# Introduction

The problem being addressed is to create marketing strategies for a national gym chain through analysing 2,000 membership records. My approach is to first perform exploratory analysis on all variables within the dataset using tables and plots. Then determine the optimal number of customer segments through the elbow method and analysing three silhouette plots. Finally, I will undertake customer segmentation through both Kmeans++ and agglomerative clustering methods. This achieving the ability to create targeted recommendations for marketing strategies based on these segments.

Exploratory Data Analysis

Age

A histogram of members age shows that the distribution of the variable is moderately right skewed, with a tail extending to approximately 75 years of age. This indicates that marketing to ages 60-75 may be ineffective in creating value, this also applying to ages 20-25.

Histogram 1: Age Distribution
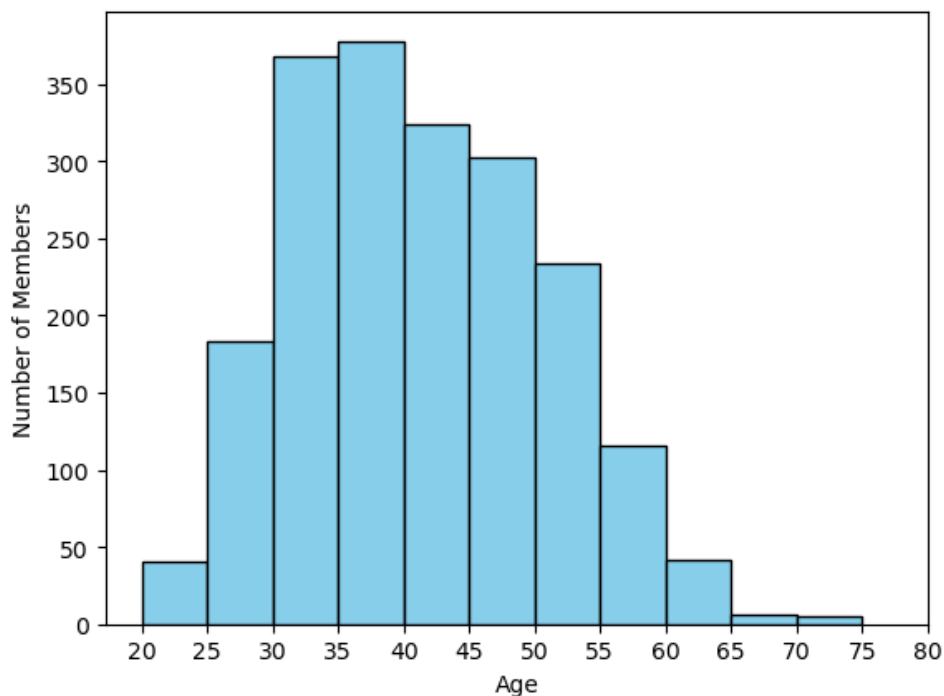
Analysing the variable further through summary statistics, the average age of a member is 41, with most falling between 33 (lower quartile) and 48 (upper quartile). This supports and highlights the necessity of targeting this range age groups within marketing efforts.

	Mean	Standard deviation	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Age	40.824	9.456	20.000	33.000	40.000	48.000	76.000

Income

The histogram of member income shows a somewhat bell-shaped distribution however is considerably skewed to the right, having a tail extending to $325,000. Suggesting marketing to high income outliers may be unnecessary. However, annual income is considerably spread out amongst their members.

Histogram 2: Income Distribution
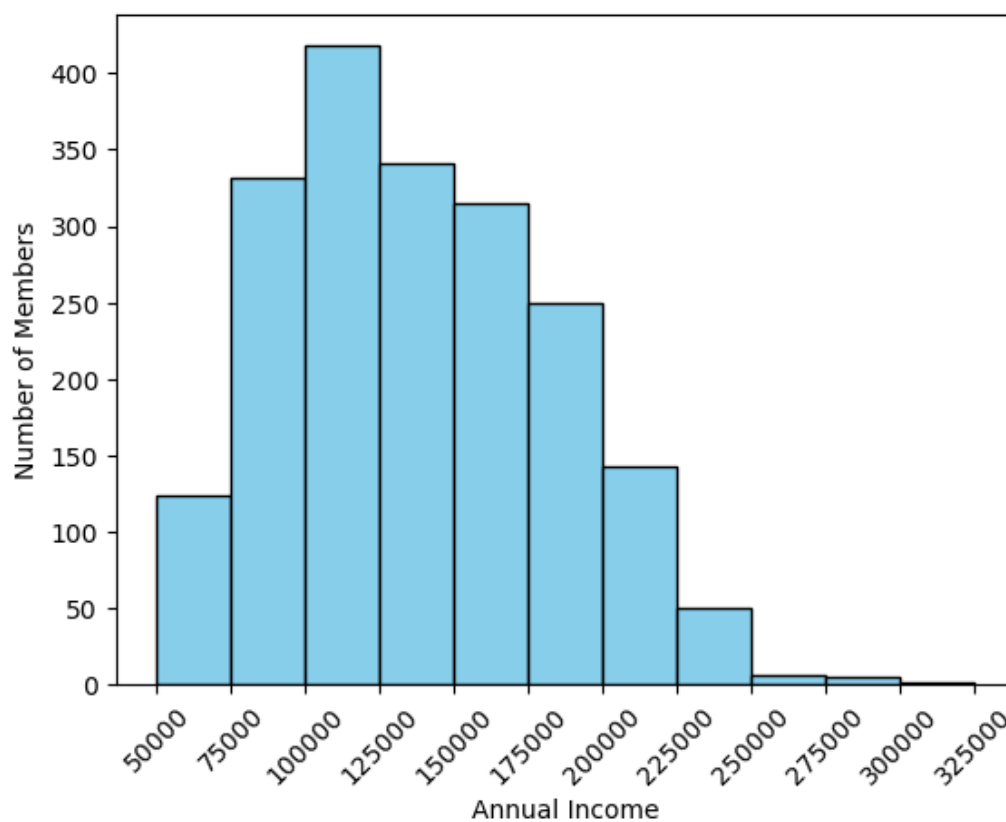
This is supported by summary statistics, showing that members average $137516.196 annually, with most falling between $171232.500 (upper quartile) and $101262.750 (lower quartile). This quantitatively showing this considerably large range of annual income. Evidently showing different marketing strategies for different income levels would be appropriate, within the $100,000 to $175,000 range

Table 2: Income Summary Statistics

	Mean	Standard deviation	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Income	$137,516.196	$46,184.297	$35,832.000	$101,262.750	$133,004.000	$171,232.500	$309,364.000

Gender

Through our pie chart it is evident that 60.450% of the gym chains customer base is female, meaning 39.550% is male. Demonstrating the dominant demographic of women within the gym chain nationally and communicates that marketing efforts should not primarily align with masculine stereotypes many gym chains commonly rely on.

Pie Chart 1: Proportion of Gender
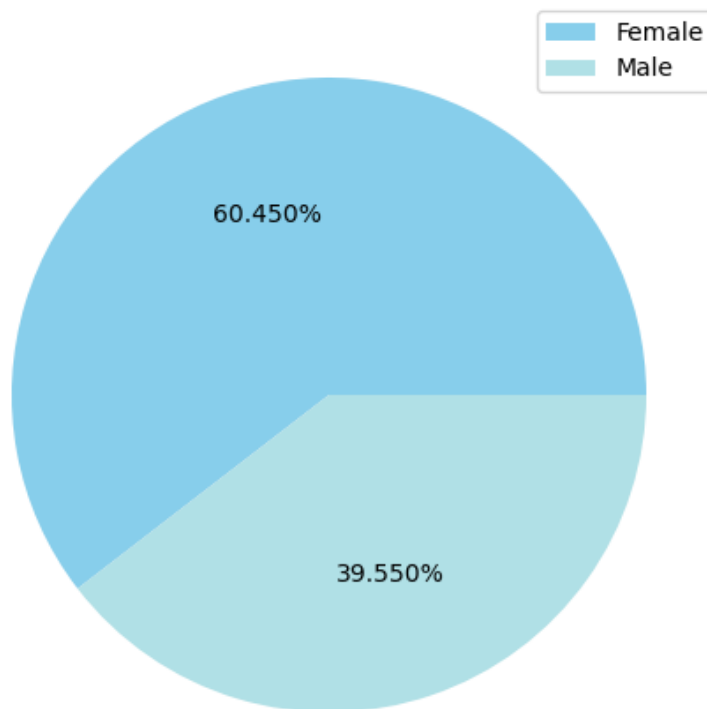
Education

Sample indicates that 43.850% of their customer base has a high school education level, 37.850% have university level education, 8.700% with graduate level education, and 9.600% with other/unknown education status. This presents an opportunity to provide educational opportunities related to fitness, that could be effectively utilized within marketing campaigns targeted to members with only high school level education.

Pie Chart 2: Proportion of Education Level
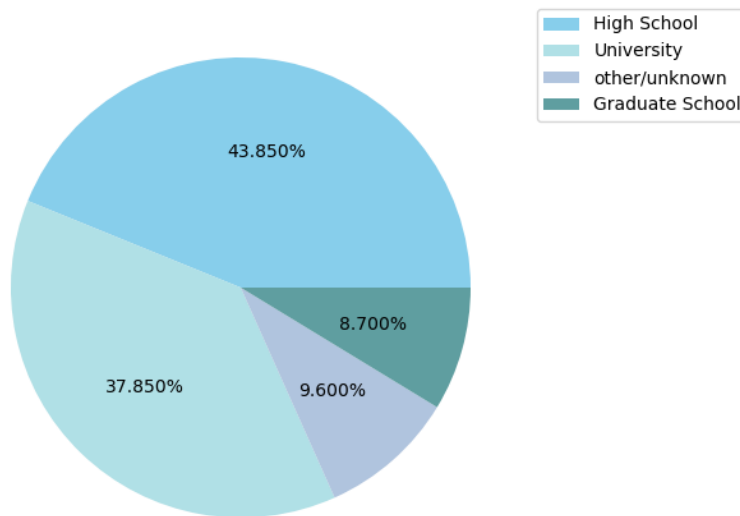
Occupation

49.600% of members are unemployed, 39.550% are skilled employees/ officials, and 10.850% are in management roles or self-employed. Highlighting an opportunity to take the large percentage of unskilled members, and market to them through providing an avenue to potentially gain employable skills.

Pie Chart 3: Proportion of Employment Status
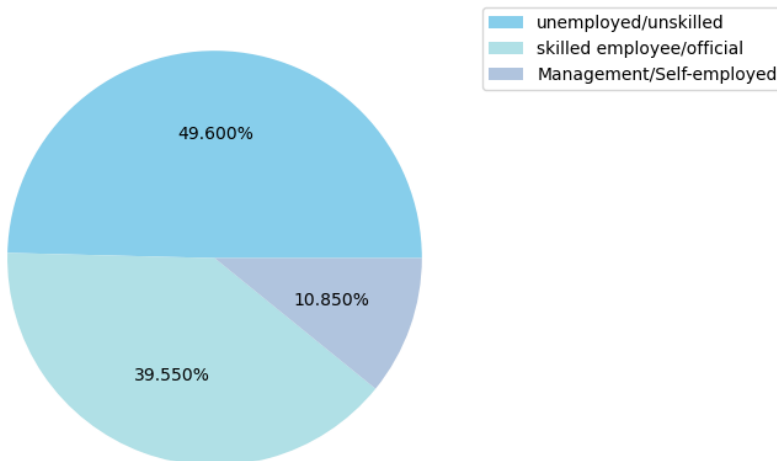
Settlement Size

56.600% of members live within a small city, 39.900% live within a big city, and 3.6% live within a mid-sized city. Showing two majorly different marketing environments in small and big cities that the business should have in consideration when creating campaigns, and one they should not prioritize in mid-sized cities.
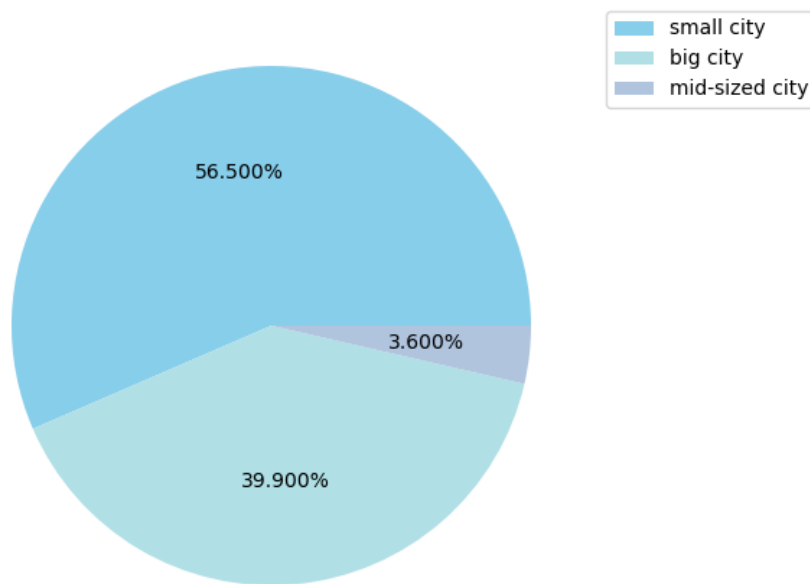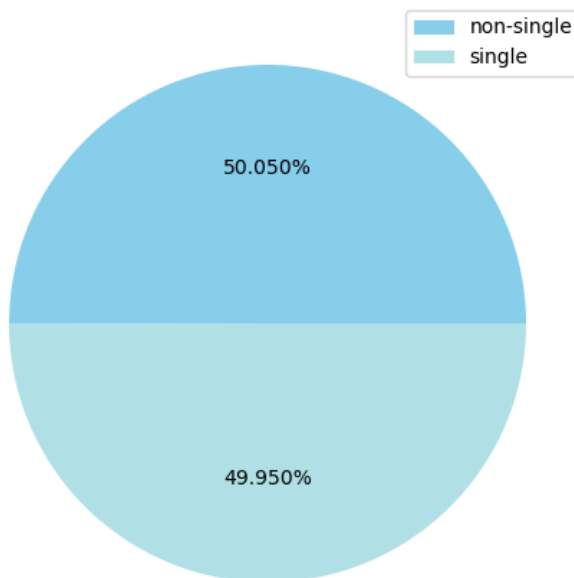
Marital Status

50.050% of members are non-single (widowed, divorced, separated or married), while 49.950% of members are single. This meaning marketing efforts targeted to both single and non-single members are equally as appropriate and should be balanced accordingly

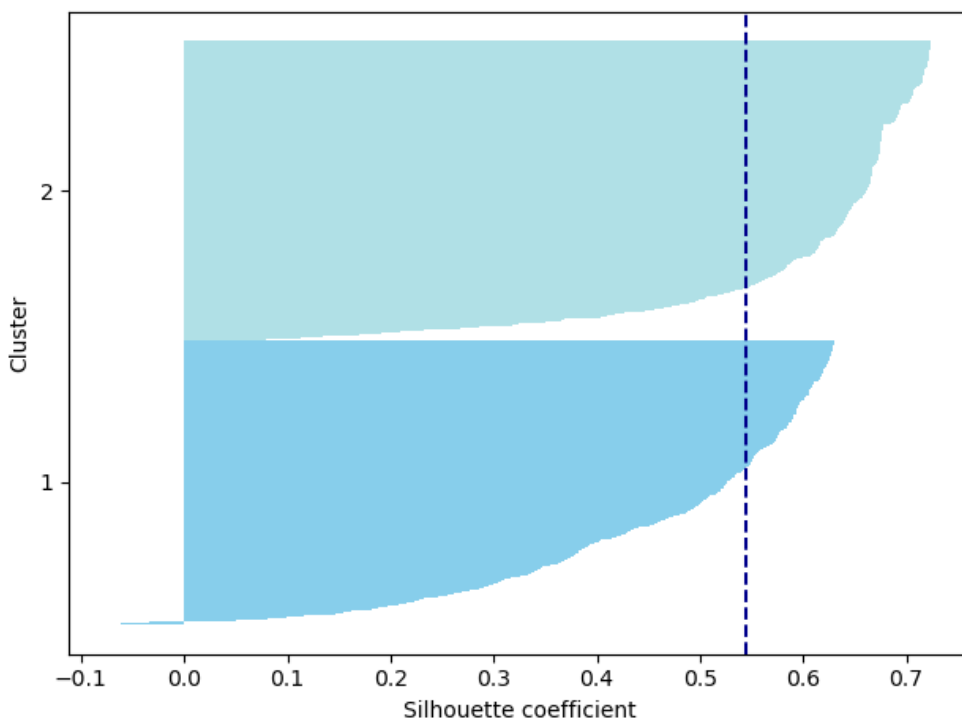*Pie Chart 5: Proportion of Member Marital Status*

Customer Segmentation

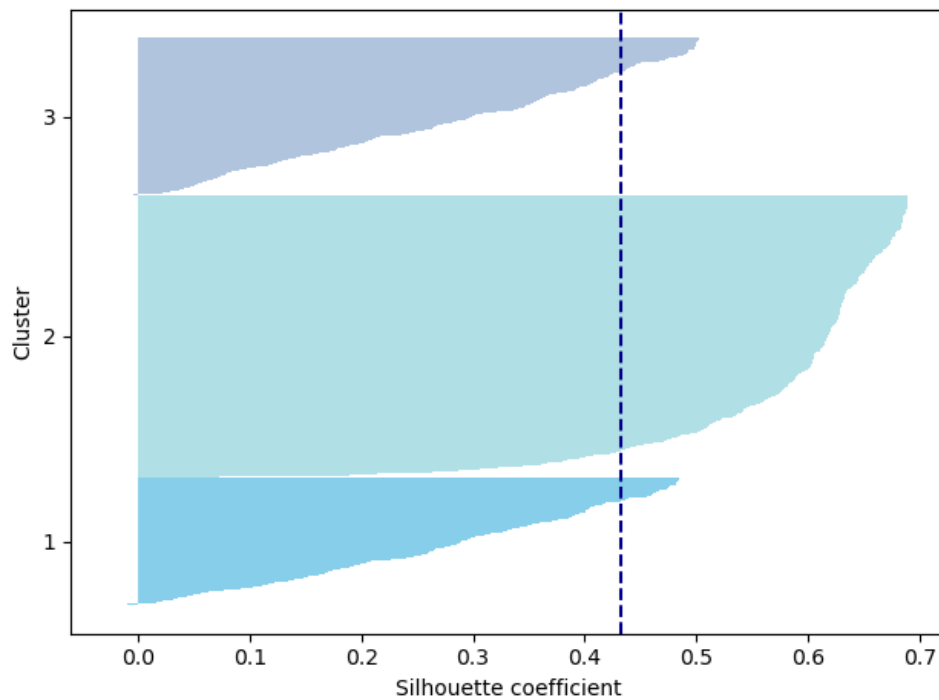*Inertia Plot: Within Cluster SSE of Different Cluster Amounts*
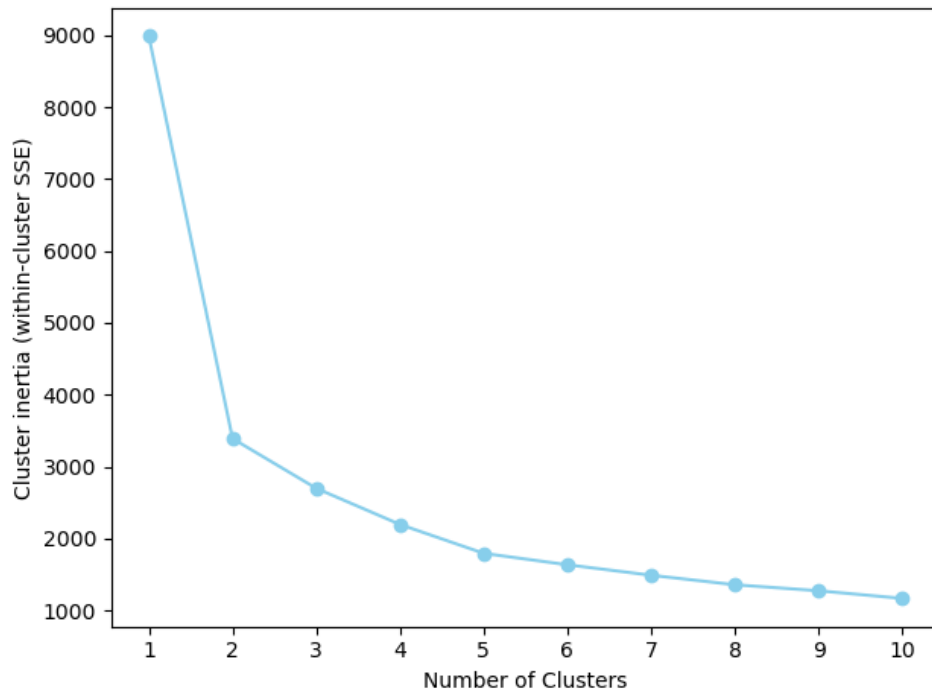
The inertia plot detailing within cluster SSE suggests an elbow point at 3 clusters. However, this is hard to say with 100% certainty as cluster SSE declines at considerable rates from clusters 2-4.
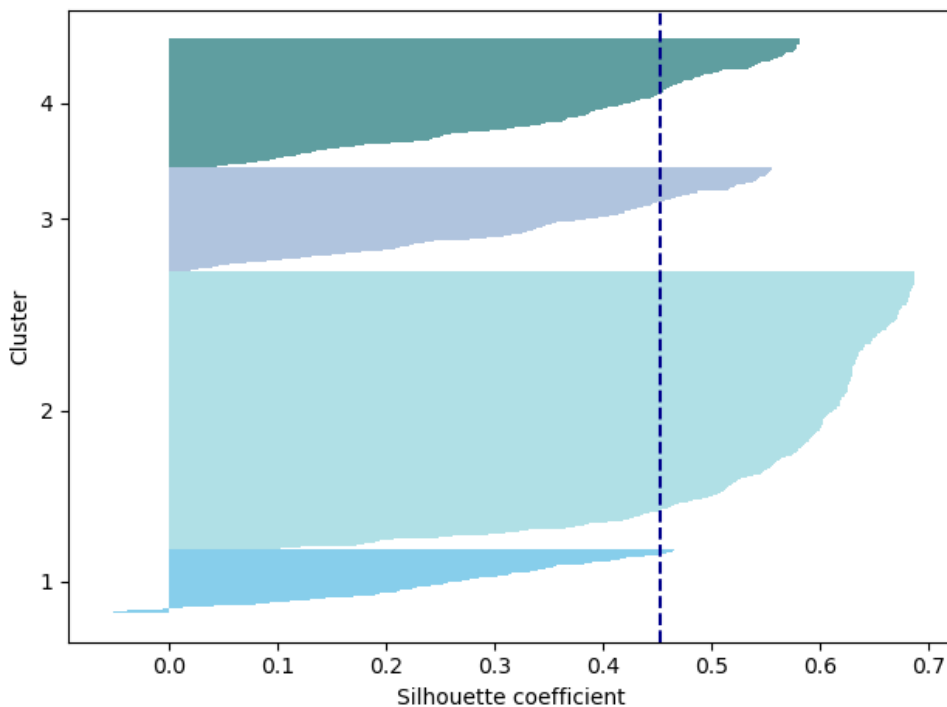
Silhouette Plot: 2 Clusters

Silhouette Plot: 3 Clusters

Silhouette Plot: 4 Clusters

Comparing the three silhouette plots confirms that two clusters provide the optimal cohesion and separation, having the highest silhouette average of approximately 0.55. In comparison, three- and four- cluster solutions have lower average score in the 0.4 – 0.5 range.

Table 3: Kmeans++ Cluster Centres

Cluster	Gender	Marital Status	Age	Education	Income	Occupation	Settlement Size	Number of Customers
1	Female	Non-Single	48.174	University	173443.318	Skilled Employee / Official	Big-City	976.000
2	Male	Single	33.831	High School	103341.616	Unemployed / Unskilled	Small City	1024.000

Table 4: Agglomerative Cluster Centres

Cluster	Gender	Marital Status	Age	Education	Income	Occupation	Settlement Size	Number of Customers
1	Female	Non-Single	47.081	University	168085.082	Skilled Employee / Official	Mid-Sized City	1163.000
2	Male	Single	32.129	High School	95041.151	Unemployed / Unskilled	Small City	837.000

Segmentation Interpretation and Naming

Kmeans++ Segments and Attributes

Mature wealthy urban women

- Non-single middle-aged females with university education and high income, located within big cities

Unskilled suburban adult men

- Single, unskilled high school educated adult males with below average income (relative to other members), living within small cities

Agglomerative Segments and Attributes

Mature wealthy women in towns

- Non-single middle-aged females with university education and high income, located within mid-sized cities

Unskilled suburban adult men

- Single, unskilled high school educated adult males with below average income (compared to other members), living within small cities

Overlap: There is significant overlap between agglomerative and Kmeans++ segments, forming the same segment 'Unskilled suburban adult men'. The techniques create different segments between 'Mature wealthy urban women' and 'Mature wealthy women in towns'. It is important to note that the number of customers also differs completely between all clusters.

Marketing Recommendations

Segment: Midlife Urban wealthy women

Recommendations

1. Billboard Marketing in High Traffic City Streets
2. Partnering and selling discounted products to members from premium fitness brands
3. Provide and promote classes focusing on less intensive workout styles such as Yoga or Pilates

Segment: Unskilled suburban adult men

Recommendations

1. Advertising personal trainer training to loyal members (members can gain skills to become employed)
2. Sponsoring suburban sports teams (Jersey placements)
3. Pitchside advertising boards for local sporting events

Conclusion

This report conducted exploratory data analysis on all variables within the gym chain's dataset, identifying key demographic attributes of their members through visualizations and summary statistics of age, income, education, occupation, settlement size and marital status. To identify appropriate customer segments, I analysed both inertia and silhouette plots, concluding that segmenting their members into two clusters supported optimal cohesion and separation of segments. I applied both Kmeans++ and agglomerative clustering methods to identify and interpret these segments. Based on a Kmeans++ cluster centre summary table, I named these segments 'Mature Wealthy Urban Women' and 'Unskilled Suburban Adult Men'. Finally, I proposed targeted marketing recommendations based upon this analysis.