Infotact Solutions

Customer Segmentation using K-Means

INTERNSHIP | **TEAM 2 - 2 Months** | DATA ANALYTICS

Problem Statement

Customer Segmentation using K-Means

- Businesses struggle with **generic marketing** → wasted ad spend.
- Customers have different spending behaviors.
- Goal: Segment customers into groups for personalized marketing.

Dataset Overview

Customer Segmentation using K-Means

- **Source**: customer_shopping_data.csv from Kaggle
- 143,622 transactions, 10 columns
- Key features: Age, Gender, Category, Quantity, Price, Payment method, Mall
- Granularity: Transaction-level → aggregated to customer-level

```
✅ Dataset loaded successfully!
Shape: (143622, 10)

First 5 rows:
```

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall
0	I138884	C241288	Female	28.0	Clothing	5	1500.4	Credit Card	5/8/2022	Kanyon
1	I317333	C111565	Male	21.0	Shoes	3	1800.51	Debit Card	12/12/2021	Forum Istanbul
2	I127801	C266599	Male	20.0	Clothing	1	300.08	Cash	9/11/2021	Metrocity
3	I173702	C988172	Female	66.0	Shoes	5	3000.85	Credit Card	16/05/2021	Metropol AVM
4	I337046	C189076	Female	53.0	Books	4	60.6	Cash	24/10/2021	Kanyon

Data Cleaning (Week 1)

Customer Segmentation using K-Means

- Removed 31,000 duplicate rows
- Fixed missing values & datatypes
- Removed invalid values (negative prices, unrealistic ages)
- Final dataset: 112,002 clean records

Handle Duplicates

We check for duplicate rows and remove them to avoid biased analysis.

```
dup_count = df.duplicated().sum()
print("Duplicate rows before:", dup_count)

# Remove duplicates
df = df.drop_duplicates()

print("Duplicate rows after:", df.duplicated().sum())
```

```
Duplicate rows before: 31615
Duplicate rows after: 0
```

```
Cleaned dataset shape: (112002, 10)
```

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall
0	I138884	C241288	Female	28.0	Clothing	5	1500.40	Credit Card	2022-08-05	Kanyon
1	I317333	C111565	Male	21.0	Shoes	3	1800.51	Debit Card	2021-12-12	Forum Istanbul
2	I127801	C266599	Male	20.0	Clothing	1	300.08	Cash	2021-11-09	Metrocity
3	I173702	C988172	Female	66.0	Shoes	5	3000.85	Credit Card	2021-05-16	Metropol AVM
4	I337046	C189076	Female	53.0	Books	4	60.60	Cash	2021-10-24	Kanyon

Feature Engineering

Customer Segmentation using K-Means

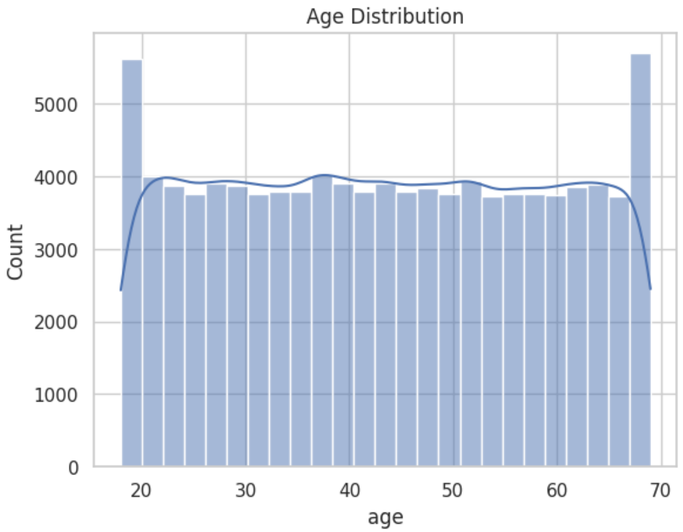
- Aggregated features per customer:
- Total spend
- Total quantity
- #Unique categories, malls
- Recency (days since last purchase)
- Invoice count
- Added line_amount = quantity × price

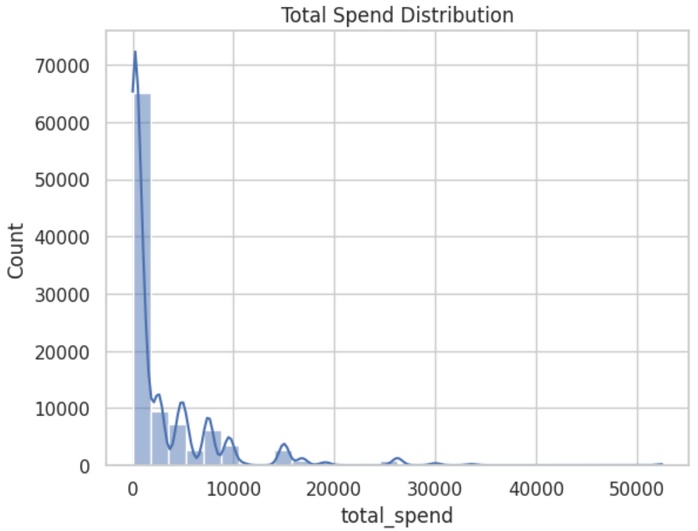
	customer_id	gender	age	total_spend	total_quantity	unique_categories	unique_malls	first_purchase	last_purchase	num_
0	C100004	Male	61.0	7502.00	5	1	1	2021-11-26	2021-11-26	
1	C100005	Male	34.0	2400.68	2	1	1	2023-03-03	2023-03-03	
2	C100006	Male	44.0	322.56	3	1	1	2022-12-01	2022-12-01	
3	C100012	Male	25.0	130.75	5	1	1	2021-08-15	2021-08-15	
4	C100019	Female	21.0	35.84	1	1	1	2021-07-25	2021-07-25	

Exploratory Data Analysis (EDA)

Customer Segmentation using K-Means

- Age Distribution: 20–60 dominates
- Spend Distribution: Skewed, few high spenders
- Gender: Balanced
- Recency: Many inactive
- Correlations: Spend ↔ Quantity strong

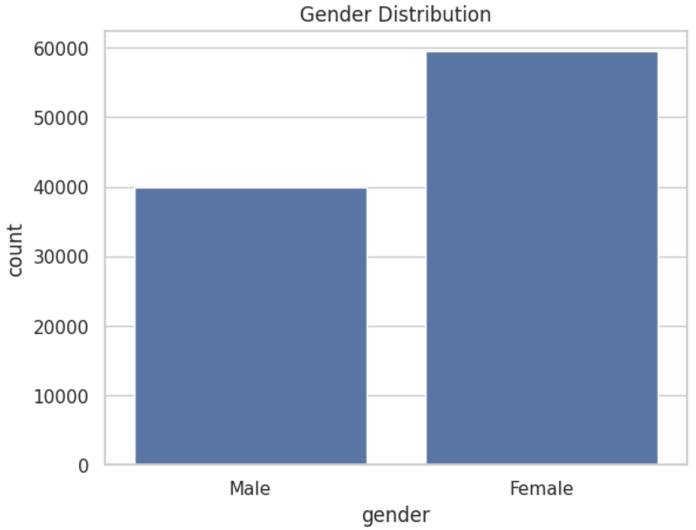


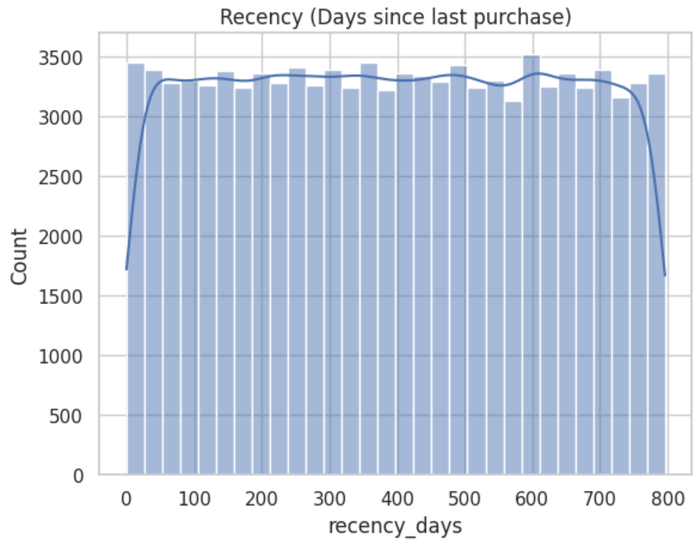


Exploratory Data Analysis (EDA)

Customer Segmentation using K-Means

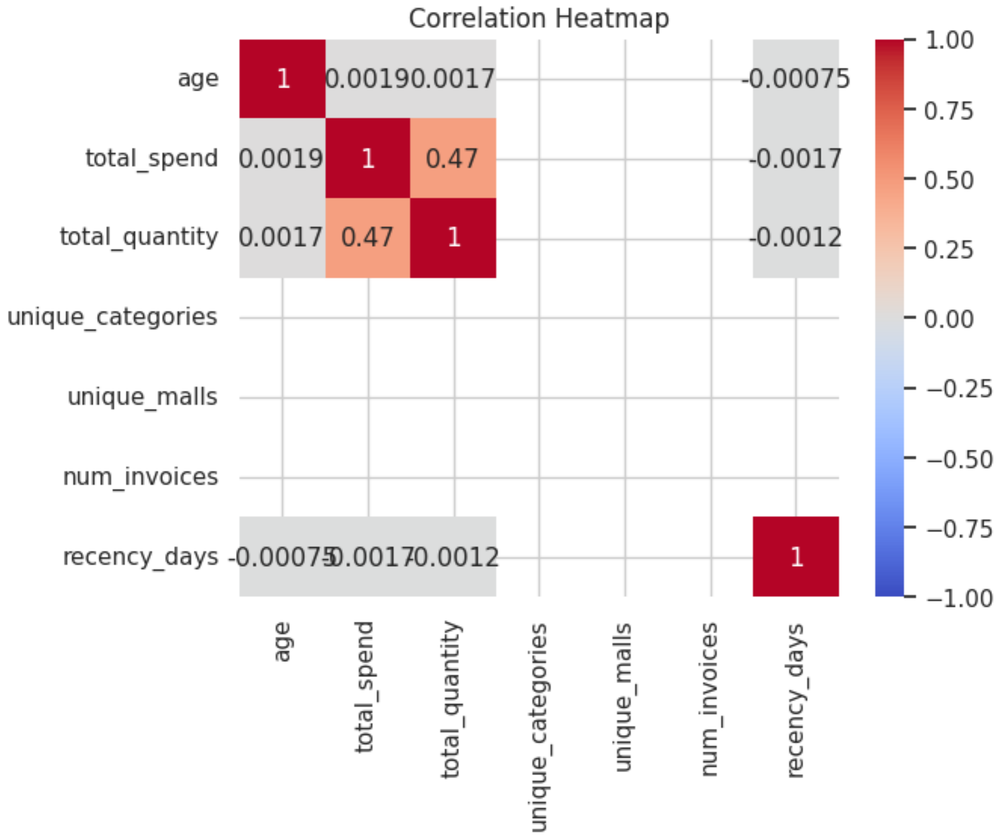
- Age Distribution: 20–60 dominates
- Spend Distribution: Skewed, few high spenders
- Gender: Balanced
- Recency: Many inactive
- Correlations: Spend ↔ Quantity strong





Exploratory Data Analysis (EDA)

Customer Segmentation using K-Means



Scaling

Customer Segmentation using K-Means

- Features on different scales (spend vs quantity vs recency).
- Applied **StandardScaler** to normalize.
- Saved: X_scaled.npy, scaler.pkl

```
Scaling complete. Files saved:
- customer_level_week1.csv
- X_scaled.npy
- scaler.pkl
```

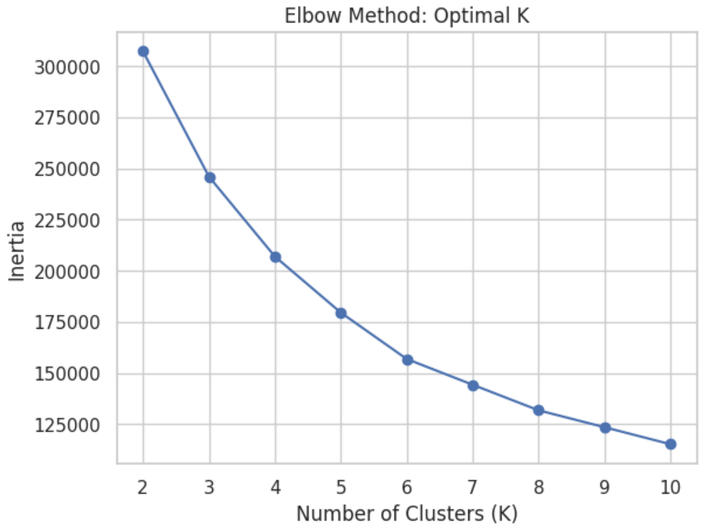
customer_level_week1

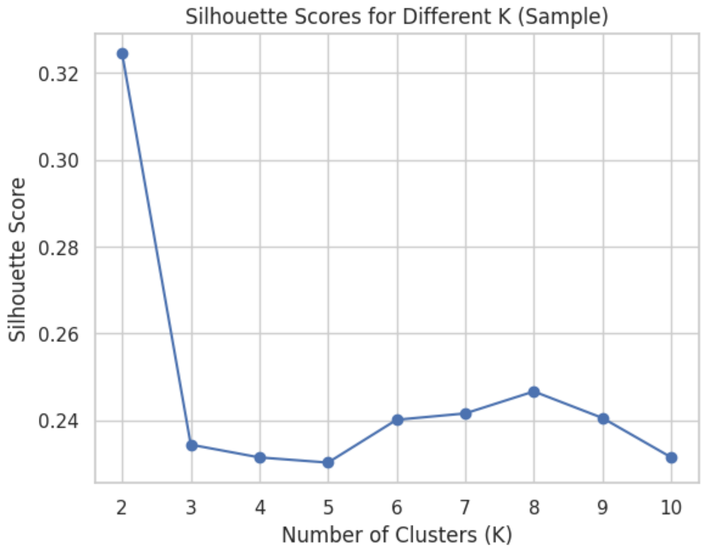
customer_id	gender	age	total_spend	total_quantity	unique_categories	unique_malls	first_purchase	last_purchase	num_invoices	recency_days
C100004	Male	61.0	7502.0	5	1	1	2021-11-26	2021-11-26	1	467
C100005	Male	34.0	2400.68	2	1	1	2023-03-03	2023-03-03	1	5
C100006	Male	44.0	322.56	3	1	1	2022-12-01	2022-12-01	1	97
C100012	Male	25.0	130.75	5	1	1	2021-08-15	2021-08-15	1	570
C100019	Female	21.0	35.84	1	1	1	2021-07-25	2021-07-25	1	591
C100025	Male	55.0	143.36	2	1	1	2021-06-03	2021-06-03	1	643
C100028	Female	21.0	15.15	1	1	1	2021-11-25	2021-11-25	1	468
C100030	Male	46.0	4801.28	4	1	1	2022-10-13	2022-10-13	1	146
C100034	Male	60.0	1200.32	2	1	1	2021-08-06	2021-08-06	1	579
C100041	Female	50.0	2700.7200000000000	3	1	1	2021-04-23	2021-04-23	1	684
C100042	Female	28.0	650.56	4	1	1	2022-01-04	2022-01-04	1	428
C100045	Male	48.0	5.23	1	1	1	2022-01-10	2022-01-10	1	422
C100066	Female	46.0	322.56	3	1	1	2022-08-02	2022-08-02	1	218
C100067	Female	62.0	600.17	1	1	1	2021-01-06	2021-01-06	1	791
C100078	Male	64.0	60.6	2	1	1	2022-12-01	2022-12-01	1	97
C100088	Female	30.0	5.23	1	1	1	2021-03-05	2021-03-05	1	733
C100090	Female	62.0	162.64	2	1	1	2022-03-07	2022-03-07	1	366
C100095	Female	31.0	130.75	5	1	1	2022-10-09	2022-10-09	1	150
C100096	Male	40.0	2700.7200000000000	3	1	1	2021-10-03	2021-10-03	1	521
C100099	Male	64.0	136.35000000000000	3	1	1	2022-03-23	2022-03-23	1	350

K-Means Clustering

Customer Segmentation using K-Means

- Tested K=2 to 10
- Elbow Method → Optimal K=4
- **Silhouette Score** → Highest at 4 clusters
- Final model: **K=4**





PCA Visualization

Customer Segmentation using K-Means

- PCA reduced 7 features → 2 dimensions
- Clear separation of clusters observed
- Cluster boundaries show good segmentation

Cluster Profiles

Customer Segmentation using K-Means

- **Cluster 0:** Young, moderate spend → Growth segment
- **Cluster 1:** Middle-aged, medium spend → Regulars
- **Cluster 2:** Older, low spend → Occasional shoppers
- **Cluster 3:** High-spending → VIP/Premium

cluster	age	total_spend	total_quantity	unique_categories	unique_malls	num_invoices	recency_days
0	29.5	1790.5	3.2	1.0	1.0	1.0	252.9
1	43.5	1822.6	3.2	1.0	1.0	1.0	638.8
2	57.5	1834.8	3.2	1.0	1.0	1.0	249.8
3	43.4	17689.7	6.5	1.0	1.0	1.0	396.2

```
✅ Clustered dataset loaded
cluster
1    34949
0    29324
2    28753
3     6427
Name: count, dtype: int64
```

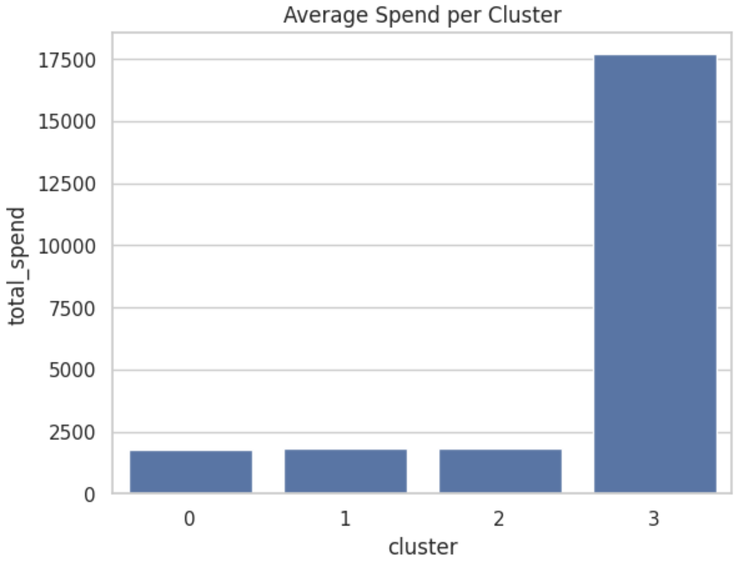
	customer_id	gender	age	total_spend	total_quantity	unique_categories	unique_malls	first_purchase	last_purchase	num_
0	C100004	Male	61.0	7502.00	5	1	1	2021-11-26	2021-11-26	
1	C100005	Male	34.0	2400.68	2	1	1	2023-03-03	2023-03-03	
2	C100006	Male	44.0	322.56	3	1	1	2022-12-01	2022-12-01	
3	C100012	Male	25.0	130.75	5	1	1	2021-08-15	2021-08-15	
4	C100019	Female	21.0	35.84	1	1	1	2021-07-25	2021-07-25	

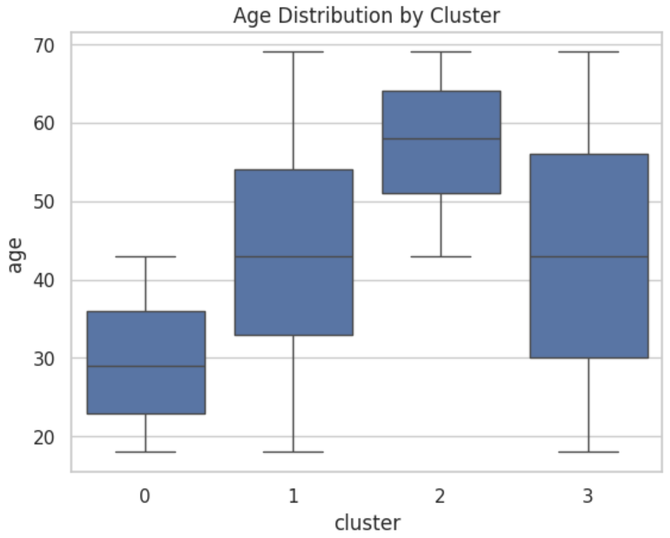
Cluster Distribution Visualizations

Customer Segmentation using K-Means

We plot visualizations to better understand cluster profiles:

- **Bar Plot** → Avg spend per cluster
- **Box Plot** → Age distribution by cluster



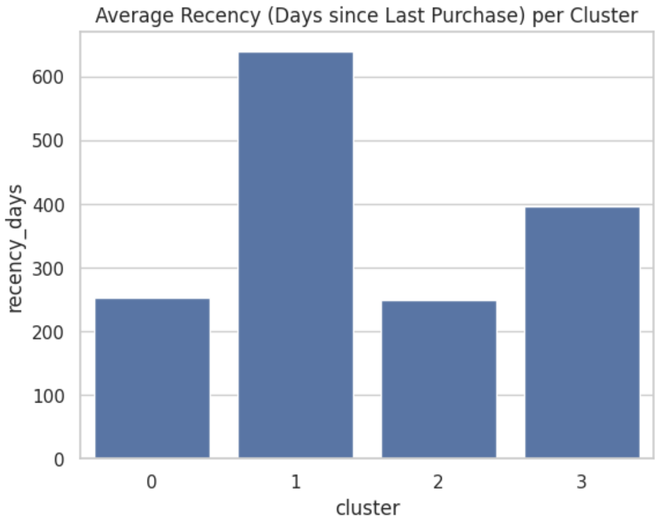


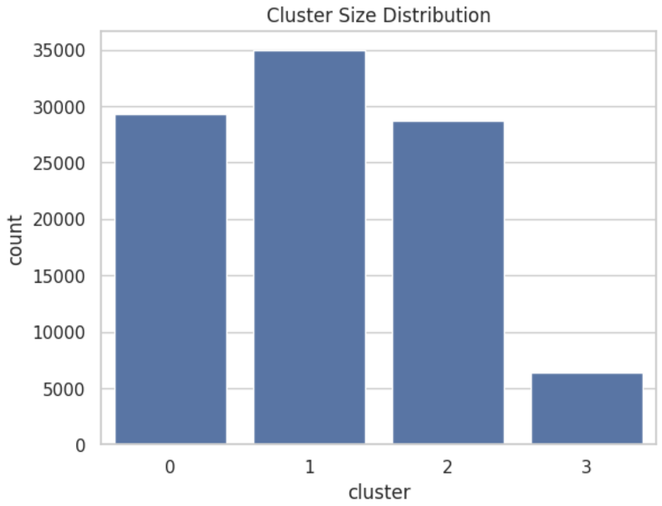
Cluster Distribution Visualizations

Customer Segmentation using K-Means

We plot visualizations to better understand cluster profiles:

- **Bar Plot** → Avg spend per cluster
- **Box Plot** → Age distribution by cluster
- **Bar Plot** → Recency by cluster
- **Count Plot** → Cluster sizes





Marketing Strategies

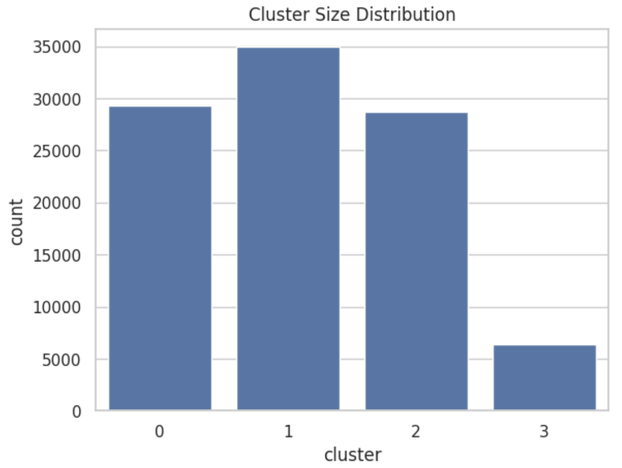
Customer Segmentation using K-Means

We plot visualizations to better understand cluster profiles:

- **Bar Plot** → Recency by cluster
- **Count Plot** → Cluster sizes

From the summary and visualizations, we can profile clusters like this (example):

- **Cluster 0**: Young, high-spending, frequent buyers → Premium customers
- **Cluster 1**: Middle-aged, medium spend, moderate frequency → Regulars
- **Cluster 2**: Older, low spend, infrequent purchases → Occasional shoppers
- **Cluster 3**: Young but low recency (haven't purchased recently) → At-risk customers

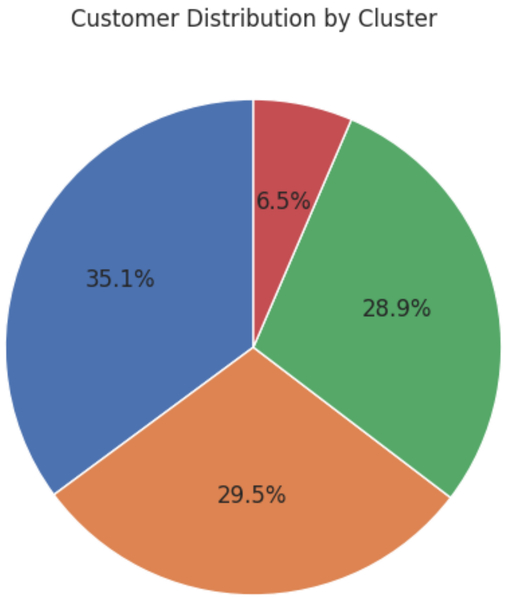


Visualizations for Report

Customer Segmentation using K-Means

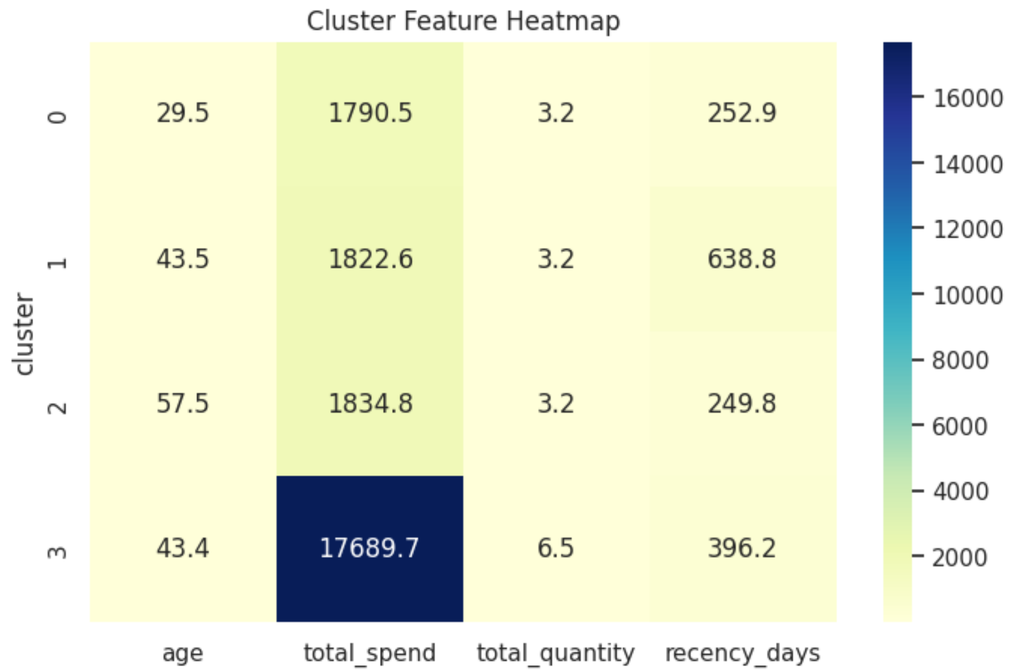
We now create professional visuals:

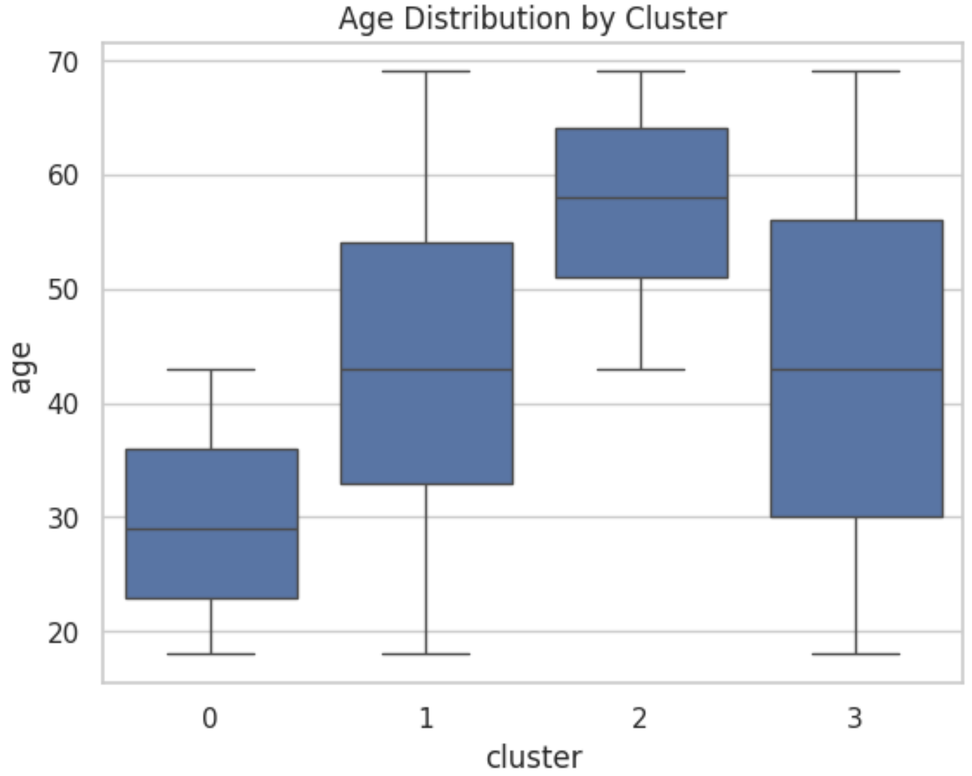
- Donut chart → % customers per cluster
- Bar chart → Avg Spend by cluster
- Box plot → Age distribution per cluster
- Heatmap → Correlation of features within clusters



Visualizations for Report

Customer Segmentation using K-Means





Final Results

Customer Segmentation using K-Means

- Segmented 99,453 customers into 4 clusters
- Generated final_customer_clusters.csv
- Ready for CRM integration

Key Findings:

- **Cluster 0** → Young, high-spending, frequent → Premium segment
- **Cluster 1** → Middle-aged, moderate spending → Regular customers
- **Cluster 2** → Older, low-spending → Occasional shoppers
- **Cluster 3** → At-risk, low recency → Need re-engagement

Marketing Recommendations:

- Premium → Loyalty rewards, VIP access
- Regular → Product bundles, upselling campaigns
- Occasional → Discount offers, seasonal promotions
- At-risk → Win-back campaigns, festive deals

This segmentation enables **targeted marketing strategies** to increase revenue & engagement.