Regional Equity and Clustering - Cost Analys Final Report

Project 2: Lachesis

Sub Team: Healthcare Cost and Outcome Analysis

Creator: Pulunuwan Jayasekera

1. Introduction

Australia's health system invests substantial resources to improve population health, yet outcomes and access differ widely across regions. Understanding how healthcare costs per person, potentially preventable hospitalizations (PPH), and socio-economic disadvantages align across Primary Health Networks (PHNs) is essential for identifying equity gaps and targeting improvement efforts.

1.1 Aim:

This study assesses regional cost effectiveness and equity by integrating PHN level outcomes (PPH per 100,000) with spending (state/territory cost per person) and socio-economic context (SEIFA IRSD). The analysis profiles regional patterns and highlights where high spending isn't reducing preventable hospitalizations, and where disadvantage is linked to higher PPH.

1.2 Approach:

Once the data were cleaned and combined, exploratory data analysis (EDA) examines distributions and relationships among key variables. Features are standardized and K-Means clustering is applied to group PHNs with similar cost outcome equity profiles. Cluster profiles and regional summaries are then interpreted to practical insights for decision makers.

1.3 Scope:

The analysis uses the most recent available datasets: AIHW Health Expenditure 2022–23, AIHW PPH rates 2022–23, and ABS SEIFA IRSD 2021. Data were cleaned, standardized, and integrated into a PHN-level dataset. Results are descriptive and comparative across regions rather than causal.

2. Data Sources and Preprocessing

2.1 Data Sources:

Three main datasets were used to build the analysis:

- **Health Expenditure Datacube**
  Source: AIHW
  Year: 2022–23
  Provides state and territory–level health expenditure per person. This dataset forms the cost indicator used to assess resource allocation.

- **Potentially Preventable Hospitalizations (PPH)**
  Source: AIHW
  Year: 2022–23
  Provides rates of hospital admissions that could have been avoided with timely and effective primary care, reported per 100,000 population at the PHN level. This dataset is the core outcome indicator of health system effectiveness.

- **Socio-Economic Indexes for Areas (SEIFA IRSD**
  Source: ABS
  Year: 2021
  Provides an index of relative socio-economic disadvantage at the SA2 level. This was mapped to PHNs using population-weighted concordance files, producing a PHN-level equity indicator.

2.2 Preprocessing and Integration:

Several steps were followed to transform the raw data into a unified, analysis-ready dataset:

- **Filtering:** limited each dataset to the most recent year (2022–23 for cost and PPH, 2021 for SEIFA). Only "All persons" demographic was retained for PPH.

- **Cleaning:** removed irrelevant rows (e.g., national totals, cross-border PHNs like Vic/NSW), standardized column names, and converted numeric fields.

- **Aggregation:** calculated PHN-level weighted averages for SEIFA using population as weights.

- **Merging:** combined cost (state-level), PPH (PHN-level), and SEIFA (PHN-level) datasets into a single baseline dataset with one row per PHN.

- **Feature scaling:** applied standardization to Cost, PPH, and SEIFA variables to allow direct comparison in clustering.

The resulting baseline dataset contained 30 PHNs with three core indicators: cost per person, PPH rate per 100,000, and SEIFA IRSD score. This served as the foundation for exploratory analysis and clustering.

# 3. Exploratory Data Analysis

3.1 Descriptive Stats Summary:

```
# Checking the statistics
eda_df[['PPH_rate_per_100k','Cost_per_person','SEIFA_IRSD_Score', 'IRSD_Decile_Mean']].describe()
```

	PPH_rate_per_100k	Cost_per_person	SEIFA_IRSD_Score	IRSD_Decile_Mean
count	30.000000	30.000000	30.000000	30.000000
mean	2049.114167	120.675400	994.258700	5.279067
std	526.648665	10.309658	36.942266	1.542697
min	1616.840000	111.426000	921.922000	3.324000
25%	1729.606000	111.426000	971.302500	3.926750
50%	1890.711500	120.072000	992.012500	5.303500
75%	2183.709000	124.859000	1016.204750	6.077500
max	4152.960000	158.075000	1085.809000	9.188000

- **Cost per person:** 111–158 (mean ≈ 120.7); low spread.

- **PPH rate per 100k:** 1617–4153 (mean ≈ 2049); wide spread.

- **SEIFA IRSD score:** 922–1086 (mean ≈ 994); moderate spread.

3.2 Distributions
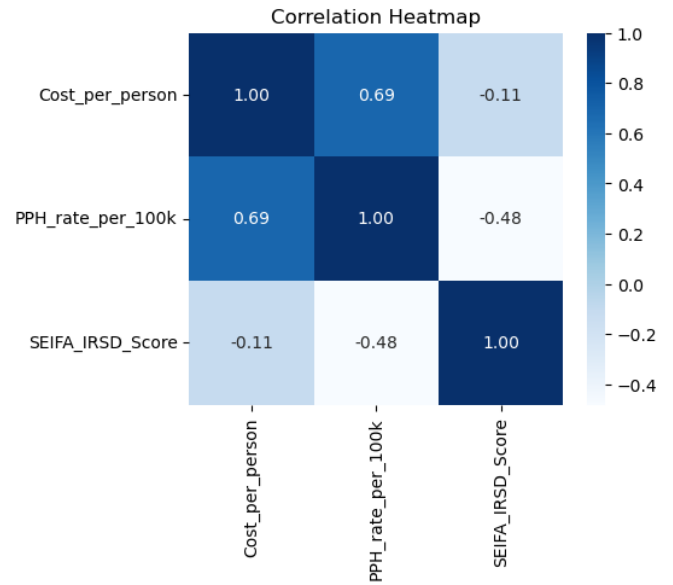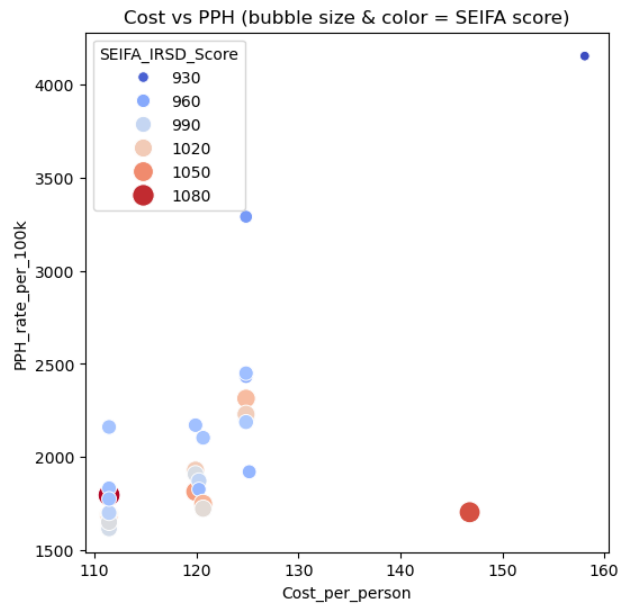
Distribution of SEIFA_IRSD_Score
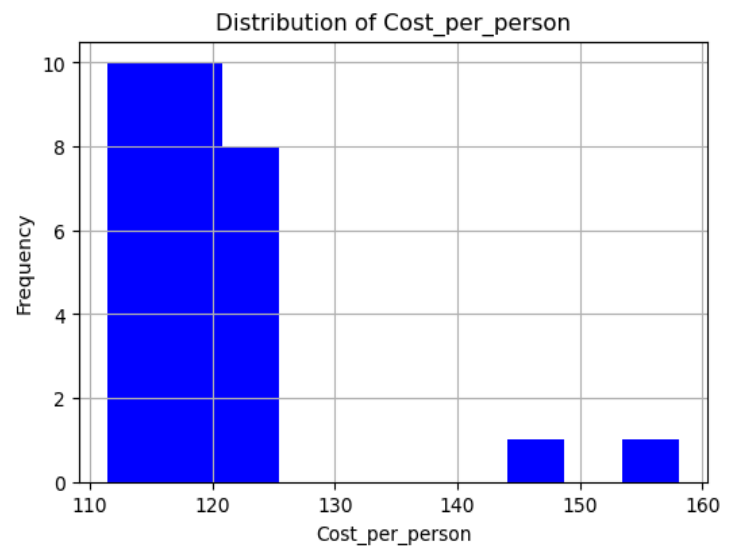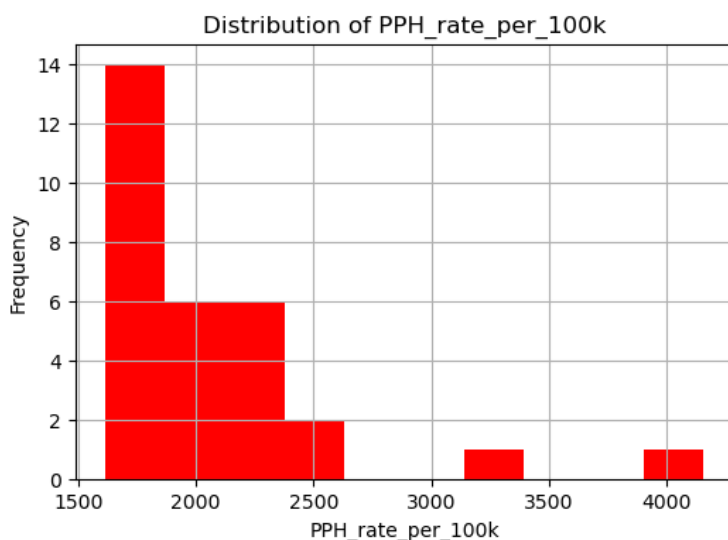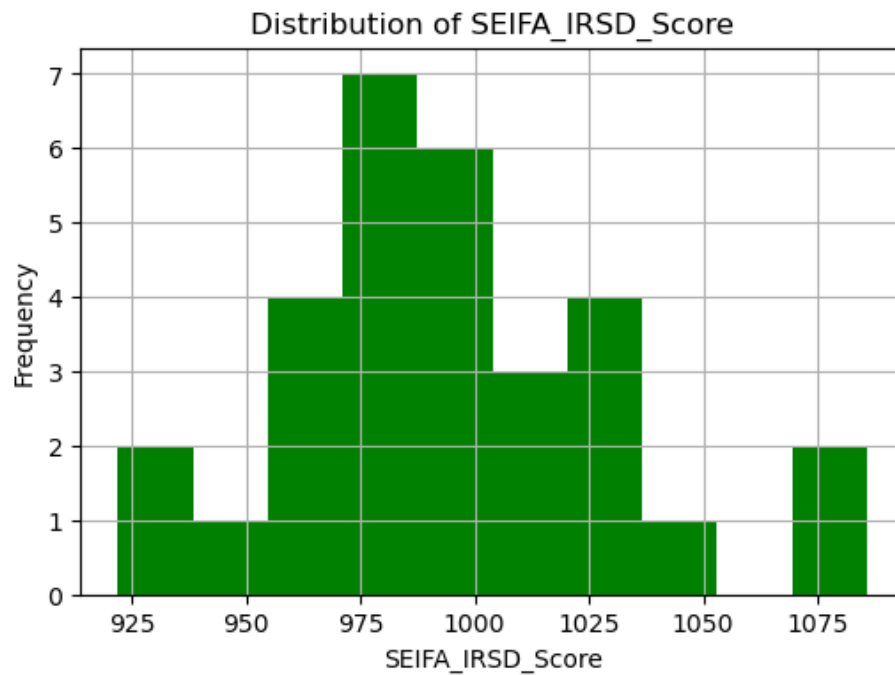
- **Cost per person:** most PHNs cluster tightly around 110–125; a few high-cost points (~150, ~158).

- **PPH rate:** most PHNs sit between 1600–2500, with two clear high outliers (>3000 and >4000).

- **SEIFA IRSD:** roughly bell-shaped around ~1000, with tails toward ~920 and ~1080.

3.3 Relationship:

- **Cost vs PPH:** moderate positive correlation (r ≈ +0.69). PHNs with higher spend often also show higher PPH—suggesting spending is reactive to need \.

- **SEIFA vs PPH:** moderate negative correlation (r ≈ −0.48). More disadvantaged areas (lower SEIFA) tend to have higher PPH.

- **SEIFA vs Cost:** near zero (r ≈ −0.11). Spending per person isn't strongly aligned with socio-economic status at the PHN level.

3.4 Outliers
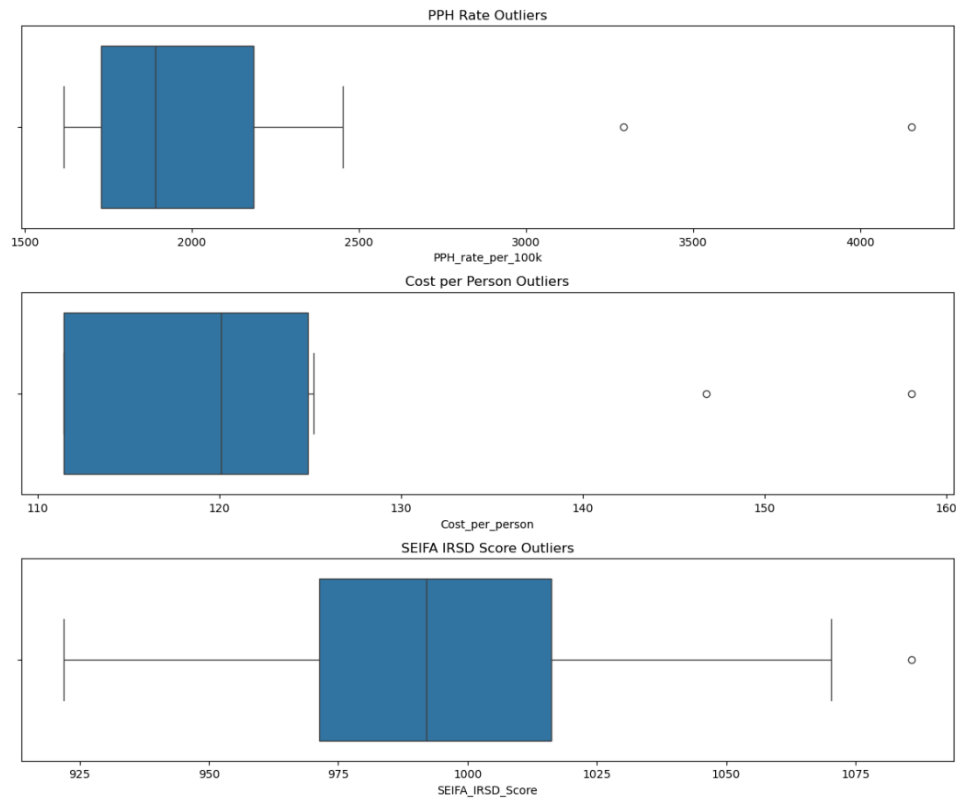
- **PPH rate:** two strong high outliers (≈3300+ and ≈4150+).

- **Cost per person:** two high outliers (~150 and ~158).

- **SEIFA IRSD:** one high outlier (≈1086).

3.5 Key EDA Insights

- PPH varies a lot more than spending. Outcomes swing widely between PHNs, while per-person spend is comparatively steady.

- High spending with high PPH likely reflects higher need. Areas with worse outcomes often attract more funding, so the pattern doesn't mean "more money automatically lowers PPH."

- Socio-economic disadvantage is meaningfully associated with worse outcomes (higher PPH), while spend does not systematically track disadvantage.

4. Clustering Analysis

4.1 Objective:

Group PHNs into peer clusters based on Cost per person (cost), PPH rate per 100k (outcome), and SEIFA IRSD score (equity). This supports like-for-like comparison, highlights outliers, and identifies "good practice" vs "at-risk" regions.)

4.2 Method:

K-Means clustering was applied to standardized features (Standard Scaling on Cost, PPH, SEIFA). The number of clusters was chosen as k = 4. The Elbow curve flattened around 4, while Silhouette favored very coarse splits (k≈2). K=4 was selected to obtain actionable peer groups rather than a binary partition.
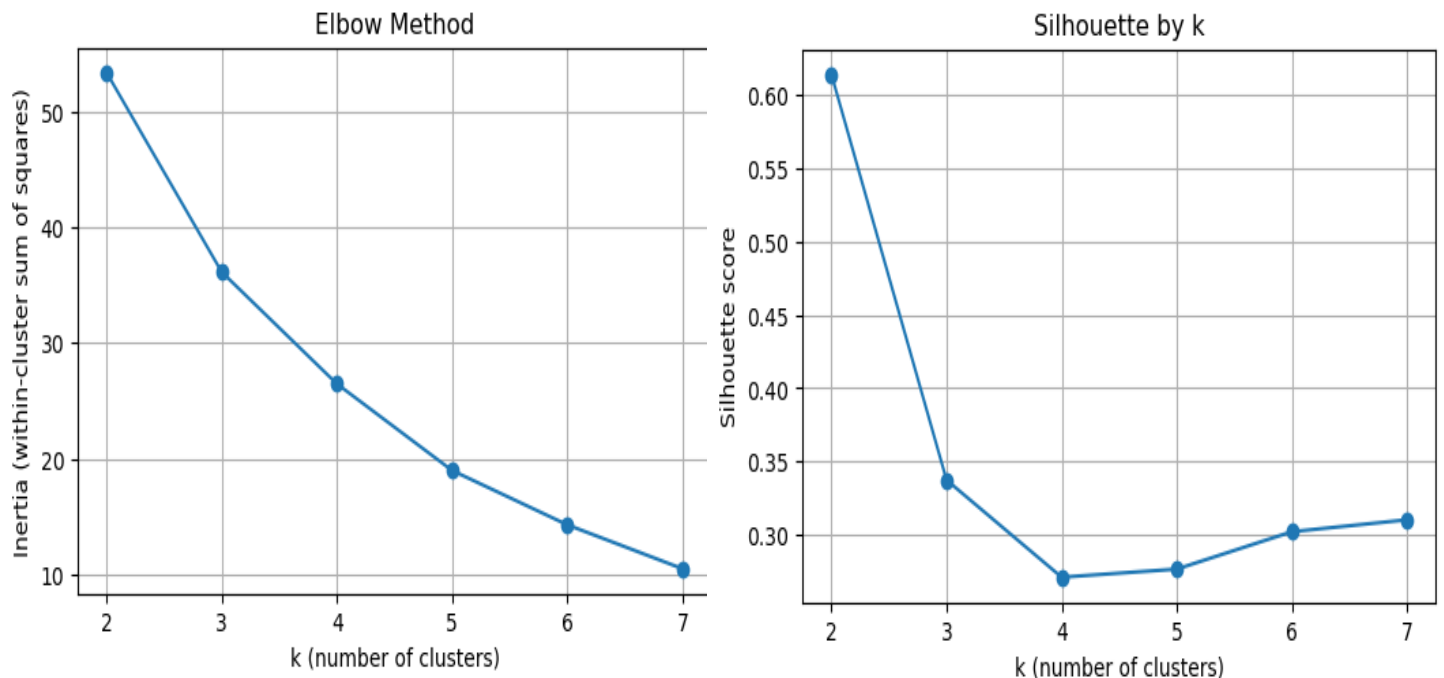
4.3 Cluster Profile Summary:

Cluster	n(PHNs)	Cost per person (mean) - AUD	PPH per 100k (mean)	SEIFA IRSD (mean)	Short Label
0	8	123.8	2345	972	Mid cost, higher PPH, more disadvantaged.
1	8	122.5	1903	1042	Mid cost, lower PPH, more advantaged.
2	1	158.1	4153	922	Critical outlier (very high cost & PPH, most disadvantaged)
3	13	114.8	1795	984	Lower cost and lower PPH(efficient baseline)

4.4 Scatter: Cost vs PPH, color by cluster, bubble = SEIFA
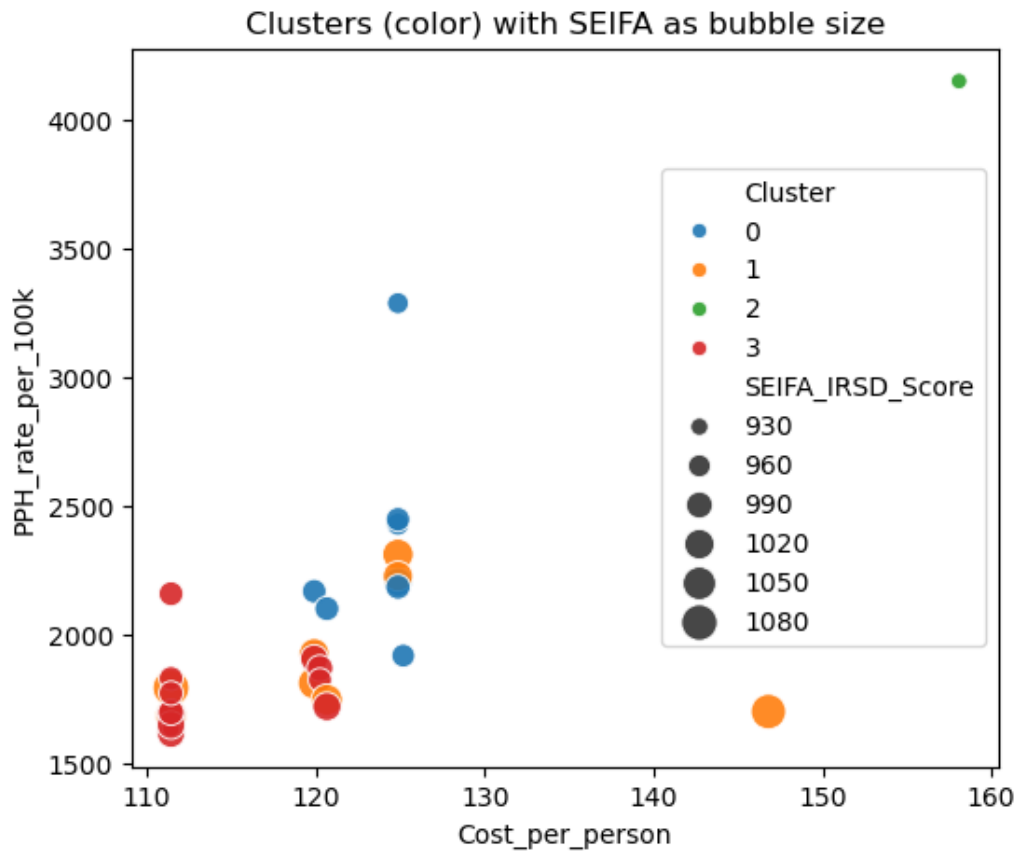
4.5 State Cluster Matrix:

State Cluster	ACT	NSW	NT	QLD	SA	TAS	VIC	WA
0	0	0	0	5	0	1	1	1
1	1	2	0	2	0	0	2	1
2	0	0	1	0	0	0	0	0
3	0	8	0	0	2	0	2	1

4.6 Cluster Highlights:

- Cluster 0 (n=8): Mid-level cost (~124), higher PPH (~2,345), below-average SEIFA (~972).

- Cluster 1 (n=8): Mid-level cost (~122), lower PPH (~1,903), higher SEIFA (~1,042).

- Cluster 2 (n=1): Very high cost (~158), extreme PPH (~4,153), low SEIFA (~922).

- Cluster 3 (n=13): Lower cost (~115), lower PPH (~1,795), SEIFA around average (~984).

5. Findings & Interpretation

5.1 Key Findings:

- **Efficiency gradient:**

  - Cluster 3 (13 PHNs) exhibits lower cost with lower PPH; indicative of comparatively efficient/healthy performance.

  - Cluster 1 has similar cost to Cluster 0 but better outcomes (lower PPH) and higher SEIFA, suggesting socio-economic advantage is associated with better outcomes at comparable spend.

- **Equity signal:**

  - Cluster 0 vs Cluster 1 shows how SEIFA relates to outcomes: at similar spend, more disadvantaged PHNs (Cluster 0) tend to have higher PPH.

- **Critical outlier:**

  - Cluster 2 (NT is the only state belongs to this cluster) has very high cost and very high PPH with low SEIFA.

- **Within-state variability:**

  - PHNs within the same state land in different clusters (e.g., NSW mostly in Cluster 3; QLD split across Clusters 0–1 and VIC split across Clusters 0,1 & 3), showing real regional differences that state averages can hide.

5.2 Interpretation:

- Spending not equal outcomes by itself. Similar per capital expenditures produces different outcomes depending on socio – economic context and service configuration.

- Lower SEIFA PHNs show higher PPH even with similar spending, pointing to gaps in access and primary care.

- Direct funding and programs to PHN peer groups rather than using one statewide approach.

6 Limitations & Data Considerations

- Australia has 31 PHNs nationally, but one cross-border PHN (Vic/NSW) was excluded due to complexity in cost/outcome assignment. This left 30 PHNs for analysis.

- SEIFA IRSD data from 2021 was combined with health expenditure and PPH data from 2022–23. This is the closest available alignment but may introduce minor inconsistencies.

- Health expenditure was available only at state/territory level, while PPH and SEIFA were PHN-level. This required mapping and assumptions that may smooth over local variations.

7. References

- Australian Institute of Health and Welfare (AIHW). (2023). Health expenditure Australia 2021–22: Data tables (HWE 101). Canberra: AIHW.
  Link

- Australian Institute of Health and Welfare (AIHW). (2023). Potentially preventable hospitalisations in Australia by small geographic areas, 2017–18 to 2021–22: Data tables (HPF 72). Canberra: AIHW.
  Link

- Australian Bureau of Statistics (ABS). (2023). Socio-Economic Indexes for Areas (SEIFA), Australia, 2021 (cat. no. 2033.0.55.001). Canberra: ABS.
  Link

8. Appendix

- The complete code including notebooks and datasets used and generated in the analysis are available at https://github.com/Redback-Operations/redback-senior-mobile/tree/main/Cost_Of_HealthCare_Analysis/Regional_Equity_and_Clustering_Analys