Final Report

(for client review)

Harnessing Big Data to Improve Financial Integrity and Operational Efficiency Project

Group 17

Feifei Ou (24490283)

Kuan Teng Lee (24331733)

Justin Lu (24326939)

Sudha Palanimalai Palanimalai (24358475)

Rafid Rizwan Sohan (24312485)

Charchil Shailesh Singh (24311765)

10 OCT 2025

**Table of Contents**

[1. Executive summary 3](#_Toc210985299)

[2. ETL Process 4](#_Toc210985300)

[3. Datawarehouse 4](#_Toc210985301)

[4. Exploratory Data Analysis 4](#_Toc210985302)

[5. Modelling 9](#_Toc210985303)

[6. Data visualization 12](#_Toc210985304)

[7. Solution and summary 13](#_Toc210985305)

[8. limitation and future improvement 15](#_Toc210985306)

# 1. Executive summary

This project investigates inconsistencies between UWA’s internal Commonwealth Supported Place (CSP) enrolment records and the Australian Government’s official funding allocations. Using three available datasets—*student enrolment data (2024)*, *indexed government rates*, and the *2024 funding agreement*—team 17 rebuilt a comparable analytical framework to improve transparency and data governance in higher-education funding.

Exploratory analysis revealed that the main discrepancy arises from different reporting logics: UWA data are recorded at the *unit* or *course* level, while government allocations are based on *Funding Cluster × Field of Education (FOE)*. To address this, the team tested three reconciliation methods. The first (student-level assumption) exposed structural inconsistencies; the second (government-aligned recalculation) achieved logical alignment with policy tables; and the third (cleaned & versioned) integrated FOE mapping, rate versioning, and handling of special cases to produce the most reliable reconciliation framework.

The recalculated total Commonwealth contribution ($161.9M) was slightly higher than the official figure ($147.2M), with differences mainly due to unmapped FOE codes (366 records), absence of grandfathered student details, and lack of student-level identifiers.

In addition, a **logistic regression model** was developed to demonstrate how data-driven methods could identify potential funding discrepancies. An *error flag* was created when the gap between actual and expected payments exceeded 10%.Although the model was built using simulated data due to the absence of detailed payment records, it achieved strong performance (accuracy 87%, F1-score 0.81) and highlighted that **unit type**, **government contribution**, and **EFTSL load** are the most influential factors behind discrepancy risks.  
 This modelling framework illustrates how predictive analytics could enhance funding audit, anomaly detection, and decision-making once complete datasets become available.

Despite these data limitations, the project successfully delivered a repeatable, auditable reconciliation process and a prototype dashboard that allows finance teams to identify and explain funding differences without altering internal systems.

# 2. ETL Process

**Key Findings and *Method*:**

* Focus on CSP Student only
* Filter out mismatched FOE codes

The FOE code 60000 and 120000 are present exclusively in the “2024 UWA Student Data” dataset, with no matching entries found in the “2024 Allocation of Units of Study” dataset.

* Filter out special FOE codes

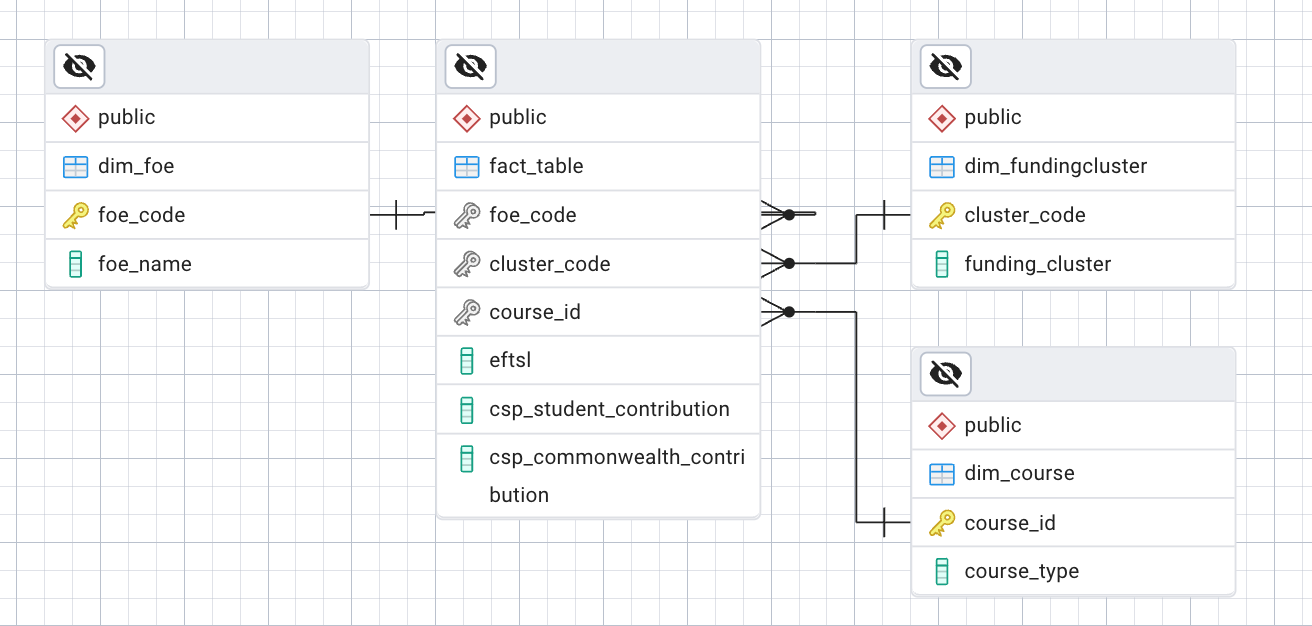
We cannot merge special FOE code without student major data - such as Funding Cluster varies for FOE depending on E312 or E392 (Yes/No).

* Simplify government contribution calculation

We cannot identify which year student enrolled in a specific unit and which contribution rate (2024 contribution rate or grandfather contribution rate) without Student detailed data. Therefore, we assume that Government payment is based on commonwealth contribution 2024 only.

# 3. Datawarehouse

A Star Schema Data Warehouse was implemented in PostgreSQL to store cleaned data, improve query performance, and provide structured analytical access for reporting and visualization.



This data warehouse was designed with one fact table and three dimension tables, including dim\_foe, dim\_fundingcluster and dim\_course. Specifically, the fact table only includes CSP Student data for further analysis.

# 4. Exploratory Data Analysis

**4.1 Key Findings:**

* The dataset is **not student-level**, but a **combination of Course × Unit × Field of Education (FOE)**.
* **EFTSL** (Equivalent Full-Time Student Load) measures study load, not student headcount.
* The **university and government apply different aggregation logics** — the government calculates funding at the **Funding Cluster + FOE** level, while UWA originally used **Course × Unit** logic.

**4.2 Data Understanding**

**Data Sources:**

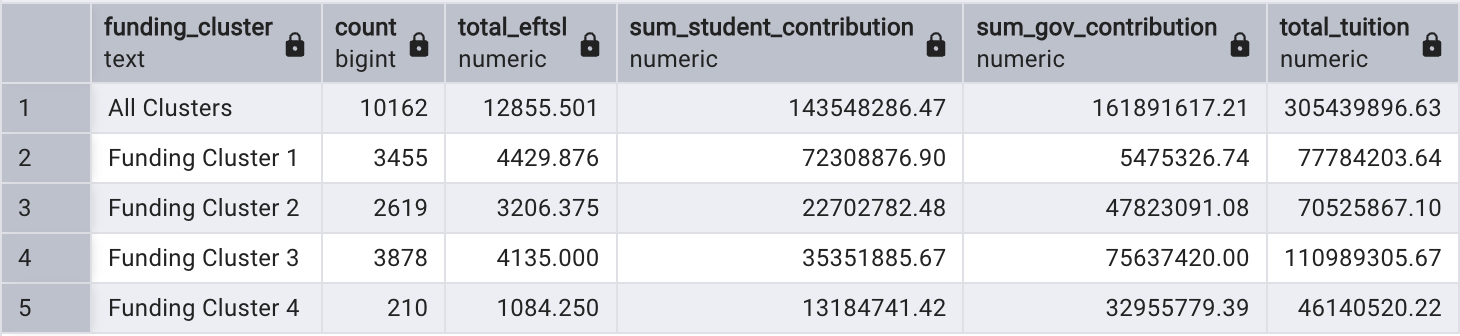
* *2024 UWA Student Data Table* (university internal records)
* *2024 CSP Allocation Table* (government mapping of FOE → Funding Cluster)
* *WA Government 2024 Funding Agreement* (official indexed contribution rates)

Key Variables:

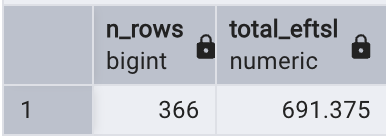
|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Notes** |
| UWACourseID | Degree or program identifier | e.g., Master of Information Systems |
| UWAUnitID | Individual unit identifier | e.g. CIITS5504 datawarehouse |
| FOE Code | Field of Education classification | e.g., Information Technology |
| EFTSL | Equivalent Full-Time Student Load | Measures academic load, not student count |

**4.3 Exploratory Data Analysis**

**4.3.1 Total EFTSL Across All Funding Clusters**



**4.3.2 Records with Special FOE Codes**

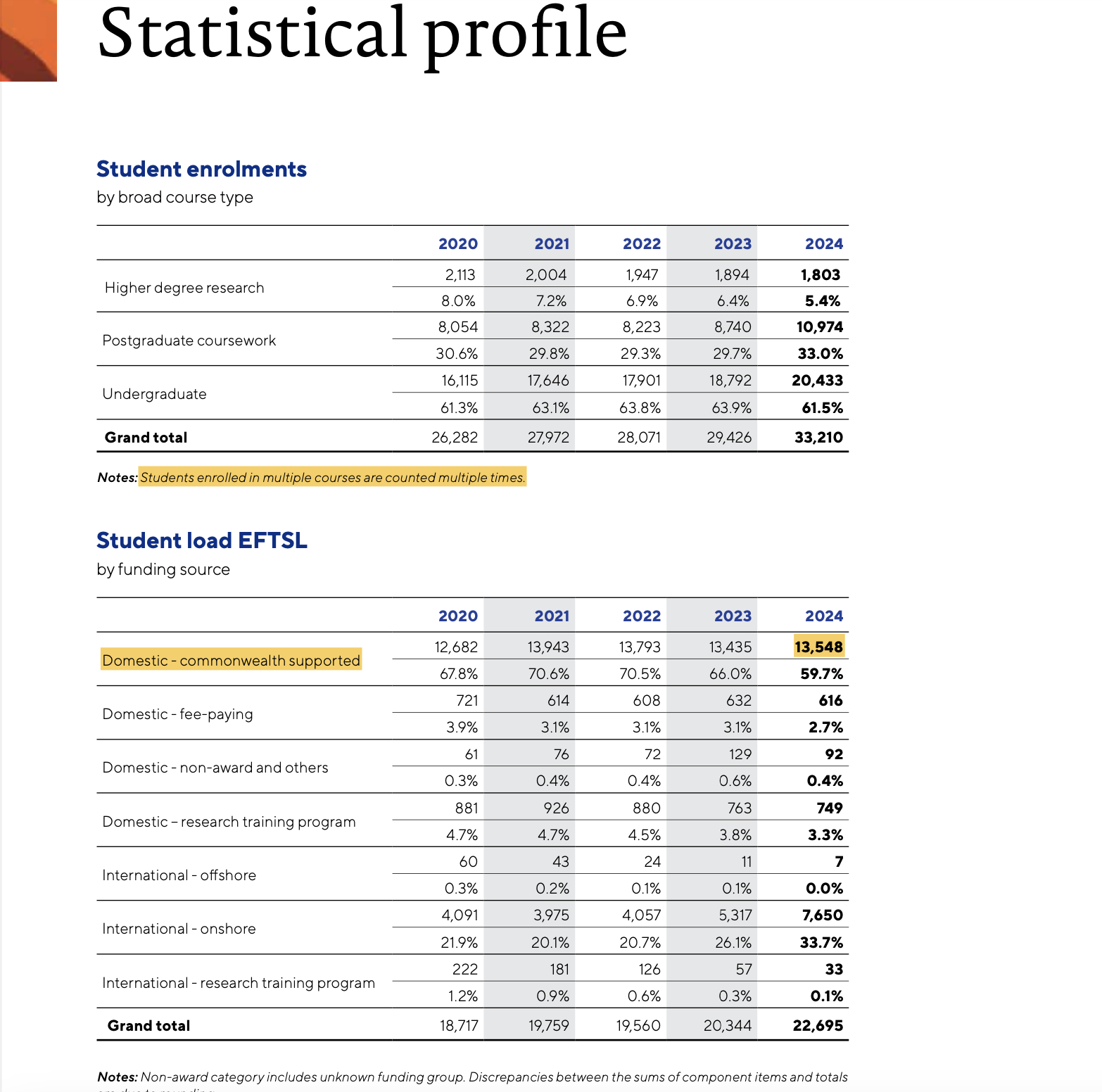


(Mostly belong to Psychology and Behavioural Science units)

Interpretation:

* Total EFTSL Across All Funding Clusters(without Special FOE Codes):12,855.5
* 366 records identified without FOE codes and Total EFTSL = **691.375**
* **Total EFTSL Across All Funding Clusters = 13546.875**
* Total Government Contribution (without Special FOE Codes group students)): $161,891,617

The record shown in UWA 2024 annual financial report about commonwealth student total EFTSL = 13548, only **1.125** points difference with our calculation.



The total EFTSL and contribution figures confirm that the dataset accurately represents aggregated study loads rather than individual enrolments.

**4.3.3 Validation Against Government Totals**

|  |  |
| --- | --- |
| **Source** | **Total Commonwealth Contribution** |
| UWA Internal (recalculated) | $161.89M |
| Government Funding Agreement 2024 | $147.20M |

**Interpretation:**

The total Commonwealth contribution recalculated from UWA’s internal enrolment data ($161.89M) exceeds the amount reported in the 2024 Government Funding Agreement ($147.20M). This discrepancy can be attributed to several data and methodological limitations identified during our reconciliation process:

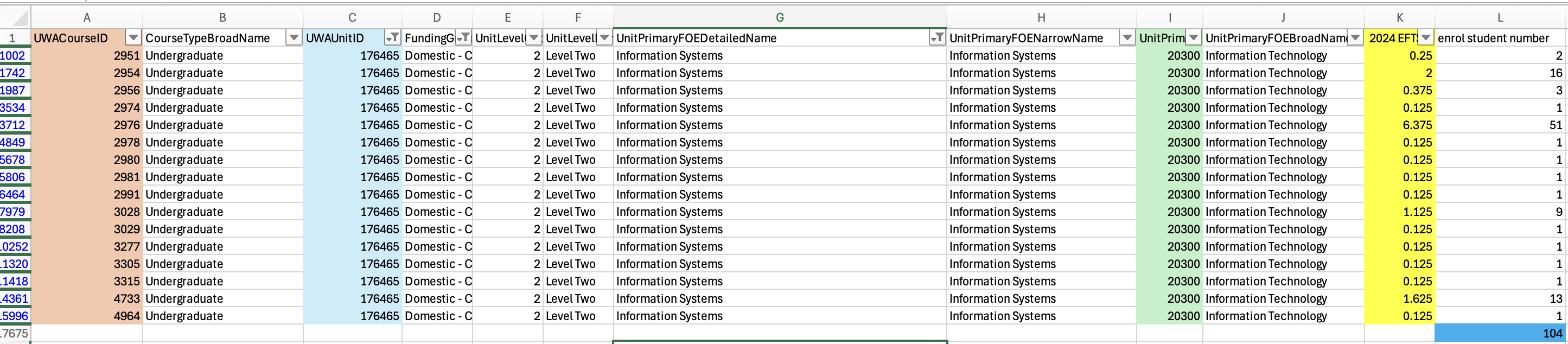
**1. Excluded special FOE cases:**  
The recalculation excluded 366 enrolment records with unmapped or special FOE codes due to missing mapping information from the client. These records are likely to contribute to the funding difference.

**2. Absence of grandfathering data:**  
The dataset provided does not include information indicating whether a student is classified as “grandfathered” (i.e., eligible to retain pre-2021 contribution rates). If such students exist, applying the grandfathering rule would lower the total recalculated Commonwealth contribution.

**3. Aggregation limitations:**  
The current dataset is structured at the unit × FOE × course × EFTSL level rather than the individual student level. As a result, potential double counting or overlapping EFTSL across units for the same student cannot be verified or adjusted, possibly inflating UWA’s internal total.

**4.3.4 Data Granularity Verification**

* Data Granularity – Each row represents a **combination of (UWACourseID × UWAUnitID × FOE Code × Level)**.



**Why the same Unit has many rows**

* A single unit can appear multiple times because:
  + It belongs to different degree programs (different UWACourseIDs).
  + It is split across different FOE codes.
  + It is separated by unit levels.

So: **1 Medical Studies unit ≠ 1 row**, it is broken into many rows.

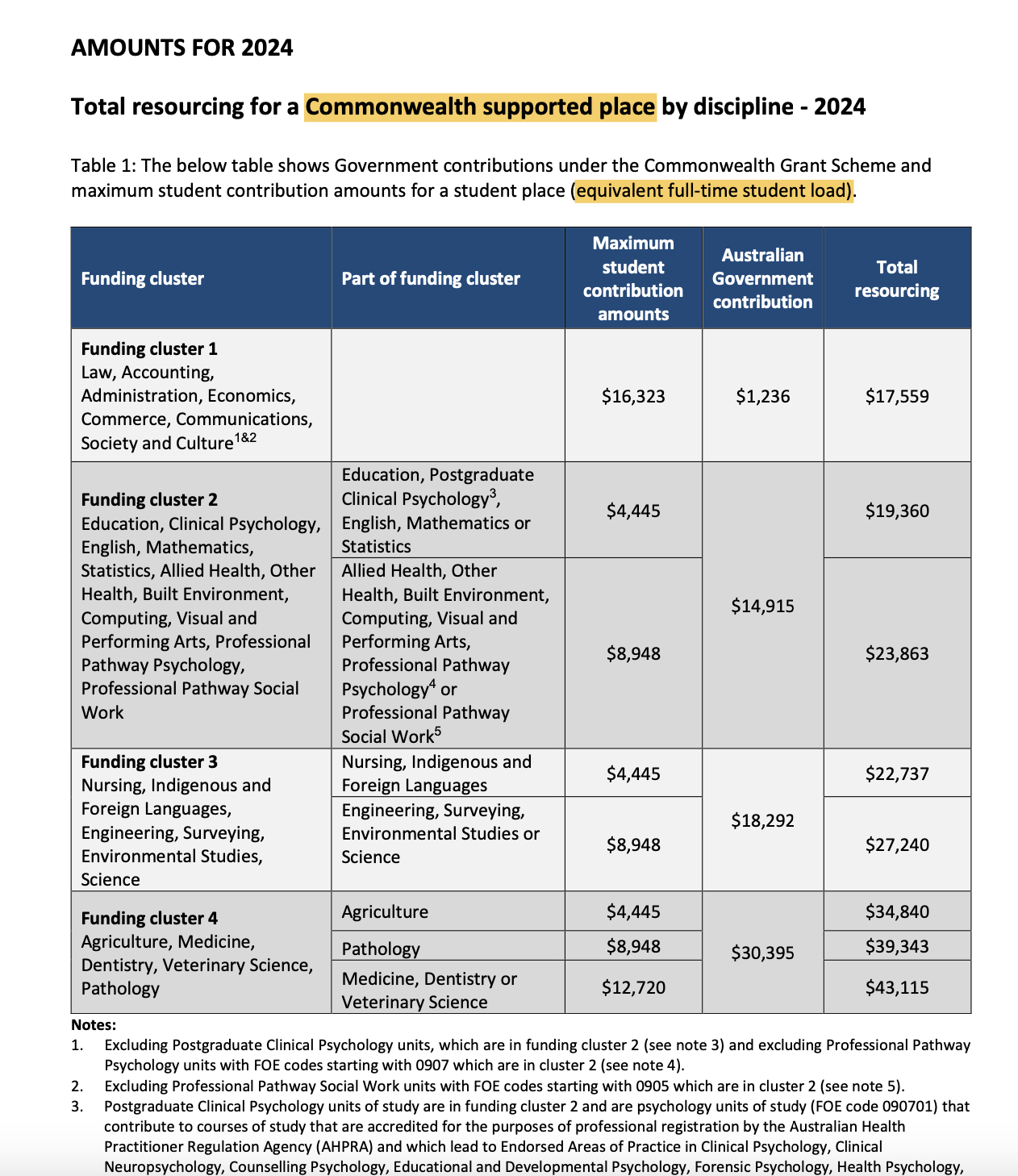
**What EFTSL means**

* **EFTSL = Equivalent Full-Time Student Load**
* It does not represent headcount; it represents “student load.”
  + 1 EFTSL ≈ one student’s full-time study load for a year.
  + 0.25 EFTSL → could mean:
    - 2 students each at 12.5% load, or
    - more students splitting that load.

**4.4 Conclusion:**

Our analysis indicates that the discrepancy between the University’s calculations and the Government’s payments arises from **differences in data granularity rather than miscalculation of EFTSL**.

**Government methodology:**



Under the Commonwealth Grant Scheme, the Government allocates funding based on **equivalent full-time student load (EFTSL)** at the level of **Funding Cluster + Field of Education (FOE) code**. In other words, the Government looks at the aggregate EFTSL for each cluster/discipline combination and applies the indexed contribution rate.

**University methodology:**

The student enrolment dataset provided by the University records EFTSL at the unit/course enrolment level. Each row represents a unit-level enrolment with an associated EFTSL value. These records are much more granular and do not directly correspond to “one student”. As a result, when these values are aggregated, the total EFTSL is correct overall, but the distribution across rows looks distorted (for example, a single row may appear to carry an extremely large EFTSL).

# 5. Modelling

We use logistic regression for data modelling with error\_flag created if discrepancy amount between actual and expected amount are more than 10% error flag will be triggered as 1 and the rest will be remained.

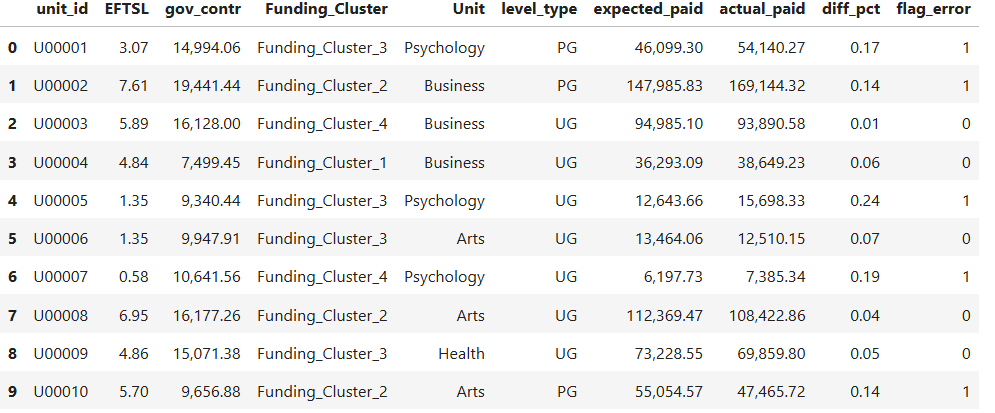
The goal of this model is to understand, clarify that when a payment discrepancy occurs specifically, which area will likely have the discrepancy amount different occurred.

Since we don’t have the specify actual amount, we can’t use one total amount to compare with expected amount in the dataset for few reasons:

* The agreement didn’t have specify contribution in clusters and units
* Need a clear dataset with clear calculation in actual amount

Instead of this, we simulate data on how logistic modelling helps us to recognize pattern if we have a clear dataset. This allowed us to demonstrate how logistic regression can be effectively used to detect and understand patterns in discrepancies once a complete and more accurate dataset becomes available for future analysis.

**5.1 Data simulation**



Each rows represent each unit with unit\_id, total amount of EFTSL with levels(etc. Postgraduate), funding cluster, gov\_contr and expected paid.

gov\_contr = contribution by government base on each cluster   
expected\_fee = gov\_contr \* total\_eftsl   
actual\_fee = actual fee from agreement or specifc dataset  
error\_flag = 1 will be triggered if the actual and expected discrepancy are more than 10%

**5.2 Model Performance**

|  |  |  |
| --- | --- | --- |
| **Actual / Predicted** | **Negative** | **Positive** |
| **Negative** | 2632 | 60 |
| **Positive** | 540 | 1268 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Class 0** | **Class 1** | **Macro Avg** | **Weighted Avg** |
| **Preicison** | 0.83 | 0.95 | 0.89 | 0.88 |
| **Recall** | 0.98 | 0.70 | 0.84 | 0.87 |
| **F1-Score** | 0.90 | 0.81 | 0.85 | 0.86 |
| **Support** | 2692 | 1808 | 4500 | 4500 |
| **Overall Accuracy** | | | 0.87 |  |

The model achieved a high overall accuracy of **87%**, indicating that it correctly predicted the payment discrepancy outcome for most cases. In particular, the model performed well in distinguishing between units with and without significant errors. The precision for identifying discrepancy cases (flag = 1) was **0.95**, meaning that 95% of the cases predicted as having an error were actually true errors. The recall for the same class was **0.70**, showing that the model successfully detected 70% of all true errors in the dataset. The F1-score of **0.81** reflects a strong balance between precision and recall. Overall, the model provides a reliable and balanced classification performance, effectively identifying potential errors while keeping false alarms low.

**5.3 Important and findings**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Coefficient (coef)** | **Odds Ratio** |
| **Unit\_Psychology** | **3.48** | **32.42** |
| **Unit\_Arts** | -1.53 | 0.22 |
| **Unit\_Business** | -1.49 | 0.22 |
| **Unit\_Health** | -0.52 | 0.59 |
| **Funding\_Cluster\_Funding\_Cluster\_4** | -0.07 | 0.94 |
| **gov\_contr** | 0.06 | 1.06 |
| **level\_type\_UG** | -0.06 | 0.94 |
| **EFTSL** | 0.05 | 1.05 |
| **level\_type\_PG** | 0.00 | 1.00 |
| **Funding\_Cluster\_Funding\_Cluster\_2** | 0.00 | 1.00 |
| **Funding\_Cluster\_Funding\_Cluster\_1** | -0.00 | 1.00 |
| **Funding\_Cluster\_Funding\_Cluster\_3** | 0.00 | 1.00 |

The analysis was conducted using a dummy dataset, the model still provides a useful demonstration of how logistic regression can identify the most influential variables affecting payment discrepancies.

In this scenario, three variables showed the strongest influence on payment discrepancies: unit type, government contribution, and student load (EFTSL). Among these, the unit type had the greatest impact Psychology units were much more likely to experience large discrepancies, while Arts and Business units appeared more stable and consistent. The government contribution showed a small positive relationship, meaning higher funding slightly increased the chance of discrepancies. Likewise, EFTSL had a mild positive effect, suggesting that larger or more intensive units were somewhat more prone to discrepancies in the simulated data.

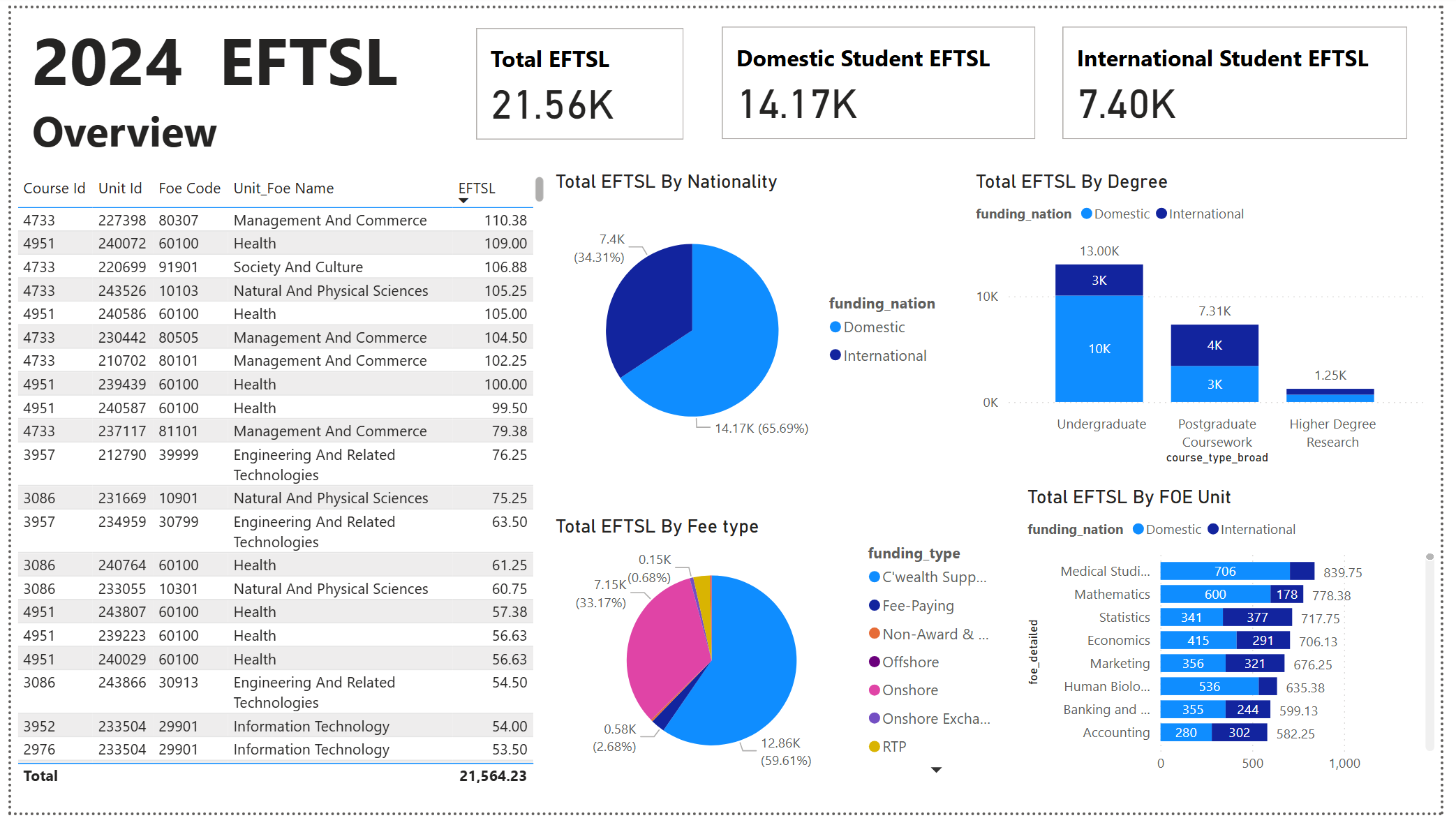
Overall, while these results do not represent real-world patterns, they illustrate how logistic regression can detect key drivers and patterns once accurate data becomes available and future purposes. With access to actual, detailed payment records, this same modelling approach could be applied to produce meaningful and actionable insights such as identifying high-risk areas for funding errors and improving payment accuracy in future analyses.

**5.4 Modelling Objective and Approach**

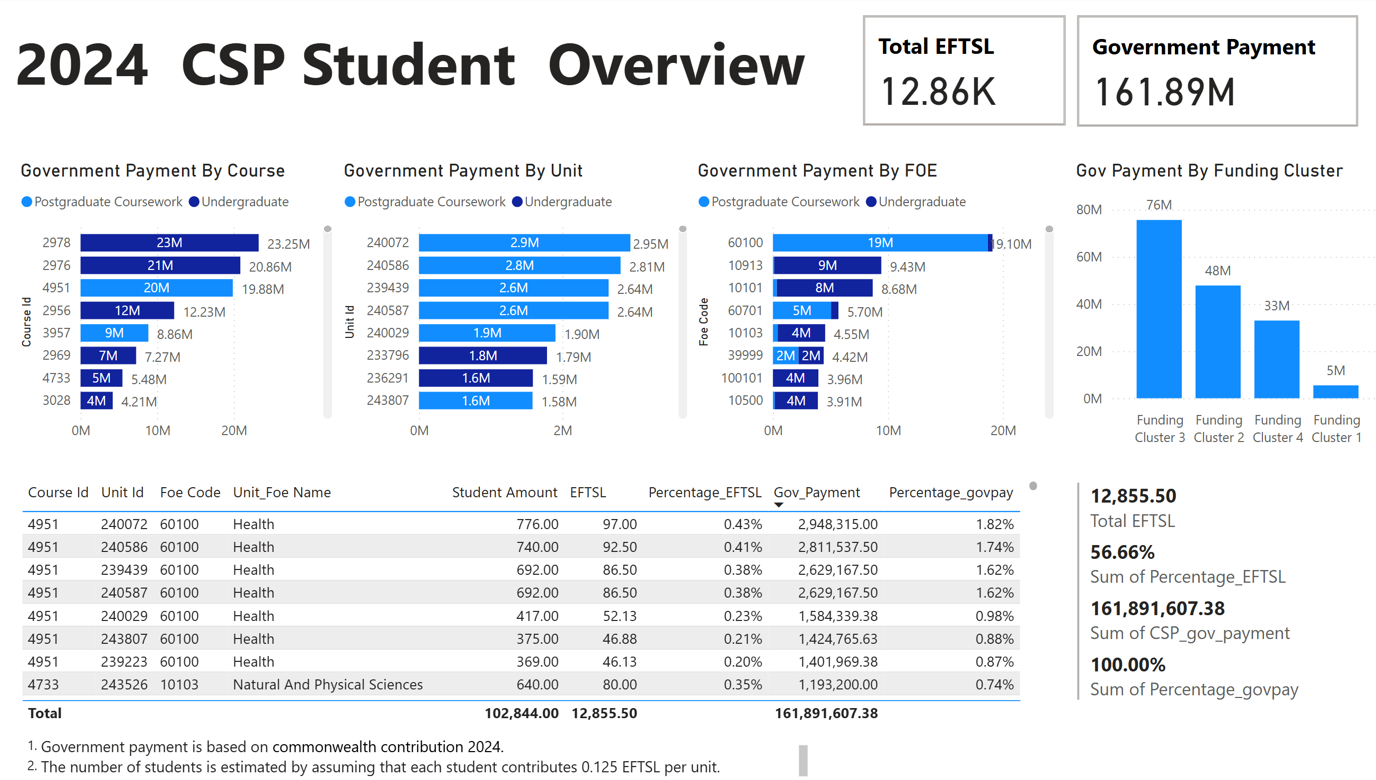
This demonstrates how logistic regression can be used to analysis and predict payment discrepancies in funding data. Even though the dataset was simulated, the model successfully illustrated on how key variables such as unit type, funding amount, and student load can influence the likelihood of payment errors. In a real-world setting, this approach could help organizations detect unusual payment patterns, prioritize audits, and improve funding accuracy across departments. For future application, obtaining complete and specific actual payment data would allow the model to be retrained and validated for real operational use, improving both its predictive accuracy and practical decision-making value.

# 6. Data visualization

**2024 EFTSL Overview:** Focuses on the total student Equivalent Full-Time Student Load (EFTSL) by key dimensions such as nationality, Degree Level, and Fee Type. This dashboard provides a comprehensive and interactive view of the EFTSL distribution.



**2024 CSP Student Overview:** Focuses on students in Commonwealth Supported Place, visualizing Government Payment distribution by Course, unit, FOE, and Funding Cluster. This is primary dashboard to identify key patterns and discrepancies between internal payment data and external data sources.



# 7. Solution and summary

**7.1 Context**

As the project team did not have access to the client’s internal database structure (UWA Finance or BI system schema), the analysis was conducted entirely based on three provided files:

1. **Student enrolment data**
2. **Indexed government funding rates**
3. **Funding agreement (2024)**

The internal reporting logic and database schema of UWA remain unknown, and it was therefore not feasible to recommend system-level interventions such as schema redesigns or database views. Instead, the proposed solutions focus on **data-level and process-level reconciliation** that can be implemented using existing datasets, Excel, ETL scripts, or annual audit workflows.

**7.2 Proposed Solutions**

**7.2.1 Data-level reconciliation**

**Goal:** Rebuild a comparable view of EFTSL and funding amounts at the *Funding Cluster × FOE* level, enabling one-to-one comparison between UWA’s internal data and the Government’s indexed rates.

**Approach:**  
At the student-level enrolment table, insert two additional fields — FOE Code and Funding Cluster — by mapping the unit’s UnitPrimaryFOECode to the Government’s FOE–Cluster reference table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Student\_id** | **UWACourseID** | **UWAUnitID** | **FundingGroupName** | **2024 EFTSL** | **UnitPrimaryFOENarrowName** | **FOE** | **Funding cluster** |
| **12345** | **2951** | **176465** | **Domestic - C'wealth Supported** | **1** | **Information Systems** | **20300** | **2** |

This enriched table allows direct calculation of:

* Government (indexed) funding per record = EFTSL × Indexed Rate;
* (If available) Difference per record = UWA Amount − Government Indexed Amount.  
  **Outcome:**

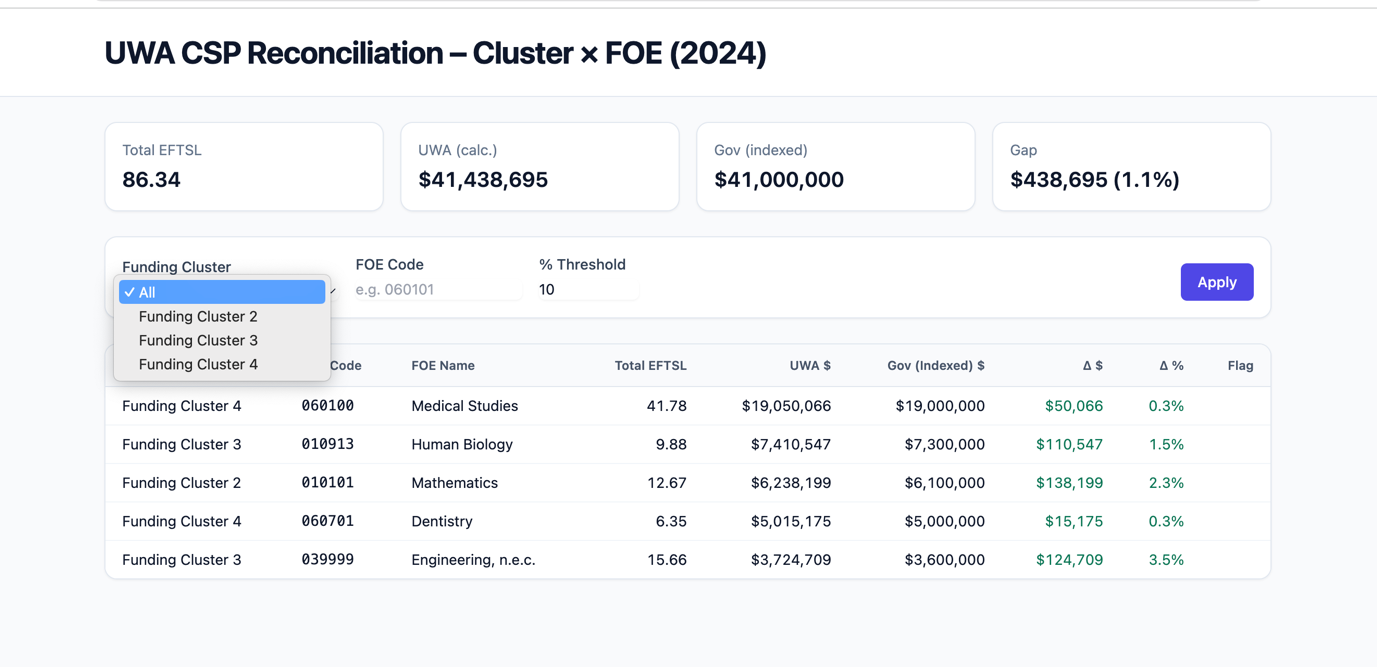
A complete row-level reconciliation dataset that can be aggregated to *Cluster × FOE* for annual comparison. Each row now contains both UWA and Government definitions of funding, allowing transparent variance tracking without any change to internal systems.

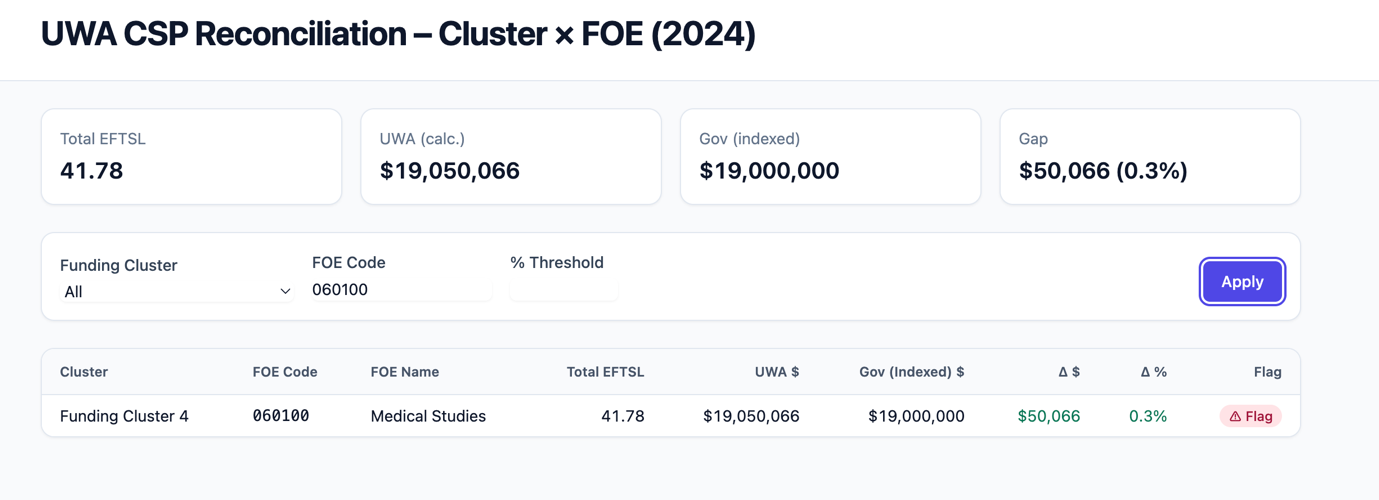
**7.2.2 System-level visualization (prototype model)**

**Goal:** Provide a lightweight, repeatable dashboard to visualize reconciliation results without modifying UWA’s BI or finance systems.

**Approach:**  
Developed an offline interactive HTML prototype that reads the aggregated *Cluster × FOE* dataset and automatically computes totals and variances.  
 Key functions include:

* KPI summary cards (Total EFTSL, UWA $, Government $, Gap %);
* Filters by Funding Cluster, FOE Code, and variance threshold;
* Conditional flagging of large deviations;
* Export capability for audit documentation.





**Outcome:**  
This prototype offers a simple, sustainable governance tool that finance staff can refresh annually by updating one CSV file. It complements existing reporting processes and clearly visualizes the source of discrepancies between UWA and Government figures.

<https://github.com/justinwonderwall/Capstone_Project/blob/Fei/uwa_csp_recon%20(1).html>

# 8. limitation and future improvement

**8.1 Limitations**

This project has several limitations that constrain the interpretation and generalisability of its findings:

1. **Limited dataset scope**  
   The available dataset only covers **2024 enrolment records**, and many special student categories were not included — such as cases with unmapped or exceptional FOE codes, *grandfathered* students under legacy funding rules, and both international and domestic fee-paying cohorts. As a result, the current analysis provides only a **limited view** of the overall funding structure.
2. **Use of synthetic or incomplete data**  
   The reconciliation model was tested on a simulated or incomplete dataset rather than the university’s live production data. Consequently, the model’s current accuracy and interpretive power are limited. With access to additional years or student-level datasets, more robust comparisons and model validation could be performed.
3. **Unknown internal data structure**  
   The internal schema and reporting logic of UWA’s finance and BI systems remain inaccessible to the team. This restricts the ability to design system-integrated solutions or to verify whether certain discrepancies stem from data warehouse transformations or reporting filters.
4. **No record-level external validation**

Without access to government record-level data, the team could not confirm whether residual gaps (e.g. $14.7M total variance) result from true funding policy differences or structural misalignment in the data aggregation process. The validation is therefore **macro-level rather than transaction-level**.

**8.2 Future Improvements**

Future iterations of this project could address these limitations by:

* Incorporating **multi-year enrolment data** and adding external datasets (e.g., fee-paying, international, and postgraduate cohorts) to achieve a full population view.
* Integrating **automated FOE mapping** and **grandfather-rule identification** to improve accuracy of funding classification.
* Extending the current reconciliation model into a **governance dashboard** with real data connections, allowing continuous monitoring rather than static annual checks.
* Comparing **alternative reconciliation models** (e.g., regression-based prediction vs. rule-based aggregation) to evaluate consistency and sensitivity across methodologies.

Overall, while current findings are indicative, future access to complete, multi-year data would enable this framework to evolve into a robust, auditable funding governance tool.