ANALYSING UNIVERSITY STUDENT ACADEMIC PERFORMANCE AT THE UNIT LEVEL

Eric Parkin, Simon Huband, David Gibson and Dirk Ifenthaler
Curtin Learning and Teaching, Curtin University, Perth, Western Australia

ABSTRACT

This paper describes the creation of a dataset to enable the analysis of student academic performance at the unit level at a large Australian University. The dataset was designed to enable academic leaders to explore research questions concerning student performance and pass rates in units. Four example research questions are presented here and explored, to demonstrate the value of the exploratory dataset, along with an overview of the methodology and tools used to synthesise the data. The data handling methodology and reflections on tools and processes may be useful for other analysts.

KEYWORDS

Data Integration, Exploratory Analysis, Hypothesis Testing, Analytics Tools and Methods, Pass Rates

1. INTRODUCTION

Universities are increasingly interested in improving their capability to make data informed decisions around Learning and Teaching. Curtin University has employed a Learning and Teaching analytics team to help enable this capability. Much the team's work involves combining and transforming institutional data to create datasets for the University's academic leaders to explore for insights, and to use to answer hypotheses about learning and teaching. The datasets are also used by the analytics team to conduct more complicated analyses at the request of academic leaders. This paper describes the creation of a dataset with information on student academic performance in units (known elsewhere as subjects). A sample analysis of four research questions follows, to demonstrate how the dataset can be used to answer hypotheses about learning and teaching.

The need for this dataset was prompted by requests from academic leaders for data on student academic performance, with statistics at the unit level. Faculties used existing data to identify units with low pass rates, but the data did not provide the detail required to explore some of their specific questions. Existing data addressed high-level reporting and strategic decision-making needs, rather than learning and teaching needs. One of the main differences in this tool, compared to other learning analytics tools used by the university, is that it enables analysis at the unit level. Other learning analytics tools used by the university have focused primarily either on students, such as by using machine learning to predict student attrition (Kevin EK Chai, 2015), or on courses (known elsewhere as degrees), by using clustering to find categories of students that attrition in courses is low, (David Gibson, 2015).

The issue of student attrition is related closely linked to academic performance in units. One study at another Australian institution found that "first year academic performance is a much stronger predictor of attrition than any demographic factors or educational background" (Andrew Harvey, 2014), and determined that "attrition is an institutional problem, but the causes often lie at the lower levels of course and subject" (Andrew Harvey, 2014). According to a recent report (Higher Education Standards Panel, 2017) it is apparent that student attrition and the factors driving it have been of concern since the Commonwealth claimed a role in higher education funding. Substantial resources have been committed over many years to exploring ways to reduce it. The consistently reported drivers of attrition, which are documented in the report, are summarised in Table 1.

Table 1. Drivers of attrition

Driver	Examples
The learning environment	Dramatic changes in learning culture from school to higher education New modes of learning (off-site, online, part-time)
The teaching ability of lecturers	Adequacy of level of training in teaching
Lack of student engagement	Helpful and unhelpful patterns of student/student and student/teacher interaction
High student staff ratios	Availability of lecturers and tutors to students
Lack of student support information and services	
Personal factors	Financial, social, emotional, health or other life events.

The adjusted attrition rate (the rate at which students do not complete their course, and do not return to study it at any Australian University) in Australia has changed little over the period from 2005 to 2014. The attrition rate fell from 15.04 per cent in 2005, down to a low of 12.48 per cent in 2009, before rising over the remainder of the period to 15.18 per cent by 2014 (Higher Education Standards Panel, 2017).

The panel report and the other historical studies of attrition do not seem to focus on why students choose either to course switch or leave, even though the rate difference between normal attrition and adjusted attrition (course or institutional switchers) is significant. For example, the three-year running national average from 2012 to 2014 is 6%.

The success rate measures units of study passed by commencing students and is understandably highly correlated with the adjusted attrition rate and adjusted retention rate, as poor academic performance is a major factor in a student's decision to discontinue studies. The success rate in 2015 was 83.72 % and has dropped from its peak in 2004 of 86.85 % (Higher Education Standards Panel, 2017).

This report also found that student characteristics alone appear to only explain a relatively small part of the overall variation in student attrition, with an adjusted R^2 of 22.55 for a full model consisting of institution, full-time or part-time, attendance mode, age group, basis of admission, field of education, socio-economic status, indigenous, non-English speaking background, and gender (Higher Education Standards Panel, 2017). The field of education, which may be the closest indicator related to units of study, had an adjusted R^2 of only 1.49, well below the leading indicator, institution, which had 18.83. This indicates that exploring unit of study success rates may need to be part of a wider array of inquiries and interventions that will impact institutional culture to have an impact on schools, faculty areas, and the larger university.

The paper is structured as follows: **Section 2** introduces the research questions used as a sample analysis, **Section 3** describes the methodology of creating the dataset and conducting the analysis, **Section 4** presents the results of the analysis, **Section 5** discusses plans for future improvements and **Section 6** concludes with a reflection on the methodology and results.

2. RESEARCH QUESTIONS

Curtin University is a large, multi-campus Western Australian University, with over 35,000 from diverse backgrounds studying undergraduate and postgraduate , including over 10,000 international students as of 2017, see Curtin's 'Office of Strategy and Planning' website for more detail (Curtin University Office of Strategy and Planning, 2017).

To demonstrate the capabilities of the dataset, the following research questions were formulated, based on typical questions the analytics team are asked by the University's academic leaders:

> **RQ1a:** Is there a difference in pass rates between international students and domestic students who studied a first-year communications unit in 2017?
> **RQ1b:** If there is a difference in pass rates, is this difference consistent with other first-year units?
> **RQ2:** Were students who attempted, but did not pass a first-year communications unit in 2017, less likely to have completed their attempt, compared to students who studied other first-year units?

RQ3: Were international students who entered via a particular 'enabling pathway' less likely to pass a first-year communications unit in 2017 than international students who entered via other pathways?
RQ4a: Is there a difference in pass rates between students on their first attempt and students on their second or subsequent attempt at a first-year communications unit in 2017?
RQ4b: If there is a difference in pass rates, is this difference consistent with other first-year units?

The 'communications units' refer to six communications-skills focused units from the University's four main Faculties (**Business and Law**, **Health Sciences**, **Humanities**, and **Science and Engineering**). At least one of these six units is compulsory for most undergraduate students.

To eliminate potential confounding factors, only student enrolments which met the criteria in Table 2 were included. To use as a control group in the analyses, a set of **'control units'** were selected, which included all enrolments that met all criteria in Table 2, excluding enrolments in communications units. Further detail about the units is shown in the results section.

Table 2. Student enrolment inclusion criteria

Category	Criteria
Year	2017 enrolments only
Campus	Enrolments at the University' main campus only
Attendance Mode	Face-to-face enrolments only (online students are excluded)
Year Level	First-year units only
Unit Level	Undergraduate units only (enabling and postgraduate units are excluded)
Result Type	Exclude units graded only as pass or fail (not marked out of 100)
Coursework or Thesis	Coursework type units only. Thesis-type units are excluded.
For Degree Enrolment	Students studying 'not for degree', e.g. doing a single unit, are excluded

3. METHODOLOGY

This section summarises the methodology used to create the dataset. The primary purpose of the dataset was to enable academic leaders to investigate unit outcomes, enabling investigation of the research questions was a secondary goal. The aim for the workflow was to rapidly develop a rapid prototype that could be quickly and easily modified and updated with new data and was easy for staff to engage with and can be summarised in the following steps:

1. **Explore Data**: find data sources and gain understanding of the data
2. **Extract Data**: extract data from the source systems
3. **Combine Data**: merge data to into a single table and derive new columns
4. **Finalise Data**: add calculated columns and aggregate data
5. **Share Data**: prepare data for dissemination and share it with academic leaders
6. **Analyse Data**: use the data to answer the research questions

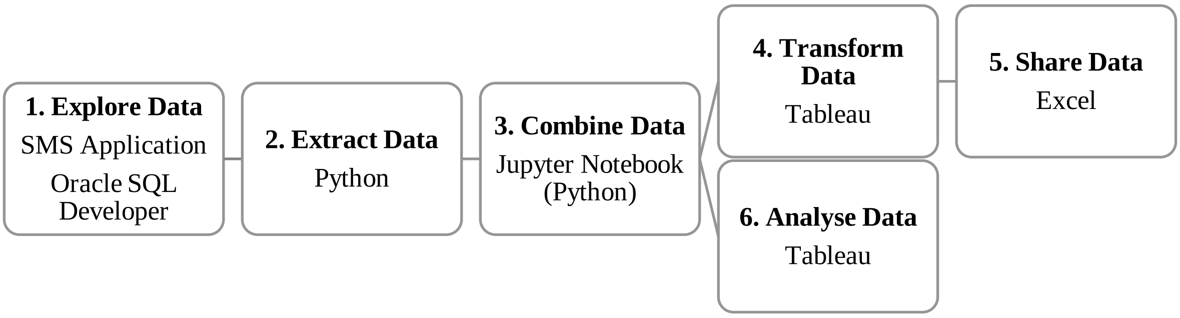

Figure 1. Data workflow

1. Explore Data

Gain an understanding data available, its structure, accuracy, and reliability, by exploring the Student Management System (SMS) application, and consulting with domain experts. Find the required tables and columns by referring to the database documentation and exploring the database using a database IDE (Oracle SQLDeveloper) in this case. The Result of this step will be a list of tables and columns, their sources, and notes/diagrams about the data, and how the tables relate to each other, i.e. primary and foreign keys.

2. Extract Data

Extract the data with Python script using the 'Pandas' 'read_sql' module to connect to the database (with login credentials), extract the tables, and save them as CSV files on a local (encrypted) storage device.

3. Combine Data

Combine the data into a single table with one row per enrolment. A Jupyter Notebook, using the 'Pandas' and 'NumPy' libraries, was the tool of choice for this step. Within in the notebook, import the CSV files, then sequentially transform and merge the tables into a single data frame. Use 'groupby' commands to derive new columns which require constructing from multiple rows (e.g. derive number of attempts by counting the number of rows with an enrolment in the same unit for the same student). Validate the data by running test commands to check for issues such as duplicates and inconsistencies between rows. Perform a 'sanity test' of the data by checking whether the data is sensible. Complete a final manual 'spot check' of the data by validating a selection of rows with the SMS Application. Correct any issues, re-execute the code, and repeat. Lastly, rename columns save as a CSV file on a local (encrypted) storage device.

Steps 1 to 3 will differ significantly for different institutions, but the combined table produced by step should be similar to the structure described in Table 3. Producing a dataset with these fields can be considered the goals of steps 1 to 3.

Table 3. Combined data structure

Group	Columns
Enrolment: (unique combination to define an enrolment in a unit)	Student ID, Unit Code (unique unit identifier), Year, Study Period (Semester/Term)
Course Details (determined by the course (degree) the student is studying	Course (degree) code and title, Undergraduate or Postgraduate course, 'Not for Degree' course flag, Basis for Admission (determines entry pathway)
Unit Details (determined by the unit code)	Unit title, Unit owning Faculty and School, Year Level, Postgraduate or Undergraduate unit, Campus, Internal or Fully Online
Enrolment Details (details about a student's enrolment in a unit)	Attempt Number, Enrolment Status (Pass/Fail/Withdraw), Grade and Mark
Student Details (unique per student)	International or Domestic Student flag

4. Transform Data

Transform the combined data into the final dataset, by connecting to the combined CSV file from a Tableau workbook and the set data types for each column. Create calculated columns to flag rows as 1 or 0, based on existing columns (e.g. 'domestic student pass', or 'zero mark'). Create a table and add rows for unit, campus, faculty, year, and semester. Set the granularity of the dataset by adding or removing rows (e.g. remove semester for a more aggregated dataset or add 'attendance mode' to create separate rows for online and face-to-face enrolments). Add filters to remove student enrolments which should not be included, e.g. 'not for degree' students. Validate the calculated columns by creating tables which flag inconsistencies, e.g. count rows which are flagged as both 'passed' and 'zero mark' (should be none).

This step could be done within other spreadsheet tools, or within the Jupyter Notebook (although this may be more time-consuming). One of the transformed datasets, focusing on students with multiple attemtps is shown in Table 4. Different columns can be used to focus on different information (e.g. international or domestic student statistics), but the 'Unit and course information' and 'Basic Enrolment Statistics' can remain.

Table 4. Transformed data structure, with data for multiple attempts

Group	Columns
Unit and Course Details	Unit Code, Unit Title, Faculty and School, Campus, Postgraduate or Undergraduate Unit, Year and Study Period
Basic Enrolment Statistics	Total Enrolments, Passes, Pass %, Fails, Fail %, Withdrawals, Withdraw %
First Attempt Statistics	No. on 1st Attempt, 1st attempt pass %, 1st attempt fail %, 1st attempt withdrawal %, 1st attempt avg. mark
Second Attempt Statistics	No. on 2nd Attempt, 2nd attempt pass %, 2nd attempt fail %, 2nd attempt withdrawal %, 1st attempt avg. mark
Third+ Attempt Statistics	No. on 3rd+ Attempt, 3rd+ attempt pass %, 3rd+ attempt fail %, 3rd+ attempt withdrawal %, 3rd+ attempt avg. mark

5. Share Data

Create a shareable from of the transformed datasets by copying data from Tableau text tables into separate tabs of a 'master' Excel workbook. Add a description tab with instructions, explanations of the data, and known issues. Create a separate workbook for each area by deleting data for other areas. Email the Excel workbooks to authorised personnel, for distribution more widely in their areas.

6. Analyse Data

The combined dataset produced in step 3 was used to conduct the analysis for this paper (with some of the transformed columns from step 4), the data was filtered to the scope in Table 2, and structured to create the aggregated tables shown in the results section, instead of to produce the more aggregated, unit level data.

4. RESULTS

This section presents the results of analysing the research questions presented in section 2. The findings presented here cannot be generalised to apply in all contexts, and can vary significantly when looking at individual units, and when looking at different types of units. These results are a sample only and are not intended to provide definitive answers to questions that are known to be inherently complex.

For context in the analyses to follow, overall enrolment and completion statistics for the communications units and control group units are shown in Table 5.

Table 5. Overall completion statistics for the communications units, and control group units

	Control Units	Communications Units	Difference
Number of Units	248	6	**-**
Enrolments	52565 **(100.0%)**	7726 **(100.0%)**	**-**
Passes	41433 **(78.8%)**	6467 **(83.7%)**	**+4.9%**
Fails	8280 **(15.8%)**	863 **(11.2%)**	**-4.6%**
Withdrawals	2842 **(5.4%)**	396 **(5.1%)**	**-0.3%**

RQ1a: Is there a difference in pass rates between international students and domestic students who studied a first-year communications unit in 2017?
RQ1b: If there is a difference in pass rates, is this difference consistent with other first-year units?

The completion statistics for international students and domestic students in the communications units and control units are shown in Table 6.

RQ1a Answer: The results show that the pass rates in the communications units is lower for international students compared to domestic students **(-3.8%)**. This lower pass rate for international students is accounted for, mainly by the higher fail rate **(+7.5pp)**, which is partly offset by the lower withdrawal rate.

Table 6. Completion statistics for international students

	Measure	Domestic Students	International Students	Difference
Communication Units	**Enrolments**	6508 **(100.0%)**	1218 **(100.0%)**	**-**
	Passes	5487 **(84.3%)**	980 **(80.5%)**	**-3.8pp**
	Fails	650 **(10.0%)**	213 **(17.5%)**	**+7.5pp**
	Withdrawals	371 **(5.7%)**	25 **(2.1%)**	**-3.6pp**
Control Units	**Enrolments**	47706 **(100.0%)**	4859 **(100.0%)**	**-**
	Passes	37725 **(79.1%)**	3718 **(76.5%)**	**-2.6pp**
	Fails	7221 **(15.1%)**	1086 **(21.8%)**	**+6.7pp**
	Withdrawals	2760 **(5.8%)**	82 **(1.7%)**	**-4.1pp**

RQ1b Answer: The results show that the pass rates are lower for international students more in the control units, however, the difference is more pronounced in the communications units **(-3.8pp** compared to **-2.6pp)**.

RQ2: Were students who attempted, but did not pass a first-year communications unit in 2017, less likely to have completed their attempt, compared to students who studied other first-year units?

'Non-completions' can be classified into four categories:

1. **'Fail (complete & low mark)'** – fail based on the student's mark (out of 100) being too low.
2. **'Fail (incomplete & non-zero mark)'** – student achieved a non-zero mark and failed on the basis of not meeting all pass criteria (e.g. student did not submit all compulsory assessments).
3. **'Fail (incomplete & zero mark)'** – student achieved a zero mark and failed on the basis of not meeting all pass criteria (these students likely have not participated in the unit, despite being enrolled).
4. **'Withdrawal'** – students decided to drop out of the unit after the cut-off date for fees (census date).

A breakdown of the non-completion statistics, as a proportion of all non-passing students, for the control units and communications units is shown in Table 7.

Table 7. Non-completion statistics

Measure	Control Units	Communications Units	Difference
All non-passes (withdrawals + fails)	11155 **(100.0%)**	1259 **(100.0%)**	**-**
Fails (complete & low mark)	5006 **(44.9%)**	344 **(27.3%)**	**-17.6pp**
Fails (incomplete & non-zero mark)	2156 **(22.6%)**	420 **(33.4%)**	**+10.8pp**
Fails (incomplete & zero mark)	786 **(7.0%)**	99 **(7.9%)**	**+0.9pp**
Withdrawals	2847 **(25.5%)**	396 **(31.4%)**	**+5.9pp**

RQ2 Answer: The results show that students who attempted but did not pass a communications unit were significantly less likely to have completed their attempt **(-17.6pp)**, than students who studied other units. The difference is accounted for by the higher rates of 'fails with incomplete attempts & non-zero mark' **(+10.8pp)**, 'fails with zero mark' **(+0.9pp)**, and withdrawals **(+5.9pp)**.

RQ3: Were international students who entered via a particular 'enabling pathway' less likely to pass a first-year communications unit in 2017 than international students who entered via other pathways?

The **'Enabling pathway'** looked at in this analysis, offers entry programs for students who do not meet the University's undergraduate entry requirements and is predominantly used by international students. The enrolment and completion statistics for these students are shown in Table 8.

Table 8. International student communications unit completion statistics by entry pathway

Measure	Other Entry Pathways	Enabling Pathway	Difference
Enrolments	1027 **(100.0%)**	191 **(100.0%)**	**-**
Passes	827 **(80.5%)**	153 **(80.1%)**	**-0.4pp**
Fails	181 **(17.6%)**	32 **(16.8%)**	**-0.8pp**
Withdraw	19 **(1.9%)**	6 **(3.1%)**	**+1.2pp**

RQ3 Answer: Of students who studied the communication units, the results show there is little difference in completion rates between international students who entered the university via the enabling pathway, and international students who entered via other pathways. Specifically, there is only a **-0.4pp** difference in pass rates, a **-0.8pp** difference in failure rate, and a **+1.2pp** difference in withdrawal rates.

RQ4a: Is there a difference in pass rates between students on their first attempt and students on their second or subsequent attempt at a first-year communications unit in 2017?
RQ4b: If there is a difference in pass rates, is this difference consistent with other first-year units?

The completion statistics for students on their first attempt, and students on their second or subsequent attempt, for both the communications units and control units are shown in Table 9.

RQ4a Answer: Students on their second or subsequent attempt at a communications unit, had a much lower pass rate than students on their first attempt **(-31.5pp)**. This is accounted for by the higher fail rate **(+22.3pp)** and higher withdrawal rate **(+9.2pp)**.

Table 9. Completion rates by first attempt and by second or subsequent attempts

	Measure	1st Attempt	2nd or Subsequent Attempt	Difference
Communications Units	**Enrolments**	7291 **(100.0%)**	435 **(100.0%)**	**-**
	Passes	6232 **(85.5%)**	235 **(54.0%)**	**-31.5pp**
	Fails	723 **(9.9%)**	140 **(32.2%)**	**+22.3pp**
	Withdrawals	336 **(4.6%)**	60 **(13.8%)**	**+9.2pp**
Control Units	**Enrolments**	48959 **(100.0%)**	3606 **(100.0%)**	**-**
	Passes	39428 **(80.5%)**	2015 **(55.9%)**	**-24.6pp**
	Fails	7114 **(14.5%)**	1166 **(32.3%)**	**+17.8pp**
	Withdrawals	2417 **(4.9%)**	425 **(11.8%)**	**+6.0pp**

RQ4b Answer: The lower pass for students on their second or subsequent attempt at a communications unit, is consistent with the control units, however, the difference is more pronounced in the communications units **(-31.5pp** compared to **-24.6pp)**.

5. FUTURE WORK

The Analytics team has plans to improve the pass rates dataset, including fixing issues, and making enhancements. These include: adding a row for student who have been left in an 'Enrolled' state, correcting for cases of duplicate enrolments present in some older source data, and combining unit codes with their replacements – allowing units that have been replaced to be analysed across multiple years. Additional columns are planned based on requests from academic leaders, including adding Tertiary Entrance Rank, and columns to flag the number of units failed in the same semester, and whether the student was ever granted a deferred assessment in a unit. Longer term goals for the analytics team are to make the dataset accessible via dashboards, and to combine the pass rates data set with the team's student retention dataset.

The analysis of the research questions could be improved by exploring the statistical significance of the results, and where there are highly significant results, investigating the data more deeply to find causal factors (e.g. attempt to find why the communications units have a greater prevalence of non-completion).

6. DISCUSSION/CONCLUSIONS

The methodology used to create the pass rates dataset, was effective in its goal of enabling the rapid creation of a prototype dataset. It required minimal computing infrastructure, and was capable of being used to investigate hypotheses about student academic performance in units. Due to the success of the project, the methodology and tools described, will likely be used by the analytics team in future projects, and is suggested for consideration by other institutions undertaking similar projects.

The issues and complexities encountered during the project, highlighted not only the importance of rigorous testing, but also of the importance of engaging with staff and of having a deep understanding of the data and how it is used within the organisation. Without such engagement and understanding, it is likely the datasets produced would have been flawed, and from previous experience, could have caused users to 'distrust' the accuracy of the data, and hence, refrain from using it. Being open about known issues in the data, and encouraging staff to report new issues, helped to improve trust.

The choice of sharing data with pre-calculated columns, in a 'flat' table structure, rather than as Pivot tables, helped lower the barrier for staff to engage with the data, and broadened its audience. The analytics team has received positive feedback on the usefulness of the data, had several requests for updated data and for improvements/additions to the dataset, evidencing the usefulness and accessibility of the data.

One major limitation of the dataset is that in its current form the data is 'static'. When academic leaders wish to view the data at a different level of aggregation, or with different filters, a new dataset would need to be created. Making the data available via dashboards could help solve this problem.

At present, no consideration has been given to release the data to external researchers for the purpose of open science. This will be considered in the future, however, ethical considerations complicate this task. Considerable care and effort would be required to cluster and anonymise the data, and was not a priority for this project.

ACKNOWLEDGEMENT

This paper was partly funded by the UNESCO Chair of Data Science in Higher Education Learning and Teaching position (Curtin University, 2018).

REFERENCES

Andrew Harvey, M. L., 2014. Beyond demographics: Predicting student attrition with the Bachelor of Arts degree. *The international Journal of the First Year in Hiher Education,* 5(1), pp. 19-29.

Curtin University Office of Strategy and Planning, 2017. *Curtin University Student Statistics 2013-2017.* [Online] Available at: https://planning.curtin.edu.au/stats/students2013-2017.cfm [Accessed 8 October 2018].

Curtin University, 2018. *UNESCO Chair of Data Science in Higher Education Learning and Teaching.* [Online] Available at: https://research.curtin.edu.au/projects-expertise/institutes-centres/unesco/ [Accessed 8 October 2018].

David Gibson, S. d. F., 2015. Exploratory Analytics in Learning Analytics. *Technology, Knowledge and Learning,* 21(1), pp. 5-19.

Higher Education Standards Panel, 2017. *Improving retention, completion and success in higher education.* [Online] Available at: https://docs.education.gov.au/system/files/doc/other/final_discussion_paper.pdf [Accessed 8 October 2018].

Kevin EK Chai, F. G., 2015. *Predicting the risk of attrition for undergraduate students with time based modelling.* s.l., 12th International Conference on Cognition and Exploratory Learning in Digital Age.

Note: This research received ethics approval by Curtin University's research integrity office, code: HRE2018-0519.