Personality Clustering

Sara Akbarzadeh

July 2025

Abstract

This project investigates personality trait clustering using the Big Five dataset. After preprocessing and dimensionality reduction with PCA, several clustering algorithms were applied, including KMeans. Evaluation metrics such as Silhouette Score and Davies-Bouldin Index were used to assess the quality of clusters. Visualization and analysis of average trait scores per cluster revealed meaningful groupings based on personality patterns. The results demonstrate how unsupervised learning can uncover underlying behavioral structures in human personality data.

1 Introduction

Personality plays a fundamental role in shaping human behavior, influencing individual preferences, actions, and social interactions. In recent years, psychological models such as the Big Five Personality Traits — Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness — have gained significant attention due to their ability to quantitatively represent diverse personality dimensions.

With the growing availability of personality-related datasets, data-driven approaches offer promising insights into understanding and categorizing human behavior. Among these approaches, unsupervised learning methods such as clustering can uncover hidden patterns without relying on predefined labels.

This project explores the application of clustering techniques, particularly KMeans, to a dataset based on the Big Five personality model. The main goal is to discover natural groupings of individuals with similar personality profiles. Dimensionality reduction via Principal Component Analysis (PCA) is employed to visualize the clusters effectively and enhance interpretability. To assess the quality of clustering results, evaluation metrics like the Silhouette Score and Davies-Bouldin Index are used. The findings demonstrate how machine learning can contribute to the psychological analysis of personality data.

2 Dataset

The dataset used in this project was obtained from Kaggle and contains responses to a personality questionnaire based on the Big Five Personality Traits. The dataset comprises 1,015,341 rows and 110 columns, making it both large-scale and feature-rich.

Each row represents an individual's responses to a set of questions designed to measure five core personality dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. In addition to these primary traits, the dataset includes auxiliary features such as demographic information, timestamps, and response metadata.

Due to its volume and complexity, this dataset provides a valuable foundation for applying clustering techniques to uncover meaningful personality-based groupings. Before performing clustering, several preprocessing steps were conducted to ensure data quality and relevance, which are discussed in the following section.

3 Data Preprocessing

Before applying clustering algorithms, the dataset underwent several preprocessing steps to ensure its quality and relevance. This included both data cleaning and careful feature selection.

3.1 Data Cleaning

The original dataset consisted of over one million records and 110 columns. To maintain the integrity of the analysis, all rows containing missing values were removed. Additionally, 60 columns deemed irrelevant were excluded. This ensured that the data used for clustering focused solely on the behavioral dimensions measured by the questionnaire.

3.2 Feature Selection

The dataset included 50 questionnaire items designed to assess the Big Five personality traits: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience. Each of these traits was measured by 10 specific items. Only these 50 features were retained for the analysis. For improved interpretability, the original feature codes were replaced with their corresponding questionnaire statements, providing clearer context in subsequent analysis and visualization.

4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in any data-driven project as it helps to understand the underlying structure, detect anomalies, and identify patterns or relationships within the data. Through visualization and summary statistics, EDA provides insights that guide subsequent modeling choices and ensure the quality of the analysis.

The distribution of the first six personality trait features is shown in Figure 1. These histograms illustrate how participants responded to various trait-related questions, revealing the spread, central tendency, and skewness of the data for each feature.

Additionally, the correlation matrix heatmap presented in Figure 2 highlights the relationships between all personality trait features. This heatmap displays the degree to which different traits are positively or negatively correlated, offering valuable insights into their interdependencies.

5 Feature Scaling and Sampling

Before applying clustering algorithms, it is essential to scale the data, particularly when using distance-based methods such as K-Means and DBSCAN. To this end, the personality trait features were standardized so that each feature has a mean of zero and a standard deviation of one. This normalization ensures that all traits contribute equally during clustering and prevents features with larger numeric ranges from disproportionately influencing the distance calculations.

Due to computational constraints of the local machine's CPU, processing the entire dataset was not feasible. After consultation and approval from the teaching assistant team, a random sample comprising 7.5% of the standardized dataset was selected for the clustering analysis. This sampling maintained a representative subset of the data while significantly reducing computational load. To guarantee reproducibility, the sampling was performed using a fixed random seed.

6 Dimension Reduction

To visualize the high-dimensional personality trait data and aid clustering analysis, dimensionality reduction techniques were applied. Principal Component Analysis (PCA) was first used to reduce the data to two components that capture the maximum variance. The two-dimensional PCA projection, shown in Figure 3, reveals the overall spread and grouping tendencies within the dataset. The first two

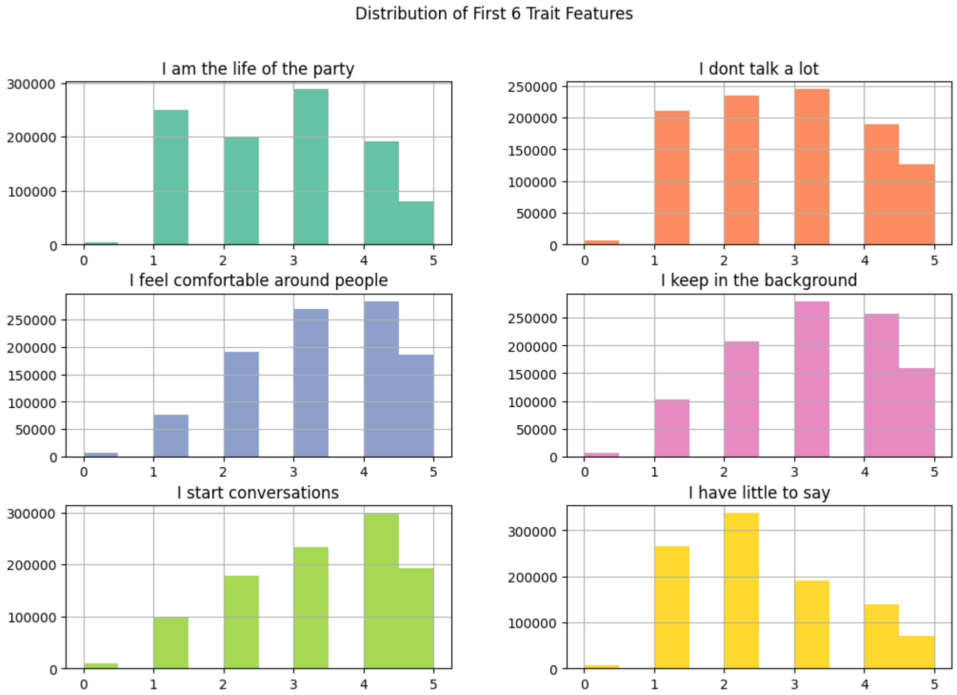

Figure 1: Distribution of the first six personality trait features.

principal components explain a substantial portion of the total variance, highlighting their effectiveness in summarizing the data.

In addition to PCA, t-Distributed Stochastic Neighbor Embedding (t-SNE) was employed to capture complex nonlinear relationships in the data. The t-SNE projection in Figure 4 offers a complementary visualization emphasizing local neighborhood structure and potential cluster separations that might not be evident with linear methods.

7 Optimal Number of Clusters

To determine the most appropriate number of clusters for the personality data, I evaluated the performance of K-Means clustering across a range of cluster counts from 2 to 7. Three internal validation metrics were used: the Silhouette Score, the Davies-Bouldin Index, and Inertia.

The Silhouette Score measures how well each data point fits within its assigned cluster, where higher values indicate better-defined clusters. The Davies-Bouldin Index assesses the average similarity between clusters, with lower values reflecting better separation. Inertia, which is central to the Elbow Method, evaluates how tightly data points are grouped around cluster centers.

As shown in Figure 5, the optimal number of clusters was identified as five. At this point, the Silhouette Score was relatively high, the Davies-Bouldin Index was low, and the Inertia plot exhibited a distinct "elbow," suggesting a balance between cluster compactness and separation. This informed the decision to proceed with five clusters in subsequent clustering algorithms.

8 Methods

This section presents the clustering algorithms applied to the reduced personality data, including their implementation details, evaluation metrics, and visualizations of the resulting clusters.

8.1 K-Means Clustering

K-Means clustering was applied with the number of clusters set to five, as determined by the optimal cluster analysis. Using the PCA-reduced data, the algorithm assigned each data point to one of five clusters. The model achieved a Silhouette Score of 0.3248 and a Davies-Bouldin Index of 0.9144, indicating moderately well-separated and compact clusters.

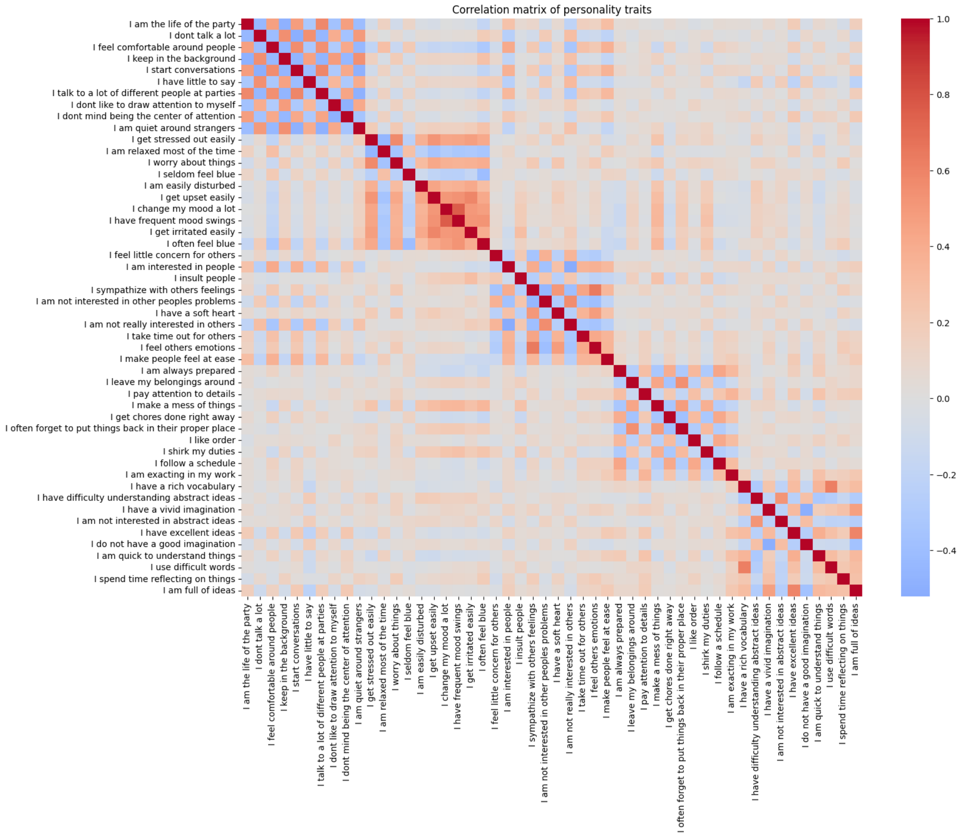

Figure 2: Correlation matrix heatmap of personality traits.

Figure 6 shows the scatter plot of the PCA components colored by K-Means cluster assignments. This visualization highlights the distinct grouping formed by the algorithm.

8.1.1 Interpretation

Based on the average standardized scores across the 50 personality items, each cluster exhibits a unique personality profile. For instance, Cluster 1 is characterized by high extraversion and openness, representing outgoing and imaginative individuals. In contrast, Clusters 3 and 4 correspond to more introverted and reserved profiles with low extraversion and openness. Cluster 2 shows high extraversion and conscientiousness, indicating sociable and organized individuals, while Cluster 0 represents a balanced personality without strong deviations.

8.2 Hierarchical Clustering

Due to memory limitations, hierarchical clustering was performed on a random subset of 8,000 samples drawn from the PCA-reduced data. The Ward linkage method was employed, and the dendrogram truncated to five levels is shown in Figure 7.

Five clusters were extracted from the dendrogram, achieving a Silhouette Score of 0.2629 and a Davies-Bouldin Index of 1.0717. Figure 8 depicts the cluster assignments projected onto the PCA components.

8.3 DBSCAN Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was also applied to the full PCA-reduced dataset using parameters $\epsilon = 0.5$ and minimum samples equal to 5. DBSCAN identified 2 clusters along with 64 noise points (outliers).

The algorithm produced a Silhouette Score of 0.4817 but a higher Davies-Bouldin Index of 2.0357, indicating that while clusters were cohesive, cluster separation was less optimal compared to K-Means and Hierarchical methods.

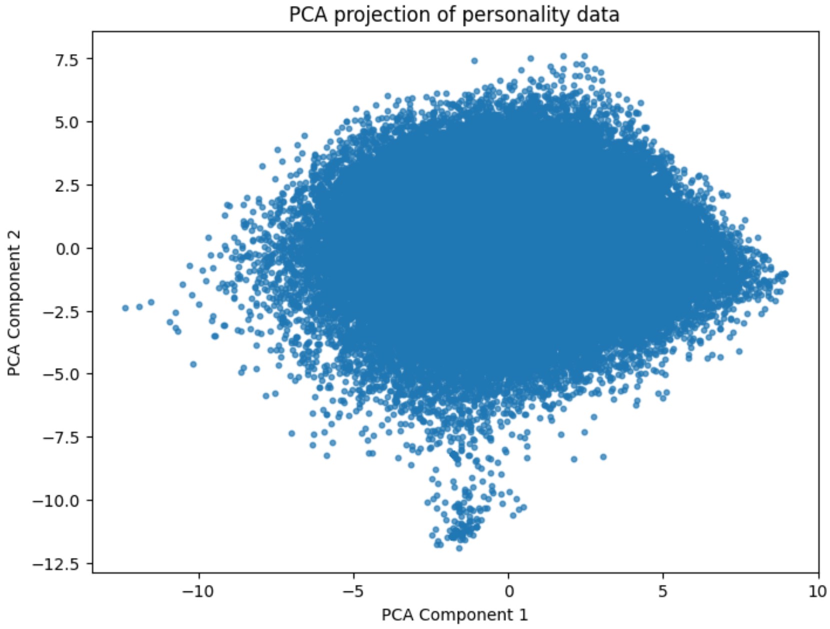

Figure 3: PCA projection of the personality data onto two principal components.

Figure 9 displays the DBSCAN clustering results. The presence of noise points reflects the algorithm's ability to detect outliers in the data.

8.4 Results Comparison

The performance of the three clustering algorithms—K-Means, DBSCAN, and Hierarchical clustering—is compared using two internal validation metrics: the Silhouette Score and the Davies-Bouldin Index.

Figure 10 illustrates the Silhouette Scores for each algorithm. DBSCAN achieved the highest Silhouette Score of 0.4817, indicating better cohesion within clusters compared to K-Means (0.3248) and Hierarchical clustering (0.2629).

Figure 11 shows the Davies-Bouldin Index for each method. K-Means scored the lowest index of 0.9144, suggesting better cluster separation, whereas DBSCAN had the highest index of 2.0357, indicating less optimal partitioning.

These results highlight a trade-off: DBSCAN clusters are more cohesive but less separated, while K-Means offers a better balance of compactness and distinctness. Hierarchical clustering performs moderately on both metrics, constrained by sampling limitations.

9 Conclusion

In this study, I applied three unsupervised clustering algorithms—K-Means, Hierarchical clustering, and DBSCAN—to personality data derived from the Big Five Personality Test. After preprocessing, scaling, and dimensionality reduction, I evaluated the models using Silhouette Scores and Davies-Bouldin Indices.

K-Means provided the most balanced clustering with well-separated and compact groups, while DBSCAN identified fewer but highly cohesive clusters and effectively detected noise points. Hierarchical clustering, performed on a sampled subset due to computational constraints, showed moderate performance.

Overall, the analysis demonstrated that the personality data naturally groups into five distinct clusters, each representing different personality profiles. These findings contribute to a better understanding of human personality traits and illustrate the effectiveness of clustering methods for psychological data analysis. Future work could explore larger datasets and alternative algorithms to further validate these results.

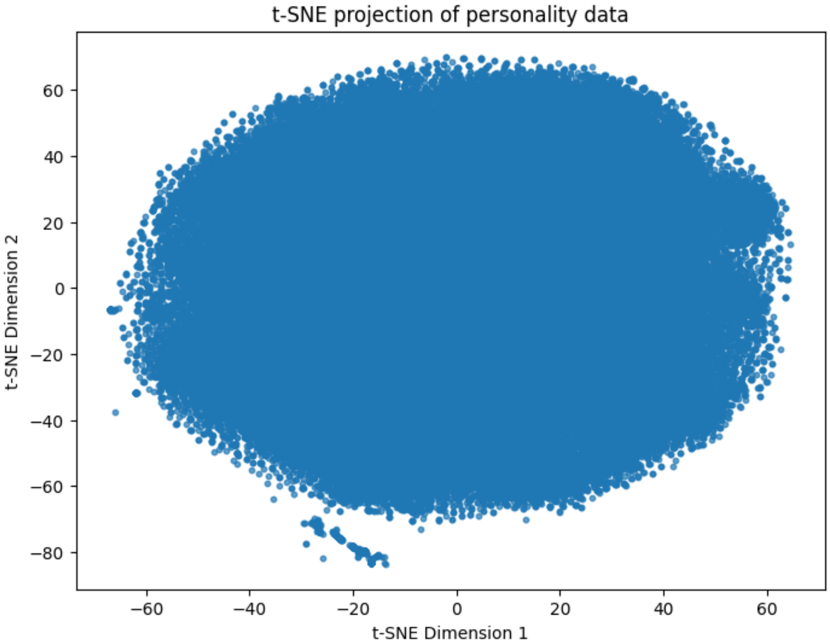

Figure 4: t-SNE projection of the personality data in two dimensions.

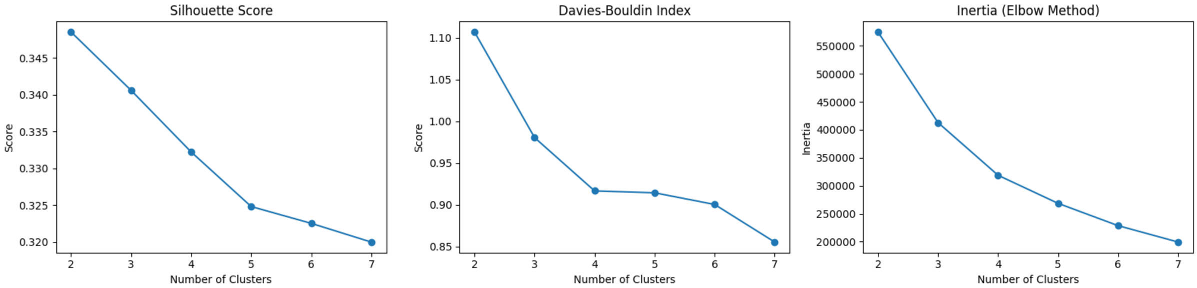

Figure 5: Evaluation metrics for different numbers of clusters: Silhouette Score, Davies-Bouldin Index, and Inertia (Elbow Method).

10 Future Work

Future research could expand on this study by incorporating larger and more diverse datasets to improve the generalizability of the clustering results. Exploring additional clustering algorithms, such as Gaussian Mixture Models or spectral clustering, may reveal different structures within the personality data. Integrating demographic or behavioral variables alongside personality traits could provide deeper insights into cluster characteristics. Furthermore, applying advanced dimensionality reduction techniques and optimizing hyperparameters using automated methods could enhance clustering performance. Finally, validating the clusters with external psychological assessments or longitudinal studies would strengthen the practical relevance of the findings.

References

[1] Tunguz, Big Five Personality Test Dataset, Kaggle, 2019.
https://www.kaggle.com/datasets/tunguz/big-five-personality-test

[2] MacQueen, J., “Some Methods for Classification and Analysis of Multivariate Observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

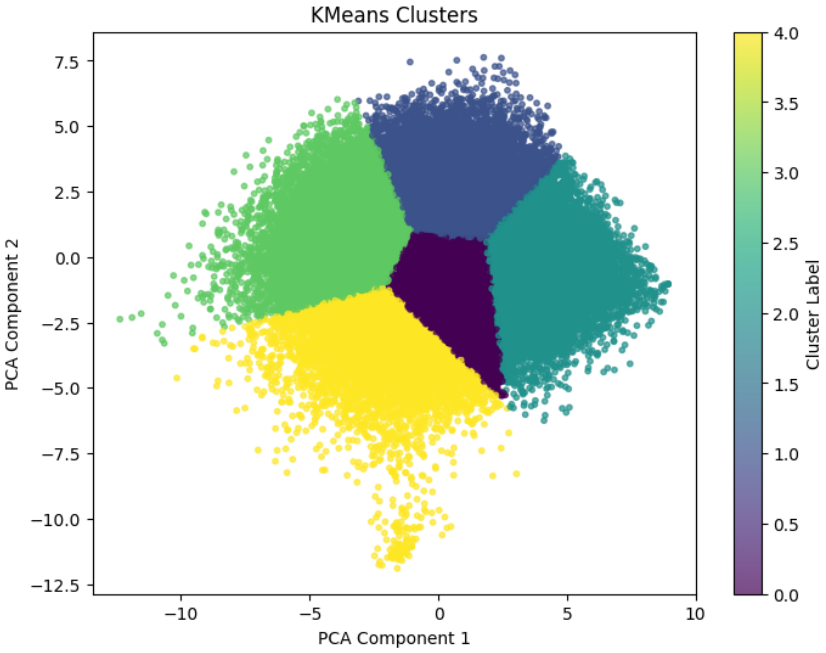

Figure 6: K-Means clustering results on PCA-reduced personality data.

[3] van der Maaten, L. and Hinton, G., "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. Nov, 2008, pp. 2579–2605.

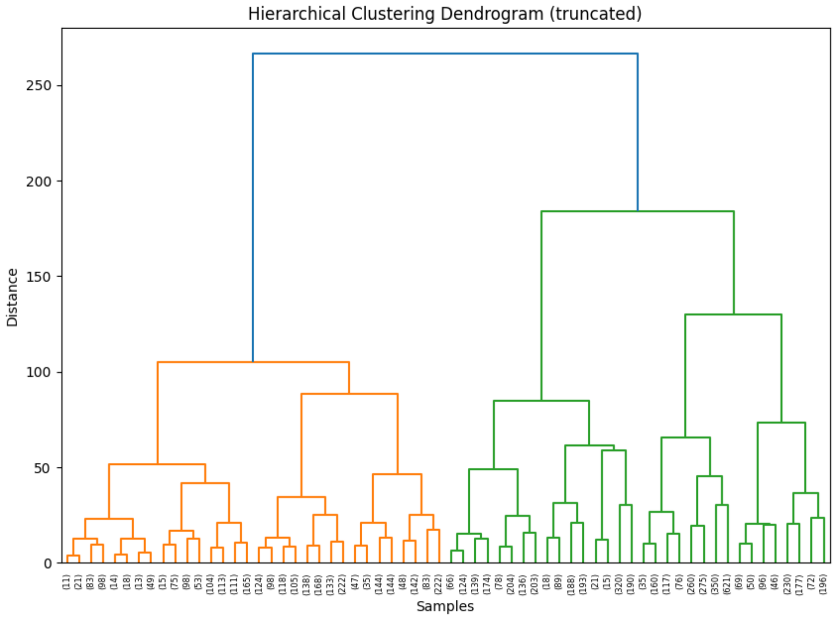

Figure 7: Truncated dendrogram of hierarchical clustering using Ward linkage on a sampled subset.

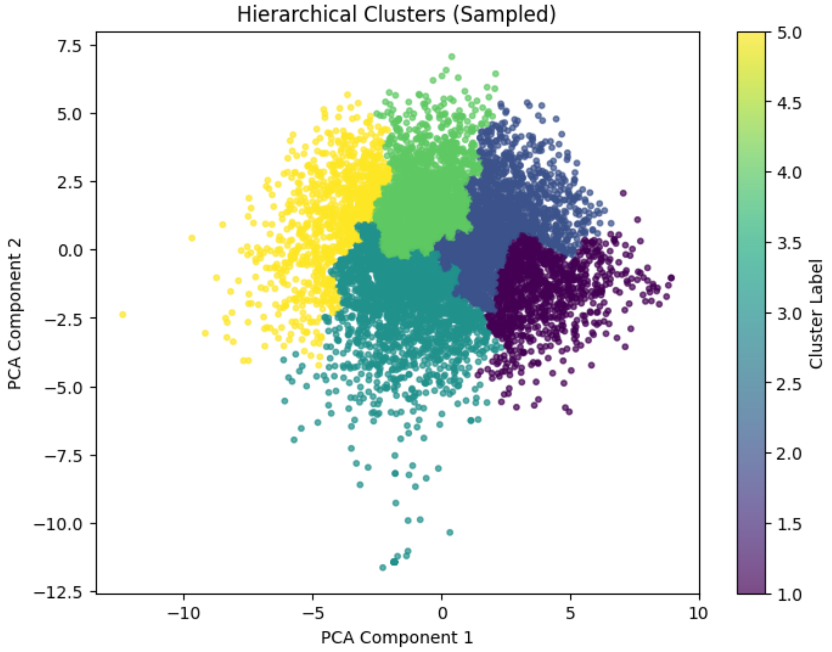

Figure 8: Hierarchical clustering results on the sampled PCA-reduced data.

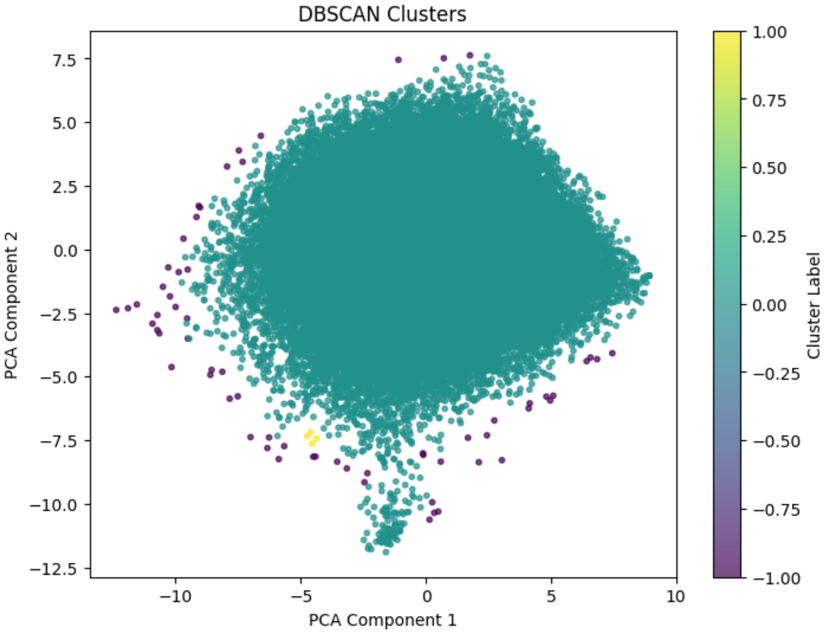

Figure 9: DBSCAN clustering results on PCA-reduced personality data, showing clusters and noise points.

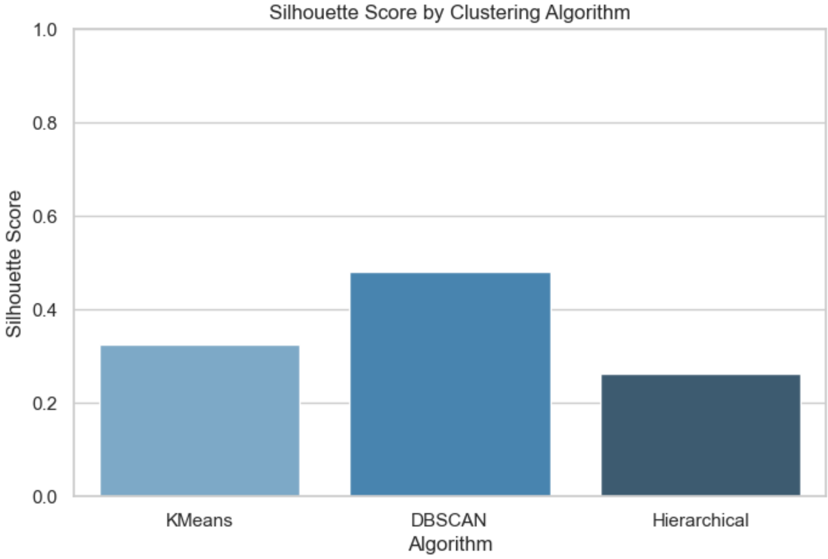

Figure 10: Silhouette Score by clustering algorithm. Higher values indicate better-defined clusters.

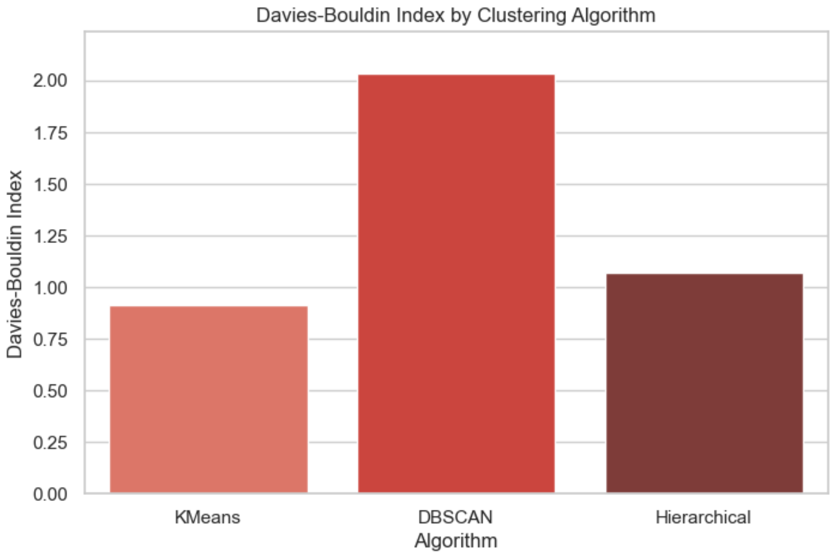

Figure 11: Davies-Bouldin Index by clustering algorithm. Lower values indicate better cluster separation.