Data Mining and Discovery Report

Student Name: *Giridhar Reddy Goddilla*
Student ID: 23067661

Submission Date: 7 April 2025
Tutor: John Evans

Topic 1: Clustering-Based Anomaly Detection
Topic 2: K-Means Clustering
Dataset Used: Sales Transactions Dataset Weekly Dataset

Clustering - Based Anomaly Detection on Sales Transactions Weekly Data

1. Introduction

This project explores two essential techniques in unsupervised learning **K-Means Clustering** and **Clustering-Based Anomaly Detection** applied to weekly sales transaction data from the UCI Machine Learning Repository. The dataset records weekly sales volumes for various products over a 52-week period, providing a rich basis for identifying patterns, grouping behaviours, and detecting outliers.
The primary goal is twofold:

1. **Group** similar products based on their weekly sales patterns.
2. **Identify anomalies**, i.e., products that deviate significantly from typical group behaviours.

These insights are valuable in many business scenarios, including inventory control, sales forecasting, and fraud detection.

2. Data Preprocessing

Before applying machine learning techniques, the dataset was preprocessed through the following steps:

- **Data Exploration**: We confirmed the dataset contains no missing values, and used *.info*() and *.describe*() to understand structure and summary statistics.
- **Feature Selection**: Only the weekly sales columns (*W0* to *W51*) were selected as input features.
- **Standardization**: All selected features were scaled using *StandardScaler* to ensure consistent contribution to clustering distance metrics.
- **Dimensionality Reduction**: Principal Component Analysis (PCA) was applied to reduce the 52-dimensional data to two dimensions for visualization, while preserving key variance.

3. Implementation of Techniques

K-Means clustering was performed on the standardized dataset to discover natural groupings. To determine the optimal number of clusters, we evaluated both inertia and silhouette scores for a range of *k* values (from 2 to 9). The best silhouette score pointed to the ideal *k*, and the clustering model was trained accordingly.
This process successfully segmented products into groups based on similar sales behaviours. For example, some clusters represented consistently selling items, while others included products with seasonal or erratic sales.

3.2 Clustering-Based Anomaly Detection

To uncover outliers, we calculated the **Euclidean distance** from each product to its respective cluster center. Products in the **top 5% of distances** were flagged as potential anomalies the assumption being that extreme deviations from group norms may indicate unusual, erroneous, or exceptional behaviour.

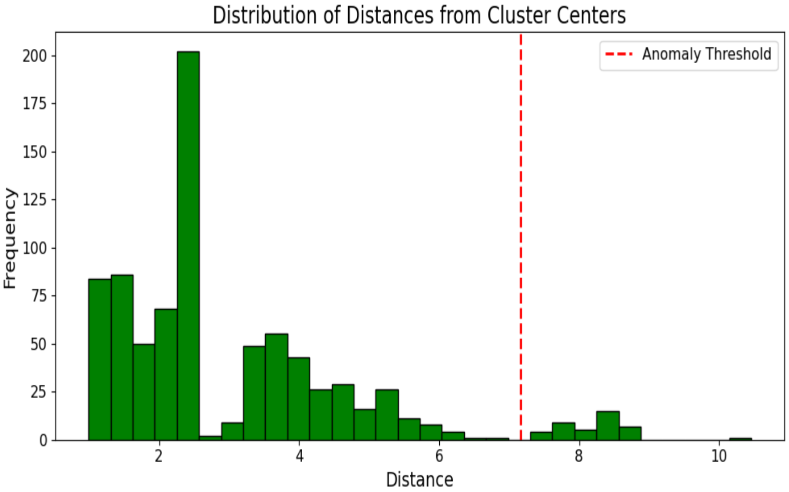

Figure 1 displays the distribution of distances across all data points, with a red dashed line marking the 95th percentile used as the threshold for identifying anomalies.

Figure 1: Distribution of distances from cluster centers.

Products to the right of the threshold are classified as anomalies.

To visualize the clustering and anomaly results, we used PCA-reduced features and plotted them in a 2D scatter plot. As shown in Figure 2, products marked as **anomalies** appear as **pink crosses**, while **regular items** are represented by **sky-blue circles**. A visual marker highlights the most extreme outlier.

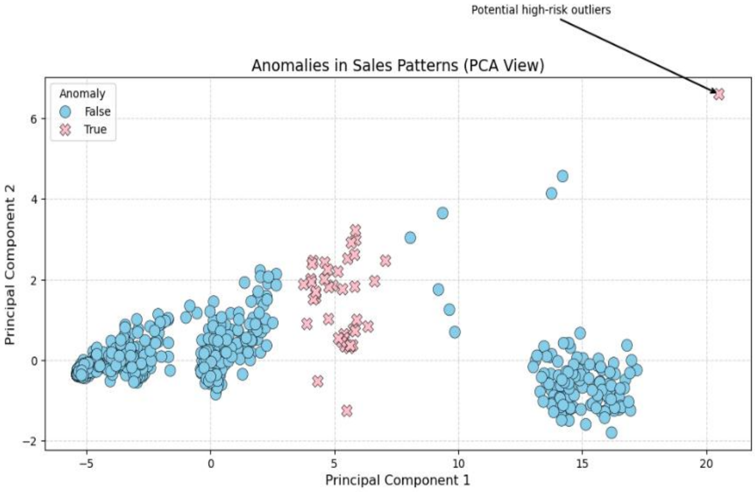

Figure 2:

PCA scatter plot of products. Sky-blue circles represent typical products; pink crosses highlight flagged anomalies.

These visualizations offer a powerful way to interpret the results and support business decision-making.

4.Comparison of Technique

Technique	**Purpose**	**Strength**
K-Means Clustering	Group similar products by sales trends	Reveals patterns, supports segmentation
Distance-Based Anomaly Detection	Flag unusually behaving products	Identifies outliers for review or intervention

While K-Means organizes the structure of the data, the anomaly detection layer helps isolate instances that deviate from the norm. PCA enhances interpretability by making these high-dimensional relationships visible in two dimensions.

5. Key Insights and Findings

- The dataset segmented cleanly into **three primary product clusters**.
- A total of **24 anomalies** were identified, many of which showed erratic or extremely low sales.
- One product in particular stood out as a **major outlier**, potentially representing an error, discontinued item, or a highly seasonal product.
- The combined use of K-Means and anomaly detection enables both trend analysis and quality control in a unified framework.

6. Ethical Considerations

This project uses a publicly available dataset that is **fully anonymized**. There is no risk of disclosing any sensitive or personal information. However, as with all unsupervised learning models, it's important to interpret anomalies with caution. Anomalous data does not always indicate something wrong it may simply represent rare or valid behaviour. Business context and expert validation should be applied before making decisions based on these findings.

7.Conclusion

This project successfully applied **K-Means Clustering** and **Clustering-Based Anomaly Detection** to a real-world retail dataset. The workflow demonstrated how unsupervised learning can reveal both underlying structure and exception cases in complex data.

Through distance-based outlier detection and effective visualization using PCA, we provided a clear interpretation of the dataset that can support business applications like sales strategy, stock planning, and data quality checks.

Going forward, future improvements could include:

- Time-aware clustering for seasonal adjustment.
- Dynamic thresholds for anomaly sensitivity.
- Integration with supervised learning models for predictive analytics.

8. References

- UCI Machine Learning Repository. (n.d.). *Sales Transactions Dataset Weekly*. Retrieved from https://archive.ics.uci.edu/dataset/396/sales+transactions+dataset+weekly
- Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.
- Scikit-learn Developers. (2023). *Scikit-learn Documentation*. Retrieved from https://scikit-learn.org/stable/
- Aggarwal, C. C. (2013). *Outlier Analysis*. Springer.
- TensorFlow Developers. (2020). *Anomaly detection with TensorFlow | Workshop* [Video]. YouTube. https://www.youtube.com/watch?v=2K3ScZp1dXQ
- Jolliffe, I. T. & Cadima, J. (2016). *Principal component analysis: A review and recent developments*. Philosophical Transactions A, 374(2065), 20150202.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). *Anomaly Detection: A Survey*. ACM Computing Surveys, 41(3), 15. https://doi.org/10.1145/1541880.1541882
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Academic lecture material on Clustering and Anomaly Detection
- Personal notes and recorded lectures from 360DigitMG training sessions