Unsupervised Clustering of White Wine

Based on Physicochemical Features

Masha S. Logan

Why Group Wine?

Wine quality assessment is critical for consumer satisfaction and economic success

- Traditional sensory analysis is subjective, inconsistent, and costly.
- Data-driven approaches offer potential for **objective**, **early**, and **cost-effective** quality prediction.

Goal: apply **unsupervised learning** to discover natural groupings of white wines based on physicochemical features *without using quality labels*.

Practical Motivation:

- Identify distinct chemical profiles.
- Predict quality trends before sensory evaluation.
- Support producers with **agile, data-driven quality control**.

Data Overview and Challenges

Dataset

- UCI Wine Quality Dataset
- 4,898 white wine samples
- 11 physicochemical properties (e.g., acidity, sugar, alcohol)
- No missing values

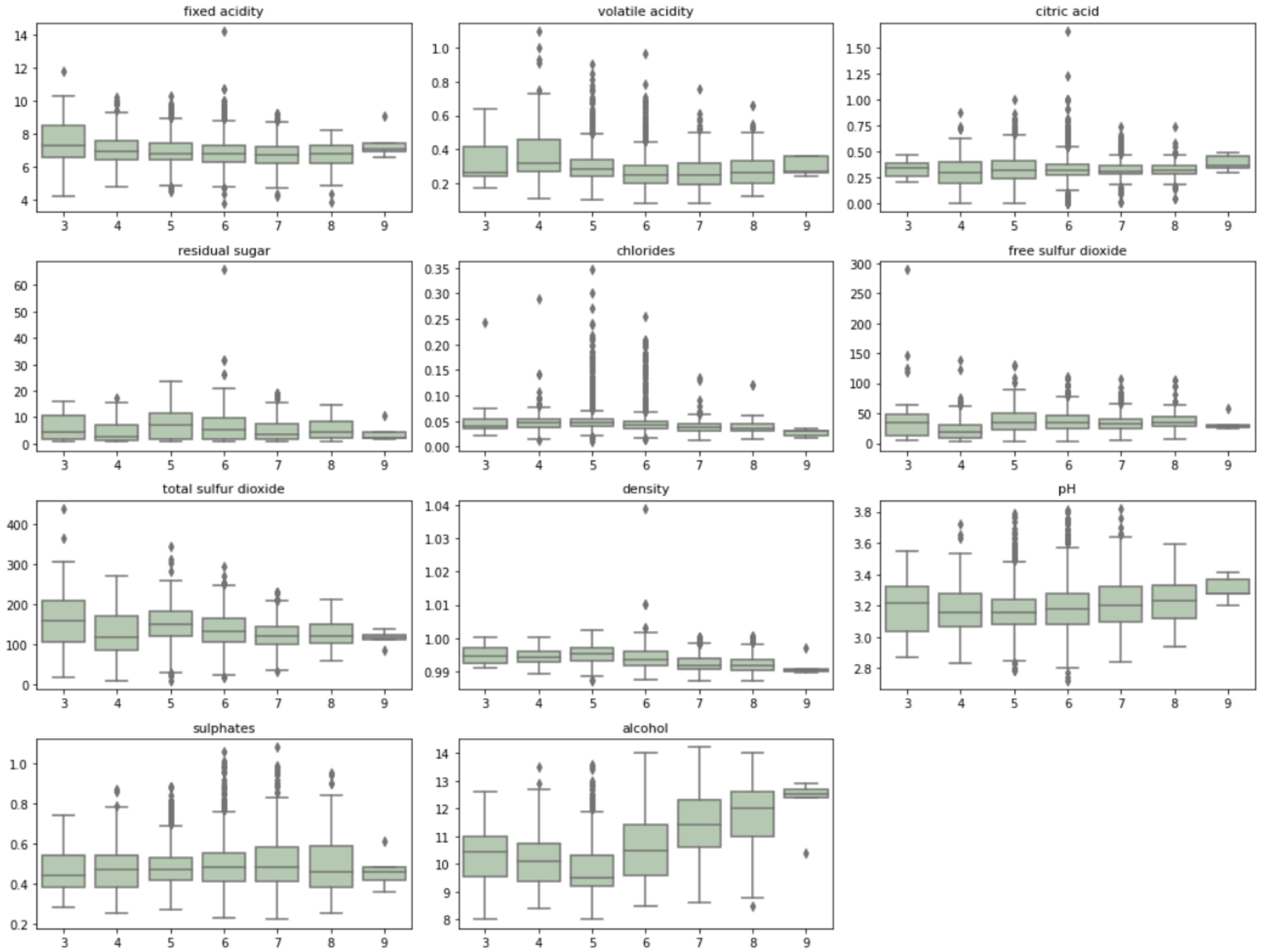
Key Challenges:

- **Class Imbalance**: Majority of wines clustered around mid-range quality scores (5 and 6).
- **Subtle Feature Variation**: Physicochemical differences between quality levels are minor and continuous, not sharply distinct.
- **Limited Scope**: Dataset lacks sensory notes, grape varieties, or vintage year — only chemical measurements available

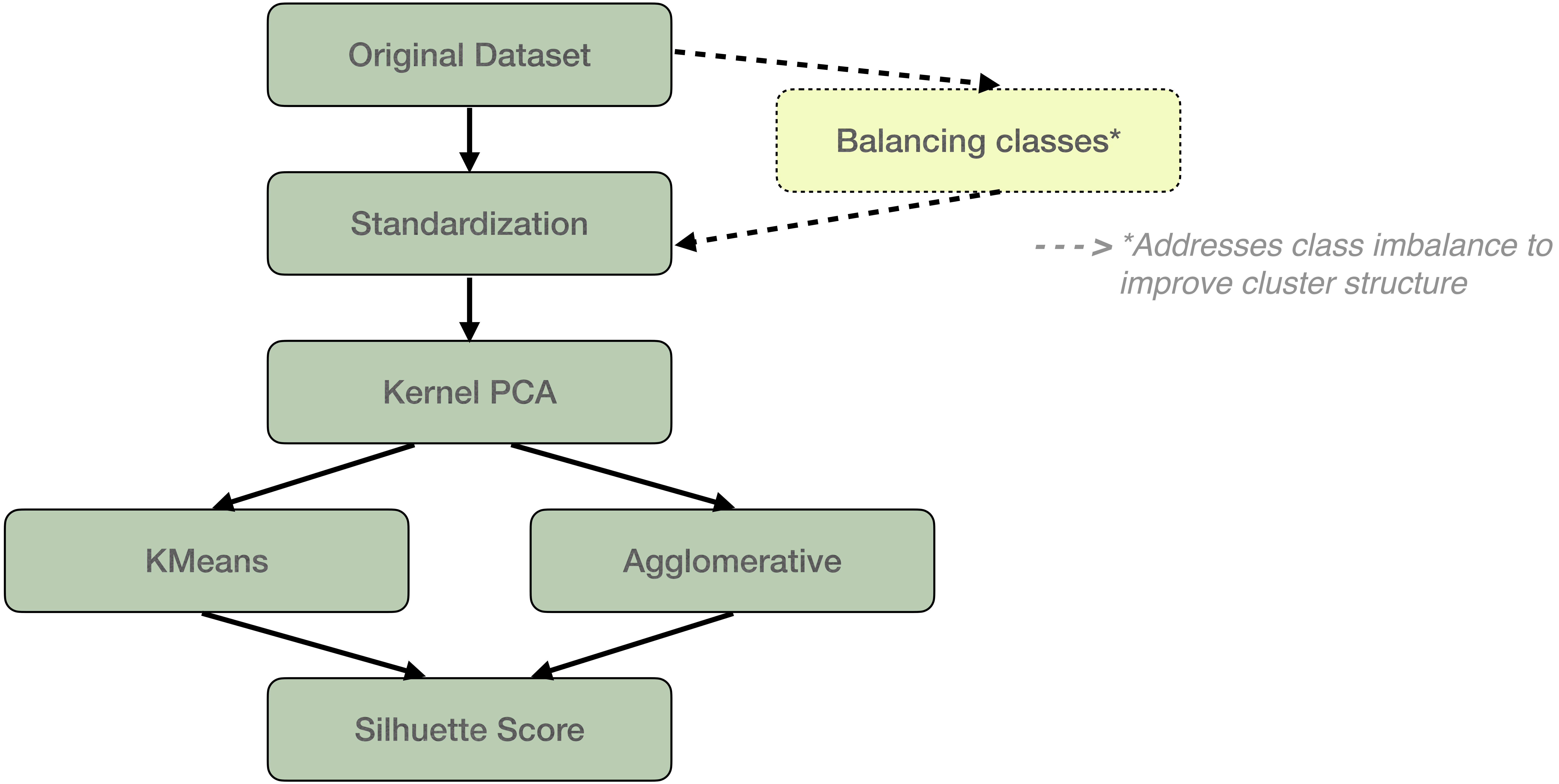
Exploratory Data Analysis

1. Wines with higher quality scores tend to have **higher alcohol levels**
2. **Lower volatile acidity** is associated with better wines
3. **Sulphates** slightly increase with quality
4. Lighter, drier wines tend to achieve better quality ratings.
5. Higher-quality wines have slightly **higher pH values** (less acidic), although the difference is subtle.
6. **Chlorides** and **free sulfur dioxide** show little visible separation across qualities

Most chemical attributes, aside from alcohol and volatile acidity, show **substantial overlap between classes**, suggesting **complex and subtle relationships** rather than sharp boundaries.



Machine Learning Models: Approach



Best Model Comparison

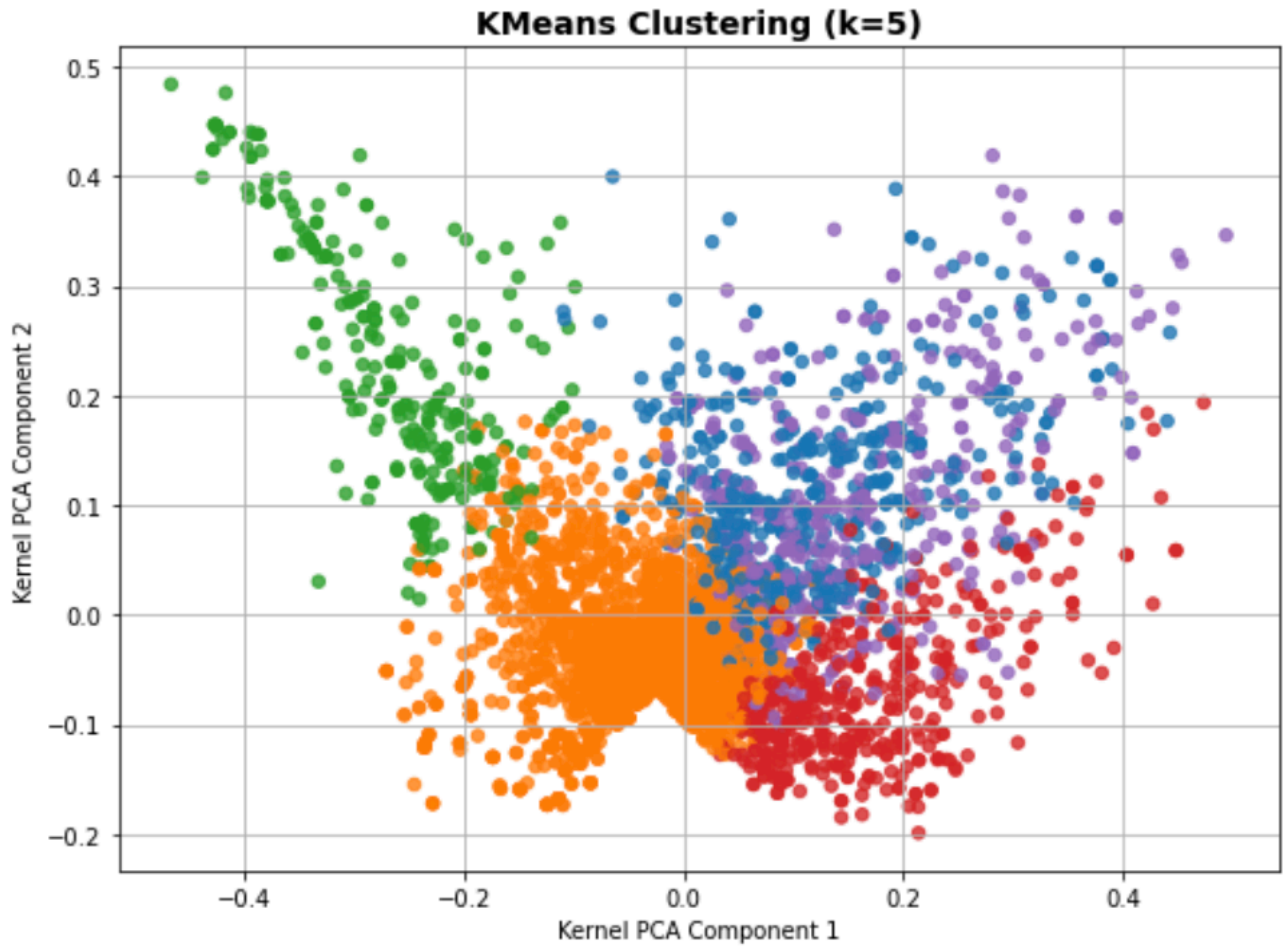
Higher silhouette does not always indicate meaningful clustering

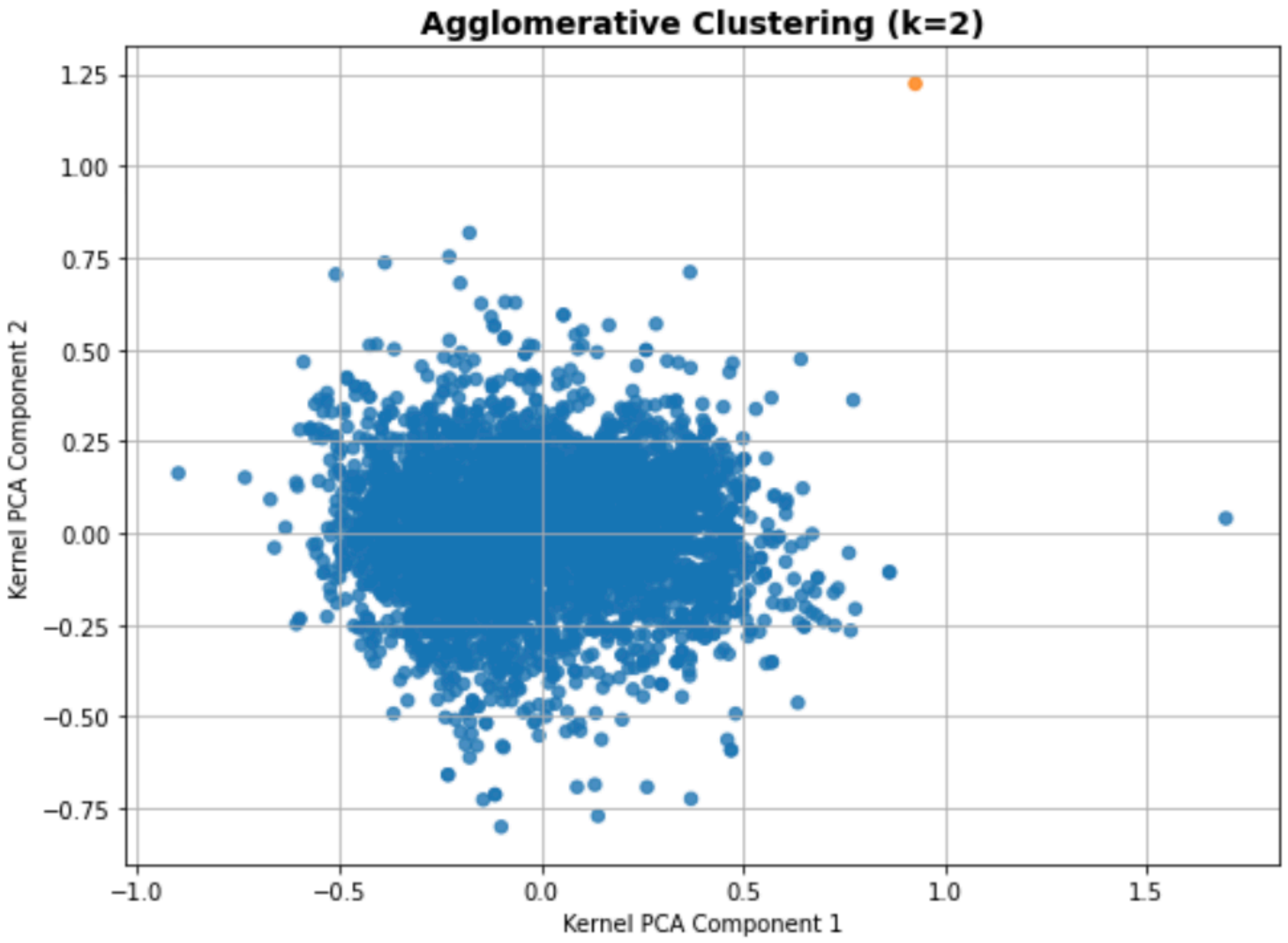
Model	Kernel	Gamma	Components	Clusters (k)	Silhouette Score
KMeans (original data)	RBF	0.5	4	5	0.457
KMeans (balanced)	RBF	0.5	4	4	0.559
Agglomerative (original)	Sigmoid	0.05	5	2	0.755
Agglomerative (balanced)	Sigmoid	0.05	5	2	0.767

- **KMeans after SMOTE** improved silhouette score (0.559)
- **Agglomerative after SMOTE** still failed - collapsed into a single cluster again, despite high silhouette (0.767)

Results

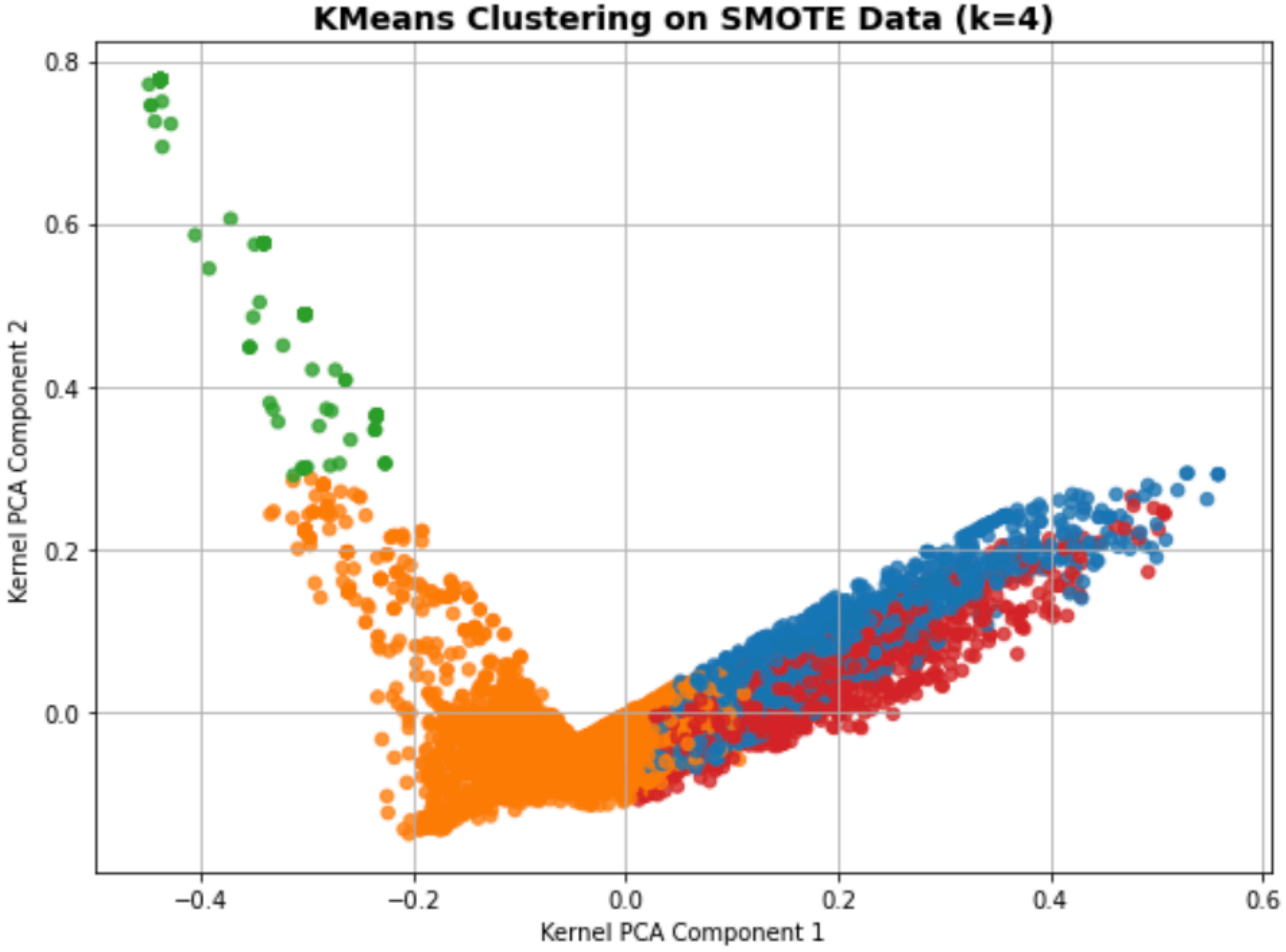
KMeans is **a better fit** for capturing subtle structure in the original data.

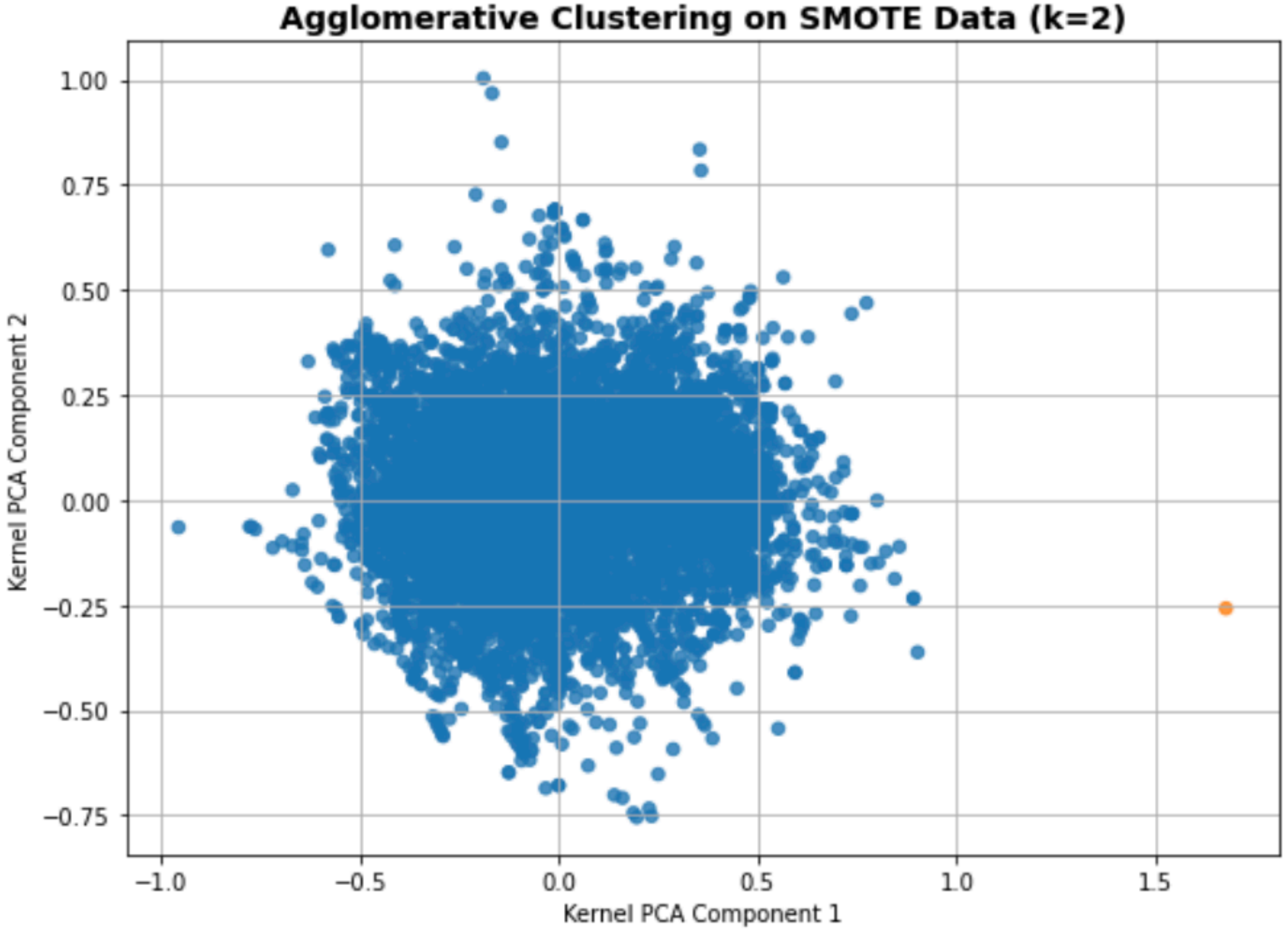




Results: After Balancing

KMeans is **a better fit** for capturing subtle structure in the balanced data.

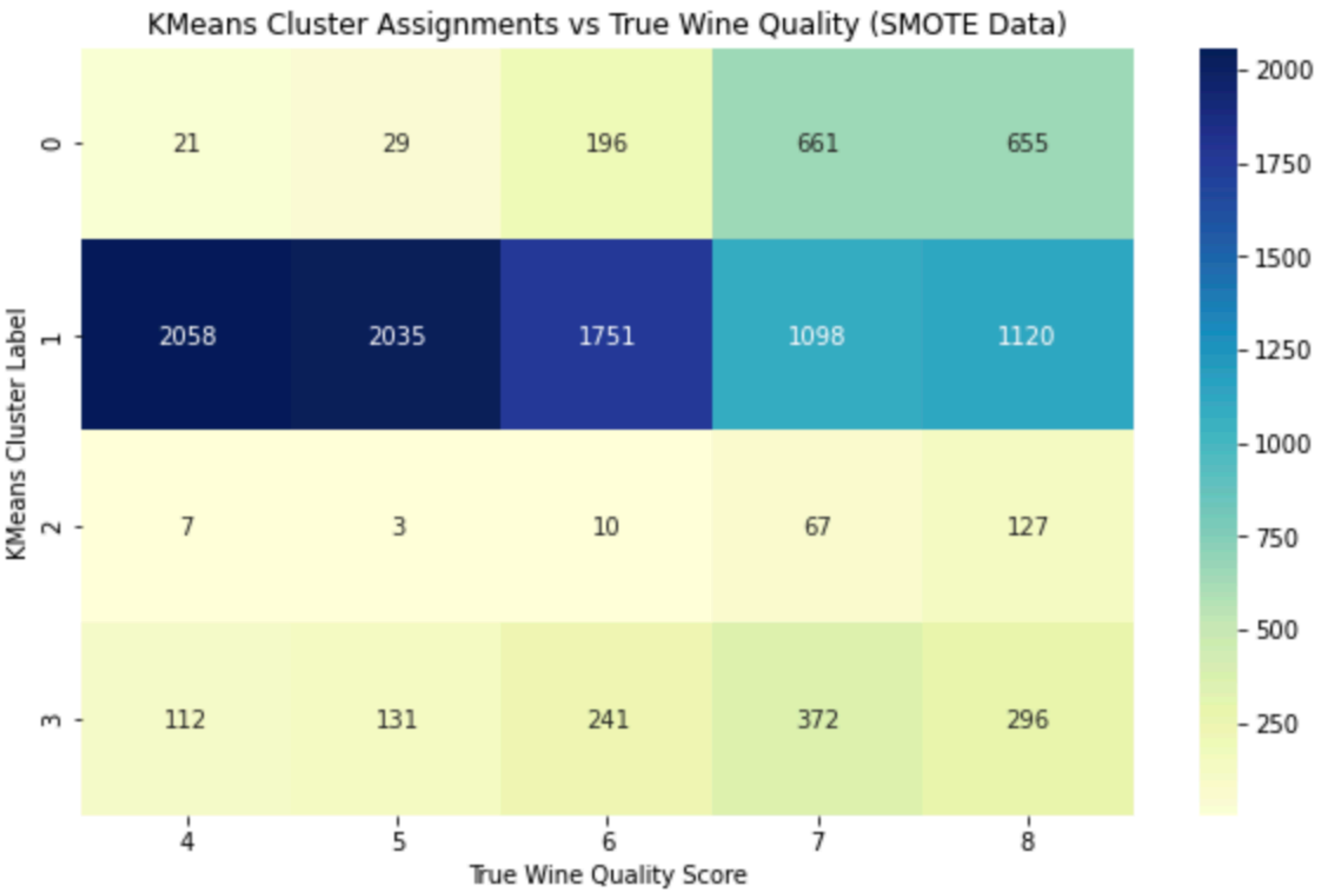




Wine Quality Explained (or not?)

Perfect separation between quality labels was **not achieved:**

- Wine quality **cannot be fully explained** by physicochemical data alone.
- Additional factors (e.g., grape variety, vintage year, fermentation practices) likely influence final quality



Conclusion and Future Work

Physicochemical data **alone is not sufficient** for fully predicting or explaining wine quality.

Unsupervised learning can **partially uncover structure** in white wine chemical profiles:

- **SMOTE balancing** improved clustering quality (higher silhouette score and better visual separation).
- **KMeans clustering** performed better
- **Agglomerative clustering** consistently collapsed into a single cluster

Future improvements:

- Incorporate **grape variety**, **vintage year**, and **production methods** into feature set.
- Explore **neural network-based clustering** approaches (e.g., deep embedded clustering)

Thank you!

Please reach out with any questions and suggestions:

LinkedIn: https://www.linkedin.com/in/mashalogan/
GitHub: https://github.com/mashuzza

Project files: https://github.com/mashuzza/python-projects/tree/main/unsupervised-learning-wine-quality-classification