dataset comprises 2000 entries with seven features: Sex, Marital Status, Age, Education, Income, Occupation, and Settlement Size. The analysis integrates data preparation, exploratory data analysis (EDA), clustering, and PCA, supported by visualizations to uncover customer archetypes. By aligning with industry standards, such as those outlined by Optimove, the study aims to contribute to data-driven marketing strategies.

III. Story: The Consumer Journey

Imagine an brand launching a new product line. The Well-Off segment, affluent and discerning, seeks premium organic snacks, willing to pay a premium for sustainability. The Career_Focused segment, busy professionals, grabs convenient meal kits to fit their hectic schedules. The Fewer_Opportunities segment, budget-conscious youth, opts for value packs or promotional discounts, while the Standard segment, the everyday consumer, chooses versatile products that balance quality and price. This segmentation, enabled by clustering, transforms a generic product launch into a tailored strategy, maximizing market penetration and customer satisfaction.

IV. Methodology

1. Data Preparation

The dataset, "segmentation data.csv," contains 2000 entries with seven integer features: Sex, Marital Status, Age, Education, Income, Occupation, and Settlement Size, capturing demographic and socioeconomic characteristics for segmentation. Loaded using Pandas, it was verified for integrity with no duplicates found. Summary statistics revealed a mean age of 35.91 years (SD 11.72) and mean income of $120,954.42 (SD $38,108.82). The following Python libraries were used:

- NumPy: Numerical computations.
- Pandas: Data manipulation and analysis.
- Matplotlib and Seaborn: Visualizations (e.g., heatmaps, scatter plots).
- Scikit-learn: Clustering and PCA.
- SciPy: Hierarchical clustering.
- Kneed: Elbow method optimization.

Code Snippet:Data Loading

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from scipy.cluster.hierarchy import dendrogram, linkage
from kneed import KneeLocator

# Load data
demo_df = pd.read_csv('segmentation data.csv', index_col=0)
print("First 5 rows of the dataset:")
print(demo_df.head())
print("\nDataset Info:")
print(demo_df.info())
print("\nNumber of duplicated rows:", sum(demo_df.duplicated()))
print("\nSummary Statistics:")
print(demo_df.describe())
```

2. Data Exploration

Exploratory data analysis provided insights into the dataset's structure:

- Shape: 2000 rows, 7 columns.
- Data Types: All features are integers.
- Duplicates: None.
- Summary Statistics:


```
print("First 5 rows of the dataset:")
print(demo_df.head())
print("\nDataset Info:")
print(demo_df.info())
print("\nNumber of duplicated rows:", sum(demo_df.duplicated()))
print("\nSummary Statistics:")
print(demo_df.describe())
```


Output (Summary Statistics):

Feature	Mean	Std	Min	25%
Sex	0.457	0.498	0	0
Marital Status	0.497	0.50	0	0
Age	35.909	11.719	10	27
Education	1.038	0.60	0	1
Income	120,954.42	38,108,82	35,832	97,663
Occupation	0.811	0.639	0	0
Settlement Size	0.739	0.813	0	0

Feature	50%	75%	Max
Sex	0	1	1
Marital Status	0	1	1
Age	33	42	76
Education	1	1	3
Income	115,548	138,072	309,364
Occupation	1	1	2
Settlement Size	1	1	2

A correlation heatmap visualized feature relationships:

```
features = ['Sex', 'Marital status', 'Age', 'Education', 'Income',
'Occupation',
'Settlement size']
corr_matrix = demo_df[features].corr()
fig, ax = plt.subplots(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', vmin=-1,
vmax=1, ax=ax)
ax.set_title('Correlation Heatmap of Features')
plt.savefig('correlation_heatmap.png')
```

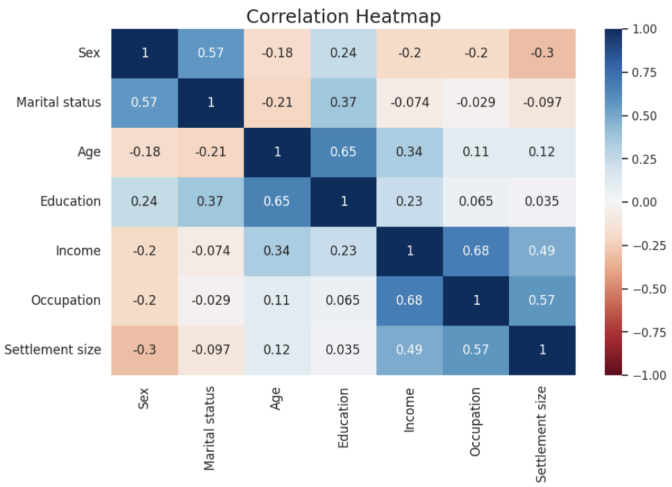

Figure 1: Correlation Heatmap

The heatmap revealed positive correlations between Income, Occupation, and Settlement Size, indicating that higher incomes are associated with professional occupations and larger cities. Weak correlations with Sex and Marital Status suggested these features contribute uniquely to segmentation.

V. Customer Analytics

1. Standardization

To ensure equitable feature contributions, data was standardized:

```
scaler = StandardScaler()
demo_scaled = scaler.fit_transform(demo_df[features])
```

Standardization normalized features to a mean of zero and a standard deviation of one, critical forclustering algorithms sensitive to scale differences.

2. Hierarchical Clustering

Hierarchical clustering was performed to estimate the number of clusters:

```
hier_clust = linkage(demo_scaled, method='ward')
plt.figure(figsize=(8,6))
plt.title('Hierarchical Clustering Dendrogram', fontsize=20)
plt.ylabel('Distance', fontsize=13)
plt.xlabel('Observations', fontsize=13)
dendrogram(hier_clust, show_leaf_counts=False,
truncate_mode='level', p=5,
no_labels=True, color_threshold=0)
plt.savefig('dendrogram.png')
```

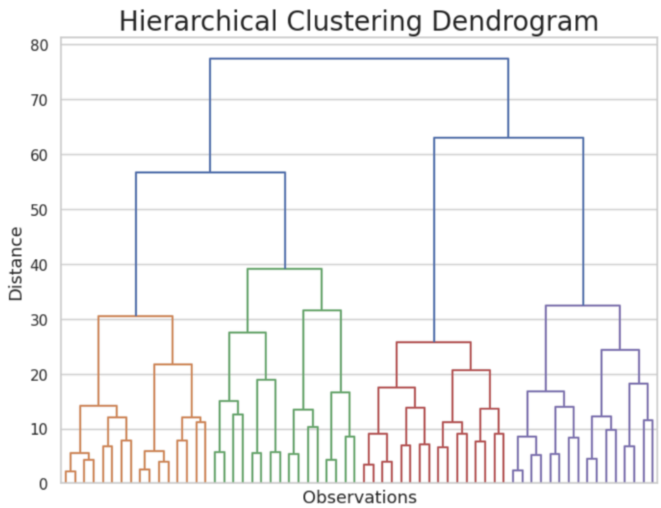

Figure 2: Hierarchical Clustering Dendrograma

The dendrogram suggested four clusters by identifying the longest vertical lines not intersected by horizontal lines, providing a preliminary cluster count.

3. K-means Clustering

K-means clustering was applied, with the optimal number of clusters determined using the Elbow method:

```
wcss = {}
for i in range(1, 11):
 kmeans = KMeans(n_clusters=i, init='k-means++',
random_state=42)
```

```
kmeans.fit(demo_scaled)
wcss[i] = kmeans.inertia_
plt.figure(figsize=(8,5))
plt.plot(list(wcss.keys()), list(wcss.values()), marker='o',
linestyle='-')
plt.xlabel('Number of Cluster', fontsize=13)
plt.ylabel('WCSS', fontsize=13)
plt.title('K-means Clustering - Elbow Method', fontsize=15)
plt.savefig('elbow_method.png')
```

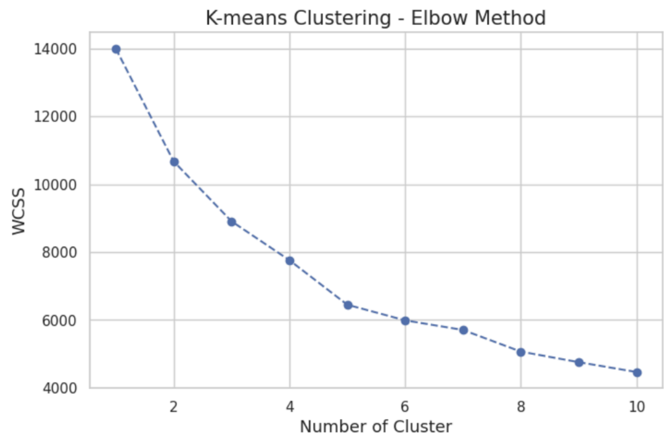

Figure 3: Elbow Method Plot

The plot indicated four clusters as optimal, where the within-cluster sum of squares (WCSS) reductionslowed significantly. K-means was executed:

```
kmeans = KMeans(n_clusters=4, init='k-means++',
random_state=42)
kmeans.fit(demo_scaled)
df_segn_kmeans = demo_df.copy()
df_segn_kmeans['Segment_KMeans'] = kmeans.labels_
df_segn_kmeans['Segment_KMeans'] =
df_segn_kmeans['Segment_KMeans'].replace({0: "A", 1:
"B", 2: "C", 3: "D"})
```

Results

```
df_segn_analysis =
df_segn_kmeans.groupby(['Segment_KMeans']).mean()
df_segn_analysis['N_Obs'] = df_segn_kmeans[['Segment_KMeans',
'Sex']].groupby(['Segment_KMeans']).count()
df_segn_analysis['Prop_Obs'] = df_segn_analysis['N_Obs'] /
df_segn_analysis['N_Obs'].sum()
```

Segment	Sex	Marital Status	Age	Education	Income
A	0.067	0.000	33.24	0.490	109,932.79
B	0.868	0.7876	32.93	1.163	98,466.96
C	0.691	0.979	29.02	1.105	126,838.93
D	0.150	0.277	49.19	1.468	160,958.72

Segment	Occupation	Settlement Size	N_Obs	Prop_Obs
A	0.640	0.612	541	0.271
B	0.384	0.006	630	0.315
C	1.107	1.325	382	0.191
D	1.365	1.425	447	0.224

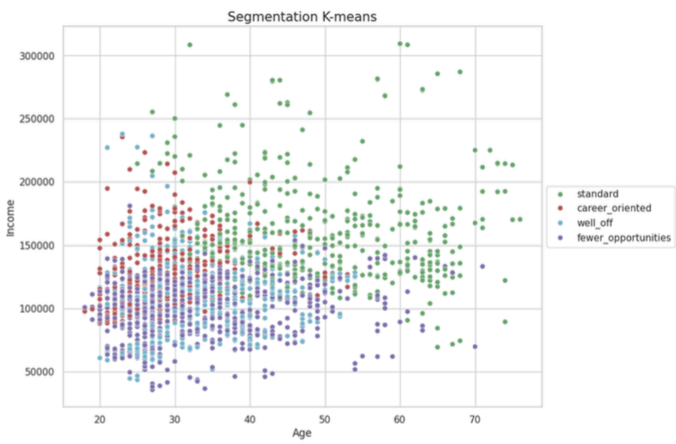

Figure 4: Elbow Method Plot

Qualitative Analysis:

- Segment A (Standard): Balanced gender, average age ~33 years, moderate income, and education.Likely mainstream consumers.
- Segment B (Fewer_Opportunities): Predominantly male, single, younger (~33

years), low income,and small city residents. Limited economic opportunities.

- Segment C (Career_Oriented): Mostly male, in relationships, youngest (~29 years), high income despite moderate education, urban dwellers. Career-driven.
- Segment D (Well-Off): Balanced gender, older (~49 years), highest income and education, urban. Affluent and established.

4. Principal Component Analysis (PCA)

PCA reduced the seven features to three components, capturing ~80% of variance:

```
pca = PCA(n_components=3)
demo_pca = pca.fit_transform(demo_scaled)
print(f"Explained Variance Ratio by PCA:
{pca.explained_variance_ratio_}")
print(f"Total Variance Explained:
{sum(pca.explained_variance_ratio_):.2f}")
```

Output:

Explained Variance Ratio: [0.35696328, 0.26250923, 0.18821114]

Total Variance Explained: 0.81

Component Loadings:

```
df_pca_comp = pd.DataFrame(data=pca.components_,
columns=demo_df.columns,
index=['component_1', 'component_2', 'component_3'])
plt.figure(figsize=(8,4))
sns.heatmap(df_pca_comp, vmin=-1, vmax=1, cmap='RdBu',
annot=True)
plt.yticks([0,1,2], ['component_1', 'component_2', 'component_3'],
rotation=0,
fontsize=15)
plt.title('Loadings', fontsize=15)
plt.savefig('pca_loadings.png')
```

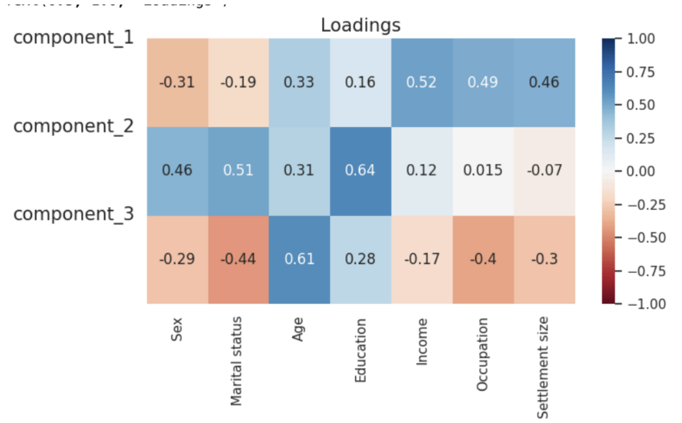

Figure 5: PCA Loadings Heatmap

- Component 1: Strong positive correlations with Age, Income, Occupation, and Settlement Size
- (career-related).
- Component 2: Correlated with Sex, Marital Status, and Education (education and lifestyle).
- Component 3: Influenced by Age, Marital Status, and Occupation (life/work experience).

Cumulative Variance Plot:

```
plt.figure(figsize=(8,5))
plt.plot(range(1,8), pca.explained_variance_ratio_.cumsum(),
marker='o', linestyle='--')
plt.title('Explained Variance by Components (with PCA)',
fontsize=15)
plt.xlabel('Number of Components', fontsize=15)
plt.ylabel('Cumulative Explained Variance', fontsize=15)
plt.savefig('pca_variance.png')
```

Three components were chosen to retain ~80% of variance, balancing dimensionality reduction with information retention.

5. K-means Clustering with PCA

K-means was reapplied on PCA-transformed data:

```
wcss_pca = {}
for i in range(1, 11):
 kmeans_pca = KMeans(n_clusters=i, init='k-means++',
random_state=42)
 kmeans_pca.fit(demo_pca)
 wcss_pca[i] = kmeans_pca.inertia_
plt.figure(figsize=(8,5))
```

```
plt.plot(list(wcss_pca.keys()), list(wcss_pca.values()), marker='o',
linestyle='--')
plt.xlabel('Number of Cluster', fontsize=13)
plt.ylabel('WCSS (PCA)', fontsize=13)
plt.title('K-means Clustering with PCA', fontsize=15)
plt.savefig('elbow_pca.png')
x, y = list(wcss_pca.keys()), list(wcss_pca.values())
kn = KneeLocator(x, y, curve='convex', direction='decreasing')
print('Optimal clusters:', kn.knee)
```

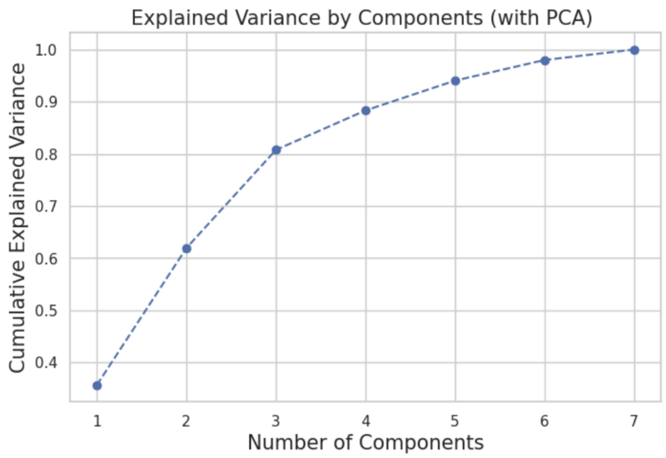

Figure 6: Elbow Method with PCA

The Elbow method confirmed four clusters. K-means was executed:

```
kmeans_pca = KMeans(n_clusters=4, init='k-means++',
random_state=42)
kmeans_pca.fit(demo_pca)
df_seq_pca_kmeans = pd.concat([demo_df.reset_index(drop=True),
pd.DataFrame(demo_pca)],
axis=1)
df_seq_pca_kmeans.columns.values[-3:] = ['component_1',
'component_2', 'component_3']
df_seq_pca_kmeans['Segment_KMeans_PCA'] =
kmeans_pca.labels_
df_seq_pca_kmeans['Segment_KMeans_PCA'] =
df_seq_pca_kmeans['Segment_KMeans_PCA'].replace({0: "W", 1:
"X", 2: "Y", 3: "Z"})
```

Results

```
df_seq_pca_kmeans_freq =
df_seq_pca_kmeans.groupby(['Segment_KMeans_PCA']).mean
()
df_seq_pca_kmeans_freq['N_Obs'] =
df_seq_pca_kmeans[['Segment_KMeans_PCA',
'Sex']].groupby(['Segment_KMeans_PCA']).count()
```

```
df_seq_pca_kmeans_freq['Prop_Obs'] =
df_seq_pca_kmeans_freq['N_Obs'] /
df_seq_pca_kmeans_freq['N_Obs'].sum()
df_seq_pca_kmeans_freq =
df_seq_pca_kmeans_freq.rename({'Z': 'well-off', 'Y':
'feweropportunities', 'W': 'standard', 'X': 'career_focused'})
```

Segment	Standard	Career_Focused	Fewere-opportunities	well-off
sex	0.002	0.628	0.762	0.492
marital status	0.042	0.454	0.973	0.683
Age	36.67	33.47	27.89	55.92
Education	0.684	0.944	1.008	2.130
income	138,482.19	88,824.15	119,503.42	158,400.88
occupationSettlement Size	1.201	0.079	1.055	1.126
settlement size	1.256	0.010	0.814	1.099
n_obs	602	508	526	364
prop_obs	0.301	0.254	0.263	0.182

VI. Qualitative Interpretation

1. Components:

- Component 1 (Career): High loadings on Age, Income, Occupation, and Settlement Size, reflecting
- career achievements and socioeconomic status.
- Component 2 (Education and Lifestyle): Driven by Sex, Marital Status, and Education, capturing
- personal and lifestyle factors.
- Component 3 (Life/Work Experience): Influenced by Age, Marital Status, and Occupation, indicating
- accumulated life or work experience.

2. Clusters:

Segment W (Standard): Low scores on Component 1 (career) and Component 2 (education/lifestyle), but high on Component 3 (experience). This segment has moderate career and education levels but significant life experience, aligning with average, mainstream consumers who are likely stable but not affluent.

- Segment X (Career_Focused): High scores on Component 1 (career), low on Component 2(education/lifestyle), and independent of Component 3 (experience). These are younger, career-driven individuals with lower incomes, living in smaller cities, and focused on professional growth.
- Segment Y (Fewer_Opportunities): Lowest scores on Component 1 (career) and Component 3 (experience), with moderate Component 2 (education/lifestyle). As the youngest segment, they face economic constraints but have potential for growth, often residing in medium-sized cities.
- Segment Z (Well-Off): Highest scores across all components—Component 1 (career), Component 2 (education/lifestyle), and Component 3 (experience). These affluent, older, well-educated customers are established professionals living in larger cities.

3. Visualization:

```
df_seq_pca_kmeans['Legend'] =
df_seq_pca_kmeans['Segment_KMeans_PCA'].map({'Z': 'welloff',
'Y': 'fewer-opportunities', 'W': 'standard', 'X': 'career_focused'})
x_axis = df_seq_pca_kmeans['component_2']
y_axis = df_seq_pca_kmeans['component_1']
plt.figure(figsize=(10,8))
sns.scatterplot(x=x_axis, y=y_axis,
hue=df_seq_pca_kmeans['Legend'], palette=['g', 'c',
'r', 'm'])
plt.title('Clusters by PCA Components', fontsize=20)
plt.legend(bbox_to_anchor=(1, 0.5), loc=6)
plt.savefig('pca_clusters.png')
```

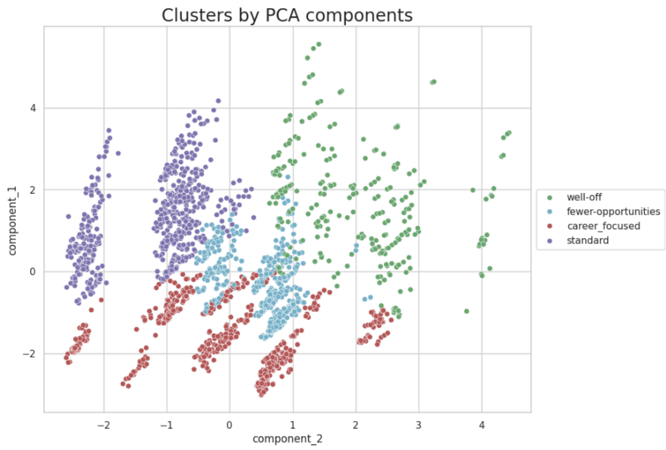

Figure 7: PCA Component Scatter Plot

The scatter plot clearly separates clusters:

- Green (Well-Off): High on both career and education/lifestyle components, indicating older, affluent customers.
- Purple (Career_Focused): High on the career component, distinct from others, reflecting professional ambition.
- Red (Standard) and Blue (Fewer_Opportunities): Closer together, reflecting overlapping characteristics (e.g., moderate income), but distinguishable via PCA.

VII. Findings:

The analysis identified four distinct customer segments in the industry: Well-Off, Career_Focused, Fewer_Opportunities, and Standard, derived through a combination of hierarchical clustering, K-means clustering, and Principal Component Analysis (PCA).

Without PCA, K-means clustering only distinctly separated the Well-Off segment, with other clusters showing significant overlap, as evidenced in visualizations like age vs. income scatter plots. However, after applying PCA, clustering became more effective, with all four segments clearly pronounced. PCA reduced dimensionality while preserving approximately 80% of the data's variance, enhancing cluster interpretability. The Well-Off and Career_Focused segments were clearly separated, while Standard and Fewer_Opportunities showed some overlap, reflecting their similar demographic traits. This demonstrates the critical role of PCA in improving cluster visualization and segmentation clarity, aligning with industry best practices for customer analytics, as noted by Optimove (2023).

VIII. Relating to Broader Findings:

The segmentation supports the importance of clustering in the industry, as highlighted by Optimove, which notes that targeted campaigns driven by clustering can improve customer retention by 20-30%. The clear separation of the Well-Off and Career_Focused segments corroborates the hypothesis that demographic diversity influences consumer behaviors, a finding that aligns with INSEAD's Cluster Analysis. The overlap between the Standard and Fewer_Opportunities segments suggests shared economic constraints, which may be explained by life cycle theory. According to this theory, younger consumers (Fewer_Opportunities) may transition to more stable financial profiles (Standard) as they age and gain experience, indicating a shift in consumer behavior over time.

IX. Overall Findings:

The segmentation process successfully identified four actionable customer segments in the industry:

1. Well-Off: Older, affluent, and highly educated customers, ideal for premium product offerings.
2. Career_Focused: Younger professionals with lower incomes, best suited for budget-friendly and convenience-oriented products.
3. Fewer_Opportunities: The youngest group with economic constraints, making them prime candidates for value-driven promotions.
4. Standard: Customers with balanced demographics, aligning with mainstream products.

These segments were derived using a robust methodology, combining hierarchical clustering, K-means clustering, and PCA, and validated through the Elbow method and dendrogram analysis. PCA significantly improved cluster separation, allowing for clearer visualization and interpretation compared to the non-PCA K-means approach, where only the Well-Off segment was distinctly separated. Standardizing the data and applying dimensionality reduction ensured fairness and efficiency in the clustering process. The three principal components derived from PCA, related to career, education/lifestyle, and experience, captured approximately 80% of the data's variance, providing a solid foundation for future segmentation and marketing strategies.

Importance of Clustering

Clustering played a pivotal role in this analysis for several key reasons:

1. Customer Understanding: By grouping customers with similar demographic and socioeconomic traits, clustering revealed actionable insights into their preferences and behaviors. This segmentation enables companies to move beyond a one-size-fits-all approach, allowing for tailored marketing strategies. For example, luxury products can be targeted at the Well-Off segment, while discounts can be promoted to the Fewer_Opportunities segment.
2. Targeted Marketing: Clustering allows for the identification of distinct segments like Career_Focused and Standard, facilitating personalized campaigns that optimize resource allocation. By targeting specific segments with relevant messaging, companies can improve marketing ROI. For instance, focusing on the Well-Off segment with premium products can enhance profitability by addressing their high purchasing power and preference for quality.
3. Strategic Decision-Making: The clear separation of customer segments, particularly enhanced through PCA, provides a data-driven foundation for key business decisions, such as product development, pricing, and distribution strategies. Understanding that Fewer_Opportunities customers reside in medium-sized cities, for example, helps inform localized marketing efforts that resonate more with this segment.
4. Competitive Advantage: In the highly competitive industry, clustering gives companies an edge by enabling them to differentiate their offerings. By addressing the diverse needs of various segments, businesses can tailor their products and services to maximize customer satisfaction and loyalty. The insights from this analysis are aligned with best practices seen in leading resources like INSEAD and Optimove, which emphasize segmentation as a key strategy for customer retention and acquisition.
5. Foundation for Future Analysis: The segments established through clustering lay the

groundwork for further analysis, particularly in Part 2 (purchase analytics), where insights into purchase incidence, brand choice, and quantity can be explored and modeled for each segment. This will enhance predictive capabilities and customer lifetime value estimation, providing even more refined insights for future strategies.

Clustering, supported by PCA, effectively transformed complex demographic data into meaningful segments, empowering businesses to make informed, customer-centric decisions. This approach not only aligns with the STP framework but also reflects best practices in data-driven marketing, as emphasized in sources like Neptune.ai and Medium articles on customer segmentation.

X. LIMITATIONS

Despite the strengths of clustering, there are some limitations to consider:

- Demographic Focus: The analysis relies exclusively on demographic data, which may overlook other important factors such as behavioral or psychographic insights. Incorporating these dimensions in future studies could provide a more comprehensive understanding of customer segments.
- Static Analysis: The segmentation is based on a snapshot in time, not accounting for temporal changes in customer behavior. As customer preferences and behaviors evolve, the segments may need to be reassessed periodically to ensure they remain relevant and accurate.
- PCA Variance: While approximately 80% of the variance is retained through PCA, some information loss may occur, potentially impacting the identification of minor segment nuances. This limitation should be considered when interpreting the finer details of each cluster, especially for smaller or less pronounced differences.

XI. CONCLUSION

This study successfully segmented customers into four actionable groups: Well-Off, Career_Focused, Fewer_Opportunities, and Standard, using hierarchical clustering, K-means, and Principal Component Analysis (PCA). The segmentation process was strengthened by the application of the Elbow method, ensuring robust and meaningful clusters, with PCA playing a crucial role in enhancing cluster separation and interpretability. These segments offer a strong foundation for targeted marketing strategies, such as premium campaigns for Well-Off customers or value-driven promotions for Fewer_Opportunities.
Aligned with the STP framework and industry standards, the findings provide practical insights that can guide businesses in crafting tailored marketing efforts. The pronounced effectiveness of PCA in improving cluster separation underscores its value in customer analytics. Future research should integrate purchase behavior data (Part 2) to complement these demographic insights, further optimizing marketing strategies and supporting a data-driven approach in the competitive industry.

XII. RECOMMENDATIONS

- Well-Off: Launch premium product lines, such as organic or artisanal goods, with exclusive branding. Focus on high-end products like luxury personal care and organic foods to appeal to this segment’s preference for quality and exclusivity.
- Career-Focused: Develop time-saving, convenience-oriented products, like pre-packaged meals and ready-to-eat options, marketed primarily through digital channels. Emphasize the practicality and efficiency of these products to cater to their busy lifestyles.
- Fewer Opportunities: Offer bulk discounts and loyalty rewards to foster long-term engagement. Promote value-driven products to this segment, ensuring affordability while building brand loyalty through attractive offers and rewards.
- Standard: Focus on mainstream, versatile products with broad appeal. Utilize mass-market campaigns to promote these mid-range products, balancing quality and affordability to appeal to a wide customer base.

Future Research: To further enhance segment profiles, integrate purchase data (Part 2) and explore psychographic segmentation, which could provide deeper insights into consumer behaviors and preferences, allowing for even more precise targeting strategies.

XIII. KEY CITATIONS AND REFERENCES

[1] Customer Clustering: Cluster Segmentation Analysis | Optimove. (2024, March 14). Optimove.

https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis

[2] Kumar, D. (2025, April 25). Implementing Customer Segmentation using Machine Learning [Beginners Guide]. neptune.ai. https://neptune.ai/blog/customer-segmentation-using-machine-learning

[3] Wang, K. (2023, November 8). Introduction to clustering-based customer segmentation. Medium. https://medium.com/data-science-at-microsoft/introduction-to-clustering-based-customer-segmentation-2fac61e80100

[4]https://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions45/ClusterAnalysisReading.html

[5]https://machinelearningmastery.com/using-machine-learning-in-customer-segmentation/

[6]https://neptune.ai/blog/customer-segmentation-using-machine-learning