Iowa shop clusterization

Lukas Sabastijauskas

Vilnius University, Faculty of Mathematics and Informatics

2025-06-17

Task presentation

Goal: Identify meaningful clusters of towns in the state of Iowa based on liquor sales patterns.

Motivation:

- The Iowa Liquor Sales dataset contains millions of transaction records from liquor stores across the state.
- By analyzing these data, I aim to uncover patterns of alcohol sales behavior at the town level.
- Clustering can reveal underlying structure in consumer behavior, regional preferences, or supply patterns.

Approach Overview:

1. Load and clean transactional liquor sales data using PySpark.
2. Engineer numeric features from the dataset (e.g., Bottle Volume, Bottles Sold).
3. Reduce dimensionality using PCA to capture key patterns.
4. Apply KMeans clustering to detect similar town-level sales profiles.
5. Visualize cluster distributions and map results geographically.

Methodology Overview

Principal Component Analysis (PCA):

- A dimensionality reduction technique that transforms correlated features into a smaller set of uncorrelated variables called principal components.
- Captures the most variance in the data with fewer components.
- Helps visualize and reduce noise before clustering.

KMeans Clustering:

- An unsupervised learning algorithm that partitions data into k clusters by minimizing within-cluster variance.
- Each point is assigned to the nearest cluster centroid.
- Requires specifying the number of clusters k.

Elbow Method:

- Used to determine the optimal number of clusters k.
- Plot the within-cluster sum of squares (WSS) against different values of k.
- The 'elbow' point indicates a suitable trade-off between model complexity and explained variance.

Process

- **Data Loading and Preprocessing:**
  - Selected meaningful numeric features for clustering.
  - Cleaned and converted data types.
- **Dimensionality Reduction with PCA:**
  - Applied PCA to reduce feature space.
  - Analyzed explained variance to select number of components.
  - Selected 2 principal components capturing most variance.
- **Choosing Number of Clusters:**
  - Used the Elbow method on Within-Cluster Sum of Squares (WSS).
  - Tested k values from 2 to 7.
  - Identified optimal clusters at $k = 5$ based on the elbow point.
- **Clustering with KMeans**
- **Results and Visualization:**

Results I - clusters information

Cluster	0	1	2	3	4
Pack	11.31	6.01	22.32	1.05	11.78
State Bottle Retail	15.05	17.13	7.37	45.03	15.02
Sold Bottles	7.92	9.38	10.42	2.41	15.26
Bottle Volume	747.17	1751.29	363.51	3048.99	1001.40
State Bottle Cost	10.01	11.38	4.92	30.02	10.01
Size in percentage	47.03%	20.62%	17.11%	14.91%	0.33%

- **Cluster 0 – Moderate Pack Buyers:** Medium pack size, moderate price and cost, with moderate bottle volume. Likely represents average customers.
- **Cluster 1 – Large Volume Buyers:** Small pack size but large bottle volume and relatively high price and cost. Possibly bulk buyers or commercial purchases.
- **Cluster 2 – Budget Shoppers:** Large packs with low price and cost. Likely cost-sensitive customers buying in value packs.
- **Cluster 3 – Premium :** Very high price and cost, large bottle volume but small pack size and sold quantity. Luxury buyers.
- **Cluster 4 – High Frequency Buyers:** Very high number of items sold. Likely frequent buyers or popular standard products.

Results II

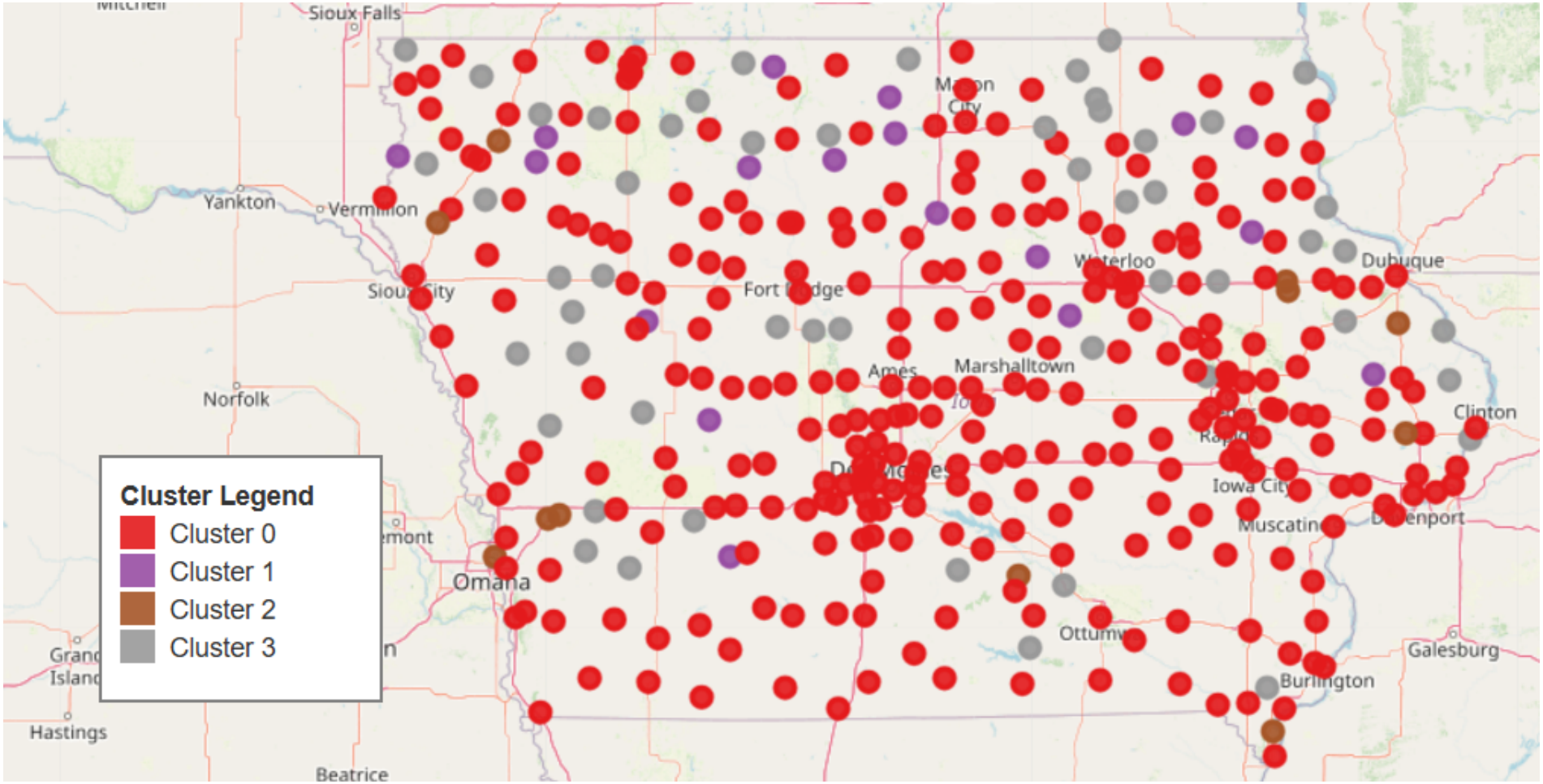

Figure: Clustered map

Conclusions

- **Efficient Data Processing:** The use of PySpark enabled fast and scalable data processing, especially valuable given the size of the Iowa Liquor Sales dataset.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied to reduce the dataset to 2 main components, capturing the most relevant variation and enabling clearer clustering.
- **Clustering Results:** KMeans clustering (with $k = 5$) successfully grouped towns based on similar sales patterns, with each cluster showing distinct characteristics.
- **Application Potential:** This clustering can be used to segment buyers and inform marketing strategies or logistics for alcohol distribution in Iowa.
- **Limitations in Visualization:** For the final city-based map, I assigned each city to its most frequent cluster label. While this offers a clean representation, it may oversimplify the full distribution of buyer types in diverse towns.
- **Overall Outcome:** The project successfully identified meaningful patterns in the sales data and produced an informative clustering and geographic overview, despite some visualization limitations.

Thank you for your attention