Detailed Analysis Report: Customer Segmentation using Clustering

1. Data Preparation and Preprocessing

Initial Data Loading

- Three datasets were imported: Transactions, Customers, and Products
- Used pandas for data manipulation and analysis
- Removed the 'Quantity' column from transactions data

Data Integration

1. Merged transaction data with product data to include category information
2. Created customer-level features:
    a. Spending by category (Books, Clothing, Electronics, Home Decor)
    b. Transaction frequency per customer

3. Final dataset included:
    - Customer spending across 4 categories
    - Transaction frequency
    - Total of 199 unique customers

Feature Standardization

- Applied StandardScaler to normalize all features
- This ensures all variables contribute equally to the clustering

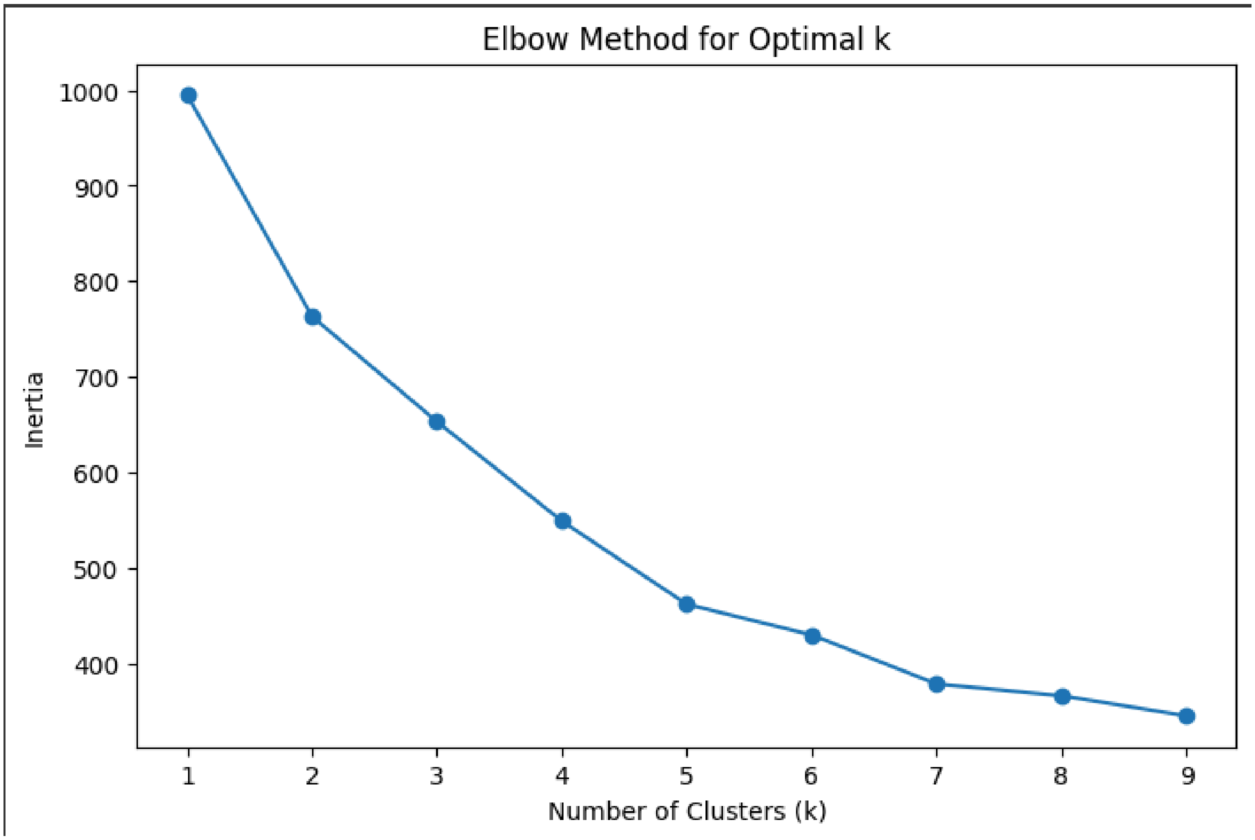
2. Clustering Analysis

Optimal Cluster Selection

Used multiple methods to determine the optimal number of clusters:

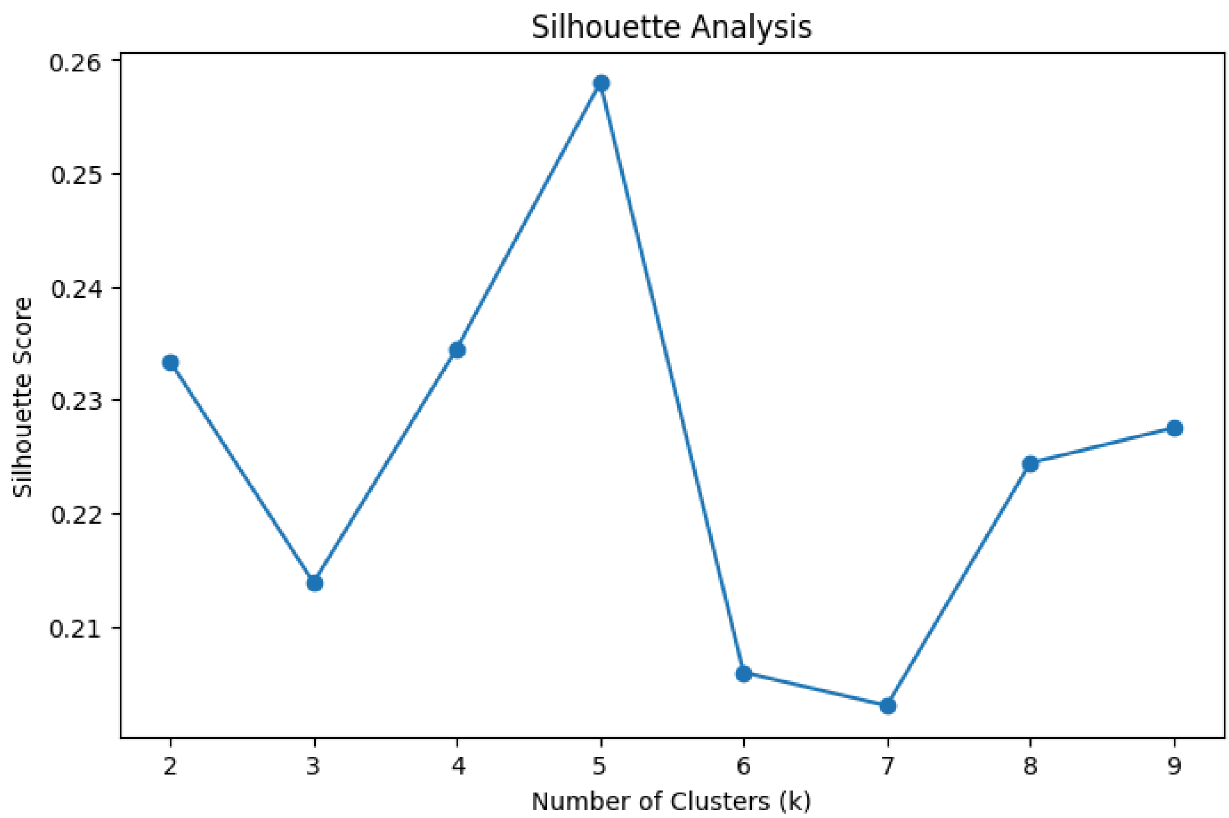
Elbow Method

- Plotted inertia (within-cluster sum of squares) for k=1 to 9
- Used to identify the point where adding more clusters provides diminishing returns



Silhouette Analysis

- Calculated silhouette scores for k=2 to 9
- Helps evaluate cluster separation and cohesion
- Optimal k was chosen as 5 based on these analyses



K-Means Clustering

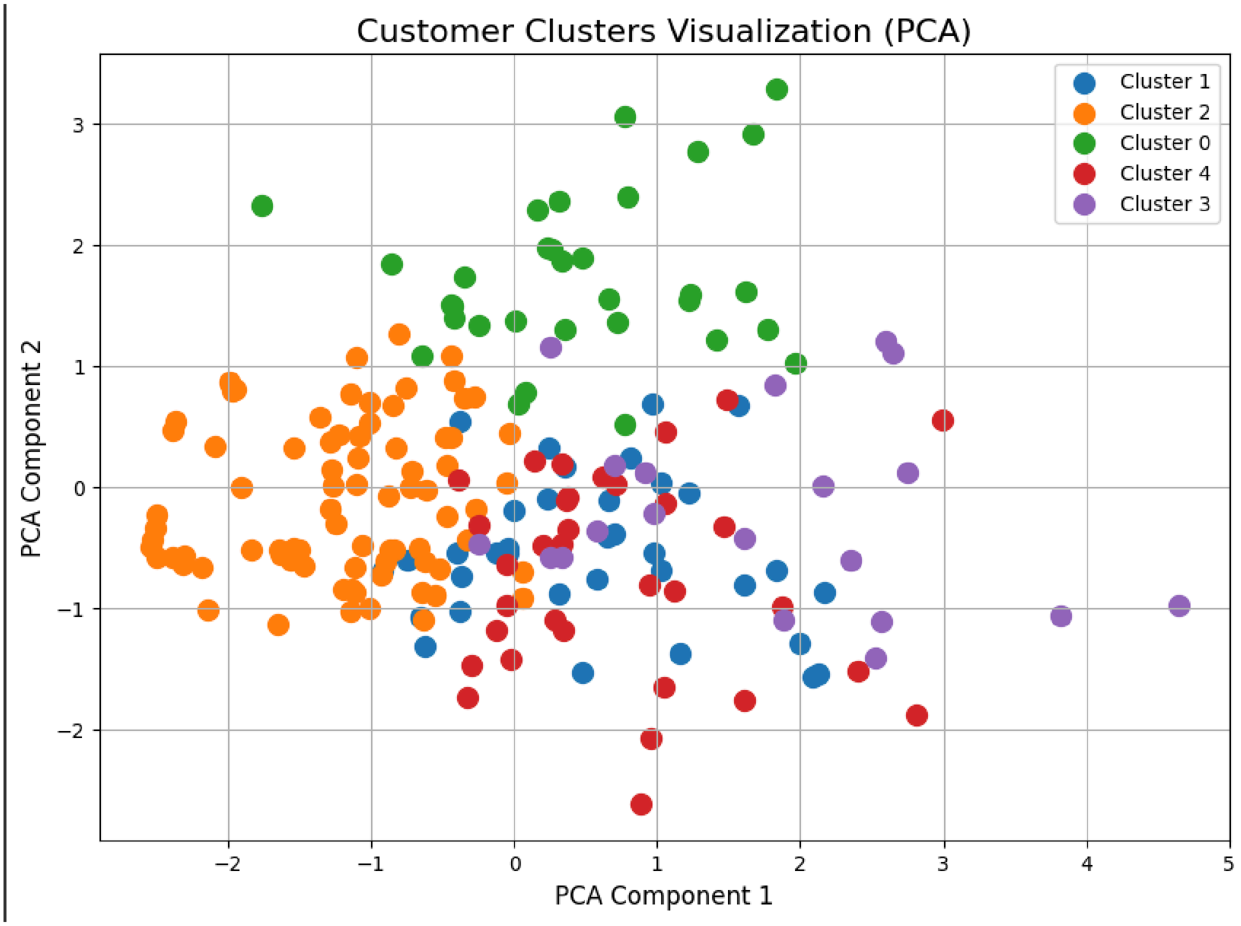
- Implemented K-means clustering with k=5
- Used random_state=42 for reproducibility
- Applied clustering to the standardized features

3. Visualization Techniques

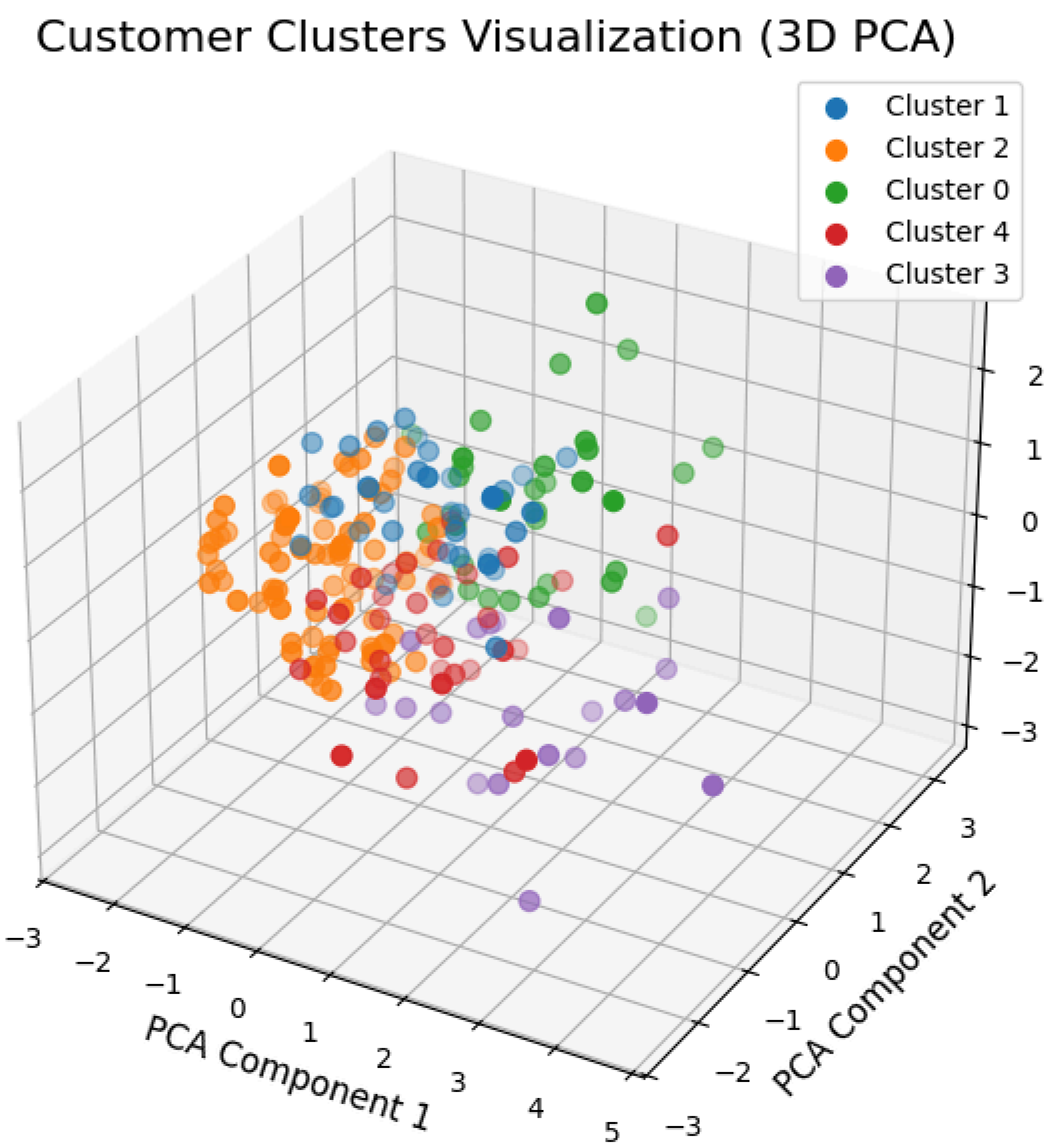
Multiple visualization approaches were used to understand the cluster distribution:

PCA (Principal Component Analysis)

1. 2D PCA Visualization
   a. Reduced dimensions to 2 components
   b. Plotted clusters with different colors
   c. Showed overall cluster separation

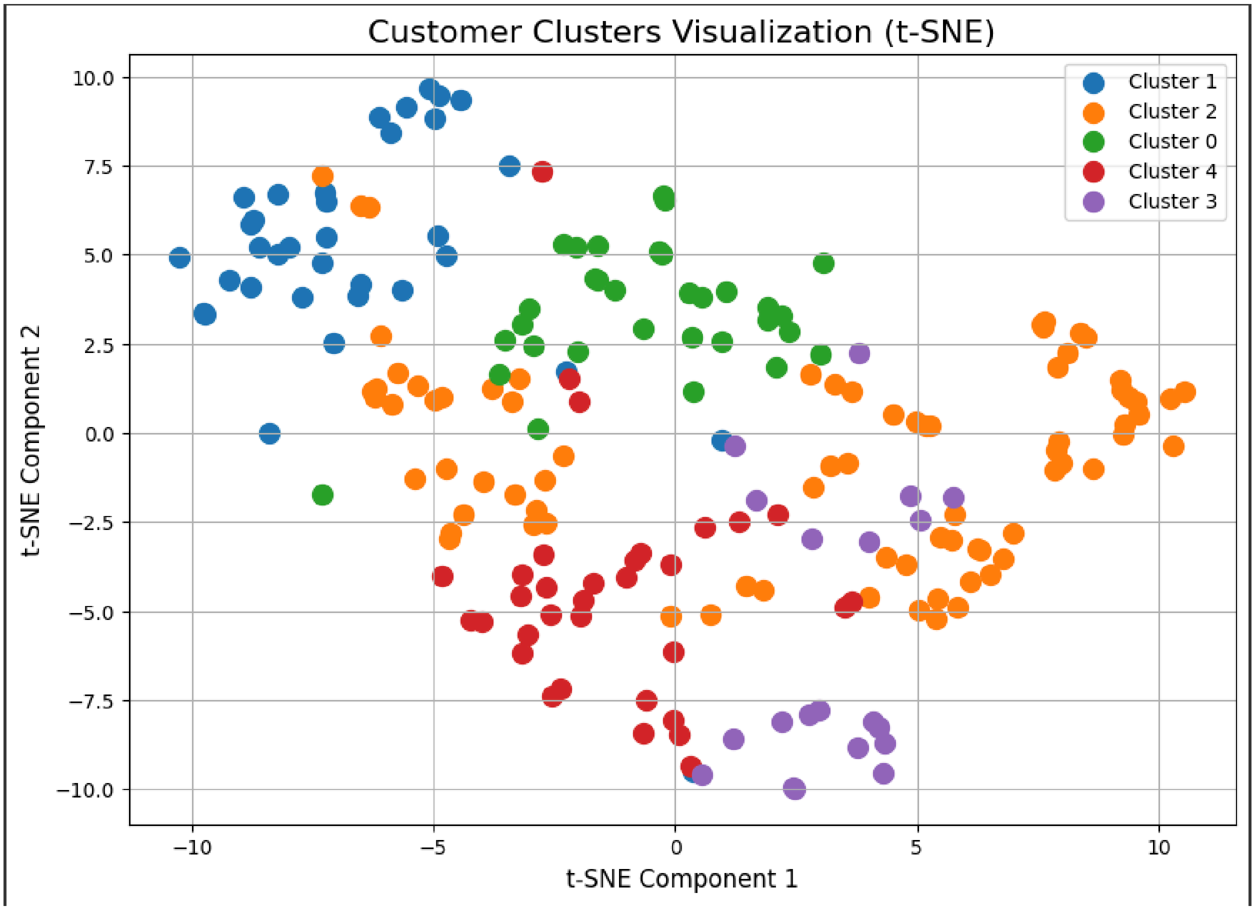


2. 3D PCA Visualization
    - Used 3 principal components
    - Created interactive 3D scatter plot
    - Provided better spatial understanding of clusters

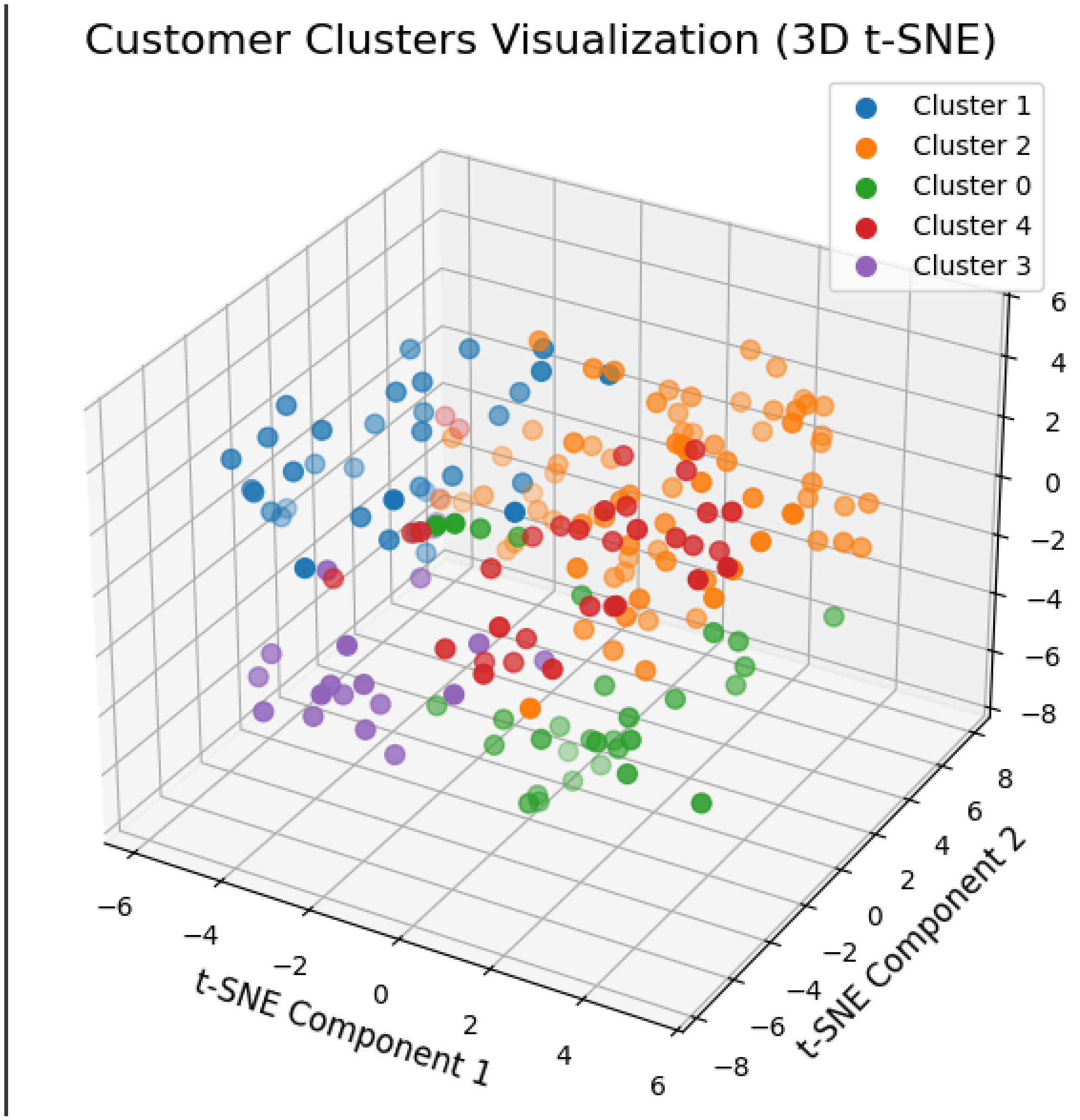


t-SNE Visualization

1. 2D t-SNE
    a. Non-linear dimensionality reduction
    b. Parameters: perplexity=30, n_iter=300
    c. Better for preserving local structure



2. 3D t-SNE
    - Extended visualization to 3 dimensions
    - Helped identify more subtle patterns



4. Cluster Validation

Several validation metrics were used to evaluate clustering quality:

Davies-Bouldin Index

- Measures average similarity between clusters
- Lower values indicate better clustering
- Davies-Bouldin Index: 1.1741396444956298

Silhouette Score

- Measures how similar objects are to their own cluster compared to other clusters
- Range: [-1, 1], higher is better

- Silhouette Score: 0.25799424578946467

Calinski-Harabasz Score
- Ratio of between-cluster dispersion and within-cluster dispersion
- Higher values indicate better defined clusters
- Calinski-Harabasz Score: 55.74712772949246

Dunn Index
- Ratio of minimum inter-cluster distance to maximum intra-cluster distance
- Higher values indicate better clustering
- Dunn Index: 0.09506783685716438

5. Technical Implementation
- Used Python's scientific computing stack (numpy, pandas)
- Leveraged scikit-learn for machine learning operations
- Visualization through matplotlib and seaborn
- 3D plotting capabilities using mpl_toolkits.mplot3d

This analysis provides a comprehensive customer segmentation that can be used for:
- Targeted marketing strategies
- Customer behavior analysis
- Product recommendations
- Resource allocation for different customer segments