This section provides a comparative analysis of HDBSCAN and Agglomerative Hierarchical Clustering (AHC) as applied to the 'Bottleneck Duration Seconds' time series data, using Dynamic Time Warping (DTW) distance.

**1. Key Differences in Approach**

|  |  |  |
| --- | --- | --- |
| **Feature** | **HDBSCAN** | **Agglomerative Hierarchical Clustering (AHC)** |
| **Algorithm Type** | Density-based, hierarchical but extracts flat clusters from hierarchy. | Connectivity-based, builds a hierarchy from individual data points. |
| **Noise Handling** | **Explicitly identifies noise points** (-1 label). | **Does not explicitly handle noise**; all points are assigned to a cluster. |
| **Cluster Shape** | Can find arbitrary shapes and varying densities. | Tends to find convex, spherical clusters (depending on linkage). |
| **Number of Clusters** | Automatically determines the optimal number of clusters based on density. | Requires a 'cut' (either fixed number of clusters or distance threshold) to determine flat clusters. |
| **Parameter Tuning** | min\_cluster\_size, min\_samples (often less sensitive). | n\_clusters (or cut threshold), linkage\_method (more sensitive to choice). |
| **Metric Input** | Can directly use a precomputed distance matrix. | Can use a precomputed distance matrix (except 'ward' linkage). |

In this analysis, both algorithms were fed with the same DTW distance matrix (or derived from the same time series sequences for AHC).

**2. Comparison of Performance Metrics**

Using the compare\_clustering\_metrics\_summary table (image\_723b39.png):

|  |  |  |
| --- | --- | --- |
| **Metric** | **HDBSCAN (DTW)** | **AHC (DTW)** |
| N\_Clusters | 6.0 | 3.0 |
| Noise Points | 12.0 | 0.0 |
| Noise Ratio | 0.28 | 0.00 |
| Silhouette Score | **0.5592** | 0.3807 |
| Calinski Harabasz | **45.4199** | 15.6888 |
| Davies Bouldin | **0.8159** | 1.1555 |
| ARI / AMI | NaN | NaN |

**Key Observations from Metrics:**

* **Number of Clusters:** HDBSCAN found 6 clusters, while AHC found 3. This indicates HDBSCAN's ability to discover a more granular structure in the data by identifying more distinct density regions.
* **Noise Handling:** HDBSCAN identified 12 noise points (28% of the data), whereas AHC, by design, assigned all points to a cluster, resulting in 0 noise.
* **Internal Validation Scores:**
  + **Silhouette Score:** HDBSCAN's score (0.5592) is significantly higher than AHC's (0.3807). This suggests that the clusters formed by HDBSCAN, when considering only the non-noise points, are much better defined, more compact, and better separated from each other in the DTW distance space.
  + **Calinski Harabasz Score:** HDBSCAN's score (45.4199) is substantially higher than AHC's (15.6888), reinforcing that HDBSCAN's clusters exhibit greater between-cluster dispersion relative to within-cluster dispersion.
  + **Davies Bouldin Score:** HDBSCAN's score (0.8159) is lower than AHC's (1.1555). A lower Davies-Bouldin score is desirable, indicating better separation between clusters and better compactness within clusters.

Based on these internal validation metrics, **HDBSCAN clearly outperforms AHC** in terms of finding a more well-structured and separated clustering solution for this time series dataset when using DTW distances.

**3. Comparison of Visual Clustering Results (2D Plots)**

Referring to image\_723a7f.png (specifically the middle plot for HDBSCAN and the right plot for AHC):

* **HDBSCAN Plot:** Shows 6 distinct colored clusters and numerous grey 'x' markers representing noise points. The clusters appear relatively tight and well-separated visually in the PCA-reduced 2D space. The presence of noise allows the algorithm to focus on forming high-density clusters, leaving ambiguous points unassigned.
* **AHC Plot:** Displays 3 distinct colored clusters, with no noise points. The clusters, while identifiable, appear somewhat less distinct or more broadly dispersed compared to HDBSCAN's clusters, and some overlap might be inferred. Since all points are forced into a cluster, potentially anomalous 'Stoppage Reason' time series are grouped with others, which might dilute cluster purity.

The visual representation reinforces the metric comparison: HDBSCAN, by intelligently handling noise and finding varying-density clusters, provides a more granular and potentially more meaningful partitioning of the time series data.

**4. Strengths and Weaknesses in this Context**

**HDBSCAN (with DTW Precomputed Metric):**

* **Strengths:**
  + **Automatic Cluster Number:** It intelligently determines the optimal number of clusters, eliminating the need for manual n\_clusters selection (as in AHC).
  + **Noise Robustness:** Its ability to identify and exclude noise points is highly beneficial for real-world time series data, which often contains outliers or anomalous patterns. This leads to cleaner and more representative clusters.
  + **Density-Based:** Can discover clusters of arbitrary shape and varying densities, which is well-suited for complex time series patterns.
  + **Strong Performance:** Demonstrated superior internal validation scores, indicating better-defined and more separated clusters.
* **Weaknesses:**
  + Interpretation of min\_cluster\_size and min\_samples can sometimes be less intuitive than n\_clusters for non-experts.
  + Results depend on the quality of the precomputed distance matrix (DTW in this case).

**Agglomerative Hierarchical Clustering (AHC with DTW Distance):**

* **Strengths:**
  + **Dendrogram Visualization:** Provides a clear hierarchical structure (dendrogram), which can be very insightful for understanding relationships between clusters at different levels of granularity.
  + **Guaranteed Cluster Assignment:** Every data point is assigned to a cluster, which might be desirable if all data must be categorized.
  + **Interpretable Parameters:** The n\_clusters parameter is straightforward to understand.
* **Weaknesses:**
  + **Requires n\_clusters:** The optimal number of clusters has to be chosen (e.g., via silhouette optimization), which can be subjective and may not always capture natural groupings as effectively as HDBSCAN.
  + **No Noise Handling:** Forces outliers into existing clusters, potentially distorting cluster shapes or characteristics. The low silhouette score and high Davies-Bouldin score suggest this might be happening.
  + **Less Optimal Performance:** Demonstrated lower internal validation scores, suggesting less distinct or more overlapping clusters compared to HDBSCAN for this dataset.
  + 'Ward' linkage method, often preferred for its balanced cluster sizes, is not compatible with precomputed DTW distances.

**5. Conclusion for Time Series (DTW) Clustering**

For the specific task of clustering 'Stoppage Reason' time series data using DTW distances, **HDBSCAN appears to be the more effective algorithm**. Its density-based nature, ability to automatically determine cluster count, and robust noise handling (which identified 28% of the data as noise) resulted in a higher-quality clustering solution as evidenced by significantly better Silhouette, Calinski-Harabasz, and Davies-Bouldin scores. The visual plots further support this, showing clearer cluster separation and the useful identification of anomalous time series as noise.

While AHC provides an insightful dendrogram, its forcing of all points into clusters and its reliance on a pre-selected n\_clusters led to a less optimal partition based on the chosen evaluation metrics. HDBSCAN's approach aligns better with the characteristics of time series data where some patterns might be truly unique and not fit well into a larger cluster.