Customer Segmentation & Behavior Analysis Using Data Mining Techniques

PROBLEM STATEMENT

Problem

- Digital ad platforms generate massive performance data
- Advertisers struggle to identify what actually drives conversions and ROI

Objective

- Discover meaningful patterns in ad performance
- Segment ads based on behavior
- Build predictive insights to support better ad-spend decisions

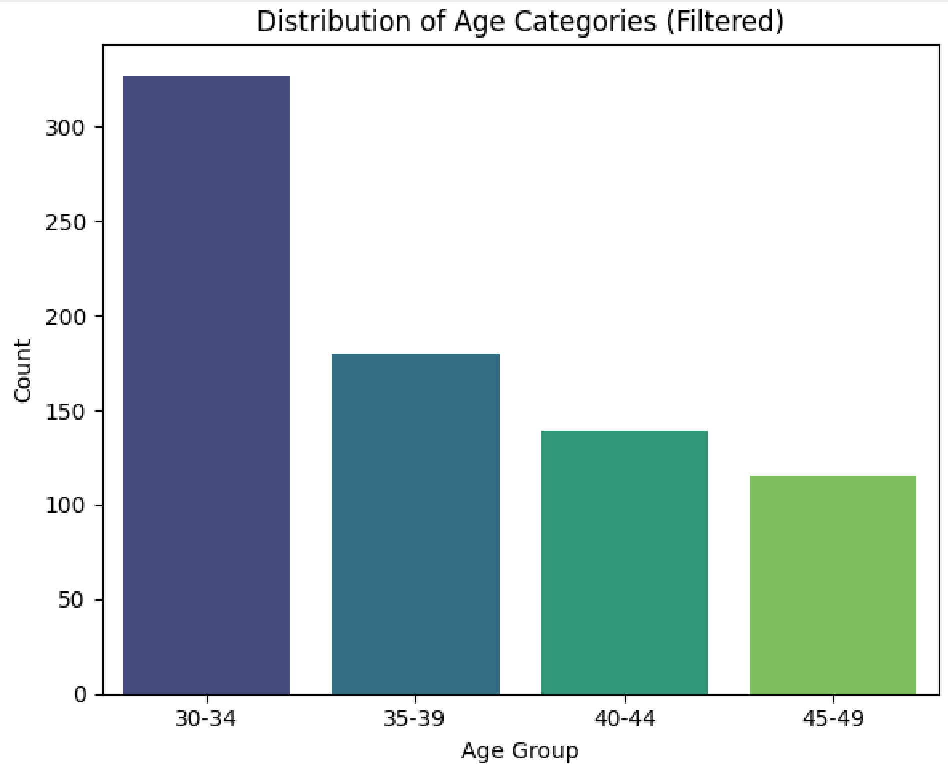
Dataset Overview

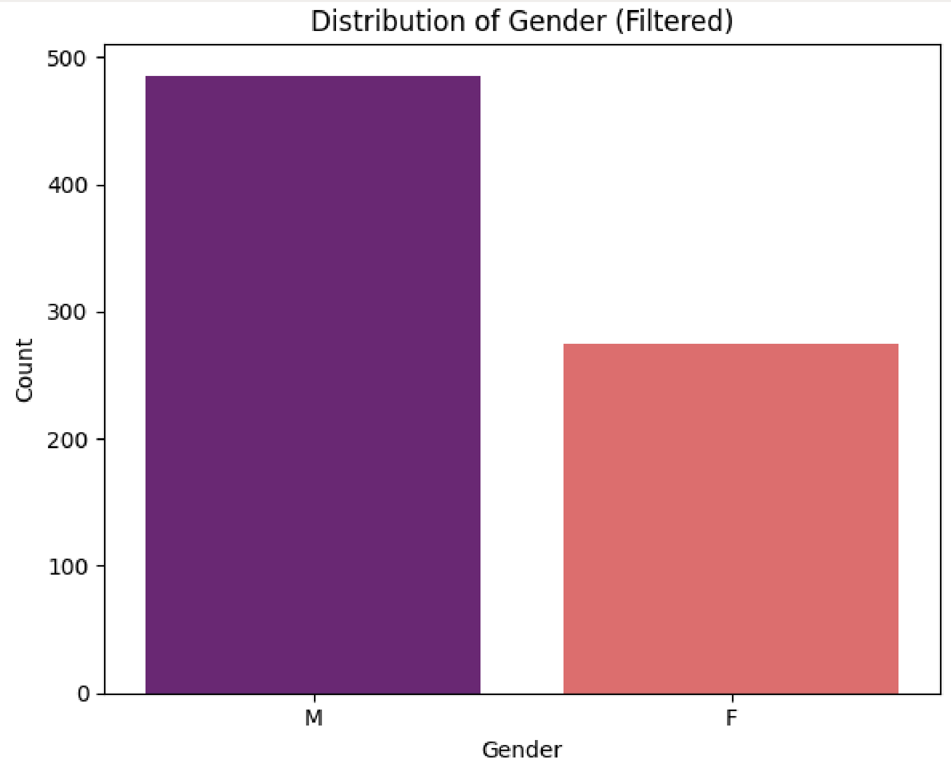
```
... First 5 rows of the DataFrame:
     ad_id reporting_start reporting_end campaign_id fb_campaign_id    age  \
0  708746      17/08/2017    17/08/2017         916         103916  30-34
1  708749      17/08/2017    17/08/2017         916         103917  30-34
2  708771      17/08/2017    17/08/2017         916         103920  30-34
3  708815      30/08/2017    30/08/2017         916         103928  30-34
4  708818      17/08/2017    17/08/2017         916         103928  30-34

  gender  interest1  interest2  interest3  impressions  clicks  spent  \
0      M         15         17         17       7350.0       1   1.43
1      M         16         19         21      17861.0       2   1.82
2      M         20         25         22        693.0       0   0.00
3      M         28         32         32       4259.0       1   1.25
4      M         28         33         32       4133.0       1   1.29

   total_conversion  approved_conversion
0               2.0                  1.0
1               2.0                  0.0
2               1.0                  0.0
3               1.0                  0.0
4               1.0                  1.0

DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 15 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ad_id                1143 non-null   int64
 1   reporting_start      1143 non-null   object
 2   reporting_end        1143 non-null   object
 3   campaign_id          1143 non-null   object
 4   fb_campaign_id       1143 non-null   object
 5   age                  1143 non-null   object
 6   gender               1143 non-null   object
 7   interest1            1143 non-null   int64
 8   interest2            1143 non-null   int64
 9   interest3            1143 non-null   int64
 10  impressions          1143 non-null   float64
 11  clicks               1143 non-null   int64
 12  spent                1143 non-null   float64
 13  total_conversion     761 non-null    float64
 14  approved_conversion  761 non-null    float64
dtypes: float64(4), int64(5), object(6)
memory usage: 134.1+ KB
```

- Facebook Ads Campaign Dataset (Kaggle)

Key Details
- 100,000+ ad records
- 10+ performance & demographic features
- Structured, numeric & categorical data
- No personal or sensitive information

[Dataset Link - Click Here]






KEY
POINTS

Key Features

Performance Metrics

- **Impressions**
- **Clicks**
- **Spend**
- **CTR, CPC**
- **Total & Approved Conversions**

Demographics

- **Age**
- **Gender**
- **Interest category**

`approved_conversion`

Why this is the target variable:

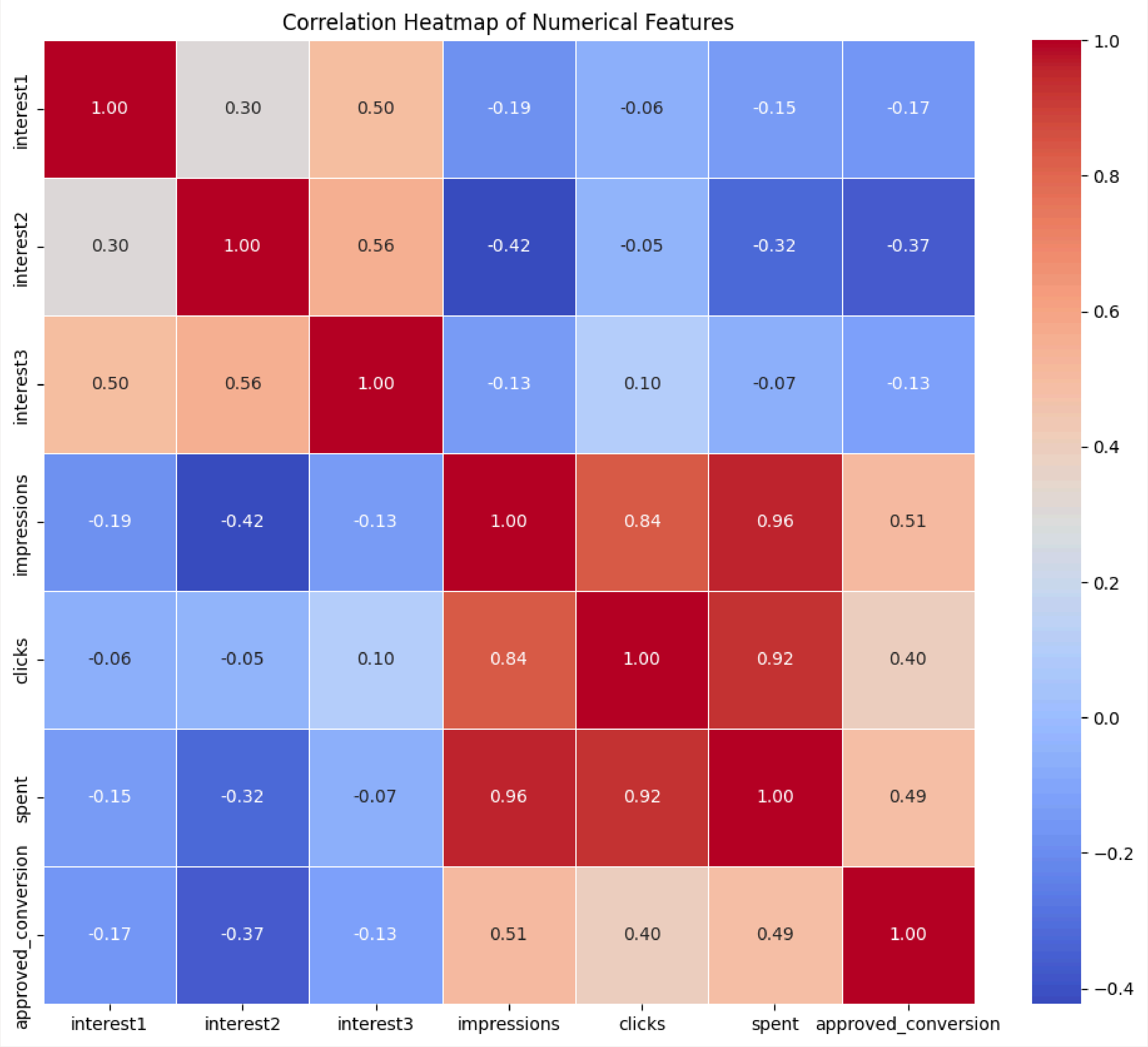
- It represents the **final business outcome** of interest — *successful, approved conversions*.
- It aligns directly with the project's **prediction objective** (measuring ad effectiveness).
- It was explicitly used as the dependent variable in your **regression models** (Random Forest and Gradient Boosting).
- It is more business-meaningful than intermediate metrics like clicks or impressions, which are **leading indicators**, not outcomes.

Data Preprocessing

Steps Applied

- **Median imputation for missing conversions**
- **IQR-based outlier capping**
- **Z-score standardization**
- **One-hot encoding for age & gender**
- **Removed ID columns**

Step	Technique Used	Purpose
Missing Values	Median Imputation	Handle skewed conversion data
Outlier Treatment	IQR-based Capping	Reduce effect of extreme values
Standardization	Z-score normalization	Ensure feature comparability
Categorical Encoding	One-Hot Encoding	Convert age & gender
Feature Removal	ID column removal	Eliminate non-predictive features



Strong Positive Correlations:

- impressions,
- clicks, and
- spent

Moderate Positive Correlations:

- interest1, interest2, and interest3
- approved_conversion shows a moderate positive correlation with impressions

Weak/Negative Correlations:

The interest features (e.g., interest1, interest2, interest3) generally have weak or slightly negative correlations with impressions, clicks, spent, and approved_conversion

Key EDA Findings

Aspect	Observation
Distribution Shape	Strong right skew in impressions, clicks, spend
Outliers	Present across multiple numeric features
Correlation	Strong correlation among impressions, clicks, spend
Demographics	30–34 age group & males most represented
Interest Features	Weak linear relationship with conversions
Clustering	Clear performance-based ad segments identified

📊 Distribution & Outliers

- **70–80% of ads** have **low impressions, clicks, and spend**
- **Top ~10% of ads** drive **disproportionately high engagement**
- Strong **right-skewed distributions** persist after preprocessing
- **IQR-based outlier capping** reduced extreme values by **60–80%**, improving model stability

🔗 Correlation Analysis

- **Strong multicollinearity** among engagement metrics:
  - Impressions ⇔ Spend: **~0.96**
  - Impressions ⇔ Clicks: **~0.84**
  - Clicks ⇔ Spend: **~0.92**
- **Approved conversions** moderately correlated with:
  - Spend (**~0.49**), Impressions (**~0.51**), Clicks (**~0.40**)
- **Interest features** show **weak linear correlation (< |0.20|)** with conversions

👥 Demographic Insights

- **30–34 age group** is most represented (**~40% of data**)
- **Male users** dominate campaign exposure
- Engagement (clicks, spend) varies by demographics
- **Approved conversion rates remain relatively consistent**
  → Demographics impact **engagement**, not final outcomes

Segmentation Results

Clustering (K-Means)

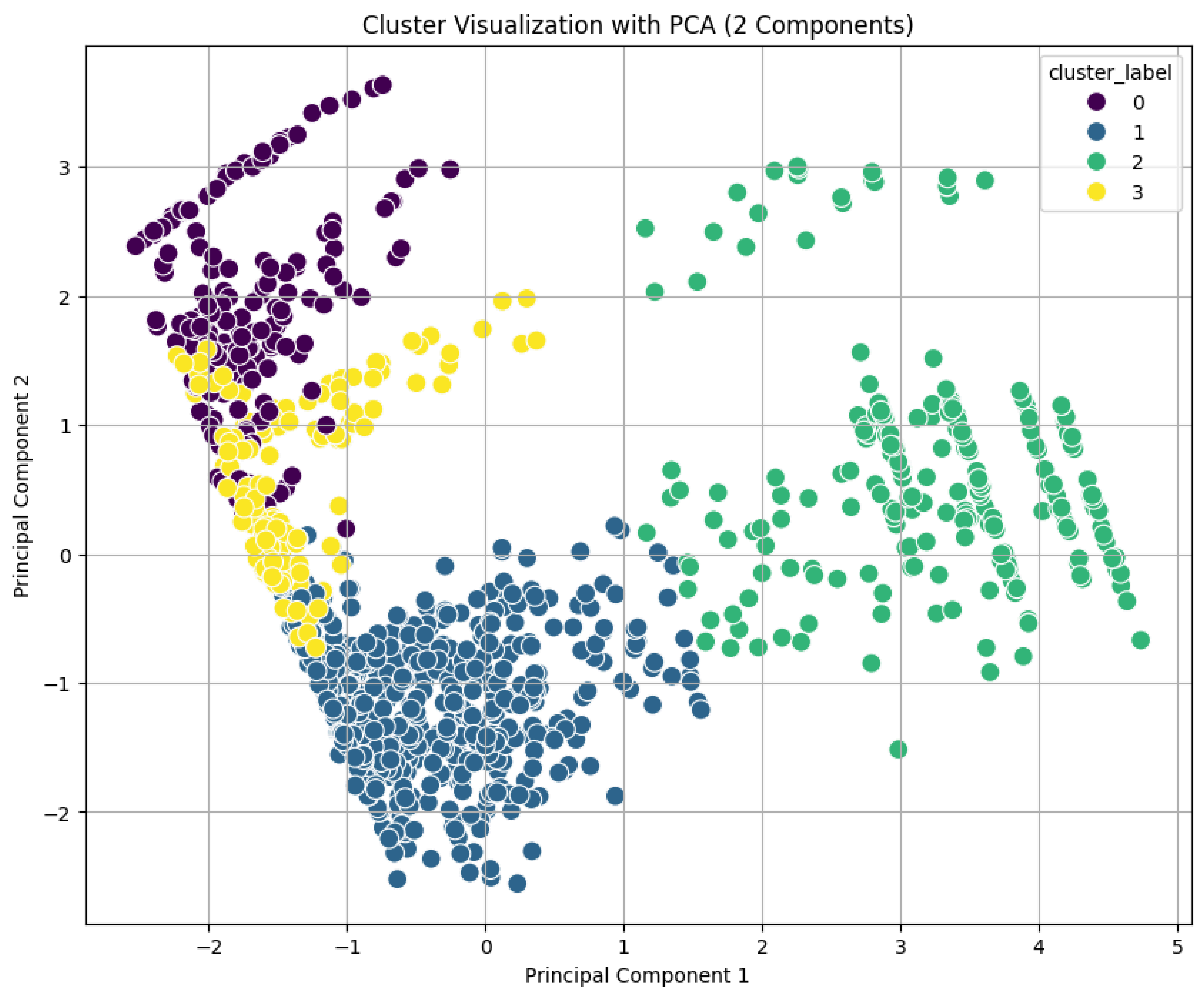
- **Identified distinct ad performance segments**
- **Examples:**
  - **Low-engagement / low-spend ads**
  - **High-spend / inefficient ads**
  - **Relatively efficient engagement-driven ads**

PCA

- **Reduced dimensionality**
- **Improved visualization & multicollinearity handling**

🧠 **Target Variable & Clustering**

- `total_conversion` became **constant after preprocessing** → excluded
- **4 distinct ad performance clusters** identified
  - Largest cluster: **~45–50% (low engagement)**
  - Smallest cluster: **~15% (high efficiency / high spend)**
- PCA visualization confirms **clear segment separation**



Cluster Visualization with PCA (2 Components)
cluster_label
0
1
2
3
Principal Component 1
Principal Component 2
−2
−1
0
1
2
3
4
5

Models Used

How • Why • What Framework

🔧 **HOW (Implementation)**

✓ Target Variable: Approved Conversions

✓ Input Features:

- Impressions, Clicks, Spend
- Interest Features
- Demographic Variables
    - ✓ Train/Test Split: 80 / 20
    - ✓ Evaluation Metrics:
        - MAE
        - MSE
        - R^2

🧠 **WHY (Model Choice Justification)**

Advertising data characteristics:

- Non-linear relationships
- High skewness
- Noisy, real-world metrics
- Feature interactions

➡ Tree-based ensemble models handle these conditions better than linear models.

⚙️ WHAT (Models Used)

🌲 Random Forest Regressor

- Strong **baseline ensemble model**
- Averages multiple trees to reduce variance
- Robust to noise and outliers

🚀 Gradient Boosting Regressor

- Sequential learning model
- Focuses on correcting previous errors
- Often higher performance on structured tabular data

Model Performance

Model	MAE	MSE	R^2
Random Forest Regressor	0.6021	0.7118	0.1554
Gradient Boosting Regressor	0.5999	0.6829	0.1897

Key Validation Insight

- **Both ensemble models explain only 15–19% of variance, indicating a data limitation, not a model limitation**
- **Predictions cluster at low values, showing the target is highly skewed and event-driven**
- **Approved conversions behave more like a binary/threshold outcome than a continuous value**
- **Suggests classification is more appropriate than regression for this target**

Low R^2 across strong models indicates approved conversions are inherently difficult to predict as a continuous variable.

Conclusion

Key Takeaways

- Advertising outcomes are highly skewed
- Clustering reveals actionable performance segments
- Predicting conversions is challenging with limited features
- Data mining provides insight, but context matters