

MIS 451 Assignment

Application of Clustering Models to Business Data

Name: Tran Phuong Thao

IRN: 2132300447

1. Business problem

A luxury clothing company aims to enhance its marketing effectiveness by launching personalized, targeted campaigns. The sales team has collected valuable customer data including age, income, annual spend, and recency of purchase (measured as days since the last transaction). However, the company currently lacks insight into how many distinct customer segments exist within its base.

The key business challenge is to identify and understand distinct customer segments based on their attributes to tailor marketing strategies for each group using data-driven techniques.

To solve this, we will apply clustering techniques (such as K-Means) to segment customers based on the given features. These insights will enable the business to design and launch more strategic, targeted marketing initiatives.

2. Some explanations for steps taken

Before applying the clustering model, we performed essential data preprocessing steps to ensure the quality and consistency of our dataset:

- **Missing Values:** Checked for null values across all features. No missing data was found, indicating a complete dataset.

```
df.isnull().sum()

income    0
age       0
days_since_purchase  0
annual_spend    0

dtype: int64
```

- **Duplicates:** None were found.

```
[34] df.duplicated().sum()

np.int64(0)
```

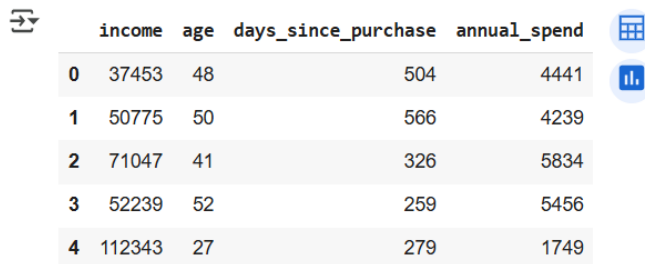
- **Data Types:** Ensured all columns used in clustering (e.g., Age, Income, Annual Spend) were numeric and suitable for analysis.

```
[32] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  -
0   income                1000 non-null   int64
1   age                   1000 non-null   int64
2   days_since_purchase    1000 non-null   int64
3   annual_spend           1000 non-null   int64
dtypes: int64(4)
memory usage: 31.4 KB
```

- **Feature Selection:** Selected only numeric columns for clustering (the dataset only includes numeric variables, but I still this step like normal).

```
[39] # Select only numerical columns for clustering
features = df.select_dtypes(include=[np.number])
features.head()
```



	income	age	days_since_purchase	annual_spend
0	37453	48	504	4441
1	50775	50	566	4239
2	71047	41	326	5834
3	52239	52	259	5456
4	112343	27	279	1749

- **Standardization:** Applied StandardScaler from scikit-learn to standardize the dataset. This step is crucial since KMeans is distance-based and sensitive to the scale of features. Without scaling, variables with larger ranges (e.g., income) could dominate the clustering process.

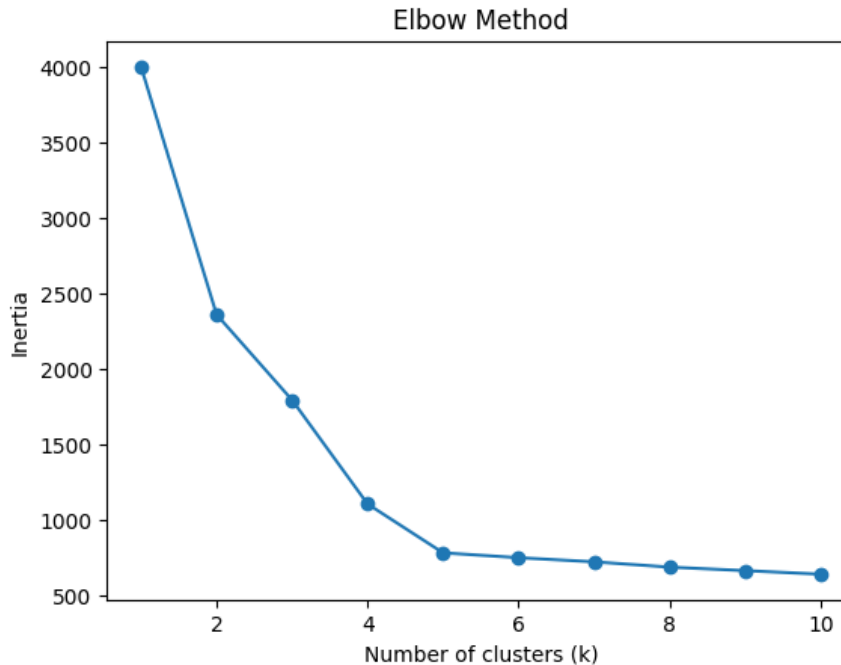
```
#Standardization is important because K-Means is distance-based.
scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)

scaled_df = pd.DataFrame(scaled_features, columns=features.columns)
scaled_df.head()
```



	income	age	days_since_purchase	annual_spend
0	-0.890426	0.604861	0.753257	0.062276
1	-0.553094	0.789354	1.306361	-0.058939
2	-0.039778	-0.040865	-0.834688	0.898182
3	-0.516023	0.973847	-1.432398	0.671353
4	1.005897	-1.332317	-1.253977	-1.553129

- **Determining Optimal Number of Clusters:** Use the Elbow Method, which involves plotting the within-cluster sum of squares (inertia) against various values of k. This inflection point suggests that 4 clusters is an appropriate choice, balancing model simplicity and accuracy. Choosing too few clusters might oversimplify the customer base, while too many could overcomplicate interpretation and strategy development.



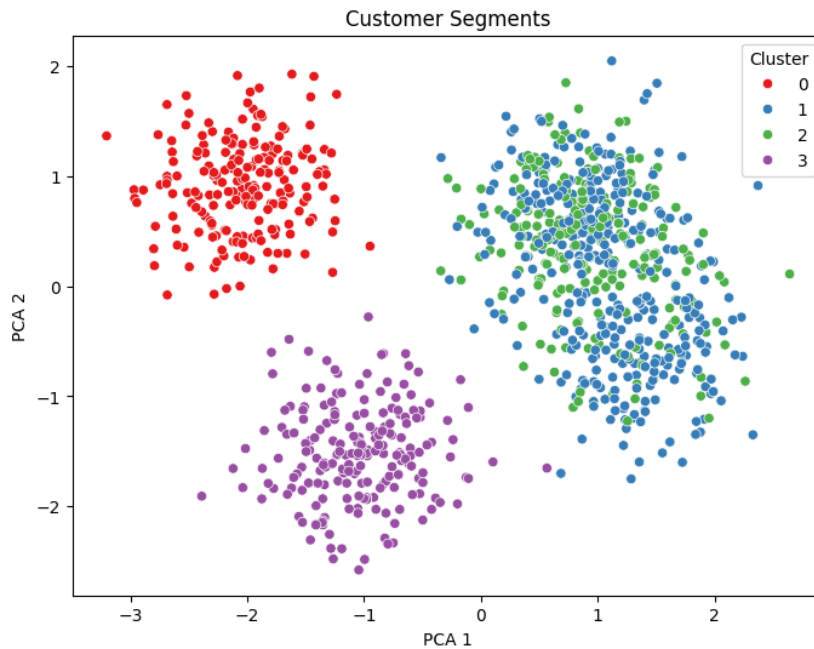
- Clustering with KMeans (k = 4):** Apply the KMeans algorithm with 4 clusters, assigning each customer a cluster label based on their standardized features (e.g., income, age, days since last purchase, annual spend). These labels segment the customer base into distinct groups with similar behavioral patterns.

```
kmeans = KMeans(n_clusters=4, random_state=42)
df['Cluster'] = kmeans.fit_predict(scaled_features)
df.head()
```

	income	age	days_since_purchase	annual_spend	Cluster	PCA1	PCA2
0	37453	48	504	4441	1	0.855025	-0.935161
1	50775	50	566	4239	1	0.963841	-1.047314
2	71047	41	326	5834	2	0.350945	0.742276
3	52239	52	259	5456	2	0.833541	0.960941
4	112343	27	279	1749	0	-2.476828	0.755844

- PCA Visualization of Clusters:** To visually validate the clustering, we applied Principal Component Analysis (PCA) to reduce the feature space to two dimensions. The resulting scatter plot shows a clear separation among the 4 clusters. Notably, Cluster 1 and Cluster 2 show some overlap, possibly indicating shared traits that may not be easily separable in

reduced dimensions - hinting at hidden patterns or gradual customer transitions between segments.



- **Cluster Distribution and Characteristics:** We analyzed the cluster sizes and feature averages to extract meaningful customer personas

```
[67] df.groupby('Cluster').mean()
```

	income	age	days_since_purchase	annual_spend	PCA1	PCA2
Cluster						
0	116310.542714	29.291457	294.125628	2501.728643	-2.028239	0.911603
1	86113.537604	49.348189	507.155989	5547.167131	1.086344	0.102994
2	41571.379167	49.233333	324.329167	5588.512500	0.934205	0.370375
3	42476.311881	30.108911	500.618812	2508.415842	-1.042509	-1.521158

3. Recommendations for Targeted Marketing

Cluster 0: Wealthy but Disengaged

- Problem: High income, low spending, moderate recency
 - Strategy: Re-engagement through exclusive loyalty perks or luxury preview events
 - Tactic: Personalized email campaigns emphasizing premium offerings and exclusivity
-

Cluster 1: Loyal High Spenders

- Problem: None; this is a highly valuable segment
 - Strategy: Focus on retention and upselling
 - Tactic: VIP programs, early access to new collections, or member-only discounts
-

Cluster 2: Budget-Conscious Big Spenders

- Problem: Low income but high spending, yet inactive recently
 - Strategy: Offer value-based bundles and reactivation incentives
 - Tactic: Targeted promotions with flexible payment options or discount thresholds
-

Cluster 3: Young, Low-Spending, Least Engaged

- Problem: Low spend, low engagement
- Strategy: Brand-building and social media campaigns to increase awareness and interest
- Tactic: Collaborations with influencers, referral programs, or student discounts