46862897_BUSA3020_Customer Segmentation Report

Understanding customer behavior is crucial for effective marketing and customer satisfaction. This analysis uses a dataset from 4,000 supermarket loyalty cardholders, including demographics and socio-economic variables like age, gender, marital status, education, occupation, settlement size, and income. We apply cluster analysis techniques, particularly K-means++ and Agglomerative clustering, to segment customers. The process involves data preprocessing, PCA, and clustering, to identify distinct customer groups and enhance marketing strategies and customer loyalty.

1. Dataset Overview

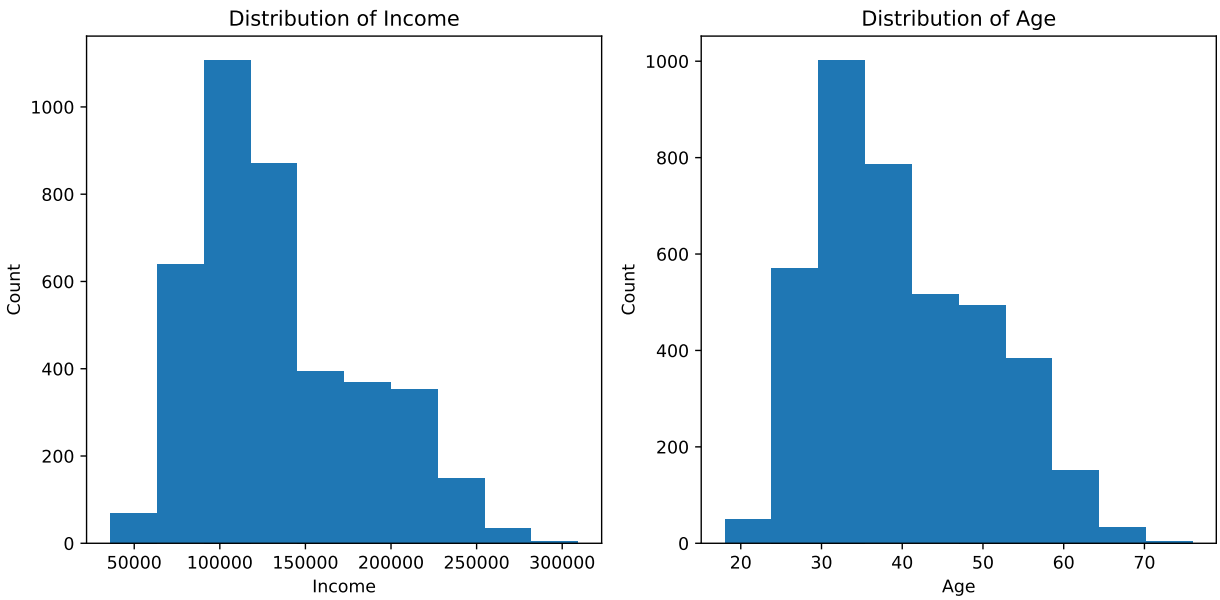
The dataset comprises 4000 observations distributed across 7 columns, with 7 duplicates and no missing values It is crucial to remove duplicates to avoid impacting the quality, accuracy, and reliability of the data, and lead to inaccurate results in later analysis.

2. Exploratory Data Analysis (EDA)

```
##             Gender  Marital Status    Education  Settlement Size   Occupation  \
## count  3993.000000     3993.000000  3993.000000      3993.000000  3993.000000
## mean      0.489356        0.510644     1.707989         1.089657     1.200601
## std       0.499949        0.499949     1.024196         0.869476     0.526435
## min       0.000000        0.000000     0.000000         0.000000     0.000000
## 25%       0.000000        0.000000     1.000000         0.000000     1.000000
## 50%       0.000000        1.000000     1.000000         1.000000     1.000000
## 75%       1.000000        1.000000     3.000000         2.000000     2.000000
## max       1.000000        1.000000     3.000000         2.000000     2.000000
##
##               Income          Age
## count    3993.000000  3993.000000
## mean   134366.283496    39.949662
## std     48533.105006    10.269135
## min     35832.000000    18.000000
## 25%     97816.000000    32.000000
## 50%    122635.000000    38.000000
## 75%    165904.000000    47.000000
## max    309364.000000    76.000000
```

Analysis:

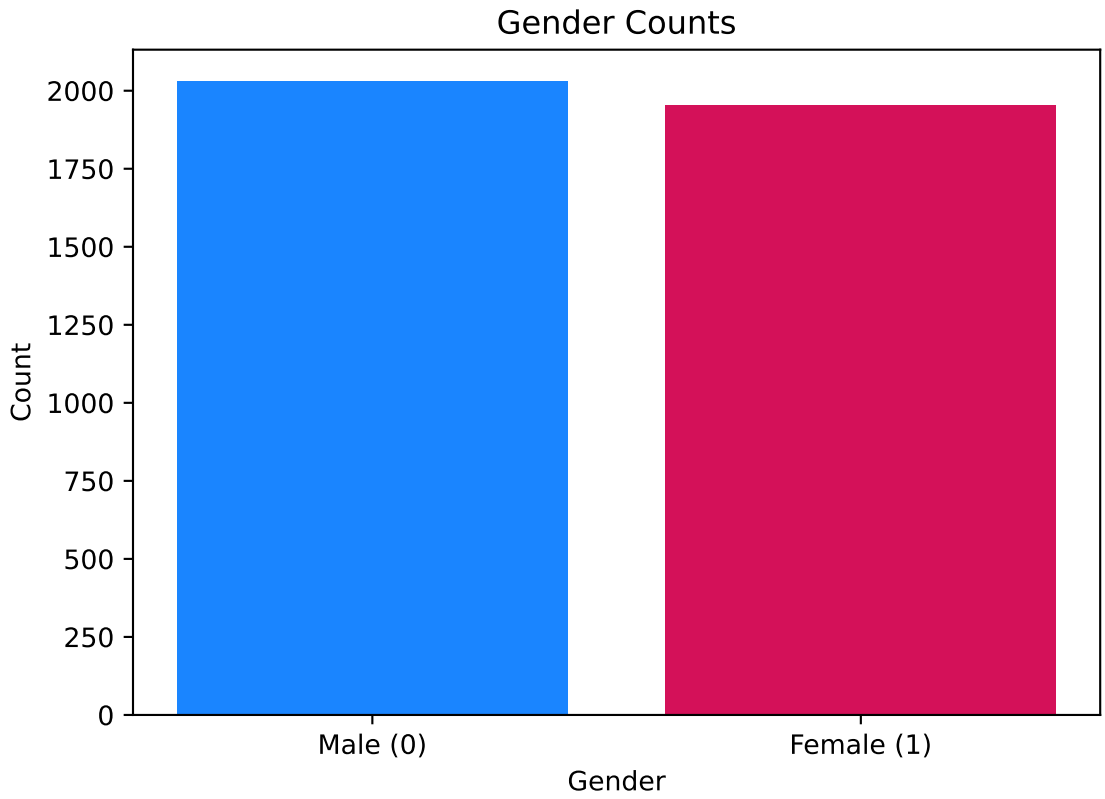
- Income: 75% of observations are under 165,904, but the highest is 309,364.
- Age: 75% are under 47, but the highest is 76.



Income and Age has normal and unimodal distributions although it appears to be slightly left-skewed.

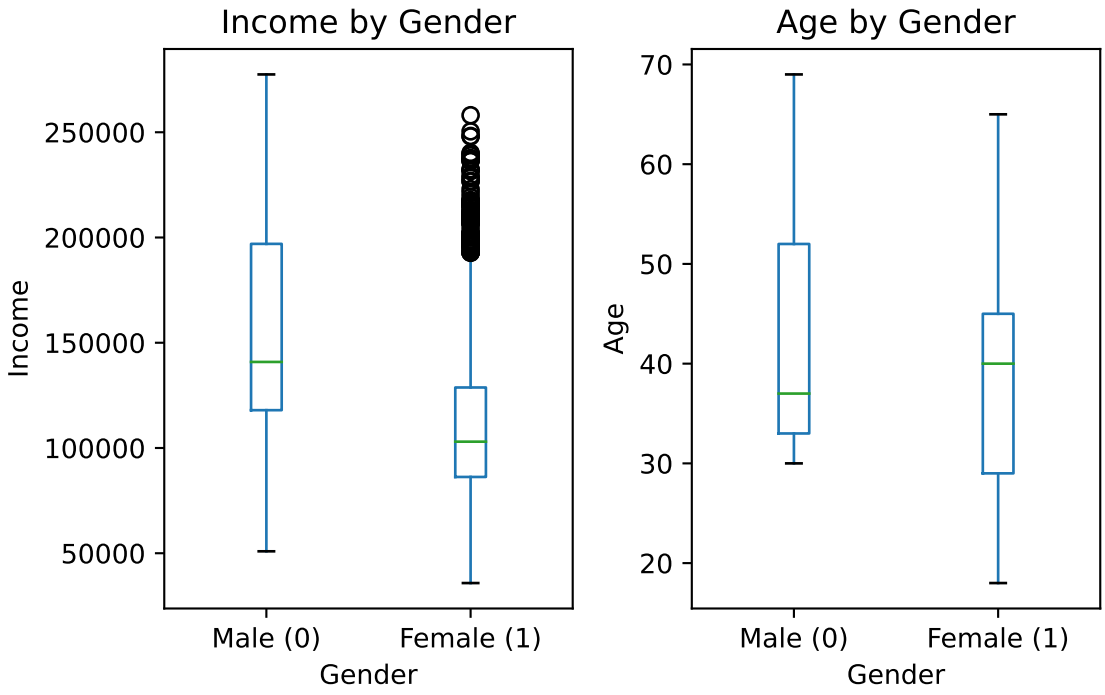
2.1 Visualization

```
## ([<matplotlib.axis.XTick object at 0x000001DFD8677BE0>, <matplotlib.axis.XTick object at 0x000001DFD8
```



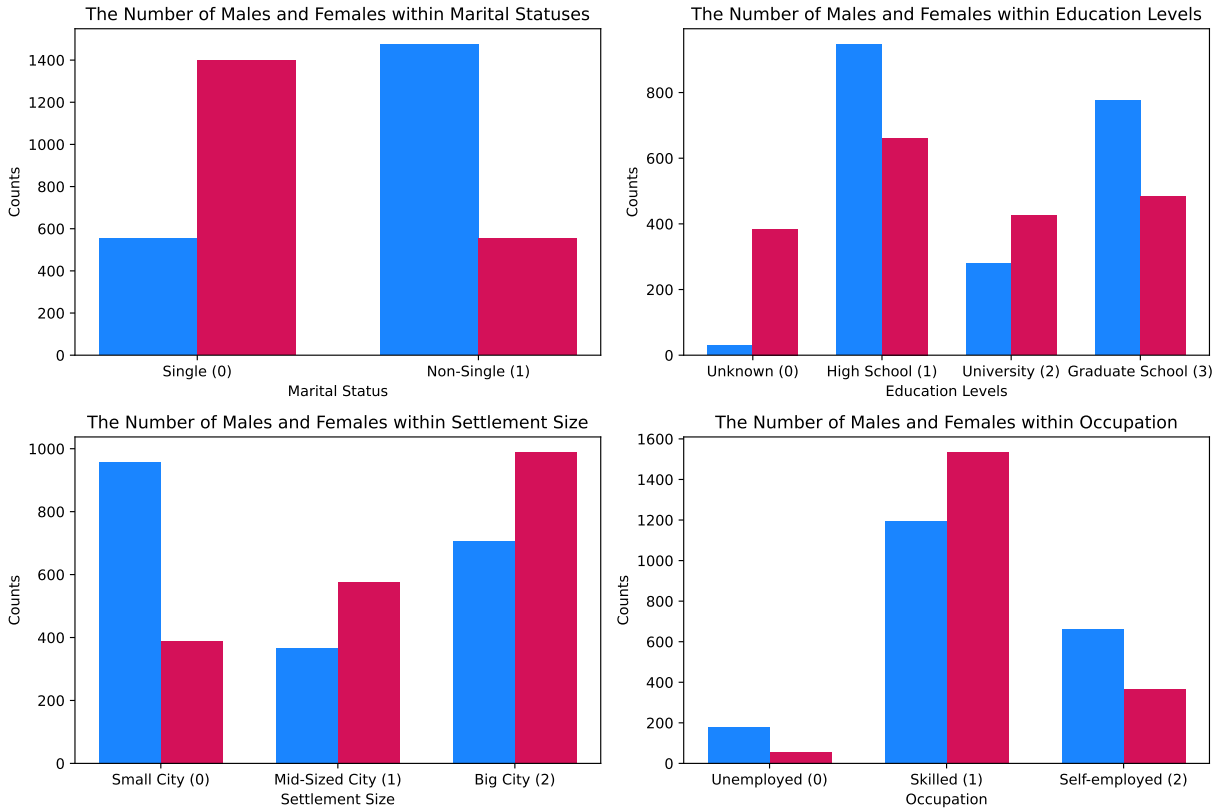
```
## 0    2030
## 1    1954
## Name: Gender, dtype: int64
```

Analysis: The number of male consumers are roughly equal to that of females, with 50.95% males and 49.05% females buying items in their supermarket chain.



Analysis:

- Men have a wider range and higher median income, indicating varied earnings with generally higher amounts.
- Women show more extreme income values despite a narrower range. Women also have a wider age range and a higher average age compared to men.



Analysis: Single females nearly double single males; non-single females are half of males. Females with unknown education outnumber males; more females at university level. Males predominate in small cities; females slightly more in mid-sized and big cities. More female skilled workers; more male unemployed and self-employed.

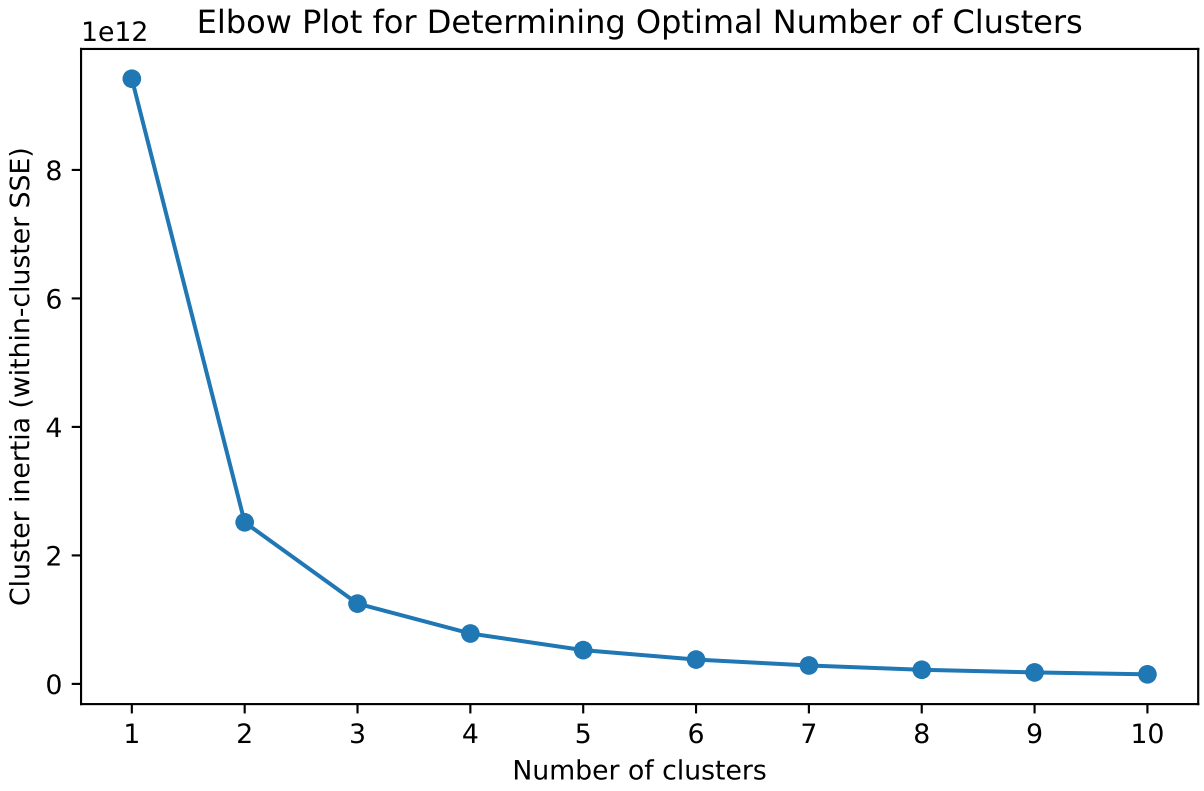
3. Clustering

3.1. K-Means++ Clustering

K-means clustering organizes data into groups by refining centroids iteratively, like assigning students to teams based on similar skills. KMeans++ improves results by better centroid selection despite higher computational cost.

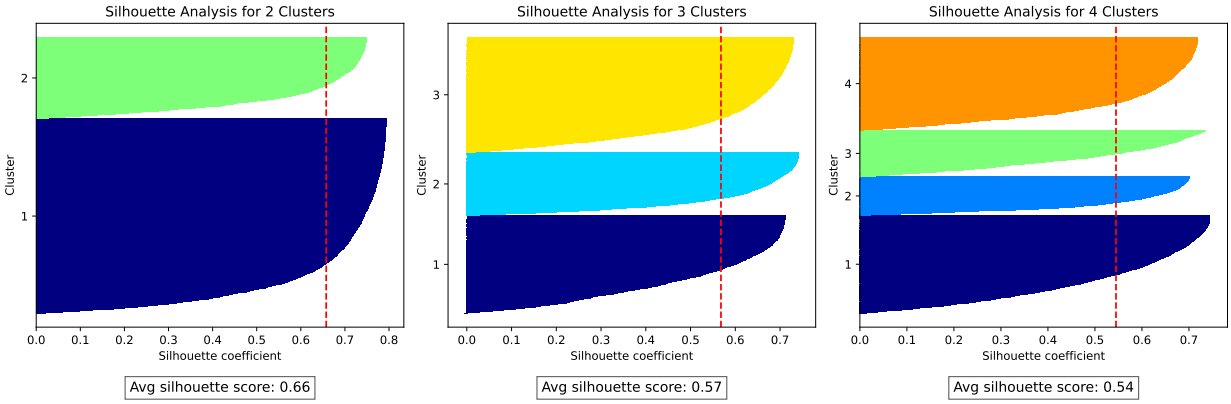
```
## KMeans(n_clusters=1, n_init=10, random_state=0)
## KMeans(n_clusters=2, n_init=10, random_state=0)
## KMeans(n_clusters=3, n_init=10, random_state=0)
## KMeans(n_clusters=4, n_init=10, random_state=0)
## KMeans(n_clusters=5, n_init=10, random_state=0)
## KMeans(n_clusters=6, n_init=10, random_state=0)
## KMeans(n_clusters=7, n_init=10, random_state=0)
## KMeans(n_init=10, random_state=0)
## KMeans(n_clusters=9, n_init=10, random_state=0)
## KMeans(n_clusters=10, n_init=10, random_state=0)
```

```
## ([<matplotlib.axis.XTick object at 0x000001DFDBE74D90>, <matplotlib.axis.XTick object at 0x000001DFDB
```



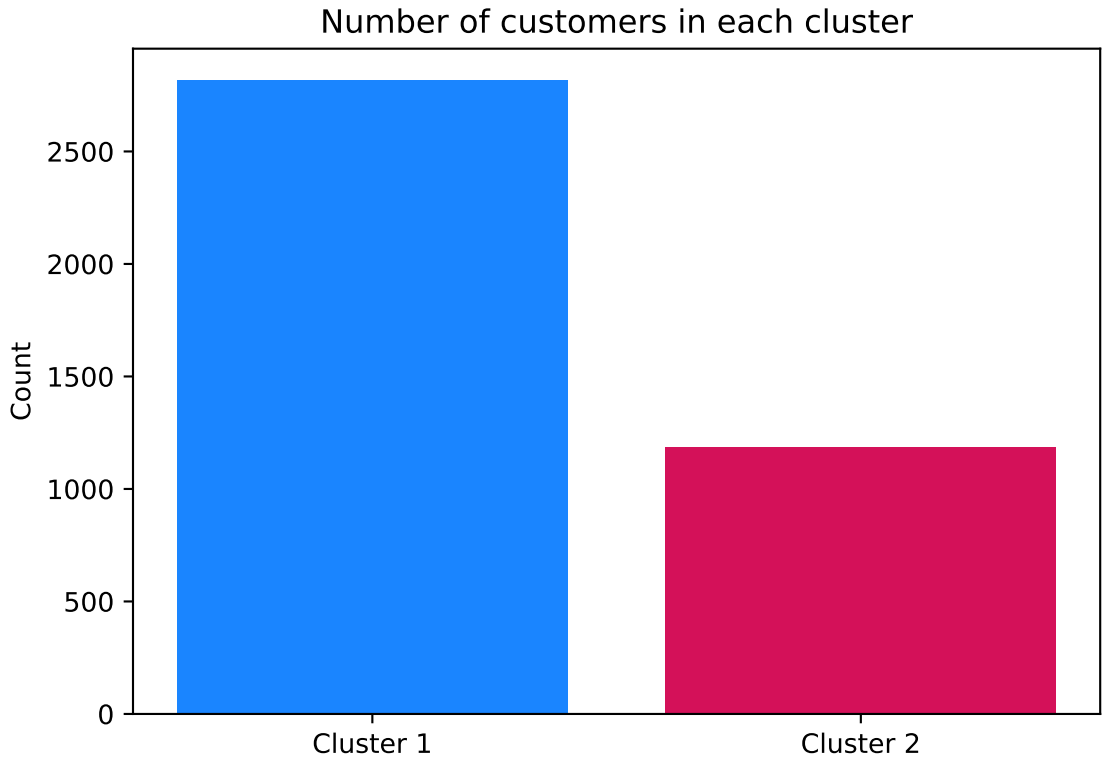
The elbow method is used to determine the optimal number of clusters by graphing data fit across different group numbers to identify the point where the improvement slows down. From the graph, there are 3 possible values (2,3, and 4).

The silhouette method evaluates cluster quality by measuring data point similarity within its cluster compared to others. Higher silhouette values indicate better clustering. The 3 values will be plotted using the method to determine how many clusters would result in better clustering.



Analysis: The 2-cluster has the highest average silhouette score (0.66), followed by 3-cluster (0.57) and 4-cluster (0.54). ***Hence, there are 2 clusters in KMeans++.***

```
## ([<matplotlib.axis.XTick object at 0x000001DFE4412590>, <matplotlib.axis.XTick object at 0x000001DFE
```




```
## 1    2816
## 2    1184
## Name: Cluster, dtype: int64
```

Analysis: The number of customers in Cluster 1 (2816 customers) is significantly higher than Cluster 2 (1184 customers), with a difference of 632 customers. This suggests that Cluster 1 is the dominant cluster, representing a larger portion of the customer base.

To analyse the meanings of each cluster using KMeans++, a descriptive statistic should be conducted.

```
## KMeans++ - Cluster 1
```

Description	Mean	Comment
Gender	0.5795	About 57.95% female, 42.05% male.
Marital Status	0.5039	About 50.39% non-single, 49.61% single.
Education	1.317	Average level between high school and university.
Settlement Size	0.9205	Mostly mid-sized city residents.
Occupation	0.9925	Predominantly skilled employees.
Income	107414	Relatively high income.
Age	34.88	Middle-aged individuals.

Conclusion: Overall, Cluster 1 represents a group of predominantly middle-aged, moderately wealthy females with education levels between high school and university living in mid-sized to big cities with skilled employee or official roles and a balanced proportion of single and non-single individuals.

```
## KMeans++ - Cluster 2
```

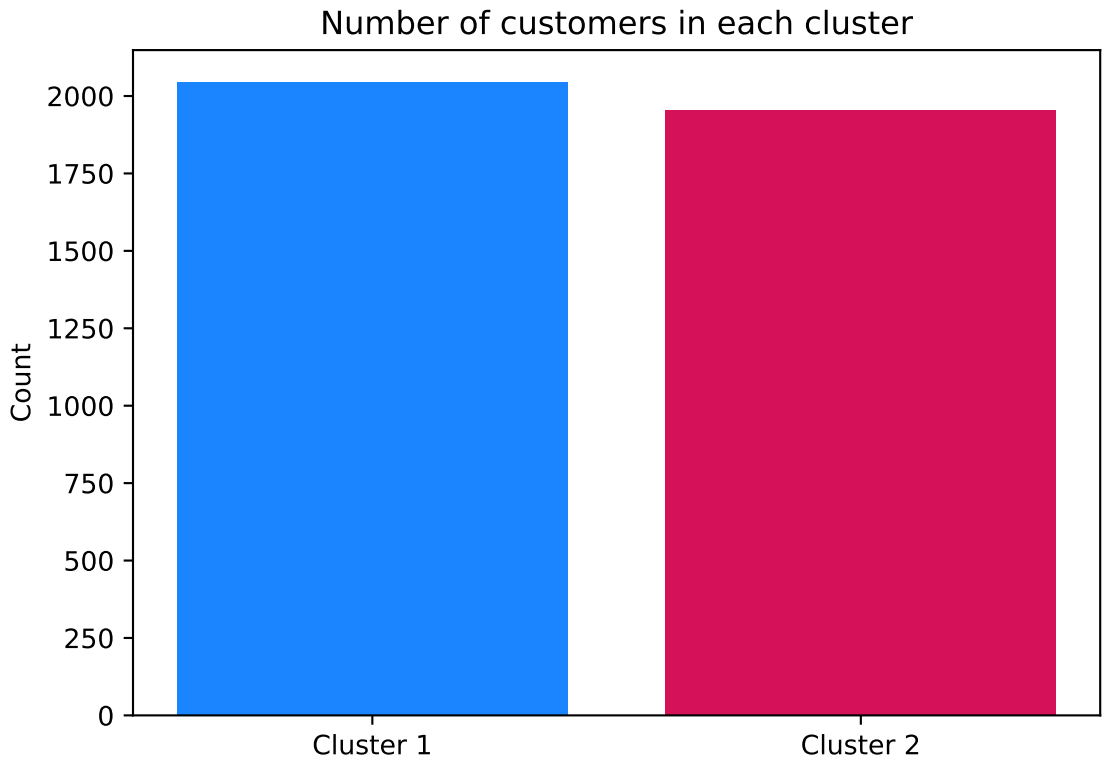
Description	Mean	Comment
Gender	0.2753	About 27.53% female, 72.47% male.
Marital Status	0.5262	About 52.62% non-single, 47.38% single.
Education	2.639	Average level close to graduate school.
Settlement Size	1.493	Mostly big city residents.
Occupation	1.696	Predominantly in management or highly qualified roles.
Income	198426	Relatively high income.
Age	51.99	Older individuals.

Conclusion: Cluster 1 represents predominantly younger, non-single, skilled female employees living in small to mid-sized cities, with education levels around high school to university and lower income levels.

4.2 Agglomerative Clustering

Agglomerative clustering groups similar items into clusters step-by-step. For example, in a library, books are gradually organized into categories like fiction and non-fiction.

```
## ([<matplotlib.axis.XTick object at 0x000001DFE42FA9B0>, <matplotlib.axis.XTick object at 0x000001DFE4
```



```
## 1    2046
## 2    1954
## Name: Cluster, dtype: int64
```

Analysis: The number of customers in Cluster 1 (approximately 2046 customers) is significantly slightly higher than Cluster 2 (around 1954 customers).

To analyse the meanings of each cluster using Agglomerative Clustering, a descriptive statistic should be conducted.

```
## Agglomerative Clustering - Cluster 1
```

Description	Mean	Comment
Gender	0.2918	About 29.18% female, 70.82% male.
Marital Status	0.6618	About 66.18% non-single, 33.82% single.
Education	1.931	Average level between high school and university.
Settlement Size	0.9511	Mostly mid-sized city residents.
Occupation	1.357	Predominantly skilled employees.
Income	171562	Relatively high income.
Age	46.5	Middle-aged individuals.

Conclusion: Overall, Cluster 1 represents a group of predominantly older, wealthy males with university level living in mid-sized to big cities with skilled employee, official, or management roles and a higher proportion of non-single individuals.

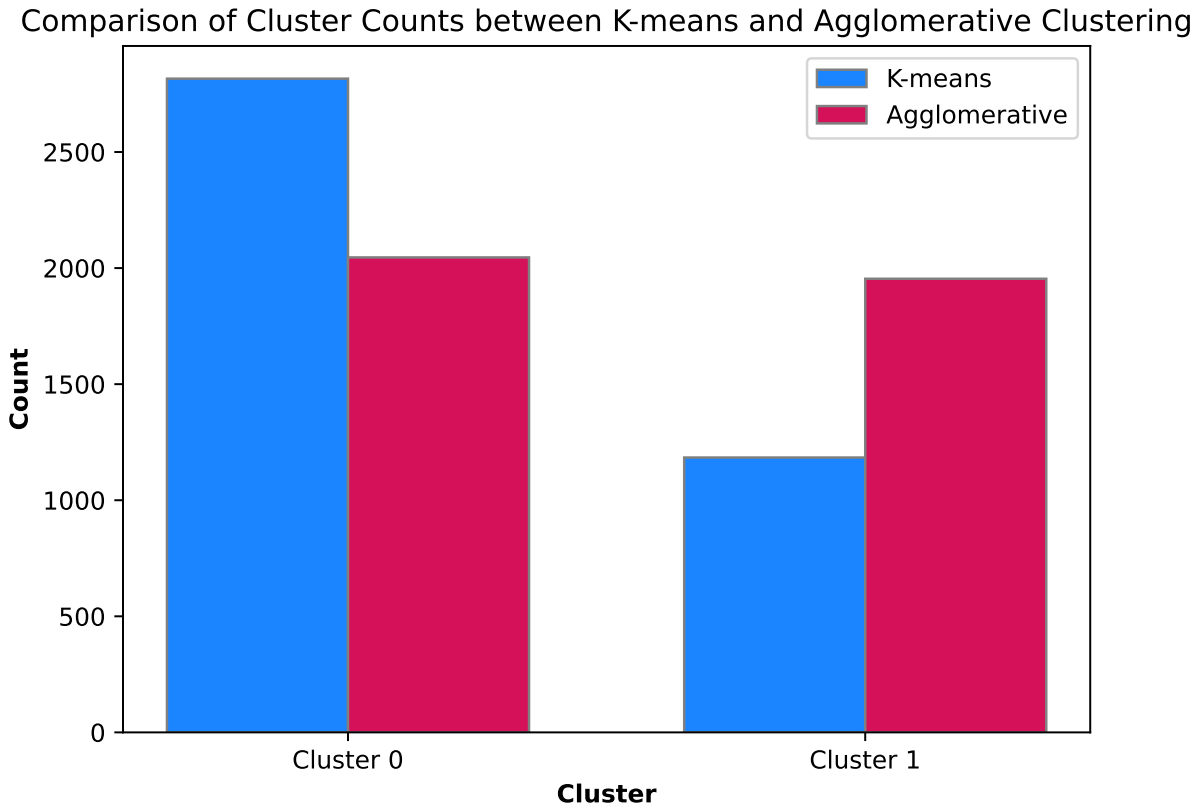
```
## Agglomerative Clustering - Cluster 2
```

Description	Mean	Comment
Gender	0.6965	About 69.65% female, 30.35% male.
Marital Status	0.3521	About 35.21% non-single, 64.79% single.
Education	1.475	Average level between high school and university.
Settlement Size	1.235	Mostly mid-sized city residents.
Occupation	1.037	Predominantly skilled employees.
Income	95394	Moderate income.
Age	33.08	Younger individuals.

Conclusion: Cluster 2 represents a group of predominantly younger, moderately wealthy females with education levels between high school and university living in mid-sized to big cities with skilled employee or official roles and a higher proportion of single individuals

4. K-Means and Agglomerative Clustering Insight Comparison

```
## ([<matplotlib.axis.XTick object at 0x000001DFE44D6140>, <matplotlib.axis.XTick object at 0x000001DFE
```



Revise what we have discovered in the outputs of the two clustering algorithms:

Table 5: Table continues below

Description	KMeans_Cluster1	KMeans_Cluster2
Age	Middle-aged	Older
Education	High school to University	University to Graduate School
Income	Moderately wealthy	Wealthy
Gender	Predominantly females	Predominantly males
Marital Status	Single and non-single	Single and non-single
Settlement Size	Mid-sized to big cities	Mid-sized to big cities
Occupation	Skilled employees	High-level occupations

Agglomerative_Cluster1	Agglomerative_Cluster2
Older	Younger
University	High school to University
Wealthy	Moderately wealthy
Predominantly males	Predominantly females
Non-single individuals	Single individuals
Mid-sized to big cities	Mid-sized to big cities
Skilled employees	Skilled employees

Overlap and Differences:

- K-Means++ Cluster 1 vs. Agglomerative Cluster 1:
  - Overlap: Settlement Size, Occupation
  - Differences: Age, Education, Income, Gender, Marital Status
- K-Means++ Cluster 2 vs. Agglomerative Cluster 2:
  - Overlap: Settlement Size, Occupation
  - Differences: Age, Education, Income, Gender, Marital Status

–> Conclusion

After the data are clustered into 2 different groups, there are obvious distinctions. In general, the first clusters identified by two techniques both have higher number of data points than the second clusters have.

The first cluster identified by Kmeans++ is totally different compared to the first cluster identified by agglomerative clustering. Looking at the comparison table, the Kmeans++ first cluster appears to be the second cluster in agglomerative clustering. This pattern also applies to Kmeans++ second cluster, which is more identical to the first cluster in agglomerative clustering. identified.

6. Suggestions tailored for the two clusters identified by Kmeans++:

6.1. Cluster 1: Middle-aged, Moderately Wealthy Females with Education Levels between High School and University Living in Mid-Sized to Big Cities with Skilled Employee or Official Roles

- Goal: Enhance work-life balance and convenience.
  - Promote time-saving products and services: Offer ready-to-eat meals, meal kits, and quick recipes that fit into their busy schedules.
  - Loyalty Programs: Introduce loyalty programs with exclusive deals on household essentials and personal care products.
  - Community Engagement: Organize local community events or online forums where they can share tips on managing work-life balance and get recommendations for related products.
  - Health and Wellness: Provide newsletters with health tips and promote fitness-related products and services, such as gym memberships or wellness workshops.

6.2. Cluster 2: Older, Wealthy Males with Education Levels between University and Graduate School Living in Mid-Sized to Big Cities with High-Level Occupations

- Goal: Focus on premium quality and luxury.
  - Premium Product Lines: Emphasize high-end product lines, such as gourmet foods, premium wines, and exclusive brands.
  - Personalized Services: Offer personalized shopping experiences, such as personal shoppers, custom orders, and home delivery services tailored to their preferences.
  - Exclusive Events: Invite them to exclusive events like wine tastings, cooking classes, and gourmet food festivals.
  - Health and Wellness: Highlight the health benefits of premium products and provide tailored health and wellness programs, including private fitness sessions and wellness retreats.

7. In Summary:

KMeans++ and Agglomerative clustering segmented a dataset of 4,000 customers using variables like age, gender, income, education, marital status, and settlement size. This resulted in two main segments, enabling targeted and effective marketing strategies to attract the right customers with the right demands at the right time.