SUPERMARKET CUSTOMER BASE SEGMENTATION AND INSIGHTS

Audrey Nathania Hermawan

Introduction

The contemporary competitive retail environment often neglects customer profiling, leading to ineffective customer targeting, limited marketing customisation, resource allocation issues, and potential income loss. Hence, this report assesses a dataset of 2000 consumers for a supermarket chain utilising K-Means and Hierarchical clustering algorithms to discover consumer segments based on specified attributes and generate tailored marketing insights. The ultimate goal is to enhance customer satisfaction, develop loyalty, boost sales, and sustain a competitive advantage.

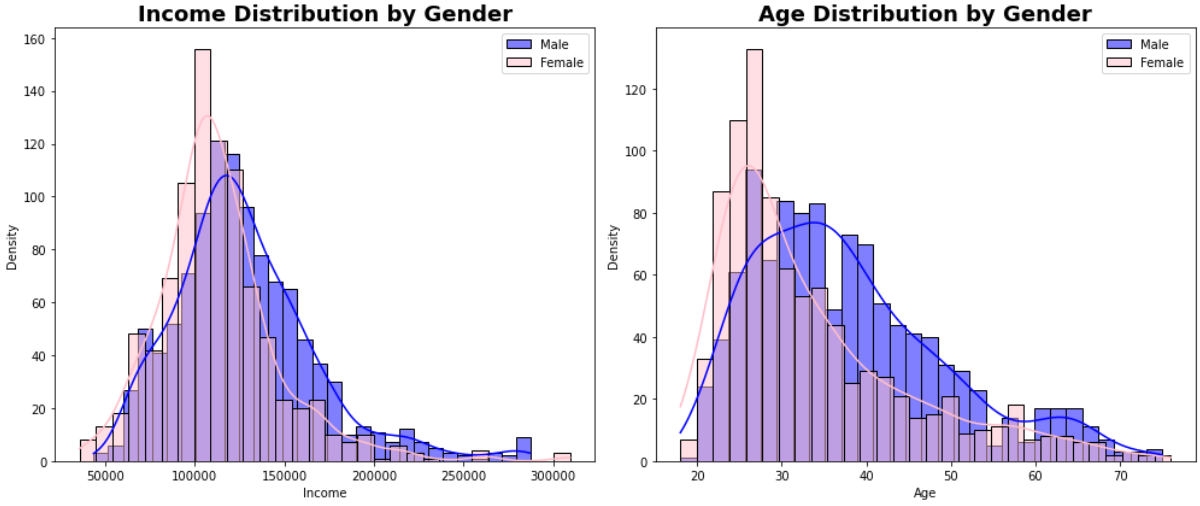
Exploratory Data Analysis

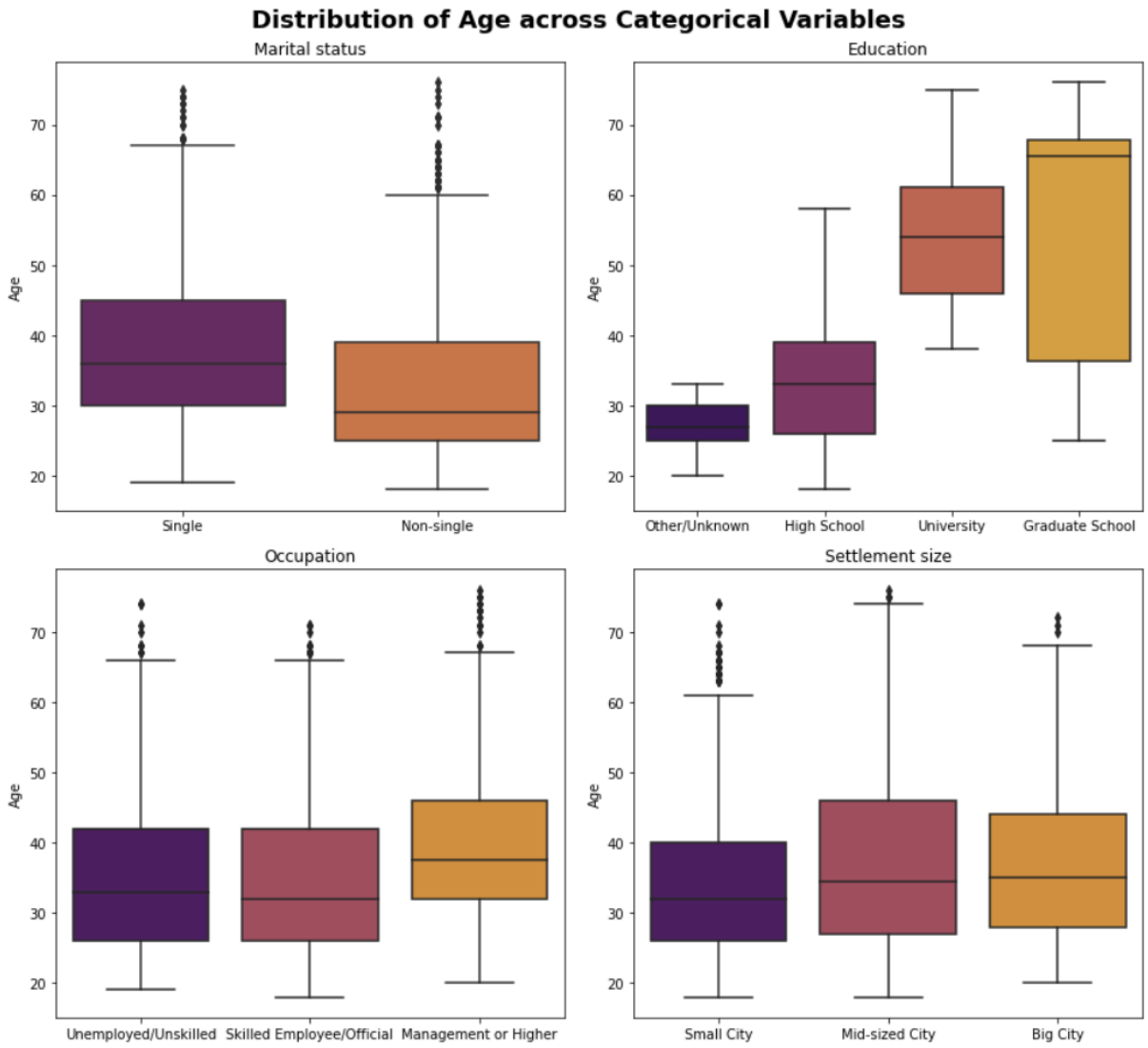
The dataset encompasses 3 and 5 numerical and categorical variables, respectively.

```
Data shape: Rows = 2000 ; Columns = 8
```

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
0	100000001	0	0	67	2	124670	1	2
1	100000002	1	1	22	1	150773	1	2
2	100000003	0	0	49	1	89210	0	0
3	100000004	0	0	45	1	171565	1	1
4	100000005	0	0	53	1	149031	1	1

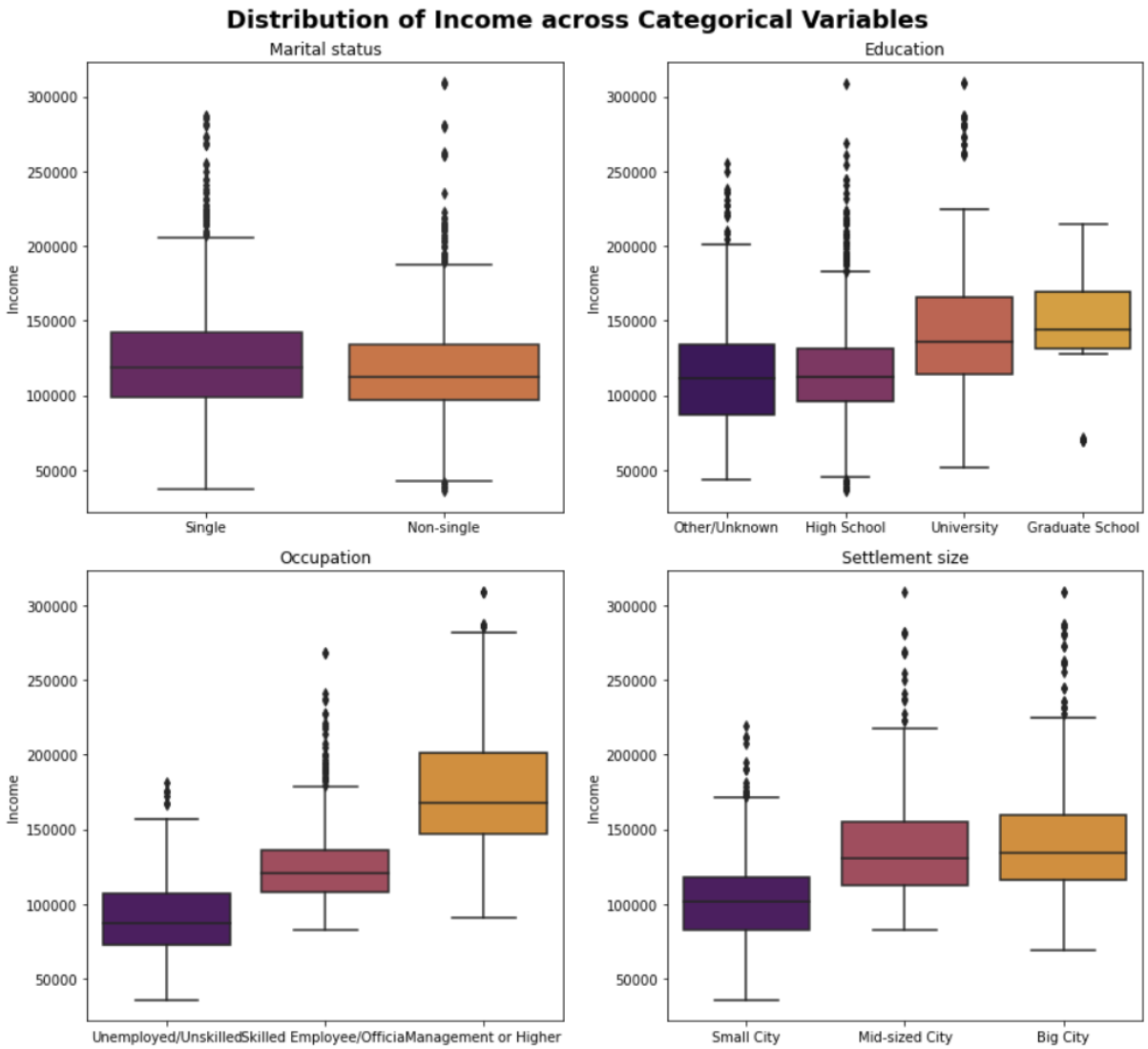
The analysis indicates that the majority of the supermarket's customers are middle-aged, with a right-skewed age distribution. Females are most prevalent in their late twenties, while males become more dominant after the age of 30. Older customers tend to have higher education levels and hold more senior occupational roles, reflecting their experience and career progression. Additionally, they are more likely to reside in mid-sized and major urban areas.



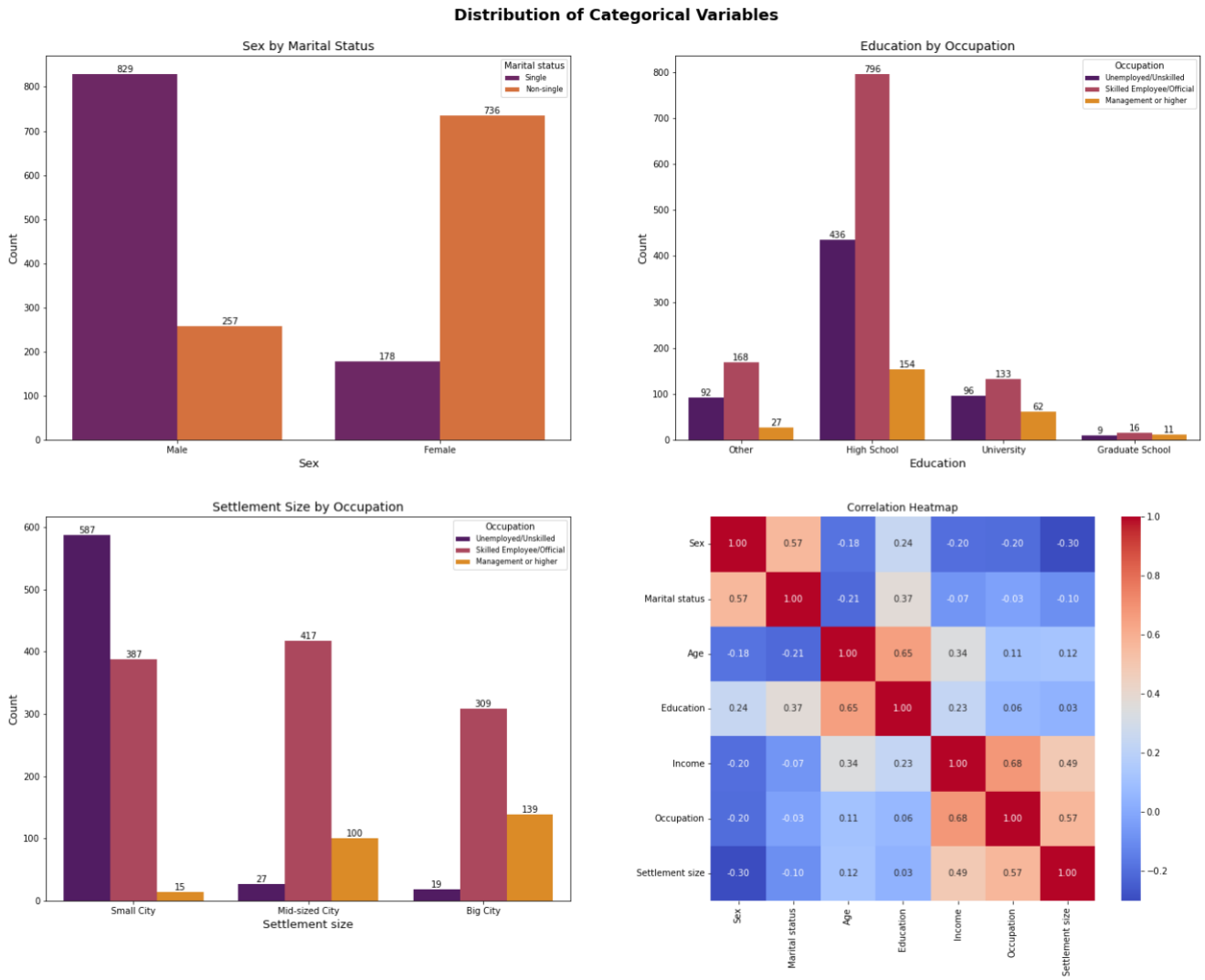


The average annual income stands at $120,954, with males predominantly occupying the higher-income brackets, which may suggest gender disparities and occupational segregation. Additionally, individuals with higher education and professional roles tend to earn more. Larger cities report the highest average incomes, driven by greater economic opportunities and access to higher-paying jobs.

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
mean	100001000.5	0.457	0.4965	35.909	1.038	120954.419	0.8105	0.739
min	100000001.0	0.000	0.0000	18.000	0.000	35832.000	0.0000	0.000
max	100002000.0	1.000	1.0000	76.000	3.000	309364.000	2.0000	2.000



Regarding categorical factors, the retailer primarily caters to non-single females and single males, with a relatively balanced distribution of gender and marital status among customers. Additionally, high school education is the dominant category, whereas skilled workers represent the most common occupation type in the supermarket. Nevertheless, small cities with limited job opportunities experience widespread unemployment, aligning with the positive correlation between settlement size and occupation. These findings serve as valuable insights for the subsequent segmentation analysis.



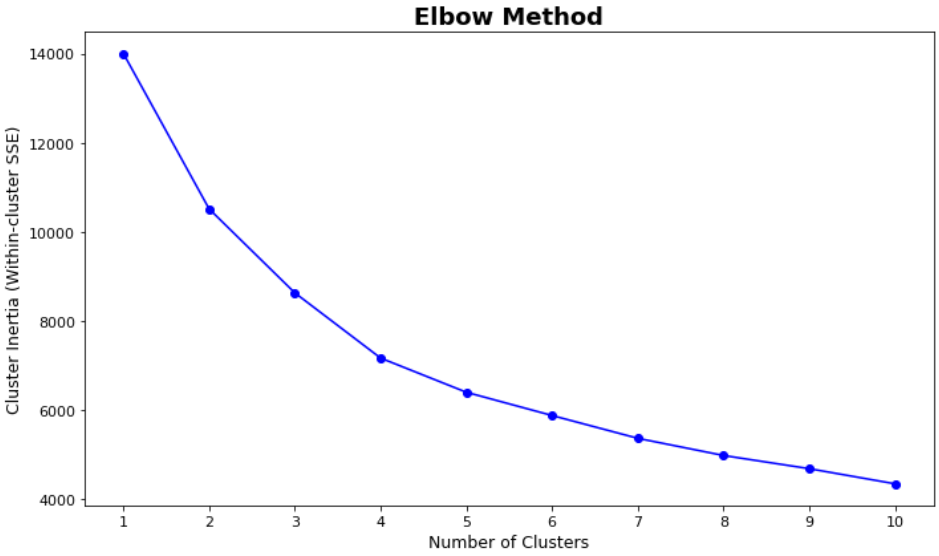
Customer Segmentation

Before performing customer segmentation, data is standardised and irrelevant feature such as ID, which lacks meaningful information is eliminated. This ensures feature consistency and comparability within the same numerical range.

K-Means Clustering

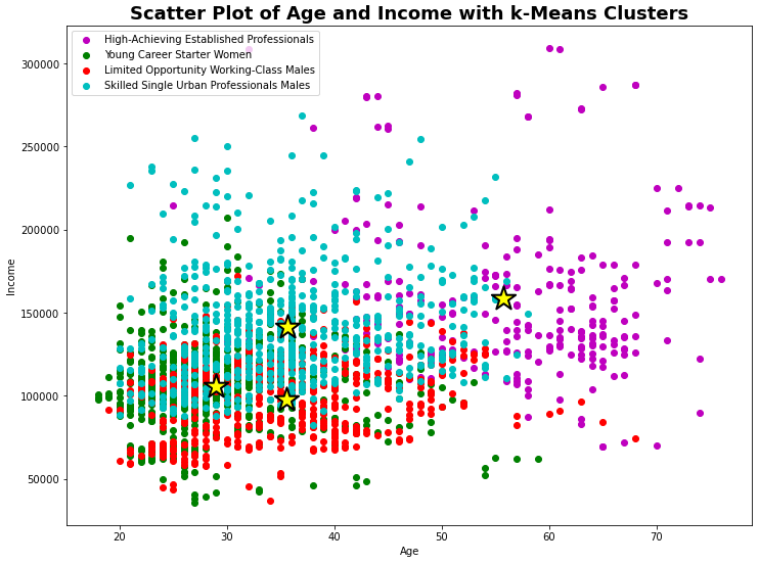
The elbow plot is utilised to determine the ideal number of clusters for the K-Means algorithm. The graph suggests that 4 clusters are optimal based on the transition from a

steep to a gradual decline. The analysis proceeds by training the k-Means clustering model on normalised data with 4 clusters, assigning each data point to the nearest centroid.



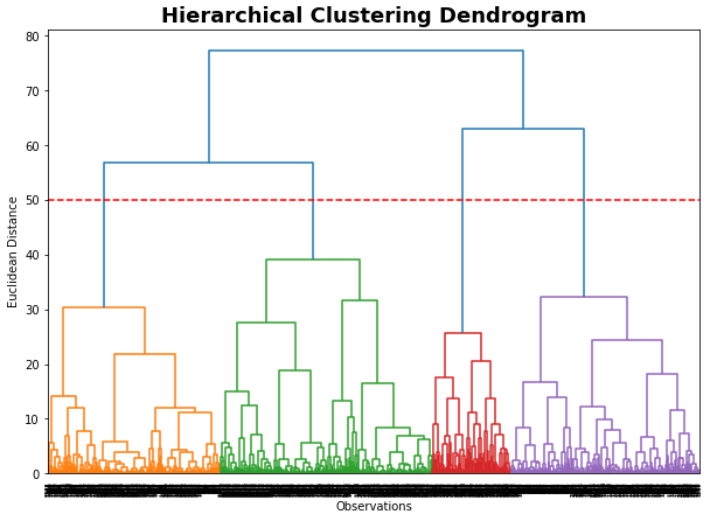
The first cluster (Young Career Starter Women) consists primarily of non-single young females residing in small cities with low incomes. In contrast, the second cluster (Limited Opportunity Working-Class Males) comprises middle-aged unemployed males living in small towns. The third cluster (Skilled Single Urban Professionals Males) is dominated by middle-thirties single males with moderate incomes dwelling in big cities. Lastly, the fourth cluster (High-Achieving Established Professionals) includes the oldest group with high income and education, residing in mid-sized cities.

KMeans_Cluster	Income	Age	Sex	Marital status	Education	Occupation	Settlement size
Young Career Starter Women	105759.12	27.0	1	1	1	1	0
Limited Opportunity Working-Class Males	97859.85	35.0	0	0	1	0	0
Skilled Single Urban Professionals Males	141218.25	35.0	0	0	1	1	2
High-Achieving Established Professionals	158338.42	57.0	1	1	2	1	1



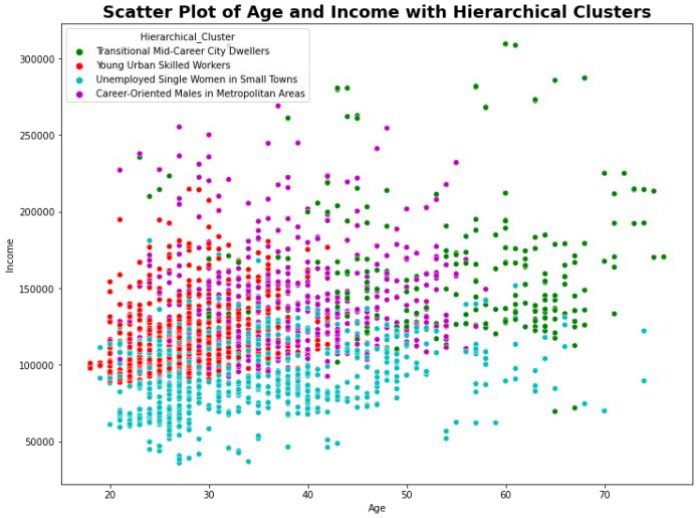
Hierarchical Clustering

The dataset can be divided into four clusters based on the hierarchical clustering dendrogram, determined by creating a cut below the longest non-intersected vertical line.



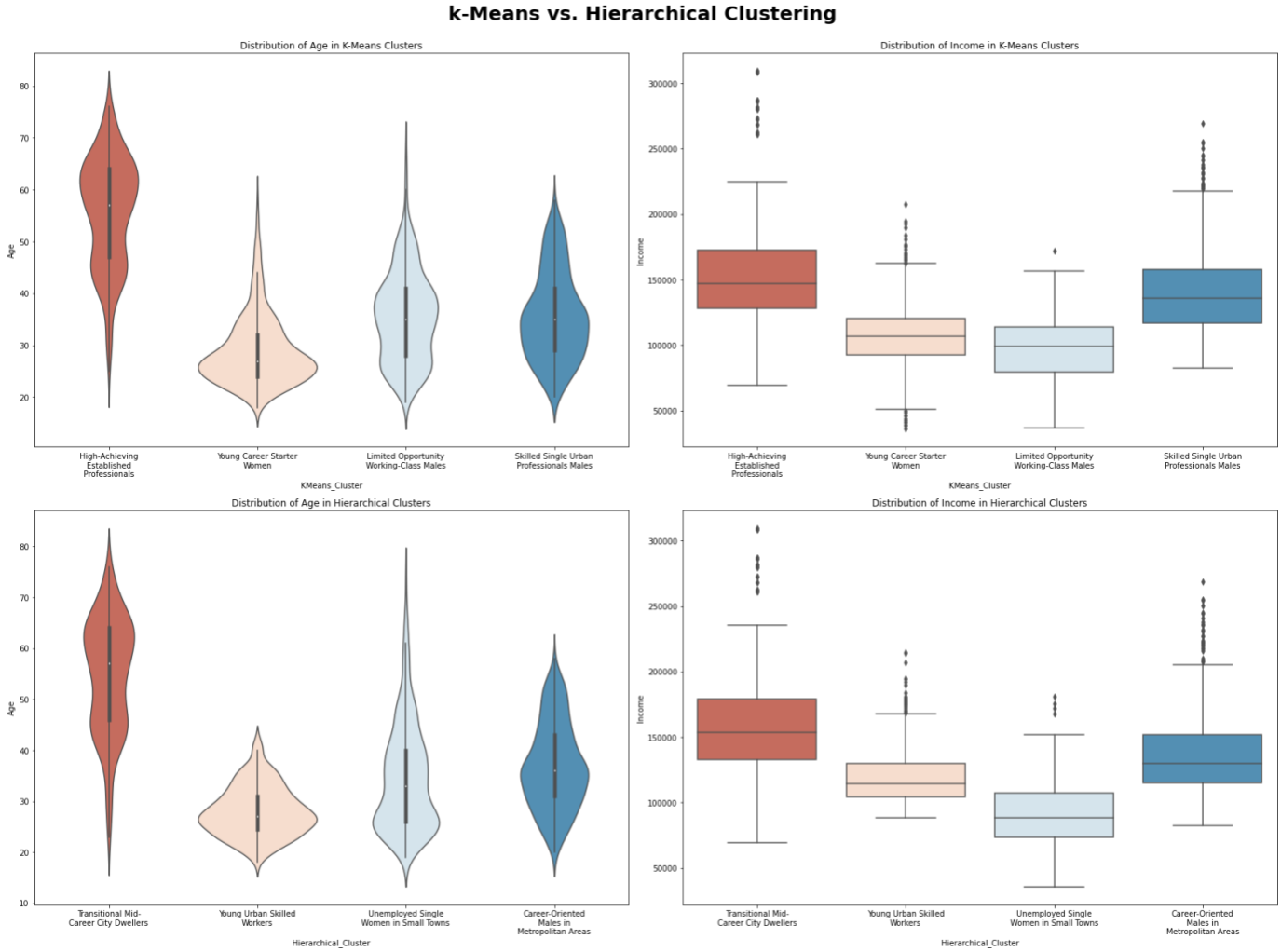
The first cluster (Unemployed Single Women in Small Towns) mainly consists of middle-aged single females in small cities with low incomes and unemployed. The second cluster (Career-Oriented Males in Metropolitan Areas) includes single males in their prime working years with medium income and career focus in big cities. The third cluster (Transitional Mid-Career City Dwellers) is dominated by non-single customers approaching retirement in mid-sized cities. The fourth cluster (Young Urban Skilled Workers) primarily comprises young adult non-single females with medium income and skilled or official employment in small cities.

Hierarchical_Cluster	Income	Age	Sex	Marital status	Education	Occupation	Settlement size
Unemployed Single Women in Small Towns	90807.33	33.0	1	0	1	0	0
Career-Oriented Males in Metropolitan Areas	137369.34	36.0	0	0	1	1	2
Transitional Mid-Career City Dwellers	163924.68	57.0	0	1	2	1	1
Young Urban Skilled Workers	120399.59	27.0	1	1	1	1	0



Comparison

Despite yielding similar results in some aspects, the K-means and hierarchical clustering algorithms demonstrate slight variations in each cluster's characteristics. Furthermore, the scatterplot of Age and Income in both algorithms fails to differentiate clusters adequately due to overlapping data points, possibly due to neglecting other distinguishing features. Hence, depending solely on numerical features may not guarantee distinct cluster separation, making it advisable to incorporate Principal Component Analysis for improved outcomes.



Recommendation

Marketing recommendations can be generated by utilising the shared dominant characteristics of each cluster discovered using Hierarchical and K-Means approaches, such as:

Cluster 1 (Low socioeconomic young females in small-sized cities)

- Promote budget-friendly products, including affordable meal bundles
- Introduce loyalty rewards programs
- Provide in-store cooking classes and nutrition workshops to highlight cost-effective yet nutritious choices.
- Boost brand visibility by collaborating with social media influencers and showcasing affordable grocery hauls.

Cluster 2 (Middle-aged males with medium income)

- Emphasise convenient meal options for busy individuals, including ready-to-eat and easy-to-prepare meals.
- Target male-oriented product categories via digital marketing and online shopping services.
- Highlight professional-oriented products with superior quality.
- Partner with fitness and sports centres to offer signup rewards such as sporting goods or apparel.

Cluster 3 (Prime-aged high-income males in mid-sized regions)

- Focus on supplying healthy, nutritious, locally sourced options, such as organic produce, whole-grain items, and protein supplements.
- Offer a selection of premium smart appliances, kitchen gadgets, and personal care items that appeal to their technological interests.
- Provide additional discounts and exclusive shopping hours for seniors, ensuring a calm and supportive atmosphere, assistance with heavy items, and personalised dietary guidance.

Cluster 4 (Married skilled high-income adults in mid-sized cities)

- Offer instant home delivery, drive-thru groceries, or personal shopper service.
- Highlight exclusive products such as organic food, specialty meats, sophisticated household items, and luxury personal care products.
- Serve complimentary fruit for children and coffee for adults, allowing customers to explore the aisles and make additional purchases conveniently.
- Implement bulk buying promotions to encourage larger purchases.

Conclusion

In conclusion, this report leverages the provided dataset to conduct a customer segmentation analysis for a supermarket chain. The analysis includes exploratory data analysis, feature selection, data standardisation, and the application of both Hierarchical and K-means clustering techniques. Based on these insights, marketing strategy recommendations are tailored to the distinct characteristics of each customer segment. Nevertheless, ongoing evaluation of customer feedback

and spending behaviours is essential to sustain marketing effectiveness in the long run. Through proactive involvement with each consumer category and offering personalised experiences, the supermarket chain may attain sustainable business growth and maintain a competitive edge.