Customer Segmentation using K-Means and PCA

# 🎯 Goal

To group similar customers together based on features like age, income, coverage amount, and premium — this helps businesses tailor marketing, pricing, and service offerings more effectively.

# ✅ Step-by-Step Breakdown

## 1. Loaded and Preprocessed the Data

- Read customer data from a CSV file.  
- Selected numerical features: ['Age', 'Income Level', 'Coverage Amount', 'Premium Amount']  
- Standardized them using StandardScaler to ensure no feature dominates due to scale.

💡 Why? Clustering algorithms like K-Means are sensitive to feature scales — scaling ensures fair contribution from all features.

## 2. K-Means Clustering (Before PCA)

- Applied K-Means with k=4 clusters on the original scaled data.  
- Visualized the clusters using two features: Income Level vs Coverage Amount.

📉 Observation: The clusters were split mostly based on income, showing poor contribution from other features. This may mean features were correlated or redundant.

## 3. Dimensionality Reduction with PCA

- Applied Principal Component Analysis (PCA) to reduce the 4 features to 2 principal components.  
- These 2 components retained ~50.6% of the original variance.

💡 Why PCA?  
- Helps remove multicollinearity between features.  
- Makes it easier to visualize the data in 2D.  
- Emphasizes the most important variance in data, reducing noise.

## 4. K-Means Clustering on PCA-Reduced Data

- Re-applied K-Means (k=4) on the PCA-transformed data.  
- Visualized clusters in 2D PCA space.

📉 Observation: Much cleaner, more balanced clusters — PCA allowed all features to influence the result fairly, leading to more meaningful segmentation.

## 5. Added Cluster Labels & Profiled Segments

- Added the new cluster labels (PCA\_Cluster) back to your DataFrame.  
- Calculated averages for each cluster group using .groupby().mean(numeric\_only=True).

💡 Why? This helps answer:  
- What defines each customer segment?  
- Which groups are high-income, low-premium, etc.?  
- How can we target each group?

# 🧠 Final Takeaways

|  |  |  |
| --- | --- | --- |
| Technique | Purpose | Benefit |
| Standardization | Normalize scales of features | Ensures fair clustering |
| K-Means | Unsupervised clustering method | Finds natural groupings in your customer base |
| PCA | Reduce dimensions, remove correlation | Improves clustering quality & enables visualization |
| Profiling | Understand characteristics of clusters | Enables actionable business decisions |