Customer Personality Analysis

[Dataset Link]

Data Analysis and Preprocessing

The initial dataset contained **2240 samples (rows)** and **29 features (columns)**.

- Transformed birth date to age (ref: 01.01.2015) and converted Dt_Customer to days since enrollment.
- Dropped unnecessary columns.
- Removed invalid marital status entries (YOLO, Absurd, Alone).
- Label-encoded Marital Status and Education Level.
- Removed outliers in Age and Income using IQR method.

After cleaning and preprocessing, the final dataset consists of **2198 rows** and **20 columns**.

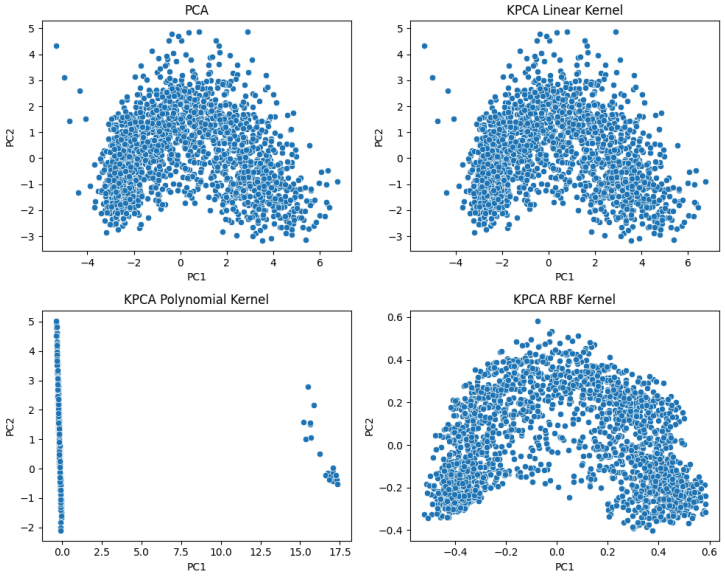
Correlation Heatmap Insights

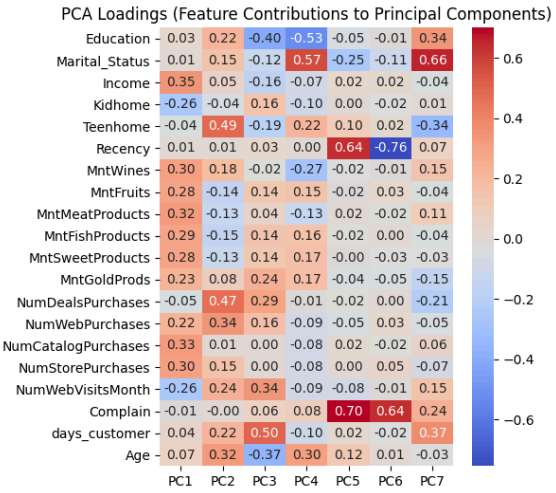
Income:

- Positively correlated with spending on all products, especially wine and meat.
- Negatively correlated with the number of children at home.
- High-income customers avoid discounts, indicating less price sensitivity.

Education: Higher education levels correlate with increased wine purchases.
Family-Oriented Customers: More children at home leads to more web visits.

Dimensionality Reduction: PCA

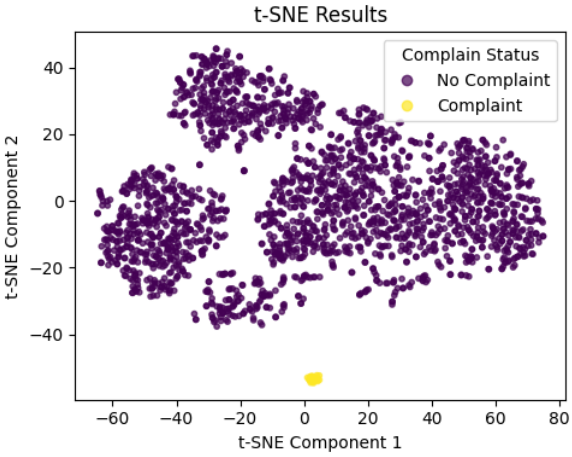


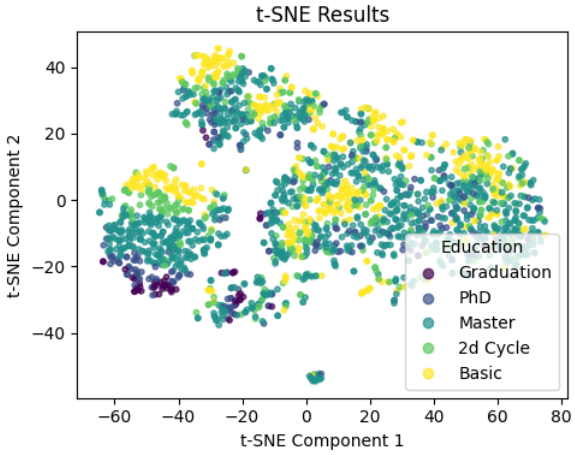


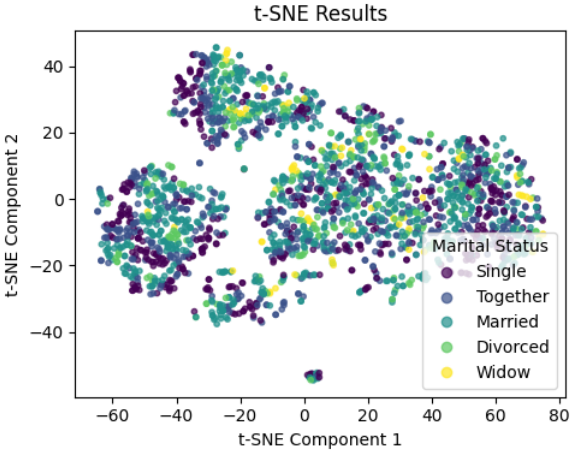
The results projected on the first 2 PCs did not reveal any distinct patterns or clusters, suggesting that PCA might not be effective for this dataset in identifying meaningful groupings. Given the limitations of PCA, t-SNE will be utilized as the next step.

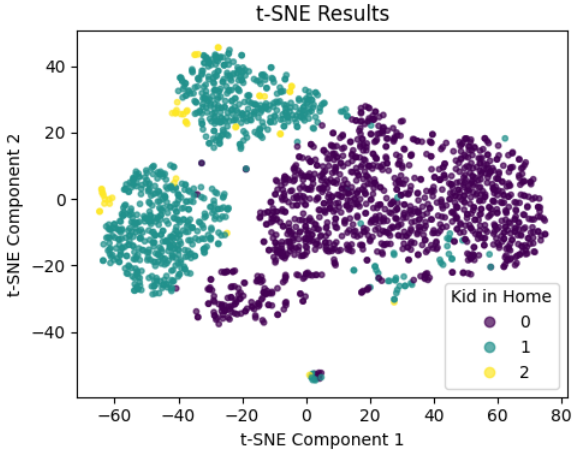
Dimensionality Reduction: t-SNE

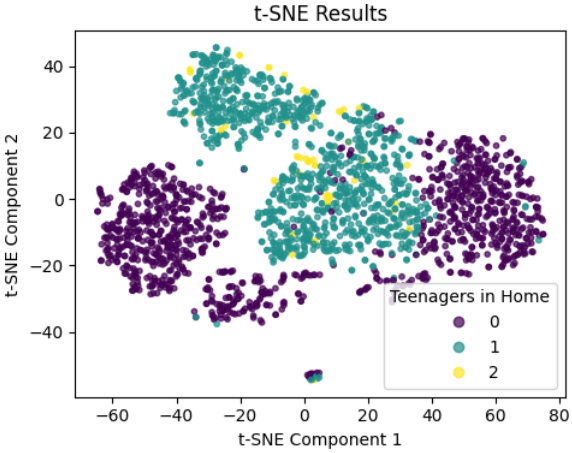
The t-SNE results provide meaningful groupings that align with specific customer traits.





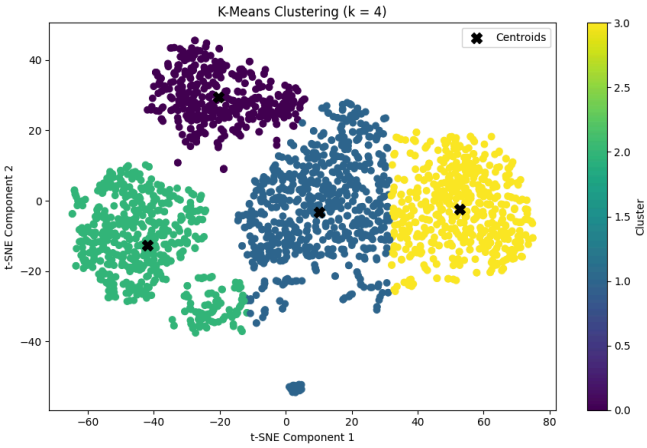


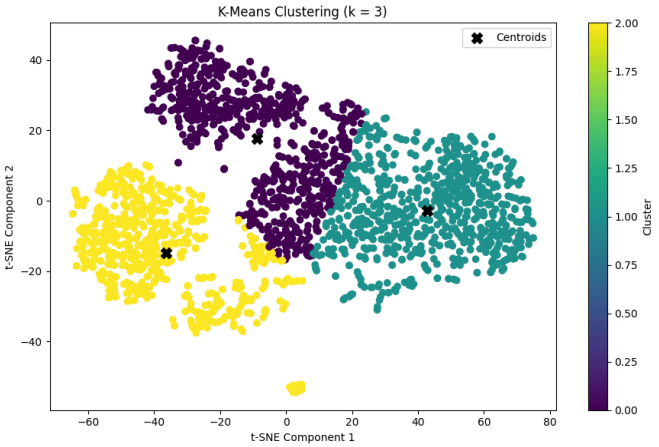


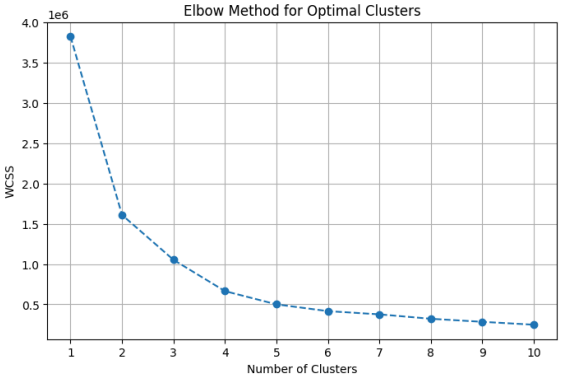


- Customers who frequently complain are grouped into small, distinct cluster
- Within each cluster, subgroups are formed based on education levels
- Marital Status may not be a key differentiator for customer segmentation
- Customers with children at home ("Kids at Home" and "Teens at Home") are grouped together within specific clusters.

Clustering: K-means

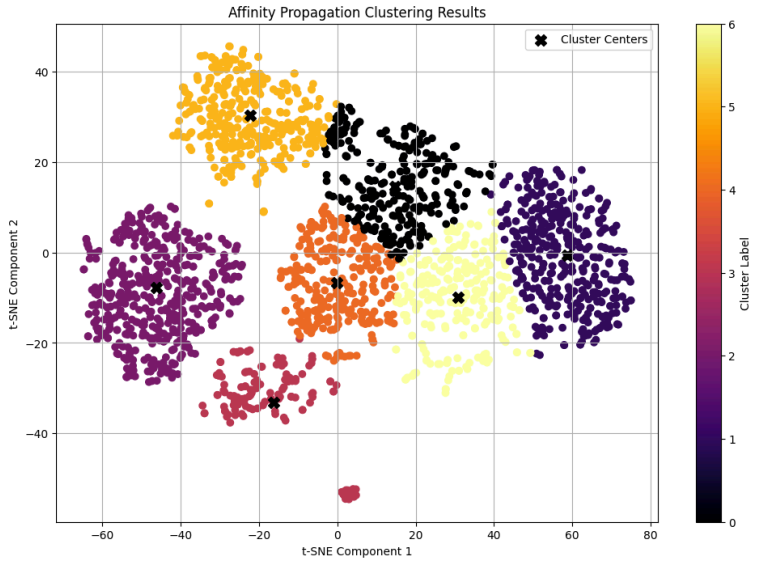






Silhouette Score k=4: 0.48
Silhouette Score k=3: 0.44

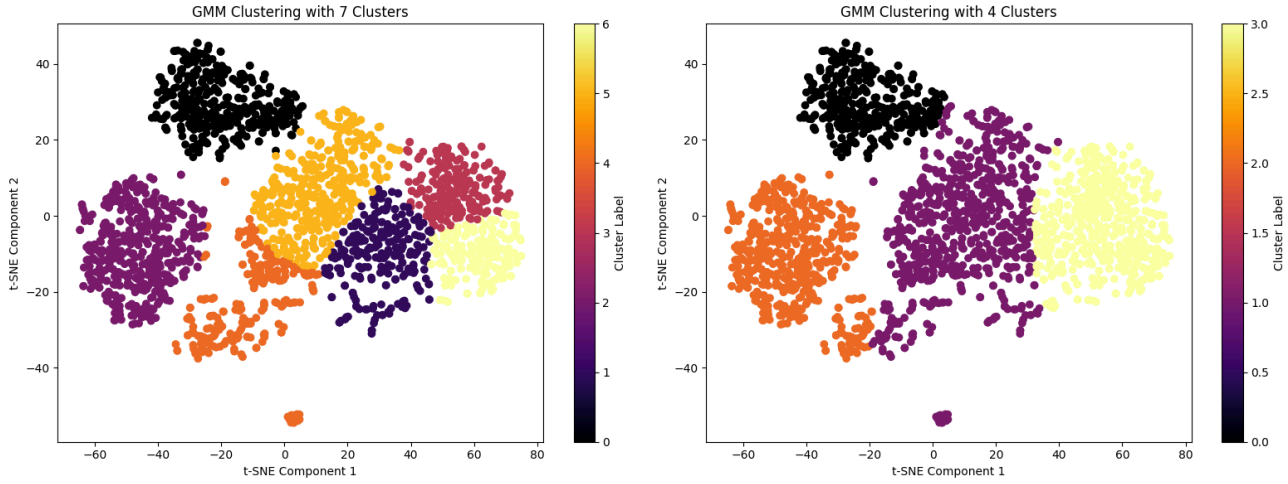
Clustering: Affinity Propagation



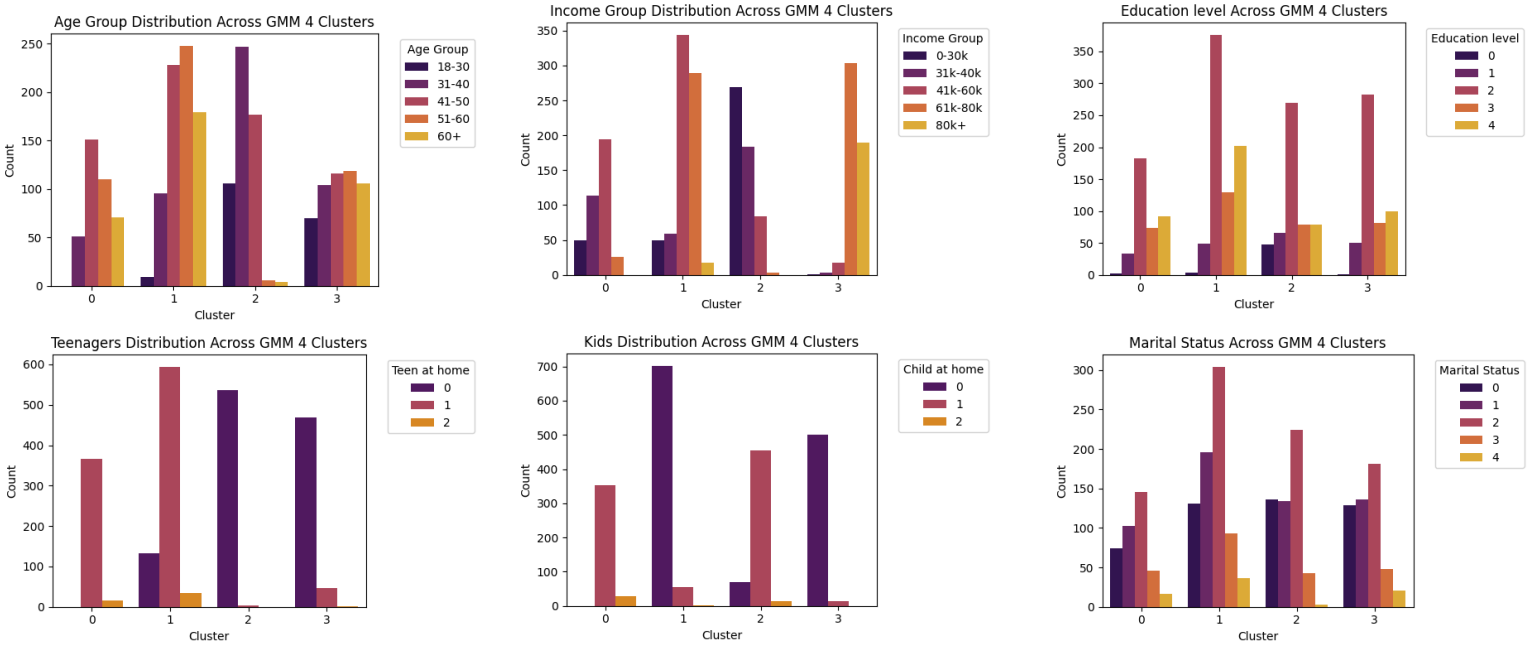
The number of clusters is defined by the algorithm.

Number of clusters: 7
Silhouette Score: 0.456

Clustering: GMM



Number of clusters: 7 to compare with Affinity Propagation
Number of clusters: 4 to compare with K-means



Note: Please refer to the Jupyter Notebook for detailed plots and visualizations summarized here. Thank you.

Summary

Cluster	Age	Income	Education	Children	Complain	Purchases	Spending
0	Middle age 40-60	Low 0-60k	Bachelor's degree	At least 1 child/teen at home	-	Don't do catalog purchases.	low
1	Middle to Old 50-60+	Average 40k-80k	The majority are BAs but a high percentage of PhDs too	No child, 1-2 teen	The only cluster with complains	Prefer web purchases and visit the web a lot.	The second cluster that spends the most after cluster 3, especially on wine
2	Young adult 18-40	Very Low 0-40k	Average, mainly Bachelor's degree	No teen, 1-2 child	-	Visit the web a lot, but don't do shopping	low
3	All age groups	Very high 60k-80k+	Average, mainly Bachelor's degree	Almost no child/teen at home	-	Don't visit web. Do catalog purchases. Avoid discounts	Customers who spend the most.