Customer Segmentation Analysis with Python

A comprehensive data-driven approach to understanding customer behavior patterns and creating actionable marketing segments using pandas, scikit-learn, and machine learning clustering techniques.

Project Overview

Dataset Foundation

- Each row represents one customer with detailed profile information.
- Includes demographics: ID, birth year, education, and marital status.
- Contains household details like number of kids and teenagers.
- Tracks spending on products such as wine, fruits, meat, fish, sweets, and gold.
- Records purchase activity (web, catalog, store), website visits, and campaign responses.

ID	Year_Birth	Education	Marital_Sta	Income	Kidhome	Teenhome
5524	1957	Graduatior	Single	58138	0	0
2174	1954	Graduatior	Single	46344	1	1
4141	1965	Graduatior	Together	71613	0	0
6182	1984	Graduatior	Together	26646	1	0
5324	1981	PhD	Married	58293	1	0
7446	1967	Master	Together	62513	0	1
965	1971	Graduatior	Divorced	55635	0	1
6177	1985	PhD	Married	33454	1	0
4855	1974	PhD	Together	30351	1	0
5899	1950	PhD	Together	5648	1	1

MntWines	MntFruits	MntMeatPr	MntFishPro	MntSweetF	MntGoldPr
635	88	546	172	88	88
11	1	6	2	1	6
426	49	127	111	21	42
11	4	20	10	3	5
173	43	118	46	27	15

Analysis Goals

- Clean and transform raw customer data
- Engineer meaningful features for segmentation
- Identify distinct customer groups using K-Means clustering
- Generate actionable insights for targeted marketing

Data Preparation Pipeline

Load Data

Import CSV with pandas and examine structure using head(), info(), and shape methods

Handle Missing Values

Check for nulls using isna().sum() and remove incomplete records with dropna()

Feature Engineering

Create new variables: Age from Year_Birth, Total_Children, Total_Spending, and Customer_Since

Validate

Verify data quality with describe() and value_counts() for categorical variables

Key Feature Engineering

Age Calculation

```
df["Age"] = 2025 - df["Year_Birth"]
```

Convert birth year to current age for demographic analysis

Total Spending

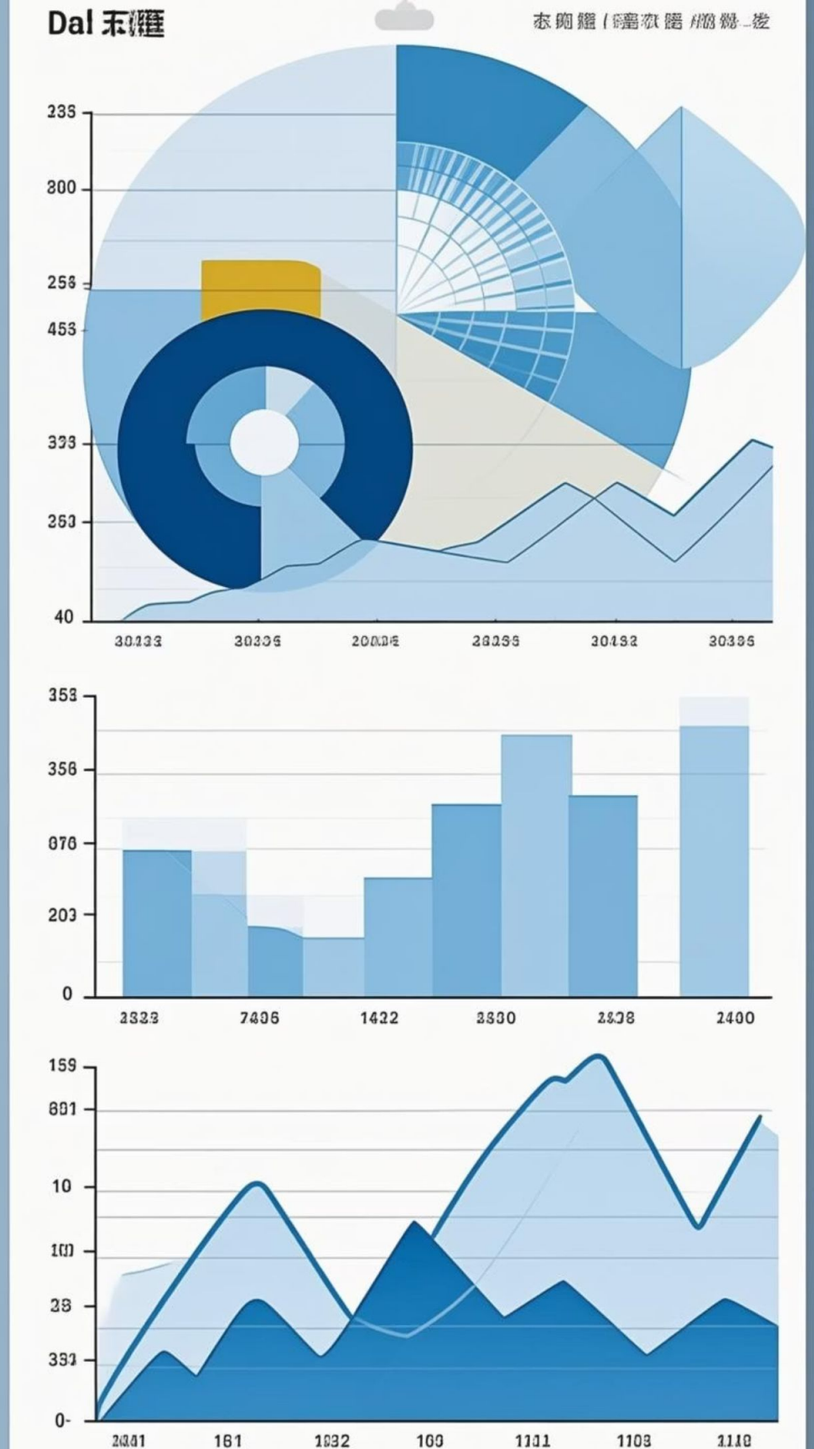
```
spend_cols = ['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']df["Total_Spending"] = df[spend_cols].sum(axis=1)
```

Aggregate spending across all product categories

Customer Tenure

```
df["Customer_Since"] = (pd.Timestamp("today") - df["Dt_Customer"]).dt.days
```

Calculate days since customer registration for loyalty analysis



Exploratory Data Analysis

Distribution Analysis

Created histograms with KDE curves for Age, Income, and Total_Spending to understand data spread and identify potential outliers or skewness patterns.

Categorical Comparisons

Used boxplots to compare Income by Education level and Total_Spending by Marital_Status, revealing key demographic spending differences.

Correlation Patterns

Generated heatmap showing relationships between Income, Age, Recency, Total_Spending, and purchase channels to identify multicollinearity.

Campaign Acceptance Analysis

Feature Creation

```
df["AcceptedAny"] = df[["AcceptedCmp1", "AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5", "Response"]].
sum(axis=1)
df["AcceptedAny"].unique()
```

0.0s Pythc

```
array([1, 0, 3, 2, 4, 5])
```

```
df["AcceptedAny"] = df["AcceptedAny"].apply(lambda x: 1 if x > 0 else 0)
df["AcceptedAny"].unique()
```

0.0s Pythc

```
array([1, 0])
```

Binary flag indicating whether customer accepted any marketing campaign,

enabling response rate analysis by demographic segments.

Grouped acceptance rates by Marital_Status revealed significant variation in campaign responsiveness across different relationship statuses,

informing targeted marketing strategies.

K-Means Clustering Implementation

1

Feature Selection

Selected 7 key features: Age, Income, Total_Spending, NumWebPurchases, NumStorePurchases, NumWebVisitsMonth, and Recency

2

Standardization

Applied StandardScaler to normalize features and ensure equal weighting in distance calculations

3

Optimal K Selection

Used Elbow Method testing k=2 to k=10, plotting WCSS to identify optimal cluster count of 6

4

Model Training

Fit KMeans with n_clusters=6 and assigned cluster labels to each customer record

5

Dimensionality Reduction

Applied PCA to reduce features to 2 components for visualization while preserving variance

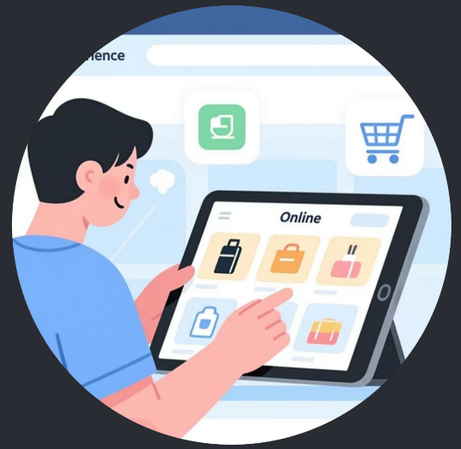
Customer Segment Profiles

Cluster 0 — Premium Customers

High-income customers who spend a lot and frequently purchase across channels.

Cluster 1 — Budget Customers

Low-income customers with very low spending and minimal engagement.



Cluster 2 — Online Shoppers

Customers who prefer online shopping and make many web purchases.

Cluster 3 — Inactive / Lost Customers

Customers who haven’t purchased recently (high recency) and show very low activity.

Cluster 4 — Average / Moderate Customers

Middle-income customers with moderate spending and balanced online + store purchases.

Cluster 5 — Potential High-Value Customers

High-income customers who currently spend less but can be targeted to increase sales.

Implementation & Next Steps

Model Deployment

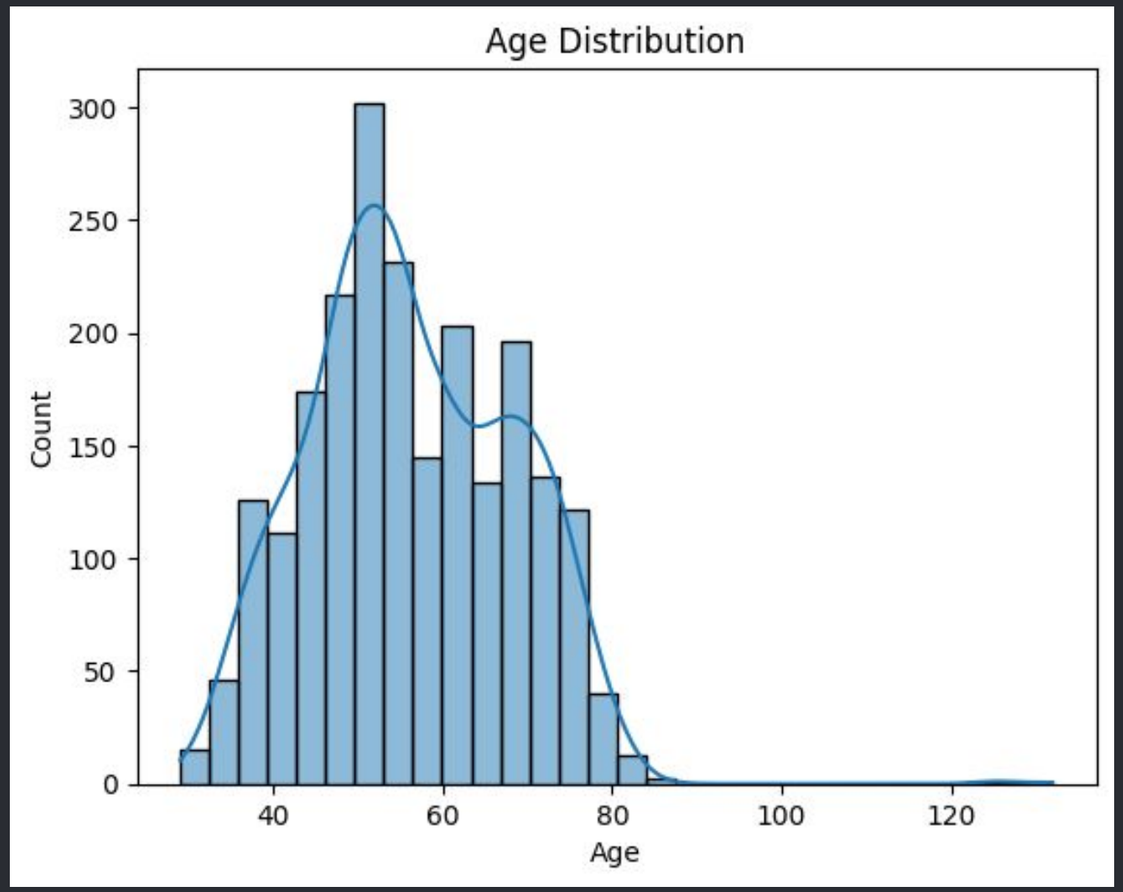
```
joblib.dump(kmeans,
"kmeans_customer_segmenta
tion_model.pkl")joblib.du
mp(scaler,
"scaler_customer_segmenta
tion.pkl")
```

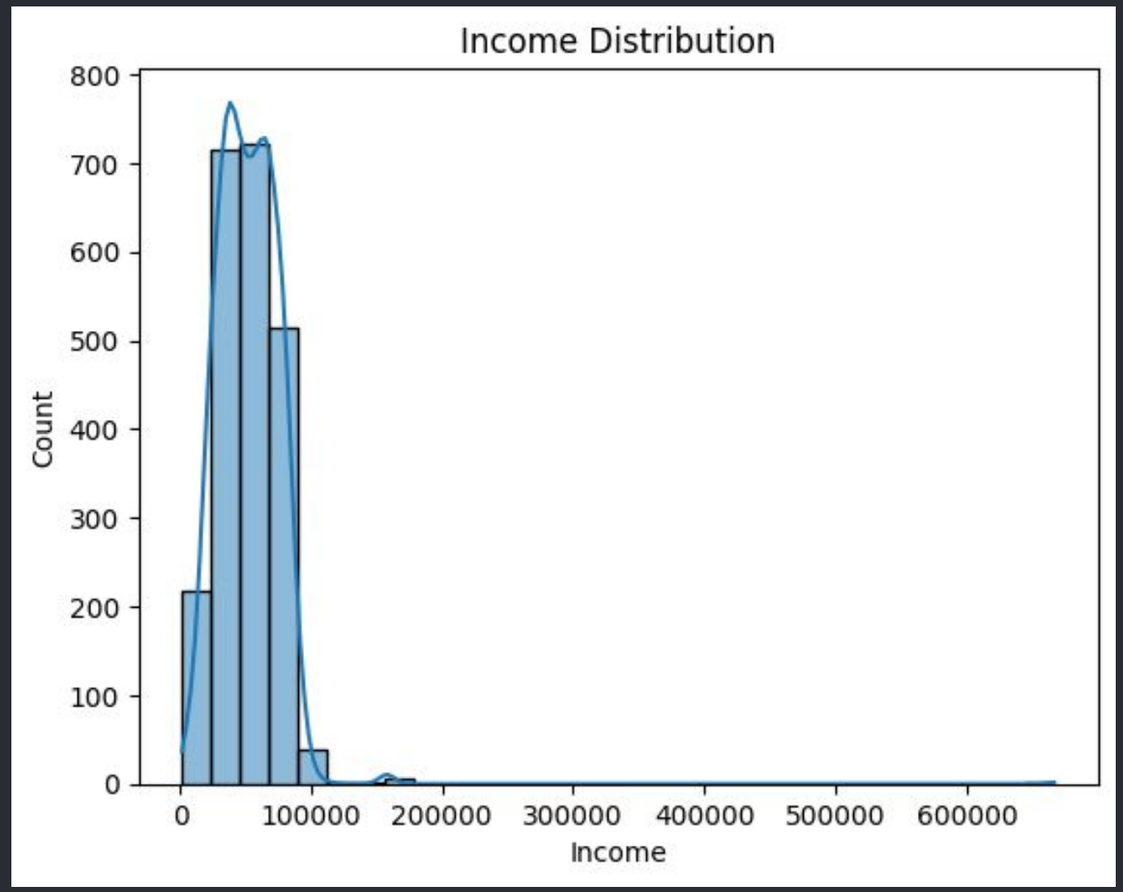
Saved trained model and scaler for production use, enabling real-time customer classification for new records.

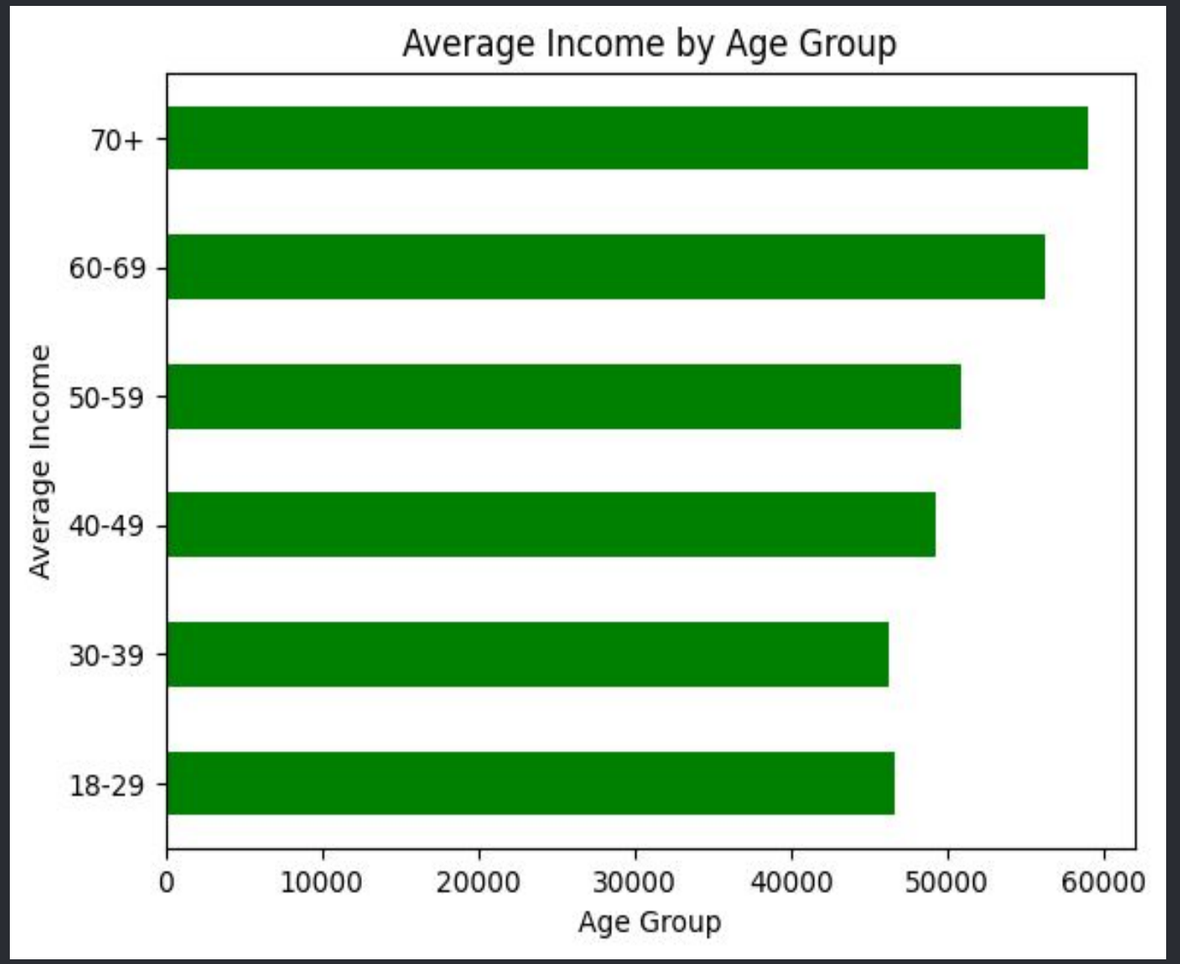
Business Applications

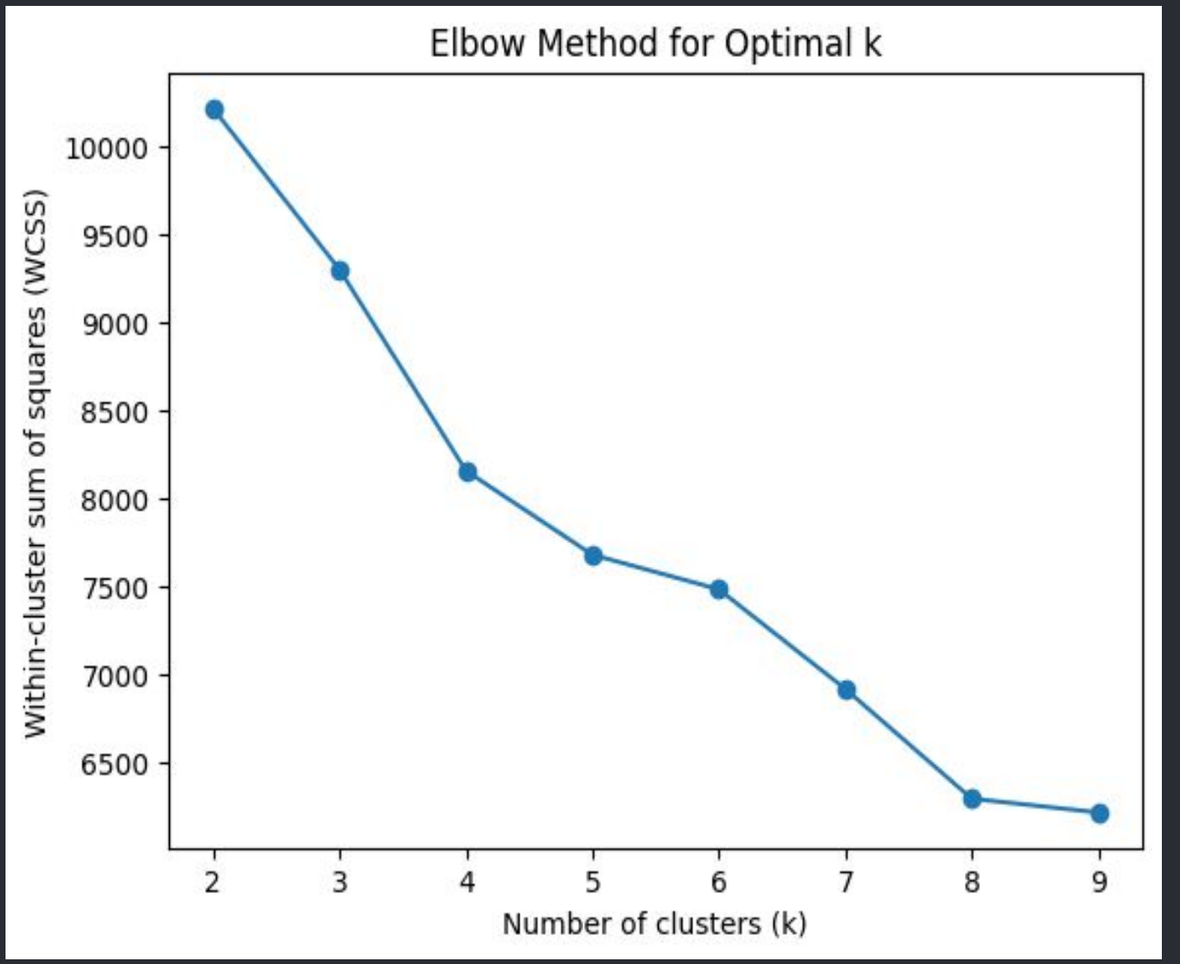
- Personalized marketing campaigns by segment
- Tailored product recommendations
- Channel-specific promotional strategies
- Customer lifetime value prediction
- Churn risk identification and prevention

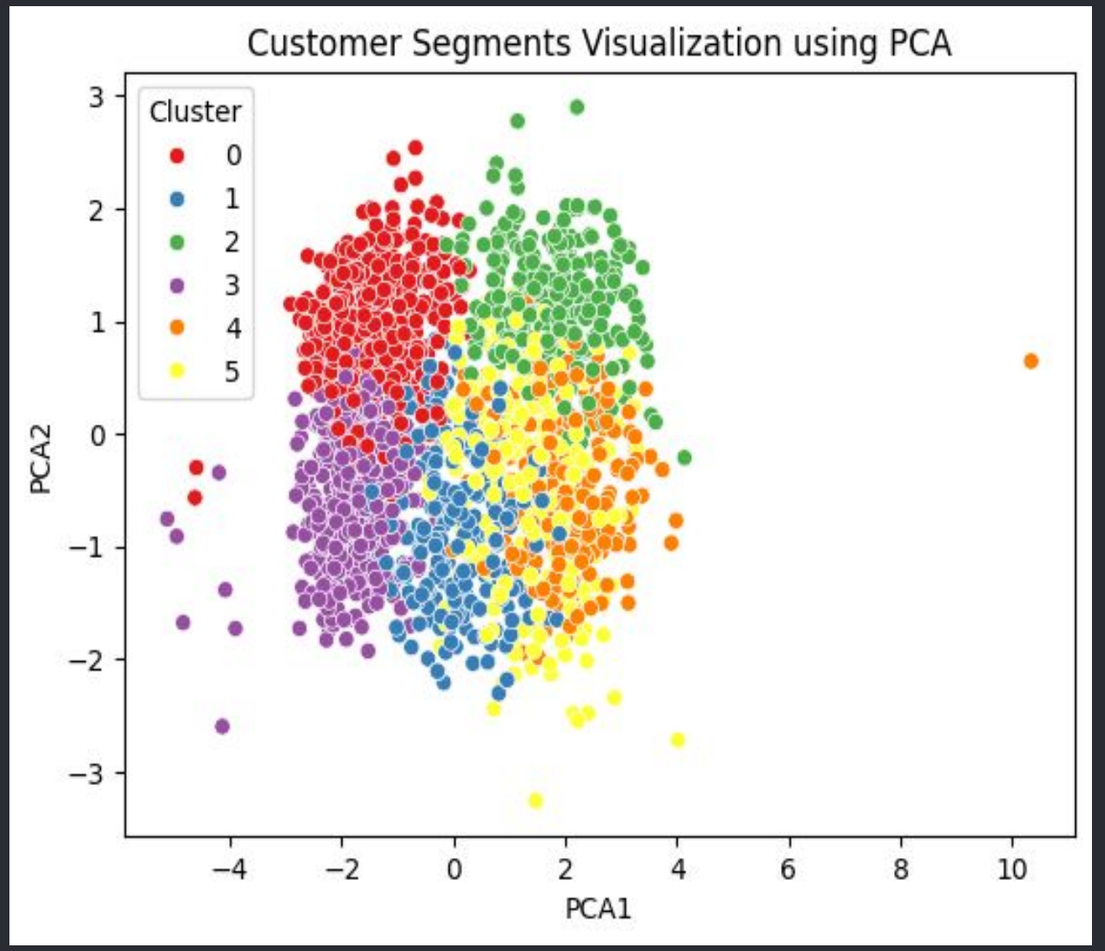
Outputs and Results






Average Income by Age Group
70+
60-69
50-59
40-49
30-39
18-29
Average Income
0
10000
20000
30000
40000
50000
60000
Age Group


Elbow Method for Optimal k
Within-cluster sum of squares (WCSS)
10000
9500
9000
8500
8000
7500
7000
6500
2
3
4
5
6
7
8
9
Number of clusters (k)



Customer Segmentation Prediction

Enter customer details to predict the segment.

Age (18-100)

30

Income (0-200000)

50000.00

Total Spending(sum of all purchases) (0-5000)

1000.00

Number of Web Purchases (0-100)

10

Number of Store Purchases (0-100)

10

Number of Web Visits (0-50)

10

Recency (days since last purchase) (0-365)

30

Predict Segment

Predicted Segment: Cluster 2