BUSA8001
Applied Predictive Analytics
Programming Task 2

MACQUARIE University
SYDNEY·AUSTRALIA

Sulaiman Yusuf Zakaria

47895810

Word Count:

Introduction	**62 Words**
Exploratory Data Analysis	**158 Words**
Customer Segmentation and Naming	**264 Words**
Segment Interpretation and Comparison	**234 Words**
Marketing Recommendations	**115 Words**
Conclusion	**114 Words**
Total	**947 Words**

1. Introduction

In the fitness industry, knowing your customers is essential for designing suitable membership programs and marketing strategies. This report segments gym members based on demographic data, including age, income, and education, across seven variables. Using Python, the process involved exploratory analysis, standardising numeric features, and applying K-Means++ and Agglomerative Clustering to uncover customer groups that support more targeted and effective business decisions.

2. Exploratory Data Analysis (EDA)

Before clustering, the dataset was explored to understand key patterns. Age is nearly normal, with most members aged 30–50. Income is right-skewed, concentrated around $100K–$150K, with a few outliers retained.

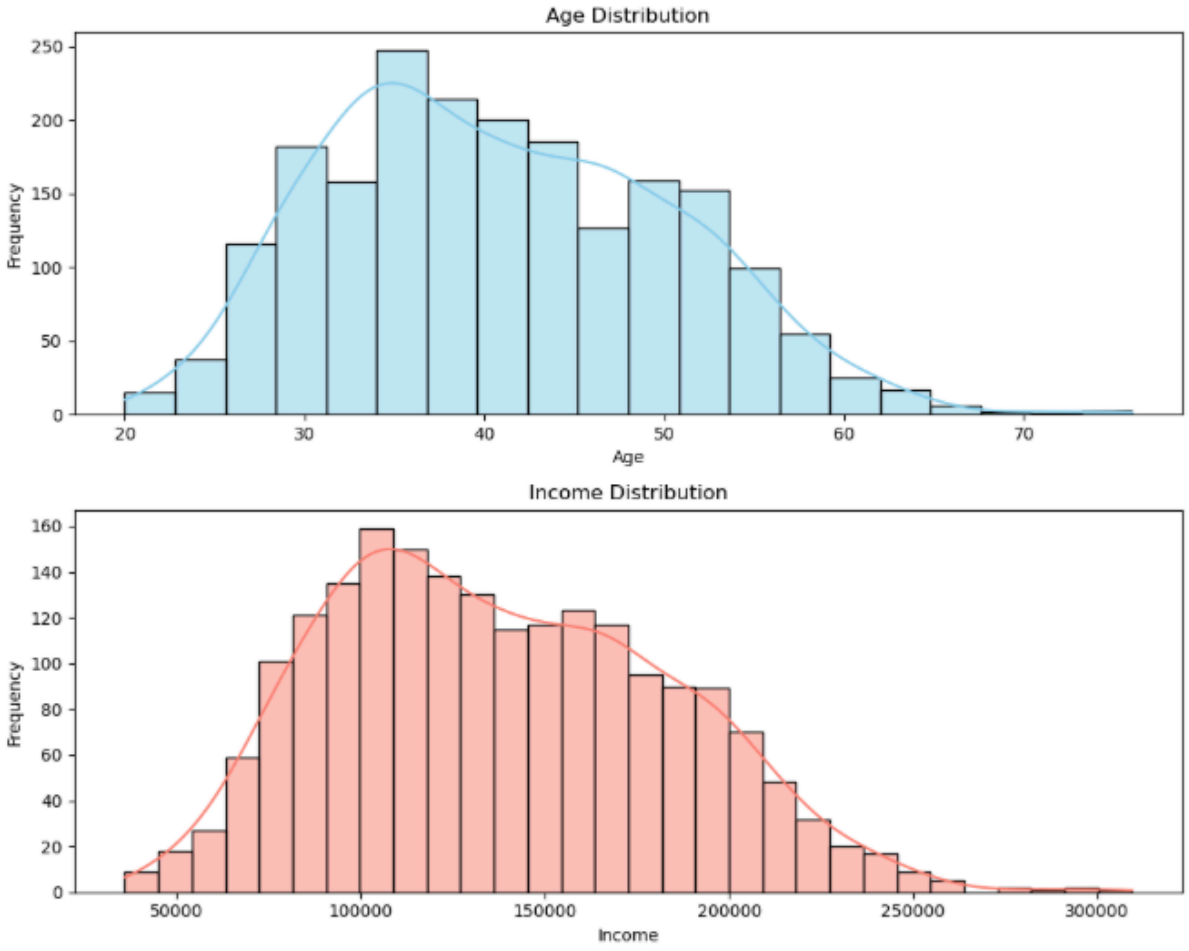

Figure 1: Age and Income Distribution

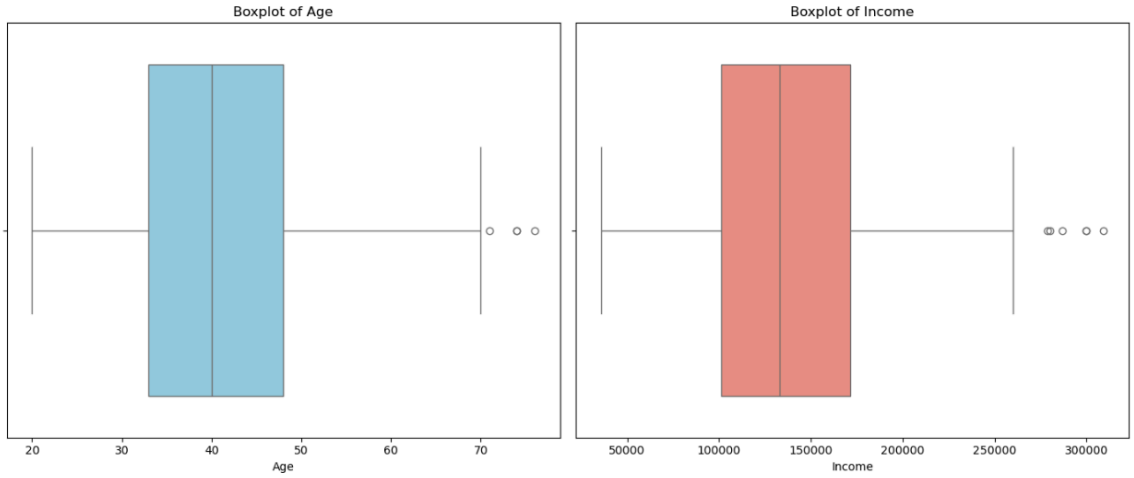

Figure 2: Boxplot of Age and Income

Categorical features reveal that 60% are female, marital status is balanced, and most members have high school or university education. Occupation shows variation: skilled and self-employed workers earn about $170K, while unskilled earn closer to $100K.

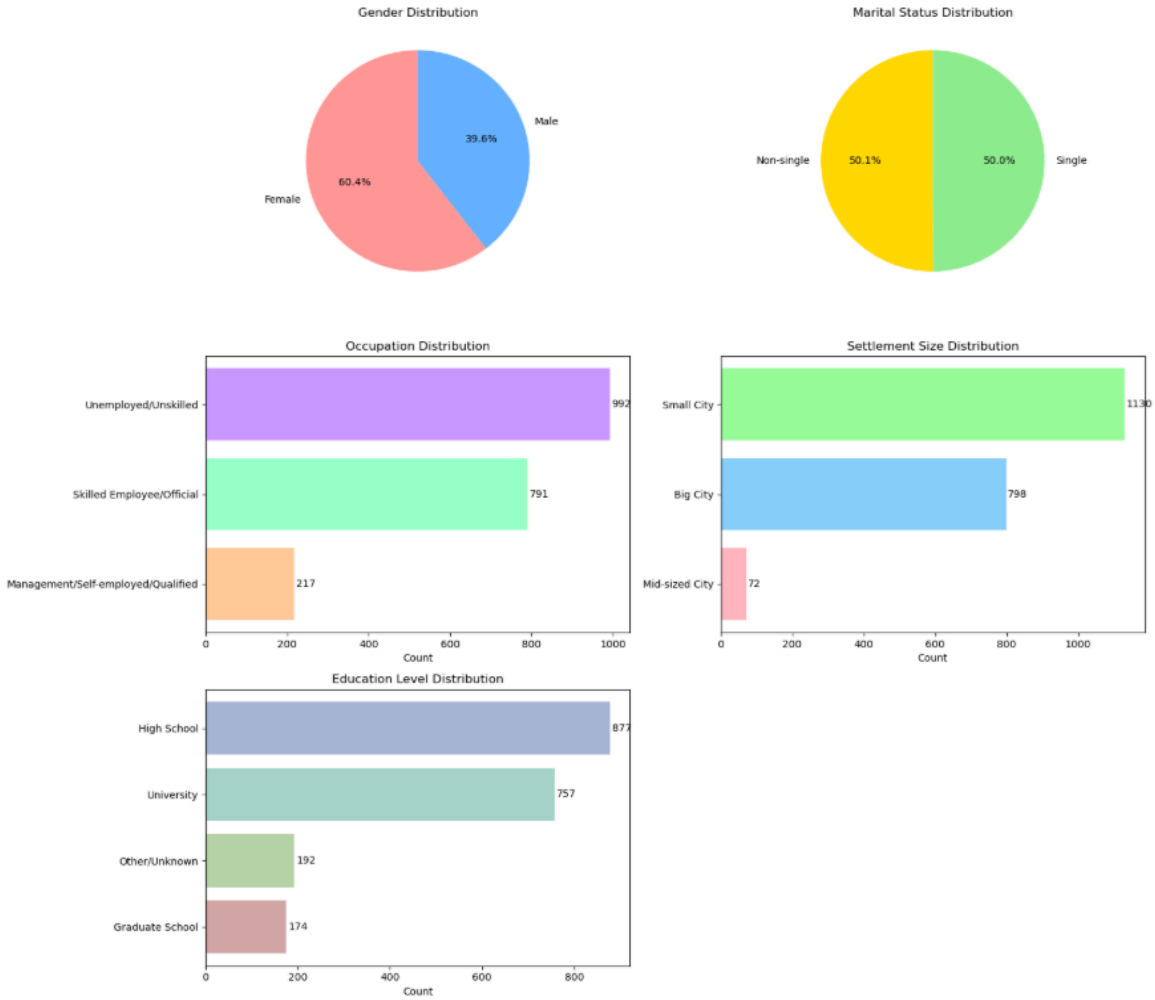

Figure 3: Distribution of Categorical Variables

Income also increases with age, particularly among those with higher education and skilled roles. A correlation matrix confirms strong age income correlation, and some links between occupation, settlement size, and marital status. These insights support variable relevance for clustering.

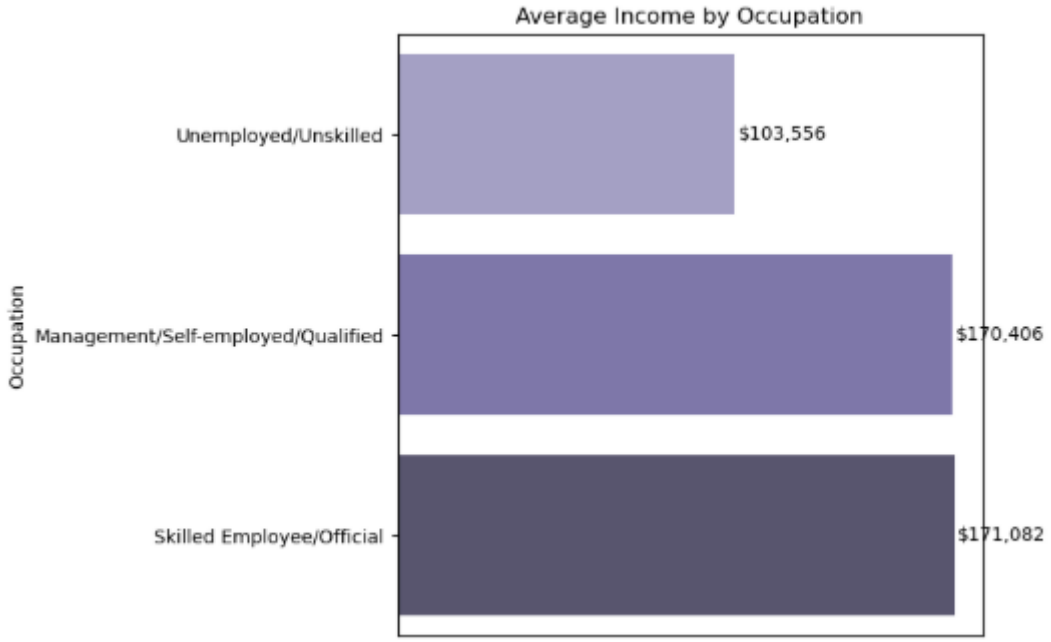

Figure 4: Income by Occupation

Income increases with age, especially among university-educated and skilled members, confirming these traits as key income predictors and potential gym engagement drivers.

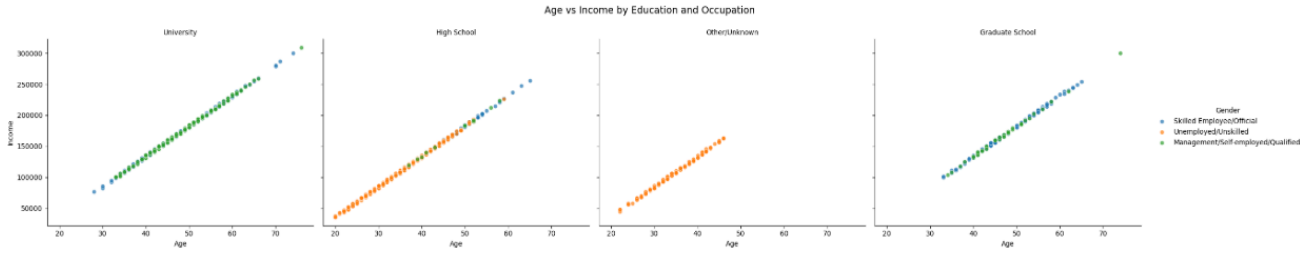

Figure 5: Age vs Income by Education and Occupation

The correlation matrix shows strong links between age and income, along with associations between occupation, marital status, and settlement size, indicating lifestyle patterns that could support more tailored segmentation strategies.

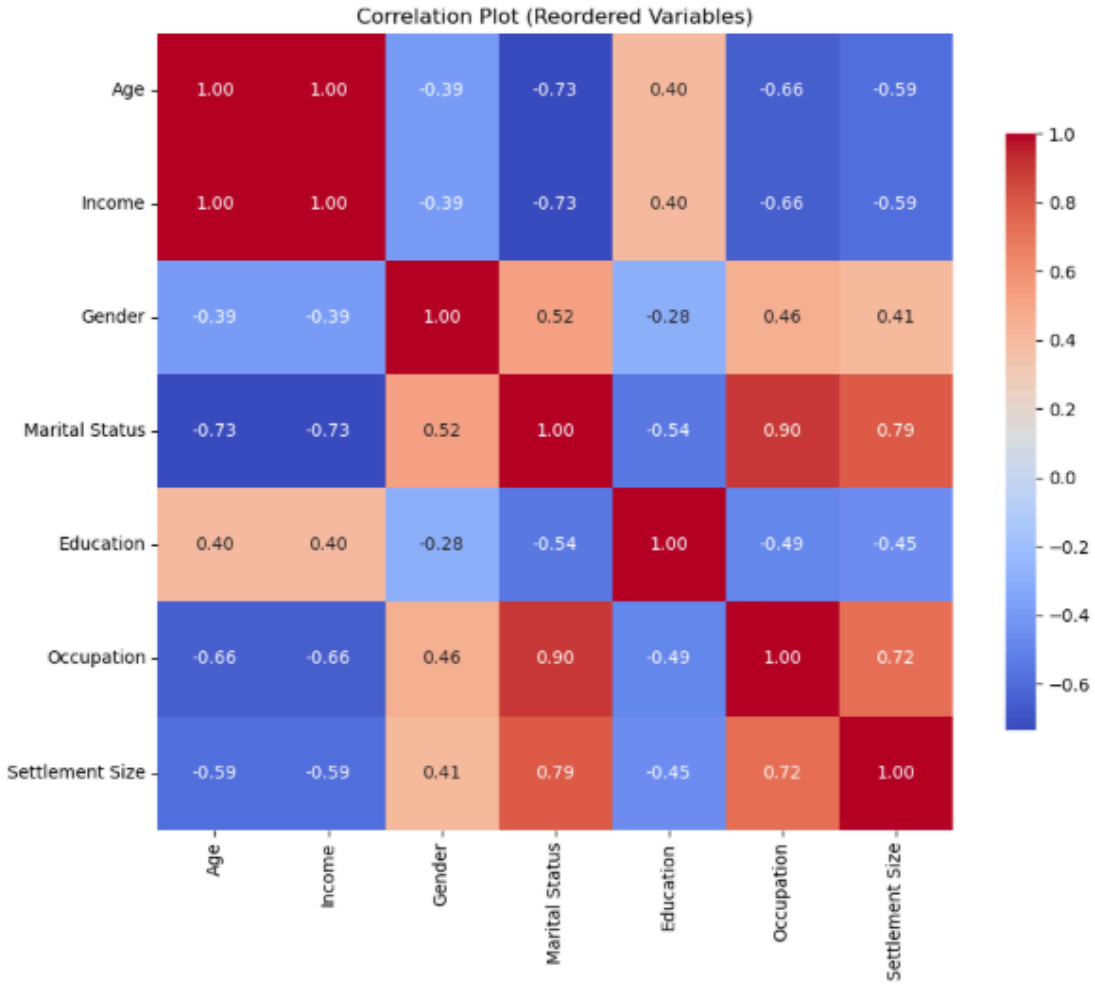

Figure 6: Correlation Matrix

3. Customer Segmentation and Naming

3.1 Standardisation

Before clustering, I standardised age and income using z-score transformation to give both variables equal weight. This is important for distance-based methods like K-Means++ and Agglomerative Clustering, as features on different scales can distort the results. I applied StandardScaler from scikit-learn.

3.2 Determining the Optimal Number of Clusters

I applied the Elbow Method and Silhouette Analysis to determine the optimal number of clusters.

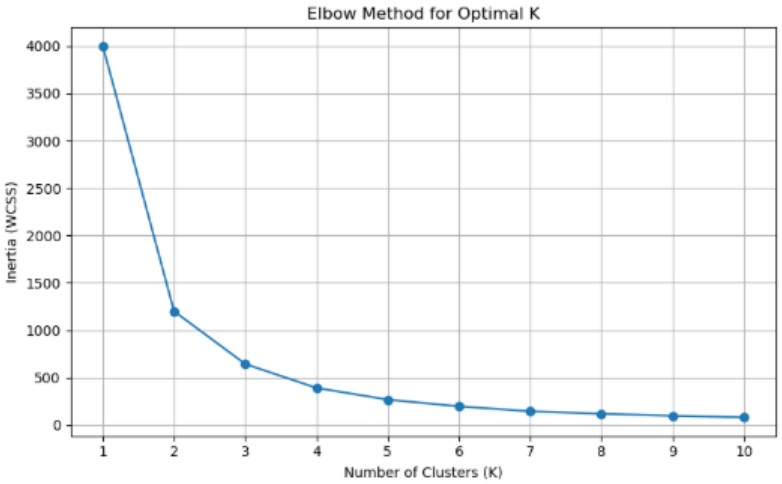

Figure 7: Elbow Method

The WCSS curve shows a clear elbow at K = 3, indicating a good trade-off between model complexity and fit.

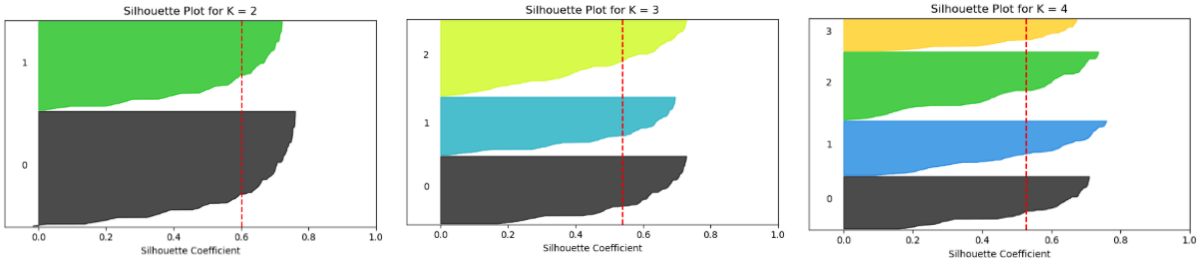

Figure 8: Silhouette Plots

K = 2 had the highest average score, but one cluster dominated. K = 4 resulted in uneven groups. K = 3 produced balanced, well-separated clusters. Based on these visual and numerical cues, I selected **K = 3**.

3.3 K-Means++ Clustering

Using K = 3, I applied K-Means++ to assign members into clusters based on standardised numerical variables and encoded categorical features.

KMeans_Cluster	Age	Income	Gender	Marital Status	Education	Occupation	Settlement Size	Count
0	30.717	88189.256	Male	Single	High School	Unemployed/Unskilled	Small City	671
1	52.709	195566.239	Female	Non-single	University	Skilled Employee/Official	Big City	577
2	40.722	136988.939	Female	Non-single	University	Unemployed/Unskilled	Small City	752

Table 1: Summary of K-Means++ Cluster Profiles

Cluster 0 contains younger, low-income members. Cluster 1 includes older, affluent professionals. Cluster 2 comprises middle-aged members with moderate income and varied backgrounds. The clusters are well-distributed in size and will be interpreted further in Section 4.

3.4 Agglomerative Clustering

I also used Agglomerative Clustering with K = 3 for comparison. This hierarchical method begins with each point as a cluster and merges them using Ward's linkage, which minimizes internal variance.

Agglo_Cluster	Age	Income	Gender	Marital Status	Education	Occupation	Settlement Size	Count
0	48.401	174516.608	Female	Non-single	University	Skilled Employee/Official	Big City	1031
1	30.717	88189.256	Male	Single	High School	Unemployed/Unskilled	Small City	671
2	37.366	120573.084	Male	Single	High School	Unemployed/Unskilled	Small City	298

Table 2: Summary of Agglomerative Clustering Results

While the same number of clusters was used, the distribution differs. Cluster 0 includes nearly half the members, while Clusters 1 and 2 are smaller. The overall groupings align broadly with K-Means++, but Agglomerative is less balanced. Still, it reveals subtle patterns that could be useful for deeper segmentation.

4. Segment Interpretation and Comparison

4.1 Interpretasi & Naming

K-Means++ produced three well-separated and balanced segments:

- **Cluster 0 – Young Budget-Conscious**: Younger, low-income members with limited education and unskilled jobs.
- **Cluster 1 – Affluent Professionals**: Older, high-income professionals with strong educational backgrounds.
- **Cluster 2 – Stable Mid-Level Members**: Middle-aged members with moderate income and diverse attributes.

Cluster	Segment Name	Description
0	Young Budget-Conscious	Younger members with low income, mainly high school graduates and unskilled workers
1	Affluent Professionals	Older members with high income, strong educational background, and professional jobs
2	Stable Mid-Level Members	Middle-aged individuals with moderate income and diverse attributes

Table 3: K-Means++ Clustering Segments

Agglomerative Clustering revealed similar group types but with less balance:

- **Cluster 0 – Value-Oriented Members**: Large group with low to moderate income.
- **Cluster 1 – Split Affluent Group**: Smaller group of older, high-income professionals.
- **Cluster 2 – Mixed Mid-to-High Earners**: Diverse mid-aged members with varied traits.

Cluster	Segment Name	Description
0	Value-Oriented Members	Large group with young to mid-aged members earning low to moderate income
1	Split Affluent Group	High-income older professionals, similar to K-Means Cluster 1, but smaller in size
2	Mixed Mid-to-High Earners	Diverse group with mid to high income and varied occupations and education

Table 4: Agglomerative Clustering Segments

4.2 Cross-tab Comparison

Table 5 shows K-Means Cluster 0 aligns with Agglo Cluster 1, and Cluster 1 maps fully to Agglo Cluster 0. Cluster 2 overlaps with Agglo Clusters 0 and 2. This suggests Agglomerative groups diverse members together, while K-Means++ offers clearer separation.

Agglo_Cluster	0	1	2
Cluster_KMeans			
0	0	671	0
1	577	0	0
2	454	0	298

Table 5: Cross-Tabulation of K-Means++ vs Agglomerative Clustering

These overlaps show that both methods identify similar segment types, but Agglomerative tends to merge diverse members into fewer large groups, while K-Means++ gives clearer, more actionable separation.

4.3 Visual Comparison

Figure 9 compares clusters by age and income. K-Means++ shows three well-separated groups, while Agglomerative forms one dominant and two less clear clusters. This supports earlier findings: K-Means++ gives better segmentation for targeting, while Agglomerative merges varied members into broader groups.

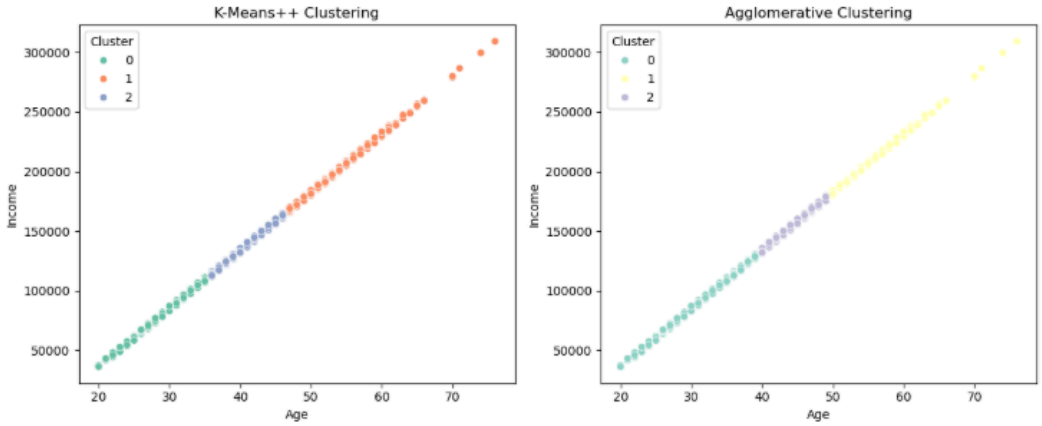

Figure 9: Scatterplot Comparison of Clustering Results

This visual supports the earlier cross-tab findings: Agglomerative tends to merge more diverse members into one large cluster, while K-Means++ provides clearer, more actionable separation.

5. Marketing Recommendations

Based on the K-Means++ clusters, the following strategies match each group's traits to improve engagement and retention.

Segment 0 – Young Budget-Conscious

Students and casual workers with low income and high digital use.

- Introduce app-based plans under $20/week
- Partner with youth-focused brands for discounts
- Promote via TikTok or Instagram Reels

Segment 1 – Affluent Professionals

Older, high-income members who value quality.

- Offer annual premium packages with PT and wellness
- Provide a dashboard to track progress
- Use milestone-triggered email reminders

Segment 2 – Stable Mid-Level Members

Routine-driven, family-focused members.

- Design flexible family plans
- Run short health challenges with rewards
- Add simple perks like free physio consults

These strategies are practical, data-driven, and tailored to what each group values.

6. Conclusion

This report used a straightforward approach to group gym members based on things like age, income, education, and occupation. After looking into the data, I applied K-Means++ and Agglomerative Clustering to find three segments with different characteristics.

K-Means++ gave better and more balanced clusters, which makes it easier to apply in business settings. Based on the results, I suggested clear and relevant marketing strategies for each group, depending on their profile and lifestyle.

Moving forward, the segmentation could be improved by including behaviour-based data, such as how often members visit the gym or which programs they attend. This would allow the gym to personalise services more effectively and improve engagement in the long term.