Table of Contents

[GitHub Repository Link 2](#_Toc207800870)

[Executive Summary 2](#_Toc207800871)

[Project Background 3](#_Toc207800872)

[Data Infrastructure & Tools 3](#_Toc207800873)

[Tooling Choices and Purpose 3](#_Toc207800874)

[Methodological Transition 4](#_Toc207800875)

[Data Analytics 5](#_Toc207800876)

[Bias & Privacy Considerations 6](#_Toc207800877)

[Limitations of Clustering Methods 6](#_Toc207800878)

[Data Engineering 7](#_Toc207800879)

[Data Preprocessing 7](#_Toc207800880)

[Visual Diagnostics 7](#_Toc207800881)

[Data Visualisation & Dashboards 8](#_Toc207800882)

[Methodology 10](#_Toc207800883)

[Impact Evaluation 11](#_Toc207800884)

[Recommendations 11](#_Toc207800885)

[Conclusion 12](#_Toc207800886)

[References 12](#_Toc207800887)

[Appendices 13](#_Toc207800888)

[Figure 1: Folder Structure 3](#_Toc207800794)

[Figure 2: Tools Used 3](#_Toc207800795)

[Figure 3: PCA Scatter plot: 4](#_Toc207800796)

[Figure 4: Cluster Interpretation 5](#_Toc207800797)

[Figure 5: Cluster Interpretation Summary 5](#_Toc207800798)

[Figure 6: Easy Read Cluster Summary 5](#_Toc207800799)

[Figure 7: Missing Value Heatmap 7](#_Toc207800800)

[Figure 8: RC1\_Adopted 7](#_Toc207800801)

[Figure 9: RC2\_Died 7](#_Toc207800802)

[Figure 10: RC5\_Transfer\_to\_another\_LA 8](#_Toc207800803)

[Figure 11: Correlation Heatmap 8](#_Toc207800804)

[Figure 12: Dendrogram 8](#_Toc207800805)

[Figure 13: Clusters in reduced dimensions 9](#_Toc207800806)

[Figure 14: Distribution of closure patterns 9](#_Toc207800807)

[Figure 15: Pairplot Closures Features 9](#_Toc207800808)

[Appendix 1: Public Dataset 13](#_Toc207800828)

[Appendix 2: Closure Reason Descriptions 14](#_Toc207800829)

[Appendix 3: Cluster Counts by Region 14](#_Toc207800830)

[Appendix 4: Dendrogram at Threshold 100 15](#_Toc207800831)

[Appendix 5: Data Guidance from CiN Census 15](#_Toc207800832)

GitHub Repository Link https://github.com/IfyQ/DS\_Hierarchical\_Clustering.git

# Executive Summary

This project aimed to explore regional patterns in case closure outcomes across local authorities in England using hierarchical clustering. The analysis uses publicly available data from the Children in Need Census (2013–2024), focusing on closure reasons (RC1–RC9). Hierarchical clustering was selected for its interpretability and suitability for small datasets where the number of clusters is not known in advance (Gupta et al., 2023). Eleven clusters were identified using Ward linkage and PCA for dimensionality reduction. Key findings revealed:

**Key Finding:**

**Cluster 1**: High early help referrals (RC8/RC9), prevalent in Outer and Inner London.

**Cluster 2**: High adoption and SGO rates (RC1/RC4), dominant in Northeast, Northwest, and Yorkshire.

**Cluster 3**: Older children’s transitions (RC3/RC6), common in Inner London and Southwest.

**Cluster 4**: Mixed closures, large authorities with diverse case types.

**Clusters 5–11:** Sparse or region-specific patterns, including outliers and authorities with low closure volumes.

These insights support benchmarking, policy evaluation, and targeted support for local authorities. Future iterations may incorporate referral sources, assessment factors, and child characteristics to deepen understanding of service pathways.

# Project Background

The Children in Need Census (2013–2024) records closure reasons for social care cases. This dataset was chosen for its relevance to my professional domain—children and adult services—and its structured, longitudinal nature.

My local authority was part of **Cluster 2**, marked by high adoption and SGO rates. These insights are valuable for benchmarking and informing early help strategies. The analysis complements work under the **Supporting Families programme**, where understanding closure outcomes supports planning, resource allocation, and performance monitoring. Linking national patterns to local practice shows how public data can support operational goals and improve outcomes.

# Data Infrastructure & Tools

The project used a structured folder system with directories for raw data, notebooks, outputs, visuals, and documentation. GitHub enabled reproducibility and transparent tracking of changes.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Folder Structure

## Tooling Choices and Purpose

The project began with KMeans clustering due to its simplicity but transitioned to hierarchical clustering due to limitations like requiring a predefined number of clusters and unstable results. Hierarchical clustering allowed dendrogram-based exploration and handled non-convex shapes and varying densities. SciPy’s clustering integrated with the pipeline, and PCA from Scikit-learn enabled visualisation.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: Tools Used

## Methodological Transition

The project initially used KMeans clustering due to its simplicity and scalability. However, several limitations emerged during implementation including requiring predefined number of clusters, which was difficult to justify with this dataset.

To address these issues, the methodology was transitioned to hierarchical clustering, which does not require specifying the number of clusters in advance, allowing flexible exploration via dendrograms.

See ***transition\_note.md*** for a detailed comparison of KMeans and hierarchical clustering and was supported by the existing tooling ecosystem. SciPy’s hierarchical clustering functions integrated seamlessly with the data pipeline, and PCA from Scikit-learn enabled effective visualisation of clusters in reduced dimensions.

# Data Analytics

Hierarchical clustering was chosen for its flexibility. Ward linkage minimised variance within clusters (Malamine, 2023). A distance threshold of 20 was used to cut the dendrogram, resulting in 11 clusters. PCA reduced dimensionality for visualisation.

A screen shot of a graph

AI-generated content may be incorrect.

Figure 3: PCA Scatter plot:

Cluster interpretation followed the Children in Need Census 2025–2026 guidance, focusing on closure reasons (RC1–RC9). Key patterns were visualised and summarised to support interpretation.A close-up of a chart

AI-generated content may be incorrect.

Figure 4: Cluster Interpretation

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: Cluster Interpretation Summary

A screenshot of a computer

AI-generated content may be incorrect.

Figure 6: Easy Read Cluster Summary

## Bias & Privacy Considerations

This analysis used publicly available data published under the Open Government Licence. No data was processed, ensuring GDPR compliance.

Bias risks:

* **Feature bias**: Closure reasons may reflect systemic or regional reporting differences.
* **Interpretation bias**: Cluster labels reflect statistical patterns, not causality.

## Limitations of Clustering Methods

KMeans required predefined clusters and assumed spherical shapes, producing unstable and unclear results. Hierarchical clustering resolved these issues by:

* Allowing flexible cluster selection via dendrograms.
* Handling varied shapes and densities.
* Providing clearer visuals.
* Being more robust to noise and missing data.

This transition improved analytical integrity and insight quality.

# Data Engineering

The raw dataset contained closure reason counts (RC1–RC7) for each local authority from 2013 to 2024. Key steps included:

## Data Preprocessing

* **Cleaning**: Replaced non-numeric entries ('c', 'z', 'x', 'u', 'k') with NaN.
* **Imputation**: Used median imputation to preserve distribution.
* **Merging**: Linked LA names and regions using new\_la\_code.
* **Scaling**: Applied StandardScaler before PCA and clustering.

## Visual Diagnostics

1. **Missing Value Heatmap -** Showed gaps across features, supporting median imputation. A purple and yellow color line

   AI-generated content may be incorrect.

Figure 7: Missing Value Heatmap

**2. Boxplots:** Revealed skewed distributions, especially for RC2 (Died) and RC3 (Child Arrangements Order). RC4 and RC5 showed useful variation.

A graph of a distribution of rc1

AI-generated content may be incorrect.

Figure 8: RC1\_Adopted

A graph of a number of objects

AI-generated content may be incorrect.

Figure 9: RC2\_Died

A graph of a distribution of rcs

AI-generated content may be incorrect.

Figure 10: RC5\_Transfer\_to\_another\_LA

**3. Correlation Heatmap:** Highlighted relationships between closure reasons. Strong correlations (e.g., RC1 and RC4) suggested similar outcomes; weak ones (e.g., RC6 vs RC1) indicated distinct pathways.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

Figure 11: Correlation Heatmap

# Data Visualisation & Dashboards

Visuals were created using Matplotlib and Seaborn, focusing on clarity and interpretability:

* **Dendrogram**: Showed hierarchical relationships.

A graph of a diagram

AI-generated content may be incorrect.

Figure 12: Dendrogram

* **PCA Scatter Plot**: Displayed clusters in reduced dimensions.

A diagram of a scatter plot of clusters

AI-generated content may be incorrect.

Figure 13: Clusters in reduced dimensions

* **Cluster Counts by Region**: Highlighted regional closure patterns.

A graph with different colored bars

AI-generated content may be incorrect.

Figure 14: Distribution of closure patterns

* **Pairplot**: Explored feature relationships.

A chart of data on a white background

AI-generated content may be incorrect.

Figure 15: Pairplot Closures Features

# Methodology

Hierarchical clustering grouped local authorities by closure reasons from the Children in Need Census. Key steps:

**Data Preparation**

Used the CiN\_Closure\_Reason\_2013\_to\_2024.csv dataset. Created:

* **Reference table**: LA names, codes, regions.
* **Clustering table**: Closure reason features (RC1–RC7), aggregated.

**Cleaning Steps**:

* Standardised column names.
* Replaced non-numeric entries with NaN.
* Imputed missing values using median.
* Converted features to numeric and scaled.

**Feature Scaling and Dimensionality Reduction**  
StandardScaler standardised feature ranges.

PCA reduced data to two components for visualisation while retaining variance.

**Hierarchical Clustering**  
Ward’s method computed linkage distances. A threshold of 20 cut the dendrogram and assigned cluster labels. This approach avoided KMeans instability and improved visual separation.

**Cluster Visualisation and Interpretation**

* **Dendrogram**: Illustrated relationships.
* **PCA Scatter Plot**: Showed cluster distribution.
* **Bar Chart**: Summarised cluster counts by region.

These visuals supported interpretation of regional patterns and closure profiles.

**Outputs**

* hierarchical\_cluster\_scatter.png: PCA scatter plot with cluster labels
* dendrogram.png: Hierarchical tree diagram
* cluster\_region\_bar.png: Cluster distribution by region
* cluster\_region\_summary.csv: Tabular summary of cluster-region relationships

This methodology supports evidence-based grouping aligned with CIN guidance.

# Impact Evaluation

The analysis provides actionable insights for children’s social care:

* **Benchmarking**: Authorities can compare closure profiles with similar peers.
* **Identifying outliers and best practices**: Unusual patterns may highlight innovation or areas needing review.
* **Informing resource allocation**: Understanding dominant pathways (e.g., adoption, SGO, transitions) helps tailor services.

My local authority, part of Cluster 2, aligns with our focus on permanency outcomes and supports our work under the **Supporting Families programme**.

# Recommendations

To enhance the value of this analysis:

* Extending the feature set to include referral sources, assessment factors, and child characteristics.
* Link clusters to outcomes such as re-referrals, Child Protection Plans (CPPs), or successful early help exits.
* Explore time trends and model drift.
* Test ensemble clustering for stability and robustness.

These steps will deepen insights and strengthen planning.

# Conclusion

Hierarchical clustering revealed meaningful patterns in case closure outcomes. Its flexibility and visual clarity made it suitable for this dataset, supporting evidence-based planning and service improvement.

# References

* DataCamp (2025) *Hierarchical clustering: concept overview with examples*. Available at: https://www.datacamp.com/tutorial/hierarchical-clustering (Accessed: 2 September 2025).
* Department for Education (2024) *Characteristics of children in need in England*. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/characteristics-of-children-in-need (Accessed: 2 September 2025).
* Department for Education (2024) Children in Need Census 2024 to 2025: Guide. Available at: https://www.gov.uk/government/publications/children-in-need-census-2024-to-2025-guide (Accessed: 2 September 2025).
* Gupta, A., Sharma, H. and Akhtar, A. (2023) ‘A comparative analysis of K-means and hierarchical clustering’, EPRA International Journal of Multidisciplinary Research. Available at: <https://eprajournals.com/IJMR/article/5796/download>(Accessed: 2 September 2025).
* Larose, D.T. and Larose, C.D. (2014) *Discovering knowledge in data: An introduction to data mining*. 2nd edn. Wiley. Available at: https://ieeexplore.ieee.org/book/10066951 (Accessed: 2 September 2025).
* **Microsoft** (2025) *Copilot (GPT-4) [Large Language Model]*. Available at: https://copilot.microsoft.com (Accessed: 2 September 2025).
* Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research*, 12 (Oct), pp. 2825–2830.
* The National Archives (2015) *Open Government Licence v3.0*. Available at: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (Accessed: 2 September 2025).
* Virtanen, P. et al. (2020) ‘SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python’, Nature Methods, 17(3), pp. 261–272.

# Appendices

A screen shot of a grid

AI-generated content may be incorrect.

Appendix 1: Public Dataset

A screenshot of a computer

AI-generated content may be incorrect.

Appendix 2: Closure Reason Descriptions

A graph of different colored bars

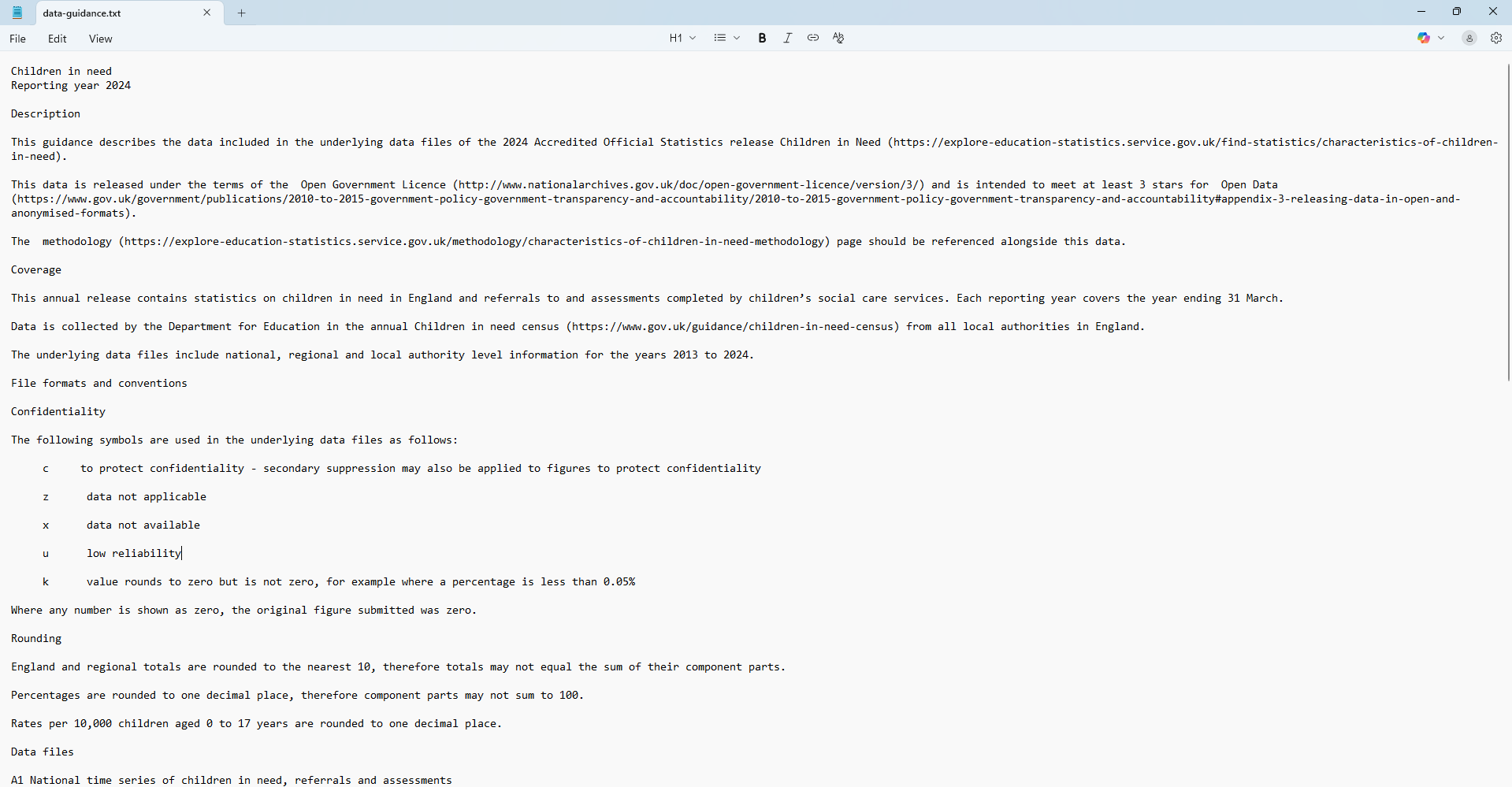
AI-generated content may be incorrect.

Appendix 3: Cluster Counts by Region

A diagram of a city

AI-generated content may be incorrect.

Appendix 4: Dendrogram at Threshold 100



Appendix 5: Data Guidance from CiN Census